Jo McEntyre • Jim Ostell

The NCBI Handbook

Last Updated: April 6, 2012

National Center for Biotechnology Information (US)
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

NLM Citation: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-.

Bioinformatics consists of a computational approach to biomedical information management and analysis. It is being used increasingly as a component of research within both academic and industrial settings and is becoming integrated into both undergraduate and postgraduate curricula. The new generation of biology graduates is emerging with experience in using bioinformatics resources and, in some cases, programming skills.

The National Center for Biotechnology Information (NCBI) is one of the world's premier Web sites for biomedical and bioinformatics research. Based within the National Library of Medicine at the National Institutes of Health, USA, the NCBI hosts many databases used by biomedical and research professionals. The services include PubMed, the bibliographic database; GenBank, the nucleotide sequence database; and the BLAST algorithm for sequence comparison, among many others.

Although each NCBI resource has online help documentation associated with it, there is no cohesive approach to describing the databases and search engines, nor any significant information on how the databases work or how they can be leveraged, for bioinformatics research on a larger scale. The NCBI Handbook is designed to address this information gap.

All of our users know how to execute a straightforward PubMed or BLAST search. However, feedback from help desk personnel and booth staff at scientific meetings suggests that people often want to know how to use our resources in a more sophisticated manner and are frequently unaware of less well-known databases that might be helpful to them. The intended audience for The NCBI Handbook is, therefore, the growing number of scientists and students who would like a more in-depth guide to NCBI resources—powerusers and aspiring powerusers.

The NCBI Handbook is focused on the relatively stable information about each resource; it is not a point-and-click user guide (this type of information can be found in the online help documents, referred to frequently but not repeated, in the Handbook). Each chapter is devoted to one service; after a brief overview on using the resource, there is an account of how the resource works, including topics such as how data are included in a database, database design, query processing, and how the different resources relate to each other. For example, the BLAST chapter briefly describes what to use BLAST for, the various varieties of the BLAST algorithm, and BLAST statistics, before discussing output formats, query processing, and tips for setting up a BLAST database. A certain amount of biological knowledge is assumed.

The online content will be updated when necessary, although major changes are not expected to occur more than once every few years. (For example, PubMed query processing does not change dramatically year after year.) We hope that The NCBI Handbook will provide a valuable reference for anyone who wants to use our resources more effectively.

Editors

Jo McEntyre

Jim Ostell

Table of Contents

Part 1 The Databases 1

Chapter 1 GenBank: The Nucleotide Sequence Database 3

History 3

International Collaboration 4

Confidentiality of Data 4

Direct Submissions 4

Bulk Submissions: High-Throughput Genomic Sequence (HTGS) 5

Whole Genome Shotgun Sequences (WGS) 7

Bulk Submissions: EST, STS, and GSS 8

Bulk Submissions: HTC and FLIC 9

Submission Tools 10

Sequence Data Flow and Processing: From Laboratory to GenBank 11

Microbial Genomes 13

Third Party Annotation (TPA) Sequence Database 15

Appendix: GenBank, RefSeq, TPA and UniProt: What's in a Name? 15

References 19

Chapter 2 PubMed: The Bibliographic Database 21

Data Sources 21

Electronic Data Submission 22

Database Management and Hardware 23

Indexing 23

How PubMed Queries Are Processed 24

Using PubMed 28

Additional PubMed Features 29

Results 30

Links from PubMed 30

How to Create Hyperlinks to PubMed 31

Customer Support 31

Chapter 3 Macromolecular Structure Databases 33
Overview 33
Content of the Molecular Modeling Database (MMDB) 36
Content of the Conserved Domain Database (CDD) 40
Finding and Viewing Structures 45
Finding and Viewing Structure Neighbors 50
Finding and Viewing Conserved Domains 52
Finding and Viewing Proteins with Similar Domain Architectures 53
Links Between Structure and Other Resources 54
Saving Output from Database Searches 56
FTP 56
Frequently Asked Questions 57
References 57

Chapter 4 The Taxonomy Project 59
Introduction 59
Adding to the Taxonomy Database 62
Using the Taxonomy Browser 63
The Taxonomy Database: TAXON 68
Nomenclature Issues 70
Taxonomy in Entrez: A Quick Tour 72
The Common Tree Viewer 74
Indexing Taxonomy in Entrez 74
The Taxonomy Statistics Page 80
Other Relevant References 80
NCBI Taxonomists 81
Contact Us 81
Appendix 1. TAXON nametypes 81
Appendix 2. Functional classes of TAXON scientific names 85
Appendix 3. Other TAXON data types 90

Chapter 5 The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation 93

Introduction 93

Searching dbSNP 95

Submitted Content 100

Computed Content (The dbSNP Build Cycle) 105

dbSNP Resource Integration 114

How to Create a Local Copy of dbSNP 115

Appendix 1. dbSNP report formats. 120

Appendix 2. Rules and methodology for mapping 122

Appendix 3 Alignment profiling function 123

Appendix 4. 3D structure neighbor analysis. 126

Chapter 6 The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository 127

Site Description 127

Design and Implementation 128

Retrieving Data 130

Depositing Data 132

Search and Integration 133

Example of Retrieving Data 135

Future Directions 137

Frequently Asked Questions 137

Acknowledgments 140

Contributors 140

References 141

Chapter 7 Online Mendelian Inheritance in Man (OMIM): A Directory of Human Genes and Genetic Disorders 143

Content and Access 143

Guide to OMIM Pages 147

FTP 148

Legal Statement 149

References 149

Chapter 8 The NCBI BookShelf: Searchable Biomedical Books 151

Content Acquisition 151

How to Use the Books 152

Technology 156

The BookShelf Data Flow 159

NCBI Book DTD 159

Frequently Asked Questions 161

Chapter 9 PubMed Central (PMC): An Archive for Literature from Life Sciences Journals 163

A PubMed Central (PMC) Site Guide 163

Participation in PMC 167

Links to Other NCBI Resources 168

PMC Architecture 169

Data Flow: 1. SGML/XML Processing 169

Data Flow: 2. Loading the Database 174

Special Characters 175

PMC DTD 176

Frequently Asked Questions 178

Chapter 10 The SKY/CGH Database for Spectral Karyotyping and Comparative Genomic Hybridization Data 179

Database Content 179

Data Analysis: Query Tools 187

Data Integration 187

Contributors 188

References 188

Chapter 11 The Major Histocompatibility Complex Database, dbMHC 191

Introduction 191

dbMHC Resources 192

Database Content 203

Integration with Other Resources 213
References 214

Part 2 Data Flow and Processing 217

Chapter 12 Sequin: A Sequence Submission and Editing Tool 219
Sequin: A Brief Overview 219
Sequence Submission 220
Packaging the Submissions 222
Viewing and Editing the Sequences 224
Computational Functions of Sequin 226
Advanced Topics 230
Conclusion 232

Chapter 13 The Processing of Biological Sequence Data at NCBI 233
Overview 233
Data Flow Components 236
Data Flow Architecture 241

Chapter 14 Genome Assembly and Annotation Process 245
Overview of the Genome Assembly and Annotation Process 246
The Input Data 248
Preparation of the Input Sequences 251
Alignment of Sequences to the Input Genomic Sequences 253
Genome Assembly 254
Annotation of Genes 257
Annotation of Other Features 262
Product Data Sets 263
Production of Maps That Display Genome Features 264
Public Release of Assembly and Models 265
Integration with Other Resources 265
Contributors 266
References 267

Part 3 Querying and Linking the Data 271

Chapter 15 The Entrez Search and Retrieval System 273

Entrez Design Principles 273

Entrez Is a Discovery System 276

Entrez Is Growing 277

How Entrez Works 277

References 279

Chapter 16 The BLAST Sequence Analysis Tool 281

Introduction 281

How BLAST Works: The Basics 282

BLAST Scores and Statistics 283

BLAST Output: 1. The Traditional Report 284

BLAST Output: 2. The Hit Table 285

BLAST Output: 3. Structured Output 287

BLAST Code 290

Appendix 1. FASTA identifiers 291

Appendix 2. Readdb API 292

Appendix 3. Excerpt from a demonstration program doblast.c 293

Appendix 4. A function to print a view of a SeqAlign: MySeqAlignPrint 294

References 295

Chapter 17 LinkOut: Linking to External Resources from Entrez Databases 297

How Is LinkOut Represented in Entrez? 297

How Does LinkOut Work? 297

Guides for LinkOut Providers 303

Communicating with LinkOut Providers 305

Chapter 18 The Reference Sequence (RefSeq) Database 307

Summary 307

Introduction 307

Database Content: Background 308
Assembling and Maintaining the RefSeq Collection 314
Access and Retrieval 324
Related Resources 327
Related Reading 327

Chapter 19 Gene: A Directory of Genes 329
Summary 329
Overview 329
Maintaining the Data 330
How to Query Gene 332
Display Formats 337
Content 338
References 341

Chapter 20 Using the Map Viewer to Explore Genomes 343
Introduction 343
Maintenance of Data 344
Methods of Access 350
Interpreting the Display 354
Customizing the Display 358
Associated Tools 361
Technical Details 362
Caveats for Using Evolving Data 364
References 365

Chapter 21 UniGene: A Unified View of the Transcriptome 367
Expressed Sequence Tags (ESTs) 367
Sequence Clusters 369
UniGene Cluster Browser 372
Protein Similarity Analysis 373
Digital Differential Display (DDD) 374

HomoloGene 375

References 377

Chapter 22 The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes 379

Introduction 379

Construction of the COGs 380

Phyletic Pattern Analysis in COGs 382

Description of the COGs Website 383

Future Directions 383

The COG Team 384

References 384

Part 4 User Support 387

Chapter 23 User Services: Helping You Find Your Way 389

The User Services Team 389

Development of User Support Materials 390

Outreach 391

Conclusion 395

References 395

Chapter 24 Exercises: Using Map Viewer 397

1. How Do I Obtain the Genomic Sequence around My Gene of Interest? 397

2. If I Have Physical and/or Genetic Mapping Data, How Do I Use the Map Viewer to Find a Candidate Disease Gene in That Region? 399

3. How Can I Find and Display a Gene with the Map Viewer? 402

4. How Can I Analyze a Gene Using the Map Viewer? 404

5. How Can I Create My Own Transcript Models with the Map Viewer? 407

6. Using the Mouse Map Viewer 409

7. How Can I Find Members of a Gene Family Using the Map Viewer? 412

8. How Can I Find Genes Encoding a Protein Domain Using the Map Viewer? 414

Glossary 417

Part 1. The Databases

Chapter 1. GenBank: The Nucleotide Sequence Database

Ilene Mizrachi

Created: October 9, 2002; Updated: August 22, 2007.

Summary

The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of an international collaboration with the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate, doubling every 10 months. Release 134, produced in February 2003, contained over 29.3 billion nucleotide bases in more than 23.0 million sequences. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

Direct submissions are made to GenBank using BankIt, which is a Web-based form, or the stand-alone submission program, Sequin. Upon receipt of a sequence submission, the GenBank staff assigns an Accession number to the sequence and performs quality assurance checks. The submissions are then released to the public database, where the entries are retrievable by Entrez or downloadable by FTP. Bulk submissions of Expressed Sequence Tag (EST), Sequence Tagged Site (STS), Genome Survey Sequence (GSS), and High-Throughput Genome Sequence (HTGS) data are most often submitted by large-scale sequencing centers. The GenBank direct submissions group also processes complete microbial genome sequences.

History

Initially, GenBank was built and maintained at Los Alamos National Laboratory (LANL). In the early 1990s, this responsibility was awarded to NCBI through congressional mandate. NCBI undertook the task of scanning the literature for sequences and manually typing the sequences into the database. Staff then added annotation to these records, based upon information in the published article. Scanning sequences from the literature and placing them into GenBank is now a rare occurrence. Nearly all of the sequences are now deposited directly by the labs that generate the sequences. This is attributable to, in part, a requirement by most journal publishers that nucleotide sequences are first deposited into publicly available databases (DDBJ/EMBL/GenBank) so that the Accession number can be cited and the sequence can be retrieved when the article is published. NCBI began accepting direct submissions to GenBank in 1993 and received data from LANL until 1996. Currently, NCBI receives and processes about 20,000 direct submission

sequences per month, in addition to the approximately 200,000 bulk submissions that are processed automatically.

International Collaboration

In the mid-1990s, the GenBank database became part of the International Nucleotide Sequence Database Collaboration with the EMBL database (European Bioinformatics Institute, Hinxton, United Kingdom) and the Genome Sequence Database (GSDB; LANL, Los Alamos, NM). Subsequently, the GSDB was removed from the Collaboration (by the National Center for Genome Resources, Santa Fe, NM), and DDBJ (Mishima, Japan) joined the group. Each database has its own set of submission and retrieval tools, but the three databases exchange data daily so that all three databases should contain the same set of sequences. Members of the DDBJ, EMBL, and GenBank staff meet annually to discuss technical issues, and an international advisory board meets with the database staff to provide additional guidance. An entry can only be updated by the database that initially prepared it to avoid conflicting data at the three sites.

The Collaboration created a Feature Table Definition that outlines legal features and syntax for the DDBJ, EMBL, and GenBank feature tables. The purpose of this document is to standardize annotation across the databases. The presentation and format of the data are different in the three databases, however, the underlying biological information is the same.

Confidentiality of Data

When scientists submit data to GenBank, they have the opportunity to keep their data confidential for a specified period of time. This helps to allay concerns that the availability of their data in GenBank before publication may compromise their work. When the article containing the citation of the sequence or its Accession number is published, the sequence record is released. The database staff request that submitters notify GenBank of the date of publication so that the sequence can be released without delay. The request to release should be sent to gb-admin@ncbi.nlm.nih.gov.

Direct Submissions

The typical GenBank submission consists of a single, contiguous stretch of DNA or RNA sequence with annotations. The annotations are meant to provide an adequate representation of the biological information in the record. The GenBank Feature Table Definition describes the various features and subsequent qualifiers agreed upon by the International Nucleotide Sequence Database Collaboration.

Currently, only nucleotide sequences are accepted for direct submission to GenBank. These include mRNA sequences with coding regions, fragments of genomic DNA with a single gene or multiple genes, and ribosomal RNA gene clusters. If part of the nucleotide sequence encodes a protein, a conceptual translation, called a CDS (coding sequence), is

annotated. The span of the CDS feature is mapped to the nucleotide sequence encoding the protein. A protein Accession number (/protein_id) is assigned to the translation product, which will subsequently be added to the protein databases.

Multiple sequences can be submitted together. Such batch submissions of non-related sequences may be processed together but will be displayed in Entrez (Chapter 15) as single records. Alternatively, by using the Sequin submission tool (Chapter 12), a submitter can specify that several sequences are biologically related. Such sequences are classified as environmental sample sets, population sets, phylogenetic sets, mutation sets, or segmented sets. Each sequence within a set is assigned its own Accession number and can be viewed independently in Entrez. However, with the exception of segmented sets, each set is also indexed within the PopSet division of Entrez, thus allowing scientists to view the relationship between the sequences.

What defines a set? Environmental sample, population, phylogenetic, and mutation sets all contain a group of sequences that spans the same gene or region of the genome. Environmental samples are derived from a group of unclassified or unknown organisms. A population set contains sequences from different isolates of the same organism. A phylogenetic set contains sequences from different organisms that are used to determine the phylogenetic relationship between them. Sequencing multiple mutations within a single gene gives rise to a mutation set.

All sets, except segmented sets, may contain an alignment of the sequences within them and might include external sequences already present in the database. In fact, the submitter can begin with an existing alignment to create a submission to the database using the Sequin submission tool. Currently, Sequin accepts FASTA+GAP, PHYLIP, MACAW, NEXUS Interleaved, and NEXUS Contiguous alignments. Submitted alignments will be displayed in the PopSet section of Entrez.

Segmented sets are a collection of noncontiguous sequences that cover a specified genetic region. The most common example is a set of genomic sequences containing exons from a single gene where part or all of the intervening regions have not been sequenced. Each member record within the set contains the appropriate annotation, exon features in this case. However, the mRNA and CDS will be annotated as joined features across the individual records. Segmented sets themselves can be part of an environmental sample, population, phylogenetic, or mutation set.

Bulk Submissions: High-Throughput Genomic Sequence (HTGS)

HTGS entries are submitted in bulk by genome centers, processed by an automated system, and then released to GenBank. Currently, about 30 genome centers are submitting data for a number of organisms, including human, mouse, rat, rice, and *Plasmodium falciparum*, the malaria parasite.

HTGS data are submitted in four phases of completion: 0, 1, 2, and 3. Phase 0 sequences are one-to-few reads of a single clone and are not usually assembled into contigs. They are

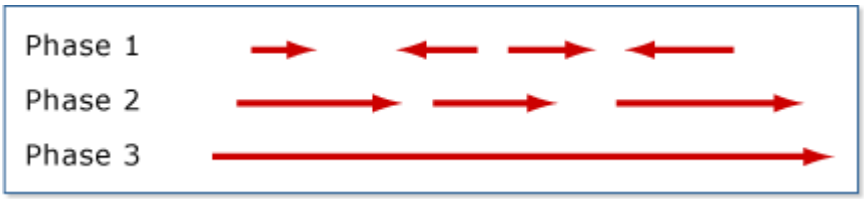


Figure 1. Diagram showing the orientation and gaps that might be expected in high-throughput sequence from phases 1, 2, and 3.

low-quality sequences that are often used to check whether another center is already sequencing a particular clone. Phase 1 entries are assembled into contigs that are separated by sequence gaps, the relative order and orientation of which are not known (Figure 1). Phase 2 entries are also unfinished sequences that may or may not contain sequence gaps. If there are gaps, then the contigs are in the correct order and orientation. Phase 3 sequences are of finished quality and have no gaps. For each organism, the group overseeing the sequencing effort determines the definition of finished quality.

Phase 0, 1, and 2 records are in the HTG division of GenBank, whereas phase 3 entries go into the taxonomic division of the organism, for example, PRI (primate) for human. An entry keeps its Accession number as it progresses from one phase to another but receives a new Accession.Version number and a new gi number each time there is a sequence change.

Submitting Data to the HTG Division

To submit sequences in bulk to the HTG processing system, a center or group must set up an FTP account by writing to htgs-admin@ncbi.nlm.nih.gov. Submitters frequently use two tools to create HTG submissions, Sequin or fa2htgs. Both of these tools require FASTA-formatted sequence, i.e., a definition line beginning with a "greater than" sign (">") followed by a unique identifier for the sequence. The raw sequence appears on the lines after the definition line. For sequences composed of contigs separated by gaps, a modified FASTA format is used. In addition, Sequin users must modify the Sequin configuration file so that the HTG genome center features are enabled.

fa2htgs is a command-line program that is downloaded to the user's computer. The submitter invokes a script with a series of parameters (arguments) to create a submission. It has an advantage over Sequin in that it can be set up by the user to create submissions in bulk from multiple files.

Submissions to HTG must contain three identifiers that are used to track each HTG record: the genome center tag, the sequence name, and the Accession number. The genome center tag is assigned by NCBI and is generally the FTP account login name. The sequence name is a unique identifier that is assigned by the submitter to a particular clone or entry and must be unique within the group's submissions. When a sequence is first submitted, it has only a sequence name and genome center tag; the Accession number is

assigned during processing. All updates to that entry must include the center tag, sequence name, and Accession number, or processing will fail.

The HTG Processing Pathway

Submitters deposit HTGS sequences in the form of Seq-submit files generated by Sequin, fa2htgs, or their own ASN.1 dumper tool into the SEQSUBMIT directory of their FTP account. Every morning, scripts automatically pick up the files from the FTP site and copy them to the processing pathway, as well as to an archive. Once processing is complete and if there are no errors in the submission, the files are automatically loaded into GenBank. The processing time is related to the number of submissions that day; therefore, processing can take from one to many hours.

Entries can fail HTG processing because of three types of problems:

1. Formatting: submissions are not in the proper Seq-submit format.
2. Identification: submissions may be missing the genome center tag, sequence name, or Accession number, or this information is incorrect.
3. Data: submissions have problems with the data and therefore fail the validator checks.

When submissions fail HTG processing, a GenBank annotator sends email to the sequencing center, describing the problem and asking the center to submit a corrected entry. Annotators do not fix incorrect submissions; this ensures that the staff of the submitting genome center fixes the problems in their database as well.

The processing pathway also generates reports. For successful submissions, two files are generated: one contains the submission in GenBank flat file format (without the sequence); and another is a status report file. The status report file, ac4htgs, contains the genome center, sequence name, Accession number, phase, create date, and update date for the submission. Submissions that fail processing receive an error file with a short description of the error(s) that prevented processing. The GenBank annotator also sends email to the submitter, explaining the errors in further detail.

Additional Quality Assurance

When successful submissions are loaded into GenBank, they undergo additional validation checks. If GenBank annotators find errors, they write to the submitters, asking them to fix these errors and submit an update.

Whole Genome Shotgun Sequences (WGS)

Genome centers are taking multiple approaches to sequencing complete genomes from a number of organisms. In addition to the traditional clone-based sequencing whose data are being submitted to HTGS, these centers are also using a WGS approach to sequence the genome. The shotgun sequencing reads are assembled into contigs, which are now being accepted for inclusion in GenBank. WGS contig assemblies may be updated as the

sequencing project progresses and new assemblies are computed. WGS sequence records may also contain annotation, similar to other GenBank records.

Each sequencing project is assigned a stable project ID, which is made up of four letters. The Accession number for a WGS sequence contains the project ID, a two-digit version number, and six digits for the contig ID. For instance, a project would be assigned an Accession number AAAX00000000. The first assembly version would be AAAX01000000. The last six digits of this ID identify individual contigs. A master record for each assembly is created. This master record contains information that is common among all records of the sequencing project, such as the biological source, submitter, and publication information. There is also a link to the range of Accession numbers for the individual contigs in this assembly.

WGS submissions can be created using tbl12asn, a utility that is packaged with the Sequin submission software. Information on submitting these sequences can be found at Whole Genome Shotgun Submissions.

Bulk Submissions: EST, STS, and GSS

Expressed Sequence Tags (EST), Sequence Tagged Sites (STSs), and Genome Survey Sequences (GSSs) sequences are generally submitted in a batch and are usually part of a large sequencing project devoted to a particular genome. These entries have a streamlined submission process and undergo minimal processing before being loaded to GenBank.

ESTs are generally short (<1 kb), single-pass cDNA sequences from a particular tissue and/or developmental stage. However, they can also be longer sequences that are obtained by differential display or Rapid Amplification of cDNA Ends (RACE) experiments. The common feature of all ESTs is that little is known about them; therefore, they lack feature annotation.

STSs are short genomic landmark sequences (1). They are operationally unique in that they are specifically amplified from the genome by PCR amplification. In addition, they define a specific location on the genome and are, therefore, useful for mapping.

GSSs are also short sequences but are derived from genomic DNA, about which little is known. They include, but are not limited to, single-pass GSSs, BAC ends, exon-trapped genomic sequences, and Alu PCR sequences.

EST, STS, and GSS sequences reside in their respective divisions within GenBank, rather than in the taxonomic division of the organism. The sequences are maintained within GenBank in the dbEST, dbSTS, and dbGSS databases.

Submitting Data to dbEST, dbSTS, or dbGSS

Because of the large numbers of sequences that are submitted at once, dbEST, dbSTS, and dbGSS entries are stored in relational databases where information that is common to all sequences can be shared. Submissions consist of several files containing the common

information, plus a file of the sequences themselves. The three types of submissions have different requirements, but all include a Publication file and a Contact file. See the dbEST, dbSTS, and dbGSS pages for the specific requirements for each type of submission.

In general, users generate the appropriate files for the submission type and then email the files to batch-sub@ncbi.nlm.nih.gov. If the files are too big for email, they can be deposited into a FTP account. Upon receipt, the files are examined by a GenBank annotator, who fixes any errors when possible or contacts the submitter to request corrected files. Once the files are satisfactory, they are loaded into the appropriate database and assigned Accession numbers. Additional formatting errors may be detected at this step by the data-loading software, such as double quotes anywhere in the file or invalid characters in the sequences. Again, if the annotator cannot fix the errors, a request for a corrected submission is sent to the user. After all problems are resolved, the entries are loaded into GenBank.

Bulk Submissions: HTC and FLIC

HTC records are High-Throughput cDNA/mRNA submissions that are similar to ESTs but often contain more information. For example, HTC entries often have a systematic gene name (not necessarily an official gene name) that is related to the lab or center that submitted them, and the longest open reading frame is often annotated as a coding region.

FLIC records, Full-Length Insert cDNA, contain the entire sequence of a cloned cDNA/mRNA. Therefore, FLICs are generally longer, and sometimes even full-length, mRNAs. They are usually annotated with genes and coding regions, although these may be lab systematic names rather than functional names.

HTC Submissions

HTC entries are usually generated with Sequin or tbl2asn, and the files are emailed to gb-sub@ncbi.nlm.nih.gov. If the files are too big for email, then by prior arrangement, the submitter can deposit the files by FTP and send a notification to gb-admin@ncbi.nlm.nih.gov that files are on the FTP site.

HTC entries undergo the same validation and processing as non-bulk submissions. Once processing is complete, the records are loaded into GenBank and are available in Entrez and other retrieval systems.

FLIC Submissions

FLICs are processed via an automated FLIC processing system that is based on the HTG automated processing system. Submitters use the program tbl2asn to generate their submissions. As with HTG submissions, submissions to the automated FLIC processing system must contain three identifiers: the genome center tag, the sequence name (SeqId), and the Accession number. The genome center tag is assigned by NCBI and is generally

the FTP account login name. The sequence name is a unique identifier that is assigned by the submitter to a particular clone or entry and must be unique within the group's FLIC submissions. When a sequence is first submitted, it has only a sequence name and genome center tag; the Accession number is assigned during processing. All updates to that entry include the center tag, sequence name, and Accession number, or processing will fail.

The FLIC Processing Pathway

The FLIC processing system is analogous to the HTG processing system. Submitters deposit their submissions in the FLICSEQSUBMIT directory of their FTP account and notify us that the submissions are there. We then run the scripts to pick up the files from the FTP site and copy them to the processing pathway, as well as to an archive. Once processing is complete and if there are no errors in the submission, the files are automatically loaded into GenBank.

As with HTG submissions, FLIC entries can fail for three reasons: problems with the format, problems with the identification of the record (the genome center, the SeqId, or the Accession number), or problems with the data itself. When submissions fail FLIC processing, a GenBank annotator sends email to the sequencing center, describing the problem and asking the center to submit a corrected entry. Annotators do not fix incorrect submissions; this ensures that the staff of the submitting genome center fixes the problems in their database as well. At the completion of processing, reports are generated and deposited in the submitter's FTP account, as described for HTG submissions.

Submission Tools

Direct submissions to GenBank are prepared using one of two submission tools, BankIt or Sequin.

BankIt

BankIt is a Web-based form that is a convenient and easy way to submit a small number of sequences with minimal annotation to GenBank. To complete the form, a user is prompted to enter submitter information, the nucleotide sequence, biological source information, and features and annotation pertinent to the submission. BankIt has extensive Help documentation to guide the submitter. Included with the Help document is a set of annotation examples that detail the types of information that are required for each type of submission. After the information is entered into the form, BankIt transforms this information into a GenBank flatfile for review. In addition, a number of quality assurance and validation checks ensure that the sequence submitted to GenBank is of the highest quality. The submitter is asked to include spans (sequence coordinates) for the coding regions and other features and to include amino acid sequence for the proteins that derive from these coding regions. The BankIt validator compares the amino acid sequence provided by the submitter with the conceptual translation of the coding region based on the provided spans. If there is a discrepancy, the submitter is requested to fix the problem, and the process is halted until the error is resolved. To prevent the deposit of

sequences that contain cloning vector sequence, a BLAST similarity search is performed on the sequence, comparing it to the VecScreen database. If there is a match to this database, the user is asked to remove the contaminating vector sequence from their submission or provide an explanation as to why the screen was positive. Completed forms are saved in ASN.1 format, and the entry is submitted to the GenBank processing queue. The submitter receives confirmation by email, indicating that the submission process was successful.

Sequin

Sequin is more appropriate for complicated submissions containing a significant amount of annotation or many sequences. It is a stand-alone application available on NCBI's FTP site. Sequin creates submissions from nucleotide and amino acid sequences in FASTA format with tagged biological source information in the FASTA definition line. As in BankIt, Sequin has the ability to predict the spans of coding regions. Alternatively, a submitter can specify the spans of their coding regions in a five-column, tab-delimited table and import that table into Sequin. For submitting multiple, related sequences, e.g., those in a phylogenetic or population study, Sequin accepts the output of many popular multiple sequence-alignment packages, including FASTA+GAP, PHYLIP, MACAW, NEXUS Interleaved, and NEXUS Contiguous. It also allows users to annotate features in a single record or a set of records globally. For more information on Sequin, see Chapter 12.

Completed Sequin submissions should be emailed to GenBank at gb-sub@ncbi.nlm.nih.gov. Larger files may be submitted by SequinMacrosend.

Sequence Data Flow and Processing: From Laboratory to GenBank

Triage

All direct submissions to GenBank, created either by Sequin or BankIt, are processed by the GenBank annotation staff. The first step in processing submissions is called triage. Within 48 hours of receipt, the database staff reviews the submission to determine whether it meets the minimal criteria for incorporation into GenBank and then assigns an Accession number to each sequence. All sequences must be >50 bp in length and be sequenced by, or on behalf of, the group submitting the sequence. GenBank will not accept sequences constructed *in silico*; noncontiguous sequences containing internal, unsequenced spacers; or sequences for which there is not a physical counterpart, such as those derived from a mix of genomic DNA and mRNA. Submissions are also checked to determine whether they are new sequences or updates to sequences submitted previously. After receiving Accession numbers, the sequences are put into a queue for more extensive processing and review by the annotation staff.

Indexing

Triaged submissions are subjected to a thorough examination, referred to as the indexing phase. Here, entries are checked for:

1. Biological validity. For example, does the conceptual translation of a coding region match the amino acid sequence provided by the submitter? Annotators also ensure that the source organism name and lineage are present, and that they are represented in NCBI's taxonomy database. If either of these is not true, the submitter is asked to correct the problem. Entries are also subjected to a series of BLAST similarity searches to compare the annotation with existing sequences in GenBank.
2. Vector contamination. Entries are screened against NCBI's UniVec database to detect contaminating cloning vector.
3. Publication status. If there is a published citation, PubMed and MEDLINE identifiers are added to the entry so that the sequence and publication records can be linked in Entrez.
4. Formatting and spelling. If there are problems with the sequence or annotation, the annotator works with the submitter to correct them.

Completed entries are sent to the submitter for a final review before release into the public database. If the submitters requested that their sequences be released after processing, they have 5 days to make changes prior to release. The submitter may also request that GenBank hold their sequence until a future date. The sequence must become publicly available once the Accession number or the sequence has been published. The GenBank annotation staff currently processes about 1,900 submissions per month, corresponding to approximately 20,000 sequences.

GenBank annotation staff must also respond to email inquiries that arrive at the rate of approximately 200 per day. These exchanges address a range of topics including:

- updates to existing GenBank records, such as new annotation or sequence changes
- problem resolution during the indexing phase
- requests for release of the submitter's sequence data or an extension of the hold date
- requests for release of sequences that have been published but are not yet available in GenBank
- lists of Accession numbers that are due to appear in upcoming issues of a publisher's journals
- reports of potential annotation problems with entries in the public database
- requests for information on how to submit data to GenBank

One annotator is responsible for handling all email received in a 24-hour period, and all messages must be acted upon and replied to in a timely fashion. Replies to previous emails are forwarded to the appropriate annotator.

Processing Tools

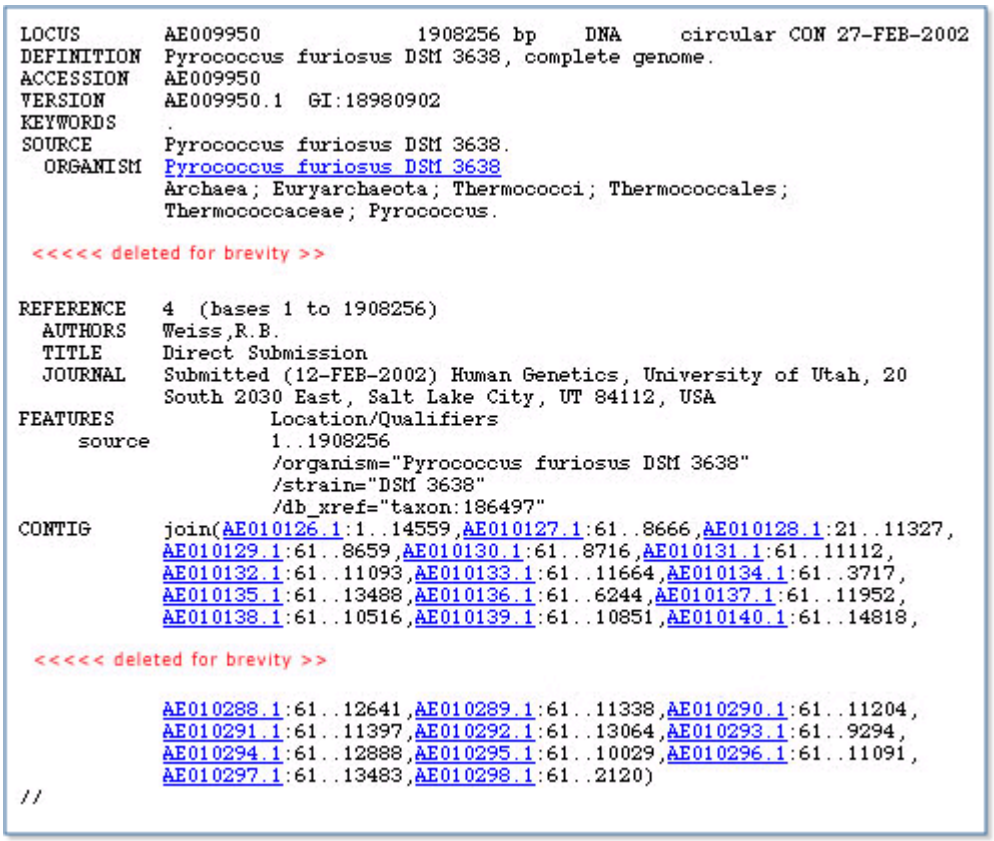
The annotation staff uses a variety of tools to process and update sequence submissions. Sequence records are edited with Sequin, which allows staff to annotate large sets of records by global editing rather than changing each record individually. This is truly a time saver because more than 100 entries can be edited in a single step (see Chapter 12 on Sequin for more details). Records are stored in a database that is accessed through a queue management tool that automates some of the processing steps, such as looking up taxonomy and PubMed data, starting BLAST jobs, and running automatic validation checks. Hence, when an annotator is ready to start working on an entry, all of this information is ready to view. In addition, all of the correspondence between GenBank staff and the submitter is stored with the entry. For updates to entries already present in the public database, the live version of the entry is retrieved from ID, and after making changes, the annotator loads the entry back into the public database. This entry is available to the public immediately after loading.

Microbial Genomes

The GenBank direct submissions group has processed more than 50 complete microbial genomes since 1996. These genomes are relatively small in size compared with their eukaryotic counterparts, ranging from five hundred thousand to five million bases. Nonetheless, these genomes can contain thousands of genes, coding regions, and structural RNAs; therefore, processing and presenting them correctly is a challenge. Currently, the DDBJ/EMBL/GenBank Nucleotide Sequence Database Collaboration has a 350-kilobase (kb) upper size limit for sequence entries. Because a complete bacterial genome is larger than this arbitrary limit, it must be split into pieces. GenBank routinely splits complete microbial genomes into 10-kb pieces with a 60-bp overlap between pieces. Each piece contains approximately 10 genes. A CON entry, containing instructions on how to put the pieces back together, is also made. The CON entry contains descriptor information, such as source organism and references, as well as a join statement providing explicit instructions on how to generate the complete genome from the pieces. The Accession number assigned to the CON record is also added as a secondary Accession number on each of the pieces that make up the complete genome (see Figure 2).

Submitting and Processing Data

Submitters of complete genomes are encouraged to contact us at genomes@ncbi.nlm.nih.gov before preparing their entries. A FTP account is required to submit large files, and the submission should be deposited at least 1 month before publication to allow for processing time and coordinated release before publication. In addition, submitters are required to follow certain guidelines, such as providing unique identifiers for proteins and systematic names for all genes. Entries should be prepared with the submission tool tbl2asn, a utility that is part of the Sequin package (Chapter 12). This utility creates an ASN.1 submission file from a five-column, tab-delimited file



```
LOCUS       AE009950             1908256 bp    DNA     circular CON 27-FEB-2002
DEFINITION  Pyrococcus furiosus DSM 3638, complete genome.
ACCESSION   AE009950
VERSION     AE009950.1  GI:18980902
KEYWORDS    .
SOURCE      Pyrococcus furiosus DSM 3638.
  ORGANISM  Pyrococcus furiosus DSM 3638
            Archaea; Euryarchaeota; Thermococci; Thermococcales;
            Thermococcaceae; Pyrococcus.

 <<<<< deleted for brevity >>

REFERENCE   4  (bases 1 to 1908256)
  AUTHORS   Weiss,R.B.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-FEB-2002) Human Genetics, University of Utah, 20
            South 2030 East, Salt Lake City, UT 84112, USA
FEATURES             Location/Qualifiers
     source          1..1908256
                     /organism="Pyrococcus furiosus DSM 3638"
                     /strain="DSM 3638"
                     /db_xref="taxon:186497"
CONTIG      join(AE010126.1:1..14559,AE010127.1:61..8666,AE010128.1:21..11327,
            AE010129.1:61..8659,AE010130.1:61..8716,AE010131.1:61..11112,
            AE010132.1:61..11093,AE010133.1:61..11664,AE010134.1:61..3717,
            AE010135.1:61..13488,AE010136.1:61..6244,AE010137.1:61..11952,
            AE010138.1:61..10516,AE010139.1:61..10851,AE010140.1:61..14818,

 <<<<< deleted for brevity >>

            AE010288.1:61..12641,AE010289.1:61..11338,AE010290.1:61..11204,
            AE010291.1:61..11397,AE010292.1:61..13064,AE010293.1:61..9294,
            AE010294.1:61..12888,AE010295.1:61..10029,AE010296.1:61..11091,
            AE010297.1:61..13483,AE010298.1:61..2120)
//
```

Figure 2. The information toward the *bottom* of the record describes how to generate the complete genome from the pieces.

containing feature annotation, a FASTA-formatted nucleotide sequence, and an optional FASTA-formatted protein sequence.

Complete genome submissions are reviewed by a member of the GenBank annotation staff to ensure that the annotation and gene and protein identifiers are correct, and that the entry is in proper GenBank format. Any problems with the entry are resolved through communication with the submitter. Once the record is complete, the genome is carefully split into its component pieces. The genome is split so that none of the breaks occurs within a gene or coding region. A member of the annotation staff performs quality assurance checks on the set of genome pieces to ensure that they are correct and representative of the complete genome. The pieces are then loaded into GenBank, and the CON record is created.

The microbial genome records in GenBank are the building blocks for the Microbial Genome Resources in Entrez Genomes.

Third Party Annotation (TPA) Sequence Database

The vast amount of publicly available data from the human genome project and other genome sequencing efforts is a valuable resource for scientists throughout the world. A laboratory studying a particular gene or gene family may have sequenced numerous cDNAs but has neither the resources nor inclination to sequence large genomic regions containing the genes, especially when the sequence is available in public databases. The researcher might choose then to download genomic sequences from GenBank and perform experimental analyses on these sequences. However, because this researcher did not perform the sequencing, the sequence, with its new annotations, cannot be submitted to DDBJ/EMBL/GenBank. This is unfortunate because important scientific information is being excluded from the public databases. To address this problem, the International Nucleotide Sequence Database Collaboration established a separate section of the database for such TPA (see Third Party Annotation Sequence Database).

All sequences in the TPA database are derived from the publicly available collection of sequences in DDBJ/EMBL/GenBank. Researchers can submit both new and alternative annotations of genomic sequence to GenBank. TPA entries can be also created by combining the exon sequences from genomic sequences or by making contigs of EST sequences to make mRNA sequences. TPA submissions must use sequence data that are already represented in DDBJ/EMBL/GenBank, have annotation that is experimentally supported, and appear in a peer-reviewed scientific journal. TPA sequences will be released to the public database only when their Accession numbers and/or sequence data appear in a peer-reviewed publication in a biological journal.

Appendix: GenBank, RefSeq, TPA and UniProt: What's in a Name?

The National Center for Biotechnology Information (NCBI) often is asked about the differences between its GenBank, RefSeq, and TPA databases and how they relate to the UniProt database. This document was prepared in response to those inquiries, and more specifically to a request from attendees at a 2006 workshop on microbial genomes held at NCBI and attended by bacterial annotation groups, sequencing centers, and members of the American Society for Microbiology (ASM). The article originally was published in the May 2007 issue of the American Society for Microbiology's journal *Microbe* (http://www.microbemagazine.org/index.php?option=com_content&view=article&id=1270:genbank-refseq-tpa-and-uniprot-whats-in-a-name&catid=347:letters&Itemid=419). While there was some input from the European Bioinformatics Institute on UniProt and Swiss-Prot for the document, it represents an NCBI perspective.

GenBank

NCBI's GenBank database is a collection of publicly available annotated nucleotide sequences, including mRNA sequences with coding regions, segments of genomic DNA with a single gene or multiple genes, and ribosomal RNA gene clusters.

GenBank is specifically intended to be an archive of primary sequence data. Thus, to be included, the sequencing must have been conducted by the submitter. NCBI does some quality control checks and will notify a submitter if something appears amiss, but it does not curate the data; the author has the final say on the sequence and annotation placed in the GenBank record. Authors are encouraged to update their records with new sequence or annotation data, but in practice records are seldom updated.

Records can be updated only by the author, or by a third party if the author has given them permission and notified NCBI. This delegation of authority has happened in a limited number of cases, generally where a genome sequence was determined by a lab or sequencing center and updating rights were subsequently given to a model organism database, which then took over ongoing maintenance of annotation.

Because GenBank is an archival database and includes all sequence data submitted, there are multiple entries for some loci. Just as the primary literature includes similar experiments conducted under slightly different conditions, GenBank may include many sequencing results for the same loci. These different sequencing submissions can reflect genetic variations between individuals or organisms, and analyzing these differences is one way of identifying single nucleotide polymorphisms.

GenBank exchanges data daily with its two partners in the International Nucleotide Sequence Database Collaboration (INSDC): the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ). Nearly all sequence data are deposited into INSDC databases by the labs that generate the sequences, in part because journal publishers generally require deposition prior to publication so that an accession number can be included in the paper.

If part of a GenBank nucleotide sequence encodes a protein, a conceptual translation – called a coding region or coding sequence (CDS) – is annotated. A protein accession number (a "protein id") is assigned to the translation product and is noted on the GenBank record. This protein id is linked to a record for the protein sequence in NCBI's protein databases. In the UniProt database, described later, these sequences are contained in the TrEMBL (Translated EMBL) portion of the database.

See the GenBank overview at http://www.ncbi.nlm.nih.gov/Genbank/index.html.

RefSeq

The Reference Sequence (RefSeq) database is a curated collection of DNA, RNA, and protein sequences built by NCBI. Unlike GenBank, RefSeq provides only one example of each natural biological molecule for major organisms ranging from viruses to bacteria to

eukaryotes. For each model organism, RefSeq aims to provide separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts. RefSeq is limited to major organisms for which sufficient data is available (almost 4,000 distinct "named" organisms as of January 2007), while GenBank includes sequences for any organism submitted (approximately 250,000 different named organisms).

To produce RefSeq records, NCBI culls the best available information on each molecule and updates the records as more information emerges. A commonly used analogy is that if GenBank is akin to the primary research literature, RefSeq is akin to the review literature.

In some cases, creation of a RefSeq record involves no more than selecting a single good example from GenBank and making a copy in RefSeq, which credits the GenBank record. In other cases, NCBI in-house staff generates and annotates the records based on the existing primary data, sometimes by combining parts of several GenBank records. Also, some records are automatically imported from other curated databases, such as the SGD database of yeast genome data and the FlyBase database of Drosophila genomes (for a list of RefSeq collaborators see www.ncbi.nlm.nih.gov/RefSeq/collaborators). The approach selected for creating a RefSeq record depends on the specific organism and the quality of information available.

When NCBI first creates a RefSeq record, the record initially reflects only the information from the source GenBank record with added links. At this point, the record has not yet been reviewed by NCBI staff, and therefore it is identified as "provisional." After NCBI examines the record – often adding information from other GenBank records, such as the sequences for the 5'UTR and 3'UTR, and providing further literature references – it is marked as "reviewed."

RefSeq records appear in a similar format as the GenBank records from which they are derived. However, they can be distinguished from GenBank records by their accession prefix, which includes an underscore, and a notation in the "comment" field that indicates the RefSeq status. RefSeq records can be accessed through NCBI's Nucleotide and Protein databases, which are among the many databases linked through the Entrez search and retrieval system. When retrieving search results, users can choose to see all GenBank records or only RefSeq records by clicking on the appropriate tab at the top of the results page. Users also can choose to search only RefSeq records, or specific types of RefSeq records (such as mRNAs), by using the "Limits" feature in Entrez. Further information about the database can be obtained at the RefSeq homepage.

Key Characteristics of GenBank versus RefSeq

GenBank	RefSeq
Not curated	Curated

Key Characteristics of continues on next page...

Key Characteristics of continued from previous page.

Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to primary literature	Akin to review articles
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

TPA

The Third Party Annotation (TPA) database contains sequences that are derived or assembled from sequences already in the INSDC databases. Whereas DDBJ, EMBL and GenBank contain primary sequence data and corresponding annotations submitted by the laboratories that did the sequencing, the TPA database contains nucleotide sequences built from the existing primary data with new annotation that has been published in a peer-reviewed scientific journal. The database includes two types of records: experimental (supported by wet-lab evidence) and inferential (where the annotation is inferred and not the subject of direct experimentation).

TPA bridges the gap between GenBank and RefSeq, permitting authors publishing new experimental evidence to re-annotate sequences in a public database as they think best, even if they were not the primary sequencer or the curator of a model organism database. These records are part of the INSDC collaboration, and thus appear in all three databases (GenBank, DDBJ and EMBL).

Like GenBank and RefSeq records, TPA records can be retrieved through the Nucleotide section of Entrez. The TPA records can be distinguished from other records by the definition line, which begins with the letters "TPA," and by the Keywords field, which states "Third Party Annotation; TPA." Users can restrict their search to TPA data by selecting the database in the Properties search field or by adding the command “AND tpa[prop]” to their query. The database is significantly smaller than GenBank, with about one record for every 12,000 in GenBank. Details about how to submit data and examples of what can and cannot be submitted to TPA are provided on the TPA homepage.

UniProt

UniProt (Universal Protein Resource) is a protein sequence database that was formed through the merger of three separate protein databases: the Swiss Institute of Bioinformatics’ and the European Bioinformatics Institute’s Swiss-Prot and TrEMBL

(Translated EMBL Nucleotide Sequence Data Library) databases, and Georgetown University's PIR-PSD (Protein Information Resource Protein Sequence Database).

Swiss-Prot and TrEMBL continue as two separate sections of the UniProt database. The Swiss-Prot component consists of manually annotated protein sequence records that have added information, such as binding sites for drugs. The TrEMBL portion consists of computationally analyzed sequence records that are awaiting full manual annotation; following curation, they are transferred to Swiss-Prot.

TrEMBL is derived from the CDS translations annotated on records in the INSDC databases, with some additional computational merging and adjustment. Given the very high rate of sequencing, and the effort it takes to do manual annotation, the Swiss-Prot component of UniProt is generally much smaller than the TrEMBL component. Because Swiss-Prot's manual annotation provides much additional information, NCBI's protein databases provide links to Swiss-Prot records, even if the sequence is the same as one or more INSDC translations.

Key Characteristics of UniProt versus GenBank and RefSeq

UniProt	GenBank and RefSeq
Produced by SIB, EBI & Georgetown U.	Produced by INSDC and NCBI
Protein data only	Protein and nucleotide data
Curated in Swiss-Prot, not in TrEMBL	Curated in RefSeq, not in GenBank

References

1. Olson M, Hood L, Cantor C, Botstein D. A common language for physical mapping of the human genome. Science. 1989;245(4925):1434–1435.

Chapter 2. PubMed: The Bibliographic Database

Kathi Canese, Jennifer Jentsch, and Carol Myers

Created: October 9, 2002; Updated: August 13, 2003.

Summary

PubMed is a database developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), one of the institutes of the National Institutes of Health (NIH). The database was designed to provide access to citations (with abstracts) from biomedical journals. Subsequently, a linking feature was added to provide access to full-text journal articles at Web sites of participating publishers, as well as to other related Web resources. PubMed is the bibliographic component of the NCBI's Entrez retrieval system.

Data Sources

MEDLINE®

PubMed's primary data resource is MEDLINE, the NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences, such as molecular biology. MEDLINE contains bibliographic citations and author abstracts from about 4,600 biomedical journals published in the United States and 70 other countries. The database contains about 12 million citations dating back to the mid-1960s. Coverage is worldwide, but most records are from English-language sources or have English abstracts.

Non-MEDLINE

In addition to MEDLINE citations, PubMed® provides access to non-MEDLINE resources, such as out-of-scope citations, citations that precede MEDLINE selection, and PubMed Central (PMC; see Chapter 9) citations. Together, these are often referred to as "PubMed-only citations." Out-of-scope citations are primarily from general science and chemistry journals that contain life sciences articles indexed for MEDLINE, e.g., the plate tectonics or astrophysics articles from *Science* magazine. Publishers can also submit citations with publication dates that precede the journal's selection for MEDLINE indexing, usually because they want to create links to older content. PMC citations are taken from life sciences journals (MEDLINE or non-MEDLINE) that submit full-text articles to PMC. In addition to the incorporation of PubMed-only citations, PubMed has been enhanced recently by the incorporation of citations from the following unique databases: HealthSTAR, AIDSLINE, HISTLINE, SPACELINE, BIOETHICSLINE, and POPLINE.

In response to new approaches to electronic publishing, PubMed can now also accommodate articles published electronically in advance of being collected into an issue. We refer to these citations as "ahead of print" or "epub" citations.

Journal Selection Criteria

All content in PubMed ultimately comes from publishers of biomedical journals, and journals that are to be included in MEDLINE are subject to a selection process. The Fact Sheet on *Journal Selection for Index Medicus®/MEDLINE®* describes the journal selection policy, criteria, and procedures for data submission.

Electronic Data Submission

Electronic data submission benefits everyone: publishers, the NLM, and users. For the NLM, it eliminates the tremendous costs associated with entering data by hand. For publishers and users, it means that newly published data appear rapidly and accurately in PubMed. Some publishers are now making pre-publication material available before it is formally published ("ahead of print" or "epub" citations); others are publishing electronic-only journals. By close collaboration with the publisher, the citations for these publications can appear in PubMed on the same day as the article is published.

Furthermore, electronic data submission allows publishers to create links from abstracts in PubMed to the full text of the appropriate articles available on their own Web site. This can be achieved using LinkOut (Chapter 17). Both subscribers to the journals and other PubMed users can access the full text according to criteria that are determined by the publishers, increasing traffic to their sites.

Although the NLM works with many publishers directly, some publishers contract with commercial data aggregators, companies that prepare and submit the publisher's data to the NLM. Many aggregators also host publisher data on their Web sites.

Electronic Data Submission Process

All electronic data are supplied via FTP to NCBI in XML format, in accordance with the NLM's specifications (document type definition, or DTD). These specifications can be found in NLM Standard Publisher Data Format document. The document includes information on XML tag descriptions, how to handle special characters (e.g., α or β), examples of tagged records, the PubMed DTD, and a FAQ section for participating or potential data providers. Publishers or other data providers who want to submit electronic data should write to: publisher@ncbi.nlm.nih.gov.

NCBI staff will guide new data providers through the approval process for file submission. New providers are asked to submit test files, which are then checked for XML formatting and syntax and for bibliographic accuracy and completeness. The files are revised and resubmitted as many times as necessary until all criteria are met. Once approved, a private account is set up on our FTP site to receive new journal issues, or in the case of online publications, individual articles as they are added to the publisher's Web site. We run a file-loading script that automatically processes the files daily, Monday through Friday at approximately 9:00 a.m. (Eastern Time). The new citations are assigned a PubMed ID number (PMID), a confirmation report is sent to the provider, and the new citations

usually become available in PubMed sometime after 11:00 a.m. the next day, Tuesday through Saturday.

After posting in PubMed, the citations are forwarded to NLM's Indexing Section for bibliographic data verification and for the addition of subject indexing terms from Medical Subject Headings [MeSH]. This process can take several weeks, after which time completed citations flow back into PubMed, replacing the originally submitted data.

Database Management and Hardware

PubMed is one of the NCBI databases within the relational database management system, Entrez (see Chapter 15). Entrez is a text-based search and retrieval system based on in-house software that uses an indexing system for rapid retrieval of information.

Requests for NCBI services, including PubMed, are first proxied through three load-balanced Dell PowerEdge 1650 servers, each with two central processing units. The proxy servers, in turn, load-balance requests forwarded on to the Web servers for PubMed and other NCBI services.

The PubMed Web servers comprise eight Dell PowerEdge 8450 servers. The Dell servers have eight central processing units, 8 GB of memory, and about 300 GB of disk space and run the Linux operating system.

The Web servers retrieve PubMed records from two Sybase SQL database servers, which run on Sun Enterprise 450s. To accommodate the data volume output by PubMed and other Web-based services, the NLM has a high-speed connection (OC-3, up to 155 Mbits/sec) to the Internet, as well as a 622 Mbits/sec connection (OC-12) to Internet2, the noncommercial network used by many leading research universities.

Indexing

PubMed Citation Status and Assignment of MeSH Terms

Citations in PubMed are assigned one of three citation status tags that display next to the PubMed ID (PMID) numbers on all PubMed citations. The citation status tags indicate the citation's stage in the MEDLINE indexing process. The three tags are:

[**PubMed - as supplied by publisher**]: This tag is displayed on citations added recently to PubMed via electronic submission from a publisher (which may or may not move on for MEDLINE MeSH indexing).

[**PubMed - in process**]: This tag is displayed on citations that have had the first stage of quality review to verify that the journal, date, volume, and/or issue are correct. They will be reviewed for other accurate bibliographic data at the article level (e.g., pagination, authors, article title, and abstract) and indexed, i.e., the articles will be reviewed and MeSH vocabulary will be assigned (if the subject of the article is within the scope of MEDLINE).

[**PubMed - indexed for MEDLINE**]: This tag is displayed on citations that have been indexed with MeSH, Publication Types, Registry Numbers, etc., and have been completely reviewed for accurate bibliographic data. This is an intellectual process of assigning controlled vocabulary terms to describe the contents of the journal article and verifying other aspects of the citation data.

Most citations that are received electronically from publishers progress through "in process" status to MEDLINE status. Those citations not indexed for MEDLINE remain tagged [PubMed - as supplied by publisher]. Citations with "in process" status proceed to MEDLINE status after MeSH terms, publication types, sequence Accession numbers, and other indexing data are added.

All records are added to PubMed Monday through Friday and become available for viewing Tuesday through Saturday. For additional information, please see the NLM Fact Sheet: *What's the Difference Between MEDLINE® and PubMed®?*

The Automatic Computer Indexing Process

The aim of the computer indexing process is to automatically create multiple machine-readable access points that refer to the different components of the journal citations for use when searching PubMed. The citations are loaded into PubMed from both the NLM Data Creation and Maintenance System (DCMS) and directly from journal publishers (Figure 1). Both sources are in XML.

During the computer indexing process, the citation information is broken down into index fields such as Journal Name, Author Name, and Title/Abstract. The words in each of the fields are checked against the corresponding index (i.e., title words in a new citation are looked up in the Title/Abstract Index). If the word already exists, the PMID of the citation is listed with that index term. If the word is a new one for the Index, it is added as a new Index term, and the PMID is listed alongside it. (In the first instance that the term already exists, the new term will have only this one citation associated with it; this is how the PubMed indexes grow.)

Each PubMed citation is, therefore, associated with several indexes, and in cases similar to the Title/Abstract Index, many different index terms can refer back to a single citation. Likewise, commonly used terms will refer to thousands of citations (the term "cell", for example, is found in the Title/Abstract of 1,092,124 citations at the time of this writing). The Field Indexes can be browsed by using PubMed's Preview/Index function.

How PubMed Queries Are Processed

Automatic Term Mapping

PubMed uses Automatic Term Mapping to process words entered in the query box by someone searching PubMed. Terms entered without a qualifier, i.e., a simple text phrase

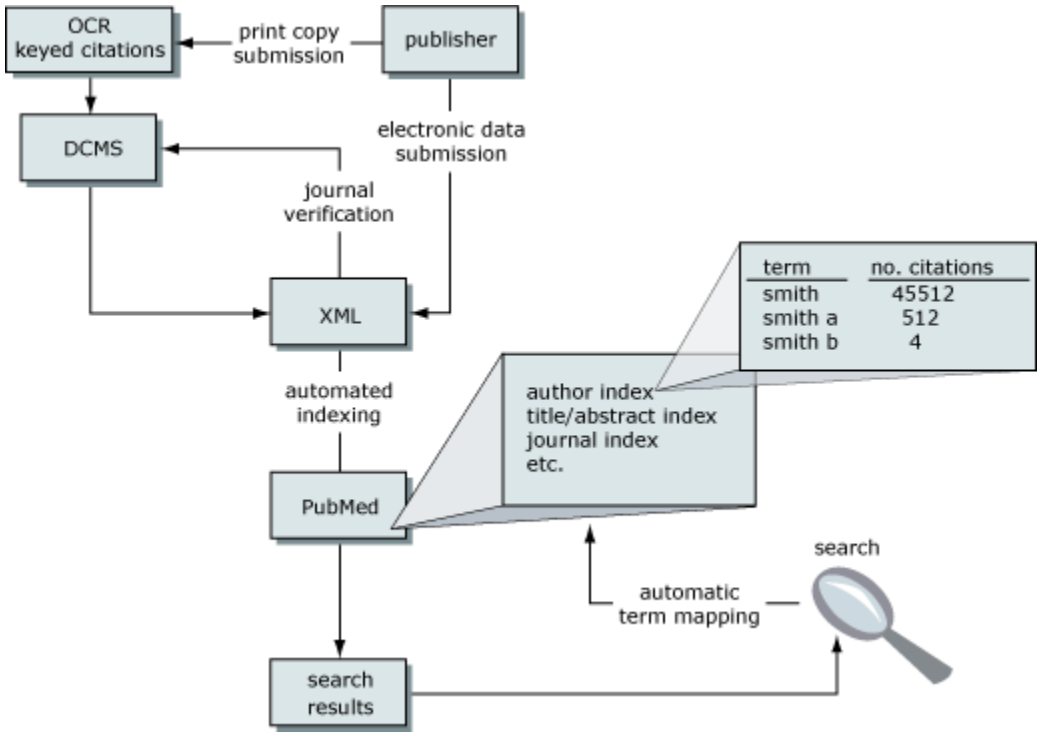


Figure 1. A schematic representation of PubMed data flow.

that does not specify a search field, are looked up against the following translation tables and indexes in a distinct order:

1. MeSH Translation Table

2. Journals Translation Table

3. Author Index

1. MeSH Translation Table

The MeSH Translation Table contains:

- MeSH Terms
- Subheadings
- See-Reference mappings (also known as entry terms) for MeSH terms
- Mappings derived from the Unified Medical Language System (UMLS) that have equivalent synonyms or lexical variants in English
- Names of Substances and synonyms to the Names of Substances (now known as Supplementary Concept Substance Names)

If the search term is found in this translation table, the term will be mapped to the appropriate MeSH term, and the Indexes will be searched as both the text word entered by the user and the MeSH term:

Search term: gallstones.

"Gallstones" is an entry term for the MeSH term "cholelithiasis" in the MeSH translation table.

Search translated to: "cholelithiasis" [MeSH Terms] OR gallstones [Text Word]

When a term is searched as a MeSH term, PubMed automatically searches that term plus the more specific terms underneath in the MeSH hierarchy:

Search term: breast cancer.

"Breast cancer" is an entry term for the MeSH term "breast neoplasms" in the MeSH translation table.

"Breast neoplasms" has the specific headings "breast neoplasms, male", "mammary neoplasms", "mammary neoplasms, experimental", and "phyllodes tumor", all of which are also searched.

2. Journals Translation Table

If the search term(s) is not found in the MeSH Translation Table, the PubMed search algorithm then looks up the term in the Journals Translation Table, which contains the full journal title, MEDLINE abbreviation, and International Standard Serial Number (ISSN):

Search term: New England Journal of Medicine.

"New England Journal of Medicine" maps to N Engl J Med.

Search translated to: "N Engl J Med" [Journal Name]

If a journal name is also a MeSH term, PubMed will search the term as both a MeSH term and as a Text Word, but not as a journal name, for a search that does not specify the "journal" field:

Search term: Cell.

Search translated as: "cells" [MeSH Terms] OR cell [Text Word]

Search term: Cell [Journal].

Search translated as: "Cell" [Journal]

3. Author Index

If the phrase is not found in MeSH or the Journals Translation Table and is a word with one or two letters after it, PubMed then checks the Author Index. The author's name should be entered in the form: Last Name (space) Initials, e.g., o'malley f, smith jp, or gomez-sanchez m.

If only one initial is used, PubMed finds all names with that first initial, and if only an author's last name is entered, PubMed will search that name in All Fields. It will not default to the Author Index because an initial does not follow the last name:

Search term: o'malley f.

Search translated as: o'malley fa, o'malley fb, o'malley fc, o'malley fd, o'malley f jr, etc.

Search term: o'malley.

Search translated as: "o'malley" [All Fields]

A history of the NLM's author indexing policy regarding the number of authors to include in a citation is outlined in Table 1.

Table 1. History of NLM author-indexing policy.

Dates	Policy
1966–1984	MEDLINE did not limit the number of authors.
1984–1995	The NLM limited the number of authors to 10, with "et al." as the eleventh occurrence.
1996–1999	The NLM increased the limit from 10 to 25. If there were more than 25 authors, the first 24 were listed, the last author was used as the 25th, and the twenty-sixth and beyond became "et al.".
2000–present	MEDLINE does not limit the number of authors.

Search Rules and Field Abbreviations

It is possible to override PubMed's Automatic Term Mapping by using search rules, syntax, and qualifying terms with search field abbreviations.

The Boolean operators AND, OR, and NOT must be entered in uppercase letters and are processed left to right. Nesting of search terms is possible by enclosing concepts in parentheses. The terms inside the set of parentheses will be processed as a unit and then incorporated into the overall strategy. Terms may be qualified using PubMed's Search Field Descriptions and Tags. Each search term should be followed by the appropriate search field tag, which indicates which field will be searched:

Search term: o'malley [au] will search only the author field. Specifying the field precludes the Automatic Term Mapping, which would result in the search o'malley[All Fields] if the field were not specified. Similarly, using the search term Cell [Journal] avoids using the MeSH Translation Table, which would interpret Cell as only a text word and MeSH term.

Using PubMed

Searching

Simple Searching

A simple search can be conducted from the PubMed homepage by entering terms in the query box and pressing Enter from the keyboard or clicking on the **Go** button on the screen.

If more than one term is entered in the query box, PubMed will go through the Automatic Term Mapping protocol described in the previous section, first looking for all the terms, as typed, to find an exact match. If the exact phrase is not found, PubMed clips a term off the end and repeats Automatic Term Mapping, again looking for an exact match, but this time to the abbreviated query. This continues until none of the words are found in any one of the translation tables. In this case, PubMed combines terms (with the AND Boolean operator) and applies the Automatic Term Mapping process to each individual word. PubMed ignores Stopwords, such as "about", "of", or "what". People can also apply their own Boolean operators (AND, OR, NOT) to multiple search terms; the Boolean operators must be in uppercase.

Search term: vitamin c common cold.

Translated as: (("ascorbic acid" [MeSH Terms] OR vitamin c [Text Word]) AND ("common cold" [MeSH Terms] OR common cold [Text Word]))

Search term: single cell separation brain.

Translated as: ((("single person" [MeSH Terms] OR single [Text Word]) AND ("cell separation" [MeSH Terms] OR cell separation [Text Word])) AND ("brain" [MeSH Terms] OR brain [Text Word]))

If a phrase of more than two terms is not found in any translation table, then the last word of the phrase is dropped, and the remainder of the phrase is sent through the entire process again. This continues, removing one word at a time, until a match is found.

If there is no match found during the Automatic Term Mapping process, the individual terms will be combined with AND and searched in All Fields.

One can see how PubMed interpreted a search by selecting **Details** from the Features Bar on the PubMed search pages after completing a search. For more information, see Details.

Complex Searching

There are a variety of ways that PubMed can be searched in a more sophisticated manner than simply typing search terms into the search box and selecting **Go**. It is possible to construct complex search strategies using Boolean operators and the various functions listed below, provided in the Features Bar:

- Limits restricts search terms to a specific search field.
- Preview/Index allows users to view and select terms from search field indexes and to preview the number of search results before displaying citations.
- History holds previous search strategies and results. The results can be combined to make new searches.
- Clipboard allows users to save or view selected citations from one search or several searches.
- Details displays the search strategy as it was translated by PubMed, including error messages.

Additional PubMed Features

The following resources are available to facilitate effective searches:

- MeSH Database allows searching of MeSH, NLM's controlled vocabulary. Users can find MeSH terms appropriate to a search strategy, obtain information about each term, and view the terms within their hierarchical structure.
- Clinical Queries is a set of search filters developed for clinicians to retrieve clinical studies of the etiology, prognosis, diagnosis, prevention, or treatment of disorders. The Systematic Reviews feature retrieves systematic reviews and meta-analysis studies by topic.
- Journal Database allows searches of journal names, MEDLINE abbreviations, or ISSN numbers for journals that are included in the Entrez system. A list of journals with links to full text is also included.
- Single Citation Matcher is a “fill-in-the-blank” form that allows a user to find the PubMed ID (PMID) number for a single article or all citations in a given journal issue by entering partial journal citation information.
- Batch Citation Matcher allows users to find PMID numbers that correspond to their own list of citations. Publishers or other database providers who want to link directly from bibliographic references on their Web sites to entries in PubMed use this service frequently.
- Cubby is a place for users to store search strategies, LinkOut preferences, and changes to the default Document Delivery Services.

Results

PubMed retrieves and displays search results in the Summary format in the order the record was initially added to PubMed, with the most recent first. (Note that this date can differ widely from the publication date.) Citations can be viewed in several other formats and can be sorted, saved, and printed, or the full text can be ordered.

Links from PubMed

A variety of links can be found on PubMed citations including:

Related Articles, which retrieves a precalculated set of PubMed citations that are closely related to the selected article. PubMed creates this set by comparing words from the title, abstract, and MeSH terms using a word-weighted algorithm.

LinkOut, which provides links to publishers, aggregators, libraries, biological databases, sequencing centers, and other Web sites. These link to the provider's site to obtain the full text of articles or related resources, e.g., consumer health information or molecular biology database records. There may be a charge to access the text or information, depending on the policy of the provider.

Books, which provides links to textbooks so that users can explore unfamiliar concepts found in search results. In collaboration with book publishers, NCBI is adapting textbooks for the Web and linking them to PubMed. The Books link displays a facsimile of the abstract, in which some words or phrases show up as hypertext links to the corresponding terms in the books available at NCBI. Selecting a hyperlinked word or phrase takes you to a list of book entries in which the phrase is found.

NCBI databases, as well as other resources, may be available from the **Links** pull-down menu to the right of each citation and from the **Display** pull-down menu. PubMed will return only the first 500 items when using the **Display** pull-down menu, from which the following links are available:

- Protein – amino acid (protein) sequences from SWISS-PROT, PIR, PRF, and PDB and translated protein sequences from the DNA sequences databases.
- Nucleotide – DNA sequences from GenBank, EMBL, and DDBJ.
- PopSet – aligned sequences submitted as a set from a population, phylogenetic, or mutation study describing such events as evolution and population variation.
- Structure – three-dimensional structures from the Molecular Modeling Database (MMDB) that were determined by X-ray crystallography and NMR spectroscopy.
- Genome – records and graphic displays of entire genomes and chromosomes for megabase-scale sequences.
- ProbeSet – gene expression data repository and online resource for the retrieval of gene expression data from any organism or artificial source.
- OMIM – directory of human genes and genetic disorders.
- SNP – dbSNP is a database of single nucleotide polymorphisms.

- Domains – The Domains database is used to identify the conserved domains present in a protein sequence.
- 3D Domains – the domains from Entrez Structure.
- PMC – PubMed Central.

How to Create Hyperlinks to PubMed

The Entrez system provides three distinct ways to create Web URL links that search and retrieve items from PubMed and the molecular biology databases: (1) by using the Entrez Programming Utilities; (2) via the URL button on the **Details** screen; and (3) by constructing URLs by hand.

The Entrez Programming Utilities can be used to create URL links directly to all Entrez data, including PubMed citations and their link information, without using the front-end Entrez query engine. These Utilities provide a fast, efficient way to search and download citation data.

Customer Support

If you need more assistance, please contact our Customer Support services by selecting the **Support Center** link displayed on all PubMed pages or by sending an email to custserv@nlm.nih.gov. You may also contact the NLM Customer Service Desk at 1-888-346-3656 [(1-888)-FINDNLM]. Hours of operation are Monday through Friday from 8:30 a.m. to 8:45 p.m. and Saturday from 9:00 a.m. to 5:00 p.m. (Eastern Time).

Additional information is also available in the PubMed Tutorial, PubMed Training Manuals, and NLM Technical Bulletin.

FAQs are available on all PubMed pages.

Chapter 3. Macromolecular Structure Databases

Eric Sayers and Steve Bryant

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The resources provided by NCBI for studying the three-dimensional (3D) structures of proteins center around two databases: the Molecular Modeling Database (MMDB), which provides structural information about individual proteins; and the Conserved Domain Database (CDD), which provides a directory of sequence and structure alignments representing conserved functional domains within proteins(CDs). Together, these two databases allow scientists to retrieve and view structures, find structurally similar proteins to a protein of interest, and identify conserved functional sites.

To enable scientists to accomplish these tasks, NCBI has integrated MMDB and CDD into the Entrez retrieval system (Chapter 15). In addition, structures can be found by BLAST, because sequences derived from MMDB structures have been included in the BLAST databases (Chapter 16). Once a protein structure has been identified, the domains within the protein, as well as domain "neighbors" (i.e., those with similar structure) can be found. For novel data not yet included in Entrez, there are separate search services available.

Protein structures can be visualized using Cn3D, an interactive 3D graphic modeling tool. Details of the structure, such as ligand-binding sites, can be scrutinized and highlighted. Cn3D can also display multiple sequence alignments based on sequence and/or structural similarity among related sequences, 3D domains, or members of a CDD family. Cn3D images and alignments can be manipulated easily and exported to other applications for presentation or further analysis.

Overview

The Structure homepage (Figure 1) contains links to the more specialized pages for each of the main tools and databases, introduced below, as well as search facilities for the Molecular Modeling Database (MMDB; Ref. 1).

MMDB is based on the structures within Protein Data Bank (PDB) and can be queried using the Entrez search engine, as well as via the more direct but less flexible Structure **Summary** search (see Figure 1). Once found, any structure of interest can be viewed using Cn3D (2), a piece of software that can be freely downloaded for Mac, PC, and UNIX platforms.

Often used in conjunction with Cn3D is the Vector Alignment Search Tool (VAST; Refs. 3, 4). VAST is used to precompute "structure neighbors" or structures similar to each MMDB entry. For people that have a set of 3D coordinates for a protein not yet in MMDB, there is also a VAST search service. The output of the precomputed VAST searches is a list of structure records, each representing one of the Non-Redundant PDB

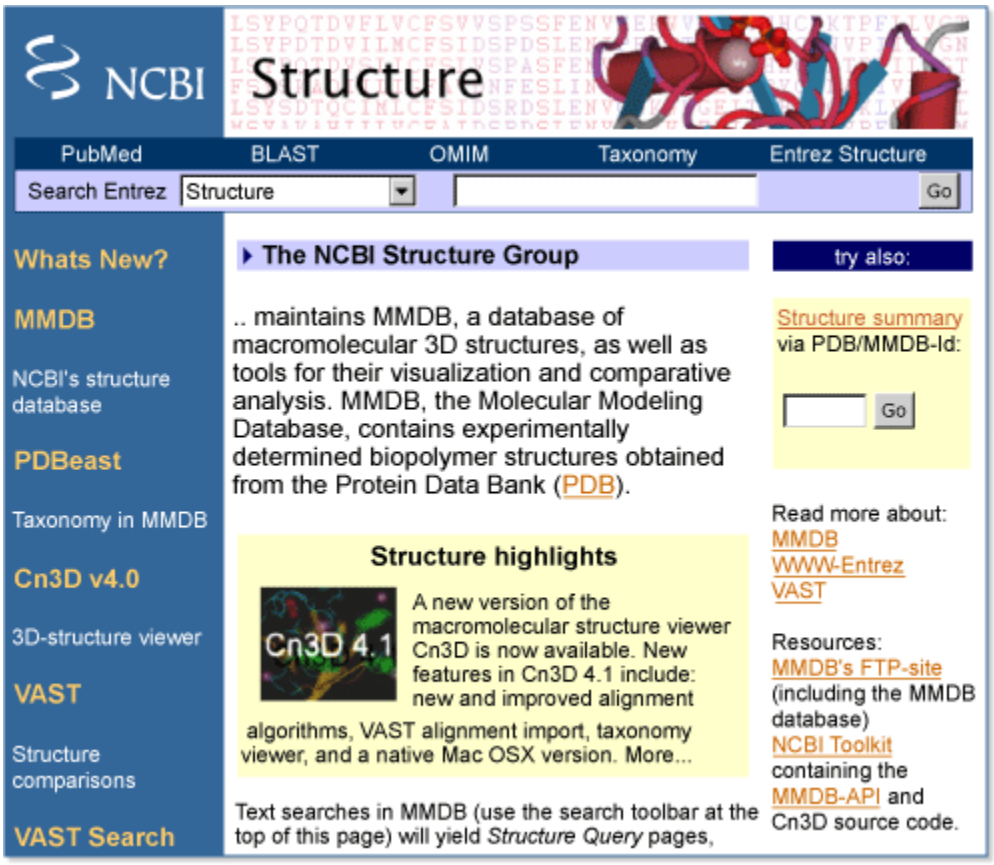


Figure 1. The Structure homepage. This page can be found by selecting the **Structure** link on the tool bar atop many NCBI Web pages. Two searches can be performed from this page, an Entrez **Structure** search or a Structure **Summary** search. Both query the MMDB database. The difference is that the **Entrez Structure** can take any text as a query (such as a PDB code, protein name, text word, author, or journal) and will result initially in a list of one or more document summaries, displayed within the Entrez environment (Chapter 15), whereas only a PDB code or MMDB ID number can be used for the Structure **Summary** search, resulting in direct display of the Structure Summary page for that record (Figure 2). Announcements about new features or updates can also be found on this page, as well as links to more specialized pages on the various Structure databases and tools.

chain sets (nr-PDB), which can also be downloaded. There are four clustered subsets of MMDB that compose nr-PDB, each consisting of clusters having a preset level of sequence similarity.

The structures within MMDB are now being linked to the NCBI Taxonomy database (Chapter 4). Known as the PDBeast project, this effort makes it possible to find: (1) all MMDB structures from a particular organism; and (2) all structures within a node of the taxonomy tree (such as lizards or *Bacillus)*, which launches the Taxonomy Browser showing the number of MMDB records in each node.

The second database within the **Structure** resources is the Conserved Domain Database (CDD; Ref. 5), originally based largely on Pfam and SMART, collections of alignments

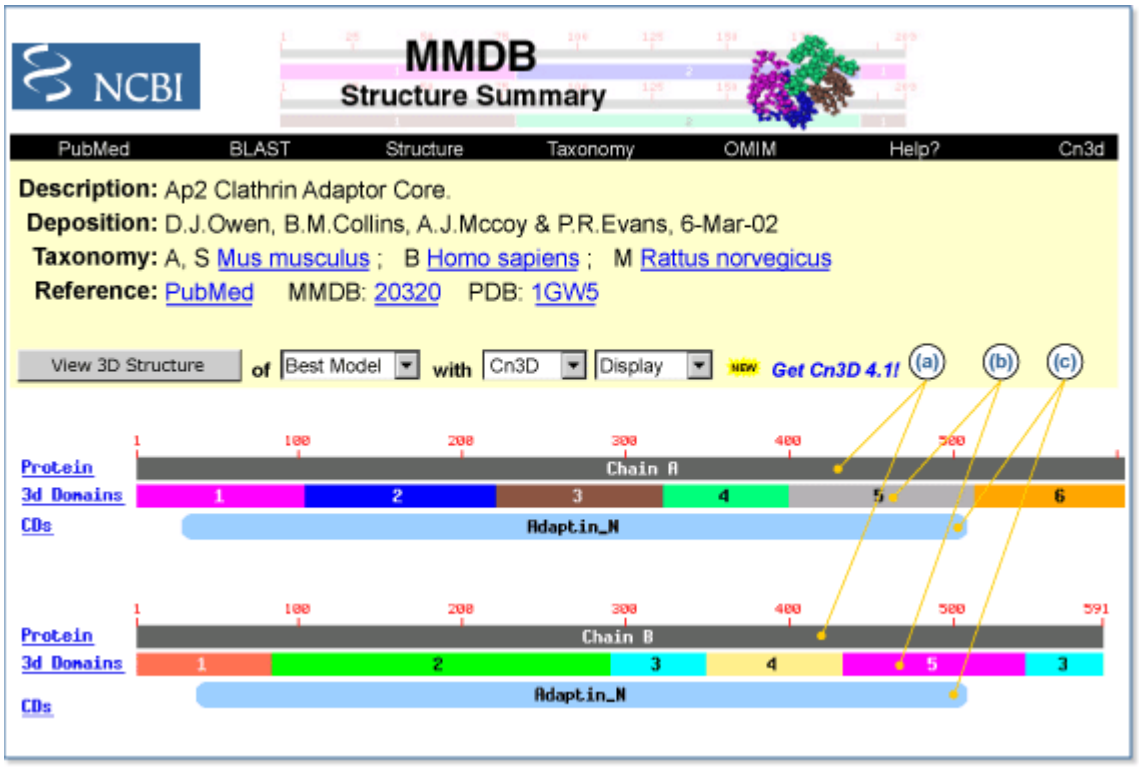


Figure 2. The Structure Summary page. The page consists of three parts: the header, the view bar, and the graphic display. The header contains basic identifying information about the record: a description of the protein (*Description:*), the author list (*Deposition:*), the species of origin (*Taxonomy:*), literature references (*Reference:*), the MMDB-ID (*MMDB:*), and the PDB code (*PDB:*). Several of these data serve as links to additional information. For example, the species name links to the Taxonomy browser, the literature references link to PubMed, and the PDB code links to the PDB Web site. The view bar allows the user to view the structure record either as a graphic with Cn3D or as a text record in either ASN.1, PDB (RasMol), or Mage formats. The latter can also be downloaded directly from this page. The graphic display contains a variety of information and links to related databases: (*a*) The Chain bar. Each chain of the molecule is displayed as a *dark bar* labeled with residue numbers. To the *left* of this bar is a **Protein** hyperlink that takes the user to a view of the protein record in Entrez Protein. The bar itself is also a hyperlink and displays the VAST neighbors of the chain. If a structure contains nucleotide sequences, they are displayed in the order contained in the PDB record. A **Nucleotide** hyperlink to their *left* takes the user to the appropriate record in Entrez Nucleotide. (*b*) The VAST (3D) Domain bar. The *colored bars* immediately below the Chain bar indicate the locations of structural domains found by the original MMDB processing of the protein. In many cases, such a domain contains unconnected sections of the protein sequence, and in such cases, discontinuous pieces making up the domain will have bars of the same color. To the *left* of the Domain bar is a 3D Domains hyperlink (*3d Domains*) that launches the 3D Domains browser in Entrez, where the user can find information about each constituent domain. Selecting a colored segment displays the VAST Structure Neighbors page for that domain. (*c*) The CD bar. Below the VAST Domain bar are *rounded, rectangular bars* representing conserved domains found by a CD-Search. The bars identify the best scoring hits; overlapping hits are shown only if the mutual overlap with hits having better scores is less than 50%. The *CDs* hyperlink to the *left* of the bar displays the CD records in Entrez Domains. Each of the colored bars is also a hyperlink that displays the corresponding CD Summary page configured to show the multiple alignment of the protein sequence with members of the selected CD.

that represent functional domains conserved across evolution. CDD now also contains the alignments of the NCBI COG database, the NCBI Library of Ancient Domains (LOAD) along with new curated alignments assembled at NCBI. CDD can be searched from the CDD page in several ways, including by a domain keyword search. Three tools have been developed to assist in analysis of CDD: (1) the CD-Search, which uses a BLAST-based algorithm to search the position-specific scoring matrices (PSSM) of CDD alignments; (2) the CD-Browser, which provides a graphic display of domains of interest, along with the sequence alignment; and (3) the Conserved Domain Architecture Retrieval Tool (CDART), which searches for proteins with similar domain architectures.

All the above databases and tools are discussed in more detail in other parts of this Chapter, including tips on how to make the best use of them.

Content of the Molecular Modeling Database (MMDB)

Sources of Primary Data

To build MMDB (1), 3D structure data are retrieved from the PDB database (6) administered by the Research Collaboratory for Structural Bioinformatics (RCSB). In all cases, the structures in MMDB have been determined by experimental methods, primarily X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. Theoretical structure models are omitted. The data in each record are then checked for agreement between the atomic coordinates and the primary sequence, and the sequence data are then extracted from the coordinate set. The resulting agreement between sequence and structure allows the record to be linked efficiently into searches and alignment displays involving other NCBI databases.

The data are converted into ASN.1 (7), which can be parsed easily and can also accept numerous annotations to the structure data. In contrast to a PDB record, a MMDB record in ASN.1 contains all necessary bonding information in addition to sequence information, allowing consistent display of the 3D structure using Cn3D. The annotations provided in the PDB record by the submitting authors are added, along with uniformly defined secondary structure and domain features. These features support structure-based similarity searches using VAST. Finally, two coordinate subsets are added to the record: one containing only backbone atoms, and one representing a single-conformer model in cases where multiple conformations or structures were present in the PDB record. Both of these additions further simplify viewing both an individual structure and its alignments with structure neighbors in Cn3D. When this process is complete, the record is assigned a unique Accession number, the MMDB-ID (Box 1), while also retaining the original four-character PDB code.

Box 1. Accession numbers.

MMDB records have several types of Accession numbers associated with them, representing the following data types:

- Each MMDB record has at least three Accession numbers: the PDB code of the corresponding PDB record (e.g., 1CYO, 1B8G); a unique MMDB-ID (e.g., 645, 12342); and a gi number for each protein chain. A new MMDB-ID is assigned whenever PDB updates either the sequence or coordinates of a structure record, even if the PDB code is retained.
- If an MMDB record contains more than one polypeptide or nucleotide chain, each chain in the MMDB record is assigned an Accession number in Entrez Protein or Nucleotide consisting of the PDB code followed by the letter designating that chain (e.g., 1B8GA, 3TATB, 1MUHB).
- Each 3D Domain identified in an MMDB record is assigned a unique integer identifier that is appended to the Accession number of the chain to which it belongs (e.g., 1B8G A 2). This new Accession number becomes its identifier in Entrez 3D Domains. New 3D Domain identifiers are assigned whenever a new MMDB-ID is assigned.
- For conserved domains, the Accession number is based on the source database:

```
Pfam:          pfam00049
SMART:         smart00078
LOAD:          LOAD Toprim
CD:            cd00101
COG:           COG5641
```

Annotation of 3D Domains

After initial processing, 3D domains are automatically identified within each MMDB record. 3D domains are annotations on individual MMDB structures that define the boundaries of compact substructures contained within them. In this way, they are similar to secondary structure annotations that define the boundaries of helical or β-strand substructures. Because proteins are often similar at the level of domains, VAST compares each 3D domain to every other one and to complete polypeptide chains. The results are stored in Entrez as a **Related 3D Domain** link.

To identify 3D domains within a polypeptide chain, MMDB's domain parser searches for one or more breakpoints in the structure. These breakpoints fall between major secondary structure elements such that the ratio of intra- to interdomain contacts remains above a set threshold. The 3D domains identified in this way provide a means to both increase the sensitivity of structure neighbor calculations and also present 3D superpositions based on compact domains as well as on complete polypeptide chains. They are not intended to

represent domains identified by comparative sequence and structure analysis, nor do they represent modules that recur in related proteins, although there is often good agreement between domain boundaries identified by these methods.

Links to Other NCBI Resources

After initially processing the PDB record, structure staff add a number of links and other information that further integrate the MMDB record with other NCBI resources. To begin, the sequence information extracted from the PDB record is entered into the Entrez Protein and/or Nucleotide databases as appropriate, providing a means to retrieve the structure information from sequence searches. As with all sequences in Entrez, precomputed BLAST searches are then performed on these sequences, linking them to other molecules of similar sequence. For proteins, these BLAST neighbors may be different than those determined by VAST; whereas VAST uses a conservative significance threshold, the structural similarities it detects often represent remote relationships not detectable by sequence comparison. The literature citations in the PDB record are linked to PubMed so that Entrez searches can allow access to the original descriptions of the structure determinations. Finally, semiautomatic processing of the "source" field of the PDB record provides links to the NCBI Taxonomy database. Although these links normally follow the genus and species information given, in some cases this information is either absent in the PDB record or refers only to how a sample was obtained. In these cases, the staff manually enters the appropriate taxonomy links.

The MMDB Record

The Structure Summary page for each MMDB record summarizes the database content for that record and serves as a starting point for analyzing the record using the NCBI structure tools (Figure 2).

VAST Structure Neighbors

Although VAST itself is not a database, the VAST results computed for each MMDB record are stored with this record and are summarized on a separate page for the whole polypeptide chain as well as for each 3D domain found in the protein (Figure 3). These pages can be accessed most easily by clicking on either the chain bar or the 3D Domain bar in the graphic display of the Structure Summary page (Figure 2).

nr-PDB

The non-redundant PDB database (nr-PDB) is a collection of four sets of sequence-dissimilar cluster PDB polypeptide chains assembled by NCBI Structure staff. The four sets differ only in their respective levels of non-redundancy. The staff assembles each set by comparing all the chains available from PDB with each other using the BLAST algorithm. The chains are then clustered into groups of similar sequence using a single-linkage clustering procedure. Chains within a sequence-similar group are automatically ranked according to the quality of their structural data. Details of the measures used to

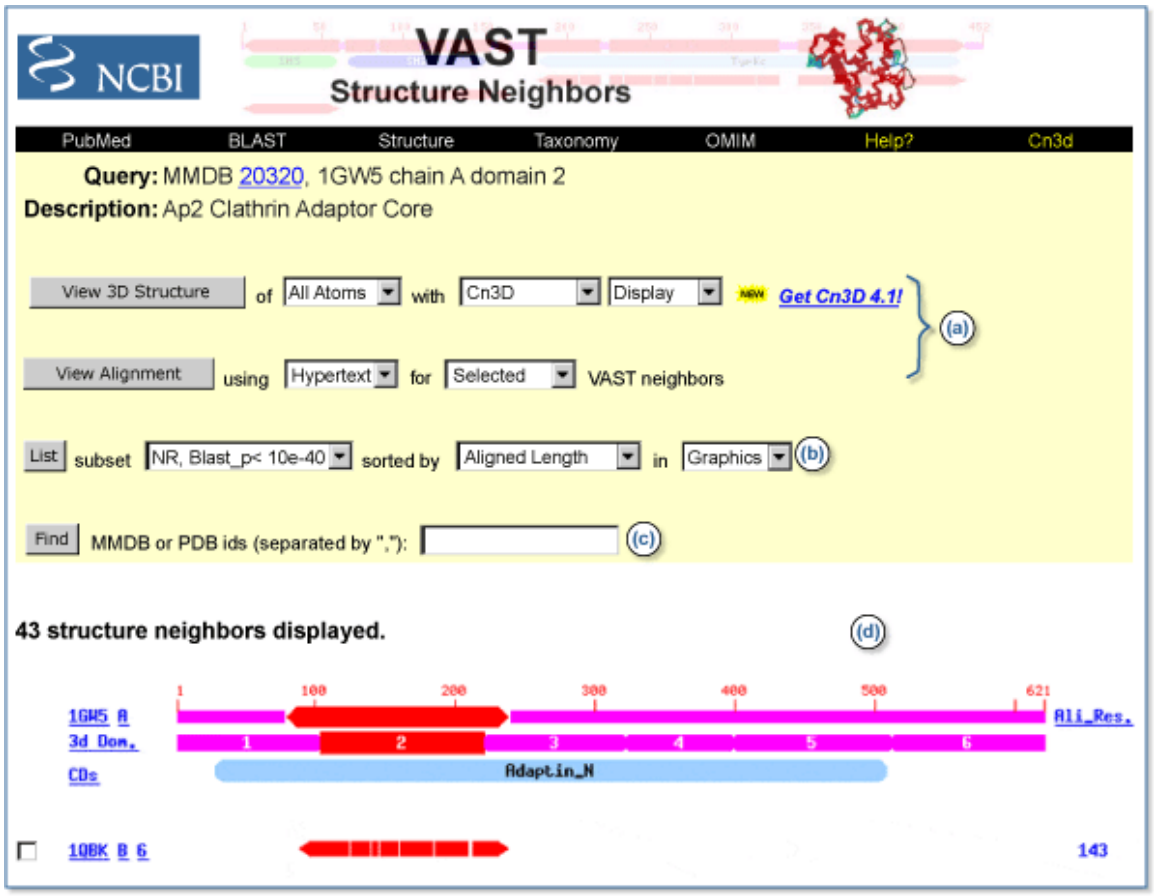


Figure 3. VAST Structure Neighbors page. The *top portion* of the page contains identifying information about the 3D Domain, along with three functional bars. (*a*) The View bar. This bar allows a user to view a selected alignment either as a graphic using Cn3D or as a sequence alignment in HTML, text, or mFASTA format. The user may select which chains to display in the alignment by checking the boxes that appear to the *left* of each neighbor in the *lower portion* of the page. (*b*) The nr-PDB bar. This bar allows a user to either display all matching records in MMDB or to limit the displayed domains to only representatives of the selected nr-PDB set. The user may also select how the matching domains are sorted in the display and whether the results are shown as graphics or as tabulated data. (*c*) The Find bar. This bar allows the user to find specific structural neighbors by entering their PDB or MMDB identifiers. (*d*) The *lower portion* of the page displays a graphical alignment of the various matching domains. The *upper three bars* show summary information about the query sequence: the *top bar* shows the maximum extent of alignment found on all the sequences displayed on the current page (users should note that the appearance of this bar, therefore, depends on which hits are displayed); the *middle bar* represents the query sequence itself that served as input for the VAST search; and the *lower bar* shows any matching CDs and is identical to the CD bar on the Structure Summary page. Listed below these three summary bars are the hits from the VAST search, sorted according to the selection in the nr-PDB bar. Aligned regions are shown in *red*, with gaps indicating unaligned regions. To the *left* of each domain accession is a check box that can be used to select any combination of domains to be displayed either on this page or using Cn3D. Moreover, each of the bars in the display is itself a link, and placing the mouse pointer over any bar reveals both the extent of the alignment by residue number and the data linked to the bar.

determine structure precision and completeness and the methodology of assembling the nr-PDB clusters can be found on the nr-PDB Web page.

Content of the Conserved Domain Database (CDD)

What Is a Conserved Domain (CD)?

CDs are recurring units in polypeptide chains (sequence and structure motifs), the extents of which can be determined by comparative analysis. Molecular evolution uses such domains as building blocks and these may be recombined in different arrangements to make different proteins with different functions. The CDD contains sequence alignments that define the features that are conserved within each domain family. Therefore, the CDD serves as a classification resource that groups proteins based on the presence of these predefined domains. CDD entries often name the domain family and describe the role of conserved residues in binding or catalysis. Conserved domains are displayed in MMDB Structure summaries and link to a sequence alignment showing other proteins in which the domain is conserved, which may provide clues on protein function.

Sources of Primary Data

The collections of domain alignments in the CDD are imported either from two databases outside of the NCBI, named Pfam (8) and SMART (9); from the NCBI COB database; from another NCBI collection named LOAD; and from a database curated by the CDD staff. The first task is to identify the underlying sequences in each collection and then link these sequences to the corresponding ones in Entrez. If the CDD staff cannot find the Accession numbers for the sequences in the records from the source databases, they locate appropriate sequences using BLAST. Particular attention is paid to any resulting match that is linked to a structure record in MMDB, and the staff substitute alignment rows with such sequences whenever possible. After the staff imports a collection, they then choose a sequence that best represents the family. Whenever possible, the staff chooses a representative that has a structure record in MMDB.

The Position-specific Score Matrix (PSSM)

Once imported and constructed, each domain alignment in CDD is used to calculate a model sequence, called a consensus sequence, for each CD. The consensus sequence lists the most frequently found residue in each position in the alignment; however, for a sequence position to be included in the consensus sequence, it must be present in at least 50% of the aligned sequences. Aligned columns covered by the consensus sequence are then used to calculate a PSSM, which memorizes the degree to which particular residues are conserved at each position in the sequence. Once calculated, the PSSM is stored with the alignment and becomes part of the CDD. The RPS-BLAST tool locates CDs within a query sequence by searching against this database of PSSMs.

Reverse Position-specific BLAST (RPS-BLAST)

RPS-BLAST (Chapter 16) is a variant of the popular Position-specific Iterated BLAST (PSI-BLAST) program. PSI-BLAST finds sequences similar to the query and uses the resulting alignments to build a PSSM for the query. With this PSSM the database is scanned again to draw in more hits and further refine the scoring model. RPS-BLAST uses a query sequence to search a database of precalculated PSSMs and report significant hits in a single pass. The role of the PSSM has changed from "query" to "subject"; hence, the term "reverse" in RPS-BLAST. RPS-BLAST is the search tool used in the CD-Search service.

The CD Summary

Analogous to the Structure Summary page, the CD Summary page displays the available information about a given CD and offers various links for either viewing the CD alignment or initiating further searches (Figure 4). The CD Summary page can be retrieved by selecting the CD name on any page.

CD Records Curated at NCBI

In 2002, NCBI released the first group of curated CD records, a new and expanding set of annotated protein multiple sequence alignments and corresponding structure alignments. These new records have Accession numbers beginning with "cd" and have been added to the default CD-Search database. Most curated CD records are based on existing family descriptions from SMART and Pfam, but the alignments may have been revised extensively by quantitatively using three-dimensional structures and by re-examining the domain extent. In addition, CDD curators annotate conserved functional residues, ligands, and co-factors contained within the structures. They also record evidence for these sites as pointers to relevant literature or to three-dimensional structures exemplifying their properties. These annotations may be viewed using Cn3D and thus provide a direct way of visualizing functional properties of a protein domain in the context of its three-dimensional structure. (See Box 3 and Figure 7.)

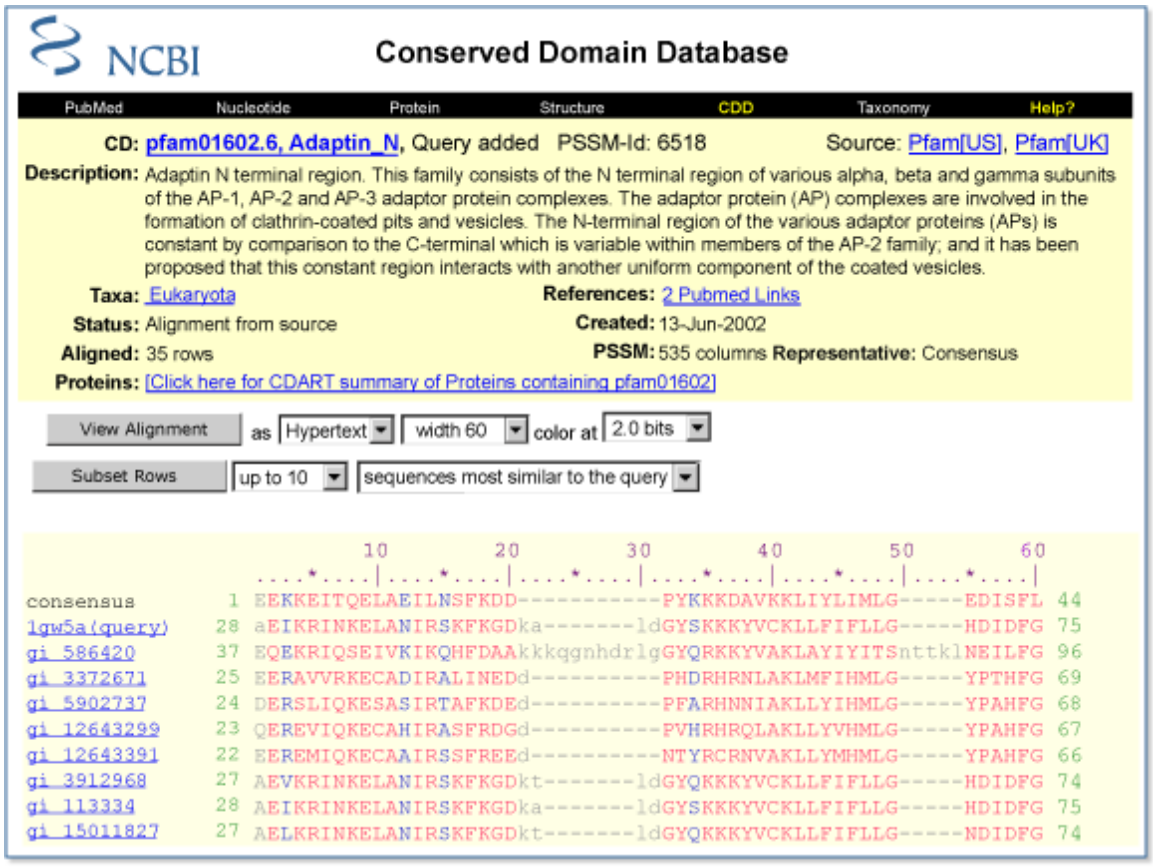


Figure 4. CD summary page. The *top* of the page serves as a header and reports a variety of identifying information, including the name and description of the CD, other related CDs with links to their summary pages, as well as the source database, status, and creation date of the CD. A taxonomic node link (*Taxa:*) launches the Taxonomy Browser, whereas a Proteins link (*Proteins:*) uses CDART to show other proteins that contain the CD. *Below* the header is the interface for viewing the CD alignment, which can be done either graphically with Cn3D (if the CD contains a sequence with structural data) or in HTML, text, or mFASTA format. It is also possible to view a selected number of the top-listed sequences, sequences from the most diverse members, or sequences most similar to the query. In addition, users may now select sequences with the NCBI Taxonomy Common Tree tool. The *lower portion* of the page contains the alignment itself. Members with a structural record in MMDB are listed first, and the identifier of each sequence links to the corresponding record.

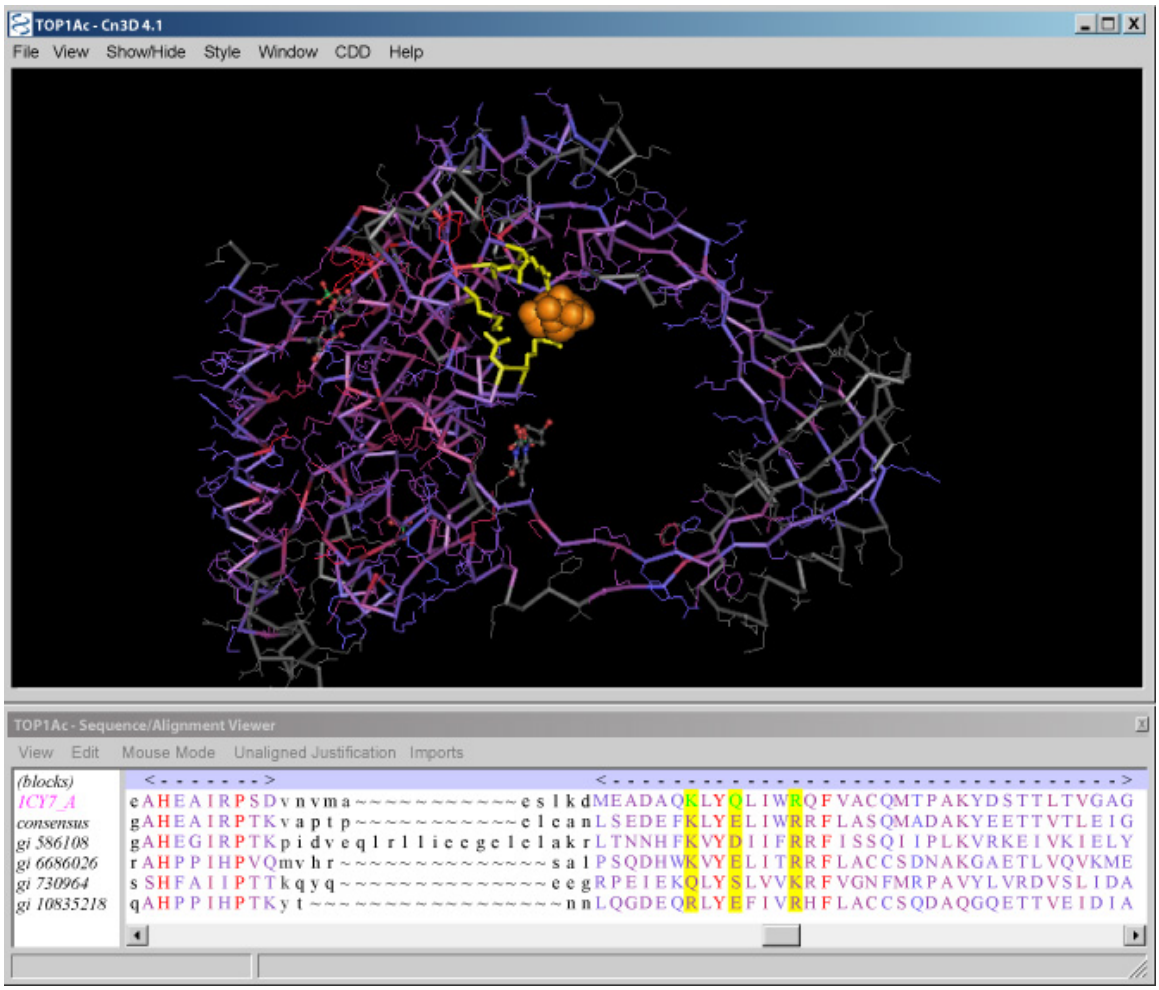


Figure 7. Sequence and structure views of the TOP1Ac conserved domain common to type III bacterial and eukaryotic DNA topoisomerases. The *upper window* displays the structure of the domain with the residues colored according to their sequence conservation, with *red* indicating high conservation and *blue* indicating low conservation. The nucleotide bound at site II is shown as an *orange* space-filling model, and the residues involved in this binding site are *yellow*. The *lower window* displays the sequence alignment for the domain with aligned residues shown as colored *capital letters*. Residues aligned to three of the binding site residues are highlighted in *yellow*. The sequence for NP_004609 (gi 10835218) occupies the *bottom row*.

Box 3. Example query: finding and viewing CDs in a protein.

Finding CDs in a Protein

Suppose that we are interested in topoisomerase enzymes and would like to find human topoisomerases that most closely resemble those found in eubacteria and thus may share a common ancestor. Further suppose that through a colleague, we are aware of a recent and particularly interesting crystal structure of a topoisomerase from *Escherichia coli* with PDB code 1I7D. How can we identify the conserved functional domains in this protein

Box 3 continues on next page...

Box 3 continued from previous page.

and then find human proteins with the same domains? From the Structure main page, we enter the PDB code 1I7D in the Structure **Summary** search box and quickly find the Structure Summary page for this record. We see that in this crystal structure, the protein is complexed with a single-stranded oligonucleotide. We also see that the protein has five 3D Domains. Two CDs align to the sequence as well, and they overlap with one another at the N-terminus of the protein in the region corresponding to the first 3D domain.

Analyzing CDs Found in a Protein

The Struture summary page displays only the CDs that give the best match to the protein sequence. To see all of the matching CDs, we can easily perform a full CD-Search. Select the **Protein** link to the left of the graphic to reveal the flatfile for the record. Then follow the **Domains** link in the **Link** menu on the right to view the results of the CD-Search. Select **Show Details** to see all CDs matching the query sequence. We find that nine CDs match this sequence, and that the statistics of each match are shown below the alignment graphic. The CD with the best hit is TopA from the COG database, and it is further clear that this domain consists of two smaller domains: TOPRIM (alignments from Pfam, SMART, and curated CD) and a topoisomerase domain (alignments from Pfam and curated CD). We can learn more about these CDs by studying the pairwise alignments at the bottom of the page and by studying their CD Summary pages, reached by selecting the links to their left.

Finding Other Proteins with Similar Domain Architecture

We now would like to find human proteins that have these same CDs. To perform a CDART search, simply select the **Show Domain Relatives** button. To limit these results to human proteins, we select the **Subset by Taxonomy** button. A taxonomic tree is then displayed, and we next check the box for **Mammal**, the lowest taxa including *Homo sapiens*. Selecting **Choose** then displays a Common Tree, and by clicking on the appropriate "scissor" icons, we can cut away all branches except the one leading to *H. sapiens*. We can execute this taxonomic restriction by selecting **Go back**, and we now find a much shorter list of CDART results. In the most similar group, we find two members, one of which is NP_004609. Selecting the **more>** link for this record shows the CD-Search results for this human protein. Interestingly, we find that the topoisomerase is very well conserved, whereas only a portion of the TOPRIM domain has been retained.

Viewing a CD Alignment with a 3D Structure

We now would like to view the alignment of the topoisomerase in the human protein to other members of this CD. On the CD-Search page, select the colored bar of this CD to see a CD-Browser window displaying the alignment. Because this is a curated CD record, we are able to view functional features of the protein domain on a structural template. The rightmost menu in the View Alignment bar shows the available features for this domain,

Box 3 continues on next page...

Box 3 continued from previous page.

whereas the topmost row in the alignment itself marks the residues involved in this feature with # symbols. The second row of the alignment is the consensus sequence of the CD record, whereas the third row contains the NP_004609 sequence, labeled "query". At the bottom of the page, buttons allow Cn3D to be launched with various structural features highlighted. For example, if we are interested in nucleotide binding site II, Cn3D will launch with the view depicted in Figure 7, showing the bound nucleotide in orange. Additonal Cn3D windows not shown in Figure 7 allow one to highlight the binding site residues yellow as shown, and these highlights also appear in the sequence window. In this figure, the NP_004609 sequence has been merged into the alignment (bottom row) using tools within Cn3D, and the result shows that this human protein closely conserves these important functional residues.

The Distinction between 3D Domains and CDs

The term "domain" refers in general to a distinct functional and/or structural unit of a protein. Each polypeptide chain in MMDB is analyzed for the presence of two classes of domains, and it is important for users to understand the difference between them. One class, called 3D Domains, is based solely on similar, compact substructures, whereas the second class, called Conserved Domains (CDs), is based solely on conserved sequence motifs. These two classifications often agree, because the compact substructures within a protein often correspond to domains joined by recombination in the evolutionary history of a protein. Note that CD links can be identified even when no 3D structures within a family are known. Moreover, 3D Domain links may also indicate relationships either to structures not included in CDD entries or to structures so distantly related that no significant similarity can be found by sequence comparisons.

Finding and Viewing Structures

For an example query on finding and viewing structures, see Box 2.

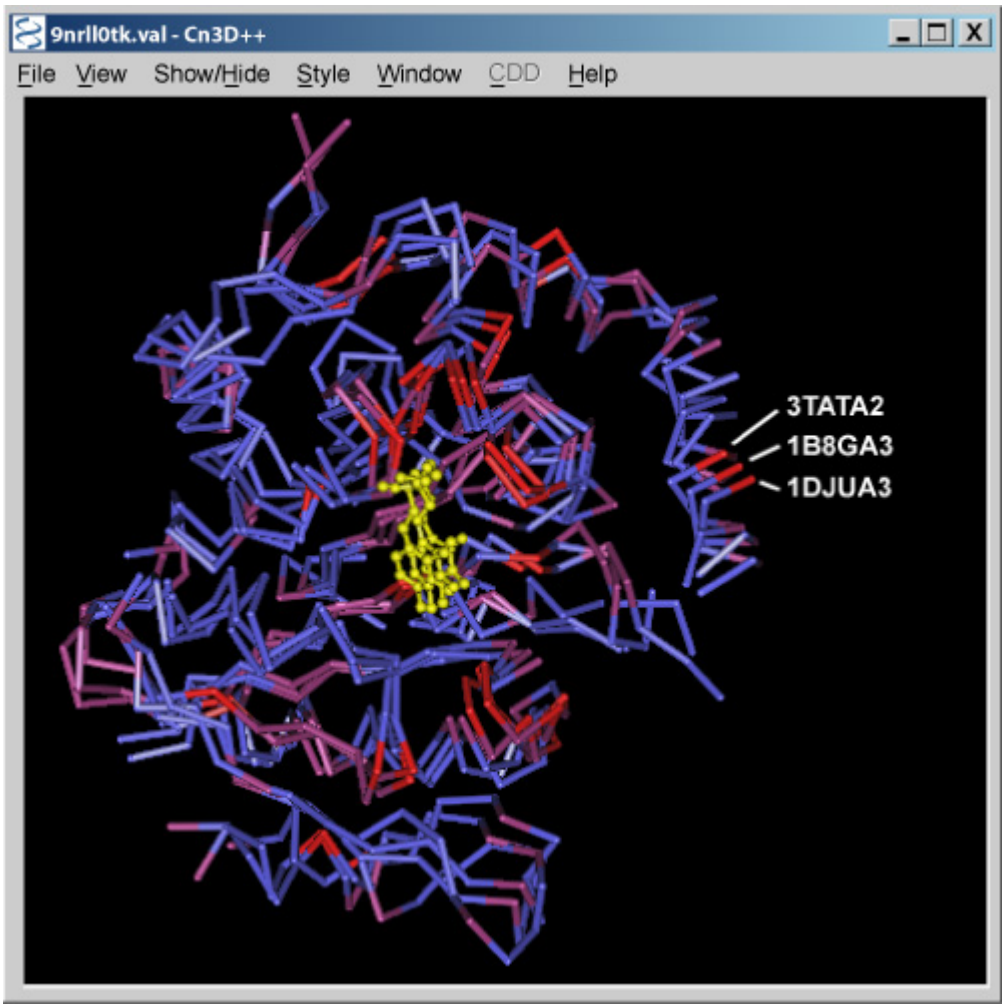


Figure 6. VAST structural alignment of 1B8GA3, 3TATA2, and 1DJUA3. The backbone atoms of the aligned residues of the three structures are shown colored according to their sequence conservation of each position in the alignment. Highly conserved positions are colored more *red*, whereas poorly conserved positions are colored more *blue*. The bound pyridoxal phosphate ligands are *yellow*.

Box 2. Example query: finding and viewing structural data of a protein.

Finding the Structure of a Protein

Suppose that we are interested in the biosynthesis of aminocyclopropanes and would like to find structural information on important active site residues in any available aminocyclopropane synthases. To begin, we would go to the **Structure** main page and enter "aminocyclopropane synthase" in the **Search** box. Pressing Enter displays a short list of structures, one of which is 1B8G, 1-aminocyclopropane-1-carboxylate synthase. Perhaps we would like to know the species from which this protein was derived. Selecting the **Taxonomy** link to the right shows that this protein was derived from *Malux x*

Box 2 continues on next page...

Box 2 continued from previous page.

domestica, or the common apple tree. Going back to the Entrez results page and selecting the PDB code (1B8G) opens the Structure Summary page for this record. The species is again displayed on this page, along with a link to the *Journal of Molecular Biology* article describing how the structure was determined. We immediately see from this page that this protein appears as a dimer in the structure, with each chain having three 3D domains, as identified by VAST. In addition, CD-Search has identified an "aminotran_1_2" CD in each chain. Now we are ready to view the 3D structure.

Viewing the 3D Structure

Once we have found the Structure Summary page, viewing the 3D structure is straightforward. To view the structure in Cn3D, we simply select the **View 3D Structure** button. The default view is to show helices in green, strands in brown, and loops in blue. This color scheme is also reflected in the Sequence/Alignment Viewer.

Locating an Active Site

Upon inspecting the structure, we immediately notice that a small molecule is bound to the protein, likely at the active site of the enzyme. How do we find out what that molecule is? One easy way is to return to the Structure Summary page and select the link to the PDB code, which takes us to the PDB Structure Explorer page for 1B8G. Quickly, we see that pyridoxal-5′-phosphate (PLP) is a HET group, or heterogen, in the structure. Our interest piqued, we would now like to know more about the structural domain containing the active site. Returning to Cn3D, we manipulate the structure so that PLP is easily visible and then use the mouse to double-click on any PLP atom. The molecule becomes selected and turns yellow. Now from the **Show/Hide** menu, we choose **Select by distance** and **Residues only** and enter 5 Angstroms for a search radius. Scanning the Sequence/Alignment Viewer, we see that seven residues are now highlighted: 117-119, 230, 268, 270, and 279. Glancing at the 3D Domain display in the Structure Summary page, we note that all of these residues lie in domain 3. We now focus our attention on this domain.

Viewing Structure Neighbors of a 3D Domain

Given that this enzyme is a dimer, we arbitrarily choose domain 3 in chain A, the accession of which is thus 1B8GA3. By clicking on the 3D Domain bar at a point within domain 3, we are taken to the VAST Structure Neighbors page for this domain, where we find nearly 200 structure neighbors.

Restricting the Search by Taxonomy

Perhaps we would now like to identify some of the most evolutionarily distant structure neighbors of domain 1B8GA3 as a means of finding conserved residues that may be associated with its binding and/or catalytic function. One powerful way of doing this is to

Box 2 continues on next page...

Box 2 continued from previous page.

choose structure neighbors from phylogenetically distant organisms. We therefore need to combine our present search with a Taxonomy search. Given that 1B8G is derived from the superkingdom Eukaryota, we would like to find structure neighbors in other superkingdom taxa, such as Eubacteria and Archaea. Returning to the Structure Summary page, select the 3D Domains link in the graphic display to open the list of 3D Domains in Entrez. Finding 1B8GA3 in the list, selecting the **Related 3D Domains** link shows a list of all the structure neighbors of this domain. From this page, we select **Preview/Index**, which shows our recent queries. Suppose our set of related 3D Domains is #5. We then perform two searches:

```
1. #5 AND "Archaea"[Organism]
2. #5 AND "Eubacteria"[Organism]
```

Looking at the Archaea results, we find among them 1DJUA3, a domain from an aromatic aminotransferase from *Pyrococcus horikoshii*. Concerning the Eubacteria results, we find among the several hundred matching domains 3TATA2, a tyrosine aminotransferase from *Escherichia coli*.

Viewing a 3D Superposition of Active Sites

Returning to the VAST Structure Neighbors page for 1B8GA3, we want to select 1DJUA3 and 3TATA2 to display in a structural alignment. One way to do this is to enter these two Accession numbers in the **Find** box and press **Find**. We now see only these two neighbors, and we can select **View 3D Structure** to launch Cn3D.

Cn3D again displays the aligned residues in red, and we can highlight these further by selecting **Show aligned residues** from the **Show/Hide** menu. The excellent agreement between both the active site structures and the conformations of the bound ligands is readily apparent. Furthermore, by selecting **Style/Coloring Shortcuts/Sequence Conservation/Variety**, we can easily see that the most highly conserved residues are concentrated near the binding site (Figure 6).

Why Would I Want to Do This?

- To determine the overall shape and size of a protein
- To locate a residue of interest in the overall structure
- To locate residues in close proximity to a residue of interest
- To develop or test chemical hypotheses regarding an enzyme mechanism
- To locate or predict possible binding sites of a ligand
- To interpret mutation studies
- To find areas of positive or negative charge on the protein surface
- To locate particularly hydrophobic or hydrophilic regions of a protein

- To infer the 3D structure and related properties of a protein with unknown structure from the structure of a homologous protein
- To study evolutionary processes at the level of molecular structure
- To study the function of a protein
- To study the molecular basis of disease and design novel treatments

How to Begin

The first step to any structural analysis at NCBI is to find the structure records for the protein of interest or for proteins similar to it. One may search MMDB directly by entering search terms such as PDB code, protein name, author, or journal in the Entrez Structure **Search** box on the Structure homepage. Alternative points of entry are shown below.

By using the full array of Entrez search tools, the resulting list of MMDB records can be honed, ideally, to a workable list from which a record can be selected. Users should note that multiple records may exist for a given protein, reflecting different experimental techniques, conditions, and the presence or absence of various ligands or metal ions. Records may also contain different fragments of the full-length molecule. In addition, many structures of mutant proteins are also available. The PDB record for a given structure generally contains some description of the experimental conditions under which the structure was determined, and this file can be accessed by selecting the PDB code link at the top of the Structure Summary page.

Alternative Points of Entry

Structure Summary pages can also be found from the following NCBI databases and tools:

- Select the Structure **links** to the right of any Entrez record found; records with Structure links can also be located by choosing **Structure links** from the **Display** pull-down menu.
- Select the **Related Sequences** link to the right of an Entrez record to find proteins related by sequence similarity and then select **Structure links** in the **Display** pull-down menu.
- Choose the PDB database from a blastp (protein-protein BLAST) search; only sequences with structure records will be retrieved by BLAST. The **Related Structures** link provides 3D views in Cn3D.
- Select the **3D Structures** button on any BLink report to show those BLAST hits for which structural data are available.
- From the results of any protein BLAST search, click on a red 'S' linkout to view the sequence alignment with a structure record.

Viewing 3D Structures

3D Domains

The 3D domains of a protein are displayed on the Structure Summary page. It is useful to know how many 3D domains a protein contains and whether they are continuous in sequence when viewing the full 3D structure of the molecule.

Secondary Structure

Knowing the secondary structure of a protein can also be a useful prelude to viewing the 3D structure of the molecule. The secondary structure can be viewed easily by first selecting the **Protein** link to the left of the desired chain in the graphic display. Finding oneself in Entrez Protein, selecting **Graphics** in the Display pull-down menu presents secondary structure diagrams for the molecule.

Full Protein Structures

Cn3D is a software package for displaying 3D structures of proteins. Once it has been installed and the Internet browser has been configured correctly, simply selecting the **View 3D Structure** button on a Structure Summary page launches the application. Once the structure is loaded, a user can manipulate and annotate it using an array of options as described in the Cn3D Tutorial. By default, Cn3D colors the structure according to the secondary structure elements. However, another useful view is to color the protein by domain (see **Style** menu options), using the same color scheme as is shown in the graphic display on the Structure Summary page. These color changes also affect the residues displayed in the Sequence/Alignment Viewer, allowing the identification of domain or secondary structure elements in the primary sequence. In addition to Cn3D, users can also display 3D structures with RasMol or Mage. Structures can also be saved locally as an ASN.1, PDB, or Mage file (depending on the choice of structure viewer) for later display.

Finding and Viewing Structure Neighbors

For an example query on finding and viewing structure neighbors, see Box 2.

Why Would I Want to Do This?

- To determine structurally conserved regions in a protein family
- To locate the structural equivalent of a residue of interest in another related protein
- To gain insights into the allowable structural variability in a particular protein family
- To develop or test chemical hypotheses regarding an enzyme mechanism
- To predict possible binding sites of a ligand from the location of a binding site in a related protein
- To identify sites where conformational changes are concentrated
- To interpret mutation studies
- To find areas of conserved positive or negative charge on the protein surface

- To locate conserved hydrophobic or hydrophilic regions of a protein
- To identify evolutionary relationships across protein families
- To identify functionally equivalent proteins with little or no sequence conservation

How to Begin

The Vector Alignment Search Tool (VAST) is used to calculate similar structures on each protein contained in the MMDB. The graphic display on each Structure Summary page (Figure 2) links directly to the relevant VAST results for both whole proteins and 3D domains:

- The 3D Domains link transfers the user to Entrez 3D Domains, showing a list of the VAST neighbors.
- Selecting the chain bar displays the VAST Structure Neighbors page for the entire chain.
- Selecting a 3D Domain bar displays the VAST Structure Neighbors page for the selected domain.

Alternative Points of Entry

- From any Entrez search, select **Related 3D Domains** to the right of any record found to view the Vast Structure Neighbors page.

Viewing a 2D Alignment of Structure Neighbors

A graphic 2D HTML alignment of VAST neighbors can be viewed as follows:

- On the lower portion of the VAST Structure Neighbors page (Figure 3), select the desired neighbors to view by checking the boxes to their left.
- On the **View/Save** bar, configure the pull-down menus to the right of the **View Alignment** button.
- Select **View Alignment**.

Viewing a 3D Alignment of Structure Neighbors

Alignments of VAST structure neighbors can be viewed as a 3D image using Cn3D.

- On the lower portion of the VAST Structure Neighbors page (Figure 3), select the desired neighbors to view by checking the boxes to their left.
- On the **View/Save** bar, configure the pull-down menus to the right of the **View 3D Structure** button.
- Select **View 3D Structure**.

Cn3D automatically launches and displays the aligned structures. Each displayed chain has a unique color; however, the portions of the structures involved in the alignment are shown in red. These same colors are also reflected in the Sequence/Alignment Viewer. Among the many viewing options provided by Cn3D, of particular use is the **Show/Hide**

menu that allows only the aligned residues to be viewed, only the aligned domains, or all residues of each chain.

Finding and Viewing Conserved Domains

For an example query on finding and viewing conserved domains, see Box 3.

Why Would I Want to Do This?

- To locate functional domains within a protein
- To predict the function of a protein whose function is unknown
- To establish evolutionary relationships across protein families
- To interpret mutation studies
- To predict the structure of a protein of unknown structure

How to Begin

Following the Domains link for any protein in Entrez, one can find the conserved domains within that protein. The CD-Search (or Protein BLAST, with CD-Search option selected) can be used to find conserved domains (CDs) within a protein. Either the Accession number, gi number, or the FASTA sequence can be used as a query.

Alternative Points of Entry

Information on the CDs contained within a protein can also be found from these databases and tools:

- From any Entrez search: select the **Domains** link to the right of a displayed record.
- From the Structure Summary page of a MMDB record: this page displays the CDs within each protein chain immediately below the 3D Domain bar in the graphic display. Selecting the **CDs** link shows the CD-Search results page.
- From an Entrez Domains search: choose **Domains** from the Entrez **Search** pull-down menu and enter a search term to retrieve a list of CDs. Clicking on any resulting CD displays the CD Summary page. To find the location of this CD in an aligned protein, select the CD link following a protein name in the bottom portion of this page.
- From the CDD page: locate CDs by entering text terms into the search box and proceed as for an Entrez CD search.
- From a BLink report: select the **CDD-Search** button to display the CD-Search results page.
- From the BLAST main page: follow the RPS-BLAST link to load the CD-Search page.

Viewing Conserved Domains

Results from a CD search are displayed as colored bars underneath a sequence ruler. Moving the mouse over these bars reveals the identity of each domain; domains are also listed in a format similar to BLAST summary output (Chapter 16). Pairwise alignments between the matched region of the target protein and the representative sequence of each domain are shown below the bar. Red letters indicate residues identical to those in the representative sequence, whereas blue letters indicate residues with a positive BLOSUM62 score in the BLAST alignment.

Viewing Multiple Alignments of a Query Protein with Members of a Conserved Domain

These can be displayed by clicking a CD bar within a MMDB Structure Summary page or from a hyperlinked CD name on a CD-Search results page.

