Gene Help

Last Updated: November 4, 2022

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

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

NLM Citation: Gene Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-.

This book contains information on Entrez Gene, a database created and maintained by the National Center for Biotechnology Information (NCBI).

Table of Contents

Gene Help: Integrated Access to Genes of Genomes in the Reference Sequence Collection

Mike Murphy, Garth Brown, Craig Wallin, Tatiana Tatusova, Kim Pruitt, Terence Murphy, and Donna Maglott

Created: September 13, 2006; Updated: November 4, 2022.

Introduction

Gene supplies gene-specific connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. Unique identifiers are assigned to genes with defining sequences, genes with known map positions, and genes inferred from phenotypic information. These gene identifiers are used throughout NCBI's databases and tracked through updates of annotation. Gene includes genomes represented by NCBI Reference Sequences (or RefSeqs) and is integrated for indexing and query and retrieval from NCBI's Entrez and E-Utilities systems. Gene comprises sequences from thousands of distinct taxonomic identifiers, ranging from viruses to bacteria to eukaryotes. It represents chromosomes, organelles, plasmids, viruses, transcripts, and millions of proteins.

Quick Starttips

Gene is accessed like any other Entrez database, namely by

- querying on any word,
- restricting the query term to a certain field, or
- applying filters or properties

Here are some representative queries:

Find genes by...	Search text
free text	human muscular dystrophy
partial name and multiple species	transporter[title] AND ("Drosophila melanogaster"[orgn] OR "Mus musculus"[orgn])
chromosome and symbol	(II[chr] OR 2[chr]) AND adh*[sym]
associated sequence accession number	M11313[accn]
gene name (symbol)	BRCA1[sym]
publication (PubMed ID)	11331580[PMID]
Gene Ontology (GO) terms or identifiers	"cell adhesion"[GO] 10030[GO]
genes with variants of medical interest	gene_clinvar[filter]

Table continued from previous page.

chromosome and species	Y[CHR] AND human[ORGN]
Enzyme Commission (EC) numbers	1.9.3.1[EC]

When you look at the URLs that underlie these links, you will see that they are constructed by combining 'http://www.ncbi.nlm.nih.gov/gene/?term=' with a query term qualified by field names (in square brackets).

How Data Are Maintained

New Records

Records are added to Gene if any of the following conditions is met:

- A RefSeq is created for a completely sequenced genome and that record contains annotated genes. In the case of prokaryotes, only reference genomes and representative genomes from well-sampled species are currently added to Gene. In the case of RNA viruses with polyprotein precursors, annotated proteins may be treated as equivalent to a "gene".
- A recognized genome-specific database provides information about genes (preferably with defining sequence) or mapped phenotypes.
- The NCBI Genome Annotation Pipeline reports model genes.
- A model organism is scheduled for sequencing, and representative sequences are identified to characterize known genes.

The minimum set of data necessary for a gene record, therefore, is: a unique identifier, or GeneID, assigned by NCBI; a preferred symbol; and either defining sequence information, map information, or official nomenclature from an authority list.

Gene records are not created for genomes which are incompletely represented by whole genome shotgun (WGS) assemblies. In terms of RefSeqs accessions, this means that genes annotated on accessions of the pattern NZ_ABCD12345678 are not submitted to Gene. Although not all existing records have been removed, loci defined by repetitive elements, endogenous retroviruses not named by nomenclature authorities, and loci identified by single transcripts with no other supporting data also are not in scope for Gene.

Numbering system

A unique GeneID is assigned to each new record. There are currently two number generators being used by Gene; one that is assigning values in the range of 7,000,000 – 99,999,999 and another that is assigning values > 100,000,000. Thus the sequence of GeneIDs is expected to have gaps.

Updates

Records are updated when new information is received. For some genomes, this may occur when a genome is re-annotated and the corresponding RefSeqs are updated. For other genomes, this may occur when any information attached to a single gene record is altered. Updates are processed daily.

Some components of the Gene record are updated automatically from other resources. Table 1 summarizes these data elements, their sources, and the update frequency. For example, GeneRIFs are processed independently of the Gene record. Most GeneRIFs are provided by the staff of the National Library of Medicine's Index Section

and are integrated weekly. Those are available with the first update to Gene of the week. Public users are also invited to submit GeneRIFs, via the 'New GeneRIF' link in the Bibliography section of a Gene report.

When any change is made to a record, the modification date is changed. This includes changes in GeneRIFs. The modification date, therefore, is the later of any update to Gene or supplemental information.

About two days are required for an update to be reflected in all reports from Gene. In some cases, the full report may be more up-to-date than the ftp site because the ftp files are regenerated after a re-index of the database, a process that may lag a day behind the update to the database itself. The last modification date is available in the ftp files.

Table 1. Data sources for Gene.

Data category	Source	Species	Update frequency
Official nomenclature	HUGO Gene Nomenclature Committee (HGNC)	Human	Daily
	Mouse Gene Nomenclature Committee (MGNC)	Mouse	Daily
	Rat Gene Nomenclature Committee	Rat	Bimonthly
	Zebrafish Nomenclature Committee (ZNC)	Zebrafish	Weekly
	Chicken Gene Nomenclature Consortium (CGNC)	Chicken	Periodically
	FlyBase	*D. melanogaster*	Data release
	MaizeGDB	*Zea mays*	Periodically
	SGD	*S. cerevisiae*	Data release
	Xenbase	*X. tropicalis, X. laevis*	Weekly
	The Arabidopsis Information Resource	Arabidopsis	Data release
	WormBase	*C.elegans*	Data release
GeneRIF	Index Section, NLM/public	All	Weekly/Daily
	HuGE Navigator	Human	Bimonthly
GO terms	Gene Ontology	Several	Weekly
Disease Names	OMIM	Human	Daily
Interactions	BIND/BOND	Several	Static
	BioGRID	Several	Monthly
	EcoCyc	*E. coli*	Data release
	HPRD	Human	Static
REACTOME	REACTOME	Several	Data release

Suppressed Records

Gene will suppress a record for several reasons:

- Review by NCBI staff and/or collaborators indicates that a record is no longer supported or in scope for Gene. An explanation for the suppression is provided by RefSeq staff.
- Review by NCBI staff and/or collaborators indicates that the original record defined only part of what is now understood to be the functional gene unit. In that event, one record is made secondary to another, and the URL to the current record is provided.

- The molecular basis for a Gene record that was previously only a mapped phenotype is discovered, and there was already a record for the causative locus or loci. The record for the mapped phenotype is made secondary to one of the causative loci and added to the phenotype section of all.

By default, all records, *i.e.*, current and suppressed, are retrieved by a query submitted with no restrictions. You can, however, restrict your results to current records. For example,

- click on Current only from the list filters in the Results filter sidebar at the left of your results display, or
- qualify your query with the phrase "AND alive[property]"

Query Tips provides additional details.

How content is selected

The content of a Gene record depends on availability of information and curatorial decisions. If you have suggestions about types of information that should be included in general, or for a specific record, please let us know by using our update form. More details about maintenance of certain types of information are provided in Gene's FAQ.

How Data Are Displayed (Display Settings/Format)

NCBI's Entrez system supports multiple display options for each of its databases. The options available can be browsed by clicking on *Display Settings* (Figure 1). The options depend on whether you are viewing a set of results, or just one record. When viewing a set of results, in Tabular or Summary format for example, *Display settings* also provides choices for controlling the number of items to display, and their order. Additional customization of display formats and filtering options is possible by configuring your My NCBI preferences.

Gene provides the following categories of formats:

- short reports of gene-specific data (Tabular, Summary, UI List)
- subsets of content for one gene (GeneTable, GeneRIF)
- comprehensive reports for one gene (Full Report, ASN.1, XML)

Short Reports

- Tabular
- Summary
- UI List

Tabular

When you process a query, the results are displayed by default in the Tabular format (Figure 2). You can see that this is the Tabular format by noting the word Tabular at the top of the results section to the right of *Display Settings*.

In the Tabular format, a check box is provided at the left of each record. The check box enables you to select which of the records in the retrieval set that you want to review in another format, according to your selection in *Display Settings*. If none is checked, all are displayed in the selected format.

The Tabular format includes the preferred gene symbol and unique identifier (the GeneID), the Description (including the complete gene name and species), the location on a genomic RefSeq for the reference assembly chromosome (if known), RefSeq accession.version, coordinates, orientation), other symbols and names, and for human only, the Mendelian Inheritance in Man (MIM) number for the gene. If a gene is not annotated on the

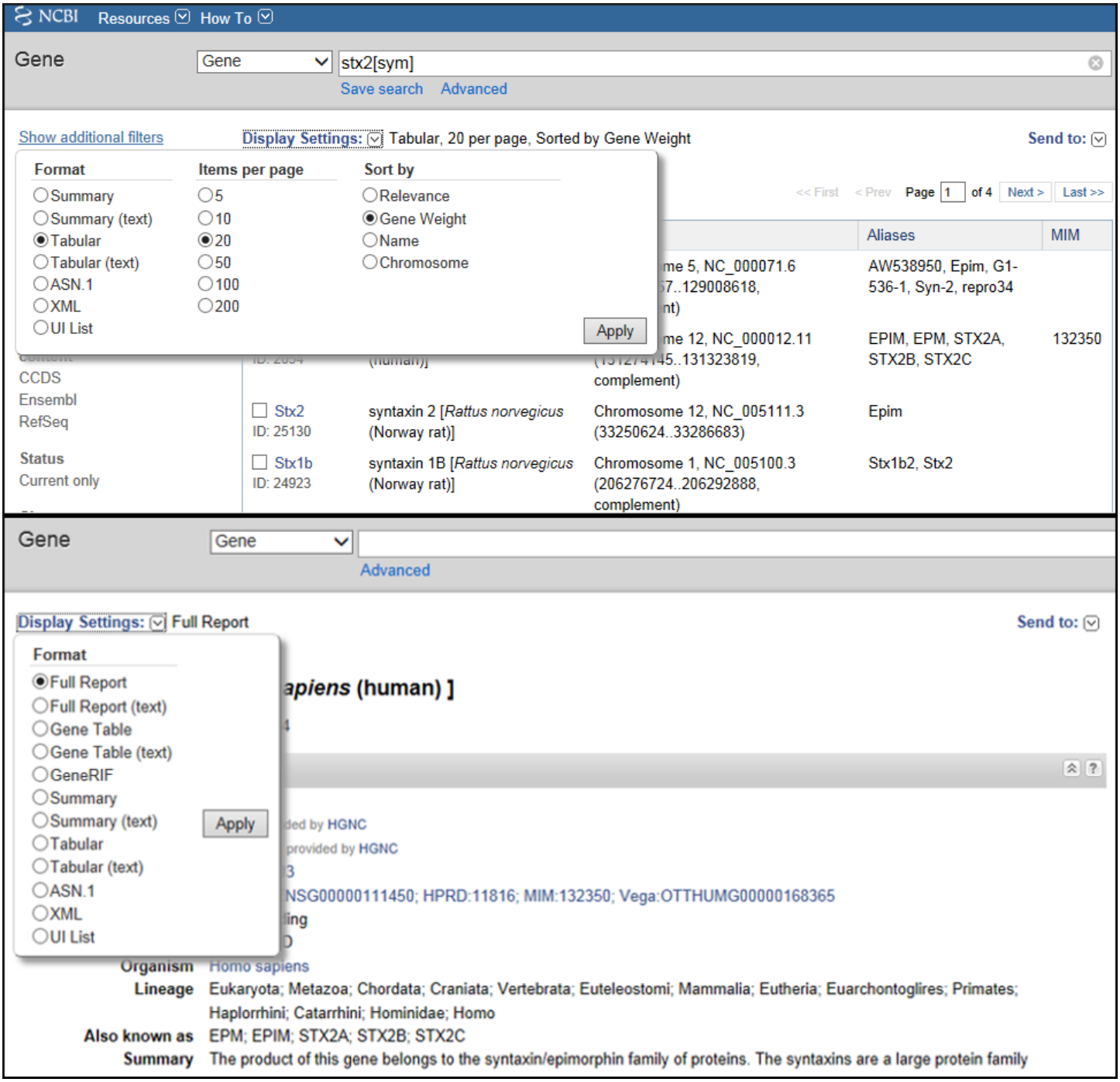

Figure 1. Display Settings.There are two types of options to configure your display: from a query result (top) or from a record-specific display (bottom). The latter has no need to offer controls on the number of records to display or how to sort them.

reference assembly, a location is not reported. Also note that if a gene is on a named plasmid, then the plasmid name is given as the location.

The order of preference for displaying a symbol as preferred is:

- Official symbol
- Locus tag
- First symbol in the set of aliases

The Tabular display is also available as tab-delimited text that includes additional columns in a more parsable format. Gene aliases and other designations are also provided. In *Display Settings*, select the Tabular (text) format. There is an upper limit of 200 records that can be returned by this mechanism; to download a complete result set in Tabular(text) format, use the **Send to:** option at the upper right, select File, and Format Tabular (text).

Summary

The functions allowed from the Summary format (also known as the 'docsum') are similar to those described for the Tabular format. The Summary format contains similar information as the Tabular(text) format, including gene aliases and other designations.

UI List

This option displays only the unique identifiers (UIs) or GeneIDS for the records retrieved by your query (without the functions supported by Entrez).

Sort by

When the Tabular or Summary options are selected, the *Display Settings* menu also allows you to reorder the results. The options are:

- **Relevance** (the current default). Relevance is calculated from Gene's assessment of what fields are the most important by which to find search results. For example, Gene assigns more value to results that match a term in the 'Gene Name' (symbol) field compared to a match in free text such as the RefSeq or GeneRIF summary. Thus if your query is the single term 'cat', then records with symbols of 'cat' will be sorted ahead of records with the term cat only elsewhere in the record.
- **Gene Weight**. Gene Weight is calculated from multiple lines of evidence geared toward evaluating how well a gene has been characterized. These lines include:
    1. Informative Gene-PubMed links. Informativeness is inversely proportional to the number of Gene records connected to a PubMed record.
    2. Informative symbols or full names. A gene with a symbol constructed as LOC+GeneID is weighted less, for example, than a gene with the symbol 'ABCA1'. A gene with a description that starts with the word 'hypothetical' is weighted less than one with a description that starts with 'cystic fibrosis'.
    3. Inclusion in HomoloGene or Protein Clusters. Genes (or their products) that are known to be conserved are weighted more highly.
    4. Inclusion in OMIM or Books.
- **Name**. Results are sorted alphabetically (case insensitive) by the symbol of the gene.
- **Chromosome**. Results are sorted in the following order:
    1. Alphabetically by organism name
    2. Numerically by chromosome
    3. Numerically by the start position on the chromosome

For example, suppose that the search results include genes for *Homo sapiens* (human) and *Mus musculus* (mouse). The human genes will all appear before those for mouse. Within the set of human genes in the results, those that are placed on chromosome 1 will appear first, followed by those placed on chromosome 2, and so on. Finally, within a chromosome, genes will be sorted according to their start positions on the chromosome. Genes that are not placed on a chromosome will appear at the end of the results. Genes that are placed on multiple chromosomes will be sorted according to the first such chromosome.

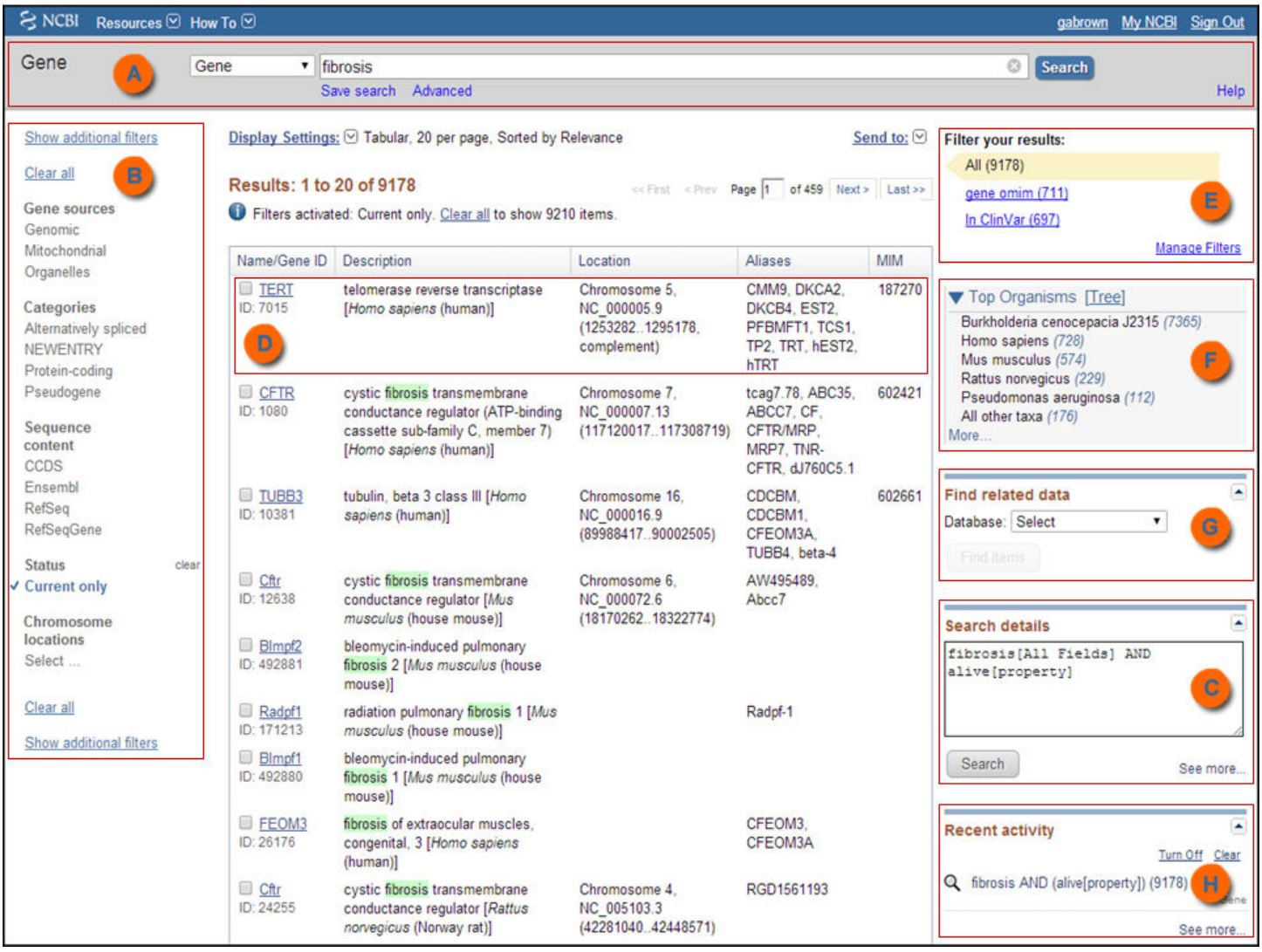

Figure 2. A representative Tabular report from Gene sorted by Relevance. **(A)** The display is the result of a query for records with the word **fibrosis**, subsequently filtered using the *Current only* Results filter sidebar **(B)** on the left of the results table to exclude records that have been discontinued or merged (compare the text in the query bar in the grey section at the top of the display to the text in the Search details box **(C)** at the lower right after selection of the filter). This figure is part of the display generated when the user gabrown was logged into My NCBI; this user selected the text display option for Links and a green highlight for words in the display matching the query. The links to Save search and Advanced below the query bar are common to other Entrez databases. The specific implementation of Advanced search in Gene is provided here. **(D)** A row of the default Tabular display. To see one of the Gene entries, click on the symbol in the Name/GeneID column. Each row of the Tabular display includes the preferred gene symbol and unique identifier (the GeneID), the Description (including the complete gene name and species), the location on a genomic RefSeq for the reference assembly (chromosome (if known), RefSeq accession.version, coordinates, orientation), other symbols and names, and for human only, the Mendelian Inheritance in Man (MIM) number for the gene. The column at the right has additional sections: **(E)** More options to filter your results, controlled via settings in you're My NCBI account. **(F)** Use the linked numbers in the Top Organisms list to select records from your search results for a specific organism. Press Tree to expand the list and display taxonomic relationships. **(G)** Look for data related to your query results in other NCBI databases. Make a selection, and press Find items. **(C)** Search details indicates your most recent search. You can edit the query in the box, and press Search to do a new search, or click on See more for even more details. **(H)** Recent activity allows you to see your recent queries or visited pages. This can be quite useful to open a display you recently viewed. For a complete listing of hints on how to use Entrez interfaces effectively, see the Entrez Help documentation.

