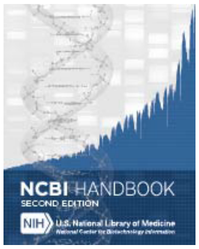
**NLM Citation:** Wagner L, Agarwala R. UniGene. 2013 Nov 14. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
**Bookshelf URL:** https://www.ncbi.nlm.nih.gov/books/


NCBI HANDBOOK
SECOND EDITION
NIH U.S. National Library of Medicine
National Center for Biotechnology Information

# UniGene

Lukas Wagner, PhD and Richa Agarwala, PhD

Created: November 14, 2013.

## Scope

UniGene is a largely automated analytical system for producing an organized view of the transcriptome. By analyzing sequences known to be expressed, and the libraries or samples from which they were derived, it is possible to organize the data into gene-specific clusters, and, in some cases, evaluate the patterns of expression by tissue, health status, and age. 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.

## History

The task of assembling an inventory of all genes of *Homo sapiens* and other organisms began more than two decades ago with large-scale survey sequencing of transcribed sequences. The resulting Expressed Sequence Tags (ESTs) continue to be an invaluable component of efforts to characterize the transcriptome of many organisms. These efforts which rely on ESTs include genome annotation (2-4), expression systems (5), and full-length cDNA cloning projects (6). In addition, 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.

Now that 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. However, many species of medical and agricultural importance do not yet have a complete annotated genome available. Furthermore, when the genomic sequence of an organism is made public, a collection of cDNA sequences provides the best tool for identifying genes within the DNA sequence. When the source material for cDNA sequences is drawn from diverse tissues, an approximate expression profile for the organism's transcriptome can be computed. This approximate expression profile can serve to at least identify transcripts of interest to researchers interested in a particular system, and at best to characterize the function of novel transcripts. Thus, we can anticipate that the sequencing of transcribed products will remain a significant area of interest well into the future.

## Data Model

The data model for UniGene is straightforward. Identify sequences of RNA molecules, the source of those sequences (species, tissue, age, health status), compute when independent sequences are derived from the same gene based on sequence similarity, and report the results. Historically this computation was based on ESTs (Extended Sequence Tags), but now the vast majority of sequences are either full-length clones or RNAseq data.

## ESTs

The basic strategy of generating ESTs involves selecting cDNA clones at random and performing a single, automated, sequencing read from one or both ends of their inserts. This class of sequence is characterized by being short (typically about 400-600 bases) and relatively inaccurate (around 2% error). 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. 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. In addition, large-scale EST projects were launched for several other organisms of experimental interest. In 1992, a database called dbEST was established to serve as a collection point for ESTs, which are then distributed to the scientific community as the EST division of GenBank.

# Dataflow

The number of transcribed sequences is large enough that interactive analysis of each sequence by a researcher is impossible. 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 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.

# Access

The UniGene website allows the user to search for particular genes of interest, or to browse UniGene entries related by expression or sequence similarity. Each UniGene Web page includes a header with a query bar and a sidebar providing links to related online resources. UniGene is also the basis for other NCBI resources:

- ProtEST, a facility for browsing protein similarities;
- Digital DifferentialDisplay (DDD), for comparison of EST-based expression profiles; and
- a library browser and display, which support exploration of cDNA libraries from

tissues of interest.

Looking for sequences expressed under a particular circumstance (a body site or developmental stage, for example) is a common method by which users identify individual genes or sets of genes that are of interest. There are several interfaces into UniGene's data to help users do this. Most broadly, there is a straightforward way to browse all cDNA libraries prepared with RNA from a particular biological source. Properties of individual libraries are summarized as well; the library submitter's description of source material and protocol, and a summary of the UniGene clusters expressed by the library's sequences.

## UniGene Cluster Browser

The UniGene Cluster page summarizes the sequences in the cluster and additional derived information that may be used to infer the identity and in some cases function 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., Gene, HomoloGene, 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 selected model organisms with an annotated geneome. For each model organism, the protein with the highest degree of sequence similarity to the nucleotide sequence is listed, with its title and GenBank accession. The sequence alignment is summarized using the percent identity and length of the aligned region. Also provided is a link (a popup menu attached to the protein accession) to other protein-oriented NCBI resources including ProtEST, which summarizes translating searches of the model organism protein against all organisms in UniGene.

The next section summarizes information on the aligned or inferred map position of the gene. For human and some other annotated genomes, the map position and link to the genomic neighborhood as represented in Map Viewer. Absent these aligned map positions, 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 not an optimal probe of gene expression, both the total number of ESTs and the tissues from which they originated are often useful. In the Gene Expression section of the cluster browser, a link to a summary of the gene's expression in cDNA libraries is available, with an example shown in Figure 2. Expression in each body site or developmental stage available in ESTs (excepting those from normalized or subtracted libraries, also excepting those which come from libraries of mixed source material) is reported, expresssed as counts of ESTs transcribed per miliion sequenced. Expression data is also available by FTP. UniGene clusters with similar expression profile are precomputed, and available under the link labelled "show more like this." This is most likely to be informative when expression differs markedly from uniform expression. Clusters that are predominantly expressed in a single body site or developmental stage are searchable in Entrez, by using the field "Restricted Expression." More specifically, these are clusters where 2/3 or more of the detected gene expression expresssed in normalized units of transcripts per million is from a single source. Links to NCBI's GEO computed from the GenBank accessions in the UniGene cluster are also present in this section of the cluster view.