Viewing CD Alignments in the Context of 3D Structure

If members of a CD have MMDB records, one of these records can be viewed as a 3D image along with the sequence alignment using Cn3D (launched by selecting the pink dot on a CD-Search results page). As in other alignment views, colored capital letters indicate aligned residues, allowing the sequence of the protein sequence of interest to be mapped onto the available 3D structure.

Finding and Viewing Proteins with Similar Domain Architectures

For an example query on finding and viewing proteins with similar domain architectures, see Box 3.

Why Would I Want to Do This?

- To locate related functional domains in other protein families
- To gain insights into how a given CD is situated within a protein relative to other CDs
- To explore functional links between different CDs
- To predict the function of a protein whose function is unknown
- To establish evolutionary relationships across protein families

How to Begin

Following the **Domain Relatives** link for any protein in Entrez, one can find other proteins with similar domain architecture. The Conserved Domain Architecture Retrieval Tool (CDART) can take an Accession number or the FASTA sequence as a query to find out the domain architecture of a protein sequence and list other proteins with related domain architectures.

Alternative Points of Entry

- From a CD-Search results page, click **Show Domain Relatives**
- From a CD-Summary page, click the **Proteins** link
- From an Entrez Domains searc, click the **Proteins** link in the Links menu

Results of a CDART Search

These are described in Figure 5. The protein "hits", which have similar domain architectures to the query sequence, can be further refined by taxonomic group, in which the results can be limited to selected nodes of the taxonomic tree. Furthermore, search results may be limited to those that contain only particular conserved domains.

Links Between Structure and Other Resources

Integration with Other NCBI Resources

As illustrated in the sections above, there are numerous connections between the Structure resources and other databases and tools available at the NCBI. What follows is a listing of major tools that support connections.

Entrez

Because Entrez is an integrated database system (Chapter 15), the links attached to each structure give immediate access to PubMed, Protein, Nucleotide, 3D Domain, CDD, or Taxonomy records.

BLAST

Although the BLAST service is designed to find matches based solely on sequence, the sequences of Structure records are included in the BLAST databases, and by selecting the PDB search database, BLAST searches only the protein sequences provided by MMDB records. A new **Related Structure** link provides 3D views for sequences with structure data identified in a BLAST search.

BLink

The BLink report represents a precomputed list of similar proteins for many proteins (see, for example, links from Entrez Gene records; Chapter 19). The **3D Structures** option on any BLink report shows the BLAST hits that have 3D structure data in MMDB, whereas the **CDD-Search** button displays the CD-Search results page for the query protein.

Microbial Genomes

A particularly useful interface with the structural databases is provided on the Microbial Genomes page (10). To the left of the list of genomes are several hyperlinks, two of which offer users direct access to structural information. The red **[D]** link displays a listing of every protein in the genome, each with a link to a BLink page showing the results of a

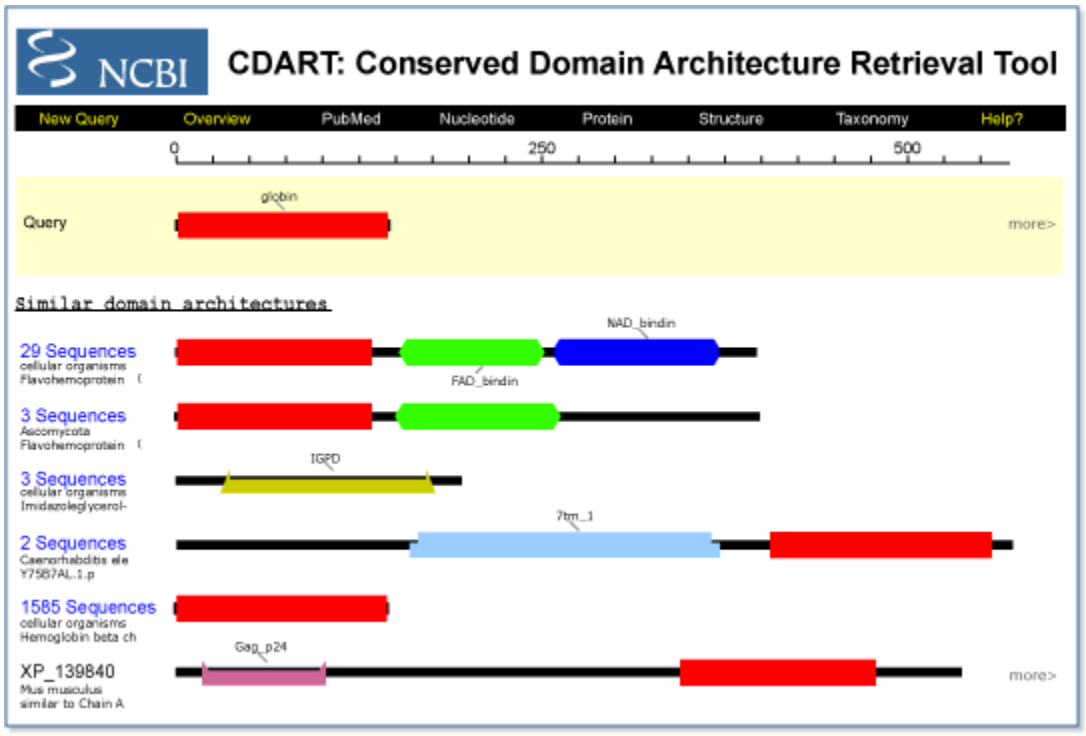


Figure 5. A CDART results page. At the *top* of the CDART results page in a *yellow box*, the query sequence CDs are represented as "beads on a string". Each CD had a unique color and shape and is labeled both in the display itself and in a legend located at the *bottom* of the page. The shapes representing CDs are hyperlinked to the corresponding CD summary page. The matching proteins to the query are listed *below* the yellow box, ranked according to the number of non-redundant hits to the domains in the query sequence. Each match is either a single protein, in which case its Accession number is shown, or is a cluster of very similar proteins, in which case the number of members in the cluster is shown. Cluster members can be displayed by selecting the logo to the *left* of its diagram. Selecting any protein Accession number displays the flatfile for that protein. To the *right* of any drawing for a single protein (either on the main results page or after expanding a protein cluster) is a **more>** link, which displays the CD-Search results page for the selected protein so that the sequence alignment, e.g., of a CDART hit with a CD contained in the original protein of interest, can be examined.

BLAST pdb search for that protein. The **[S]** link displays a similar protein list for the selected genome, but now with a listing of the conserved domains found in each protein by a CD-Search.

Links to Non-NCBI Resources

The Protein Data Bank (PDB)

As stated elsewhere, all records in the MMDB are obtained originally from the Protein Data Bank (PDB) (6). Links to the original PDB records are located on the Structure Summary page of each MMDB record. Updates of the MMDB with new PDB records occur once a month.

Pfam and SMART

The CDD staff imports CD collections from both the Pfam and SMART databases. Links to the original records in these databases are located on the appropriate CD Summary page. Both Pfam and SMART are updated several times per year in roughly bimonthly intervals, and the CDD staff update CDD accordingly.

Saving Output from Database Searches

Exporting Graphics Files from Cn3D

Structures displayed in Cn3D can be exported as a Portable Network Graphics (PNG) file from within Cn3D (the Export PNG command in the **File** menu). The structure file itself, in the orientation currently being viewed, can also be saved for later launching in Cn3D.

Saving Individual MMDB Records

Individual MMDB records can be saved/downloaded to a local computer directly from the Structure Summary page for that record. **Save File** in the **View** bar downloads the file in a choice of three formats: ASN.1 (select **Cn3D**); PDB (select **RasMol**); or Mage (select **Mage**).

Saving VAST Alignments

Alignments of VAST neighbors can be saved/downloaded from the VAST Structure Neighbors page of any MMDB record. By selecting options in the **View Alignment** pull-down menu, the alignment data can be saved, formatted as HTML, text, or mFASTA, and then saved.

FTP

MMDB

Users can download the NCBI Structure databases from the NCBI FTP site: ftp://ftp.ncbi.nih.gov/mmdb. A Readme file contains descriptions of the contents and information about recent updates. Within the mmdb directory are four subdirectories that contain the following data:

- mmdbdata: the current MMDB database (NOTE: these files can not be read directly by Cn3D).
- vastdata: the current set of VAST neighbor annotations to MMDB records
- nrtable: the current non-redundant PDB database
- pdbeast: table listing the taxonomic classification of MMDB records

CDD

CDD data can be downloaded from ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd. A Readme file contains descriptions of the data archives. Users can download the PSSMs for each CD record, the sequence alignments in mFASTA format, or a text file containing the accessions and descriptions of all CD records.

Frequently Asked Questions

- Cn3D
- VAST searches
- CDD

References

1. Wang Y, Anderson JB, Chen J, Geer LY, He S, Hurwitz DI, Liebert CA, Madej T, Marchler GH, Marchler-Bauer A, Panchenko AR, Shoemaker BA, Song JS, Thiessen PA, Yamashita RA, Bryant SH. MMDB: Entrez's 3D-structure database. Nucleic Acids Res. 2002;30:249–252. PubMed PMID: 11752307.
2. Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH. Cn3D: sequence and structure views for Entrez. Trends Biochem Sci. 2000;25:300–302. PubMed PMID: 10838572.
3. Madej T, Gibrat J-F, Bryant SH. Threading a database of protein cores. Proteins. 1995;23:356–369. PubMed PMID: 8710828.
4. Gibrat J-F, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol. 1996;6:377–385. PubMed PMID: 8804824.
5. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 2002;30:281–283. PubMed PMID: 11752315.
6. Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm W, Weissig H, Greer DS, Bourne PE, Berman HM. The Protein Data Bank: unifying the archive. Nucleic Acids Res. 2002;30:245–248. PubMed PMID: 11752306.
7. Ohkawa H, Ostell J, Bryant S. MMDB: an ASN.1 specification for macromolecular structure. Proc Int Conf Intell Syst Mol Biol. 1995;3:259–267. PubMed PMID: 7584445.
8. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL. The Pfam proteins family database. Nucleic Acids Res. 2002;30:276–280. PubMed PMID: 11752314.
9. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. SMART: a Web-based tool for the study of genetically mobile domains. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res. 2002;30:242–244. PubMed PMID: 11752305.
10. Wang Y, Bryant S, Tatusov R, Tatusova T. Links from genome proteins to known 3D structures. Genome Res. 2000;10:1643–1647. PubMed PMID: 11042161.

Chapter 4. The Taxonomy Project

Scott Federhen

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The NCBI Taxonomy database is a curated set of names and classifications for all of the organisms that are represented in GenBank. When new sequences are submitted to GenBank, the submission is checked for new organism names, which are then classified and added to the Taxonomy database. As of April 2003, there were 176,890 total taxa represented.

There are two main tools for viewing the information in the Taxonomy database: the Taxonomy Browser, and Taxonomy Entrez. Both systems allow searching of the Taxonomy database for names, and both link to the relevant sequence data. However, the Taxonomy Browser provides a hierarchical view of the classification (the best display for most casual users interested in exploring our classification), whereas Entrez Taxonomy provides a uniform indexing, search, and retrieval engine with a common mechanism for linking between the Taxonomy and other relevant Entrez databases.

Introduction

Organismal taxonomy is a powerful organizing principle in the study of biological systems. Inheritance, homology by common descent, and the conservation of sequence and structure in the determination of function are all central ideas in biology that are directly related to the evolutionary history of any group of organisms. Because of this, taxonomy plays an important cross-linking role in many of the NCBI tools and databases.

The NCBI Taxonomy database is a curated set of names and classifications for all of the organisms that are represented in GenBank. When new sequences are submitted to GenBank, the submission is checked for new organism names, which are then classified and added to the taxonomy database. As of April 1, 2003, there were 4,653 families, 26,427 genera, 130,207 species, and 176,890 total taxa represented.

Of the several different ways to build a taxonomy, our group maintains a phylogenetic taxonomy. In a phylogenetic classification scheme, the structure of the taxonomic tree approximates the evolutionary relationships among the organisms included in the classification (the “tree of life”; see Figure 1).

Our classification represents an assimilation of information from many different sources (see Box 1). Much of the success of the project is attributable to the flood of new molecular data that has revolutionized our understanding of the phylogeny of many groups, especially of previous poorly understood groups such as Bacteria, Archea, and Fungi. Users should be aware that some parts of the classification are better developed

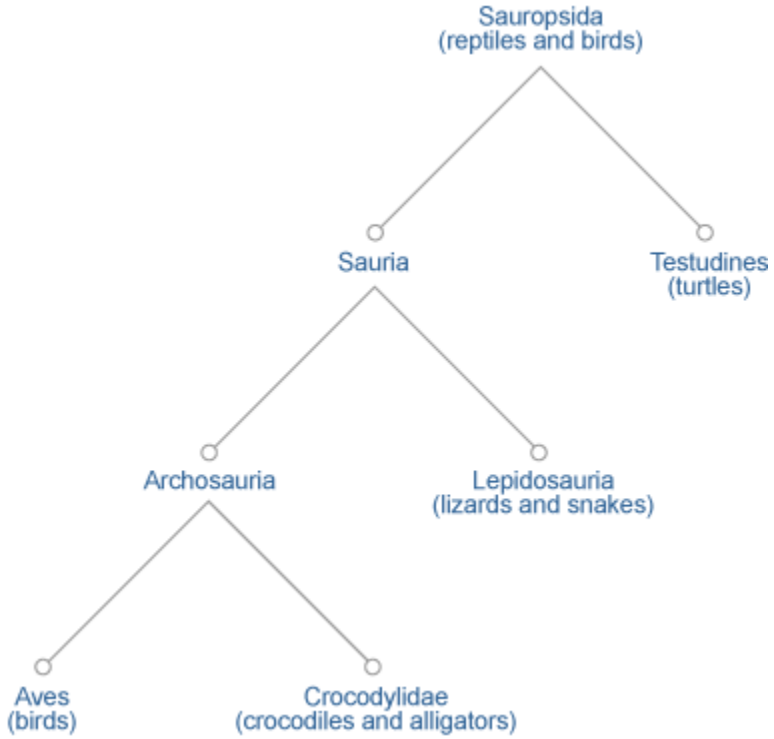


Figure 1. A phylogenetic classification scheme. If two organisms (A and B) are listed more closely together in the taxonomy than either is to organism C, the assertion is that C diverged from the lineage leading to A +B earlier in evolutionary history, and that A and B share a common ancestor that is not in the direct line of evolutionary descent to species C. For example, the current consensus it that the closest living relatives of the birds are the crocodiles; therefore, our classification does not include the familiar taxon Reptilia (turtles, lizards and snakes, and crocodiles), which excludes the birds, and would break the phylogenetic principle outlined above.

than others and that the primary systematic and phylogenetic literature is the most reliable information source.

We do not rely on sequence data alone to build our classification, and we do not perform phylogenetic analysis ourselves as part of the taxonomy project. Most of the organisms in GenBank are represented by only a snippet of sequence; therefore, sequence information alone is not enough to build a robust phylogeny. The vast majority of species are not there at all, although about 50% of the birds and the mammals are represented. We therefore also rely on analyses from morphological studies; the challenge of modern systematics is to unify molecular and morphological data to elucidate the evolutionary history of life on earth.

Box 1. History of the Taxonomy Project.

By the time the NCBI was created in 1988, the nucleotide sequence databases (GenBank, EMBL, and DDBJ) each maintained their own taxonomic classifications. All three classifications derived from the one developed at the Los Alamos National Lab (LANL) but had diverged considerably. Furthermore, the protein sequence databases (SWISS-PROT and PIR) each developed their own taxonomic classifications that were very different from each other and from the nucleotide database taxonomies. To add to the mix, in 1990 the NCBI and the NLM initiated a journal-scanning program to capture and annotate sequences reported in the literature that had not been submitted to any of the sequence databases. We, of course, began to assign our own taxonomic classifications for these records.

The Taxonomy Project started in 1991, in association with the launch of Entrez (Chapter 15). The goal was to combine the many taxonomies that existed at the time into a single classification that would span all of the organisms represented in any of the GenBank sources databases (Chapter 1).

To represent, manipulate, and store versions of each of the different database taxonomies, we wrote a stand-alone, tree-structured database manager, TaxMan. This also allowed us to merge the taxonomies into a single composite classification. The resulting hybrid was, at first, a bigger mess than any of the pieces had been, but it gave us a starting point that spanned all of the names in all of the sequence databases. For many years, we cleaned up and maintained the NCBI Taxonomy database with TaxMan.

After the initial unification and clean-up of the taxonomy for Entrez was complete, Mitch Sogin organized a workshop to give us advice on the clean-up and recommendations for the long-term maintenance of the taxonomy. This was held at the NCBI in 1993 and included: Mitch Sogin (protists), David Hillis (chordates), John Taylor (fungi), S.C. Jong (fungi), John Gunderson (protists), Russell Chapman (algae), Gary Olsen (bacteria), Michael Donoghue (plants), Ward Wheeler (invertebrates), Rodney Honeycutt (invertebrates), Jack Holt (bacteria), Eugene Koonin (viruses), Andrzej Elzanowski (PIR taxonomy), Lois Blaine (ATCC), and Scott Federhen (NCBI). Many of these attendees went on to serve as curators for different branches of the classification. In particular, David Hillis, John Taylor, and Gary Olsen put in long hours to help the project move along.

In 1995, as more demands were made on the Taxonomy database, the system was moved to a SyBase relational database (TAXON), originally developed by Tim Clark. Hierarchical organism indexing was added to the Nucleotide and Protein domains of Entrez, and the Taxonomy browser made its first appearance on the Web.

In 1997, the EMBL and DDBJ databases agreed to adopt the NCBI taxonomy as the standard classification for the nucleotide sequence databases. Before that, we would see new organism names from the EMBL and DDBJ only after their entries were released to

Box 1 continues on next page...

Box 1 continued from previous page.

the public, and any corrections (in spelling, or nomenclature, or classification) would have to be made after the fact. We now receive taxonomy consults on new names from the EMBL and DDBJ before the release of their entries, just as we do from our own GenBank indexers. SWISS-PROT has also recently (2001) agreed to use our Taxonomy database and send us taxonomy consults.

Adding to the Taxonomy Database

Currently, more than 100 new species are added to the database daily, and the rate is accelerating as sequence analysis becomes an ever more common component of systematic research and the taxonomic description of new species.

Sources of New Names

The EMBL and DDBJ databases, as well as GenBank, now use the NCBI Taxonomy as the standard classification for nucleotide sequences (see Box 1). Nearly all of the new species found in the Taxonomy database are via sequences submitted to one of these databases from species that are not yet represented. In these cases, the NCBI taxonomy group is consulted, and any problems with the nomenclature and classifications are resolved before the sequence entries are released to the public. We also receive consults for submissions that are not identified to the species level (e.g., "Hantavirus" or "Bacillus sp.") and for anything that looks confusing, incorrect, or incomplete to the database indexers. All consults include information on the problem organism names, source features, and publication titles (if any). The email addresses for the submitters are also included in case we need to contact them about the nomenclature, classification, or annotation of their entries.

The number and complexity of organisms in a submission can vary enormously. Many contain a single new name, others may include 100 species, all from the same familiar genus, whereas others may include 100 names (only half identified at the species level) from 100 genera (all of which are new to the Taxonomy database) without any other identifying information at all.

Some new organism names are found by software when the protein sequence databases (SWISS-PROT, PIR, and the PRF) are added to Entrez; because most of the entries in the protein databases have been derived from entries in the nucleotide database, this is a small number. The NCBI structure group may also find new names in the PDB protein structure database. Finally, because we made the Taxonomy database publicly accessible on the Web, we have had a steady stream of comments and corrections to our spellings and classification from outside users.

More on Submission

We often receive consults on submissions with explicitly new species names that will be published as part of the description of a new species. These sequence entries (like any other) may be designated "hold until published" (HUP) and will not be released until the corresponding journal article has been published. These species names will not appear on any of our taxonomy Web sites until the corresponding sequence entries have been released.

Occasionally, the same new genus name is proposed simultaneously for different taxa; in one case, two papers with conflicting new names had been submitted to the same journal, and both had gone through one round of review and revision without detection of the duplication. Although these duplications would have been discovered in time, the increasingly common practice of including some sequence analysis in the description of a new species can lead to earlier detection of these problems. In many cases, the new species name proposed in the submitted manuscript is changed during the editorial review process, and a different name appears in the publication. Submitters are encouraged to inform us when their new descriptions have been published, particularly if the proposed names have been changed.

We strongly encourage the submission of strain names for cultured bacteria, algae, and fungi and for sequences from laboratory animals in biochemical and genetic studies; of cultivar names for sequences from cultivated plants; and of specimen vouchers (something that definitively ties the sequence to its source) for sequences from phylogenetic studies. There are many other kinds of useful information that may be contained within the sequence submission, but these data are the bare minimum necessary to maintain a reliable link between an entry in the sequence database and the biological source material.

Using the Taxonomy Browser

The Taxonomy Browser (TaxBrowser) provides a hierarchical view of the classification from any particular place in the taxonomy. This is probably the display of choice for most casual users (browsers) of the taxonomy who are interested in exploring our classification. The TaxBrowser displays only the subset of taxa from the taxonomy database that is linked to public sequence entries. About 15% of the full Taxonomy database is not displayed on the public Web pages because the names are from sequence entries that have not yet been released.

TaxBrowser is updated continuously. New species will appear on a daily basis as the new names appear in sequence entries indexed during the daily release cycle of the Entrez databases. New taxa in the classification appear in TaxBrowser on an ongoing basis, as sections of the taxonomy already linked to public sequence entries are revised.

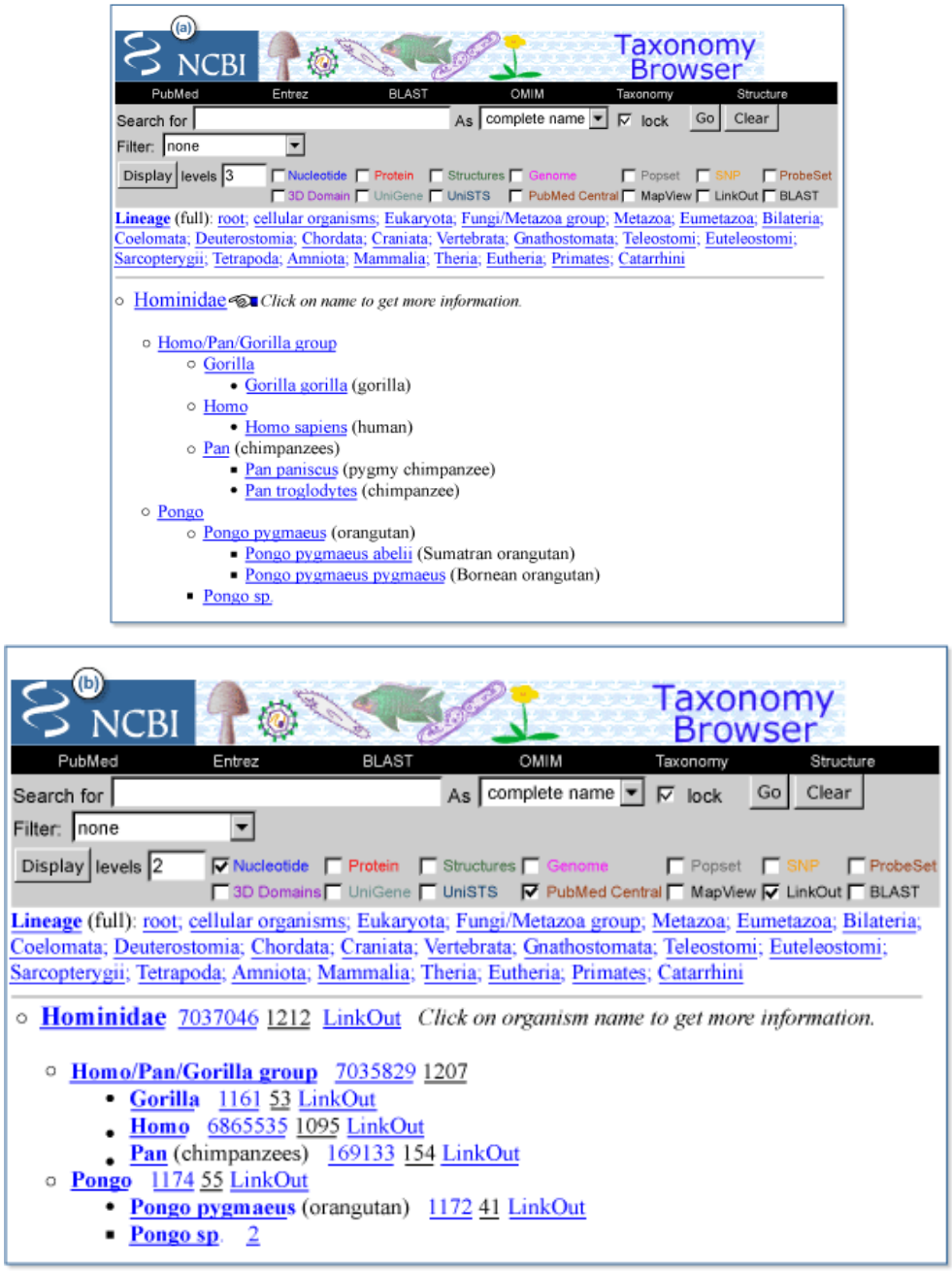


Figure 2. The Taxonomy Browser hierarchical display for the family Hominidae. (*a*) There are four genera listed in this family (Gorilla, Homo, Pan, and Pongo) with six species-level names (*Gorilla gorilla*, *Homo sapiens*, *Pan paniscus*, *Pan troglodytes*, *Pongo pygmaeus*, and *Pongo* sp.) and 2 subspecies. Common names are shown in *parentheses* if they are available in the Taxonomy database. The lineage above Hominidae is shown in the *line* at the *top* of the display; selecting the word **Lineage** will toggle back and forth between the abbreviated lineage (the display used in GenBank flatfiles) and the full lineage (as it appears in the Taxonomy database). Selecting any of the taxa *above* Hominidae (in the lineage) or *below* Hominidae (in the hierarchical display) will refocus the browser on that taxon instead of the Hominidae. Selecting Hominidae itself, however, will display the taxon-specific page for the Hominidae. (*b*) The default setting displays three levels of the classification on the hierarchy pages. To change this, enter a different number in the **Display levels** box and select **Accept**. If any of the check boxes to the *right* of the **Display levels** box are selected (i.e., *Nucleotide, Protein, Structures, ...*), the numbers of records in the corresponding Entrez database that are associated with that taxon will appear as hyperlinks. Selecting a link retrieves those records.

The Hierarchical Display

The browser produces two different kinds of Web pages: (1) hierarchy pages, which present a familiar indented flatfile view of the taxonomic classification, centered on a particular taxon in the database; and (2) taxon-specific pages, which summarize all of the information that we associate with any particular taxonomic entry in the database. For example, “hominidae” as a search term from the TaxBrowser homepage finds our human family (Figure 2).

The Taxon-specific Display

The taxon-specific browser display page shows all of the information that is associated with a particular taxon in the Taxonomy database and some information collected through links with related databases (Figure 3).

There are two sets of links to Entrez records from the Taxonomy Browser. The "subtree links" are accumulated up the tree in a hierarchical fashion; for example, there are 16 million nucleotide records and a half million protein records associated with the Chordata (Figure 3a). These are all linked into the taxonomy at or below the species leves and can be retrieved en masse via the subtree hotlink.

"Direct links" will retrieve Entrez records that are linked directly to this particular node in the taxonomy database. Many of the Entrez domains (e.g., sequences and structures) are linked into the taxonomy at or below the species level; it is a data error when a sequence entry is directly linked into the taxonomy at a taxon somewhere above the species level. For other Entrez domains (e.g., literature and phylogenetic sets), this is not the case. A journal article may talk about several different species but may also refer directly to the phylum Chordata. We have searched the full text of the articles in the PubMed Central archive with the scientific names from the taxonomy database. Twenty-seven articles in PubMed Central refer directly to the phylum Chordata; 9,299 articles are linked into the taxonomy somewhere in the Chordata subtree. The PopSet domain contains population studies, phylogenetic sets, and alignments. We have recently changed the way that we index phylogenetic sets in Entrez. The five "Direct links" at the Chordata will retrieve the phylogenetic sets that explicitly span the Chordata; the "Subtree links" will also include phylogenetic sets that are completely contained within the Chordata.

The taxon-specific browser pages now also show the NCBI LinkOut links to external resources. These include links to a broad range of different kinds of resources and are provided for the convenience of our users; the NCBI does not vouch for the content of these resources, although we do make an effort to ensure that they are of good scientific quality. A complete list of external resources can be found here. Groups interested in participating in the LinkOut program should visit the LinkOut homepage.

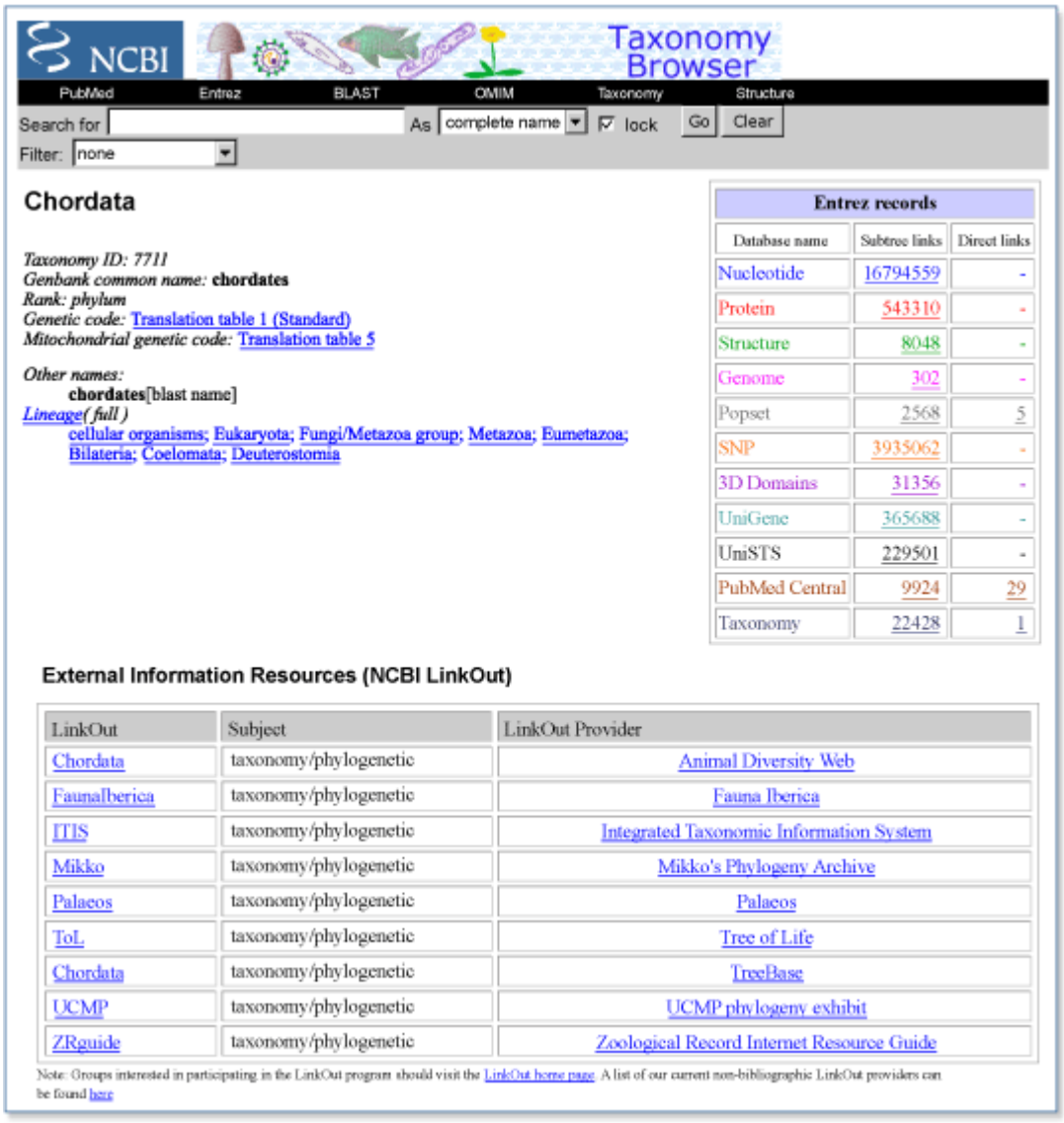


Figure 3. The Taxonomy Browser taxon-specific display. The name, taxid, rank, genetic codes, and other names (if any) associated with this taxon are all listed. The full lineage is shown; selecting the word **Lineage** toggles between the abbreviated and full versions. There may also be citation information and comments hyperlinked to the appropriate sources. The numbers of Entrez database records that link to this Taxonomy record are displayed and can be retrieved via the hotlinked entry counts. LinkOut links to external resources appear at the bottom of the page.

Search Options

There are several different ways to search for names in the Taxonomy database. If the search results in a terminal node in our taxonomy, the taxon-specific browser page is displayed; if the search returns with an internal (non-terminal) node, the hierarchical classification page is displayed.

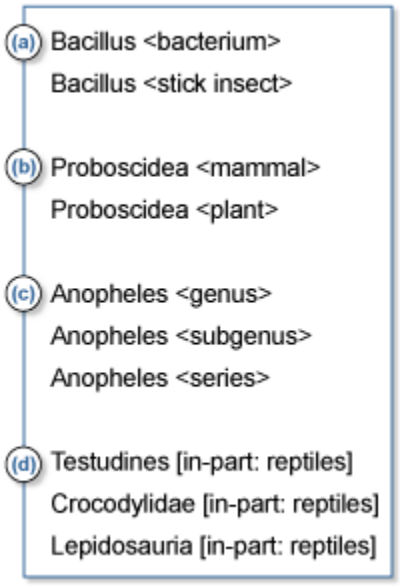


Figure 4. Examples of search results for duplicated names. Searches for *Bacillus* (*a*), *Proboscidea* (*b*), *Anopheles* (*c*), and reptiles (*d*) result in the options shown above. If the duplicated name is not the primary (scientific) name of the node [as in (*d*)], the primary name is given first, followed by the nametype and the duplicated name in *square brackets*.

Complete Name. By default, TaxBrowser looks for the complete name when a term is typed into the search box. It looks for a case-insensitive, full-length string match to all of the nametypes stored in the Taxonomy database. For example, *Homo sapiens, Escherichia, Tetrapoda,* and *Embryophyta* would all retrieve results.

Names can be duplicated in the Taxonomy database, but the taxonomy browser can only be focused on a single taxon at any one time. If a complete name search retrieves more than one entry from the taxonomy, an intermediate name selection screen appears (Figure 4). Each duplicated name includes a manually curated suffix that differentiates between the duplicated names.

Wild Card. This is a regular expression search, *, with wild cards. It is useful when the correct spelling of a scientific name is uncertain or to find ambiguous combinations for abbreviated species names. For example, *C* elegans* results in a list of 16 species and subspecies (Box 2). Note: there is still only one *H. sapiens*.

Token Set. This treats the search string as an unordered set of tokens, each of which must be found in one of the names associated with a particular node. For example, "sapiens" retrieves:

```
Homo sapiens
Homo sapiens neanderthalensis
```

Phonetic Name. This search qualifier can be used when the user has exhausted all other search options to find the organism of interest. The results using this function can be patchy, however. For example, "drozofila" and "kaynohrhabdietees" retrieve respectable results; however, "seenohrabdietees" and "eshereesheeya" are not found.

Taxonomy ID. This allows searching by the numerical unique identifier (taxid) of the NCBI Taxonomy database, e.g., 9606 or 666.

Box 2. TaxBrowser results from using the wild card search, C* elegans.

Cunninghamella elegans

Caenorhabditis elegans

Codonanthe elegans

Cyclamen coum subsp. elegans

Cestrum elegans

Chaerophyllum elegans

Chalara elegans

Chrysemys scripta elegans

Ceuthophilus elegans

Carpolepis elegans

Cylindrocladiella elegans

Coluria elegans

Cymbidium elegans

Coronilla elegans

Gymnothamnion elegans (synonym: *Callithamnion elegans*)

Centruroides elegans

How to Link to the TaxBrowser

There is a help page that describes how to make hyperlinks to the Taxonomy Browser pages.

The Taxonomy Database: TAXON

The NCBI Taxonomy database is stored as a SyBase relational database, called TAXON. The NCBI taxonomy group maintains the database with a customized software tool, the

Taxonomy Editor. Each entry in the database is a "taxon", also referred to as a "node" in the database. The "root node" (taxid1) is at the top of the hierarchy. The path from the root node to any other particular taxon in the database is called its "lineage"; the collection of all of the nodes beneath any particular taxon is called its "subtree". Each node in the database may be associated with several names, of several different nametypes. For indexing and retrieval purposes, the nametypes are essentially equivalent.

The Taxonomy database is populated with species names that have appeared in a sequence record from one of the nucleotide or protein databases. If a name has ever appeared in a sequence record at any time (even if it is not found in the current version of the record), we try to keep it in the Taxonomy database for tracking purposes (as a synonym, a misspelling, or other nametype), unless there are good reasons for removing it completely (for example, if it might cause a future submission to map to the wrong place in the taxonomy).

Taxids

Each taxon in the database has a unique identifier, its taxid. Taxids are assigned sequentially. When a taxon is deleted, its taxid disappears and is not reassigned (Table 1; see the FTP for a list of deleted taxids). When one taxon is merged with another taxon (e.g., if the names were determined to be synonyms or one was a misspelling), the taxid of the node that has disappeared is listed as a "secondary taxid" to the taxid of the node that remains (see the merged taxid file on the FTP site). In either case, the taxid that has disappeared will never be assigned to a new entry in the database.

Table 1. Files on the taxonomy FTP site.

File	Uncompresses to	Description
taxdump.tar.Z[a]	readme.txt	A terse description of the dmp files
	nodes.dmp	Structure of the database; lists each taxid with its parent taxid, rank, and other values associated with each node (genetic codes, etc.)
	names.dmp	Lists all the names associated with each taxid
	delnodes.dmp	Deleted taxid list
	merged.dmp	Merged nodes file
	division.dmp	GenBank division files
	gencode.dmp	Genetic codes files
	gc.prt	Print version of genetic codes
gi_taxid_nucl.dmp.gz	gi_taxid_nucl.dmp	A list of gi_taxid pairs for every live gi-identified sequence in the nucleotide sequence database

a For non-UNIX users, the file taxdmp.zip includes the same (zip compressed) data.

Table 1 continues on next page...

Table 1 continued from previous page.

File	Uncompresses to	Description
gi_taxid_prot.dmp.gz	gi_taxid_prot.dmp	A list of gi_taxid pairs for every live gi-identified sequence in the protein sequence database
gi_taxid_nucl_diff.dmp	gi_taxid_nucl_diff	List of differences between latest gi_taxid_nucl and previous listing
gi_taxid_prot_diff.dmp	gi_taxid_prot_diff	List of differences between latest gi_taxid_prot and previous listing

a For non-UNIX users, the file taxdmp.zip includes the same (zip compressed) data.

Nomenclature Issues

TAXON Nametypes

There are many possible types of names that can be associated with an organism taxid in TAXON. To track and display the names correctly, the various names associated with a taxid are tagged with a nametype, for example "scientific name", "synonym", or "common name". Each taxid **must have one** (and only one) scientific name but may have zero or many other names (for example, several synonyms, several common names, along with only one "GenBank common name").

When sequences are submitted to GenBank, usually only a scientific name is included; most other names are added by NCBI taxonomists at the time of submission or later, when further information is discovered. For a complete description of each nametype used in TAXON, see Appendix 1.

Classes of TAXON Scientific Names

Scientific names, the only required nametype for a taxid, can be further qualified into different classes. Not all "scientific names" that accompany sequence submissions are true Linnaean Latin binomial names; if the taxon is not identified to the species level, it is not possible to assign a binomial name to it. For indexing and retrieval purposes, TAXON needs to know whether the scientific name is a Latin binomial name, or otherwise. A full listing of the classes of TAXON scientific names can be viewed in Appendix 2.

Duplicated Names

The treatment of duplicated names was discussed briefly in the section on the Taxonomy browser. For our purposes, there are four main classes of duplicated scientific names: (1) real duplicate names, (2) structural duplicates, (3) polyphyletic genera, and (4) other duplicate names.

Real Duplicate Names

There are several main codes of nomenclature for living organisms: the Zoological Code (International Code of Zoological Nomenclature, ICZN; for animals), the Botanical Code

(International Code of Botanical Nomenclature, ICBN; for plants), the Bacteriological Code (International Code of Nomenclature of Bacteria, ICNB; for prokaryotes), and the Viral Code (International Code of Virus Classification and Nomenclature, ICVCN; for viruses). Within each code, names are required to be unique. When duplicate names are discovered within a code, one of them is changed (generally, the newer duplicate name). However, the codes are complex, and not all names are subject to these restrictions. For example, *Polyphaga* is both a genus of cockroaches and a suborder of beetles, and the damselfly genus *Lestoidea* is listed within the superfamily Lestoidea.

There is no real effort to make the scientific names of taxa unique among Codes, and among the relatively small set of names represented in the NCBI taxonomy database (20,000 genera), there are approximately 200 duplicate names (or about 1%), mostly at the genus level.

Early in 2002, the first duplicate species name was recorded in the Taxonomy database. *Agathis montana* is both a wasp and a conifer. In this case, we have used the full species names (with authorities) to provide unambiguous scientific names for the sequence entries. (The conifer is listed as *Agathis montana* de Laub; the wasp, *Agathis montana* Shest).

Structural Duplicates

In the Zoological and Bacteriological Codes, the subgenus that includes the type species is required to have the same name as the genus. This is a systematic source of duplicate names. For these duplicates, we use the associated rank in the unique name, e.g., *Drosophila* <genus> and *Drosophila* <subgenus>. Duplicated genera/subgenera also occur in the Botanical Codes, e.g. *Pinus* <genus> and *Pinus* <subgenus>.

Polyphyletic Genera

Certain genera, especially among the asexual forms of Ascomycota and Basidiomycota, are polyphyletic, i.e., they do not share a common ancestor. Pending taxonomic revisions that will transfer species assigned to "form" genera such as *Cryptococcus* to more natural genera, we have chosen to duplicate such polyphyletic genera in different branches of the Taxonomy database. This will maintain a phylogenetic classification and ensure that all species assigned to a polyphyletic genus can be retrieved when searching on the genus. Therefore, for example, the basidiomycete genus *Sporobolomyces* is represented in three different branches of the Basidiomycota: *Sporobolomyces* <Sporidiobolaceae>, *Sporobolomyces* <Agaricostilbomycetidae>, and *Sporobolomyces* <Erythrobasidium clade>.

Other Duplicate Names

We list many duplicate names in other nametypes (apart from our preferred "scientific name" for each taxon). Most of these are included for retrieval purposes, common names or the names of familiar paraphyletic taxa that we have not included in our classification, e.g. Osteichthyes, Coelenterata, and reptiles.

Other TAXON Data Types

Aside from names, there are several optional types of information that may be associated with a taxid. These are (1) rank, such as species, genus or family; (2) genetic code, for translating proteins; (3) GenBank division; (4) literature citations; and (5) abbreviated lineage, for display in GenBank flat files. For more details on these data types, see Appendix 3.

Taxonomy in Entrez: A Quick Tour

The TAXON database is a node within the Entrez integrated retrieval system (Chapter 15) that provides an important organizing principle for other Entrez databases. Taxonomy provides an alternative view of TAXON to that of TaxBrowser. Entrez adds some very powerful capabilities (for example, Boolean queries, search history, and both internal and external links) to TAXON, but in many ways it is an unnatural way to represent such hierarchical data in Taxonomy. (TaxBrowser is the way to view the taxonomy hierarchically.)

Taxonomy was the first Entrez database to have an internal hierarchical structure. Because Entrez deals with unordered sets of objects in a given domain, an alternative way to represent these hierarchical relationships in Entrez was required (see the section Hierarchy Fields, below).

The main focus of the Entrez Taxonomy homepage is the search bar but also worth noting are the Help and TaxBrowser hotlinks that lead to Entrez generic help documentation and the Taxonomy browser, respectively.

The default Entrez search is case insensitive and can be for any of the names that can be found in the Taxonomy database. Thus, any of the following search terms, Homo sapiens, homo sapiens, human, or Man, will retrieve the node for *Homo sapiens*.

As for other Entrez databases, Taxonomy supports Boolean searching, a **History** function, and searches limited by field. The Taxonomy fields can be browsed under **Preview/Index**, some are specific to Taxonomy (such as **Lineage** or **Rank**), and others are found in all Entrez databases (such as Entrez **Date**).

Each search result, listed in document summary (DocSum) format, may have several links associated with it. For example, for the search result Homo sapiens, the **Nucleotide** link will retrieve all the human sequences from the nucleotide databases, and the **Genome** link will retrieve the human genome from the Genomes database.

Search Tips and Tricks

A helpful list follows:

1. A search for Hominidae retrieves a single, hyperlinked entry. Selecting the link shows the structure of the taxon. On the other hand, a search for Hominidae[subtree] will retrieve a nonhierarchical list of all of the taxa listed within the Hominidae.

2. A search for species[rank] yields a list of all species in the Taxonomy database (108,020 in May 2002).

3. Find the Taxonomy update frequency by selecting Entrez **Date** from the pull-down menu under **Preview/Index**, typing "2002/02" in the box and selecting **Index**. The result:

```
2002/01 (5176)
2002/01/03 (478)
2002/01/08 (2)
2002/01/10 (2260)
2002/01/14 (7)
2002/01/16 (239)
```

shows that in January 2002, 5,176 new taxa were added, the bulk of which appeared in Entrez for the first time on January 10, 2002. These taxa can be retrieved by selecting 2002/01/10, then selecting the **AND** button above the window, followed by **Go**.

4. An overview of the distribution of taxa in the DocSum list can be seen if **Summary** is changed to **Common Tree**, followed by selecting **Display**.

5. To filter out less interesting names from a DocSum list, add some terms to the query, e.g., 2002/01/10[date] NOT uncultured[prop] NOT unspecified[prop].

Displays in Taxonomy Entrez

There are a variety of choices regarding how search results can be displayed in Taxonomy Entrez.

Summary. This is the default display view. There are as many as four pieces of information in this display, if they are all present in the Taxonomy database: (1) scientific name of the taxon; (2) common name, if one is available; (3) taxonomic rank, if one is assigned; and (4) BLAST name, inherited from the taxonomy, e.g., Homo sapiens (human), species, mammals.

Brief. Shows only the scientific names of the taxa. This view can be used to download lists of species names from Entrez.

Tax ID List. Shows only the taxids of the taxa. This view can be used to download taxid lists from Entrez.

Info. Shows a summary of most of the information associated with each taxon in the Taxonomy database (similar to the TaxBrowser taxon-specific display; Figure 3). This can be downloaded as a text file; an XML representation of these data is under development.

Common Tree. A special display that shows a skeleton view of the relationships among the selected set of taxa and is described in the section below.

LinkOut. Displays a list of the linkout links (if any) for each of the selected sets of taxa (see Chapter 17).

Entrez Links. The remaining views follow Entrez links from the selected set of taxa to the other Entrez databases (Nucleotide, Protein, Genome, etc.) The **Display** view allows all links for a whole set of taxa to be viewed at once.

The Common Tree Viewer

The Common Tree view shows an abbreviated view of the taxonomic hierarchy and is designed to highlight the relationships between a selected set of organisms. Figure 5 shows the Common Tree view for a familiar set of model organisms.

If there are more than few dozen taxa selected for the common tree view, the display becomes visually complex and generally less useful. When a large list of taxa is sent to the Common Tree display, a summary screen is displayed first. For example, we currently list 727 families in the Viridiplantae (plants and green algae) (Figure 6).

There are several formatting options for saving the common tree display to a text file: text tree, phylip tree, and taxid list.

Hyperlinks to a common tree display can be made in two ways: (1) by specifying the common tree view in an Entrez query URL (for example, this link, which displays the common tree view of all of the taxonomy nodes with LinkOut links to the Butterfly Net International Web site); or (2) by providing a list of taxids directly to the common tree cgi function (for example, this link, which will display a live version of Figure 5).

Using Batch Taxonomy Entrez

The Batch Entrez page allows you to upload a file of taxids or taxon names into Taxonomy Entrez.

Indexing Taxonomy in Entrez

As for any Entrez database, the contents are indexed by creating term lists for each field of each database record (or taxid). For TAXON, the types of fields include name fields, hierarchy fields, inherited fields, and generic Entrez fields.

Name Fields

There are five different index fields for names in Taxonomy Entrez.

All names, [name] in an Entrez search – this is the default search field in Taxonomy Entrez. This is different from most Entrez databases, where the default search field is the composite [All Fields].

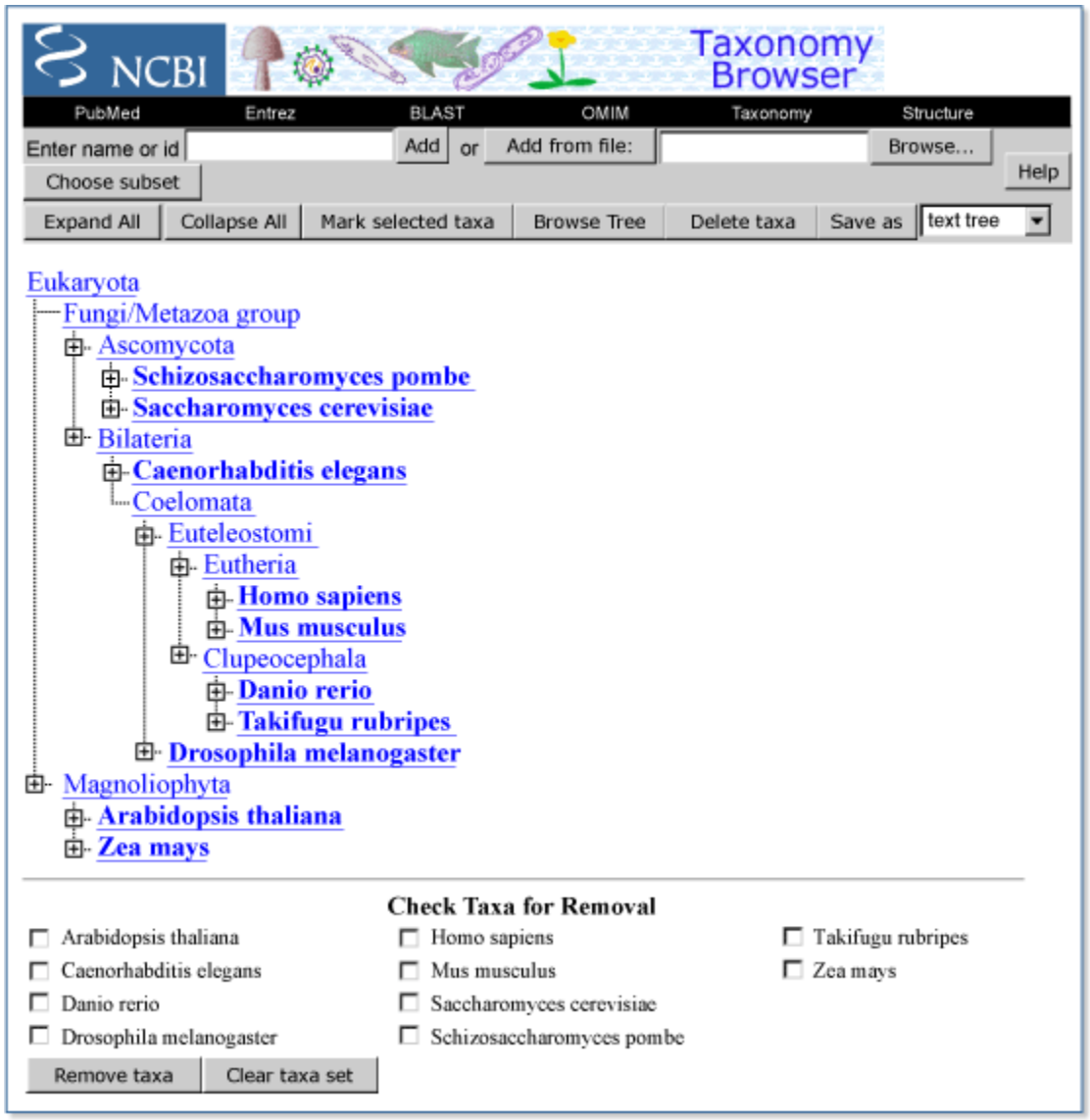


Figure 5. The Common Tree view for some model organisms. The ten species shown in *bold* are the ones that were selected as input to the Common Tree display. The other taxa displayed show the taxonomic relationships between the selected taxa. For example, *Eutheria* is included because it is the smallest taxonomic group in our classification that includes both *Homo sapiens* and *Mus musculus*. A "+" box in the tree indicates that part of the taxonomic classification has been suppressed in this abbreviated view; selecting the "+" will fill in the missing lineage (and change the "+" to a "-"). The **Expand All** and **Collapse All** buttons at the *top* of the display will do this globally. The **Search for** box at the *top* of the display can be used to add taxa to the Common Tree display; taxa can be removed using the list at the *bottom* of the page.

Scientific name, [sname] – using [sname] as a qualifier in a search restricts it to the nametype "scientific name", the single preferred name for each taxon.

Common name, [cname] – restricts the search to common names.

Synonym, [synonym] – restricts the search to the "synonym" nametype.

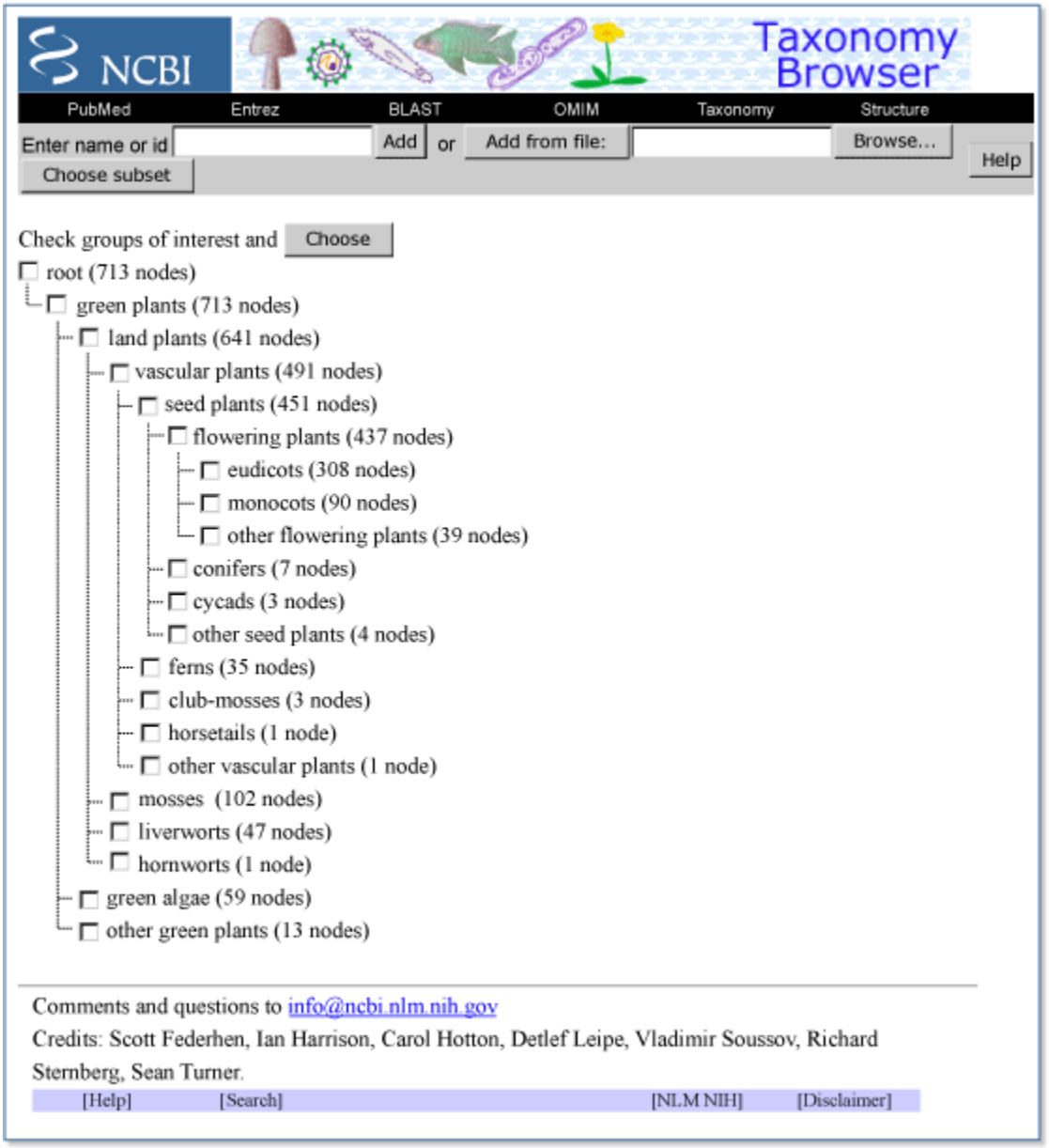


Figure 6. The Common Tree summary page for the plants and green algae. The taxa are aggregated at the predetermined set of nodes in the Taxonomy database that have been assigned "BLAST names". This serves an informal, very abbreviated, vernacular classification that gives a convenient overview. The BLAST names will often not provide complete coverage for all species at all levels in the tree. Here, not all of our *flowering plants* are flagged as *eudicots* or *monocots*. The Common Tree summary display recognizes cases such as these and lists the remaining taxa as *other flowering plants*. The full Common Tree for some or all of the taxa can be seen by selecting the check box next to *monocots* on the summary page and then **Choose**. This will display the full Common Tree view for 108 families of monocots.

Taxid, [uid] – restricts the search to taxonomy IDs, the unique numerical identifiers for taxa in the database. Taxids are not indexed in the other Entrez name fields.

Hierarchy Fields

The [lineage] and [subtree] index fields are a way to superimpose the hierarchical relationships represented in the taxonomy on top of the Entrez data model. For an example of how to use these field limits for searching Taxonomy Entrez, see Box 3.

Lineage. For each node, the [lineage] index field retrieves all of the taxa listed at or above that node in the taxonomy. For example, the query Mammalia[lineage] retrieves 18 taxa from Entrez.

Subtree. For each node, the [subtree] index field retrieves all of the taxa listed at or below that node in the taxonomy. For example, the query Mammalia[lineage] retrieves 4,021 taxa from Entrez (as of March 9, 2002).

Next level. Returns all of the direct children of a given taxon.

Rank. Returns all of the taxa of a given Linnaean rank. The query Aves[subtree] AND species[rank] retrieves all of the species of birds with public sequence entries (there are 2,459, approximately half of the currently described species of extant birds).

Box 3. Examples of combining the subtree and lineage field limits with Boolean operators for searching Taxonomy Entrez.

(1) Mammalia[subtree] AND Mammalia[lineage] returns the taxon Mammalia.

(2) Mammalia[subtree] OR Mammalia[lineage] returns all of the taxa in a direct parent–child relationship with the taxon Mammalia.

(3) root[subtree] NOT (Mammalia[subtree] OR Mammalia[lineage]) returns all of the taxa not in a direct parent–child relationship with the taxon Mammalia.

(4) Sauropsida[subtree] NOT Aves[subtree] will retrieve the members of the classical taxon Reptilia, excluding the birds.

Inherited Fields

The genetic code [gc], mitochondrial genetic code [mgc], and GenBank division [division] fields are all inherited within the taxonomy. The information in these fields refers to the genetic code used by a taxon or in which GenBank division it resides. Because whole families or branches may use the same code or reside in the same GenBank division, this property is usually indexed with a taxon high in the taxonomic tree, and the information is inherited by all those taxa below it. If there is no [gc] field associated with a taxon in the database, it is assumed that the standard genetic code is used. A genetic code may be referred to by either name or translation table number. For example, the two equivalent queries, standard[gc] and translation table 1[gc], each retrieves the set of organisms that use the standard genetic code for translating genomic sequences. Likewise, these two queries echinoderm mitochondrial[mgc] and translation

table 9[mgc] will each retrieve the set of organisms that use the echinoderm mitochondrial genetic code for translating their mitochondrial sequences.

Generic Entrez Fields

The remaining index fields are common to most or all Entrez domains, although some have special features in the taxonomy domain. For example, the field text word, [word], indexes words from the Taxonomy Entrez name indexes. Most punctuation is ignored, and the index is searched one word at a time; therefore, the search "homo sapiens[word]" will retrieve nothing.

Several useful terms are indexed in the properties field, [prop], including functional nametypes and classifications, the rank level of a taxon, and inherited values. See Box 4 for a detailed discussion of searches using the [prop] field.

More information on using the generic Entrez fields can be found in the Entrez Help documents.

Box 4. The properties [prop] field of Taxonomy in Entrez.

There are several useful terms and phrases indexed in the [prop] field. Possible search strategies that specify the prop field are explained below.

(1) Using functional nametypes and classifications

unspecified [prop] not identified at the species level

uncultured [prop] environmental sample sequences

unclassified [prop] listed in an "unclassified" bin

incertae sedis [prop] listed in an "incertae sedis" bin

We do not explicitly flag names as "unspecified" in TAXON; rather, we rely on heuristics to index names as "unspecified" in the properties field. Many are missed. Taxa are indexed as "uncultured" if they are listed within an environmental samples bin or if their scientific names begin with the word "uncultured".

(2) Using rank level of taxon

All of these search strategies below are valid. Taxonomy Entrez displays only taxa that are linked to public sequence entries, and because sequence entries are supposed to correspond to the Taxonomy database at or below the species level, the Entrez query: terminal [prop] NOT "at or below species level" [prop] should only retrieve problem cases.

above genus level [prop]

Box 4 continues on next page...

Box 4 continued from previous page.

above species level [prop]

"at or below species level" [prop] (needs explicit quotes)

below species level [prop]

terminal [prop]

non terminal [prop]

(3) Inherited value assignment points

genetic code [prop]

mitochondrial genetic code [prop]

standard [prop] invertebrate mitochondrial [prop]

translation table 5 [prop]

The query "genetic code [prop]" retrieves all of the taxa at which one of the genomic genetic codes is explicitly set. The second query retrieves all of the taxa at which one of the mitochondrial genetic codes is explicitly set, and so on.

division [prop]

INV [prop]

invertebrates [prop]

The above terms index the assignments of the GenBank division codes, which are divided along crude taxonomic categories (see Chapter 1). We have placed the division flags in the database so as to preserve the original assignment of species to GenBank divisions.

Taxonomy Fields in Other Entrez Databases

Many of the Entrez databases (Nucleotide, Protein, Genome, etc.) include an **Organism** field, [orgn], that indexes entries in that database by taxonomic group. All of the names associated with a taxon (scientific name, synonyms, common names, and so on) are indexed in the **Organism** field and will retrieve the same set of entries. The **Organism** field will retrieve all of the entries below the term and any of their children.

To not retrieve such "exploded" terms, the unexploded indexes should be used. This query will only retrieve the entries that are linked directly to *Homo sapiens*: Homo sapiens[orgn:noexp]. This query will not retrieve entries that are linked to the subordinate node *Homo sapiens neanderthalensis*.

Taxids are indexed with the prefix txid: txid9606 [orgn].

Source organism modifiers are indexed in the [properties] field, and such queries would be in the form: src strain[prop], src variety[prop], or src specimen voucher[prop]. These queries will retrieve all entries with a strain qualifier, a variety qualifier, or a specimen_voucher qualifier, respectively.

All of the organism source feature modifiers (/clone, /serovar, /variety, etc.) are indexed in the text word field, [text word]. For example, one could query GenBank for: "strain k-12" [text word]. Because strain information is inconsistent in the sequence databases (as in the literature), a better query would be: "strain k 12"[word] OR "strain k12"[word]. Note: explicit double-quotes may be necessary for some of these queries.

The Taxonomy Statistics Page

The Taxonomy Statistics page displays tables of counts of the number of taxa in the public subset of the Taxonomy database. The numbers displayed are hyperlinks that will retrieve the corresponding set of entries. The table can be configured to display data based on three criteria: Entrez release date, rank, and taxa. The default setting shows the counts by rank for a pre-selected set of taxa (across all dates).

The checkboxes **unclassified**, **uncultured**, and **unspecified** will exclude the corresponding sets of taxa from the count. These work by appending the terms "NOT unclassified[prop]", etc., to the statistics query. Checking **uncultured** and **unspecified** removes about 20% of taxa in the database and gives a much better count of the number of formally described species. As of April 1, 2003, the count was as follows:

```
Archaea: 82 genera, 364 species
Bacteria: 1163 genera, 9927 species
Eukaryota: 24939 genera, 74832 species
```

Selecting one of the **rank** categories (e.g., species) loads a new table that shows, in this example, the number of new species added to each taxon each year, starting with 1993. The **Interval** pull-down menu shows release statistics in finer detail. The list of taxa in the display can be customized through the **Customize** link.

Other Relevant References

Taxonomy FTP

A complete copy of the public NCBI taxonomy database is deposited several times a day on our FTP site. See Table 1 for details.

Tax BLAST

Taxonomy BLAST reports (Tax BLAST) are available from the BLAST results page and from the BLink pages. Tax BLAST post-processes the BLAST output results according to the source organisms of the sequences in the BLAST results page. A help page is available

that describes the three different views presented on the Tax BLAST page (Lineage Report, Organism Report, and Taxonomy Report).

Toolkit Function Libraries

The function library for the taxonomy application software in the NCBI Toolkit is ncbitxc2.a (or libncbitxc2.a). The source code can be found in the NCBI Toolkit Source Browser and can be downloaded from the toolbox directory on the FTP site.

NCBI Taxonomists

In the early years of the project, Scott Federhen did all of the software and database development. In recent years, Vladimir Soussov and his group have been responsible for software and database development.

Scott Federhen (1990–present)

Andrzej Elzanowski (1994–1997)

Detlef Leipe (1994–present)

Mark Hershkovitz (1996–1997)

Carol Hotton (1997–present)

Mimi Harrington (1999–2000)

Ian Harrison (1999–2002)

Sean Turner (2000–present)

Rick Sternberg (2001–present)

Contact Us

If you have a comment or correction to our Taxonomy database, perhaps a misspelling or classification or if something looks wrong, please send a message to info@ncbi.nlm.nih.gov.

Appendix 1. TAXON nametypes

Scientific Name

Every node in the database is required to have exactly one "scientific name". Wherever possible, this is a validly published name with respect to the relevant code of nomenclature. Formal names that are subject to a code of nomenclature and are associated with a validly published description of the taxon will be Latinized uninomials above the species level, binomials (e.g. *Homo sapiens*) at the species level, and trinomials for the formally described infraspecific categories (e.g., *Homo sapiens neanderthalensis*).

For many of our taxa, it is not possible to find an appropriate formal scientific name; these nodes are given an informal "scientific name". The different classes of informal names are discussed in Appendix 2. Functional Classes of TAXON Scientific Names.

The scientific name is the one that will be used in all of the sequence entries that map to this node in the Taxonomy database. Entries that are submitted with any of the other names associated with this node will be replaced with this name. When we change the scientific name of a node in the Taxonomy database, the corresponding entries in the sequence databases will be updated to reflect the change. For example, we list *Homo neanderthalensis* as a synonym for *Homo sapiens neanderthalensis*. Both are in common use in the literature. We try to impose consistent usage on the entries in the sequence databases, and resolving the nomenclatural disputes that inevitably arise between submitters is one of the most difficult challenges that we face.

Synonym

The "synonym" nametype is applied to both synonyms in the formal nomenclatural sense (objective, nomenclatural, homotypic *versus* subjective, heterotypic) and more loosely to include orthographic variants and a host of names that have found their way into the taxonomy database over the years, because they were found in sequence entries and later merged into the same taxon in the Taxonomy database.

Acronym

The "acronym" nametype is used primarily for the viruses. The International Committee on Taxonomy of Viruses (ICTV) maintains an official list of acronyms for viral species, but the literature is often full of common variants, and it is convenient to list these as well. For example, we list HIV, LAV-1, HIV1, and HIV-1 as acronyms for the human immunodeficiency virus type 1.

Anamorph

The term "anamorph" is reserved for names applied to asexual forms of fungi, which present some special nomenclatural challenges. Many fungi are known to undergo both sexual and asexual reproduction at different points in their life cycle (so-called "perfect" fungi); for many others, however, only the asexually reproducing (anamorphic or mitosporic) form is known (in some, perhaps many, asexual species, the sexual cycle may have been lost altogether). These anamorphs, often with simple and not especially diagnostic morphology, were given Linnaean binomial names. A number of named anamorphic species have subsequently been found to be associated with sexual forms (teleomorphs) with a different name (for example, *Aspergillus nidulans* is the name given to the asexual stage of the teleomorphic species *Emericella nidulans*). In these cases, the teleomorphic name is given precedence in the GenBank Taxonomy database as the "scientific" name, and the anamorphic name is listed as an "anamorph" nametype.

Misspelling

The “misspelling” nametype is for simple misspellings. Some of these are included because the misspelling is present in the literature, but most of them are there because they were once found in a sequence entry (which has since been corrected). We keep them in the database for tracking purposes, because copies of the original sequence entry can still be retrieved. Misspellings are not listed on the TaxBrowser pages nor on the Taxonomy Entrez Info display views, but they are indexed in the Entrez search fields (so that searches and Entrez queries with the misspelling will find the appropriate node).

Misnomer

“Misnomer” is a rarely used nametype. It is used for names that might otherwise be listed as “misspellings” but which we want to appear on the browser and Entrez display pages.

Common Name

The “common name” nametype is used for vernacular names associated with a particular taxon. These may be found at any level in the hierarchy; for example, “human”, “reptiles”, and “pale devil's-claw” are all used. Common names should be in lowercase letters, except where part of the name is derived from a proper noun, for example, “American butterfish” and “Robert's arboreal rice rat”.

The use of common names is inherently variable, regional, and often inconsistent. There is generally no authoritative reference that regulates the use of common names, and there is often not perfect correspondence between common names and formally described scientific taxa; therefore, there are some caveats to their use. For scientific discourse, there is no substitute for formal scientific names. Nevertheless, common names are invaluable for many indexing, retrieval, and display purposes. The combination “*Oecomys roberti* (Robert's arboreal rice rat)” conveys much more information than either name by itself. Issues raised by the variable, regional, and inexact use of common names are partly addressed by the “genbank common name” nametype (below) and the ability to customize names in the GenBank flatfile.

BLAST Name

The “BLAST names” are a specially designated set of common names selected from the Taxonomy database. These were chosen to provide a pool of familiar names for large groups of organisms (such as “insects”, “mammals”, “fungi”, and others) so that any particular species (which may not have an informative common name of its own) could inherit a meaningful collective common name from the Taxonomy database. This was originally developed for BLAST, because a list of BLAST results will typically include entries from many species identified by Latin binomials, which may not be familiar to all users. BLAST names may be nested; for example, “eukaryotes”, “animals”, “chordates”, “mammals”, and “primates” are all flagged as “blast names”.

Blast names are now used in several other applications, for example the Tax BLAST displays, the Summary view in Entrez Taxonomy, and in the Summary display of the Common Tree format.

In-part

The "in-part" nametype is included for retrieval terms that have a broader range of application than the taxon or taxa at which they appear. For example, we list reptiles and Reptilia as in-part nametypes at our nodes *Testudines* (the turtles), *Lepidosauria* (the lizards and snakes), and *Crocodylidae* (the crocodilians).

Includes

The "includes" nametype is the opposite of the in-part nametype and is included for retrieval terms that have a narrower scope of application than the taxon at which they appear. For example, we could list "reptiles" as an "includes" nametype for the *Amniota* (or at any higher node in the lineage).

Equivalent Name

The "equivalent name" nametype is a catch-all category, used for names that we would like to associate with a particular node in the database (for indexing or tracking purposes) but which do not seem to fit well into any of the other existing nametypes.

GenBank Common Name

The "genbank common name" was introduced to provide a mechanism by which, when there is more than one common name associated with a particular node in the taxonomy, one of them could be designated to be the common name that should be used by default in the GenBank flatfiles and other applications that are trying to find a common name to use for display (or other) purposes. This is not intended to confer any special status or blessing on this particular common name over any of the other common names that might be associated with the same node, and we have developed mechanisms to override this choice for a common name on a case-by-case basis if another name is more appropriate or desirable for a particular sequence entry. Each node may have at most one "genbank common name".

GenBank Acronym

There may be more than one acronym associated with a particular node in the Taxonomy database (particularly if several virus names have been synonymized in a single species). Just as with the "genbank common name", the "genbank acronym" provides a mechanism to designate one of them to be the acronym that should be used for display (or other) purposes. Each node may have at most one "genbank acronym".

GenBank Synonym

The "genbank synonym" nametype is intended for those special cases in which there is more than one name commonly used in the literature for a particular species, and it is informative to have both names displayed prominently in the corresponding sequence record. Each node may have at most one "genbank synonym". For example,

```
SOURCE Takifugu rubripes (Fugu rubripes)
ORGANISM Takifugu rubripes
```

GenBank Anamorph

Although the use of either the anamorph or teleomorph name is formally correct under the International Code of Botanical Nomenclature, we prefer to give precedence to the telemorphic name as the "scientific name" in the Taxonomy database, both to emphasize their commonality and to avoid having two (or more) taxids that effectively apply to the same organism. However, in many cases, the anamorphic name is much more commonly used in the literature, especially when sequences are normally derived from the asexual form of the species. In these cases, the "genbank anamorph" nametype can be used to annotate the corresponding sequence records with both names. Each node may have at most one "genbank anamorph". For example:

```
SOURCE Emericella nidulans (anamorph: Aspergillus nidulans)
ORGANISM Emericella nidulans
```

Appendix 2. Functional classes of TAXON scientific names

Formal Names

Whenever possible, formal scientific names are used for taxa. There are several codes of nomenclature that regulate the description and use of names in different branches of the tree of life. These are: the International Code of Zoological Nomenclature (ICZN), the International Code of Botanical Nomenclature (ICBN), the International Code of Nomenclature for Cultivated Plants (ICNCP), the International Code of Nomenclature of Bacteria (ICNB), and the International Code of Virus Classification and Nomenclature (ICVCN).

The viral code is less well developed than the others, but it includes an official classification for the viruses as well as a list of approved species names. Viral names are not Latin binomials (as required by the other codes), although there are some instances (e.g., *Herpesvirus papio* or *Herpesvirus sylvilagus*). When possible, we try to use ICTV-approved names for viral taxa, but new viral species names appear in the literature (and therefore in the sequence databases) much faster than they are approved into the ICTV lists. We are working to set up taxonomy LinkOut links (see Chapter 17) to the ICTV database, which will make the subset of ICTV-approved names explicit.

The zoological, botanical, and bacteriological codes mandate Latin binomials for species names. They do not describe an official classification (such as the ICTV), with the exception that the binomial species nomenclature itself makes the classification to the genus level explicit. If a genus is found to be polyphyletic, the classification cannot be corrected without formally renaming at least some of the species in the genus. (This is somewhat reminiscent of the "smart identifier" problem in computer science.)

The fungi are subject to the botanical code. The cyanobacteria (blue-green algae) have been subject to both the botanical and the bacteriological codes, and the issue is still controversial.

Authorities

"Authorities" appear at the end of the formal species name and include at least the name or standard abbreviation of the taxonomist who first described that name in the scientific literature. Other information may appear in the authority as well, often the year of description, and can become quite complicated if the taxon has been transferred or amended by other taxonomists over the years. We do not use authorities in our taxon names, although many are included in the database listed as synonyms. We have made an exception to this rule in the case of our first duplicated species name in the database, *Agathis montana*, to provide unambiguous names for the corresponding sequence entries.

Subspecies

All three of the codes of nomenclature for cellular organisms provide for names at the subspecies level. The botanical and bacteriological codes include the string "subsp." in the formal name; the zoological code does not, e.g., *Homo sapiens neanderthalensis*, *Zea mays subsp. mays*, and *Klebsiella pneumoniae subsp. ozaenae*.

Varietas and Forma

The botanical code (but none of the others) provides for two additional formal ranks beneath the subspecies level, varietas and forma. These names will include the strings "var." and "f.", respectively, e.g., *Marchantia paleacea var. diptera, Penicillium aurantiogriseum var. neoechinulatum, Salix babylonica f. rokkaku*, or *Fragaria vesca subsp. Vesca f. alba*

Other Subspecific Names

We list taxa with other subspecific names where it seems useful and appropriate and where it is necessary to find places for names in the sequence databases. For indexing purposes in the Genomes division of Entrez, it is convenient to have strain-level nodes for bacterial species with a complete genome sequence, particularly when there are two or more complete genome sequences available for different strains of the same species, e.g., *Escherichia coli* K12, *Escherichia coli* O157:H7, *Escherichia coli* O157:H7 EDL933, *Mycobacterium tuberculosis* CDC1551, and *Mycobacterium tuberculosis* H37Rv.

Several other classes of subspecific groups do not have formal standing in the nomenclature but represent well-characterized and biologically meaningful groups, e.g., serovar, pathovar, forma specialis, and others. In many cases, these may eventually be promoted to a species; therefore, it is convenient to represent them independently from the outset, e.g., *Xanthomonas campestris* pv. *campestris, Xanthomonas campestris* pv. *vesicatoria*, *Pneumocystis carinii* f. sp. *hominis, Pneumocystis carinii* f. sp. *mustelae, Salmonella enterica* subsp. *enterica* serovar Dublin, and *Salmonella enterica* subsp. *enterica* serovar Panama.

Many other names below the species level have been added to the Taxonomy database to accommodate SWISS-PROT entries, where strain (and other) information is annotated with the organism name for some species.

Informal Names

In general, we try to avoid unqualified species names such as *Bacillus* sp., although many of them exist in the Taxonomy database because of earlier sequence entries. *Bacillus* sp. is a particularly egregious example, because *Bacillus* is a duplicated genus name and could refer to either a bacterium or an insect. In our database, *Bacillus* sp. is assumed to be a bacterium, but *Bacillus* sp. P-4-N, on the other hand, is classified with the insects.

When entries are not identified at the species level, multiple sequences can be from the same unidentified species. Sequences from multiple different unidentified species in the same genus are also possible. To keep track of this, we add unique informal names to the Taxonomy database, e.g., a meaningful identifier from the submitters could be used. This could be a strain name, a culture collection accession, a voucher specimen, an isolate name or location—anything that could tie the entry to the literature (or even to the lab notebook). If nothing else is available, we may construct a unique name using a default formula such as the submitter's initials and year of submission. This way, if a formal name is ever determined or described for any of these organisms, we can synonymize the informal name with the formal one in the Taxonomy database, and the corresponding entries in the sequence databases will be updated automatically. For example, AJ302786 was originally submitted (in November 2000) as *Agathis* sp. and was added to the Taxonomy database as *Agathis* sp. RDB-2000. In January 2002, this wasp was identified as belonging to the species *Agathis montana*, and the node was renamed; the informal name *Agathis* sp. RDB-2000 was listed as a synonym. A separate member of the genus, *Agathis sp.* DMA-1998, is still listed with an informal name.

Here are some examples of informal names in the Taxonomy database:

```
Anabaena sp. PCC 7108
Anabaena sp. M14-2
Calophyllum sp. 'Fay et al. 1997'
Scutellospora sp. Rav1/RBv2/RCv3
Ehrlichia-like sp. 'Schotti variant'
Gilia sp. Porter and Heil 7991
```

```
Camponotus n. sp. BGW-2001
Camponotus sp. nr. gasseri BGW-2001
Drosophila sp. 'white tip scutellum'
Chrysoperla sp. 'C.c.2 slow motorboat'
Saranthe aff. eichleri Chase 3915
Agabus cf. nitidus IR-2001
Simulium damnosum s.l. 'Kagera'
Amoebophrya sp. ex Karlodinium micrum
```

We use single quotes when it seems appropriate to group a phrase into a single lexical unit. Some of these names include abbreviations with special meanings.

"n. sp." indicates that this is a new, undescribed species and not simply an unidentified species. "sp. nr." indicates "species near". In the example above, this indicates that this is similar to Camponotus gasseri. "aff.", affinis, related to but not identical to the species given. "cf.", confer; literally, "compare with" conveys resemblance to a given species but is not necessarily related to it. "s.l.", sensu lato; literally, "in the broad sense". "ex", "from" or "out of" the biological host of the specimen.

Note that names with cf., aff., nr., and n. sp. are not unique and should have unique identifiers appended to the name.

Cultured bacterial strains and other specimens that have not been identified to the genus level are given informal names as well, e.g., Desulfurococcaceae str. SRI-465; crenarchaeote OlA-6.

Names such as Camponotus sp. 1 are avoided, because different submitters might easily use the same name to refer to different species. See Box 4 for how to retrieve these names in Taxonomy Entrez.

Uncultured Names

Sequences from environmental samples are given "uncultured" names. In these studies, nucleotide sequences are cloned directly from the environment and come from varied sources, such as Antarctic sea ice, activated sewer sludge, and dental plaque. Apart from the sequence itself, there is no way to identify the source organisms or to recover them for further studies. These studies are particularly important in bacterial systematics work, which shows that the vast majority of environmental bacteria are not closely related to laboratory cultured strains (as measured by 16S rRNA sequences). Many of the deepest-branching groups in our bacterial classification are defined only by anonymous sequences from these environmental samples studies, e.g., candidate division OP5, candidate division Termite group 1, candidate subdivision kps59rc, phosphorous removal reactor sludge group, and marine archael group 1.

These samples vary widely in length and in quality, from short single-read sequences of a few hundred base pairs to high-quality, full-length 16S sequences. We now give all of these samples anonymous names, which may indicate the phylogenetic affiliation of the sequence, as far is it may be determined, e.g., uncultured archaeon, uncultured

crenarchaeote, uncultured gamma proteobacerium, or uncultured enterobacterium. See Box 4 for how to retrieve these names in Taxonomy Entrez.

Candidatus Names

Some groups of bacteria have never been cultured but can be characterized and reliably recovered from the environment by other means. These include endosymbiotic bacteria and organisms similar to the phytoplasmas, which can be identified by the plant diseases that they cause. We do not give these "uncultured" names, as above. These represent a special challenge for bacterial nomenclature, because a formal species description requires the designation of a cultured type strain. The bacteriological code has a special provision for names of this sort, Candidatus, e.g., Candidatus Endobugula or Candidatus Endobugula sertula; Candidatus Phlomobacter or Candidatus Phlomobacter fragariae. These often appear in the literature without the Candidatus prefix; therefore, we list the unqualified names as synonyms for retrieval purposes.

Informal Names above the Species Level

We allow informal names for unranked nodes above the species level as well. These should all be phylogenetically meaningful groups, e.g., the Fungi/Metazoa group, eudicotyledons, Erythrobasidium clade, RTA clade, and core jakobids. In addition, there are several other classes of nodes and names above the species level that explicitly do not represent phylogenetically meaningful groups. These are outlined below.

Unclassified Bins

We are expected to add new species names to the database in a timely manner, preferably within a day or two. If we are able to find only a partial classification for a new taxon in the database, we place it as deeply as we can and list it in an explicit "unclassified" bin. As more information becomes available, these bins are emptied, and we give full classifications to the taxa listed there. In general, we suppress the names of the unclassified bins themselves so they do not appear in the abbreviated lineages that appear in the GenBank flatfiles, e.g., unclassified Salticidae, unclassified Bacteria, and unclassified Myxozoa.

Incertae Sedis Bins

If the best taxonomic opinion available is that the position of a particular taxon is uncertain, then we will list it in an "incertae sedis" bin. This is a more permanent assignment than for taxa that are listed in unclassified bins, e.g., *Neoptera incertae sedis, Chlorophyceae incertae sedis,* and *Riodininae incertae sedis.*

Mitosporic Bins

Fungi that were known only in the asexual (mitosporic, anamorphic) state were placed formerly in a separate, highly polyphyletic category of "imperfect" fungi, the Deuteromycota. Spurred especially by the development of molecular phylogenetics,

current mycological practice is to classify anamorphic species as close to their sexual relatives as available information will support. Mitosporic categories can occur at any rank, e.g., *mitosporic Ascomycota, mitosporic Hymenomycetes, mitosporic Hypocreales,* and *mitosporic Coniochaetaceae*. The ultimate goal is to fully incorporate anamorphs into the natural phylogenetic classification.

Other Names

The requirement that the Taxonomy database includes names from all of the entries in the sequence database introduces a number of names that are not typically treated in a taxonomic database. These are listed in the top-level group "Other". Plasmids are typically annotated with their host organism, using the /plasmid source organism qualifier. Broad-host-range plasmids that are not associated with any single species are listed in their own bin. Plasmid and transposon names from very old sequence entries are listed in separate bins here as well. Plasmids that have been artificially engineered are listed in the "vectors" bin.

Appendix 3. Other TAXON data types

Ranks

We do not require that Linnaean ranks be assigned to all of our taxa, but we do include a standard rank table that allows us to assign formal ranks where it seems appropriate. We do not require that sibling taxa all have the same rank, but we do not allow taxa of higher rank to be listed beneath taxa of lower rank. We allow unranked nodes to be placed at any point in our classification.

The one rank that we particularly care about is "species". We try to ensure that all of the sequence entries map into the Taxonomy database at or below a species-level node.

Genetic Codes

The genetic codes and mitochondrial genetic codes that are appropriate for translating protein sequences in different branches of the tree of life are assigned at nodes in the Taxonomy database and inherited by species at the terminal branches of the tree. Plastid sequences are all translated with the standard genetic code, but many of the mRNAs undergo extensive RNA editing, making it difficult or impossible to translate sequences from the plastid genome directly. The genetic codes are listed on our Web site.

GenBank Divisions

GenBank taxonomic division assignments are made in the Taxonomy database and inherited by species at the terminal branches of the tree, just as with the genetic codes.

References

The Taxonomy database allows us to store comments and references at any taxon. These may include hotlinks to abstracts in PubMed, as well as links to external addresses on the Web.

Abbreviated Lineage

Some branches of our taxonomy are many levels deep, e.g., the bony fish (as we moved to a phylogenetic classification) and the drosophilids (a model taxon for evolutionary studies). In many cases, the classification lines in the GenBank flatfiles became longer than the sequences themselves. This became a storage and update issue, and the classification lines themselves became less helpful as generally familiar taxa names became buried within less recognizable taxa.

To address this problem, the Taxonomy database allows us to flag taxa that should (or should not) appear in the abbreviated classification line in the GenBank flatfiles. The full lineages are indexed in Entrez and displayed in the Taxonomy Browser.

Chapter 5. The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation

Adrienne Kitts and Stephen Sherry

Created: October 9, 2002; Updated: February 2, 2011.

Summary

Sequence variations exist at defined positions within genomes and are responsible for individual phenotypic characteristics, including a person's propensity toward complex disorders such as heart disease and cancer. As tools for understanding human variation and molecular genetics, sequence variations can be used for gene mapping, definition of population structure, and performance of functional studies.

The Single Nucleotide Polymorphism database (dbSNP) is a public-domain archive for a broad collection of simple genetic polymorphisms. This collection of polymorphisms includes single-base nucleotide substitutions (also known as single nucleotide polymorphisms or SNPs), small-scale multi-base deletions or insertions (also called deletion insertion polymorphisms or DIPs), and retroposable element insertions and microsatellite repeat variations (also called short tandem repeats or STRs). Please note that in this chapter, you can substitute any class of variation for the term SNP. Each dbSNP entry includes the sequence context of the polymorphism (i.e., the surrounding sequence), the occurrence frequency of the polymorphism (by population or individual), and the experimental method(s), protocols, and conditions used to assay the variation.

dbSNP accepts submissions for variations in any species and from any part of a genome. This document will provide you with options for finding SNPs in dbSNP, discuss dbSNP content and organization, and furnish instructions to help you create your own (local) copy of dbSNP.

Introduction

The dbSNP has been designed to support submissions and research into a broad range of biological problems. These include physical mapping, functional analysis, pharmacogenomics, association studies, and evolutionary studies. Because dbSNP was developed to complement GenBank, it may contain nucleotide sequences (Figure 1) from any organism.

Physical Mapping

In the physical mapping of nucleotide sequences, variations are used as positional markers. When mapped to a unique location in a genome, variation markers work with the same logic as Sequence Tagged Sites (STSs) or framework microsatellite markers. As is

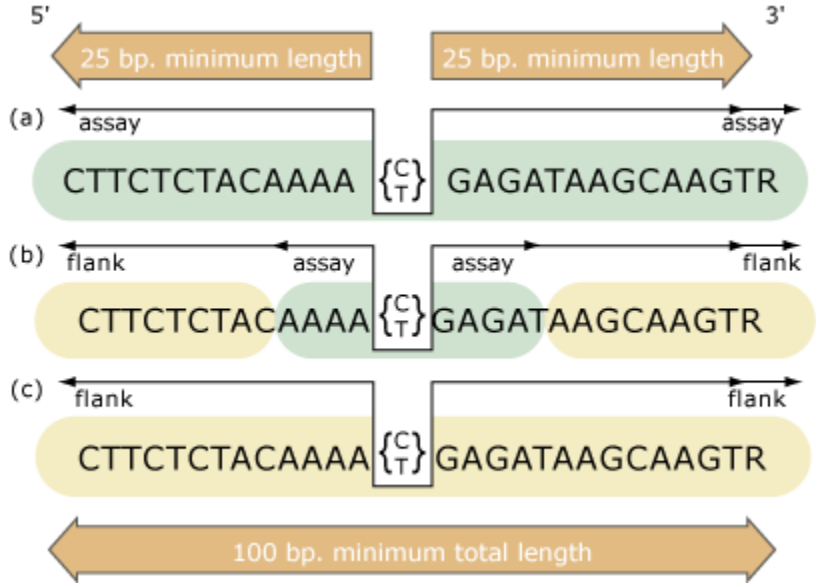


Figure 1. The structure of the flanking sequence in dbSNP is a composite of bases either assayed for variation or included from published sequence. We make the distinction to distinguish regions of sequence that have been experimentally surveyed for variation (*assay*) from those regions that have not been surveyed (*flank*). The minimum sequence length for a variation definition (SNPassay) is 25 bp for both the 5′ and 3′ flanks and 100 bp overall to ensure an adequate sequence for accurate mapping of the variation on reference genome sequence. (*a*) Flanking sequence completely surveyed for variation. Both 5′ and 3′ flanking sequence can be defined with 5′_assay and 3′_assay fields, respectively, when all flanking sequence was examined for variation. This can occur in both experimental contexts (e.g., denaturing high-pressure liquid chromatography or DNA sequencing) and computational contexts (e.g., analysis of BAC overlap sequence). (*b*) Partial survey of flanking sequence can occur when detection methods examine only a region of sequence surrounding the variation that is shorter than either the 25 bp per flank rule or the 100 bp overall length rule. In these experimental designs (e.g., chip hybridization, enzymatic cleavage), we designate the experimental sequence 5′_assay or 3′_assay, and you can insert published sequence (usually from a gene reference sequence) as 5′_flank or 3′_flank to construct a sequence definition that will satisfy the length rules. (*c*) Unknown or no survey of flanking sequence can occur when variations are captured from published literature without an indication of survey conditions. In these cases, the entire flanking sequence is defined as 5′_flank and 3′_flank.

the case for STSs, the position of a variation is defined by its unique flanking sequence, and hence, variations can serve as stable landmarks in the genome, even if the variation is fixed for one allele in a sample. When multiple alleles are observed in a sample pedigree, pedigree members can be tested for variation genotypes as in traditional physical mapping studies.

Functional Analysis

Variations that occur in functional regions of genes or in conserved non-coding regions might cause significant changes in the complement of transcribed sequences. This can lead to changes in protein expression that can affect aspects of the phenotype such as

metabolism or cell signaling. We note possible functional implications of DNA sequence variations in dbSNP in terms of how the variation alters mRNA transcripts.

Association Studies

The associations between variations and complex genetic traits are more ambiguous than simple, single-gene mutations that lead to a phenotypic change. When multiple genes are involved in a trait, then the identification of the genetic causes of the trait requires the identification of the chromosomal segment combinations, or haplotypes, that carry the putative gene variants.

Evolutionary Studies

The variations in dbSNP currently represent an uneven but large sampling of genome diversity. The human data in dbSNP include submissions from the SNP Consortium, variations mined from genome sequence as part of the human genome project, and individual lab contributions of variations in specific genes, mRNAs, ESTs, or genomic regions.

Null Results Are Important

Systematic surveys of sequence variation will undoubtedly reveal sequences that are invariant in the sample. These observations can be submitted to dbSNP as NoVariation records that record the sequence, the population, and the sample size that were used in the survey.

Searching dbSNP

The SNP database can be queried from the dbSNP homepage (Figures 2a and 2b), by using Entrez SNP, or by using the links to the six basic dbSNP search options located just below the text box at the top of the dbSNP homepage. Each of these six search options is described below.

Entrez SNP

dbSNP is a part of the Entrez integrated information retrieval system (Chapter 15) and may be searched using either qualifiers (aliases) or a combination of 25 different search fields. A complete list of the qualifiers and search fields can be found on the Entrez SNP site.

Single Record (Search by ID Number) Query in dbSNP

Use this query module to select SNPs based on dbSNP record identifiers. These include reference SNP (refSNP) cluster ID numbers (rs#), submitted SNP Accession numbers (ss#), local (or submitter) IDs, Celera IDs, Genbank accession numbers, and STS accession numbers.

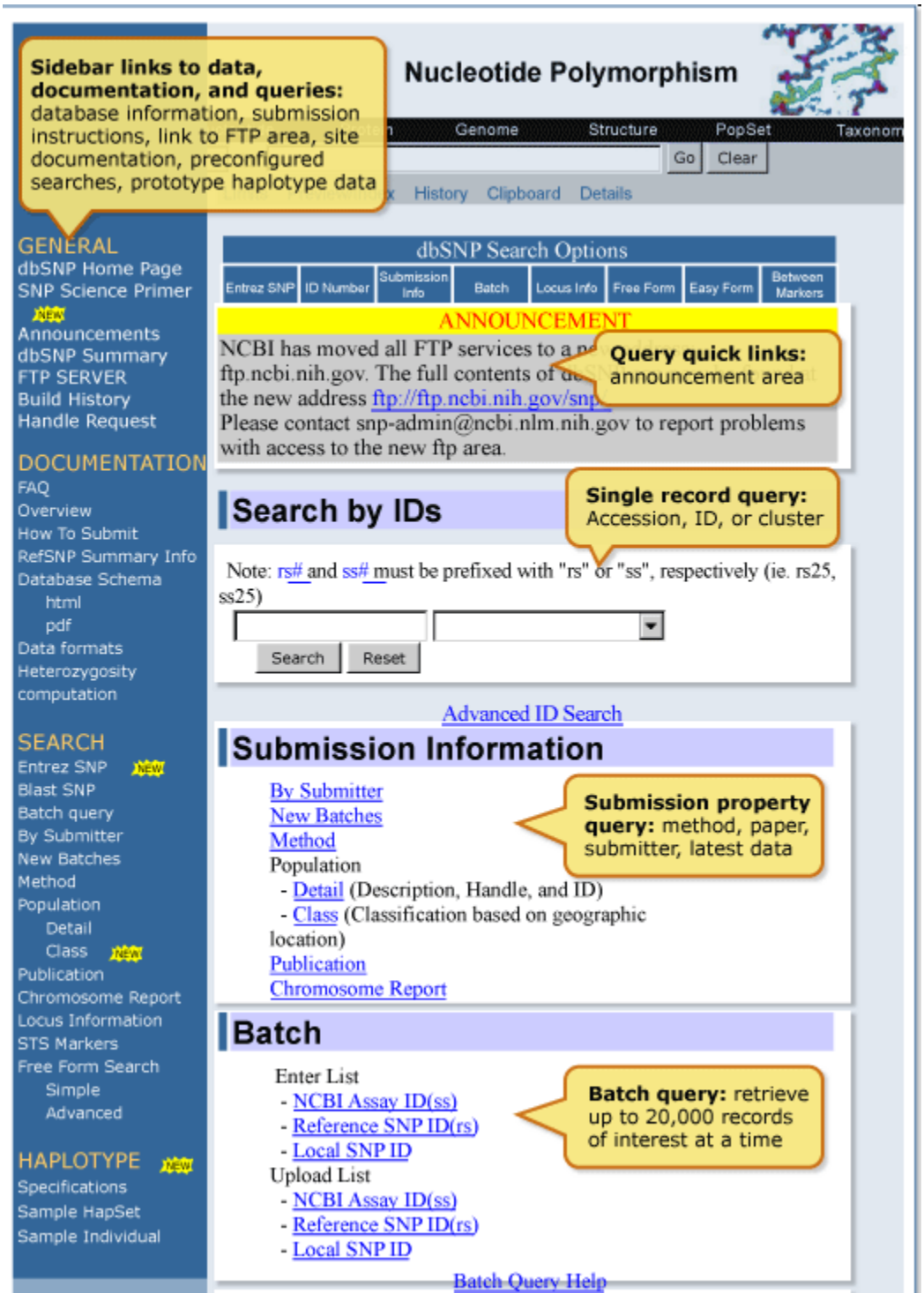


Figure 2a. We organized the dbSNP homepage with links to documentation, FTP, and sub-query pages on the *left sidebar* and a selection of query modules on the *right sidebar*.

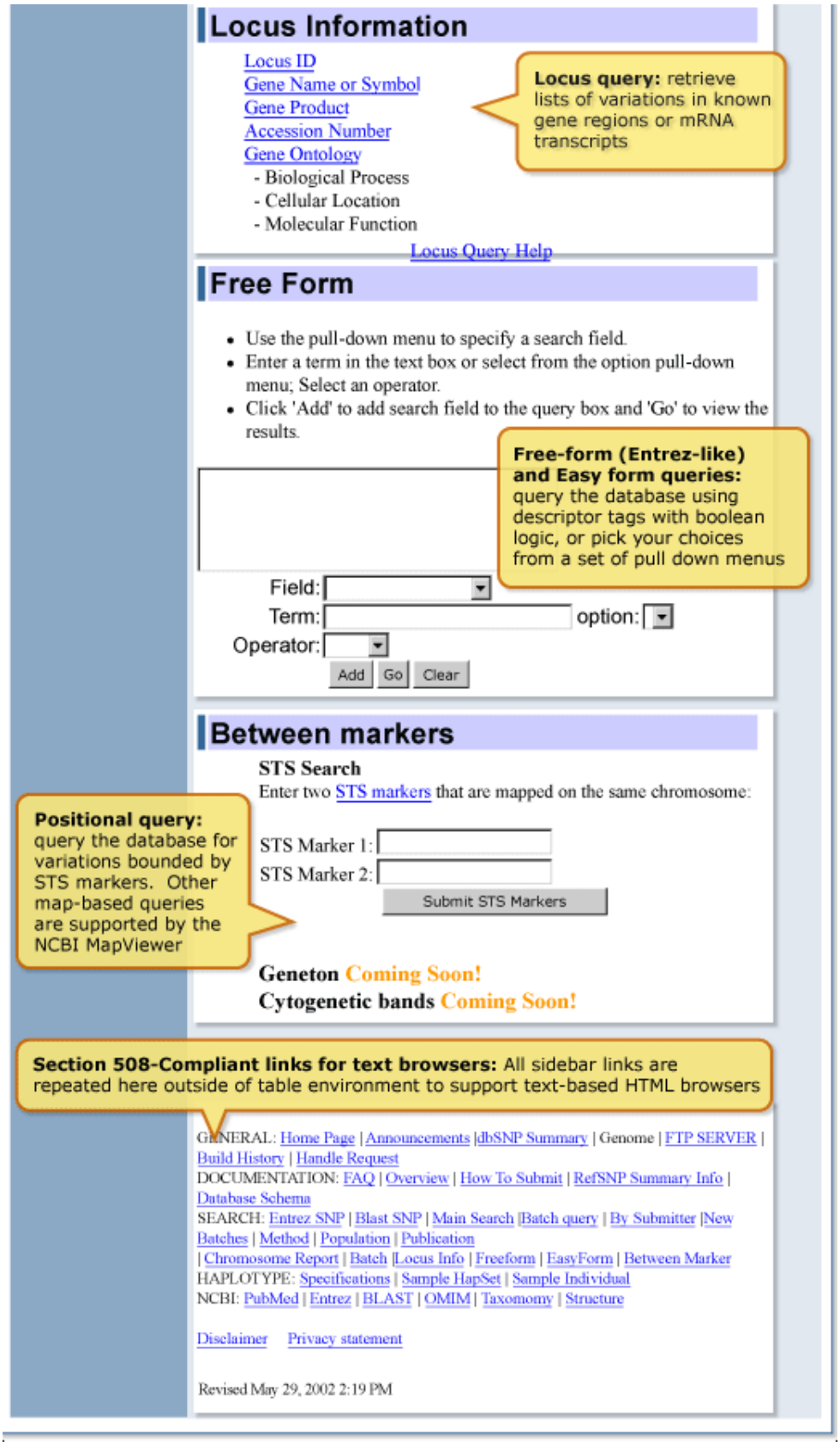


Figure 2b. We organized the dbSNP homepage with links to documentation, FTP, and sub-query pages on the left sidebar and a selection of query modules on the right sidebar.

SNP Submission Information Queries

Use this module to construct a query that will select SNPs based on submission records by laboratory (submitter), new data (called "new batches" — this query limitation is more recent than a user-specified date), the methods used to assay for variation (Table 1) populations of interest (Table 2), and publication information.

Table 1. Method classes organize submissions by a general methodological or experimental approach to assaying for variation in the DNA sequence.

Method class	Class code in Sybase, ASN.1, and XML
Denaturing high pressure liquid chromatography (DHPLC)	1
DNA hybridization	2
Computational analysis	3
Single-stranded conformational polymorphism (SSCP)	5
Other	6
Unknown	7
Restriction fragment length polymorphism (RFLP)	8
Direct DNA sequencing	9

Table 2. Population classes organize population samples by geographic region.

Population class	Description	Population class in Sybase, ASN.1, and XML
Central Asia	Samples from Russia and its satellite Republics and from nations bordering the Indian Ocean between East Asia and the Persian Gulf regions.	8
Central/South Africa	Samples from nations south of the Equator, Madagascar, and neighboring island nations.	4
Central/South America	Samples from Mainland Central and South America and island nations of the western Atlantic, Gulf of Mexico, and Eastern Pacific.	10
East Asia	Samples from eastern and south eastern Mainland Asia and from Northern Pacific island nations.	6
Europe	Samples from Europe north and west of Caucasus Mountains, Scandinavia, and Atlantic islands.	5

Table 2. continues on next page...

Table 2. continued from previous page.

Multi-National	Samples that were designated to maximize measures of heterogeneity or sample human diversity in a global fashion. Examples include OEFNER|GLOBAL and the CEPH repository.	1
North America	All samples north of the Tropic of Cancer, including defined samples of United States Caucasians, African Americans, Hispanic Americans, and the NHGRI polymorphism discovery resource (NCBI| NIHPDR).	9
North/East Africa and Middle East	Samples collected from North Africa (including the Sahara desert), East Africa (south to the Equator), Levant, and the Persian Gulf.	2
Pacific	Samples from Australia, New Zealand, Central and Southern Pacific Islands, and Southeast Asian peninsular/island nations.	7
Unknown	Samples with unknown geographic provinces that are not global in nature.	11
West Africa	Sub-Saharan nations bordering the Atlantic north of the Congo River and central/southern Atlantic island nations.	3

dbSNP Batch Query

Use sets of variation IDs (including RefSNP (rs) IDs, Submitted SNP (ss) IDs, and Local SNP IDs) collected from other queries to generate a variety of SNP reports.

Locus Information Query

This search was originally accomplished by LocusLink, which has now been replaced by Entrez Gene. Entrez Gene is the successor to LocusLink and has two major differences that differentiate it from Locus Link: Entrez Gene is greater in scope (more of the genomes represented by NCBI Reference Sequences or RefSeqs) and Entrez Gene has been integrated for indexing and query in NCBI's Entrez system.

Between-Markers Positional Query

Use this query approach if you are interested in retrieving variations that have been mapped to a specific region of the genome bounded by two STS markers. Other map-based queries are available through the NCBI Map Viewer tool.

ADA Section 508-compliant Link

All links located on the left sidebar of the dbSNP homepage are also provided in text format at the bottom of the page to support browsing by text-based Web browsers. Suggestions for improving database access by disabled persons should be sent to the dbSNP development group at snp-admin@ncbi.nlm.nih.gov.

Submitted Content

The SNP database has two major classes of content: the first class is submitted data, i.e., original observations of sequence variation; and the second class is computed content, i.e., content generated during the dbSNP "build" cycle by computation on original submitted data. Computed content consists of refSNPs, other computed data, and links that increase the utility of dbSNP.

A complete copy of the SNP database is publicly available and can be downloaded from the SNP FTP site (see the section *How to Create a Local Copy of dbSNP*). dbSNP accepts submissions from public laboratories and private organizations. (There are online instructions for preparing a submission to dbSNP.) A short tag or abbreviation called Submitter HANDLE uniquely defines each submitting laboratory and groups the submissions within the database. The 10 major data elements of a submission follow.

Flanking Sequence Context DNA or cDNA

The essential component of a submission to dbSNP is the nucleotide sequence itself. dbSNP accepts submissions as either genomic DNA or cDNA (i.e., sequenced mRNA transcript) sequence. Sequence submissions have a minimum length requirement to maximize the specificity of the sequence in larger contexts, such as a reference genome sequence. We also structure submissions so that the user can distinguish regions of sequence actually surveyed for variation from regions of sequence that are cut and pasted from a published reference sequence to satisfy the minimum-length requirements. Figure 1 shows the details of flanking sequence structure.

Alleles

Alleles define variation class (Table 3). In the dbSNP submission scheme, we define single-nucleotide variants as G, A, T, or C. We do not permit ambiguous IUPAC codes, such as N, in the allele definition of a variation. In cases where variants occur in close proximity to one another, we permit IUPAC codes such as N, and in the flanking sequence of a

variation, we actually encourage them. See Table 3 for the rules that guide dbSNP post-submission processing in assigning allele classes to each variation.

Table 3. Allele definitions define the class of the variation in dbSNP.

dbSNP variation class[a, b]	Rules for assigning allele classes	Sample allele definition	Class code in Sybase, ASN.1, and XML[c]
Single Nucleotide Variations (SNVs)[a]	Strictly defined as single base substitutions involving A, T, C, or G. Formerly called "SNP". Name changed to "SNV" to emphasize that the dbSNP database contains both rare and polymorphic variants.	A/T	1
Deletion/Insertion Variations (DIVs)[a]	Designated using the full sequence of the insertion as one allele, and either a fully defined string for the variant allele or a "-" character to specify the deleted allele. This class will be assigned to a variation if the variation alleles are of different lengths or if one of the alleles is deleted ("-"). Formerly called "DIP". Name changed to "DIV" to emphasize that the dbSNP database contains both rare and polymorphic variants.	T/-/CCTA/G	2
Heterozygous sequence[a]	The term heterozygous is used to specify a region detected by certain methods that do not resolve the polymorphism into a specific sequence motif. In these cases, a unique flanking sequence must be provided to define a sequence context for the variation.	(heterozygous)	3

Seven of the classes apply to both submissions of variations (submitted SNP assay, or ss#) and the non-redundant refSNP clusters (rs#'s) created in dbSNP.
The "Mixed" class is assigned to refSNP clusters that group submissions from different variation classes.
Class codes have a numeric representation in the database itself and in the export versions of the data (ASN. 1 and XML).

Table 3. continues on next page...

Table 3. continued from previous page.

Microsatellite or short tandem repeat (STR)[a]	Alleles are designated by providing the repeat motif and the copy number for each allele. Expansion of the allele repeat motif designated in dbSNP into full-length sequence will be only an approximation of the true genomic sequence because many microsatellite markers are not fully sequenced and are resolved as size variants only.	(CAC)8/9/10/11	4
Named variant[a]	Applies to insertion/deletion polymorphisms of longer sequence features, such as retroposon dimorphism for Alu or line elements. These variations frequently include a deletion "-" indicator for the absent allele.	(alu) / -	5
No-variation[a]	Reports may be submitted for segments of sequence that are assayed and determined to be invariant in the sample.	(NoVariation)	6
Mixed[b]		Mix of other classes	7
Multi-Nucleotide Variation (MNV)[a]	Assigned to variations that are multi-base variations of a single, common length. Formerly called "MNP". Name changed to "MNV" to emphasize that the dbSNP database contains both rare and polymorphic variants.	GGA/AGT	8

Seven of the classes apply to both submissions of variations (submitted SNP assay, or ss#) and the non-redundant refSNP clusters (rs#'s) created in dbSNP.
The "Mixed" class is assigned to refSNP clusters that group submissions from different variation classes.
Class codes have a numeric representation in the database itself and in the export versions of the data (ASN.1 and XML).

Method

Each submitter defines the methods in their submission as either the techniques used to assay variation or the techniques used to estimate allele frequencies. We group methods by method class (Table 1) to facilitate queries using general experimental technique as a query field. The submitter provides all other details of the techniques in a free-text

description of the method. Submitters can also use the METHOD_EXCEPTION_ field to describe changes to a general protocol for particular sets of data (batch-specific details). Submitters generally define methods only once in the database.

Population

Each submitter defines population samples either as the group used to initially identify variations or as the group used to identify population-specific measures of allele frequencies. These populations may be one and the same in some experimental designs. We assign populations a population class (Table 2) based on the geographic origin of the sample. These broad categories provide a general framework for organizing the approximately 700 (as of this writing) sample descriptions in dbSNP. Similar to method descriptions, population descriptions minimally require the submitter to provide a Population ID and a free-text description of the sample.

Sample Size

There are two sample-size fields in dbSNP. One field is called SNPASSAY SAMPLE SIZE, and it reports the number of chromosomes in the sample used to initially ascertain or discover the variation. The other sample size field is called SNPPOPUSE SAMPLE SIZE, and it reports the number of chromosomes used as the denominator in computing estimates of allele frequencies. These two measures need not be the same.

Population-specific Allele Frequencies

Alleles typically exist at different frequencies in different populations; a very common allele in one population may be quite rare in another population. Also, allelic variants can emerge as private polymorphisms when particular populations have been reproductively isolated from neighboring groups, as is the case with religious isolates or island populations. Frequency data are submitted to dbSNP as allele counts or binned frequency intervals, depending on the precision of the experimental method used to make the measurement. dbSNP contains records of allele frequencies for specific population samples defined by each submitter (Table 4).

Table 4. Validation status codes summarize the available validation data in assay reports and refSNP clusters.

Validation evidence	Description	Code in database for ss#	Code in FTP dumps for ss#	Code in database for rs#	Code in FTP dumps for rs#
Not validated	For ss#, no batch update or validation data, no frequency data (or frequency is	0	Not present	0[a]	Not present

If the rs# has a single ss# with code 1, then rs# is set to code 0.
For a single member rs where the ss# validation status = 1, the rs# validation status is set to 0.

Table 4. continues on next page...

Table 4. continued from previous page.

	0 or 1). rs# status code is OR'd from the ss# codes.				
Multiple reporting	Status = 1 for an rs# with at least two ss# numbers; having at least one ss# is validated by a non-computational method. For a ss#, status = 1 if the method is non-computational.	1	1[b]	1,0[b]	1
With frequency	Frequency data is present with a value between 0 and 1.	2	2	2	2
Both frequency	For ss#, the method is non-computational and frequency data is present. If the ss# is a single cluster member, then the rs# code is set to 2.	3	3	3/2	3
Submitter validation	Submission of a batch update or validation section that reports a second validation method on the assay.	4	4	4	4

If the rs# has a single ss# with code 1, then rs# is set to code 0.
For a single member rs where the ss# validation status = 1, the rs# validation status is set to 0.

Population-specific Genotype Frequencies

Similar to alleles, genotypes have frequencies in populations that can be submitted to dbSNP.

Population-specific Heterozygosity Estimates

Some methods for detection of variation (e.g., denaturing high-pressure liquid chromatography or DHPLC) can recognize when DNA fragments contain a variation without resolving the precise nature of the sequence change. These data define an empirical measure of heterozygosity when submitted to dbSNP.

Individual Genotypes

dbSNP accepts individual genotypes for samples from publicly available repositories such as CEPH or Coriell. Genotypes reported in dbSNP contain links to population and

method descriptions. General genotype data provide the foundation for individual haplotype definitions and are useful for selecting positive and negative control reagents in new experiments.

Validation Information

dbSNP accepts individual assay records (ss numbers) without validation evidence. When possible, however, we try to distinguish high-quality validated data from unconfirmed (usually computational) variation reports. Assays validated directly by the submitter through the VALIDATION section show the type of evidence used to confirm the variation. Additionally, dbSNP will flag an assay as validated (Table 4) when we observe frequency or genotype data for the record.

Computed Content (The dbSNP Build Cycle)

We release the content of dbSNP to the public in periodic "builds" that we synchronize with the release of new genome assemblies for each organism (Chapter 14). During each build, we map both the data submitted since the last build and the current refSNP set to the genome. The following 7 tasks define the sequence of steps in the dbSNP build cycle (Figure 3).

Submitted SNPs and Reference SNP Clusters

Once a new SNP is submitted to dbSNP, it is assigned a unique submitted SNP ID number (ss#). Once the ss number is assigned, we align the flanking sequence of each submitted SNP to its appropriate genomic contig. If several ss numbers map to the same position on the contig, we cluster them together, call the cluster a "reference SNP cluster", or "refSNP", and provide the cluster with a unique RefSNP ID number (rs#). If only one ss number maps to a specific position, then that ss is assigned an rs number and is the only member of its RefSNP cluster until another submitted SNP is found that maps to the same position.

A refSNP has a number of summary properties that are computed over all cluster members (Figure 4), and are used to annotate the variations contained in other NCBI resources. We export the entire dbSNP refSNP set in many report formats on the FTP site, and deliver them as sets of results when a user conducts a dbSNP batch query. We also maintain both refSNPs and submitted SNPs in FASTA databases for use in BLAST searches of dbSNP.

New Submissions and the Start of a New Build

Each build starts with a "close of data" that defines the set of new submissions that will be mapped to genome sequence by multiple cycles of BLAST and MegaBLAST for subsequent annotation and grouping of the SNPs into refSNPs. The set of new data entering each build typically includes all submissions received since the close of data in the previous build.

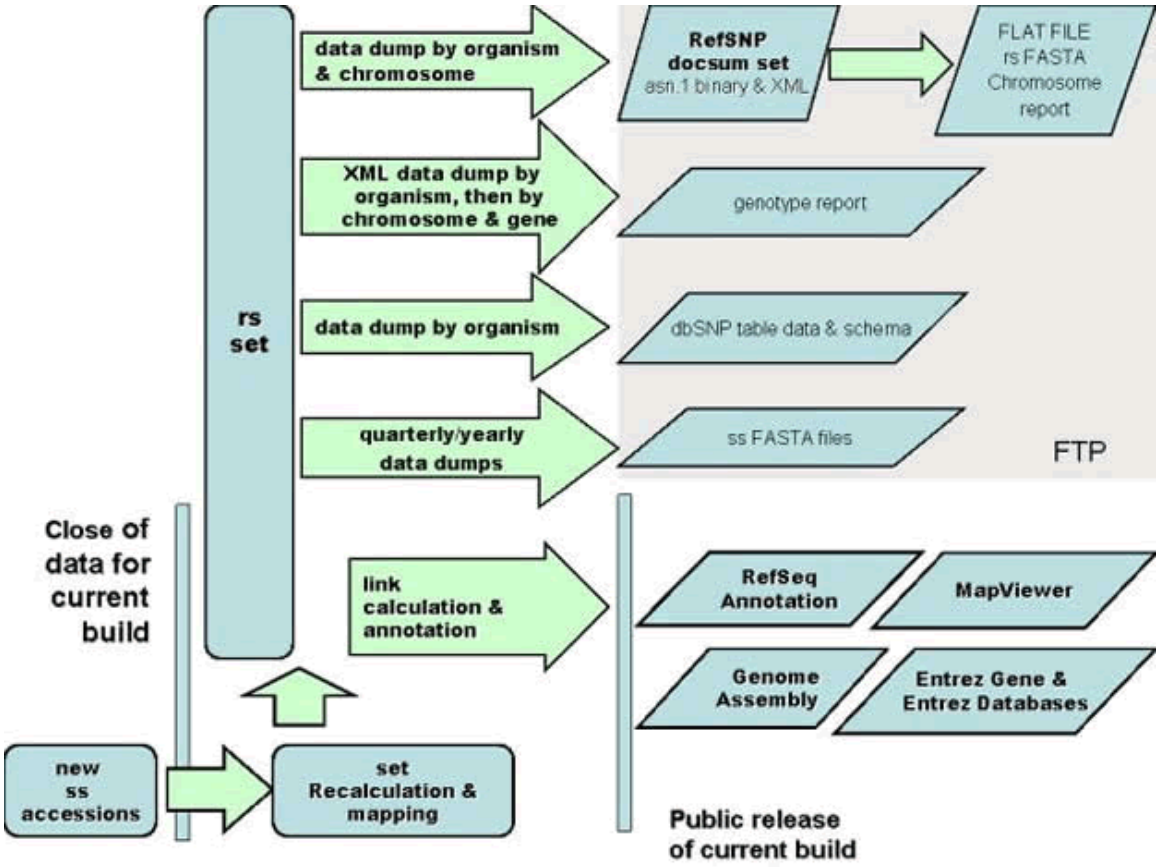


Figure 3. The dbSNP build cycle starts with close of data for new submissions. We map all data, including existing refSNP clusters and new submissions, to reference genome sequence if available for the organism. Otherwise, we map them to non-redundant DNA sequences from GenBank. We then use map data on co-occurrence of hit locations to either merge submissions into existing clusters or to create new clusters. We then annotate the new non-redundant refSNP (rs) set on reference sequences and dump the contents of dbSNP into a variety of comprehensive formats on the dbSNP FTP site for release with the online build of the database.

Mapping to a Genome Sequence

When a new genome build is ready, dbSNP obtains the FASTA files for submitted SNPs that were submitted prior to the "close of data", as well as the FASTA files for the refSNPs in the current build, and then maps the submitted SNPs and refSNPs to the genome sequence using the procedure described in Appendix 2. The refSNP set is also mapped to the previous genome assembly to support users who require older mapping data (e.g. during the production cycle for dbSNP human build 126, The refSNP set was mapped to both human build 36.1 and human build 35.1).

It should also be mentioned that during a build cycle, some organisms have refSNPs mapped to multiple assemblies (e.g. human has two major assemblies: the NCBI Reference Genome assembly and the Celera assembly).

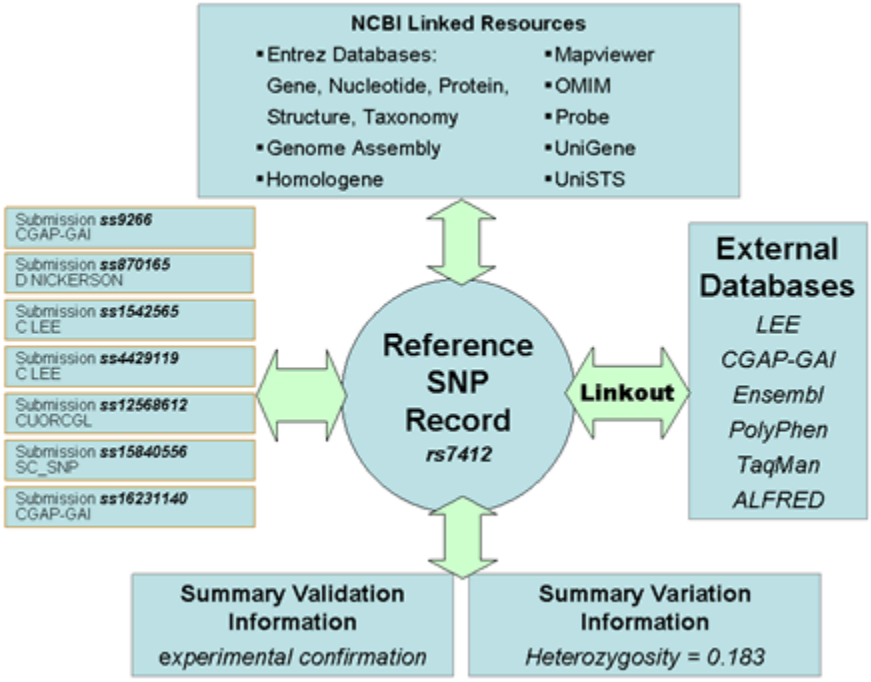


Figure 4. rs7412 has an average heterozygosity of 18.3% based on the frequency data provided by seven submissions, and the cluster as a whole is validated because at least one of the underlying submissions has been experimentally validated. rs7412 is annotated as a variation feature on RefSeq contigs, mRNAs, and proteins. Pointers in the refSNP summary record direct the user to additional information on six additional Web sites, through linkout URLs. These Web sites may contain additional data that were used in the initial variation call, or it may be additional phenotype or molecular data that indicate the function of the variation.

refSNP Clustering and refSNP Orientation

Since submitters to dbSNP can arbitrarily define variations on either strand of DNA sequence, submissions in a refSNP cluster can be reported on the forward or reverse strand. The orientation of the refSNP, and hence its sequence and allele string, is set by a cluster exemplar. By convention, the cluster exemplar is the member of a cluster that has longest sequence. In subsequent builds, this sequence may be in reverse orientation to the current orientation of the refSNP. When this occurs, we preserve the orientation of the refSNP by using the reverse complement of the cluster exemplar to set the orientation of the refSNP sequence.

Once the clustering process determines the orientation of all member sequences in a cluster, it will gather a comprehensive set of alleles for a refSNP cluster.

Hint:
When the alleles of a submission appear to be different from the alleles of its parent refSNP, check the orientation of the submission for reverse orientation.

Re-Mapping and refSNP Merging

RefSNPs are operationally defined as a variation at a location on an ideal reference chromosome. Such reference chromosomes are the goal of genome assemblies. However, since work is still in progress on many of the genome projects, and since even the'finished' genomes are updated to correct past annotation errors, we currently define a refSNP as a variation in the interim reference sequence. Every time there is a genomic assembly update, the interim reference sequence may change, so the refSNPs must be updated or re-clustered.

The re-clustering process begins when NCBI updates the genomic assembly. All existing refSNPs (rs) and newly submitted SNPs (ss) are mapped to the genome assembly using multiple BLAST and MegaBLAST cycles as delineated in Appendix 2. We then cluster SNPs that co-locate at the same place on the genome into a single refSNP. Newly submitted SNPs can either co-locate to form a new refSNP cluster, or may instead cluster with an already existing refSNP. When newly submitted SNPs cluster among themselves, they are assigned a new refSNP number, and when they cluster with an already existing refSNP, they are assigned to that refSNP cluster.

Sometimes an existing refSNP will co-locate with another refSNP when dbSNP begins using an improved clustering algorithm, or when genome assemblies change between builds. A refSNP co-locates with another refSNP only if the mapped chromosome positions of the two refSNPs are identical. So when dbSNP uses an improved clustering algorithm that enhances our ability to more precisely place refSNPs, if the new placement of a refSNP is identical in location to another refSNP, the two refSNPs co-locate. Similarly, if a change in a genome assembly alters the position of a refSNP so that it is identical with the position of another refSNP, the two refSNPs co-locate. When two existing refSNPs co-locate, the refSNP with the higher refSNP number is retired (which means we never reuse it), and all the submitted SNPs in that higher refSNP cluster number are re-assigned to the refSNP with the lower refSNP number. The re-assignment of the submitted SNPs from a higher refSNP number to a refSNP cluster with a lower refSNP number is called a "merge", and occurs during the "rs merge" step of the dbSNP mapping process. Merging is only used to reduce redundancy in the catalog of rs numbers so that each position has a unique identifier. All "rs merge" actions that occur are recorded and tracked.

