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.

NLM Citation: Pontius JU, Wagner L, Schuler GD. UniGene: A Unified View of the Transcriptome. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 21.

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.

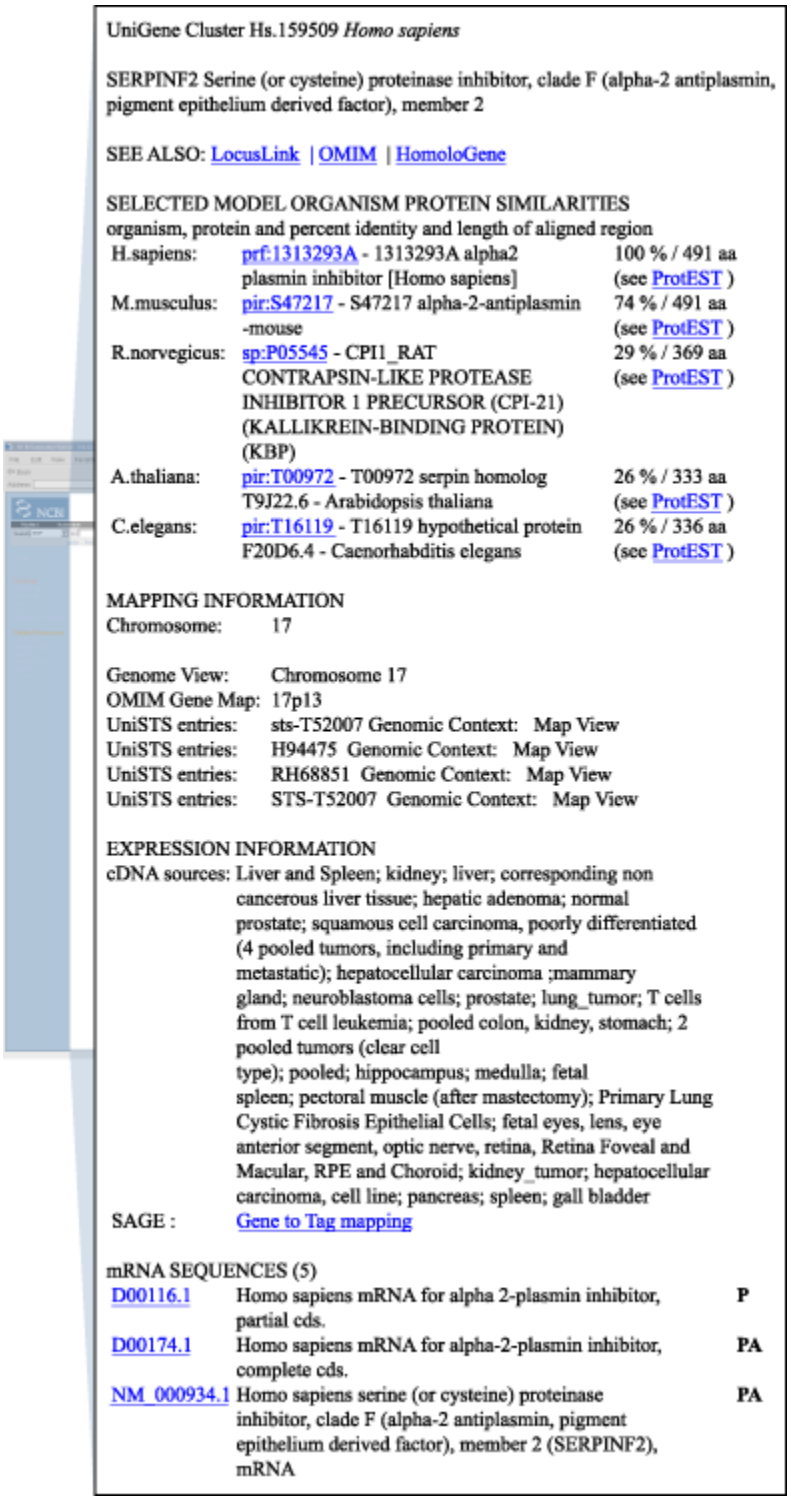
One avenue to gene discovery is to use a database search tool, such as BLAST (11), to perform a sequence similarity search against dbEST. The query for such a search would be a gene or protein sequence, perhaps from a model organism, that is expected to be related to the human gene of interest. Because 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



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.

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.

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.