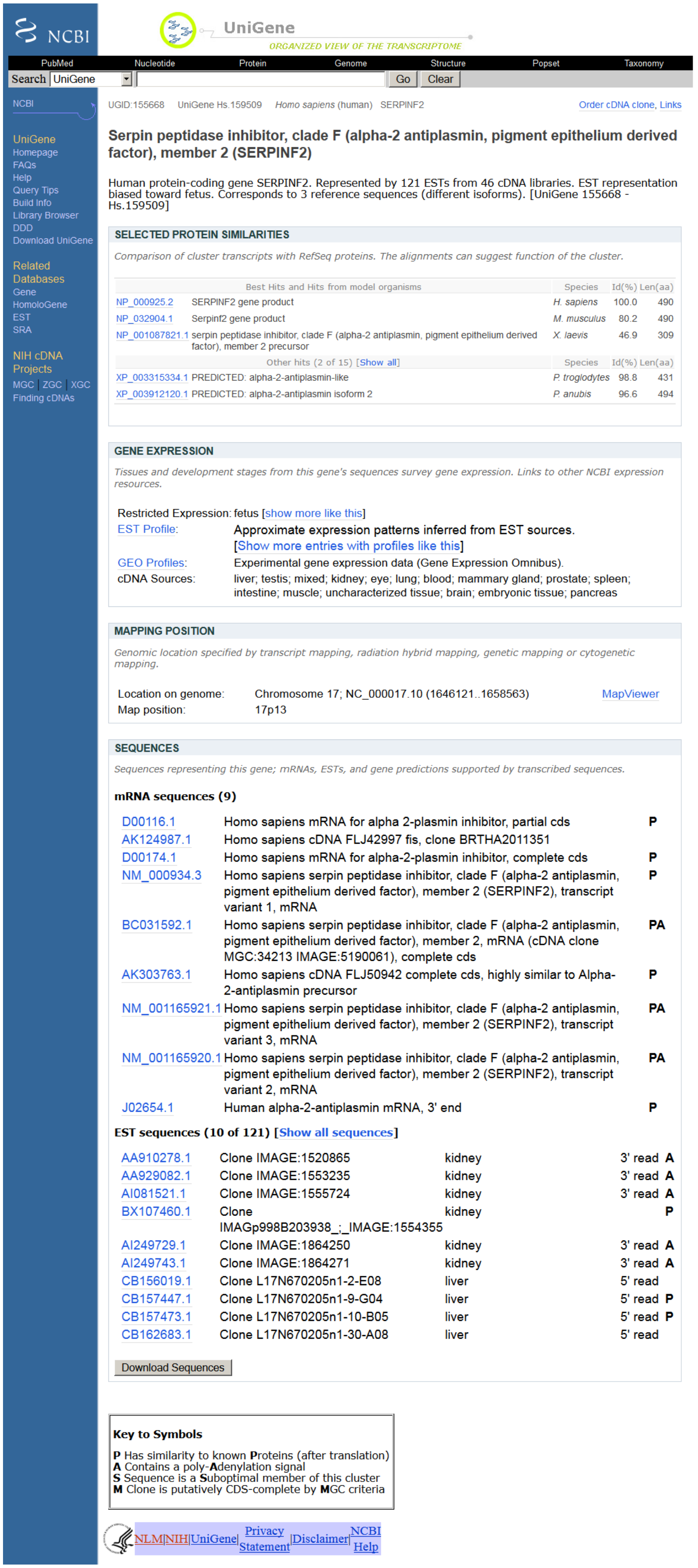
The component sequences of the cluster are listed, with a brief description of each one and a link to its UniGene Sequence page. The Sequence page provides more detailed information about the individual sequence, and in the case of ESTs, includes a link to its corresponding UniGene Library page. On the 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 a button for users to download the sequences of the cluster in FASTA format.

An FTP representation of UniGene is available as well. Sequence sets as FASTA (both aggregate and best representative per cluster), a summary of the mapping of sequences to UniGene clusters with library of origin for ESTs, and an expression summary. A common use of UniGene is to use a single representative sequence from each cluster for primer design or as a BLAST database. In this case, researchers are advised to retain both the sequence accession number as well as the cluster identifier for later reference. This is because the cluster identifier is not guaranteed to be indefinitely stable. While most UniGene builds differ only through incremental changes to existing clusters or the addition of newly represented transcripts, new sequences or new genome mapping can provide information that leads to substantial reorganization of previously identified clusters.

# Related Tools

## Protein Similarity Browser

The ProtEST section of UniGene allows the user to explore precomputed alignments for a selected protein to the cDNA sequences found in any cluster. Especially for cases where looking at alignments to the same protein of transcripts from multiple organisms, this interface provides a single concise overview. In the cluster viewer's protein similarity section, this overview is under the "Protein/EST matches” link in the popup menu that appears on mouseover of a protein accession; this popup is shown in Figure 3. BLASTX has been used to compare each sequence in UniGene to selected protein sequences drawn from model organisms with a fully annotated genome. By default, only alignments to the organisms in the same broad taxonomic group as the original



UGID:155668 UniGene Hs.159509 *Homo sapiens* (human) SERPINF2

**Serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2)**

Human protein-coding gene SERPINF2. Represented by 121 ESTs from 46 cDNA libraries. EST representation biased toward fetus. Corresponds to 3 reference sequences (different isoforms). [UniGene 155668 - Hs.159509]

**SELECTED PROTEIN SIMILARITIES**

*Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.*

| Best Hits and Hits from model organisms | | Species | Id(%) | Len(aa) |
|---|---|---|---|---|
| NP_000925.2 | SERPINF2 gene product | *H. sapiens* | 100.0 | 490 |
| NP_032904.1 | Serpinf2 gene product | *M. musculus* | 80.2 | 490 |
| NP_001087821.1 | serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 precursor | *X. laevis* | 46.9 | 309 |
| Other hits (2 of 15) [Show all] | | Species | Id(%) | Len(aa) |
| XP_003315334.1 | PREDICTED: alpha-2-antiplasmin-like | *P. troglodytes* | 98.8 | 431 |
| XP_003912120.1 | PREDICTED: alpha-2-antiplasmin isoform 2 | *P. anubis* | 96.6 | 494 |

**GENE EXPRESSION**

*Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.*

Restricted Expression: fetus [show more like this]

EST Profile: Approximate expression patterns inferred from EST sources. [Show more entries with profiles like this]

GEO Profiles: Experimental gene expression data (Gene Expression Omnibus).

cDNA Sources: liver; testis; mixed; kidney; eye; lung; blood; mammary gland; prostate; spleen; intestine; muscle; uncharacterized tissue; brain; embryonic tissue; pancreas

**MAPPING POSITION**

*Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.*

Location on genome: Chromosome 17; NC_000017.10 (1646121..1658563) MapViewer

Map position: 17p13

**SEQUENCES**

*Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.*

**mRNA sequences (9)**

| | | |
|---|---|---|
| D00116.1 | Homo sapiens mRNA for alpha 2-plasmin inhibitor, partial cds | P |
| AK124987.1 | Homo sapiens cDNA FLJ42997 fis, clone BRTHA2011351 | |
| D00174.1 | Homo sapiens mRNA for alpha-2-plasmin inhibitor, complete cds | P |
| NM_000934.3 | Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2), transcript variant 1, mRNA | P |
| BC031592.1 | Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2, mRNA (cDNA clone MGC:34213 IMAGE:5190061), complete cds | PA |
| AK303763.1 | Homo sapiens cDNA FLJ50942 complete cds, highly similar to Alpha-2-antiplasmin precursor | P |
| NM_001165921.1 | Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2), transcript variant 3, mRNA | PA |
| NM_001165920.1 | Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2), transcript variant 2, mRNA | PA |
| J02654.1 | Human alpha-2-antiplasmin mRNA, 3' end | P |