There is an important exception to the merge process described above; this exception occurs when a co-locating SNP meets certain clinical and publication criteria. A refSNP meeting these criteria is termed "precious" and will keep its original refSNP number (the refSNP number will NOT retire as discussed above) if it co-locates with a SNP that has a lower refSNP number. The purpose of having "precious" SNPs is to maintain refSNP number continuity for those SNPs that have been cited in the literature and are clinically important.

Once the clusters are formed, the variation of a refSNP is the union of all possible alleles defined in the set of submitted SNPs that compose the cluster.

Please Note: dbSNP only merges rs numbers that have an identical set of mappings to a genome and the same allele type (e.g. both must be the same variation type and share one allele in common). We therefore would not merge a SNP and an indel (insertion/deletion) into a single rs number (different variation classes) since they represent two different types of mutational "events".

RefSNP Number Stability

The stability of a refSNP number depends on what is meant by “stable”. If a refSNP number has been merged into another refSNP number, it is very easy to use a retired refSNP number to find the current refSNP number (see hint below) — so in this case, one could consider a refSNP number to be stable since merged refSNP numbers are always associated with, and can always retrieve the current refSNP number.

Hint:
There are three ways the you can locate the partner numbers of a merged refSNP: • If you enter a retired rs number into the “Search for IDs” search text box on the dbSNP home page, the response page will state that the SNP has been merged, and will provide the new rs number and a link to the refSNP page for that new rs number. • You can retrieve a list of merged rs numbers from Entrez SNP. Just type “mergedrs” (without the quotation marks) in the text box at the top of the page and click the “go” button.). You can limit the output to merged rs numbers within a certain species by clicking on the “Limits” tab and then selecting the organism you wish from the organism selection box. Each entry in the returned list will include the old rs numbers that has merged, and the new rs number it has merged into (with a link to the refSNPpage for the new rs number). • You can also review the RsMergeArch table for the merge partners of a particular species of interest, as it tracks all merge events that occur in dbSNP. This table is available on the dbSNP FTP site, a full description of of it can be found in the dbSNP Data Dictionary, and the column definitions are located in the dbSNP_main_table.sql.gz, which can be found in the shared_schema directory of the dbSNP FTP site.

If, however, what is meant by “stable” is that the refSNP number of a particular variation always remains the same, then one should not consider a refSNP entirely stable, as a refSNP number may change if two refSNP numbers merge as described above. Merging is more likely to happen, however, if the submitted flanking sequence of the refSNP exemplar is short, is of low quality, or if the genome assembly is immature. A refSNP number may also change if:

- All of the submitted SNP (ss) numbers in a refSNP cluster are withdrawn by the submitter(a less than 1 in 100 occurrence)
- dbSNP breaks up an existing cluster and re-instantiates a retired rs number based on a reported conflict from a dbSNP user (a less than 1 in 1,000,000 occurrence)

Functional Analysis

Variation Functional Class

We compute a functional context for sequence variations by inspecting the flanking sequence for gene features during the contig annotation process, and do the same for RefSeq/GenBank mRNAs.

Table 5 defines variation functional classes. We base class on the relationship between a variation and any local gene features. If a variation is near a transcript or in a transcript interval, but not in the coding region, then we define the functional class by the position of the variation relative to the structure of the aligned transcript. In other words, a variation may be near a gene (locus region), in a UTR (mRNA-utr), in an intron (intron), or in a splice site (splice site). If the variation is in a coding region, then the functional class of the variation depends on how each allele may affect the translated peptide sequence.

Typically, one allele of a variation will be the same as the contig (contig reference), and the other allele will be either a synonymous change or a non-synonymous change. In some cases, one allele will be a synonymous change, and the other allele will be a non-synonymous change. If either allele in the variation is a non-synonymous change, then the variation is classified as non-synonymous; otherwise, the variation is classified as a synonymous variation. The primary functional classifications are as follows:

- The functional class is noted as Contig Reference when the allele is identical to the contig (contig reference), and hence causes no change to the translated sequence.
- The functional class is noted as synonymous substitution when an allele that is substituted for the reference sequence yields a new codon that encodes the same amino acid.
- The functional class is noted as non-synonymous substitution when an allele that is substituted for the reference sequence yields a new codon that encodes a different amino acid.
- The functional class is noted as coding when a problem with the annotated coding region feature prohibits conceptual translation. The coding notation is based solely on position in this case.

Because functional classification is defined by positional and sequence parameters, two facts emerge: (*a*) if a gene has multiple transcripts because of alternative splicing, then a variation may have several different functional relationships to the gene; and (*b*) if multiple genes are densely packed in a contig region, then a variation at a single location in the genome may have multiple, potentially different, relationships to its local gene neighbors.

Table 5. Function codes for refSNPs in gene features. [a]

Functional class	Description	Database code
Locus region	Variation is within 2 Kb 5′ or 500 bp 3′ of a gene feature (on either strand), but the variation is not in the transcript for the gene. This class is indicated with an "L" in graphical summaries. **As of build 127, function code 1 has been modified into a two digit code** that will more precisely indicate the location of a SNP. The two digit code has function code 1as the first digit, which will keep the meaning as described above, and 3 or 5 as the second digit, which will indicate whether the SNP is 3' or 5' to the region of interest. See function codes 13 and 15 in this table.	1
Coding	Variation is in the coding region of the gene. This class is assigned if the allele-specific class is unknown. This class is indicated with a C in graphical summaries. **This code was retired as of dbSNP build 127**.	2
Coding-synon	The variation allele is synonymous with the contig codon in a gene. An allele receives this class when substitution and translation of the allele into the codon makes no change to the amino acid specified by the reference sequence. A variation is a synonymous substitution if all alleles are classified as contig reference or coding-synon. This class is indicated with a C in graphical summaries.	3
Coding-nonsynon	The variation allele is nonsynonymous for the contig codon in a gene. An allele receives this class when substitution and translation of the allele into the codon changes the amino acid specified by the reference sequence. A variation is a nonsynonymous substitution if any alleles are classified as coding-nonsynon. This class is indicated with a C or N in graphical summaries. **As of build 128, function code 4 has been modified into a two digit code** that will more precisely indicate the nonsynonymous nature of the SNP. The two digit code has function code 4 as the first digit, which will keep its original meaning, and 1, 2, or 4 as the second digit, which will indicate whether the SNP is nonsense, missense, or frameshift. See function codes 41, 42, and 44 in this table	4
mRNA-UTR	The variation is in the transcript of a gene but not in the coding region of the transcript. This class is indicated by a "T" in graphical summaries. **As of build 127, function code 5 has been modified into a two digit code** that will more precisely indicate the location of a SNP. The two digit code has function code 5 as the first digit, which will keep its original meaning, and 3 or 5 as the second digit, which will indicate whether the SNP is 3' or 5' to the region of interest. See function codes 53 and 55 in this table.	5

Most gene features are defined by the location of the variation with respect to transcript exon boundaries. Variations in coding regions, however, have a functional class assigned to each allele for the variation because these classes depend on allele sequence.

Table 5. continues on next page...

Table 5. continued from previous page.

Intron	The variation is in the intron of a gene but not in the first two or last two bases of the intron. This class is indicated by an L in graphical summaries.	6
Splice-site	The variation is in the first two or last two bases of the intron. This class is indicated by a "T" in graphical summaries. **As of build 127, function code 7 has been modified into a two digit code** that will more precisely indicate the location of a SNP. The two digit code has function codes 7 as the first digit, which will keep its original meaning, and 3 or 5 as the second digit, which will indicate whether the SNP is 3' or 5' to the region of interest. See function codes 73 and 75 in this table.	7
Contig-reference	The variation allele is identical to the contig nucleotide. Typically, one allele of a variation is the same as the reference genome. The letter used to indicate the variation is a C or N, depending on the state of the alternative allele for the variation.	8
Coding-exception	The variation is in the coding region of a gene, but the precise location cannot be resolved because of an error in the alignment of the exon. The class is indicated by a C in graphical summaries.	9
NearGene–3	Function Code 13, where: 1=locus region (see function code 1 in this table) 3=SNP is 3'to and 0.5kb away from gene	13
NearGene–5	Function Code 15, where: 1=locus region (see function code 1 in this table) 5= SNP is 5' to 2kb away from gene	15
Coding-nonsynonymous nonsense	Function Code 41, where: 4 =Coding- nonsynonymous (see function code 4 in this table) 1 = Nonsense (changes to the Stop codon)	41
Coding-nonsynonymous missense	Function Code 42, where: 4 =Coding- nonsynonymous (see function code 4 in this table) 2 = missense (alters codon to make an altered amino acid in protein product)	42
Coding-nonsynonymous frameshift	Function Code 44, where: 4 =Coding- nonsynonymous (see function code 4 in this table) 4 = frameshift (alters codon to make an altered amino acid in protein product)	44
UTR–3	Function code 53, where: 5= UTR (untranslated region: see function code 5 in this table) 3= SNP located in the 3' untranslated region	53

Most gene features are defined by the location of the variation with respect to transcript exon boundaries. Variations in coding regions, however, have a functional class assigned to each allele for the variation because these classes depend on allele sequence.

Table 5. continues on next page...

Table 5. continued from previous page.

UTR–5	Function code 55, where: 5= UTR (untranslated region: see function code 5 in this table) 5(as the second digit)= SNP located in the 5' untranslated region	55
Splice–3	Function code 73, where: 7=splice site (see function code 7 in this table) 3=3' acceptor dinucleotide	73
Splice–5	Function code 75, where: 7=splice site (see function code 7 in this table) 5=5' donor dinucleotide	75

Most gene features are defined by the location of the variation with respect to transcript exon boundaries. Variations in coding regions, however, have a functional class assigned to each allele for the variation because these classes depend on allele sequence.

SNP Position in 3D Structure

When a SNP results in amino acid sequence change, knowing where that amino acid lies in the protein structure is valuable. We provide this information using the following procedure: To find the location of a SNP within a particular protein, we attempt to identify similar proteins whose structure is known by comparing the protein sequence against proteins from the PDB database of known protein structures using BLAST. Then, if we find matches, we use the BLAST alignment to identify the amino acid in the protein of known structure that corresponds to the amino acid containing the SNP. We store the position of the amino acid on the 3D structure that corresponds to the amino acid containing the SNP in the dbSNP table SNP3D.

Population Diversity Data

The best single measure of a variation's diversity in different populations is its average heterozygosity. This measure serves as the general probability that both alleles are in a diploid individual or in a sample of two chromosomes. Estimates of average heterozygosity have an accompanying standard error based on the sample sizes of the underlying data, which reflects the overall uncertainty of the estimate. dbSNP's computation of average heterozygosity and standard error for RefSNP clusters is available online. Please note that dbSNP computes heterozygosity based on the submitted allele frequency for each SNP. If the frequency data for a SNP is not submitted, we cannot compute the heterozygosity value, and therefore the refSNP report will show no heterozygosity estimate.

Additional population diversity data include population counts, individuals sampled for a variation, genotype frequencies, and Hardy Weinberg probabilities.

Build Resource Integration

We annotate the non-redundant set of variations (refSNP cluster set) on reference genome sequence contigs, chromosomes, mRNAs, and proteins as part of the NCBI RefSeq project (Chapter 18). We compute summary properties for each refSNP cluster, which we then

use to build fresh indexes for dbSNP in the Entrez databases, and to update the variation map in the NCBI Map Viewer. Finally, we update links between dbSNP and dbMHC, OMIM, Homologene, the NCBI Probe database, PubMed, UniGene, and UniSTS.

Annotating GenBank and Other RefSeq Records

GenBank records can be annotated only by their original authors. Therefore, when we find high-quality hits of refSNP records to the HTGS and non-redundant divisions of GenBank, we connect them using LinkOut (Chapter 17).

We annotate RefSeq mRNAs with variation features when the refSNP has a high-quality hit to the mRNA sequence. If the variation is in the coding region of the transcript and has a non-synonymous allele that changes the protein sequence, we also annotate the variation on the protein translation of the mRNA. The alleles in protein annotations are the amino acid translations of the affected codons.

NCBI Map Viewer Variation and Linkage Maps

The Map Viewer (Chapter 20) can show multiple maps of sequence features in common chromosome coordinates. The variation map shows all variation features that we annotate on the current genome assembly. There are two ways to see the variation data. The default graphic mode shows the data as tick marks on the vertical coordinate scale. When variation is selected as the master map from the Maps and Options drop-down menu, and the user zooms in on the map to the individual RefSNP level, a summary of map quality, variation quality warning, functional relationships to genes, average heterozygosity with standard error, and validation information are provided. If genotype, haplotype, or LinkOut data are available, the master map will also contain links to this information.

Public Release

Public release of a new build involves an update to the public database and the production of a new set of files on the dbSNP FTP site. We make an announcement to the dbsnp-announce mailing list when the new build for an organism is publicly available.

dbSNP Resource Integration

Links from SNP Records to Submitter Websites

The SNP database supports and encourages connections between assay records (submitted SNP ID numbers, or ss numbers) and supplementary data on the submitter's Web site. This connection is made using the LINKOUT field in the SNPassay batch header. LinkOut URLs are base URLs to which dbSNP can append the submitter's ID for the variation to construct a complete URL to the specific data for the record. We provide LinkOut pointers in the batch header section of SNP detail reports and in the refSNP report cluster membership section.

Links within NCBI

We make the following connections between refSNP clusters and other NCBI resources during the contig annotation process:

Entrez Gene

There are two methods by which we localize variations to known genes: (*a*) if a variation is mapped to the genome, we note the variation/gene relationship (Table 5) during functional classification and store the locus_id of the gene in the dbSNP table SNPContigLocusId; and (*b*) if the variation does not map to the genome, we look for high-quality blast hits for the variation against mRNA sequence. We note these hits with the protein_ID (PID) of the protein (the conceptual translation of the mRNA transcript). Entrez Gene scans this table nightly and updates the table MapLinkPID with the locus_id for the gene when the protein is a known product of a gene.

UniSTS

When an original submitted SNP record shows a relationship between a SNP and a STS, we share the data with dbSTS and establish a link between the SNP and the STS record. We also examine refSNPs for proximity to STS features during contig annotation. When we determine that a variation needs to be placed within an STS feature, we note the relationship in the dbSNP table SnpInSts.

UniGene

The contig annotation pipeline relates refSNPs to UniGene EST clusters based on shared chromosomal location. We store Variation/UniGene cluster relationships in the dbSNP table UnigeneSnp.

PubMed

We connect individual submissions to PubMed record(s) of publications cited at the time of submission. To view links from PubMed to dbSNP, select “linkouts” as a PubMed query result.

dbMHC

dbSNP stores the underlying variation data that define HLA alleles at the nucleotide level. The combinations of alleles that define specific HLA alleles are stored in dbMHC. dbSNP points to dbMHC at the haplotype level, and dbMHC points to dbSNP at both the haplotype and variation level.

How to Create a Local Copy of dbSNP

dbSNP is a relational database that contains hundreds of tables. Since the inception of build 125, the design dbSNP has been altered to a ”hub and spoke” model, where the dbSNP_Main_Table acts as the hub of a wheel, storing all of the central tables of the

database, while each spoke of the wheel is an organism-specific database that contains the latest data for a specific organism. dbSNP exports the full contents of the database for the public to download from the dbSNP FTP site.

Due to security concerns and vendor endorsement issues, we cannot provide users with direct dumps of dbSNP. The task of creating a local copy of dbSNP can be complicated, and should be left to an experienced programmer. The following sections will guide you in the process of creating a local copy of dbSNP, but these instructions assume knowledge of relational databases, and were not written with the novice in mind.

If you have problems establishing a local copy of dbSNP, please contact dbSNP at snp-admin@ncbi.nlm.nih.gov.

Schema: The dbSNP Physical Model

A schema is a necessary part of constructing your own copy of dbSNP because it is a visual representation of dbSNP that shows the logical relationship between data in dbSNP. It is available as a printable PDF file from the dbSNP FTP site.

Data in dbSNP are organized into "subject areas" depending on the nature of the data. The data dictionary currently includes a description of all the tables in dbSNP as well as tables of columns and their properties. Foreign keys are not enforced in the physical model because they make it harder to load table data asynchronously. In the future, we will add descriptions of individual columns. The data dictionary is also available online from the dbSNP Web site.

Resources Required for Creating a Local Copy of dbSNP

Software:

- **Relational database software**. If you are planning to create a local copy of dbSNP, you must first have a relational database server, such as Sybase, Microsoft SQL server, or Oracle. dbSNP at NCBI runs on an MSSQL server version 2000, but we know of users who have successfully created their local copy of dbSNP on Oracle.
- **Data loading tool**. Loading data from the dbSNP FTP site into a database requires a bulk data-loading tool, which usually comes with a database installation. An example of such a tool is the bcp (bulk-copy) utility that comes with Sybase, or the "bulkinsert"command in the MSSQL server.
- **winzip/gzip to decompress FTP files**. Complete instructions on how to uncompress *.gz and *.Z files can be found on the dbSNP FTP site.

Hardware:

- **Computer platforms/OS**. Databases can be maintained on any PC, Mac, or UNIX with an Internet connection.
- **Disk space**. Currently, a complete copy of dbSNP that will include all organisms contained in dbSNP requires 500 GB of space. Depending on the organism you are

interested in, you can simply create a local database that only includes data for the organism of your interest. Please allow room for growth.
- **Memory**. The current sql server for dbSNP has 4GB of memory.
- **Internet connection**. We recommend a high-speed connection to download such large database files.

dbSNP Data Location

The FTP database directory in the dbSNP FTP site contains the schema, data, and SQL statements to create the tables and indices for dbSNP:

- The shared_schema subdirectory contains the schema DDL (SQL Data Definition Language) for the dbSNP_main_table.
- The shared_data subdirectory contains data housed in the dbSNP_main_table that is shared by all organisms.
- The organism_schema sub-directory contains links to the schema DDL for each organism specific database.
- The organism_data sub-directory contains links to the data housed in each organism specific database. The data organized in tables, where there is one file per table. The file name convention is: <tablename>.bcp.gz. The file name convention for the mapping table also includes the dbSNP build ID number and the NCBI genome build ID number. For example, B125_SNPContigLoc_35_1 means that during dbSNP build 125, this SNPContigLoc table has SNPs mapped to NCBI contig build 35 version 1. The data files have one line per table row. Fields of data within each file are tab delimited.

dbSNP uses standard SQL DDL(Data Definition Language) to create tables, views for those tables, and indexes. There are many utilities available to generate table/index creation statements from a database.

Hint
If your firewall blocks passive FTP, you might get an error message that reads: “Passive mode refused. Turning off passive mode. No control connection for command: No such file or directory”. If this happens, try using a "smart" FTP client like NCFTP (available on most UNIX machines). Smart FTP clients are better at auto-negotiating active/passive FTP connections than are older FTP clients (e.g. Sun Solaris FTP).

Stepwise Procedure for Creating a Local Copy of dbSNP

1 **Prepare the local area.**

(check available space, etc.)

2 **Download the schema files.**

a Download the following files from the dbSNP shared_schema sub-directory: dbSNP_main_table, dbSNP_main_index_constraint, and all the files in the

shared_data sub-directory. Together, the files from both of these sub-directories will allow you to create tables and indices for the dbSNP_main_table.

b Go to the organism_schema subdirectory, and select the organism for which you wish to create a database. For the purpose of this example, human_9606 has been selected. Once human_9606 is selected, you will be directed to the human organism_schema sub-directory. Download all of the files contained in this subdirectory.

c Go to the organism_data subdirectory, and select the organism for which you wish to create a database. For the purpose of this example, human_9606 has been selected. Once you select human_9606, you will be directed to the human organism_data sub-directory. Download all of the files contained in this subdirectory.

A user must always down load the files located in the most recent versions of the shared_schema and shared_data sub-directories in addition to any organism specific content.

Save all the files in your local directory and decompress them.

Hint:
On a UNIX operating system, use gunzip to decompress the files: dbSNP_main_table and dbSNP_main_index_constraint. The files on the SNP FTP site are UNIX files. UNIX, MS-DOS and Macintosh text files use different characters to indicate a new line. Load the appropriate new line conversion program for your system before using bcp.

3 **Create the dbSNP_main_table**

a From the shared_schema sub-directory, use the dbSNP_main_table file to create tables, and use the dbSNP_main_index_constraint files to create indices for the dbSNP main database.

b Load all of the bcp files located in the shared_data sub-directory into the dbSNP_main_table you just created using the data-loading tool of your database server (e.g., bcp for Sybase). See the sample FTP protocol and sample Unix C Shell script (below) for directions.

c Create indices by opening the dbSNP_main_index_constraint.sql file. If you are using a database server that provides the isql utility, then use the following command:

```
isql -S <servername> -U usename -P password -i dbSNP_main_index_constraint.sql
```

Hint:
The ".bcp" files in the shared_data and organism_data sub-directories may be loaded into most spreadsheet programs by setting the field delimiter character to "tab".

4 **Create the organism specific database**

Once the dbSNP_main_table has been created, create the organism specific database using the files in your specific organism's organism_schema and organism_data subdirectories. Human_9606 will be used for the purpose of this example:

a Create the human_9606 database using the following files found in the human_9606 organism_schema: human_9606_table.sql.gz, human_9606_view.sql.gz, human_9606_index_constraint.sql.gz,

and human_9606_foreign_key.sql.gz

b Load all of the bcp files located in the shared_data sub-directory into the human_9606 database you just created using the data-loading tool of your database server (e.g., bcp for Sybase). See the sample FTP protocol and sample Unix C shell script (below) for directions.

Hint:
Use "**ftp -i**" to turn off interactive prompting during multiple file transfers to avoid having to hit "yes" to confirm transfer hundreds of times.

Hint:
To avoid an overflow of your transaction log while using the bcp command option (available in Sybase and SQL servers), select the "batch mode" by using the command option: -b number of rows. For example, the command option -b 10000 will cause a commit to the table every 10,000 rows.

5 **Sample FTP Loading protocol.**

a. Type ftp -i ftp.ncbi.nih.gov (Use "anonymous" as user name and your email as your password).
b. Type: cd snp/database
c. To get dbSNP_main for shared tables and shared data: Type ls to see if you are in the directory with the right files. Then type "cd shared_schema" to get schema file for dbSNP_main, and finally, type "cd shared_data" to get the data for dbSNP_main.
d. Type binary (to set binary transfer mode).
e. Type mget *.gz (to initiate transfer). Depending on the speed of the connection, this may take hours since the total transfer size is gigabytes in size and growing.
f. To decompress the *.gz files, type gunzip *.gz. (Currently, the total size of the uncompressed bcp files is over 10 GB).

6 **Use scripts to automate data loading.**

a Located in the loadscript subdirectory of the dbSNP FTP site, there is a file called cmd.create_local_dbSNP.txt that provides a sample UNIX C shell script for creating a local copy of dbSNP_main and a local copy of a specific organism database using files in the shared_schema, and the organism_schema sub-directories.

b Also in the the loadscript subdirectory of the dbSNP FTP site, there is a file called cmd.bulkinsert.txt that provides a sample UNIX C shell script for loading tables with files located in shared_data and organism_data sub-directories.

7 **Data integrity (creating a partial local copy of dbSNP).**

dbSNP is a relational database. Each table has either a unique index or a primary key. Foreign keys are not reinforced. There are advantages and a disadvantage to this approach. The advantages are that this approach makes it easy to drop and re-create the table using the dbSNP_main_table, which then makes it possible to create a partial local copy of dbSNP. For example, if you are interested only in the original submitted SNP and their population frequencies, and not in their map locations on NCBI genome contigs or GenBank Accession numbers (both are huge tables), then these tables can be skipped (i.e., SNPContigLoc and MapLink). Please remember that mapping tables such as SNPContigLoc will have a build ID prefix and suffix included in its file name. (e.g. SNPContigLoc will be b125_SNPContigLoc_35_1for SNP build 125, and NCBI contig build 35 version 1). Of course, to select tables for a particular query, the contents of each table and the dbSNP entity relationship (ER) diagram need to be understood. The disadvantage of un-reinforced references is that either the stored procedures or the external code needs to be written to ensure the referential integrity.

Appendix 1. dbSNP report formats.

ASN.1

The docsum_2005.asn file is the ASN structure definition file for ASN.1 and is located in the /specs subdirectory of the dbSNP FTP site. The 00readme file, located in the main dbSNP FTP directory, provides information about ASN.1 data structure and data exchange. ASN.1 text or binary output can be converted into one or more of the following formats: flatfile, FASTA, docsum, chromosome report, RS/SS, and XML. To convert from ASN.1 to another format, request ASN.1 output from either the dbSNP FTP site or the dbSNP batch query pages, and use dstool (located in the "bin" directory of the dSNP FTP site) to locally convert the output into as many alternative formats as needed.

XML

The XML format provides query-specific information about refSNP clusters,

as well as cluster members in the NCBI SNP Exchange (NSE) format. The XML schema is located in the docsum_2005.xsd file, which is housed in the /specs sub-directory of the

dbSNP FTP site. A human-readable text form of the NSE definitions can be found in docsum_2005.asn, also located in the /specs sub-directory of the dbSNP FTP site.

FASTA: ss and rs

The FASTA report format provides the flanking sequence for each report of variation in dbSNP, as well as all the submitted sequences that have no variation. ss FASTA contains all submitted SNP sequences in FASTA format, whereas rs FASTA contains all the reference SNP sequences in FASTA format. The FASTA data format is typically used for sequence comparisons using BLAST. BLAST SNP is useful for conducting a few sequence comparisons in the FASTA format, whereas multiple FASTA sequence comparisons will require the construction of a local BLAST database of FASTA formatted data and the installation of a local stand-alone version of BLAST.

rs docsum Flatfile

The rs docsum flatfile report is generated from the ASN.1 datafiles and is provided in the files "/ASN1_flat/ds_flat_chXX.flat". Files are generated per chromosome (chXX in file name),as with all of the large report dumps. Because flatfile reports are compact, they will not provide you with as much information as the ASN.1 binary report, but they are useful for scanning human SNP data manually because they provide detailed information at a glance. A full description of the information provided in the rs docsum flatfile format is available in the 00readme file, located in the SNP directory of the SNP FTP site.

Chromosome Reports

The chromosome reports format provides an ordered list of RefSNPs in approximate chromosome coordinates. Chromosome reports is a small file to download but contains a great deal of information that might be helpful in identifying SNPs useful as markers on maps or contigs because the coordinate system used in this format is the same as that used for the NCBI genome Map Viewer. It should also be mentioned that the chromosome reports directory might contain the multi/ file and/or the noton/ files. These files are lists (in chromosome report format) of SNPs that hit multiple chromosomes in the genome and those that did not hit any chromosomes in the genome, respectively. A full description of the information provided in the chromosome reports format is available in the 00readme file, located in the SNP directory of the SNP FTP site.

Genotype Report

The dbSNP Genotype report shows strain-specific genotype information for model organisms, and contains a genotype detail link as well as a genotype XML link. The genotype detail link will provide the user with submitter and genotype information for each of the submitted SNPs in a refSNP cluster of interest, and the genotype XML link will allow the user to download the reported data in the Genotype Exchange XML format, which can be read by either Internet Explorer or Netscape browsers. XML dumps via the

dbSNP ftp server provide the same content for all genotype data in dbSNP by organism and chromosome.

Appendix 2. Rules and methodology for mapping

A cycle of MegaBLAST and Blast alignment to the NCBI genome assembly of an organism is initiated either by the appearance of FASTA-formatted genome sequence for a new build of the assembly or by the significant accrual of newly submitted SNP data for that organism.

Organism-specific Genome Mapping

The refSNP(rs) and submitted SNP (ss) mapping process is a multi-step, computer-based procedure that begins when refSNP and submitted SNP FASTA sets are aligned to the most recent genome assembly using BLAST or MegaBLAST. Repeat masking during this process is accomplished automatically using the BLAST/MegaBLAST "Dust" option. To increase alignment stringency, multiple cycles of BLAST/MegaBLAST are employed, where the word size limit is reduced from 64 in the first cycle to16 by the final cycle. MegaBLAST parameters are set to a default position with the exception of a seeding hit suppression parameter(e.g. Parameter="-U F -F 'mL' -J T -X 10 -r 1 -q -3 -W 64 -m 11 -e 0.01")that suppresses seeding hits but allows extension through regions of lowercase sequence. The quality of each alignment is determined using an Alignment Profiling Function developed specifically for this purpose. Alignments are selected based on a variable stringency threshold that ranges from 70% alignment to 50% alignment. If the alignment profiling function indicates the quality of the alignment is below a 70% alignment threshold, the alignment is discarded, although an alignment threshold of 50% is sometimes used in case there are gaps in the sequence.

The BLAST/MegaBLAST output of ASN.1 binary files of local alignments is then analyzed by an algorithm ("Globalizer") that sorts those local alignments that do not fit the dbSNP alignment profile criteria (defined by position and proximity to one another) to create a "Global Alignment" — a group of local alignments that lay close to one another on a sequence. If the global alignment is greater than or equal to a pre-determined percentage of the flanking sequence, it is accepted as a true alignment between the refSNP or submitted SNP and the genome assembly. The Globalizer is especially helpful when refSNPs or submitted SNPs based on mRNA sequence are being aligned to the genome assembly. In such a case, the MegaBLAST/BLAST ASN.1 binary output contains many small alignments that will not map to the genome assembly unless they undergo the "globalization" process.

The ASN.1 binary output of "Globalizer" is then processed by a program called "Hit Analyzer". This program defines the alleles and LOC types for each hit, and also determines the map position by using the closest map positions on either side of the SNP to establish the hit location. The text output of "Hit Analyzer" is then processed by the "Hit Filter", which filters out paralogous hits and uses multiple strategies to select only

those SNPs that have the greatest degree of alignment to a particular contig. The output from the "Hit Filter" is then placed into a map.bcp file and is processed by the "SnpMapInfo" program, which creates an MD5 signature for each SNP that is representative of all the positional information available for that SNP. The MD5 signature is then placed in the SNP MAP INFO file, which is then loaded into dbSNP.

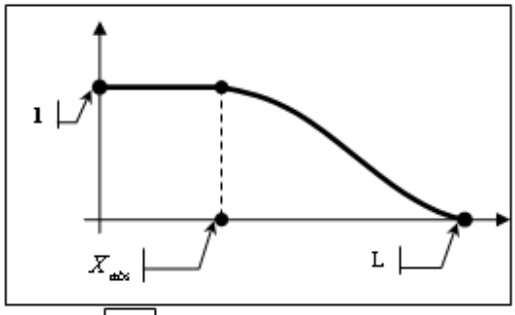
RefSNPs and submitted SNPs are analyzed against GenBank mRNA, RefSeq mRNA, and GenBank clone accessions using a similar procedure to that described in the above paragraphs.

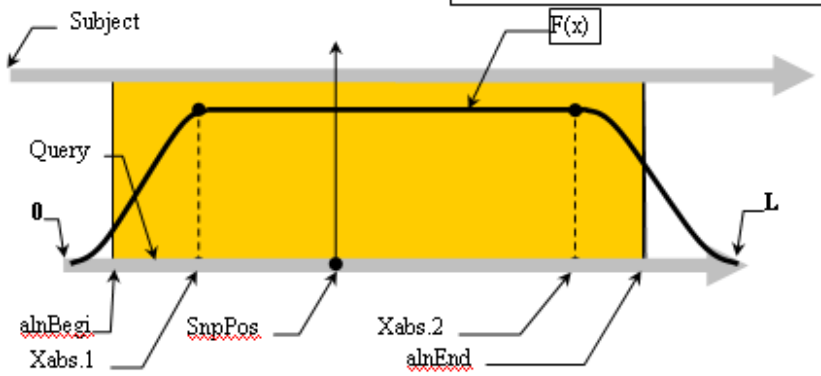
Once the all the results from previous steps are loaded into dbSNP, we perform cluster analysis using a program called "SNPHitCluster" which analyses SNPs having the same signatures to find candidates for clustering. If an MD5 signature for a particular SNP is different from the MD5 signature of another SNP, then the hits for those two SNPs are different, and therefore, the SNPs are unique and need not be clustered. If an MD5 signature of a particular SNP is the same as that of another SNP, the two SNPs may have the same hit pattern, and if after further analysis, the hit patterns are shown to be the same, the two SNPs will be clustered.

Appendix 3 Alignment profiling function

Mismatch weights are not equal along the flanking sequence, and should therefore be assigned according to a profiling function. Because of the nature of the sequencing process, it is common to have errors concentrated along the flanking sequence tails; we must, therefore, be mindful of this consideration and not disregard alignments in the tails of the query sequence just because of the high concentration of errors found there. Let us assume, therefore, that the distribution of errors follows the rule of natural distribution starting on some point within the flank. This can be approximated with the function F(x):

$$F(x) = \begin{cases} 0, \forall x \in [-\infty..0] \\ K, \forall x \in [0..X_{abs}] \\ \frac{1}{2} \bullet \left(\cos\left(\pi \bullet \frac{x - X_{abs}}{L - X_{abs}} \right) + 1 \right), \forall x \in [X_{abs}..L] \\ 0, \forall x \in [L..\infty] \end{cases}$$





Alignment Quality, "Q", can be calculated using the following equation:

$$Q_{absolute} = \int_{alignBegin}^{Xabs.1} F(x)dx + (X_{abs.2} - X_{abs.1}) + \int_{xAbs.2}^{alignEnd} F(x)dx =$$

$$\int_{aln.Begin}^{Xabs.1} \frac{1}{2}\left(1 - \cos(\pi \frac{x}{X_{abs.1}})\right)dx + (X_{abs.2} - X_{abs.1}) + \int_{xAbs.2}^{alignEnd} \frac{1}{2}\left(\cos(\pi \frac{x - X_{abs.2}}{L - X_{abs.2}}) + 1\right)dx =$$

$$\frac{\int_{aln.Begin}^{Xabs.1}\left(1 - \cos(\pi \frac{x}{X_{abs.1}})\right)dx + 2(X_{abs.2} - X_{abs.1}) + \int_{xAbs.2}^{alignEnd}\left(\cos(\pi \frac{x - X_{abs.2}}{L - X_{abs.2}}) + 1\right)dx}{2}$$

Having:

$$F_1(x) = \int_{alignBegin}^{Xabs.1} \left(1 - \cos(\pi \frac{x}{X_{abs.1}})\right) dx = x - \int_{alignBegin}^{Xabs.1} \cos(\frac{\pi}{X_{abs.1}} x) dx =$$

$$\left(x - \frac{X_{abs.1}}{\pi} \sin\left(\pi \frac{x}{X_{abs.1}}\right)\right)_{alignBegin}^{Xabs.1};$$

and

$$F_2(x) = \int_{Xabs.2}^{alignEnd} \left(\cos(\pi \frac{x - X_{abs2}}{L - X_{abs2}}) + 1\right) dx = \int_{Xabs2}^{alignEnd} \cos(x \frac{\pi}{L - X_{abs2}} - \frac{\pi X_{abs2}}{L - X_{abs2}}) dx + x =$$

$$= \left(\frac{L - X_{abs2}}{\pi} \bullet \sin(\pi \bullet \frac{x - X_{abs2}}{L - X_{abs2}}) + x\right)_{Xabs2}^{alignEnd};$$

The optimistic identity rate (so named since it doesn't include mismatches) can be calculated by the following function:

$$I_{abs} = \frac{2 \bullet (X_{abs.2} - X_{abs.1}) + \left(F_1(x)\right)_{a\ln Begin}^{X_{abs.1}} + \left(F_2(x)\right)_{X_{abs.2}}^{a\ln End}}{2 \bullet (X_{abs.2} - X_{abs.1}) + \left(F_1(x)\right)_{0}^{X_{abs.1}} + \left(F_2(x)\right)_{X_{abs.2}}^{L}};$$

Mismatches will affect the numerator of the above function. A function to describe mismatches will contain parts of unmovable discontinuations. Strictly speaking, we must take the integral of this function in order to determine the mismatch effect, but due to the corpuscular nature of the alignment, we can easily replace it with the sum of the elementary function:

$$M = \sum_{i=1}^{N} F(x_i); \forall i \in m,$$

where m is the mismatch position vector.

Thus, the final function:

$$I_{abs} = \frac{2 \bullet (X_{abs.2} - X_{abs.1}) + \left(F_1(x)\right)_{a\ln Begin}^{X_{abs.1}} + \left(F_2(x)\right)_{X_{abs.2}}^{a\ln End} - M}{2 \bullet (X_{abs.2} - X_{abs.1}) + \left(F_1(x)\right)_{0}^{X_{abs.1}} + \left(F_2(x)\right)_{X_{abs.2}}^{L}};$$

Appendix 4. 3D structure neighbor analysis.

When a protein is known to have a structure neighbor, dbSNP projects the RefSNPs located in that protein sequence onto sequence structures.

First, contig annotation results provide the SNP ID (snp_id), protein accession (protein_acc), contig and SNP amino acid residue (residue), as well as the amino acid position (aa_position) for a particular RefSNP. These data can be found in the dbSNP table, SNPContigLocusId. FASTA sequence is then obtained for each protein accession using the program idfetch, with the command line parameters set to:

```
-t 5 -dp -c 1 -q
```

We BLAST these sequences against the PDB database using "blastall" with the command line parameters set to:

```
-p blastp -d pdb -i protein.fasta -o result.blast -e 0.0001 -m 3 -I T -v 1 -b 1
```

Each SNP position in the protein sequence is used to determine its corresponding amino acid and amino acid position in the 3D structure from the BLAST result. These data are stored in the SNP3D table.

Chapter 6. The Gene Expression Omnibus (GEO): A Gene Expression and Hybridization Repository

Ron Edgar and Alex Lash

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The Gene Expression Omnibus (GEO) project was initiated at NCBI in 1999 in response to the growing demand for a public repository for data generated from high-throughput microarray experiments. GEO has a flexible and open design that allows the submission, storage, and retrieval of many types of data sets, such as those from high-throughput gene expression, genomic hybridization, and antibody array experiments. GEO was never intended to replace lab-specific gene expression databases or laboratory information management systems (LIMS), both of which usually cater to a particular type of data set and analytical method. Rather, GEO complements these resources by acting as a central, molecular abundance–data distribution hub. GEO is available on the World Wide Web at http://www.ncbi.nih.gov/geo.

Site Description

High-throughput hybridization array- and sequencing-based experiments have become increasingly common in molecular biology laboratories in recent years (1–4). These techniques are used to measure the molecular abundance of mRNA, genomic DNA, and proteins in absolute or relative terms. The main attraction of these techniques is their highly parallel nature; large numbers of simultaneous molecular sampling events are performed under very similar conditions. This means that time and resources are saved, and complex biological systems can be represented in a more holistic manner. Furthermore, the development of tissue arrays means that it is possible to analyze, in parallel, the gene expression of large numbers of tumor tissue samples from patients at different stages of cancer development (5).

Because of the plethora of measuring techniques for molecular abundance in use, our primary goal in creating the Gene Expression Omnibus (GEO) was to cover the broadest possible spectrum of these techniques and remain flexible and responsive to future trends, rather than choosing only one of these techniques or setting rigid requirements and standards for entry. In taking this approach, however, we recognize that there are obvious, inherent limitations to functionality and analysis that can be provided on such heterogeneous data sets.

This chapter is both more current and more detailed than the previous literature report on GEO (6). However, more detailed descriptions, tools, and news releases are available on the GEO Web site.

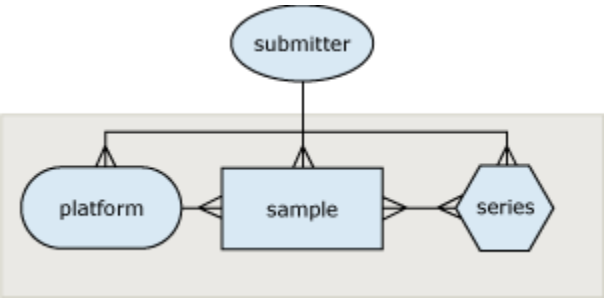


Figure 1. GEO design. The entity–relationship diagram for GEO.

Design and Implementation

The three principle components (or entities) of GEO are modeled after the three organizational units common to high-throughput gene expression and array-based methodologies. These entities are called *platforms*, *samples* and *series* (Figure 1; Table 1). A *platform* is, essentially, a list of probes that defines what set of molecules may be detected in any experiment using that platform. A *sample* describes the set of molecules that are being probed and references a single platform used to generate molecular abundance data. Each sample has one, and only one, parent platform that must be defined previously. A *series* organizes samples into the meaningful data sets that make up an experiment and are bound together by a common attribute.

The GEO repository is a relational database, which required that some fundamental implementation decisions were made:

(*a*) GEO does not store raw hybridization-array image data, although "reference" images of less than 100 Kb may be stored. This decision was based on an assertion that most users of the data within the GEO repository would not be equipped to use raw image data (7); although some may disagree, this means that repository storage requirements are reduced roughly by a factor of 20.

(*b*) We decided to use a different storage mechanism for data and metadata. Within the GEO repository, metadata are stored in designated fields within the database table. However, data from the entire set of probe attributes (for each platform) and molecular abundance measurements (for each sample) are stored as a single, text-compressed BLOB. This mode of data storage allows great flexibility in the amount and type of information stored in this BLOB. It allows any number of supplementary attributes or measurements to be provided by the submitter, including optional or submitter-defined information. For example, a microarray (the platform) consisting of several thousand spots (the probes) would have a set of probe attributes, some of which are defined by GEO. The GEO-defined attributes include, for each probe, the position within the array and biological reagent contents of each probe such as a GenBank Accession number, open reading frame (ORF) name, and clone identifier, as well as any number of submitter-defined columns. As another example, the set of probe-target measurements given in the data from a sample

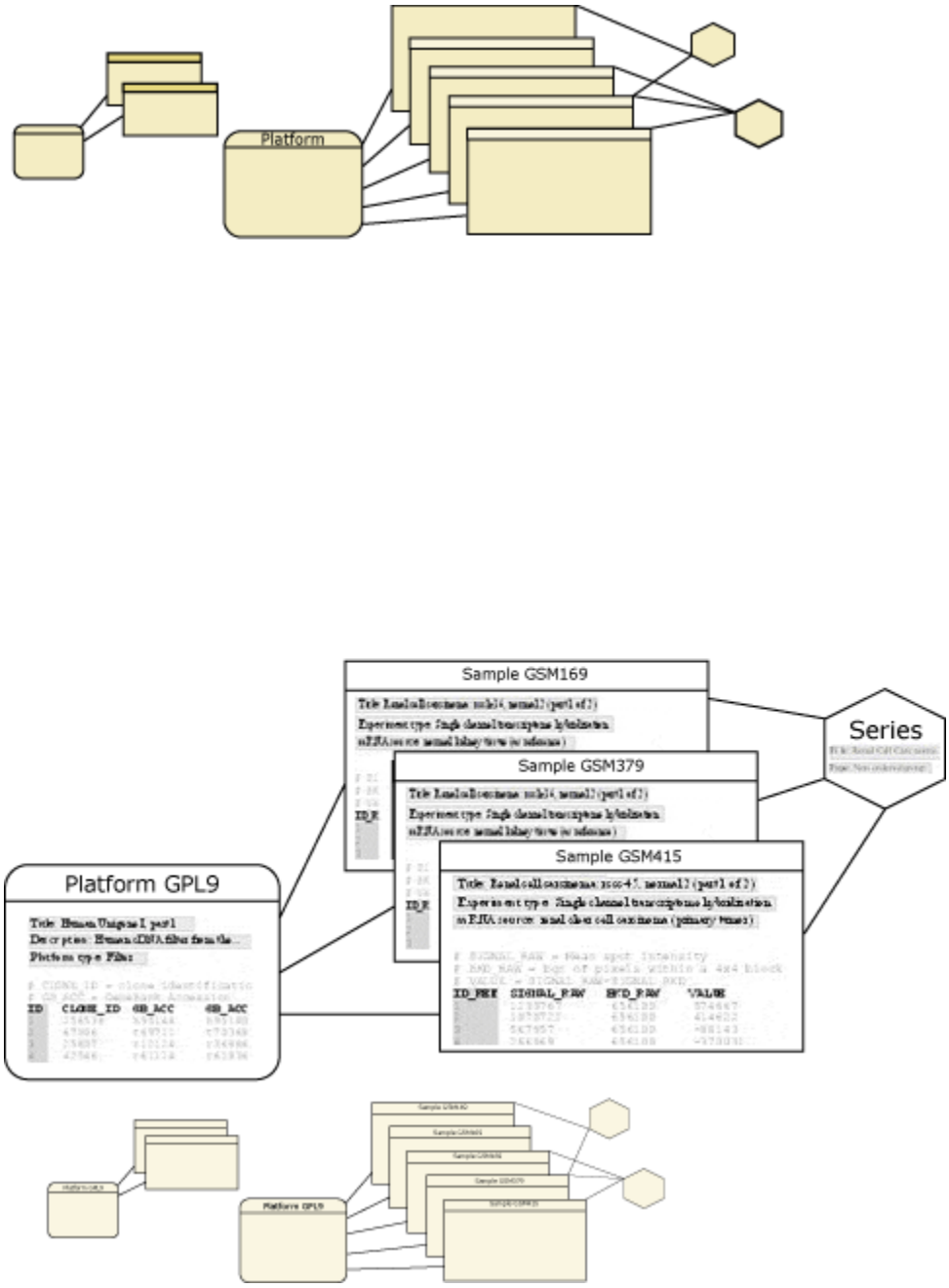


Figure 2. GEO implementation example. An actual example of three samples referencing one platform and contained in a single series.

may contain the final, relevant abundance value of the probe defined in its platform, as well as any other GEO-defined (e.g., raw signal, background signal) and submitter-defined data.

Once a platform, sample, or series is defined by a submitter, an Accession number (i.e., a unique, stable identifier) is assigned (Figure 2). Whether a GEO Accession number refers

to a platform, sample, or series can be understood by the Accession number "prefix". Platforms have the prefix GPL, samples have the prefix GSM, and series have the prefix GSE.

Table 1. Entity prefixes, types, and subtypes in the GEO database.

Accession prefix	Entity type	Subtype	Description
GPL	Platform	Commercial nucleotide array	Commercially available nucleotide hybridization array
		Commercial tissue array	Commercially available tissue array
		Commercial antibody array	Commercially available antibody array
		Non-commercial nucleotide array	Nucleotide array that is not commercially available
		Non-commercial tissue array	Tissue array that is not commercially available
		Non-commercial antibody array	Antibody array that is not commercially available
GSM	Sample	Dual channel	Dual mRNA target sample hybridization
		Single channel	Single mRNA target sample hybridization
		Dual channel genomic	Dual DNA target sample hybridization, e.g., array CGH
		SAGE	Serial analysis of gene expression
GSE	Series	Time–course	Time–course experiment, e.g., yeast cell cycle
		Dose–response	Dose–response experiment, e.g., response to drug dosage
		Other ordered	Ordered, but unspecified
		Other	Unordered

Retrieving Data

A GEO Accession number is required to retrieve data from the GEO repository database (Figure 3). An Accession number may be acquired in any number of ways, including direct reference, such as from a publication citing data deposited to GEO, or through a query interface, such as through NCBI's Entrez ProbeSet interface (covered below).

Given a valid GEO Accession number, the Accession Display tool available on the GEO Web site provides a number of options for the retrieval and display of repository contents (see Box 1).

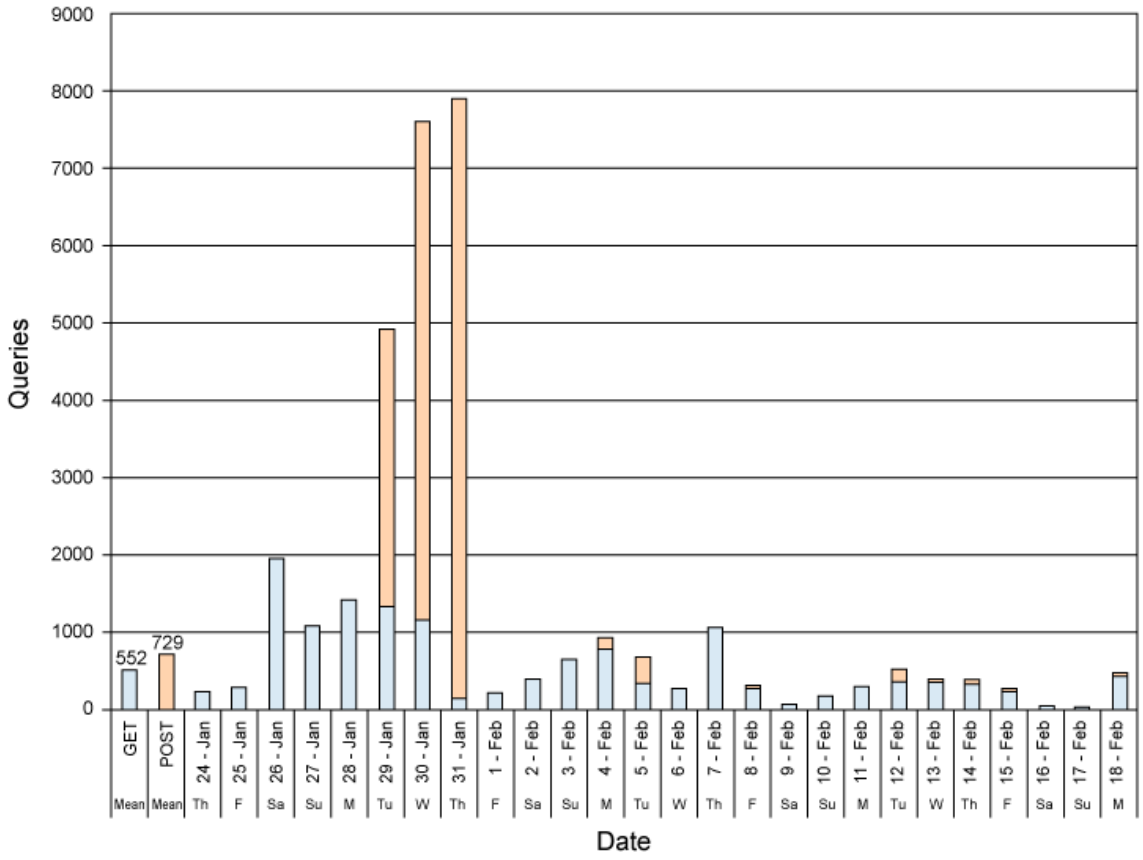


Figure 3. GEO retrieval statistics. Daily usage statistics evaluated over a 4-week period January 24 to February 20, 2002. Web server *GET* (*blue*) and *POST* (*magenta*) calls are evaluated for URL http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi. *GET* calls correspond roughly to links being followed from other Web pages, most likely following Entrez ProbeSet queries. *POST* calls roughly correspond to direct queries by Accession number. The spike of activity seen from January 29 to January 31 represents retrievals by one IP address and most likely represent an automated "Web crawler" pull.

Box 1. GEO Web site Accession Display tool.

It is very easy to use the **Accession Display** tool:

1. Type in a valid public or private[a] GEO Accession number in the top **GEO accession** box.
2. Select desired display options.

Box 1 continues on next page...

a To view one's own private, currently unreleased accessions, login with username and password at the bottom **login** bar.

Box 1 continued from previous page.

3. Press the **Go** button.

Three types of display options are currently available:

- **Scope** allows you to display the GEO accession(s) that you want to target for display. You may display the GEO accession, which is typed into the **GEO accession** box itself (**Self**), or any (**Platform**, **Samples**, or **Series**) or all (**Family**) of the accessions related to an accession. GEO platforms (GPL prefix) may have related samples and, through those related samples, related series. GEO samples (GSM prefix) will always have one related platform and may have multiple, related series. GEO series (GSE prefix) will have at least one related sample and, through those related samples, will have at least one related platform. The **Family** setting will retrieve all accessions (of different types) related to self (including self).
- **Format** allows you to display the GEO accession in human-readable, linked HTML form or in machine-readable, SOFT form (Box 2).
- **Amount** allows you to control the amount of data that you will see displayed. **Brief** displays the accession's metadata only. **Quick** displays the accession's metadata and the first 20 rows of its data set. **Full** displays the accession's metadata and the full data set. **Data** omits the accession's metadata, showing only the links to other accessions as well as the full data set.

Box 2. SOFT.

Simple Omnibus Format in Text (SOFT) is a line-based, ASCII text format that allows for the representation of multiple GEO platforms, samples, and series in one file. In SOFT, metadata appear as label-value pairs and are associated with the tab-delimited text tables of platforms and samples. SOFT has been designed for easy manipulation by readily available line-scanning software and may be quite readily produced from, and imported into, spreadsheet, database, and analysis software. More information about SOFT and the submission process is available from the GEO Web site.

Depositing Data

There are several formats in which data can be deposited and retrieved from GEO. For deposit: (1) a file containing an ASCII-encoded text table of data can be uploaded, and metadata fields can be interactively entered through a series of Web forms; or (2) both data and metadata for one or more platforms, samples, or series can be uploaded directly in a format we call Simple Omnibus Format in Text, or SOFT (Box 2).

Interactive and direct modes of communication are available for new data submissions and updating data submissions. The interactive Web form route is straightforward and most suited for occasional submissions of a relatively small number of samples. Bulk

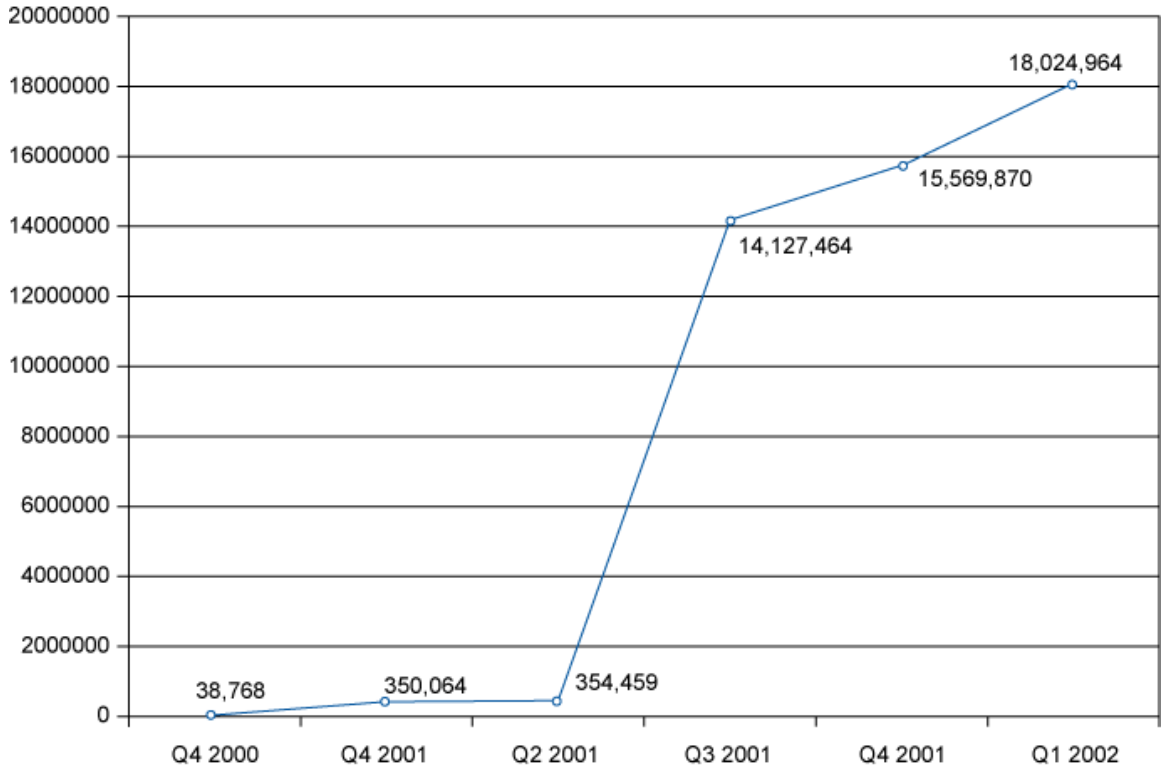


Figure 4. GEO submission statistics. Cumulative individual sample measurements submitted to GEO are shown. Data are presented by quarter since operations began on July 25, 2000.

submissions of large data sets may be rapidly incorporated into GEO via direct deposit of SOFT formatted data.

Submissions may be held private for a maximum of 6 months; this policy allows data release concordant with manuscript publication. Such submissions are given a final Accession number at the time of submission, which may be quoted in a publication.

Currently, submissions are validated according to a limited set of criteria (see the GEO Web site for more details). Submissions are scanned by our staff to assure that the submissions are organized correctly and include meaningful information. It is entirely up to the submitter to make the data useful to others.

A quarterly, cumulative graph of the number of individual molecular abundance measurements in public submissions made through the first quarter of 2002 is shown in Figure 4.

Search and Integration

Extensive indexing and linking on the data in GEO are performed periodically and can be queried through Entrez ProbeSet (Box 3). Many users of Entrez will recognize this

interface as similar to that of other popular NCBI resources such as PubMed and GenBank. As with any Entrez database, a Boolean phrase may be entered and restricted to any number of supported attribute fields (Table 2). Matches are linked to the full GEO entry as well as to other Entrez databases, currently Nucleotide, Taxonomy, and PubMed, as well as related Entrez ProbeSet entries. (See Chapter 15 for more details.) Entrez ProbeSet is accessible through the Entrez Web site as one of the drop-down menu selections.

Table 2. Entrez ProbeSet fields.

Field name	Description
Accession	GEO accession identifier
Author	Author of GEO sample
CloneID	Clone identifier of GEO sample's platform
Country	Country of GEO sample's submitter
Email	email of submitter
GBAcc	GenBank Accession of GEO sample's platform
Institute	Institute of GEO sample's submitter
Keyword	Keyword of GEO sample
ORF	Open reading frame (ORF) designation of GEO sample's platform
Organism	Organism of GEO sample and its parent taxonomic nodes
RefSeq	RefSeq accession of GEO sample's platform
SAGEtag	Serial analysis of gene expression (SAGE) 10-bp tag of GEO sample
Subtype	Subtype of GEO sample
Target ref	Target reference of GEO sample
Target src	Target source of GEO sample
Text Word	Word from description of GEO sample or sample's platform, and word from the titles of sample and its platform
Title	Titles of GEO sample and its platform

Box 3. Entrez ProbeSet indexing and linking process.

The basic unit (defined by a unique identifier, or UID, in Entrez parlance) in Entrez ProbeSet is the GEO sample, fused with its affiliated platform and series information. The indexing process iterates through all platforms in the GEO database, extracting metadata and the data table and fishing for any sequence-based identifiers such as GenBank Accession, ORFs, Clone IDs, or SAGE tags. Each sample belonging to that platform is in turn assigned a new UID and indexed with the above platform information plus any related series metadata (Table 2).

Box 3 continues on next page...

Box 3 continued from previous page.

GenBank Accessions, PubMed references, and taxonomy information are also linked to the appropriate Entrez databases for cross-reference and appear in the **Links** section of the display. Neighbors (related intra-Entrez database links) are generated for UIDs sharing the same GEO platform or series.

Example of Retrieving Data

Because samples are oftentimes organized into meaningful data sets within series, an example of retrieving a series and all the data of its associated samples and platform(s) is illustrative of the retrieval capabilities of the GEO Web site. For this example, to select a series of interest, we scan down a list of series in the GEO repository. However, to arrive at our series of interest, we could have just as well performed an Entrez ProbeSet query and followed GEO accession links to a sample and then to its related series, or followed links from PubMed to Entrez ProbeSet, and then to GEO. A step-by-step example of selecting a series of data and retrieving the data for this series from the GEO repository follows:

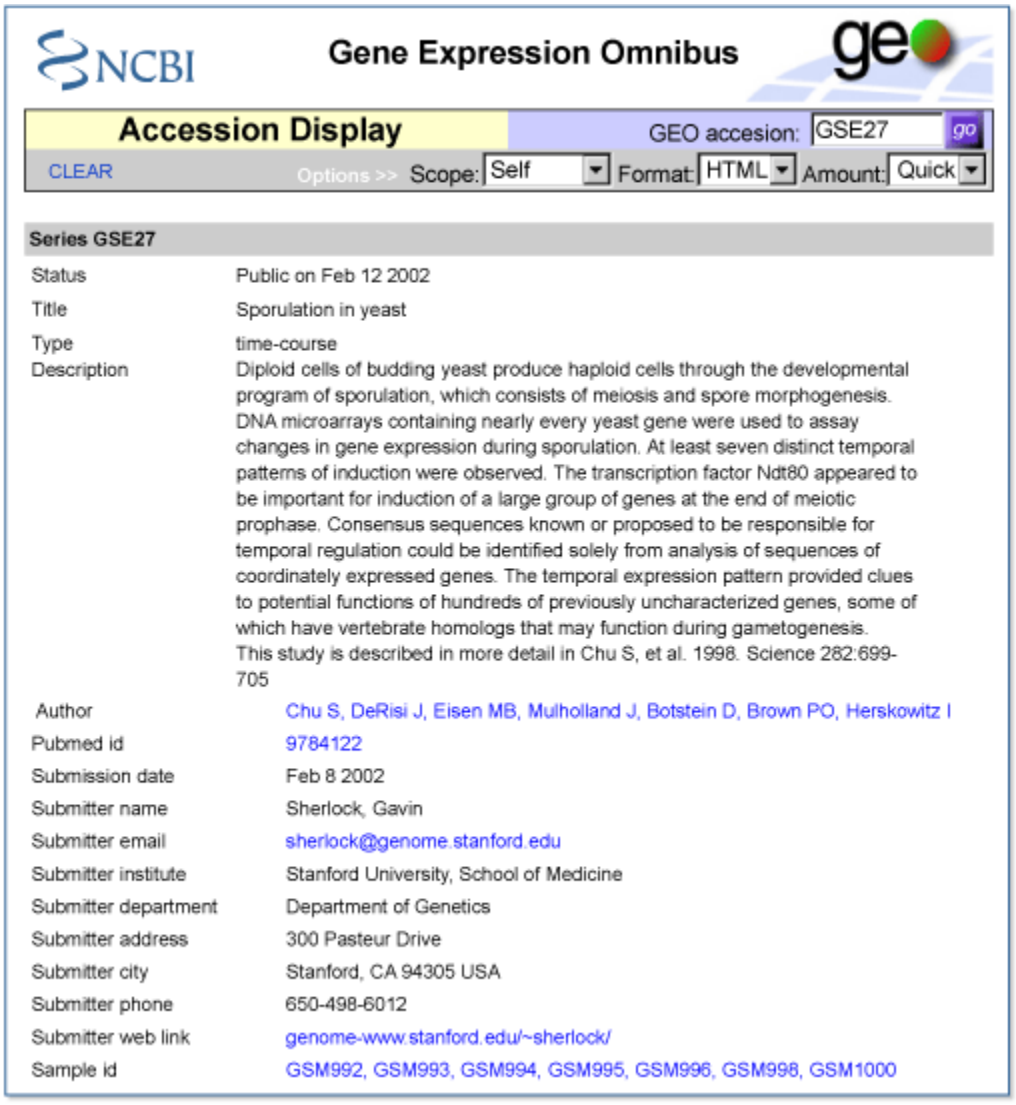
1. Select the linked number of public series from the table of Repository Contents given on the GEO homepage:

Repository contents	
Platforms	105 (120 Mb)
Samples	2361 (1747 Mb)
Series	79
Tue Sep 24 14:45:37 2002 EDT	

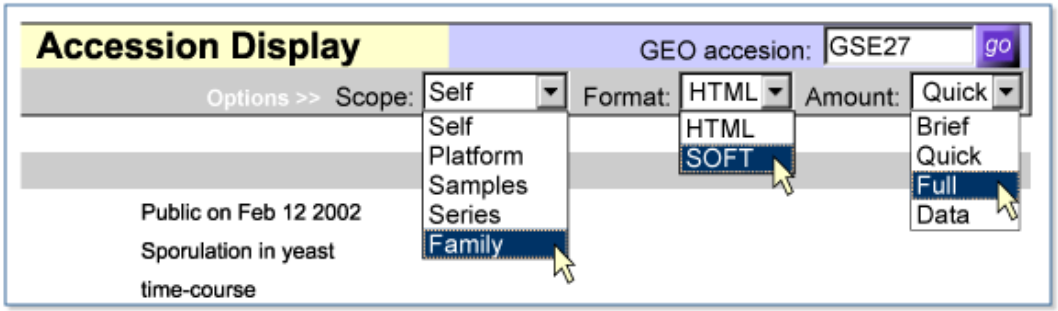
2. Scan down the list of public series in the GEO repository and select GSE27, on sporulation in yeast:

GSE26	Non-ordered group	6	Gavin Sherlock	Copper regulon in S. cerevisiae
GSE27	Time course series	7	Gavin Sherlock	Sporulation in yeast
GSE28	Time course series	7	Gavin Sherlock	Diauxic shift
GSE29	Non-ordered group	4	Gavin Sherlock	Adaptive evolution in yeast

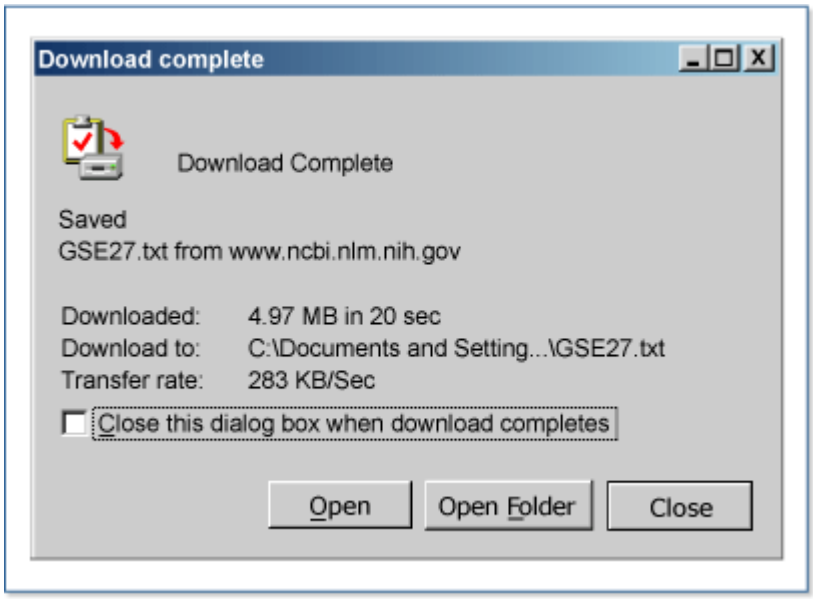
3. The description of GSE27 on the Accession Display allows a summary assessment of the data. The data set can be downloaded in SOFT format:



4. In the **Accession Display** options, select **Scope:**Family, **Format:**SOFT, and **Amount:**Full and then press the **go** button:



5. A browser dialog states that it took 19 seconds to download the 5 MB SOFT file of data and metadata for one series (GSE27), seven samples (GSM992 to GSM1000), and one platform (GPL67).



Future Directions

The GEO resource is under constant development and aims to improve its indexing, linking, searching, and display capabilities to allow vigorous data mining. Because the data sets stored within GEO are from heterogeneous techniques and sources, they are not necessarily comparable. For this reason, we have defined a ProbeSet to be a collection of GEO samples that contains comparable data. The selection of GEO samples into ProbeSets is necessary before integrating data in the GEO repository into other NCBI resources (see Chapter 15, Chapter 16, and Chapter 20), as well as for developing useful display tools for these data (Figure 5).

Frequently Asked Questions

1. How do I submit my data?

To submit data, an identity within the GEO resource must first be established. On first login, authentication and contact information must be provided. Authentication information (username and password) is used to identify users making submissions and updates to submissions. Contact information is displayed when repository contents are retrieved by others. This information is entered only once and can be updated at any time.

2. Is there a "hold until date" feature in GEO?

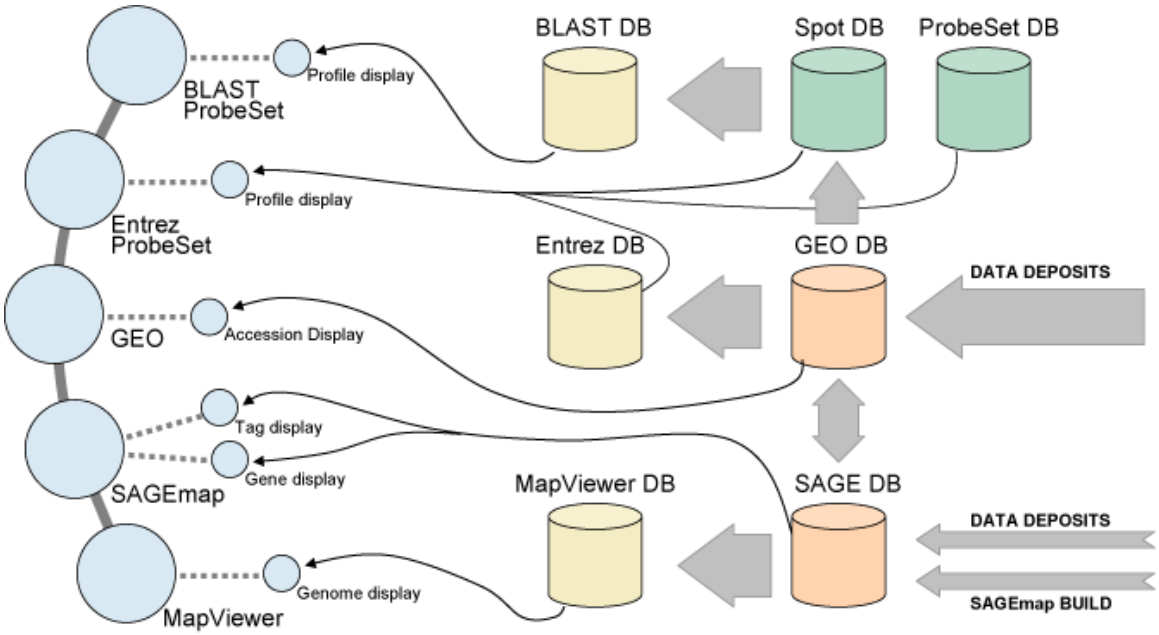


Figure 5. Constellation of NCBI gene expression resources. Anticipated development of gene expression resources at NCBI is shown. *Blue spheres* represent Web sites, *orange cylinders* represent primary NCBI databases, *green cylinders* represent secondary databases, and *yellow cylinders* represent tertiary NCBI interface databases. *Arrows* represent data flow, and *lines* represent Web site links.

Yes. This feature allows a submitter to submit data to GEO and receive a GEO Accession number before the data become public. There is currently a 6-month limit to this hold period. All private data are publicly released eventually.

3. What kinds of data will GEO accept?

GEO was designed around the common features of most of the high-throughput gene expression and array-based measuring technologies in use today. These technologies include hybridization filter, spotted microarray, high-density oligonucleotide array, serial analysis of gene expression, and Comparative Genomic Hybridization (CGH) and protein (antibody) arrays but may be expanded in the future.

4. Does GEO archive raw data images?

No. However, a reference image will be optionally accepted (limited to 100 Kb in size in JPEG format). In combination with optional references to horizontal and vertical coordinates, this image can be used to provide the user of the data with a qualitative assessment of the data.

5. Are there any Quality Assurance (QA) measurements that are required by GEO?

Not at this time. These requirements may be added in the future.

6. How can I submit QA measurements to GEO?

QA measurements are currently optional. If QA measurements are performed at the image-analysis step, these can be submitted as additional sample data.

7. How can I make corrections to data that I have already submitted?

By logging in with a username and password, an option to update a previous submission or your contact information is given. Accession updates can also be made through a link from the Accession Display after logging in. Updating the data of an already existing and valid GEO Accession number will cause a new version of that data element to be created. Alterations of metadata will not create a new version. All of the various versions of a data element will remain in the database.

8. How are submitters authenticated?

In their first submission to GEO, submitters will be asked to select a username and password. This username and password can be used to submit additional data in the future without reentering contact information, as well as to authenticate the submitter when updating or resubmitting data elements under an existing GEO Accession number.

9. How do I get data from GEO?

You need not login to retrieve data. All the data are available for downloading. NCBI places no restrictions on the use of data whatsoever but does not guarantee that no restrictions exist from others. You should carefully read NCBI's data disclaimer, available on the GEO Web site.

10. What kind of queries and retrievals will be possible in GEO?

Currently, there are three ways to retrieve submissions. One way is by entering a valid GEO Accession number into the query box on the header bar of this page; this will take you to the Accession Display. Another is to use the platform, sample, and series lists, located on the GEO Statistics page. Sophisticated queries of GEO data and linking to other Entrez databases can be accomplished by using Entrez ProbeSet.

11. What does Scope mean in the Accession Display?

GEO platforms (GPL prefix) may have related samples and, through those related samples, related series. GEO samples (GSM prefix) will always have one related platform and may have multiple, related series. GEO series (GSE prefix) will have at least one related sample and, through those related samples, will have at least one related platform. The **Family** setting will retrieve all accessions (of different types) related to self (including Self). Please see Box 1 for more details.

12. What is SOFT?

SOFT stands for Simple Omnibus Format in Text. SOFT is an ASCII text format that was designed to be a machine-readable representation of data retrieved from, or submitted to, GEO. SOFT output is obtained by using the Accession Display, and SOFT can be used to submit data to GEO. Please see Box 2 for more details.

13. What does the word "taxon" mean?

The NCBI's Taxonomy group has constructed and maintains a taxonomic hierarchy based upon the most recent information, which is described in Chapter 4 of this Handbook.

Acknowledgments

We gratefully acknowledge the work of Vladimir Soussov, as well as the entire NCBI Entrez team, especially Grisha Starchenko, Vladimir Sirotinin, Alexey Iskhakov, Anton Golikov, and Pramod Paranthaman. We thank Jim Ostell for guidance, Lou Staudt for discussions during our initial planning for GEO, and the extreme patience shown by Brian Oliver, Wolfgang Huber, and Gavin Sherlock when making the first data submissions. Admirable patience was also exhibited by Al Zhong during the development of the direct deposit validator. Special thanks go to Manish Inala and Wataru Fujibuchi for their continuing work on future features and tools.

Contributors

Table 3 shows a collection of data sets from various sources. Ron Edgar, Michael Domrachev, Tugba Suzek, Tanya Barrett, and Alex E. Lash contributed to this NCBI resource.

Table 3. Selective data set survey.

Source	Accessions	Description
NHGRI melanoma study	GSE1	This series represents a group of cutaneous malignant melanomas and unrelated controls that were clustered based on correlation coefficients calculated through a comparison of gene expression profiles.
Stanford Microarray Database	GSE4 to GSE9, and GSE18 to GSE29	These series represent microarray studies from the public collection of the Stanford Microarray Database (SMD).
Cancer Genome Anatomy Project	GSE14	This series represents the Cancer Genome Anatomy Project SAGE library collection. Libraries contained herein were either produced through CGAP funding or donated to CGAP.
Affymetrix Gene Chips™	GPL71 to GPL101	These platforms represent the latest probe attributes of the commercially

Table 3 continues on next page...

Table 3 continued from previous page.

Source	Accessions	Description
		available Affymetrix Gene Chips™ high density oligonucleotide arrays.
National Children's Medical Center Microarray Center	GSM1131 to GSM1345	These samples represent direct deposits of data derived from Affymetrix Gene Chip™ arrays and come from the Microarray Center database at the National Children's Medical Center.

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995;270:467–470. PubMed PMID: 7569999.
2. Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques. 1995;19:442–447. PubMed PMID: 7495558.
3. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995;270:484–487. PubMed PMID: 7570003.
4. Emili AQ, Cagney G. Large-scale functional analysis using peptide or protein arrays. Nat Biotechnol. 2000;18:393–397. PubMed PMID: 10748518.
5. Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med. 1998;4:844–847. PubMed PMID: 9662379.
6. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 2002;30:207–210. PubMed PMID: 11752295.
7. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001;29:365–371. PubMed PMID: 11726920.

Chapter 7. Online Mendelian Inheritance in Man (OMIM): A Directory of Human Genes and Genetic Disorders

Donna Maglott, Joanna S. Amberger, and Ada Hamosh

Created: October 9, 2002.

Summary

Online Mendelian Inheritance in Man (OMIM™*) is a timely, authoritative compendium of bibliographic material and observations on inherited disorders and human genes. It is the continuously updated electronic version of *Mendelian Inheritance in Man* (MIM). MIM was last published in 1998 (1) and is authored and edited by Dr. Victor A. McKusick and a team of science writers, editors, scientists, and physicians at The Johns Hopkins University and around the world (2). Curation of the database and editorial decisions take place at The Johns Hopkins University School of Medicine. OMIM provides authoritative free text overviews of genetic disorders and gene loci that can be used by clinicians, researchers, students, and educators. In addition, OMIM has many rich connections to relevant primary data resources such as bibliographic, sequence, and map information.

Content and Access

OMIM Entries

OMIM comprises descriptive, full-text MIM entries, a tabular Synopsis of the (Human Gene Map) that includes the Morbid Map, clinical synopses, and mini-MIMs.

OMIM entries are authored and edited by experts in the field and by the OMIM staff, based on information in the published literature. All entries are assigned a unique, stable, six-digit ID number and provide names and symbols used for the disorder and/or gene, a literature-based description, citations, contributor information, and creation and editing dates. Because MIM is derived from the primary literature, the text is replete with citations and links to PubMed. OMIM authors create entries for each unique gene or Mendelian disorder for which sufficient information exists and do not wittingly create more than one entry for each gene locus.

MIM is organized into autosomal, X-linked, Y-linked, and mitochondrial catalogs, and MIM numbers are assigned sequentially within each catalog (Table 1). The kinds of

* Trademark status. OMIM™ and Online Mendelian Inheritance in Man™ are trademarks of The Johns Hopkins University.

information that may be included in MIM entries are approved name and symbol (obtained from the HUGO Nomenclature Committee), alternative names and symbols in common use, and a text description of the disease or gene. Many of the longest entries and most new entries in MIM have headings within the text that may include Clinical Features, Inheritance, Population Genetics, Heterogeneity, Genotype/Phenotype Correlations, Cloning, Gene Structure, Gene Function, Mapping, and more. Information on selected disease-causing mutations is contained in the Allelic Variant section of the entry describing the gene.

With the increasing complexity of biological information, OMIM makes a critical contribution by distilling what is known about a gene or disease into a single, searchable entry. The rich text of the OMIM entry, along with the source reference citations, make it easy to retrieve data of interest. The OMIM entry can then serve as a gateway to other sources of related information via the many curated and computed links within each entry.

Table 1. The OMIM numbering system.

MIM number range[a]	Explanation
100000-199999	Autosomal loci or phenotypes (entries created before May 15, 1994)
200000-299999	Autosomal loci or phenotypes (entries created before May 15, 1994)
300000-399999	X-linked loci or phenotypes
400000-499999	Y-linked loci or phenotypes
500000-599999	Mitochondrial loci or phenotypes
600000-	Autosomal loci or phenotypes (entries created after May 15, 1994)

a MIM numbers are frequently preceded by a symbol. An asterisk (*) before a MIM number indicates that the entry describes a distinct gene or phenotype and that the mode of inheritance of the phenotype has been proved (in the judgment of the authors and editors) and that the phenotype described is not known to be determined by a gene represented by other asterisked entries in MIM.

A number sign (#) before a MIM number describing a phenotype indicates that the phenotype is caused by mutation in a gene represented by another entry and usually in any of two or more genes represented by other entries. The number sign is also used for phenotypes that result from specific chromosomal aberrations, such as Down syndrome, and for contiguous gene syndromes, such as Langer-Giedion syndrome. Whenever a number sign is used, the reason is stated at the outset of the entry.

The absence of an asterisk (or other sign) preceding the number indicates that the distinctness of the phenotype as a mendelizing entity or the characterization of the gene in the human is not established.

The OMIM Gene Map

The OMIM Synopsis of the Human Gene Map is a tabular listing of the genes and loci represented in MIM ordered pter to qter from chromosome 1–22, X, and Y. The information in the map includes the cytogenetic location, symbol, title, MIM number,

method of mapping, comments, associated disorders and their MIM numbers, and the map location of the mouse ortholog. Links are provided from the cytogenetic location to the human Map Viewer, from MIM numbers to OMIM entries, and from mouse map locations to the Mouse Genome Database (MGD).

The Synopsis of the human gene map has also been sorted alphabetically by disorder and is referred to as the Morbid Map.

Access

OMIM can be found either by direct query via Entrez, from other data resources within NCBI that connect to OMIM directly (for example, LocusLink or UniGene), or through Entrez cross-indexing (for example, from a PubMed abstract of an article cited in an OMIM entry). OMIM is indexed for retrieval in Entrez using a weighted system so that if the query term(s) appears in the title of a MIM entry, it will appear at the top of the retrieval list. Field restrictions are supported for some types of information, or one can use the **Limits** page to restrict a search (Figure 1). There are also several format options for viewing a retrieval set that may affect which entries are displayed (Box 1).

Queries can also be entered in the query box shown on all Entrez database pages, selecting OMIM as the database in the **Search** option. The **Preview/Index, History**, **Clipboard**, or Cubby functions can also be used for OMIM, as for the rest of Entrez.

The Gene Map can be accessed from an individual OMIM entry via the cytogenetic location displayed when appropriate under the entry titles. The Gene Map may also be queried directly. When queried directly, the first entry that matches the query is shown in the top row of the table, followed by 19 entries ordered by cytogenetic location. The **Find Next** button can be used to find additional gene map entries that match the query.

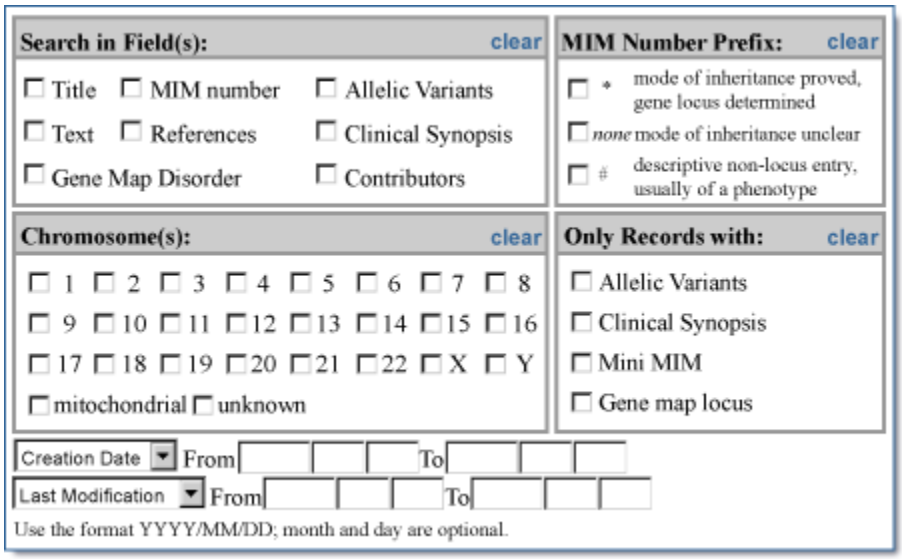


Figure 1. **The Limitsoptions for searching OMIM in Entrez.**

Box 1. Display options for viewing OMIM query results.

Title (default)

Details

Clinical Synopsis

Allelic Variants

mini-MIM

ASN.1

LinkOut

Related Entries

Genome Links

Nucleotide Links

Protein Links

PubMed Links

SNP Links

Structure Links

UniSTS Links

Box 1 continues on next page...

Box 1 continued from previous page.

Obtaining multiple views of a query result:

1. Enter query term or terms (example: renal failure hypertension).

2. Default display is **Titles**.

3. Select **Clinical Synopsis** and click on **Display** at the left to see the **Clinical Synopsis** section of all entries that have them.

4. Similarly, select **mini-MIM** or **Allelic Variants**.

NOTE: In the same bar, the number of entries to display and the format in which to display them can be configured by use of the **Show** and **Text** buttons, respectively.

Guide to OMIM Pages

Query Bar

OMIM is queried via a standard Entrez query bar. The mechanics of selecting entries to display, how to display them, and identifying related entries either within Entrez or from external resources is also according to the Entrez/LinkOut standard. The display options (Box 1) allow the user to format results in several ways. A useful function particular to OMIM is the option to display **Allelic Variants**, **Clinical Synopsis**, or **mini-MIM** views for a retrieval set.

OMIM Navigation

The OMIM homepage and the search results pages share the same navigational links to the advanced query page (Gene Map, Morbid Map), Help documents, FAQs, statistics, and related resources.

When viewing the text of an OMIM entry, however, the navigational links serve as an electronic table of contents. The section headings within the entry are listed similar to a table of contents, and selecting one moves the display to that section. Within an entry, selecting the **MIM #** link takes you back to the top of the entry.

OMIM staff actively contributes to the curation of data in LocusLink. Thus, if the MIM number is represented in LocusLink, a reciprocal LocusLink link is provided in the section to the left. Other links provided by the LocusLink collaboration may also be listed in this section, e.g., links to Nomenclature, Reference Sequences, or UniGene clusters that are specific to the subject OMIM entry.

The sequence links in the LocusLink section may be different from the Entrez indexing links available via the **Links** link at the top right of an entry. The Entrez indexing links result indirectly from the references in the OMIM entry and may include related sequences in other species, for example. Thus, OMIM pages allow two levels of sequence

connection: the specific ones in the left section under LocusLink and the indirect but still informative ones through Entrez indexing link at the upper right. More information on OMIM link types can be found in the Help documents.

OMIM entries may also contain a link to LinkOut for resources external to NCBI (see Chapter 17). Some of these external resouces are curated by OMIM staff, in which case they are displayed by name. Others can be seen either by selecting **LinkOut** in the **Links** pull-down menu or by selecting the **LinkOut** display option in the query bar.

Entrez Links

At the top of any report page, or associated with each entry in the query result page, are the links to related data generated from Entrez (Chapter 15). Here, PubMed links to the PubMed abstracts of the reference citations in the entry. **Related Entries** are to all other OMIM entries referenced in the subject entry. **Nucleotide**, **Protein**, and **LinkOut** connections are as documented in the previous section.

The OMIM Entry

Each OMIM entry has a unique number and is given a primary title and symbol. This is the title that is displayed in the document retrieval list. Alternative designations are listed below the primary title. Some entries contain information that is related but not synonymous to the primary topic and is not addressed in another entry (e.g., splice variants, phenotypic variants, etc.). This information is set off by the word "included" in the title. The first "included" title is displayed in the document retrieval list. The cytogenetic map location when known is given for each entry. When a disease shows genetic heterogeneity, more than one map location may be given. The "light bulb" icon at the end of text paragraphs links to related articles in PubMed. References within the text are linked to the complete citation at the end of the entry. There, the PubMed ID is linked to the PubMed abstract.

Some entries contain an **Allelic Variants** section, which lists noteworthy mutations for the gene. Allelic variants are given a 10 digit number: the 6-digit number of the parent locus, followed by a decimal point and a unique 4-digit variant number. Criteria for inclusion include the first mutation to be discovered, high population frequency, distinctive phenotype, historic significance, unusual mechanism of mutation, unusual pathogenetic mechanism, and distinctive inheritance (e.g., dominant with some mutations, recessive with other mutations in the same gene). Most of the allelic variants represent disease-producing mutations. A few polymorphisms are included, many of which show a statistically significant association with specific common disorders.

FTP

The OMIM data are available for bulk transfer, but it should be noted that there are **restrictions on use**.

Legal Statement

OMIM is funded by a contract from the National Library of Medicine and the National Human Genome Research Institute and by licensing fees paid to the Johns Hopkins University by commercial entities for adaptations of the database. The terms of these licenses are being managed by the Johns Hopkins University in accordance with its conflict of interest policies.

References

1. McKusick VA, et al. Mendelian Inheritance in Man. 12th ed. Baltimore: Johns Hopkins University Press; 1998.
2. Hamosh A,et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2002;30:52–55. PubMed PMID: 11752252.

Chapter 8. The NCBI BookShelf: Searchable Biomedical Books

Bart Trawick, Jeff Beck, and Jo McEntyre

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The BookShelf is a collection of biomedical books that can be searched directly in Entrez or found via keyword links in PubMed abstracts. Books have been added to the BookShelf in collaboration with authors and publishers, and the complete content (including all figures and tables) is free to use for anyone with an Internet connection.

The online books are displayed one section at a time, with navigation provided to other parts of the current chapter or to other chapters within the book. Many of the books on the BookShelf can be browsed without any restriction at all; others have less flexibility for navigating the complete content. The publisher (or the owner of the content) defines the rules for access.

The books are linked to PubMed through research papers citations within the text. In the future, more links may be established between the BookShelf and other resources at NCBI, such as gene and protein sequences, genomes, and macromolecular structures.

Content Acquisition

Basis for Inclusion

The BookShelf provides a venue through which publishers and authors can make the full text of biomedical books available to the scientific community. As the BookShelf grows, we welcome proposals from authors, editors, or publishers of any "in-scope" texts, from undergraduate textbooks to more specialized publications, including collections of review articles or workshop proceedings. The scope of the BookShelf is broadly biomedical, including clinical works and those concerning basic biological and chemical sciences.

Information for Authors, Editors, and Publishers

All books made available on the BookShelf have been provided in electronic format to NCBI. The publisher must provide the content *in full*. Because the complete content is required for display and indexing purposes, the authors, editors, and publisher of a book should all be a part of the decision to participate in the BookShelf. Interested parties can contact us at books@ncbi.nlm.nih.gov for more information or to make a proposal for the inclusion of a book. There is a simple contract that specifies the terms of use of the content. A sample contract can also be obtained by request at the above email address.

The complete contents of each book will be converted into XML according to the NCBI Book Document Type Definition (DTD), a public domain DTD developed at NCBI for

this project (see below). Books may be submitted to conform to this DTD, or NCBI will convert the source data to validate against the Book DTD. If a conversion needs to be done, the content must be in a format robust enough to meet the needs of the BookShelf publishing system and DTD.

Any book that was printed from SGML or XML should allow for a straightforward conversion. We have had success converting books from Word, XYWrite, PDF, and Quark Express formats, and we anticipate that we would also be able to convert from other desktop publishing packages. HTML and PDF formats are less desirable because the data formats are less detailed.

Figures should be supplied in TIFF format, although GIF and JPEG formats may be accepted. The submitted text files are converted into XML according to the NCBI Book DTD; graphic files are converted into GIF and JPEG formats. Three hard copies of the book are also required, along with the electronic files.

The XML files are stored in a database. When a reader requests a book, chapter, or section, the XML is retrieved from the database and converted into HTML on the fly using Extensible Stylesheet Language Transformations (XSLT) and Cascading Style Sheets (CSS).

How to Use the Books

There are three ways to access the content in BookShelf:

1. Through hyperlinked terms in PubMed abstracts

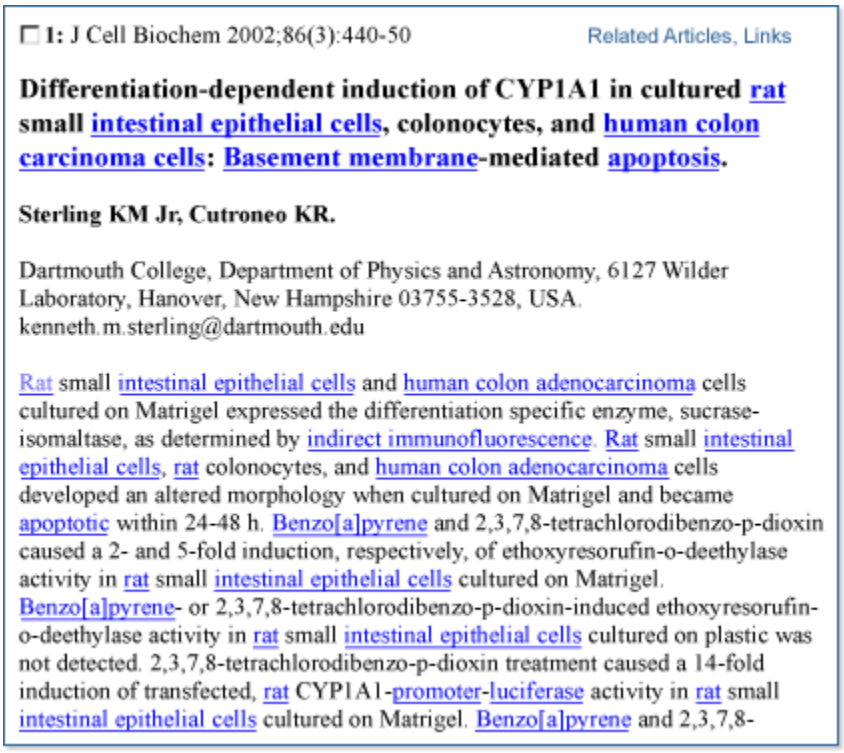
2. By a direct search using search terms or phrases (in the same way as the bibliographic database of PubMed is searched)

3. Through the Table of Contents of the book (note: some publishers restrict browsing through the entire book by disabling hyperlinks in the Table of Contents)

Links from PubMed

The BookShelf can be accessed from all PubMed abstract pages. When viewing a full PubMed abstract, select the **Books** hyperlink in the upper right-hand corner. This generates a version of the abstract in which certain phrases and terms appear as hypertext links (see Figure 1). The linked term may be one or more words in length. If a word or a phrase is linked, it means that the exact phrase also appears in at least one book. Selecting a linked phrase retrieves a list of books that contain that term.

A statistical weighting system based on the frequency of each phrase in a book section, relative to the rest of the book, is used to identify "good" phrases. A phrase that appears repeatedly in only a few sections and rarely in other parts of the book indicates a definitive phrase for those few sections; therefore, it ranks highly. Furthermore, the appearance of a phrase in the title, for example, has a greater value in the weighting system than one appearing solely in the text.


1: J Cell Biochem 2002;86(3):440-50 Related Articles, Links

Differentiation-dependent induction of CYP1A1 in cultured rat small intestinal epithelial cells, colonocytes, and human colon carcinoma cells: Basement membrane-mediated apoptosis.

Sterling KM Jr, Cutroneo KR.

Dartmouth College, Department of Physics and Astronomy, 6127 Wilder Laboratory, Hanover, New Hampshire 03755-3528, USA. kenneth.m.sterling@dartmouth.edu

Rat small intestinal epithelial cells and human colon adenocarcinoma cells cultured on Matrigel expressed the differentiation specific enzyme, sucrase-isomaltase, as determined by indirect immunofluorescence. Rat small intestinal epithelial cells, rat colonocytes, and human colon adenocarcinoma cells developed an altered morphology when cultured on Matrigel and became apoptotic within 24-48 h. Benzo[a]pyrene and 2,3,7,8-tetrachlorodibenzo-p-dioxin caused a 2- and 5-fold induction, respectively, of ethoxyresorufin-o-deethylase activity in rat small intestinal epithelial cells cultured on Matrigel. Benzo[a]pyrene- or 2,3,7,8-tetrachlorodibenzo-p-dioxin-induced ethoxyresorufin-o-deethylase activity in rat small intestinal epithelial cells cultured on plastic was not detected. 2,3,7,8-tetrachlorodibenzo-p-dioxin treatment caused a 14-fold induction of transfected, rat CYP1A1-promoter-luciferase activity in rat small intestinal epithelial cells cultured on Matrigel. Benzo[a]pyrene and 2,3,7,8-

Figure 1. A PubMed abstract showing terms linked to books. This view was generated by selecting **Links** found in the *top right corner* of the abstract and then selecting **Books** from the drop-down menu that appears.

Each PubMed abstract can thus be linked to the appropriate book pages. This method allows two very dissimilar types of text—the dense, focused PubMed abstracts and the more descriptive book text—to find common ground.

Direct Search of Books

Book contents may be searched directly from the BookShelf homepage by using the search boxes located in the Table of Contents and navigation bars of books or by selecting **Books** from the pull-down menu in any Entrez database search bar (see Chapter 15). Search terms may be combined using Boolean operators that conform to PubMed syntax (see Chapter 2). The BookShelf also allows search fields to be specified. A complete list of BookShelf database fields can be found in Table 1.

Table 1. Field limits for use in the BookShelf.

Field[a]	Use
[Author]	Search for the authors of books or chapters.

a Filters are applied immediately following a search term, with no separating spaces, e.g., watson[author] AND cmed[book].

Table 1 continues on next page...

Table 1 continued from previous page.

Field[a]	Use
[Book]	Typically used with a Boolean expression to limit a search to a particular book.
[PmId]	Locate a journal article citation in a book by its PubMed ID.
[Rid]	Locate a particular book element (such as a figure or table) by its reference ID.
[Secondary Text]	Search for secondary text, e.g., units (mg/l, etc.)
[Title]	Search for words used in any title (book, chapter, section, subsection, figure, etc.).
[Type]	Locate a division of a book such as a section, chapter, or figure group.

a Filters are applied immediately following a search term, with no separating spaces, e.g., watson[author] AND cmed[book].

Interpreting Results from a Search or PubMed Link

Results are shown as a list of books in which the term is found, along with the number of sections, figures, and tables that contain the term (Figure 2). The book that contains the most hits appears at the top of the list. Choosing the hyperlinked number of items that is associated with a particular book will then display a document summary list of the individual sections, figures, and tables found. (When a term or phrase is found fewer than 20 times within the BookShelf, the document summary page is shown directly, without first displaying the results clustered by book.)

The document summary list is sorted with the most relevant documents shown at the top of the page. The sorting makes use of scores allocated to phrases as a measure of how relevant they are to a given section (a part of the statistical weighting system also used for linking PubMed abstracts to the books). For each book section found, the title, along with some context regarding the hierarchy of the section location (e.g., the chapter and book), is given. An icon is used to distinguish figure legends and tables from text sections (Figure 3). Selecting a hyperlinked section title displays the part of the book that contains the search term. From this point, the user may be able to navigate further throughout the book content, according to the policy of the publisher (see *Navigating Book Content*).

Navigating Book Content

Each HTML page of content seen in a Web browser represents one section of one chapter of a book, i.e., all of the content (including subsections and so on) within the first-level heading of a chapter. The amount of content this represents varies according to the structure of the original book. Some books have very long sections, some short, some a mixture; although on the whole, most chapters are divided into 3-10 sections.

The top of every page contains links to both short and detailed Tables of Contents and a description of the current location within the book (Figure 4). The hierarchal elements

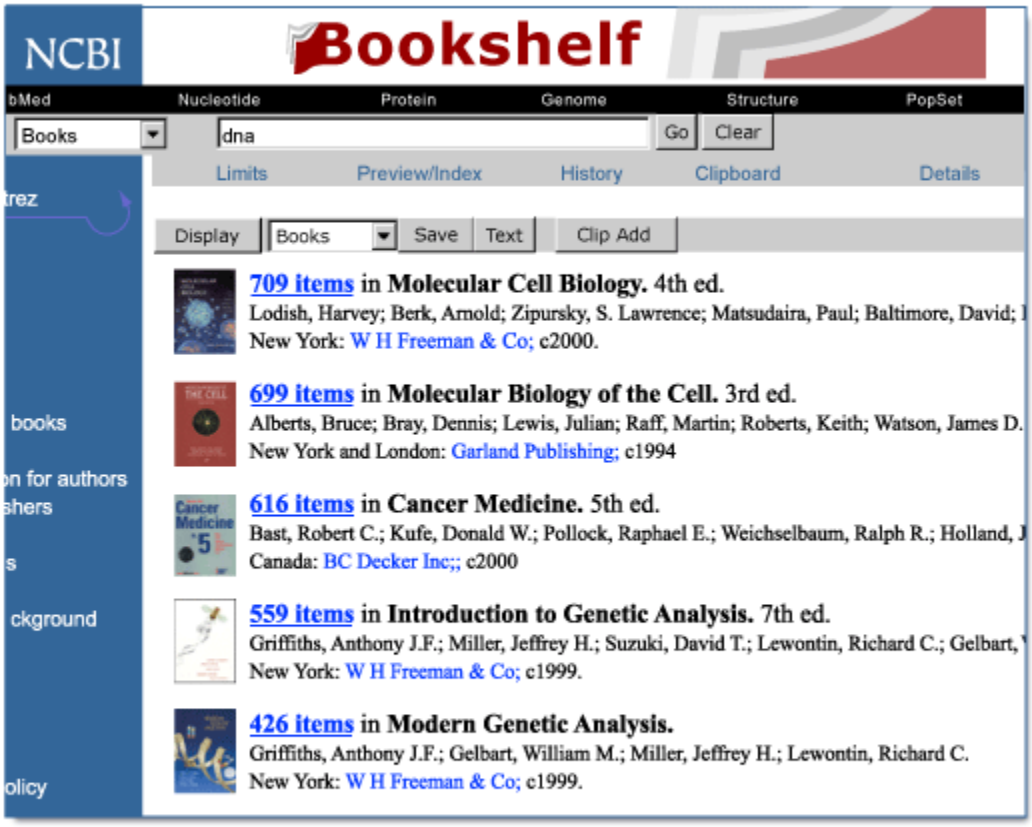


Figure 2. A results page from a BookShelf search. If more than 20 sections, tables, and/or figures are found that contain the query term, a summary page, such as the one above, is displayed.

that describe the current location are hyperlinked and may be used to travel up the organizational levels of a book. Additionally, a navigation sidebar shows the current section among its peers and lists the figures and tables found within the current section (Figure 4). Reference citations in the text are linked where possible to PubMed abstracts by the Citation Matcher. References internal to the book, e.g., to other chapters or sections, figures, tables, and boxes, are also hyperlinked. Further navigation from the current page to other parts of the book depends on the access policy of the publisher.

Figure 3. A document summary list of book sections. The most relevant sections to the query appear at the top of the list. Note the icons that designate figure and table hits. The list may also be displayed in a brief format that lists only the section names that contain the term by choosing **Brief** from the drop-down menu to the immediate *right* of the **Display** button.

Technology

Text Conversion and XML

All book content submitted to the BookShelf is converted to XML according to the public NCBI Book DTD.

Any files submitted in SGML or XML are converted to the Book DTD using XSLT. Books submitted in desktop publishing or word processing formats are converted by a contractor

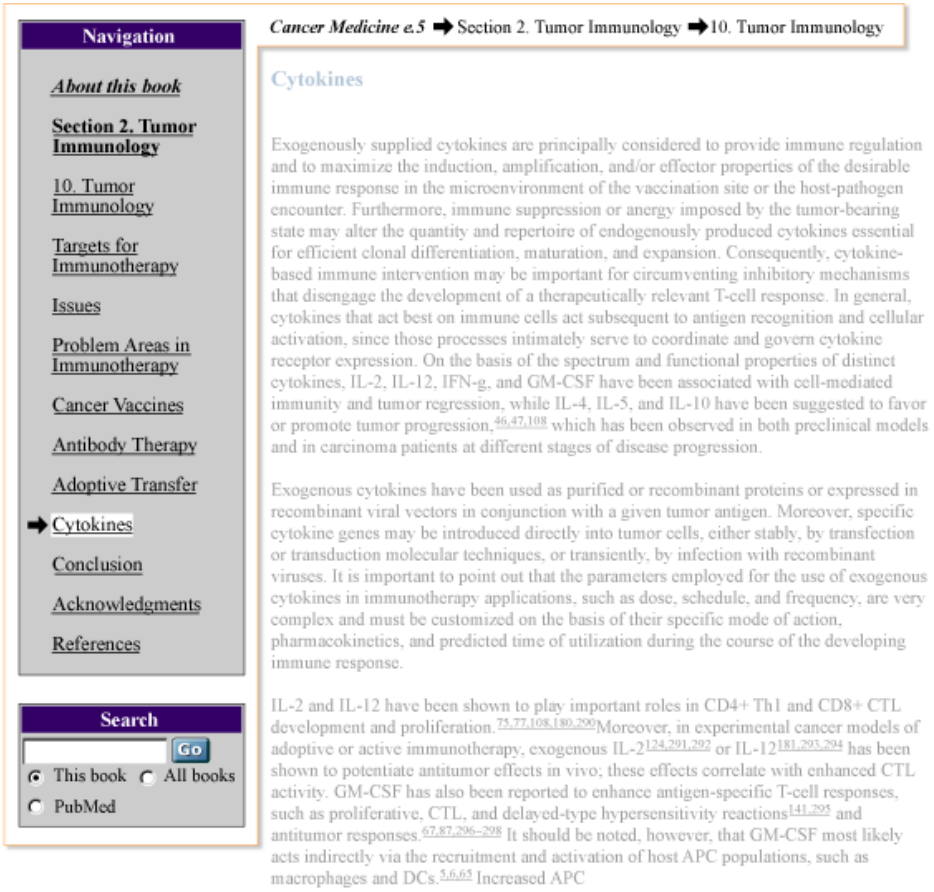


Figure 4. Navigation elements that appear on every book HTML page. Hyperlinks to various sections within a chapter appear within a navagation bar to the *left* of each page. Hyperlinks may be disabled within some books at the request of the publisher. A **Search** box is located *below* the navigation bar. At the *top* of each page is a hyperlinked, hierarchical tree that illustrates a page's relative position within a book.

using proprietary technologies. Once the book XML is valid against the DTD, the book is ready to be loaded into the database.

The XML for a book is generally a set of files. Each chapter and appendix is an independent file, as is the frontmatter. The book is pulled together by the book.xml file, which defines the structure of the book. For example, a book with two chapters and a bibliography at the end of the book would be structured as follows:

```
<!DOCTYPE book SYSTEM "ncbi-book.dtd" [
<!-- graphics -->
<!ENTITY % Graphics SYSTEM "graphics.xml">
%Graphics;
<!ENTITY frontmatter SYSTEM "fm.xml">
```

```
<!ENTITY chapter1 SYSTEM "ch1.xml">
<!ENTITY chapter2 SYSTEM "ch2.xml">
<!--back-->
<!ENTITY biblist SYSTEM "biblist.xml">
]>
<book>
&frontmatter;
<body>
&chapter1;
&chapter2;
</body>
<back>
&biblist;
</back>
</book>
```

The book.xml file is composed of two parts. The first part defines all of the components that are required to build the book. These definitions occur within the <!DOCTYPE [ ] > tag. The second part builds the structure of the book. The root element is <book>. <book> contains whatever is in &frontmatter;, <body>, and <back>.

The book.xml example above refers to five external files: fm.xml, which contains all of the frontmatter of the book; ch1.xml, which contains chapter 1; ch2.xml, which contains chapter 2; biblist.xml, which contains the bibliography for the book; and graphics.xml, which defines the images. If any of these files is not valid according to the DTD or if the files are not found where they are defined in their <!ENTITY > declaration, then the book will not be valid.

Images

All of the images, including figures, icons, and book-specific character graphics (see Special Characters below), are called out in the text as entities. The entities are defined in the graphics.xml file.

<!ENTITY ch2fu6 SYSTEM "data/mga/pictures/ch2/ch2fu6.gif" NDATA GIF>

<!ENTITY ch2fu7 SYSTEM "data/mga/pictures/ch2/ch2fu7.jpg" NDATA JPG>

<!ENTITY ch2fu8 SYSTEM "data/mga/pictures/ch2/ch2fu8.gif" NDATA GIF>

<!ENTITY ch2fu9 SYSTEM "data/mga/pictures/ch2/ch2fu9.gif" NDATA GIF>

<!ENTITY ch2fu10 SYSTEM "data/mga/pictures/ch2/ch2fu10.gif" NDATA GIF>

<!ENTITY ch2e1 SYSTEM "data/mga/pictures/ch2/ch2e1.gif" NDATA GIF>

<!ENTITY ch2e2 SYSTEM "data/mga/pictures/ch2/ch2e2.gif" NDATA GIF>

Graphic files are converted into GIF and JPEG formats and optimized for display on the Web. The images are not loaded into the database; they are retrieved from a file server when called by the HTML page.

Math and Formulae

Math expressions and chemical formulae and structures are handled as images.

Special Characters

The BookShelf uses the same character sets that PubMed Central uses (see Chapter 9). These include a number of standard ISO character sets (8879 and 9573), along with a set of characters that has been defined to accommodate characters not in the standard set. The ISO Standard Character sets referenced are listed in Box 1 of Chapter 9. Special characters are converted to the BookShelf/PubMed Central (PMC) character set during conversion into XML. Characters created for one book (book-specific characters) are called out in the XML as images. To provide for the most flexibility in displaying characters across platforms, BookShelf uses UTF-8 encoding whenever possible. Because not all browsers support the same subset of UTF-8 characters and some characters cannot be represented in UTF-8, the BookShelf displays characters as a combination of GIFs and UTF-8 characters, depending on the Browser/OS combination and the character to be displayed.

The BookShelf Data Flow

The XML files for each book are stored in a Sybase database. To show the best representation to each reader, the reader's browser and operating system are noted and passed to the rendering software. When a reader requests a book, chapter, or section, the XML and the character images or UTF-8 characters appropriate to the reader's system are retrieved from the database.

The XML is converted to HTML using XSLT stylesheets. The look of the HTML pages is controlled further using CSS, which allow manipulation of colors, fonts, and typefaces (Figure 5).

The Table of Contents for a book is created from actual elements within the book content, rather than from the Table of Contents given in the book frontmatter of the hard copy. This ensures that the Table of Contents represents the content accurately as it is organized on BookShelf.

NCBI Book DTD

History

In the first version of the NCBI BookShelf project, Quark files were converted directly to HTML for display online. The result was effective, illustrating the value of having a textbook online and linked to PubMed; however, it was labor intensive, limiting, and not scaleable.

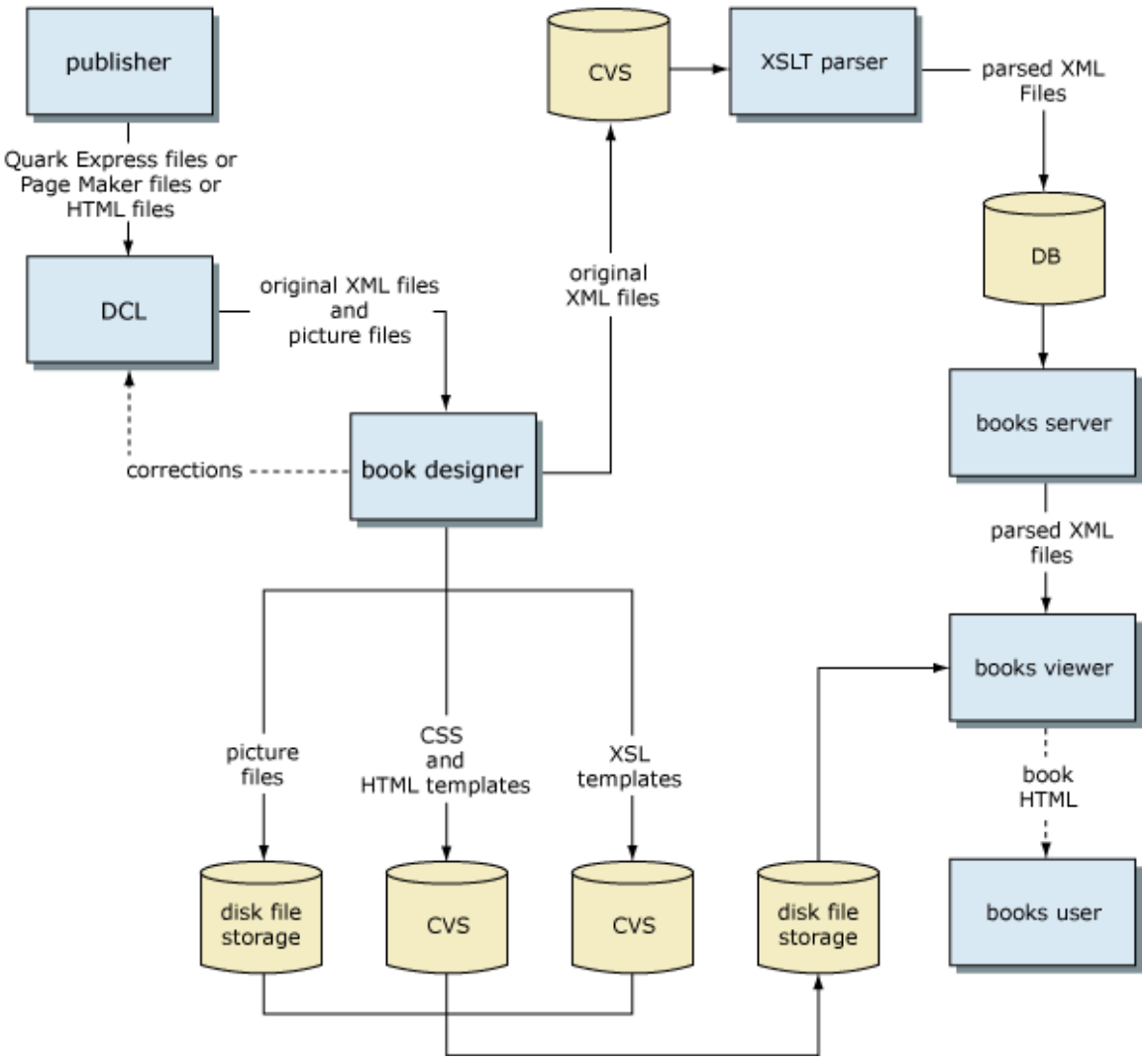


Figure 5. **Processing and data flow of books.**

To simplify the delivery of books online and to allow for the expansion of linking within the Entrez system, NCBI decided to convert all content into a centralized XML format. The normalized XML content is easier to render, allows added value such as the addition of links to other NCBI databases, and simplifies the addition of new volumes.

PMC created a new DTD for the BookShelf project, which was based on the ISO-12803 DTD. As more books were converted to the NCBI Book DTD, changes had to be made to accommodate the data.

The NCBI Book DTD is a public DTD available on request from books@ncbi.nlm.nih.gov.

Frequently Asked Questions

1. How do I access the books at NCBI?

The online books can be accessed by direct searching in Entrez or through PubMed abstracts. After performing a general PubMed search, click on the author name of one of the search results to view the abstract. A hypertext link called **Links** is displayed to the right of the abstract title. This link contains a drop-down menu consisting of various choices, depending upon the specific abstract. Choosing **Books** from this drop-down menu will highlight keywords in the abstract that, when selected, initiate a search of all BookShelf content for that particular term.

2. Which books are available at NCBI?

The book list is updated on a regular basis and can be viewed on the BookShelf homepage: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books.

3. Can I search the books at NCBI?

Yes, the books can be searched either as a complete collection or as a single, selected book (restricted using search options found under **Limits**).

4. Can I browse the whole book?

The system has been designed so that the user is delivered to the most relevant book sections for a particular term or concept. Although navigation is possible in the immediate vicinity of the page to which you are delivered, it may not be possible to browse the complete book on BookShelf. The range of navigation for each book is determined on a case-by-case basis, in agreement with the publisher.

5. I am the publisher/author/editor of a book. How can I participate?

Please email books@ncbi.nlm.nih.gov to discuss potential projects.

Chapter 9. PubMed Central (PMC): An Archive for Literature from Life Sciences Journals

Jeff Beck and Ed Sequeira

Created: October 9, 2002; Updated: August 13, 2003.

Summary

PubMed Central (PMC) is the National Library of Medicine's digital archive of full-text journal literature. Journals deposit material in PMC on a voluntary basis. Articles in PMC may be retrieved either by browsing a table of contents for a specific journal or by searching the database. Certain journals allow the full text of their articles to be viewed directly in PMC. These are always free, although there may be a time lag of a few weeks to a year or more between publication of a journal issue and when it is available in PMC. Other journals require that PMC direct users to the journal's own Web site to see the full text of an article. In this case, the material will always be available free to any user no more than 1 year after publication but will usually be available only to the journal's subscribers for the first 6 months to 1 year.

To increase the functionality of the database, a variety of links are added to the articles in PMC: between an article correction and the original article; from an article to other articles in PMC that cite it; from a citation in the references section to the corresponding abstract in PubMed and to its full text in PMC; and from an article to related records in other Entrez databases such as Reference Sequences, OMIM, and Books.

A PubMed Central (PMC) Site Guide

The PMC homepage has a list of all journals available in PMC and the earliest available issue for each journal (Figure 1). From here, a table of contents for the latest available issue of a journal or a list of all issues of the journal available through PMC can be viewed.

Every article citation in a table of contents includes one or more links (Figure 2). Articles for which the full text is available directly in PMC generally have links to an Abstract view, a Full Text view, and a PDF (printable view). Where applicable, they also have links to Corrections and to supplementary data that may be available for the article. In cases where the full text is available only at the journal publisher's site, there is only one link, to a PubLink page (described below).

In addition to the header information for the article itself, the upper part of a Full Text or PubLink page contains a variety of links, including links to other forms of the article, to related information in PubMed and other Entrez databases, and to corrections or "cited-in" lists where these apply (Figure 3). The sidebar in the body of a Full Text page (Figure 4) has links to tables and thumbnail images of any figures in the article, which when selected will display the full figure. Figures and tables may also be opened directly from the point in the text where they are referenced. Citations in the References section of an

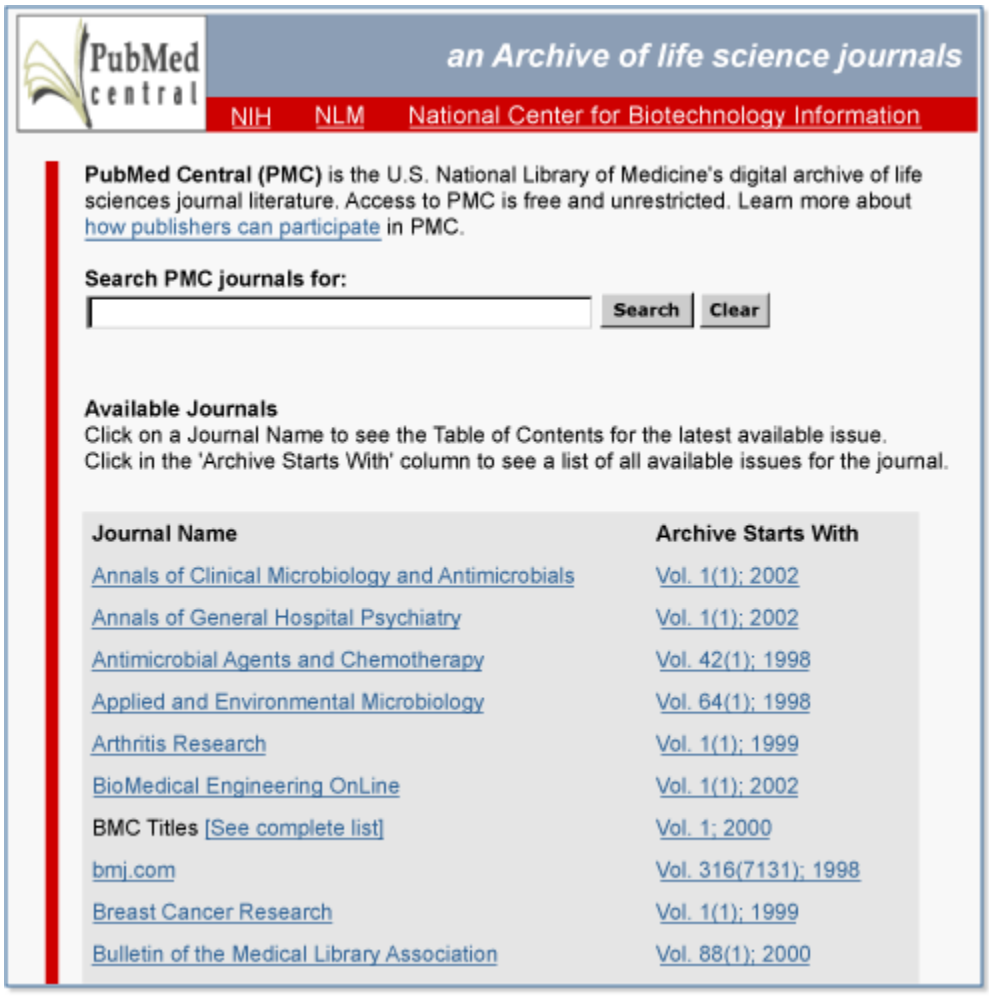


Figure 1. **The PMC journal list.**

article frequently include a link to the corresponding PubMed abstract and sometimes also have a link to the full text of the referenced article in PMC (Figure 5).

An Abstract page is identical to a Full Text page that has been cut off at the end of the abstract.

A PMC PubLink page (Figure 6) is similar to an Abstract page, except that it does not have links to alternate forms (Full Text or PDF) of the article in PMC. Instead, it contains a link to the full text at the publisher's site and information about when it will be freely available.

When an article has been cited by other articles in PMC, a “cited-in” link displays just under the article header information on both the Abstract and Full Text pages. Selecting this cited-in link gives you a list of the articles that have referenced the subject article (Figure 7).