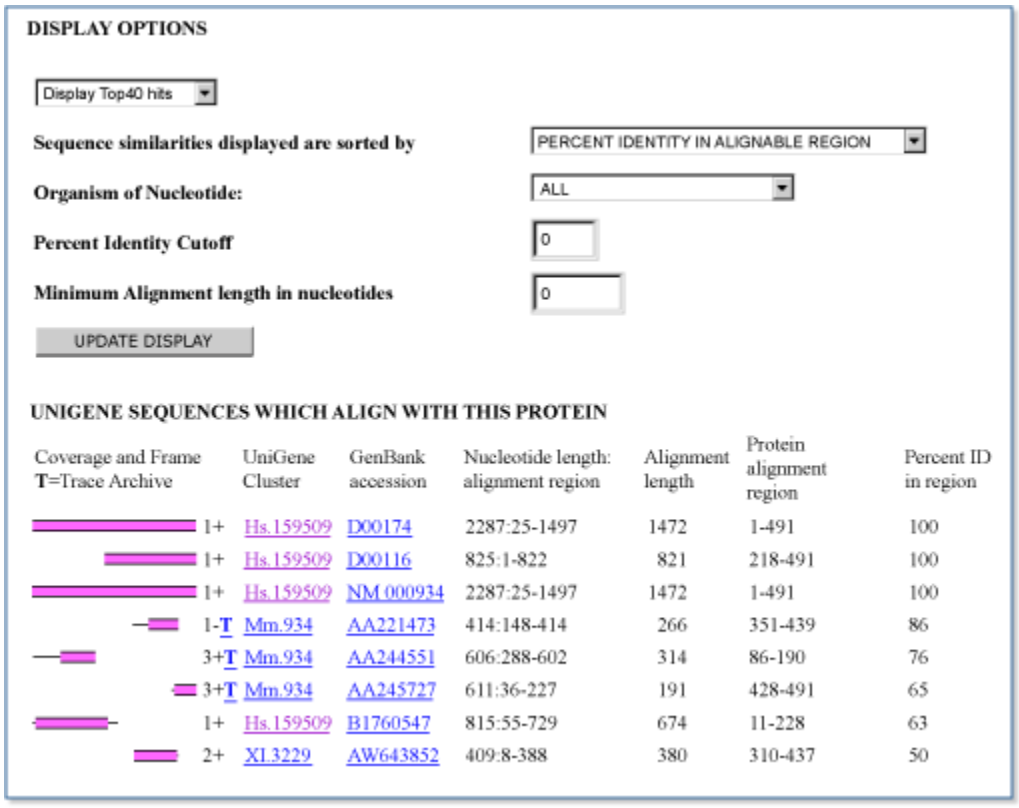
Although ESTs are a poor probe of gene expression, both the total number of ESTs and the tissues from which they originated are often useful. Both of these are displayed in the cluster browser. The tissues are listed under Expression Information, which includes the tissue source of libraries of the component 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.



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.

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.

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.

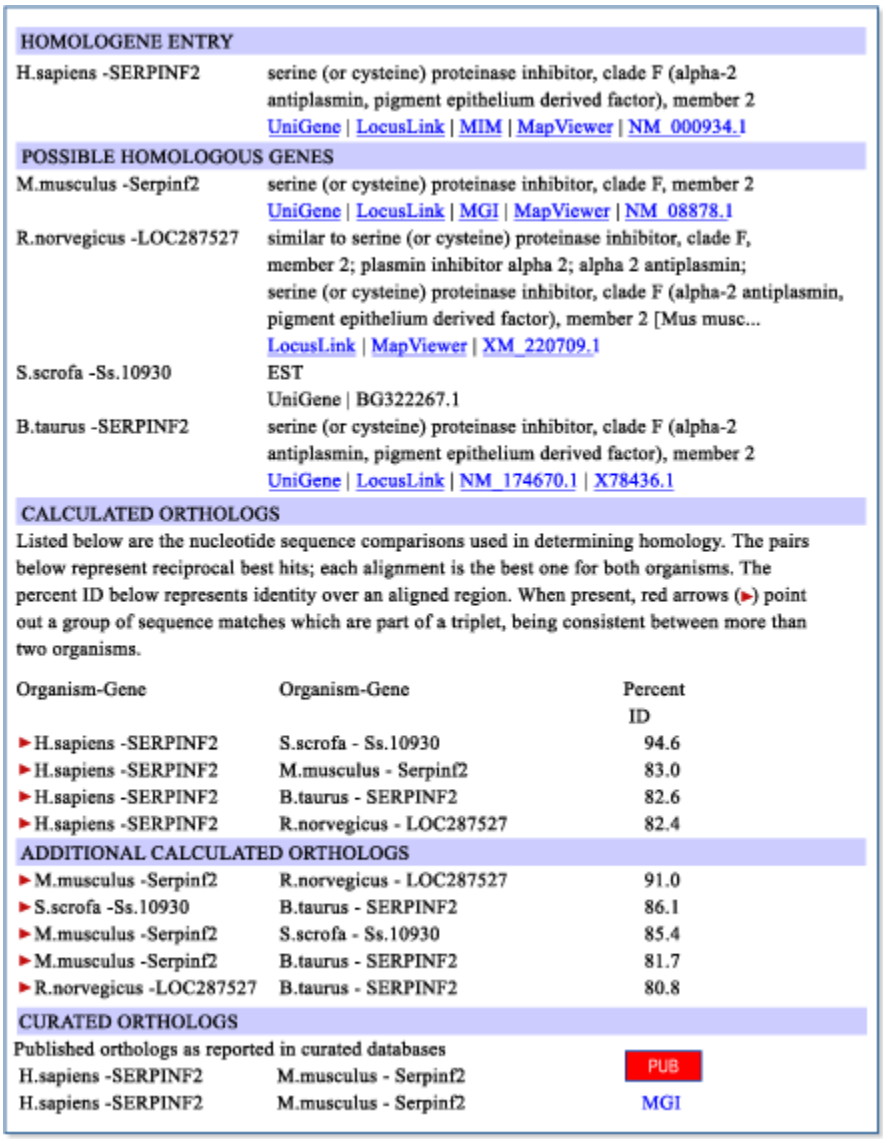
DDD uses the Fisher Exact test to restrict the output to statistically significant differences ($P \leq 0.05$). The analysis is also restricted to deeply sequenced libraries; only those with over 1000 sequences in UniGene are included in DDD. These requirements place limitations on the capabilities of 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



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.

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 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.