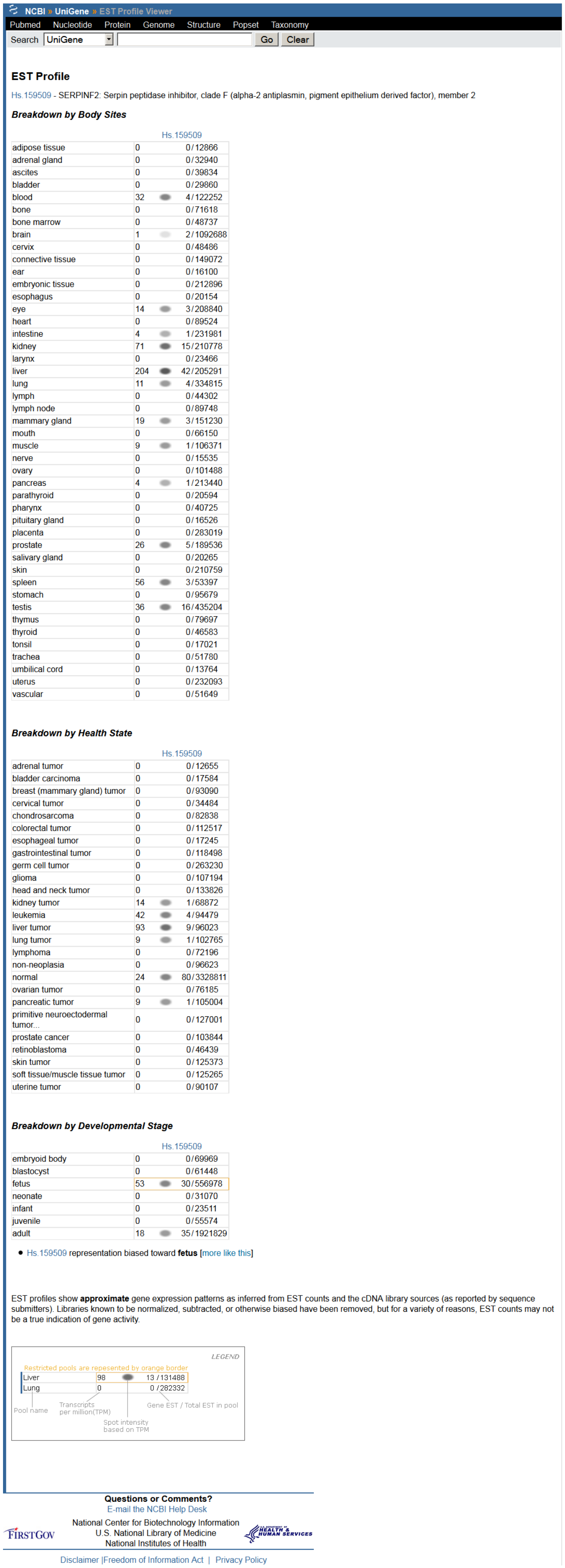
**EST sequences (10 of 121) [Show all sequences]**

| | | | | |
|---|---|---|---|---|
| AA910278.1 | Clone IMAGE:1520865 | kidney | 3' read | A |
| AA929082.1 | Clone IMAGE:1553235 | kidney | 3' read | A |
| AI081521.1 | Clone IMAGE:1555724 | kidney | 3' read | A |
| BX107460.1 | Clone IMAGp998B203938_;_IMAGE:1554355 | kidney | | P |
| AI249729.1 | Clone IMAGE:1864250 | kidney | 3' read | A |
| AI249743.1 | Clone IMAGE:1864271 | kidney | 3' read | A |
| CB156019.1 | Clone L17N670205n1-2-E08 | liver | 5' read | |
| CB157447.1 | Clone L17N670205n1-9-G04 | liver | 5' read | P |
| CB157473.1 | Clone L17N670205n1-10-B05 | liver | 5' read | P |
| CB162683.1 | Clone L17N670205n1-30-A08 | liver | 5' read | |

**Key to Symbols**

**P** Has similarity to known **P**roteins (after translation)
**A** Contains a poly-**A**denylation signal
**S** Sequence is a **S**uboptimal member of this cluster
**M** Clone is putatively CDS-complete by **M**GC criteria

**Figure 1.** Web view of a UniGene cluster.


NCBI » UniGene » EST Profile Viewer
Pubmed Nucleotide Protein Genome Structure Popset Taxonomy
Search UniGene Go Clear

## EST Profile

Hs.159509 - SERPINF2: Serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2

***Breakdown by Body Sites***

| | Hs.159509 | | |
|---|---|---|---|
| adipose tissue | 0 | | 0/12866 |
| adrenal gland | 0 | | 0/32940 |
| ascites | 0 | | 0/39834 |
| bladder | 0 | | 0/29860 |
| blood | 32 | ● | 4/122252 |
| bone | 0 | | 0/71618 |
| bone marrow | 0 | | 0/48737 |
| brain | 1 | ● | 2/1092688 |
| cervix | 0 | | 0/48486 |
| connective tissue | 0 | | 0/149072 |
| ear | 0 | | 0/16100 |
| embryonic tissue | 0 | | 0/212896 |
| esophagus | 0 | | 0/20154 |
| eye | 14 | ● | 3/208840 |
| heart | 0 | | 0/89524 |
| intestine | 4 | ● | 1/231981 |
| kidney | 71 | ● | 15/210778 |
| larynx | 0 | | 0/23466 |
| liver | 204 | ● | 42/205291 |
| lung | 11 | ● | 4/334815 |
| lymph | 0 | | 0/44302 |
| lymph node | 0 | | 0/89748 |
| mammary gland | 19 | ● | 3/151230 |
| mouth | 0 | | 0/66150 |
| muscle | 9 | ● | 1/106371 |
| nerve | 0 | | 0/15535 |
| ovary | 0 | | 0/101488 |
| pancreas | 4 | ● | 1/213440 |
| parathyroid | 0 | | 0/20594 |
| pharynx | 0 | | 0/40725 |
| pituitary gland | 0 | | 0/16526 |
| placenta | 0 | | 0/283019 |
| prostate | 26 | ● | 5/189536 |
| salivary gland | 0 | | 0/20265 |
| skin | 0 | | 0/210759 |
| spleen | 56 | ● | 3/53397 |
| stomach | 0 | | 0/95679 |
| testis | 36 | ● | 16/435204 |
| thymus | 0 | | 0/79697 |
| thyroid | 0 | | 0/46583 |
| tonsil | 0 | | 0/17021 |
| trachea | 0 | | 0/51780 |
| umbilical cord | 0 | | 0/13764 |
| uterus | 0 | | 0/232093 |
| vascular | 0 | | 0/51649 |