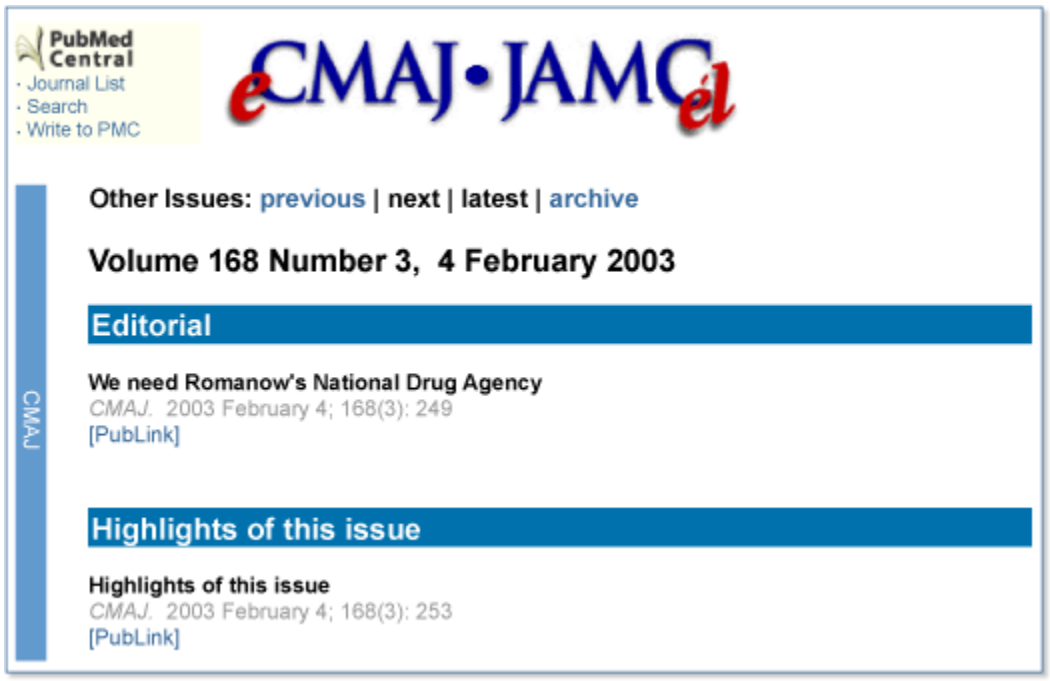


Figure 2. **PMC table of contents.**

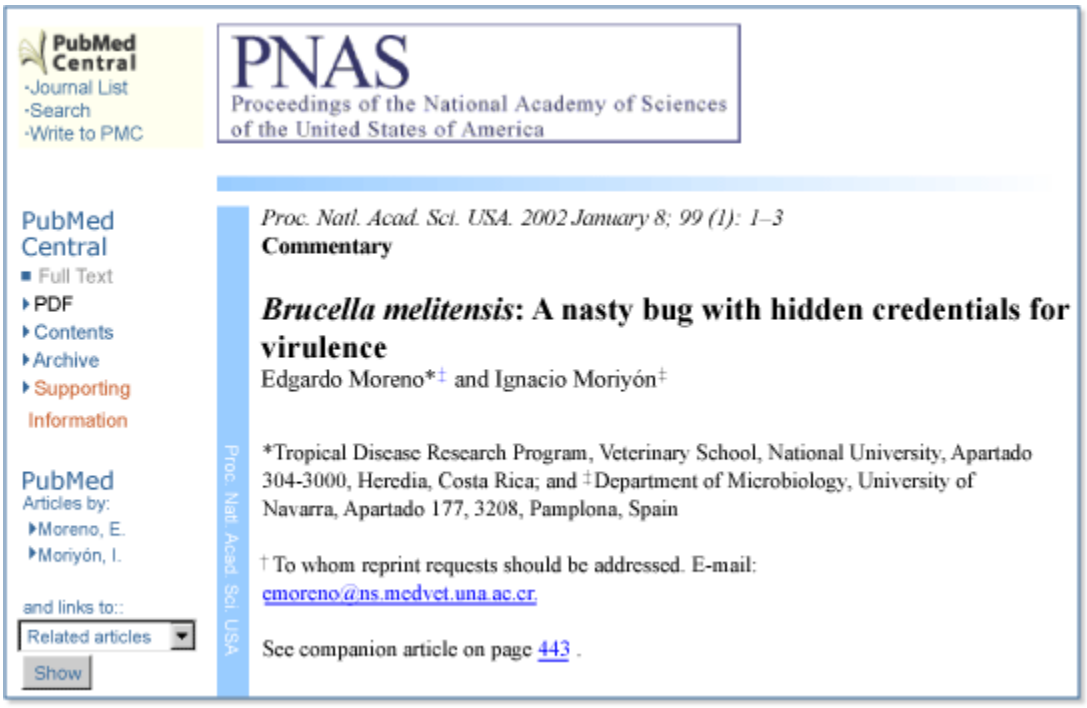


Figure 3. **Header of Abstract and Full Text pages.**

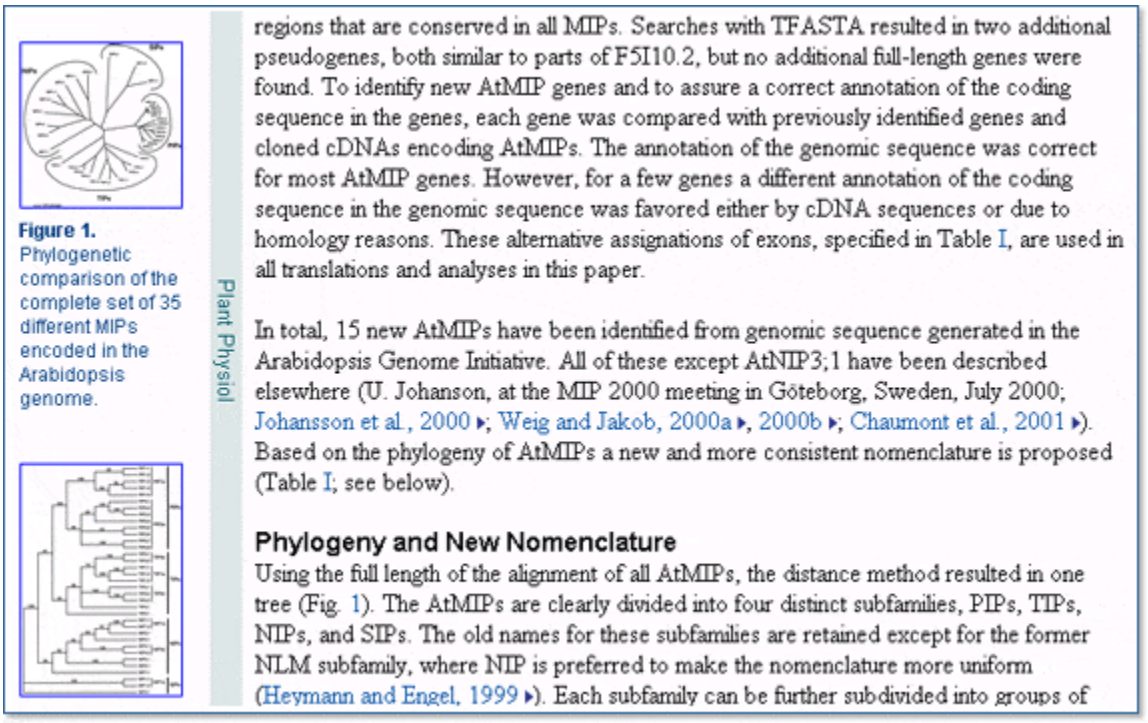


Figure 1.
Phylogenetic comparison of the complete set of 35 different MIPs encoded in the Arabidopsis genome.

Plant Physiol

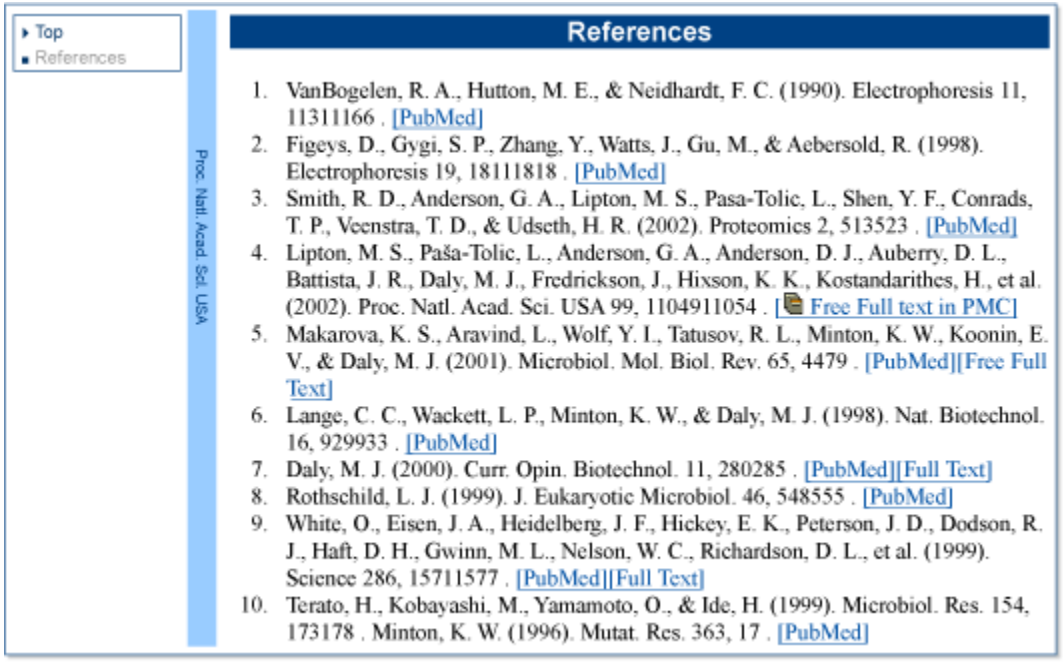
regions that are conserved in all MIPs. Searches with TFASTA resulted in two additional pseudogenes, both similar to parts of F5I10.2, but no additional full-length genes were found. To identify new AtMIP genes and to assure a correct annotation of the coding sequence in the genes, each gene was compared with previously identified genes and cloned cDNAs encoding AtMIPs. The annotation of the genomic sequence was correct for most AtMIP genes. However, for a few genes a different annotation of the coding sequence in the genomic sequence was favored either by cDNA sequences or due to homology reasons. These alternative assignations of exons, specified in Table I, are used in all translations and analyses in this paper.

In total, 15 new AtMIPs have been identified from genomic sequence generated in the Arabidopsis Genome Initiative. All of these except AtNIP3;1 have been described elsewhere (U. Johanson, at the MIP 2000 meeting in Göteborg, Sweden, July 2000; Johansson et al., 2000 ▸; Weig and Jakob, 2000a ▸, 2000b ▸; Chaumont et al., 2001 ▸). Based on the phylogeny of AtMIPs a new and more consistent nomenclature is proposed (Table I; see below).

Phylogeny and New Nomenclature

Using the full length of the alignment of all AtMIPs, the distance method resulted in one tree (Fig. 1). The AtMIPs are clearly divided into four distinct subfamilies, PIPs, TIPs, NIPs, and SIPs. The old names for these subfamilies are retained except for the former NLM subfamily, where NIP is preferred to make the nomenclature more uniform (Heymann and Engel, 1999 ▸). Each subfamily can be further subdivided into groups of

Figure 4. **Body of Full Text page.**



▸ Top
▪ References

Proc. Natl. Acad. Sci. USA

References

1. VanBogelen, R. A., Hutton, M. E., & Neidhardt, F. C. (1990). Electrophoresis 11, 11311166 . [PubMed]
2. Figeys, D., Gygi, S. P., Zhang, Y., Watts, J., Gu, M., & Aebersold, R. (1998). Electrophoresis 19, 18111818 . [PubMed]
3. Smith, R. D., Anderson, G. A., Lipton, M. S., Pasa-Tolic, L., Shen, Y. F., Conrads, T. P., Veenstra, T. D., & Udseth, H. R. (2002). Proteomics 2, 513523 . [PubMed]
4. Lipton, M. S., Paša-Tolic, L., Anderson, G. A., Anderson, D. J., Auberry, D. L., Battista, J. R., Daly, M. J., Fredrickson, J., Hixson, K. K., Kostandarithes, H., et al. (2002). Proc. Natl. Acad. Sci. USA 99, 1104911054 . [Free Full text in PMC]
5. Makarova, K. S., Aravind, L., Wolf, Y. I., Tatusov, R. L., Minton, K. W., Koonin, E. V., & Daly, M. J. (2001). Microbiol. Mol. Biol. Rev. 65, 4479 . [PubMed][Free Full Text]
6. Lange, C. C., Wackett, L. P., Minton, K. W., & Daly, M. J. (1998). Nat. Biotechnol. 16, 929933 . [PubMed]
7. Daly, M. J. (2000). Curr. Opin. Biotechnol. 11, 280285 . [PubMed][Full Text]
8. Rothschild, L. J. (1999). J. Eukaryotic Microbiol. 46, 548555 . [PubMed]
9. White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., et al. (1999). Science 286, 15711577 . [PubMed][Full Text]
10. Terato, H., Kobayashi, M., Yamamoto, O., & Ide, H. (1999). Microbiol. Res. 154, 173178 . Minton, K. W. (1996). Mutat. Res. 363, 17 . [PubMed]

Figure 5. **References.**

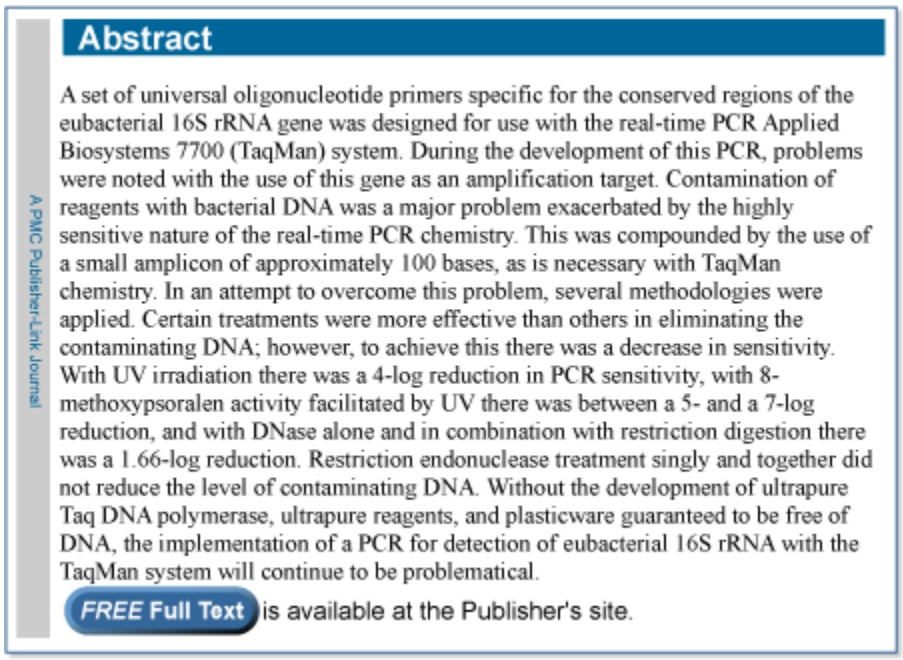


Figure 6. **PubLink page.**

PMC article citations may also be retrieved by doing a search in PMC or through a PubMed search. (In PubMed, use the **subsets** limit if you want to find only articles that are available in PMC.)

Participation in PMC

Participation by publishers in PMC is voluntary, although participating journals must meet certain editorial standards. A participating journal is expected to include all of its peer-reviewed primary research articles in PMC. Journals are encouraged to also deposit other content such as review articles, essays, and editorials. Review journals, and similar publications that have no primary research articles, are also invited to include their contents in PMC. However, primary research papers without peer review are not accepted.

Journals that deposit material in PMC may make the full text viewable directly in PMC or may require that PMC link to the journal site for viewing the complete article. In the latter case, the full text must be freely available at the journal site no more than 1 year after publication. In the case of full text that is viewable directly in PMC, which by definition is free, the journal may delay the release of its material for more than 1 year after publication, although all current journals have delays of 1 year or less.

In either case, the journal must provide SGML or XML for the full text, along with any related high-resolution image files. All data must meet PMC standards for syntactically correct and complete data.

Figure 7. **Cited-in list.**

The rationale behind the insistence on free access is that continued use of the material, which is encouraged by open access, serves as the best test of the durability and utility of the archive as technology changes over time. PMC does not claim copyright on any material deposited in the archive. Copyright remains with the journal publisher or with individual authors, whichever is applicable.

Refer to Information for Publishers to learn more about participating in PMC.

Links to Other NCBI Resources

From Abstract and Full Text pages in PMC are links to related articles in PubMed and to related records in other Entrez databases, such as Nucleotides or Books. These are identical to the links between databases that you can find in any Entrez record.

PMC Architecture

PubMed Central is an XML-based publishing system for full-text journal articles. All journal content in the archive was either supplied in, or has been converted to, a Document Type Definition (DTD) written at NCBI for the publication and storage of full-text articles.

The content is displayed dynamically on the PMC site by journal, volume, and issue (if applicable). XML, Web graphics, PDFs, and supplemental data are stored in a Sybase database. When a reader requests an article, the XML is retrieved from the database, and it is converted to HTML using XSLT stylesheets. The look of the HTML pages is controlled further by using Cascading Style Sheets (CSS), which allow manipulation of colors, fonts, and typefaces.

Data Flow: 1. SGML/XML Processing

We receive journal content either directly from publishers or from publishers' vendors. This content includes:

- SGML or XML of the articles to be deposited
- High-resolution images
- Supplemental data associated with the articles
- PDF versions of the articles

All of the text is converted to a central DTD, the PMC DTD, and the images are converted to Web format (GIF and JPEG). These files, along with any supplemental data or PDFs, are loaded into the database for linking, indexing, and retrieval (Figure 8). The source text in SGML or XML format is parsed against the source DTD. If the source files do not conform to the DTD, they are returned to the publisher or vendor for correction.

Once all of the files in an issue or batch have been validated, they are converted to PMC XML (referred to as a PMC XML file or PXML) using XSLT (Figure 9). Because XSLT is an XML conversion tool, SGML source files must first be converted to XML. This is done using SX, which is available from James Clark. The transformation will close any empty elements and insert ending tags for any element that is not closed.

For each publisher that submits SGML, an XML version of the DTD is created. This is used to parse the output of SX before the XML conversion is started. The XML version of the publisher's DTD is not used to validate the source data (because this has already been done using the original DTD).

XSLT requires that the input file be valid XML. If a DTD is not available for validation, the parser will check the syntax; it will also replace all of the character entities with the appropriate UTF-8 representation. This can cause a problem because the relationship between the characters in the input file and their UTF-8 representations may not be one

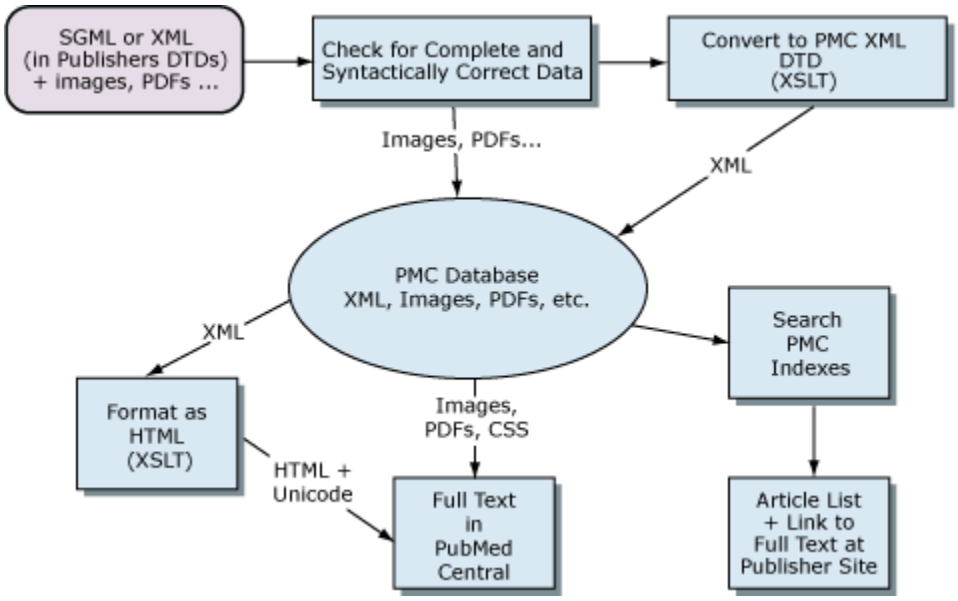


Figure 8. **Data flow.**

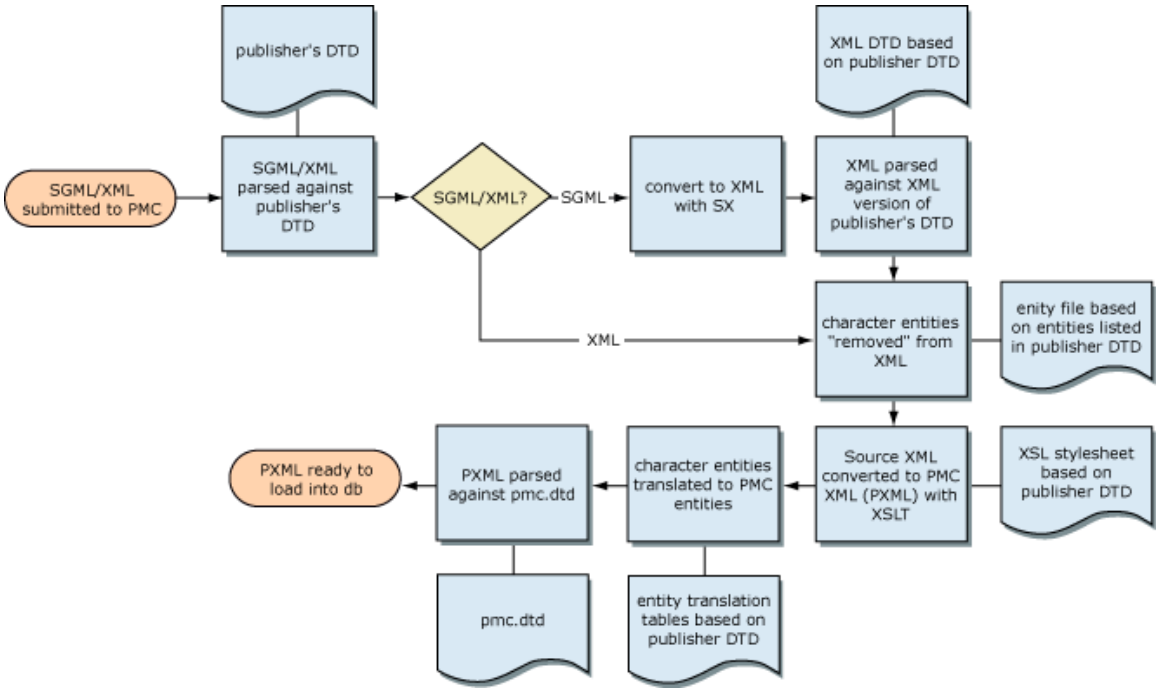


Figure 9. **Text conversion flow.**

to one. This means that characters translated to UTF-8 might not translate back to the original character entity accurately.

After the XSLT conversion, the original character entities are converted to character entities that are valid under the PMC DTD. Character translation tables for each source DTD regulate this conversion.

The resulting XML is then validated against the PMC DTD.

Several other items are created along with the PXML file. These are:

1. An entity file (articlename.ent). This file lists all of the character entities (from the PMC DTD-defined entity sets) that are in the article. One entity file is created for each article. This information is loaded into the database and is used to prepare the final HTML file for display. A sample is below:

```
agr
deg
Dgr
ldquo
lsqb
lt
pmc811
mdash
```

2. A PubMedID file (articlename.pmid). This file includes a set of reference citations in the format

```
Journal_Title|year|volume|first_page|AuthorName|Refno|pubmedid
```

When the article is converted, this information is collected from each journal citation in the bibliography and sent in a query to PubMed (using the Citation Matcher utility). If a value is returned, it is written in the last field. If no value is returned, an error message is written in this field. This information is saved so that if the article ever needs to be reconverted, the PubMed IDs will not need to be looked up again. A sample is below:

```
Nucleic Acids Res|1992|20|2673|Murray JM|13:2:493:16|1319571
J Biol Chem|1999|274|35297|Naumann M|13:2:493:17|10585392
EMBO J|2000|19|3475|Osaka F|13:2:493:18|10880460
Nature|2000|405|662|Osterlund MT|13:2:493:19|0
FASEB J|1998|12|469|Seeger M|13:2:493:20|9535219
```

3. A source node file (articlename.src). When an article is passed through the XSLT conversion, a list is made of each node, or named piece of information, that is included in the file. As the conversion is running, each node that is being processed is recorded. When the conversion is complete, the processed node list is compared with the list of nodes in the source file, and any piece of information that was not processed is reported in a conversion log. A sample is below:

```
/ART
```

```
/ART/@AID
/ART/@DATE
/ART/@ISS
/ART/BM/FN/P/EXREF
/ART/BM/FN/P/EXREF/@ACCESS
/ART/BM/FN/P/EXREF/@TYPE
/ART/FM
/ART/FM/ABS
/ART/FM/ABS/P
/ART/FM/ABS/P/EMPH
/ART/FM/ACC
/ART/FM/ATL
/ART/FM/ATL/EMPH
/ART/FM/AUG
/ART/FM/PUBFRONT/CPYRT/CPYRTNME/COLLAB
/ART/FM/PUBFRONT/CPYRT/DATE
/ART/FM/PUBFRONT/CPYRT/DATE/YEAR
/ART/FM/PUBFRONT/DOI
/ART/FM/PUBFRONT/EXTENT
/ART/FM/PUBFRONT/FPAGE
/ART/FM/PUBFRONT/ISSN
/ART/FM/PUBFRONT/LPAGE
/ART/FM/RE
/ART/FM/RV
```

Image Processing

To accommodate the archiving requirements of the PubMed Central project, it is important that figures be submitted in the greatest resolution possible, in TIFF or EPS format. Figures in these formats will be available for data migration when formats change in the future, and PubMed Central will be able to keep all of the figures current.

For display on the PMC site, two copies of each figure are made: a GIF thumbnail (100 pixels wide) and a JPEG file that will be displayed with the figure caption when the figure is requested.

Supplemental Data Processing

Supplemental data include any supporting information that accompanies the article but is not part of the article. They may be text files, Word document files, spreadsheet files, executables, video, and others.

Sometimes a journal has a Web site where all of this supplemental information is stored. In this case, PubMed Central establishes links from the article to the supplemental information on the publisher's site.

In other cases, the supplemental data files are submitted with the article to be loaded into the PMC database. Either way, the information concerning this supplemental data is collected in a Supplemental Data file, which includes the location of the supplemental

file(s), the type of information that is available, and how the link should be built from the article. PMC does not validate any of the supplemental data files that are supplied.

Mathematics

Mathematical symbols and notations can be difficult to display in HTML because of built-up expressions and unusual characters. For the most part, expressions that are simple enough to display using HTML are not handled as math unless they are tagged as math specifically. Publishers that supply content to PMC handle math expressions in one of two ways: supplied images or encoding in SGML.

1. Math Images

Any expression that cannot be tagged by the source DTD is supplied as an image. In this case, PMC will pass the image callout through to the PXML file and display the supplied image in the HTML file.

2. Math in SGML

Several of the Source DTDs used by publishers to submit data to PMC are robust enough to allow coding of almost any mathematical expression in SGML. Most of these were derived from the Elsevier DTD; therefore, many of the elements are similar.

During article conversion, any items that are recognized as math are translated into Tex. This would include any expression tagged specifically as a "formula" or "display formula," as well as any free-standing expression that cannot be represented in HTML. These expressions include radicals, fractions, and anything with an overbar (other than accented characters). For example:

> "$x + y = 2z$" would not be recognized as a math expression, but "<formula>$x + y = 2z$</formula>" would be.
>
> "1/2" would not be recognized as a math expression, but <fraction><numerator>1</numerator><denominator>2 </denominator></fraction> would be.
>
> "<radical>2x</radical> would be recognized as a math expression, as would "<overbar>47X</overbar>."

The SGML:

```
<FD ID="E2">I<SUP><UP>o</UP>
</SUP><INF><UP>f</UP></INF>&cjs1134;I<INF><UP>f</UP></INF>=&phgr;
<SUP><UP>o</UP></SUP><INF><UP>f</UP></INF>&cjs1134;&phgr;<INF>
```

```
<UP>f</UP></INF>=1&plus;K<INF><UP>sv</UP></INF>&lsqb;<UP>Q
</UP>&rsqb;</FD>
```

Converts to:

```
<fd id="E2"><math mathtype="tex"
id="M2">\documentclass[12pt]{minimal}
\usepackage{wasysym}
\usepackage[substack]{amsmath}
\usepackage{amsfonts}
\usepackage{amssymb}
\usepackage{amsbsy}
\usepackage[mathscr]{eucal}
\usepackage{mathrsfs}
\DeclareFontFamily{T1}{linotext}{}
\DeclareFontShape{T1}{linotext}{m}{n} { <-> linotext }{}
\DeclareSymbolFont{linotext}{T1}{linotext}{m}{n}
\DeclareSymbolFontAlphabet{\mathLINOTEXT}{linotext}
\begin{document}
\[
I^{o}_{f}/I_{f}={\phi}^{o}_{f}/{\phi}_{f}=1+K_{sv}[Q]
\]
\end{document}
</math></fd>
```

When the articles are loaded into the database, the equation markup is written into an Equation table. This table will also include the equation image, which will be created from the Tex markup.

The image for the equation shown above in SGML and PXML (with Tex) is:

$$I_f^o/I_f = \phi_f^o/\phi_f = 1 + K_{sv}[Q]$$

Data Flow: 2. Loading the Database

Because all of the content in PubMed Central is in the same format—PMC DTD—loading articles into the database is relatively straightforward. Once an article is loaded into the production (public) database, it will retain its ArticleId (article ID number) in perpetuity. On loading, each article is validated against the PMC DTD. Also, any external files that are referenced by the XML are checked. If any file, such as a figure, is missing, the loading will be aborted.

The database loading software and daily maintenance programs perform several other tests to ensure the accuracy and vitality of the archive:

1. Journal identity. The Journal title being loaded is verified against the ISSN number in the PXML to verify that the journal identity is correct.
2. Duplicate articles. An article may not be loaded more than once. Any changes to the article must be submitted as a replacement article, which will use the same ArticleId.

3. Publication date/delay. Rules for delay of publication embargo can be set up in the database to ensure that an issue will not be released to the public before a certain amount of time has passed since the publisher made the issue available.
4. PubMed IDs. PubMedIDs for the article being loaded or any bibliographic citation in the article that are not defined in the PXML are looked up upon loading.
5. Link updates. Links between related articles and from articles to external sources are updated daily.

The database has been designed to allow multiple versions of articles. In addition to article information, the database also stores information on content suppliers and publishers and journal-specific information.

Special Characters

PubMed Central uses a number of standard ISO character sets (8879 and 9573), along with a set of characters that has been defined to accommodate characters not in the standard set. The ISO Standard Character sets referenced are listed in Box 1.

Each publisher DTD defines a set of characters that may be used in their articles. Generally, these publisher DTDs use the same standard ISO character sets listed in Box 1. Any character that cannot be represented by the standard ISO sets is defined in a publisher-specific character set. These publisher-specified characters are converted into characters in the PMC entity list during conversion (see *SGML/XML Processing*). The PMC entity list is publicly available.

The supplied data also include groups of entities that are to be combined in the final document. Sometimes these are grouped in a tag such as:

```
<A><AC>&alpha;</AC><AC>&acute;</AC></A>
```

and sometimes they are just positioned next to each other in the text. These combined entities must be mapped either to an ISO character or to a character in the PMC character set.

For the most flexibility in displaying characters across platforms, PMC uses UTF-8 encoding whenever possible. Because not all browsers support the same subset of UTF-8 characters and some characters cannot be represented in UTF-8, PMC displays characters as a combination of GIFs and UTF-8 characters, depending on the Browser/OS combination and the character to be displayed.

Box 1. ISO Standard Character sets used by PMC.

```
<!ENTITY % ISOlat1 PUBLIC "ISO 8879-1986//ENTITIES Added Latin 1//EN">
<!ENTITY % ISOlat2 PUBLIC "ISO 8879-1986//ENTITIES Added Latin 2//EN">
<!ENTITY % ISOnum PUBLIC "ISO 8879-1986//ENTITIES Numeric and Special
```

Box 1 continues on next page...

Box 1 continued from previous page.

```
Graphic//EN">
<!ENTITY % ISOpub PUBLIC "ISO 8879-1986//ENTITIES Publishing//EN">
<!ENTITY % ISOgrk1 PUBLIC "ISO 8879-1986//ENTITIES Greek Letters//EN">
<!ENTITY % ISOgrk2 PUBLIC "ISO 8879-1986//ENTITIES Monotoniko Greek//
EN">
<!ENTITY % ISOtech PUBLIC "ISO 8879-1986//ENTITIES General Technical//
EN">
<!ENTITY % ISOdia PUBLIC "ISO 8879-1986//ENTITIES Diacritical Marks//
EN">
<!ENTITY % ISOAMSO PUBLIC "ISO 9573-13:1991//ENTITIES Added Math
Symbols: Ordinary //EN">
<!ENTITY % ISOAMSR PUBLIC "ISO 9573-13:1991//ENTITIES Added Math
Symbols: Relations //EN">
<!ENTITY % ISOamsr PUBLIC "ISO 8879-1986//ENTITIES Added Math Symbols:
Relations//EN">
<!ENTITY % ISOamsn PUBLIC "ISO 8879-1986//ENTITIES Added Math Symbols:
Negated Relations//EN">
<!ENTITY % ISOAMSA PUBLIC "ISO 9573-13:1991//ENTITIES Added Math
Symbols: Arrow Relations //EN">
<!ENTITY % ISOAMSB PUBLIC "ISO 9573-13:1991//ENTITIES Added Math
Symbols: Binary Operators //EN">
<!ENTITY % ISOamsc PUBLIC "ISO 8879-1986//ENTITIES Added Math Symbols:
Delimiters//EN">
<!ENTITY % ISOmopf PUBLIC "ISO 9573-13:1991//ENTITIES Math Alphabets:
Open Face//EN">
<!ENTITY % ISOmscr PUBLIC "ISO 9573-13:1991//ENTITIES Math Alphabets:
Script//EN">
<!ENTITY % ISOmfrk PUBLIC "ISO 9573-13:1991//ENTITIES Math Alphabets:
Fraktur//EN">
<!ENTITY % ISObox PUBLIC "ISO 8879:1986//ENTITIES Box and Line Drawing//
EN">
<!ENTITY % ISOcyr1 PUBLIC "ISO 8879:1986//ENTITIES Russian Cyrillic//
EN">
<!ENTITY % ISOcyr2 PUBLIC "ISO 8879:1986//ENTITIES Non-Russian
Cyrillic//EN">
<!ENTITY % ISOGRK3 PUBLIC "ISO 9573-13:1991//ENTITIES Greek Symbols //
EN">
<!ENTITY % ISOGRK4 PUBLIC "ISO 9573-13:1991//ENTITIES Alternative Greek
Symbols //EN">
<!ENTITY % ISOTECH PUBLIC "ISO 9573-13:1991//ENTITIES General
Technical //EN">
```

PMC DTD

History

In the first version of the PMC project, the SGML and XML were loaded into a database in its native format. The HTML rendering software was then required to convert content

from different sources into normalized HTML on the fly when a reader requested an article.

This was slow and cumbersome on the rendering side and was not scaleable. At that time, PMC was receiving content for about five journals in two DTDs, the keton.dtd from HighWire Press and the article.dtd from BioMed Central. The set-up for a new journal was difficult, and it soon became obvious that this solution would not scale easily.

To satisfy the archiving requirement for the PMC project and to simplify the delivery of articles online, PubMed Central decided to convert all content into a centralized format. The normalized content is easier to render, allows enhanced value such as links to other NCBI databases to be added, and simplifies content archiving.

PMC created a new DTD, which was strongly influenced by the BioMedCentral article.dtd and the keton.dtd. The original emphasis was on simplicity. As more and more articles from more and more journals were converted to the PMC DTD, changes had to be made to accommodate the data. The PMC DTD is publicly available.

Review and Revision of the PMC DTD

Because the PMC DTD grew rapidly, it was feared that the original "simplicity" of its design would lead to confusing data structures. With more and more publishers inquiring about submitting content directly in the PMC DTD, PubMed Central decided that an independent review was necessary. Mulberry Technologies, Inc., an electronic publishing consultancy specializing in SGML- and XML-based systems, reviewed the DTD and created a modified version.

At approximately the same time, under the auspices of a Mellon Grant to explore ejournal archiving, Harvard University Library contracted with Inera, Inc. to review a variety of DTDs from selected publishers, PMC included. The study focused on two key questions:

1. Can a common DTD be designed and developed into which publishers' proprietary SGML files can be transformed to meet the requirements of an archiving institution?
2. If such a structure can be developed, what are the issues that will be encountered when transforming publishers' SGML files into the archive structure for deposit into the archive?

The requirement of the archival article DTD was defined as the ability to represent the intellectual content of journal articles. This study is available and suggestions from the study were used in the NLM Archiving DTD Suite.

The NLM Archiving DTD will not be backwards-compatible with the pmc-1.dtd. It should be publicly available by the end of 2002, along with complete documentation for publishers and authors. A draft version is available (http://www.pubmedcentral.nih.gov/pmcdoc/dtd/nlm_lib/0.1/documentation/HTML/index.html), along with a draft version of the documentation.

Frequently Asked Questions

Please refer to the PMC site for answers to frequently asked questions.

Chapter 10. The SKY/CGH Database for Spectral Karyotyping and Comparative Genomic Hybridization Data

Turid Knutsen,[1] Vasuki Gobu,[2] Rodger Knaus,[2] Thomas Ried,[1] and Karl Sirotkin[2]

Created: October 9, 2002; Updated: August 13, 2003.

Summary

Spectral Karyotyping (SKY) (1–7) and Comparative Genomic Hybidization (CGH) (8–11) are complementary fluorescent molecular cytogenetic techniques that have revolutionized the detection of chromosomal abnormalities. SKY permits the simultaneous visualization of all human or mouse chromosomes in a different color, facilitating the detection of chromosomal translocations and rearrangements (Figure 1). CGH uses the hybridization of differentially labeled tumor and reference DNA to generate a map of DNA copy number changes in tumor genomes.

The goal of the SKY/CGH database is to allow investigators to submit and analyze both clinical and research (e.g., cell lines) SKY and CGH data. The database is growing and currently has a total of about 700 datasets, some of which are being held private until published. Several hundred labs around the world use this technique, with many more looking at the data they generate. Submitters can enter data from their own cases in either of two formats, public or private; the public data is generally that which has already been published, whereas the private data can be viewed only by the submitters, who can transfer it to the public format at their discretion. The results are stored under the name of the submitter and are listed according to case number. The homepage includes a basic description of SKY and CGH techniques and provides links to a more detailed explanation and relevant literature.

Database Content

Detailed information on how to submit data either to the SKY or CGH sectors of the database can be found through links on the homepage. What follows is a brief outline.

Spectral Karyotyping

The submitter enters the written karyotype, the number of normal and abnormal copies for each chromosome, and the number of cells for each clone. Each abnormal chromosome segment is then described by typing in the beginning and ending bands, starting from the top of the chromosome (Figure 2); the computer then builds a colored

[1] National Cancer Institute (NCI). [2] National Center for Biotechnology Information (NCBI).

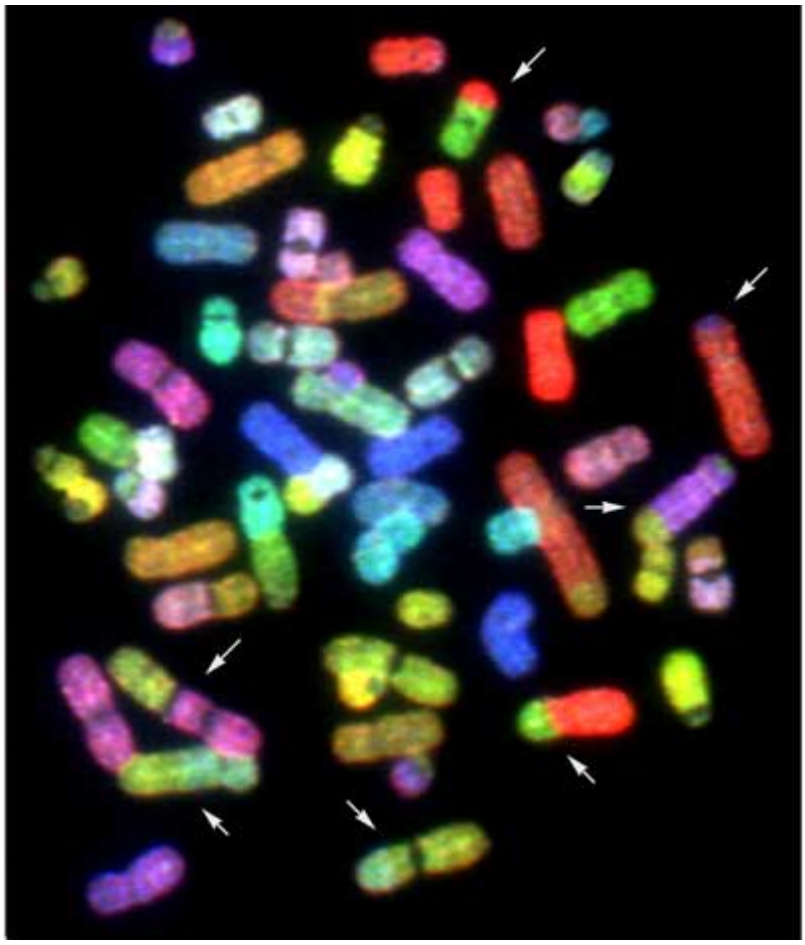


Figure 1. A metaphase spread after SKY hybridization. The RGB image demonstrates cytogenetic abnormalities in a cell from a secondary leukemia cell line. *Arrows* indicate some of the many chromosomal translocations in this cell line.

ideogram of this chromosome and eventually a full karyotype (SKYGRAM) with each normal and abnormal chromosome displayed in its unique SKY classification color, with band overlay (Figure 3). Each breakpoint submitted is automatically linked by a button marked FISH to the human Map Viewer (Figure 4; Chapter 20), which provides a list of genes at that site and available FISH clones for that breakpoint.

Comparative Genomic Hybridization

The CGH database displays gains, losses, and amplification of chromosomes and chromosome segments. The data are entered either by hand or automatically from various CGH software programs. In the manual format, the submitter enters the band information for each affected chromosome, describing the start band and stop band for each gain or loss, and the computer program displays the final karyotype with vertical bars to the left or right of each chromosome, indicating loss or gain, respectively (Figure 5).

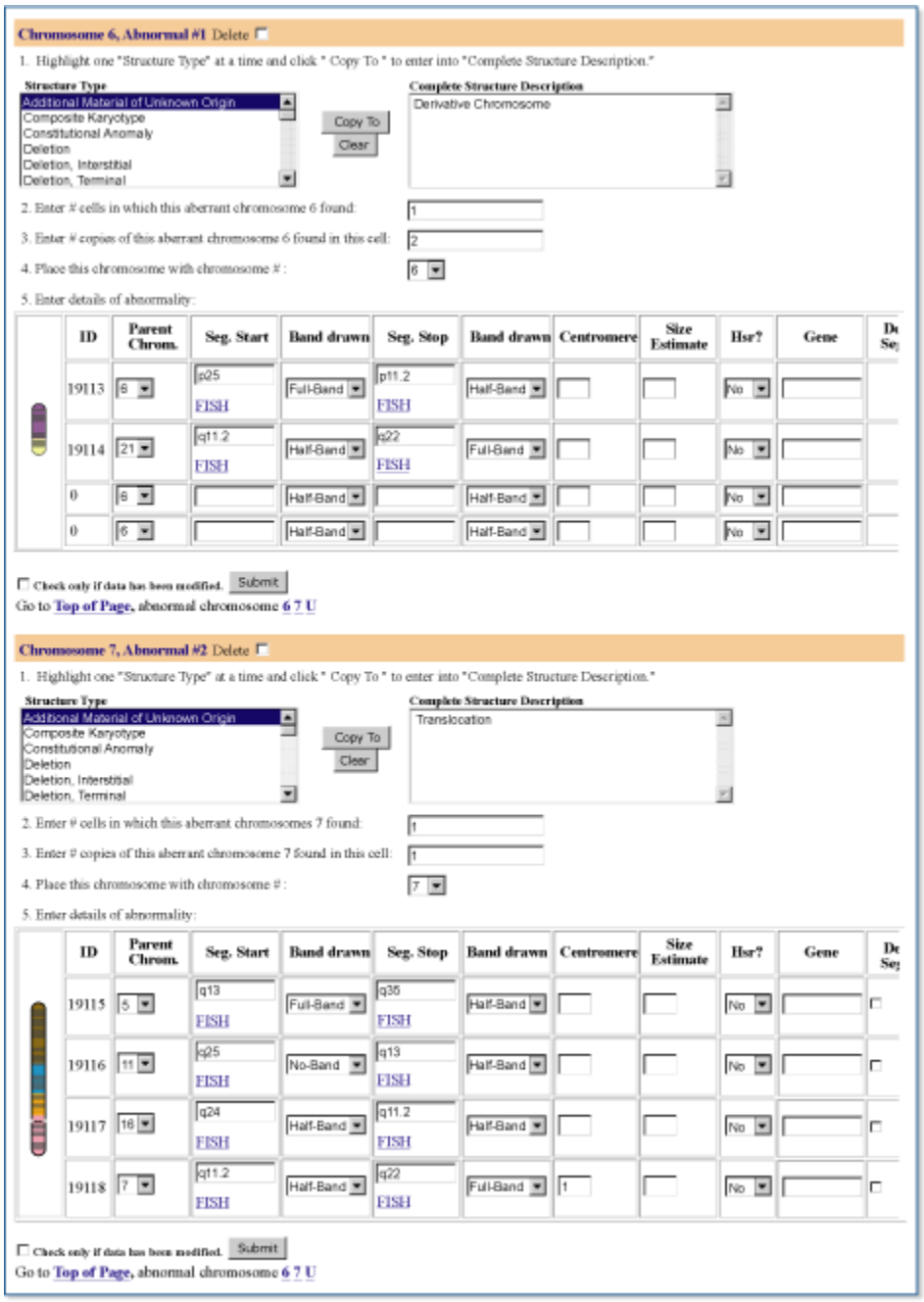


Figure 2. SKY data entry form for two different abnormal chromosomes, built segment by segment, for the SKYGRAM image. Chromosome images on the *left* are the result of entering the start (*top*) and stop (*bottom*) band for each segment.

Case Information

Clinical data submitted include case identification, World Health Organization (WHO) disease classification code, diagnosis, organ, tumor type, and disease stage. To obtain the

Figure 3. A SKYGRAM from the SKY/CGH database demonstrating cytogenetic abnormalities in a transitional cell carcinoma of the bladder cell line. The written karyotype and a one-line case summary are also provided.

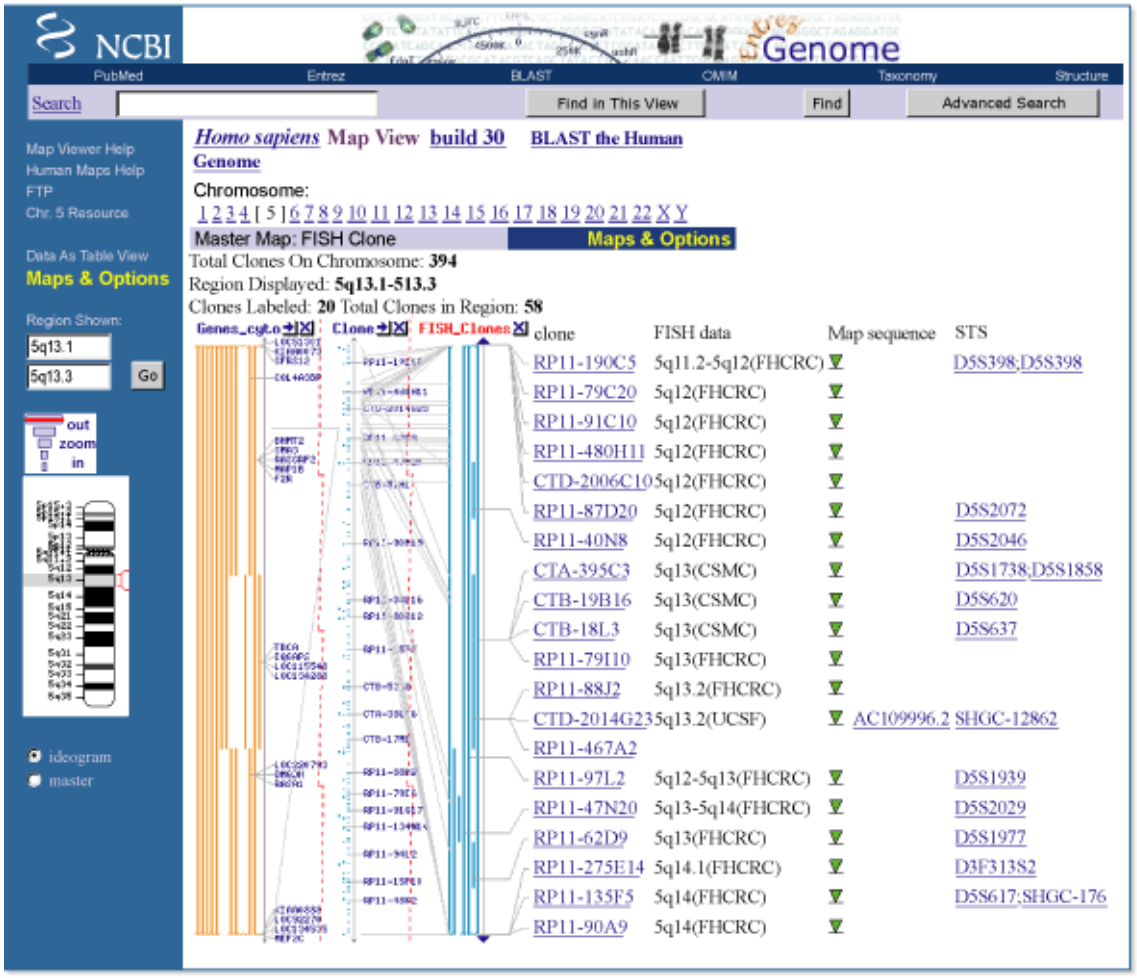


Figure 4. Map Viewer image depicting the information on genes, clones, FISH clones, map sequences, and STSs for a specific chromosomal breakpoint (5q13) identified in a SKYGRAM image.

correct classification code, a link is provided to the NCI's Metathesaurus ^{TM} site, which includes, among its many systems, the codes developed by the WHO and NCI, and published as the International Classification of Diseases, 3rd edition (ICD-O-3).

Reference Information

The references for the published cases are entered into the Case Information page and are linked to their abstracts in PubMed.

Tools for Data Entry

SKYIN

A colored karyotype with band overlay is presented to the submitter, who then builds each aberrant chromosome by cutting and pasting (by clicking with the mouse at appropriate

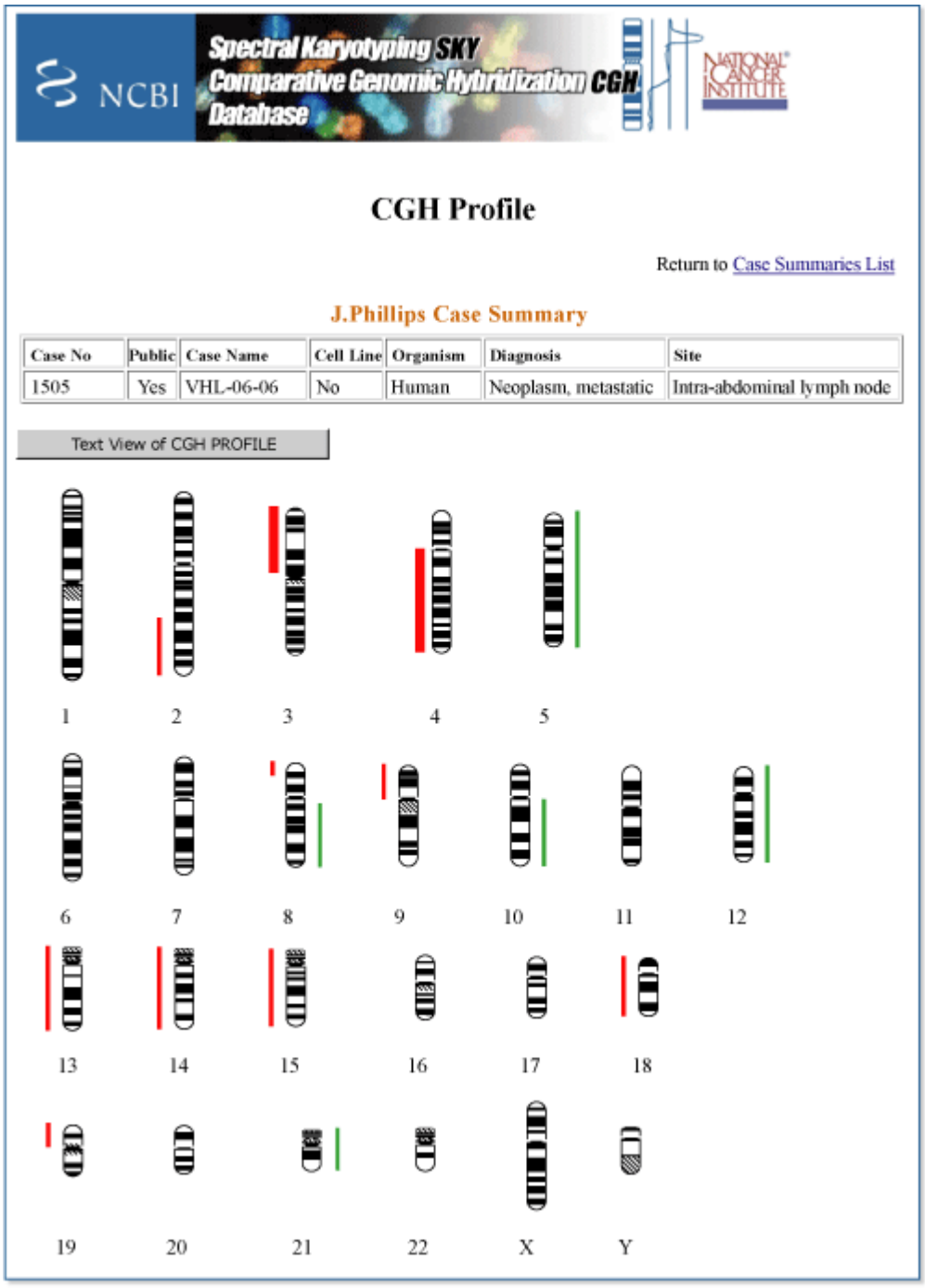


Figure 5. A CGH profile from the SKY/CGH database demonstrating copy number gains and losses in a tumor cell line established from a metastatic lymph node in von Hippel-Lindau renal cell carcinoma.

breakpoints) (Figure 6). Each aberrant chromosome is then inserted into the full karyotype, which is displayed as a SKYGRAM.

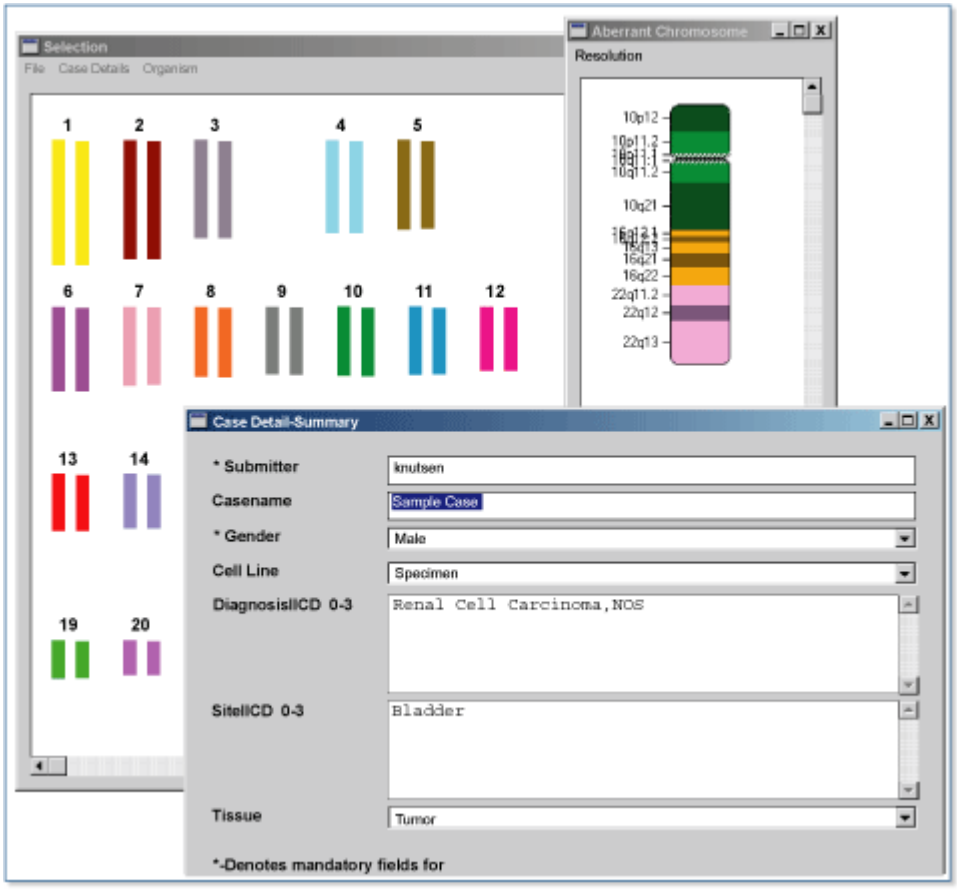


Figure 6. SKYIN format. Clicking on a chromosome brings up that chromosome with band overlay. Using the cursor, the operator cuts and pastes together each abnormal chromosome. The abnormal chromosome shown is a combination of chromosomes 10, 16, and 22. *Inset,* an example of the clinical information entered for a case.

Karyotype Parser

To speed up the entry of cytogenetic data into the database, NCBI has built a computer program to automatically read short-form karyotypes, extract the information therein, and insert it into the SKY database (Figure 7). Karyotypes are written according to specific rules described in *An International System for Human Cytogenetic Nomenclature (1995)* (12). Using these rules, the parser (1) breaks the karyotype into small syntactic components, (2) assembles information from these components into an information structure in computer memory, (3) transforms this information into the formats required for an application, and (4) uses the information in the application, i.e., inserts it into the database. To accomplish this, the syntactic parser first extracts the information out of each piece of the input; the pieces are then put directly into a tree structure that represents karyotype semantics. For insertion into the SKY database, the karyotype information is transformed into ASN.1 structures that reflect the design of the database.

Figure 7. Karyotype Parser. The short-form written karyotype, entered in the karyotype field in (*a*), has been converted into a modified long-form karyotype (*b*), which describes each abnormal chromosome from *top* to *bottom*. Both short and long terms use standardized symbols and abbreviations specified by ISCN.

NCI Metathesaurus

Data submitters must use the same terminology for diagnosis (morphology) and organ site (topography) to permit comparison or combination of the data in the SKY/CGH database. From the many different disease classification systems, the *International Classification of Diseases for Oncology*, 3rd edition (ICD-O-3)(13) was selected as the database's standard. It contains a morphology tree and a topography tree. In most cases, the submitter must select one term from each tree to fully classify a case. To find and select the correct ICD-O-3 morphology and topography terms, the user is referred to NCI Metathesaurus, a comprehensive biomedical terminology database, produced by the NCI

Center for Bioinformatics Enterprise Vocabulary Service. This tool facilitates mapping concepts from one vocabulary to other standard vocabularies.

Data Analysis: Query Tools

Quick Search Format

Quick Search can be found at the top of the SKY/CGH homepage and can be used for several types of information in the database; these are defined in Searchable Topics in the Help section. Topics include cytogenetic information (whole chromosome, chromosome arm, or chromosome breakpoint), submitter name, case name, cell line by name, diagnosis, site of disease, treatment, hereditary disorders, mouse strain, and genotype. One or more terms can be entered, and there are options to search SKY alone, CGH alone, SKY AND CGH, or the default, SKY OR CGH.

The query results page displays information on all relevant cases, clones, and cells, along with details of SKY and/or CGH studies and clinical information for each case.

Advanced Search Format

All of the public clinical and cytogenetic information can be searched. This format is currently under development.

The CGH Case Comparison Tool

This tool compares and summarizes the CGH profiles from multiple cases on one ideogram. There are numerous criteria that can be used for comparison, such as diagnosis, tumor site, mouse strain, and gain or loss of specific chromosomes, chromosome arms, or chromosome bands.

Data Integration

Integration with the NCBI Map Viewer

All chromosomal bands, including breakpoints involved in chromosomal abnormalities, are linked to the Map Viewer database (Figure 4 ; see also Chapter 20) The Map Viewer provides graphical displays of features on NCBI's assembly of human genomic sequence data as well as cytogenetic, genetic, physical, and radiation hybrid maps. Map features that can be seen along the sequence include NCBI contigs, the BAC tiling path, and the location of genes, STSs, FISH mapped clones, ESTs, GenomeScan models, and variation (SNPs; see Chapter 5).

SKY/CGH Database Links

Links are provided to related Web sites including: chromosome databases (e.g., the Mitelman database); other NCI (e.g., CGAP and CCAP) and NCBI [e.g., the Map Viewer

(Chapter 20), Entrez Gene (Chapter 19) resources; and PubMed (Chapter 2)] sites; The Jackson Laboratory; and several other CGH sites.

Contributors

NCBI: Karl Sirotkin, Vasuki Gobu, Rodger Knaus, Joel Plotkin, Carolyn Shenmen, and Jim Ostell

NCI: Turid Knutsen, Hesed Padilla-Nash, Meena Augustus, Evelin Schröck, Ilan R. Kirsch, Susan Greenhut, James Kriebel, and Thomas Ried

References

1. Schröck E, du Manoir S, Veldman T, Schoell B, Wienberg J, Ferguson-Smith MA, Ning Y, Ledbetter DH, Bar-Am I, Soenksen D, Garini Y, Ried T. Multicolor spectral karyotyping of human chromosomes. Science. 1996;273:494–497. PubMed PMID: 8662537.
2. Liyanage M, Coleman A, du Manoir S, Veldman T, McCormack S, Dickson RB, Barlow C, Wynshaw-Boris A, Janz S, Wienberg J, Ferguson-Smith MA, Schröck E, Ried T. Multicolour spectral karyotyping of mouse chromosomes. Nat Genet. 1996;14:312–315. PubMed PMID: 8896561.
3. Ried T, Liyanage M, du Manoir S, Heselmeyer K, Auer G, Macville M, Schröck E. Tumor cytogenetics revisited: comparative genomic hybridization and spectral karyotyping. J Mol Med. 1997;75:801–814. PubMed PMID: 9428610.
4. Weaver ZA, McCormack SJ, Liyanage M, du Manoir S, Coleman A, Schröck E, Dickson RB, Ried T. A recurring pattern of chromosomal aberrations in mammary gland tumors of MMTV–c-myc transgenic mice. Genes Chromosomes Cancer. 1999;25:251–260. PubMed PMID: 10379871.
5. Knutsen T, Ried T. SKY: a comprehensive diagnostic and research tool. A review of the first 300 published cases. J Assoc Genet Technol. 2000;26:3–15.
6. Padilla-Nash HM, Heselmeyer-Haddad K, Wangsa D, Zhang H, Ghadimi BM, Macville M, Augustus M, Schröck E, Hilgenfeld E, Ried T. Jumping translocations are common in solid tumor cell lines and result in recurrent fusions of whole chromosome arms. Genes Chromosomes Cancer. 2001;30:349–363. PubMed PMID: 11241788.
7. Phillips JL, Ghadimi BM, Wangsa D, Padilla-Nash H, Worrell R, Hewitt S, Walther M, Linehan WM, Klausner RD, Ried T. Molecular cytogenetic characterization of early and late renal cell carcinomas in Von Hippel-Lindau (VHL) disease. Genes Chromosomes Cancer. 2001;31:1–9. PubMed PMID: 11284029.
8. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258:818–821. PubMed PMID: 1359641.
9. Heselmeyer K, Schröck E, du Manoir S, Blegen H, Shah K, Steinbeck R, Auer G, Ried T. Gain of chromosome 3q defines the transition from severe dysplasia to invasive

carcinoma of the uterine cervix. Proc Natl Acad Sci U S A. 1996;93:479–484. PubMed PMID: 8552665.
10. Ried T, Liyanage M, du Manoir S, Heselmeyer K, Auer G, Macville M, Schröck E. Tumor cytogenetics revisited: comparative genomic hybridization and spectral karyotyping. J Mol Med. 1997;75:801–814. PubMed PMID: 9428610.
11. Forozan F, Karhu R, Kononen J, Kallioniemi A, Kallioniemi OP. Genome screening by comparative genomic hybridization. Trends Genet. 1997;13:405–409. PubMed PMID: 9351342.
12. ISCN: An International System for Human Cytogenetic Nomenclature. In: Mitelman F, editor. Basel: S. Karger; 1995.
13. Fritz A, Percy C, Jack Andrew, Shanmugaratnam K, Sobin L, Parkin DM, Whelan S, editors. International Classification of Diseases for Oncology, 3rd ed. Geneva: World Health Organization; 2000.

Chapter 11. The Major Histocompatibility Complex Database, dbMHC

Adrienne Kitts, Michael Feolo, and Wolfgang Helmberg

Created: May 27, 2003; Updated: August 13, 2003.

Summary

One of the most intensely studied regions of the human genome is the Major Histocompatibility Complex (MHC), a group of genes that occupies approximately 4–6 megabases on the short arm of chromosome 6. The MHC genes, known in humans as Human Leukocyte Antigen (HLA) genes, are highly polymorphic and encode molecules involved in the immune response. The MHC database, dbMHC, was designed to provide a neutral platform where the HLA community can submit, edit, view, and exchange MHC data. It currently consists of an interactive Alignment Viewer for HLA and related genes, an MHC microsatellite database (dbMHCms), a sequence interpretation site for Sequencing Based Typing (SBT), and a Primer/Probe database. dbMHC staff are in the process of creating a new database that will house a wide variety of HLA data including:

- Detailed single nucleotide polymorphism (SNP) mapping data of the HLA region
- KIR gene analysis data
- HLA diversity/anthropology data
- Multigene haplotype data
- HLA/disease association data
- Peptide binding prediction data

The MHC database is fully integrated with other NCBI resources, as well as with the International Histocompatibility Working Group (IHWG) Web site, and provides links to the IMmunoGeneTics HLA (IMGT/HLA) database.

This chapter provides a detailed description of current dbMHC resources and reviews dbMHC content and data computation protocols.

Introduction

The most important known function of HLA genes is the presentation of processed peptides (antigens) to T cells, although there are many genes in the HLA region yet to be characterized, and work continues to locate and analyze new genes in and around the MHC (1–4). The vast amount of work required to elucidate the MHC led to a series of International Histocompatibility Workshop/Conferences (IHWCs). These conferences were the genesis for the IHWG, a group of researchers actively involved in projects that lead to shared data resources for the HLA community (5–16). dbMHC is a permanent public archive that was conceived by the IHWG and is implemented and hosted by NCBI. This database is intended to be expandable; if you have ideas for additional resources, please send an email to helmberg@ncbi.nlm.nih.gov.

The staff of the MHC database are currently accepting online submissions for the Primer/Probe/Mix component of dbMHC. Submissions can include typing data from any of the following: Sequence Specific Oligonucleotides (SSOs), Sequence Specific Primers (SSPs), SSO and/or SSP mixes, HLA typing kits, and Sequencing Based Typing (SBT). dbMHC allows submitters to edit their submissions online at any time.

dbMHC Resources

Box 1 contains information on setting up accounts for using dbMHC.

Box 1. dbMHC guests and dbMHC accounts.

A user can access dbMHC resources as a guest, that is, without having or being a member of an account. However, dbMHC guests will be unable to submit data to dbMHC or edit existing data. Data from a guest session will not be saved from session to session or even from frame to frame. Guests do have the option to download data from a particular session.

To create a dbMHC account, select the "Create an Account" from the left sidebar of the dbMHC homepage. Here, you will provide institutional information and specify an account administrator. Only the account administrator is allowed to do the following:

- Enter new users.
- View or edit existing users.
- Change user permissions. This includes permission to modify allele reactivities and to enter new primers/probes/mixes or typing kits and to modify existing ones.
- Edit institutional information.

Alignment Viewer

The Alignment Viewer is designed to display pre-compiled allele sequence alignments in an aligned or FASTA format for selected loci. It offers an interactive display of the alignment, where users can select alleles, highlight SNPs within the alignment, and change between a Codon display and a Decade display (blocks of 10 nucleotides). Users can also switch the type of display, from one where the sequence is completely written out to one where only the differences between selected sequences and a pre-selected reference sequence are seen. All sequences displayed in the Alignment Viewer can be downloaded and formatted as alignment, FASTA, or XML. If the **Alignment** option is selected, it is up to the user to define how the alignment should be organized. The **Alignment** option allows the user to download the entire alignment, or just a section of it, and also allows the user to specify how the sequence will be displayed (e.g., in groups of 10 nucleotides or amino acids).

dbMHC can be used as a tool to help design primers/probes, to evaluate the reactivity patterns of potential probes, and to evaluate the polymorphism content of a particular gene.

MHC Microsatellite Database (dbMHCms)

dbMHCms contains many of the known microsatellites across the HLA region and is designed to search for descriptive information, when available, about these highly heterozygotic short tandem repeats (STRs). The MHC microsatellite information that users will be able to extract using dbMHCms includes:

- Physical location
- Number and length of known alleles
- Allelic motifs
- Informativity
- Heterozygosity
- Primer sequences

Users will find the markers shown on the dbMHCms Web page useful because they provide evidence for genetic linkage and/or association of a genomic region with disease susceptibility. This resource was developed by NCBI in collaboration with A. Foissac, M. Salhi, and Anne Cambon-Thomsen, who provided the original data on MHC microsatellites in a series of updates (17–19). dbMHCms will be expanded to include the STR markers used in the 13th IHWG workshop.

Sequencing Based Typing (SBT) Interface

The SBT interface is designed to analyze sequence information provided by users and provides an interpretation of the sequence that includes potential allele assignments and division of the sequence into exons and introns. Sequence can be submitted to the SBT interface by copying and pasting or by uploading from a file, either as haploid strands or heterozygote sequence. The current format restricts input to text or FASTA-formatted sequence.

Submitted sequences are aligned to reference locus sequences located in the dbMHC allele database based on a user-defined degree of nucleotide mismatch. After a brief analysis, the SBT interface will display exons, introns, and untranslated regions for each sequence. Allele assignments are listed according to matching order. Mismatched nucleotide positions are listed separately (in the lower frame of the SBT interface) after selecting **Alignment**. Note that the SBT interface analyzes one or more sequences for a single locus. If sequences from multiple loci that have identical group-specific amplifications are submitted, the SBT interface must be used to analyze sequences from only one locus at a time.

Primer/Probe Database

This database is designed to provide a comprehensive and standardized characterization of individual typing primers and probes, as well as primer and/or probe mixes, and encompasses the following technologies:

1. Probe hybridization using Sequence Specific Oligonucleotides (SSOs).
2. DNA amplification using Sequence Specific Primers (SSPs).
3. Sequencing Based Typing (SBT) protocols. The Primer/Probe database can be used as a reference resource for probe design and is intended to provide necessary tools for the exchange of DNA typing data based on the above technologies. This database is composed of a Primer/ Probe Interface and a Typing Kit Interface. Each interface has a series of interactive frames that allow the user to access different database functions.

Primer/Probe Interface

The Primer/Probe Interface provides information on individual reagents used for the typing of MHC or MHC-related loci. Users can access this information through the multiple functional frames within the interface. The words "primers" and "probes" represent SSOs, SSPs, SSP mixes, SSO mixes, and their combinations, which include nested PCR or SSO hybridization on group-specific amplifications. Nested PCR and group-specific amplification are both based on a primary amplification with an SSP mix.

Selection of Primers/Probes

The process of selecting a primer/probe begins by entering information in the upper frame of the Primer/Probe Interface page. Users enter search options for primers, probes, and mixes, which are grouped by type (SSO, SSP, SSO mix, and SSP mix), locus, and source (submitting institution). Primers/probes can also be selected by entering either the global or local name into the search field. Once this information is entered, the lower frame fills in with the Primer/Probe Listing. Primers/probes that are then selected by the user within the Primer/Probe Listing will be displayed in the large scroll box (in the upper frame) and can be downloaded in XML format.

Primer/Probe Listing

This function is found in the lower frame of the Primer/Probe Interface page. Results of a Primer/Probe selection are displayed here and will include a unique global ID, local name, source, locus, and type for each primer/probe result. The Global Name (global ID) of each result provides a link to the Primer/Probe View or Mix View functions, and the check box associated with each primer/probe/mix can be used to generate a list of items in the Primer/Probe Interface for subsequent download in XML format.

Box 2 contains information on global IDs.

Box 2. Global IDs.

Upon submission, dbMHC creates a unique global ID number for each primer/probe/mix/kit submitted. This global ID consists of three letters for the submitting institution and seven digits.

dbMHC uses the global ID number to store the entire history of a reagent. The global ID number will never change, but every time a user edits the sequence of a submitted primer/probe/mix/kit, that edit is given an incremental version number. Thus, all previous versions of a particular primer/probe/mix/kit are consequently accessible, and submitted primers/probes/mixes/kits can never be deleted.

A "local name" is the identifier given by submitters to their primers and probes.

Primer/Probe View

The Primer/Probe View is located in the upper frame of the Primer/Probe Interface and appears as a result of selecting a global ID. This view will display the entire set of data of the SSP and/or SSO primers and/or probes that the user had selected in the Primer/Probe Interface. The data displayed include:

- Local name
- Locus as specified by the submitter
- Global ID
- Date of last change ("last modified")
- Probe orientation as specified by the submitter
- Corresponding allele sequence (Allele Seq:)
- Reagent type
- Optional filter, i.e., pre-amplification
- Version number
- Probe sequence (Probe Seq:)
- The annealing position of the 3' end of the probe as specified by the submitter
- Probe stringency

Primer/Probe View offers links that will enable a user to edit primer/probe information, change primer/probe stringency, and allow users to list alleles detected by a probe with a sequence alignment. If a list of alleles detected by a primer/probe without sequence alignment is wanted, Primer/Probe View can do this as well. Select **Listed**, and the results will be displayed in the Primer/Probe Allele Reactivities Listing.

Primer/Probe Edit

The primer/probe edit function allows users to enter new primers/probes into dbMHC or allows users to edit existing primer/probe data. All but the following three fields can be edited (these fields are set by dbMHC):

- Global ID

- Version number
- Date of last change

Within the primer/probe edit frame, there is an alternative input field for primer/probe sequence called Allele Sequence. If primer or probe orientation is set to "reverse", the Allele Sequence field will reverse complement the allele sequence as displayed in the alignment to generate the appropriate primer/probe sequence.

Before submitting primer/probe data, a user should define the matching stringency of the primer or probe in the Annealing Stringency field. Once the user has selected the matching stringency, the probe can be submitted. Submission of a primer/probe triggers dbMHC to begin an allele reactivity calculation that is based on the primer/probe sequence and the selected matching stringency. The result of this calculation is a list of alleles that might be detected by the submitted primer/probe. The user will find this detectable allele list in the Reactive Allele Listing (lower frame).

Alignment, on the Primer/Probe View, opens a sequence alignment in the lower frame, which is the reactive allele alignment.

Reactive Allele Listing

The Reactive Allele Listing is located in the lower frame of the Primer/Probe Interface. It lists the alleles that might be detected by a selected primer/probe or mix (Boxes 3 and 4).

Box 3. Primer/probe reactivity.

Reactivity scores characterize the reactivity between a primer/probe and an allele for primers and probes and are listed below:

- Positive: probe anneals with allele.
- Weak: probe anneals sometimes with allele.
- Unclear: annealing cannot be predicted, no empirical information, allele has not been sequenced at the annealing position.
- Negative: probe does not anneal.

  dbMHC calculates primer/probe reactivity based on sequence similarity only and does not take into account laboratory experimental conditions such as magnesium concentration, temperature, etc.

Box 4. Submitter's reactivity scores vs. dbMHC reactivity scores.

A submitter's reactivity score can differ from the dbMHC's calculated reactivity score. If there is disparity between the submitter's score and the dbMHC score, users should regard

Box 4 continues on next page...

Box 4 continued from previous page.

the submitter's score as reliable. In cases of unexplainable disagreement, however, users should contact the submitting institution for further information.

Reactive Allele Alignment

Selecting **Aligned** from the Primer/Probe View results in a display of the reactive allele alignment in the lower frame of the Primer/Probe Interface. It displays the alleles that might be detected by a certain primer/probe or mix in alignment with that primer/probe sequence. By default, the first 20 alleles detected by the primer/probe or mix will be displayed. The frame for the reactive allele alignment also allows a submitter to set or edit the Allele Reactivity Score of a primer/probe in this frame. If a submitter chooses to not set a reactivity score, then dbMHC sets the submitter reactivity score to "not edited". dbMHC will also calculate a system reactivity score for the primer/probe based on the primer/probe sequence and matching stringency. Colored option boxes indicate dbMHC's reactivity score for the primer/probe, whereas dots within the option boxes indicate the submitter's reactivity score. If the submitter's score (see Box 5) is set and differs from dbMHC's score, the dbMHC score box changes to a warning color.

The alignment position of the 3' end of the probe is recorded for each allele. Both the sense and the reverse alignment positions are displayed if the alignment represents an SSP mix.

Only users with permission from an account administrator will be able to edit or add additional alleles to a Reactive Allele Listing for a particular primer/probe or mix. To edit the allele reactivity list, mark the "edit reactivity list" check box in the reactive allele alignment frame.

Box 5. Setting allele reactivity scores.

A submitter can set individual allele reactivity scores either one-by-one or in a batch. The allele reactivity score for all alleles or for individual alleles that are unedited can be set the same as dbMHC's proposed allele reactivity score.

If the user chooses to set the reactivity score as a batch, be aware that alleles within the batch that are not currently displayed will be scored as well. If the user makes a mistake in the scoring, **Reset** will reset the reactivity scores to the values present at the beginning of the session.

Submit stores the edited scores in the database. Once submitted, allele reactivity scores cannot be automatically reset to their prior value.

Box 5 continues on next page...

Box 5 continued from previous page.

Submitting allele reactivity scores will trigger a new dbMHC allele reactivity calculation. If the submitted primer/probe sequence is shorter than 10 nucleotides, dbMHC will use the score information of the alleles to extend the probe sequence.

Mix View

The Mix View function is located in the upper frame of the Primer/Probe Interface and can be accessed via the link on the global ID of the Primer/Probe Listing page.

It displays an entire set of mix data for selected SSO and SSP mixes:

- Local name
- Mix type
- Locus as specified by the submitter
- Optional filter, i.e., pre-amplification
- Global ID
- Version number
- Date of last change
- List of probes as mix elements
- Mix stringency

The Mix View function contains links to the Mix Element function, where users can change or add elements of a mix. Users will also find links to the Reactive Allele Listing, where they may list alleles detected by a selected mix or selected individual elements of a mix that do not have a sequence alignment. Finally, users will find links to the reactive allele alignment function, where they can view alleles with sequence alignments that are detected by a selected mix or selected individual elements of a mix.

Mix Edit

The Mix Edit function is located in the upper frame of the Primer/Probe Interface and allows users to enter new mixes or to edit existing mix data. Users can edit all but the following three fields (these fields are set by dbMHC):

- Global ID
- Version number
- Date of last change

Annealing stringency defines the cumulative matching stringency of the elements of the mix. For SSP mixes, both the sense and the reverse primer must react with the allele with at least the defined stringency.

Mix Element

The Mix Element function is located in the lower frame of the Primer/Probe Interface and allows users to add elements to a mix or edit existing elements of a mix. To use the Mix

Element function, the user must first specify the mix to be altered in the Mix Edit function and define a source for the primers/probes that the user wants listed. The mix element function displays only SSO probes for SSO mixes and displays SSP probes from SSP mixes in a sense column and a reverse orientation column. For mixes that contain probes from different sources, users must enter the mix elements separately.

Typing Kit Interface

The Typing Kit Interface is the gateway to information on individual typing kits used for typing MHC or MHC-related loci. Users can access this information through multiple functional frames within the interface. Typing kits contained within the Typing Kit Interface consist of SSOs, SSO mixes, or SSP mixes. Elements of typing kits may interact with unamplified DNA, pre-amplified DNA, or several distinct groups of pre-amplified DNA within one locus (e.g., two different amplifications of a certain exon using a distinct variation). The elements of each typing kit will react in characteristic patterns with individual alleles. These patterns can be used to determine allelic variants or a group of allelic variants within a locus (see Box 6).

Box 6. Versions of a typing kit.

All typing kits are identified by a global ID that is created by dbMHC, which will store the entire history of changes made to a typing kit. An incremental version number is given to every editing session of a typing kit; thus, even if some or all elements of a kit were deleted or altered by a user during an editing session, previous versions of the kit and kit elements will still be available.

Selection of the Typing Kit

Use the Typing Kit Interface to enter search parameters for typing kits. Typing kits are grouped by:

- Type: SSO, SSP
- Locus
- Source: the submitting institution

Typing kits can also be selected by typing either the global or local name into the search field.

Search results, based on user-selected parameters, will be displayed in the Typing Kit Listing, which appears in the lower frame of the Typing Kit Interface. Typing kits selected from the Typing Kit Listing will be displayed in the large scroll box (of the upper frame). They either can be used for combined pattern interpretation of multiple kits or downloaded in XML format.

Typing Kit Listing

The Typing Kit Listing is displayed in the lower frame of the Typing Kit Interface. On the basis of the criteria selected, kits are listed with their unique global ID, local name, source, locus, and type. The global ID for each kit provides a link to the Typing Kit View (upper frame of the Typing Kit Interface). The check boxes associated with each typing kit are used to generate a list of items in the Kit Select view for subsequent interpretation of a pattern or for download in XML format.

Typing Kit View

The Typing Kit View (upper frame of the Typing Kit Interface) lists the entire set of primer/probe data for selected SSO and SSP kits. Typing kit probe data include:

- Local name
- Kit type
- Locus (with or without group-specific pre-amplifications)
- Global ID
- Version number
- Batch

If **List Elements** from the Typing Kit View is selected, the Typing Kit Elements list provides links to the kit components. Users will also find links to the Edit Kit Locus Groups and Elements page, where they can edit kit information, or they can use **Save kit as...** to access the New Typing Kit function, where they can create a new kit based on a currently displayed one (see Box 7).

Box 7. Creating a virtual kit.

If a primer/probe within an existing typing kit malfunctions and a user created a modification of the primer/probe, that modified primer/probe is considered by dbMHC to be an entirely new probe.

The kit that contains this modified primer/probe is considered by dbMHC as an entirely new kit. Therefore, when entering a sequence modification of primer/probe within a kit, a submitter must create a new kit by using **Save kit as** located within New Typing Kit.

Save kit as will create a copy of the existing kit. The user can then rename the copy of the existing kit, virtually creating a new kit. The user can then go to the kit element frame, Edit Kit Locus Groups and Elements, to remove the old primer/probe from the new kit and replace it with the modified primer/probe.

Typing Kit Elements

Typing Kit Elements is displayed in the lower frame of the Typing Kit Interface and is used to display all of the elements (components) within a particular typing kit. This function will group the kit elements according to the kit locus groups and will display

them as an ordered list. The global ID of each kit element serves as a link to either the Mix View or the Primer/Probe View of this element.

Edit Kit Locus: Groups and Elements

The Typing Kit Interface **Locus:** function (upper frame) can be considered a switchboard for assembling a typing kit for dbMHC, although kit name, batch, and type must be edited using the New Typing Kit function, accessed from **New Kit** on the Typing Kit Interface. The kit's global ID number, version number, and the date of last change are set by dbMHC. Users wanting to edit elements of a particular kit must select a particular locus first and then select **Add locus group** or **Edit locus group**, which will open the Edit Locus Group Elements page in the lower frame of the browser (see Box 8).

Box 8. Typing kit locus "groups".

HLA typing kits are usually used to detect alleles in one or more loci. Within a typing kit, a particular locus and an optional pre-amplification define what is termed a "group". Thus, one typing kit can contain several groups, with each group either consisting of the same locus and a different pre-amplification (or no pre-amplification) or consisting of different loci (with or without pre-amplification).

Edit Locus Group Elements

The kit group function is located in the lower frame of the Typing Kit Interface and allows a user to add single elements to, or remove single elements from, a kit group. The kit group frame displays the elements for a particular locus selected by the user in the **kit locus** function. If the typing kit of interest is an SSO kit, only SSO or SSO mixes will be displayed. If the typing kit is an SSP kit, only SSP mixes will be displayed. Users can then add a primer/probe to the displayed group elements by clicking on a particular primer/probe in the left column and can remove a primer/probe from the group by selecting that element and clicking **Remove**.

New Typing Kit

Access to the kit edit function, **New Kit**, is found in the upper frame of the Typing Kit Interface and allows a user to enter or modify the name, batch, and type of typing kit. Users may change the kit type so long as the kit does not yet contain any element or group.

Typing Kit Interpretation

Access to the typing kit interpretation function, **Interpret**, is found on the Typing Kit Interface. It provides an online typing pattern interpretation tool. Typing kit reactivity patterns can be analyzed one at a time. Several typing kits that are used to type one sample can be analyzed in combination. An order number represents each element of a typing kit. This number provides a link to the probe reactivity alignment view of each element. The

display also indicates whether, and for which elements, an allele-specific amplification has been used. Reactivity patterns can be entered either as a string or via the graphical display. Both the graphical display and the text field will update each other. The string entry accepts "1", "+" for positive reactivity, "0","-" for negative reactivity, "n" for not tested, "w" for weak reactivity, and "?" for undetermined. The graphic entry symbols are green for positive reactivity, orange for weak reactivity, yellow for undetermined, white for negative reactivity, and gray for "not tested". Users can preset the main reactivity by selecting one reactivity option and clicking on **Set All**. If the option **cycle** has been selected, repeated clicking on one reactivity field of a kit will cycle through all possible reactivities.

The reactivity string or pattern entered will be interpreted as a heterozygote allele combination. Multiple kits can be combined. Users can set the degree of tolerance, which limits the number of false-positive or false-negative reactivities per locus. Each locus is analyzed separately. Allele assignments for each locus are listed according to the number of false-positive/false-negative reactivities.

Typing Kit Pattern

Selecting **Pattern** from the Typing Kit Interface results in a display (in the lower frame) of a cross-tab view of allele reactivities of an individual typing kit. Alleles are listed in rows, and kit elements are listed in columns. Each element of a typing kit is represented by the respective order number. This number provides a link to the probe reactivity alignment view of each element. The display also indicates whether and for which elements an allele-specific amplification has been used.

A "+" with a green background signals a detection of an allele by a kit element, a "w" with an orange background signals a weak detection, a "?" with a yellow background signals a lack of information, and an "r" with a red background signals a rejected interaction, although originally suggested by the prediction algorithm. If a kit is designed to detect only a subset of alleles, the display will be limited to this subset (see Box 9).

Box 9. Some dbMHC/browser limitations.

- Because of operating system limitations, the Alignment Viewer can only display up to 300 nucleotides in one line if the browser is Internet Explorer, whereas Netscape is not restricted by this limitation.
- Several essential parts of dbMHC are based on Javascript interaction and dynamic text generation within a page. Users must be aware that many browsers are unable to properly interpret and display Javascript-generated text.
- Netscape version 4.76 does not check for browser content size changes; therefore, users must manually resize to trigger the correct size recognition. Users may resize contents by using a "post" command, which will lead to a new download of the initial request instead of simply resizing the window.

Box 9 continues on next page...

Box 9 continued from previous page.

- Netscape 4.7 may sometimes cause fatal errors and does not allow users to copy and paste sequences from the alignment to the probe sequence field in the probe edit function.
- Internet Explorer version 5.5 and Netscape version 6.2 will correctly interface with dbMHC.

Database Content

The data submitted to dbMHC are stored in a Microsoft SQL (MSSQL) relational database. Table 1 is a data dictionary that defines dbMHC's database tables or record sets. For each dbMHC table, the data dictionary provides the table name, a column name, the data type in a particular table, and a summary comment.

The relationships between dbMHC database tables are depicted in Figure 1.

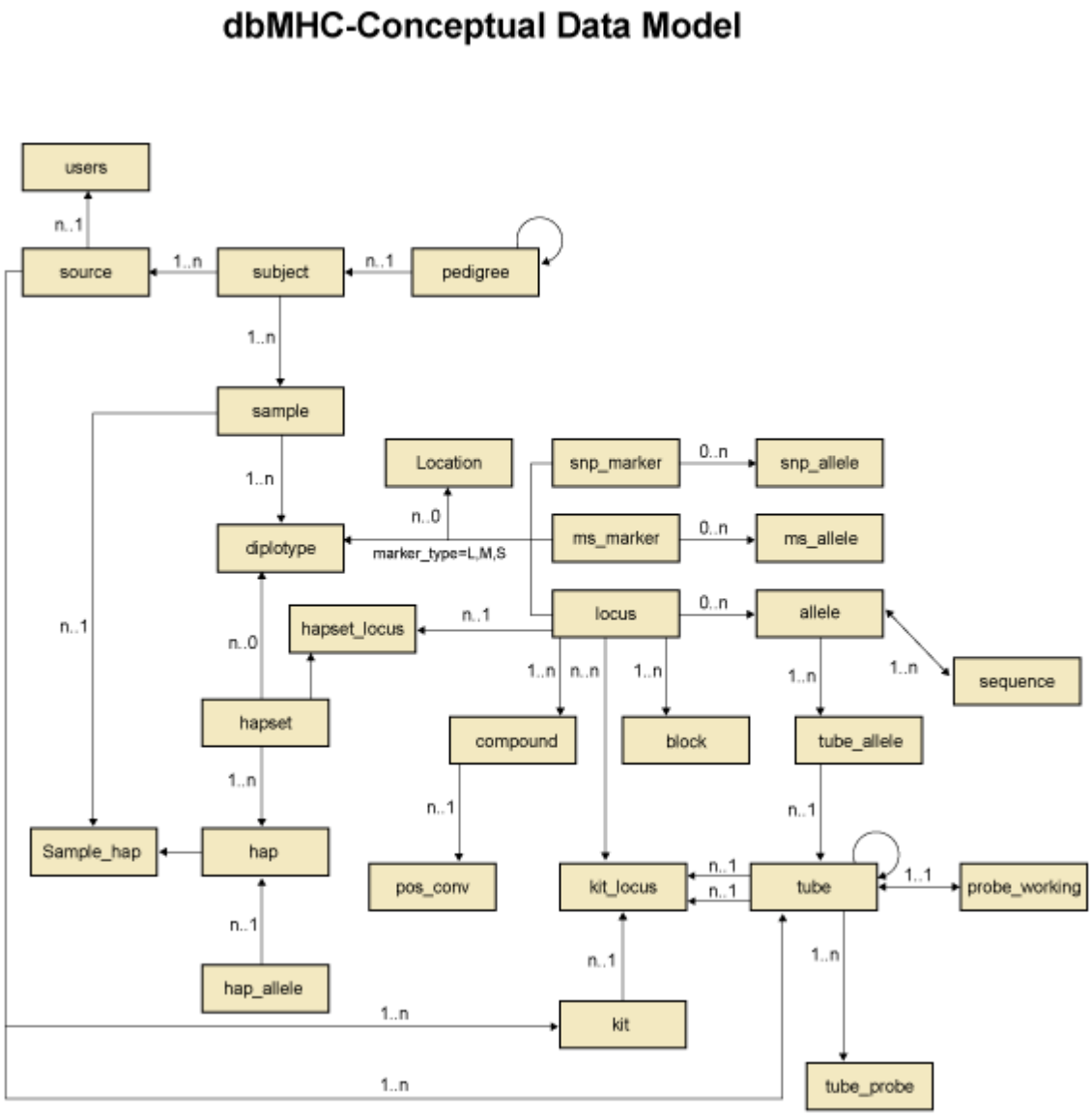


Figure 1. **Physical model of the relationships among dbMHC database tables.**

Table 1. **Data dictionary.**

Table name	Column name	Data type	Column comment
allele	allele_nr	int	Unique identifier allele instance.
	allele_id	int	Identifier for each allele.
	allele_name	varchar(30)	Full allele name defined by IHWG.
	allele_short	varchar(30)	Common use allele name.
	allele_group	int	Group associated with allele.

Table 1 continues on next page...

Table 1 continued from previous page.

Table name	Column name	Data type	Column comment
	compound_nr	int	Number identifying to which compound allele is a member.
	locus_id	int	Number identifying to which locus allele is a member.
	active	tinyint	Status labeling current active allele for this allele_id.
	db_version_nr	int	IHWG database version to which this allele belongs.
	user_id	int	Identifier or user who created/updated allele.
	user_date	datetime	Date when record was created/updated.
block	block_nr	int	Unique identifier for block instance.
	block_id	int	Identifier for each block.
	block_ord	int	Sequence order for each block.
	block_type	varchar(1)	Defines block type.
	locus_id	int	Identifier for locus for which block is a member.
	ref_pos	int	Start reference position.
	ref_length	int	Reference length.
	working_pos	int	Start working position.
	working_length	int	Working length.
	block_name	varchar(20)	Common name.
	active	tinyint	Status of row.
	db_version_nr	int	IHWG database version to which this block belongs.
compound	locus_id	int	Locus identifier.
	compound_nr	int	Compound number.
job_trans	trans_nr	int	Unique identifier for transaction.
	app_name	varchar(30)	Application name that processes transaction.
	app_arg	varchar(30)	Arguments to processing application.
	create_date	datetime	Time created.
	start_date	datetime	Time application started processing transaction.
	end_date	datetime	Time application completed processing transaction.
	status	int	Current status of transaction.
	priority	int	Defines priority level.
	message	varchar(255)	Any message that the processing application wants to record.

Table 1 continues on next page...

Table 1 continued from previous page.

Table name	Column name	Data type	Column comment
kit	kit_nr	int	Unique identifier for kit instance.
	kit_id	int	Kit identifier.
	kit_local_id	int	Submitter identifier.
	version_nr	int	Version number of kit.
	kit_name	varchar(30)	Submitter name.
	kit_global_id	varchar(30)	NCBI-defined name.
	kit_batch	varchar(30)	Submitter batch.
	type	int	Type.
	active	tinyint	Status labeling current active kit for this kit_id.
	source_id	int	Source identifier.
	user_id	int	User identifier.
	user_date	datetime	Date created/updated.
kit_locus	kit_locus_nr	int	Unique identifier for kit locus instance.
	kit_nr	int	
	kit_id	int	Kit identifier.
	locus_id	int	Locus identifier.
	locus_order_nr	int	Order for loci to be used.
	filter_id	int	Filter identifier.
	tube_id	int	Tube identifier.
	tube_order_nr	int	Order for tube to be used.
	active	tinyint	Status labeling current active kit locus for this kit id.
	user_id	int	User identifier.
	user_date	datetime	Date created/updated.
locus	locus_id	int	Identifier for locus.
	locus_NCBI_id	varchar(30)	NCBI locus identifier.
	locus_name	varchar(30)	Common name.
	display_id	int	Identifier of allele that should be used as the display reference.
	ref_id	int	Identifier of reference allele.
	locus_MIM_id	int	MIM locus identifier.
	locus_pub_id	int	PubMed locus identifier.
	display_order	int	Order number for display.

Table 1 continues on next page...

Table 1 continued from previous page.

Table name	Column name	Data type	Column comment
mhc_snp	locus_id	int	Identifier for locus.
	ref_pos	int	Reference position.
	ref_offset	int	Reference offset.
	subsnp_id	int	dbSNP's ss number.
	snp_length	int	SNP's length.
pos_conv	compound_nr	int	Compound number.
	working_pos	int	Working position.
	ref_pos	int	Reference position.
	ref_offset	int	Reference offset.
	blast_pos	int	BLAST position.
probe_working	probe_working_nr	int	Unique identifier for each row.
	tube_nr	int	Unique identifier for tube instance.
	probe_id	int	Tube (probe) identifier.
	user_id	int	User identifier.
	user_date	datetime	Date created/updated.
	active	tinyint	Active status.
	sequence	varchar(255)	Working sequence.
sequence	seq_type	int	Sequence type.
	seq_nr	int	Sequence number.
	seq_ord	int	Sequence order.
	sequence	varchar(255)	Character sequence.
session	cur_session_nr	int	Current session number.
	next_session_nr	int	Next avaliable session number.
	user_id	int	User identifier.
	create_date	datetime	Date created/updated.
source	source_nr	int	Unique identifier source instance.
	source_id	int	Source identifier.
	source_code	varchar(3)	Source code for use in naming tubes and kits.
	institution	varchar(50)	Source name.
	admin_id	int	User identifier for administrator for source.

Table 1 continues on next page...

Table 1 continued from previous page.

Table name	Column name	Data type	Column comment
	address	varchar(50)	Address.
	city	varchar(50)	City.
	state	varchar(50)	State.
	postal_code	varchar(20)	Postal code.
	country	varchar(50)	Country.
	phone	varchar(30)	Phone number.
	fax	varchar(30)	Fax number.
	email	varchar(50)	Email address.
	active	tinyint	Status labeling current active source for this source id.
	user_id	int	User identifier.
	user_date	datetime	Date created/updated.
	options	int	Not used.
	info	varchar(255)	Comments/general information.
tube	tube_nr	int	Unique identifier for tube instance.
	tube_id	int	Tube identifier.
	tube_name	varchar(30)	Submitter name.
	tube_global_id	varchar(30)	NCBI-defined name.
	filter_id	int	Filter identifier.
	source_id	int	Source identifier.
	source_date	datetime	Not used.
	type	int	Type.
	sense	tinyint	Orientation.
	locus_id	int	Locus identifier.
	ref_pos	int	Submitted reference position.
	ref_offset	int	Submitted reference offset.
	stringency	int	Stringency set for annealing results.
	active	tinyint	Status labeling current active tube for this tube_id.
	user_id	int	User identifier.
	user_date	datetime	Date created/updated.
	recalc_date	datetime	Not used.
	db_version	int	IHWG database version to which this tube belongs.
	info	varchar(255)	Comments/general information.

Table 1 continues on next page...

Table 1 continued from previous page.

Table name	Column name	Data type	Column comment
	sequence	varchar(255)	Submitted sequence.
	source_nr	int	
tube_allele	tube_allele_nr	int	Unique identifier for tube allele instance.
	tube_id	int	Tube identifier.
	locus_id	int	Locus identifier.
	allele_id	int	Allele identifier.
	user_status	int	User status.
	system_status	int	System status.
	ref_pos	int	Reference position.
	ref_offset	int	Reference offset.
	block_nr	int	Not used.
	score	real	Annealing score.
	working_pos	int	Working position.
	for_tube_id	int	Forward tube identifier.
	for_working_pos	int	Forward working position.
	rev_tube_id	int	Reverse tube identifier.
	rev_working_pos	int	Reverse working position.
	sense	tinyint	Orientation.
	seq_known	int	Sequence known (sequence fully defined).
	active	tinyint	Status labeling current active tube allele for this tube_id allele_id combination.
	user_id	int	User identifier.
	user_date	datetime	Date created/updated.
tube_probe	tube_probe_nr	int	Unique identifier for tube probe instance.
	tube_nr	int	Unique identifier for tube instance.
	probe_id	int	Tube identifier for sub tube in mix.
	tube_id	int	Tube identifier.
	active	tinyint	Status labeling current active tube probe for this tube_id probe_id combination.
users	user_id	int	User identifier.
	user_date	datetime	Date created/updated.
	user_nr	int	Unique identifier for user instance.

Table 1 continues on next page...

Table 1 continued from previous page.

Table name	Column name	Data type	Column comment
	user_id	int	User identifier.
	first_name	varchar(30)	First name.
	last_name	varchar(30)	Last name.
	login	varchar(30)	Login.
	pwd	varchar(30)	Password.
	source_id	int	Source identifier.
	phone	varchar(30)	Phone number.
	fax	varchar(30)	Fax number.
	email	varchar(50)	Email address.
	create_probe	tinyint	Permission to create probe.
	create_kit	tinyint	Permission to create kit.
	modify_probe	tinyint	Permission to modify probe.
	modify_react	tinyint	Permission to modify reactivity.
	modify_kit	tinyint	Permission to modify kit.
	info	varchar(255)	Comments/general information.
	active	tinyint	Status labeling current active user for this user_id.
	createdby_id	int	Created by identifier (another user_id).
	created_date	datetime	Date created/updated.
	source_admin	tinyint	Is user_id an administrator for source?

Primer/Probe Database

dbMHC's primer/probe database or interface is not curated. As such, the accuracy of the information presented in this database is dependent entirely upon the accuracy of the data submitted to dbMHC.

We suggest that each primer or probe submitted should be characterized by its complete sequence. In cases where submission of the complete sequence is impossible because the sequence information is considered proprietary, dbMHC offers the option to submit partial sequences. Submitted primer/probe specifications should comply with American Society of Histocompatibility and Immunogenetics (ASHI) (20) and European Federation of Immunogenetics (EFI) (21) standards for primers/probes used in histocompatibility DNA testing. If the submitted sequence contains fewer than 10 nucleotides, the submitter must also provide the position of the 3' end of the probe within the alignment of a locus. This additional information is necessary because the likelihood of a primer/probe reactivity calculation producing multiple erroneous alignment positions becomes too great if dbMHC calculates it with fewer than 10 nucleotides.

Reagent Allele Reactivity Prediction

The allele reactivity prediction tool for primers and probes available through dbMHC presents calculations based solely on sequence similarity and can therefore offer only suggestions for the possible allele reactivities of individual primers and probes. Users should not consider this tool to be 100% accurate. A reliable prediction of primer/probe allele reactivities requires complete sequence information, as well as reaction data that include annealing temperature and magnesium concentration. Because these data are not consistently available to dbMHC, we are unable to offer more than a suggestion for primer/probe allele reactivity.

Computation of Primer/Probe–Allele Interaction Stringency

dbMHC uses sequence comparisons to create a sequence interaction match or stringency grade that is compiled within a penalizing system. dbMHC's starting point for grading the interaction between a primer/probe and an allele is 100%, with each difference in sequence between the primer/probe and the allele causing a reduction in the remaining match grade by a certain percentage. The primary factors dbMHC uses to compute stringency grading include:

- Nucleotide differences
- Nucleotide position and primer/probe type

Nucleotide Differences

dbMHC divides nucleotide interactions into five categories: perfect match, high match, medium match, poor match, and no match. dbMHC defines these categories by using the purine and pyrimidine interaction between the allele and the primer/probe, as well as the number of hydrogen bonds affected during the virtual pairing of the allele with the primer/probe. See Table 2.

Table 2. Penalties for nucleotide mismatches.

Probe	DNA template			
	A	T	G	C
A	0.56	0	0.42	0.69
T	0	0.54	0.42	0.65
G	0.67	0.63	0.56	0
C	0.9	0.94	0	1

Penalties calculated according to Peyret et al. (22).

Nucleotide Position and Primer/Probe Type

dbMHC uses a penalty system based on nucleotide position in its calculation of SSO/SSP interactions with different alleles. The system dbMHC uses indicates the extent to which an individual nucleotide position will affect the interaction stringency grade in a worst-

case scenario, where guanines are in opposing sequence positions or cytosines are in opposing sequence positions. The maximum penalty given in the system is in the middle of an SSO and at the 3' end of an SSP, as shown in Tables 3 and 4.

If the initial interaction stringency grade for a particular primer/probe–allele interaction is 100% and a mismatch occurs in a position that carries a penalty of 100%, then dbMHC will reduce the interaction stringency grade to 0%. If the mismatch occurs in a position that carries a penalty of 95%, dbMHC will reduce the interaction stringency grade to 5%. A more detailed example of how dbMHC computes primer/probe interaction stringency is available online.

dbMHC's reactive allele alignment algorithm searches for sequences with a match grade above the stringency level set by the submitter of each probe. Currently, the search algorithm searches within all loci that are part of dbMHC. Within each locus, the algorithm constructs compound sequences that represent the combined polymorphic positions of alleles of the same length. Alleles with insertions or deletions are handled separately. If a primer or probe matches with a certain position within the compound, all contributing alleles are checked for that position, whether or not they match.

The reactive allele alignment algorithm will check within a specified locus for sequences that have a match grade in accordance with primer/probe specifications at a user-indicated position. The algorithm then extends SSPs toward the 5' end of the probe, to a maximum of 10 nucleotides, and extends both sides of SSOs a maximum of 15 nucleotides, observing all polymorphic positions in matching alleles. The resulting probe extension is then used by the alignment algorithm to check for cross-reactivities within other loci.

If the submitted primer/probe sequence is shorter than 10 nucleotides and the user submits a score that accepts certain alleles and rejects others, the reactive allele alignment algorithm will use this information to refine the probe extension. If the probe submitter rejects all alleles containing a unique sequence motif in the vicinity of the probe sequence, the algorithm will generate the extended probe sequence such that it does not match the unique sequence motif.

Table 3. Position penalties for SSP nucleotide mismatches.

5'	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	3'
	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.07	0.12	0.17	0.21	0.25	0.30	0.50	0.90	
Primer	T	T	T	C	T	T	C	A	C	C	T	C	C	G	T	G	T	C	
DNA template	T	T	T	C	T	T	C	A	C	A	T	C	C	G	T	G	T	C	

Match score calculation for SSPs.
An 18-mer primer anneals to a mismatched DNA. Primer and template are shown in the sense orientation. The substitution of C with A leads to the mismatch C-T with a penalty of 0.94 (refer to Table 2).
This position has a 6% influence on the overall probe reactivity.
The overall probe score is $1 - (0.06 \times 0.94) = 0.94$.

Table 4. Position penalties for SSO nucleotide mismatches.

5'	10	9	8	7	6	5	4	3	2	1	1	2	3	4	5	6	7	8	9	10	3'
	0.22	0.22	0.34	0.34	0.70	0.70	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.70	0.70	0.34	0.34	0.22	0.22	
Probe	T	T	T	C	T	T	C	A	C	A	T	C	C	G	T	G	T	C	C	C	
DNA template	T	T	T	C	T	C	C	A	C	A	T	C	C	G	T	G	T	C	C	C	

Match score calculation for SSO probes.
A 20-mer SSO anneals to a mismatched DNA. Both probe and template are shown in the sense orientation. The substitution of T with C leads to the probe-template mismatch T-G with a penalty of 0.42 (refer to Table 2).
This position has a 70% influence on the overall probe reactivity.
The overall probe score is $1 - (0.7 \times 0.42) = 0.71$.

Integration with Other Resources

NCBI Links

The dbMHC provides a set of internal links to other NCBI resource homepages through either the black “quick link” bar located at the top of the dbMHC homepage or through the rotating HLA molecule located at the top left-hand corner of the dbMHC homepage. These links will take users to the following NCBI resources:

- PubMed
- Nucleotide
- Protein
- OMIM
- BLAST
- Molecular Modeling database (MMdb)

The dbMHC Alignment Viewer provides locus-level linkage to NCBI's LocusLink and to dbSNP at the individual SNP level.

Graphic View

The Graphic View page is a hyperlink-enabled representation of the MHC region on chromosome 6. Locus-level links from the Graphic View page include:

- dbSNP (at the SNP haplotype level)
- Map View
- LocusLink
- OMIM
- Nucleotide
- Protein
- PubMed
- Structure
- Books

The header and link column sections of dbMHC's Graphic View are the same as those on the main page. The content section of this page contains a selection box, called Choose Linked Resource, and three horizontally arranged sections, called Chromosome 6, HLA Class I, and HLA Class II. Genes listed in the HLA Class II section act as hyperlinks to the Web resource selected in Choose Linked Resource.

External Links

Currently, dbMHC provides links (from the left sidebar) to the following external sites:

The MHC haplotype project.

The International Histocompatibility Working Group (IHWG). Links to individual IHWG projects are listed in the text portion of the homepage.

The IMmunoGeneTics (IMGT)/HLA database. Links to the IMGT/HLA database are also located on the allele selection page. Users can access the allele selection page by selecting **Alleles** on the dbMHC Alignment Viewer. When **Alleles** is selected, each of the listed allele names on the allele selection page links directly to the IMGT/HLA allele-specific information page at the gene–allele level.

References

1. Bodmer J G, Marsh S G E, Albert E D, Bodmer W F, Bontrop R E, Dupont B, Erlich H A, Hansen J A, Mach B, Mayr W R, Parham P, Petersdorf E W, Sasazuki T, Schreuder G M T, Strominger J L, Svejgaard A, Terasaki P I. Nomenclature for factors of the HLA system, 1998. Tissue Antigens. 1999;53(4):407–446. PubMed PMID: 10321590.
2. Robinson J, Bodmer J G, Malik A, Marsh S G E. Development of the International Immunogenetics HLA Database. Hum Immunol. 1998;59(Suppl):1–17.
3. Robinson J, Marsh S G E, Bodmer J G. Eur J Immunogenet. 1999;26:75.
4. Robinson J, Bodmer J G, Marsh S G E. The American Society for Histocompatibility and Immunogenetics 25th annual meeting. New Orleans, Louisiana, USA. October 20-24, 1999. Hum Immunol. 1999;60:S1. PubMed PMID: 10549324.
5. Histocompatibility testing: report of a conference and workshop. Washington, DC: National Academy of Sciences - National Research Council, 1965.
6. Histocompatibility testing 1965. Copenhagen: Munksgaard, 1965.
7. Curtoni ES, Mattiuz PL, Tosi RM, eds. Histocompatibility testing 1967. Copenhagen: Munksgaard, 1967.
8. Terasaki PI, ed. Histocompatibility testing 1970. Copenhagen: Munksgaard, 1970.
9. Dausset J, Colombani J, eds. Histocompatibility testing 1972. Copenhagen: Munksgaard, 1973.
10. Kissmeyer-Nielsen F, ed. Histocompatibility testing 1975. Copenhagen: Munksgaard, 1975.
11. Bodmer WF, Batchelor JR, Bodmer JG, Festenstein H, Morris PJ, eds. Histocompatibility testing 1977. Copenhagen: Munksgaard, 1978.

12. Terasaki PI, ed. Histocompatibility testing 1980. Los Angeles: UCLA Tissue Typing Laboratory, 1980.
13. Albert ED, Baur MP, Mayr WR, eds. Histocompatibility testing 1984. Heidelberg: Springer-Verlag, 1984.
14. Dupont B, ed. Immunobiology of HLA. Vol. I. Histocompatibility testing 1987, and Vol. II. Immunogenetics and histocompatibility. New York: Springer-Verlag, 1988.
15. Tsuji K, Aizawa M, Sasazuki T, eds. HLA 1991. New York: Oxford University Press, 1992.
16. Charron, D, ed. Genetic diversity of HLA. Functional and medical implications. Paris: EDK Publishers, 1996.
17. Foissac A, Salhi M, Cambon-Thomsen A. Microsatellites in the HLA region: 1999 update. Tissue Antigens. 2000;55(6):477–509. PubMed PMID: 10902606.
18. Foissac A, Cambon-Thomsen A. Microsatellites in the HLA region: 1998 update. Tissue Antigens. 1998;52(4):318–352. PubMed PMID: 9820597.
19. Foissac A, Crouau-Roy B, Faure S, Thomsen M, Cambon-Thomsen A. Microsatellites in the HLA region: an overview. Tissue Antigens. 1997;49(3 Pt 1):197–214. PubMed PMID: 9098926.
20. ASHI Governance. Standards for histocompatibility testing, 1998, P2.151.
21. European Federation for Immunogenetics. Modifications of the EFI standards for histocompatibility testing, 1996, P2.3100.
22. Peyret N, Seneviratne P A, Allawi H T, SantaLucia J Jr. Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A. C.C, G.G, and T.T mismatches. Biochemistry. 1999;38:3468–3477. PubMed PMID: 10090733.

Part 2. Data Flow and Processing

Chapter 12. Sequin: A Sequence Submission and Editing Tool

Jonathan Kans

Created: October 9, 2002; Updated: August 13, 2003.

Summary

Sequin is a stand-alone sequence record editor, designed for preparing new sequences for submission to GenBank and for editing existing records. Sequin runs on the most popular computer platforms found in biology laboratories, including PC, Macintosh, UNIX, and Linux. It can handle a wide range of sequence lengths and complexities, including entire chromosomes and large datasets from population or phylogenetic studies. Sequin is also used within NCBI by the GenBank and Reference Sequence indexers for routine processing of records before their release.

Sequin has a modular construction, which simplifies its use, design, and implementation. Sequin relies on many components of the NCBI Toolkit and thus acts as a quality assurance that these functions are working properly.

Detailed information on how to use Sequin to submit records to GenBank or edit sequence records can be found in the Sequin Quick Guide. Although this chapter will make frequent reference to that help document, the focus will be mostly on the underlying concepts and software components upon which Sequin is built.

Sequin: A Brief Overview

As input, Sequin takes a biological sequence(s) from a scientist wanting to submit or edit sequence data. The sequence (or set of sequences) can be new information that has not yet been assigned a GenBank Accession number, or it can be an existing GenBank sequence record. If Sequin is being used to submit a sequence(s) to GenBank, then the scientist is prompted to include his/her contact information, information about other authors, and the sequence, at the start of the submission process. Once all the necessary information has been entered, it is then possible to view the sequence in a variety of displays and edit it using Sequin's suite of editing tools.

Sequin is designed for use by people with different levels of expertise. Thus, it has several built-in functions that can, for example, ensure that a new user submits a valid sequence record to GenBank, or it can be prompted to automatically generate a sequence definition line. At the other end of the scale, for computer-literate users, Sequin can be customized by the addition of more (perhaps research-specific) analysis functions. Furthermore, there are some extremely powerful functions built into Sequin that are only available to NCBI Indexing staff. These are switched off by default in the public download version of Sequin because they include the ability to make the kinds of changes to a sequence record that

can also completely destroy it, if handled incorrectly. These various built-in Sequin functions are discussed further below.

Sequin's versatility is based on its design: (*a*) Sequin holds the sequence(s) being manipulated in memory, in a structured format that allows a rapid response to the commands initiated by the person who is using Sequin; and (*b*) it makes use of many standard functions found in the NCBI Toolkit for both basic data manipulations and as components of Sequin-specific tasks. In particular, Sequin makes heavy use of the Toolkit's object manager, a "behind the scenes" support system that keeps track of Sequin's internal data structures and the relationships of each piece of information to others. This allows many of Sequin's functions to operate independently of each other, making data manipulation much faster and making the program easier to maintain.

Sequence Submission

Sequin is used to edit and submit sequences to GenBank and handles a wide range of sequence lengths and complexities. After downloading and installing Sequin, a scientist wanting to submit or edit a sequence(s) is led through a series of forms to input information about the sequence to be submitted. The forms are "smart", and different forms will appear, customized to the type of submission. Detailed information on how to fill in these forms can be found within the **Help** feature of the Sequin application itself and in the online Help documents.

Sequin expects sequence data in FASTA formatted files, which should be prepared as plain text before uploading them into Sequin. Population, phylogenetic, and mutation studies can also be entered in PHYLIP, NEXUS, MACAW, or FASTA+GAP formats.

A sequence in FASTA format consists of a definition line, which starts with a ">", and the sequence itself, which starts on a new line directly below the definition line. The definition line should contain the name or identifier of the sequence but may also include other useful information. In the case of nucleotides, the name of the source organism and strain should be included; for proteins, it is useful to include the gene and protein names. Given all this information, Sequin can automatically assemble a record suitable for inclusion in GenBank (see below). Detailed information on how to prepare FASTA files for Sequin can be found in the Quick Guide.

Single Sequences

For single nucleotide sequence submissions to GenBank, the submitter supplies Sequin with the nucleotide sequence and any translated protein sequence(s). For example, a submission consisting of a nucleotide from mouse strain BALB/c that contains the β-hemoglobin gene, encoding the adult major chain β-hemoglobin protein, would have two sequences with the following definition lines, where "BALB23g" and "BALB23p" are nucleotide and protein IDs provided by the submitter:

```
> BALB23g [organism=Mus musculus] [strain=BALB/c]
> BALB23p [gene=Hbb-b1] [protein=hemoglobin, beta adult major chain]
```

The organism name is essential to make a legal GenBank flatfile. It can be included in the definition line as shown above, for the convenience of the submitter, or one of the Sequin submission forms will prompt for its clarity.

Although it is not necessary to include a protein translation with the nucleotide submission, scientists are strongly encouraged to do so because this, along with the source organism information, enables Sequin to automatically calculate the coding region (CDS) on the nucleotide being submitted. Furthermore, with gene and protein names properly annotated, the record becomes informative to other scientists who may retrieve it through a BLAST or Entrez search (see also Chapter 15 and Chapter 16).

Segmented Nucleotide Sets

A segmented nucleotide entry is a set of non-contiguous sequences that has a defined order and orientation. For example, a genomic DNA segmented set could include encoding exons along with fragments of their flanking introns. An example of an mRNA segmented pair of records would be the 5′ and 3′ ends of an mRNA where the middle region has not been sequenced. To import nucleotides in a segmented set, each individual sequence must be in FASTA format with an appropriate definition line, and all sequences should be in the same file.

High-Throughput Genomic Sequences

Genome Sequencing Centers use automated sequencing machines to rapidly produce large quantities of "unfinished" DNA sequence, called high-throughput genomic sequence (HTGS). These sequences are not usually annotated with any features, such as coding regions, at all, and in the initial phases are not of high ("finished") quality.

The sequencing machines produce intensity traces for the four fluorescent dyes that correspond to the four bases adenine, cytosine, guanine, and thymine. Software such as PHRED and PHRAP convert these raw traces into the sequence letters A, C, G, or T. PHRED is a base-calling program that "reads" the sequences of the DNA fragments and produces a quality score. With multiple overlapping reads to work on, PHRAP assembles the DNA fragments using the quality scores of PHRED, itself producing a quality score for each base. The resulting file, which PHRAP outputs in ".ace" format, consists of the sequence itself plus the associated quality scores. Sequin can use these files as input and assemble valid GenBank records from them. Further information on using Sequin to prepare a HTGS record can be found here.

Feature Tables

Some Genome Centers now analyze their sequences and record the base positions of a number of sequence features such as the gene, mRNA, or coding regions. Sequin can

capture this information and include it in a GenBank submission as long as it is formatted correctly in a feature table. Sequin can read a simple, five-column, tab-delimited file in which the first and second columns are the start and stop locations of the feature, respectively, the third column is the type of feature (the feature key—gene, mRNA, CDS, etc.), the fourth column is the qualifier name (e.g., "product"), and the fifth the qualifier value (e.g., the name of the protein or gene). The features for an entire bacterial genome can be read in seconds using this format. A sample features table is shown below.

```
>Feature sde3g
240     4048    gene
                        gene        SDE3
240     1361    mRNA
1450    1614
1730    3184
3275    4048
                        product     RNA helicase SDE3
579     1361    CDS
1450    1614
1730    3184
3275    3880
                        product     RNA helicase SDE3
```

Alignments

Population, phylogenetic, and mutation studies all involve the alignment of a number of sequences with each other so that regions of sequence similarity are emphasized. Sometimes it is necessary to introduce gaps into the sequences to give the best alignment. Sequin reads several output formats from sequence and phylogenetic analysis programs, including PHYLIP, NEXUS, PAUP, or FASTA+GAP.

The submitted sequence alignment represents the relationship between sequences. This inferred relationship allows Sequin to propagate features annotated on one sequence to the equivalent positions on the remaining sequences in the alignment. Feature propagation is one of the many editing functions possible in Sequin. Using this tool significantly reduces the time required to annotate an alignment submission.

Automated Submission

tbl2asn is a program that automates the submission of sequence records to GenBank. It uses many of the same functions as Sequin but is driven entirely by data files, and records need no additional manual editing before submission. Entire genomes, consisting of many chromosomes with feature annotation, can be processed in seconds using this method.

Packaging the Submissions

Sequences given to Sequin in the input data formats described in this chapter are retained within Sequin memory, allowing them to be manipulated in real time. For example, for

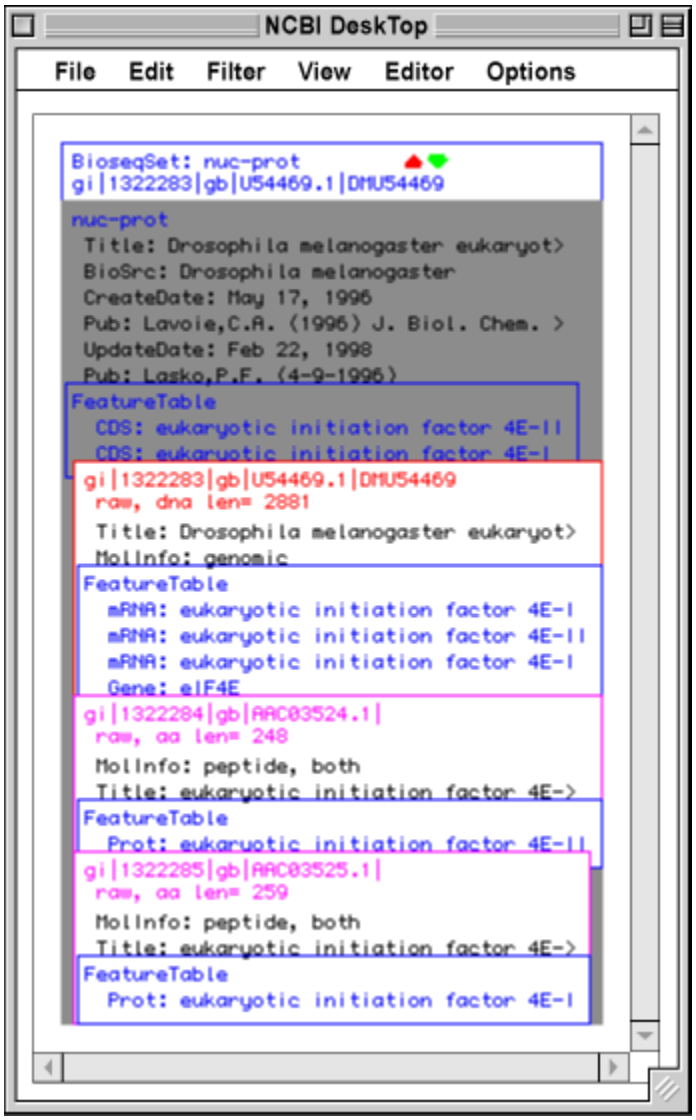


Figure 1. The internal structure of a sequence record in Sequin, as seen in the Desktop window. The display can be understood as a Venn diagram. Selecting the up or down arrow expands or contracts, respectively, the level of detail shown. In a typical submission of a protein-coding gene, a BioseqSet (of class "nuc-prot") contains two Bioseqs, one for the nucleotide and one for the protein. Descriptors, such as BioSrc, can be packaged on the set and thus apply to all Bioseqs within the set. Features allow annotation on specific regions of a sequence. For example, the CDS location provides instructions to translate the DNA sequence into the protein product.

submission to GenBank, the sequence is transformed from the Sequin internal structure to Abstract Syntax Notation 1 (ASN.1), the data description language in which GenBank records are stored. This is the format transmitted over the Internet when submitting to GenBank. Sequin can also output information in other formats, such as GenBank flatfile or XML, for saving to a local file.

Most sequence submissions are packaged into a BioseqSet, which contains one or more sequences (Bioseqs), along with supporting information that has been included by the submitter, such as source organism, type of molecule, sequence length, and so on (Figure 1). There are different classes of BioseqSets; thus, a simple single nucleotide submission is called a nuc-prot set (a BioseqSet of class nuc-prot) containing the nucleotide and protein Bioseqs. Similarly, population, phylogenetic, and mutation sequences, along with alignments, are packaged into BioseqSets of classes pop-set, phy-set, and mut-set, respectively. The alignment information is extracted into a Seq-align, which is packaged as annotation (Seq-annot) associated with the BioseqSet. In the case of PHRAP quality scores, these are converted into a Seq-graph, which, similar to alignment information, is packaged in a Seq-annot; however, in this case it is associated with the nucleotide sequence and not the higher-level BioseqSet. The Seq-graph of PHRAP scores can be displayed in Sequin's Graphical view.

Features are usually packaged on the sequence indicated by their location. For example, the gene feature is packaged on the nucleotide Bioseq, and a protein feature is packaged on the protein Bioseq. Proteins are real sequences, and features such as mature peptides are annotated on the proteins in protein coordinates (although they can be mapped to nucleotide coordinates for display in a GenBank flatfile). A CDS (coding region) feature location points to the nucleotide, but the feature product points to the protein. For historical reasons, the CDS is usually packaged on the nuc-prot set instead of on the nucleotide sequence.

Viewing and Editing the Sequences

After the record has been constructed, the features can be viewed in a variety of display formats (Table 1). These include the traditional GenBank or GenPept flatfiles, a graphical overview, the feature spans displayed over the actual sequence letters, and ASN.1. These formats are generated by components of the NCBI Toolkit.

The different format generators all work independently from one another. When Sequin starts up, it registers a set of function procedures used to generate each display format. While issuing Sequin commands during manipulation of the sequence, appropriate messages (for example, “generate the view from the internal sequence record”, “highlight this feature”, “export the view to the clipboard”, etc.) are sent to the viewer by calling one of these procedures. Separate lists of registered formats are maintained for nucleotide, protein, and genome record types.

Just as the different format generators do not need to know about each other, Sequin's viewer windows do not need to know about other Sequin viewer or editor windows that are active at the same time. When editing a sequence, the user may have several different views of the same sequence open at the same time (for example, a GenBank flatfile and a graphical view). Clicking on a feature in the graphical view will select the same feature in the GenBank flatfile, and double-clicking on a feature launches the specific editor for that

feature. This type of communication between different windows is orchestrated by the NCBI Toolkit's object manager.

Table 1. The display formats available in Sequin.