Subset of data content

Gene Table

The Gene Table display represents the gene structure as annotated on the indicated genomic RefSeq. The default report is based on the reference assembly, but the selection menu in the top box (Figure 3) allows you to generate reports from other RefSeq genomic sequences.

The report provides information about the intron/exon organization of each transcript, and, if an mRNA, the region of each exon that contains coding sequence. It does this in two ways:

- graphically, by repeating the display included in the Full Report
- in a table, by reporting the position of any exon or coding region, and reporting the length of exons, coding regions, and introns

The Gene Table display supports retrieval of gene-related sequence, as summarized in Table 2.

Please note that Gene Table is not supported when the gene has not (yet) been annotated on any of NCBI's Genomic RefSeqs.

The sequence being retrieved is from the indicated genomic sequence, not the RNA. This means that the length of any non-aligning nucleotides, including a poly(A) tail or vector sequence, is not included in the GeneTable report.

Unaligned tails can be displayed graphically in the Sequence Viewer; follow the Open Full View link on Gene's Full Report, click Configure on the right side of the Graphical Panel, and add RefSeq Alignments from the Alignments Track tab. Unaligned tails are displayed as boxes with the number of aligned bases shown above. Note that RefSeq transcripts with perfect alignments (excluding poly(A) tail) are NOT displayed in the RefSeq Alignment track. More information on how features are rendered in the Sequence Viewer is available from the Graphical Panel Legend section of the Sequence Viewer Help document.

When following a link from GeneTable to the sequence-specific nucleotide or protein record, use the Display Settings options there to generate the format you prefer (*e.g.* GenBank).

Because Gene Table reflects the annotation on the current genomic sequence, for bulk access you may prefer to use one of the General Feature Format (GFF, version 3) files in the species-specific GFF subdirectory. For example:

- Human
- Mouse

Please note that RefSeq may update annotation on sequences representing a genome less frequently than updates to gene-specific RefSeqs. This means that if the version of a RefSeq RNA has changed, or if the number of transcript variants has changed, the GeneTable display will be out of date with respect to the Reference Sequences section of the full Gene report. Please check also the Reference Sequences section of the Gene record to determine whether updates have occurred (new versions and/or more variants and/or suppression resulting from review).

Please see Table 2 for a summary of how to access gene-specific sequence information via Gene.

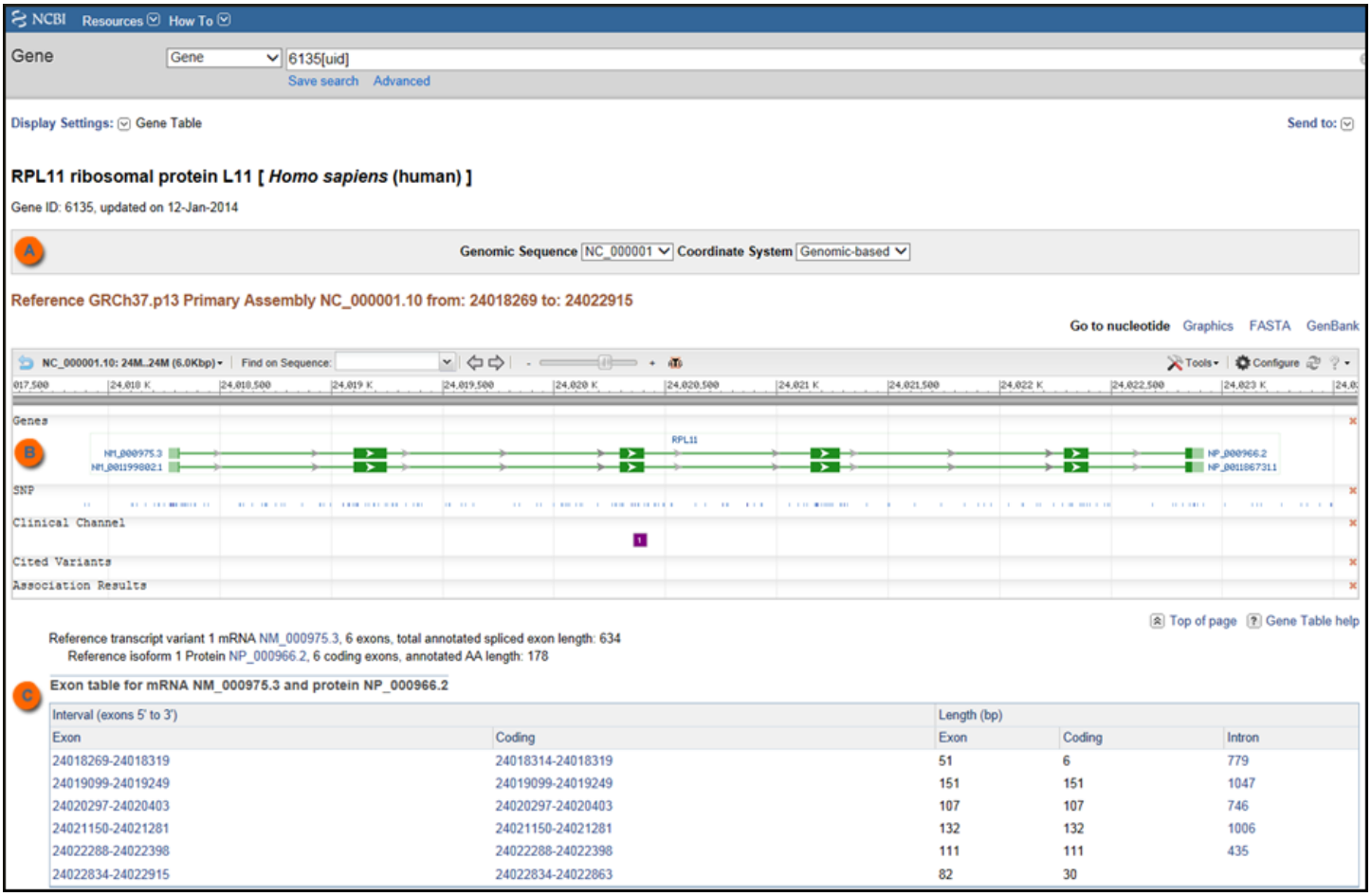

Figure 3. Gene Table display. Use the Genomic Sequence and Coordinate System options in section **A** to select the sequence and the numbering system you want to use to generate your report. The embedded graphic **(B)** allows you to see what other elements are in the range of the gene and either navigate within the page or in the full nucleotide graphics display. The table section **(C)** will be repeated for each transcript variant annotated on the selected sequence (not shown here for NM_001199802.1).

Table 2. Access to Gene-specific sequence information from Gene.

Scope	Link to use
Complete gene feature (genomic sequence)	From the graphic display, hover over the bar indicating the position of the gene. Click on the (i) icon at the top right of the tool tip, and scroll down. Links are provided to the sequence of the gene.
Complete RNA	The RefSeq RNAs link in the Related information menu at the right of the GeneTable and Full Report display options take you to the Nucleotide database where you can retrieve and analyze each RNA.
Complete protein	The RefSeq Proteins link in the Related information menu at the right of the GeneTable and Full Report display options take you to the Protein database where you can retrieve and analyze each protein (*e.g.*, use BLink to compare any sequence to others in the Protein sequence database).
Single intron, exon, or coding sequence (CDS)	Display in FASTA format from the range shown in the corresponding table column. In the Nucleotide database, Display settings can be used to reformat the display.

GeneRIF

The GeneRIF display for a Gene can also be accessed by a URL constructed as:

http://www.ncbi.nlm.nih.gov/gene/GENEID/?report=generif, where a GeneID replaces GENEID.

Example for GeneID 1059: http://www.ncbi.nlm.nih.gov/gene/1059/?report=generif

This display lists the text of the GeneRIF (which anchors a link to PubMed), the title of the paper, and the authors.

The PubMed (GeneRIF) display provides a listing of all the PubMed uids that are associated with GeneRIFs AND interaction data for a GeneID. Thus the count of GeneRIFs displayed for a gene may differ from the number of results in PubMed when the *PubMed (GeneRIF)* link is used.

Full Reports

All of the content that Gene provides is defined by the ASN.1 file. The Full Report display is of the HTML transformation of that ASN.1 and includes navigation tools (Table of contents and Related information), discovery elements, diagrams, and text. Some gene-specific information is **not** maintained in Gene but is maintained in more specialized databases such as GEO and HomoloGene. Access to the additional information maintained in other resources within NCBI or external to NCBI is provided by the listings under Related information (on the right beneath the Table of contents) and by other HTML anchors within the page.

The Full Report display is divided into the gray Search bar (explained in Query tips), navigation and discovery functions at the right, and content elements divided by horizontal separators that display or hide that subsection.

- Navigation/Discovery column
  - Table of contents
  - Genome Browsers
  - Related information
  - Links to other resources
  - General information
  - Related sites
  - Feedback
  - Subscription
  - Recent activity
- Content elements
  - Title
  - Summary
  - Genomic context
  - Genomic regions, transcripts, and products
  - Expression
  - Bibliography
  - Phenotypes
  - Variation
  - HIV-1 Interactions
  - Interactions
  - General gene information
  - General protein information
  - NCBI Reference Sequences (RefSeq)
  - Related sequences
  - Additional links
  - Gene LinkOut

Navigation/Discovery column

The menu at the right of the Gene report supports navigation to multiple sites of interest. Each submenu can be expanded and compressed by clicking on the down ( ) or up ( ) arrows, respectively.

Table of contents

lists the subcategories (or content elements) of information available for a gene. Clicking on a subcategory name takes you to that portion of the gene record.

Genome Browsers

provides links to NCBI and non-NCBI genome browsers, such as Genome Data Viewer, Map Viewer, Variation Viewer, 1000 Genomes Browser, the Ensembl browser, and the UCSC browser.

Related information

indicates other Entrez databases (or report types) that reference Gene. Each line anchors a link to gene-specific data in those databases/reports (Figure 4).

Links to other resources

names sites or files outside of the Entrez system that contain information specific to a gene record. When the identifier for that database needs to be displayed, the link may be repeated in other sections of the full report, such as the Summary or Additional Links sections.

General information

enumerates resources that may help you find and understand the information in Gene. The Help link goes to the default help document. The default help document is also accessed by the question marks ( ) in the horizontal section separators.

Related sites

provides links to home pages of a subset of Entrez databases likely of interest to users of Gene.

Feedback

enumerates several sites where you can comment on or add data to Gene and/or RefSeq.

Subscription

provides a link where you can subscribe to a mailing list to receive announcements about updates to RefSeq.

Recent activity

displays your recent database searches and document views. You can click on any to return to the results of that query or that document.

Content elements

Each content element is divided by a horizontal separator. The arrow at the left end of the separator allows you to open ( ) or close ( ) the display of that section. The arrows pointing up at the right end of the separator ( ) will return you to the top of the page should you want to make a different selection from the Table of contents. A link to this Help document is also provided ( ).

Figure 4. Representative Related information section. The names of these links indicate both the name of the target NCBI database and, in many cases, a subset of records or displays at that target. Details about some of these links are provided in this section. Complete documentation of gene-specific links is provided here.

Title

The section immediately below Display Settings/Send to: (Figure 5) provides the preferred symbol and descriptive name in bold font, followed by the italicized binomial in brackets. If there is a recognized authority for the gene nomenclature of a species, then that authority is the source for these values.

The second line of this section contains the NCBI GeneID and the last date a record was changed. The date is in the format day-month-year. Change is defined as any modification to the content of the record, including

ancillary changes such as the URL for a displayed link. If a record was merged or discontinued, that information is provided also.

Summary

The section (Figure 5) may include several categories of information, namely:

Official Symbol: and Name: Nomenclature provided by the named external authority.

Primary source: Identifier and link to the major resource outside of NCBI that provided information about this gene. For some taxa, this resource may be the nomenclature authority; in other taxa it may be the group that defines genes and submits annotation to public sequence databases.

Locus tag corresponds to the systematic feature qualifier used by the international sequence collaboration (INSDC, DDBJ/EMBL/GenBank) and can be assigned by sequence submitters as a unique, systematic gene descriptor. When such a value is not available from submitted sequence, the identifier from a collaborating model organism database is used. Locus tag is often used to anchor a link to a database other than Gene. Locus tag may also be used as the preferred symbol if an official symbol has not been identified for a gene.

See related: A listing of other identifiers for this gene, provided as database name/value pairs.

Gene type: Possible values are tRNA, rRNA, snRNA, scRNA, snoRNA, miscRNA, ncRNA, protein coding, pseudo, other, and unknown. These are indexed as properties of a gene. Descriptions of these gene types are detailed at properties.

Feature type(s): Feature types annotated on RefSeq(s) associated with genes with a biological region Gene type. Annotated INSDC features are listed along with feature classes or controlled vocabularies in the feature_type: feature_class or feature_type: controlled vocabulary, where each INSDC feature_type is listed on a separate line, and multiple feature_class or controlled_vocabulary terms associated with each feature_type are provided in a comma-separated list following the colon, e.g.:

misc_feature: conserved_region

regulatory: TATA box, locus_control_region, promoter, transcriptional_cis_regulatory_region

These are indexed as properties of a gene, and related *feattype* properties are listed in the Properties section below.

RefSeq status: Any of the set of status descriptions defined by RefSeq. The aim is to describe the gene-level curation status for a given locus, defined as the best RefSeq status found on any of the RefSeq records (NM_, NR_, NG_, XM_, XR_ accession records) associated with the gene, ranked in the order: reviewed > validated > provisional > inferred > predicted > model. In particular, note that an individual locus may be represented by both known (NM_, NR_) and model (XM_, XR_) RefSeq records, and the Gene RefSeq status is based on the known RefSeq records. In this case, the models are provided as supplemental information. Further information about RefSeq statuses and record curation is available on the RefSeq site.

Organism: The binomial, and strain when appropriate, with a link to NCBI's Taxonomy resource.

Lineage: Binomial and lineage from the Taxonomy database.

Also known as: Unofficial symbols and descriptions that have been used for this gene and its products. If there is no official symbol, and no locus_tag, the symbol at the top of the display is repeated in this section. These names are integrated from several sources, including model organism databases, annotation on sequence records, and interactive curation from the published literature.

Figure 5. Representative Title and Summary sections of a Full Report.

Annotation information: Information about annotation oddities for a gene on the reference assembly. May be a report from NCBI's genome annotation pipeline or a comment written by a RefSeq curator to explain how a gene is (or is not) represented in NCBI's annotation. Not provided if the RefSeq group does not provide annotation for a genome or if there are no problems in the annotation.

Summary: Descriptive text about the gene, its cellular localization, its function, its expression, and its effect on phenotype. Records with a summary section can be retrieved by use of the property has_summary (Table 3).

Expression: A teaser sentence briefly describing tissue-specific expression of the gene, based on data in the *Expression* section (see below). A gene is considered to be expressed in a particular sample if it is at a level >=5% of the expression seen in the most strongly expressing sample. Please note that for organisms with expression data from multiple projects, the teaser sentence and indexing are only available on data in the primary expression dataset.

Orthologs: Orthologous genes as determined from genome annotation pipeline data. These are also reported in the file gene_orthologs.gz available by FTP.

Table 3. Other properties in Gene (excluding those related to genetype, rnatype, source, srcdb refseq, and feattype).

Property name	Explanation
alive discontinued replaced	The *alive* property is set when the record is current and primary (i.e., not secondary or discontinued). The term *secondary* is applied to any record that has been merged into another. This occurs most often when multiple genes are defined based on incomplete data, and these are later discovered to be parts of the same gene. One gene record then becomes secondary to the other. The *discontinued* property is set when the record is no longer current, and it has not been made secondary to any other gene record. The *replaced* property is set when the record is no longer current because it has been made secondary to another gene record.
annotated gene	A gene that is annotated on RefSeq chromosome or contig accessions.
expression category ubiquitous expression	A gene that is expressed in all samples.
expression category broad expression	A gene that is expressed in >=50% of samples.

Table 3. continued from previous page.

Property name	Explanation
expression category restricted expression	A gene that is expressed in more than 1 and less than 50% of samples.
expression category biased expression	A gene that is expressed in only 1 sample of the primary dataset.
expression category low expression	A gene that is not expressed above 1.0 RPKM in any sample.
generif	A record having one or more GeneRIF annotations attached.
has ccds	A gene that encodes a protein sequence that is a member of the Consensus CDS (CCDS) set. See http://www.ncbi.nlm.nih.gov/projects/CCDS/.
has expression data	A record having associated expression data.
has lrg	A record having an associated gene-specific genomic Locus Reference Genomic (LRG) sequence.
has ortholog	A gene identified by the annotation pipeline as having orthologs
has pseudogene	A record with one or more related pseudogene records.
has refseqgene	A record having an associated gene-specific genomic RefSeq in the RefSeqGene class.
has summary	A record with summary text (gene.summary)
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.
hiv1 protein interactions hiv1 replication interactions hiv1 interactions	Genes with curated HIV-1:human protein interaction data. See http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html. Genes with curated HIV-1:human replication interaction data. Data is from https://southernresearch.org/. Genes with either curated HIV-1:human protein interaction data or curated HIV-1:human replication interaction data.
interaction	A record with data in the Interaction section.
matches ensembl	A record that matches Ensembl annotation based on comparison of mRNA and protein features. Both the mRNA and protein must meet the matching criteria. If there are multiple possible matches, only the best is reported. Matches are determined as follows: For all organisms, for a protein to be identified as a match between RefSeq and Ensembl, there must be at least 80% overlap between the two, and either 60% or more of the splice sites match or there is at most one splice site mismatch. For non-coding transcripts, the RefSeq and Ensembl transcript features must meet a minimum overlap threshold of 50% and either 60% or more of the splice sites match or there is at most one splice site mismatch.
officially named	A record with official nomenclature.
phenotype	A record including a named phenotype.
phenotype only	A record for a mapped phenotypic trait for which the molecular basis in not known.
readthrough readthrough child readthrough parent	A gene that is sometimes transcribed with another gene. The *readthrough* property will find any gene in a readthrough relationship. The other two properties will find only child or parent genes in a readthrough relationship, respectively.

Genomic Context

The Genomic Context section (Figure 6) reports the location of the gene on the chromosome in non-sequence coordinates. The section also provides information on the primary sequence location of the gene, which is the location(s) on the primary assembly of the current reference assembly, or the location(s) on alternate loci or

alternate assemblies if the gene is not annotated on the reference primary assembly. For many genes, including those annotated using NCBI's Eukaryotic Genomic Annotation Pipeline, the sequence location information is provided as a table that includes the Annotation Release number and Assembly name. Note that the Assembly column includes a link, by accession, to NCBI's Assembly database.