***Breakdown by Health State***

| | Hs.159509 | | |
|---|---|---|---|
| adrenal tumor | 0 | | 0/12655 |
| bladder carcinoma | 0 | | 0/17584 |
| breast (mammary gland) tumor | 0 | | 0/93090 |
| cervical tumor | 0 | | 0/34484 |
| chondrosarcoma | 0 | | 0/82838 |
| colorectal tumor | 0 | | 0/112517 |
| esophageal tumor | 0 | | 0/17245 |
| gastrointestinal tumor | 0 | | 0/118498 |
| germ cell tumor | 0 | | 0/263230 |
| glioma | 0 | | 0/107194 |
| head and neck tumor | 0 | | 0/133826 |
| kidney tumor | 14 | ● | 1/68872 |
| leukemia | 42 | ● | 4/94479 |
| liver tumor | 93 | ● | 9/96023 |
| lung tumor | 9 | ● | 1/102765 |
| lymphoma | 0 | | 0/72196 |
| non-neoplasia | 0 | | 0/96623 |
| normal | 24 | ● | 80/3328811 |
| ovarian tumor | 0 | | 0/76185 |
| pancreatic tumor | 9 | ● | 1/105004 |
| primitive neuroectodermal tumor... | 0 | | 0/127001 |
| prostate cancer | 0 | | 0/103844 |
| retinoblastoma | 0 | | 0/46439 |
| skin tumor | 0 | | 0/125373 |
| soft tissue/muscle tissue tumor | 0 | | 0/125265 |
| uterine tumor | 0 | | 0/90107 |

***Breakdown by Developmental Stage***

| | Hs.159509 | | |
|---|---|---|---|
| embryoid body | 0 | | 0/69969 |
| blastocyst | 0 | | 0/61448 |
| fetus | 53 | ● | 30/556978 |
| neonate | 0 | | 0/31070 |
| infant | 0 | | 0/23511 |
| juvenile | 0 | | 0/55574 |
| adult | 18 | ● | 35/1921829 |

- Hs.159509 representation biased toward **fetus** [more like this]

EST profiles show **approximate** gene expression patterns as inferred from EST counts and the cDNA library sources (as reported by sequence submitters). Libraries known to be normalized, subtracted, or otherwise biased have been removed, but for a variety of reasons, EST counts may not be a true indication of gene activity.

*LEGEND*

Restricted pools are repesented by orange border

| | | | |
|---|---|---|---|
| Liver | 98 | ● | 13 / 131488 |
| Lung | 0 | | 0 / 282332 |

Pool name
Transcripts per million(TPM)
Spot intensity based on TPM
Gene EST / Total EST in pool

**Questions or Comments?**
E-mail the NCBI Help Desk

National Center for Biotechnology Information
U.S. National Library of Medicine
National Institutes of Health

**Figure 2.** Expression profile view of a UniGene cluster.

organism are shown (primates, rodents, etc), though alignments to a broader set of organisms can be selected from the protEST pulldown menu. These alignments include alignments both to RefSeqs, which are based on a sequenced and annotated mRNA, as well as RefSeqs, which are gene predictions.

The sequence alignments in ProtEST are summarized in tabular form (Figure 4). The first column is a schematic representation of the nucleotideprotein 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 the UniGene cluster ID, the GenBank accession, and a summary of the aligned region and percent identity.

## Digital Differential Display (DDD)

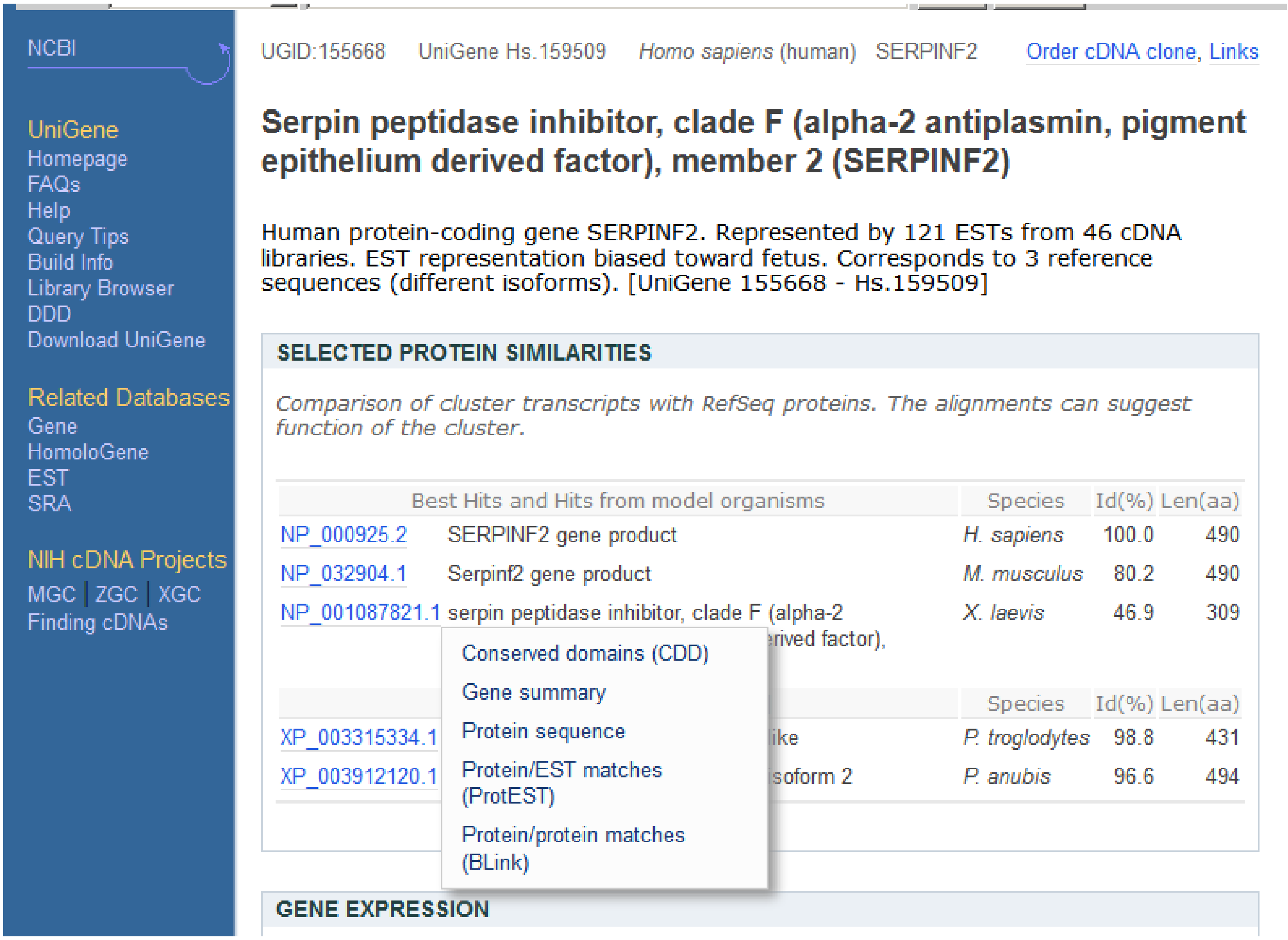
DDD is a tool for comparing EST-based expression profiles among the various libraries, or pools of libraries, represented in UniGene. These comparisons allow the identification of those genes that differ among libraries of different tissues, making it possible to determine which genes may be contributing 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 is UniGene, the DDD resource is organism specific and is available from the UniGene website 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. An example comparing two pools of libraries with similar sequence counts from human muscle and human brain is shown in Figure 5.