Format	Notes
Alignment	For sets of aligned sequences
ASN.1	Abstract Syntax Notation 1 format
Desktop	The internal structure of a record
EMBL	As the record would appear in the EMBL database
FASTA	FASTA format
GBSeq	XML structured representation of GenBank format
GenBank	As the record would appear in GenBank or DDBJ
GenPept	Flatfile view of a protein
Graphic	Graphical representation of the sequence (several styles are available)
Quality	Displays the quality scores for each base in biological order
Sequence	Nucleotide sequence as letters plus any annotated features
Summary	Similar to graphical view but with no labels
Table	Sequin's 5-column feature table format
XML	eXtensible Markup Language, representation of ASN.1 data

The Sequence Editor

The sequence editor is used like a text editor, with new sequence added at the position of the cursor. Furthermore, the sequence editor automatically adjusts the biological feature intervals as editing proceeds. For example, if 60 bases are pasted or typed onto the 5′ end of a sequence record, the sequence editor will shift all the features by 60 bases. This means that interval correction does not need to be done by hand. Prior to Sequin, it was usually easier to resubmit from scratch than to edit all of the feature intervals manually.

The Feature Editors

Feature editor windows have a common structure, organized by tabs (Figures 2–4). The first tab is for elements specific to the given feature (Figure 2). For example, in a coding region editor, the first tab has controls for entering the coding region and reading frame. The second tab is for elements common to all sequence features (Figure 3). These include an exception flag, which allows explanations to be given for unusual events (e.g., RNA editing or non-consensus splice sites) and a comment (a free-text statement shown as "/note" in the flatfile). The last tab is a location spreadsheet allowing multiple feature intervals to be entered (Figure 4). For a coding region, these would reflect the boundaries of the exons used to encode the protein.

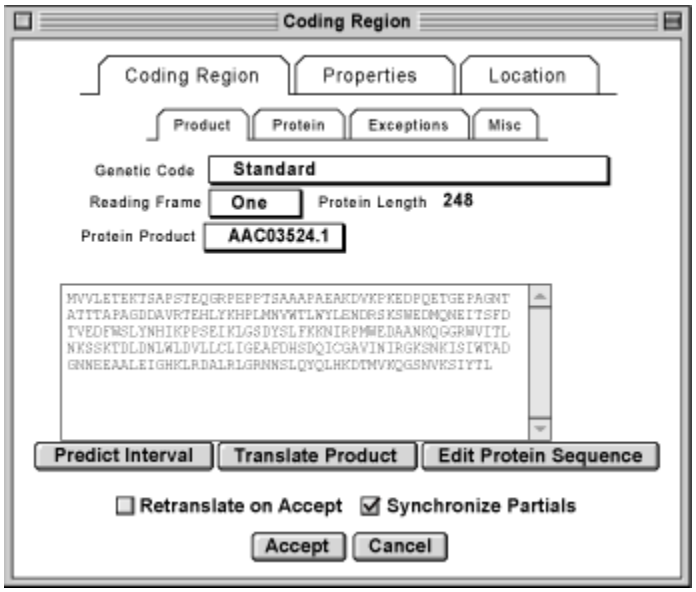


Figure 2. **The first feature editor tab, Coding Region, allows entry of information specific to the given type of feature.**

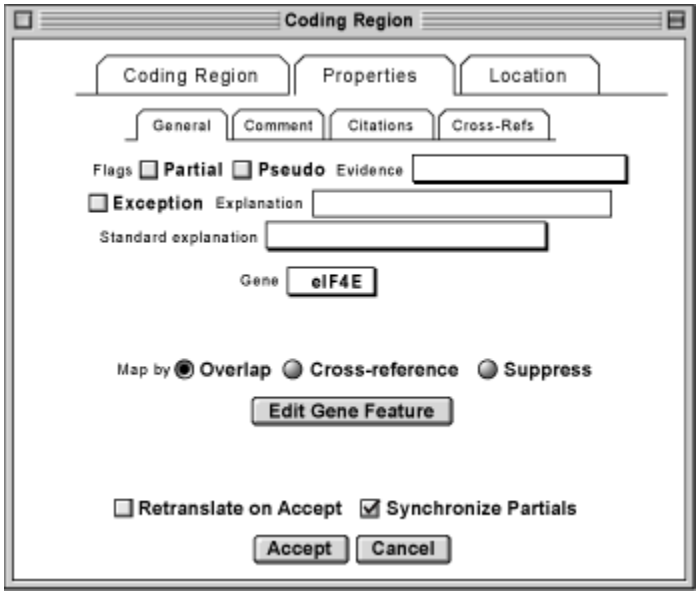


Figure 3. **The Properties tab is for entry of information common to all types of features.**

Computational Functions of Sequin

Sequin combines several NCBI Toolkit functions to perform many useful computations on the data in Sequin's memory.

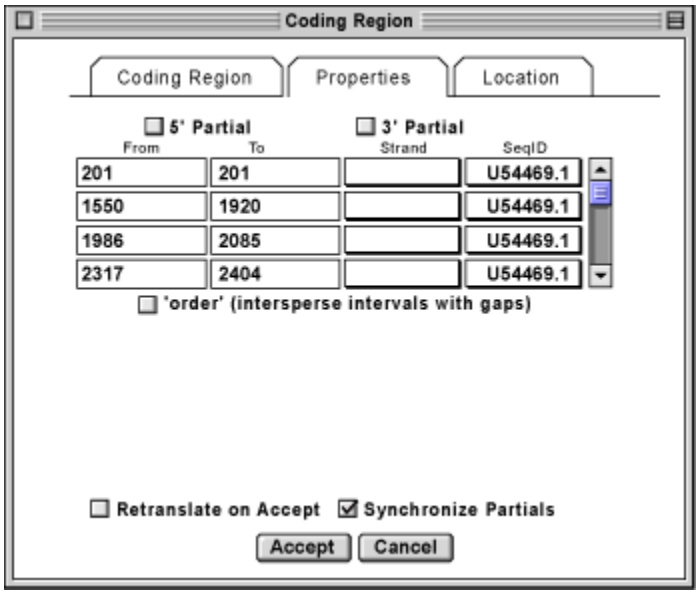


Figure 4. **The Location tab has a spreadsheet for entry of feature locations.**

Automatic Annotation of Coding Regions

When a nucleotide is submitted to GenBank using Sequin, it is essential to give the name of the source organism. The submitter is also strongly encouraged to supply the translated protein sequence(s) for the nucleotide.

Supplying the organism name allows Sequin to automatically find and use the correct genetic code for translating the nucleotide sequence to protein for the most frequently sequenced organisms. On the basis of only the genetic code and sequences of the nucleotide and protein products, Sequin will then calculate the location of the protein-coding region(s) on the nucleotide sequence. This is an extremely powerful function of Sequin. The ability to do this automatically, instead of by hand, has made sequence submission much faster and less error-prone.

Sequin uses a reverse translating alignment algorithm, called Suggest Intervals, to locate the protein-coding region(s) on the nucleotide sequence. The algorithm builds a table of the positions of all possible stretches of three amino acids in the protein. It then translates the nucleotide in all six reading frames and searches for a match to one of these triplets. When it finds one, it attempts to extend the match on each side of the initial hit. If the extension hits a mismatch or an intron, it stops. Given these candidate regions of matching, Sequin then tries to find the best set of other identical regions that will generate a complete protein. While doing this, the algorithm takes splice sites into account when deciding where to start an intron in eukaryotic sequences and fuses regions split by a single amino acid mismatch.

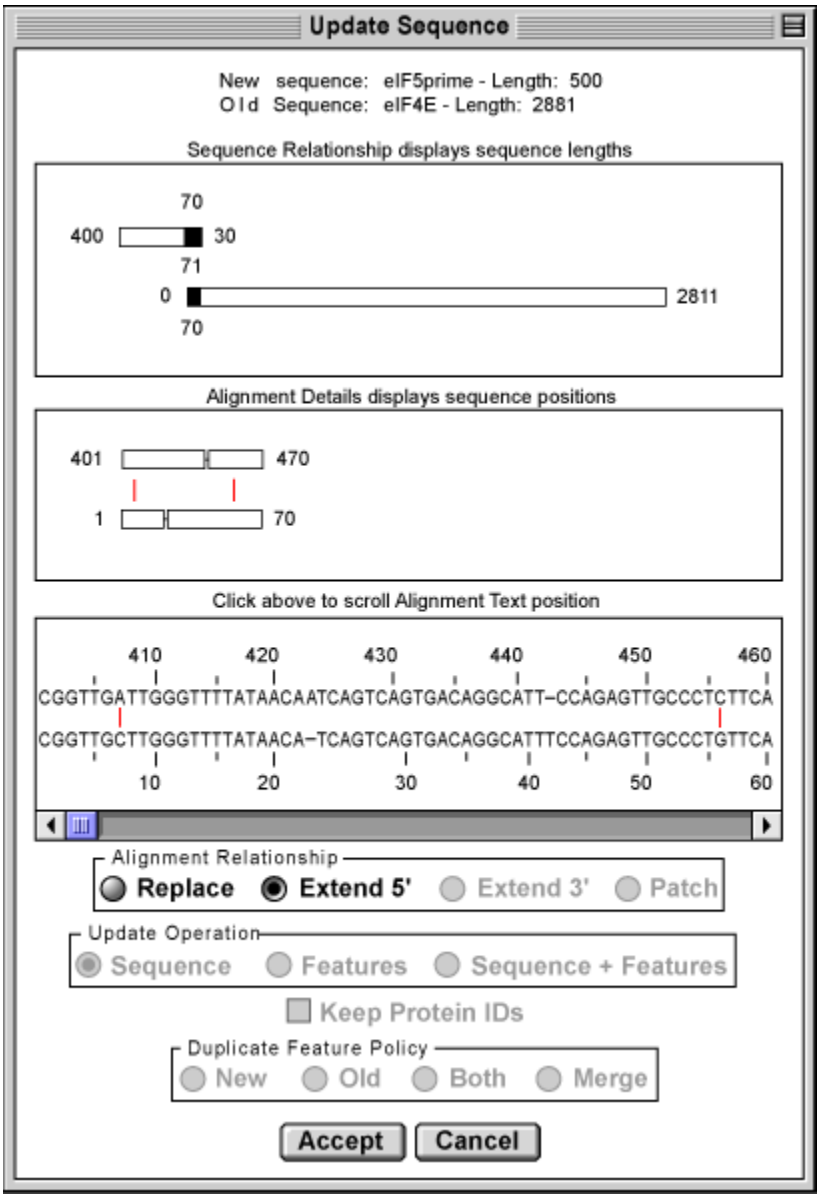


Figure 5. The Update Sequence window. The *top panel* shows the overall relationship between the two sequences, including the parts that align and any parts that do not align. From these data, Sequin determines whether the user is updating with a 5′ overlap, 3′ overlap, or full replacement sequence, and it presets radio buttons to indicate the relationship. The *second panel* shows a simple graphical view of the positions of gaps and base mismatches in the old and new sequences. The *third panel* shows the same information but with the actual sequence letters. Clicking on the second panel scrolls to the same place in the third panel.

Updating Sequence Records

The ability to propagate features through an alignment and the way the sequence editor can adjust feature positions as the sequence is edited are combined in Sequin to provide a simple and automatic method for updating an existing sequence.

The **Update Sequence** function allows overlapping sequence or a replacement sequence to be included in an existing sequence record. Sequin makes an alignment (using the BLAST functions in the NCBI Toolkit), merges the sequence if necessary, and propagates features onto the new sequence in the new positions. This effectively replaces the old sequence and features. The **Update Sequence** function is based on the NCBI Toolkit's alignment indexing, which allows Sequin to produce several displays that help the user to confirm that the correct sequence is in fact being processed (Figure 5).

The Validator

The final version of the sequence, complete with all the annotated features, can be checked using the validator. This function checks for consistency and for the presence of required information for submission to GenBank. The validator searches for missing organism information, incorrect coding-region lengths (compared to the submitted protein sequence), internal stop codons in coding regions, mismatched amino acids, and non-consensus splice sites. Double-clicking on an item in the error report launches an editor on the “offending” feature. The NCBI Toolkit has a program (testval), which is a stand-alone version of the validator.

The validator also checks for inconsistency between nucleotide and protein sequences, especially in coding regions, the protein product, and the protein feature on the product. For example, if the coding region is marked as incomplete at the 5′ end, the protein product and protein feature should be marked as incomplete at the amino end. (Unless told otherwise, the CDS editor will automatically synchronize these anomalies, facilitating the correction of this kind of inconsistency.)

Additional checks include ensuring that all features are annotated within the range of the sequence, all feature location intervals are noted on the same DNA strand, tRNA codons conform to the given genetic code, and that there are no duplicate features or different genes with the same names. The validator even checks that the sequence letters are valid for the indicated alphabet (e.g., the letter "E" may appear in proteins but not in nucleotides).

In cases where an exception has been flagged in a feature editor, specific validator tests can be disabled. For example, if the reason given for an exception is “RNA editing”, this turns off CDS translation checking in the validator. Likewise, “ribosomal slippage” disables exon splice checking, and “trans splicing” suppresses the error message that usually appears when feature intervals are indicated on different DNA strands.

Automatic Definition Line Generator

NCBI has a preferred format for the definition line in the GenBank format. It starts with the organism name, then the names of the protein products of coding regions (with the gene name in parentheses), with "complete" or "partial" at the end:

```
DEFINITION Human T-cell lymphotropic virus type 1 isolate ES-TMD envelope
           glycoprotein (env) gene, partial cds.
```

There is also a standard style for explaining alternative splice products. Sequin's Automatic Definition Line Generator collects CDS, RNA, and exon features in the order that they appear on the nucleotide sequence, finds the relevant genes (usually by location overlap), and prepares a definition line that conforms to GenBank policy.

Recalculation of Multiple CDS Features

In spite of all the safeguards built into Sequin, a submitter sometimes uses an incorrect genetic code for an organism. This means that the protein products of CDS translations may be incorrect. Sequin can retranslate all CDS features with a single command. Even so, if the sequence being edited is large, for example, a whole chromosome, this can be a time-consuming operation. To speed it up, the NCBI Toolkit uses a finite state machine (an efficient pattern search algorithm) for rapid translation. The machine is primed with a given genetic code, and then nucleotide sequence letters are fed into the algorithm one at a time, in the order they appear in the sequence. This allows all six frames (three frames on each strand) to be translated in the least possible time. The **Open Reading Frame** search in the NCBI Toolkit's tbl2asn program also uses this function.

Advanced Topics

The Special Menu

The Special Menu of Sequin encompasses a powerful set of tools that are available to GenBank and Reference Sequence indexers only. The Special Menu is not available to the public in the standard release because without a thorough understanding of the NCBI Data Model, use of the functions can cause irreparable damage to a record. It allows indexers to globally edit features, qualifiers, or descriptors in all sequences in a record, so that the same correction does not have to be made at each occurrence of the error. For example, all CDS features with internal stop codons can be converted to pseudogenes. Another common error made by submitters is to enter a repeat unit (a rpt_unit; e.g., ATTGG) in the repeat-type field (a rpt_type; e.g., tandem repeat). The Special Menu allows indexing staff to convert rpt_type to rpt_unit throughout the record.

NCBI Desktop Window

Although Sequin has editors for changing specific fields and Special Menu functions for doing bulk changes on several features, it is not possible to anticipate all of the

manipulations NCBI indexers might need to do to clean up a problem record. The NCBI Desktop window shows the internal structure of a record (Figure 1), i.e., how Bioseqs are packaged within Bioseq-sets, and where features, alignments, graphs, and descriptors are packaged on the sequences or sets. Objects (sets, sequences, features, etc.) can be dragged out of the record or moved to a different place in the record. Such manipulations could break the validity of the sequence record; therefore, great care must be taken when using it.

For technically adept Sequin users, the Desktop is where additional analysis functions can be added to Sequin without building a complicated user interface. With a feature or sequence selected, items in the Filter menu perform specific analyses on the selected objects. The standard filters include reverse, complement, and reverse complement of a sequence, and reverse complement of a sequence and all of its features. These are needed to repair the occasional record that came in on the wrong strand or in 3′→5′ direction. Adding new filter functions requires adding code to one of the Sequin source code files (one is provided with no other code in it for this purpose) and recompiling the program.

Network Analysis Functions

The functions of Sequin can be expanded by the addition of a configuration file that specifies the URLs for other programs (CGI scripts) available from the Internet. For example, tRNAscan-SE, a program by Sean Eddy and colleagues at Washington University (St. Louis, MO), can be used on sequences in Sequin in this way.

At a minimum, the CGI should be able to read FASTA format and should return either sequence data or the five-column feature table (discussed above) as its result. The external programs that Sequin knows about appear in the **Analysis** menu. When one of these analyses is triggered, Sequin sends a message to the URL and checks for the result with a timer so that the user can continue to work while the (Web) server is processing the request. The code for the CGI that chaperones data from Sequin to tRNAscan-SE and converts the tRNAscan results to a five-column feature table is in the demo directory of the public NCBI Toolkit.

Network Fetch Functions

As sequencing of complete bacterial genomes and eukaryotic chromosomes becomes more commonplace, the demand to break up long sequences into more manageable bites has increased (although Sequin is perfectly capable of editing these large records). Some genomes at NCBI are thus represented as segmented (or delta) Bioseqs, which are composed of pointers to other raw sequences in GenBank. Obtaining the entire sequence and all of the features requires fetching the individual components from a network service.

The object manager allows Sequin to know about different fetch functions that can be used. When a sequence is needed, these functions will be called until one of them satisfies the request. For example, the lsqfetch configuration file can be edited to point to a

directory containing sequence files on a user's disk. The SeqFetch function calls a network service at NCBI to obtain sequences and to look up Accession numbers given gi numbers or gi numbers given Accession numbers.

When used internally by NCBI indexers, Sequin can also fetch records from the DirSub and TMSmart databases. To ensure the confidentiality of pre-released records, this access requires the indexer to have a database password and to be working from a computer within NCBI. For additional protection, the paths to the database scripts are stored in a configuration file and are not encoded in the public Sequin source code.

Conclusion

Sequin has an important dual role as a primary submission tool and as a full-featured sequence record editor. It is designed to be modular on several levels, which simplifies the design and implementation of its components. Sequin sits at the top of the NCBI software Toolkit, relying on many of the underlying components, and thus acts as a quality assurance that these functions are working properly.

Chapter 13. The Processing of Biological Sequence Data at NCBI

Karl Sirotkin, Tatiana Tatusova, Eugene Yaschenko, and Mark Cavanaugh

Created: October 9, 2002; Updated: March 14, 2006.

The biological sequence information that builds the foundation of NCBI's databases and curated resources comes from many sources. How are these data managed and processed once they reach NCBI? This chapter discusses the flow of sequence data, from the management of data submission to the generation of publicly available data products.

Overview

The central dogma of molecular biology asserts that sequences flow from DNA to RNA to protein. In Entrez, DNA and RNA sequences are retrieved together as nucleotides and then integrated, along with proteins, into the NCBI system. Once in the system nucleotides and proteins are both available for public use in at least three ways:

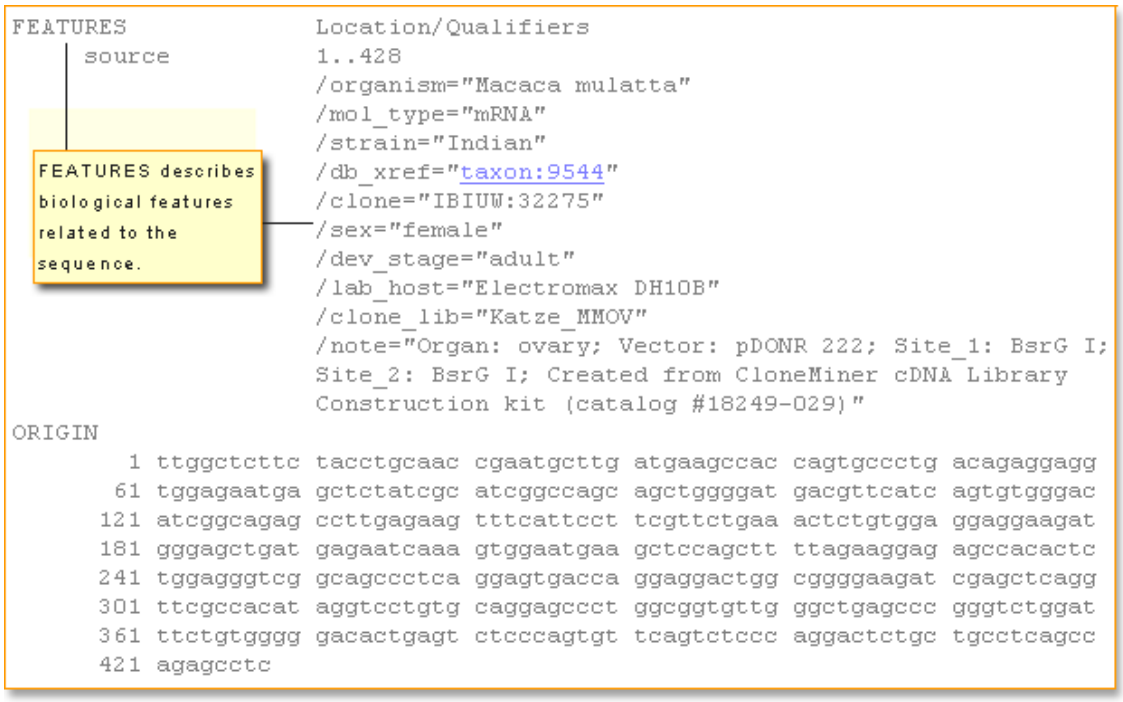
1. The Entrez system (Chapter 15) retrieves nucleotide and protein sequences according to text queries that are entered into the search box. Text queries can be followed by search fields, such as author, definition line, and organism (for example, "homo sapiens"[orgn]), and are used to further define raw sequence data being used for retrieval.
2. The sequences themselves can be searched directly by using BLAST (Chapter 16), which uses a sequence as a query to find similar sequences.
3. Large subsets of sequences can be downloaded by FTP.

There are many sources for both nucleotide and protein sequences. Sequences submitted directly to GenBank (Chapter 1) or replicated from one of our two collaborating databases, the European Molecular Biology Laboratory (EMBL) Data Library and the DNA Data Bank of Japan (DDBJ), are the major sources. The Reference Sequence collection (Chapter 18) and the UniProt database, which incorporates data from SWISS-PROT, are yet additional sources.

An information management system that consists of two major components, the ID database and the IQ database, underlies the submission, storage, and access of GenBank, BLAST, and other curated data resources (such as the Reference Sequences (Chapter 18), the Map Viewer (Chapter 20), or Entrez Gene (Chapter 19)). Whereas ID handles incoming sequences and feeds other databases with subsets to suit different needs, IQ holds links between sequences stored in ID and between these sequences and other resources.

Abstract Syntax Notation 1 (ASN.1) Is the Data Format Used by the ID System

ASN.1 is the data description language in which all sequence data at NCBI are structured. ASN.1 allows a detailed description of both the sequences and the information associated with them, such as author names, source organism, and biological features (known as "features"). The image below shows **FEATURES** as displayed in GenBank format.




```
FEATURES             Location/Qualifiers
     source          1..428
                     /organism="Macaca mulatta"
                     /mol_type="mRNA"
                     /strain="Indian"
                     /db_xref="taxon:9544"
                     /clone="IBIUW:32275"
                     /sex="female"
                     /dev_stage="adult"
                     /lab_host="Electromax DH10B"
                     /clone_lib="Katze_MMOV"
                     /note="Organ: ovary; Vector: pDONR 222; Site_1: BsrG I;
                     Site_2: BsrG I; Created from CloneMiner cDNA Library
                     Construction kit (catalog #18249-029)"
ORIGIN
        1 ttggctcttc tacctgcaac cgaatgcttg atgaagccac cagtgccctg acagaggagg
       61 tggagaatga gctctatcgc atcggccagc agctggggat gacgttcatc agtgtgggac
      121 atcggcagag ccttgagaag tttcattcct tcgttctgaa actctgtgga ggaggaagat
      181 gggagctgat gagaatcaaa gtggaatgaa gctccagctt ttagaaggag agccacactc
      241 tggagggtcg gcagccctca ggagtgacca ggaggactgg cggggaagat cgagctcagg
      301 ttcgccacat aggtcctgtg caggagccct ggcggtgttg ggctgagccc gggtctggat
      361 ttctgtgggg gacactgagt ctcccagtgt tcagtctccc aggactctgc tgcctcagcc
      421 agagcctc
```

In the ASN.1 format, the organism information is presented as shown below. You can also see a complete ASN.1 record.

```
orgname { name binomial { genus "Macaca" , species "mulatta" } ,
```

Maintaining all data in the same structured format simplifies data parsing, manipulation, and quality assurance, and eases the task of data integration and software development for sequence analysis. All of the various divisions of GenBank can be downloaded in ASN.1 from the NCBI FTP site. In the ID data management system, data are stored as ASN.1 blobs, minimizing the amount of biological information that is captured and updated in the relational database schema.

Similar to an XML DTD, ASN.1 has an associated file that contains the description of the legal data structure. This file is called asn.all and is available as part of the "C" toolkit in an archive named "ncbi.tar.gz" located in the FTP directory. When unpacked, the directory "/demo", found in the "ncbi.tar.gz" archive, contains the asn.all file. In the same "/demo" directory is testval.c, a tool that validates the data against asn.all. Additionally, a set of

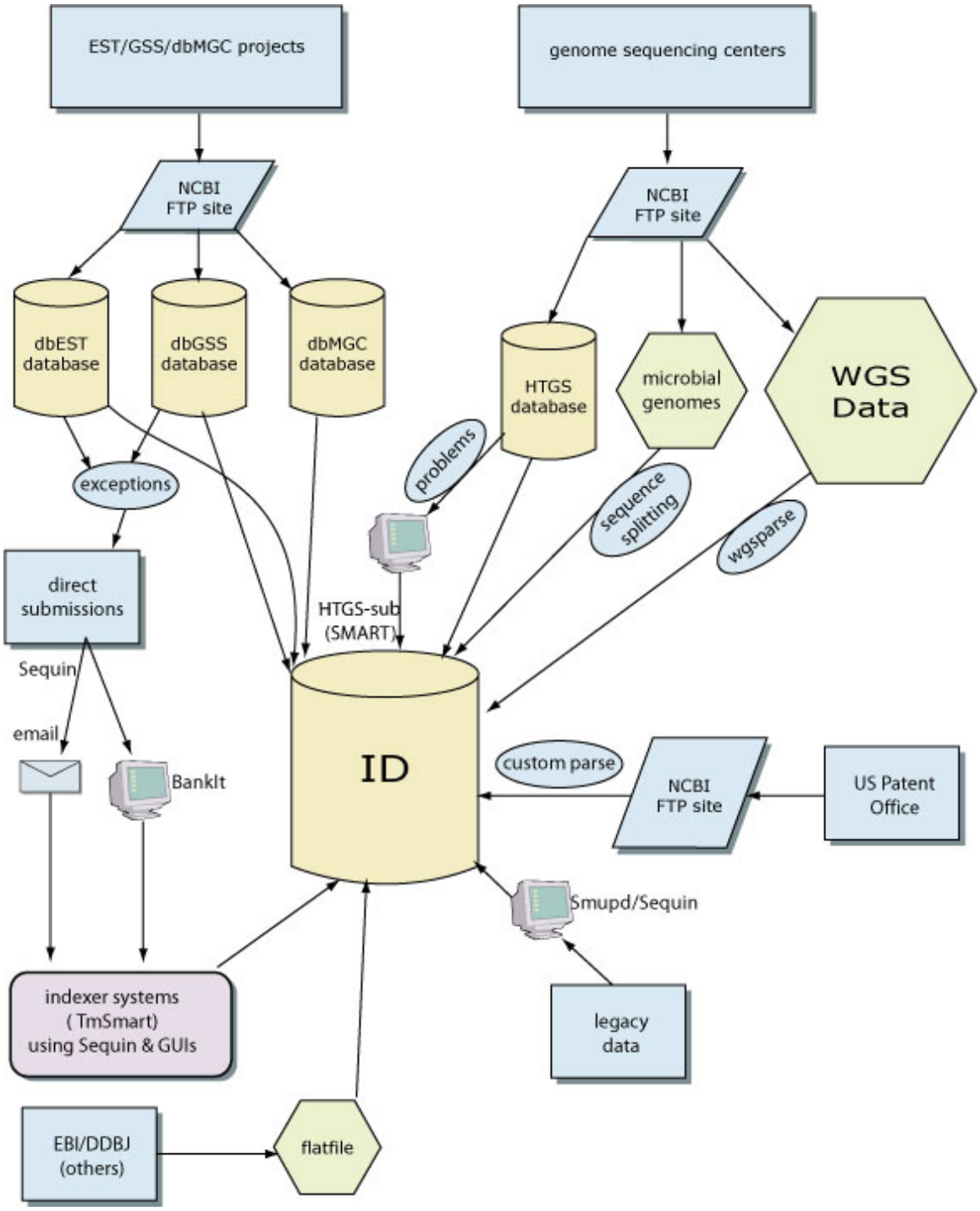


Figure 1. Sources of sequence data available at NCBI.

utilities for producing ASN.1 while programming in "C" is found in the subutil.c file of the "/api" directory, which is unpacked from the same "ncbi.tar.gz" archive.

Sources of Sequence Data

The sequence data available at NCBI comes from many different sources (Figure 1). In summary, the data consist of:

- GenBank sequences (Chapter 1)
- Reference sequences (Chapter 18)
- sequences from other databases, such as SWISS-PROT, PIR, PRF, and PDB
- sequences from the United States patents

The submission pathway depends on the data source (see Figure 1) and volume. HTGS and other large-volume submitters use FTP, usually after converting their data to ASN.1 with tools such as tabl2asn. Small-volume submitters typically use either BankIt (Chapter 1) or Sequin (Chapter 12) to prepare the ASN.1 for submission.

The data received are then subjected to some quality control by the submission tools BankIt, Sequin, and fa2htgs. These tools have built-in validation mechanisms to check if the data submitted have the correct structure and contain the essential information. The work of the GenBank indexing staff, who uses Sequin, adds one more layer of quality control and provides assistance to submitters. The staff also helps with the use of Sequin for complex submissions

Data Flow Components

The ID Database

The ID database is a group of standard relational databases that holds both ASN.1 objects and sequence identifier-related information. ASN.1 objects follow the specifications in the asn.all file for NCBI sequence data objects. ID holds data for GenBank and the many databases in the Entrez system. Details of the architecture of relational ID databases and the software associated with them are described later in this chapter. All of the sequences from the International Nucleotide Sequence Database Collaboration (INSDC)are in GenBank, and they all have Accession numbers assigned to them. Accession numbers point to sequences and their associated biological information and annotation.

In the ID database, blobs are added into a single column of a relational database. Although the columns behave as in a relational database, the information that makes each blob, such as biological features, raw sequence data, and author information, are neither parsed nor split out. In this sense, the ID database can be considered as a hybrid database that stores complex objects.

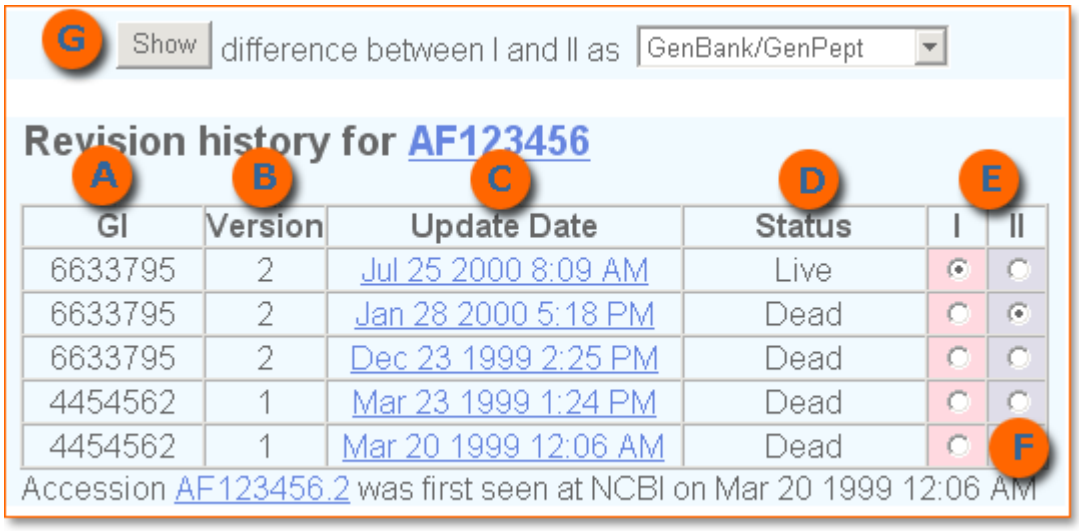
Note: Blob stands for Binary Large Object (or binary data object) and refers to a large piece of data, a large structured data object that can be stored as a unit and processed by software that knows the structure. For more information, check the Glossary.

Versions, GIs, Annotation Changes, and Takeovers

Every time a change is made to a sequence, a new version of the sequence is produced. This new version has a new GI number (GI or GenInfo Identifier is a sequence identification number for a nucleotide sequence) assigned to it (**A** and **B** in the image below). When a change is made to the annotation associated with a sequence, a new blob is produced, but no new version or GI is assigned. This series of events marks the history of the sequence since its first days in GenBank.

You can track annotation and sequence changes, as well as the "takeover" of one record by another by using the Sequence Revision History tool. The tool can be accessed from the side blue bar in Entrez Nucleotide and Entrez Protein and is used to highlight differences in sequence versions and annotations. To understand how the History tool works, let's examine the history of the Gallus gallus doublesex and mab-3 related transcription factor 1 mRNA (Accession AF123456), which was first added to GenBank March 20, 1999.

Click on **Check sequence revision history** in the blue side bar of Entrez Nucleotide or Entrez Protein to be directed to the **Sequence Revision History** page. Enter the Accession or GI numbers or the FASTA-style Sequence IDs (**SeqIds**) into the **Find** box. The **Revision history** for AF123456 is displayed.



The Update Date column (**C** in the image above) contains the date of every update to AF123456. Some involve sequence changes, others involve only annotation changes. Click on a date in the column to retrieve AF123456 as it existed at that point in time. The status column (**D**) reports which version is live and which ones are dead. Columns I and II (**E**) are used to compare two different sequences.

Notice that on **Mar 23 1999**, at 1:24 PM, a new ASN.1 blob was produced for Accession AF12345. However, no new GI number (**A**) or version (**B**) was assigned because the changes were limited to the annotation and biological features of the sequence, with no changes made to the sequence data. On December 23, 1999, Accession AF123456 gained a

new GI (**6633795**) and version (**Version 2**) because in this case a change was made to the sequence data.

Compare the two blobs produced on March 23, 1999 and December 23, 1999 to see the difference between them.

- Start by accessing the Revision history for AF12345.
- Select one sequence in each column (I or II) as shown in the image above (**E**).
- Push the **Show** button at the upper left of the page to display the two blobs (**G**).

The differences between blobs are highlighted, with each blob displaying a different color. Compare ASN.1 blobs produced on March 20, 1999 and March 23, 1999 and you will see that the differences between the two are limited to the annotation and biological features described in the blobs, whereas the sequence data remain the same.

The understanding of the biological features related to a sequence can change with or without a change in the underlying genetic sequence. For example, the sequence revision history of J00179 reveals that although the annotation changed four times, there has been only one sequence version (**J00179**) with one GI (**183807**). J00179 can still be retrieved in Entrez by searching its Accession or GI number, but this record has been replaced by Accession U01317 and therefore is no longer indexed. The version number assigned to the "take over" record U01317 is 1, whereas the replaced version of this record (J00179) remains as **Version 0**. All sequences deposited before February 1999 received no sequence version, that's why J00179 is version zero. In February 1999, the use of a sequence version was implemented, and all sequences deposited in GenBank at that time received a version number 1. Since then, ordinals assigned to sequence versions have increased every time a change is made to the sequence data.

The use of both systems, Version and GI, leads to two parallel ways of tracking sequence versions for an object. In the GenBank flatfile, the Accession Version provides the ordinal instance (version) of the sequence. Within ID, each unique sequence is assigned a GI number; and therefore the instances of an Accession can be tracked by checking its chain of GI numbers. Note that Accession and Accession Version are different things, with the former been used to designate a DNA sequence of some molecule or piece of some molecule deposited in GenBank and the latter to indicate the version of that sequence. A single Accession can have many GIs that are assigned every time the sequence changes, whereas an Accession Version has only one GI.

Within the ID relational databases, there is a chain identifier that can be used to link these GI numbers. Not all sequences within ID are in GenBank and not all have sequence versions, but all sequences have a chain of GI numbers. For this reason, internally, the GI number is the universal pointer to a particular sequence, as opposed to the Accession Version, which would work only for versioned sequences. The ID database is also the controller for allowed "takeovers" of one Accession by another. In the example above, GI 4454562 is taken over by GI 6633795. A takeover can also occur when the sequences of

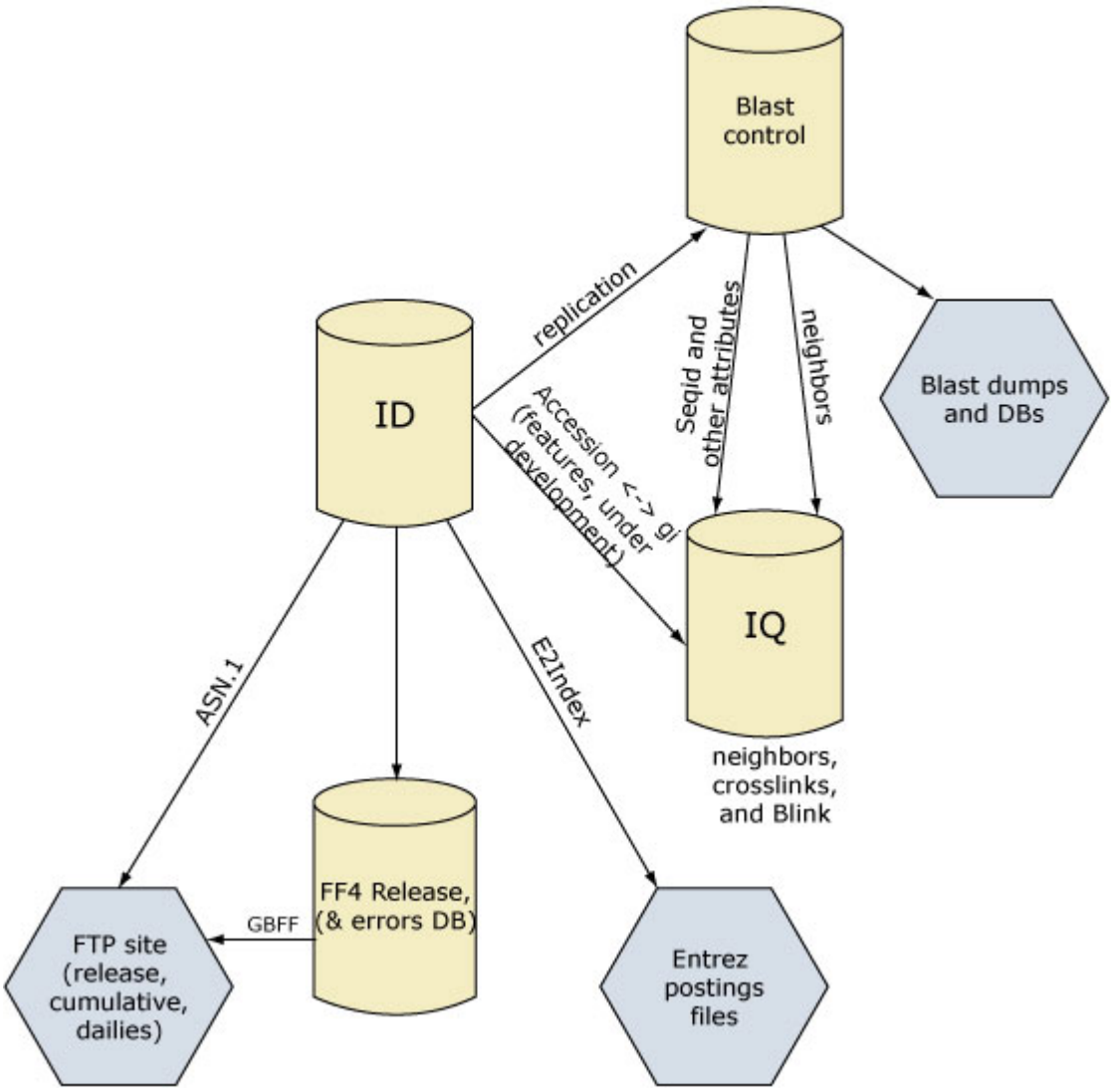


Figure 2. Products of the ID system.

two clones are merged into a single clone. One or several of the Accessions of older clones can be taken over by a new Accession.

Output of Data from the ID System

Once all incoming data have been converted to ASN.1 format and entered into ID, the data are then replicated into several different servers and transformed into several different formats (Figure 2). The replication is necessary for a number of reasons: (i) it separates the "incoming" data system (ID) from the "outgoing" data which is the data

used in response to scientific queries by users; (ii) it helps balance the load of queries, thus providing quicker response times and allowing different servers to specialize in different functions; and (iii) it protects against data loss should one server fail. The details of the internal structure of the ID system and how the structure is replicated are discussed in the Data Flow Architecture section.

The IQ Database

The IQ database is a Sybase data-warehousing product that preserves its SQL language interface but which inverts its data by storing it by column, not by row. Its strength is in its ability to speed up results from queries based on the anticipated indexing. This non-relational database holds links between many different objects.

For example, as part of the processing of incoming sequences, each protein and nucleotide sequence is searched for similar sequences (Chapter 16) against the rest of the database. Users can then select the **Related Sequences** link that is displayed next to each record in Entrez Nucleotide and Entrez Protein (Chapter 15) to see a set of similar sequences, sometimes known as "neighbors". The IQ database keeps track of the neighbors for any given sequence. These relationships are all pre-computed to save users' time.

IQ stores the relationships between similar nucleotide sequences and between similar protein sequences and which proteins are coded for by which nucleotides and also holds information on the links between entries in different Entrez databases. This might include, for example, information on the publications cited within sequence records, which links to PubMed or to an organism in the Taxonomy database. Some of this information comes from the analysis of the ASN.1 in ID by e2index, a tool that extracts terms from NCBI sequence ASN.1 during "indexing" for Entrez.

The BLAST Control Database

The BLAST Control database receives information from ID that is used to generate BLAST databases (Chapter 16) for the BLAST query service and for stand-alone BLAST users. The information is used internally to generate the sequence neighbors stored in IQ.

The GenBank Flatfile and Error Capture Databases

Many NCBI users think of the GenBank flatfile as the archetypal sequence data format (see an example of a GenBank flatfile). However, within NCBI and especially within the ID internal data flow system, ASN.1 is considered the original format from which reports such as the GenBank flatfile can be generated (see an example of an ASN.1 file).

Although the GenBank flatfile is usually generated on demand from the ASN.1, for certain products such as complete GenBank releases, a GenBank flatfile image is made for each active sequence. This flatfile is stored in a database called FF4Release, which consists of the latest transformation of ASN.1 to the GenBank flatfile format.

The FF4Release database is also a place where internal error reports are captured. The reports can be analyzed and displayed for different time points in the data processing pathway:

- ASN.1 itself can be validated using the testval (or its replacement, asnval) tool—syntax checking is not necessary, because the underlying ASN.1 libraries enforce proper syntax according to the definition file.
- Errors can be discovered during conversion to the GenBank flatfile format.
- Through a reparse from the GenBank flatfile format to ASN.1. This is done as a further check for legality of the ASN.1, and our current software for producing GenBank format reports from it.

Entrez Postings Files

When sequences are submitted to GenBank or one of our collaborating databases, additional information about the sequence is often included. This might be a brief description of a gene in the definition line, along with annotated sequence features such as the source organism name. To make this information searchable via Entrez, these words have to be indexed. They are extracted from the ASN.1 using e2index and then stored in the Entrez posting files, which are optimized for Boolean queries by the Entrez system (see Chapter 15).

All of these products from the ID system are listed in Table 1. NCBI also generates weekly "LiveLists" for public, collaborator, and in-house use. LiveLists show all Accession numbers currently in use. Accession numbers that have been replaced or otherwise removed from circulation because of error or submitter request are not in the LiveList.

Table 1. Products of the ID system.

Type	Source	ASN.1	GBFF [a]	Qscore	GenPept	Protein FASTA
Cumulative	GenBank	X		X	X	X
Incremental	GenBank	X		X	X	X
Incremental	GenBank[b]		X	X		
Cumulative	RefSeq	X	X		X	X
Incremental	RefSeq	X	X		X	X

[a] GBFF, GenBank flatfile; Qscore, sequencing quality score; GenPept, GenBank Gene Products.
[b] NCBI records only.

Data Flow Architecture

Sequences enter ID when a client (internal to NCBI) loads data into the system. The ASN.1 data can be loaded either through a stand-alone program or a client API. In both cases, the data are submitted to ID through IDProdOS, an open server (commonly called "middleware") that sits between the clients and the database system. An overview of the

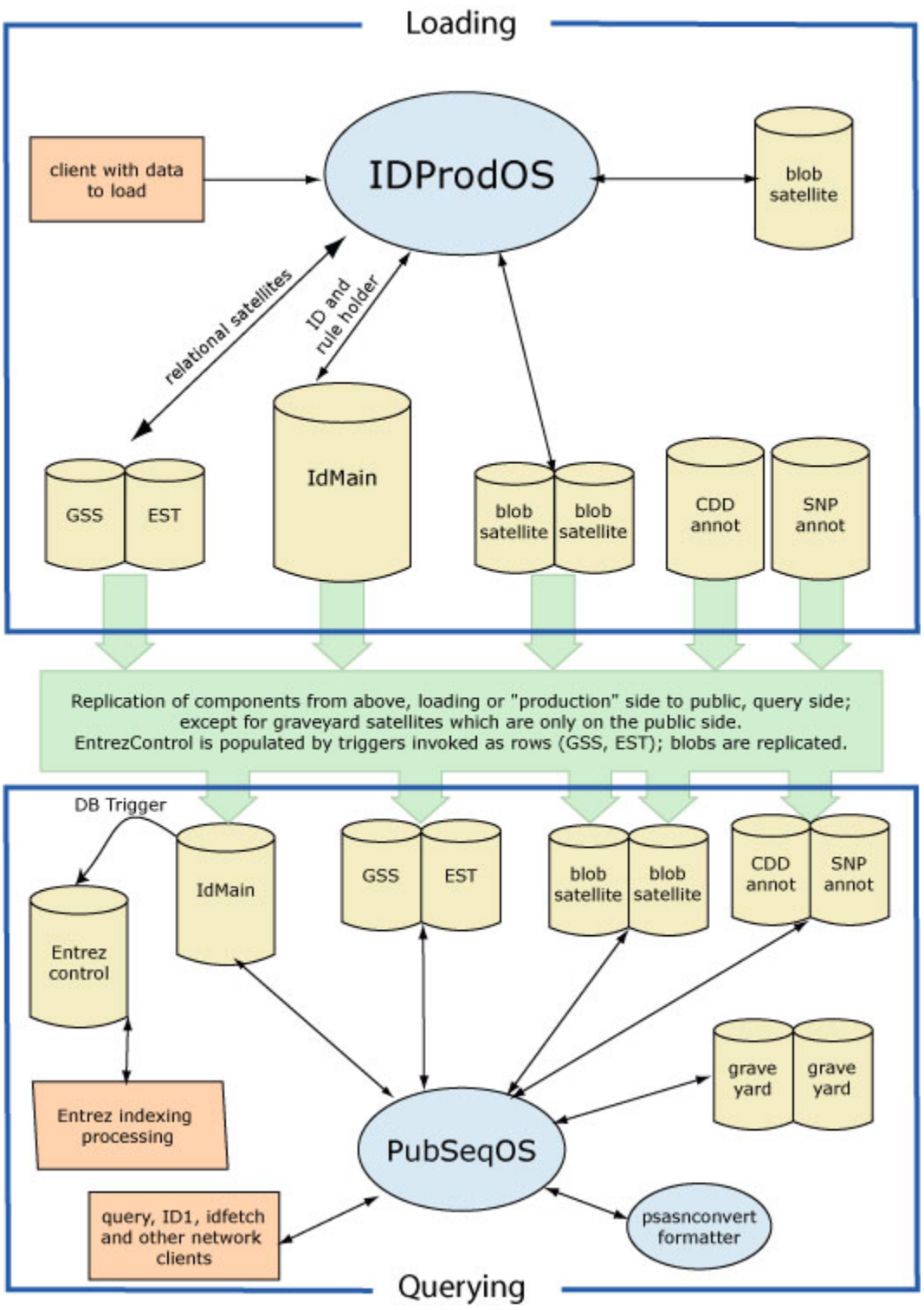


Figure 3. The ID system architecture.

flow of sequence data through the ID architecture with its multiple components is shown in Figure 3 and discussed below.

IDProdOS hides details of the underlying complexity from the client API, which was shown to be useful when the previous version of the ID system (a single database and an open server) was converted to the current system without requiring any changes to the clients.

IDProdOS does an initial check of the actions required by the load. For example, in a record that has DNA and protein sequences, including annotation and sequence identifiers, the identifier on the protein has to be unique. The same identifier should not be given to an outdated DNA sequence and a current sequence, unless the current sequence has replaced the old one. That's because proteins, generally, are not allowed to move between GenBank records, although proteins moving between segments of a complete genome submission are sometimes allowed.

Additional checking is performed by stored procedures in the IdMain database. The details of what is allowed vary according to the source of the ASN.1, which includes direct submissions from collaborators and the NCBI RefSeq project. These procedures check (i) which sequence identifiers may be used, (ii) which sequences may be replaced by which other sequences, and (iii) which sequence version may be used in a record.

If the sequences pass all these checks, three things happen: (i) IDProdOS changes the SeqId pointers in the blob to GI numbers, which are now used as sequence-specific pointers, (ii) IdMain retains the sequence identifier information that was also used for the checking, and (iii) IDProdOS loads the ASN.1 blobs to the blob satellites.

The IdMain database contains the sequence identifiers for each of the sequence records, including all those for ASN.1 blobs that contain multiple sequences. It enforces sequence version rules, among other rules.

Relational satellite databases are fully normalized databases that hold records for which there is only one sequence per intended ASN.1 blob. Few, if any, features are allowed on records intended for relational satellite databases (the PubSeqOS produces the ASN.1 by converting the data extracted from relational tables). This contrasts with the Blob satellite databases, from which ASN.1 is retrieved as-is. Blob satellite databases, different from relational databases, contain ASN.1 objects as unnormalized data objects.

Recently, annotation-only satellite databases have been added to the ID system. These satellites contain annotation to be added to Bioseqs, linked by GI number. Because there are multiple such annotation satellite databases, more than one set of additional annotation may be added to a Bioseq.

The SnpAnnot database contains feature information that is limited to simple mutation information from dbSNP (Chapter 5). The CDD Annotation database contains feature information that is limited to protein domains for the protein sequences known to ID. In

both cases, these features might be added to NCBI-curated records by the PubSeqOS when the records are requested.

To visualize the role of replication, the rectangle in the middle of Figure 3 represents the use of the Sybase Replication Server to copy information from the loading side of the system to the query side.

Similar to IDProdOS, PubSeqOS is a open server (also called "middleware") that sits between the clients and the database system. It hides details of the underlying complexity from the client API. It actually has an almost identical code base as IDProdOS because they both serve similar functions. When a record is requested in a format other than ASN.1, psansconvert is called to do the conversion. This distinct *child* process allows both insulation from any possible instability and allows for use of multiple central processing units (CPUs) in a natural way.

Note: The *child* process is a technical term used to describe a process that is owned by and completely dependent on a parent process that initiated it.

At the query side are all records in Entrez, plus graveyards and EntrezControl, a special database that is not queried by the public. EntrezControl is used to control the indexing of blobs for Entrez. Its rows are initiated by a trigger that fires when rows are added by replication to the IdMan database. A trigger is a special, database-stored procedure that responds to changes in a database table.

The graveyards are databases that contain blobs that were replaced or taken over and therefore no longer indexed in Entrez. Once replaced or taken over, blobs do not change—which is the reason why they are limited to the query side—but they are still retrievable by GI or other sequence identifier.

Chapter 14. Genome Assembly and Annotation Process

Paul Kitts

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The primary data produced by genome sequencing projects are often highly fragmented and sparsely annotated. This is especially true for the Human Genome Project as a result of its policy of releasing sequence data to the public sequence databases every day (1, 2). So that individual researchers do not have to piece together extended segments of a genome and then relate the sequence to genetic maps and known genes, NCBI provides annotated assemblies of public genome sequence data. NCBI assimilates data of various types, from numerous sources, to provide an integrated view of a genome, making it easier for researchers to spot informative relationships that might not have been apparent from looking at the primary data. The annotated genomes can be explored using Map Viewer (Chapter 20) to display different types of data side-by-side and to follow links between related pieces of data.

This chapter describes the series of steps, the "pipeline", that produces NCBI's annotated genome assembly from data deposited in the public sequence databases. A variant of the annotation process developed for the human genome is used to annotate the mouse genome, and similar procedures will be applied to other genomes (Box 1).

NCBI constantly strives to improve the accuracy of its human genome assembly and annotation, to make the data displays more informative, and to enhance the utility of our access tools. Each run through the assembly and annotation procedure, together with feedback from outside groups and individual users, is used to improve the process, refine the parameters for individual steps, and add new features. Consequently, the details of the assembly and annotation process change from one run to the next. This chapter, therefore, describes the overall human genome assembly and annotation process and provides short descriptions of the key steps, but it does not detail specific procedures or parameters. However, sufficient detail is provided to enable users of our assembly and annotations to become familiar with the complexities and possible limitations of the data we provide.

Box 1. Annotation of other genomes.

NCBI may assemble a genome prior to annotation, add annotations to a genome assembled elsewhere, or simply process an annotated genome to produce RefSeqs and maps for display in Map Viewer (Chapter 20).

The basic procedures used to annotate other eukaryotic genomes are essentially the same as those used to annotate the human genome. However, the overall process is adjusted to

Box 1 continues on next page...

Box 1 continued from previous page.

accommodate the different types of input data that are available for each organism. Genes can be annotated on any genome for which a significant number of mRNA, EST, or protein sequences are available. Other features, such as clones, STS markers, and SNPs, can also be annotated whenever the relevant data are available for an organism.

For example, genes and other features are placed on the mouse Whole Genome Shotgun (WGS) assembly from the Mouse Genome Sequencing Consortium (MGSC) by skipping the assembly steps used in the human process but following the annotation steps with relatively minor adjustments. A variation of the human process is also used to assemble and annotate genomic contigs from finished mouse clone sequences (see the Map Viewer display of the mouse genome).

Overview of the Genome Assembly and Annotation Process

Figure 1 shows how the main steps in the human genome assembly and annotation process are organized and also shows the most significant interdependencies between the steps. The pipeline is not linear, because whenever possible, steps are performed in parallel to reduce the overall time taken to produce an annotated assembly from a new set of data. Some of the steps are run incrementally on a timetable that is independent from that of the main pipeline to produce a new assembly more quickly.

Data Freeze

New sequence data that could be used to improve the genome assembly and annotation become available on a daily basis. Since the assembly and annotation process takes several weeks to complete, the data are "frozen" at the start of the build process by making a copy of all of the data available for use at that time. Freezing the data provides a stable set of inputs for the remainder of the build process. Additional or revised data that become available during the period taken to complete the process are not used until the next build.

The Build Cycle

A build begins with a freeze of the input data and ends with the public release of an annotated assembly of genomic sequences (Figure 1). There are usually a few months between builds so that the latest build can be evaluated and improvements can be made.

Steps Run Incrementally

The early steps of the genome assembly and annotation pipeline involve many computationally intensive processes, including masking the repetitive sequences and aligning each genomic sequence to the other genomic sequences, mRNAs, and Expressed Sequence Tags (ESTs). Running these steps incrementally minimizes the time between starting a new build and being ready to start the assembly and annotation steps.

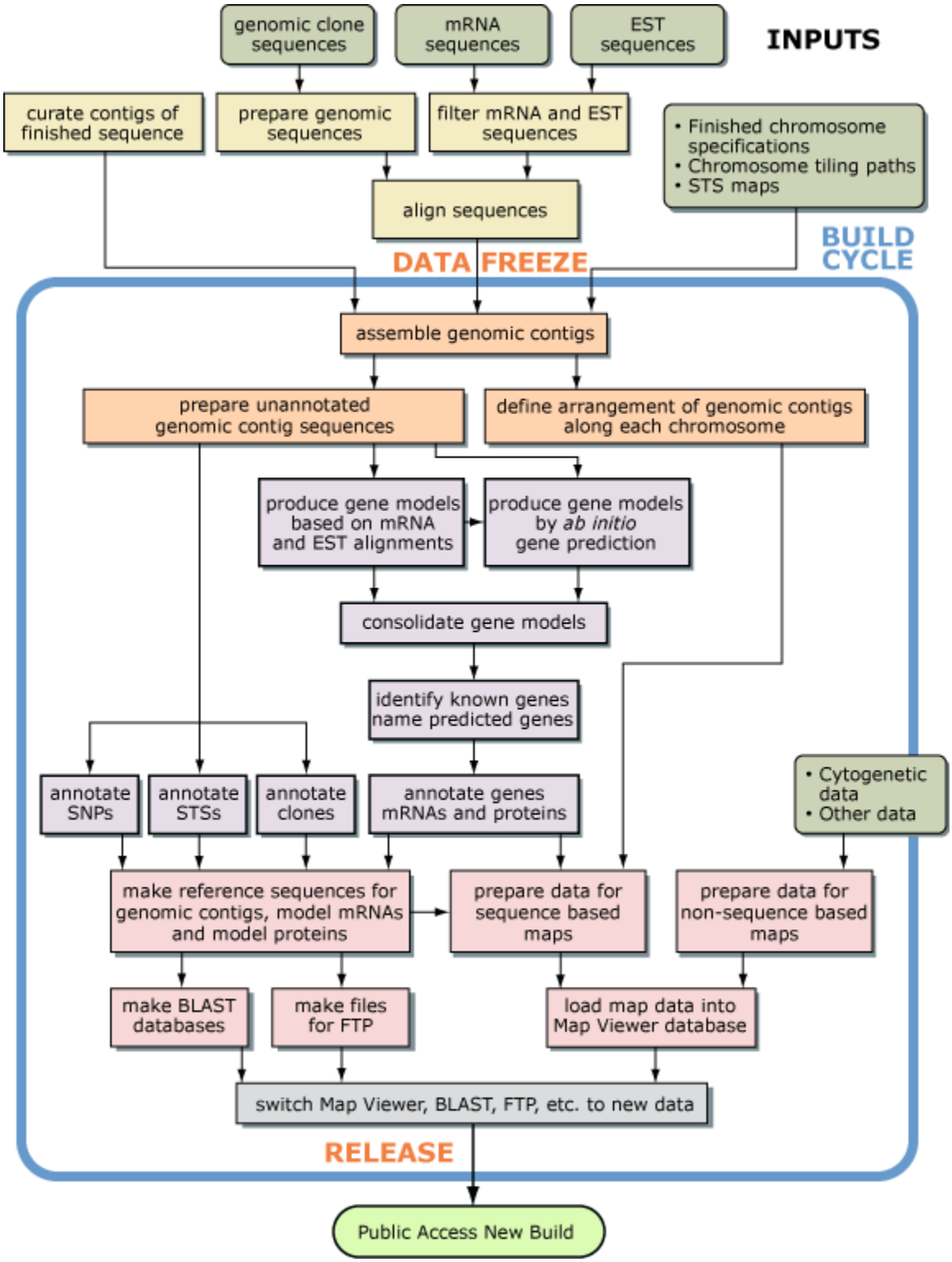


Figure 1. **The human genome assembly and annotation process.**

Approximately once a week, independent from the build cycle, new or updated sequences that have been deposited in GenBank are retrieved for processing. Periodically, old versions of the sequences are purged from the set of accumulated data files.

The manual refinement of the set of assembled genomic contigs produced entirely from finished sequence is another time-consuming step that is carried out incrementally, approximately once a week.

Steps Run Irregularly

Because some data change infrequently, some relatively quick steps are executed on time frames that are not tied to the build cycle. For example, the list of special cases used to override the automatic process is updated whenever the need becomes apparent.

The Input Data

The main inputs for the genome assembly and annotation process are genomic sequences, transcript sequences, and Sequence Tagged Site (STS) maps.

Human Genomic Sequences

Genomic Sequences Used for Assembly

Genomic sequences from the following five data sets are processed for use in the assembly:

High-Throughput Genomic Sequences. Human high-throughput genomic sequences (HTGSs) are retrieved using the Entrez query system (Chapter 15). The query used returns sequences for all entries that contain the HTG keyword, regardless of whether the sequence is finished or is in any of the unfinished draft phases.

Finished Chromosome Sequences. The center that coordinates the sequencing of a finished chromosome submits a specification regarding how to build the sequence of that chromosome from its component clone sequences to a data repository at the European Bioinformatics Institute (EBI). The sequences specified for any finished chromosomes are retrieved from GenBank and included in the set of genomic sequences to be processed for an assembly.

Genomic Sequences from the Tiling Paths of Individual Chromosomes. The human genome sequencing centers use a variety of experimental evidence to compile an ordered list of clones they believe provides the best coverage for each chromosome. At least once every 2 months, the sequencing centers submit an updated minimal tiling path for each chromosome to a data repository at the EBI. These tiling path files (TPFs) include Accession numbers if sequence for a clone is available. The tiling path repository is checked each day for the Accession numbers from any tiling path that has been updated. Any secondary Accession numbers are replaced with the corresponding primary Accession numbers, and any invalid Accession numbers are flagged to prevent those

sequences from being used for assembly. The latest version of the sequence for each Accession number in the most recent clone tiling paths is retrieved from GenBank and included in the set of genomic sequences to be processed for assembly.

Assembled Blocks of Contiguous Finished Genomic Sequence. As the sequences for individual clones are finished, they are merged with overlapping finished sequences to form contigs (3). The primary source for identifying neighboring clones is the clone tiling path for each chromosome. Additional information is obtained from some GenBank entries that contain annotation specifying the neighboring clones. BLAST (4) is used to align the sequences of candidate pairs of clones, and a merged sequence is produced automatically if the expected overlap is confirmed by the sequence alignment. When the automatic processes do not find an expected overlap, there is a manual review to find the correct overlap, refining the clone order if necessary. The most recent set of finished contigs are processed for assembly.

Additional Genomic Sequences. A few other specific human genomic sequences are added to the assembly set because they contain genes that may not be represented in the genomic sequences from the other sources.

Genomic Sequences Used for Ordering and Orienting

Sequences known to come from both ends of the same cloned genomic fragment provide valuable linking information that helps to order and orient sequence contigs in the assembly step (5, 6). The SNP Consortium sequenced the ends of the inserts in several million plasmid clones containing small (0.8–6 Kbp) fragments of human genomic DNA. In many cases, both ends of the same insert were sequenced (Table 4 in Ref. 7), thereby providing a set of plasmid paired-end sequences.

Genomic Sequences Used for Annotation

Curated Genomic Regions. The Reference Sequence project (Chapter 18; Refs. 8, 9) provides reviewed annotated sequences for a number of genomic regions that are difficult to annotate correctly by automated processes (e.g., immunoglobulin gene regions). These Reference Sequences (RefSeqs) are aligned to the assembled genome so that the curated annotation can be transferred to the assembled genomic sequence. RefSeqs for known pseudogenes are also aligned, not only to enable transfer of the correct annotation but, more importantly, to prevent prediction of erroneous model transcripts and proteins.

BAC End Sequences. Sequences from the ends of human genomic inserts in Bacterial Artificial Chromosome (BAC) clones are used to help map the location of specific clones onto the assembled genome sequence. The BAC end sequences are obtained from dbGSS (see Chapter 1). The clone names are extracted and converted to a standardized format to facilitate linking of the BAC end sequences with mapping data and additional sequences for the same clone, when these are available.

Genomic Sequences Used for Alignment

Human genomic sequences that are not used for either assembly or annotation are processed so that their relationships to the assembled genome can be displayed in Map Viewer (Chapter 20). Most genomic sequences deposited in GenBank by individual scientists will not be HTG and therefore will not be used for assembly; however, they are used for alignment. The exceptions are a few non-HTG sequences that are used for assembly, because they are included in the clone tiling path of an individual chromosome or in an assembled block of finished sequence. Any sequences intended for assembly but not used, either because they are redundant or are rejected by one of the quality screens in the assembly process, are also used for alignment.

Human Transcript Sequences

Human transcript sequences are used to help order and orient genomic fragments in the assembly step, for feature annotation and also to produce maps that show the locations of the transcripts on the assembled genome. Transcripts used include: (*a*) human mRNA RefSeqs (8, 9), except model transcripts produced from previous rounds of genome annotation; (*b*) human mRNA sequences deposited in GenBank by individual scientists, except those mRNAs produced after a translocation or other rearrangement of the genome; and (*c*) a nonredundant set of EST sequences from the BLAST FTP site. Additional information relating to these EST sequences is obtained from UniGene (Chapter 21).

Transcripts from Other Organisms

Transcripts from other organisms may be aligned to the genome being processed. These data may reveal the location of potential genes not identified by other means. RefSeq mRNAs, GenBank mRNAs, and ESTs, obtained from the same sources that provide the human transcripts, are used. Their alignments are processed for display in Map Viewer but are not used in the assembly step or for feature annotation.

Sequence Tagged Site (STS) Maps

Genetic linkage maps, radiation hybrid (RH) maps, and a YAC map are used to help avoid assembling genomic contigs incorrectly and to help place the contigs along the chromosomes. The positions of STS markers on various maps (Table 1) are transformed into a common format that allows us to compare the maps to each other during the assembly process. Additional maps are processed so that they can be displayed in Map Viewer.

The maps listed in Table 1 are static and are not updated with additional markers. Any new STS maps are added to our data set soon after they are released.

Table 1. STS maps used for assembly or display.

Map type	Map	Contig assembly	Contig placement	Display
Genetic linkage	Genethon (20)	X	X	X
Genetic linkage	Marshfield (21)	X	X	X
Genetic linkage	Decode (22)			X
Radiation hybrid	GeneMap99-G3 (23, 24)	X	X	X
Radiation hybrid	GeneMap99-GB4 (23, 24)	X	X	X
Radiation hybrid	NCBI RH (25)			X
Radiation hybrid	Stanford G3	X	X	X
Radiation hybrid	Stanford TNG (26)		X	X
Radiation hybrid	Whitehead-RH	X	X	X
YAC	Whitehead-YAC	X	X	X

Special Cases

Our own review of previous genome assemblies or feedback from users sometimes identify particular cases in which bad data or overlooked data prevent the automated processes from producing the best possible assembly of a particular segment of the genome. To help guide the assembly process, a list of such special cases is maintained. The list is used to provide supplemental data that override the automatic processes that assign a particular input genomic sequence to a chromosome or determine whether it is used for assembly.

Preparation of the Input Sequences

The raw input genomic sequences are screened for contaminants, the repetitive sequences are masked, and the draft genomic sequences are split into fragments in preparation for alignment to other sequences. The input transcript sequences are also screened for contaminants before they are aligned to the genomic sequences. The STS content of the input genomic sequences is determined.

Preparation of Genomic Sequences

Removing Contaminants

Draft-quality HTGSs sometimes contain segments of sequence derived from foreign sources, most commonly the cloning vector or bacterial host. Finished sequences are usually, but not always, free of such contaminants. Common contaminants introduce artificial blocks of homologous sequence that can give rise to misleading alignments between two unrelated genomic sequences. MegaBLAST (10) is used to compare the raw genomic sequences to a database of contaminant sequences (including the UniVec database of vector sequences, the *Escherichia coli* genome, bacterial insertion sequences,

and bacteriophage). Any foreign segments are removed from draft-quality sequence or masked in finished sequence to prevent them from participating in alignments.

Masking of Repetitive Sequences

Sequences that occur in many copies in the genome will align to many different clones. Such repetitive sequences include interspersed repeats (SINEs, LINEs, LTR elements, and DNA transposons), satellite sequences, and low-complexity sequences (7, 11, 12). Matches between repetitive sequences on unrelated clones make it difficult to identify alignments that indicate a genuine overlap between clones. To eliminate the confounding matches that are based only on repetitive sequences, the genomic sequences are run through RepeatMasker to identify known repeats. Repeats are masked by converting the sequence to lowercase letters so that they do not initiate alignments.

Fragmentation of Draft Sequences

Draft HTGSs consist of a set of sequence contigs derived from a particular clone artificially linked together to form a single sequence. The masked, draft genomic sequences are split at the gaps between their constituent contigs to create separate sequence fragments that can be aligned independently. Vector sequences and other contaminants are also removed at this stage by trimming or further splitting the sequence fragments.

Determination of STS Content

Any STS markers contained within the input genomic sequences are identified by e-PCR (13) using the UniSTS database. The resulting data are used primarily to relate the genomic sequences to independently derived STS maps (genetic, radiation hybrid) but are also used to identify some foreign sequences.

Filtering

Sequences from other clones being sequenced at the same institution can occasionally cross-contaminate draft HTGSs. The contaminating sequences may come from another clone from the same organism or from another organism. The raw genomic sequences are screened in several ways to detect cross-contamination: (*a*) they are compared with the genome sequences from completely sequenced organisms using MegaBLAST (10); (*b*) they are screened for the presence of organism-specific interspersed repeats using RepeatMasker; and (*3*) they are screened for the presence of mapped STS markers from other organisms using e-PCR (13). Any input sequence that contains foreign sequences, repeats, or markers is flagged for removal from the data set used for assembly. Draft sequences longer than the maximum insert length expected for a genomic clone are also rejected because it is likely they are contaminated with sequences from at least one other clone.

At this stage, draft sequences composed of fragments that are too small to contribute significantly to the assembly or that are tagged with the HTGS_CANCELLED keyword

are also flagged for removal. Another filter rejects sequences annotated as being from another organism or as being RNA, erroneously included in the input sequences.

Chromosome Assignment

To improve assembly of the genomic sequences, the input genomic sequences are assigned to a specific chromosome before attempting to merge the sequences. Genomic sequences that appear on any of the chromosome tiling paths are automatically assigned to the designated chromosome. Other genomic sequences are assigned to a chromosome based on: (*a*) annotation on the submitted GenBank record; (*b*) the presence of multiple STS markers that have been mapped to the same chromosome; (*c*) fluorescence *in situ* hybridization (FISH) mapping (14, 15); or (*d*) personal communication from a scientist with specialized knowledge. If there is no assignment, or the assignments are conflicted, the sequences are treated as unassigned and assembled without constraint by chromosome.

Filtering of Transcript Sequences

Transcript sequences that contain sequences derived from vectors or other common contaminants can produce artificial alignments to the assembled genomic sequence. The input transcript sequences are therefore compared with a database of contaminants using MegaBLAST (10), as described for genomic sequences. Any transcripts with significant matches to the sequence of a contaminant are excluded from the set of transcript sequences used for genome assembly or annotation.

mRNA sequences shorter than 300 bases are excluded from the set of sequences that are aligned to the genomic sequences because they are too small to contribute significantly to genome assembly or annotation. Also excluded are any mRNA sequences flagged because they do not represent the true sequence of a transcript, e.g., those that are chimeric or contain genomic sequences.

Alignment of Sequences to the Input Genomic Sequences

Alignment of the input genomic sequences to each other and to various other sequences is essential for both genome assembly and genome annotation. All relevant sequences are initially aligned to the unassembled genomic sequences because this means that the computationally intensive alignment processes can be run incrementally at an early stage in the pipeline. If necessary, these alignments are remapped to the sequence of the assembled genome at a later stage by a process that requires relatively little computation.

Alignment of Genomic Sequences to Each Other

Assembly of the genomic sequences from individual clones into longer contiguous sequences (contigs) requires knowledge of which sequences overlap. The overlaps between genomic sequences are evaluated by aligning the sequences from individual genomic clones to each other. After masking of repeats, decontamination and fragmentation, each

fragment of genomic sequence is aligned pairwise to all of the other fragments using MegaBLAST (10). Alignments that are sufficiently long and of sufficiently high percentage identity are saved for consideration in the assembly step.

Alignment of Clone End Sequences to the Genomic Sequences

The pairs of short genomic sequences derived from the ends of plasmid clones help to order and orient sequence fragments in the assembly step. These clone end sequences are aligned to the processed genomic sequences, as described for *Alignment of Genomic Sequences to Each Other*.

Alignment of Transcripts to the Genomic Sequences

Annotation of genes requires knowledge of where the sequences for known transcripts align to the assembled genomic sequences. RefSeq RNA sequences, mRNA sequences from GenBank, and EST sequences from dbEST are aligned to the processed genomic sequences, as described for *Alignment of Genomic Sequences to Each Other*. Later, the alignments are remapped to the assembled genomic sequence.

Alignment of Curated Genomic Regions to the Genomic Sequences

Curated genomic regions provide accurate annotation for regions of the genome that are difficult to annotate correctly by automated processes. Sequences from curated genomic regions are initially aligned to the unassembled genomic sequences and later remapped to the sequence of the assembled genome, as described for *Alignment of Transcripts to the Genomic Sequences*.

Alignment of Translated Genomic Sequences to Proteins

Homologies between the polypeptides encoded by the genomic sequences and known proteins/conserved protein domains provide hints for the gene prediction process. The repeat-masked genomic sequences are compared with a non-redundant database of vertebrate proteins and to the NCBI Conserved Domain Database (CDD; Ref. 16) using different versions of BLAST (4) (BLASTX and RPS-BLAST, respectively). Significant alignments are saved for use in the gene prediction step.

Genome Assembly

The input genomic sequences are assembled into a series of genomic sequence contigs. These are then ordered, oriented with respect to each other, and placed along each chromosome with appropriately sized gaps inserted between adjacent contigs. The resulting genome assembly thus consists of a set of genomic sequence contigs and a specification for how to arrange the sequence contigs along each chromosome.

Finished Chromosomes

A chromosome sequence is considered finished when any gaps that remain cannot be closed using current cloning and sequencing technology. In practice, therefore, the sequence for a finished chromosome usually consists of a small number of genomic sequence contigs. These are assembled from their component clone sequences according to the specification provided by the center responsible for sequencing that chromosome. This specification also prescribes the order, orientation, and estimated sizes for the gaps between contigs.

Unfinished Chromosomes

Genomic sequence contigs for unfinished chromosomes are assembled and laid out based largely on the clone tiling path. However, the tiling paths do not specify the orientation of the clone sequences or how they should be joined; therefore, data on the alignment of the input genomic sequences to each other and to other sequences are also used to guide the assembly. Genomic sequences that augment the initial set of genomic contigs based on the tiling path clones are also incorporated.

Resolution of Conflicts in the Chromosome Tiling Paths

Before the tiling paths are used in the genome assembly, the order of the finished clone sequences included in the tiling paths is compared with the specifications used to assemble the curated contigs of finished sequence. Discrepancies are resolved before proceeding with the assembly. Sequence from any clone should appear at just one place in the assembled genome; therefore, if a clone is listed more than once in the tiling paths, only the location with the best evidence is used in the assembly step.

Genomic Sequences Excluded from the Assembly Step

Clone sequences that consist only of unassembled reads (HTGS_PHASE0) or that were flagged because of suspected cross-contamination or other problem detected in the pre-assembly screens are not used in the assembly step.

Assembly of the Genomic Sequence Contigs

Adjacent, finished clone sequences from the chromosome tiling path that have good sequence overlap are merged. Tiling path draft sequences that are adjacent to and overlap the finished clone sequences or other draft clone sequences are added to extend the initial genomic sequence contigs. After that, genomic sequences from clones not on any chromosome tiling path are added, provided they have good overlaps with the assembled tiling path clones. Genomic sequences from additional clones may be added if they provide the sequence for a known gene that is missing from the existing genomic sequence contigs. Finally, the individual fragments of draft sequences are ordered and oriented.

Assembly of Finished Sequences from Tiling Path Clones. The quality of any overlaps between finished clone sequences that are adjacent in the clone tiling path are assessed using the alignments between pairs of genomic sequences that were produced in advance. Sequences that have high-quality overlaps, or that are known from annotation or other data to abut, are merged to form a genomic sequence contig. Clone sequences that have no good overlaps are retained as separate contigs.

Addition of Draft Sequences from Tiling Path Clones. The procedure used for merging draft sequence from tiling path clones is similar to that described for merging finished sequences, except that the minimum-overlap quality required for merging is different. An overlap involving a draft sequence can contain more mismatches, but must be longer, than an overlap between two finished sequences. Preference is given to finished sequences so that a contig made by merging finished and draft sequences will contain the finished sequence for the overlapping portion. Draft clone sequences that have no good overlaps are retained as separate contigs.

Addition of Sequences from Other Genomic Clones. Genomic sequences from clones that are not on any chromosome tiling path are used to close, or extend into, gaps in the backbone of genomic contigs assembled based on the tiling paths. Genomic sequences that are fully contained within the existing genomic contigs are not used. Any remaining genomic sequences that were either assigned to the relevant chromosome or could not be assigned to any chromosome are evaluated, and sequences that have good-quality overlaps with genomic contigs are merged in to extend a contig or to join two adjacent contigs, if two additional conditions are met: (1) the gap must be sufficiently large to accommodate the additional sequence; and (2) the Sequence Tagged Site (STS) marker content of the additional clone sequence must be compatible with that of the flanking clone sequences when compared with various STS maps.

After all of the chromosomes have been assembled, any remaining genomic clone sequence that contains a known gene not present in the other contigs is added to the assembly as a separate contig.

Ordering and Orienting Draft Sequence Fragments. The order and orientation of the fragments of HTGS_PHASE1 draft sequence need to be defined before sequence made from contigs that include this category of draft sequence can be completed. Some fragments may be ordered and oriented by overlaps with sequences from adjacent clones. Many more can be defined by aligning them with mRNAs, ESTs, or plasmid paired-end sequences. Any fragments whose order and orientation remain undefined are placed in the nearest open gap and given an arbitrary orientation. Fragments of draft sequence are connected to flanking sequences and to each other by runs of 100 unknown bases (Ns), which represent an arbitrarily sized gap in the sequence.

Placement of the Genomic Contigs

After the genomic sequence contigs are assembled, they are oriented and placed in order along each chromosome with appropriately sized gaps inserted between adjacent contigs.

The chromosome tiling paths specify the order of the clone sequences and the sizes for some gaps. Therefore, the order and orientation of most of the genomic sequence contigs are derived from the tiling paths. Many of the remaining contigs are placed by comparison of the STS marker content of the contigs, as determined by e-PCR (13) using the UniSTS database, to various STS maps. There are some contigs that can be assigned to a specific chromosome but cannot be placed along that chromosome. Others cannot even be assigned to a specific chromosome and therefore remain unplaced within the genome assembly.

Gaps between the clone contigs laid out in the chromosome tiling paths are arbitrarily set at 50 Kbp, and 3 Mbp for the centromere, unless another gap size is specified in the tiling path. Any remaining gaps between genomic sequence contigs are arbitrarily set to 10 Kbp.

Preparing a Provisional Genome Assembly

A set of sequences and data files is produced to represent the provisional assembly. This set includes: (*a*) sequences for each genomic contig in FASTA format; (*b*) specifications describing how to assemble each genomic contig from its components; and (*c*) a specification for how to arrange the contigs along each chromosome. A raw RefSeq entry is also made for each genomic sequence contig.

Quality Control

The provisional assembly is checked for consistency with the chromosome tiling paths and various STS maps. The order in which the component clone sequences appear in the assembled chromosomes is compared with their order in the tiling paths on which the assembly was based. The STS marker order along each chromosome in the provisional assembly, as determined by e-PCR (13) using the UniSTS database, is checked for consistency with a set of STS maps. Haussler et al. at the University of California at Santa Cruz (UCSC) also perform a set of independent quality checks on the provisional assembly. In addition to comparing the assembly to the chromosome tiling paths and to various STS maps, they also look for potentially misassembled contigs using alignments of BAC end sequences. Any serious errors in the assembly may be corrected by repeating the assembly steps using different parameters or by manually editing the assembly.

Annotation of Genes

Identification of genes within the genome assembly reveals the functional significance of particular stretches of genomic sequence. Genes are found using three complementary approaches: (*a*) known genes are placed primarily by aligning mRNAs to the assembled genomic contigs; (*b*) additional genes are located based on alignment of ESTs to the assembled genomic contigs; and (*c*) previously unknown genes are predicted using hints provided by protein homologies. Whenever possible, predicted genes are identified by homology between the protein they encode and known protein sequences.

Generation of Transcript-based Gene Models

Alignments between known transcripts and the assembled genomic sequences are processed to produce gene models. Each gene model consists of an ordered series of exons. The transcripts defining each gene model are used as evidence to support that model.

Alignment of Transcripts to the Assembled Genome

The alignments between RefSeq RNA sequences, mRNA and EST sequences from GenBank and the component genomic sequences are remapped to produce alignments of these transcripts to the assembled genomic contigs.

Production of Candidate Gene Models

A candidate gene model is produced from each set of alignments between a particular transcript and one strand of a particular genomic contig as follows: (1) putative exons are identified by looking for mRNA splice sites near the ends of those alignments that satisfy minimum length and percentage identity criteria; (2) a mutually compatible set of exons for the model is selected by applying rules, such as restrictions on the size of an intron, that define plausible exon–intron structures; and (3) BLAST (4) may be used to produce additional alignments to try to identify exons that were missed because they were too short to be represented in the initial set of transcript alignments. Candidate gene models are only retained if good-quality alignments between their exons and the defining transcript cover either more than half the length of the transcript or more than 1 Kbp.

Selection of the Best RefSeq RNA-based Gene Models

Each RefSeq RNA represents a distinct transcript produced from a particular gene (Chapter 18; Refs. 8, 9). Hence, there should not be more than one gene model corresponding to any given RefSeq RNA. Therefore, all gene models based on a particular RefSeq RNA are compared, and the best one is selected. Because the RefSeq RNA is taken to be the best representation of a particular transcript, this gene model is preserved without any further modification. Any extra models may represent paralogs; therefore, they are included with the mRNA- and EST-based models for further processing. Between builds, RefSeq RNAs are refined based on a review of related gene models and transcript alignments produced during the genome annotation process.

Exon Refinement

Many gene models may be produced for the same gene because the input data set frequently contains multiple EST or mRNA sequences representing the same transcript. This redundancy is used to refine the splice sites defining a particular exon. Similar exons are clustered, and splice sites may be adjusted in some models to match those used by the majority of models containing the same exon. Inconsistent models may be discarded at this stage, unless they have sufficient support to be retained as likely splice variants.

Chaining of Transcript-based Gene Models

Many of the mRNAs and most of the ESTs used to generate the initial gene models provide sequence for only part of the native transcript. Overlapping gene models that are compatible with each other are combined into an extended model. This chaining step produces models more likely to represent the full gene.

Ab Initio Gene Prediction

GenomeScan, an *ab initio* gene prediction program, is used to provide models for genes inferred from the genomic sequence using hints provided by protein homologies (17). The genes predicted by GenomeScan are combined with the transcript-based gene models, but they are also retained as a distinct set of models that can be viewed or searched separately.

Dividing the Genomic Sequences into Segments

GenomeScan produces better results when long genomic sequences are broken into shorter segments at putative gene boundaries. The locations of gene models based on RefSeq RNA alignments are, therefore, used to divide the assembled genomic contigs into segments. Repetitive sequences are masked by remapping the repeats found in the component genomic sequences.

Producing Protein Hints for GenomeScan

GenomeScan can use data on protein homologies to improve its gene predictions (17). The locations of genomic sequences that potentially code for polypeptides with homology to other proteins are obtained from three sources. Significant alignments between translated genomic segments and vertebrate proteins are obtained by filtering and remapping the precomputed alignments. Significant alignments between translated genomic sequences and conserved protein domains are obtained in the same manner. A third set of alignments comes from running GenomeScan without any hints. The proteins predicted by this initial run are aligned to proteins from SWISS-PROT (18) and NCBI RefSeq proteins (8, 9) using blastp (4). The eukaryotic protein with the best match is then aligned to the genomic sequence segments using tblastn (4). These three sets of data are converted into the format required by GenomeScan and merged to produce a single set of protein hints.

Predicting Genes Using GenomeScan

Each segment of genomic sequence is processed by GenomeScan using the combined set of protein homology-based hints as an additional input. This produces one model containing all of the predicted exons for each putative gene. Models with coding sequences shorter than 90 amino acids are discarded. Each remaining model is aligned to proteins from SWISS-PROT and NCBI RefSeq proteins using blastp. The eukaryotic protein with the best match to any model is used as evidence for that model and to provide a clue as to the possible function of that model.

Consolidation of Gene Models

Consolidation of the transcript-based gene models and the predicted gene models forms a single set of models. Models are clustered into genes if they share one or more exons or if Entrez Gene (Chapter 19; Refs. 8, 9) indicates that the transcripts used as evidence for the models come from the same gene. If a model is entirely contained within a longer model, it is redundant and, therefore, eliminated. Sets of identical models are reduced to a single representative model linked to all of the supporting evidence. For sets of very similar models, a single model is picked as a representative, giving preference to models based on RefSeq RNAs or on GenBank mRNAs. Predicted gene models that significantly overlap transcript-based models but that are not sufficiently similar to consolidate are discarded.

Pruning of Gene Models

Some gene models are discarded because: superior gene annotation is available from a curated genomic region, they are likely to represent pseudogenes, or they are incompatible with other gene models.

Gene Models Superceded by Curated Genomic Regions

The manually reviewed annotations from curated genomic region RefSeqs are used in preference to any corresponding gene models generated by automated processing. The curated genomic regions are aligned to the assembled genomic contigs by remapping the alignments between these RefSeqs and the component genomic sequences. Any gene model that significantly overlaps a segment of the assembled sequence that corresponds to a curated genomic region is discarded.

Gene Models Likely to Be Pseudogenes

When transcripts from a particular gene are aligned to the genomic sequences, they will align not only to the active copy of the gene but also to any segment of the genome containing a pseudogene derived from the active gene. Because model transcripts or model proteins that represent nontranscribed pseudogenes are undesirable, an attempt is made to identify and remove such models.

Whenever possible, alignments of RefSeqs for pseudogenes, either curated genomic regions or RNAs, are used to annotate pseudogenes. Some additional models derived from pseudogenes that are not yet represented by RefSeqs are eliminated by the following mechanism. All models based on the same supporting mRNA are compared with respect to the percent identity of the alignments and the number of exons. Only the model with the strongest evidence is retained.

Conflicting Gene Models

When two gene models are found to have an extensive overlap, then in general only the model with the stronger evidence is retained. However, models based on RefSeqs are

always retained. Whereas any model not based on a RefSeq is discarded if it overlaps a model that is RefSeq based, two RefSeq-based models that overlap are both retained.

Location of Model Coding Regions

Initially, the longest open reading frame from each gene model is annotated as the protein coding sequence. This annotation can be revised if evidence associated with that model provides support for an alternative coding region. The protein coding sequence from any transcript used as evidence for a gene model is compared with the longest open reading frame in that model using BLAST (4). If the two do not match, the conflict is noted, and the annotation is revised if there is evidence to support an alternative coding region. For example, the coding sequence from the transcript evidence may indicate that an alternate translation start site is used, or that the model contains a premature termination codon. Models with coding regions less than 90 amino acids long are discarded, unless they are based on a RefSeq.

Relating Gene Models to Known Genes, Transcripts, and Proteins

The set of gene models produced by the preceding steps is a mixture of models for predicted genes and for known genes. To help identify models representing known genes, the model transcripts are compared with known transcripts. To help name the predicted genes, the proteins encoded by the models are also compared with known proteins.

Relating the Model Transcripts to Known Transcripts

To provide continuity from build to build and to identify genes based on their predicted transcripts, MegaBLAST (10) is used to compare model RNAs to: (*a*) RefSeq RNAs; (*b*) mRNAs from GenBank; and (*c*) model RNAs from the previous build. These comparisons are reported as reciprocal best hits if: (*a*) they produce a significant hit; (*b*) no other model has a better hit to that particular RNA; and (*c*) no other RNA has a better hit to that particular model.

Relating the Model Proteins to Known Proteins

The eukaryotic proteins with the best match to each protein predicted by the annotation process are used to identify the best model for a possible gene and to assign a name to gene models that are novel. The proteins encoded by the models are aligned to proteins from SWISS-PROT (18), NCBI RefSeq proteins (8, 9), and the NCBI non-redundant protein database using blastp (4). The name of the eukaryotic protein with the best match, its sequence identifier, and match score are recorded for each predicted protein with a significant hit.

Assigning Gene Identifiers to Models

Gene models are attributed to known genes whenever the correspondence is clear. If a model RNA has a reciprocal best hit with a known RNA, then the annotation of the known RNA is used to identify the gene. The first models to be assigned to genes are those

that have reciprocal best hits with RefSeq RNAs. This is followed by assignment of those models that have reciprocal best hits to models from the previous build or to GenBank mRNAs. Gene data for models that match a mRNA not yet represented by a RefSeq are obtained from NCBI gene-specific databases (currently Entrez Gene, Chapter 19). If the mRNA is associated with an entry in one of these databases, then the information attached to that gene record (e.g., symbols, names, and database cross-references) is used in the annotation. If the correspondence with known genes is ambiguous, as may occur if there are undocumented paralogs, then an interim gene identifier is assigned.

Selection of Transcript Models to Represent Each Gene

Multiple models based on alternative transcripts for some genes may be produced. In most of these cases, one transcript model is selected to represent the product of the gene for annotation purposes. Any homology between eukaryotic proteins and proteins encoded by the models guides the choice between alternative models. Multiple transcripts are annotated only if the models are based on RefSeq mRNAs representing alternative transcripts from the same gene.

Although alternative transcript models are not annotated, the alignments between the transcripts that represent alternative splicing and genomic contigs are processed for display in Map Viewer, Evidence Viewer, and Model Maker (see Chapter 20).

Naming of Gene Products

The transcripts and protein products of any models that have been assigned to a known gene are given the product names that appear in the LocusLink entry for that gene. The gene products from other genes are named based on any significant homology to other eukaryotic proteins, provided that the matching protein has a meaningful name (i.e., names such as “Hypothetical...” are ignored).

Annotation of the Assembled Genomic Contigs

The genomic contig RefSeqs are annotated with features that provide information about the location of genes, mRNAs, and coding regions. Features from curated genomic region RefSeqs are copied to the contigs based on the alignment between the curated sequence and the corresponding contig. Protein domains from the Conserved Domain Database (CDD; Ref. 16) are identified using reverse position-specific BLAST (RPS-BLAST; Ref. 4), and their locations are annotated. A description of the evidence supporting those RNAs and proteins that are not curated RefSeqs, i.e., those that are models, is also recorded.

Annotation of Other Features

Reference sequences produced by the genome assembly process are annotated with features that provide landmarks valuable for making connections between maps based on different coordinate systems and for associating genes with diseases.

Annotation of STSs

Placement of STSs on the genome assembly allows sequence-based data to be integrated with non-sequence-based maps that contain STS markers, such as genetic and radiation hybrid maps. STSs are identified by using e-PCR (13) to find sequences that match the STS primer pairs from UniSTS, the spacing of which is consistent with the reported PCR product size. The number of times that each STS appears in the assembled genome is recorded so that only those STSs that appear at only one or two locations in the assembled genome are annotated.

Annotation of Clones

Placement on the genome assembly of clones that have been mapped to cytogenetic bands by FISH provides the means to determine the correspondence between the sequence and cytogenetic coordinate systems (14, 15). Knowing this correspondence allows the integration of sequence-based data with cytogenetic data. For human, only those clones mapped by fluorescence *in situ* hybridization (FISH) by the human BAC Resource Consortium (see the Human BAC Resource) are annotated. Clones are placed using three types of sequence tags. Clones that have sequence for the genomic insert, either draft or finished, with a GenBank Accession number are localized by remapping the alignment between the clone sequence and other genomic clones to the assembled genomic contigs. Similarly, clones that have BAC end sequences are localized by remapping the alignment between the BAC end sequences and genomic clone sequences to the assembled genomic contigs. Clones that have STS markers confirmed by PCR or hybridization experiments are mapped using the locations in the assembled contigs of STS markers that were identified by e-PCR. The number of places that each clone appears in the assembled genome is recorded so that only those clones that either have a unique placement in the assembled genome or are placed twice on the same chromosome are annotated.

Annotation of Sequence Variation

Placement of Single Nucleotide Polymorphisms (SNPs) and other variations on the genome provides numerous landmarks that are valuable for associating genes with diseases (Chapter 5). Variations from dbSNP (19) are placed in their genomic contexts using the sequences that flank the variation. Flanking sequences are first run through RepeatMasker to mask repetitive sequences and then aligned to the assembled genomic sequence contigs using MegaBLAST (10). The resulting matches are classified as either high or low confidence, depending on the quality of the alignment, and the number of matches for each SNP is recorded so that only those SNPs that map to one or two locations in the assembled genome are annotated.

Product Data Sets

The products of our assembly and annotation process are made available to the public as RefSeqs of assembled chromosome sequences, genomic sequence contigs, model

transcripts, and model proteins. RefSeqs are produced in alternative formats so that they can be retrieved by Entrez, BLAST, or FTP.

RefSeqs

A fully annotated Refseq entry is made for each genomic sequence contig. Separate RefSeq model RNA and protein entries are also made for any of the transcripts and coding regions annotated on genomic contigs not identified as existing RefSeqs. Finally, a RefSeq entry is made for each chromosome by combining the annotated sequences of the genomic contigs in the appropriate order and with the appropriate spacing. The resulting RefSeqs can be retrieved through Entrez.

BLAST Databases

The assembled genomic contig RefSeqs are formatted as a BLAST database (Chapter 16). Separate BLAST databases are also produced from the set of transcripts and the set of proteins annotated on the assembled genome. These databases include both known and model RefSeqs. In addition, separate BLAST databases are produced from the complete sets of transcripts and proteins predicted by GenomeScan.

Data Files for FTP

The annotated genomic sequence contig, model transcript, and model protein RefSeqs are saved in GenBank flatfile and ASN.1 formats. The same sets of sequences that are used to make BLAST databases are also saved in FASTA format. All of these data files, together with files that specify the construction of the genomic contigs and their arrangement along the chromosomes, are made available for download by FTP.

Production of Maps That Display Genome Features

We produce many maps showing the locations of various features annotated on our genome assembly. Maps containing whatever combination of features that interests the user can be selected and displayed side-by-side using Map Viewer (Chapter 20). Detailed descriptions of the maps available for each genome are available in the relevant Genome Map Viewer help document.

Preparation of Map Data

Basic map data are prepared for each map to identify each feature, delineate its position on the chromosome, and specify how it is to be displayed. For many maps, supplemental data are prepared to provide more information about each feature. Map Viewer displays this map-specific supplemental information when users select a particular map as the Master Map (Chapter 20).

Maps Based on Sequence Coordinates

Maps that display those features annotated on the genomic sequence contigs (genes, STSs, clones, and SNPs) are generated by translating the positions of the features on the contigs into chromosome coordinates. Contig coordinates are translated into chromosome coordinates using the positions of the contigs along each chromosome, as determined in the genome assembly step. Using this same method, alignments between various sequences and the genomic contigs are translated into chromosome coordinates to produce additional maps that show the locations of the aligned sequences on the chromosomes. Maps generated from sequence alignments include maps that show the genomic positions of mRNA plus EST sequences, or genomic sequences from GenBank. The specifications used to build each genomic contig are also translated into chromosome coordinates to produce one map that shows the component sequences used to assemble each contig and another that simply shows the finished and draft sections of the contigs.

Maps Based on Other Coordinate Systems

Cytogenetic maps, genetic linkage maps, and radiation hybrid maps use different coordinate systems that are not based on sequence. To generate data for these types of maps, the locations of the map elements are listed in the coordinate system appropriate to each map. Map Viewer can scale maps defined in different coordinate systems so that they can de displayed side by side.

Making the Map Data Available for Use

All of the map data for the new genome assembly are loaded into the Map Viewer database. Next, the objects in the new maps are indexed so that users can search for and then display specific features (Chapter 20). The data from the Map Viewer database are exported to produce a set of map data files that is made available via FTP.

Public Release of Assembly and Models

To ensure that a consistent view of the annotated genome assembly is presented, the release of databases and FTP files is coordinated. When everything is ready for release, the Map Viewer display is switched to the new build, the BLAST databases are swapped, and the files on the FTP site are replaced. Several associated databases are then refreshed, including LocusLink, dbSNP, and UniSTS, so that the data they contain reflect the new build. Finally, the Web pages that provide statistics for the build and record changes to the genome assembly and annotation process are updated.

Integration with Other Resources

The products of the genome assembly and annotation process are linked extensively to various NCBI resources. These links provide different views of the data and more information for researchers as they follow a particular line of investigation.

Links between Map Viewer and Other Resources

The maps displayed by Map Viewer have embedded links between map objects and relevant NCBI resources (Table 2). Many of these resources also have reciprocal links back to Map Viewer, allowing, for example, a gene in LocusLink to be displayed in its genomic context.

Table 2. Links from Map Viewer objects to other NCBI resources.

Map object	Linked NCBI resource	Resource description
Accession number	Entrez	Chapter 15
Clone	Clone Registry	http://www.ncbi.nlm.nih.gov/genome/clone/
Disease gene	OMIM	Chapter 7
EST or mRNA	UniGene	Chapter 21
Gene	LocusLink	Chapter 19
Gene or transcript	Evidence Viewer	Chapter 20
Gene or transcript	Model Maker	Chapter 20
Gene	Human–mouse homology map	http://www.ncbi.nlm.nih.gov/Homology/
STS	UniSTS	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unists
Variation (SNP)	dbSNP	Chapter 5

Links between Reference Sequences and Other Resources

During the production of RefSeqs, links between the annotated features (clones, genes, SNPs, and STSs) and the relevant resources listed in Table 2 are created. Links are also made between the genomic contig RefSeqs and the RefSeqs for the model transcripts and proteins that they encode.

Integration with BLAST

A customized BLAST Web page allows the comparison of any sequence to a BLAST database of model transcript, model protein, or genomic contig RefSeqs. Users can choose to view any hits that result from such a search on a diagram showing the chromosomal location of the hits, with each hit linked to a Map Viewer display of the region encompassing the sequence alignment.

Contributors

Richa Agarwala, Jonathan Baker, Hsiu-Chuan Chen, Vyacheslav Chetvernin, Deanna Church, Cliff Clausen, Dmitry Dernovoy, Olga Ermolaeva, Wratko Hlavina, Wonhee Jang, Philip Johnson, Jonathan Kans, Paul Kitts, Alex Lash, David Lipman, Donna Maglott, Jim Ostell, Keith Oxenrider, Kim Pruitt, Sergei Resenchuk, Victor Sapojnikov, Greg Schuler,

Steve Sherry, Andrei Shkeda, Alexandre Souvorov, Tugba Suzek, Tatiana Tatusova, Lukas Wagner, and Sarah Wheelan

References

1. Bently DR. Genomic sequence information should be released immediately and freely in the public domain. Science. 1996;274:533–534. PubMed PMID: 8928006.
2. Guyer M. Statement on the rapid release of genomic DNA sequence. Genome Res. 1998;8:413. PubMed PMID: 9582183.
3. Jang W, Chen HC, Sicotte H, Schuler GD. Making effective use of human genomic sequence data. Trends Genet. 1999;15:284–286. PubMed PMID: 10390628.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402. PubMed PMID: 9254694.
5. Zhao S, Malek J, Mahairas G, Fu L, Nierman W, Venter JC, Adams MD. Human BAC ends quality assessment and sequence analyses. Genomics. 2000;63:321–332. PubMed PMID: 10704280.
6. Mahairas GG, Wallace JC, Smith K, Swartzell S, Holzman T, Keller A, Shaker R, Furlong J, Young J, Zhao S, Adams MD, Hood L. Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. Proc Natl Acad Sci U S A. 1999;96:9739–9744. PubMed PMID: 10449764.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser

J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, Szustakowki J, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. PubMed PMID: 11237011.

8. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 2000;16:44–47. PubMed PMID: 10637631.
9. Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 2001;29:137–140. PubMed PMID: 11125071.
10. Zhang Z, Schwartz S, Wagner L, Miller W. A GREEDY algorithm for aligning DNA sequences. J Comput Biol. 2000;7:203–214. PubMed PMID: 10890397.
11. Jurka J. Repeats in genomic DNA: mining and meaning. Curr Opin Struct Biol. 1998;8:333–337. PubMed PMID: 9666329.
12. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 1999;9:657–663. PubMed PMID: 10607616.
13. Schuler GD. Sequence mapping by electronic PCR. Genome Res. 1997;7:541–550. PubMed PMID: 9149949.
14. Kirsch IR, Green ED, Yonescu R, Strausberg R, Carter N, Bentley D, Leversha MA, Dunham I, Braden VV, Hilgenfeld E, Schuler G, Lash AE, Shen GL, Martelli M, Kuehl WM, Klausner RD, Ried T. A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. Nat Genet. 2000;24:339–340. PubMed PMID: 10742091.
15. Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, Furey TS, Kim UJ, Kuo WL, Olivier M, Conroy J, Kasprzyk A, Massa H, Yonescu R, Sait S, Thoreen C, Snijders A, Lemyre E, Bailey JA, Bruzel A, Burrill WD, Clegg SM, Collins S, Dhami P, Friedman C, Han CS, Herrick S, Lee J, Ligon AH, Lowry S, Morley M, Narasimhan S, Osoegawa K, Peng Z, Plajzer-Frick I, Quade BJ, Scott D, Sirotkin K, Thorpe AA, Gray JW, Hudson J, Pinkel D, Ried T, Rowen L, Shen-Ong GL, Strausberg RL, Birney E, Callen DF, Cheng JF, Cox DR, Doggett NA, Carter NP, Eichler EE, Haussler D, Korenberg JR, Morton CC, Albertson D, Schuler G, de Jong PJ, Trask BJ. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature. 2001;409:953–958. PubMed PMID: 11237021.
16. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-

dimensional structure. Nucleic Acids Res. 2002;30:281–283. PubMed PMID: 11752315.
17. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. Genome Res. 2001;11:803–816. PubMed PMID: 11337476.
18. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res. 1998;26:38–42. PubMed PMID: 9399796.
19. Sherry ST, Ward M, Sirotkin K. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 1999;9:677–679. PubMed PMID: 10447503.
20. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature. 1996;380:152–154. PubMed PMID: 8600387.
21. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet. 1998;63:861–869. PubMed PMID: 9718341.
22. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K. A high-resolution recombination map of the human genome. Nat Genet. 2002;31:241–247. PubMed PMID: 12053178.
23. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ.et al. A gene map of the human genome. Science. 1996;274:540–546. PubMed PMID: 8849440.
24. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJ, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Bentley DR.et al. A physical map of 30,000 human genes. Science. 1998;282:744–746. PubMed PMID: 9784132.
25. Agarwala R, Applegate DL, Maglott D, Schuler GD, Schaffer AA. A fast and scalable radiation hybrid map construction and integration strategy. Genome Res. 2000;10:350–364. PubMed PMID: 10720576.
26. Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, Brady SD, Bushard JM, Bustos VI, Chu A, Chung TR, De Witte A, Denys ME, Dominguez R, Fang NY, Foster BD, Freudenberg RW, Hadley D, Hamilton LR, Jeffrey TJ, Kelly L, Lazzeroni L, Levy MR, Lewis SC, Liu X, Lopez FJ, Louie B, Marquis JP, Martinez RA, Matsuura MK, Misherghi NS, Norton JA, Olshen A, Perkins SM, Perou AJ, Piercy C, Piercy M, Qin F, Reif T, Sheppard K, Shokoohi V, Smick GA, Sun WL, Stewart EA, Fernando J, Tejeda, Tran NM, Trejo T, Vo NT, Yan SC, Zierten DL, Zhao S, Sachidanandam R, Trask BJ, Myers RM, Cox DR. A high-resolution radiation hybrid map of the human genome draft sequence. Science. 2001;291:1298–1302. PubMed PMID: 11181994.

Part 3. Querying and Linking the Data

Chapter 15. The Entrez Search and Retrieval System

Jim Ostell

Created: October 9, 2002; Updated: August 13, 2003.

Summary

Entrez is the text-based search and retrieval system used at NCBI for all of the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others. Entrez is at once an indexing and retrieval system, a collection of data from many sources, and an organizing principle for biomedical information. These general concepts are the focus of this chapter. Other chapters cover the details of a specific Entrez database (e.g., PubMed in Chapter 2) or a specific source of data (e.g., GenBank in Chapter 1).

Entrez Design Principles

History

The first version of Entrez was distributed by NCBI in 1991 on CD-ROM. At that time, it consisted of nucleotide sequences from GenBank and PDB; protein sequences from translated GenBank, PIR, SWISS-PROT, PDB, and PRF; and associated citations and abstracts from MEDLINE (now PubMed and referred to as PubMed below). We will use this first design to illustrate the principles behind Entrez.

Entrez Nodes Represent Data

An Entrez "node" is a collection of data that is grouped together and indexed together. It is usually referred to as an Entrez database. In the first version of Entrez, there were three nodes: published articles, nucleotide sequences, and protein sequences. Each node represents specific data objects of the same type, e.g., protein sequences, which are each given a unique ID (UID) within that logical Entrez Proteins node. Records in a node may come from a single source (e.g., all published articles are from PubMed) or many sources (e.g., proteins are from translated GenBank sequences, SWISS-PROT, or PIR) (Figure 1).

Note that the UID identifies a single, well-defined object (i.e., a particular protein sequence or PubMed citation). There may be other information about objects in nodes, such as protein names or EC numbers, that may be used as index terms to find the record, but these pieces of information are not the central organizing principle of the node. Each data object represents a stable, objective observation of data as much as possible, rather than interpretations of the data, which are subject to change or confusion over time or across disciplines. For example, barring experimental error, a particular mRNA sequence report is not likely to change over the years; however, the given name, position on the chromosome, or function of the protein product may well change as our knowledge

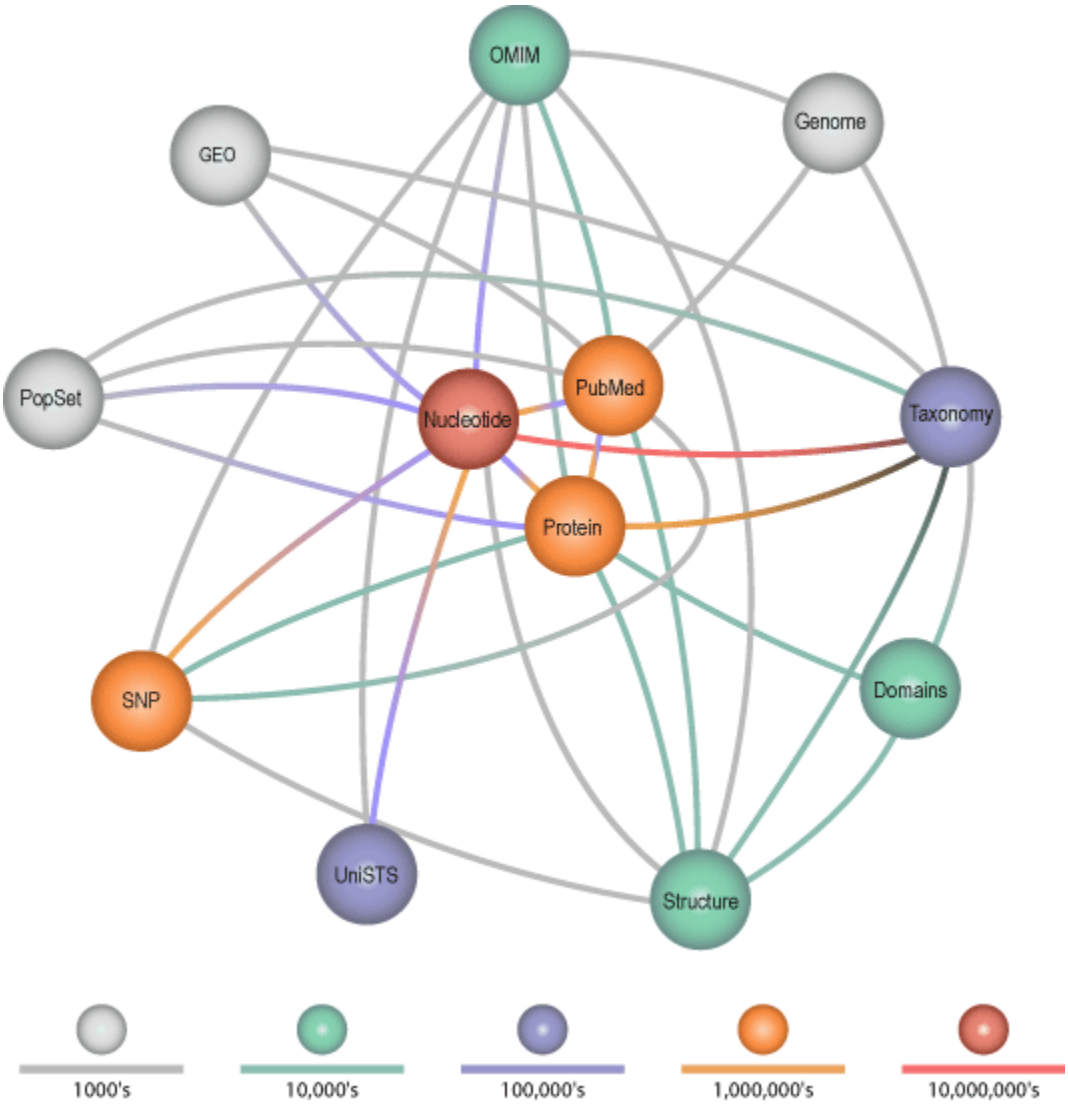


Figure 1. The original version of Entrez had just 3 nodes: nucleotides, proteins, and PubMed abstracts. Entrez has now grown to nearly 20 nodes.

develops. Even a published article is a stable observation. The fact that the article was published at a certain time and contained certain words will not change over time, although the importance of the article topic may change many times.

Entrez Nodes Are Intended for Linking

Another criterion for selecting a particular data type to be an Entrez node is to enable linking to other Entrez nodes in a useful and reliable way. For example, given a protein sequence, it is very useful to quickly find the nucleotide sequence that encodes it. Or given a research article, it is useful to find the sequences it describes, if any.

Links between Nodes

One way to achieve this is to put all of the information into one record. For example, many GenBank records contain pertinent article citations. However, PubMed also contains the article abstract and additional index terms (e.g., MeSH terms); furthermore, the bibliographic information is also more carefully curated than the citation in a GenBank entry. It therefore makes much more sense to search for articles in PubMed rather than in GenBank.

When a subset of articles has been retrieved from PubMed, it may be useful to link to sequence information associated with the abstracts. The article citation in the GenBank record can be used to establish the link to PubMed and, conversely, to make the reciprocal link from the PubMed article back to the GenBank record. Treating each Entrez node separately but enabling linking between related data in different nodes means that the retrieval characteristics for each node can be optimized for the characteristics and strengths of that node, whereas related data can be reached in nodes with different strengths.

This approach also means that new connections between data can be made. In the example above, the GenBank record cited the published article, but there was no link from that article in PubMed to the sequence until Entrez made the reciprocal link from PubMed. Now, when searching articles in PubMed, it is possible to find this sequence, although no PubMed records have been changed. Because of this design principle, the Entrez system is richly interconnected, although any particular association may originate from only one record in one node.

Links within Nodes

Another type of linking in Entrez is between records of the same type, often called "neighbors", in sequence and structure nodes. Most often these associations are computed at NCBI. For example, in Entrez Proteins, all of the protein sequences are "BLASTed" against each other, and the highest-scoring hits are stored as indexes within the node. This means that each protein record has associated with it a list of highly similar sequences, or neighbors.

Again, associations that may not be present in the original records can be made. For example, a well-annotated SWISS-PROT record for a particular protein may have fields that describe other protein or GenBank records from which it was derived. At a later date, a closely related protein may appear in GenBank that will not be referenced by the SWISS-PROT record. However, if a scientist finds an article in PubMed that has a link to the new GenBank record, that person can look at the protein and then use the BLAST-computed neighbors to find the SWISS-PROT record (as well as many others), although neither the SWISS-PROT record nor the new GenBank record refers to each other anywhere.

Entrez Nodes Are Intended for Computation

There are many advantages to establishing new associations by computational methods (as in the GenBank–SWISS-PROT example above), especially for large, rapidly changing data sets such as those in biomedicine.

As computers get faster and cheaper, this type of association can be made more efficiently. As data sets get bigger, the problem remains tractable or may even improve because of better statistics. If a new algorithm or approach is found to be an improvement, it is possible to apply it over the whole data set within a practical timescale and by using a reasonable number of resources. Any associations that require human curation, such as the application of controlled vocabularies, do not scale well with rapidly growing sets of data or evolving data interpretations. Although these manual kinds of approaches certainly add value, computational approaches can often produce good results more objectively and efficiently.

Entrez Is a Discovery System

A data-retrieval system succeeds when you can retrieve the same data you put in. A discovery system is intended to let you find more information than appears in the original data. By making links between selected nodes and making computed associations within the same node, Entrez is designed to infer relationships between different data that may suggest future experiments or assist in interpretation of the available information, although it may come from different sources.

The ability to compare genotype information across a huge range of organisms is a powerful tool for molecular biologists. For example, this technique was used in the discovery of a gene associated with hereditary nonpolyposis colon cancer (HNPCC). The tumor cells from most familial cases of HNPCC had altered, short, repeated DNA sequences, suggesting that DNA replication errors had occurred during tumor development. This information caused a group of investigators to look for human homologs of the well-characterized *Escherichia coli* DNA mismatch repair enzyme, MutS (1). Mutants in a *MutS* homolog in yeast, *MSH2*, showed expansion and contraction of dinucleotide repeats similar to the mutation found in the human tumor cells. By comparing the protein sequences between the yeast MSH2, the *E. coli* MutS, and a human gene product isolated and cloned from HNPCC colorectal tumor, the researchers could show that the amino acid sequences of all three proteins were very similar. From this, they inferred that the human gene, which they called *hMSH2*, may also play a role in repairing DNA, and that the mutation found in tumors negatively affects this function, leading to tumor development.

The researchers could connect the functional data about the yeast and bacterial genes with the genetic mapping and clinical phenotype information in humans. Entrez is designed to support this kind of process when the underlying data are available electronically. In PubMed, the research paper about the discovery of *hMSH2* (1) has links to the protein

sequence, which in turn has links to "neighbors" (related sequences). There are lots of records for this protein and its relatives in many organisms, but among them are the proteins from yeast and *E. coli* that prompted the study. From those records there are links back to the PubMed abstracts of articles that reported these proteins. PubMed also has a "neighbor" function, **Related articles**, that represents other articles that contain words and phrases in common with the current record. Because phrases such as "*Escherichia coli*", "mismatch repair", and "MutS" all occur in the current article, many of the articles most related to this one describe studies on the *E. coli* mismatch repair system. These articles may not be directly linked to any sequence themselves and may not contain the words "human" or "colon cancer" but are relevant to HNPCC nonetheless, because of what the bacterial system may tell us.

Entrez Is Growing

The original three-node Entrez system has evolved over the past 10 years to include more nodes (Figure 1). These include:

1. Taxonomy, which is organized around the names and phylogenetic relationships of organisms
2. Structure, organized around the three-dimensional structures of proteins and nucleic acids
3. Genomes, in which each record represents a chromosome of an organism
4. *Online Mendelian Inheritance in Man* (OMIM), a text-based resource organized around human genes and their phenotypes
5. PopSet, consisting of collections of aligned sequences from a single population study
6. Books, representing published books in biomedicine

More nodes are planned for addition in the near future. Each one of these nodes is richly connected to others. Each offers unique information and unique new relationships among its members. The combination of new links and new relationships increases the chances for discovery. The addition of each new node creates different paths through the data that may lead to new connections, without more work on the old nodes.

How Entrez Works

Entrez integrates data from a large number of sources, formats, and databases into a uniform information model and retrieval system. The actual databases from which records are retrieved and on which the Entrez indexes are based have different designs, based on the type of data, and reside on different machines. These will be referred to as the "source databases". A common theme in the implementation of Entrez is that some functions are unique to each source database, whereas others are common to all Entrez databases.

A Division of Labor: Basic Principles

Some of the common routines and formats for every Entrez node include the term lists and posting files (i.e., the retrieval engine) used for Boolean queries, the links within and between nodes, and the summary format used for listing search results in which each record is called a DocSum. Generally, an Entrez query is a Boolean expression that is evaluated by the common Entrez engine and yields a list of unique ID numbers (UIDs), which identify records in an Entrez node. Given one or more UIDs, Entrez can retrieve the DocSum(s) very quickly. The links made within or between Entrez nodes from one or more UIDs is also a function across all Entrez source databases.

The software that tracks the addition of new or updated records or identifies those that should be deleted from Entrez may be unique for each source database. Each database must also have accompanying software to gather index terms, DocSums, and links from the source data and present them to the common Entrez indexer. This can be achieved through either a set of C++ libraries or by generating an XML document in a specific DTD that contains the terms, DocSums, and links. Although the common engine retrieves a DocSum(s) given a UID(s), the retrieval of a full, formatted record is directed to the source database, where software unique to that database is used to format the record correctly. All of this software is written by the NCBI group that runs the database.

This combination of database-specific software and a common set of Entrez routines and applications allows code sharing and common large-retrieval server administration but enables flexibility and simplicity for a wide variety of data sources.

Software

Although the basic principles of Entrez have remained the same for almost a decade, the software implementation has been through at least three major redesigns and many minor ones.

Currently, Entrez is written using the NCBI C++ Toolkit. The indexing fields (which for PubMed, for example, would be Title, Author, Publication Date, Journal, Abstract, and so on) and DocSum fields (which for PubMed are Author, Title, Journal, Publication Date, Volume, and Page Number) for each node are defined in a configuration file; but for performance at runtime, the configuration files are used to automatically generate base classes for each database. These are the basic pieces of information used by Entrez that can also be inherited and used by more database-specific, hand-coded features. The term indexes are based on the Indexed Sequential-Access Method (ISAM) and are in large, shared, memory-mapped files. The postings are large bitmaps, with one bit per document in the node. Depending on how sparsely populated the posting is, the bit array is adaptively compressed on disk using one of four possible schemes. Boolean operations are performed by using AND or OR postings of bit arrays into a result bit array. DocSums are small, fielded data structures stored on the same machines as the postings to support rapid retrieval.

The Web-based Entrez retrieval program, called *query*, is a fast cgi application that uses the Web application framework from the NCBI C++ Toolkit. One aspect of this framework is a set of classes that represents an HTML page. These classes allow the combination of static template pages, on the fly, with callbacks to class methods at tagged parts of the template. The Web page generated in an Entrez session contains elements from static templates and elements generated dynamically from common Entrez classes and from classes unique to one or a few Entrez nodes. Again, this design supports a common core of robust, common functionality maintained by one group, with support for customizations by diverse groups within NCBI.

Boolean query processing, DocSum retrieval, and other common functions are supported on a number of load-balanced "front-end" UNIX machines. Because Entrez can support session context (for example, in the use of query history, NCBI Cubby, Filters, etc.), a "history server" has been implemented on the front-end machines so that if a user is sent to machine "A" by the load balancer for their first query but to machine "B" for the second query, Entrez can quickly locate the user's query history and obtain it from machine "A". Other than that, the front-end machines are completely independent of each other and can be added and removed readily from *query* support. Retrieval of full documents comes from a variety of "back-end" databases, depending on the node. These might be Sybase or Microsoft SQL Server relational databases of a variety of schemas or text files of various formats. Links are supported using the Sybase IQ database product.

References

1. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell. 1993;75:1027–1038. PubMed PMID: 8252616.

Chapter 16. The BLAST Sequence Analysis Tool

Tom Madden

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The comparison of nucleotide or protein sequences from the same or different organisms is a very powerful tool in molecular biology. By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships. Now that whole genomes are being sequenced, sequence similarity searching can be used to predict the location and function of protein-coding and transcription-regulation regions in genomic DNA.

Basic Local Alignment Search Tool (BLAST) (1, 2) is the tool most frequently used for calculating sequence similarity. BLAST comes in variations for use with different query sequences against different databases. All BLAST applications, as well as information on which BLAST program to use and other help documentation, are listed on the BLAST homepage. This chapter will focus more on how BLAST works, its output, and how both the output and program itself can be further manipulated or customized, rather than on how to use BLAST or interpret BLAST results.

Introduction

The way most people use BLAST is to input a nucleotide or protein sequence as a query against all (or a subset of) the public sequence databases, pasting the sequence into the textbox on one of the BLAST Web pages. This sends the query over the Internet, the search is performed on the NCBI databases and servers, and the results are posted back to the person's browser in the chosen display format. However, many biotech companies, genome scientists, and bioinformatics personnel may want to use "stand-alone" BLAST to query their own, local databases or want to customize BLAST in some way to make it better suit their needs. Stand-alone BLAST comes in two forms: the executables that can be run from the command line; or the Standalone WWW BLAST Server, which allows users to set up their own in-house versions of the BLAST Web pages.

There are many different variations of BLAST available to use for different sequence comparisons, e.g., a DNA query to a DNA database, a protein query to a protein database, and a DNA query, translated in all six reading frames, to a protein sequence database. Other adaptations of BLAST, such as PSI-BLAST (for iterative protein sequence similarity searches using a position-specific score matrix) and RPS-BLAST (for searching for protein domains in the Conserved Domains Database, Chapter 3) perform comparisons against sequence profiles.

This chapter will first describe the BLAST architecture—how it works at the NCBI site—and then go on to describe the various BLAST outputs. The best known of these outputs is the default display from BLAST Web pages, the so-called "traditional report". As well as

obtaining BLAST results in the traditional report, results can also be delivered in structured output, such as a hit table (see below), XML, or ASN.1. The optimal choice of output format depends upon the application. The final part of the chapter discusses stand-alone BLAST and describes possibilities for customization. There are many interfaces to BLAST that are often not exploited by users but can lead to more efficient and robust applications.

How BLAST Works: The Basics

The BLAST algorithm is a heuristic program, which means that it relies on some smart shortcuts to perform the search faster. BLAST performs "local" alignments. Most proteins are modular in nature, with functional domains often being repeated within the same protein as well as across different proteins from different species. The BLAST algorithm is tuned to find these domains or shorter stretches of sequence similarity. The local alignment approach also means that a mRNA can be aligned with a piece of genomic DNA, as is frequently required in genome assembly and analysis. If instead BLAST started out by attempting to align two sequences over their entire lengths (known as a global alignment), fewer similarities would be detected, especially with respect to domains and motifs.

When a query is submitted via one of the BLAST Web pages, the sequence, plus any other input information such as the database to be searched, word size, expect value, and so on, are fed to the algorithm on the BLAST server. BLAST works by first making a look-up table of all the “words” (short subsequences, which for proteins the default is three letters) and “neighboring words”, i.e., similar words in the query sequence. The sequence database is then scanned for these “hot spots”. When a match is identified, it is used to initiate gap-free and gapped extensions of the “word”.

BLAST does not search GenBank flatfiles (or any subset of GenBank flatfiles) directly. Rather, sequences are made into BLAST databases. Each entry is split, and two files are formed, one containing just the header information and one containing just the sequence information. These are the data that the algorithm uses. If BLAST is to be run in “stand-alone” mode, the data file could consist of local, private data, downloaded NCBI BLAST databases, or a combination of the two.

After the algorithm has looked up all possible "words" from the query sequence and extended them maximally, it assembles the best alignment for each query–sequence pair and writes this information to an SeqAlign data structure (in ASN.1 ; also used by Sequin, see Chapter 12). The SeqAlign structure in itself does not contain the sequence information; rather, it refers to the sequences in the BLAST database (Figure 1).

The BLAST Formatter, which sits on the BLAST server, can use the information in the SeqAlign to retrieve the similar sequences found and display them in a variety of ways. Thus, once a query has been completed, the results can be reformatted without having to re-execute the search. This is possible because of the QBLAST system.