To facilitate working with previous assembly versions, the sequence coordinates from the last annotation of the previous assembly version may also be listed. This feature is currently limited to human, where the location on the GRCh37.p13 assembly is provided, but will be expanded to more organisms with future assembly updates. A link to NCBI's Genome Data Viewer resource is provided in the upper right hand corner of this section.

If the gene is included in the current genome annotation, the section also diagrams neighboring genes and indicates their orientations. If the name of a gene is too long to use for a label, it is truncated and marked with an ellipsis (...). The gene being shown on the diagram is in maroon. All other diagrams and labels anchor links to specific Gene pages, supporting quick navigation to review neighboring genes by clicking in the area of the symbol/arrow.

The diagram shows the gene's placement on any and all chromosomes in the current genome annotation. Otherwise, the diagram will show another genomic placement in the current genome annotation in this order of precedence: reference contig; reference genomic region (NG); alternate assembly chromosome; alternate assembly contig. The location information for all current placements will be provided in the ASN.1 of the record and in the Reference Sequences Section. If a gene is not included in the current genome annotation, no diagram is provided.

Genomic Regions, Transcripts, and Products

The Genomic Regions, Transcripts, and Products section (Figure 6) is provided when a gene has been annotated on a genomic RefSeq, in other words, when the intron/exon/coding region information, or the position of a pseudogene, is available in some genomic coordinate system. The display in this section is generated from NCBI's Sequence Viewer, the same software that drives the Graphics sequence display option available from the sequence databases, and provides some of the navigation features. A legend describes how annotated features are rendered in this display, and a link in the top right hand corner of the sequence panel provides complete Help documentation.

Depending on the data that are available, you can add data tracks to the display using the Configure button in the top right hand corner of the graphical panel to:

- view the intron/exon/coding region organization of a gene and its RNA product(s), or the placement of a pseudogene, on a genomic RefSeq
- identify the RefSeqs that correspond to any RNA or protein product and see an overview of the exons they represent
- evaluate expression under different experimental conditions by adding RNA-seq tracks
- view variants in dbSNP, dbVar, or ClinVar
- explore differences between genome assemblies

You may also:

- alter the zoom level of the display (more…)
- hover over a feature to display information about it via a tool tip
- move upstream and downstream of the sequence being displayed (more…)
- navigate to a full display of the genomic context via the link to **Graphics**
- navigate to the genomic sequence of the gene in **FASTA** format
- navigate to the genomic sequence of the gene in **GenBank** format

Figure 6. Genomic context and Genomic regions, transcripts, and products sections of the Full Report display. These sections provide diagrams of the gene and its neighbors, the gene's intron/exon organization, and the RefSeqs that are used to represent RNA products and their coding regions. These sections also anchor links to the sequence in FASTA and GenBank format, the full nucleotide Graphics viewer, and the NCBI Genome Data Viewer. Additional tracks may be displayed based on available data. For example, variant tracks, Ensembl annotation, and RNA-Seq coverage tracks may be displayed for human genes.

- change the display of the genomic sequence on which the gene is annotated. The default display is the chromosome of the reference assembly; for some taxa there are alternate assemblies. For human, the RefSeqGene can also be selected

Each position of a gene product, when represented by a RefSeq RNA and/or protein, is provided relative to the genomic accession on which it is annotated. For some species, including human and other vertebrates, the genomic RefSeqs are updated independently of the annotated product RNAs, with the latter being updated more frequently. This means that several kinds of discrepancies between the diagram and the current RefSeq RNAs may result.

- The diagram may be labeled with an mRNA accession for a predicted model transcript (of the format XM_123456 or XM_123456789), yet display of that accession from Entrez Nucleotide indicates that the accession is no longer primary. That means that a curated mRNA (accession of the format NM_123456 or

NM_123456789) has been generated to replace the previous model accession. The new "NM" accession will be reported in the Reference Sequences section of Gene.

- The diagram may be labeled with curated RNA accession numbers (of the format NM_123456 or NM_123456789 or NR_123456) different from those listed in the Reference Sequences section. This will result if curation after the submission of the annotated genome identified more transcript variants, which therefore are listed only in the Reference Sequence section and not in the diagram. It will also result if curation after submission of the annotated genome identified an error in the annotated product, and the accession for that product was suppressed. In that case, the Genomic regions, transcripts and products section will indicate a transcript not listed in the Reference Sequences section of the Gene report.
- The diagram may be labeled with a curated RNA accession number that represents a previous version of the accession. A version number change (e.g., NM_321321.1 -> NM_321321.2) occurs to a RefSeq record when there is any update to the sequence of that record. Sequence updates include the alteration, addition, or removal of nucleotides or amino acids from a record. Older RefSeq records (NM_321321.1) may be labeled on the diagram but updated RefSeq records (NM_321321.2) will be reported in the Reference Sequences section of Gene. The diagram shows the RefSeq records that were annotated in the last release while the Reference Sequences section shows the current version of the RefSeq records. The diagram is updated upon a new annotation release. Between releases, BLAST2SEQ can be used to determine sequence differences between older and newer RefSeq records.

Changing the zoom level in the display

- **Select and display only a subsequence.** Left click in the white section with the coordinates and ruler, and drag to select your region of interest. Then, right click, select *zoom on range*, and the display will refresh to provide the region of interest.
- **Use the in/out zoom functions.** Right click, and select either zoom in or zoom out. The display will refresh and change the region displayed by a factor of 2.

Move upstream and downstream

- A single left click anywhere in the display other than the ruler section, followed by a drag, results in a shift to display upstream and downstream sequence.

Expression

Expression data for some genes are now being displayed graphically after the Genomic regions, transcripts, and products section. Records with an expression data section can be retrieved by use of the property “has expression data”, while other properties exist to look for different expression patterns (Table 3). The data are computed from RNA-seq alignments compared to the most recent RefSeq gene models on the reference genome, and then normalized by RPKM (Reads Per Kilobase of transcript per Million mapped reads).

Because methodologies differ between RNA-seq projects, the data are binned by specific BioProject to reduce variability. If more than one BioProject is available, they can be accessed separately using the toggle bar at the top of the section.

Many BioProjects include a number of replicates per sample. In these cases, the data are shown per sample and are reported as an average \+/- standard deviation. However, the full data are provided in a table accessible by going to the See details link at the top right corner of the Expression section. The table lists the data by sample, and each sample has a + button that can be used to display the replicate data for each sample.

The data for the single BioProject shown can be downloaded by selecting the Download button in the top right corner of the See details page. The full dataset in XML format is available from the Gene FTP site.

Note: NCBI's staff have chosen studies which sample a range of representative tissues, with internal replicates of each tissue where possible. The expression profiles displayed provide a reliable starting point for assessing the expression of genes in the body. For interpretation of the data, researchers should consider examining several independent relevant studies rather than relying on any single assay. Please see PMID: 26996076 for a discussion of the challenges of making RNA-seq expression data relevant to the clinical setting.

Bibliography

The Bibliography section (Figure 7) may have two components:

A. An embedded display of a subset of PubMed citations.
B. An embedded display of a subset of GeneRIFs.

The approach in both components is to display a limited number of records within the full display (5 for PubMed, 10 for GeneRIF), provide a count of the total records available, and support links to a display of all records. The GeneRIFs component also provides a link to submit a new GeneRIF for the gene, or to submit a request to the RefSeq curators to review information in the record.

What is a GeneRIF?

A GeneRIF is a concise phrase describing a function or functions of a gene, with the PubMed citation supporting that assertion. The majority of GeneRIFs have been provided by a collaboration between the NLM's Index Section and NCBI. There is no constraint on the number of independent submissions of GeneRIFs per PubMed id, although those from non-NLM sources are reviewed by RefSeq staff. The GeneRIF homepage provides more information about the project, including how general users can make submissions. If more than one GeneRIF for a gene has the same text but a different citation, the link to PubMed (icon at the left) will result in a display of all citations.

Each species has a GeneID with the symbol **NEWENTRY**. When staff of the NLM indexer sections cannot identify the gene to which a publication belongs, the GeneRIF is connected to the NEWENTRY, which is a placeholder for all the 'unconnected' GeneRIFs for a species. The GeneRIF text remains associated with the **NEWENTRY** GeneID until a RefSeq curator can identify or create the specific gene or genes to which the submission should be connected.

The full display of GeneRIFs for a gene can be generated at any time by selecting GeneRIF as the format from Display Settings.

Phenotypes

This section reports the effect of the gene on phenotype, especially disease. For human genes (Figure 8), the first row links to the NIH Genetic Testing Registry (GTR), a central location for genetic test information that is submitted voluntarily by test providers. The second row links to the Phenotype-Genotype Integrator, (PheGenI, pronounced FEE-GEE-NEE), a web portal providing a tabular display of genome-wide association study results relating the gene and/or its expression to a phenotype. PheGenI includes links to Genotype-Tissue Expression (GTex) results and viewers to display the relationships among genetic variants at the nucleotide level. Subsequent rows of the Phenotypes section may display the following:

Professional guidelines: As professional practice guidelines, position statements, and recommendations are identified that relate to a disorder, gene, or variation, staff at NCBI connect them to the appropriate records. An alphabetical list of many of these guidelines can be found here: MedGen summary of professional guidelines

You can also identify all conditions associated with guidelines via this URL: http://www.ncbi.nlm.nih.gov/medgen?term="has%20guideline"[Properties]

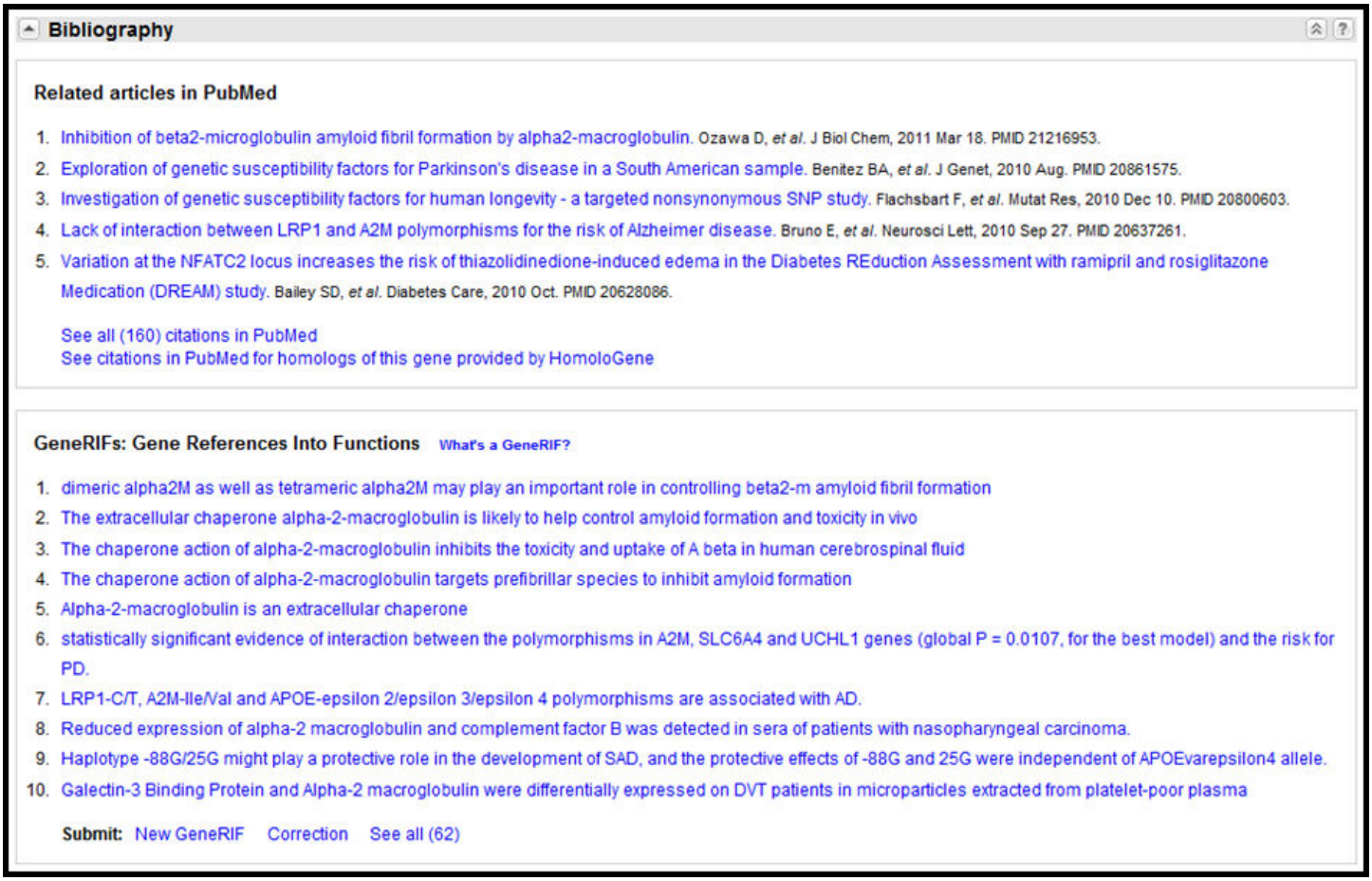

Figure 7. Representative Bibliography section displaying articles in PubMed and GeneRIFs. If the number of citations exceeds 5 (PubMed) or (10) GeneRIFs, the first 5 or 10 are displayed, along with the total count and a link to the display of all records.

Associated conditions: each row of a named phenotype provides links to more information, as available. In the case of human disease, this may include links to MedGen, OMIM, and GeneReviews; a link to the NIH Genetic Testing Registry (GTR) comparing laboratories offering the test may also be provided.

Copy number response: provides evidence of dosage sensitivity (either haploinsufficiency or triplosensitivity) as determined by the ClinGen group (https://www.clinicalgenome.org/).

NHGRI GWAS Catalog: provides a link to the SNP-trait associations reported in the NHGRI Catalog of Genome-Wide Association Studies, and the associated PubMed citation.

Variation

The section is designed to make it easier to navigate to gene-specific reports of sequence variation in NCBI's major variation resources, namely (1) dbSNP for variations of length less than approximately 50 bp, (2) dbVar for longer variations, including complex rearrangements, and (3) ClinVar, for the subset of both types of variation that may have medical relevance. ClinVar is available only for human. For human genes where variation may be related to a condition, and as practice guidelines, position statements, and recommendations are developed, links to Professional guidelines may be provided in the Phenotypes section.

The links that are provided to ClinVar and dbVar are equivalent to the links provided to those resources in the Related information section at the right.

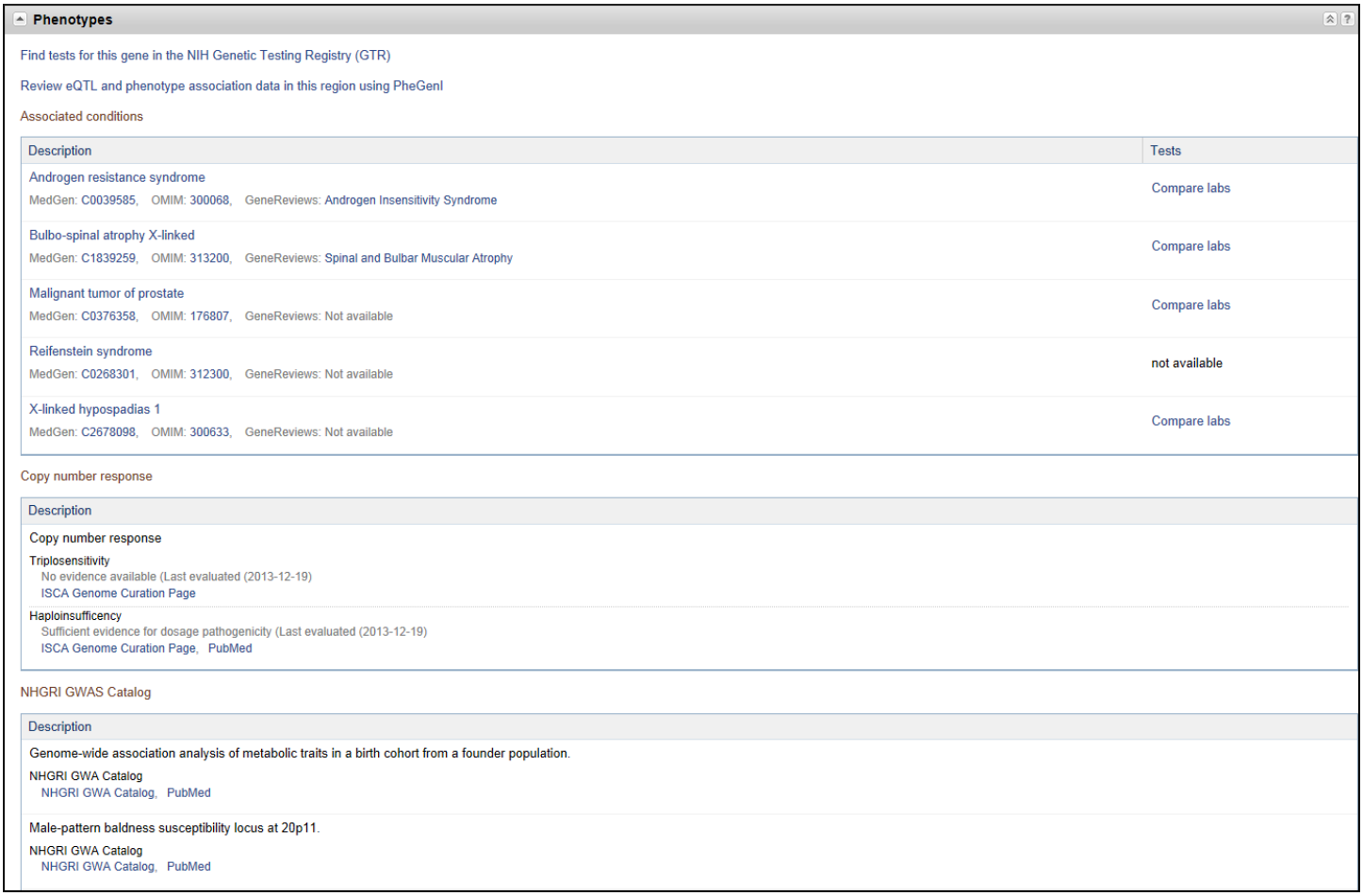

Figure 8. Representative Phenotypes section in the Full Report display. This section reports the effect of a gene on phenotype, particularly disease, when known. For some human diseases, links to the NIH Genetic Testing Registry (GTR) and Phenotype-Genotype Integrator are available at the top of the display.

To view, search, and navigate the human variations in dbSNP, dbVar, and ClinVar in a genomic context, follow the links to 'See Variation Viewer …'. Links to the GRCh38 and GRCh37.p13 assemblies are available.

There are several types of links provided for data in dbSNP:

- See SNP Geneview Report is equivalent to the link named SNP: GeneView in the Related information section. It displays by default only the variants in the coding region (note that cSNP is checked). To see all variations, select 'in gene region' instead. Note that this page also supports downloads.
- See SNP Genotype Report is equivalent to the link named SNP: Genotype in the Related information section. It displays information about populations and submitters of genotype data in the region of gene. An LD plot is also provided.
- See SNP Variation Viewer report is equivalent to the link named SNP: VarView in the Related information section and is available only for human. This display makes it easier to display both medically relevant and all short variations submitted to dbSNP in the region of a gene.

HIV-1 interactions

This subcategory is divided further into **Replication interactions** and **Protein interactions**.