DDD uses the Fisher Exact test to restrict the output to statistically significant differences (P 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.

## cDNA Library Browser and UniGene cDNA Library Display

Researchers frequentlly wish to identify particular cDNA libraries that interest them. In addition to the gene-oriented resources described above, UniGene offers an overview of all libraries from an organism of interest in the library browser. Libraries are grouped by their source material (body site or developmental stage where these are described by the library’s submitter), with the Web interface for browsing shown in Figure 6. For individual libraries, the library summary aggregates information provided by the submitter with the UniGene clusters that contain sequences from a library, with the Web interface for an individual library summary shown in Figure 7. Researchers may download all sequences from the library in FASTA format whether they are in any UniGene cluster or not from this page.



UGID:155668 UniGene Hs.159509 *Homo sapiens* (human) SERPINF2

**Serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2)**

Human protein-coding gene SERPINF2. Represented by 121 ESTs from 46 cDNA libraries. EST representation biased toward fetus. Corresponds to 3 reference sequences (different isoforms). [UniGene 155668 - Hs.159509]

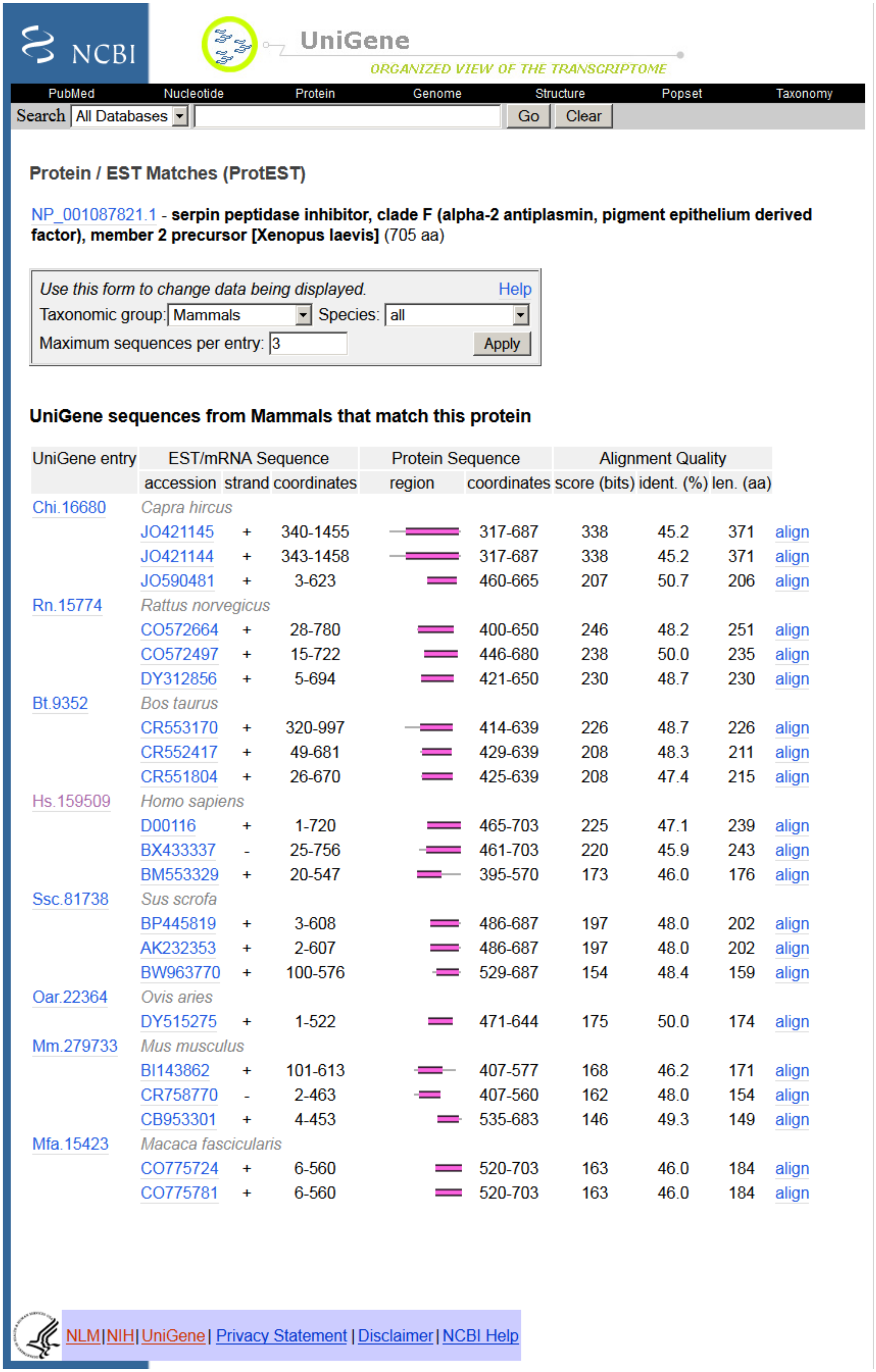
**SELECTED PROTEIN SIMILARITIES**

*Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.*

| Best Hits and Hits from model organisms | | Species | Id(%) | Len(aa) |
|---|---|---|---|---|
| NP_000925.2 | SERPINF2 gene product | *H. sapiens* | 100.0 | 490 |
| NP_032904.1 | Serpinf2 gene product | *M. musculus* | 80.2 | 490 |
| NP_001087821.1 | serpin peptidase inhibitor, clade F (alpha-2<br>rived factor), | *X. laevis* | 46.9 | 309 |
| | | Species | Id(%) | Len(aa) |
| XP_003315334.1 | ike | *P. troglodytes* | 98.8 | 431 |
| XP_003912120.1 | soform 2 | *P. anubis* | 96.6 | 494 |

**GENE EXPRESSION**

**Figure 3.** Popup menu in UniGene cluster browser providing links to information about proteins similar to the transcript sequences in the cluster.