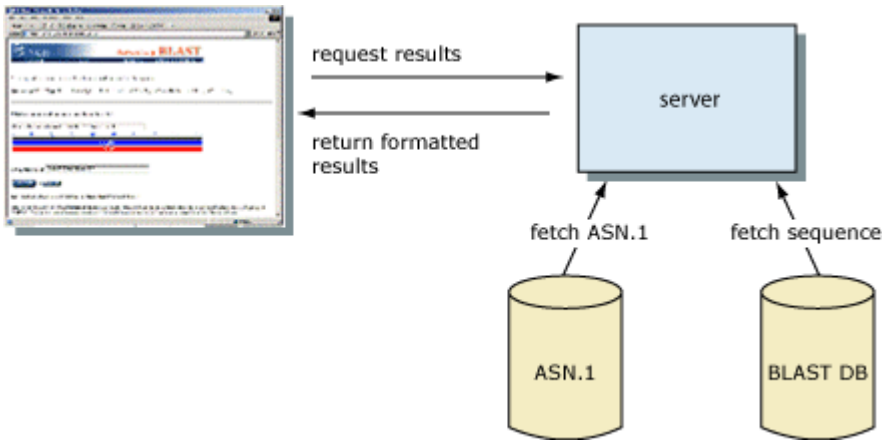


Figure 1. How the BLAST results Web pages are assembled. The QBLAST system located on the BLAST server executes the search, writing information about the sequence alignment in ASN.1. The results can then be formatted by fetching the ASN.1 (*fetch ASN.1*) and fetching the sequences (*fetch sequence*) from the BLAST databases. Because the execution of the search algorithm is decoupled from the formatting, the results can be delivered in a variety of formats without re-running the search.

BLAST Scores and Statistics

Once BLAST has found a similar sequence to the query in the database, it is helpful to have some idea of whether the alignment is "good" and whether it portrays a possible biological relationship, or whether the similarity observed is attributable to chance alone. BLAST uses statistical theory to produce a bit score and expect value (E-value) for each alignment pair (query to hit).

The bit score gives an indication of how good the alignment is; the higher the score, the better the alignment. In general terms, this score is calculated from a formula that takes into account the alignment of similar or identical residues, as well as any gaps introduced to align the sequences. A key element in this calculation is the "substitution matrix ", which assigns a score for aligning any possible pair of residues. The BLOSUM62 matrix is the default for most BLAST programs, the exceptions being blastn and MegaBLAST (programs that perform nucleotide–nucleotide comparisons and hence do not use protein-specific matrices). Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used.

The E-value gives an indication of the statistical significance of a given pairwise alignment and reflects the size of the database and the scoring system used. The lower the E-value, the more significant the hit. A sequence alignment that has an E-value of 0.05 means that this similarity has a 5 in 100 (1 in 20) chance of occurring by chance alone. Although a statistician might consider this to be significant, it still may not represent a biologically meaningful result, and analysis of the alignments (see below) is required to determine "biological" significance.

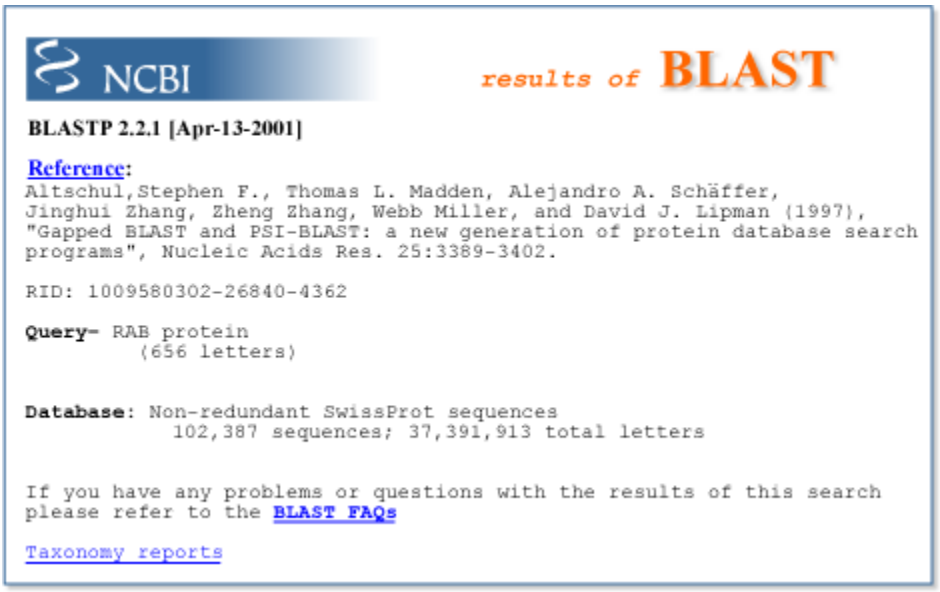


Figure 2. The BLAST report header. The *top line* gives information about the type of program (in this case, *BLASTP*), the version (*2.2.1*), and a version release date. The research paper that describes BLAST is then cited, followed by the request ID (issued by QBLAST), the query sequence definition line, and a summary of the database searched. The **Taxonomy reports** link displays this BLAST result on the basis of information in the Taxonomy database (Chapter 4).

BLAST Output: 1. The Traditional Report

Most BLAST users are familiar with the so-called "traditional" BLAST report. The report consists of three major sections: (1) the header, which contains information about the query sequence, the database searched (Figure 2). On the Web, there is also a graphical overview (Figure 3); (2) the one-line descriptions of each database sequence found to match the query sequence; these provide a quick overview for browsing (Figure 4); (3) the alignments for each database sequence matched (Figure 5) (there may be more than one alignment for a database sequence it matches).

The traditional report is really designed for human readability, as opposed to being parsed by a program. For example, the one-line descriptions are useful for people to get a quick overview of their search results, but they are rarely complete descriptors because of limited space. Also, for convenience, there are several pieces of information that are displayed in both the one-line descriptions and alignments (for example, the E-values, scores, and descriptions); therefore, the person viewing the search output does not need to move back and forth between sections.

New features may be added to the report, e.g., the addition of links to Entrez Gene records (Chapter 19) from sequence hits, which result in a change of output format. These are easy for people to pick up on and take advantage of but can trip programs that parse this BLAST output.

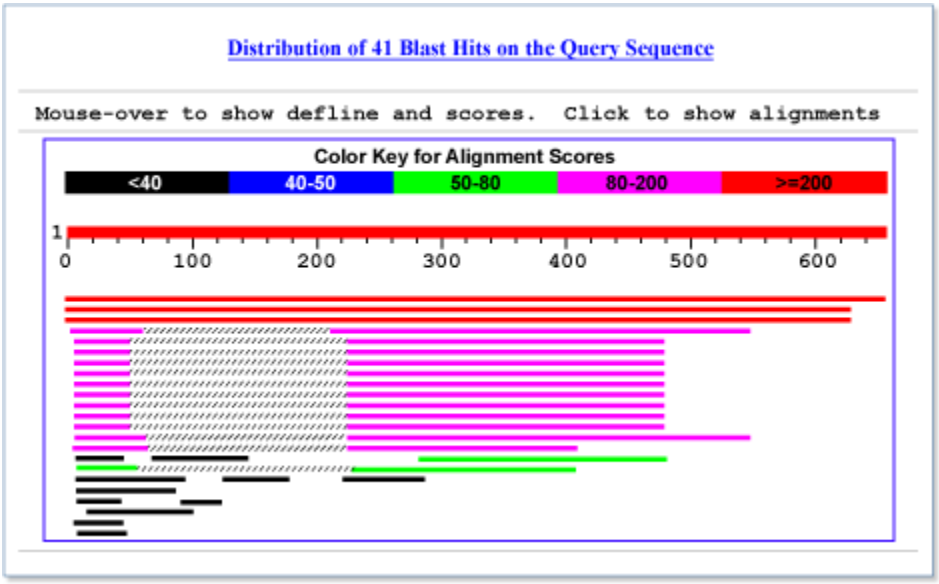


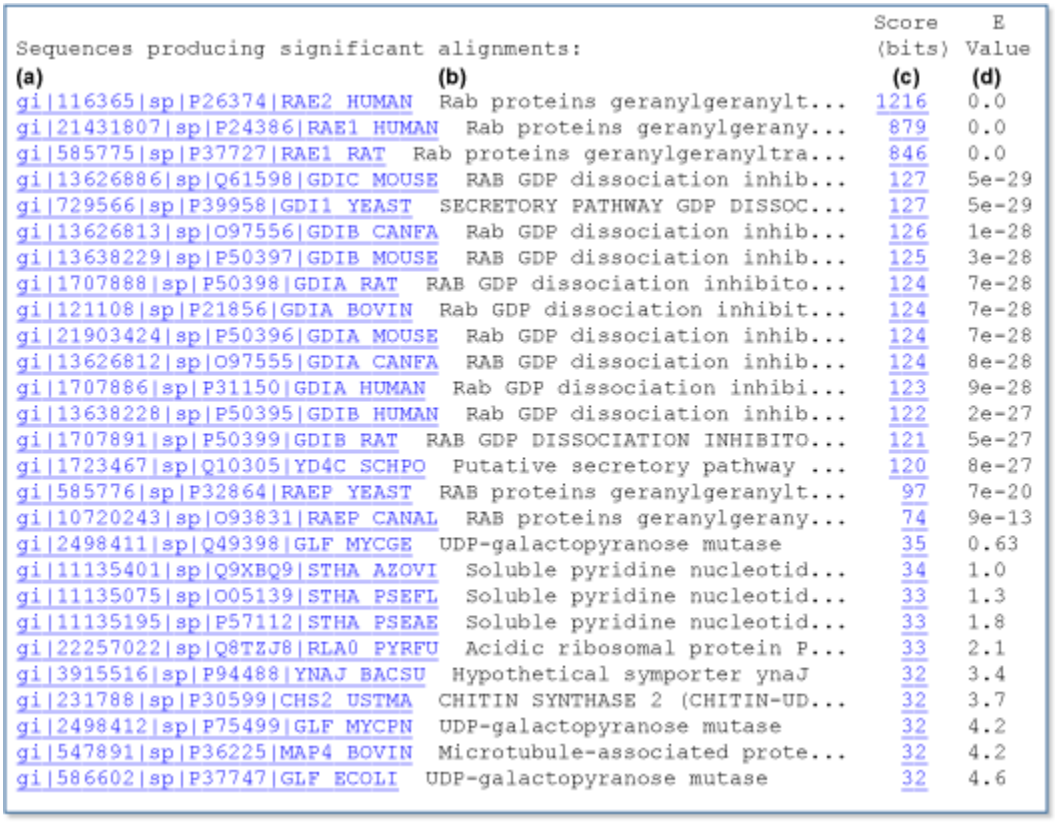
Figure 3. Graphical overview of BLAST results. The query sequence is represented by the *numbered red bar* at the *top* of the figure. Database hits are shown aligned to the query, *below* the red bar. Of the aligned sequences, the most similar are shown closest to the query. In this case, there are three high-scoring database matches that align to most of the query sequence. The next twelve bars represent lower-scoring matches that align to two regions of the query, from about residues 3–60 and residues 220–500. The *cross-hatched parts* of the these bars indicate that the two regions of similarity are on the same protein, but that this intervening region does not match. The remaining bars show lower-scoring alignments. Mousing over the bars displays the definition line for that sequence to be shown in the window above the graphic.

By default, a maximum of 500 sequence matches are displayed, which can be changed on the advanced BLAST page with the **Alignments** option. Many components of the BLAST results display via the Internet and are hyperlinked to the same information at different places in the page, to additional information including help documentation, and to the Entrez sequence records of matched sequences. These records provide more information about the sequence, including links to relevant research abstracts in PubMed.

BLAST Output: 2. The Hit Table

Although the traditional report is ideal for investigating the characteristics of one gene or protein, often scientists want to make a large number of BLAST runs for a specialized purpose and need only a subset of the information contained in the traditional BLAST report. Furthermore, in cases where the BLAST output will be processed further, it can be unreliable to parse the traditional report. The traditional report is merely a display format with no formal structure or rules, and improvements may be made at any time, changing the underlying HTML. The hit table format provides a simple and clean alternative (Figure 6).

The screening of many newly sequenced human Expressed Sequence Tags (ESTs) for contamination by the *Escherichia coli* cloning vector is a good example of when it is



Sequences producing significant alignments:		Score (bits)	E Value
(a)	(b)	(c)	(d)
gi|116365|sp|P26374|RAE2_HUMAN	Rab proteins geranylgeranylt...	1216	0.0
gi|21431807|sp|P24386|RAE1_HUMAN	Rab proteins geranylgerany...	879	0.0
gi|585775|sp|P37727|RAE1_RAT	Rab proteins geranylgeranyltra...	846	0.0
gi|13626886|sp|Q61598|GDIC_MOUSE	RAB GDP dissociation inhib...	127	5e-29
gi|729566|sp|P39958|GDI1_YEAST	SECRETORY PATHWAY GDP DISSOC...	127	5e-29
gi|13626813|sp|O97556|GDIB_CANFA	Rab GDP dissociation inhib...	126	1e-28
gi|13638229|sp|P50397|GDIB_MOUSE	RAB GDP dissociation inhib...	125	3e-28
gi|1707888|sp|P50398|GDIA_RAT	RAB GDP dissociation inhibito...	124	7e-28
gi|121108|sp|P21856|GDIA_BOVIN	Rab GDP dissociation inhibit...	124	7e-28
gi|21903424|sp|P50396|GDIA_MOUSE	Rab GDP dissociation inhib...	124	7e-28
gi|13626812|sp|O97555|GDIA_CANFA	RAB GDP dissociation inhib...	124	8e-28
gi|1707886|sp|P31150|GDIA_HUMAN	Rab GDP dissociation inhibi...	123	9e-28
gi|13638228|sp|P50395|GDIB_HUMAN	Rab GDP dissociation inhib...	122	2e-27
gi|1707891|sp|P50399|GDIB_RAT	RAB GDP DISSOCIATION INHIBITO...	121	5e-27
gi|1723467|sp|Q10305|YD4C_SCHPO	Putative secretory pathway ...	120	8e-27
gi|585776|sp|P32864|RAEP_YEAST	RAB proteins geranylgeranylt...	97	7e-20
gi|10720243|sp|O93831|RAEP_CANAL	RAB proteins geranylgerany...	74	9e-13
gi|2498411|sp|Q49398|GLF_MYCGE	UDP-galactopyranose mutase	35	0.63
gi|11135401|sp|Q9XBQ9|STHA_AZOVI	Soluble pyridine nucleotid...	34	1.0
gi|11135075|sp|O05139|STHA_PSEFL	Soluble pyridine nucleotid...	33	1.3
gi|11135195|sp|P57112|STHA_PSEAE	Soluble pyridine nucleotid...	33	1.8
gi|22257022|sp|Q8TZJ8|RLA0_PYRFU	Acidic ribosomal protein P...	33	2.1
gi|3915516|sp|P94488|YNAJ_BACSU	Hypothetical symporter ynaJ	32	3.4
gi|231788|sp|P30599|CHS2_USTMA	CHITIN SYNTHASE 2 (CHITIN-UD...	32	3.7
gi|2498412|sp|P75499|GLF_MYCPN	UDP-galactopyranose mutase	32	4.2
gi|547891|sp|P36225|MAP4_BOVIN	Microtubule-associated prote...	32	4.2
gi|586602|sp|P37747|GLF_ECOLI	UDP-galactopyranose mutase	32	4.6

Figure 4. One-line descriptions in the BLAST report. Each line is composed of four fields: (*a*) the gi number, database designation, Accession number, and locus name for the matched sequence, separated by vertical bars (Appendix 1); (*b*) a brief textual description of the sequence, the definition. This usually includes information on the organism from which the sequence was derived, the type of sequence (e.g., mRNA or DNA), and some information about function or phenotype. The definition line is often truncated in the one-line descriptions to keep the display compact; (*c*) the alignment score in bits. Higher scoring hits are found at the top of the list; and (*d*) the E-value, which provides an estimate of statistical significance. For the first hit in the list, the gi number is *116365*, the database designation is *sp* (for SWISS-PROT), the Accession number is *P26374*, the locus name is *RAE2_HUMAN*, the definition line is *Rab proteins*, the score is *1216*, and the E-value is *0.0*. Note that the first 17 hits have very low E-values (much less than 1) and are either RAB proteins or GDP dissociation inhibitors. The other database matches have much higher E-values, 0.5 and above, which means that these sequences may have been matched by chance alone.

preferable to use the hit table output over the traditional report. In this case, a strict, high E-value threshold would be applied to differentiate between contaminating *E. coli* sequence and the human sequence. Those human ESTs that find very strong, near-exact *E.coli* sequence matches can be discarded without further examination. (Borderline cases may require further examination by a scientist.)

For these purposes, the hit table output is more useful than the traditional report; it contains only the information required in a more formal structure. The hit table output contains no sequences or definition lines, but for each sequence matched, it lists the

```
>gi|116365|sp|P26374|RAE2_HUMAN  Rab proteins geranylgeranyltransferase component A 2 (Rab escort
          protein 2) (REP-2) (Choroideraemia-like protein)
         Length = 656

 Score =  846 bits (2186), Expect = 0.0
 Identities = 432/632 (68%), Positives = 489/632 (77%), Gaps = 13/632 (2%)

Query: 1   MADNLPTEFDVVIIGTGLPESILAAACSRSGQRVLHIDSRSYYGGNWASFSFSGLLSWLK 60
           MADNLP++FDV++IGTGLPESI+AAACSRSGQRVLH+DSRSYYGGNWASFSFSGLLSWLK
Sbjct: 1   MADNLPSDFDVIVIGTGLPESIIAAACSRSGQRVLHVDSRSYYGGNWASFSFSGLLSWLK 60

Query: 61  EYQQNNDIGEESTVVWQDLIHETEEAITLRKKDETIQHTEAFPYASQDMEDNVEEIGALQ 120
           EYQ+NND+  E++ +WQ+ I E EEAI L  KD+TIQH E F YASQD+  +VEE GALQ
Sbjct: 61  EYQENNDVVTENS-MWQEQILENEEAIPLSSKDKTIQHVEVFCYASQDLHKDVEEAGALQ 119

Query: 121 KNPSLGVS----NTFTEVLDSALPEESQLSYFNSDEMPAKHTQKSDTEISLEVTDVEESV 176
           KN +   S          S LP   +     S E+PA+ +Q    E S EV D E +
Sbjct: 120 KNHASVTSAQSAEAAEAAETSCLPTAVEPLSMGSCEIPAEQSQCPGPESSPEVNDAEATG 179

Query: 177 EKEKYCGDKTCMHTVXXXXXXXXXXXXTVEDKADEPIRNRITYSQIVKEGRRFNIDLVSK 236
           +KE      +                V+D  + P +NRITYSQI+KEGRRFNIDLVS+
Sbjct: 180 KKENSDAKSS--------TEEPSENVPKVQDNTETPKKNRITYSQIIKEGRRFNIDLVSQ 231

Query: 237 LLYSQGLLIDLLIKSDVSRYVEFKNVTRILAFREGKVEQVPCSRADVFNSKELTMVEKRM 296
           LLYS+GLLIDLLIKS+VSRY EFKN+TRILAFREG VEQVPCSRADVFNSK+LTMVEKRM
Sbjct: 232 LLYSRGLLIDLLIKSNVSRYAEFKNITRILAFREGTVEQVPCSRADVFNSKQLTMVEKRM 291

Query: 297 LMKFLTFCLEYEQHPDEYQAFRQCSFSEYLKTKKLTPNLQHFVLHSIAMTSESSCTTIDG 356
           LMKFLTFC+EYE+HPDEY+A+    +FSEYLKT+KLTPNLQ+FVLHSIAMTSE++  T+DG
Sbjct: 292 LMKFLTFCVEYEEHPDEYRAYEGTTFSEYLKTQKLTPNLQYFVLHSIAMTSETTSCTVDG 351

Query: 357 LNATKNFLQCLGRFGNTPFLFPLYGQGEIPQGFCRMCAVFGGIYCLRHKVQCFVVDKESG 416
           L ATK FLQCLGR+GNTPFLFPLYGQGE+PQ FCRMCAVFGGIYCLRH VQC VVDKES
Sbjct: 352 LKATKKFLQCLGRYGNTPFLFPLYGQGELPQCFCRMCAVFGGIYCLRHSVQCLVVDKESR 411

Query: 417 RCKAIIDHFGQRINAKYFIVEDSYLSEETCSNVQYKQISRAVLITDQSILKTDLDQQTSI 476
           +CKA+ID FGQRI +K+FI+EDSYLSE TCS VQY+QISRAVLITD S+LKTD DQQ SI
Sbjct: 412 KCKAVIDQFGQRIISKHFIIEDSYLSENTCSRVQYRQISRAVLITDGSVLKTDADQQVSI 471

Query: 477 LIVPPAEPGACAVRVTELCSSTMTCMKDTYLVHLTCSSSKTAREDLESVVKKLFTPYTET 536
           L VP  EPG+  VRV ELCSSTMTCMK TYLVHLTC SSKTAREDLE VV+KLFTPYTE
Sbjct: 472 LAVPAEEPGSFGVRVIELCSSTMTCMKGTYLVHLTCMSSKTAREDLERVVQKLFTPYTEI 531

Query: 537 EINEEELTKPRLLWALYFNMRDSSGISRSSYNGLPSNVYVCSGPDCGLGNEHAVKQAETL 596
           E   E++ KPRLLWALYFNMRDSS ISR  YN LPSNVYVCSGPD GLGN++AVKQAETL
Sbjct: 532 EAENEQVEKPRLLWALYFNMRDSSDISRDCYNDLPSNVYVCSGPDSGLGNDNAVKQAETL 591

Query: 597 FQXXXXXXXXXXXXXXXXXXXDGDDKQPEAP 628
           FQ                  DGD  Q E P
Sbjct: 592 FQQICPNEDFCPAPPNPEDIVLDGDSSQQEVP 623
```

Figure 5. A pairwise sequence alignment from a BLAST report. The alignment is preceded by the sequence identifier, the full definition line, and the length of the matched sequence, in amino acids. Next comes the bit score (the raw score is in *parentheses*) and then the E-value. The following line contains information on the number of identical residues in this alignment (*Identities*), the number of conservative substitutions (*Positives*), and if applicable, the number of gaps in the alignment. Finally, the actual alignment is shown, with the query on *top*, and the database match is labeled as *Sbjct*, below. The numbers at *left* and *right* refer to the position in the amino acid sequence. One or more dashes (–) within a sequence indicate insertions or deletions. Amino acid residues in the query sequence that have been masked because of low complexity are replaced by Xs (see, for example, the *fourth* and *last* blocks). The line between the two sequences indicates the similarities between the sequences. If the query and the subject have the same amino acid at a given location, the residue itself is shown. Conservative substitutions, as judged by the substitution matrix, are indicated with +.

sequence identifier, the start and stop points for stretches of sequence similarity (offset by one residue), the percent identity of the match, and the E-value.

BLAST Output: 3. Structured Output

There are drawbacks to parsing both the BLAST report and even the simpler hit table. There is no way to automatically check for truncated or otherwise corrupted output in cases when a large number of sequences are being screened. (This may happen if the disk

```
# BLASTN 2.2.1 [Aug-1-2001]
# Database: ecoli
# Query:gi|4730899|dbj|AP000130.1|Homo sapiens genomic DNA of 21q22.1, GART and AML, f43D11-119B8 region, segment 5/10.
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit
gi|4730899|dbj|AP000130.1|AP000130     gi|2367099|gb|AE000133.1|AE000133     100.00   1198    0       0       52913   54110   10943
gi|4730899|dbj|AP000130.1|AP000130     gi|1789919|gb|AE000427.1|AE000427     100.00   1198    0       0       52913   54107   2347
gi|4730899|dbj|AP000130.1|AP000130     gi|1789607|gb|AE000401.1|AE000401     100.00   1198    0       0       52913   54107   6037
gi|4730899|dbj|AP000130.1|AP000130     gi|1788338|gb|AE000294.1|AE000294     100.00   1198    0       0       52913   54107   5700
gi|4730899|dbj|AP000130.1|AP000130     gi|1787588|gb|AE000231.1|AE000231     100.00   1198    0       0       52913   54107   4146
gi|4730899|dbj|AP000130.1|AP000130     gi|1786875|gb|AE000170.1|AE000170     100.00   1198    0       0       52913   54107   2321
gi|4730899|dbj|AP000130.1|AP000130     gi|1786751|gb|AE000160.1|AE000160     100.00   1198    0       0       52913   54107   9133
gi|4730899|dbj|AP000130.1|AP000130     gi|1788508|gb|AE000308.1|AE000308     99.92    1198    1       0       52913   54107   11740
gi|4730899|dbj|AP000130.1|AP000130     gi|2367181|gb|AE000381.1|AE000381     99.83    1198    2       0       52913   54107   2030
gi|4730899|dbj|AP000130.1|AP000130     gi|1788298|gb|AE000291.1|AE000291     99.58    1198    5       0       52910   54107   5290
gi|4730899|dbj|AP000130.1|AP000130     gi|1787633|gb|AE000234.1|AE000234     93.21    1105    72      3       52912   54014   1146
```

Figure 6. BLAST output in hit table format. This shows the results of a search of an *E. coli* database using a human sequence as a query. The lines starting with a # sign should be considered comments and ignored. The *last comment line* lists the fields in the table.

is full, for example.) Also, there is no rigorous check for syntax changes in the output, such as the addition of new features, which can lead to erroneous parsing. Structured output allows for automatic and rigorous checks for syntax errors and changes. Both XML and ASN.1 are examples of structured output in which there are built-in checks for correct and complete syntax and structure. (In the case of XML, for example, this is ensured by the necessity for matching tags and the DTD.) For text reports, there is often no specification, but perhaps a (incomplete) description of the file is written afterward.

ASN.1 Is Used by the BLAST Server

As well as the hit table and traditional report shown in HTML, BLAST results can also be formatted in plain text, XML, and ASN.1 (Figure 7), and what's more, the format for a given BLAST result can be changed without re-executing the search.

A change in BLAST format without re-executing the search is possible because when a scientist looks at a Web page of BLAST results at NCBI, the HTML that makes that page has been created from ASN.1 (Figure 7). Although the formatted results are requested from the server, the information about the alignments is fetched from a disk in ASN.1, as are the corresponding sequences from the BLAST databases (see Figure 1). The formatter on the BLAST server then puts these results together as a BLAST report. The BLAST search itself has been uncoupled from the way the result is formatted, thus allowing different output formats from the same search. The strict internal validation of ASN.1 ensures that these output formats can always be produced reliably.

Information about the Alignment Is Contained within a SeqAlign

SeqAlign is the ASN.1 object that contains the alignment information about the BLAST search. The SeqAlign does not contain the actual sequence that was found in the match but does contain the start, stop, and gap information, as well as scores, E-values, sequence identifiers, and (DNA) strand information.

As mentioned above, the actual database sequences are fetched from the BLAST databases when needed. This means that an identifier must uniquely identify a sequence in the

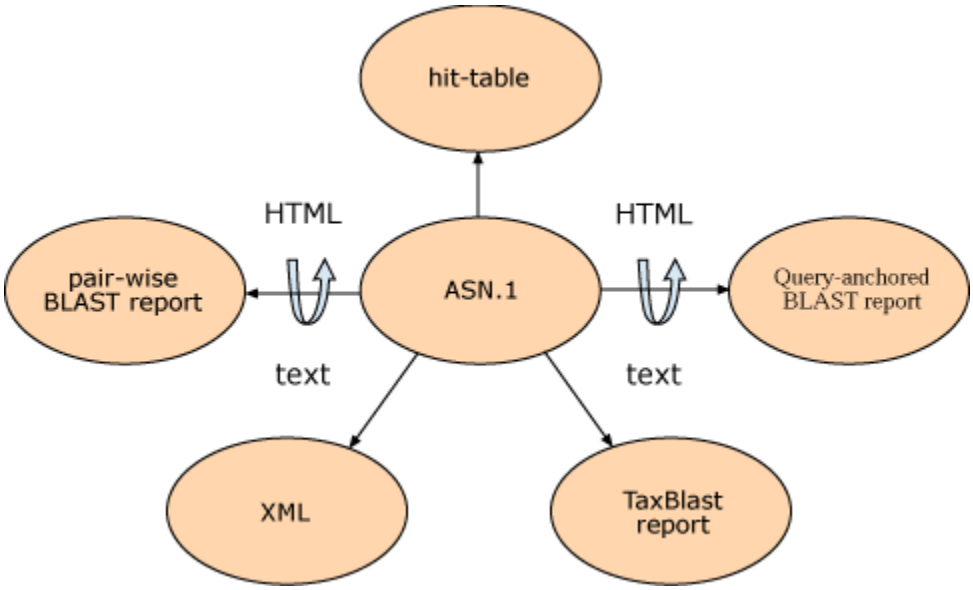


Figure 7. The different output formats that can be produced from ASN.1. Note that some nodes can be viewed as both HTML and text. XML is also structured output but can be produced from ASN.1 because it has equivalent information.

database. Furthermore, the query sequence cannot have the same identifier as any sequence in the database unless the query sequence itself is in the database. If one is using stand-alone BLAST with a custom database, it is possible to specify that every sequence is uniquely identified by using the **–O** option with formatdb (the program that converts FASTA files to BLAST database format). This also indexes the entries by identifier. Similarly, the **–J** option in the (stand-alone) programs blastall, blastpgp, megablast, or rpsblast certifies that the query does not use an identifier already in the database for a different sequence. If the **–O** and **–J** options are not used, BLAST assigns unique identifiers (for that run) to all sequences and shields the user from this knowledge.

Any BLAST database or FASTA file from the NCBI Web site that contains gi numbers already satisfies the uniqueness criterion. Unique identifiers are normally a problem only when custom databases are produced and care is not taken in assigning identifiers. The identifier for a FASTA entry is the first token (meaning the letters up to the first space) after the > sign on the definition line. The simplest case is to simply have a unique token (e.g., 1, 2, and so on), but it is possible to construct more complicated identifiers that might, for example, describe the data source. For the FASTA identifiers to be reliably parsed, it is necessary for them to follow a specific syntax (see Appendix 1).

More information on the SeqAlign produced by BLAST can be found here or be downloaded as a PowerPoint presentation, as well as from the NCBI Toolkit Software Developer's handbook.

XML

XML and ASN.1 are both structured languages and can express the same information; therefore, it is possible to produce a SeqAlign in XML. Some users do not find the format of the information in the SeqAlign to be convenient because it does not contain actual sequence information, and when the sequence is fetched from the BLAST database, it is packed two or four bases per byte. Typically, these users are familiar with the BLAST report and want something similar but in a format that can be parsed reliably. The XML produced by BLAST meets this need, containing the query and database sequences, sequence definition lines, the start and stop points of the alignments (one offset), as well as scores, E-values, and percent identity. There is a public DTD for this XML output.

BLAST Code

The BLAST code is part of the NCBI Toolkit, which has many low-level functions to make it platform independent; the Toolkit is supported under Linux and many varieties of UNIX, NT, and MacOS. To use the Toolkit, developers should write a function "Main", which is called by the Toolkit "main". The BLAST code is contained mostly in the tools directory (see Appendix 2 for an example).

The BLAST code has a modular design. For example, the Application Programming Interface (API) for retrieval from the BLAST databases is independent of the compute engine. The compute engine is independent from the formatter; therefore, it is possible (as mentioned above) to compute results once but view them in many different modes.

Readdb API

The readdb API can be used to easily extract information from the BLAST databases. Among the data available are the date the database was produced, the title, the number of letters, number of sequences, and the longest sequence. Also available are the sequence and description of any entry. The latest version of the BLAST databases also contains a taxid (an integer specifying some node of the NCBI taxonomy tree; see Chapter 4). Users are strongly encouraged to use the readdb API rather than reading the files associated with the database, because the the files are subject to change. The API, on the other hand, will support the newest version, and an attempt will be made to support older versions. See Appendix 2 for an example of a simple program (db2fasta.c) that demonstrates the use of the readdb API.

Performing a BLAST Search with C Function Calls

Only a few function calls are needed to perform a BLAST search. Appendix 3 shows an excerpt from a Demonstration Program doblast.c.

Formatting a SeqAlign

MySeqAlignPrint (called in the example in Appendix 3) is a simple function to print a view of a SeqAlign (see Appendix 4).

Appendix 1. FASTA identifiers

The syntax of the FASTA definition lines used in the NCBI BLAST databases depends upon the database from which each sequence was obtained (see Chapter 1 on GenBank). Table 1 shows how the sequence source databases are identified.

For example, if the identifier of a sequence in a BLAST result is gb|M73307|AGMA13GT, the gb tag indicates that sequence is from GenBank, M73307 is the GenBank Accession number, and AGMA13GT is the GenBank locus.

The bar (|) separates different fields. In some cases, a field is left empty, although the original specification called for including this field. To make these identifiers backwards-compatible for older parsers, the empty field is denoted by an additional bar (||).

A gi identifier has been assigned to each sequence in NCBI's sequence databases. If the sequence is from an NCBI database, then the gi number appears at the beginning of the identifier in a traditional report. For example, gi|16760827|ref|NP_456444.1 indicates an NCBI reference sequence with the gi number 16760827 and Accession number NP_456444.1. (In stand-alone BLAST, or when running BLAST from the command line, the **–I** option should be used to display the gi number.)

The reason for adding the gi identifier is to provide a uniform, stable naming convention. If a nucleotide or protein sequence changes (for example, if it is edited by the original submitter of the sequence), a new gi identifier is assigned, but the Accession number of the record remains unchanged. Thus, the gi identifier provides a mechanism for identifying the exact sequence that was used or retrieved in a given search. This is also useful when creating crosslinks between different Entrez databases (Chapter 15).

Table 1. Database identifiers in FASTA definition lines.

Database name	Identifier syntax
GenBank	gb|accession|locus
EMBL Data Library	emb|accession|locus
DDBJ, DNA Database of Japan	dbj|accession|locus
NBRF PIR	pir||entry
Protein Research Foundation	prf||name

a gnl allows databases not included in this list to use the same identifying syntax. This is used for sequences in the **trace databases**, e.g., gnl|ti|53185177. The combination of the second and third fields should be unique.

Table 1 continues on next page...

Table 1 continued from previous page.

Database name	Identifier syntax
SWISS-PROT	sp|accession|entry name
Brookhaven Protein Data Bank	pdb|entry|chain
Patents	pat|country|number
GenInfo Backbone Id	bbs|number
General database identifier[a]	gnl|database|identifier
NCBI Reference Sequence	ref|accession|locus
Local Sequence identifier	lcl|identifier

a gnl allows databases not included in this list to use the same identifying syntax. This is used for sequences in the **trace databases**, e.g., gnl|ti|53185177. The combination of the second and third fields should be unique.

Appendix 2. Readdb API

A simple program (db2fasta.c) that demonstrates the use of the readdb API.

```
Int2 Main (void)
{
    BioseqPtr bsp;
    Boolean is_prot;
    ReadDBFILEPtr rdfp;
    FILE *fp;
    Int4 index;
if (! GetArgs ("db2fasta", NUMARG, myargs))
{
    return (1);
}
    if (myargs[1].intvalue)
    is_prot = TRUE;
    else
    is_prot = FALSE;
    fp = FileOpen("stdout", "w");
    rdfp = readdb_new(myargs[0].strvalue, is_prot);
    index = readdb_acc2fasta(rdfp, myargs[2].strvalue);
    bsp = readdb_get_bioseq(rdfp, index);
    BioseqRawToFasta(bsp, fp, !is_prot);
    bsp = BioseqFree(bsp);
    rdfp = readdb_destruct(rdfp);
    return 0;
}
```

Note that:

1. Readdb_new allocates an object for reading the database.
2. Readdb_acc2fasta fetches the ordinal number (zero offset) of the record given a FASTA identifier (e.g., gb|AAH06776.1|AAH0676).

3. Readdb_get_bioseq fetches the BioseqPtr (which contains the sequence, description, and identifiers) for this record.
4. BioseqRawToFasta dumps the sequence as FASTA.

Note also that Main is called, rather than "main", and a call to GetArgs is used to get the command-line arguments. db2fasta.c is contained in the tar archive ftp://ftp.ncbi.nih.gov/blast/demo/blast_demo.tar.gz.

Appendix 3. Excerpt from a demonstration program doblast.c

```
/* Get default options. */
options = BLASTOptionNew(blast_program, TRUE);
if (options == NULL)
    return 5;

options->expect_value = (Nlm_FloatHi) myargs [3].floatvalue;

/* Perform the actual search. */
seqalign = BioseqBlastEngine(query_bsp, blast_program, blast_database, options,

    NULL, NULL, NULL);

/* Do something with the SeqAlign... */
MySeqAlignPrint(seqalign, outfp);

/* clean up. */
seqalign = SeqAlignSetFree(seqalign);
options = BLASTOptionDelete(options);

sep = SeqEntryFree(sep);
FileClose(infp);
FileClose(outfp);
```

The main steps here are:

1. BLASTOptionNew allocates a BLASTOptionBlk with default values for the specified program (e.g., blastp); the Boolean argument specifies a gapped search.
2. The expect_value member of the BLASTOptionBlk is changed to a non-default value specified on the command-line.
3. BioseqBlastEngine performs the search of the BioseqPtr (query_bsp). The BioseqPtr could have been obtained from the BLAST databases, Entrez, or from FASTA using the function call FastaToSeqEntry.

The BLASTOptionBlk structure contains a large number of members. The most useful ones and a brief description for each are listed in Table 2.

Table 2. The most frequently used BLAST options in the BLASTOptionBlk structure.

Type[a]	Element	Description
Nlm_FloatHi	expect_value	Expect value cutoff
Int2	wordsize	Number of letters used in making words for lookup table
Int2	penalty	Mismatch penalty (only blastn and MegaBLAST)
Int2	reward	Match reward (only blastn and MegaBLAST)
CharPtr	matrix	Matrix used for comparison (not blastn or MegaBLAST)
Int4	gap_open	Cost for gap existence
Int4	gap_extend	Cost to extend a gap one more letter (including first)
CharPtr	filter_string	Filtering options (e.g., L, mL)
Int4	hitlist_size	Number of database sequences to save hits for
Int2	number_of_cpus	Number of CPUs to use

a The types are given in terms of those in the NCBI Toolkit. Nlm_FloatHi is a double, Int2/Int4 are 2- or 4-byte integers, and CharPtr is just char*.

Appendix 4. A function to print a view of a SeqAlign: MySeqAlignPrint

```
#define BUFFER_LEN 50

/*
    Print a report on hits with start/stop. Zero-offset is used.
*/
static void MySeqAlignPrint(SeqAlignPtr seqalign, FILE *outfp)
{
    Char query_id_buf[BUFFER_LEN+1], target_id_buf[BUFFER_LEN+1];
    SeqIdPtr query_id, target_id;
    while (seqalign)
    {
        query_id = SeqAlignId(seqalign, 0);
        SeqIdWrite(query_id, query_id_buf, PRINTID_FASTA_LONG,
BUFFER_LEN);

        target_id = SeqAlignId(seqalign, 1);
        SeqIdWrite(target_id, target_id_buf, PRINTID_FASTA_LONG,
BUFFER_LEN);

        fprintf(outfp, "%s:%ld-%ld\t%s:%ld-%ld\n",
            query_id_buf, (long) SeqAlignStart(seqalign, 0), (long)
```

```
SeqAlignStop(seqalign, 0),
            target_id_buf, (long) SeqAlignStart(seqalign, 1),
(long)     SeqAlignStop(seqalign, 1));
        seqalign = seqalign->next;
    }
return;
}
```

Note that:

1. SeqAlignId gets the sequence identifier for the zero-th identifier (zero offset). This is actually a C structure.
2. SeqIdWrite formats the information in query_id into a FASTA identifier (e.g., gi|129295) and places it into query_buf.
3. SeqAlignStart and SeqAlignStop return the start values of the zero-th and first sequences (or first and second).

All of this is done by high-level function calls, and it is not necessary to write low-level function calls to parse the ASN.1.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. J Mol Biol. 1990;215:403–410. PubMed PMID: 2231712.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402. PubMed PMID: 9254694.

Chapter 17. LinkOut: Linking to External Resources from Entrez Databases

Kathy Kwan

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The power of linking is one of the most important developments that the World Wide Web offers to the scientific and research community. By providing a convenient and effective means for sharing ideas, linking helps scientists and scholars promote their research goals.

LinkOut is a powerful linking feature of the Entrez search and retrieval system (Chapter 15). It is designed to provide Entrez users with links from database records to a wide variety of relevant online resources, including full-text publications, biological databases, consumer health information, and research tools. (See Sample Links for examples of LinkOut resources.) The goal of LinkOut is to facilitate access to relevant online resources beyond the Entrez system to extend, clarify, or supplement information found in the Entrez databases. By branching out to relevant resources on the Web, LinkOut expands on the theme of Entrez as an information discovery system.

How Is LinkOut Represented in Entrez?

Any Entrez database record, e.g., a nucleotide sequence, a taxonomic record, a protein structure, or a PubMed abstract, can be linked to Web resources external to NCBI via LinkOut. The LinkOut homepage contains up-to-date documentation about LinkOut. The page that lists LinkOut resources associated with a given record can be accessed in a variety of ways (Figures 1–3). In the case of PubMed, the full-text article and other resources related to the abstract being viewed may be accessed directly by icon buttons above the abstract (Figure 2). The LinkOut display can be customized by using the LinkOut preferences in the Cubby.

How Does LinkOut Work?

Design Overview

The URLs to LinkOut resources are all provided by the person or organization that owns or created the resource. Links can be provided in any URL syntax, and providers of links may choose as much or as little access to their resource as they wish. Providers use one format to submit links to all Entrez databases.

LinkOut is in itself an Entrez database that holds all the linking information to external resources. The separation of the data records (e.g., PubMed abstracts) from the external linking information (e.g., URLs to journal articles on a publisher's Web site) enables both

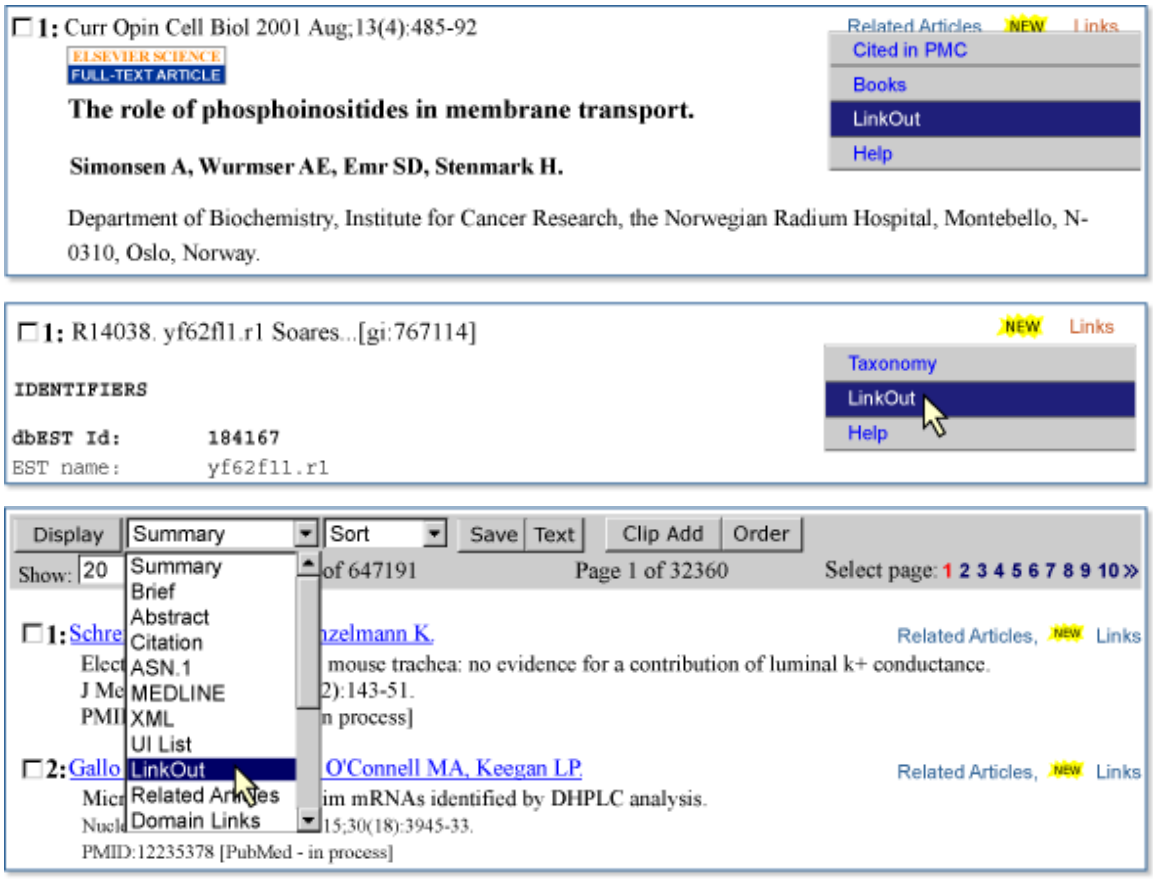


Figure 1. The LinkOut display can be accessed by selecting LinkOut from a PubMed record (*top panel*), from other Entrez databases (*middle panel*), or from the Display list (*lower panel*).

the external link providers and NCBI to manage linking in a flexible manner. This means that if links to external resources change, such as in the case of a Web site redesign, this will not affect the Entrez database records, and linking information can be updated as frequently as necessary.

The LinkOut database contains information on the relationship between a link and all of the applicable unique Entrez ID numbers (UIDs). By taking advantage of the interconnectivity among Entrez nodes, the linking information is presented seamlessly and efficiently.

LinkOut DTD and XML Files

LinkOut information is submitted in XML, defined by the LinkOut Document Type Definition (DTD).

Linking information is supplied in two elements: the Provider element, which specifies information about a link provider; and the LinkSet element, which describes information

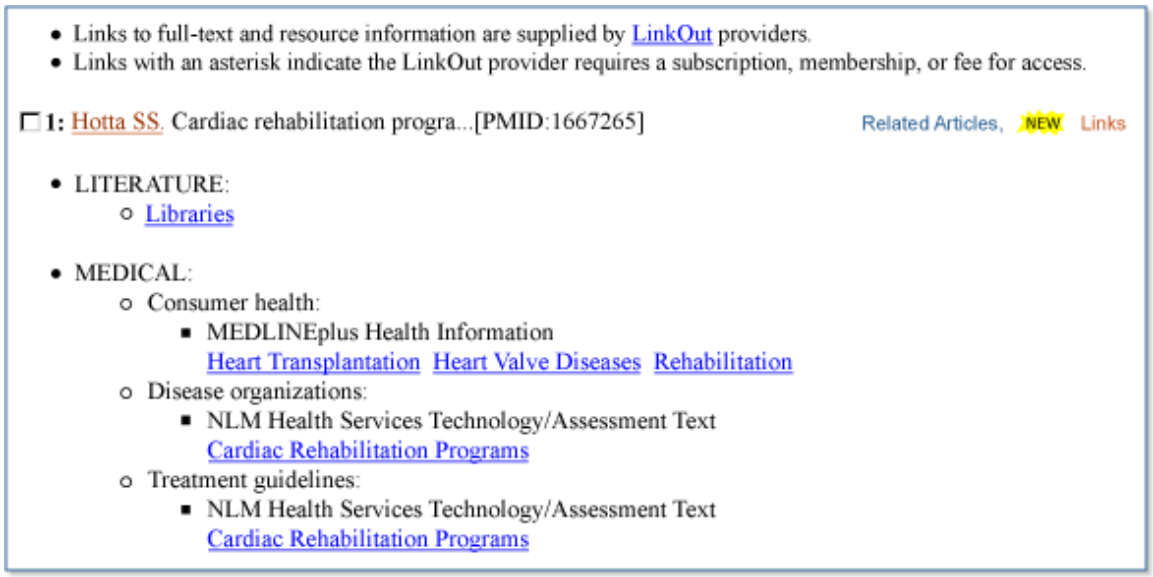


Figure 3. Links to external resources are listed in the LinkOut Display of an Entrez record.

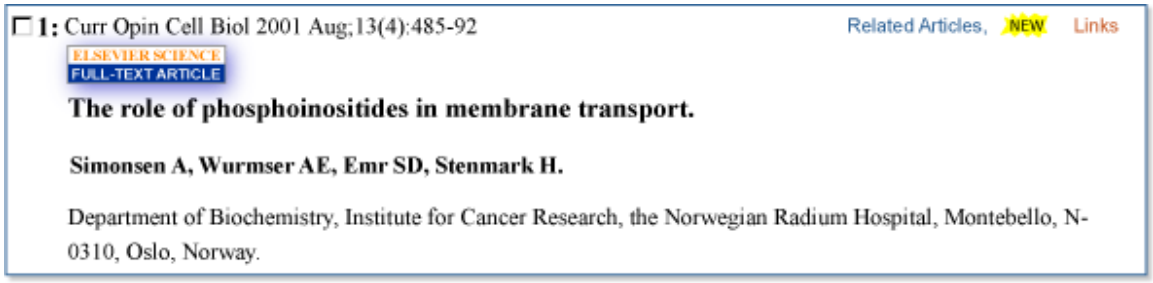


Figure 2. From PubMed, the links to the full text of research articles are also managed by LinkOut and can be accessed through an icon from PubMed Abstracts, highlighted here in *purple*, as well as from the associated list of LinkOut resources in Figure 3.

about the link. Each element should be submitted to NCBI in a separate file. Identity files contain the Provider element, and Resource files contain the LinkSet element.

The Identity file is always called providerinfo.xml. It describes the identity of a provider, including an ID (ProviderId) and an abbreviated name (NameAbbr) assigned by NCBI, the provider's name, and other general information about the provider. There should be only one providerinfo.xml file for each provider (see Box 1 for an example of an Identity file).

The Resource file, which contains the LinkSet information, specifies a set of Entrez records with a valid Entrez query, a specific rule to build the link to an external resource, and description of the resource using the SubjectType, Attribute, and UrlName fields. There is no standard for naming the LinkSet files, except that they must use the .xml

extension. There may be any number of LinkSet files associated with a ProviderId. (See Box 2 for an example of a Resource file.)

Terms used in SubjectType and Attribute elements are controlled to describe LinkOut resources in a systematic manner. This is because resources are presented to users by SubjectType on the LinkOut display page (and within the Cubby system), making it easier to browse and access available resources. Attributes can be used to describe the nature of a LinkOut resource (i.e., whether the resource requires a subscription or registration to access the content). A short text string may be used in the UrlName element to describe a resource. UrlName is typically used when the allowed SubjectType and Attribute terms cannot describe the resource adequately or when multiple links are available from one provider for a single Entrez record.

Box 1. Example of an identity file.

```
<?xml version="1.0"?>
<!DOCTYPE Provider PUBLIC "-//NLM//DTD LinkOut 1.0//EN" "LinkOut.dtd">
<Provider>
    <ProviderId>777</ProviderId>
    <Name>WebDatabase Co.</Name>
    <NameAbbr>WebDB</NameAbbr>
    <SubjectType>gene/protein/disease-specific</SubjectType>
    <Attribute>registration required</Attribute>
    <Url>http://www.webdatabase.com</Url>
    <IconUrl>http://www.webdatabase.com/images/webdb.gif</IconUrl>
    <Brief>On-line publisher of biomedical databases and other Web
resources</Brief>
</Provider>
```

Identity File Elements

Provider: root element of the identity file.

ProviderId: unique ID assigned by NCBI.

Name: full name of the resource provider.

NameAbbr: short, one-word name of the provider assigned by NCBI. May only include alpha and numeric characters; spaces and special characters such as hyphens are not allowed.

SubjectType, Attribute: descriptions of the resources and relationship of the provider to the resources listed in the resource file. SubjectType and Attribute values appearing in the identity file will apply to all of the resources listed by that provider.

Url: URL of the provider's Web site, used in the LinkOut Providers list in Cubby.

IconUrl: logo of the provider, used to display the link from Entrez records.

Box 1 continues on next page...

Box 1 continued from previous page.

Brief: short (up to 256 characters) description of the provider.

Box 2. Example of a resource file.

```
<?xml version="1.0"?>
<!DOCTYPE LinkSet PUBLIC "-//NLM//DTD LinkOut 1.0//EN" "LinkOut.dtd"
[<!ENTITY icon.url "http://www.webdatabase.com/images/webdb.gif">
<!ENTITY base.url "http://www.webdatabase.com/cgi-bin/elegans?">]>
<LinkSet>
    <Link>
        <LinkId>1</LinkId>
        <ProviderId>777</ProviderId>
        <IconUrl>&icon.url;</IconUrl>
        <ObjectSelector>
            <Database>Nucleotide</Database>
            <ObjectList>
                <Query>Caenorhabditis elegans [orgn]</Query>
            </ObjectList>
        </ObjectSelector>
        <ObjectUrl>
            <Base>&base.url;</Base>
            <Rule>an_lookup=&lo.pacc;</Rule>
            <UrlName>Caenorhabditis elegans</UrlName>
            <SubjectType>organism-specific</SubjectType>
        </ObjectUrl>
    </Link>
</LinkSet>
```

Resource File Elements

LinkSet: the root element of the resource file.

Link: an element that describes a specific set of resources grouped together by access characteristics or for convenience. A resource file may have multiple Link elements.

LinkId: an identifier assigned by the provider for its own reference. It may be any character string. Each Link should have a unique LinkId within each LinkSet or file.

ProviderId: the identifier number assigned to the provider by NCBI and listed in the providerinfo.xml file.

IconUrl: the URL to the icon that will be displayed on the PubMed Citation and Abstract Displays.

ObjectSelector: an element containing sub-elements in which providers will specify which Entrez records are being linked from by a <Link> element.

Box 2 continues on next page...

Box 2 continued from previous page.

Database: a sub-element of <ObjectSelector>. Databases available for linking include: PubMed, Protein, Nucleotide, Genome, Structure, PopSet, Taxonomy, and OMIM.

ObjectList: a sub-element of <ObjectSelector> containing either the <Query> or <ObjectID> that specifies the Entrez records from which the resource will be linked.

Query: a sub-element of <ObjectList> that contains any valid Entrez search, used to select the Entrez records being linked from.

ObjId: a sub-element of <ObjectList> that contains an Entrez record unique identifier (UID).

ObjUrl: an element that contains the necessary information for the Entrez system to construct URLs to link to the provider's resources.

Base: a sub-element of <ObjUrl> that is the base of the URL for the provider's records.

Rule: a sub-element of <ObjUrl> that specifies the construction of the remainder of the URL, based upon the specification of systems where the resources reside.

UrlName: a short (two- or three-word) description of the link. This may be used when multiple links are available for a single Entrez record. This may also be used if the allowed terms in SubjectType and Attribute cannot meet the need of a provider.

SubjectType, Attribute: sub-elements of <ObjectUrl>, used to describe the subject(s) of the provider's resources, barriers (if any) to using the resources, and relationship of the provider to the resources listed in the resource file. The SubjectType(s) and Attribute(s) will be applied to the resources provided within a <Link>.

XML File Processing and Indexing

All links from Entrez are generated on a daily basis so that new or modified Entrez records will have accurate LinkOut resources connected to them. Once a day, all LinkOut files are parsed according to the LinkOut DTD, and the LinkOut database is rebuilt, relating the Entrez UIDs with the link information specified in the LinkSet XML files.

A LinkOut record consists of a link and the associated information, including its URL and all descriptive terms (SubjectType, Attribute, and UrlName) pertaining to the link. The Entrez UIDs applicable to the link are indexed to associate this information to the corresponding Entrez databases. As explained in Chapter 15, LinkOut information is interconnected with all related Entrez records.

LinkOut Filters

To facilitate search and retrieval of LinkOut resources, there are a number of filters in the LinkOut-enabled Entrez databases. These filters, although not part of the LinkOut database, use the result generated in the LinkOut indexing process.

The filters are all prefixed with **lo**. Filters are available for all allowable SubjectType and Attribute terms and the NameAbbr of a provider. Some examples include:

- **loprov** LinkOut Provider
- **loattr** LinkOut Attribute
- **losubj** LinkOut SubjectType
- **loall** all LinkOut resources in an Entrez database

To use these filters to retrieve a set of Entrez records with LinkOut resources, the filter term can be entered as a search. For example, in PubMed, searching

```
"loattrfull text online"[Filter]
```

will retrieve all records with LinkOut resources that have an attribute "full-text online". The **Preview/Index** section in PubMed can also be used to select LinkOut filters by first selecting **Filter** and then typing in "lo" and selecting **Index** to browse through all of the filters related to LinkOut.

Guides for LinkOut Providers

LinkOut resources should be directly relevant to specific subjects of the Entrez records to which they will be linked, thus providing further research resources for Entrez users. The information and its delivery system should be of high quality and must not, through typographic or factual errors, omissions, or other flaws or inconsistencies, mislead, hinder, or frustrate the research efforts of Entrez users. The resources should be easy to use and navigate. Resources from professional societies, government agencies, educational institutions, or individuals and organizations that have received grants from major funding organizations are preferred.

Participation in LinkOut is voluntary. Providers need to submit two types of files to describe the LinkOut resources, Identity files and Resource files (see Boxes 1 and 2). These files include the necessary information for the Entrez system to construct an appropriate URL to access specific resources.

A list of Frequently Asked Questions is available to address questions that potential LinkOut providers may have. Current lists of LinkOut providers can also be browsed.

Submission Procedures

Step 1. Initial Contact. A prospective provider can write to linkout@ncbi.nlm.nih.gov, indicating interest in creating links from Entrez records to the providers' Web-accessible online resources. Please include the name, email address, and phone number of an

individual who will act as a designated contact. The email should also include a LinkOut Identity file (providerinfo.xml) based on the specifications described above.

Step 2. File Evaluation. NCBI staff will evaluate the resources before a ProviderId and NameAbbr are assigned. NCBI will also provide assistance with setting up an appropriate Resource file to describe the LinkOut resources.

Step 3. File Submission. An FTP account will be assigned to a provider for submission. Files must have been validated by the LinkOut Validation utility before uploading. Providers may transfer new versions of current files or add new Resource files at their own discretion. Providers are responsible for keeping their files current and valid. Links in Entrez databases are regenerated each day based on the files in each provider's directory; therefore, providers must delete obsolete files from their holdings directory.

Step 4. Representation in Entrez. Once a provider's LinkOut files are processed, the resources described in the file will be available in the LinkOut display of a relevant Entrez record as described in the above section, *How Is LinkOut Represented in Entrez?*. In PubMed, publishers of the abstract can choose to display a "button" on the Abstract and Citation displays of the PubMed record by adding the parameter "holding=NameAbbr" to the basic PubMed URL. For example, to activate the icon of WebDatabase Co, the URL would be provided as:

```
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=WebDB
```

Furthermore, multiple NameAbbr parameters may be used in a URL to activate more than one icon. For example, to display icons for both WebDB and MyDB, the following URL should be provided:

```
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=WebDB,MyDB
```

A provider's icon can be activated if the provider is selected from the LinkOut Preferences in Cubby.

All access restrictions will still apply. For example, if access to a database is limited by the user IP address, access will only be allowed via computers within an approved IP range; if access is password protected, the password must still be entered.

Detailed Guides

Interested parties can consult the following guides for more details:

- LinkOut and Non-Bibliographic Resources – written for providers of general LinkOut resources in all Entrez databases.
- LinkOut and Publisher Holdings – written for publishers and others that provide full-text links to PubMed records.
- LinkOut and Library Holdings – written for libraries to indicate information about their electronic full-text subscription and print holdings.

Auxiliary Tools

A number of tools are available to facilitate participation in LinkOut:

- Library LinkOut Files Submission Utility utility developed for Libraries to generate and manage their LinkOut files. Libraries simply check off their electronic journal collections from a list of journals that participate in LinkOut. With this utility, libraries can provide correct holdings information easily, and staff do not have to construct LinkOut files by hand.
- LinkOut File Validation utility to be used by providers of links to parse their LinkOut files, ensuring the accuracy of the files before submission. Besides validating the file syntax against the LinkOut DTD, this tool will ensure that only allowable SubjectType and Attribute terms have been provided.

Additional tools are being developed to assist other groups of providers. Interested parties can subscribe to announcement lists described in *Communicating with LinkOut Providers* (next section) to be informed of new developments.

Communicating with LinkOut Providers

LinkOut resource providers can communicate with NCBI's LinkOut team in a number of ways. Users and providers can write to linkout@ncbi.nlm.nih.gov to ask questions about LinkOut. There are also three announcement lists where development related to LinkOut will be communicated to link providers:

1. Linkout-news is for general announcements on LinkOut.
2. Library-linkout is for announcements on development related to library LinkOut participants.
3. Tax-linkout is for announcements relevant to linking to taxonomic resources on the Web.

Chapter 18. The Reference Sequence (RefSeq) Database

Kim Pruitt, Garth Brown, Tatiana Tatusova, and Donna Maglott

Created: October 9, 2002; Updated: April 6, 2012.

Summary

NCBI's Reference Sequence (RefSeq) database is a collection of taxonomically diverse, non-redundant and richly annotated sequences representing naturally occurring molecules of DNA, RNA, and protein. Included are sequences from plasmids, organelles, viruses, archaea, bacteria, and eukaryotes. Each RefSeq is constructed wholly from sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC). Similar to a review article, a RefSeq is a synthesis of information integrated across multiple sources at a given time. RefSeqs provide a foundation for uniting sequence data with genetic and functional information. They are generated to provide reference standards for multiple purposes ranging from genome annotation to reporting locations of sequence variation in medical records. The RefSeq collection is available without restriction and can be retrieved in several different ways, such as by searching or by available links in NCBI resources, including PubMed, Nucleotide, Protein, Gene, and Map Viewer, searching with a sequence via BLAST, and downloading from the RefSeq FTP site.

This chapter describes:

- The database content
- How data are assembled and maintained
- How RefSeqs can be accessed and retrieved

Introduction

NCBI's Reference Sequence (RefSeq) collection is a freely accessible database of naturally occurring DNA, RNA, and protein sequences. It is a unique resource because it provides a large, multi-species, curated sequence database representing separate but explicitly linked records from genomes to transcripts and translation products, as appropriate. Unlike the sequence redundancy found in the public sequence repositories that comprise the INSDC, (*i.e.*, NCBI's GenBank, the European Nucleotide Archive [ENA], and the DNA Data Bank of Japan [DDBJ]), the RefSeq collection aims to provide, for each included species, a complete set of non-redundant, extensively cross-linked, and richly annotated nucleic acid and protein records. It is recognized, however, that the coverage and finishing of public sequence data varies from organism to organism so intermediate genomic records are provided in some circumstances.

The non-redundant nature of the RefSeq collection facilitates database inquiries based on genomic location, sequence, or text annotation. Be aware, however, that the RefSeq

collection does include alternatively spliced transcripts encoding the same protein or distinct protein isoforms, in addition to orthologs, paralogs, and alternative haplotypes for some organisms, which will affect the outcome of a database query.

RefSeq records are based on sequence records submitted to the INSDC. However, the RefSeq collection is a distinct database. The public archival databases house sequences and annotations supplied by original authors and cannot be altered by others. The RefSeq collection differs from the archival databases in the same way that a review article differs from a related collection of primary research articles on the same subject. Each RefSeq record represents a synthesis, by a person or group, of the primary information that was generated and submitted by others. Other organizing principles or standards of judgment are possible, which is why the work is attributed to the synthesizing "editors". The RefSeq dataset is curated on an ongoing basis by collaborating groups and by NCBI staff. Sequence records are presented in a standard format and subjected to computational validation. The INSDC source of the RefSeq record, the curation status, and attribution to the curation group are also indicated.

The RefSeq collection establishes a useful baseline for integrating diverse data types, including sequence, genetic, expression, and functional information, into one consistent framework with a uniform set of conventions and standards. The RefSeq collection supports the following activities:

- genome annotation
- gene characterization
- comparative genomics
- reporting sequence variation, and
- expression studies

Database Content: Background

The May 2011 RefSeq collection (Release 47) includes sequences from more than 12,000 distinct taxonomic identifiers, ranging from viruses to bacteria to eukaryotes. It represents chromosomes, organelles, plasmids, viruses, transcripts, and more than 12.6 million proteins. Every sequence has a stable accession number, a version number, and an integer identifier (gi) assigned to it. Outdated versions are always available if a sequence is updated. RefSeq records can be distinguished from INSDC records by the inclusion of an underscore ("_") at the third position of the accession number. The RefSeq accession prefix has an implied meaning in terms of the type of molecule it represents, as outlined in Table 1.

Table 1. RefSeq accession numbers and molecule types.

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NZ_[b]	Genomic	Complete genomes and unfinished WGS data
NM_	mRNA	Protein-coding transcripts (usually curated)
NR_	RNA	Non-protein-coding transcripts
XM_[c]	mRNA	Predicted model protein-coding transcript
XR_[c]	RNA	Predicted model non-protein-coding transcript
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	Annotated on genomic molecules without an instantiated transcript record
XP_[c]	Protein	Predicted model, associated with an XM_ accession
WP_	Protein	Non-redundant across multiple strains and species

[a] Whole Genome Shotgun sequence data.
[b] An ordered collection of WGS sequence for a genome.
[c] Computed.

Updates

RefSeq updates are provided daily. These include new records added to the collection, and records updated to reflect sequence or annotation changes, including complete re-annotation of a genome. New and updated records are made available in Entrez and BLAST databases as soon as possible. The RefSeq FTP site also provides daily update information.

Flat File Format and Annotated Features

RefSeq records appear similar in format to GenBank records. Attributes novel to RefSeq records include a unique accession prefix followed by an underscore (Table 1) and a **COMMENT** field that indicates the RefSeq status and the INSDC source of the sequence information (Figures 1A, 1B, 1C, and 1D). For human RefSeqs, the **COMMENT** field also indicates whether the RefSeq is a reference standard from the RefSeqGene project. Some RefSeq records may include feature annotations or database cross-references (db_xrefs) that are not seen in the underlying INSDC record. This annotation is provided by computation and by manual curation. For example, nucleotide variation, STS, and tRNA features are computed for a subset of RefSeq entries using the data available in dbSNP

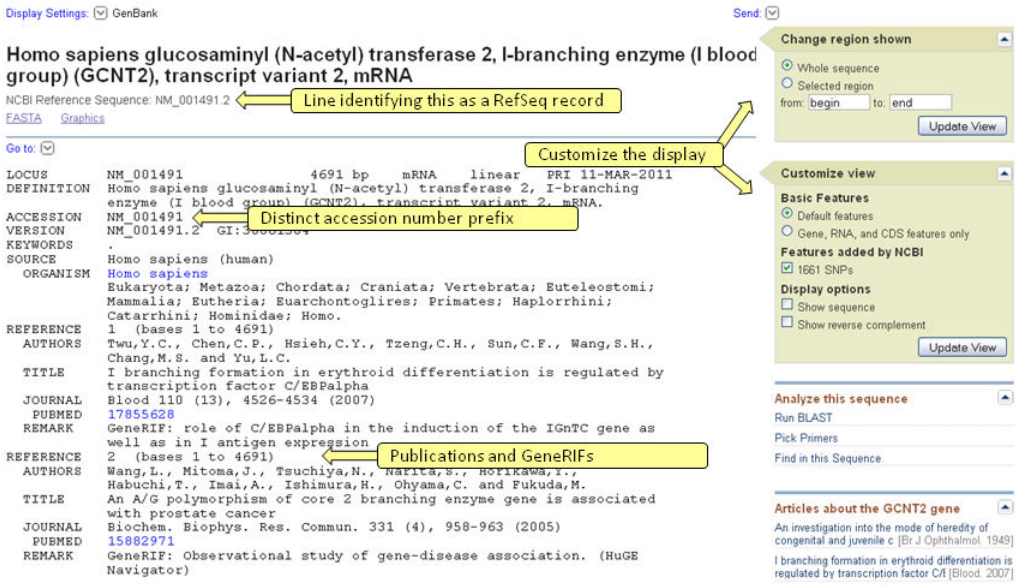


Figure 1A. Features of a RefSeq record. The beginning of a RefSeq record when displayed in the GenBank flat file format is shown.

(Chapter 5), UniSTS, and through tRNA-scan prediction (Lowe and Eddy, 1997). For human and mouse, exon feature annotation is also calculated for RefSeq transcript and non-transcribed pseudogene records. Db_xrefs provide links to Gene, nomenclature authorities, such as the HUGO Gene Nomenclature Committee (HGNC) for human RefSeq records, and to the Consensus CDS (CCDS) project. RefSeq proteins also report conserved domains computed by NCBI's Conserved Domain Database (Chapter 3). Additional protein features are propagated from the corresponding UniProtKB/Swiss-Prot records for a subset of species. Other nucleotide and protein features, publications, and comments may be added by collaborating groups or NCBI staff.

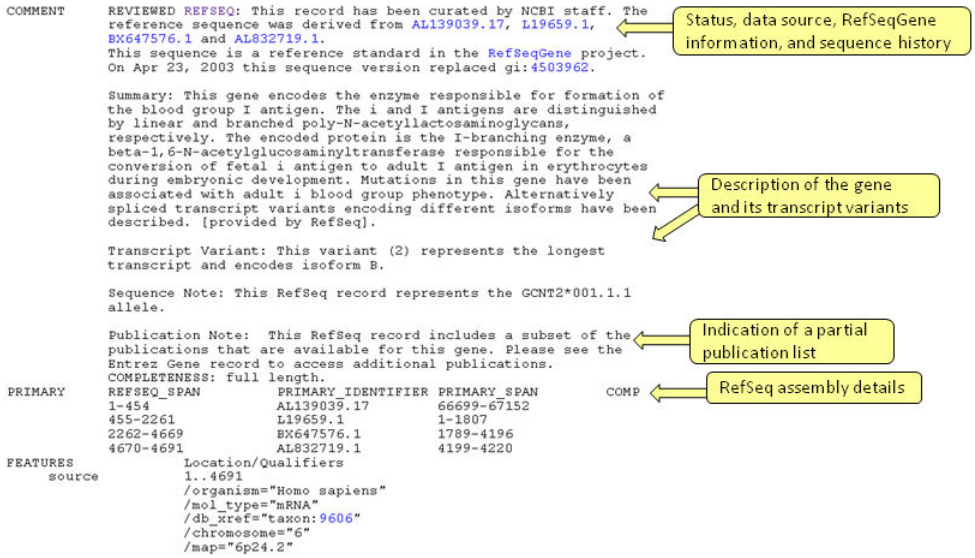


Figure 1B. The COMMENT and PRIMARY sections. The gene Summary is provided for RefSeqs with a **REVIEWED** status only. The PRIMARY block, providing the RefSeq assembly details, is displayed for vertebrate records predominantly.

Figure 1C. The FEATURES section. Only a subset of the available feature annotation is shown.

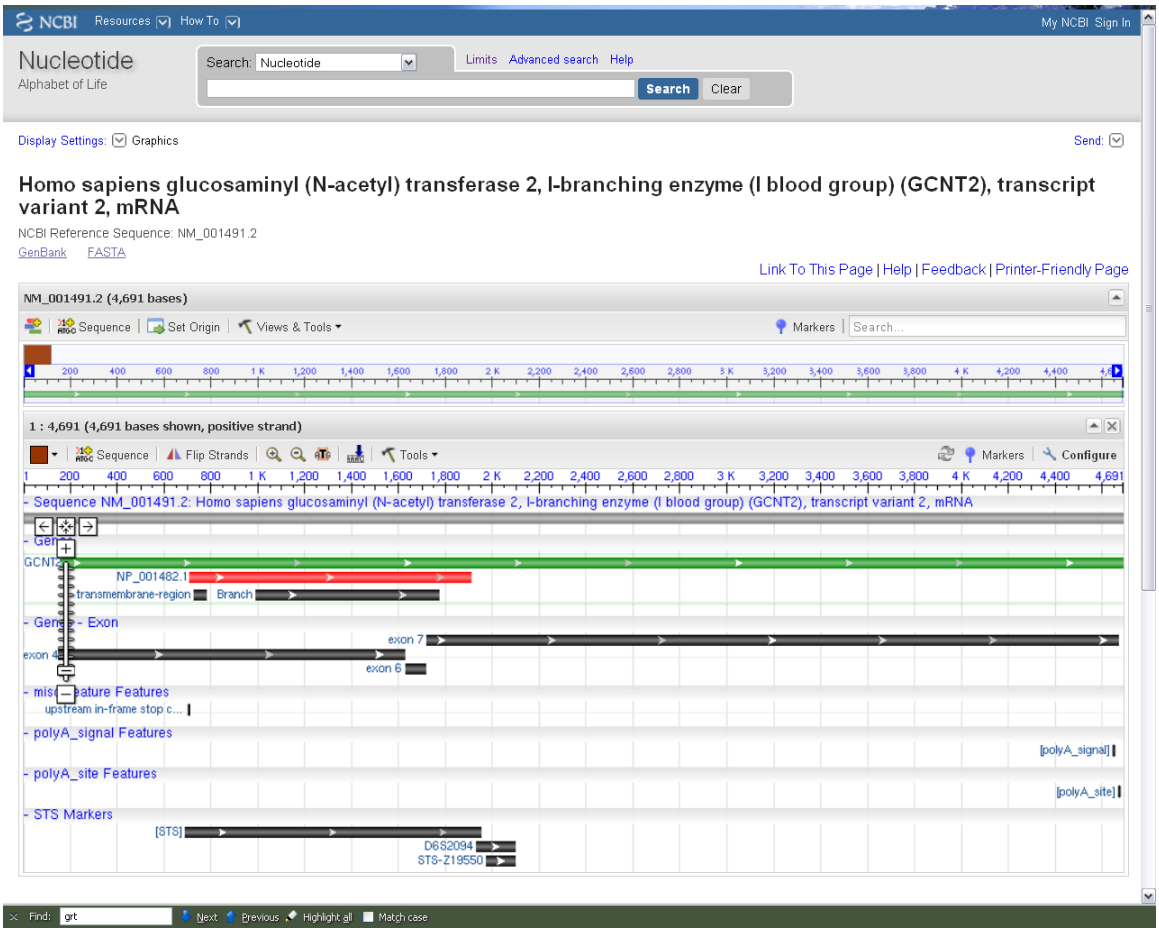


Figure 1D. NCBI's Sequence Viewer. The annotated features on a RefSeq record can be displayed in a graphical format (note the link 'Graphics' in Figure 1A). The display can be modified by following the 'Configure' link. The Help document provides additional information about the display and includes the Graphical View Legend, which provides details on how features are rendered.

Table 2. RefSeq status codes.

Code	Description
MODEL	The RefSeq record is provided by the NCBI Genome Annotation pipeline and is not subject to individual review or revision between annotation runs.
INFERRED	The RefSeq record has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.
PREDICTED	The RefSeq record has not yet been subject to individual review, and some aspect of the RefSeq record is predicted.
PROVISIONAL	The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff.

Table 2. continues on next page...

Table 2. continued from previous page.

Code	Description
REVIEWED	The RefSeq record has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature. Some RefSeq records may incorporate expanded sequence and annotation information.
VALIDATED	The RefSeq record has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review at which time additional functional information may be provided.
WGS	The RefSeq record is provided to represent a collection of whole genome shotgun sequences. These records are not subject to individual review or revisions between genome updates.

Assembling and Maintaining the RefSeq Collection

Summary

The RefSeq collection is the result of data extraction from INSDC submissions, curation, and computation, combined with extensive collaboration with authoritative groups. Each molecule is annotated as accurately as possible with the organism name, strain (or breed, ecotype, cultivar, or isolate), gene symbol for that organism, and informative protein name. Collaborations with authoritative groups outside of NCBI provide a variety of information, including curated sequence data, nomenclature, feature annotations, and links to external organism-specific resources. When no collaboration has been established, NCBI staff assembles the data from the INSDC submission. Each record has a **COMMENT**, indicating the level of curation that it has received (Table 2), and attribution of the collaborating group. Thus, a RefSeq record may be an essentially unchanged, validated copy of the original INSDC submission, or include updated or additional information supplied by collaborators or NCBI staff.

If multiple INSDC submissions represent the same molecule for an organism, the "best" sequence is chosen to represent as the RefSeq record. Known mutations, sequencing errors, cloning artifacts and erroneous annotation are avoided. Sequences are validated to confirm that the genomic sequence corresponding to an annotated mRNA feature matches the mRNA sequence record, and that coding region features translate into the corresponding protein sequence.

Working groups using distinct process pipelines compile the RefSeq collection for different organisms (Figure 2). RefSeq records are provided via several distinct approaches including:

- collaboration
- extraction from GenBank
- computational genome annotation pipeline
- curation by NCBI staff

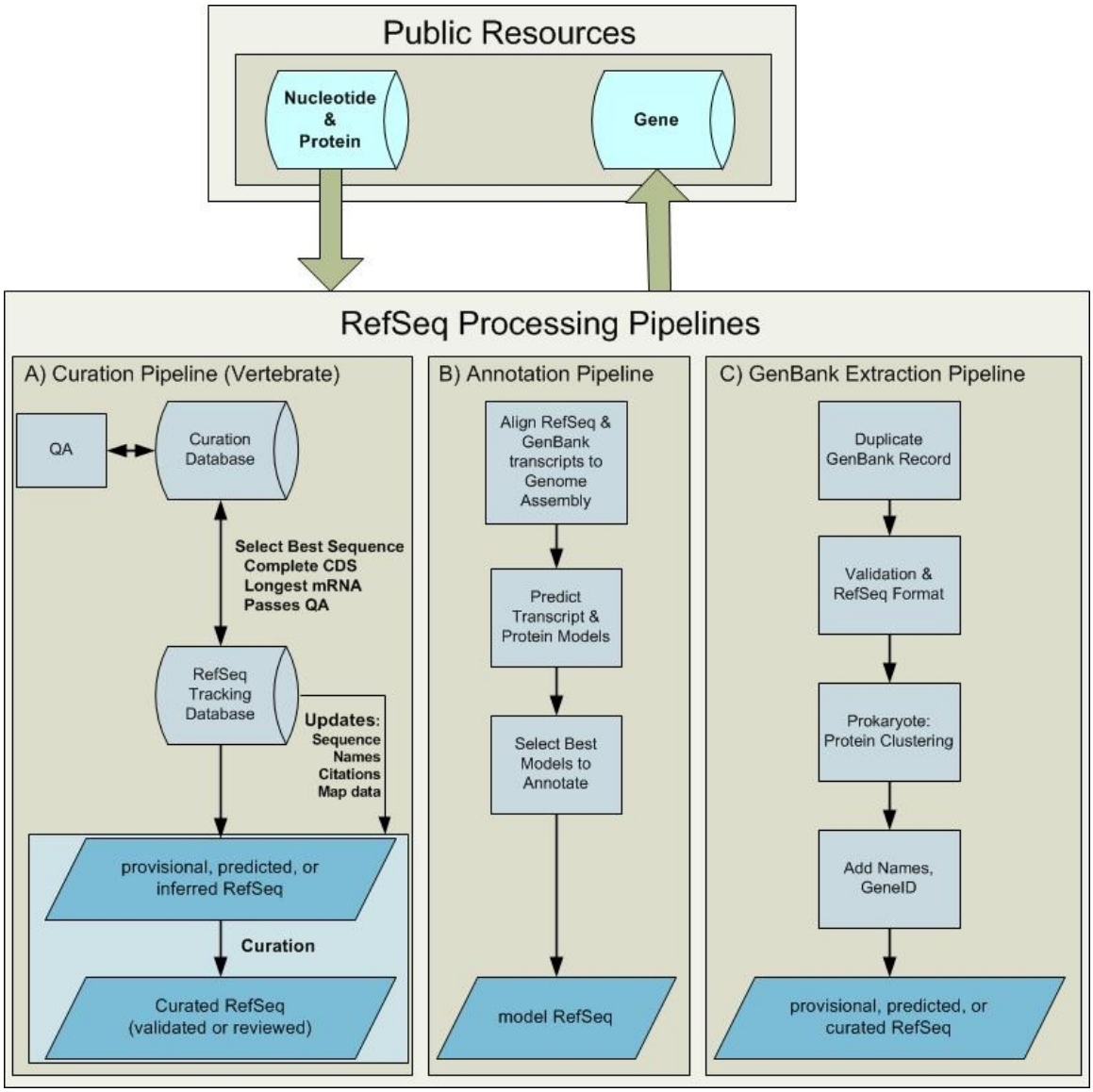


Figure 2. RefSeq Processing Pipelines. Sequence data deposited in the public archival databases is available for RefSeq processing. Processing pipelines include the vertebrate curation pipeline, the computational genome annotation pipeline, and extraction from GenBank. These pipelines generate new and updated RefSeq records that become publicly available in Entrez Nucleotide, Protein, and Gene databases. (A) Once a gene is defined and associated with sufficient sequence information in an internal curation database, it can be pushed into the RefSeq pipeline. The RefSeq process is initiated by selecting the longest mRNA annotated with a complete coding sequence for each locus. This RefSeq record has a status of **PROVISIONAL**, **PREDICTED**, or **INFERRED**. Subsequent curation may result in a sequence or annotation update and a RefSeq status of VALIDATED or REVIEWED. Records are updated if the underlying INSDC submission is updated or if other associated data are updated, including nomenclature, publications, or map location. (B) Available RefSeq and INSDC data are aligned to an assembled genome, *ab initio* gene prediction that uses the alignment data is performed, and an analysis program integrates all available data to define the annotation models. New **MODEL** RefSeq records are generated by this pipeline. (C) When a complete, annotated genome becomes available in the INSDC, a set of corresponding RefSeq records are generated by duplicating the GenBank records, followed by validation and addition of cross-references to Gene (via a db_xref citing the GeneID) and more informative and standardized protein names, when available.

Collaboration

RefSeq welcomes collaborations with authoritative groups outside of NCBI that are willing to provide sequences, nomenclature, annotation, or links to phenotypic or organism-specific resources. The RefSeq feedback form can be used to provide corrections or to initiate collaboration. The extent of collaboration may vary. For some species, the sequences and annotation of the entire RefSeq collection is provided by a collaborating authoritative group (see Table 3 for examples). For others, most notably the human and mouse RefSeq collections, numerous collaborations with individual scientists contribute to the representation of specific genes or complete gene families. Nomenclature for human and mouse is also provided via collaboration with the HUGO Gene Nomenclature Committee (HGNC) and the Mouse Genome Informatics group (MGI), respectively; Table 4 provides additional examples. Other collaborations extend across entire sets of organisms; for example, a board of Viral Genomes Advisors supports curation of the viral RefSeq collection. Thus, RefSeq records may contain information provided by an external authoritative source and/or analyses and curation at NCBI. The collaborating group is identified on the record.

Processing of RefSeq records supplied entirely by an external group is largely automated. The sequence and/or annotation is periodically submitted, validated to detect conflicts in the annotation, and modified slightly to format the submission as a RefSeq record, including addition of db_xrefs to Gene. NCBI staff do not directly curate the annotation or modify the sequence of RefSeq records provided by collaborating groups. Any problems identified by the validation process or by the scientific community are reported to the submitting group, and any update made to the annotation or sequence is reflected in a future RefSeq release.

Table 3. Examples of collaborators who contribute RefSeq records.

Organism	Collaborator
Saccharomyces cerevisiae	Saccharomyces Genome Database (SGD)
Arabidopsis thaliana	The Arabidopsis Information Resource (TAIR)
Pseudomonas aeruginosa	*Pseudomonas aeruginosa* Community Annotation Project (PseudoCAP)
Drosophila melanogaster	FlyBase
multiple invertebrates	VectorBase

Table 4. Examples of collaborating groups

FlyBase
HUGO Gene Nomenclature Committee (HGNC)
Microbial genomes
Mouse Genome Informatics (MGI)

Table 4. continues on next page...

Table 4. continued from previous page.

Online Mendelian Inheritance in Man (OMIM)
Rat Genome Database (RGD)
VectorBase
Viral Genome Advisors
XenBase
Zebrafish Information Network (ZFIN)

Extraction from GenBank records

Complete genome data for viruses, organelles, prokaryotes, and some eukaryotes is propagated to RefSeq records from the whole genome sequence data and annotation available in GenBank (also in the ENA and DDBJ public archives). Generally, an initial validation step is performed before the RefSeq record is made public. The resulting RefSeq record is a copy of the GenBank submission but may contain some additional annotations as a result of the validation step. In particular, transcripts are provided as separate RefSeq records for most eukaryotic organisms; the GenBank submission of the genome sequence from which the RefSeq record is propagated instantiates the protein only, not the transcript.

This process flow is supported by the BioProject and Genome databases. The BioProject database tracks the status of whole-genome sequencing projects submitted to GenBank, other types of large-scale projects, and provides an overview of the organism and links to data and other resources. The resulting genomic RefSeq data is represented in the Genome database, which includes bacteria, archaea, eukaryotes, viroids, viruses, plasmids, and organelles. The Genome website provides custom displays, analysis, and tools for prokaryotic and some eukaryotic genomes (see Table 5).

Note that processing of most eukaryotic genomes is more complex, requires more than basic extraction from GenBank, and occurs independently, largely because the volume of data is significantly greater.

Extraction of GenBank whole genome data for processing into RefSeq records falls into four primary categories: chromosomes, microbial genomes, small complete genomes, and targeted loci.

Table 5. Selected Entrez Genome resources.

Web Page	Web Site
Genome homepage	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Eukaryotes	http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi

Table 5. continues on next page...

Table 5. continued from previous page.

Web Page	Web Site
Prokaryotes	`http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi`
Viral Genomes	http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239
Organelles	http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html
Plant Genomes	http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html

Chromosomes

Complete chromosome sequence assembled from individual clones (that are themselves available from the INSDC) is propagated into a RefSeq record. For some genomes, the RefSeq representation uses a unit of interest to the research community; for example, some of the RefSeq genomic records for *Drosophila melanogaster* represent chromosome arms rather than complete chromosomes. RefSeq records may also be available for some genomes that are not yet fully sequenced but for which complete sequence is available for individual chromosomes. These complete chromosome RefSeq records may be annotated by the NCBI computational annotation pipeline, or they may be curated by an organism-specific collaborating group and undergo NCBI validation before being released.

Microbial genomes

For microbial species, historically all complete and draft genomes submitted to GenBank were propagated to the RefSeq collection. This is no longer tenable because of the volume of genomic data being generated, so additional RefSeq records are created from new GenBank submissions only to span the taxonomic diversity; this means in general, one genomic RefSeq per species is provided. If significant sequence diversity exists, or if subspecies or subgroups require representation as determined by NCBI staff, more than one RefSeq may exist for a given species.

Small complete genomes

RefSeq records representing organelle, viral, and plasmid genomes are based on single GenBank records. For organelle and viral genomes, if more than one GenBank submission is available for a species, typically only one is chosen to propagate to the RefSeq collection. Various factors, including the level of annotation, strain information, and community input are considered when deciding which GenBank submission to represent. There is no plasmid taxonomy; a GenBank submission is propagated to the RefSeq collection if it is part of a larger registered genome sequencing project, or if it exhibits significant sequence divergence when compared to other plasmids.

Targeted loci

The RefSeq Targeted Loci Project is a collaborative effort to curate and maintain molecular markers of use in the identification and classification of organisms. The initial

focus is on ribosomal RNAs, although expansion to other informative sequences is anticipated. From GenBank submissions, the project creates RefSeq records for the small subunit of ribosomal RNA (16S in prokaryotes and 18S in eukaryotes) and the large subunit ribosomal RNA (23S in prokaryotes and 28S in eukaryotes). As of November 2010, there are 3331 16S rDNA RefSeq records from bacteria and archaea and 137 18S rDNA, and 97 28S rDNA RefSeq records from fungi.

Computational Genome Annotation Pipeline

NCBI computes annotation of genomic sequence data for some genomes including some microbes, vertebrates (*e.g.*, human, mouse, rat, cow, and zebrafish, and others) and invertebrates (*e.g.*, honey bee, acorn worm, and pea aphid). The annotation pipeline is automated and yields genomic, transcript, and protein (when appropriate) RefSeq records. Names annotated on the transcript and protein products are based on sequence similarity. Annotation data are refreshed periodically, and records generated from this process flow are not curated or updated between annotation runs (see Chapter 14 for more information on the eukaryotic genome annotation pipeline; information about NCBI's prokaryotic annotation pipeline is also available). For some species, including human, RefSeq records may be provided by a mixture of methods. In other words, there may be a set of curated transcript and protein records (see the following section) in addition to a set of records generated computationally. RefSeq records that are processed by NCBI's pipelines are displayed in the NCBI Map Viewer (Chapter 20), included in Gene, and are available in NCBI's sequence databases.

Curation by NCBI Staff

A portion of the RefSeq dataset is curated by NCBI staff. This subset includes viral, mitochondrial, vertebrate, and some invertebrate organisms. Most bacterial, plant, and fungal records are provided either by collaboration or by processing the annotated genome data submitted to the INSDC; however, a small number of bacterial genomes are annotated and curated by NCBI staff.

Curation of Microbial, Viral, and Mitochondrial RefSeqs

Microbial, viral, and metazoan mitochondrial RefSeq records are validated for content propagated from the original GenBank submission, including taxonomy, publications, and annotation, prior to becoming public. This content may be modified, augmented, or deleted by NCBI curation staff.

For microbial genomes, a set of minimal annotation standards (described here) are automatically provided on all legacy and new RefSeq records. These include ribosomal RNAs, transfer RNAs, and protein-coding genes with locus_tags. Ribosomal RNAs are predicted using BLASTn tools against an RNA sequence database and/or using Infernal (Eddy, 2002) and Rfam models (Griffiths-Jones, et al, 2003). Transfer RNAs are predicted using tRNAscan-SE (Lowe and Eddy, 1997). Other annotation above the minimum standards may be added based on an external source or literature review. Annotation

associated with the NCBI's Protein Clusters database is also propagated to the RefSeq records (both proteins and genes) at selected intervals. The Protein Clusters database is a collection of RefSeq proteins from complete genomes broadly organized into the following groups: archeal and bacterial genomes and plasmids, viruses, protists, plants, and chloroplasts and mitochondria, and annotated based on sequence similarity and protein function. This clustering allows the entire group to be curated as a single set, permitting well characterized proteins to seed the annotation of less studied ones within the same cluster. NCBI staff use literature and information from other databases, including UniProtKB/Swiss-Prot, to annotate each cluster with standardized protein names, biochemical descriptions, and other data, which is then transferred to individual proteins within the relevant RefSeq records. A microbial genome RefSeq record typically has a **PROVISIONAL** review status.

Annotation of viral genomes relies on an established group of Viral RefSeq Genome Advisors, members of the International Committee on the Taxonomy of Viruses, and other experts outside of NCBI. For example, the HIV-1 RefSeq (NC_001802) was curated by NCBI staff in collaboration with the authors of the book Retroviruses, and many of the adenovirus and herpesvirus records have been curated by outside experts. Based on literature review, NCBI curators may modify the CDS and RNA annotation compared to the GenBank submission, as was done for the Measles virus RefSeq record (NC_001498). Additional NCBI resources used during the curation of viral RefSeq records include the Protein Clusters database and PASC, a virus classification tool used to validate the taxonomy of virus RefSeq records across a number of taxonomic families. NCBI also maintains several specialized annotation pipelines for use in the Virus Variation and Influenza Virus resources. Manually curated viral RefSeq records are annotated with a status of **REVIEWED** or **VALIDATED** in the RefSeq COMMENT block.

For metazoan mitochondrial RefSeq records, standardized protein, gene, and RNA names are annotated independent of species-specific nomenclature guidelines. Additional curation may include adding common names or missing tRNAs and adjusting the coding region spans based on the Protein Clusters database. Curated metazoan mitochondrial records are annotated with a status of **REVIEWED**. Non-metazoan and plant chloroplast RefSeq records are not curated, are derived entirely from the original INSDC submission, and have a status of **PROVISIONAL**.

For targeted loci, vector or primer sequence from the GenBank submission is excluded from the RefSeq record. Any feature annotation may be modified to represent a standard format, and collection identifiers and publications referencing the original GenBank submission may be added.

Curation of Vertebrate and Invertebrate Records

Curation of higher eukaryotic organisms is focused on mammalian genomes, especially human and mouse, but also includes many other species with existing or planned genome assemblies. The RefSeq processing for these organisms provides transcripts and protein records as well as some genomic region records representing gene clusters or

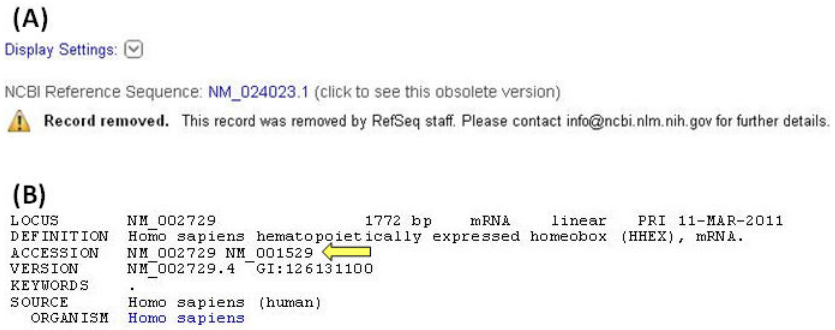


Figure 3. Suppressed or redundant RefSeq records. (A) A standard text statement is included on the Entrez document summary for suppressed RefSeq records. (A) If redundant RefSeq records are merged, then both accession numbers appear on the flat file **ACCESSION** line (yellow arrow). The first **ACCESSION** number listed is the primary identifier and all others listed are "secondary" accession numbers.

pseudogenes; these genomic region records facilitate genome-wide annotation. Because RefSeq uses evidence independent of a genome assembly to represent RNAs and proteins, the dataset can represent sequence not currently part of that genome assembly. RefSeq processing integrates the official nomenclature and other information, including alternate names, Gene Ontology (GO) terms, and literature and GeneRIFs available in Gene. Multiple collaborations support the collection of this descriptive information (Table 4; see also Chapter 19).

Sequences enter RefSeq curation processing by a combination of computational analysis, collaboration, and in-house curation. As illustrated in Figure 2, generation of the initial RefSeq record depends on identifying a representative sequence for a gene. New genes and sequence data are added to the in-house version of the Gene database by RefSeq curators, collaborators, NCBI's genome annotation pipeline, and NCBI-based mining of UniGene, cDNA alignments, and INSDC submissions. Quality assessment (QA) processes are executed regularly to identify questionable data for review. These assessments include analysis of nomenclature, sequence similarity, genomic placement, and potential cloning

errors (*e.g.*, chimeras). The QA steps also leverage data from other NCBI resources, including HomoloGene, Map Viewer, and GenBank related sequences. Data conflicts must be resolved before the INSDC submission is used to generate a RefSeq record.

A sequence record unambiguously associated with a Gene record may be propagated into a RefSeq record. The completeness of the sequence (*e.g.*, complete vs. partial CDS) and the category of the gene (*e.g.*, protein coding, pseudogene) determine whether a RefSeq will be made, and if so, of what type (DNA, RNA, mRNA plus protein). RefSeq records are not made for incomplete proteins, transposable elements, or those loci for which the product type is uncertain (*e.g.*, protein coding or not). It should be noted, however, that the RefSeq collection does include partial transcripts and proteins that are provided by collaborating groups or when the RefSeq is based on an annotated whole genome sequence submitted to the INSDC.

Once a suitable "source" sequence is identified, the RefSeq record is generated using the sequence data from the INSDC submission and the annotation data from the in-house version of the Gene database. Information from Gene includes the GeneID, cross-references to other databases, official nomenclature, aliases, alternate descriptive names, map location, and citations, including those submitted as GeneRIFs. RefSeq records are also subject to programmatic validation to identify annotation format errors and to provide annotation in a more consistent format. Records at this stage have a **PROVISIONAL**, **PREDICTED**, or **INFERRED** status depending on the evidence existing in support of the Gene record.

RefSeq processing for non-protein-coding RNA loci uses the longest defining transcript record associated with the Gene record. For non-transcribed loci (such as non-transcribed pseudogenes), the RefSeq record is typically derived from a region of a larger genomic sequence. Curation of these types of records is minimal because the current focus is on curation of protein-coding loci; however, these records provide an important reagent for the computational annotation pipeline and support annotation of non-protein-coding genes that might otherwise be missed or misrepresented as a predicted protein-coding gene.

Other RefSeq records are provided to represent larger genomic regions, including RefSeqGene sequences, gene clusters, genes requiring rearrangement to express a product (immunoglobulins and T-cell receptors), and haplotypes with known differences in gene content. These genomic region records are annotated by NCBI curation staff, often in collaboration with scientific experts, and are not provided by automatic processing.

RefSeqGene, a partner of the international Locus Reference Genomic (LRG) collaboration, provides stable reference standard genomic, RNA, and protein RefSeqs for medically important genes. These standards support the HGVS expressions used to describe sequence variation in medical records, and thus are constructed to represent standard alleles. The RefSeqGene usually represents a single gene, on the positive strand of the sequence, beginning 5 Kb upstream and extending 2 kb downstream. RefSeqGene

records also include alignments of the RefSeq transcripts for the gene. All sequences annotated on the RefSeqGene have a review status of **VALIDATED** or **REVIEWED**.

Additional curation of vertebrate and some invertebrate RefSeq records occurs at the request of public users and collaborators, or as indicated by in-house QA analyses. QA analyses focus on, but are not restricted to, HomoloGene-based reporting of inconsistent protein lengths, identification of RefSeqs with repeat elements, questions about gene-to-sequence associations or potentially redundant genes, and reports of genes annotated at one time on a genome but not during subsequent re-annotation of that genome. Additionally, alignment-based tests are conducted for human and mouse that identify RefSeq records with poor quality alignment to the genome, non-consensus splicing, or very short or very long exons. Review of these records by skilled curators results in the most current and complete representation of the nucleotide and protein sequence and feature annotation available at that time. Sequence review may allow removal of vector and linker sequence, extension of the UTRs to define the full-length transcript, modification of the CDS annotation associated with the original INSDC source accession, or the creation of additional RefSeq records to represent the products of alternative splicing. A variety of feature annotations can be added to the RefSeq transcript and protein records. For nucleotide records, these include an indication of the transcript completeness, location of poly(A) signal and site, and sites of sequence variation and RNA editing. Exon annotation is provided for RefSeq transcripts and non-transcribed pseudogenes of human and mouse only; for transcripts, exon annotation is determined from the alignment of the transcript to the reference genome assembly using Splign, and, for non-transcribed pseudogenes, from the Splign alignment of the functional gene to the pseudogene genomic region. For protein records, feature annotations may include alternate or non-AUG initiating codons, Enzyme Commission (EC) numbers, mature peptide products, protein domains, and selenocysteine residues. Finally, literature review is another source of alternate names, aliases, and functional information, the latter which may be used to construct a Reference Sequence Summary on the RefSeq record. A RefSeq record that has undergone the complete review process has a **REVIEWED** status. Note that for many genes, intermediate levels of manual curation may address issues concerning the RefSeq sequence alone; these records have a review status of **VALIDATED** pending full review.

The review process may result in updating a RefSeq record, providing new RefSeq records, modifying sequence-to-gene associations, merging Gene records, or discontinuing a RefSeq, GeneID, or both. A RefSeq record is suppressed if it is found to represent a transcribed repeat element, to be derived from the wrong organism (*i.e.*, the INSDC sequence it was based on has incorrect organism annotation), or not to represent a "gene". Records determined to represent an incomplete sequence, such as a partial protein sequence or an incompletely spliced transcript, are temporarily suppressed until more complete sequence data are available. Suppressed records can still be retrieved and will have a disclaimer appearing on the query result document summary (Figure 3a). A suppressed record is not included in BLAST databases, in the calculation of related sequences, in the BLink display (BLink are pre-computed protein BLAST results), or in

RefSeq FTP releases. If a RefSeq is found to be redundant with another public RefSeq, then one is retained and the other becomes secondary (Figure 3b). If the sequences were associated with two different Gene records, then the records are merged so that a query of Gene with either of the original GeneIDs will retrieve the remaining single record.

We welcome input from the research community to improve the quality of the RefSeq collection. Interested parties are invited to contact us by sending an email to the NCBI Help Desk (info@ncbi.nlm.nih.gov) or by using our feedback form.

Access and Retrieval

RefSeq records can be accessed by direct query, BLAST, FTP download, or indirectly through links provided from several NCBI resources, including Gene, Genome, BioProject, and Map Viewer (Table 6). In addition, RefSeq records are included in some computed resources and so links may be found from those pages to individual RefSeq records. Some links from Entrez databases to RefSeq records are based on Gene associations (e.g., links from OMIM; Chapter 7), whereas others are based on sequence similarity or RefSeq annotation content, including links from PubMed. RefSeq records are easy to distinguish in these resources by their unique accession number format (Table 1).

How to access and retrieve RefSeq records is described below.

Table 6. NCBI resources with links to RefSeq records.

BioSystems	**Gene Expression Omnibus (Chapter 6)**
BLAST results (Chapter 16)	Genome
BLink (pre-computed BLASTp)	BioProject
Bookshelf (Chapter 8)	HomoloGene
Consensus CDS project	Map Viewer (Chapter 20)
dbSNP (Chapter 5)	Probe
dbVar	Protein Clusters
Entrez (Chapter 15)	PubMed Central (Chapter 9)
Epigenomics	UniGene (Chapter 21)
Gene (Chapter 19)	UniSTS

Entrez Query Access

RefSeq records can be retrieved from the Entrez system (Chapter 15) by querying with an accession number, symbol or locus_tag, name, or by using Entrez Limits and Property terms. All RefSeqs can be found in the Entrez Nucleotide or Protein databases; both RefSeq and INSDC submissions will be included but a filter is provided at the top right hand corner of the results page to allow display of only the RefSeq accessions, if desired. Filters can be configured using the MyNCBI interface. Alternatively, a query can be restricted to retrieve only RefSeq-specific results using the Limits page or by querying

with a Property, such as "srcdb_refseq[property]", or others listed in Table 7. Limits and Properties can also be used to restrict results to molecule type, such as DNA versus mRNA. The Entrez Help document provides additional information about querying.

Gene contains the majority of the RefSeq collection and also supports querying using all the above strategies. RefSeq-to-Gene connections are also provided by direct links; RefSeq records include a link to the Gene report page via the GeneID **db_xref** link on the gene and CDS features (Figure 1C). Gene reports the RefSeq accession numbers in the RefSeq section of the report, with links to the Nucleotide or Protein records. The Links menu in Gene also provides distinct links to RefSeq RNAs, RefSeq proteins, and RefSeqGene. Gene reports may include a graphical depiction of genome annotation data in the **Genomic regions, transcripts, and products** section, with links to Nucleotide and Protein displays. When this graphical section is provided, an additional report is available with details about exon and intron boundaries and length. You can change the display format from **Full Report** to **Gene Table** to access this report. Note that RefSeq records representing assembled environmental samples (with an NS_ accession prefix) are not included in Gene but can be found in the Genome and Nucleotide databases.

RefSeq records in the Genome or BioProject databases can be retrieved using an accession number for a complete genomic molecule (NC_ accession prefix) or organism name. The BioProject database can also be queried using the property restriction "srcdb_refseq[property]".

RefSeq records belonging to the RefSeqGene set can be retrieved from the Entrez system using "RefSeqGene[keyword]".

Table 7. Entrez queries to retrieve sets of RefSeq records.

Query	Accession prefix	RefSeq status retrieved
srcdb_refseq[prop]	All RefSeq accessions	All
srcdb_refseq_known[prop]	NC_, AC_, NG_, NM_, NR_, NP_, AP_	REVIEWED, PROVISIONAL, PREDICTED, INFERRED, and VALIDATED
srcdb_refseq_reviewed[prop]	NC_, AC_,NG_, NM_, NR_, NP_, AP_	REVIEWED
srcdb_refseq_validated[prop]	NC_, NM_,NR_,NP_	VALIDATED
srcdb_refseq_provisional[prop]	NC_, AC_, NG_, NM_, NR_, NP_, AP_	PROVISIONAL
srcdb_refseq_predicted[prop]	NM_, NR_, NP_	PREDICTED
srcdb_refseq_inferred[prop]	AC_, AP_, NM_,NR_,NP_	INFERRED
srcdb_refseq_model[prop][a]	NT_, NW_, XM_, XR_, XP_, ZP_	Genome annotation models

BLAST

RefSeq transcript records are included in the Nucleotide non-redundant (nr) and the RefSeq mRNA sequences databases. RefSeq protein records are included in the Protein database. Accessions in the results set, either RefSeq or GenBank, that are associated with a Gene record are indicated by a small blue **G** icon, which is linked to the Gene report. RefSeq genomic records (whole chromosome or scaffold RefSeq records and RefSeqGene records) are provided in the Reference genomic sequences database or via organism-specific genome BLAST databases, which can be accessed via Map Viewer, BioProject reports, or the Genomic Biology webpage. RefSeqGene records are also retrieved from the nr database in BLAST results and in a dedicated RefSeqGene database.

Map Viewer

The NCBI Map Viewer supports queries by RefSeq and RefSeqGene accession numbers if the annotated genome is available in that resource.

FTP

RefSeq data are available in three FTP areas:

- Configured RefSeq BLAST databases are available for download from the BLAST FTP site; separate databases are provided for genomic, transcript, and protein records.
- Organism-specific sequence files are provided in the Genomes FTP site. This area includes RefSeq records that are generated by, or used in, Map Viewer and Genomes processing. NCBI's annotation of genomic RefSeqs is also available; a file in the latest specification (version 1.20) of Generic Feature Format version 3 (GFF3) is provided in a GFF subdirectory for the latest assembly of many organisms.
- The full RefSeq collection, including the human RefSeqGene set, is available from the RefSeq FTP site, with the exception of the NS_ accession series environmental sample records. The RefSeq collection is provided as comprehensive bi-monthly releases in addition to daily updates for records that are new or updated between RefSeq release cycles. The comprehensive release provides data in multiple file formats, including flat file and FASTA, organized into primary taxonomic groups in addition to the complete dataset. For organisms with more frequent updates to curated records, including human and mouse, subdirectories containing weekly comprehensive releases of transcript and protein RefSeq records are provided also. Information about the RefSeq release is documented on the RefSeq FTP site in the release-notes subdirectory. The availability of new releases is announced on the RefSeq website, on NCBI's Facebook and Twitter accounts, to subscribers of the refseq-announce email list, and in the NCBI Newsletter.

Related Resources

The Consensus Coding Sequence (CCDS) Project

The CCDS project aims to provide a complete set of high quality annotations of protein-coding genes on the human and mouse genomes. It leverages the computational annotation pipelines of NCBI and Ensembl, and expert curation provided predominantly by the Havana team of the Wellcome Trust Sanger Institute and NCBI's RefSeq staff, to track identical protein annotations on the reference assemblies of the human and mouse genomes, and to ensure they are consistently and accurately represented in public resources. The CCDS set includes coding regions that are annotated as full-length (with an initiating AUG and valid stop-codon), can be translated from the genome without frameshifts, and use consensus splice-sites. Annotated genes in the CCDS set are associated with a unique identifying number and version. The version number will change with a change to the CDS structure or to the underlying genomic sequence, although any change requires collaborative agreement. See PubMed ID 19498102 for more information.

Related Reading

Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT; Mouse Genome Database Group. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucl. Acids Res. 2011;39:D842–8. (PubMed ID). PubMed PMID: 21051359.

Coffin JM, Hughes SH, and E Varmus. (1997) Retroviruses. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Dwinell MR, Worthey EA, Shimoyama M, Bakir-Gungor B, DePons J, Laulederkind S, Lowry T, Nigram R, Petri V, Smith J, Stoddard A, Twigger SN, Jacob HJ, Team RGD. The Rat Genome Database 2009: variation, ontologies and pathways. Nucl. Acids Res. 2009;37:D744–9. (PubMed). PubMed PMID: 18996890.

Eddy SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics. 2002;3:18. (PubMed ID). PubMed PMID: 12095421.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucl. Acids Res. 2003;31:439–441. (PubMed ID). PubMed PMID: 12520045.

Amberger, J., Bocchini, C. and Hamosh, A. (2011), A new face and new challenges for online mendelian inheritance in man (OMIM®). Human Mutation, 32:n/a. doi: 10.1002/humu.21466.. (PubMed ID). PubMed PMID: 21472891.

Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 1997;25:955–964. (PubMed ID). PubMed PMID: 9023104.

Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucl. Acids Res. 2011;39:D52–7. (PubMed ID). PubMed PMID: 21115458.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, Deweese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko MV, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucl. Acids Res. 2011;39:D225–9. (PubMed ID). PubMed PMID: 21109532.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2008;19(7):1316–1323. (PubMed ID). PubMed PMID: 19498102.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucl. Acids Res. 2009;37:D32–36. (PubMed ID). PubMed PMID: 18927115.

Tatusova TA, Karsch-Mizrachi I, Ostell JA. Complete genomes in WWW Entrez: data representation and analysis. Bioinformatics. 1999;15:536–43. (PubMed ID). PubMed PMID: 10487861.

Available at: http://www.ncbi.nlm.nih.gov/pubmed/18996890Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, Haendel M, Howe D, Mani P, Ramachandran S, Schaper K, Segerdell E, Song P, Sprunger B, Taylor S, Van Slyke C, and M Westerfield. (2006) The Zebrafish Information Network: the zebrafish model organism database. Nucl. Acids Res. 34:D581-D585 (PubMed ID). PubMed PMID: 16381936.

Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. genenames.org: the HGNC resources in 2011. Nucleic Acids Res. 2011;39:D519–9. (PubMed ID). PubMed PMID: 20929869.

Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, Zhang Z, and The FlyBase Consortium. (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. Nucl. Acids Res. 37: D555-D559 (PubMed ID). PubMed PMID: 18948289.

Chapter 19. Gene: A Directory of Genes

Donna Maglott, Kim Pruitt, and Tatiana Tatusova

Created: March 3, 2005; Updated: December 12, 2011.

Summary

A major goal of genomic sequencing projects is to identify and characterize genes. Gene (1) has been implemented at the National Center for Biotechnology Information (NCBI) (2) to organize information about genes, serving as a major node in the nexus of genomic map, sequence, expression, protein structure, function, and homology data. Each Gene record is assigned a unique identifier, the GeneID, that can be tracked through revision cycles. Gene records are established for known or predicted genes, which are defined by nucleotide sequence or map position. Not all taxa are represented, and the current scope matches that of NCBI's Reference Sequences group (3) and NIH's Mammalian Gene Collection (4).

Gene provides several improvements over its predecessor, LocusLink (5). These include a broader taxonomic scope, better integration with other databases in NCBI, and enhanced options for query and retrieval provided by NCBI's Entrez (6) system. Identifiers established by LocusLink (known as LocusID) have been retained in Gene as the GeneID.

This chapter describes

- how data are maintained in Gene
- query strategies
- record content and displays
- technical information for the power user

Overview

Gene is one of the several gene-centered resources at NCBI. Others include the Gene Expression Omnibus (GEO), HomoloGene, Online Mendelian Inheritance in Man (OMIM), and UniGene. The taxonomic scope of these resources differs. For example UniGene has clustered transcript information for some species that Gene does not, and Gene has records not cross-referenced in UniGene. Gene is solely responsible for providing the unique GeneID that is used to identify information for genes and other types of loci.

On a regular basis, model organism databases and other contributing groups are checked for novel information. If the record already exists in Gene, new information is added and outdated information is corrected. Otherwise, a new record is created.

Gene can be considered curated because many of the contributing databases are curated. Additionally, records in Gene may be reviewed by NCBI staff. However, Gene does not always attempt to reconcile genes defined by various annotation pipelines that may differ in levels of curatorial review and rules about what constitutes a gene.

Gene serves as a hub of information for databases both within and external to NCBI. Records are processed either gene-by-gene or as part of the submission of an annotated genome or chromosome. Gene identifiers, and associated names and sequence accessions, provide a common frame of reference for many databases.

For some genomes (e.g. human, mouse, rat, chicken, dog), Gene records are updated continuously. For other genomes, updates to Gene depend on the re-submission of genomic sequence annotation from an external group.

Gene includes records for confirmed genes and for genes predicted by annotation processes. The evidence for a gene can be inferred from the status of the RefSeq that defines it (information on status definitions can be found at http://www.ncbi.nlm.nih.gov/RefSeq/key.html#status). For example, RefSeqs that are termed as predicted or model have less supporting evidence than those in the validated, provisional, or reviewed categories. However, new sequence information is submitted to the public databases daily, and the status of a gene may not reflect current knowledge. New information on related sequences can be checked from Gene through the links to Entrez Nucleotide, Entrez Protein, and BLAST Link (BLink).

Gene does not claim to be comprehensive; rather, it serves as a guide to additional information in other databases. For example, a gene can be represented by multiple sequences, but not all are reported explicitly from Gene. Instead, connections are supplied from Gene to Entrez Nucleotide, Entrez Protein, and Blink, where more sequences with significant similarity can be retrieved. In addition to the multiple links to NCBI databases, LinkOuts submitted to Gene from external databases support ready navigation to more gene-specific information. The central functions of Gene are to establish unique identifiers for genes that can be tracked and, in so doing, support accurate connections with the defining sequences, nomenclature and other descriptors. With this infrastructure, it is possible to:

- support the NCBI annotation pipeline based on placement of sequences with known GeneIDs
- provide a species-independent frame of reference for genes and all their attributes
- support identification of the genes represented by sequences in the public databases

Maintaining the Data

New Records

Records are added to Gene if any of the following conditions is met:

- A RefSeq is created for a genome that has been completely sequenced and that record contains annotated genes. In the case of RNA viruses with polyprotein precursors, annotated proteins are treated as equivalent to a gene.
- A recognized, genome-specific database provides information about genes (preferably with defining sequence), mapped phenotypes, or sequences that are

treated as markers for incompletely characterized genes (e.g. expressed sequence tags and gene traps).
- The NCBI annotation pipeline identifies potential genes (models).
- A sequence submitted to public databases defines a new gene. For some genomes, the processing in Gene depends on UniGene's clustering process to identify a single representative sequence.

The minimum set of data necessary for a record in Gene, therefore, is a unique identifier or GeneID assigned by NCBI, a preferred symbol, and any of sequence information, map information, or nomenclature from a recognized authority.

Updating Data

Existing records are updated when new information is received. The staff of Gene collaborates with curators of organism-specific databases, nomenclature authorities, international annotation groups, other groups in NCBI, and other valued contributors to resolve discrepancies and improve the data. When a record is updated, its modification date changes. For some genomes, this may occur when the genome is re-annotated and converted into an updated RefSeq. For others, it may occur when any information attached to a gene record is altered. Other changes include adding, updating, or deleting sequence information, GeneRIFs, nomenclature, publications, and key identifiers such as numbers assigned to records in Mendelian Inheritance in Man (MIM numbers) and IDs from model organism databases.

Suppressing Records

From time to time it is necessary to combine Gene records or suppress ones created in error. Current or previous records can be retrieved from Gene by the GeneID. When a secondary GeneID has been replaced with another, a URL to the current record is provided.

Supplementary Information

Filters: information in other Entrez databases

Much of the power of querying Gene comes from mining its connections with other databases. Changes in these relationships are not captured in the modification date on the Gene record. For example, if information about new single nucleotide polymorphisms (SNPs) in a gene is submitted to the Single Nucleotide Polymorphism database and this information is now connected to Gene, that change is not reflected in the modification date of the record in Gene. In other words, a query to Gene based on records that have connections to dbSNP (using filters, as described below in “How to query Gene”) will return a different set of records, although there is no change in the modification date in any of the Gene records.

Filters: LinkOut to information in non-NCBI databases

Databases external to NCBI's Entrez system can submit and update links at any time. Users logged into My NCBI may elect to display any LinkOut with a standard icon. Changes in these connections will not be reflected in the modification date on the record in Gene.

Note: Database providers are encouraged to review the documentation about supplying LinkOuts (for more information see http://www.ncbi.nlm.nih.gov/entrez/linkout/doc/nonbiblinkout.html). This is a powerful method to attract users of Gene to your own database.

How to Query Gene

As with all databases accessed via Entrez, records can be retrieved from Gene based on:

- information anywhere in the record
- information in specified fields (Box 1)
- information on properties of the record (Box 2)
- the relationship of any record to other records in the Entrez system or on providers of external links (filters, Box 3)

Queries can be as simple as a single word or as complex as a combination of terms qualified by boolean operators using field restriction, properties, and filters. Several functions standard to Entrez are available to help users query Gene efficiently. Descriptions of these functions are below:

- **Limits** supports restricting results by combinations of species, by a value in one field, and by the modification date on the record.
- **Preview/Index** provides a comprehensive list of fields, filters, and properties currently used by Gene. It also reports the number of occurrences and values stored in each field, filter, and property, and it allows you to combine any term by boolean operators with existing queries. This is a key interface to test robust query strategies.
- **History** offers a review of recent queries and menus that can be used to combine these queries to selected sets of interest.
- **Clipboard** hold records of interest for up to 8 hours.
- **Details** shows how a query was processed. A query can then be refined and resubmitted.
- **My NCBI** allows users to save searches, customize filters, and schedule document delivery.
- **Entrez Utilities** allows users to retrieve records in other programs based on the same queries used interactively.

More details on using these functions are in the Entrez help document and FAQ pages.

Specificity in query results can be improved by making judicious use of fields, properties, and filters (Boxes 1, 2 and 3). To help you decide which of these to use, think of a field as a

subcategory of information, a property as a keyword or a term that may apply to many Gene records, and a filter as a representation of how Gene relates to other databases in the NCBI website. To select what filter to use, it might be helpful to know that NCBI names many filters by the pairs of names of the databases carrying common information. For Gene, the first database name is **gene**. Thus the filter representing common information in Gene and UniSTS is named "gene unists", common information in Gene and GEO is named "gene geo", etc. Properties may have the same name in multiple Entrez databases. For example, the property srcdb_refseq_known used in Entrez Nucleotide and Protein is interpreted from Gene as *"There are associated sequence data where the source database (srcdb) is RefSeq and the type of RefSeq is known"*.

To clarify these standards, consider the following examples:

> Example 1: Find human and mouse genes not annotated on the genome but having reviewed RefSeq records. First, you have to know that if a gene is annotated on the most recent genomic annotation, the filter "gene nucleotide pos" is set. Then you need to restrict your query by species and by the type of RefSeq.

If you typed this interactively, the query would be:

```
(Human[organism] OR mouse[organism]) AND "srcdb refseq
reviewed"[Properties] NOT "gene nucleotide pos"[Filter]
```

A much simpler approach is: to use Limits to set the species; preview/index to find the appropriate properties (reviewed RefSeqs, a characteristic of multiple Gene records); and a filter to find those not annotated on a genome (based on lack of links to contig or chromosome-based RefSeqs).

The steps you might follow are:

1. Click on Limits and check both human and mouse in the mammals section.
2. Click on Preview/Index, select properties, click on Index, scroll until you see "srcdb refseq reviewed", select it, and click on AND.
3. Still in Preview/Index, select fillters, click on Index, scroll until you see gene nucletide pos, select it, and click on NOT.

> Example 2: Find all Gene records from fungi that have expression data in UniGene or GEO.

If you typed this interactively, the query would be:

```
fungi[organism] AND ( "gene unigene"[filter] OR "gene geo"[filter])
```

A much simpler approach is to use Limits to set the taxonomic group and preview/index to find the appropriate filters and combine them correctly

The steps you might follow are:

- Click on Limits and check fungi.
- Click on Preview/Index, select filters, click on Index, scroll until you see "gene unigene" select it, and click on AND.
- Still in Preview/Index, select filters, click on Index, scroll until you see "gene geo", select it, and click on OR, and click on GO.

More sample queries are provided from the Gene help documents.

Box 1: Some fields used to index Gene.

A comprehensive list, with examples, is maintained in Gene's help documentation.

Field name
Chromosome
Creation date
Default map location
Disease or phenotype
Domain name
EC/RN number
Gene name
Gene Ontology (GO terms and values)
Gene/protein name
MIM
Modification Date
Nucleotide Accession
Nucleotide UID
Nucleotide or protein Accession
Organism

Box 2: Some properties indexed by Gene.

A current list can be displayed from Gene at any time by clicking on Preview/Index, selecting Properties from the pull-down menu, and clicking on Index. Definitions of all RefSeq types are maintained at the RefSeq homepage.

continues on next page...

continued from previous page.

Property name	Explanation
alive	a current, primary record (i.e., not secondary or discontinued). The term secondary means a record that has been merged into another.
GeneRIF	a record having one or more GeneRIF annotations attached
genetype miscrna	gene encodes an RNA not in any of the specifics below
genetype other	of know type, but not any of the specific known categories
genetype protein coding	encodes a protein
genetype pseudo	pseudogene
genetype rrna	encodes ribosomal RNA
genetype scrna	encodes small cytoplasmic RNA
genetype snorna	encodes small nucleolar RNA
genetype snrna	encodes small nuclear RNA
genetype trna	encodes transfer RNA
genetype unknown	the type of gene is not known
has transcript variants	a record having two or more associated RefSeq transcripts, i.e. splice variants. NOTE: this is limited to RefSeq annotation and should NOT be used to identify all genes exhibiting alternative splicing, promoter usage, and/or polyadenylation signals.
phenotype	has an associated phenotype
phenotype only	only method of defining this gene is by phenotype
source extrachomosomal	located extrachromosomally
source genomic	located on a chromosome
source mitochondrion	located in the mitochondrion
source other	location not included in other specifics
source organelle	located in an organelle (includes mitochondrion and plastid)
source plasmid	located in a plasmid
source plastid	located in a plastid
source proviral	located in a provirus
source virion	located in a virion
srcdb refseq	has an associated RefSeq
srcdb refseq inferred	has an associated RefSeq of type inferred
srcdb refseq known	has an associated RefSeq of type known
srcdb refseq model	has an associated RefSeq of type model
srcdb refseq predicted	has an associated RefSeq of type predicted

continues on next page...

continued from previous page.

srcdb refseq provisional	has an associated RefSeq of type provisional
srcdb refseq reviewed	has an associated RefSeq of type reviewed
srcdb refseq validated	has an associated RefSeq of type validated
Property name	Explanation

Box 3: Some filters in Gene.

The Entrez system uses the term *filters* to connote the function that subsets a query or retrieval set by attributes of the record. Here are some common filters available from Gene. This is a report you can generate by selecting "Filters" from the blue sidebar within 'My NCBI'. The display from the preview/index menu is more concise.

You may select these commonly requested filters or use Browse to see all filters for this database.

Configure > Gene

Commonly Requested Filters

Gene records annotated on partial or complete chromosomal RefSeqs (Genes Genomes). Gene records with explicit links to RefSeq chromosome or contig accessions.

Gene records associated with citations in PubMed (gene pubmed). Gene records with explicit links to Entrez PubMed. Useful to identify genes that have associated publications.

Gene records associated with expression data in UniGene (gene unigene). Gene records with explicit links to Entrez UniGene. Calculated from common mRNA sequence data.

Gene records associated with PCR-based markers in UniSTS (gene unists). Gene records associated with PCR-based marker data in Entrez UniSTS. Associations are calculated by e-PCR or curated submissions.

Gene records associated with protein sequence (gene protein). Gene records with explicit links to Entrez Protein. Includes links to GenPept, RefSeq, and SwissProt accessions.

Gene records associated with variation information in dbSNP (gene snp). Gene records with explicit links to Entrez dbSNP. Supports finding genes with variation information available in dbSNP.

Gene records shown in Map Viewer (gene mapview). Gene records known to be on a current annotation of a genome.

continues on next page...

continued from previous page.

Gene records with expression data in GEO (gene geo). Gene records with additional data in Gene Expression Omnibus (GEO), based on common sequence information.