Replication interactions

This section reports human proteins shown to be required for HIV-1 infectivity and replication. The interaction data are provided by the Southern Research Institute (https://southernresearch.org/) based on published whole genome screens that used small interfering RNAs. The data is provided without review by Gene staff; if you identify an incorrect or missing interaction, please contact the external source directly for correction. The display on human records reports:

- a concise description of the interaction
- links to papers in PubMed that support the described interaction

Protein interactions

The HIV-1, Human Protein Interaction Database is funded by the Division of Acquired Immunodeficiency Syndrome (DAIDS) of the National Institute of Allergy and Infectious Diseases (NIAID). As the title indicates, this project focuses on the human proteins that have been shown to interact with proteins from HIV-1. Interaction data is provided solely by the HIV-1, Human Protein Interaction Database without review by Gene staff; if you identify an incorrect or missing interaction, please contact the external source directly for correction. The format of this section is different for the human and HIV-1 gene reports. For human, the display consists of:

- the HIV-1 protein, linked to the sequence record in the Protein database
- the HIV-1 gene, linked to the Gene record for that gene product
- a concise description of the interaction
- links to papers in PubMed that support the described interaction

For HIV-1, the display is subdivided by peptide name and includes:

- a key word categorizing the interaction
- the full name of the human gene, linked to the Gene record
- links to papers in PubMed that support the described interaction

Please note that there are separate reports from this section that are available for download, both from the HIV-1, Human Protein Interaction Database homepage and the GeneRIF subdirectory of the Gene FTP site.

Interactions

The general interactions in this section are provided, without review by Gene staff, by the external sources listed in ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interaction_sources. If you identify an incorrect or missing interaction, please contact the external source directly for correction. Interactions are reported as pairs where the product of the gene that is part of the interaction is given in the first column. If there are more than 25 pairs, pagination is provided. Depending on the type of interaction, the rest of the display may report:

- the other interactant, anchoring a link to more information
- the gene name of the other interactant, anchoring a link to that record in Gene
- the complex to which the interactant(s) belongs
- the source of these data, anchoring a link to the record at that source
- links to papers in PubMed that support the described interaction
- a concise description of the interaction, if available

General Gene Information

This section includes several subcategories of information, including:

Homology: a partial listing of orthologs in other species reported from different sources, including HomoloGene. NCBI's Eukaryotic Genome Annotation Pipeline provides the *Orthologs from Annotation Pipeline* report calculated using a combination of protein sequence similarity and local synteny information. Orthology is determined between a genome being annotated and a reference genome, typically human, and the set of pairwise orthologs tracked as a group and reported here. Links to a comparative display in Genome Data Viewer and to the OrthoDB catalog of eukaryotic orthologs also may be provided.

GeneOntology (GO): Specific GO terms provided by the Gene Ontology Annotation Database and listed by category and term, with evidence information and links to supporting publications. Each GO term supports a link to the AmiGO browser. Abbreviations in the Evidence Code column indicate the level of support for assigning a GO term to a gene. Explanations for these abbreviations are provided by the Gene Ontology website.

Gene does not alter the associations provided by a model organism database, nor does Gene recapitulate the directed acyclic graph structure provided by GO. Thus, Gene does not support retrieval of all genes associated with a specific GO term based on that term's parent. If you identify a GO term that is inappropriate for a gene, please contact the model organism database directly. ftp.ncbi.nlm.nih.gov/gene/DATA/go_process.xml documents the authorities Gene uses to connect GO terms to GeneIDs.

Genotypes: Links to various reports from dbSNP about allele frequencies in one or more populations, all variations for a gene, or disease-associated variations.

Markers: An enumeration of the markers that are related to this gene. The relationship is reported based on direct reports. Links are provided to the NCBI Probe database.

Readthrough: Information about genes that are sometimes transcribed with others. More information about readthrough transcription and how these events are represented in Gene are described in a FAQ.

Related gene/pseudogenes: If a gene, provides a link to view the records of pseudogenes related to the functional gene. If a pseudogene, provides a link to the functional gene.

Related region gene/members: Region records in Gene define officially named loci that are composed of multiple parts or represent clusters of related genes. If the record defines a region, provides a link to all members of the region. If a member of the region, provides a link to the region record.

Relationships: This section reports *some* of the public sequences that were used to support the prediction of the indicated RefSeq model. The report is not comprehensive and is provided only for those genomes for which NCBI calculates annotation, and only for those genes where there is not a supporting curated RefSeq.

The above relationships between two or more genes are reported in the file gene_group.gz available by FTP.

General Protein Information

This section applies only to genes that encode proteins. It reports the name or names that have been assigned to proteins encoded by the gene and provides other descriptive text. The names are as annotated on the RefSeq protein, when that protein is available. The sources of these names include model organism databases, annotation on public sequence databases, and curation by RefSeq staff.

NCBI Reference Sequences (RefSeqs)

This section describes the gene-specific NCBI reference sequences (RefSeqs) that have been established for this gene. In addition to enumerating the accession numbers and providing links to the appropriate Entrez sequence database, this section may also include descriptions of each transcript variant, accession numbers of the public sequences used to support any transcript, links to matching related Ensembl transcripts and proteins, and a

listing of computed domains in an encoded protein. The text provided in this section therefore supports retrieving gene records based on descriptions of conserved domains.

The Reference Sequence group uses several approaches in maintaining information. These can be broadly categorized as:

1. *RefSeqs maintained independently of Annotated Genomes* (Figure 9). RefSeqGene and RefSeq RNA and protein sequences are updated continuously, independently of any comprehensive reannotation of a genome. Because these reference sequences are curated independently of the genome annotation cycle, their versions may not match the RefSeq versions in the current genome build. You can identify updates by comparing versions in this section to versions in the Genomic regions, transcripts, and products section. *GenBank* and *FASTA* and *Sequence Viewer (Graphics)* anchor links to sequence in the given formats
2. *RefSeqs of Annotated Genomes* (Figure 10). This section reports genomic RefSeqs from all assemblies on which this gene is annotated, such as RefSeqs for chromosomes and scaffolds (contigs) from both reference and alternate assemblies. The position and strand of the gene feature is provided (offset 1). *GenBank* and *FASTA* and *Sequence Viewer (Graphics)* anchor links to sequence in the given formats. Model RNAs and proteins are also reported here.
3. *Genome Annotation*. RefSeq RNA and protein sequence are provided only through the process of genome/chromosome annotation.
4. *Suppressed Reference Sequence(s).* Accession numbers listed in this section were suppressed for the cited reason(s). Suppressed RefSeqs do not appear in BLAST databases, related sequence links, or BLAST links (BLink) but may still be retrieved by from the Nucleotide or Protein databases, and by clicking on the hyperlinked accession.version.

Related Sequences

This section has two subsections, one in which the nucleotide sequence is primary and one for protein sequences only (GenPept or UniProtKB). It contains sequence accessions that are related to the gene and provides links to the appropriate sequence record in Entrez Nucleotide, Entrez Protein or UniProtKB. It is not intended to be a comprehensive list of all sequences related to any gene; such sequences can more explicitly be found by using BLAST to query sequence databases or by using pre-calculated reports of related sequences via Entrez Nucleotide, Entrez Protein, or BLink. The sequence accessions in this section are provided in a tab-delimited format in the gene2accession.gz file in the DATA directory of the Gene FTP site.

Depending on the genome of the gene being reported, the sequences included may or may not be restricted to the same subspecies or strain.

Gene purposely lists protein accessions on records represented as not protein-coding. The intent is to make the connection between sequence annotation and Gene's current representation of the type of gene. For example, a nomenclature group may call a gene protein-coding or UniProt may create a sequence record for a protein based on an open reading frame, but RefSeq staff may judge the evidence to be weak based on a lack of cross-species homology or experimental support. Gene will report the protein sequences derived from the locus but will represent the gene as not protein-coding consistent with the RefSeq curation decision. Records of this type are reviewed periodically as new evidence is made available.

Users with evidence indicating that the Gene record should be reviewed are encouraged to contact RefSeq staff.

Accessions are reported as related sequences based on several criteria:

- mRNAs with unique best placement on a genome coinciding with an annotated gene
- cDNA/cDNA sequence relatedness (calculated based on criteria of identity, length of overlap to known accessions, and coverage of the novel accession)

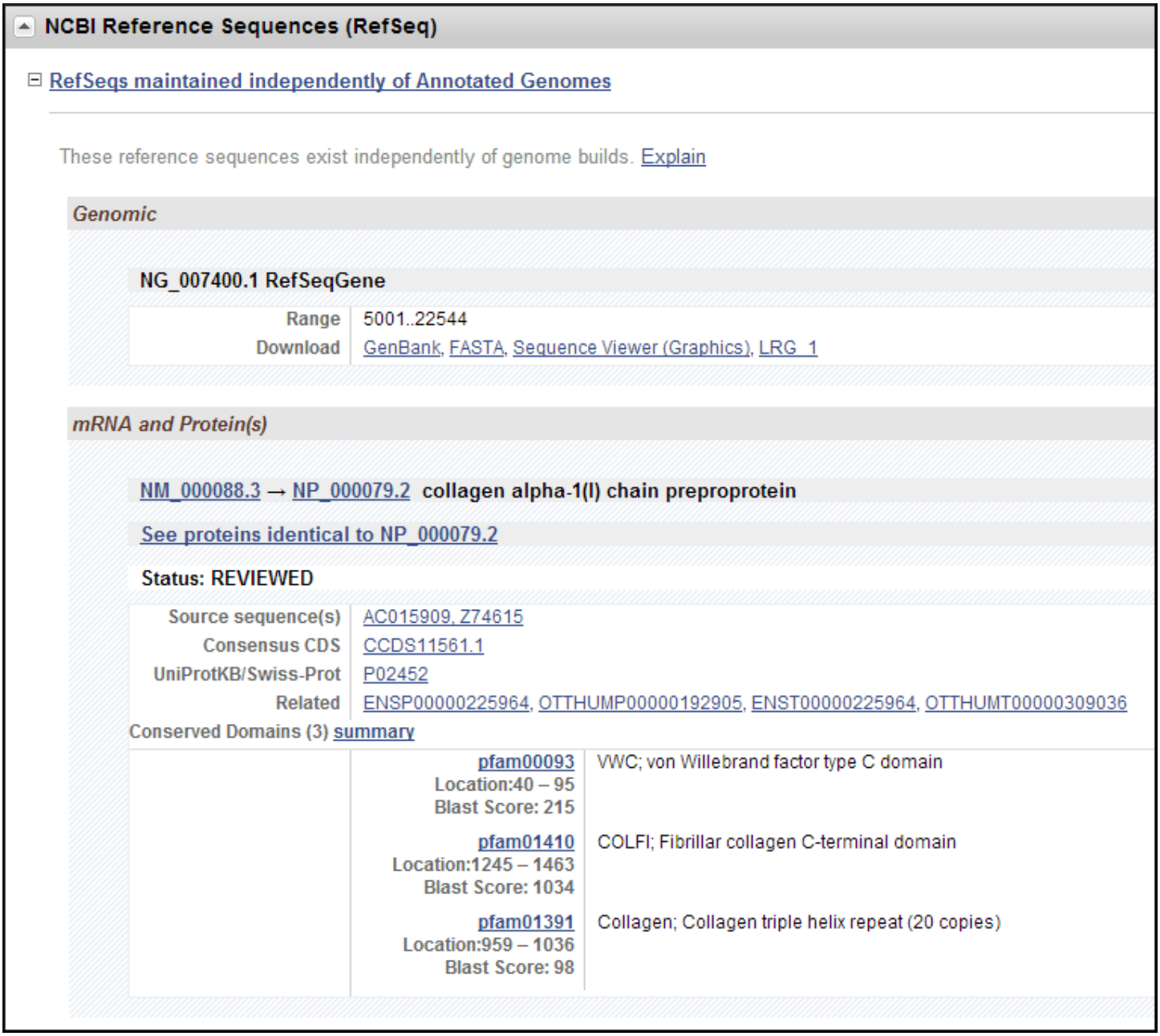

Figure 9. Representative NCBI Reference Sequences (RefSeq) section in the Full Report display. This section includes two subsections: RefSeqs maintained independently of Annotated Genomes (this figure), and RefSeqs of Annotated Genomes (Figure 10). RefSeqs maintained independently of Annotated Genomes includes: The Genomic accession number of the associated RefSeqGene record, when available, and mRNA and Protein(s) accession numbers followed by a Description of the transcript, links to the Source sequence(s) from which any Reference Sequence was derived, links to the Consensus CoDing Sequence database (Consensus CDS), links to related records in UniProtKB and Ensembl, and links to the Conserved Domains database.

- submissions from model organism databases or nomenclature authorities
- identification of proteins with identical sequences
- curation by RefSeq staff
- annotated GeneIDs from the ORFeome Collaboration or Celera

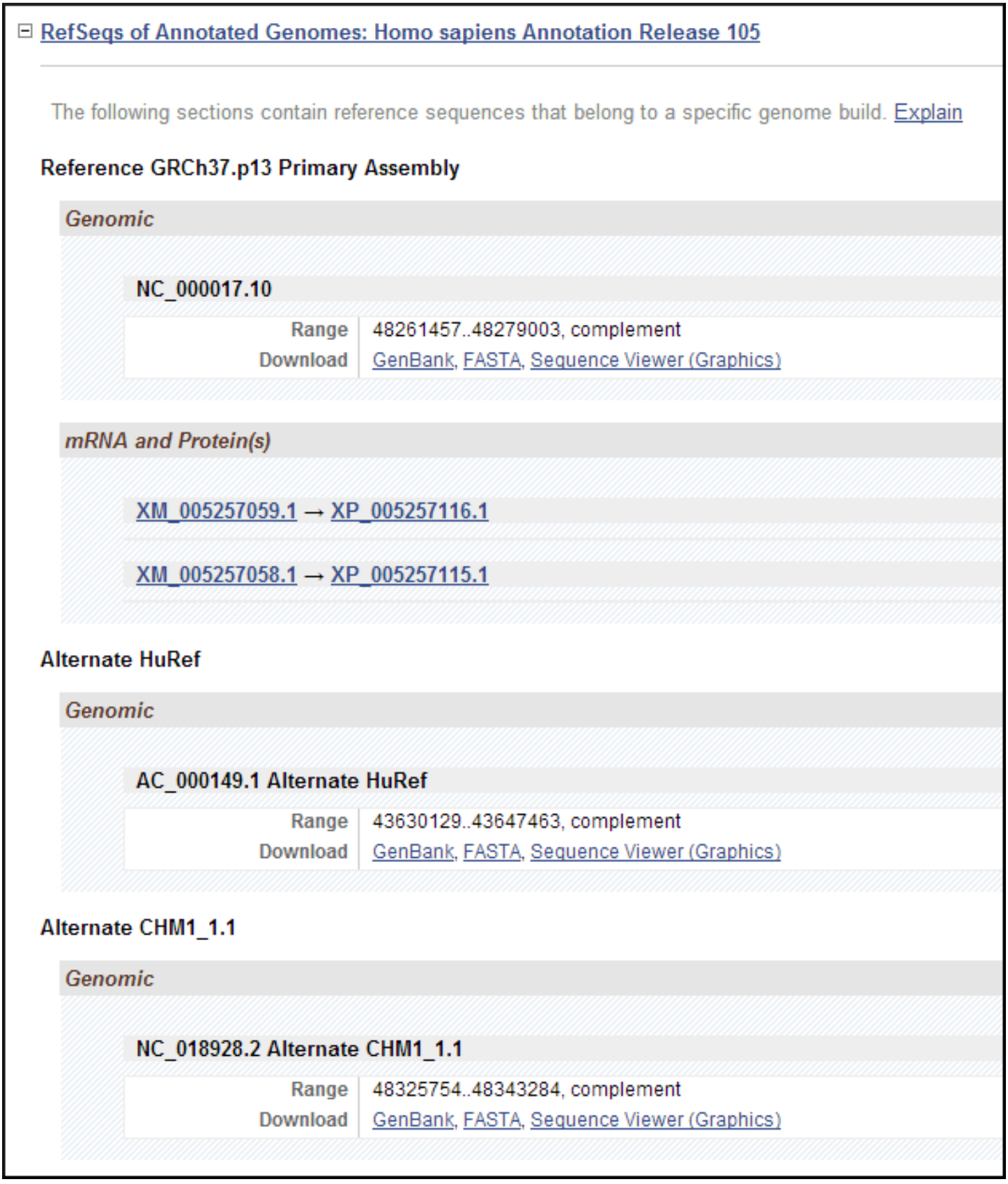

Figure 10. Representative subsection RefSeqs of Annotated Genomes in the NCBI Reference Sequences (RefSeq) section of a Full Report display. This subsection follows RefSeqs maintained independently of Annotated Genomes (Figure 9). It includes the accession numbers of the available Genomic assemblies, including Reference and Alternate assemblies, with links to multiple display formats. In this example, the gene is annotated on the complementary strand. Model RNAs and proteins are also reported.

Additional Links

This section provides a view of a subset of links to information both within and external to NCBI. Some of these links overlap those included in the Related information menu. The intent of this section is to provide a printable report of, for example, MIM numbers, and gene- or gene family-specific websites.

Gene LinkOut

LinkOut provides easy access to relevant online resources outside of the Entrez system. These connections, and their groupings, are maintained by the external database.

ASN.1

The ASN.1 display provides gene records structured according to the Gene specification. An XML transformation of the ASN.1 is also available. Detailed information about the specification is provided in the Tips for Programmers section.

XML

Any record or selected set of records can be displayed in XML format. The XML is generated automatically from the ASN.1 record that is used to support the display, with the names of the tags defined by the ASN.1 specification. Detailed information about the specification is provided in the Tips for Programmers section.

Query Tips: How to submit detailed queries, and more...

Gene uses functions common to other NCBI databases. Most functions of the Entrez indexing and query engine are used by Gene. This section summarizes only how to use the tools in the context of the Gene database. Entrez Help and PubMed Help provide general information on how to save searches, use the Clipboard, history, and Advanced Search. For general information about Entrez, see Entrez Help.

Each Entrez database provides a query bar where you can select a database to interrogate, and enter a search term or terms. If a simple query is not powerful enough, there are options available to construct Advanced search queries.