### Protein / EST Matches (ProtEST)

NP_001087821.1 - **serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 precursor [Xenopus laevis]** (705 aa)

*Use this form to change data being displayed.* Help

Taxonomic group: Mammals Species: all

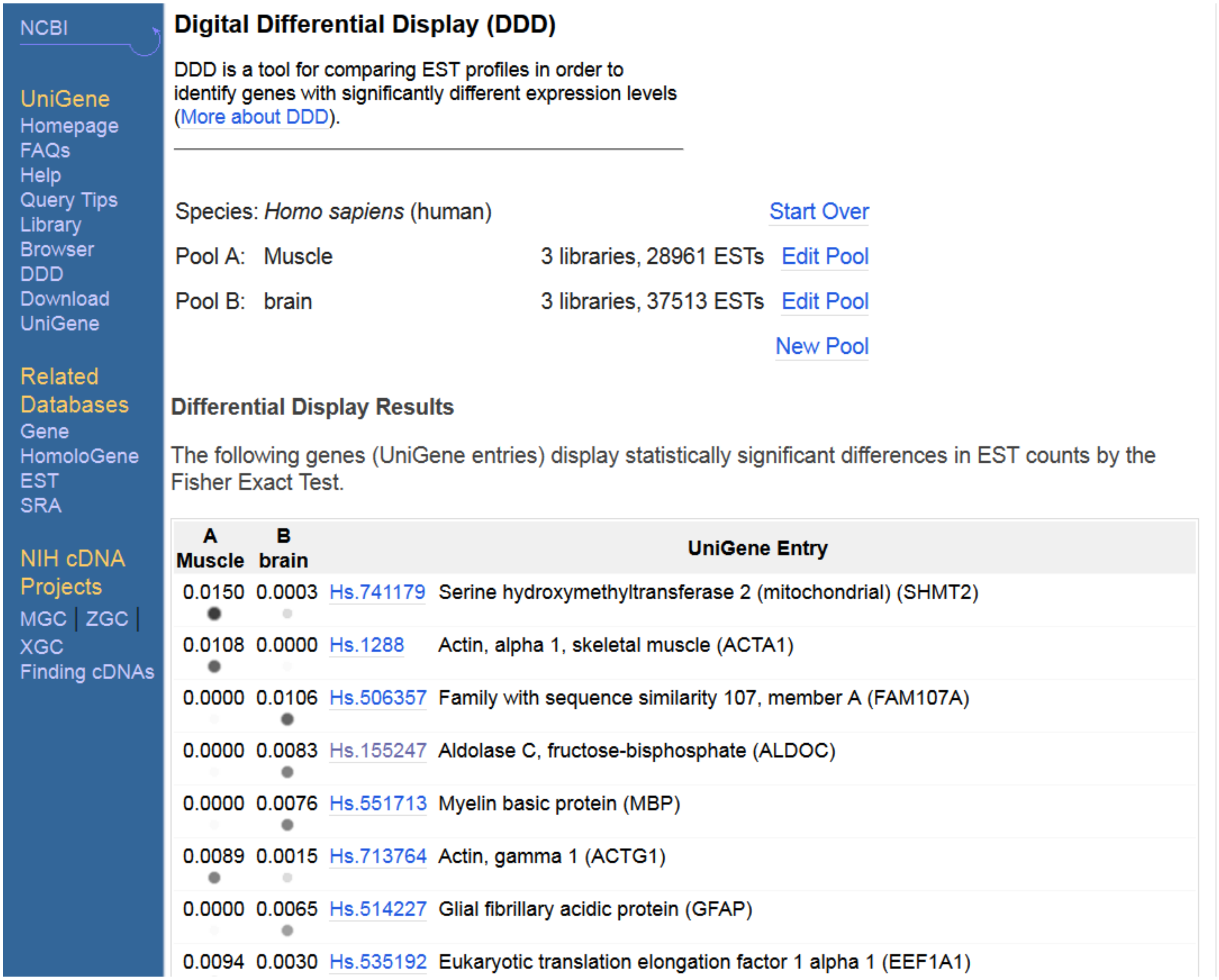
Maximum sequences per entry: 3 Apply

**UniGene sequences from Mammals that match this protein**

| UniGene entry | EST/mRNA Sequence | | | Protein Sequence | | Alignment Quality | | | |
|---|---|---|---|---|---|---|---|---|---|
| | accession | strand | coordinates | region | coordinates | score (bits) | ident. (%) | len. (aa) | |
| Chi.16680 | *Capra hircus* | | | | | | | | |
| | JO421145 | + | 340-1455 | | 317-687 | 338 | 45.2 | 371 | align |
| | JO421144 | + | 343-1458 | | 317-687 | 338 | 45.2 | 371 | align |
| | JO590481 | + | 3-623 | | 460-665 | 207 | 50.7 | 206 | align |
| Rn.15774 | *Rattus norvegicus* | | | | | | | | |
| | CO572664 | + | 28-780 | | 400-650 | 246 | 48.2 | 251 | align |
| | CO572497 | + | 15-722 | | 446-680 | 238 | 50.0 | 235 | align |
| | DY312856 | + | 5-694 | | 421-650 | 230 | 48.7 | 230 | align |
| Bt.9352 | *Bos taurus* | | | | | | | | |
| | CR553170 | + | 320-997 | | 414-639 | 226 | 48.7 | 226 | align |
| | CR552417 | + | 49-681 | | 429-639 | 208 | 48.3 | 211 | align |
| | CR551804 | + | 26-670 | | 425-639 | 208 | 47.4 | 215 | align |
| Hs.159509 | *Homo sapiens* | | | | | | | | |
| | D00116 | + | 1-720 | | 465-703 | 225 | 47.1 | 239 | align |
| | BX433337 | - | 25-756 | | 461-703 | 220 | 45.9 | 243 | align |
| | BM553329 | + | 20-547 | | 395-570 | 173 | 46.0 | 176 | align |
| Ssc.81738 | *Sus scrofa* | | | | | | | | |
| | BP445819 | + | 3-608 | | 486-687 | 197 | 48.0 | 202 | align |
| | AK232353 | + | 2-607 | | 486-687 | 197 | 48.0 | 202 | align |
| | BW963770 | + | 100-576 | | 529-687 | 154 | 48.4 | 159 | align |
| Oar.22364 | *Ovis aries* | | | | | | | | |
| | DY515275 | + | 1-522 | | 471-644 | 175 | 50.0 | 174 | align |
| Mm.279733 | *Mus musculus* | | | | | | | | |
| | BI143862 | + | 101-613 | | 407-577 | 168 | 46.2 | 171 | align |
| | CR758770 | - | 2-463 | | 407-560 | 162 | 48.0 | 154 | align |
| | CB953301 | + | 4-453 | | 535-683 | 146 | 49.3 | 149 | align |
| Mfa.15423 | *Macaca fascicularis* | | | | | | | | |
| | CO775724 | + | 6-560 | | 520-703 | 163 | 46.0 | 184 | align |
| | CO775781 | + | 6-560 | | 520-703 | 163 | 46.0 | 184 | align |

NLM | NIH | UniGene | Privacy Statement | Disclaimer | NCBI Help

**Figure 4.** Protein-transcript alignment summary.



NCBI

UniGene
Homepage
FAQs
Help
Query Tips
Library Browser
DDD
Download UniGene

Related Databases
Gene
HomoloGene
EST
SRA

NIH cDNA Projects
MGC | ZGC | XGC
Finding cDNAs

**Digital Differential Display (DDD)**

DDD is a tool for comparing EST profiles in order to identify genes with significantly different expression levels (More about DDD).