Gene records with Gene Genotype reports in dbSNP(gene genotype). Gene records with reports of genotypes in the dbSNP database.

Gene records with homology data (gene homologene). Gene records with explicit links to Entrez HomoloGene. Useful to find genes that appear to be conserved.

Gene records with MIM (Mendelian Inheritance in Man) numbers (gene omim). Gene records with explicit links to OMIM. Includes links to both disease and "gene" records.

Gene records with nucleotide sequence data (gene nucleotide). Gene records with explicit links to Entrez nucleotide, excluding RefSeq chromosome or contig accessions. Useful to find genes that have nucleotide sequence information.

Gene records with proteins calculated to contain conserved domains (gene cdd). Gene records with RefSeq proteins calculated to contain conserved domains by comparison to the CDD database.

Display Formats

Gene provides several displays differing in content and format to help you find and report the information you want. There are two default displays: the summary HTML page returned in response to a query, and the complete (Graphic) HTML display returned after a single record is selected. All HTML displays include the Links function that indicates what other resources contain additional information. Some of these links are based on information managed directly from Gene. For example, links to Entrez Nucleotide, Entrez Protein, PubMed, and OMIM are based on the sequences, citations, and MIM numbers contained in a record. Other links are managed from databases other than Gene or from information shared by other databases. For example, links to dbSNP, GEO, HomoloGene, UniGene, and UniSTS are based on shared nucleotide sequence data. Links to CDD are based on shared protein sequence. Links to Map Viewer indicate that information about the position of the gene is available.

Another useful display format is the Gene Table. If a gene has been annotated on any genomic RefSeq, the intron/exon organization of each transcript is summarized. In the case of an mRNA, the translated region of each exon is summarized. Gene Table facilitates access to other gene-related sequences, such as the complete RNA, protein, specific exons, introns, or coding regions. Other display formats include XML and ASN1- specifications for each can be found in the Gene help document.

Content

The content of an Gene record fits into several sub-categories. Those listed here correspond roughly to what is seen in the default full (Graphic) display.

Nomenclature

Gene uses official symbols and full names and reports the nomenclature authority when available. Otherwise, symbols and names are selected from the defining sequence record. For example, if sequence and positional homology (synteny) suggest that a nameless locus in one species is orthologous to a named gene in another, the symbol from the ortholog may be used. If no symbol is identified, and the genome is processed gene-by-gene rather than as a complete re-annotation, the letters LOC are prepended to the GeneID. Once a meaningful symbol is identified, the contrived "LOC" symbol is removed (because the record will still be searchable and identified by the GeneID itself).

In addition to official symbols and full names, Gene provides others seen in publications and sequence records. These alternative names are not meant to be comprehensive and often are identified only when the RefSeq is being reviewed.

Several NCBI databases use the nomenclature maintained by Gene. These names are incorporated based primarily on the name-GeneID-sequence relationships that Gene reports. These data are reported in several files on Gene's FTP site, including DATA/gene_info.gz and DATA/gene2accession.gz.

Overview

Some of the components of the Gene record describe key characteristics of the gene, its function, and its products. The Summary, written by RefSeq staff and/or by external contributors such as OMIM or Rat genome Database (RGD), provides a quick synopsis of what is known about the gene, the function of its encoded protein or RNA products, disease associations, spatial and temporal distribution, and so on. The gene type is assigned from a list of options defined in the Gene data model.

The value of RefSeqStatus indicates the maximum level of review that has been provided to the set of gene-specific accessions.

Map Data

Several types of map information may be included in an Gene record. One type is the description of location in units commonly used for a given genome. Genetic and physical map positions are incorporated from the published maps used in Map Viewer. Rather than report all position data for any gene in any coordinate system, this information can be obtained through links to Map Viewer. Information can also be accessed through marker names, which are linked to the UniSTS record.

When no independent map data are available and the gene has been placed on a genomic assembly, map position may be inferred by a calculated correspondence between sequence and other map units, such as cytogenetic bands. One example is the calculation of cytogenetic position according to the algorithm developed by Furey and Haussler (7). With each re-assembly of a genome, genes might be moved to other chromosomes with which better alignments are identified. If marker and other data are consistent with but distinct from the published map location, then the Gene record is modified to be consistent with current information.

Markers are reported in Gene either as a gene or as a marker that has a calculated or curated relationship with a gene. Gene does not store all of the markers available for a genome; that is the function of UniSTS. The marker data in Gene come from any of the following: a report from a genome-specific database; calculations based on e-PCR that indicates that an mRNA is associated with the gene; and e-PCR based localization on the genome within a region beginning 2 kb upstream of the gene and ending 0.5 kb downstream. In queries initiated from Gene, genes that have PCR-based markers can be identified by the query "gene unists"[filter].

When a gene has been annotated on a genomic RefSeq, map information is also presented by the graphic display of neighboring genes. An arrow indicates the direction of transcription. If the name of a gene is too long to be used as a label, truncation is indicated by an ellipsis (...). The gene specific to the displayed record is highlighted. The arrows and labels anchor links to the records for those genes, supporting quick navigation. If a gene is annotated on more than one genomic RefSeq, only one is used for the graphic display. The location data for each RefSeq are provided in the ASN.1 of the full Gene record.

Map data are also supported by named links to Map Viewer in the Links menu. Because links are provided by the Map Viewer database, changes in these links are not reflected in the modification date on the record. For genomes where comparative maps are available in Map Viewer, links to Map Viewer are also provided for those views.

Sequence-related Data

Sequence information is presented in multiple forms in Gene:

- graphical displays of the intron/exon organization of splice variants
- reports of intron/exon organization of each variant in the Gene Table display
- reports of RefSeq accessions and their domain content
- reports of accessions from DDBJ, EMBL, GenBank and Swiss-Prot
- links to the genomic sequence, in standard formats, for the genomic sequence of the gene, individual introns or exons, and the transcripts (Gene Table display)
- links to related records via the Conserved Domain database
- links to the BLink viewer of protein neighbors

Sequence information (accessions and links) is distributed throughout the Gene record. For example, the Transcripts and Products diagram is provided when a gene has been annotated on a genomic RefSeq, in other words when the intron/exon/coding region information is available in genomic coordinates. Each position of a gene product, when represented by a RefSeq RNA and/or protein, is provided relative to the genomic DNA. Each RefSeq Accession number (genomic, mRNA and protein) anchors a link to different formats of the sequence in Entrez Nucleotide or Entrez Protein (the link can be found over the diagram). The link from the Accession number for the genomic sequence displays only gene-specific region. The anchor on the protein accessions also facilitates retrieval of specific BLink, CDD, or COG displays.

The NCBI Reference Sequences (RefSeqs) section lists nucleotide and protein accessions that are related to the gene and provides links to the appropriate sequence record in Entrez Nucleotide or Entrez Protein. Conserved domains are reported by name, location on the sequence, and the BLAST score substantiating the assignment.

“Related sequences” lists nucleotide and protein accessions that are related to the gene and provides links to the appropriate sequence record in Entrez Nucleotide or Protein. If the protein sequence record is not part of a set of a nucleotide record and the protein it encodes, the word 'none' is printed in the nucleotide column. The type of nucleotide record is printed before the nucleotide accession, and the strain is printed after the protein accession, as applicable.

Function

Gene uses several approaches to describe the function of a gene and its encoded products. These include:

- explicit descriptive statements (RefSeq Summary and GeneRIF)
- names of genes, products, and pathways
- associated ontologies (GO)
- reports of interactions
- Enzyme Commission (EC) numbers
- inferences from domain content
- descriptions of diseases or allele-specific phenotypes
- links to other databases (OMIM, HomoloGene, PubMed)

Many of these categories include links to additional information in other databases. Links to the data sources are provided. We appreciate the cooperation of the resources that have made their data freely available.

Variation

Gene does not report variation information directly. Rather it provides three types of links to dbSNP, where these variation data are stored. These types are implemented by the filters gene snp, gene snp gene genotype, and gene snp geneview (Box 3).

Homology

Except for indicating the availability of comparative maps (limited at the time of this writing to Gene records from human, mouse, and rat), Gene provides information about homology only by displaying links to HomoloGene and/or COG. It also provides links to resources that display pre-computed sequence relationships such as BLink.

Expression

The qualitative assessment of whether a gene is expressed is captured in the Gene type and in the types of sequence accessions associated with the Gene record. The quantitative and spatio-temporal aspects of expression are stored in other databases, including GEO, and UniGene at NCBI.

Other Sites of Interest

Gene provides information about other sites of interest both within a record and via the LinkOut mechanism. As more data providers submit their LinkOuts to Gene, the second method will be increasingly powerful. Users can take advantage of the LinkOut connections, and other filters, by registering for My NCBI and customizing the display.

References

1. Maglott D.et al. *Gene: gene-centered information at NCBI.* Nucleic Acids Res. 2005;33(Database issue):D54–8.
2. Wheeler D.L.et al. *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res. 2005;33(Database issue):D39–45.
3. Pruitt K.D., Tatusova T., Maglott D.R. *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res. 2005;33(Database issue):D501–4.
4. Gerhard D.S.et al. *The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC).* Genome Res. 2004;14(10B):2121–7.
5. Pruitt K.D.et al. *Introducing RefSeq and LocusLink: curated human genome resources at the NCBI.* Trends Genet. 2000;16(1):44–7.
6. Schuler G.D.et al. *Entrez: molecular biology database and retrieval system.* Methods Enzymol. 1996;266:141–62.
7. Furey T.S., Haussler D. *Integration of the cytogenetic map with the draft human genome sequence.* Hum Mol Genet. 2003;12(9):1037–44.

Chapter 20. Using the Map Viewer to Explore Genomes

Susan M. Dombrowski and Donna Maglott

Created: October 9, 2002; Updated: August 13, 2003.

Summary

There are many different approaches to starting a genomic analysis. These include literature searching, searching databases for gene names and other genomic features, performing sequence comparisons, or using map data to find gene information by position relative to other landmarks. The NCBI Map Viewer has been developed to facilitate this latter approach.

The purpose of this chapter is to provide a foundation for gaining maximum benefit from using the Map Viewer and related resources at NCBI. It is important to note that in this document, the term "map" refers to a position of a particular type of object in a particular coordinate system. This means, for example, that there is not one sequence map but a set of maps in sequence coordinates. Readers interested in precisely how sequence-based maps are annotated and assembled should refer to Chapter 14.

Introduction

First launched with the release of the sequence of *Drosophila melanogaster* in March 2000, Map Viewer is now used to present genetic, radiation hybrid (RH), cytogenetic, breakpoint, sequence-based, and clone maps for many genomes. The availability of whole genome sequences means that objects such as genes, markers, clones, sites of variation, and clone boundaries can be positioned by aligning defining sequence from these objects against the genomic sequence. This position information can then be compared to information about order obtained by other means, such as genetic or physical mapping. The results of sequence-based queries (e.g., BLAST) can also be viewed in genomic context. Our view of the genomes of a variety of organisms is constantly being improved through the increase in underlying data.

Map Viewer integrates map and sequence data from a variety of sources. The basic architecture and principle of Map Viewer can be applied to any complete or incomplete genome as long as map data exist to support it. Map Viewer is a powerful tool because it provides: (1) a mechanism to compare maps in different coordinate systems; (2) a robust query interface; (3) diverse options for configuring the display; (4) multiple functions to report and download maps and annotated information; (5) tools to manipulate nucleotide sequence such as ModelMaker (for constructing mRNAs from putative exon sequences); (6) connections to comprehensive data files for transfer by FTP; and (7) detailed descriptions of the objects displayed on the maps.

Maintenance of Data

Data Sources

Non-Sequence-based Maps. Sources of maps that are not based directly on sequence include published maps in genetic, radiation hybrid, cytogenetic, and ordinal coordinate systems (where ordinal refers to clone order). The primary sources of each map are described in the online help documentation of each genome-specific Map Viewer. We are indebted to the researchers who make their mapping results so freely available. When a new version of any map becomes available, the data are also updated in the appropriate NCBI database.

Sequence-based Maps. The sequence-based maps shown through Map Viewer can be supplied by external sources and/or supplied from features computed within NCBI. For example, when the annotated sequence for a complete genome is submitted to the sequence databases (GenBank/EMBL/DDBJ), a copy of the data may also be accessioned as Reference Sequences (RefSeqs; see Chapter 18). The gene, transcript, and other feature annotations of the submitted complete genome are processed for display in the Map Viewer. NCBI staff may then calculate and display the position of other types of features, such as marker position or points of variation, as separate maps (Table 1).

Some of the annotation of genomic sequence carried out by NCBI is included in the genomic reference sequences (NC, NT, and NW Accession number format); however, other annotation is represented only in the Map Viewer and in the associated reports (Table 1). This latter type of annotation is based on information in several NCBI databases (Table 2) and is particularly important for attaching biological information to sequence data. Links to these resources are provided in Map Viewer to provide further information about each annotated object. It should be noted, however, that although sequence features may be placed in a genomic context automatically, there are curation steps that affect the final displays. For example, for the human and mouse genomes, sequences defining genes and pseudogenes are reviewed by collaborators and NCBI staff and, whenever possible, used as the basis of RefSeq records (NG, NM, and NR Accession number format).

Feature annotation is computed primarily in two ways: (1) by alignment of the defining sequence to the genome; or (2) for sequence tagged sites (STSs), by e-PCR (1). In some genomes, gene placement is based primarily on the alignment of mRNA [Expressed Sequence Tags (ESTs) and cDNAs], but only when an encoded protein is predicted. In other cases, where transcription evidence is weaker, more weight is given to identification of protein-coding regions. Gene identification is also constrained in that a known gene cannot be placed more than once in a haplotype (except for pseudo-autosomal regions) or on an incorrect chromosome. Thus, if any reference haplotype retains inappropriately redundant sequence that encodes a gene, only one copy will be annotated as that gene. Others will be assigned interim IDs (see Chapter 14). Some *ab initio* methods may also be used for gene prediction. The predicted genes, as well as the mRNAs, are supplied as separate maps (gene, RNA, or GenomeScan maps).

In some cases, the position of these features may suggest the location of other genomic regions of interest. For example, the position of STS markers can help define the position of phenotypes such as quantitative trait loci (QTL). Although the best annotation of a gene or region is always through annotation by an expert researcher, automated annotation of genomes and comparison to that provided by experts can provide significant useful information. Experts interested in analyzing or assisting with genome annotation should contact us at info@ncbi.nlm.nih.gov.

Table 1. Types of Map Viewer annotation provided by NCBI.

Feature[a]	Coordinate system[b]	Representative maps[c]
STS	Sequence (Mb), Radiation hybrid (cRay), Genetic (cM), Clone content (ordinal), Cytogenetic	STS, STSnw, G3, GM4, GeneMap'99, TNG, Marshfield, Genethon, deCode, Whitehead YAC, phenotype maps such as Quantitative Trait Loci (QTL)
Clones	Sequence, Cytogenetic	Clone, BES, Components
Expression	Sequence	SAGE tag, UniGene
Genes	Sequence (Mb), Cytogenetic (band names)	Genes_seq, Genes_cyto
Gene-related	Sequence, Cytogenetic	UniGene, GenomeScan, Mitelman recurrent breakpoint, morbid
Variation	Sequence (Mb)	Variation
Published accessions	Sequence (Mb)	GenBank
Phenotype	Cytogenetic, Cytogenetic (abnormalities), Sequence	OMIM's morbid map, Mitelman's recurrent breakpoint, QTL (in progress)
Source clones	Sequence (Mb)	Component
Homology	Sequence (Mb)	Indirectly via LocusLink or UniGene. For mouse and human, through the homology (hm) link to the mouse–human homology map

a The feature column lists the types of objects annotated on maps seen in Map Viewer. Those features in bold type are annotated on the RefSeqs; the rest are provided only from the Map Viewer, and the files are available for FTP transfer.
b The different map types and coordinate systems that may contain a particular type of feature.
c A partial enumeration of named maps that represents positions of this feature type.

Table 2. NCBI data resources used in NCBI-generated annotation.

Resource	Description
Clone Registry	Clone sequencing sequence status, STS content, and availability
dbSNP	Single Nucleotide Polymorphisms (SNPs), polymorphisms, small-scale insertions/deletions, polymorphic repetitive elements

Table 2 continues on next page...

Table 2 continued from previous page.

Resource	Description
Genome Guides	Directory of key resources for the genome, with links to related resources and tutorials. The directory to guide pages is available from Genomic Biology.
LocusLink	Locus-specific data for a subset of organisms with extensive links to related resources and sequence data
OMIM	Human genes and Mendelian disorders
RefSeq	NCBI's curated, non-redundant RefSeqs
UniGene	Computed clusters of cDNA and Expressed Sequence Tag (EST) sequences from the same gene, with tissue expression information and links to related resources
UniSTS	Unified, nonredundant database of sequence tagged sites (STSs)

Relationships among Coordinate Systems

In addition to supporting the display of multiple maps in the same coordinate system (e.g., multiple sequence-based maps), Map Viewer also displays maps in different coordinate systems by calculating the correspondances among them (e.g., sequence to genetic). This is accomplished by: (*a*) identifying features that have been placed on maps in different coordinate systems; and (*b*) using general conversion factors. In the first case, placement of STSs on the genome is critical for the integration of sequence data with other, non-sequence-based maps, such as genetic and RH maps. The integration of cytogenetic data with sequence data is achieved through alignment of sequence from clones that have been placed cytogentically, such as the human fluorescence *in situ* hybridization (FISH)-mapped clones from the Bacterial Artificial Chromosome (BAC) Resource Consortium (2). The integration of non-sequence-based maps with the sequence provides a powerful mechanism to access portions of sequence on the basis of marker or cytogenetic data. Many features, such as Single Nucleotide Polymorphisms (SNPs), ESTs, mRNAs, whole genome shotgun reads, and clones can be placed on the genome assembly by using standard DNA sequence alignment methods such as BLAST.

The identification of known genes within the genome assembly provides critical landmarks and functional context to the sequence data, which in turn makes it easier to traverse to other rich sources of gene and protein information, including publications, OMIM, RefSeq, Conserved Domain Database (CDD), and LocusLink.

The power of calculating correspondances between coordinate systems may be more apparent when considering a common application of Map Viewer, i.e., identifying candidate genes within a region defined by genetic markers. When markers are palced on both genetic and sequence maps, it is then possible to use the gene-related maps (gene, UniGene/EST, or *ab initio* predictions) to identify possible genes of interest. For more details on how to do this, see the Map Viewer Exercises in Chapter 24.

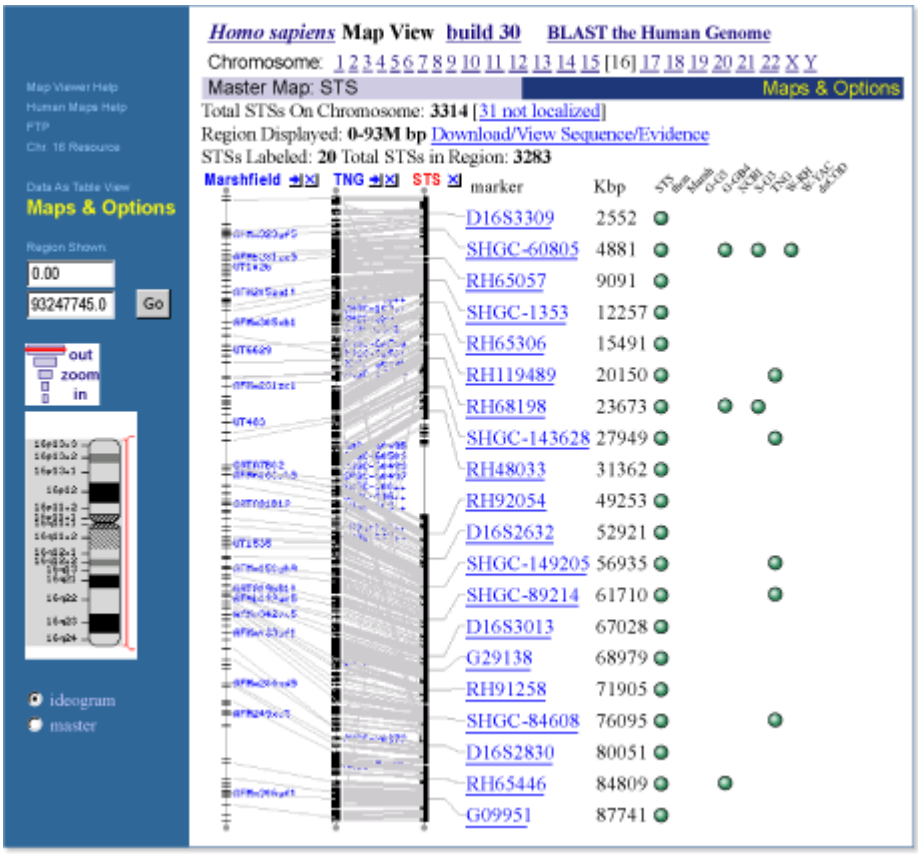


Figure 1. Evaluation of a chromosome sequence (STS) map. Potential inconsistencies in the order or orientation of sequence blocks can be investigated by displaying a genetic map (*Marshfield*), radiation hybrid map (*TNG*), and sequence map (*STS*) together and checking the **Show connections** box in the **Maps & Options** window. Note that some of the *gray lines* (connecting the same marker on different maps) are *crossed*, indicating that either the placement is incorrect on a map or the chromosome sequence is not ordered and oriented consistently with all map data.

A Work in Progress

For many genomes, identifying and positioning chromosomes and genes within sequence blocks is an ongoing process. In those cases, the Map Viewer can be used to evaluate the evidence that supports the current representation of the sequence and visualize possible conflicts. Inconsistencies in map order or in the placement of any object can be seen in the Map Viewer; this is assisted in some cases by the use of color coding (Figures 1 and 2).

For some genomes, the color-coded contig map displays whether the annotation is based on sequence assembled from draft or finished clones (blue, finished; green, whole genome shotgun; orange, draft). This is helpful when evaluating the level of confidence in the completeness of the annotation of a gene and/or its coding region.

Map Viewer also uses color coding or diagrams to represent the level of confidence in the placement of any mapped object. For example, SNPs or STSs that are placed at more that

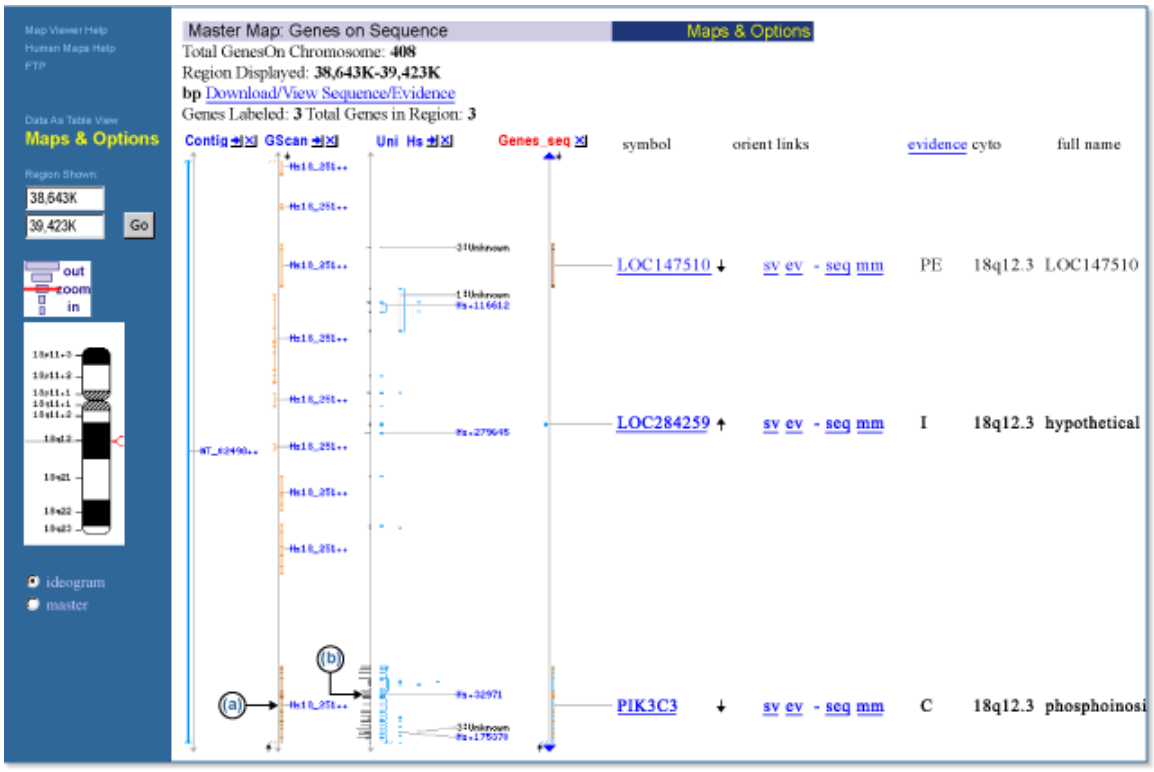


Figure 2. Evaluation of gene localization and annotation. A comparison of cDNA alignments (UniGene, RNA) and gene predictions (GenomeScan) to the genomic contig annotation can be achieved by displaying three maps simultaneously. The genomic contig (NT_024981.9) annotation is shown in the *Genes_seq* map and is displayed with the GenomeScan predictions (the *GScan* map) and the EST/mRNA alignments labeled by human UniGene clusters (the *UniG_Hs* map). Note that in this case, there are two sequence objects not included in the contig annotation: one is an *ab initio* prediction (the last model in the GScan map) (*a*); and the other is either some small gene or an alternative 3′ exon for PIK3C3 from the UniG_Hs map (*b*). This approach is especially useful when reviewing BLAST results in a genomic context.

one position in a given map are noted by color (yellow) in the detailed labels (Figure 3a). Annotated genes are shown in different colors, based on the source and level of confidence in the annotation or the model (Figure 3b).

Frequency of Updates

Although maps provided from external sources are updated when new data are available, the maps dependent on NCBI's annotation process are updated periodically in versions called “builds”. Thus, mRNA or other supporting evidence that becomes available after the data “freeze” date for one build will not be incorporated into the display until the next build. However, some of the supporting databases linked from the Map Viewer may have more updated information. For example, UniSTS may provide more recent e-PCR results, or LocusLink may show a newer name or additional sequence data. dbSNP may make major data releases between builds; in this case, the variation map is updated.

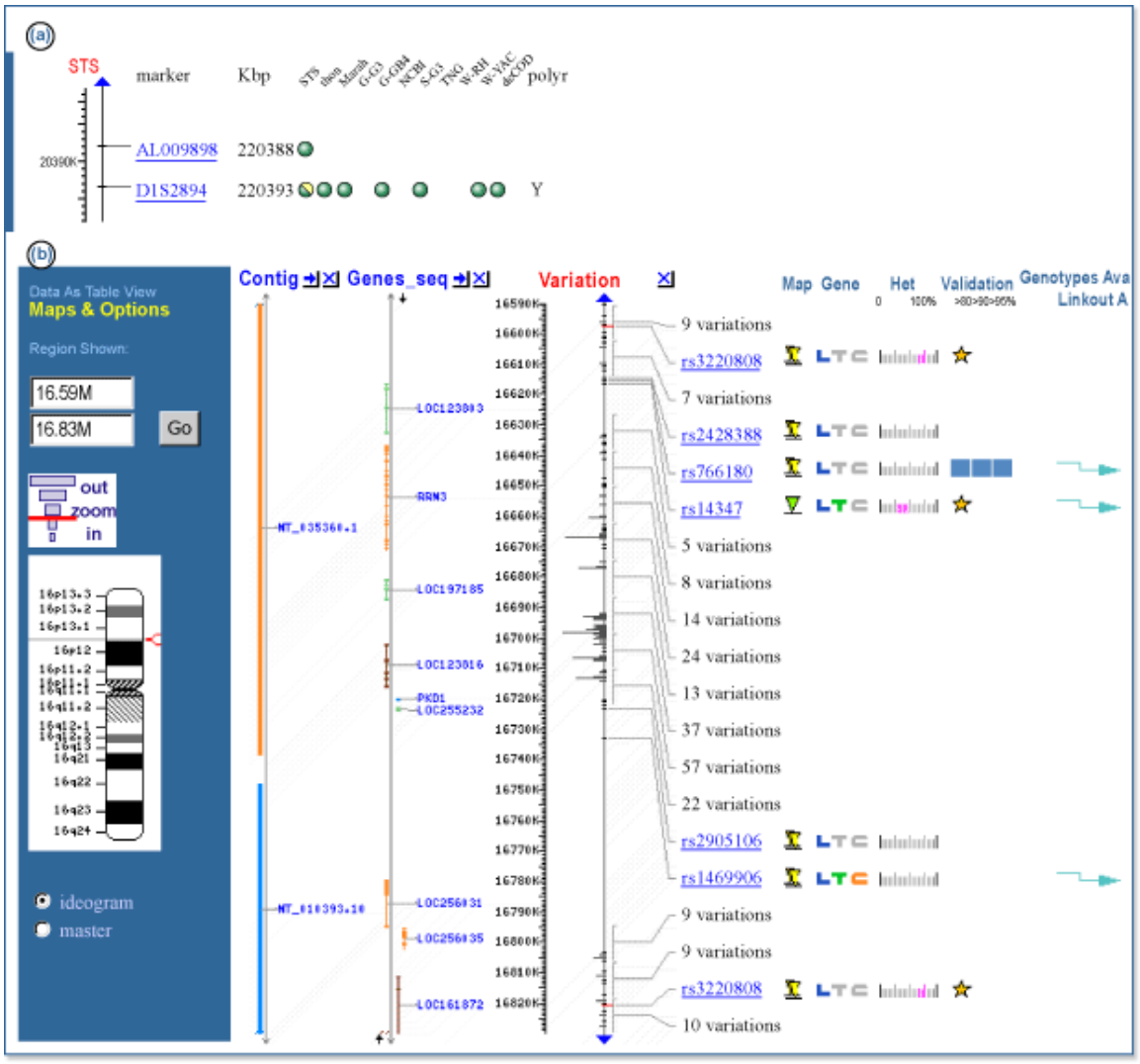


Figure 3. Representation of ambiguity. (*a*) The marker D1S2894 is found on several maps. Note that for the first map (*STS*), the circle is diagonally split with two colors. The diagonal means that the marker has been placed more than once; the two colors mean that the placements are not on the same chromosome. (*b*) A Map Viewer display of a region of chromosome 16. SNPs that are placed more than once on the chromosome are designated by a *yellow triangle*. From the *Contig* map, it appears that at least one of these SNPs (rs3220808) is placed both on draft sequence (*orange*) and on finished sequence (*blue*). This may be an artifact resulting from misassembly or perhaps a region of segmental duplication. This diagram also illustrates the use of color to indicate the source and level of confidence in annotated genes. *Blue* indicates a confirmed gene with no conflicts; *light green* indicates EST evidence only; *dark brown* indicates a GenomeScan prediction with protein homology; *orange* means that there is a conflict between the annotated gene and the mRNA evidence. (*Ab initio* predictions from GenomeScan are categorized into two types, based on presence or absence of sequence similarity to vertebrate proteins or protein domains.)

Methods of Access

Although most of this chapter discusses the human genome Map Viewer, there is a growing number of organisms for which there is Map Viewer access to the genome. To identify the taxa that have Map Viewer access to the genome, query the taxonomy database by typing "loprovmapviewer"[filter] into the query box on the Entrez Taxonomy homepage; or more simply, review the options provided on the Map Viewer homepage.

Links from NCBI Resources

Many NCBI databases are now integrated into Map Viewer (Table 2); therefore, database records are often linked to Map Viewer displays. If a sequence in the public databases was released before the date of the current Map Viewer data freeze, then the position of this sequence may be displayed within Map Viewer. For example, Entrez Nucleotide, UniGene, UniSTS, and LocusLink records for sequences annotated on the human genome provide links directly to the appropriate region of the genome via links called **Map Viewer** (in the **Links** menu), **Nucleotide**, **Map View**, or **mv** links, respectively (Figure 4). It should be noted that such links are only precomputed if at least 50% of the sequence aligns with an identity of greater than 90%.

Genome-specific resource pages also support queries via chromosome diagrams (Figure 5).

Sequence Similarity Searches

Genome-specific BLAST pages that restrict a search to a specific genome are provided for several organisms and allow the results of the search to be displayed in a genomic context (provided by Map Viewer). Genome-specific BLAST searches can be accessed from the BLAST homepage, the Map Viewer pages of individual organisms (e.g., human, mouse), and the genome-specific resource pages of individual organisms. If the reference genome (the default) is selected as the database to be searched, the **Genome View** button (Figure 6) will appear on the BLAST results display page.

Direct Query

Simple Searches

When already at a genome-specific Map Viewer page, any combination of query terms can be entered into a Map Viewer **Search for** box (Figure 7). Boolean operators (AND, OR, and NOT) and the use of * as a wild card (applied to the right of any term) are supported. The **Search for** and **Help** document hyperlinks provide current details about query options. An advanced search is available for some genomes.

Queries may include any unique identifier for a database record, e.g., a sequence Accession number or OMIM (MIM) number, or a text term or phrase, e.g., a gene symbol (BRCA2) or descriptor (p53-binding), or disease name (lung cancer). The Boolean AND operator is used automatically if multiple terms are entered. Therefore, a query for

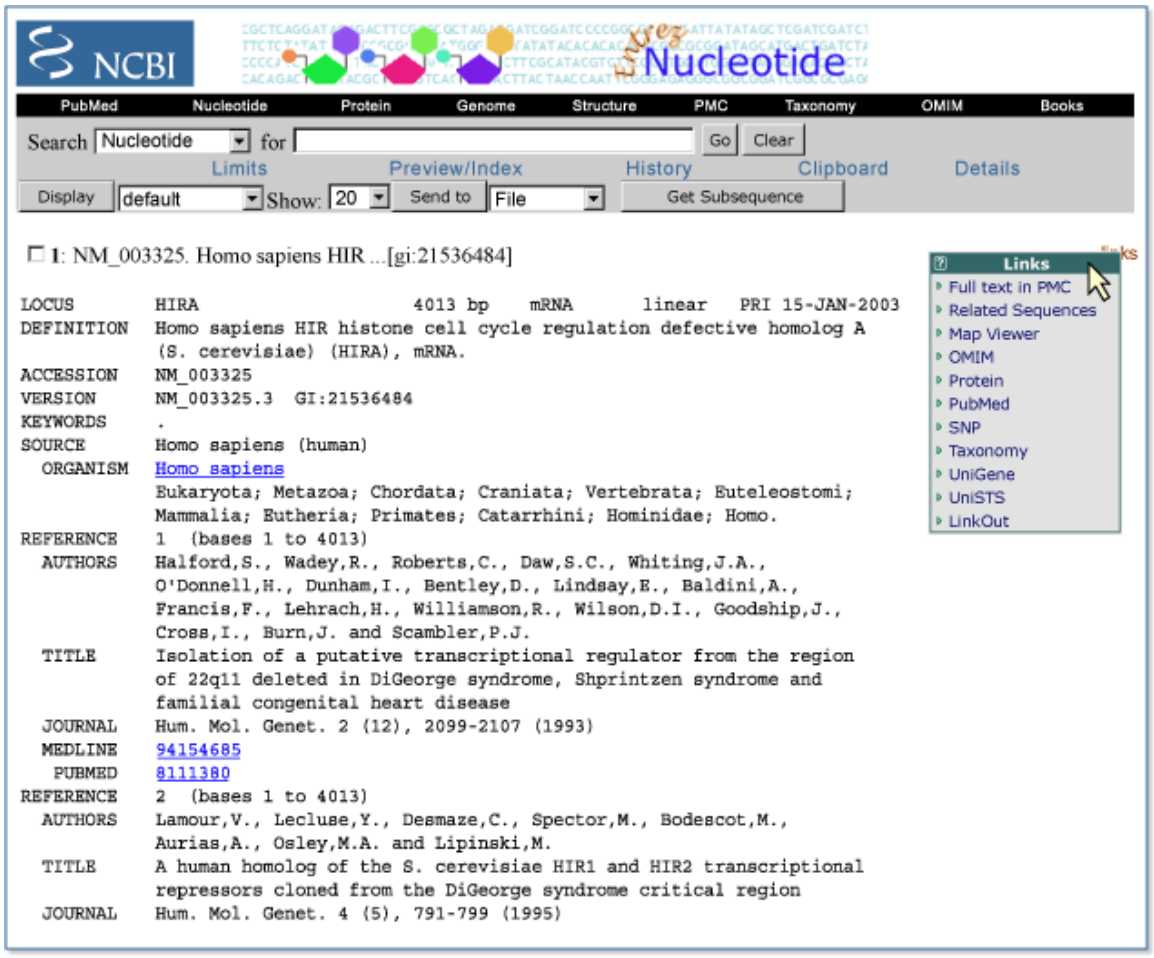


Figure 4. Connecting to Map Viewer from Entrez Nucleotide, using the Links menu in Entrez Nucleotide to connect from a record to Map Viewer.

"fanconi anemia" will automatically be interpreted as "fanconi AND anemia". The wildcard operator (*) provides a convenient mechanism to retrieve genes that share a common symbol or name, as is often found for gene families. For example, a query for ABC* will return matches to the ATP-binding cassette superfamily.

The advanced query page, accessed by checking the **Advanced search** box, provides additional options to refine a query. These additional options, which may vary from genome to genome, are useful for restricting queries to a particular search field or map type. The advanced query page also includes predefined search options to restrict the search to data with certain properties, e.g., to only find genes associated with a known disease or with sequence variation (SNPs). Additional refinements to queries against the variation map can also be made, for example, to search for variation markers known to be in a gene or coding region.

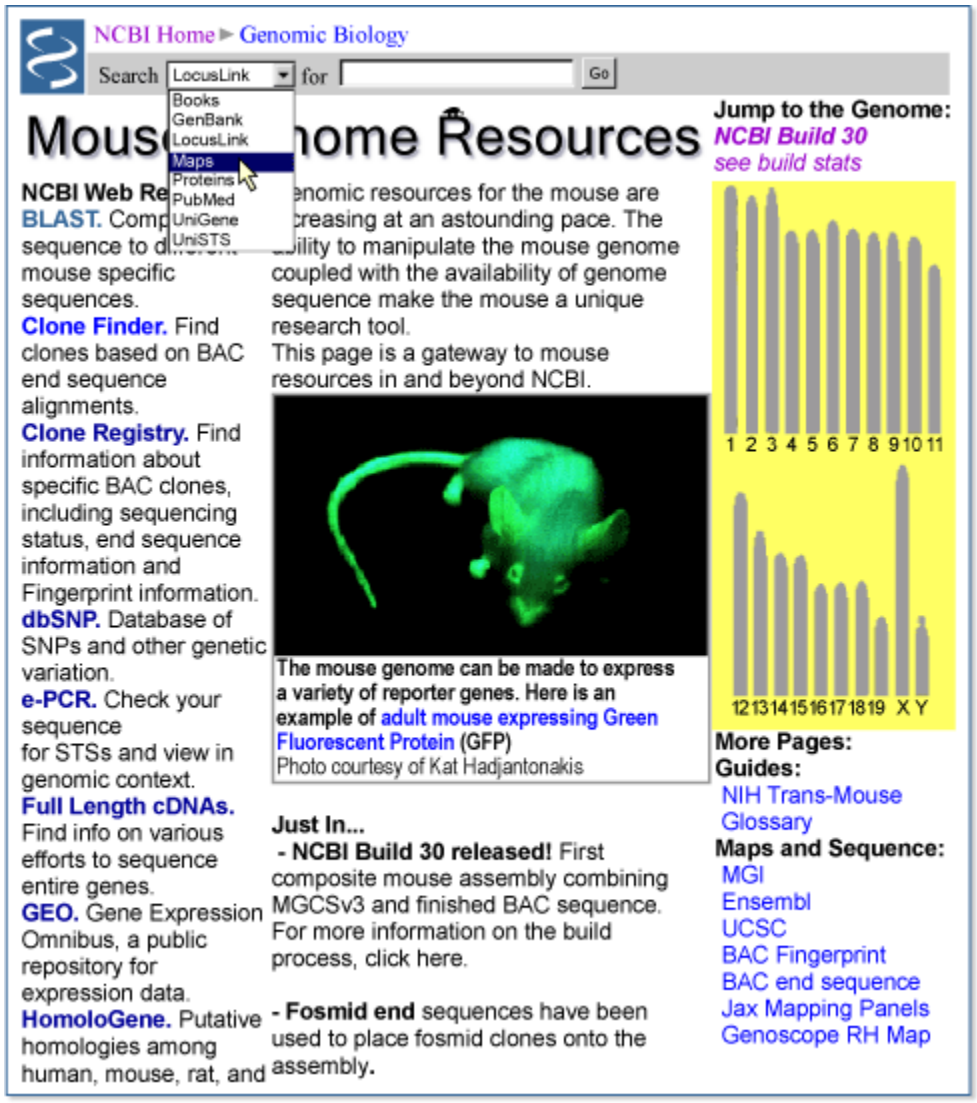


Figure 5. Example of a genome-specific resource page supporting queries to Map Viewer. Note that there are two ways to connect: (1) by selecting **Maps** from the pull-down menu in the *gray bar* at the *top* of the page and entering a query term in the **Search** box; or (2) by selecting a chromosome in the genome diagram on the *right* of the page (*yellow* background).

The same options for wild cards and Boolean operators for your query term(s) apply when starting at the Map Viewer homepage. At present, however, you must select a genome to which to restrict your search. An option to query across multiple genomes is under development.

Position-based Access

To use Map Viewer to display a particular section of a genome by using a range of positions as a query, it is first necessary to select a particular chromosome for display from either a genome-specific Map Viewer page or a Genome Guide page.

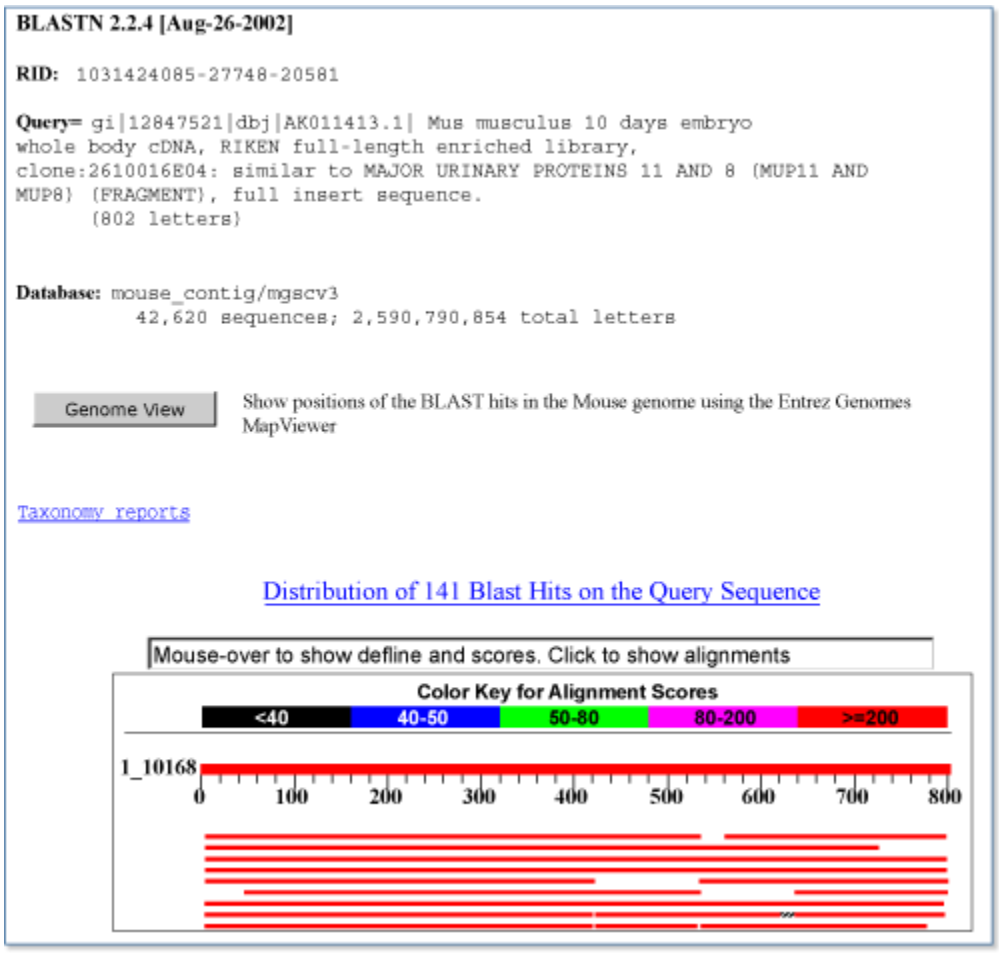


Figure 6. Accessing the Map Viewer display from a genome-specific BLAST results page. Selecting the **Genome View** button shows all of the BLAST hits on the genome.

Once a single chromosome is displayed, position-based queries can be defined by: (1) entering a value into the **Region Shown** box. This could be a numerical range (base pairs are the default if no units are entered), the names of clones, genes, markers, SNPs, or any combination. The screen will be refreshed with only that region shown. If the first entry cannot be resolved, the display will extend to the top of the map; if the second entry cannot be resolved, the display will extend to the bottom of the map. Both of these navigational aids are found on the left of the page; and (2) using the **Maps & Options** controls. One of the options in this menu is to define the region shown. Here it may be clearer that the region selected will be in the coordinates of the rightmost, or Master, map, which may also be adjusted in this menu. The values that can be used to specify the range are the same as those described in (1), above. (See *Customizing the Display* for more details on fine-tuning.)

Tutorials in Chpater 23, particularly #2, provide more examples of querying Map Viewer by position.

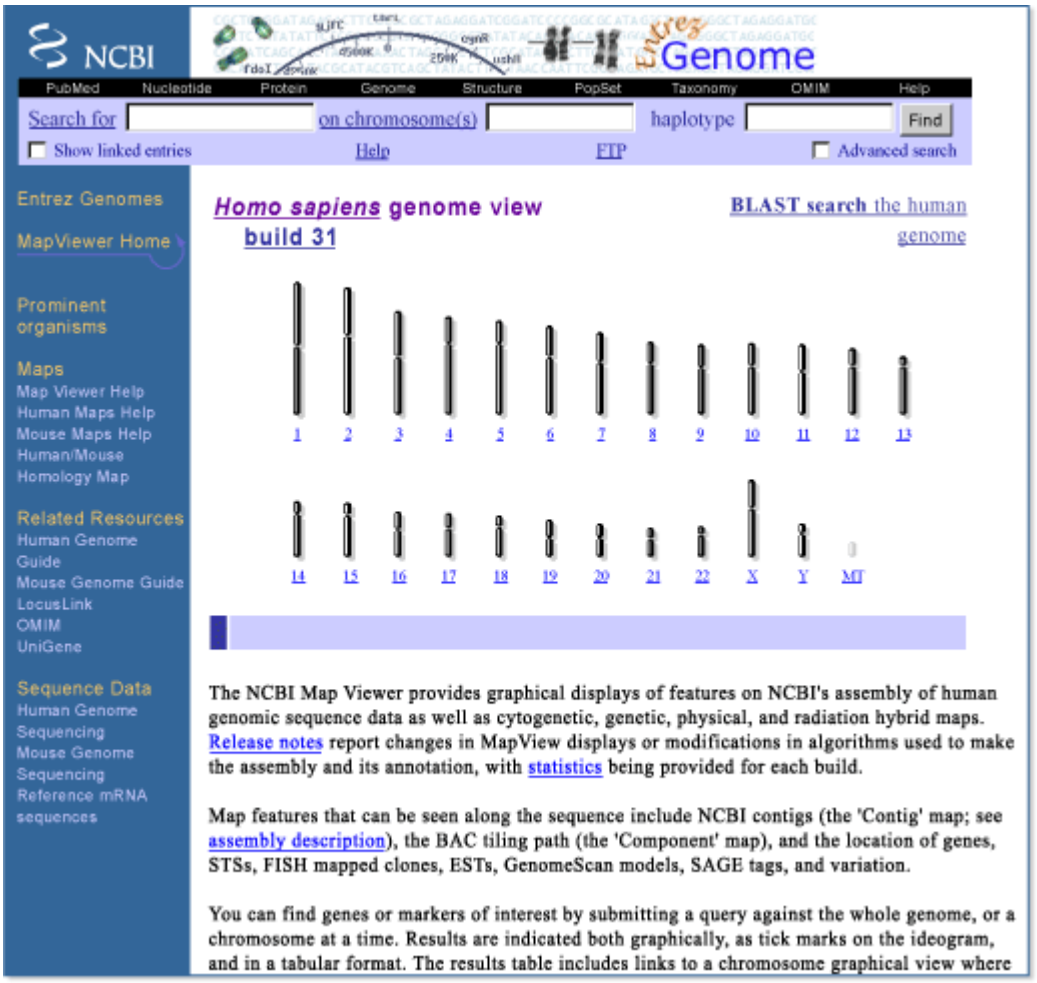


Figure 7. Representative of species-specific homepage. Note the links to the help documentation and related resources. Also note the check boxes to use the advanced query page and/or to display objects calculated to have links to any object returned by the query. A link to the genome-specific BLAST site is also provided at the *top* of the form.

Interpreting the Display

Map Viewer Summary Results

The results from a query are displayed both graphically and in a summary table (Figure 8). When the query is executed by BLAST, the graphical view is color-coded according to the BLAST score, and the table summarizes the scores and the RefSeq Accession numbers that have matches. Clicking on the RefSeq Accession number (i.e., those beginning with NT_ or NW_) displays that BLAST result in the Map Viewer.

Viewing the Maps

The Graphical Display

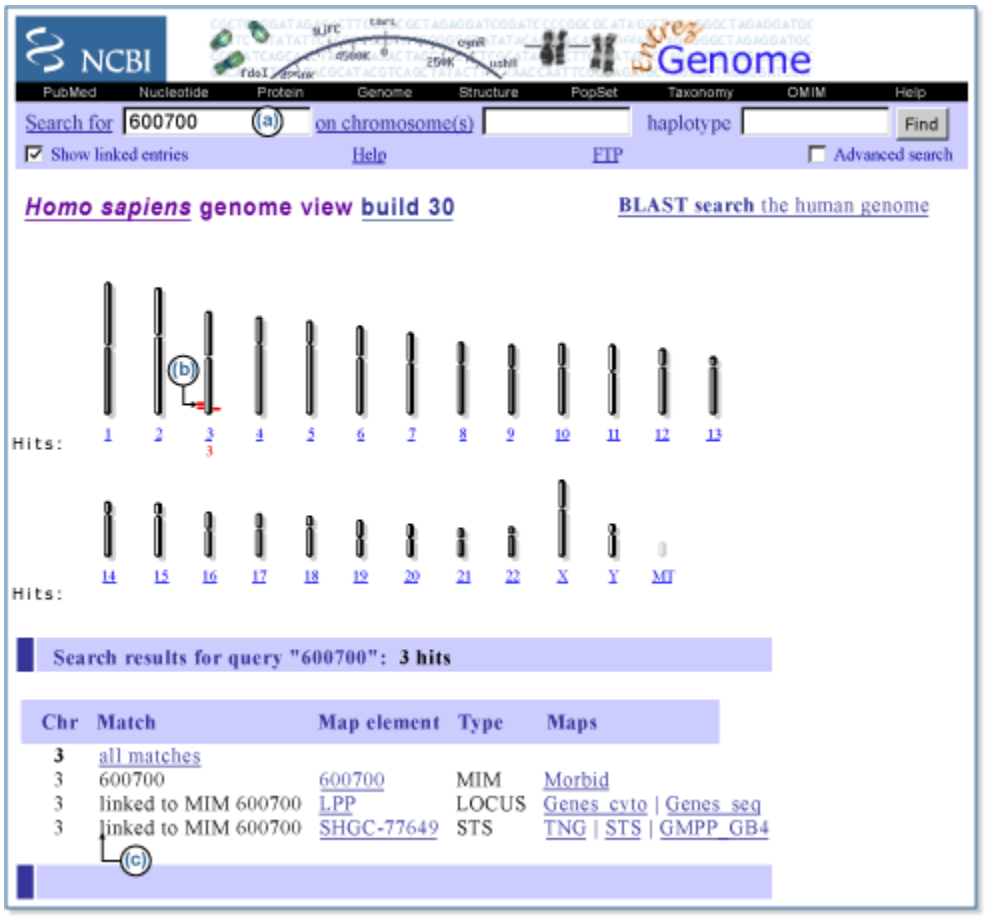


Figure 8. Map Viewer query results. (*a*) Note that an OMIM (MIM) number was entered as a query, and the **Show linked entries** box was checked. (*b*) The *red tick marks* next to the chromosome diagram indicate where the results appear to be placed on the chromosome. The left/right placement does not indicate strand; it allows more resolution between tick marks. (*c*) When a result is returned as the result of a link, the data in the **Match** column of the table begin with **linked to**. The complete results for a particular chromosome can be displayed by selecting the name of the chromosome on the graphical portion of the display or selecting **all matches** in the summary table. The number of matches per chromosome is reported under each chromosome in the graphical overview, and the total number of results returned is indicated below that. Additional pages are provided if the query returns over 100 results. The table indicates the chromosomal location, the match found, the map element returned, the type of match found, and the specific map(s) that contains the query match. Only the first 40 characters are shown in the *Match* column, and therefore the portion of text that matches the query may not be displayed. All maps that contain any object are viewed by selecting the name of the object in the *Map element* column. To see only one map, select the name of that map. In some cases, the resultant display will contain related maps in the same sequence coordinates. For example, selecting a sequence-based gene map may result in the display of mRNA alignments, labeled with UniGene cluster designations, and *ab initio* predictions.

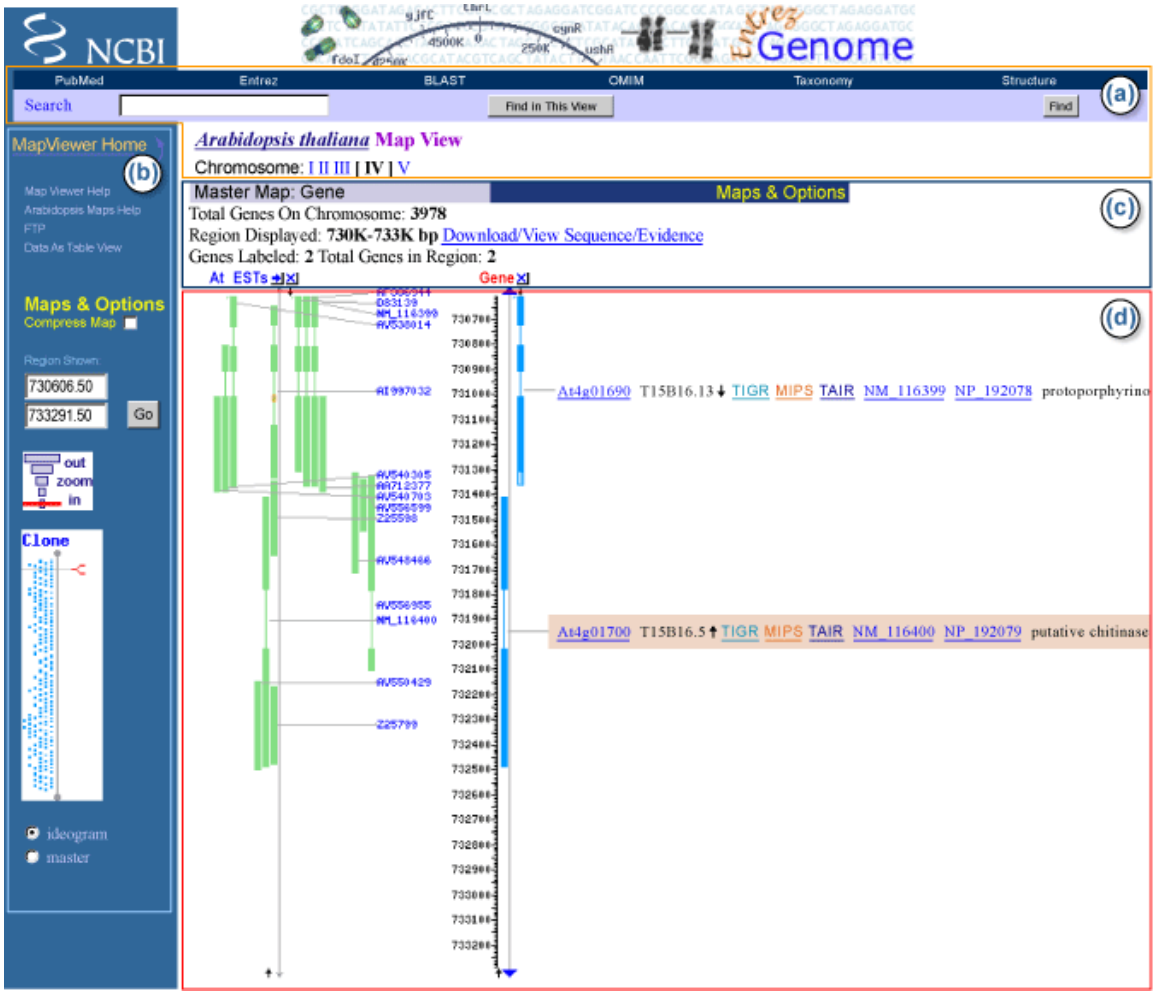


Figure 9. Representative map display. (*a*) Search and find options. (*b*) **Compress Map** allows you to select whether to display labels on maps other than the rightmost one. The **Region Shown:** and **zoom** graphic allow you to reset the range displayed either explicitly or by chromosome fractions, respectively. (*c*) Explanation and tools. This section reminds you of your chitin* query and provides statistics about the elements mapped on the Master Map for that chromosome. In this example, 3978 genes are annotated on chromosome IV, and 2 of 2 in the region are being displayed. The range shown is summarized, and links are provided to tools to download a region of the sequence and use ModelMaker and Evidence Viewer. The labels at the *top* of each map are connected to the documentation for that map, the -> makes the map become the master, and the **X** removes the map from the viewer. (*d*) Map. This section contains a representation of the order of elements in the range displayed and provides information about each, either by mousing over the label (uncompressed) or circle (compressed) of the non-master map, or by clicking on the links provided on the labels for the Master Map. This example shows that there is an option to display a ruler for any map. As described in the text and other figures, color is used in this display to convey information as well.

Text or Position Queries

General information on the chromosome being viewed is summarized at the top of the map page: the species and chromosome currently being viewed, the query term, and the name of the focal map, termed the Master Map (Figure 9c).

The summary also includes the following statistics concerning the number of objects on the Master Map, which are:

- the number of objects localized (positioned) on the chromosome
- the number of objects not localized but present on the chromosome
- the number of objects localized in the region displayed (i.e., the number decreases as you zoom in)
- the number of objects for which text descriptions are shown (dependent on user-defined page length)

A thumbnail map on the left of the page provides a coarse indication of the region displayed; by default, this is a cytogenetic map, although the Master Map can be selected (Figure 9b).

Maps are displayed vertically, with the name of each map hyperlinked to a description of it (Figure 9d). Features displayed on the Master Map have brief descriptive labels; information on features on the non-Master Maps can be found by mousing over an object. The labels on the Master Map depend on the type of object and genome being explored but can provide: (*a*) links to resources defining the mapped element, some of which may not be at NCBI; (*b*) indicators of the confidence in the placement or naming or sequence in the region; (*c*) biological features of the element (for SNPs, this includes position in a gene or effects on the coding region); (*d*) direction of transcription for genes; and (*e*) links to tools to facilitate reviewing of the sequence (**sv**), downloading a subsequence of interest (**seq**), the mRNA alignments in a region (**ev**), homology maps (**hm**), or to create cDNA sequences in real time (**mm**). (See the section on *Associated Tools* for more information.).

Sequence (BLAST) Queries

The positions of BLAST hits are highlighted on the Contig map, and a text summary of the BLAST hit is provided with links to regional alignment reports. All of the options described previously for configuring your display are still available. Thus, it is possible to evaluate the sequence match by the location (possible intron/exon structure, percent identity) as well as to determine whether the matching genomic region contains all of the query sequence in the expected order. Adding other maps to the display using the **Maps&Options** window provides a powerful mechanism to determine how the query sequence corresponds to existing annotation, such as genes, gene predictions, STS markers, or SNPs. For more hints, see the tutorial section on querying the human genome by sequence.

The Tabular Display (View Data as Table/Download)

A tabular report of the region and maps being displayed can be generated by selecting the **Data as Table View** link (Figure 9b). The default report is restricted to maps that were in the previous graphical display. Tables indicating the object name, or other identifier, and chromosome coordinates are provided for each map, along with many of the links seen in the graphical display. If the region being displayed on the map includes more than 1000 features per map, a warning message is displayed that points to the FTP site as an alternative for large-scale access.

If any of the maps are in sequence coordinates, an option is presented to report data for any sequence map in the region. Note: Links are provided for downloading tab-delimited files for any or all maps.

Customizing the Display

The Map Viewer display can be customized with regard to the region shown, the number and coordinate systems of maps, the number of objects labeled on the Master Map, and whether to show connections between objects. Each of these will be described in this section.

Selecting the Region to Display

The Map Viewer provides zoom, navigation, and other map display controls. These can be found on the display page itself and in the **Maps&Options** window (Figure 10).

As the resolution of a view is changed, the chromosome diagram is updated. The view automatically centers on a highlighted query term, or on the middle of the chromosome if browsing only. The chromosome view can be moved up and down or zoomed in and out. Zooming can be achieved in several ways: (1) by using the zoom control, located in the left column; (2) by providing a range or bounding markers in the **Region Shown** text boxes; or (3) by selecting part of the map to display a menu with predefined zoom levels. Most menu-based zooms should be carried out in two or three steps to avoid missing the region of interest. It is also possible to scroll or reposition the display by selecting **recenter** from the menu that pops up when you click on the chromosome diagram at the left or on a map or by clicking on the arrows at the top and bottom of a map.

Selecting the Maps (Tracks) to Display

Maps are categorized by the coordinate system as well as type of feature. The maps available for a genome can be seen by scrolling through the **Maps** menu in the **Maps&Options** window or in the genome-specific help documentation. For display and query purposes, different types of features annotated on sequence coordinates are treated as different maps. The maps in sequence coordinates are comparable because all of the sequence maps are based on the reference to a standard genome assembly. Thus, one can

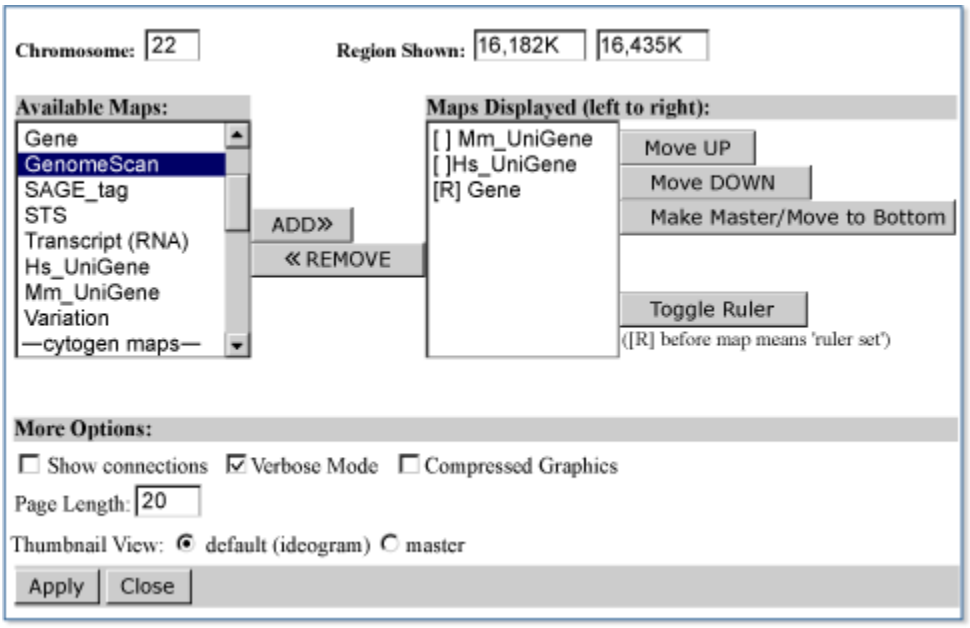


Figure 10. Representative Maps&Options window. The menus and boxes in this window allow definition of the range of the chromosome to display (**Region Shown:**), selection of the new map(s) to add to the display (**Available Maps:**), establishment of the order and ruler options [**Maps Displayed (left to right):**], control of the display of lines connected to related objects on different maps (**Show connections** box), control of the length of the label on the rightmost (Master) map (**Verbose Mode** box), compression of the graphic (**Compressed Graphics** box), control of the number of labels on the Master map (**Page Length:** box), and control of the diagram in the **Thumbnail View:**. To add map(s) to the display, select the map name(s) and then click on **ADD>>**. To remove map(s) from the display, select the map name(s) in the **Maps Displayed** box and then click on **<<REMOVE**. Please note that this is an example of a configuration that might be useful in displaying gene-related information, i.e., maps of UniGene, Gene, and GenomeScan.

display the SNP map (at high zoom level) next to the Gene, UniGene, or GenomeScan map to ascertain the number and location of polymorphisms in a region.

Some basic map controls are available directly on the display including removal of a map from the display by clicking on the **X** over the map and moving a secondary map to the Master Map position by clicking on the arrow next to the map label.

The **Maps&Options** window provides advanced options to: (*a*) add a ruler to any map; (*b*) reset the page length to display more (or less) information; (*c*) define region to display by providing coordinates or marker name in **Region Shown** boxes (also available directly on the Map Viewer display); (*d*) display direct connections between maps by checking the **Show connections** box; (*e*) optionally view text in **Verbose** or **Condensed** mode by selecting the checkbox. These user-defined preferences will be maintained for additional queries on different regions or chromosomes, until reset.

There has been considerable effort to integrate data on the sequence-based maps with data from non-sequence-based maps. Map connections provide a unique and powerful mechanism to identify features in a relevant region of the sequence map when starting

with information from a different coordinate system (see *Relationships among Coordinate Systems*).

The features that are available with Map Viewer are summarized in Box 1.

Box 1. Map Viewer-associated functions.

Query:

- Text
- Text, advanced
- Nucleotide query (by alignment or Accession number)
- Protein query (by alignment)
- By position in genome

Display Data:

- Graphical
- Tabular
- Assembled sequence
- Annotated feature sequence

Download:

- Sequence region
- Other map data for region
- Custom model (Model Maker)

Change Display Configuration:

- Zoom
- Scroll along chromosome
- Add/Remove tracks/maps
- Scalebar (ruler)
- Change order of track/map
- Specify coordinates to view
- Jump to different chromosome

Box 1 continues on next page...

Box 1 continued from previous page.

Show links

Alter number of rows displayed

FTP:

Assembled sequence

Model mRNA sequence

Model protein sequence

Contig/chromosome conversion tables

Map location, sequence-based

Map location, non-sequence-based

Links:

Help documentation

Statistics

FAQs

Associated Tools

Map Viewer provides links to several tools to display, download, or manipulate the sequence in a user-defined region. Whenever a sequence-based map is the master (the one at the right), the link **Download/View Sequence/Evidence** is provided above the map display. This opens a window that provides access to the **seq**, **ev**, and **mm** tools described below. In addition, when the annotated object is a gene (sequence or cytogenetic maps) or the species-specific UniGene cluster, the label may include these links.

The Evidence Viewer (**ev**) displays graphically the GenBank and RefSeq cDNAs that align to the genome in a particular region, along with a density plot for ESTs. The positions of any mismatches or insertions/deletions are marked, the multiple pairwise sequence alignments are provided, and computed translations are shown.

The Sequence Viewer (**sv**) is the Entrez graphical display option for any nucleotide sequence, focused on the gene indicated. By default, a 2-kb section of sequence is shown below the representation of the features, but that limit can be increased at the bottom of the page. It is also possible to zoom and navigate in the display.

Sequence Download (**seq**) provides the same function as the **Download/View Sequence** link provided at the top of the Maps page. The scope of the sequence passed to the tool corresponds to what is being viewed on the page. When connected to a gene feature, the

scope corresponds to that gene. The tool allows the user to alter the sequence scope and to select a report format (e.g., FASTA, GenBank, ASN.1). For the human and mouse genomes, a link is also provided to the Human–Mouse Homology Map (**hm**).

Model Maker (**mm**) displays the evidence for exons in a genomic region by diagramming the exons predicted from the alignment of cDNAs, from *ab initio* models (the default), and from alignment of ESTs (after an explicit selection). To facilitate construction of your own model transcript or transcripts, the splice junctions and the exons they connect are displayed, and the coding potential of any combination of exons can quickly be evaluated using ORFfinder. The sequence can also be edited, and the results can be saved or downloaded.

Technical Details

Data Access

The data displayed in Map Viewer are freely available. In addition to the view-specific reports, all of the data are available by FTP. README files document the content and format of each file. Genomic data are also available by chromosome; this includes genomic contigs (NT_ or NW_ Accession numbers) built from finished and unfinished sequence data. The contig data are available in various formats, including ASN.1, FASTA, GenBank, and GenPept. Also available in this directory are the RNAs (NM_, XM_ , and XR_ Accession numbers) and proteins (NP_, XP_).

Constructing URLs to Generate Specific Displays

Dynamic links to Map Viewer can be generated by constructing URLs with arguments that define the species, chromosome, range, types of maps (with or without units), display order, number of labels, query string, how to center a display around a query result, and the type of label for the display. The most current documentation is provided in the online help. The examples in Box 2, however, may illustrate the flexibility of the approach. Please note that the argument of the map in the URL is processed as an ordered list, with the order in the list controlling the left-to-right order in the display. Additional qualifiers control the display of a ruler and the range on the chromosome. If a query term is included as a part of the URL and that value cannot be identified on any of the maps in the list, that map will not be displayed.

Box 2. Examples of URL construction.

(a) Find the neighborhood (zoom=2) of the *HIRA* gene (chromosome 22) on all gene-containing human maps plus an ideogram, with the sequence map (loc) as the master map. Provide the detailed description of the genes (verbose=on). URL: http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?org=hum&chr=22&query=HIRA&zoom=2&maps=ideogr,morbid,gene,loc&verbose=on

Box 2 continues on next page...

Box 2 continued from previous page.

(b) Find human FISH-mapped clones (fish) in a cytogenetic region (coordinates are added to define the region) and also on the sequence map (clone). URL: http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?org=hum&chr=1&maps=clone,fish[1pter-p31]

(c) Show comparable regions on the human contig (cntg), component (comp), gene (gene,loc), and STS (sts) maps between the markers D7S726 and D7S2686. Show the ruler for the STS map (-r) and highlight the query terms. URL: http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?org=hum&chr=7&maps=cntg,comp,gene,loc,sts[D7S726:D7S2686]-r&query=D7S726+D7S2686

(d) Show potential genes (RefSeqs, ESTs, GenomeScan models) on a human genomic contig (NT_), with corresponding GenBank Accession numbers used to build the contig (displayed on the comp map) and FISH-mapped clones (on the clone map). URL: http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?ORG=hum&gnl=NT_023567&maps=cntg-r,clone,comp,scan,est,loc&query=NT_023567&cmd=focus

(e) Display mouse chromosome 6 on the radiation hybrid (rh) and genetic (mgi, wigen) maps and highlight the query term, D6Mit113. Zoom into 30% of the chromosome, with A2m in the center of that region. URL: http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?org=mouse&chr=6&maps=rh,mgi,wigen&query=D6Mit113&zoom=30

Implementation

Query terms are indexed for retrieval using the Entrez system. Thus, wild cards, Boolean operators, filters, and properties are managed as for other Entrez databases.

Each distinct object on the map is assigned a unique identifier that is specific to a particular build. Each object may have other secondary identifiers, such as IDs, in the sequence, Clone Repository, dbSNP, LocusLink, UniGene, or UniSTS databases. All descriptors are indexed as text. In addition, some are indexed by specific field values or by pre-identified properties, such as genes with associated diseases, SNPs with heterozygosity values in pre-defined ranges, or evidence type for genes. These field names or properties can be applied to restrict a query either in the Web-based query form or within a URL. The complete listings of current implementations for field qualifiers and properties are provided in the online help documentation.

Data for each map are retrieved for display from a relational database based on the IDs returned from the Entrez query. The database is used only to support display; it is refreshed with each NCBI build or update of any other map but not to track changes from build to build. Data from previous builds are archived at NCBI, but direct access is not currently supported.

Caveats for Using Evolving Data

Map Viewer displays represent the current synthesis of information available at the time of the data freeze (Table 3). It is important to understand that the underlying data may change from build to build, as our view of a genome becomes more refined. The data presented should always be critically reviewed, with a view to assessing the reliability of the assembly and annotation.

Means of reviewing reliability include: (*a*) noting the color coding of the contigs according to whether the sequence is draft or finished (this primarily applies to the human sequence); (*b*) noting the descriptions of the genes, STS, or SNPs to determine whether the element has been placed more than once; (*c*) checking that the STS order is the same on different maps; and (*d*) viewing features from different coordinate systems on the same map, e.g., showing STS features on the sequence (nucleotide coordinates), RH (cRay coordinates), and genetic maps (centiMorgan coordinates) to check for ambiguities. For more information, see the Pipeline FAQ /genome/guide/BuildFAQ.html.

Table 3. Web sites of interest.

Map Viewers	
Ensembl	www.ensembl.org
NCBI MapViewer	www.ncbi.nlm.nih.gov/mapview
UCSC Genome Browser	www. genome.ucsc.edu
Sequencing Information	
NHGRI Sequencing Information	www.nhgri.nih.gov/Data/
Celera Genomics	www.celera.com
Analysis tools	
BLAT	http://genome.ucsc.edu/cgi-bin/hgBlat?command=start
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
e-PCR	http://www.ncbi.nlm.nih.gov/ sts/epcr.cgi
Sim4 (mRNA to genomic alignment tool)	http://globin.cse.psu.edu/
Spidey (mRNA to genomic alignment tool)	http://www.ncbi.nlm.nih.gov/spidey
SSAHA	http://www.sanger.ac.uk/Software/analysis/SSAHA/
RepeatMasker	http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker
Maps	
BAC FingerPrint Map	http://genome.wustl.edu/gsc/human/Mapping/

Table 3 continues on next page...

Table 3 continued from previous page.

Other Annotation Sources and Viewers	
Celera Genomics	http://www.celera.com
DAS	http://www.biodas.org
DoubleTwist	http://www.doubletwist.com
The Genome Channel	http://compbio.ornl.gov/channel/
Incyte Genomics	http://www.incyte.com
FTP Sites	
Ensembl	ftp.ensembl.org/pub/current/data/
NCBI	ftp.ncbi.nlm.nih.nih.gov/genomes/H_sapiens/
UCSC	ftp.genome.cse.ucsc.edu/goldenPath

References

1. Schuler GD. Electronic PCR: bridging the gap between genome mapping and genome sequencing. Trends Biotechnol. 1998;16(11):456–459. PubMed PMID: 9830153.
2. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. Nature. 2001;409:953–958. PubMed PMID: 11237021.

Chapter 21. UniGene: A Unified View of the Transcriptome

Joan U. Pontius, Lukas Wagner, and Gregory D. Schuler

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The task of assembling an inventory of all genes of *Homo sapiens* and other organisms began more than a decade ago with large-scale survey sequencing of transcribed sequences. The resulting Expressed Sequence Tags (ESTs) were a gold mine of novel gene sequences that provided an infrastructure for additional large-scale projects, such as gene maps, expression systems, and full-length cDNA projects. In addition, untold numbers of targeted gene-hunting projects have benefited from the availability of these sequences and the physical clone reagents. However, the high level of redundancy found among transcribed sequences, not to mention a variety of common experimental artifacts, made it difficult for many people to make effective use of the data. This problem was the motivation for the development of UniGene, a largely automated analytical system for producing an organized view of the transcriptome. In this chapter, we discuss the properties of the input sequences, the process by which they are analyzed in UniGene, and some pointers on how to use the resource.

Expressed Sequence Tags (ESTs)

At a time when the genomes of many species have been sequenced completely, a fundamental resource expected by many researchers is a simple list of all of an organism's genes. A gene list, together with associated physical reagents and electronic information, allows one to begin to investigate the ways in which many genes interact in the complex system of the organism. However, many species of medical and agricultural importance have not yet been prioritized for genomic sequencing, and expressed cDNAs have provided the primary source of gene sequences. Furthermore, when the genomic sequence of an organism becomes available, a collection of cDNA sequences provides the best tool for identifying genes within the DNA sequence. Thus, we can anticipate that the sequencing of transcribed products will remain a significant area of interest well into the future.

The era of high-throughput cDNA sequencing was initiated in 1991 by a landmark study from Venter and his colleagues (1). The basic strategy involves selecting cDNA clones at random and performing a single, automated, sequencing read from one or both ends of their inserts. They introduced the term EST to refer to this new class of sequence, which is characterized by being short (typically about 400–600 bases) and relatively inaccurate (around 2% error). The use of single-pass sequencing was an important aspect of making the approach cost effective. In most cases, there is no initial attempt to identify or characterize the clones. Instead, they are identified using only the small bit of sequence

data obtained, comparing it to the sequences of known genes and other ESTs. It is fully expected that many clones will be redundant with others already sampled and that a smaller number will represent various sorts of contaminants or cloning artifacts. There is little point in incurring the expense of high-quality sequencing until later in the process, when clones can be validated and a non-redundant set selected.

Despite their fragmentary and inaccurate nature, ESTs were found to be an invaluable resource for the discovery of new genes, particularly those involved in human disease processes (2, 3). After the initial demonstration of the utility and cost effectiveness of the EST approach, many similar projects were initiated, resulting in an ever-increasing number of human ESTs (4–8). In addition, large-scale EST projects were launched for several other organisms of experimental interest. In 1992, a database called dbEST (9) was established to serve as a collection point for ESTs, which are then distributed to the scientific community as the EST division of GenBank (10). The EST division continues to dominate GenBank, accounting for roughly two-thirds of all submissions. The 20 organisms with the largest numbers of ESTs in the public database (as of March 7, 2002) are shown in Table 1.

One avenue to gene discovery is to use a database search tool, such as BLAST (11), to perform a sequence similarity search against dbEST. The query for such a search would be a gene or protein sequence, perhaps from a model organism, that is expected to be related to the human gene of interest. Because clone identifiers are carried with the sequence tags, it is possible to obtain the original material to generate a more accurate sequence or to use as an experimental reagent. For many EST projects, the IMAGE consortium (12) has been particularly instrumental in collecting the cDNA libraries, arraying the clones, and making the clones available for sequencing and redistribution.

For EST sequencing to be maximally productive, certain details of the library construction require some attention. For example, normalization procedures have been used to reduce the abundance of highly expressed genes so as to favor the sampling of rarer transcripts (13). More recently, subtraction techniques have been used to construct libraries depleted of clones already subjected to EST sampling (14). Although these techniques make it more efficient to find transcripts that are at low abundance in a particular tissue, it is possible that a small number of genes will still be missed because they are simply not expressed in tissues, cell types, and developmental stages that have been sampled.

Although ESTs are a useful way to identify clones of interest and provide guidance in identifying gene structure, a full-insert sequence of cDNA clones is preferable for both purposes. High-throughput full-insert cDNA sequencing projects have been the source of over 80,000 sequence submissions accessioned to date (August 2002). The full-insert cDNA sequence can allow identification of the translation product of the sequenced transcript, as well as potentially providing evidence for gene structure. Moreover, for the investigator wanting to use the clone as a reagent, having the accurate and complete sequence of the clone's insert at hand makes complete resequencing unnecessary, if the full-insert cDNA sequencing project makes clones available. Verifying that the full-insert

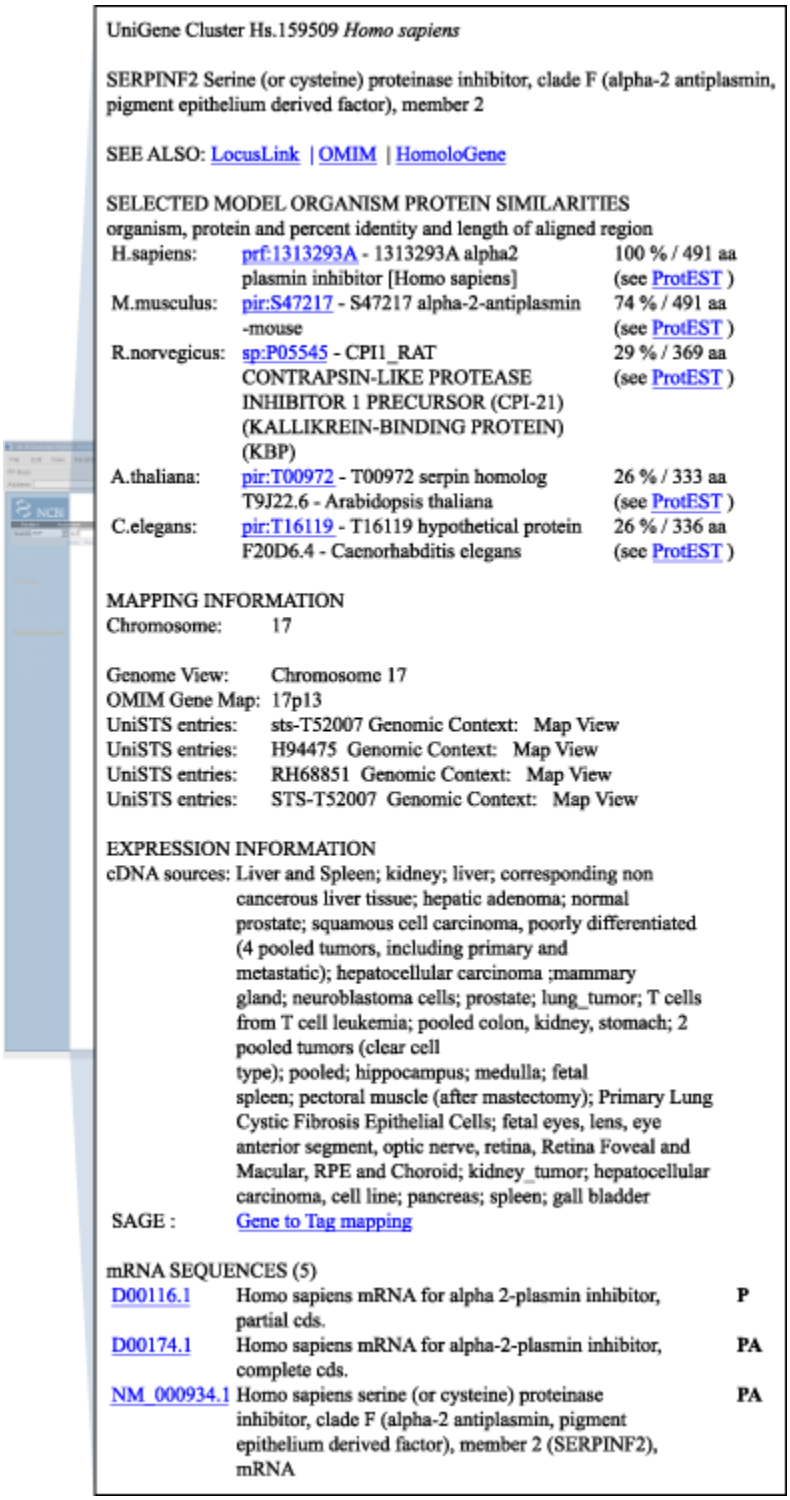
sequence corresponds to either the complete transcript of interest or to its complete, uncorrupted coding sequence is possible without committing laboratory resources and time to a clone that produced an EST. cDNA libraries do not generally include the entire transcript sequence; therefore, many full-insert sequences do not contain the entire transcription unit. Large transcripts (>6 kb) are particularly difficult to obtain.

Table 1. Top 20 organisms in dbEST (as of March 7, 2002).

Organism	ESTs
Homo sapiens (human)	4,070,035
Mus musculus (mouse)	2,522,776
Rattus norvegicus (rat)	326,707
Drosophila melanogaster (fruit fly)	255,456
Glycine max (soybean)	234,900
Bos taurus (cow)	230,256
Danio rerio (zebrafish)	197,630
Xenopus laevis (African clawed frog)	197,565
Caenorhabditis elegans (nematode)	191,268
Lycopersicon esculentum (tomato)	148,338
Zea mays (maize)	147,658
Medicago truncatula (barrel medic)	137,588
Arabidopsis thaliana (thale cress)	113,330
Chlamydomonas reinhardtii	112,489
Hordeum vulgare (barley)	104,803
Oryza sativa (rice)	104,284
Sus scrofa (pig)	103,321
Anopheles gambiae (mosquito)	88,963
Ciona intestinalis (sea squirt)	88,742
Sorghum bicolor (sorghum)	84,712

Sequence Clusters

The sheer number of transcribed sequences is extraordinary, indeed for most organisms much larger than the number of genes. A major challenge is to make putative gene assignments for these sequences, recognizing that many of these genes will be anonymous, defined only by the sequences themselves. Computationally, this can be thought of as a clustering problem in which the sequences are vertices that may be coalesced into clusters by establishing connections among them.



UniGene Cluster Hs.159509 *Homo sapiens*

SERPINF2 Serine (or cysteine) proteinase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2

SEE ALSO: LocusLink | OMIM | HomoloGene

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

H.sapiens:	prf:1313293A - 1313293A alpha2 plasmin inhibitor [Homo sapiens]	100 % / 491 aa (see ProtEST)
M.musculus:	pir:S47217 - S47217 alpha-2-antiplasmin -mouse	74 % / 491 aa (see ProtEST)
R.norvegicus:	sp:P05545 - CPI1_RAT CONTRAPSIN-LIKE PROTEASE INHIBITOR 1 PRECURSOR (CPI-21) (KALLIKREIN-BINDING PROTEIN) (KBP)	29 % / 369 aa (see ProtEST)
A.thaliana:	pir:T00972 - T00972 serpin homolog T9J22.6 - Arabidopsis thaliana	26 % / 333 aa (see ProtEST)
C.elegans:	pir:T16119 - T16119 hypothetical protein F20D6.4 - Caenorhabditis elegans	26 % / 336 aa (see ProtEST)

MAPPING INFORMATION
Chromosome: 17

Genome View: Chromosome 17
OMIM Gene Map: 17p13
UniSTS entries: sts-T52007 Genomic Context: Map View
UniSTS entries: H94475 Genomic Context: Map View
UniSTS entries: RH68851 Genomic Context: Map View
UniSTS entries: STS-T52007 Genomic Context: Map View

EXPRESSION INFORMATION
cDNA sources: Liver and Spleen; kidney; liver; corresponding non cancerous liver tissue; hepatic adenoma; normal prostate; squamous cell carcinoma, poorly differentiated (4 pooled tumors, including primary and metastatic); hepatocellular carcinoma ;mammary gland; neuroblastoma cells; prostate; lung_tumor; T cells from T cell leukemia; pooled colon, kidney, stomach; 2 pooled tumors (clear cell type); pooled; hippocampus; medulla; fetal spleen; pectoral muscle (after mastectomy); Primary Lung Cystic Fibrosis Epithelial Cells; fetal eyes, lens, eye anterior segment, optic nerve, retina, Retina Foveal and Macular, RPE and Choroid; kidney_tumor; hepatocellular carcinoma, cell line; pancreas; spleen; gall bladder
SAGE : Gene to Tag mapping

mRNA SEQUENCES (5)

D00116.1	Homo sapiens mRNA for alpha 2-plasmin inhibitor, partial cds.	P
D00174.1	Homo sapiens mRNA for alpha-2-plasmin inhibitor, complete cds.	PA
NM_000934.1	Homo sapiens serine (or cysteine) proteinase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2), mRNA	PA

Figure 1. Cluster view. A Web view of the UniGene cluster representing the human serine proteinase inhibitor gene *SERPINF2* is shown.

Experience has shown that it is important to eliminate low-quality or apparently artifactual sequences before clustering because even a small level of noise can have a large corrupting effect on a result. Thus, procedures are in place to eliminate sequences of foreign origin (most commonly *Escherichia coli*) and identify regions that are derived from the cloning vector or artificial primers or linkers. At present, UniGene focuses on protein-coding genes of the nuclear genome; therefore, those identified as rRNA or mitochondrial sequence are eliminated. Through the NCBI Trace Archive, an increasing number of EST sequences now have base-level error probabilities that are used to identify the highest quality segment of each sequence. Repetitive sequences sometimes lead to false alignments and must be treated with caution. Simple repeats (low-complexity regions) are identified using a word-overrepresentation algorithm called DUST, and transposable repetitive elements are identified by comparison with a library of known repeats for each organism. Rather than eliminating them outright, subsequences classified as repetitive are "soft-masked", which is to say that they are not allowed to initiate a sequence alignment, although they may participate in one that is triggered within a unique sequence. For a sequence to be included in UniGene, the clone insert must have at least 100 base pairs that are of high quality and not repetitive.

With a given a set of sequences, a variety of different sources of information may be used as evidence that any pair of them is or is not derived from the same gene. The most obvious type of relationship would be one in which the sequences overlap and can form a near-perfect sequence alignment. One dilemma is that some level of mismatching should be tolerated because of known levels of base substitution errors in ESTs, whereas allowing too much mismatching will cause highly similar paralogous genes to cluster together. One way to improve the results is to require that alignments show an approximate "dovetail" relationship, which is to say that they extend about as far to the ends of the sequences as possible. Values of specific parameters governing acceptable sequence alignments are chosen by examining ratios of true to false connections in curated test sets. It is important to note that the resulting clusters may contain more than one alternative-splice form.

Multiple incomplete but non-overlapping fragments of the same gene are frequently recognized in hindsight when the gene's complete sequence is submitted. To minimize the frequency of multiple clusters being identified for a single gene, UniGene clusters are required to contain at least one sequence carrying readily identifiable evidence of having reached the 3′ terminus. In other words, UniGene clusters must be anchored at the 3′ end of a transcription unit. This evidence can be either a canonical polyadenylation signal (15) or the presence of a poly(A) tail on the transcript, or the presence of at least two ESTs labeled as having been generated using the 3′ sequencing primer. Because some clusters do not contain such evidence (typically, they are single ESTs), not all uncontaminated sequences in dbEST appear in UniGene clusters. Of course, alternatively spliced terminal 3′ exons will appear as distinct clusters until sequence that spans the distinct splice forms is submitted. With the availability of genome sequence, a more stringent test of 3′ anchoring is possible, because internal priming can be recognized. Clusters that satisfy this more-stringent requirement can be identified by adding the term "has_end" to any

query. Specific query possibilities such as this one are listed under the rubric Query Tips on the UniGene homepage.

The UniGene Web site allows the user to view UniGene information on a per cluster, per sequence, or per library basis. Each UniGene Web page (Figure 1) includes a header with a query bar and a sidebar providing links to related online resources. UniGene is also the basis for three other NCBI resources: ProtEST, a facility for browsing protein similarities; Digital Differential Display (DDD), for comparison of EST-based expression profiles; and HomoloGene, which provides information about putative homology relationships.

UniGene Cluster Browser

The UniGene Cluster page summarizes the sequences in the cluster and a variety of derived information that may be used to infer the identity of the gene. Figure 1 shows an example of such a view for the human *SERPINF2* gene. When available, links are provided to a corresponding entry in other NCBI resources (e.g., LocusLink, OMIM) or external databases [e.g., Mouse Genome Informatics (MGI) at the Jackson Laboratory and the Zebrafish Information Network (ZFIN) at the University of Oregon]. Additional sections on the page provide protein similarities, mapping data, expression information, and lists of the clustered sequences.

Possible protein products for the gene are suggested by providing protein similarities between one representative sequence from the cluster and protein sequences from eight selected model organisms. For each organism, the protein with the highest degree of sequence similarity to the nucleotide sequence is listed, with its title and GenBank Accession number. The sequence alignment is described using the percent identity and length of the aligned region. Also provided is a link to ProtEST, which summarizes the UniGene protein similarities on a per protein basis.

The next section summarizes information on the inferred map position of the gene. In some cases, chromosome assignments can be drawn from other databases, such as OMIM or MGI. In other cases, radiation hybrid (RH) maps have been constructed using Sequence Tagged Site (STS) markers derived from ESTs. In these cases, the UniGene cluster can be associated with a marker in the UniSTS database, and a map position can be assigned from the RH map. More recently, map positions have been derived by alignment of the cDNA sequences to the finished or draft genomic sequences present in the NCBI MapViewer. For example, the *SERPINF2* gene in Figure 1 has a link to human chromosome 17 in the Map Viewer. The map is initially shown with a few selected tracks that are likely to be of interest, but others may be added by the user.

Although ESTs are a poor probe of gene expression, both the total number of ESTs and the tissues from which they originated are often useful. Both of these are displayed in the cluster browser. The tissues are listed under Expression Information, which includes the tissue source of libraries of the component sequences and, for human, links to the SAGE resource. Moreover, if genomic sequence is available, the UniGene map view displays expression for each exon (more precisely, for each portion of genome similar to a

transcript; because incompletely processed mRNAs are not unheard of, the presence of a transcript is insufficient to identify an exon).

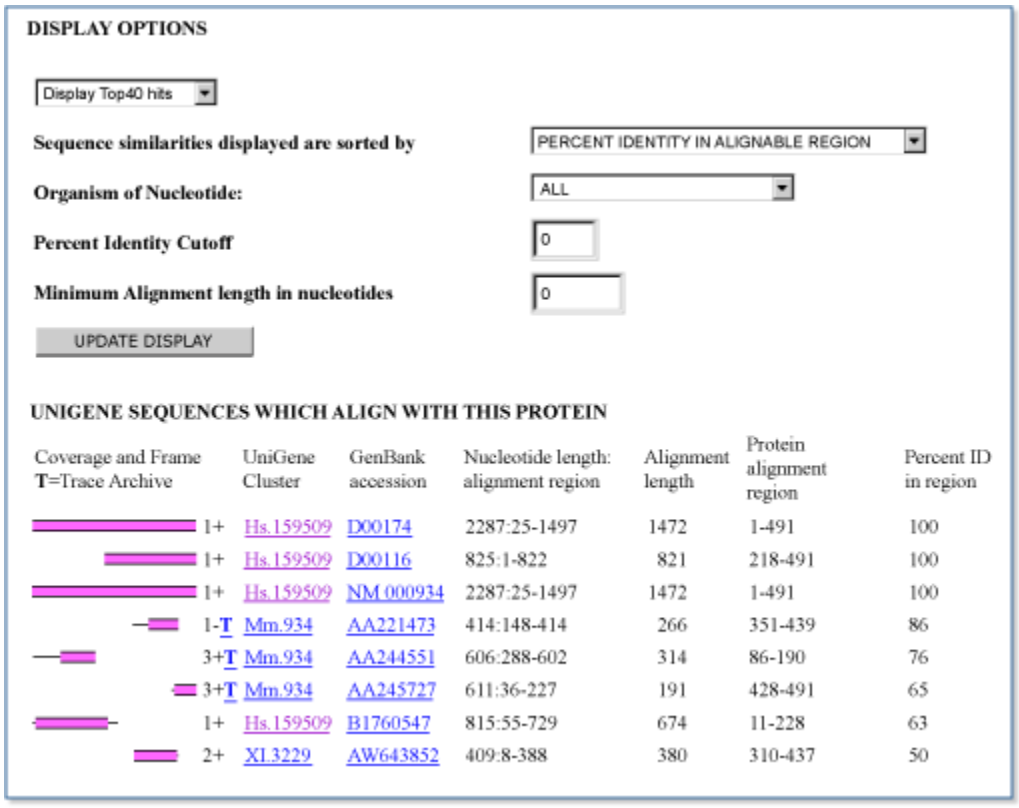
The component sequences of the cluster are listed, with a brief description of each one and a link to its UniGene Sequence page. The Sequence page provides more detailed information about the individual sequence, and in the case of ESTs, includes a link to its corresponding UniGene Library page. On the Cluster page, the EST clones that are considered by the Mammalian Gene Collection (MGC) project to be putatively full length are listed at the top, whereas others follow in order of their reported insert length. At the bottom of the UniGene Cluster page is an option for the user to download the sequences of the cluster in FASTA format.

Protein Similarity Analysis

The ProtEST section of UniGene allows the user to explore precomputed protein similarities for the cDNA sequences found in a cluster. The BLASTX program has been used to compare each sequence in UniGene to selected protein sequences drawn from eight model organisms: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*. These species were chosen as spanning a variety of taxonomic classes, as well as being well represented in the protein databases. To exclude proteins that are strictly conceptual translations and models, the proteins used in ProtEST are those originating from the RefSeq, SWISS-PROT, PIR, PDB, or PRF databases.

The ProtEST Web site has three features: information describing the amino acid sequence; information describing the nucleotide–protein alignments; and the ability for the user to modify various display options. The sequence alignments in ProtEST are summarized in tabular form (Figure 2). The first column is a schematic representation of the nucleotide–protein alignment. The width of the column represents the entire length of the protein, whereas the unaligned nucleotide sequence is represented as a thin gray line and the aligned region is represented as a thick magenta bar. The alignment representation is a hyperlink to the full alignment regenerated on-the-fly using BLAST. Other information in the table includes the frame and strand of the alignment, a link to the corresponding trace as provided in the NCBI Trace Archive, the UniGene cluster ID, the GenBank Accession number, and columns that describe the aligned region and percent identity.

To further refine the view, the sequence alignments in the table can be sorted by: (*a*) percent identity; (*b*) alignment length; (*c*) beginning coordinate of the alignment; (*d*) ending coordinate of the alignment; (*e*) UniGene cluster ID; or (*f*) GenBank Accession number. It is also possible to omit various rows of the table by restricting the display to a chosen organism or by choosing a cut-off value for the percent identity of the alignment and the length of the alignment.


DISPLAY OPTIONS

Display Top40 hits

Sequence similarities displayed are sorted by: PERCENT IDENTITY IN ALIGNABLE REGION

Organism of Nucleotide: ALL

Percent Identity Cutoff: 0

Minimum Alignment length in nucleotides: 0

UPDATE DISPLAY

UNIGENE SEQUENCES WHICH ALIGN WITH THIS PROTEIN

Coverage and Frame T=Trace Archive	UniGene Cluster	GenBank accession	Nucleotide length: alignment region	Alignment length	Protein alignment region	Percent ID in region
1+	Hs.159509	D00174	2287:25-1497	1472	1-491	100
1+	Hs.159509	D00116	825:1-822	821	218-491	100
1+	Hs.159509	NM_000934	2287:25-1497	1472	1-491	100
1-T	Mm.934	AA221473	414:148-414	266	351-439	86
3+T	Mm.934	AA244551	606:288-602	314	86-190	76
3+T	Mm.934	AA245727	611:36-227	191	428-491	65
1+	Hs.159509	B1760547	815:55-729	674	11-228	63
2+	XI.3229	AW643852	409:8-388	380	310-437	50

Figure 2. ProtEST view. A view of protein similarities for the human *SERPINF2* gene, found by BLASTX searching of a selected subset of the protein database, is shown.

Digital Differential Display (DDD)

DDD is a tool for comparing EST-based expression profiles among the various libraries, or pools of libraries, represented in UniGene. These comparisons allow the identification of those genes that differ among libraries of different tissues, making it possible to determine which genes may be contributing to a cell's unique characteristics, e.g., those that make a muscle cell different from a skin or liver cell. Along similar lines, DDD can be used to try to identify genes for which the expression levels differ between normal, premalignant, and cancerous tissues or different stages of embryonic development.

As in UniGene, the DDD resource is organism specific and is available from the UniGene Web site for that organism. For those libraries that have sequences in UniGene, DDD lists the title and tissue source and provides a link to the UniGene Library page, which gives additional information about the library. From the libraries listed, the user can select two for comparison. DDD then displays those genes for which the frequency of the transcript is significantly different between the two libraries. The output includes, for each gene, the frequency of its transcript in each library and the title of the gene's corresponding UniGene cluster. Results are sorted by significance, with the genes having the largest differences in frequencies displayed at the top. Libraries can be added sequentially to the

analysis, and DDD will perform an analysis on each possible library–gene pair combination. Similarly, groups of libraries can be pooled together and compared with other pools or single libraries.

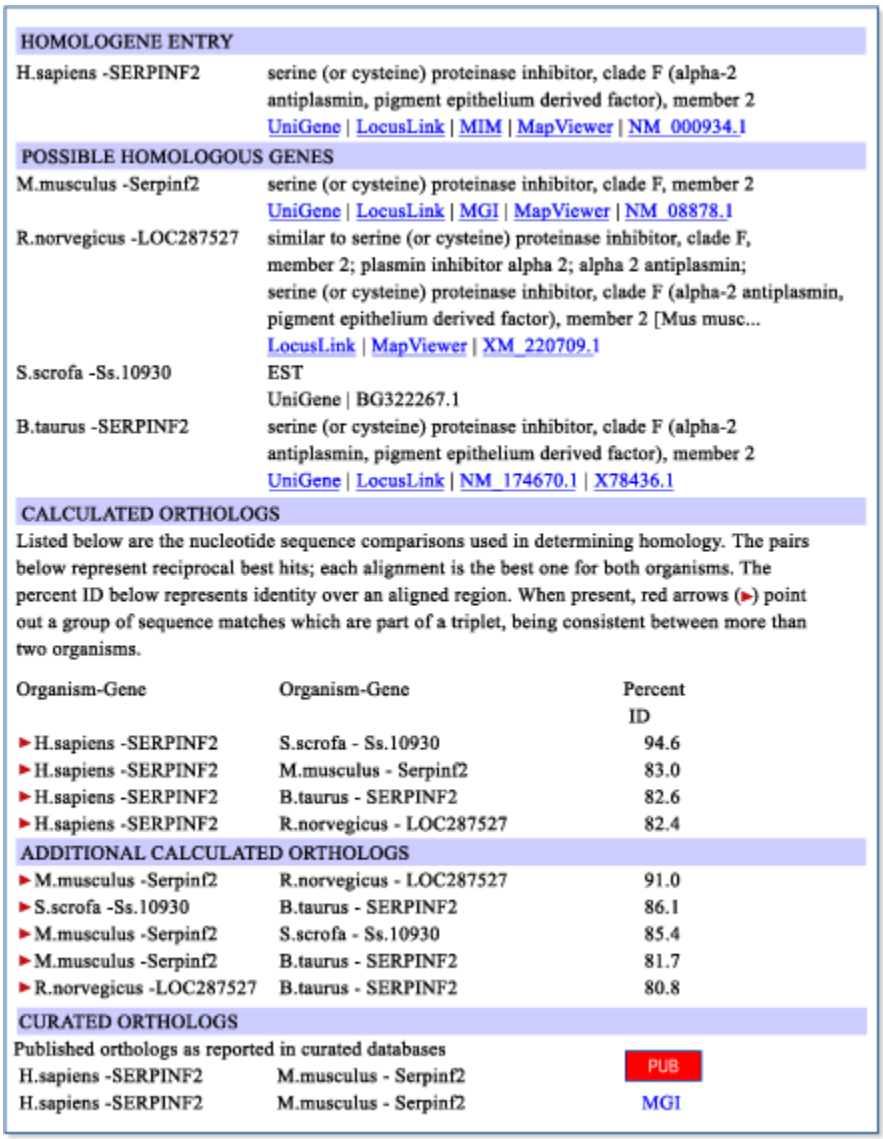
DDD uses the Fisher Exact test to restrict the output to statistically significant differences ($P \leq 0.05$). The analysis is also restricted to deeply sequenced libraries; only those with over 1000 sequences in UniGene are included in DDD. These requirements place limitations on the capabilities of the analysis. Unless there are a large number of sequences in each pool, the frequencies of genes are generally not found to be statistically significant. Furthermore, the wide variety of tissue types, cell types, histology, and methods of generating the libraries can make it difficult to attribute significant differences to any one aspect of the libraries. These issues underscore the need for more libraries to be made public and the need for the comparisons to be made using proper controls. Libaries generated by the Cancer Genome Anatomy Project (CGAP) will become especially valuable to this end. This project has resulted in a plethora of human libraries made from a variety of tissue types and generated using a variety of methods.

HomoloGene

HomoloGene is a resource for exploring putative homology relationships among genes, bringing together curated homology information and results from automated sequence comparisons. UniGene clusters, supplemented by data from genome sequencing projects, have been used as a source of gene sequences for automated comparisons.

Homology relationships, according to the experts who judge these, have been obtained from several sources. Collaborations with MGI and ZFIN at the University of Oregon have provided a large body of literature-derived data centered around *M. musculus* and *D. rerio*, respectively. Ortholog pairs involving sequences from *H. sapiens* and *M. musculus* have been imported from the NCBI Human–Mouse Homology Map. Additional information has been extracted from the literature by NCBI staff specifically for the HomoloGene project.

MegaBLAST (16) is used to perform cross-species sequence alignments and to identify those sequence pairs that share high degrees of nucleotide similarity. For each sequence, its best alignment with the sequences of the other organisms is retained. However, the best match for a sequence is not necessarily the best match for its partner sequence. For example, if there are several more sequences representing a particular gene in one organism than in the other organism, several sequences in one organism might have the same best match in the less well-represented organism. Similarly, if there are several paralogous genes in one species, they may find one identical homologous gene in another species. HomoloGene discriminates "one-way best matches" from cases where two sequences are each other's best match, or "reciprocal best matches", and only these reciprocal best matches are used. These sequence pairs are then used to find cross-species homologies between UniGene clusters. When reciprocal best matches are consistent



HOMOLOGENE ENTRY

H.sapiens -SERPINF2	serine (or cysteine) proteinase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 UniGene | LocusLink | MIM | MapViewer | NM_000934.1

POSSIBLE HOMOLOGOUS GENES

M.musculus -Serpinf2	serine (or cysteine) proteinase inhibitor, clade F, member 2 UniGene | LocusLink | MGI | MapViewer | NM_08878.1
R.norvegicus -LOC287527	similar to serine (or cysteine) proteinase inhibitor, clade F, member 2; plasmin inhibitor alpha 2; alpha 2 antiplasmin; serine (or cysteine) proteinase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 [Mus musc... LocusLink | MapViewer | XM_220709.1
S.scrofa -Ss.10930	EST UniGene | BG322267.1
B.taurus -SERPINF2	serine (or cysteine) proteinase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 UniGene | LocusLink | NM_174670.1 | X78436.1

CALCULATED ORTHOLOGS

Listed below are the nucleotide sequence comparisons used in determining homology. The pairs below represent reciprocal best hits; each alignment is the best one for both organisms. The percent ID below represents identity over an aligned region. When present, red arrows (►) point out a group of sequence matches which are part of a triplet, being consistent between more than two organisms.

Organism-Gene	Organism-Gene	Percent ID
►H.sapiens -SERPINF2	S.scrofa - Ss.10930	94.6
►H.sapiens -SERPINF2	M.musculus - Serpinf2	83.0
►H.sapiens -SERPINF2	B.taurus - SERPINF2	82.6
►H.sapiens -SERPINF2	R.norvegicus - LOC287527	82.4

ADDITIONAL CALCULATED ORTHOLOGS

►M.musculus -Serpinf2	R.norvegicus - LOC287527	91.0
►S.scrofa -Ss.10930	B.taurus - SERPINF2	86.1
►M.musculus -Serpinf2	S.scrofa - Ss.10930	85.4
►M.musculus -Serpinf2	B.taurus - SERPINF2	81.7
►R.norvegicus -LOC287527	B.taurus - SERPINF2	80.8

CURATED ORTHOLOGS

Published orthologs as reported in curated databases

H.sapiens -SERPINF2	M.musculus - Serpinf2	PUB
H.sapiens -SERPINF2	M.musculus - Serpinf2	MGI

Figure 3. HomoloGene view. Homology information for the mouse *Serpinf2* gene, with curated homologies for mouse and computed homologies extending to rat, zebrafish, and cow, is shown.

between three or more organisms, the pair is described as being part of a "consistent triplet".

The connections made by these methods result in a complex web of relationships. To simplify the Web view, it is useful to have each report page focus on an individual gene, called the "key gene", and to show connections that follow from it. An example of the report for the *M. musculus Serpinf2* gene is shown in Figure 3. The title of this key gene is shown at the top of the page, followed by genes from other species that show reciprocal best match relationships to the key gene. Each of these may have hypertext links to provide additional biological information about the gene. This is followed by a section

providing the curated homology information (if any), with links to the source of the data. Reciprocal best-match relationships are listed in the next two sections, first those directly involving the key gene and then those from a second round of walking that may be of interest. In each case, the description includes the sequence identifiers and percent identity of the alignment, with a hyperlink to reproduce a full alignment using BLAST.

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde RF, Moreno RF.et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991;252(5013):1651–1656. PubMed PMID: 2047873.
2. Sikela JM, Auffray C. Finding new genes faster than ever. Nat Genet. 1993;3(3):189–191. PubMed PMID: 8485571.
3. Boguski MS, Tolstoshev CM, Bassett DE Jr. Gene discovery in dbEST. Science. 1994;265(5181):1993–1994. PubMed PMID: 8091218.
4. Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsuba K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet. 1992;2(3):173–179. PubMed PMID: 1345164.
5. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nat Genet. 1993;4(4):373–380. PubMed PMID: 8401585.
6. Houlgatte R, Mariage-Samson R, Duprat S, Tessier A, Bentolia S, Larry B, Auffray C. The Genexpress Index: a resource for gene discovery and the genic map of the human genome. Genome Res. 1995;5(3):272–304. PubMed PMID: 8593614.
7. Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, Hawkins M, Hultman M, Kucaba T, Lacy M, Le M, Le N, Mardis E, Moore B, Parsons J, Prange C, Rifkin L, Rohlfing T, Schellenberg K, Marra M.et al. Generation and analysis of 280,000 human expressed sequence tags. Genome Res. 1996;6(9):807–828. PubMed PMID: 8889549.
8. Krizman DB, Wagner L, Lash A, Strausberg RL, Emmert-Buck MR. The Cancer Genome Anatomy Project: EST sequencing and the genetics of cancer progression. Neoplasia. 1999;1(2):101–106. PubMed PMID: 10933042.
9. Boguski MS, Lowe TM, Tolstoshev CM. dbEST: database for "expressed sequence tags". Nature Genet. 1993;4:332–333. PubMed PMID: 8401577.
10. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res. 2002;30(1):17–20. PubMed PMID: 11752243.
11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–3402. PubMed PMID: 9254694.
12. Lennon G, Auffray C, Polymeropoulos M, Soares MB. The I.M.A.G.E. consortium: an integrated molecular analysis of genomes and their expression. Genomics. 1996;33:151–152. PubMed PMID: 8617505.

13. Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci U S A. 1994;91(20):9228–9232. PubMed PMID: 7937745.
14. Bonaldo M, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. 1996;6:791–806. PubMed PMID: 8889548.
15. Wahle E, Keller W. The biochemistry of polyadenylation. Trends Biochem Sci. 1996;21(7):247–250. PubMed PMID: 8755245.
16. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7(1-2):203–214. PubMed PMID: 10890397.

Chapter 22. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes

Eugene V. Koonin

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The protein database of Clusters of Orthologous Groups (COGs) is an attempt to phylogenetically classify the complete complement of proteins (both predicted and characterized) encoded by complete genomes. Each COG is a group of three or more proteins that are inferred to be orthologs, i.e., they are direct evolutionary counterparts. The current release of the COGs database consists of 4,873 COGs, which include 136,711 proteins (~71% of all encoded proteins) from 50 bacterial genomes, 13 archaeal genomes, and 3 genomes of unicellular eukaryotes, the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, and the microsporidian *Encephalitozoon cuniculi*. The COG database is updated periodically as new genomes become available. The COGs for complete eukaryotic genomes are in preparation. The COGs can be applied to the task of functional annotation of newly sequenced genomes by using the COGnitor program, which is available on the COGs homepage.

Introduction

The recent progress in genome sequencing has led to a rapid enrichment of protein databases with an unprecedented variety of deduced protein sequences, most of them without a documented functional role. Computational biology strives to extract the maximal possible information from these sequences by classifying them according to their homologous relationships, predicting their likely biochemical activities and/or cellular functions, three-dimensional structures, and evolutionary origin. This challenge is daunting, given that even in *Escherichia coli*, arguably the best-studied organism, only about 40% of the gene products have been characterized experimentally. However, computational analysis of complete microbial genomes has shown that prokaryotic proteins are, in general, highly conserved, with about 70% of them containing ancient conserved regions shared by homologs from distantly related species. This allows one to use functional information from experimentally characterized proteins to suggest function in their homologs from poorly studied organisms. For such functional predictions to be reliable, it is critical to infer orthologous relationships between genes from different species. Orthologs are evolutionary counterparts related by vertical descent (i.e., they have evolved from a common ancestor) as opposed to paralogs, which are genes related by duplication (1, 2). Typically, orthologous proteins have the same domain architecture and the same function, although there are significant exceptions and complications to this generalization, particularly among multicellular eukaryotes.

The COGs database has been designed as an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept (3–5). The COGs reflect one-to-many and many-to-many orthologous relationships as well as simple one-to-one relationships (hence, orthologous groups of proteins). In addition to the classification itself, the COGs Web site includes the COGnitor program, which assigns proteins from newly sequenced genomes to COGs that already exist and to several functionalities that allow the user to select and analyze various subsets of COGs.

Construction of the COGs

COGs have been identified on the basis of an all-against-all sequence comparison of the proteins encoded in complete genomes using the gapped BLAST program (6; see also Chapter 16) after masking low-complexity (7) and predicted coiled-coil (8) regions. The COG construction procedure is based on the simple notion that any group of at least three proteins from distant genomes that are more similar to each other than they are to any other proteins from the same genomes are most likely to form an orthologous set (Figure 1). This prediction holds even if the absolute level of sequence similarity between the proteins in question is relatively low, and thus the COG approach accommodates both slow-evolving and fast-evolving genes.

Briefly, COG construction includes the following steps:

1. Perform the all-against-all protein sequence comparison.
2. Detect and collapse obvious paralogs, i.e., proteins from the same genome that are more similar to each other than to any proteins from other species.
3. Detect triangles of mutually consistent, genome-specific best hits (BeTs), taking into account the paralogous groups detected at step 2.
4. Merge triangles with a common side to form COGs.
5. Perform a case-by-case analysis of each COG. This analysis serves to eliminate false-positives and to identify groups that contain multidomain proteins by examining the pictorial representation of the BLAST search outputs. The sequences of detected multidomain proteins are split into single-domain segments, and steps 1–4 are repeated with the resulting shorter sequences, which assigns individual domains to COGs in accordance with their distinct evolutionary affinities.
6. Examine large COGs that include multiple members from all or several of the genomes using phylogenetic trees, cluster analysis, and visual inspection of alignments. As a result, some of these groups are split into two or more smaller ones that are included in the final set of COGs.

By the design of this procedure, a minimal COG includes three genes from distinct phylogenetic lineages; protein sets from closely related species were merged before COG construction. The approach used for the construction of COGs does not supplant a comprehensive phylogenetic analysis. Nevertheless, it provides a fast and convenient shortcut to delineate a large number of families that most likely consist of orthologs.

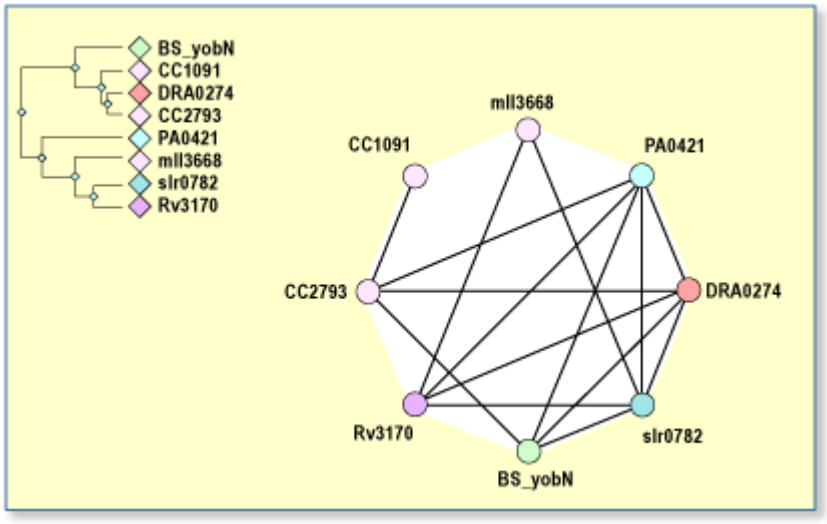


Figure 1. Example of a COG: monoamine oxidase. The COG for monoamine oxidase currently contains eight proteins from seven different organisms: one each from *Deinococcus radiodurans* (DRA0274), *Mycobacterium tuberculosis* (Rv3170), *Bacillus subtilis* (BS_yobN), *Synechocystis* (slr0782), *Pseudomonas aeruginosa* (PA0421), and *Mesorhizobium loti* (mll3668), and two paralogs from *Caulobacter crescentus* (CC2793 and CC1091). This is the only COG in the COGs database that has this phylogenetic pattern. In humans, monoamine oxidase is an enzyme of the mitochondrial outer membrane that seems to be involved in the metabolism of antibiotics and neurologically active agents and is a target for one class of antidepressant drugs.

The COGnitor Program

New proteins can be assigned to the COGs using the COGnitor program, the principal tool associated with the COGs database. COGnitor "BLASTs" the query sequences against all protein sequences encoded in the genomes that are classified in the current release of the COG system. To assign proteins to COGs, COGnitor applies the same principle that is embedded in the COG construction procedure, i.e., the consistency of genome-specific BeTs. For any given query protein, if the number of BeTs for a particular COG exceeds a predefined cut-off (three by default; the cut-off value can be changed by the user), the query protein is assigned to that COG; in cases where there are more than three BeTs to two different COGs, an ambiguous result is reported.

The Current State of the COGs Database, Updates, and Additional Classification of the COGs

Once the COGs have been identified using the above procedure, new members can be added using the COGnitor program. The assignments are further checked and curated by hand to eliminate potential false-positives. It has been shown that 95–97% of the COGnitor assignments typically require no correction (9). Once the proteins from a new genome are assigned to the appropriate pre-existing COGs through this combination of COGnitor and manual refinement, the remaining proteins from this genome are

compared to the proteins from non-COG proteins from previously available genomes, and an attempt is made to construct new COGs using the original procedure. In addition, when new sequences are added to an exisiting COG, the COG is examined for the possibility of a split (isolation of a new COG) by inspecting BLAST search outputs for all COG members and, in some cases, phylogenetic tree analysis. Thus, the number of COGs continuously grows through the construction of new COGs that typically include just a small number of species, whereas the number of proteins in the COG system increases primarily through the addition of new members to pre-existing COGs.

In bacterial and archaeal genomes, approximately 70% of the proteins typically belong to the COGs. Because each COG includes proteins from at least three distantly related species, this reveals the generally high level of evolutionary conservation of protein sequences, making the COGs a powerful tool for functional annotation of uncharacterized proteins. The COGs were classified into 18 functional categories that loosely follow those introduced by Riley (10) and also include a class for which only a general functional prediction (e.g., that of biochemical activity) was feasible, as well as a class of uncharacterized COGs. A significant majority of the COGs could be assigned to one of the well-defined functional categories, but the single largest class includes the functionally uncharacterized COGs. Additionally, the COGs were clustered according to the common metabolic pathways and macromolecular complexes.

Phyletic Pattern Analysis in COGs

A phyletic pattern is the pattern of species that are represented or not represented in a given COG; alternatively, phyletic patterns can be described in terms of the sets of COGs that are represented in a given range of species. The COGs show a broad diversity of phyletic patterns; only a small fraction are universal COGs, i.e., they are represented in all sequenced genomes, whereas COGs present in only three or four species are most abundant. This patchy distribution of phyletic patterns probably reflects the major role of horizontal gene transfer and lineage-specific gene loss in the evolution of prokaryotes, as well as the rapid evolution of certain genes in specific lineages, which may be linked to functional changes. Phyletic patterns are informative not only as indicators of probable evolutionary scenarios but also functionally; most often, different steps of the same pathway are associated with proteins that have the same phyletic pattern, whereas on some occasions, complementary patterns indicate that distinct (sometimes unrelated) proteins are responsible for the same function in different sets of species. The COG system includes a simple phyletic pattern search tool that allows the selection of COGs according to any given pattern of species. This tool effectively provides the functionality of "differential genome display" (for example, allowing the selection of all COGs that are present in one, but not the other, of a pair of genomes of interest) and can be helpful for delineating sets of candidate proteins for a particular range of functional features, e.g., virulence or hyperthermophily.

Description of the COGs Website

The main COGs Web page contains the following principal features: (*a*) a list of all COGs organized by the (predicted) functional category; (*b*) separate lists of COGs for each functional category and for a variety of major pathways and functional systems; (*c*) a table of co-occurrences of genomes in COGs; (*d*) a list of COGs organized by phyletic patterns; (*e*) the phyletic patterns search tool; (*f*) the COGnitor program; (*g*) a search engine to search COGs for gene names, COG numbers, and arbitrary text; and (*h*) Help, which covers the principal subjects related to COGs.

The individual COG pages can be reached from any of the COG lists mentioned above or by searching the site (see, for example, the COG for exonuclease I). Each of the COG pages shows the respective phyletic pattern in a table that also gives the ID number for the contributing sequence(s), a cluster dendrogram generated using the BLAST scores as the measure of similarity between proteins, and a graphical representation of BeTs for the given COG (not shown for the largest COGs). Also, each of the COG pages is hyperlinked to: (*a*) pictorial representations of BLAST search outputs for each member of the COG, which also includes links to the respective GenBank and Entrez-Genomes entries (see, for example, the link from XF2022, the protein from *Xyella fastidiosa* in the exonuclease I COG); (*b*) a multiple alignment of the COG members produced automatically using the ClustalW program (11); (*c*) a FASTA library of the protein sequences that belong to the COG (represented by the floppy disc icon); (*d*) the respective functional category of COGs and pathway (functional system) if applicable (in this exonuclease I example, the functional category L represents proteins involved in DNA replication, recombination, and repair); (*e*) a COG information page that includes functional, evolutionary, and structural information on the COG and its members (many of these pages are still under construction); (*f*) other COGs that include distinct domains of multidomain proteins that belong to the given COG through one of their domains; and (*g*) the Genome Context tool that shows the gene neighborhood around the given COG for all genomes that encode proteins of the given COG.

The COG data set and the COGnitor program also are available by anonymous ftp at ftp://ftp.ncbi.nih.gov/pub/COG.

Future Directions

Substantial evolution of the COGs is expected in the near future in terms of both growth by adding more genomes and the addition of new functionalities and layers of presentation. Quantitatively, the main forthcoming addition is the COGs for eukaryotic genomes, which are expected to approximately double the size of the COG system. Many of the COGs include paralogous proteins, and this will be addressed by introducing hierarchical organization into the COG system, whereby related COGs will be unified at a higher level. In addition, partial integration of the COGs with the NCBI's Conserved Domains Database (CDD) is expected (Chapter 3), which will result in a more flexible

and informative representation of the domain organization of proteins and of structural information that is available for COG members.

The COG Team

The COG system is developed and maintained by a team of programmers and expert biologists.

Project leader: Eugene V. Koonin.

The programming group: Roman L. Tatusov (group leader), Boris Kiryutin, Victor Smirnov, and Alexander Sverdlov (student)

The annotation group: Darren A. Natale (group leader), Natalie Fedorova, Anastasia Nikolskaya, Aviva Jacobs, Jodie Yin, B. Sridhar Rao, Dmitri M. Krylov, Sergei Mekhedov, John Jackson, Raja Mazumder, and Sona Vasudevan

References

1. Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970;19:99–106. PubMed PMID: 5449325.
2. Fitch WM. Homology: a personal view on some of the problems. Trends Genet. 2000;16:227–231. PubMed PMID: 10782117.
3. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997;278:631–637. PubMed PMID: 9381173.
4. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28:33–36. PubMed PMID: 10592175.
5. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001;29:22–28. PubMed PMID: 11125040.
6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–3402. PubMed PMID: 9254694.
7. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 1996;266:554–571. PubMed PMID: 8743706.
8. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science. 1991;252:1162–1164. PubMed PMID: 2031185.
9. Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV. Genome annotation using clusters of orthologous groups of proteins (COGs)—towards understanding the first genome of a Crenarchaeon. Genome Biol.*in press*. PubMed PMID: 11178258.
10. Riley M. Functions of the gene products of *Escherichia coli*. Microbiol Rev. 1993;57:862–952. PubMed PMID: 7508076.

11. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–4680. PubMed PMID: 7984417.

Part 4. User Support

Chapter 23. User Services: Helping You Find Your Way

David Wheeler and Barbara Rapp

Created: October 9, 2002; Updated: August 13, 2003.

Summary

The User Services team is the primary liaison between the public and the resources and data at NCBI. User Services disseminates information through outreach training programs and exhibits at scientific conferences and responds to incoming questions by email and telephone assistance. The team instructs people in the use of NCBI resources, responds to a wide range of questions, receives comments and suggestions, and coordinates with the NCBI resource developers to implement suggestions from users. In addition, User Services develops documentation, tutorials, and other support materials; produces the NCBI News; and publishes articles on NCBI resources.

The User Services Team

User Services consists of a staff of scientists and information specialists with diverse backgrounds and experiences. Scientist members of the staff hold Masters or Ph.D. degrees in an area of molecular biology, biochemistry, or biotechnology. Information specialists have Masters degrees in Library and Information Science and extensive experience using online databases of scientific information.

The Help Desk

Help Desk assistance is available from 8:30 a.m. to 5:30 p.m. Eastern Time, Monday through Friday. Two email addresses are available, info@ncbi.nlm.nih.gov and blast-help@ncbi.nlm.nih.gov. The info@ncbi.nlm.nih.gov address is for any type of inquiry, including questions about services; how to get started using the NCBI tools for a particular research problem; reports of technical problems; press inquiries; and comments or suggestions about NCBI resources. The blast-help@ncbi.nlm.nih.gov address is for questions and comments regarding sequence similarity searches using BLAST tools and databases. The Help Desk phone number is 301-496-2475.

email questions are answered as expeditiously as possible, usually within a day of receipt of the question. However, those that require extended investigation may take longer. Questions are usually handled directly by members of the User Services staff, although some are referred to a specific database development team for attention.

Examples of question topics include: data submission protocols, including the use of BankIt and Sequin (Chapter 12); finding summary information about a gene or disease; using Entrez (Chapter 15) to find sequence records for genes and proteins; choosing the best BLAST service to use for a particular application (Chapter 16); how to interpret

BLAST results; how to display and manipulate three-dimensional structures with Cn3D (Chapter 3); how to display genome data using the Map Viewer (Chapter 20); how to print or save search output; how to set up sequence databases and install NCBI software locally; and which databases to use for a specific research question. We also accept reports of possible data errors, suggestions of new features or content to include in databases, and reports of system bugs or Web problems. The NCBI also receives a number of press inquiries about the research applications of our services, bioinformatics in general, and various genome projects. Occasionally, a high-school student will submit questions for a classroom assignment, providing a special outreach opportunity to young scientists.

Because of the genetic focus of many of NCBI resources, we receive a number of questions from the general public regarding medical issues. The NCBI Help Desk staff can neither provide direct answers to medical questions nor give medical advice or guidance. However, we do provide suggestions on how to search our resources for information on the gene or condition of interest and refer users to the National Library of Medicine (NLM) customer service group for further assistance with PubMed (Chapter 2), MEDLINEplus, and ClinicalTrials.gov. We also refer them to outside organizations that can provide information on such topics as support groups and sources of medical advice.

Questions about PubMed are handled by a separate customer service group within the NLM. Their direct address is custserv@nlm.nih.gov, and their phone number is 1-888-FIND-NLM. PubMed questions that are received at the NCBI Help Desk are forwarded to NLM.

Development of User Support Materials

Because of its ongoing personal contact with our users, the User Services group plays an important role in communicating with database development and production teams, making suggestions, testing new releases and new features, and keeping them informed of problems that people are having with the services. The team also collaborates with developers in creating help documents, frequently asked questions (FAQs), tutorials, and workshop materials.

Tutorials on the Web

Web-based tutorials for BLAST, Entrez, Cn3D, and PubMed are currently available, with additional topics under development. Tutorials are produced on a collaborative basis by database development and User Services staff.

About NCBI

In keeping with the "plain language initiative" at NIH, the *About NCBI* section of the NCBI Web site presents many fundamentals of NCBI's bioinformatics tools and databases, including a science primer covering such topics as molecular genetics, genome mapping, Single Nucleotide Polymorphisms (SNPs) (Chapter 5), and microarray technology (Chapter 6). A model organism guide presents various model organisms and

their uses in laboratory settings. As an introduction and orientation to NCBI's multifaceted Web site, the *About NCBI* section appeals to the general public, educators, and researchers alike.

NCBI Site Map

The NCBI Site Map serves as a guide to NCBI resources. It provides a comprehensive, linked list of resources, along with a brief description of each resource. An effective way to locate a resource of interest within the Site Map is to perform a Find in Page search, a function that is built into all commonly used Web browsers.

Publications

The *NCBI News* is a quarterly newsletter that includes articles on new services, new features, and basic research at NCBI, as well as how to use selected resources for common applications. The newsletter is available free of charge and is offered online and by print subscription.

The User Services group also prepares fact sheets, brochures, and other public information materials to describe and illustrate NCBI services. A list of available materials is provided in the *About NCBI* section of the Web site, under News.

Overview articles entitled *GenBank* and *Database Resources of the NCBI* have also been published recently in the annual database issues of *Nucleic Acids Research* (1-3).

Outreach

NCBI's continuing emphasis on outreach to the scientific community is evident in its multifaceted program that includes exhibiting its services at scientific meetings, offering a variety of training courses, and developing Web-based tutorials and workshops.

Exhibits at Scientific Meetings

NCBI exhibits at approximately 15 scientific meetings per year, providing an opportunity for a wide range of researchers, students, and teachers to see demonstrations of NCBI resources and interact directly with NCBI staff. The current exhibit schedule is posted on the NCBI Web site in the *About NCBI* section, under *NCBI at a Glance*.

Workshops are offered at select scientific meetings and include the standing workshops described below in the Training section, but workshops also can be customized for particular audiences. Meeting organizers who would like to invite NCBI to offer a workshop are encouraged to do so.

Training Courses

NCBI has a growing training program consisting of full-day, half-day, and two-hour courses that are usually a combination of lecture and computer-based formats. There are also advanced courses that are given over a more extended time period. Each is described

briefly below, and further information on the training programs can be found in the Education section of the Web site, under *NCBI Courses*.

A Field Guide to GenBank and Other NCBI Resources

A Field Guide to GenBank and NCBI Molecular Biology Resources is a training course offered in a lecture format, followed by hands-on computer sessions. It is designed as a basic but broad introduction to NCBI tools and resources.

Field Guide topics include the following: description and scope of the primary database, GenBank (Chapter 1); derivative databases, such as UniGene (Chapter 21), Entrez Gene (Chapter 19), and Reference Sequence (RefSeq) (Chapter 18); effective database searching using Entrez; NCBI structure databases and the structure viewer, Cn3D; sequence similarity searching using the BLAST programs; the Conserved Domain Database (CDD) and associated search engine; and genome resources, including the NCBI assembly of the draft human genome, access to both finished and unfinished microbial genomes, and the genome Map Viewer.

The course is offered by invitation at academic institutions as well as at selected scientific conferences. If you are interested in hosting a course at your institution or conference, write to info@ncbi.nlm.nih.gov, and your request will be routed to the course coordinator.

The course is also offered four times a year at the NLM on the NIH campus in Bethesda, Maryland, and is free and open to anyone who would like to attend.

More information on this course is available in the Education section of the NCBI Web site, under *NCBI Courses*. The Web site includes the course handout, slide presentations, and problem sets with answers. A schedule of planned courses at NLM and elsewhere is posted under *Upcoming Courses*.

Molecular Biology Information Resources

This course is designed primarily for medical and science librarians or other professionals who are providing support services for molecular biology information resources. It provides an introduction to four categories of molecular biology information available from NCBI: nucleotide sequences, protein sequences, three-dimensional structures, and complete genomes and maps. An overview of search systems available at the NCBI, particularly Entrez and BLAST, emphasizes how search skills related to other types of information resources also apply to molecular biology databases. The course concludes with a discussion of various levels of molecular biology information services provided by librarians.

The Medical Library Association approves this course for eight continuing education credit hours. The course has been given at 24 locations since May 1997. Because of the increase in NCBI services, courses are being revised and are not being scheduled at this time.

More information on this course is available in the Education section of the NCBI Web site, under *NCBI Courses*. The Web site includes the course materials used for the lecture and a set of exercises.

NCBI Advanced Workshop for Bioinformatics Information Specialists

A new 5-day advanced course on NCBI resources has been developed as part of a collaborative project with a group of scientists and librarians who currently provide bioinformatics support services at their universities. The course provides detailed descriptive information as well as hands-on experience with handling a wide range of user questions. The course is designed for bioinformatics support staff based in university medical libraries so that they can, in turn, assist students, faculty, staff, and clinicians at their institutions in the use of molecular biology information resources. Additional information on this course is available in the Education section of the NCBI Web site.

Specialized Mini-Courses

The Service Desk staff also offers four mini-courses: BLAST QuickStart, Unmasking Genes in the Human Genome, Making Sense of DNA and Protein Sequences, and GenBank and PubMed Searching. Each is described briefly below. The purpose of the mini-courses is to focus on specific research application areas and address how to use multiple NCBI resources together to answer a research question. Additional problem-oriented mini-courses are under development.

The courses are 2 hours each in length. An overview is given during the first hour in lecture format, followed by a 1-hour hands-on session. Although primarily given on the NIH campus, NCBI is beginning to offer these workshops at outside institutions as well. Although the mini-courses were originally designed to be presented by an instructor, they are constructed in an online notebook format; therefore, it is possible to take the course on your own. Revisions to augment the online notebooks with lecture material and make the courses completely self-guided are currently under way.

BLAST QuickStart!

This mini-course is a practical introduction to the BLAST family of sequence-similarity search programs. Exercises range from simple searches to creative uses of the BLAST programs.

Unmasking Genes in the Human Genome

This mini-course covers how to find genes, promoters, and transcription factor-binding sites in human DNA sequences. It is designed around a program developed within User Services called Greengene, which integrates the output of several gene-finding tools and allows a coding sequence and accompanying protein translation to be assembled from the exons detected by these programs. Because the output of several programs is integrated, there is increased reliability in exon selection.

Making Sense of DNA and Protein Sequences

In this course, participants find a gene within a eukaryotic DNA sequence. They then predict the function of the derived protein by seeking sequence similarities to proteins with documented function using BLAST and other tools. Finally, a 3D modeling template is located for the protein sequence using the Conserved Domain Search (CDD-Search).

During the first hour, an instructor walks the class through an analysis of an uncharacterized *Drosophila melanogaster* genomic sequence from a GenBank record. During the second hour, participants perform the same analysis independently, using a different genomic sequence.

GenBank and PubMed Searching

This mini-course provides an overview of literature searching and sequence retrieval using the PubMed and Entrez database search interfaces. Exercises illustrate advanced search tips for using Entrez, many of which explore the use of the **Preview/Index** options for specifying parameters to limit the search results. The course also features 21 self-scoring exercises for GenBank.

The NCBI Learning Center

In addition to communication by email and phone provided through the Help Desk, a regular research consultation service provides one-on-one support for researchers in the NIH community. The consults are available by appointment and are provided in 1-hour time slots at the NIH Library as well as the NCBI training facility. Because of the success of the program, this type of service may be offered by appointment at selected scientific meetings in the future.

CoreBio

An innovative training program that began in 2001 aims to train molecular biologists for a new type of career as bioinformatics specialists who provide institutional support for users of computational biology tools. The NCBI Core Bioinformatics Facility (referred to as the CoreBio program) currently functions to train and support a network of bioinformatics specialists serving individual Institutes at NIH. NCBI's CoreBio facility trains Core members identified by their respective institutes in the use of its bioinformatics tools. The Core members, in turn, support the use of NCBI tools and databases by researchers at their institutes.

The training is provided over a 9-week period, with students attending lectures and completing practical exercises in the morning and returning to their regular workplace in the afternoon. The coursework centers on one major topic each week and follows the rough schedule given below:

WEEK 1: Introduction to the Sequence Databases

WEEK 2: BLAST

WEEK 3: The Human Genome

WEEK 4: Genomic Biology

WEEK 5: Molecular Modeling

WEEK 6: Web Page Development

WEEK 7: Setting Up a BLAST Web Server

WEEK 8: Interaction with Users

WEEK 9: Practicum

During week 9, the students pursue an institute-related project with the assistance of NCBI instructors. These projects run the gamut from the compilation of specialized datasets and data mining to the creation of novel BLAST interfaces and the construction of new data display tools. Students also develop a Web page to support the services they are developing for the respective Institutes at the NIH.

Although currently a NIH-based program, other organizations are welcome to consider using the program as a model for development of similar initiatives to meet their bioinformatics support needs.

Conclusion

At NCBI, we encourage our users to contact us with questions, suggestions, and requests for training or presentations on NCBI services. We invite feedback on tutorials, FAQs, and other support materials and welcome suggestions regarding additional materials that would be useful in guiding users through the wide range of services offered by NCBI.

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. GenBank. Nucleic Acids Res. 2002;30:17–20. PubMed PMID: 11752243.
2. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res. 2002;30:13–16. PubMed PMID: 11752242.
3. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2001;29:11–16. PubMed PMID: 11125038.

Chapter 24. Exercises: Using Map Viewer

David Wheeler, Kim Pruitt, Donna Maglott, Susan Dombrowski, and Andrei Gabrelian

Created: November 4, 2002; Updated: August 13, 2003.

Introduction

This chapter contains tutorials for using Map Viewer. Step-by-step instructions are provided for several common biological research problems that can be addressed by exploiting the whole-genome and positional perspectives of Map Viewer. Please be aware that the examples in these tutorials may return different results when you execute them, because the underlying data may have been updated, but we hope that the framework for obtaining, interpreting, and processing your results will be sufficiently clear if that happens. Most of the examples are for human genes, but the same logic applies to other genomes as well.

Please note that each of these tutorials is accompanied by a figure. If you are using this tutorial on the Web, we suggest that you open another browser window so that you can view the figure as you are reading the text. You are also encouraged to use Map Viewer interactively.

We welcome any suggestions that you have for improving the existing tutorials or for adding new ones.

1. How Do I Obtain the Genomic Sequence around My Gene of Interest?

There are many instances in molecular biological research when you may have only a cDNA sequence but need to have the nucleotide sequence that lies 5′ or 3′ to a gene or the introns for additional analyses. Because genomic sequence available from the public database may not have this annotation or may be so large as to make it difficult to retrieve only a region of interest, tools have been added to Map Viewer to make it easier to define, view, and download genomic sequence in multiple formats.

Direct Sequence Information via Map Viewer

From the Map Viewer homepage, select **Search** Homo sapiens (human) from the pull-down menu, enter FMR1 in the box labeled **for**, and then select **Go**. On the results page, four entries are returned (*4 hits*). *FMR1* has been mapped on three different maps: *Genes_cyto, Genes_seq,* and *Morbid*. Select the **Genes_seq** map to see the structure of the *FMR1* gene. Two links to the right of the Genes_seq map are of use in retrieving the 5′ and 3′ flanking DNA, the **sv** and **seq** links (*boxed*). At the top of the page, **Download/ View Sequence/Evidence** can also be used.

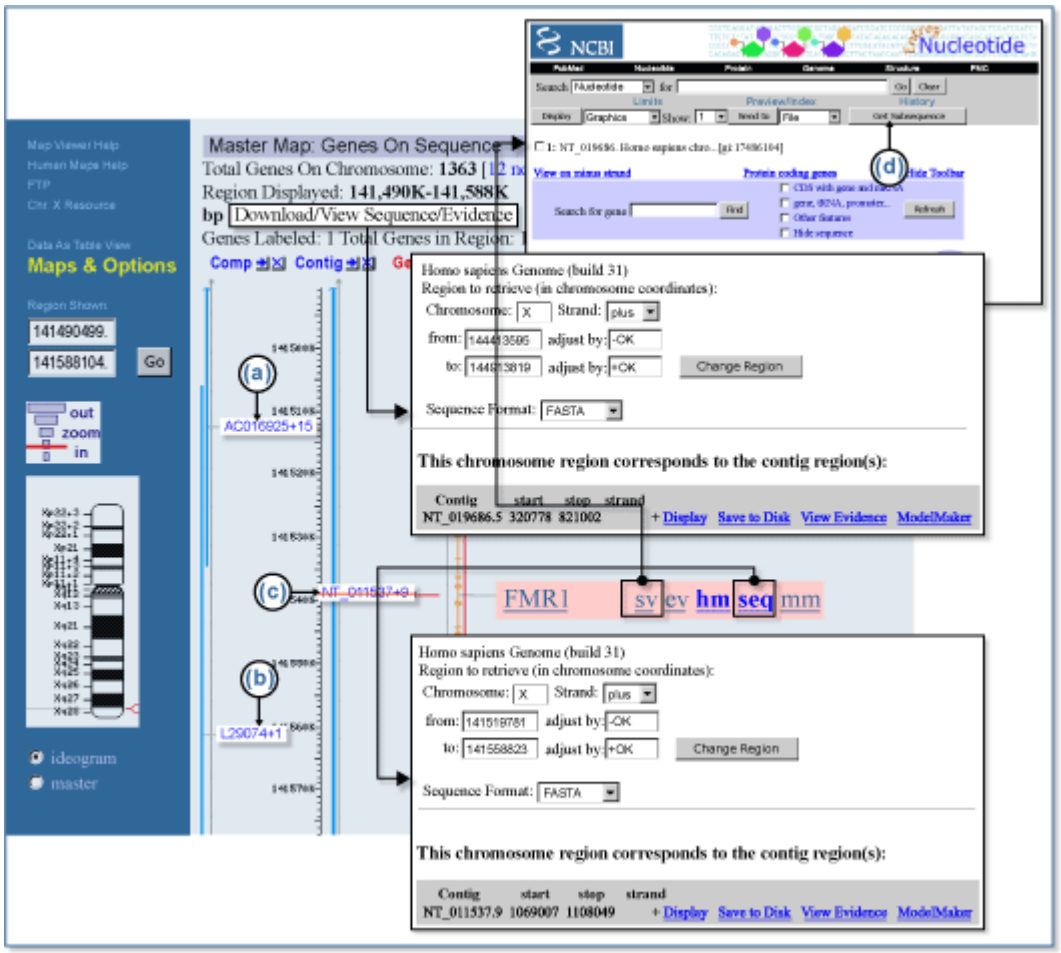


Figure 1. **Making a master map.**

The most informative maps, for the purpose of this example, are the Contig, Component, and Genes_seq maps. Selecting the **Genes_seq** map will display the current gene model. To start our search for flanking DNAs, display the **Contig** and **Component** maps. To do so, select **Maps & Options**, and from the available maps, choose the **Contig** map by left-clicking with the mouse. Select **ADD>>**. Now add the **Component** map. Next, make the Gene_seq map the master map in the display by left-clicking on the **Gene** map under the **Maps Displayed** (*left* to *right*) and selecting **Make Master/Move to Bottom:**. Finally, select the **Contig** map and choose the **Toggle Ruler** to add a ruler (*[R]*) to the display. This will guide you in finding your region of interest. Select **Apply**. From Figure 1, we can see that FMR1 has been annotated on a finished contig, NT_011537.9 (*c*), and that the contig is built from two components in this region, AC016925.15 and L29074.1 [(*a*) and (*b*), respectively].

There are two links on the Map Viewer display that are used to view and download of the region of interest: the **seq** link and the **Download/View Sequence/Evidence** link (*boxed*). The **seq** link is displayed only when the Gene_seq map is made the master map in the

display. Selecting the **seq** link will open a window, prompting the user to enter a region to retrieve. This region can then be refined further by adjusting the position, in kilobase pairs. Selecting **Display Region** will change the start and stop positions on the contig, where the gene has been annotated. The region can then be displayed and saved locally in FASTA or GenBank formats.

Selecting the **Download/View Sequence/Evidence** link from the main display page will generate the same window, but the initial coordinates are those spanning the entire region displayed by the Map Viewer rather than the region of a particular gene, as in the case above.

Let us assume that we would like to download 5.0 kb of upstream DNA and 1.0 kb of downstream DNA. To define this region, we will need to follow the **seq** link, which will open a new display showing the chromosome coordinates for *FMR1* and the corresponding position on the contig. To adjust the region, simply enter the amount of desired upstream DNA and downstream DNA into the two **adjust by:** input boxes provided, and select **Change Region**. Notice that the corresponding region on the contig has adjusted to reflect this change in position. Now we can either display the data in GenBank or FASTA formats and save the data to a disk.

Using the Sequence Viewer

Another tool for obtaining the desired sequence is the Sequence Viewer, available via the **sv** link when the Gene_seq map is the master map. The Sequence Viewer presents a graphical view of the gene within the contig. The sequence is also annotated with the coding regions, RNA and gene features, Sequence Tagged Sites (STSs), and single nucleotide polymorphisms. *Blue arrows* at the *top* and *bottom* of the display allow the user to navigate upstream or downstream. The **Get Subsequence** link (*large, open arrow*) at the *top* of the **sv** page allows the user to change the sequence range on the contig and will also display the reverse complement of the sequence. The specified region can then be displayed and saved locally in FASTA format.

2. If I Have Physical and/or Genetic Mapping Data, How Do I Use the Map Viewer to Find a Candidate Disease Gene in That Region?

In this example, we will use the Map Viewer to look for human candidate genes in a region. The types of queries that can be posted to the Map Viewer that will address this type of question are queries by genetic marker or STS.

Please note that Map Viewer supports queries by any named object positioned on a map so that it is possible to query by gene symbol or GenBank Accession number or any other object that might define your range of interest.

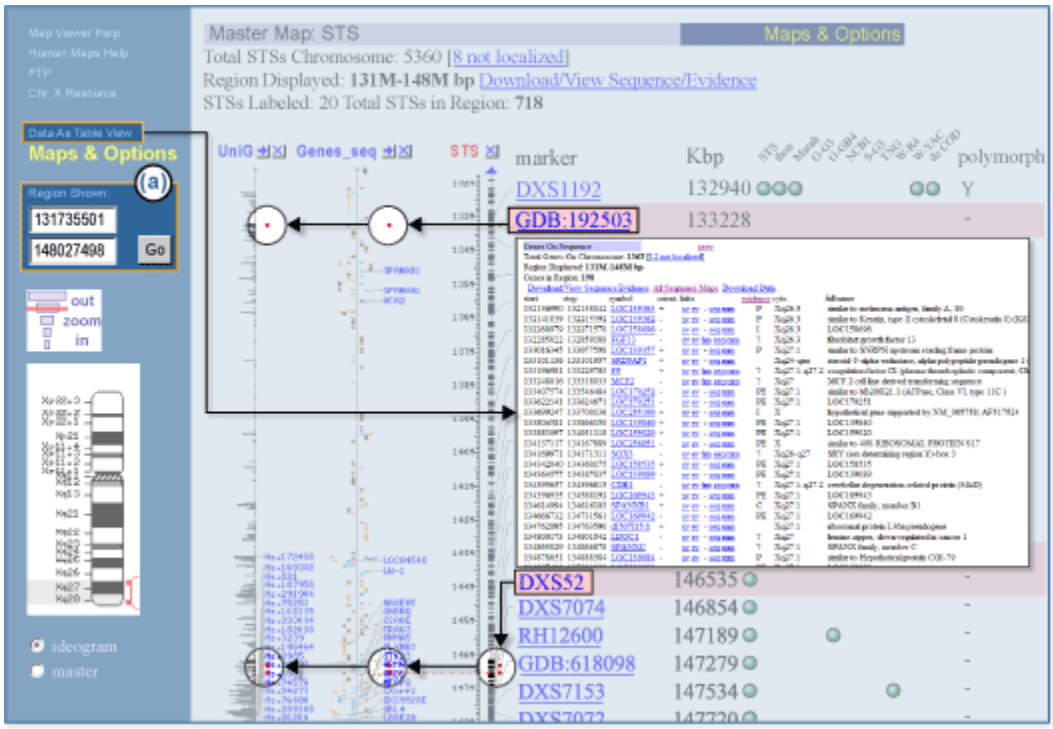


Figure 2. **Master Map: STS.**

Querying by STSs

To refine our search, we will enter the names of two STSs. In the text box, enter "sWXD113 OR DXS52" on chromosome X. Select **Find**. We can see that these two STSs map to the distal region of the long arm of the X chromosome, Xq, by the *red tick marks* that appear alongside the schematic of the X chromosome. These two STS markers have been mapped on several other maps that are also represented in the results. Select the **X** chromosome *above* the *red 3* to see both markers in the same display. The Map Viewer page now displays three different maps, each showing the physical location of these two markers.

The maps that are displayed include the UniG_Hs, Genes_seq, and STS maps. The UniG_Hs map shows the density of ESTs and mRNAs that align to the current assembly of the human genome. The Genes_seq map displays known and predicted genes that are annotated on the genomic contigs. The *rightmost* map, the STS map in this case, is termed the master map and contains descriptive information about each map element (Figure 2). The two STSs for which we are searching (*GDB:192503* and *DXS52*) are highlighted in *pink*. To the *right* of the display is a grid indicating other maps upon which these STSs are located. The *red dots* (*circled*) to the *left* of each highlighted STS show the relative position of these STSs in the context of the two other maps. By default, a ruler is displayed alongside the STS map so that the region of interest can be localized further. Notice on the

far left of the page that there is an area where you can enter the region that you would like to display [(*a*)].

At the current resolution, it is not possible to view all of the information displayed on the three maps. Therefore, some adjustment will be necessary.

Identifying a Candidate Gene

Narrow the region further using the ruler adjacent to the STS map as a guide. It is not necessary to display a ruler alongside the other two maps because they are all on the same coordinate system. In instances where sequence, genetic, cytogenetic, or radiation hybrid maps are being displayed in the same view, it is advisable to display additional rulers because the different maps show the mapped element on different coordinate systems (Kbp, cM, banding position, or centiRays, respectively). Enter the range 133.0 M to147.0 M in the *boxes* to the *left* of the page, in the **Region Shown** [(*a*)]. Select **Go**. This is a slightly better view; however, there are many more genes in the region that can be displayed.

To see all of the genes in the region defined by the two markers, select the **Data as Table View** link (*boxed*). The table that is generated lists all 144 genes in this region in a format easily read by people or computers. The table also preserves the links to additional gene-related resources seen in the graphical display, as well as reporting other objects in your displayed region. Please note that links are also provided to make it easy for you to download reports, not only of the objects in your map display, but other objects within the region defined by your display. This feature is especially useful if you are looking for other gene markers in your region of interest.

We can also change the page length under **Maps & Options** to a number large enough so that all genes are displayed graphically on a single page. By default, Map Viewer will show 20 map elements on a page. When the Genes_seq map is made the master map, there will be information at the *top* of the page, indicating how many genes have been labeled (n = 20) and how many genes are in the region (n = 144). To see all of the candidate genes in the specified region, go to **Maps & Options** and change the page length to the number of genes in the region (n = 144) and then select **Apply**.

Interpreting Your Results

At this point, you can now browse the description of the genes that are being displayed. Each gene or locus name is hyperlinked to LocusLink, where a detailed report about the gene or locus is provided. If the gene or locus of interest has supporting EST and mRNA data, then you can select the UniGene cluster number and link to UniGene, where more detailed information is provided about this gene, including its pattern of expression. LocusLink also provides connections to BLink and thus indirectly to reports of related proteins in the protein database and to viewers of protein structure, if your protein of interest is related to a protein for which the structure is known (see also Exercise 8 in this chapter).

Other Ways to Query

The example above summarizes the approach taken when defining a region of interest by entering names of markers in the query box. Gene symbols, reference SNP names, and GenBank Accession numbers for ESTs could also be used. It should also be noted that when a chromosome is displayed, you may also submit a query using the **Region Shown** boxes in the bar at the left side.

3. How Can I Find and Display a Gene with the Map Viewer?

In this example, we will locate and display the human gene implicated in Fragile X syndrome using the Map Viewer. We can find the gene beginning with several types of data. Refer to Figure 3. For these examples, we assume you are starting from the human-specific Map Viewer. If you instead are starting from the homepage, please remember to select "Homo sapiens (human)" as the species.

By Gene Symbol

If we are fortunate enough to know the official gene symbol for the Fragile X gene, *FMR1*, or an alternative symbol such as *FRAXA*, we can type the symbol into the search box at the *top* of the page and press the **Find** button. This gene has been annotated on the genome, and the gene symbol *FMR1* appears on the **Genes_seq, Genes_cyto,** and the **Morbid** maps. To generate a Map Viewer display that includes all three maps, select the link to **all matches**.

By Linkage to a Disease

Genes that are linked to a disease in Online Mendelian Inheritance in Man (OMIM) are referenced on the Morbid map and can be found by searching with a disease name or phenotype. In our case, "fragile-X" can be used. Using this query, we pick up hits to genes related to *FMR1*, as well as our intended gene. Selecting the **FMR1** link generates a display of the gene.

By Physical Marker

The *FMR1* gene contains the STS, STS-X69962 [(*b*)]; therefore, we can also use the name of this STS to find it, as in the case of a gene symbol. The search yields a table of hits showing that STS-X69962 appears on the STS map. Selecting the **STS** link gives us a Map Viewer display of the STS we found but not of the gene we sought. To get the gene into the Map Viewer display, we can add the Genes_seq map to the display. Although *FMR1* is located on the Genes_seq map rather than the STS map, because the coordinate systems used in different maps are synchronized, we can see the *FMR1* gene if we ask for the Genes_seq track in the region corresponding to the hit on the STS map. To do this, we can select the **Maps & Options** link [(*a*)], highlight the **Gene** map from the list of **available maps** in the *left-hand box*, and select the **ADD>>** button to add this map to the list of displayed maps. After selecting the **Apply** button, the Gene map is added to the Map

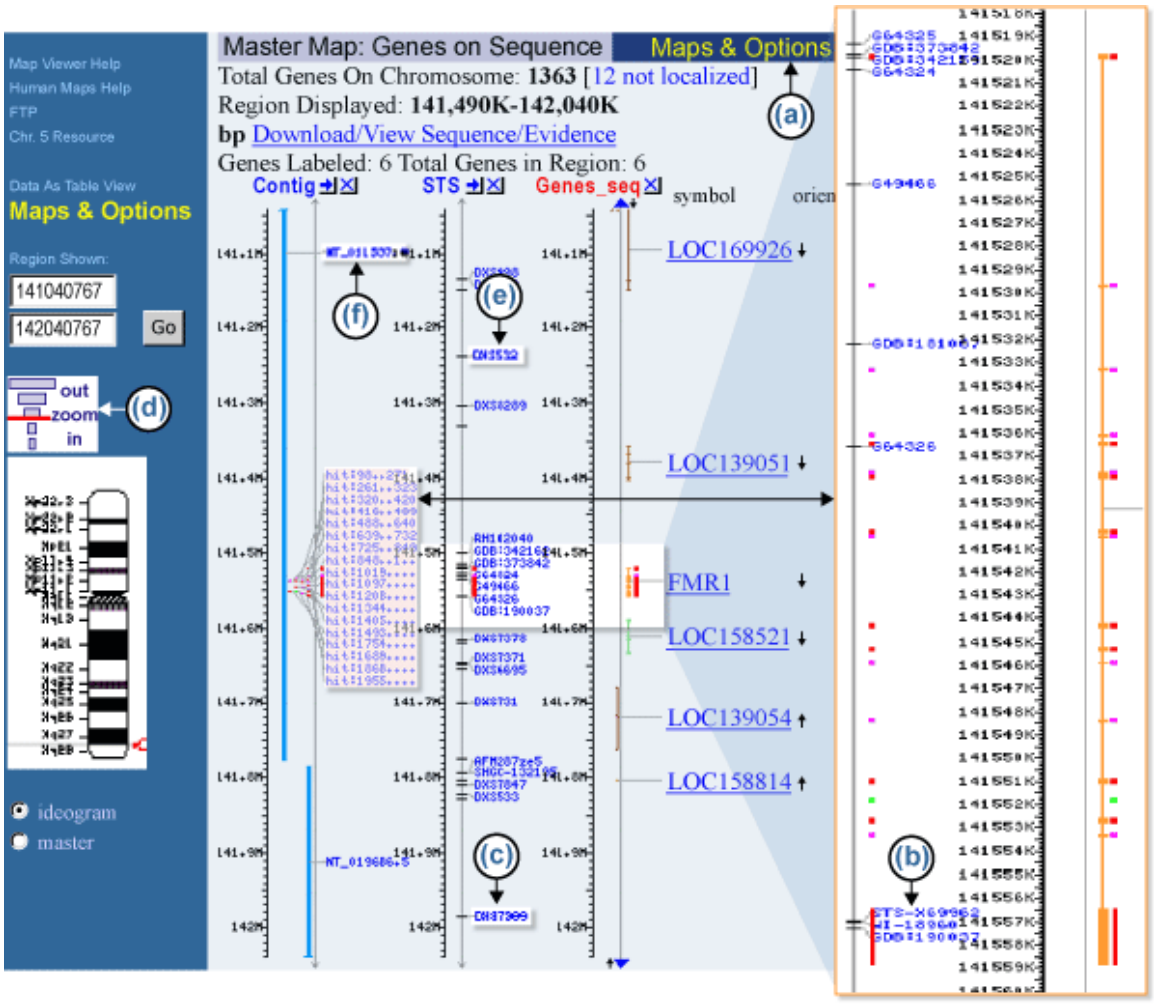


Figure 3. Location and display of the human gene implicated in Fragile X syndrome.

Viewer display. Note, however, that the view is limited to a very small 200-base pair portion of the gene. This is because we are still focused on the STS returned by our initial search. To see an expanded view, we must zoom out using the zoom control located directly over the thumbnail chromosome map in the *blue sidebar* [(*d*)]. Mousing over the control indicates that we are viewing 1/10,000th of chromosome X. We can click further up on the control to view 1/1,000th of the chromosome and see most of the *FMR1* gene. The *FMR1* gene is now centered in the Map Viewer display, and our STS hit is marked in *red*.

By Region

Often, a gene is known to reside only in a particular region. Suppose that we know only that *FMR1* resides somewhere between markers DXS532 and DXS7389 [(*e*) and (*c*), respectively]. We can use a query containing a Boolean OR to force the Map Viewer to search for both markers simultaneously. In this case, a query to Map Viewer of "DXS532

OR DXS7389" generates a number of hits to various physical maps, all to a region on chromosome X, which is marked in *red* under the **X** chromosome graphic. If we select the chromosome **X** link under the chromosome graphic, the Map Viewer display shows marker hits on several physical maps, highlighted in *red*, over a fairly large sequence region from about 141 to 142 megabases. To generate a tabular listing of all the genes in this region, select the **Data as Table View** link to the *left* of the Map Viewer display. With a genetic map, such as the Genethon map, as the master map, it is also possible to define or refine the region of display using coordinates in centimorgans. In the case of *FMR1*, entering a range of 176–198 into the **Region Shown** boxes generates a display of the genes falling within that range.

By Sequence Homology

Suppose that we have the sequence of the mouse homolog of the human *FMR1* gene and want to locate it on the human genome assembly. We might consult the Human/Mouse homology map at NCBI as the most direct approach for mapped genes, but let us assume that the human homolog of the *FMR1* gene is unmapped. In this case, we can perform a BLAST sequence similarity search with the mouse sequence to attempt to locate the corresponding human gene. We will use the mouse *Fmr1* mRNA sequence, taken from NCBI's LocusLink database (Accession number NM_002024), as our probe and follow the link to **BLAST search the human genome** located at the *top* of the Map Viewer search page; type the above Accession number into the BLAST form, and press the **Search** button. Such searches are extremely fast because they make use of an NCBI program called MegaBLAST, designed especially for this purpose. In the MegaBLAST results, the **Genomic View** button near the *top* of the page provides an entry point into a Map Viewer display. The MegaBLAST hits are indicated by a *red mark* on chromosome X, and links to hits are provided in the table at the *bottom* of the page, as in the searches by gene symbol or marker. In the case of any MegaBLAST search, all hits are to sequences [(*f*)] on the **Contig** map; therefore, it is the contig link (Accession number beginning with NT_) that we follow to view the hits in their genomic context. Following this link brings us to a display of the *FMR1* gene, with the BLAST hits indicated as highlighted "hits" on the contig map and *colored ticks* on the other maps (*large, expanded view*).

Other Ways to Query

Please note that other resources within NCBI also support querying for genes. Consider also LocusLink, UniGene, and Entrez Nucleotide. When a record of interest has been retrieved, each of these provides links to Map Viewer.

4. How Can I Analyze a Gene Using the Map Viewer?

We will analyze the *FMR1* gene. To find this gene in the human Map Viewer, enter FMR1 into the query box and select the **Genes_seq** map link in the table of search results.

To begin the analysis, we can select the link in the *blue sidebar* entitled **Data as Table View** to see a tabular listing of the chromosomal coordinates of the features visible in the

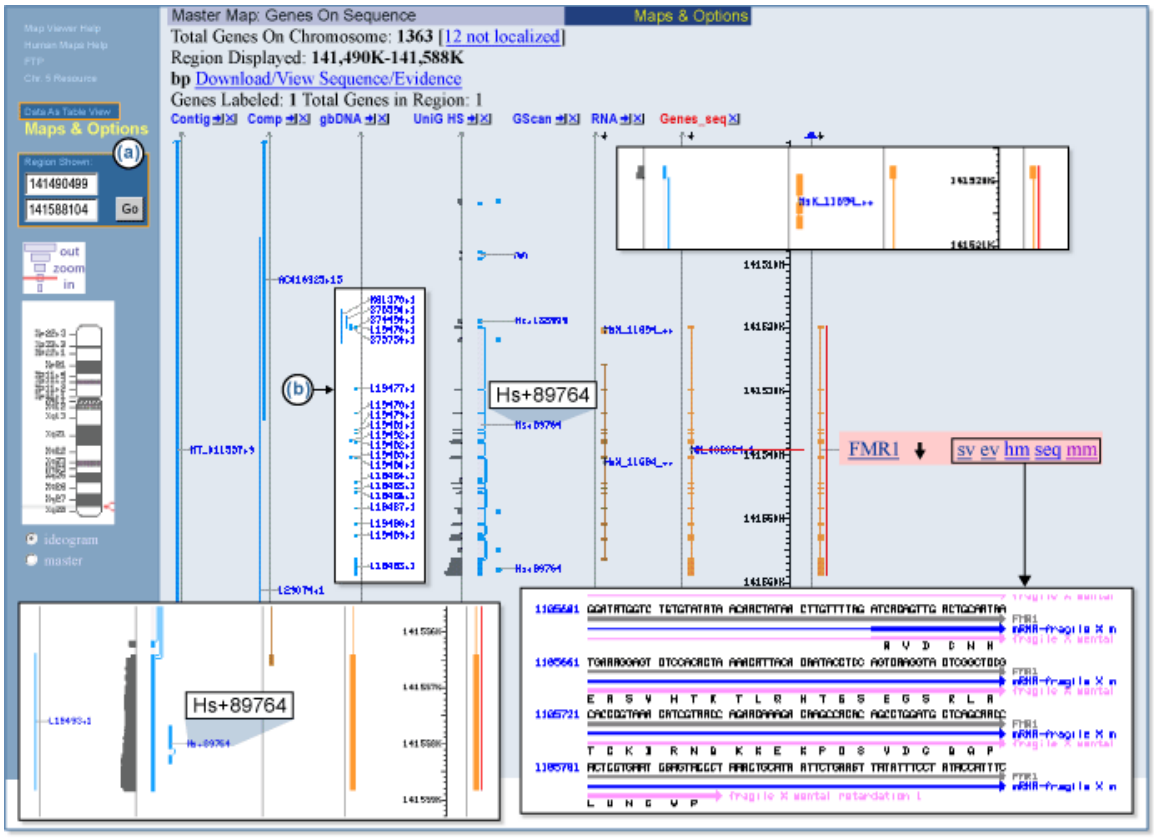


Figure 4. **Set of maps.**

current Map Viewer display. In the section of the table giving features on the Genes_seq map, we find that the *FMR1* gene extends from 141519781 to 141558823, a span of about 40,000 base pairs. We can now return to the graphical Map Viewer display and limit our view to this region by entering this range into the **Region Shown** boxes [Figure 4, (*a*)] in the *blue sidebar* and pressing the **Go** button. The coordinate system operative in the **Region Shown** boxes is that of the *rightmost* map in the Map Viewer display, called the master map; be sure that the Genes_seq map is the master before changing the coordinates for the region shown.

We are now ready to select the maps to display. The maps displayed will depend on the sort of analysis intended; however, one useful set of maps includes the Genes_seq, Contig, Comp, GScan, UniG_Hs, RNA, and gbDNA maps. This set of maps can be selected for viewing using the panel invoked by the **Maps & Options** link (*large, open arrow*). A Map Viewer display of the *FMR1* gene using this set of maps is given in Figure 4.

The Genes_seq Track

In Figure 4, the Genes_seq map, which displays annotated genes, is the master map and is the first one we will examine. The *FMR1* gene comprises 18 exons, represented by *thick lines*, interspersed over about 40,000 base pairs of sequence. The gene is drawn to the *right*

of the Genes_seq track line and therefore runs from the *top* of the display to the *bottom* and is coded on the "plus" strand in the human genome assembly. Genes located on the opposite strand run from the *bottom* of the display upward and are shown to the *left* of the Genes_seq track. The *FMR1* gene is displayed in *orange*, which indicates that the alignment between the genomic sequence and the *FMR1* transcript sequences that were used to produce the gene model was not perfect; *blue* alignments are of the best quality. The 3′-most exon (*bottom inset*) of the gene is exceptionally large and probably includes a significant untranslated region. To verify this, we can select the **sv** link (*boxed*) to the *right* of the Genes_seq map to invoke the Sequence Viewer, which shows the sequence of the *FMR1* gene. In the Sequence Viewer, we can navigate to the display for the last exon, number 18, to see that the coding sequence ends toward the beginning of the exon and that the majority of the exon is indeed untranslated.

The RNA Track

The RNA or Transcript map shows the alignment of a single mRNA sequence to the genome, and in this case, the pattern of exons produced matches exactly that shown on the Genes_seq map. If additional splice variants are sequenced, multiple alignments will be shown on the RNA track, and the gene model given on the Genes_seq track will be a composite model made up of all the exons implied by these alignments.

The GScan Track

The GenomeScan track shows gene predictions made using GenomeScan that are independent of supporting mRNA alignments. The GenomeScan model for *FMR1* is very similar to the model shown on the Genes_seq track; however, there are differences, and the alignment-based model shown on the Genes_seq track covers exons that are part of two different GenomeScan models (*topmost inset*). Note also that the GenomeScan model covers only the initial translated portion of the large 3′ exon of the transcript-based model (*bottom inset*).

The UniG_Hs Track

Both the predicted model (GScan) and the alignment-based model (Genes_seq) can be compared to the mapping of ESTs on the UniG_Hs track. The *bars* extending to the *left* of the UniG_Hs track line depict EST mapping density, whereas the *lines* to the *right* connect ESTs arising from common UniGene clusters. From the UniG_Hs track, it is clear that most of the exons arising from either of the two gene models have some EST support, and that most of the ESTs that map to these regions are members of UniGene cluster Hs.89764 (*boxed, center*). Selecting the **Hs.89764** link leads to the *FMR1* UniGene cluster.

The UniG_Mm Track

This map is comparable to the UniG_Hs track but is based instead on alignment of mouse cDNA sequences (conventional and EST). In this example, the exons suggested by the alignment of mouse cDNAs is comparable to that based on alignment of human cDNAs.

SAGE_tag

SAGE_tag provides another view of expression levels and connections to more information about the tissue of origin of the expressed sequences. The SAGE_tag map also provides a histogram of expression, and each tag is connected to a tag-specific report page.

The Contig Track

Looking across to the Contig track, we can see that the gene maps to a contig that is drawn in *blue*, indicating that the contig is derived from high quality, finished sequence. If we consult the Component map (*Comp*), we can see that the portion of the contig containing the *FMR1* gene is composed of two overlapping finished sequences, also drawn in *blue*. Because the sequence underlying the *FMR1* gene is finished, rather than draft sequence, the *FMR1* sequence and structure are likely to remain stable in future human genome assemblies.

The gbDNA Track

Additional GenBank sequences that align to the genome but were not part of the assembly are shown on the gbDNA track. In this case, a number of short sequences for the individual exons of *FMR1*, the product of intense research on this gene, are aligned to the genome (Figure 4b).

5. How Can I Create My Own Transcript Models with the Map Viewer?

The Map Viewer displays the alignment of transcripts, such as mRNA GenBank sequences and RefSeqs, to genomic sequence and shows the positions of predicted genes, but it does not stop there. By using a utility called the ModelMaker, it is possible to combine the alignment evidence with the results of gene prediction to construct novel transcripts.

Beginning with the standard Map Viewer display for the gene *FMR1*, we can display the Gscan and Genes_seq maps in parallel, as shown in Figure 5. The GScan map shows gene predictions made by the GenomeScan program. The Genes_seq track shows the exons implied by the composite alignment of transcripts, such as NCBI mRNA RefSeqs, to the genome. There are two *boxed regions* in the Map Viewer display. The *upper boxed* region shows that the GenomeScan model for *FMR1* begins at a point that is part of an intron in the transcript-based model. Furthermore, there is a separate GenomScan model upstream of the first that overlaps with the initial exon of the transcript-based model. It would be of interest to investigate whether the two GenomeScan models could be fused to produce a longer transcript. In the *lower boxed* region, we see that the transcript-based model includes an exon lacking in the GenomeScan model. Perhaps we can create a model transcript based on the fusion of the two GenomeScan models that also includes the extra exon seen in the transcript-based model.

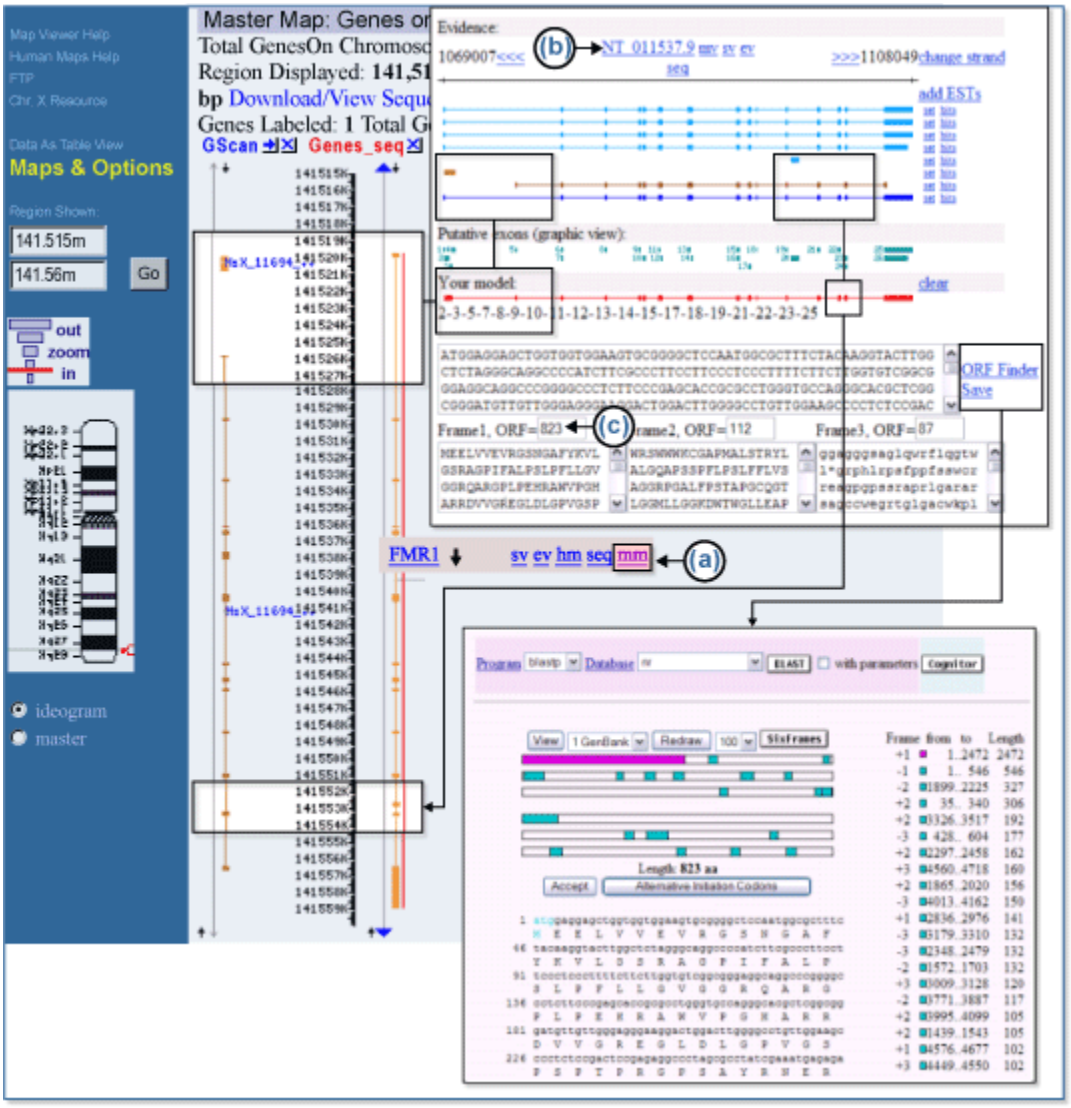


Figure 5. Use of ModelMaker to test alternative cDNA XSmodels based on GenomeScan predictions of mRNA alignments.

To attempt this synthesis, we first select the **mm** link [(*a*)] to the *right* of the *FMR1* gene link on the Genes_seq track to invoke the ModelMaker. A number of alignments between transcript or model sequences and the genomic contig upon which *FMR1* lies, NT_011537 [(*b*)], are given at the *top* of the ModelMaker display, and the implied exons resulting from these alignments are shown just *below*, numbered sequentially in the **Putative exons** pallet. We may choose any of these exons for inclusion in our model by selecting it. We can also choose a complete set of exons from an existing alignment by selecting the **set** link next to an alignment. Because we plan to begin with the GenomeScan model, we can select the **set** link next to the second alignment from the *bottom* to start. This gives us an initial model identical to that of the GenomeScan prediction and yields, as its longest Open Reading Frame (ORF), an ORF of 600 amino

acids. We can also see the second, small, GenomeScan-predicted model at the *far left* of the ModelMaker display. We hope to fuse this model with the larger model. The second model comprises three closely spaced exons, resembling a single exon in the display, numbered 2–4 in the **Putative exons** pallet. Selecting exons 2 and 3 adds them to the model (*first boxed pair* in the ModelMaker display). At this point, the longest ORF detected in the transcript has increased to 762 amino acids. If we try to include exon 4, however, we drop back to 600 amino acids because of the introduction of an internal stop codon; therefore, we can remove exon 4 from our model with another click. Finally, we can select exon 22, which is the exon from the alignment-based model that we want to include, and notice that the longest ORF detected has risen to 823 amino acids [(*c*)]. Because long ORFs without stop codons occur rarely by chance, this transcript model is promising. To explore further, we can select the **ORF Finder** link to generate a graphical view of all ORFs found in the transcript, including the longest ORF, and subject the translation of the latter to a BLAST search. In this case, we find that the majority of the predicted protein matches the *FMR1* gene product but that we have introduced some novel peptide sequence at the amino-terminal end as well as some near the carboxy terminus.

In this example, there were multiple, putative, full-length mRNAs. Please note that ESTs can be added to the display by selecting **add ESTs** (Figure 5, upper right-hand corner). Additional exons and splicing patterns may then be available to be considered in your model. This feature may be of particular importance if most of the evidence for splicing and exons comes from ESTs rather than complete mRNAs.

6. Using the Mouse Map Viewer

This assembly can be displayed by using Map Viewer for the mouse. In this particular case, the official symbol for the human and mouse genes is the same; therefore, a query by symbol returns the expected result. If this were not the case, however, it is also possible to search the mouse genome by the human sequence using BLAST the Mouse Genome. Simulating this, you can see that the best match for the human sequence is for a gene labeled LOC207836. If you check this in LocusLink, you will see that this is now annotated as Fmr1.

By Human/Mouse Homology Map

Consider an example in which we want to know whether there is a human/mouse synteny region that includes our favorite gene, *FMR1*. To begin, enter FMR1 into the Map Viewer search box and press the **Find** button. Selecting the gene name in the results table leads us to a Map Viewer display of the *FMR1* gene with the Genes_seq map, UniG_Hs map, and Genes_cyto maps shown. In Figure 6, the Genes_cyto map has been replaced with the UniG_Mm map. Selecting the **hm** link (*boxed*) to the *right* of the Genes_seq map leads to the human/mouse homology maps. One can choose between two slightly different variants of human genome assembly (NCBI and UCSC). Three sources of mouse mapping data are available: the Mouse Genome Database (MGD) map, Jackson Lab map, and the

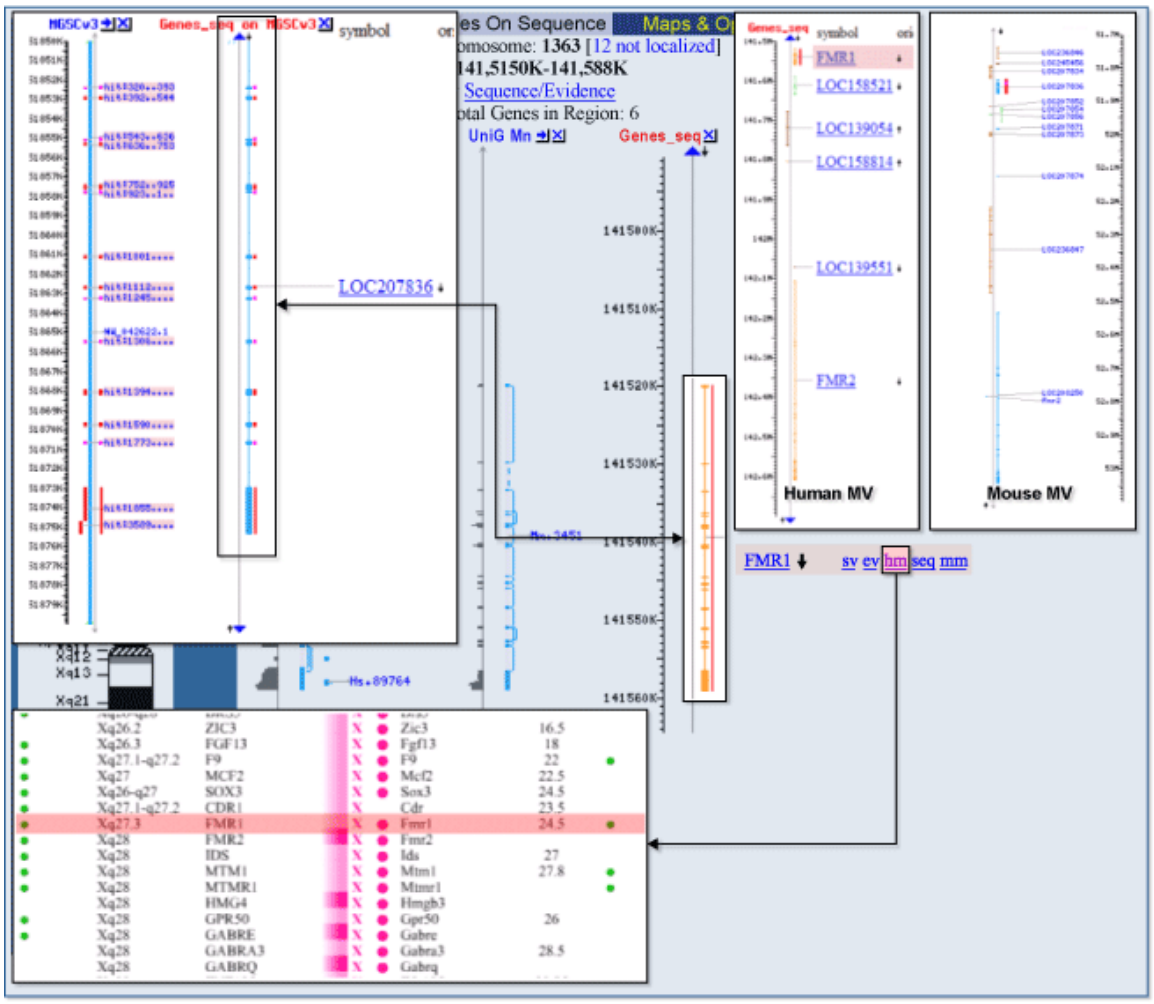


Figure 6. The human/mouse synteny region.

Whitehead/MRC Radiation Hybrid map. Either of the two human maps may be compared with either of the two mouse maps.

Let us choose the NCBI *versus* MGD variant and then check for synteny in the region of the *FMR1* homologs in the two species. Because our primary interest is the human gene *FMR1*, we select NCBI *versus* MGD_Hs, and the comparative map will appear. The row corresponding to the *FMR1* gene is highlighted (*inset*). The mouse homolog, called *Fmr1*, is also positioned on chromosome X, and the order of genes surrounding it is conserved in both genomes. Available STS data are indicated by a *green dot* that is a hyperlink to UniSTS. Links to the Map Viewer (select **Cytogenetic map position**) and LocusLink (select **Gene Symbol**) are available for each pair of genes. The user can examine the pairwise BLAST alignment that is accessible by selecting a chromosome color-coded dot preceding the mouse gene name. A dot-matrix similarity plot on the pairwise alignment page makes it easy to visually estimate the differences in sequences of the two genes. It is

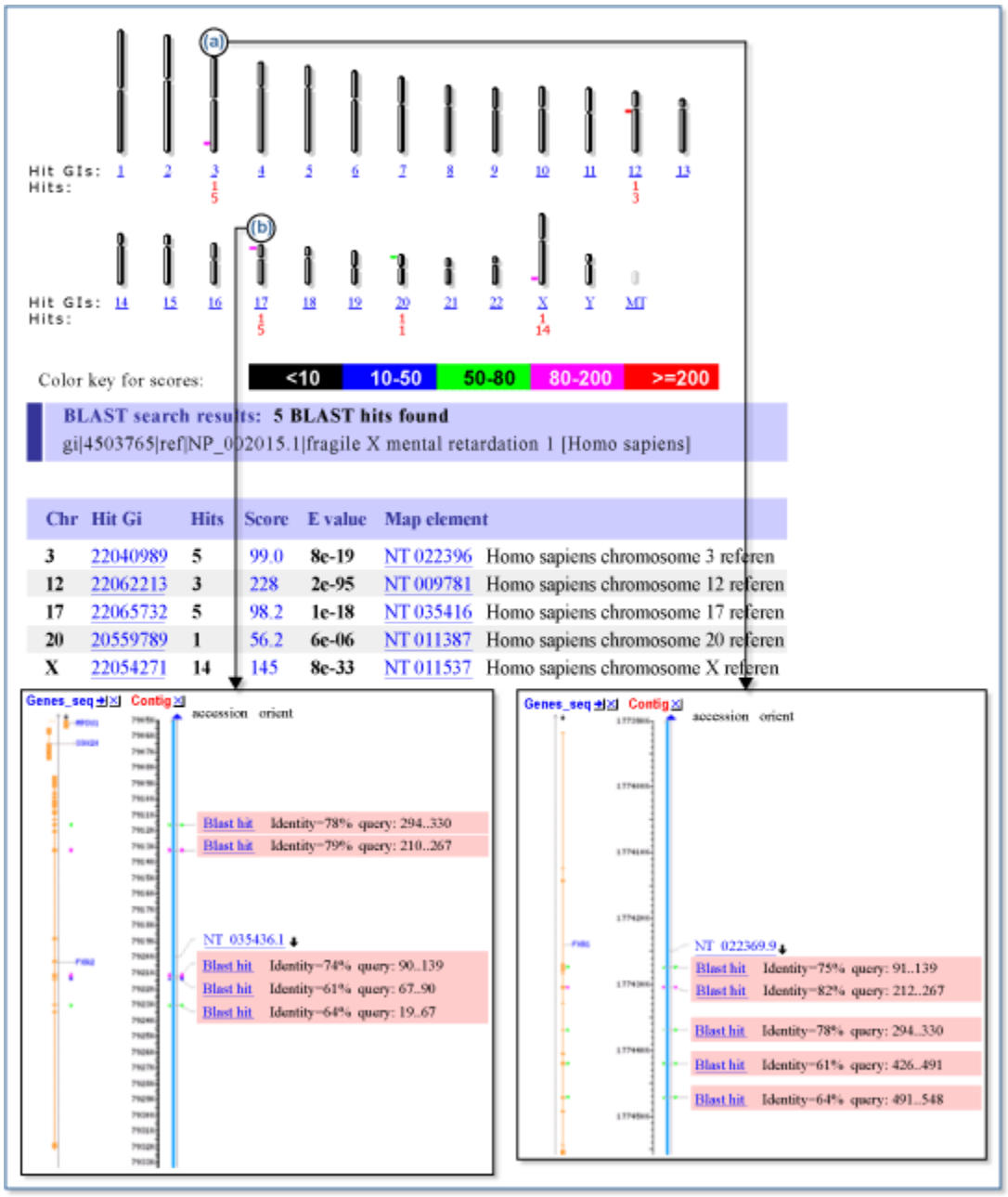


Figure 7. Review of results of a tblastn query against the mouse genome using a human protein sequence. The summary of significant BLAST hits is shown in the *top* graphic. Sections (*a*) and (*b*) show expanded views of hits to the related genes *FXR1* and *FXR2* on chromosomes 3 and 17, respectively.

interesting to see that the similarity between *FMR1* and *Fmr1* actually extends past the coding regions.

Using the Mouse UniGene Map

Figure 6 shows a Map Viewer display of human *FMR1* using the Genes_seq map, the UniG_Mm map, and the UniG_Hs map. It is apparent that the EST mapping patterns on the two UniGene maps are similar for *FMR1*. This indicates a similarity at the expressed sequence level but does not indicate a similar gene structure because both sets of ESTs are being mapped to the same human genomic sequence. To investigate similarities in gene structure, it is necessary to use the mouse version of the Map Viewer.

Using the Mouse Map Viewer

The most complete assembly of the mouse genome available at NCBI is the Mouse Genome Consortium Version 3 WGS assembly. This assembly can be displayed using the mouse version of the Map Viewer; however, the mouse homolog of *FMR1* is not labeled as such in this assembly; therefore, we cannot find it using a simple search by gene name. We can overcome this obstacle by searching the mouse genome with the human mRNA sequence using mouse genome BLAST. The result of such a search is shown in the *leftmost inset* in Figure 6 and indicates that a gene model labeled **LOC207836** is the probable homolog. The *double-arrow* links the human and mouse Map Viewer displays of the corresponding genes in the two species; the structures are not identical, but they do share a rather large 3′ exon. If we zoom out a bit in the two displays (*two right insets*) to see the surrounding genes, we can observe that the organization of the two genomes is similar in the region of *FMR1* to the extent that a second gene called *FMR2* lies downstream of *FMR1*, whereas the mouse version, *Fmr2*, likewise lies downstream of the mouse *Fmr1* homolog.

7. How Can I Find Members of a Gene Family Using the Map Viewer?

Finding members of a gene family is not straightforward by any means. However, the Map Viewer can be used to flag sets of genes that are related, either by nomenclature or by sequence similarity.

By Common Annotation

Consider the gene *FMR1*. Let us assume we do not know much about genes from this family, but we suppose (recognizing that we may have cause to regret this supposition) that they all share the common root name FMR. We start our search from the main Map Viewer page by entering FMR* (the asterisk is a wild-card symbol) into the **Search** box and pressing the **Find** button. This search results in several hits, and some obviously do not belong to the *FMR1* family (cytoplasmic FMR1 interacting proteins 1 and 2, FMRFAL); but two genes on chromosome X (*FMR2* and *FMR3*) and one on chromosome 17 (*FMR1L2*) look promising. By selecting the chromosome X **all matches** link, we go to the graphical representation of the genomic region containing three FMR genes.

Selecting the gene name (the rows for the FMR* query hits are highlighted) invokes a corresponding LocusLink page that serves as a portal to available information for the gene, including the precomputed results of a similarity search against the nr database. Go to the NCBI Reference Sequences section of the LocusLink page and select the BLAST Link (**BL**) link. BLink displays a schematic representation of BLAST alignments with links to displays of the best hit from each organism, protein domains found in the query sequence, or sequences similar to the query that have known 3D structures.

When we look at the BLAST summary for *FMR1*, we find neither *FMR2* nor *FMR3*. We did not expect *FMR3* to show up because it had not been mapped on the Genes_seq map, which suggested that its sequence is not yet known. However, why do we not see *FMR2*? If we use LocusLink to retrieve the Reference Sequences for the *FMR1* and *FMR2* gene products (NP_002015 and NP_002016) and perform a pairwise BLAST comparison, we find that there is no significant similarity between the two sequences. Apparently, the names of "FMR" genes do not reflect common sequence features but rather a physiological condition, "fragile X mental retardation" syndrome, associated with this gene. In this sense, the two are members of a group or family, but they show sequence similarity only in the pathological trinucleotide repeats $(CGG)_n$ that are often found upstream of their coding regions.

By Precomputed Sequence Similarity

The BLink page lists two annotated human homologs of the *FMR1* gene, called "fragile X mental retardation", autosomal homologs 1 and 2 (FXR1 and FXR2). They show a significant similarity to the FMR1 protein and possess the same functional domains (K-homology RNA-binding domains documented in the LocusLink reports). Selecting the **Score column** link invokes a page with the results of pairwise BLAST, as well as a visual representation of similarity between the two proteins (a variant of the dot matrix similarity plot).

By Sequence Similarity to a Query

To see whether there are undocumented homologs of *FMR1* in the genome, return to the Map Viewer maps page for the *FMR1* gene and select the **BLAST the Human Genome** link. Enter the Accession number for the FMR1 protein taken from the LocusLink report (NP_002015) into the **Search** window, select **Genome** as the database, **tblastn** as the BLAST program, and search. A tblastn search takes a protein sequence as a query and translates a nucleotide database in all reading frames to find any coding regions, documented or undocumented, that might code for a protein similar to the query. Such a search is very sensitive because it is tolerant of differences in codon usage as well as of insertions and deletions.

The results of such a search are shown in Figure 7. BLAST hits to regions on four chromosomes are shown in the genomic overview, indicated by *small tick marks*. There are 14 hits to chromosome X, clustered near the end of the q-arm. These hits are to *FMR1*, which is located in band Xq27.3. There are also 5 hits apiece to chromosomes 3 and 17

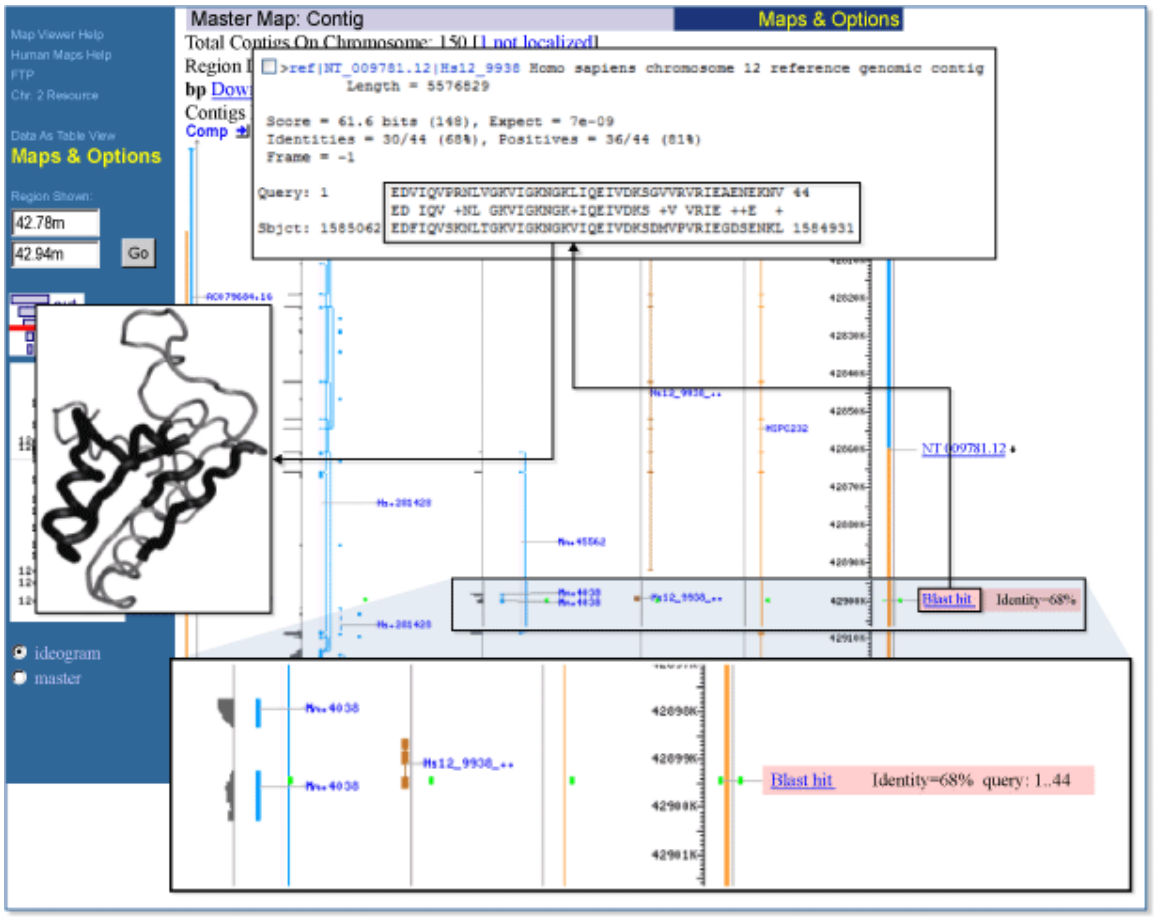


Figure 8. **BLAST hits.**

[(*a*) and (*b*), respectively]. These hits are to known homologs of *FMR1, FXR1,* and *FXR2.* Selecting the links below the chromosome graphic or on the appropriate contig link in the table below leads to a graphical display of the hits on the contig, as shown in the two *insets*. Note that the BLAST hits track the exons of the genes. The hit on chromosome 12 is to a hypothetical protein and may be worth further investigation.

8. How Can I Find Genes Encoding a Protein Domain Using the Map Viewer?

The Map Viewer displays a graphic representation of genomic information with links to related resources that allow it to serve as a springboard for many types of analyses.

Finding Protein Domains in a Gene Product

For example, one might be interested in the domain structure of the gene product. Let us use the human *FMR1* gene as an example. On the main Map Viewer page, enter FMR1 into the **Search for** window and press the **Find** button. Select the **FMR1** link to display

three maps (cyto, UniG_Hs, and Genes_seq) of the *FMR1* locus on chromosome X. Selecting the gene name next to the gene model (*FMR1*) leads to the LocusLink page, which is not only a compilation of genomic, genetic, and reference data related to the gene but also a gateway to the external resources and NCBI tools and databases. The LocusLink report for *FMR1* is linked to precomputed BLAST results for the *FMR1* gene product. Selecting **BL** (BLink, for BLAST link) invokes a page with a schematic view of the BLAST comparison of the FMR1 protein against the non-redundant (nr) database. The BLink page lists database hits to the FMR1 protein, sorted according to their BLAST similarity scores. Results may be reformatted by selecting **Sort by Taxonomy Proximity** to cluster hits from the same species.

To see the functional domains that have been identified in the FMR1 protein, select the Conserved Domains Database (CDD) **Search** button. The resulting page shows two hits to the KH-domain from the SMART database and one hit to the KH domain in the Pfam database. Notice that one of the SMART domain hits is a partial hit, as indicated by the *jagged edge* in the schematic representation.

Visualizing 3D structures

We can easily see whether there exists a three-dimensional structure that includes these conserved domains. The *pink dot* preceding the domain name in the BLink **Description of Alignments** section leads to a display of the corresponding three-dimensional structure using Cn3D, the NCBI macromolecular viewer available over the Web. The Pfam and SMART domains are linked to two 3D structures (1K1G_A and 1J4W_A). The CDD page also lists other sequences that have the same domain and shows their multiple sequence alignments.

Searching for Similar Domains in Genomic Sequence

Cut and paste the sequence of the KH-domain from the FMR1 protein and run a genome-specific BLAST search. We will use the tblastn program to compare the 44-amino acid sequence of the KH domain to the nucleotide sequence of the human genome. The results will show us other regions of the genome with the potential to code for this domain. Of course, we already know from the BLink page that there are some autosomal homologs, but we do not know whether they actually contain the KH RNA-binding domain. The obvious caveat is that some pseudogenes might contain the domain as well. Our tblastn search returns 4 hits, one being the *FMR1* gene itself on the X chromosome, and three others on chromosomes 3, 12, and 17. Select the **Genome View** button in the Genome BLAST results to see the positions of the hits on the chromosomes. One may then select the names of sequences producing significant alignments to invoke a Map Viewer display that shows the corresponding maps and report the names of the loci. As expected, hits to chromosomes 3 and 17 correspond to the autosomal homologs FXR1 and FXR2. The hit to chromosome 12 corresponds to the hypothetical protein HSPC232, and the Map Viewer display for this BLAST hit is shown (Figure 8). The hit is to a segment of intronic sequence, rather than to an exon, and is without supporting human EST alignments.

There are, however, mouse ESTs mapping to this region (*lower inset*), and there is also a GenomeScan model that covers the hit; therefore, the BLAST hit may indeed represent coding sequence. Selecting the **Blast hit** link (*boxed*) leads to an alignment (*upper inset*) that indicates a good match between our 44-amino acid domain sequence and a protein translation of the genomic sequence. If we map this 44-amino acid sequence onto the structure of 1K1G using Cn3D, we see that it covers a module consisting of two alpha helices and a three-stranded beta sheet (*leftmost inset*). It appears reasonable that our domain hit may represent an exon of the *HSPC232* gene. To follow this line of analysis, the next step might be to produce a transcript model incorporating this new exon using the Model Maker (see the Model Maker exercise in this series).

Glossary

3-D or 3D — Three-dimensional.

Accession number — An Accession number is a unique identifier given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, DDBJ). The initial deposition of a sequence record is referred to as version 1. If the sequence is updated, the version number is incremented, but the Accession number will remain constant.

Alu — The *Alu* repeat family comprises short interspersed elements (SINES) present in multiple copies in the genomes of humans and other primates. The *Alu* sequence is approximately 300 bp in length and is found commonly in introns, 3′ untranslated regions of genes, and intergenic genomic regions. They are mobile elements and are present in the human genome in extremely high copy number. Almost 1 million copies of the *Alu* sequence are estimated to be present, making it the most abundant mobile element. The *Alu* sequence is so named because of the presence of a recognition site for the *Alu*I endonuclease in the middle of the *Alu* sequence. Because of the widespread occurrence of the *Alu* repeat in the genome, the *Alu* sequence is used as a universal primer for PCR in animal cell lines; it binds in both forward and reverse directions. The *Alu* universal primer sequence is as follows: 5′-GTG GAT CAC CTG AGG TCA GGA GTT TC-3′ (26-mer).

allele — One of the variant forms of a gene at a particular locus on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant one) may be expressed more than another form (the recessive one). When "genes" are considered simply as segments of a nucleotide sequence, allele refers to each of the possible alternative nucleotides at a specific position in the sequence. For example, a CT polymorphism such as CCT[C/T]CCAT would have two alleles: C and T.

API — Application Programming Interface. An API is a set of routines that an application uses to request and carry out lower-level services performed by a computer's operating system. For computers running a graphical user interface, an API manages an application's windows, icons, menus, and dialog boxes.

ASN.1 — Abstract Syntax Notation 1 is an international standard data-representation format used to achieve interoperability between computer platforms. It allows for the reliable exchange of data in terms of structure and content by computer and software systems of all types.

BAC — Bacterial Artificial Chromosome. A BAC is a large segment of DNA (100,000–200,000 bp) from another species cloned into bacteria. Once the foreign DNA has been cloned into the host bacteria, many copies of it can be made.

BankIt — BankIt is a tool for the online submission of one or a few sequences into GenBank and is designed to make the submission process quick and easy. (BankIt also

automatically uses VecScreen to identify segments of nucleic acid sequence that may be of vector, adapter, or linker origin to combat the problem of vector contamination in GenBank.)

bit score — The value S′ is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. By normalizing a raw score using the formula: $S' = \frac{\lambda S - \ln K}{\ln 2}$ a "bit score" S′ is attained, which has a standard set of units, and where K and *lambda* are the statistical parameters of the scoring system. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

BLAST — Basic Local Alignment Search Tool (Altschul et al., J Mol Biol 215:403-410; 1990). A sequence comparison algorithm that is optimized for speed and used to search sequence databases for optimal local alignments to a query. See the BLAST chapter (Chapter 15) or the tutorial or the narrative guide to BLAST.

blastn — nucleotide–nucleotide BLAST. blastn takes nucleotide sequences in FASTA format, GenBank Accession numbers, or GI numbers and compares them against the NCBI Nucleotide databases.

blastp — protein–protein BLAST. blastp takes protein sequences in FASTA format, GenBank Accession numbers, or GI numbers and compares them against the NCBI Protein databases.

BLAT — A DNA/Protein sequence analysis program to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. BLAT is not BLAST. (See the BLAT web page.)

BLink — BLAST Link. BLink displays the results of BLAST searches that have been done for every protein sequence in the Entrez Protein data domain. It can be accessed by following the BLink link displayed beside any hit in the results of an Entrez Protein search. In contrast to Entrez's **Related Sequences** feature, which lists the titles of similar sequences, BLink displays the graphical output of precomputed blastp results against the non-redundant (nr) protein database. The output includes the positions of up to 200 BLAST hits on the query sequence, scores, and alignments. BLink offers a variety of display options, including the distribution of hits by taxonomic grouping, the best hit to each organism, the protein domains in the query sequence, similar sequences that have known 3D structures, and more. Additional options allow you to specify from which taxa you would like to exclude, increase, or decrease the BLAST cutoff score or filter the BLAST hits to show only those from a specific source database, such as RefSeq or SWISS-PROT. See the BLink help document for additional information.

BLOB — Binary Large Object (or binary data object). BLOB refers to a large piece of data, such as a bitmap. A BLOB is characterized by large field values, an unpredictable table size, and data that are formless from the perspective of a program. It is also a keyword designating the BLOB structure, which contains information about a block of data.

BLOSUM 62 — Blocks Substitution Matrix. A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM 62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. Sequences more identical than 62% are represented by a single sequence in the alignment to avoid overweighting closely related family members (Henikoff and Henikoff, Proc Natl Acad Sci U S A 89:10915-10919; 1992).

Boolean — This term refers to binary algebra that uses the logical operators AND, OR, XOR, and NOT; the outcomes consist of logical values (either TRUE or FALSE). The keyword boolean indicates that the expression or constant expression associated with the identifier takes the value TRUE or FALSE. The logical-AND (&&) operator produces the value 1 if both operands have nonzero values; otherwise, it produces the value 0. The logical-OR (||) operator produces the value 1 if either of its operands has a nonzero value. The logical-NOT (!) operator produces the value 0 if its operand is true (nonzero) and the value 1 if its operand is FALSE (0). The exclusive OR (XOR) operator yields TRUE only if one of its operands are TRUE and the other is FALSE. If both operands are the same (either TRUE or FALSE), the operation yields FALSE.

build — A run of the genome assembly and annotation process of the set of products generated by that run.

CCAP — Cancer Chromosome Aberration Project. CCAP was designed to expedite the definition and detailed characterization of the distinct chromosomal alterations that are associated with malignant transformation. The project is a collaboration among the NCI, the NCBI, and numerous research labs.

CD — Conserved Domain. CD refers to a domain (a distinct functional and/or structural unit of a protein) that has been conserved during evolution. During evolution, changes at specific positions of an amino acid sequence in the protein have occurred in a way that preserve the physico-chemical properties of the original residues, and hence the structural and/or functional properties of that region of the protein.

CDART — Conserved Domain Architecture Retrieval Tool. When given a protein query sequence, CDART displays the functional domains that make up the protein and lists proteins with similar domain architectures. The functional domains for a sequence are found by comparing the protein sequence to a database of conserved domain alignments, CDD using RPS-BLAST.

CDD — Conserved Domain Database. This database is a collection of sequence alignments and profiles representing protein domains conserved during molecular evolution.

cDNA — complementary DNA. A DNA sequence obtained by reverse transcription of a messenger RNA (mRNA) sequence.

CDS — coding region, coding sequence. CDS refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon, inclusively, if complete. A partial CDS lacks part of the complete CDS (it may lack either or both the start and stop codons). Successful translation of a CDS results in the synthesis of a protein.

CEPH — Centre d'Etude du Polymorphism Humain

CGAP — Cancer Genome Anatomy Project. CGAP is an interdisciplinary program to identify the human genes expressed in different cancerous states, based on cDNA (EST) libraries, and to determine the molecular profiles of normal, precancerous, and malignant cells. The project is a collaboration among the NCI, the NCBI, and numerous research labs.

CGH — Comparative Genomic Hybidization. CGH is a fluorescent molecular cytogenetic technique that identifies chromosomal aberrations and maps these changes to metaphase chromosomes. CGH can be used to generate a map of DNA copy number changes in tumor genomes. CGH is based on quantitative two-color fluorescence *in situ* hybridization (FISH). DNA extracted from tumor cells is labeled in one color (e.g., green) and mixed in a 1:1 ratio with DNA from normal cells, which is labeled in a different color (e.g., red). The mixture is then applied to normal metaphase chromosomes. Portions of the genome that are equally represented in normal and tumor cells will appear orange, regions that are deleted in the tumor sample relative to the normal sample will appear red, and regions that are present in higher copy number in the tumor sample (because of amplification) will appear green. Special image analysis tools are necessary to quantitate the ratio of green-to-red fluorescence to determine whether a given region is more highly represented in the normal or in the tumor sample.

CGI — Common Gateway Interface. A mechanism that allows a Web server to run a program or script on the server and send the output to a Web browser.

cluster — A group that is created based on certain criteria. For example, a gene cluster may include a set of genes whose similar expression profiles are found to be similar according to certain criteria, or a cluster may refer to a group of clones that are related to each other by homology.

Cn3D — "See in 3-D" is a structure and sequence alignment viewer for NCBI databases. It allows viewing of 3-D structures and sequence–structure or structure–structure alignments. Cn3D can work as a helper application to the browser or as a client–server application that retrieves structure records from the Molecular Modeling Database (MMDB, see below) directly from the internet. The Cn3D homepage provides access to information on how to install the program, a tutorial to get started, and a comprehensive help document.

codon — Sequence of three nucleotides in DNA or mRNA that specifies a particular amino acid during protein synthesis; also called a triplet. Of the 64 possible codons, 3 are stop codons, which do not specify amino acids.

COGs — Clusters of Orthologous Groups (of proteins) were delineated by comparing protein sequences from completely sequenced genomes. Each COG consists of individual proteins or groups of paralogs from at least three lineages and thus corresponds to an ancient conserved domain.

consensus sequence — The nucleotides or amino acids found most commonly at each position in the sequences of homologous DNAs, RNAs, or proteins.

contig — A contiguous segment of the genome made by joining overlapping clones or sequences. A clone contig consists of a group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome. A sequence contig is an extended sequence created by merging primary sequences that overlap. A contig map shows the regions of a chromosome where contiguous DNA segments overlap. Contig maps provide the ability to study a complete and often large segment of the genome by examining a series of overlapping clones, which then provide an unbroken succession of information about that region.

Coriell — Coriell Institute of Aging Cell Repository

CPU — Central Processing Unit. The CPU is the computational and control unit of a computer, the device that interprets and executes instructions.

CSS — Cascading Style Sheets. CSS specify the formatting details that control the presentation and layout of HTML and XML elements. CSS can be used for describing the formatting behavior and text decoration of simply structured XML documents but cannot display structure that varies from the structure of the source data.

Cubby — A tool of Entrez, the Cubby stores search strategies that may be updated at any time, stores LinkOut preferences to specify which LinkOut providers have to be displayed in PubMed, and changes the default document delivery service.

DCMS — Data Creation and Maintenance System

DDBJ — DNA Data Bank of Japan

definition line — A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The definition line or description line is distinguished from the sequence data by a "greater than" (>) symbol in the first column (see example); also DEFLINE, as in a flatfile.

DNA — Deoxyribonucleic acid is the chemical inside the nucleus of a cell that carries the genetic instructions for making living organisms. DNA is composed of two anti-parallel strands, each a linear polymer of nucleotides. Each nucleotide has a phosphate group linked by a phosphoester bond to a pentose (a five-carbon sugar molecule, deoxyribose), that in turn is linked to one of four organic bases, adenine, guanine, cytosine, or thymine, abbreviated A, G, C, and T, respectively. The bases are of two types: purines, which have two rings and are slightly larger (A and G); and pyrimidines, which have only one ring (C and T). Each nucleotide is joined to the next nucleotide in the chain by a covalent

phosphodiester bond between the 5′ carbon of one deoxyribose group and the 3′ carbon of the next. DNA is a helical molecule with the sugar–phosphate backbone on the outside and the nucleotides extending toward the central axis. There is specific base-pairing between the bases on opposite strands in such a way that A always pairs with T and G always pairs with C.

domain — A "domain" refers to a discrete portion of a protein assumed to fold independently of the rest of the protein and which possesses its own function.

draft sequence — Draft sequence refers to DNA sequence that is not yet finished but is generally of high quality (i.e., an accuracy of greater than 90%). Draft sequence data are mostly in the form of 10,000 base pair-sized fragments, the approximate chromosomal locations of which are known. The following keywords are associated with draft sequence: phase 0, light-pass coverage of a clone, generally only 1× coverage; phase 1, 4–10× coverage of a BAC clone (order and orientation of the fragments are unknown); and phase 2, 4–10× coverage of a BAC clone (order and orientation of the fragments are known). Phase 3 refers to the completely finished sequence.

DTD — Document Type Definition. The DTD is an optional part of the prolog of an XML document that defines the rules of the document. It sets constraints for an XML document by specifying which elements are present in the document and the relationships between elements, e.g., which tags can contain other tags, the number and sequence of the tags, and attributes of the tags. The DTD helps to validate the data when the receiving application does not have a built-in description of the incoming data.

DUST — A program for filtering low-complexity regions from nucleic acid sequences.

E-value — Expect value. The E-value is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially with the score (S) that is assigned to a match between two sequences. Essentially, the E-value describes the random background noise that exists for matches between sequences. For example, an E-value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size, one might expect to see one match with a similar score simply by chance. This means that the lower the E-value, or the closer it is to "0", the higher is the "significance" of the match. However, it is important to note that searches with short sequences can be virtually identical and have relatively high E-value. This is because the calculation of the E-value also takes into account the length of the query sequence. This is because shorter sequences have a high probability of occurring in the database purely by chance. For more information, see the following tutorial.

EC number — A number assigned to a type of enzyme according to a scheme of standardized enzyme nomenclature developed by the Enzyme Commission of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). EC numbers may be found in ENZYME, the Enzyme nomenclature database, maintained at the ExPASy molecular biology server.

EMBL — European Molecular Biology Laboratory

Entrez — Entrez is a retrieval system for searching several linked databases. It provides access to the following NCBI databases: PubMed, GenBank, Protein, Structure, Genome, PopSet, OMIM, Taxonomy, Books, ProbeSet, 3D Domains, UniSTS, SNP, and CDD. (See the Entrez chapter or the Entrez web page.)

Entrez Gene — (formerly known as LocusLink). Entrez Gene provides tracked, unique identifiers for genes (GeneIDs) and reports information associated with those identifiers for unrestricted public use. See the Entrez Gene chapter or web page.)

EST — Expressed Sequence Tag. ESTs are short (usually approximately 300–500 base pairs), single-pass sequence reads from cDNA. Typically, they are produced in large batches. They represent the genes expressed in a given tissue and/or at a given developmental stage. They are tags (some coding, others not) of expression for a given cDNA library. They are useful in identifying full-length genes and in mapping.

e-PCR — Electronic PCR is used to compare a query sequence to mapped sequence-tagged sites (STSs) to find a possible map location for the query sequence. e-PCR finds STSs in DNA sequences by searching for subsequences that closely match the PCR primers present in mapped markers. The subsequences must have the correct order, orientation, and spacing that they could plausibly prime the amplification of a PCR product of the correct molecular weight.

epub citation — "Ahead-of-print" citation. PubMed now accepts citations from publishers for articles that have been published electronically ahead of the printed issue. PubMed displays the category "[epub ahead of print]" in the part of the citation where the volume and pagination would ordinarily display. For example: Proc Natl Acad Sci U S A. 2000 May 2 [epub ahead of print].

ExoFish — Exon Finding by Sequence Homology. Exofish is a tool based on homology searches for the rapid and reliable identification of human genes. It relies on the sequence of another vertebrate, the pufferfish *Tetraodon nigroviridis* (similar to Fugu), to detect conserved sequences with a very low background. The genome of *T. nigroviridis* is eight times more compact than the human genome and has been used in the comparative identification of human genes from the rough draft of the human genome (Roest Crollius et al., Nat Genet 25:235-238; 2000).

exon — Refers to the portion of a gene that encodes for a part of that gene's mRNA. A gene may comprise many exons, some of which may include only protein-coding sequence; however, an exon may also include 5' or 3' untranslated sequence. Each exon codes for a specific portion of the complete protein. In some species (including humans), a gene's exons are separated by long regions of DNA (called introns or sometimes "junk DNA") that often have no apparent function but have been shown to encode small untranslated RNAs or regulatory information. (See also splice sites.)

exon-trapped — Exon trapping is a technique for cloning exon sequences from genomic DNA by selecting for functional splice sites, relying on the cellular splicing machinery. The genomic DNA containing the putative exon(s) is cloned into an exon-trap vector,

which has a promoter, polyadenylation signals, and splice sites, and then transfected into a cell line. If there are functional splice sites in the genomic DNA fragment, the segments of DNA between the splice sites will be removed. Total RNA is isolated and reverse-transcribed. After cDNA synthesis and PCR amplification, the exon of interest is cloned.

ExPASy — Expert Protein Analysis System is a proteomics server of the Swiss Bioinformatics Institute (SIB).

FASTA — The first widely used algorithm for similarity searching of protein and DNA sequence databases. The program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later, the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable, which specifies the size of a "word" (Pearson and Lipman). Also refers to a format for a nucleic acid or protein sequence.

fingerprint — The pattern of bands on a gel produced by a clone when restricted by a particular enzyme, such as *Hin*dIII.

finished sequence — High-quality, low-error DNA sequence that is free of gaps. To qualify as a finished sequence, only a single error out of every 10,000 bases (i.e., an accuracy of 99.999%) is allowed.

FISH — Fluorescence *in situ* hybridization. In this technique, fluorescent molecules are used to label a DNA probe, which can then hybridize to a specific DNA sequence in a chromosome spread so that the site becomes visible through a microscope. FISH has been used to highlight the locations of genes, subchromosome regions, entire chromosomes, or specific DNA sequences. It has been used for mapping and the detection of genomic rearrangements, as well as studies on DNA replication.

flatfile or flat file — A flat file is a data file that contains records (each corresponding to a row in a table); however, these records have no structured relationships. To interpret these files, the format properties of the file should be known. For example, a database management system may allow the user to export data to a comma-delimited file. Such a file is called a flat file because it has no inherent information about the data, and interpretation requires additional information. Files in a database management system have more complex storage structures.

freeze — To copy changing data so as to preserve the dataset as it existed at a particular point in time. Also used to refer to the resulting set of frozen data.

FTP — File Transfer Protocol. A method of retrieving files over a network directly to the user's computer or to his/her home directory using a set of protocols that govern how the data are to be transported.

gap — A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many

gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment. (See the figure for more information.)

GB — gigabytes

GBFF — GenBank Flat File. Refers to a format .gbff.

GenBank — GenBank is a database of nucleotide sequences from more than 100,000 organisms. Records that are annotated with coding region features also include amino acid translations. GenBank belongs to an international collaboration of sequence databases that also includes EMBL and DDBJ. [See the GenBank chapter (Chapter 1) or the GenBank web page.]

GeneID — GeneID is a unique identifier that is assigned to a gene record in Entrez Gene. It is an integer and is species specific. In other words, the integer assigned to dystrophin in human is different from that in any other species. For genomes that had been represented in LocusLink, the GeneID is the same as the LocusID. The GeneID is reported in RefSeq records as a 'db_xref' (e.g. /db_xref="GeneID:856646", in GenBank format).

genetic code — The instructions in a gene that tell the cell how to make a specific protein. A, T, G, and C are the "letters" of the DNA code; they stand for the chemicals adenine, thymine, guanine, and cytosine, respectively, that make up the nucleotide bases of DNA. Each gene's code combines the four chemicals in various ways to spell out three-letter "words" that specify which amino acid is needed at every position for making a protein.

GenomeScan — A gene identification algorithm that is used to identify exon–intron structures in genomic DNA sequence.

genotype — The genetic identity of an individual that does not show as outward characteristics. The genotype refers to the pair of alleles for a given region of the genome that an individual carries.

GEO — Gene Expression Omnibus. GEO is a gene expression data repository and online resource for the retrieval of gene expression data from any organism or artificial source. Many types of gene expression data from platform types, such as spotted microarray, high-density oligonucleotide array, hybridization filter, and serial analysis of gene expression (SAGE) data, are accepted, accessioned, and archived as a public dataset. [See the GEO chapter (Chpater 6) or the GEO web page.]

GI — The GenInfo Identifier is a sequence identification number for a nucleotide sequence. If a nucleotide sequence changes in any way, a new GI number will be assigned. A separate GI number is also assigned to each protein translation within a nucleotide sequence record, and a new GI is assigned if the protein translation changes in any way. GI sequence identifiers run parallel to the new accession.version system of sequence identifiers (see the description of Version).

GSS — Genome Survey Sequences are analogous to ESTs except that the sequences are genomic in origin, rather than cDNA (mRNA). The GSS division of GenBank contains (but is not limited to) the following types of data: random "single-pass read" genome survey sequences, cosmid/BAC/YAC end sequences, exon-trapped genomic sequences, and *Alu*-PCR sequences.

heterozygosity — The probability that a diploid individual will have two different alleles at a particular genome locus. These individuals are defined as heterozygous, whereas individuals who have two identical alleles at the locus are defined as homozygous. The probability can be estimated by sampling a representative number of individuals from the population and dividing the number of heterozygotes by the total number sampled.

HIV — Human Immunodeficiency Virus. HIV-1 is a retrovirus that is recognized as the causative agent of AIDS (Acquired Immunodeficiency Syndrome).

HNPCC — Hereditary nonpolyposis colon cancer

homogeneously staining region — A region of the chromosome identified cytologically by DNA staining or the FISH technique because of the presence of multiple copies of a subchromosomal region resulting from amplification.

homologous — The term refers to similarity attributable to descent from a common ancestor. Homologous chromosomes are members of a pair of essentially identical chromosomes, each derived from one parent. They have the same or allelic genes with genetic loci arranged in the same order. Homologous chromosomes synapse during meiosis.

HTGS — High-Throughput Genomic Sequences. The source of HTGS are large-scale genome sequencing centers; unfinished sequences are in phases 0, 1, and 2, and finished sequences are in phase 3.

HTGS_CANCELLED — A keyword added to GenBank entries by sequencing centers to indicate that work has stopped on a clone and that the existing sequence will not be finished. Sequencing centers may stop work because the clone is redundant or for various other reasons.

HTGS_PHASE0, HTGS_PHASE1, HTGS_PHASE2, HTGS_PHASE3 — Keywords added to GenBank entries by sequencing centers to indicate the status (phase) of the sequence (see phase definitions described under draft sequence).

HTML — Hypertext Markup Language. HTML is derived from SGML. It is a text-based mark-up language and is used to primarily display information using a web browser and to link pieces of information via hyperlinks. The tags used in an HTML document provide information only on how the content is to be displayed but do not provide information about the content they encompass.

HUP — Hold Until Published. HUP refers to the category for data that is electronically submitted for when it should be released to the public.

ICBN — International Code of Botanical Nomenclature

ICD — International Classification of Diseases

ICD-O-3 — International Classification of Diseases for Oncology, 3rd edition

ICNB — International Code of Nomenclature of Bacteria

ICNCP — International Code of Nomenclature for Cultivated Plants

ICTV — International Committee on Taxonomy of Viruses

ICVCN — International Code of Virus Classification and Nomenclature

ICZN — International Code of Zoological Nomenclature

ideogram — A diagrammatic representation of the karyotype of an organism.

IMAGE Consortium — Integrated Molecular Analysis of Genomes and their Expression. A consortium of academic groups that share high-quality, arrayed cDNA libraries and place sequence, map, and expression data of the clones in these arrays into the public domain. With the use of this information, unique clones can be rearrayed to form a "master array", with the aim of ultimately having a representative cDNA from every gene in the genome under study. To date, human, mouse, rat, zebrafish, and *Xenopus laevis* genomes have been studied.

intron — Refers to that portion of the DNA sequence that is present in the primary transcript and that is removed by splicing during RNA processing and is not included in the mature, functional mRNA, rRNA, or tRNA. Also called an intervening sequence. (See also splice sites.)

ISAM — Indexed Sequential-Access Method. ISAM is a database access method. It allows data records in a database to be accessed either sequentially (in the order in which they were entered) or randomly (using an index). In the index, each record has a unique key that enables its rapid location. The key is the field used to reference the record.

ISCN — International System for Human Cytogenetic Nomenclature

ISO — International Organization for Standardization

ISSN — International Standard Serial Number. The ISSN is an eight-digit number that identifies periodical publications, including electronic serials.

karyotype — The particular chromosome complement of an individual or a related group of individuals, as defined by both the number and morphology of the chromosomes, usually in mitotic metaphase, and arranged by pairs according to the standard classification.

LANL — Los Alamos National Lab

LIMS — Laboratory Information Management Systems. LIMS comprise software that helps biological and chemical laboratories handle data generation, information management, and data archiving.

LinkOut — A registry service to create links from specific articles, journals, or biological data in Entrez to resources on external web sites. Third parties can provide a URL, resource name, brief description of their web sites, and specification of the NCBI data from which they would like to establish links. The specification can be written as a valid Boolean query to Entrez or as a list of identifiers for specific articles or sequences. Entrez PubMed users can then select which external links are visible in their searches through the NCBI Cubby service (see above). (See the LinkOut chapter or web page.)

locus — In a genomic contect, locus refers to position on a chromosome. It may, therefore, refer to a marker, a gene, or any other landmark that can be described.

MACAW — Multiple Alignment Construction and Analysis Workbench. MACAW is a program for locating, analyzing, and editing blocks of localized sequence similarity among multiple seqences and linking them into a composite multiple alignment.

Map Viewer — The Map Viewer is a software component of Entrez Genomes that provides special browsing capabilities for a subset of organisms. It allows one to view and search an organism's complete genome, display chromosome maps, and zoom into progressively greater levels of detail, down to the sequence data for a region of interest. If multiple maps are available for a chromosome, it displays them aligned to each other based on shared marker and gene names and, for the sequence maps, based on a common sequence coordinate system. The organisms currently represented in the Map Viewer are listed in the Entrez Map Viewer help document, which provides general information on how to use that tool. The number and types of available maps vary by organism and are described in the "data and search tips" file provided for each organism.

MB — megabytes

MEDLINE — MEDLINE is NLM's database of indexed journal citations and abstracts in the fields of biomedicine and healthcare. It encompasses nearly 4,500 journals published in the United States and more than 70 other countries. (For more information, see the Fact Sheet.)

MegaBLAST — MegaBLAST is a program for aligning sequences that differ slightly as a result of sequencing or other similar "errors". When larger word size is used, it is up to 10 times faster than more common sequence-similarity programs. MegaBLAST is also able to efficiently handle much longer DNA sequences than the blastn program of the traditional BLAST algorithm. It uses the GREEDY algorithm for a nucleotide sequence alignment search.

MeSH — Medical Subject Headings. MeSH refers to the controlled vocabulary of NLM used for indexing articles in PubMed. MeSH terminology provides a consistent way to

retrieve information that may use different terminology for the same concepts. (See the MeSH homepage.)

Metathesaurus — Metathesaurus is a National Cancer Institute browser containing different biomedical vocabularies, including the International Classification of Diseases for Oncology ICD-O-3.

mFASTA — Multi-FASTA format.

MGC — Mammalian Gene Collection. MGC is a project of the NIH to provide a complete set of full-length (open reading frame) sequences and cDNA clones of expressed genes for human and mouse.

MGD — Mouse Genome Database. MGD contains information on mouse genetic markers, molecular segments, phenotypes, comparative mapping data, experimental mapping data, and graphical displays for genetic, physical, and cytogenetic maps.

MGI — Mouse Genome Informatics. MGI houses a database that provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse.

microsatellite — Repetitive stretches of short sequences of DNA used as genetic markers to track inheritance in families (e.g., CC[TATATATA]CCCT). Also known as short tandem repeats (STRs).

MIM — Mendelian Inheritance in Man. First published in 1966, *Mendelian Inheritance in Man (MIM)* is a genetic knowledge base that serves clinical medicine and biomedical research, including the Human Genome Project.

minimal tiling path — An ordered list or map that defines the minimal set of overlapping clones needed to provide complete coverage of a chromosome or other extended segment of DNA (compare with tiling path).

MMDB — Molecular Modeling Database. MMDB is a database of three-dimensional biomolecular structures derived from X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.

MMDB-ID — Molecular Modeling Database Accession number.

mRNA — messenger RNA. mRNA describes the section of a genomic DNA sequence that is transcribed, and can include the 5' untranslated region (5'UTR), CDS, and 3' untranslated region (3'UTR). Successful translation of the CDS section of an mRNA results in the synthesis of a protein.

mutation — A permanent structural alteration in DNA. In most cases, DNA changes have either no effect or cause harm, but occasionally a mutation can improve an organism's chance of surviving, and the beneficial change is passed on to the organism's descendants. Typically, mutations are more rare than polymorphisms in population samples because natural selection recognizes their lower fitness and removes them from the population.

NCBI — National Center for Biotechnology Information

NCBI Toolkit — Contains supported software tools from the Information Engineering Branch (IEB) of the NCBI. The NCBI Toolkit describes the three components of the ToolBox: data model, data encoding, and programming libraries. Provides access to documentation for the DataModel, C Toolkit, C++ Toolkit, NCBI C Toolkit Source Browser, XML Demo Program, XML DTDs, and the FTP site.

NCI — National Cancer Institute

NEXUS — NEXUS refers to a file format designed to contain data for processing by computer programs. NEXUS files should end with .nxs or .nex for purposes of clarity (Maddison et al., Syst Biol 46:590-621; 1997).

NIH — National Institutes of Health

NLM — National Library of Medicine

NMR — Nuclear Magnetic Resonance. NMR is a spectroscopic technique used for the determination of protein structure.

nr-PDB — non-redundant Protein Data Bank

OMIM — Online Mendelian Inheritance in Man. OMIM is a directory of human genes and genetic disorders, with links to literature references, sequence records, maps, and related databases.

ortholog — Orthology describes genes in different species that derive from a single ancestral gene in the last common ancestor of the respective species.

orthology — Orthology describes genes in different species that derive from a common ancestor, i.e., they are direct evolutionary counterparts.

paralog — A paralog is one of a set of homologous genes that have diverged from each other as a consequence of gene duplication. For example, the mouse α-*globin* and β-*globin* genes are paralogs. The relationship between mouse α-*globin* and chick β-*globin* is also considered paralogous.

paralogy — Paralogy describes the relationship of homologous genes that arose by gene duplication.

PCR — Polymerase Chain Reaction. A technique for amplifying a specific DNA segment in a complex mixture. Also present in the DNA mixture are short oligonucleotide primers to the DNA segment of interest and reagents for DNA synthesis. PCR relies on the ability of DNA to separate into its two complementary strands at high temperature (a process called denaturation) and for the two strands to anneal at an optimal lower temperature (annealing). The annealing phase is followed by a DNA synthesis step at an optimal temperature for a heat-stable DNA polymerase. After multiple rounds of denaturation,

annealing, and DNA synthesis, the DNA sequence specified by the oligonucleotide primers is amplified.

PDB — Protein Data Bank. The PDB is a database for 3D macromolecular structure data.

Pfam — Pfam is a database housing a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains.

phenotype — The observable traits or characteristics of an organism, e.g., hair color, weight, or the presence or absence of a disease. Phenotypic traits are not necessarily genetic.

PHRAP — A computer program that assembles raw sequence into sequence contigs (see above) and assigns to each position in the sequence an associated "quality score", on the basis of the PHRED scores of the raw sequence reads. A PHRAP quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRAP quality score of 30 corresponds to 99.9% accuracy for a base in the assembled sequence.

PHRED — A computer program that analyses raw sequence to produce a "base call" with an associated "quality score" for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read.

phyletic pattern — Pattern of presence–absence of a cluster of orthologs (COG) in different species.

PHYLIP — PHYLogeny Inference Package. A package of programs for various computer platforms to infer phylogenies or evolutionary trees, freely available from the Web.

PIR — Protein Information Resource

PMC — PubMed Central. NLM's digital archive of life sciences journal literature.

PMID — PubMed ID number

PNG — Portable Network Graphics. An extensible file format for the lossless, well-compressed storage of raster images (images that are composed of horizontal lines of pixels, such as those created by a computer screen). Compression of image, media, and application files is necessary to reduce the transmission time across the web. The technique of lossless compression reduces the size of the file without sacrificing any original data, and the image after expansion is exactly as it was before compression. PNG overcomes the patent issues of GIF (Graphic Interchange Format) and can replace many common uses of TIFF (Tagged Image File Format). Several features such as indexed color, grayscale, and truecolor are supported, as well as an optional alpha-channel. PNG is designed to work well in online viewing applications and is supported as an image standard by the WWW.

poly A — A string of adenylic acid residues that are added to the 3′ end of the primary mRNA transcript. Poly(A) polymerase is the enzyme that adds the poly A tail, which is between 100 and 250 bases long.

polymorphism — A common variation in the sequence of DNA among individuals. Genetic variations occurring in more than 1% of the population would be considered useful polymorphisms for genetic linkage analysis.

polypeptide — Linear polymer of amino acids connected by peptide bonds. Proteins are large polypeptides, and the two terms are commonly used interchangeably.

PRF — Protein Research Foundation

private polymorphism — Variations that are only common in specific populations. Usually such populations are reproductively isolated from other, larger groups. These variations may be completely absent in other groups.

ProtEST — A database of protein sequences from eight organisms: human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*), fruitfly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), yeast (*Saccharomyces cerevisiae*), plant (*Arabidopsis thaliana*), and bacteria (*Escherichia coli*). (See the ProtEST web page.)

PROW — Protein Reviews On the Web. An online resource that features PROW Guides —authoritative, short, structured reviews on proteins and protein families. The Guides provide approximately 20 standardized categories of information (abstract, biochemical function, ligands, references, etc.) for each protein.

pseudogene — A sequence of DNA that is very similar to a normal gene but that has been altered slightly so that it is not expressed. Such genes were probably once functional but, over time, acquired one or more mutations that rendered them incapable of producing a protein product.

PSI-BLAST — Position-Specific Iterated BLAST. PSI-BLAST (Altschul et al., J Mol Biol 215:403-410; 1990) is used for iterative protein–sequence similarity searches using a position-specific score matrix (PSSM). It is a program for searching protein databases using protein queries to find other members of the same protein family. All statistically significant alignments found by BLAST are combined into a multiple alignment, from which a PSSM is constructed. This matrix is used to search the database for additional significant alignments, and the process may be iterated until no new alignments are found.

PSSM — Position-Specific Score Matrix. The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence.

PubMed — A retrieval system containing citations, abstracts, and indexing terms for journal articles in the biomedical sciences. It includes literature citations supplied directly to NCBI by publishers as well as URLs to full text articles on the publishers' web sites. PubMed contains the complete contents of the MEDLINE and PREMEDLINE databases.

It also contains some articles and journals considered out of scope for MEDLINE, based on either content or on a period of time when the journal was not indexed and, therefore, is a superset of MEDLINE.

PXML — PubMed Central XML file

QBLAST — A queuing system to BLAST that allows users to retrieve their results at their convenience and format their results multiple times with different formatting options.

QTL — Quantitative Trait Locus. A QTL is a hypothesis that a certain region of the chromosome contains genes that contribute significantly to the expression of a complex trait. QTLs are generally identified by comparing the linkage of polymorphic molecular markers and phenotypic trait measurements. The density of the linkage map is important in the accurate and precise location of QTLs; the higher the map density, the more precise the location of the putative QTL, although there is increased likelihood that false positives will be detected. Once QTLs have been mapped to a relatively small chromosomal region, other molecular methods can be used to isolate specific genes.

RCSB — Research Collaboratory for Structural Bioinformatics. RCSB is a nonprofit consortium that works toward the elucidation of biological, macromolecular, 3-D structures.

Reciprocal best hits — Reciprocal best hits are proteins from different organisms that are each other's top BLAST hit, when the proteomes from those organisms are compared to each other. For example, proteins A–Z in organism 1 are compared against proteins AA–ZZ in organism 2. If protein A has a best hit to protein RR, and RR's best hit, when it is compared to all the proteins in organism 1, also turns out to protein A, then A and RR are reciprocal best hits. However, if RR's best hit is to B rather than to A, then A and RR are not reciprocal best hits.

RefSeq — RefSeq is the NCBI database of reference sequences; a curated, non-redundant set including genomic DNA contigs, mRNAs and proteins for known genes, and entire chromosomes.

RepeatMasker — Program that screens DNA sequences for interspersed repeats and low-complexity DNA sequences.

RFLP — Restriction Fragment Length Polymorphism. Genetic variations at the site where a restriction enzyme cuts a piece of DNA. Such variations affect the size of the resulting fragments. These sequences can be used as markers on physical maps and linkage maps. RFLP is also pronounced “rif lip”.

RH map — Radiation Hybrid map. A genome map in which STSs are positioned relative to one another on the basis of the frequency with which they are separated by radiation-induced breaks. The frequency is assayed by analyzing a panel of human–hamster hybrid cell lines. These hybrids are produced by irradiating human cells, which damages the cells and fragments the DNA. The dying human cells are fused with thymidine kinase negative (TK–) live hamster cells. The fused cells are grown under conditions that select against

hamster cells and favor the growth of hybrid cells that have taken up the human *TK* gene. In the RH maps, the unit of distance is centirays (cR), denoting a 1% chance of a break occurring between two loci.

RNA — Ribonucleic Acid. A single-stranded nucleic acid, similar to DNA, but having a ribose sugar, instead of deoxyribose, and uracil instead of thymine as one of its bases.

RPS-BLAST — Reverse Position-Specific BLAST. A program used to identify conserved domains in a protein query sequence. It does this by comparing a query protein sequence to position-specific score matrices (PSSM)s that have been prepared from conserved domain alignments. RPS-BLAST is a "reverse" version of position-specific iterated BLAST (PSI-BLAST); however, RPS-BLAST compares a query sequence against a database of profiles prepared from ready-made alignments, whereas PSI-BLAST builds alignments starting from a single protein sequence.

SAGE — Serial Analysis of Gene Expression. An experimental technique designed to quantitatively measure gene expression.

Sequin — Sequin is a stand-alone software tool developed by the NCBI for submitting and updating entries to the GenBank, EMBL, or DDBJ sequence databases. It is capable of handling simple submissions that contain a single, short mRNA sequence and complex submissions containing long sequences, multiple annotations, segmented sets of DNA, or phylogenetic and population studies.

SGD — Saccharomyces Genome Database. A database for the molecular biology and genetics of *Saccharomyces cerevisceae*, also known as baker's yeast.

SGML — Standard Generalized Markup Language. The international standard for specifying the structure and content of electronic documents. SGML is used for the markup of data in a way that is self-describing. SGML is not a language but a way of defining languages that are developed along its general principles. A subset of SGML called XML is more widely used for the markup of data. HTML (Hypertext Markup Language) is based on SGML and uses some of its concepts to provide a universal markup language for the display of information and the linking of different pieces of that information.

SKY — Spectral Karyotyping. SKY is a technique that allows for the visualization of all of an organism's chromosomes together, each labeled with a different color. This is achieved by using chromosome-specific, single-stranded DNA probes (each labeled with a different fluorophore) to hybridize or bind to the chromosomes of a cell; resulting in each chromosome being painted a different color. This technique is useful for identifying chromosome abnormalities because it is easy to spot instances where a chromosome painted in one color has a small piece of another chromosome, painted in a different color, attached to it. (Also see FISH, CGH.)

SKYGRAM — 1. A software tool to automatically convert the short-form karyotype into an image representation of a cell or clone, with each chromosome displayed in a different

color, with band overlay. The program will also incorporate the number of cells for each structural abnormality, which is displayed in brackets. 2. The full ideogram or a cell or clone, with each chromosome displayed in a different color, with band overlay.

SMART — Simple Modular Architecture Research Tool. A tool to allow automatic identification and annotation of domains in user-supplied protein sequences. For example, the SWISS-PROT database is an extensively annotated and nonredundant collection of protein sequences. SWISS-PROT annotations have been mined for SMART-derived annotations of alignments.

SMD — Stanford Microarray Database. SMD stores raw and normalized data from microarray experiments, as well as their corresponding image files. In addition, the SMD provides interfaces for data retrieval, analysis, and visualization. Data are released to the public at the researcher's discretion or upon publication.

SNP — Common, but minute, variations that occur in human DNA at a frequency of 1 every 1,000 bases. An SNP is a single base-pair site within the genome at which more than one of the four possible base pairs is commonly found in natural populations. Several hundred thousand SNP sites are being identified and mapped on the sequence of the genome, providing the densest possible map of genetic differences. SNP is pronounced "snip".

SOFT — Simple Omnibus Format in Text. SOFT is an ASCII text format that was designed to be a machine-readable representation of data retrieved from, or submitted to, the Gene Expression Omnibus (GEO). SOFT is also a line-based format, making it easy to parse, using commonly available text processing and formatting languages. (For examples of SOFT, see the guide.)

splice sites — Refers to the location of the exon-intron junctions in a pre-mRNA (i.e., the primary transcript that must undergo additional processing to become a mature RNA for translation into a protein). Splice sites can be determined by comparing the sequence of genomic DNA with that of the cDNA sequence. In mRNA, introns (non-protein coding regions) are removed by the splicing machinery; however, exons can also be removed. Depending on which exons (or parts of exons) are removed, different proteins can be made from the same initial RNA or gene. Different proteins created in this way are "splice variants" or "alternatively spliced".

SSAHA — Sequence Search and Alignment by Hashing Algorithm. SSAHA is a software tool for very fast matching and alignment of DNA sequences and is used for searching databases containing large amounts (gigabases) of genome sequence. It achieves its fast search speed by converting sequence information into a "hash table" data structure, which can then be searched very rapidly for matches (Ning et al., Genome Res 11:1725-1729; 2001).

SSLP — Simple Sequence Length Polymorphisms. SSLPs are markers based on the variation in the number of short tandem repeats in DNA.

STS — A short DNA segment that occurs only once in the human genome, the exact location and order of bases of which are known. Because each is unique, STSs are helpful for chromosome placement of mapping and sequencing data from many different laboratories. STSs serve as landmarks on the physical map of the human genome.

substitution matrix — A substitution matrix containing values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids. Such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of amino acids. If the sample is large enough to be statistically significant, the resulting matrices should reflect the true probabilities of mutations occurring through a period of evolution. (See also BLOSUM 62.)

SWISS-PROT — SWISS-PROT is a curated protein sequence database that provides a high level of annotation (such as the description of protein function, domain structures, post-translational modifications, variants, etc.), a minimal level of redundancy, and high level of integration with other databases.

Sybase — A trademarked family of products that include databases, development tools, integration middleware, enterprise portals, and mobile and wireless servers.

synteny — On the same strand. The phrase "conserved synteny" refers to conserved gene order on chromosomes of different, related species.

Tax BLAST — BLAST Taxonomy Reports page. Tax BLAST groups BLAST hits by source organism, according to information in NCBI's Taxonomy database. Species are listed in order of sequence similarity with the query sequence, the strongest match listed first.

taxID — Taxonomy Identifier. The taxID is a stable unique identifier for each taxon (for a species, a family, an order, or any other group in the taxonomy database). The taxID is seen in the GenBank records as a "source" feature table entry; for example, /db_xref="taxon:<9606>" is the taxID for *Homo sapiens*, and the line is therefore found in all recent human sequence records.

taxid — See taxID.

termination codon or stop codon — One of three codons that do not specify any amino acid and hence causes translation of mRNA into protein to be terminated. These codons mark the end of a protein coding sequence.

TIGR — The Institute for Genomic Research

tiling path — An ordered list or map that defines a set of overlapping clones that covers a chromosome or other extended segment of DNA.

TPA — Third-Party Annotation

TPF — Tiling Path Format. A table format used to specify the set of clones that will provide the best possible sequence coverage for a particular chromosome, the order of the clones along the chromosome, and the location of any gaps in the clone tiling path. Also

used to refer to a file (Tiling Path File) in which the minimal tiling path of clones covering a chromosome is specified in Tiling Path Format or to the minimal tiling path of clones so defined.

translation start site — The position within an mRNA at which synthesis of a protein begins. The translation start site is usually an AUG codon, but occasionally, GUG or CUG codons are used to initiate protein synthesis.

UID — Unique Identifier

UMLS — Unified Medical Language System. A project of the National Library of Medicine for the development and distribution of multipurpose, electronic "Knowledge Sources", and associated lexical programs. The purpose of the UMLS is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources and to make it easy for users to link disparate information systems, including computer-based patient records, bibliographic databases, factual databases, and expert systems.

unfinished sequence — See draft sequence.

UniGene cluster — ESTs and full-length mRNA sequences organized into clusters such that each represents a unique known or putative gene within the organism from which the sequences were obtained. UniGene clusters are annotated with mapping and expression information when possible (e.g., for human) and include cross-references to other resources. Sequence data can be downloaded by cluster through the UniGene web pages, or the complete dataset can be downloaded from the repository/UniGene directory of the FTP site.

UniSTS — UniSTS presents a unified, non-redundant view of sequence-tagged sites (STSs). UniSTS integrates marker and mapping data from a variety of public resources. If two or more markers have different names but the same primer pair, a single STS record is presented for the primer pair, and all the marker names are shown.

UNIX — UNIX is an operating system that was developed by Dennis Ritchie and Kenneth Thompson at Bell Labs more than 30 years ago. It allows multitasking and multiuser capabilities and offers portability with other operating systems. It comes with hundreds of programs that are of two types: integral utilites, such as the command line interpreter; and tools such as email, which are not necessary for the operation of UNIX but provide additional capabilities to the user. It is functionally organized at three levels: the kernel, which schedules tasks and manages storage; the shell, which connects and interprets user's commands, calls programs from memory, and executes them; and tools and applications, which offer additional functionality to the operating system, such as word processing and business applications. UNIX® was registered by Bell Laboratories as a trademark for computer operating systems. Today, this mark is owned by The Open Group.

URL — Uniform Resource Locator. The address of a resource on the Internet. URL syntax is in the form of protocol://host/localinfo, where "protocol" specifies the means of fetching the object (such as HTTP, used by WWW browsers and servers to exchange information, or FTP), "host" specifies the remote location where the object resides, and "localinfo" is a string (often a file name) passed to the protocol handler at the remote location. Also called Uniform Resource Identifier (URI).

UTF-8 — UCS (Universal Character Set) Transformation Format. An AscII-preserving encoding method for Unicode (a standard to provide a unique number for every character irrespective of the platform, program, or language).

UTR — Untranslated Region. The 3′ UTR is that portion of an mRNA from the position of the last codon that is used in translation to the 3′ end. The 5′ UTR is that portion of an mRNA from the 5′ end to the position of the first codon used in translation.

VAST — Vector Alignment Search Tool. A computer algorithm used to identify similar protein 3D structures.

weight — An assignment of importance to a term in a search query. If a term in a search query is found to match a word in a document, that word is given a "weight". The exact weight of the word will depend on the emphasis given to the word by the author or its position in the document. For example, a word that occurs in a chapter title will have a higher weight than the same word if it occurs in the body of the chapter. Similarly, words that occur in data collections are also assigned weights, depending on how frequently the terms occur in the collection.

WGS sequence — Whole Genome Shotgun sequence. In this semi-automated sequencing technique, high-molecular-weight DNA is sheared into random fragments, size selected (usually 2, 10, 50, and 150 kb), and cloned into an appropriate vector. The clones are then sequenced from both ends. The two ends of the same clone are referred to as mate pairs. The distance between two mate pairs can be inferred if the library size is known and has a narrow window of deviation. The sequences are aligned using sequence assembly software. Proponents of this approach argue that it is possible to sequence the whole genome at once using large arrays of sequencers, which makes the whole process much more efficient than the traditional approaches.

WHO — World Health Organization

WWW — World Wide Web. A consortium (W3C) that develops technologies such specifications, guidelines, software, and tools for the internet.

XML — Extensible Markup Language. XML describes a class of data objects called XML documents and partially describes the behavior of computer programs that process them. XML is a subset of SGML, and XML documents are conforming SGML documents. XML documents are made up of storage units called entities, which contain either parsed or unparsed data. Parsed data is made up of characters (a unit of text), some of which form character data, and some of which form markup. Markup includes tags that provide

information about the data, i.e., a description of the structure and content of the document. Character data comprises all the text that is not markup. XML provides a mechanism to impose constraints on the storage layout and logical structure.

XSL — Extensible Stylesheet Language. XSL is used for the transformation of XML-based data into HTML or other presentation formats, for display in a web browser. This is a two-part process. First, the structure of the input XML tree must be transformed into a new tree (e.g., HTML), allowing reordering of the elements, addition of text, and calculations —all without modification to the source document. This process is described by XSLT. Second, XSL-FO (XSL Formatting Objects, an XML vocabulary for formatting) is used for formatting the output, defining areas of the display page and their properties. In this way, the source XML document can be maintained from the perspective of "pure content" and can be separated from the presentation. An XML document can be delivered in different formats to different target audiences by simply switching style sheets.

XSLT — Extensible Stylesheet Language: Transformations. XSLT is a language for transforming the structure of an XML document. XSLT is designed for use as part of XSL, the stylesheet language for XML. A transformation expressed in XSLT describes a sequence of template rules for transforming a source tree into a result tree; elements from the source tree can be filtered and reordered, and a different structure can be added. A template rule has two parts: a pattern that is matched against nodes in the source tree; and a template that can be instantiated to form part of the result tree. This makes XSLT a declarative language because it is possible to specify what output should be produced when specific patterns occur in the input, which distinguishes it from procedural programming languages, where it is necessary to specify what tasks have to be performed in what order. XSLT makes use of the expression language defined by XPath (a language for addressing the parts of an XML document) for selecting elements for processing, for conditional processing, and for generating text.

YAC — Yeast Artificial Chromosome. Extremely large segments of DNA from another species spliced into the DNA of yeast. YACs are used to clone up to one million bases of foreign DNA into a host cell, where the DNA is propagated along with the other chromosomes of the yeast cell.

ZFIN — Zebrafish Information Network. ZFIN is a database for the zebrafish model organism that holds information on wild-type stocks, mutants, genes, gene expression data, and map markers.