Advanced search

The Advanced Search Builder (Figure 11) accessed from the query bar or http://www.ncbi.nlm.nih.gov/gene/advanced/ is a powerful resource to construct useful queries and to view terms that have been indexed under any field name. Table 5 describes the fields used in indexing the records and provides some representative queries using those fields. This section describes:

- filters in general and how they can be used to search Gene for records of interest
- properties assigned to Gene records with examples of how to use them
- text phrases and how they differ from text words

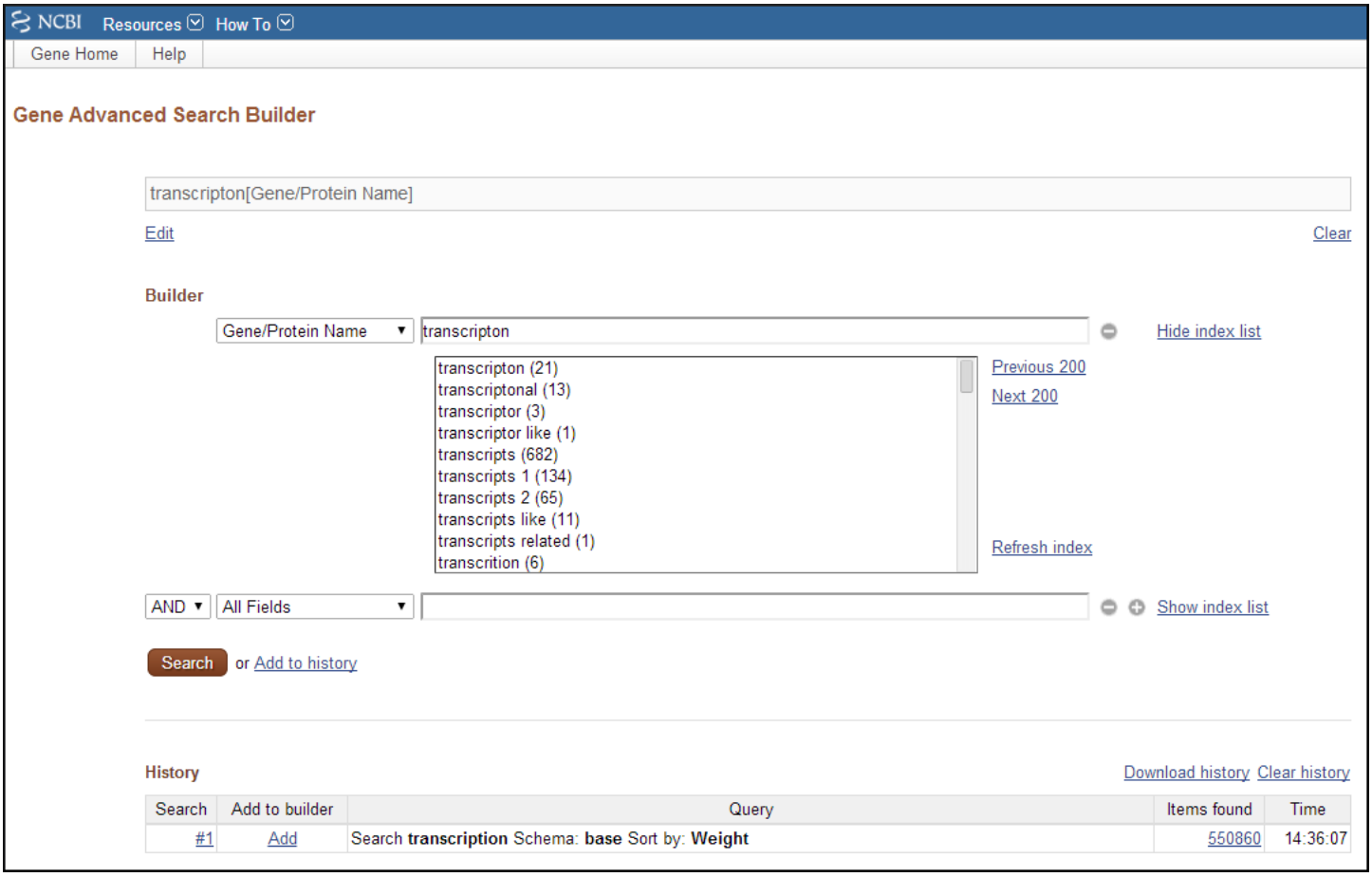

Figure 11. Advanced Search. Shown is an advanced search in which the user had queried Gene using the word transcription. Then, given the large number of results, whose counts are displayed in the History section, the user decided to explore the indices (by clicking on Show index list) for terms in the 'Gene/Protein Name' field that start with transcription. The display provides the term and the number of records with that term in the Gene database. The next step might include selecting one of the terms and clicking the Search button to display the query results, or continuing to refine the query using (or excluding) additional fields available in the index list. More information on using the Advanced Search Builder is available by following the Help link at the top of the page.

Table 5. Fields used to categorize information in Gene records.

Field name	Definition [including field abbreviations]	Examples
Name subcategory		
Disease name or phenotype of mutants	Disease or phenotype associated with the record. [DIS]	Find the genes that contribute to SCID. SCID[dis]
Gene name	A symbol for the gene. Includes preferred symbols, aliases, and locus tags. [SYM][SYMB][GN][GENE NAME]	Genes with a symbol starting with smt. smt*[sym]
Preferred symbol	The preferred symbol for the gene, not including aliases or locus tags. [PREF]	Genes with a preferred symbol starting with smt. smt*[pref]
Gene full name	Only the full name of the gene. [GFN][GENEFULLNAME]	Find genes with a full gene name of tumor protein p53: tumor protein p53[gene full name]
Gene/protein name	The short or full name of the gene or any of its protein products (when applicable). [TITL][TITLE][TF][TI][Protein]	Find genes that have the word kinase in GO annotation but do not have the word kinase in the name. kinase[gene ontology] NOT kinase[gene/protein name]

Table 5. continued from previous page.

Field name	Definition [including field abbreviations]	Examples
Protein full name	Only the full name of the protein products. [PFN][PROTEINFULLNAME]	Find genes with a full protein name of glutathione S-transferase M1: glutathione S-transferase M1[protein full name]
Location subcategory		
Base position	Base position, relative to a genomic accession. This is supported only for the reference assembly. To specify the genomic accession, the chromosome and organism (or Taxonomic ID) must be included, or the chromosome accession itself. Unplaced/unlocalized scaffold accessions may also be queried by base position. The query should define a range of at least 100 kb, and the range must be specified as a pair of integers. The query results will include genes that lie either partly or completely within the range. [CHRPOS,CPOS,CPOSITION]	9606[taxid] AND 12[chromosome] AND 9100000:9200000[chrpos] NC_000012[nucl_accn] AND 9100000:9200000[chrpos] NW_004668236[nucl_accn] AND 900000:1000000[chrpos]
Chromosome	Chromosome location of the gene. The value used is according to the convention of the source genome. In other words, if III is used, III but not 3 will be indexed in this field. [CHRM][CHR][CHROMOSOME]	Retrieve records containing the word kinase, and the gene is located on chromosome III: kinase AND III[chr] Retrieve records containing the words zinc and finger that are of human origin but not on chromosome 19: zinc finger NOT 19[chr] AND "Homo sapiens"[orgn]
Default map location	A map location in the units standard for the genome. For example, for human it is the cytogenetic band, for mouse it is the MGI map (centiMorgans). This is processed as a text field, so range queries are not implemented. For range queries, use Genome Data Viewer.	Rat genes mapped to 18 q: rat[orgn] AND 18q[default map location]
Sequence subcategory In Gene: This means searching by sequence identifier, not by the sequence itself, which is managed by BLAST .		
Nucleotide accession	An accession for a nucleotide sequence. [NACC][NUCL_ACCN]	There are instances where the same accession is applied to both nucleotide and protein sequences. To restrict an accession to nucleotide, use this field. (Accession numbers beginning with BC are not in this category.) BC052629[NACC]
Protein accession	An accession for a protein sequence. [PACC][PROT_ACCN]	There are instances where the same accession is applied to both nucleotide and protein sequences. To restrict an accession to protein, use this field. (Accession numbers beginning with three letters are not in this category.) AAH52629[PACC]
Nucleotide or Protein accession	A sequence accession of any type. [ACCN]	Find all the genes encoded in accession AE003828: AE003828
Miscellaneous subcategory (alphabetical)		

Table 5. continued from previous page.

Field name	Definition [including field abbreviations]	Examples
Assembly-specific gene annotations	Find records annotated on one assembly but not on another. [Assembly Name]	1 Find cow genes annotated on UMD_3.1 but not on Btau_4.6.1: Bos_taurus_UMD_3.1[assembly name] NOT Btau_4.6.1[assembly name] AND alive[property] 2. Find human genes annotated on GRCh38 but not on HuRef: GRCh38[assembly name] NOT HuRef[assembly name] AND alive[property] 3. Find human genes on alternate GRCh38 assemblies but not on the primary assembly: (txid9606[orgn] AND alt_ref_loci_*[assembly name]) NOT "primary assembly"[assembly name] AND alive[property]
Creation date	Date the record was created. [cd][cdat][creation date]	Records containing the word xenopus created between February 5, 2004 and February 12, 2004: 2004/2/5:2004/2/12[cd] AND xenopus[orgn]
Date Discontinued	The date on which the record was discontinued [DIS_DATE][DDAT][DISCONTINUED][DISDATE]	Records discontinued between January 1, 2006 and December 31, 2006: 2006/1/1:2006/12/31[disdate]
Domain Name	Conserved domain and protein family names. [DOMAINNAME][DOM]	Retrieve records associated with the A2M family: A2M[domainname]
EC/RN number	Enzyme commission identifier for a product of the gene. Indexed without the EC prefix. [ECNO][EC]	Retrieve records where proteins have an E.C. number of 1.9.3.1: 1.9.3.1[ECNO]
Exon count	The number of distinct, non-overlapping RefSeq exons annotated for all RNA products of this gene interval, based on annotation in this priority: reference assembly first, alternate assembly second. This field can be queried by either a single integer value or a range. [XC][NUMEXONS]	Retrieve human records with one exon: human[orgn] AND 1[exoncount] Retrieve all records with a range of exons: 10:20[exoncount]
Filter	Find records with a relationship to other data in Gene. For more examples of use of filters, see the Preview/ Index section.	Retrieve records of mouse kinase genes with expression data stored in GEO: mouse[orgn] AND gene_geoprofiles[filter] AND kinase
Gene ID	Gene identifier. This field can be queried by either a single integer value or a range. [UID][ID][GeneID]	Many integer identifiers have overlapping number spaces. To find the gene record that corresponds to the human BRCA1 gene by GeneID, use this field: 672[GeneID]
Gene length	Gene length based on annotation in this priority: reference assembly first, alternate assembly second. If there are multiple placements, only on non-reference assemblies, then the longest value on non-reference assemblies is used. This field can be queried by either a single integer value or a range. [GL][GENELEN]	Retrieve all records with a gene span less than or equal to 5kb: 1:5000[genelength]

Table 5. continued from previous page.

Field name	Definition [including field abbreviations]	Examples
Gene Ontology	GO terms applied to this gene AND the GO identifer as the integer. The terms include the component, function, and process categories. [GO][GENE ONTOLOGY]	Rat genes with GO terms starting with "kinase signaling" kinase signaling*[gene ontology] rat[orgn] Any gene with the GO id of GO:0004872: 4872[GO]
Group	Query terms to retrieve a set of genes with a specified relationship to another gene	Pseudogenes related to the same functional gene with GeneID = 11727 "related functional gene 11727"[Group]
MIM	Identifier assigned to human genes and phenotypes by OMIM [MIM]	Retrieve records that contain the MIM number 181510: 181510[MIM]
Modification date	Last date the record was modified. [MODDATE][MDAT][LMOD][DATE][UPDATED] [MD]	Retrieve records for genes from eubacterial genomes last modified after March 10, 2004: eubacteria[orgn] AND 2004/3/10:2010/1/1[md] Retrieve records from sea urchins modified in the last 30 days: echinoidea[orgn]+AND+"last 30 days"[mdat]
Organism	Scientific and common names of organism [ORGN]	Find all records in Gene for the pig: pig[organism]
Property	An attribute of a Gene record based on its content See Properties. [prop][property]	Mouse records with transcript variants: mouse[orgn] AND "has transcript variants"[property]
PubMed UID	PubMed id. [PMID]	Many integer identifiers have overlapping number spaces. To find the gene record(s) that corresponds to a paper in PubMed from Gene, use this field: 12477932[PMID]
Taxonomy ID	Identifier for the species or strain in the NCBI taxonomy database. HINT: txid{value} also works, e.g., txid9606. [TAXID][TID]	Find all records in Gene for the pig: 9823[taxid] Alternatively: txid9823
Text Word	Any word in the record. [TEXT][WORD][AB][TXT]	Retrieve records that contain "32" in a record that also contains threonine, serine, and kinase: serine AND threonine AND kinase AND 32[TEXT]

Filter

The term *filter* is used in this context to describe categories of records that are grouped according to their relationship either to other Entrez databases or to external resources that have submitted LinkOut connections. If the former, the filter is named according to the pattern "gene other_Entrez_database", such as "gene protein". If the latter, the first two letters of the filter's name are "lo", for LinkOut. For a comprehensive listing of filters valid for the Gene database and the number of records in each, follow these steps:

1. Click on the Advanced Search on the query bar.
2. Use the pull-down menu named All Fields and select Filter.
3. Click on Show Index under the open box to show the names of filter and the number of instances of each.

Filters are powerful tools to retrieve records of interest. For example, to retrieve all records for human genes that are associated with OMIM (*i.e.,* have connections to OMIM) and have links to Entrez GEO, use the "AND"

operator with both "gene omim" and "gene geo". Table 4 provides a partial list of filters for Gene; the complete list is available here.

Table 4. Filter sets (partial).

Filter name	Definition
all	Total records, current or not
gene all	All current records
gene books	Gene records with explicit links to Entrez Books
gene ccds	Genes that are represented in the CCDS collaboration
gene cdd	Genes having proteins with conserved domains identified by the Conserved Domain Database
gene genereviews	Genes mentioned in GeneReviews
gene homologene	Gene records with explicit links to Entrez HomoloGene
gene nucleotide	Gene records with explicit links to Entrez Nucleotide
gene omim	Gene records with explicit links to Entrez OMIM, and thus includes links to both disease and "gene" records in OMIM
gene probe	Gene records with explicit links to Entrez Probe
gene protein	Gene records with explicit links to Entrez Protein, and thus includes links to GenPept and SwissProt accessions
gene pubmed	Gene records with explicit links to Entrez PubMed
gene snp	Gene records with explicit links to Entrez dbSNP, and thus supports finding gene variation information available in dbSNP
gene taxonomy	Gene records with explicit links to Entrez Taxonomy
coronavirus related	Gene records with explicit links to Entrez PubMed

Properties

In general, properties are assigned to Gene records based on content rather than relationship to other database records, which is the role of filters (see Filter). There is however a small amount of redundancy between properties and filters. Many of the properties assigned to Gene records fall into these major categories:

- Type of gene: Property named as *genetype name_of_type*.
- Type of RNA: Property named as *rnatype name_of_type.*
- Source of the gene: Property named as *source name_of_source*.
- Type of RefSeq provided for the gene: Property named as *srcdb refseq type_of_refseq*.
- Type of feature annotation associated with the gene: Property named as *feattype name_of_feattype.*

The *genetype* option follows the conventions for mol_type used in the feature table of the International Nucleotide Sequence Databases (INSDC). The values should be self-explanatory, except perhaps for *miscrna*, *other*, and *unknown*. The *genetype miscrna* (misc_rna, miscellaneous RNA) is assigned to any gene that encodes an RNA product not included specifically at ncRNA vocab. The *genetype other* property is applied to loci of known type, but a specific category has not yet been applied in the Gene data model (*e.g.,* immunoglobulin and TCR gene segments). The *genetype unknown* property is applied to probable genes for which the type is still under review. This category is frequently used when the defining sequence has uncertain coding propensity. We appreciate your suggestions for any improvements.

To summarize, the *genetype* property values are:

- genetype biological region (experimentally validated non-genic genomic regions that are in scope for RefSeq Functional Elements representation, including gene regulatory elements, known structural elements, and well-characterized DNA replication origins, DNA recombination regions, and sites of genomic instability)
- genetype miscrna (miscellaneous RNA)
- genetype ncrna (non-coding RNA; includes all ncRNA classes except for snRNA, snoRNA, and scRNA [which have their own gene types]. The largest counts are from miRNA and lncRNA. ncRNA classes are documented at ncRNA vocab.)
- genetype other (when the type is known, but there is no specific enumeration for it; includes immunoglobulin and TCR gene segments, repetitive elements, and regions)
- genetype protein coding
- genetype pseudo (pseudogene)
- genetype rrna (ribosomal RNA)
- genetype scrna (small cytoplasmic RNA)
- genetype snorna (small nucleolar RNA)
- genetype snrna (small nuclear RNA)
- genetype trna (transfer RNA)
- genetype unknown (when the type of gene is uncertain)

The *rnatype* property values identify the types of RNAs that are represented on the gene:

- rnatype mirna (micro RNA)
- rnatype miscrna (miscellaneous RNA)
- rnatype mrna (messenger RNA)
- rnatype ncrna (non-coding RNA)
- rnatype other
- rnatype other genetic
- rnatype pre rna
- rnatype rnase p rna
- rnatype rrna (ribosomal RNA)
- rnatype snorna (small nucleolar RNA)
- rnatype snrna (small nuclear RNA)
- rnatype srp rna

The *source* property values should be self-explanatory, with the exception of *source other* used where a specific category has not yet been applied in the Gene data model. Values are:

- source extrachromosomal
- source genomic
- source mitochondrion
- source organelle
- source other
- source plasmid
- source plastid
- source proviral
- source virion

The *srcdb refseq* values are as enumerated by RefSeq and will not be duplicated here.

The *feattype* property values are derived from feature annotation on RefSeq(s) associated with the gene, predominately genes with a *genetype biological region*. Further information about these *feattype* values can be

found in the Feature Annotation Glossary for RefSeq Functional Elements, including links to INSDC feature specifications and controlled vocabularies, as well as links to equivalent terms in the Sequence Ontology. The values are:

- feattype caat signal
- feattype cage cluster
- feattype chromosome breakpoint
- feattype conserved region
- feattype dnase i hypersensitive site
- feattype enhancer
- feattype enhancer blocking element
- feattype epigenetically modified region
- feattype gc signal
- feattype imprinting control region
- feattype insulator
- feattype locus control region
- feattype matrix attachment region
- feattype meiotic
- feattype micrococcal nuclease hypersensitive site
- feattype misc feature
- feattype misc recomb
- feattype misc structure
- feattype mitotic
- feattype mobile element
- feattype non allelic homologous
- feattype nucleotide cleavage site
- feattype nucleotide motif
- feattype promoter
- feattype protein bind
- feattype recombination hotspot
- feattype regulatory
- feattype rep origin
- feattype repeat instability region
- feattype repeat region
- feattype replication regulatory region
- feattype replication start site
- feattype response element
- feattype sequence alteration
- feattype sequence comparison
- feattype sequence feature
- feattype silencer
- feattype stem loop
- feattype tata box
- feattype transcription start site
- feattype transcriptional cis regulatory region

Other properties used to categorize Gene records are explained in Table 3.

Text Phrases

A *text phrase* is a special type of text search that uses two or more words to form a phrase. An ordinary text search of two or more words will find gene records that contain all of the specified words anywhere in the gene record. By contrast, a text phrase search will find gene records that contain all of the specified words *together* and *in the specified order*.

A text phrase search is constructed by placing double quotes around the phrase. A list of certain phrases that can be used to find records of interest in gene is in Table 6.

Table 6. Text phrases.

Text phrase	Explanation
"Annotation Information"	For genomes that NCBI annotates, Gene represents information about the annotation of each current GeneID. Text phrases will be attached to the gene data if the gene is not annotated well, or if annotation has changed in a complex way. Text phrases will also be attached if there is no defining cDNA or genomic sequence for the gene, or if the GeneID was created after the most recent genome annotation. The goal is to facilitate retrieval of Gene records where the annotation on the RefSeq genomic records, if it exists, should be interpreted with caution. Records that are not known to have annotation issues can be retrieved by including the following in the query: NOT "Annotation Information" [Text] Specific sub-categories of annotation information are described below.
Sub-categories of annotation information	
"partial on reference assembly"	The annotated gene, as suggested by the defining cDNA, is not complete.
"spans an assembly gap"	There is a gap in the reference assembly where the defining cDNA should align.
"suggests misassembly"	There are order/orientation issues in the reference assembly suggested by the cDNA alignment.
"not annotated on reference assembly"	This gene is not annotated on the reference assembly.
"not in current annotation release"	This gene is not annotated on any assembly of the current annotation release.
"only annotated on alternate loci in reference assembly"	This gene is only annotated on one or more alternate locus assembly-units of the reference assembly.
"only annotated on patches unit in reference assembly"	This gene is only annotated on the PATCHES assembly-unit of the reference assembly.
"only annotated on alternate loci and patches unit in reference assembly"	This gene is only annotated on one or more alternate locus assembly-units and the PATCHES assembly-unit of the reference assembly
Other text phrases	
"Orthologs from Annotation Pipeline"	The Homology section for many genes features a link for "Orthologs from Annotation Pipeline." This dataset is computed as part of NCBI's Eukaryotic Genome Annotation Pipeline using a combination of protein sequence similarity and local synteny information. The pipeline determines orthology between the genome assembly that is being annotated and a reference genome, typically human. The collection of pairwise orthology calls is then tracked as a group which may be further supplemented by manual curation. This process provides ortholog information more quickly for newly annotated genomes, and supplements the content available in HomoloGene.
"involved in immune response or antiviral activity"	Related to COVID-19
"involved in cytokine storm inflammatory response"	Related to COVID-19
"involved in SARS-CoV-2 infection"	Related to COVID-19

Table 6. continued from previous page.