Species: *Homo sapiens* (human) Start Over

Pool A: Muscle 3 libraries, 28961 ESTs Edit Pool

Pool B: brain 3 libraries, 37513 ESTs Edit Pool

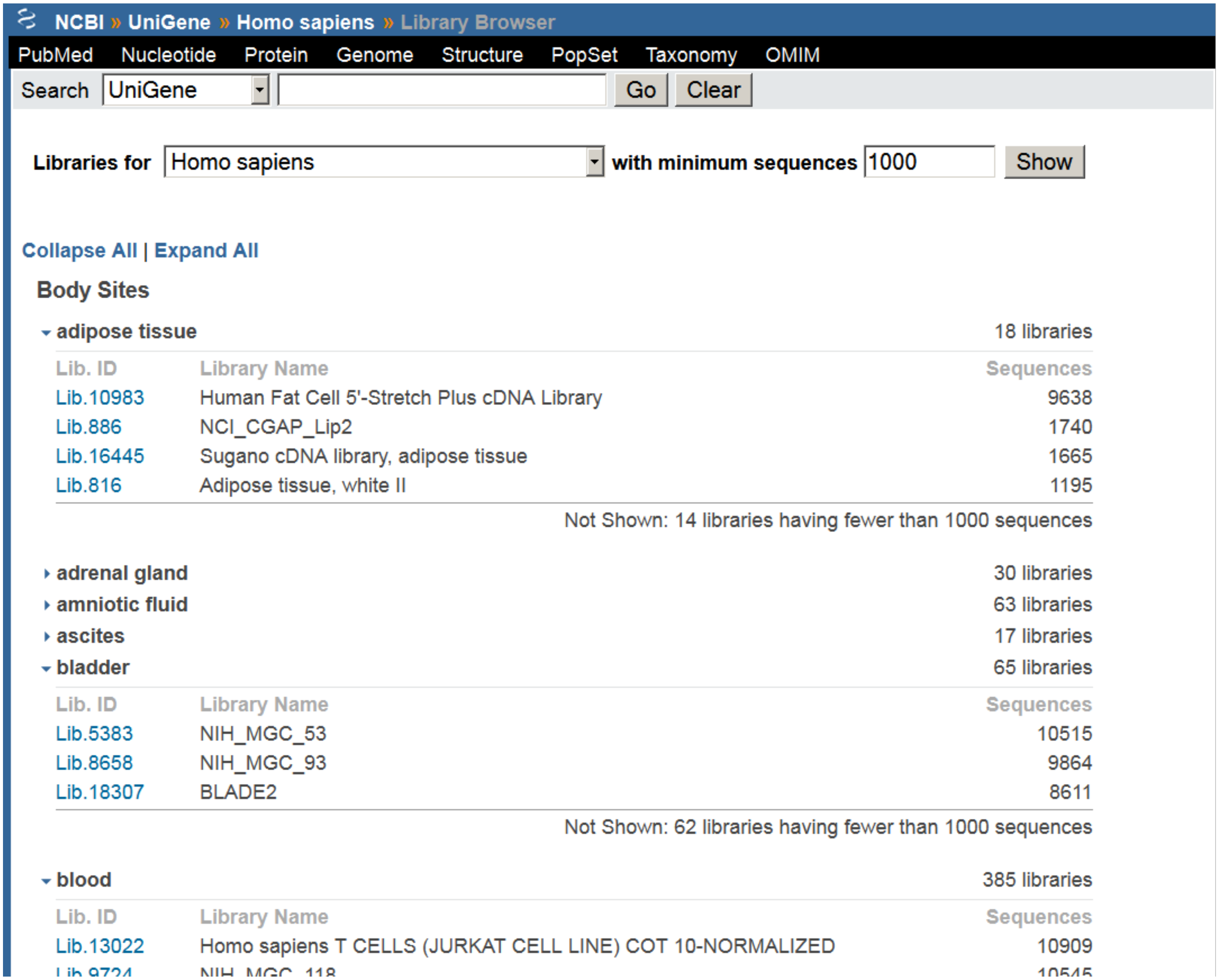
New Pool

**Differential Display Results**

The following genes (UniGene entries) display statistically significant differences in EST counts by the Fisher Exact Test.

| A<br>Muscle | B<br>brain | UniGene Entry | |
|---|---|---|---|
| 0.0150<br>● | 0.0003<br>● | Hs.741179 | Serine hydroxymethyltransferase 2 (mitochondrial) (SHMT2) |
| 0.0108<br>● | 0.0000 | Hs.1288 | Actin, alpha 1, skeletal muscle (ACTA1) |
| 0.0000 | 0.0106<br>● | Hs.506357 | Family with sequence similarity 107, member A (FAM107A) |
| 0.0000 | 0.0083<br>● | Hs.155247 | Aldolase C, fructose-bisphosphate (ALDOC) |
| 0.0000 | 0.0076<br>● | Hs.551713 | Myelin basic protein (MBP) |
| 0.0089<br>● | 0.0015<br>● | Hs.713764 | Actin, gamma 1 (ACTG1) |
| 0.0000 | 0.0065<br>● | Hs.514227 | Glial fibrillary acidic protein (GFAP) |
| 0.0094 | 0.0030 | Hs.535192 | Eukaryotic translation elongation factor 1 alpha 1 (EEF1A1) |

**Figure 5.** Differential expression assessment comparing libraries from muscle and from brain.



**Libraries for** Homo sapiens **with minimum sequences** 1000

### Body Sites

**adipose tissue** — 18 libraries

| Lib. ID | Library Name | Sequences |
|---|---|---|
| Lib.10983 | Human Fat Cell 5'-Stretch Plus cDNA Library | 9638 |
| Lib.886 | NCI_CGAP_Lip2 | 1740 |
| Lib.16445 | Sugano cDNA library, adipose tissue | 1665 |
| Lib.816 | Adipose tissue, white II | 1195 |

