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.

NLM Citation: Maglott D, Pruitt K, Tatusova T. Gene: A Directory of Genes. 2005 Mar 3 [Updated 2011 Dec 12]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 19.

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 If the record already exists in Gene, new information is added and outdated information is corrected.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

continues on next page...

continued from previous page.

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.

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

continues on next page...

continued from previous page.

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
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.

continues on next page...

continued from previous page.

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.

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. PubMed PMID: 15608257.
2. Wheeler D.L.et al. *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res. 2005;33(Database issue):D39–45. PubMed PMID: 15608222.
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. PubMed PMID: 15608248.
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. PubMed PMID: 15489334.
5. Pruitt K.D.et al. *Introducing RefSeq and LocusLink: curated human genome resources at the NCBI.* Trends Genet. 2000;16(1):44–7. PubMed PMID: 10637631.

6. Schuler G.D.et al. *Entrez: molecular biology database and retrieval system.* Methods Enzymol. 1996;266:141–62. PubMed PMID: 8743683.
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. PubMed PMID: 12700172.