Text phrase	Explanation
“relevant for COVID-19 prognosis”	Related to COVID-19
“relevant for COVID-19 treatment”	Related to COVID-19
“involved in host gene regulation”	Related to COVID-19
“involved in host gene recombination”	Related to COVID-19
“relevant for disease process”	Related to COVID-19

Please note that the double quotes are included in the text phrases shown here because they are mandatory when performing a text phrase search.

Finding subsets of your results; the ‘Results filter sidebar’ and ‘Filter your results’ options

When reviewing a query result in HTML format (not text), there are two options that allow you to display only a subset of the results:

- Using the ‘Results filter sidebar’
- ‘Filter your results’ using My NCBI

Using the ‘Results filter sidebar’

The Results filter sidebar (Figure 2) is displayed to the left of your search results and is used to narrow the search results. Clicking a sidebar filter activates that filter, and all subsequent searches will be filtered until the selected filter(s) is cleared.

A check mark is located next to an active filter and the *Filters activated:* message is displayed above the Results table. The *Search details* box on the right side displays the updated query. Selecting more than one filter narrows the search further (equivalent to using a Boolean AND). A search can be expanded by replacing AND with OR in the *Search details* box.

Turn off the sidebar filters in any of these ways:

- Use the ‘Clear all’ link at the top
- Use the ‘clear’ link next to a filter group to clear the filters within that group
- Click on a check mark to clear an individual filter

Sidebar filter groups (described below) include Gene sources, Categories, Sequence content, Status, Chromosome locations, and Search fields. Within a filter group, only filter options valid for the current search results are listed. Use the ‘Show additional filters’ link to add or remove a filter group from the sidebar. A filter group with a greyed check mark in the ‘Additional filters’ menu cannot be removed.

To filter by organism, use the ‘Top Organisms’ section at the upper right of the results page. Additional filters are available but are managed through your My NCBI account; see ‘Filter your results’ using My NCBI.

Sidebar filter groups include:

Gene sources

Filter your search results based on the type of gene in the results set.

- **Genomic**: genes encoded by chromosomes or the major genomic macromolecule for the taxon.
- **Mitochondrial**: genes encoded by mitochondria.

- **Organelles**: genes encoded by organelles, including mitochondria, plastids, and macronuclei
- **Plastids**: genes encoded by plastids.

Categories

Filter your search results based on the existence of alternatively spliced RefSeqs, or on protein-coding capacity. **NEWENTRY** records support submission of GeneRIFs, by species, for a gene not currently in Gene.

Sequence content

Filter your search results based on these properties:

- **CCDS**: records that encode a protein sequence belonging to a Consensus CDS (CCDS) set. See http://www.ncbi.nlm.nih.gov/projects/CCDS/.
- **Ensembl**: records that match Ensembl annotation based on comparison of mRNA and protein features. See Table 3 for more information.
- **RefSeq**: records with an associated RefSeq record.
- **RefSeqGene**: records with an associated gene-specific genomic RefSeq in the RefSeqGene class.

Status

Restrict your search results for records that are 'Current Only'. This is a particularly useful filter that removes discontinued or replaced records from the result set. It is equivalent to submitting a query that contains the expression 'AND alive[property]'.

Chromosome locations

Restrict your search results by Organism, reference assembly chromosome or organelle, and location.

Search fields

Restrict your search results using any of the listed search fields. Table 5 summarizes these search fields (grouped into sub-categories) used to categorize information in Gene records. The table also provides examples of how to use these entities effectively to retrieve records.

'Filter your results' using My NCBI

In addition to the sidebar filters Gene provides by default, you can take advantage of any of the standard filters for Gene available via My NCBI. For example, if you are interested in Gene records that have a record in OMIM, you can use My NCBI to define "Gene records with MIM (Mendelian Inheritance in Man) numbers" as one of your standard filters. These filter results will be shown at the upper right of the query results screen. In addition to the standard filters, My NCBI also provides a button to 'Create custom filters'. See Working with Filters for more information.

Words Excluded From Queries

Common, but uninformative, words and terms (also known as stopwords) are automatically eliminated from searches. However, a search term that is a stopword will be included if the term is explicitly qualified by a field name. For example, if you want to search for the term *was*, you could use:

- was [All Fields]

Enclosing the term in double quotes would have the same effect.

A list of stopwords used in Gene is in Table 7.

Table 7. Stopwords.

	Stopwords
A	a, about, again, all, almost, also, although, always, among, an, and, another, any, are, as, at
B	be, because, been, before, being, between, both, but, by
C	can, could
D	did, do, does, done, due, during
E	each, either, enough, especially, etc
F	for, found, from, further
G	Gene
H	had, has, have, having, here, how, however
I	i, if, in, into, is, it, its, itself
J	just
K	kg, km
M	made, mainly, make, may, mg, might, ml, mm, most, mostly, must
N	nearly, neither, no, nor
O	obtained, of, often, on, our, overall
P	perhaps, pmid, protein
Q	quite
R	rather, really, regarding
S	seem, seen, sequence, several, should, show, showed, shown, shows, significantly, since, so, some, such
T	than, that, the, their, theirs, them, then, there, therefore, these, they, this, those, through, thus, to
U	upon, use, used, using
V	various, very
W	was, we, were, what, when, which, while, with, within, without, would

Finding Data Related to Gene in Other Databases

The Related information menu, in the right column of the Full Report display, supports the function to retrieve information in other Entrez databases related to your result set. This function is supported by the links provided by NCBI's Entrez system. The calculation of Entrez links is documented here. If you navigate to that documentation on the web, click on Gene to navigate quickly to the description of gene-specific links. Navigation in the Related information menu is based on the same infrastructure in Entrez that supports navigation to records related to a set of query results. The following provides more details about some of these links.

3D structures

3D structures provide experimentally resolved structures of proteins, RNA, and DNA derived from the Protein Data Bank, and include links to literature, related sequences, and more. To retrieve all records in Gene in this category, try the query "gene structure"[filter] from the Gene Search bar.

Books

An increasing number of Gene records are annotated specifically in books and monographs provided in Bookshelf. One example, restricted to human genes, is the GeneReviews book provided in collaboration with the GeneTests group of the University of Washington. To retrieve all records in Gene in this category, try the query "gene books"[filter] from the Gene Search bar. To retrieve only genes referenced in GeneReviews, use gene_genereviews[filter].

CCDS

Genes for which the protein products were accessioned by the Consensus CDS (CCDS) project. To retrieve all records in Gene in this category, try the query "gene ccds"[filter] from the Gene Search bar.

ClinVar

ClinVar maintains information about the relationships among human variations and phenotypes, including supporting evidence. ClinVar records associated with your Gene search results can be retrieved by using this display option. To retrieve all records in Gene with variations registed in ClinVar, try the query "gene clinvar"[filter] from the Gene Search bar.

Conserved Domains

Protein sequences are routinely compared to canonical sequences for domains in the Conserved Domain Database. Domain records connected to protein associated with records associated with your Gene search results can be retrieved by using this display option. To retrieve all records in Gene in this category, try the query "gene cdd"[filter] from the Gene Search bar.

Gene neighbors

The Genomic Context diagram of the Full Report display shows a gene's genomic placement along with neighboring genes. Gene Neighbors contains the raw data corresponding to the Genomic Context diagram. To retrieve all records in Gene in this category, try the query "gene gene neighbors"[filter] from the Gene Search bar.

Genome

Genome maintains information about chromosomes and complete genomes. Genome records associated with your Gene search results can be retrieved by using this display option. To retrieve all records in Gene in this category, try the query "gene genome"[filter] from the Gene Search bar.

Genetic Testing Registry (GTR)

The GTR is a central repository for the voluntary submission of genetic test information by test providers. To retrieve all records in Gene in this category, try the query "gene gtr"[filter] from the Gene Search bar.

HomoloGene

HomoloGene compares protein-coding genes in several key genomes to identify homologs. HomoloGene records associated with r Gene search results can be retrieved by using this display option. To retrieve all records in Gene in this category, try the query "gene homologene"[filter] from the Gene Search bar.

NIH cDNA clone

Some of the mRNAs associated with your Gene search results are available from NIH-supported cDNA repositories. Reports of clones in the Nucleotide database associated with your Gene search results can be retrieved by using this display option.

Nucleotide

Nucleotide sequences associated with your Gene search results can be retrieved by using this display option. To retrieve all records in Gene with nucleotide sequence information, try the query "gene nucleotide"[filter] from the Gene Search bar.

MedGen

MedGen is NCBI's portal to information related to medical genetics, including clinical features, available tests, up-to-date literature, practice guidelines, and consumer resources. To retrieve all records in Gene in this category, try the query "gene medgen diseases"[filter] from the Gene Search bar..

OMIM

OMIM records associated with your Gene search results can be retrieved by using this display option. To retrieve all records in Gene in this category, try the query "gene omim"[filter] from the Gene Search bar.

PubChem BioAssay

Genes with products having screening results reported in the PubChem BioAssay database. To retrieve all records in Gene in this category, try the query "gene pcassay"[filter] from the Gene Search bar.

PubChem Compound

Genes with products having screening results in the PubChem Compound database. To retrieve all records in Gene in this category, try the query "gene pccompound"[filter] from the Gene Search bar.

PubChem Substance

Genes with products having screening results in the PubChem Substance database. To retrieve all records in Gene in this category, try the query "gene pcsubtance"[filter] from the Gene Search bar.

PMC

Publications available as full text from PubMedCentral may include explicit references to Gene. Publications may also be connected to Gene via a PubMed ID. PubMedCentral records associated with your Gene search results can be retrieved by using this display option. To retrieve all records in Gene in this category, try the query "gene pmc"[filter] from the Gene Search bar.

Probe

Probe records, such as those for resequencing primers or RNAi sequences, related to your Gene search results can be retrieved by using this display option. To retrieve all records in Gene in this category, try the query "gene probe"[filter] from the Gene Search bar.

Protein

Protein sequences associated with your Gene search results can be retrieved by using this display option. To retrieve all records in Gene with protein sequence information, try the query "gene protein"[filter] from the Gene Search bar.

PubMed

PubMed citations associated with your Gene search results can be retrieved by using this display option. Those that were generated from GeneRIFs, including interaction data, are indicated by the PubMed (GeneRIF) option. To retrieve all records in Gene with citations in PubMed, try the query "gene pubmed"[filter] from the Gene Search bar.

SNP

Use these display options to navigate to information about variation reported in the dbSNP database for the gene records in your search results. To retrieve all records in Gene with reported variation, try the query "gene snp"[filter] from the Gene Search bar.

Taxonomy

Use this display option to navigate to information about the taxonomy of the genomes in which the gene records in your search results are found.

Constructing Powerful Queries

Constructing queries based on free text, filters, and properties can be quite powerful in retrieving records of interest from Gene. Table 8 summarizes some of these approaches by describing:

- Scope: The intent of a query.
- Query: How to construct a query that meets that intent.
- Notes: How usage of Gene to retrieve these data may compare to other gene-related resources, namely HomoloGene or Genome Data Viewer.

Although these examples use field restriction (see Table 5 for the comprehensive list of fields used to index the information in Gene records), free text can also be submitted. Gene then weights the retrievals based on the field in which a result was found. For example, if your query matches a gene symbol in one record and arbitrary text in another, the record where the match is on the symbol will be displayed before the other in the results. Thus Gene controls the default order in which results are returned by evaluating what fields are more critical to matching your query. This default sorting order is termed 'relevance'.

Table 8. Constructing queries.

Scope	Query	Notes
Find genes mapped to *Arabidopsis thaliana* chromosome 3 that have orthologs reported in HomoloGene	arabidopsis thaliana[orgn] AND 3[chr] AND gene_homologene[filter]	[orgn] is used to restrict "Arabidopsis thaliana" to the organism field. [chr] is used to restrict "3" to the chromosome field. gene homologene[filter] is used to restrict records to those processed by HomoloGene. This query is not currently able to be processed by Genome Data Viewer, because the relationship to HomoloGene is not processed for indexing at present, nor by HomoloGene, because the chromosome data are not captured in HomoloGene.
Find genes also being processed by OMIM but for which there is not currently a RefSeq of the type "known"	gene omim[filter] NOT srcdb refseq known[prop]	gene omim[filter] is used to find all Gene records with relationships to OMIM. srcdb refseq known[prop] is used (as the Boolean NOT) to find all such records that do not have RefSeqs of the accession format NM_000000, NG_000000, or NR_000000.
Find genes from genomes other than mammals that are classified by the GO consortium to have some relationship to the cytoskeleton	cytoskeleton[go] NOT mammalia[orgn]	[go] is used to restrict to the field "Genome Ontology". [orgn] is used to restrict (as the Boolean NOT) to species not classified as mammals. Queries based on GO terms are not supported in either Genome Data Viewer or HomoloGene. Please note that Gene does not recapitulate tree-based searching for GO annotation; this retrieval is based solely on the existence of the word in any GO category. Links are provided to the GO website to support more specific searches.
Find genes expressed in human placenta but not prostate	human[orgn] AND ("placenta"[expression/tissues] NOT "prostate"[expression/tissues])	[orgn] is used to restrict "human" to the organism field. [expression/tissues] is used to restrict records to those expressed in placenta while the Boolean NOT also restricts it to those not expressed in prostate.
Find genes expressed in mouse liver	mouse[orgn] liver[expression/tissues]	[orgn] is used to restrict "mouse" to the organism field. [expression/tissues] is used to restrict records to those expressed in liver at any of the 4 developmental stages represented in the primary mouse dataset. Expression bins are indexed both on the full string and individual words.
Find genes expressed in mouse liver at developmental stage E14.5	mouse[orgn] "liver E14.5"[expression/tissues]	[orgn] is used to restrict "mouse" to the organism field. [expression/tissues] is used to restrict records to those expressed in liver only at developmental stage E14.5. Expression bins are indexed both on the full string and individual words.

Tips for Programmers

The Gene Data Model and DTD

The data model for Gene is documented in the Gene specification. It combines several definitions used by other NCBI databases, such as seqfeat, but also establishes definitions specific to Gene. Of special note is the Gene-commentary, which is used to represent many descriptors of genes. Each Gene-commentary is defined by type and supports specific representation of such elements as sequence database accession numbers (accession, version), citations (refs), external or internal resources defining the data (source), and position information. Heading, label, and text are used for general data, with the choice influenced by display in the Gene viewers.

The DTD for Gene is available from NCBI's DTD directory and is called NCBI Gene.dtd.

Entrez Programming Utilities and Gene

The full power of Entrez Programming Utilities (e-Utils) can be used to extract information from Gene programmatically. The basic strategy is to identify the query that will return the desired records and then submit that query via ESearch. The GeneIDs identified by that search can then be submitted to another function, such as ESummary or EFetch. Examples for Gene are ESummary.

Extracting Gene Summaries and other information from Gene's Document Summary

The Summary text provided via Gene and on RefSeq records can be extracted by taking advantage of the following:

- the text of the Summary is included in the Document Summary (docsum) from Gene.
- genes with Summary text can be identified by the has_summary property.

In other words:

1. use eSearch to find all GeneIDs with the has_summary property
2. use eSummary to retrieve the Summary text (*e.g.*http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=gene&id=672&retmode=xml)
3. Extract the string in the Summary tag.

Table 9 lists the name attributes of Gene's docsum that can be extracted in a similar manner. An example docsum is provided here:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=gene&id=4267&filter=asis

Table 9. The Name attributes of Gene's Document Summary (docsum).

Name attribute	Description
Name	preferred symbol (same as official, if provided)
Description	preferred gene description (same as official, if provided)
Orgname	binomial, and strain when appropriate, from NCBI's Taxonomy database
Status	0, live; 1, replaced; 2, discontinued; 3, used for GeneRIF processing
CurrentID	0, for GeneIDs without a replacement; otherwise, the new GeneID if the one being displayed has been replaced
Chromosome	chromosome on which this gene has been reported or annotated.

Table 9. continued from previous page.

Name attribute	Description
GeneticSource	genomic, mitochondrion, chloroplast, plasmid. The type of genome on which this gene occurs. Genome is used for chromosome.
MapLocation	cytogenetic or genetic map location
OtherAliases	alternate acronyms
OtherDesignations	alternate full descriptions
NomenclatureSymbol	official symbol provided by a named authority
NomenclatureName	official name provided by a named authority
NomenclatureStatus	'official' if provided by a named authority; otherwise 0
TaxID	unique identifier from NCBI's Taxonomy Database
Mim	refers to Online Mendelian Inheritance in Man (OMIM)
int	unique identifier(s) from Online Mendelian Inheritance in Man (OMIM)
GenomicInfo	refers to a genomic RefSeq for the reference assembly; otherwise null
ChrLoc	chromosome name. It should be the same as Chromosome but may not be.
ChrAccVer	refSeq accession.version of the reference assembly corresponding to ChrLoc
ChrStart	nucleotide location of the 5' end of the gene as last annotated (position value is 0-based)
ChrStop	nucleotide location of the 3' end of the gene as last annotated (position value is 0-based)
ExonCount	see Exon count description in Table 5
GeneWeight	described above in "How Data Are Displayed (Display Settings/Format)"
Summary	descriptive text about the gene
ChrSort	padded string to support ascii sorting of the chromosome
ChrStart	the smaller of ChrStart and ChrStop in the GenomicInfo section
Organism	refers to the taxon from NCBI's Taxonomy Database
ScientificName	binomial, and strain when appropriate, from NCBI's Taxonomy database
CommonName	preferred (GenBank) common name
TaxID	unique identifier from NCBI's Taxonomy Database
LocationHist	historically annotated location(s) on top-level genomic sequence only, by annotation release and assembly

Extracting Gene Neighbors

Gene Neighbors can be queried programmatically using the Entrez Links function of E-utilities. For example, to find all neighbors of GeneID 672, use this:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=gene&dbto=gene&linkname=gene_gene_neighbors&from_uid=672

Note that gene neighbors identified by this method are not associated with a specific genomic placement but with all reported genomic placements. In many cases, a gene's neighbors are the same for all genomic placements. However, in some cases, a gene's neighbors may differ from one genomic placement to another, for example, between the reference assembly and an alternate assembly.

Gene FTP Site

The FTP site for Gene (README) has three major subdirectories: DATA, GeneRIF, and Tools.