Not Shown: 14 libraries having fewer than 1000 sequences

**adrenal gland** — 30 libraries

**amniotic fluid** — 63 libraries

**ascites** — 17 libraries

**bladder** — 65 libraries

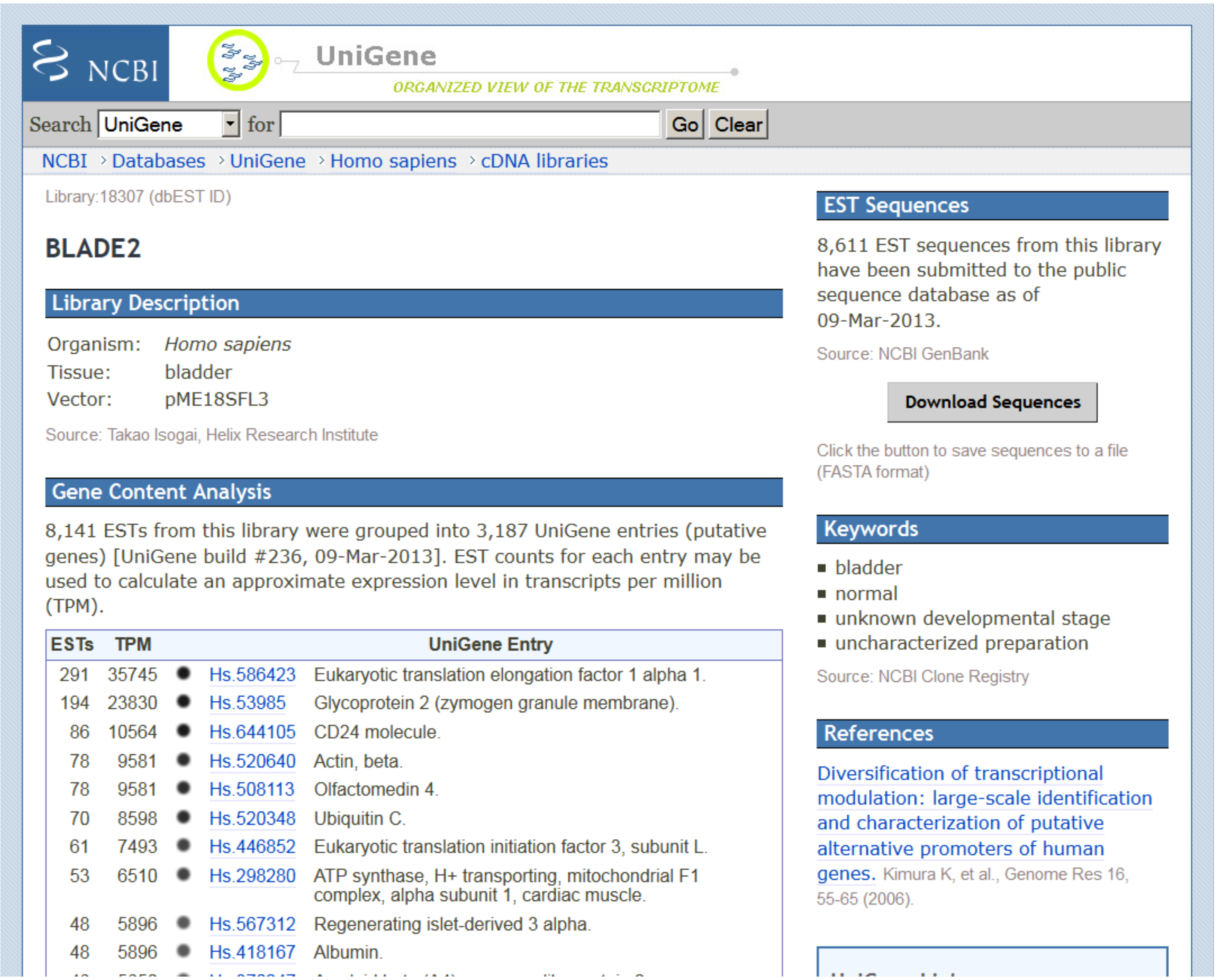
| Lib. ID | Library Name | Sequences |
|---|---|---|
| Lib.5383 | NIH_MGC_53 | 10515 |
| Lib.8658 | NIH_MGC_93 | 9864 |
| Lib.18307 | BLADE2 | 8611 |

Not Shown: 62 libraries having fewer than 1000 sequences

**blood** — 385 libraries

| Lib. ID | Library Name | Sequences |
|---|---|---|
| Lib.13022 | Homo sapiens T CELLS (JURKAT CELL LINE) COT 10-NORMALIZED | 10909 |
| Lib.9724 | NIH_MGC_118 | 10545 |

**Figure 6.** UniGene library browser.


NCBI
UniGene
ORGANIZED VIEW OF THE TRANSCRIPTOME
Search UniGene for
Go
Clear
NCBI › Databases › UniGene › Homo sapiens › cDNA libraries

Library:18307 (dbEST ID)

BLADE2

Library Description

| | |
|---|---|
| Organism: | *Homo sapiens* |
| Tissue: | bladder |
| Vector: | pME18SFL3 |

Source: Takao Isogai, Helix Research Institute

Gene Content Analysis

8,141 ESTs from this library were grouped into 3,187 UniGene entries (putative genes) [UniGene build #236, 09-Mar-2013]. EST counts for each entry may be used to calculate an approximate expression level in transcripts per million (TPM).

| ESTs | TPM | | UniGene Entry | |
|---|---|---|---|---|
| 291 | 35745 | ● | Hs.586423 | Eukaryotic translation elongation factor 1 alpha 1. |
| 194 | 23830 | ● | Hs.53985 | Glycoprotein 2 (zymogen granule membrane). |
| 86 | 10564 | ● | Hs.644105 | CD24 molecule. |
| 78 | 9581 | ● | Hs.520640 | Actin, beta. |
| 78 | 9581 | ● | Hs.508113 | Olfactomedin 4. |
| 70 | 8598 | ● | Hs.520348 | Ubiquitin C. |
| 61 | 7493 | ● | Hs.446852 | Eukaryotic translation initiation factor 3, subunit L. |
| 53 | 6510 | ● | Hs.298280 | ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle. |
| 48 | 5896 | ● | Hs.567312 | Regenerating islet-derived 3 alpha. |
| 48 | 5896 | ● | Hs.418167 | Albumin. |

EST Sequences

8,611 EST sequences from this library have been submitted to the public sequence database as of 09-Mar-2013.

Source: NCBI GenBank

Click the button to save sequences to a file (FASTA format)

Keywords

- bladder
- normal
- unknown developmental stage
- uncharacterized preparation

Source: NCBI Clone Registry

References

Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Kimura K, et al., Genome Res 16, 55-65 (2006).

**Figure 7.** UniGene library summary.

# 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. Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, Su XZ. cDNA sequences reveal considerable gene prediction inaccuracy in the Plasmodium falciparum genome. BMC Genomics. 2007;8:255. PubMed PMID: 17662120.
3. Shangguan L, Han J, Kayesh E, Sun X, Zhang C, Pervaiz T, Wen X, Fang J. Evaluation of genome sequencing quality in selected plant species using expressed sequence tags. PLoS One. 2013 Jul 29;8(7) PubMed PMID: 23922843.
4. Head and Neck Annotation Consortium. Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. Cancer Res. 2005 Mar 1;65(5):1693–9. PubMed PMID: 15753364.
5. Zhang YE, Landback P, Vibranovski MD, Long M. Accelerated recruitment of new brain development genes into the human genome. PLoS Biol. 2011 Oct;9(10) PubMed PMID: 22028629.

6. MGC Project Team. The completion of the Mammalian Gene Collection (MGC). Genome Res. 2009 Dec;19(12):2324–33. PubMed PMID: 19767417.