DATA

DATA contains files that provide key attributes of genes, including:

- all associated accession numbers, including RefSeqs (gene2accession.gz)
- matching Ensembl annotation (gene2ensembl.gz)
- GO terms (gene2go.gz)
- citations (gene2pubmed.gz)
- associated RefSeq accession numbers (gene2refseq.gz)
- relationships to other genes (gene_group.gz)
- genes that are no longer current (gene_history.gz)
- nomenclature, ID, and map data (gene_info.gz)
- neighboring genes (gene_neighbors)
- MIM numbers and records in MedGen (mim2gene_medgen)
- relationship to UniProtKB proteins (gene_refseq_uniprotkb_collab.gz)
- ortholog data (gene_orthologs.gz)

Details of the construction of these files are reported in the (README) file.

DATA also contains the ASN_BINARY subdirectory. This path contains both a comprehensive extraction from Gene (All_Data.ags.gz), several subsets categorized by source (Organelles, Plasmids), and subdirectories grouped broadly by taxonomy. Records of genes from species that are requested frequently are also provided in species-specific files, for example these mammals. The format of these extractions is compressed binary ASN.1. The program gene2xml is available to convert these files to XML or ASN.1 text. Be aware that the converted files will take approximately 100-fold more space than the original compressed binary ags.gz files.

The GENE_INFO subdirectory of DATA provides subsets of the gene_info file grouped broadly by taxonomy. This directory structure mirrors that of the ASN_BINARY path. Thus if you want the type of information provided in gene_info, but do not want to have to process the complete text, you can use one of the files in the appropriate subdirectory, for example these plants.

GeneRIF

GeneRIF contains files that provide supplemental information about gene functions, either from the GeneRIF pipeline (generifs_basic.gz) or the HIV-1, Human Protein Interaction Database (hiv_interactions.gz). The tab-delimited files are not subdivided by species of interest. All files except the file reporting GeneID/PubMedID relationships (gene2pubmed.gz) have a column with the ID from the NCBI Taxonomy database to facilitate the extraction of a subset of the data from the file by species.

Tools

Gene_tools provides or points to programs and scripts to mine data from Gene. Of particular interest is gene2xml, which can be used to convert the binary ASN.1 in the ASN_BINARY directory to XML or to ASN.1 in text format (README).

Connecting Users of Gene to Your Website

Gene can serve as a gateway to information on your website served from your local database. Users of Gene will discover your website if you participate in our LinkOut system and become a LinkOut provider. Any Entrez

database will support LinkOut. Linkout Help's Information for Other Resource Providers explains the details of this opportunity.

There are many benefits to becoming a LinkOut provider. If you want access to your database to be apparent from Gene, you can control the description of your resource, the update cycle, and the icon to anchor links to your site. In other words, you do not have to wait for NCBI staff to go to your site to obtain and process information and match to Gene records. You know your site best—you can identify which records are related to Gene records and provide the most accurate and informative URL to connect that Gene record to your site. If you already provide LinkOuts to other Entrez databases, such as Nucleotide or Protein, you do not have to re-register as a provider; you need only notify LinkOut staff and start to submit a new resource file.

With the implementation of My NCBI, it is even more advantageous to become a LinkOut provider. One of the options registered users of My NCBI can select is to display the icons for any LinkOut provider at the top of a record. The presence of your familiar logo would invite users of Gene to go to your site.

Connecting your site to Gene

URLs can be constructed to query Gene, or to display a specific record if you know the GeneID. For example, if your site maintains the identifiers (GeneID) used by Gene, you can construct a link from your site to Gene by combining this base

http://www.ncbi.nlm.nih.gov/gene/

with the GeneID. For example, to link to GeneID 1, use this URL:

http://www.ncbi.nlm.nih.gov/gene/1

URLs that query Gene are constructed by adding ?term=[search term]

For example, to find records in Gene containing the phrase 'immunoglobulin domain', use this URL

http://www.ncbi.nlm.nih.gov/gene/?term=immunoglobulin_domain

More examples of queries are provided on Gene's Home page, and general rules for building URLs to query Entrez databases are provided in the Creating a Web Link to the Entrez Databases chapter of this book. The valid display options are also documented in that chapter.

Historical Information about LocusLink

This version of Gene's help document removed detailed information about LocusLink. If you have any question about the history of LocusLink, please use this form.

Gene Frequently Asked Questions

Created: April 21, 2008; Updated: April 9, 2018.

For General Users

For Programmers and Database Developers

General Questions

1. Nomenclature. How and when are gene symbols and names assigned?
2. How can I obtain the genomic sequence for a gene?
    - From the Reference Sequences Section
    - From the Genomic regions, transcripts, and proteins section
    - From Map Viewer
    - From RefSeqGene
    - From Command Line (for bulk downloads)
3. Notification of changes in Gene
4. Differing Representations of RefSeqs
    - Display of RefSeqs in Transcripts and Products *vs*. in the Reference Sequences (RefSeq) section
    - The Gene Table display *vs*. Entrez Nucleotide.
    - Multiple chromosomal locations
    - Representation of nucleotide position
5. Gene and OMIM
6. How does Gene maintain certain types of information?
    - Conserved domains of encoded proteins
    - GeneRIFs
        1. How are they maintained?
        2. How are they reported from the web?
        3. How are they reported on the ftp site?
    - GO terms
    - Interactions
7. Why can I sometimes display a record, but then cannot retrieve it by a query?
8. How can I identify genes with/without a known function?
9. In what order are exons presented in ASN.1 and XML?
10. How are wild cards (*) processed?
11. Why are links from Gene to EST not comprehensive?
12. How does Gene represent genes spanning the origin of replication of a circular genome?
13. What is a readthrough locus and how is it represented?
14. How can I determine the position of genes and exons for my species of interest?
15. How can I retrieve all records for my species of interest?
16. How can I identify genes that have related pseudogenes?
17. How can I find all genes located within a specific region of a chromosome?
18. Why does the number of GeneRIFs displayed in the Bibliography section differ from the number of PubMed IDs reported using the PubMed(GeneRIF) link?
19. Why did many bacterial GeneIDs disappear?

Nomenclature

This section includes more details about sources, updates, and conventions for genes of uncertain function (LOC symbols).

Sources

The names (symbols) and full descriptions used in Gene come from 5 major sources:

1. Species-specific nomenclature committees, with great appreciation, as enumerated here and here.
2. The gene name (symbol) and protein names provided in submissions used as sources for RefSeq records.
3. Symbols and full descriptions submitted by contributors of information about loci not defined by sequence.
4. Curation by NCBI staff
5. NCBI's annotation pipeline

If there is a nomenclature committee for a species, those names have precedence.

Updates and access

Gene attempts to maintain current nomenclature. Updates to names in Gene are not propagated immediately to all other resources in NCBI. You may notice, for example, that symbols in genomic RefSeq annotation, Genome Data Viewer, HomoloGene or UniGene, and their respective ftp sites, are not the same as those you see in Gene. RefSeq, for example, does not resubmit the full annotation of a genomic sequence to the nucleotide database each time a symbol changes. The symbols seen in Genome Data Viewer and RefSeqs for contigs, scaffolds, and chromosomes, however, should be the same, because all are updated only with each major re-annotation of a genome.

It may help to consider that the Gene GeneID is unique across all taxa. You can therefore convert any GeneID into its current names by using the definitions provided in the file available as ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz. For example, if you transferred the gene_info.gz file to a unix or linux file system, the command

gzcat gene_info.gz | cut -f2,3,5,9,13

will give you

1. the GeneID
2. the current official symbol or database identifier if no official symbol is available
3. a pipe-delimited set of aliases
4. the full name
5. the nomenclature status of the name, where
    - 0 = official from a nomenclature committee,
    - I = interim from a nomenclature committee,
    - - = NCBI-supplied.

If a GeneID is no longer current, it will not be reported in the file gene_info.gz. The file gene_history.gz in the same ftp directory can be used to determine if there is a replacement GeneID, for which the current names can then be determined as above.

Conventions

Uniqueness. Gene does not enforce uniqueness in preferred symbols. If the same symbol has been assigned to different genes, and a nomenclature committee has not provided a unique name for these genes, Gene will not impose its own solution. In other words, please consider use of the GeneID rather than a symbol as the stable identifier of a gene.

Symbols beginning with LOC. When a published symbol is not available, and orthologs have not yet been determined, Gene will provide a symbol that is constructed as 'LOC' + the GeneID. This is not retained when a replacement symbol has been identified, although queries by the LOC term are still supported. In other words, a record with the symbol LOC12345 is equivalent to GeneID = 12345. So if the symbol changes, the record can still be retrieved on the web using LOC12345 as a query, or from any file using GeneID = 12345.

Names beginning with 'similar to'. When NCBI automatically annotates a genome, it predicts both mRNAs and the proteins they encode. The protein sequences are compared to public protein sequence records from several model organisms. If a significant match is found, and the name is informative, then the automatic annotation process previously constructed the name of the model by combining 'similar to ' and the name of the matching protein. Because the sequences represented by NCBI's predictions are provided in accessions beginning with XM_ or XP_ or XR_, you might assume that all accessions with that format would have names beginning with 'similar to '. This is not necessarily the case:

- NCBI will generate XM_ and XP_ or XR_ accessions for genes identified outside of the annotation pipeline, but annotated by the annotation pipeline, but for which curated (NM and NP or NR) accessions are not available. These genes, and the RefSeq accessions that represent them, will **not** have names beginning with **similar to**.
- The method for assigning names to models has changed. The current method appends '-like' to the end of the name of the record with the best match. Until all genomes are re-annotated, names beginning with 'similar to' will occur.
- **Other cases of uncertainty**. When the name that should be assigned to the gene or protein is uncertain, sources use different conventions. The terms that are used {'hypothetical' (often from RefSeq), 'similar to ' (from NCBI's annotation pipeline), 'putative', 'unknown', 'novel' (from original submitters)} should not be construed to indicate different types of uncertainty. The terms can be considered equivalent, and reflect primarily the source of the naming. Gene and RefSeq encourage all data submitters to conform to the suggestions from major sequence databases.

NOTE: To the greatest extent possible, each protein-coding gene in mitochondria has been assigned the same name (symbol) and full description across species. In some instances, this is at variance with the symbol assigned by species-specific nomenclature committees. In those cases, the species-specific nomenclature is provided, but not as the default. The official name is reported in the comprehensive gene_info file on the FTP site (note also the species-restricted ones in the GENE_INFO subdirectory.

Obtaining genomic sequence

From Gene's *Reference Sequences* section of the full report

1. Scroll to the section(s) labeled 'Genomic'
2. Click on FASTA
3. If you want to adjust the range to capture, modify the values in the **Change region shown** tool on the FASTA display and click on **Update View**

From Gene's diagram in the *Genomic regions, transcripts, and products* section of the Full Report or Gene Table report

When a gene is annotated on a RefSeq for a chromosome or scaffold, there is an embedded display of the annotation of that gene. This display is similar to the one obtained by retrieving the sequence of the RefSeq from the Nucleotide database and selecting the 'Graphics' display option. To get the genomic sequence in FASTA format

1. Scroll to the section(s) labeled 'Genomic regions, transcripts, and proteins'

2. Click on Go to nucleotide FASTA
3. To adjust the range to capture, modify the values in the **Change region shown** tool on the FASTA display and click on **Update View**
4. A YouTube video describing how to obtain genomic sequence in this manner is also available.

From Map Viewer

From Gene, you can navigate to Map Viewer to use the download functions there.

1. Select Map Viewer from the Genome Browsers list in the right margin of the Gene record.
2. Click on Download/View Sequence/Evidence in the upper right of Map Viewer display, or click on **dl** in the label for the gene.
3. Adjust the range and strand if you like and press enter or **Change Region/Strand**.
4. Select a format (FASTA is the default).
5. Save

From Entrez Nucleotide (note: position values are one-offset)

Within Entrez Nucleotide, feature names anchor URLs. Clicking on 'gene' results in a display (in GenBank format) of that subsequence. To save the sequence, change the display format to FASTA and save as described above.

From RefSeqGene

For a limited number of genes in the human genome, gene-specific genomic RefSeqs, termed RefSeqGenes, have been created. These have a RefSeq accession beginning with **NG_** and can be retrieved from the Nucleotide database using the query refseqgene[keyword].

In the Links menu, you can also get to the sequence by clicking on the **RefSeqGene** link.

From Command Line (for bulk downloads)

Run:

esearch -db gene -query 2 -field uid | esummary | xtract -pattern GenomicInfoType -element ChrAccVer -1-based ChrStart ChrStop | xargs -n 3 sh -c 'efetch -db nuccore -id "$0" -seq_start "$1" -seq_stop "$2" -format fasta'

Notification of changes in Gene

Gene maintains an RSS feed that is used to notify subscribers of current or future changes in Gene and any of its reports. If this is of interest to you, please subscribe.

Differing representations of RefSeqs

Display from the Nucleotide or protein databases

The Links section at the right of the Full Report, GeneTable, and GeneRIF display formats provides links labeled:

- RefSeq proteins
- RefSeq RNAs
- RefSeqGene

These links result in a display of RefSeqs specific to the gene in the Nucleotide or Protein databases, as appropriate. Those databases support many tools to format sequence records and analyze them by tools such as BLAST or BLink.

Display of RefSeqs in Transcripts and Products *vs.* in the Reference Sequences (RefSeq) section

The diagram of the placement of RefSeq transcripts in the Transcripts and Products Section is based on the annotation of the positions of exons and coding sequences on the indicated RefSeq. In most cases, this RefSeq is for the chromosome record of the reference assembly. If there are alternate assemblies, they can be selected for display from the Gene Table display.

For some genomes, the genomic RefSeqs are updated independently of the annotated product RNAs, with the latter being updated more frequently. This means that several kinds of discrepancies between the diagram and the current RefSeq RNAs may result.

- The diagram may be labeled with an mRNA accession (for a predicted transcript) of the format XM_123456, yet clicking on that accession results in an entry in Entrez Nucleotide that indicates this accession is no longer primary. That means that a curated mRNA (accession of the format NM_123456 or NM_123456789) has been generated to replace the previous model accession. This new "NM" accession will be reported in the Reference Sequences section, in the subsection entitled **RefSeqs maintained independently of Annotated Genomes.**
- The diagram may be labeled with curated RNA accessions (of the format NM_123456 or NM_123456789 or NR_123456) different from those listed in the RefSeq section. This will result if curation after the submission of the annotated genome identified more transcript variants, which therefore are listed only in the Reference Sequence section but not in the diagram. It will also result if curation after submission of the annotated genome identified an error in the annotated product, and the accession for that product was suppressed. In that case, the Transcripts and Products section will indicate a transcript not listed in the RefSeq section of the Gene report. A comment explaining why the record was suppressed is also provided.
- The diagram may be labeled with a version of an mRNA or protein accession (for example, NM_123456.1) different from that listed in the RefSeq section (for example, NM_123456.2). This will result if the sequence has been changed in any way, such as extending the 5' or 3' ends, or removing mismatches between the cDNA sequence and the reference assembly.
- The diagram in the full report display represents only one annotated assembly. There may be some RefSeqs that align only to an alternate assembly, and thus will not appear in the full report graphic but will be visible in the Gene Table display and in the Reference Sequences section.

The Gene Table display *vs.* Entrez Nucleotide.

RefSeq RNA records are often based on cDNA sequences submitted to GenBank. They therefore can differ from the reference genomic sequence, either for biological reasons (variation or RNA editing) or some unresolved sequence discrepancy. The report of intron/exon organization in the Gene Table display is based on the placement of exons and CDS on the genomic sequence. If the independently determined RefSeq mRNA cannot be aligned perfectly to the genome, the lengths given in the Gene Table display may differ from that of the mRNA sequence itself. As discussed in the section above, it is also possible that the sequence of the RefSeq RNA was updated after it was aligned to, and used to annotate, the reference sequence. This also might result in discrepancies between the annotation on the genomic sequence, and the current RefSeq RNA.

Multiple chromosomal locations

At times, one gene record may be merged into another gene record. If genes are merged after an annotation is released, there may be more than one location reported on a genomic sequence per GeneID in the Summary report, each resulting from the annotation before the merge.

Representation of nucleotide positions

NCBI uses two conventions to represent the position of features in a sequence.

- offset 0 or 0-based or zero-offset
- offset 1 or 1-based or one-offset

The names are self-explanatory. In the sequence AAAATGCCC, the position of the start codon ATG is 3 in zero-offset and 4 in one-offset. If you find a difference in position information that is 'off-by-one', please review the conventions used in each file.

The zero-offset convention is used in the ASN.1 representation of sequence databases. The ASN.1 of Gene, and the derivative tab-delimited files gene2refseq.gz and gene2accession.gz in the DATA subdirectory of Gene's ftp site also use the convention of 0 offset.

Reports designed for browsing use the convention of one-offset. Thus the position data seen in default HTML views of Gene (and Entrez Nucleotide) are always one greater than that reported in the ASN.1 display.

NOTE: The files in the Map Viewer subdirectories in the genomes path that give position information for genes (seq_gene.md.gz) and other features are one-based. Please be aware of this when processing these files.

Gene and OMIM

Gene integrates information from OMIM, and creates links to OMIM, at two levels:

1. the gene
2. associated disorders or phenotypes

Links provided from the Links menu in the upper right-hand part of the Gene record are based on both types of MIM numbers. Within the body of the record, the MIM number associated with the gene is reported in the **See Related** and **Additional links** sections; a MIM number associated with a disease may be reported in the **Phenotypes** section, along with the name of the condition. Symbols used by OMIM for genes and diseases are intermingled in Gene's **Gene aliases** section.

The gene_info.gz file provided from the Gene ftp site includes the MIM number associated with the gene. If that gene is associated with Mendelian disorders that have a different MIM number, that MIM number will not be provided in gene_info.gz.

Both types of MIM numbers associated with Gene records are reported in the ftp file mim2gene. Data are also provided by OMM at http://omim.org/downloads.

How Gene maintains certain types of information

Conserved Domains

As sequence records are added to or updated in the Protein database, they are compared to records in the Conserved Domain Database (CDD) to identify likely domain content. The results of these analyses for RefSeq proteins are indexed for retrieval in Gene, are displayed when a Gene record is retrieved from Entrez, and are integrated into the ASN.1 that is provided for ftp transfer. The sequence of events is therefore:

- new sequence added to the protein database
- analyzed by the CDD group
- Gene re-indexed

Thus it may require a few days for a new RefSeq accession to display domain information in Gene.

To extract domain information directly for any protein sequence, consider using E-utilities. The url to fetch domain data based on a protein gi follows the pattern:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=gnl|ANNOT:CDD|[put the gi here]&retmode=xml.

Example URL for efetch for CDD:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=gnl|ANNOT:CDD|6978425&retmode=xml

GeneRIFs – How are they maintained?

GeneRIFs are established by three primary methods.

1. Extraction from the published literature by staff of the National Library of Medicine.
2. Summary reports from HuGE Navigator
3. User submissions from an Gene record.

In the first case, the records are updated weekly. In the second case, Gene processes information about how a citation in PubMed is related to a GeneID, and converts that to a standard text. In the last case, RefSeq staff reviews the submission before release, and contacts the submitter if questions arise. User-submitted data should be public within a week.

GeneRIFs – How are they reported on the web?

GeneRIFs are reported from the full report in the Bibliography section. A scrolling window provides unique text of a GeneRIF; the citation or citations that support that statement are available by clicking on the document icon at the left of the GeneRIF. Because the text of GeneRIFs submitted from HuGE Navigator is computed, it is likely that more than one citation will be displayed in PubMed to support that text. Please be certain to note the report of the number of records return by the query, and scroll through the web page to review all the citations.

GeneRIFs – How are they reported on the ftp site?

GeneRIFs are reported from this subdirectory: ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/. In these files, each GeneRIF is reported separately. If there are multiple records for the same gene with the same text, each will be reported from one line in the file. If there are multiple records for the same gene with different text but the same PubMed id, each will be reported from one line in the file.

GO terms

NCBI reports GO terms appropriate for a GeneID by integrating information from the following sources:

- The ftp site of the GO consortium here
- For human only, the GOA ftp site here
- Data provided in sequence submissions.

For all genomes but human, a species-specific gene-identifier (FBgn id, MGI id, RGD ID) is converted to the GeneID. For human, the connection is made from common protein accessions. Most current gaps in the human set, therefore, result from lags in matching protein accessions to GeneIDs. According to Gene's current data flow, any association of a protein accession with more than one gene record must be reviewed by a curator. This multiplicity can be frequently with gene families where multiple genes encode the same protein sequence.

Gene currently reports, and uses for indexed queries, only the explicit GO term or terms assigned to any gene. It does not support querying at any node of the GO graph, nor retrieving all genes that match terms at more specific nodes based on a query at a higher node.

Interactions

Gene represents interaction data as pairs. Gene staff does not curate these data, but does validate identifiers supplied with the source files.

Discontinued records

The full content of discontinued records is indexed for retrieval in Gene. Often, a comment is provided in the summary section indicating why a record was discontinued. If the record is now secondary to another, the link to the current record is provided.

To retrieve all discontinued records, use this query all[filter] NOT alive[prop]

Why can I sometimes display a record, but then cannot retrieve it by a query?

There are two methods by which a gene record can be accessed:

- Directly by a public GeneID
- A query via the Entrez indexing system which returns the list of GeneIDs that satisfy your query.

For recent records, it is possible that the record itself is public, but the indexing of that record is not yet complete so retrieval by Entrez search returns no results. Because Gene re-indexes daily, this discrepancy should last no more than 24 hours.

How can I identify genes with/without a known function?

There are several qualifiers that you might consider using to determine if the function is known or not known. Gene is currently allowing the user to decide which criteria to use, rather than making that decision unilaterally.

- Does the gene encode a protein with a conserved domain?
  Use gene_cdd[filter] to identify those that do or do not.
- Has a GeneRIF been submitted for the gene?
  Use generif[prop] to identify those that do or do not
- If human, is the gene also discussed in the OMIM database?
  Use gene_omim[filter] to identify records also described (or not) in OMIM
- How is the gene named?
  If the full name starts with 'hypothetical', no group has decided how to name this. If the preferred symbol starts with NCRNA, nomenclature groups believe this gene produces a non-coding RNA of unknown function.

  Hypothetical*[title]
  Ncrna*[preferred symbol]

Examples:

1 To find mouse protein-coding genes of unknown function. This query uses the first part of the title of the gene (predicted* or hypothetical*), and excludes those that have a GeneRIF submitted.

mouse[orgn] AND "genetype protein coding"[Properties] AND (hypothetical*[title] OR predicted*[title]) AND alive[prop] NOT generif[prop]

2. To find protein-coding genes from Drosophila melanogaster that do not have a product with a conserved domain in NCBI's conserved domain database:

"drosophila melanogaster"[orgn] AND "genetype protein coding"[Properties] NOT gene_cdd[filter] AND alive[prop]

3. To find non-coding RNAs of unknown function

ncrna*[Preferred Symbol] AND alive[prop]

In What Order Are Exons Presented in ASN.1 and XML?

NCBI's new standard is to report exon location in exon order, i.e. first exon 1, then exon 2, and so on. For genes annotated on the minus strand, this means that the location of the first exon will have a numerical position greater that the second exon, etc.

Example:

```
int {
  from 5140696,
  to 5140737,
  strand minus,
  id gi 62750820
},
int {
  from 5134517,
  to 5134601,
  strand minus,
  id gi 62750820
},
```

This differs from previous reporting in which locations were ordered by sequence position, so that on the minus strand, the last exon was reported first. As genomes are re-annotated, the newer representation will be used, and reporting of exons in sequence order rather than exon order will be deprecated.

For each exon, the range will continue to be reported according to the standard of seq-interval *from* less than seq-interval *to*.

How are wild cards (*) processed?

For Gene, the wild card (*) search is processed by finding the first 5000 terms that match and then ORs them together into a single query. So if you submit a query like 'LOC*', which will match more than 5000 records, you will get a result, but with the warning that not all matches were found:

Wildcard search for 'loc*' used only the first 5000 variations. Lengthen the root word to search for all endings.

Use of wildcards on common word parts consumes many resources, so please use wildcards wisely.

Why are links from Gene to EST not comprehensive?

Gene is not intended to be a comprehensive resource for related nucleotide or protein sequences. As such, only the subset of ESTs that are directly connected to a Gene record are displayed by following the EST link. This connection is made only when an EST is used as a component of a RefSeq RNA or when an EST is used by a

model organism database as the defining Gene sequence. A comprehensive connection between Gene and EST sequences is available by following the UniGene link, and by a regular BLAST query of the EST database.

How does Gene represent genes spanning the origin of replication of a circular genome?

When a gene crosses the origin of replication of a circular genome, the complete genomic RefSeq is displayed in the Genomic Context section of the Full Report as a linear molecule opened at the origin. The appropriate portion of the gene (colored maroon) is shown at each end. The hemE gene (GeneID 4402322) of Rhizobium leguminosarum bv. viciae 3841 is an example, depicted in Figure 1. Each genomic location (A and B) is provided separately in the Genomic regions, transcripts, and products section. A consequence of this rendering is that the proximity of neighboring genes may not be apparent from the Genomic context display. This applies only to the gene spanning the origin of replication; the Genomic context display for the gene's neighbors is not affected by the rendering. Please note that at present, if you select **Open Full View** in the Genomic regions, transcripts, and products section, the gene spanning the origin is depicted across the entire sequence.

What is a readthrough locus and how is it represented?

Gene defines a readthrough locus when transcription continues through the normal transcription termination signal of one locus into an adjacent locus on the same strand. The mature transcript may retain some exons of either locus, and novel exons from the intergenic region may be included. Readthrough transcripts may be non-coding due to nonsense-mediated decay (NMD), may encode a fusion protein derived from exons from one or both loci, or may encode a novel protein product that has no similarity to the proteins of the upstream or downstream loci. Note that Gene elects to use the term readthrough rather than conjoined because loci in this category that have official names provided by nomenclature committees include the word readthrough.

Readthrough events supported by at least two independent lines of transcript and/or publication evidence are generally represented by three GeneIDs: one to represent each of the upstream and downstream loci and one to represent the readthrough products. The third GeneID is represented because the readthrough transcript may not itself be represented accurately by either the upstream or downstream locus alone. Readthrough events are represented by only two GeneIDs if the upstream locus produces an RNA but not a protein product (and is not a pseudogene), and the downstream locus is protein-coding; in this case, the readthrough transcript may encode the same protein as the downstream protein-coding locus and so is considered a transcript variant of the downstream locus.

Genes involved in a readthrough event are reported in the *General gene information* section of a Gene record. If the record represents the readthrough locus (termed the "parent"), links to the included "child" loci are provided. If the record represents a child locus, links to other child ("sibling") loci are reported, and to the readthrough parent, when appropriate. The usage of parent and child terminology in Gene is opposite that used by the ConJoinG database.

Genes involved in readthrough events can be retrieved from Gene by one of the queries:

- readthrough[property]
- readthrough parent[property]
- readthrough child[property]

Readthrough events represented by only two GeneIDs for the reasons described above, and readthrough events lacking sufficient transcript and/or publication evidence to be represented by three GeneIDs, can be retrieved using the query:

- potential readthrough child[property]

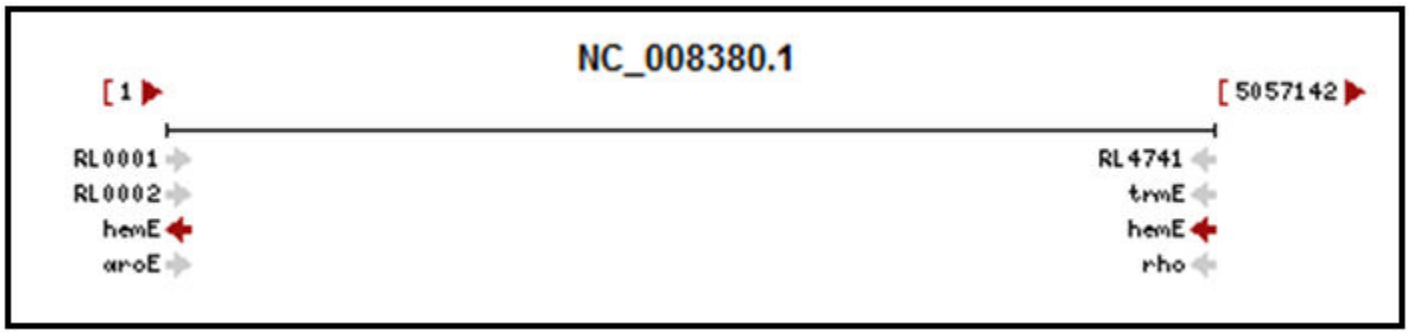

Figure 1. The Genomic Context section for hemE of Rhizobium leguminosarum bv. viciae 3841 (GeneID 4402322), a gene that spans the origin of replication of a circular genome. The complete genomic RefSeq accession is shown as a linearized molecule opened at the origin. The spanning gene is colored in maroon. Note that the relationship between neighboring genes is affected by this rendering when compared to the circular genome.

The gene_group file provided by FTP from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ also reports each pair of loci in a readthrough relationship.

How can I determine the position of genes and exons for my species of interest?

NCBI currently computes the position of genes and exons when an annotation is released. The results are available from the Genomes FTP site, ftp://ftp.ncbi.nlm.nih.gov/genomes/. A file in the latest specification (version 1.20) of Generic Feature Format version 3 (GFF3) is provided for the latest assembly of many organisms. For example, GFF3 files providing the latest annotation of the human genome may be found at ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/GFF/.

How can I retrieve all records for my species of interest?

To find all current (alive) records for a species, query Gene with:

species[organism] AND alive[property], *e.g.*, human[organism] AND alive[property]

Either the species binomial or common name can be used.

If desired, a list of the retrieved GeneIDs can be generated. Use the 'Send to:' feature near the top right hand corner of the results display to output the file; select 'UI' list from the Format menu. Alternatively, this information is recalculated daily and available from Gene's FTP site. Some species, including human, have a species-specific gene_info file:

ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

The README in the DATA directory provides more information about the contents of the gene_info file.

This information can also be obtained programmatically using E-Utilities. Combine the use of ESearch (to obtain the set of GeneIDs matching your query) with EFetch or ESummary to extract the desired data.

How can I identify genes that have related pseudogenes?

Use the property "has_pseudogene" to query Gene. For example, to find current (alive) human genes having related pseudogenes:

Human[organism] AND alive[property] AND has_pseudogene[property]

Either the species binomial or common name can be used.

How can I find all genes located within a specific region of a chromosome?

Several options exist.

1 Query Gene, including the two location subcategory fields, **chromosome [chr]** and **base position [chrpos],** in the query. To find current genes located from base position 1 to 500000 on human chromosome 1, try:

1[chr] AND 1:500000[chrpos] AND human[orgn] AND alive[prop]

Note that base position is supported only for genomic accessions of an organisms' reference genome assembly, and only for genomes where chromosome coordinates are defined. See Table 5 in Gene Help for additional information on using these fields.

2. Use Limits. In the 'Limit by Chromosomal Region' section, select 'Homo sapiens' and enter the desired values in the Chromosome and From/To boxes that appear. Choose any other Limits desired and click the Search button at the bottom of the page.

3. Use the ESearch function of E-utilities. For example,

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=1[chr]+AND+1:500000[chrpos]+AND+human[orgn]+AND+alive[prop]

Why does the number of GeneRIFs displayed in the Bibliography section differ from the number of PubMed IDs reported using the PubMed(GeneRIF) link?

The number of GeneRIFs displayed in the Bibliography section excludes those that describe interactions, which are provided separately in the Interactions section. The PubMed(GeneRIF) link provides a listing of all PubMed IDs that are associated with GeneRIFs AND interaction data for the GeneID. Additionally, a 1:1 relationship between PubMed IDs and GeneRIFs cannot be assumed; some PubMed IDs are referenced by more than one GeneRIF, and some GeneRIFs refer to more than one PubMed ID.

Why did many bacterial GeneIDs disappear?

The scope definition for annotating prokaryotic genomes with NCBI GeneIDs was changed resulting in the suppression of a large number of entries. Gene continues to provide current entries for all prokaryotic reference genomes. For bacteria, Gene includes the best supported sub-set of representative genomes for which there are >= 10 sequenced genomes for the clade. We are still refining the Gene policy for archaeal genomes. Suppressed Gene entries can still be accessed and we are supplementing the current set of suppressed records with information to facilitate navigating to the replacement non-redundant RefSeq protein. The embedded graphical display will continue to show annotation of the genomic coordinates that the Gene entry represents. If that RefSeq genome was re-annotated, then the display in Gene will automatically show the updated annotation for the accession.version:from-to coordinates associated with the Gene record. Thus, while the Gene entry may have initially displayed a CDS annotation associated with a YP_ accession number (still reported in the Reference Sequences section of the record), it may now display a CDS annotation associated with a non-redundant WP_ accession number. These records will not be subject to any further update.

For Programmers and Database Developers

1. How to connect your database to Gene--Using LinkOut
2. How to construct URLS to connect to Gene
3. Relationship of LocusID to GeneID
4. The Gene ftp site
5. Gene-related ftp sites
6. Extracting Gene in XML format
7. Unzipping Compressed ASN.1 Binary Format FTP Files
8. How to extract the Summary text from records in Gene

Using LinkOut

Because Gene is an Entrez database, database providers can now use the LinkOut mechanism to direct users of Gene to related sites providing more information about a particular record. The benefits to data providers are several:

- The provider controls making and removing connections between Gene and the provider's web site.
- The provider's web site may receive additional traffic because of links from users of Gene.

This area of LinkOut's documentation provides instructions geared more to the non-bibliographic data providers.

How to construct URLs to link to Gene

Because Gene is an Entrez database, URLs can be constructed using standard Entrez methods. The standard URL format consists of the base URL for the database followed by options that can specify the record to be displayed, display options, and search terms. To construct a URL to display a specific Gene record, combine this base URL

http://www.ncbi.nlm.nih.gov/gene/

with the GeneID. For example, to link to GeneID 1, use this URL:

http://www.ncbi.nlm.nih.gov/gene/1

This displays the Gene record in the default Full Report display setting. Additional display options, including Gene Table, GeneRIF, and Summary (docsum), can be requested using the **?report** retrieval parameter. For example, to link to the Gene Table report format for GeneID 1, use this URL:

http://www.ncbi.nlm.nih.gov/gene/1/?report=gene_table

To construct a URL that queries Gene,**?term=[search term]** is added to the base URL. For example, to search Gene for the term 'hypertension', use this URL:

http://www.ncbi.nlm.nih.gov/gene/?term=hypertension

For complex combinations of query terms, it may be helpful to use the Advanced search to help you build the query, and then save the URL that query generates. Try the following steps:

1. use the Advanced search link at the top of a Gene page to build a complex query
2. when completed, follow the **Details** link at the right of the Search Box
3. click the **URL** button
4. use the full URL provided in your web b<rowser's navigation bar

To view the complete list of Gene-specific **Properties** and **Filters** used to build more complex queries, including current counts for each in the database, follow these steps:

1. use the Advanced search link at the top of a Gene page
2. select **Properties** or **Filters** from the All fields menu
3. click on **Index** and navigate through the options.

Gene ftp site

The Gene ftp site provides two major types of reports:

- tab-delimited files matching GeneIDs to citation, accession, and name information
- a comprehensive extraction

The README file in the **gene** directory provides more detailed information. See also the Gene-OMIM faq above for more information about MIM numbers provided in gene_info.gz and mim2gene.

The comprehensive extraction is provided in ASN.1format in the DATA/ASN_BINARY directory. In addition to the comprehensive file **All_Data.ags.gz**, there are subdirectories divided by taxonomic nodes. Each of these sub-directories contains a comprehensive extraction for that node but may also contain some species-specific files. For example, Mammalia contains these files:

File name	Content
All_Mammalia.ags.gz	Gene records for mammals, including mitochondria.
Bos_taurus.ags.gz	Gene records for *Bos taurus*, including mitochondria.
Canis_familiaris.ags.gz	Gene records for *Canis familiaris*, including mitochondria.
Homo_sapiens.ags.gz	Gene records for *Homo sapiens*, including mitochondria.
Mus_musculus.ags.gz	Gene records for *Mus musculus*, including mitochondria.
Pan_troglodytes.ags.gz	Gene records for *Pan troglodytes*, including mitochondria.
Rattus_norvegicus.ags.gz	Gene records for *Rattus norvegicus*, including mitochondria.
Sus_scrofa.ags.gz	Gene records for *Sus scrofa*, including mitochondria.

GeneRIFs and Interaction data

Data associated with GeneRIFs, HIV-1 Interactions, and General Interactions are available from the GeneRIF ftp site.

Gene-related ftp sites

There are other ftp sites at NCBI that contain gene-related information. These include:

1. Map Viewer
   Within a genome-specific directory in the path ftp://ftp.ncbi.nlm.nih.gov/genomes/, click on maps, then mapview, then the folder for the current build. In that directory you should find the file seq_gene.md. The gene lines in this file give the ranges for the gene in chromosome (as applicable) and contig coordinates. For example, a command like
   gzcat seq_gene.md | egrep "GENE.*reference" will extract the 'GENE' lines for the reference assembly.
   - The first line in the file names the columns.
   - chrStart, chrEnd and orientation refer to the chromosome.
   - cnt_start, cnt_stop, cnt_orient refer to the contig

2. UniGene
3. UniSTS

Extracting Gene in XML format

If you prefer to use reports formated in XML rather than ASN.1, you have several options:

1. E-Utilities
2. gene2xml
3. Web Entrez

E-Utilities

Try the robust functions provided via E-utilities. A common approach is to combine use of ESearch to obtain a set of GeneIDs of interest with EFetch to retrieve records by GeneID. The document EFetch for Sequence and other Molecular Biology Databases provides more information about how to set the parameters for extracting information from Entrez databases. It is as simple as:

- defining **db** as gene
- defining **retmode** as xml, if needed
- defining **id** as the GeneID of interest

Example using EFetch to retrieve the full XML record for GeneID 2:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=gene&id=2&retmode=xml

Example using ESummary to retrieve the document summary (docsum) in XML format for a list of GeneIDs (by default, retmode=xml):

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=gene&id=19,11303,313210,373945,378973,464631

Example using ESearch to search for genes by symbol (by default, retmode=xml):

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=BRCA1&sort=relevance

The results from ESearch are not sorted by default, so use sort=relevance to sort the results the same as the default sort order used in searching Gene on the web. Other sort options are sort=weight, sort=name, and sort=chromosome.

A representative perl script using both ESearch and retrieval from ESummary is provided from the ftp site as taxidToGeneNames.pl. It uses NCBI's Taxonomy database identifier to support species-specific extraction of information incorporated in the Gene Summary display format.

Examples:

- taxidToGeneNames.pl -t 9606 -o xml --reports data from the summary for human genes with output as XML
- taxidToGeneNames.pl -t 10090 -o tab --reports GeneID, symbol, full name from the summary for mouse in tab-delimited output

gene2xml

The tool gene2xml, described here, converts the ASN.1 provided in binary set format (in the ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ASN_BINARY directory), into XML. It also converts ASN in the binary format into concatenated text.

A new version of gene2xml is provided when there are changes in the Gene ASN.1 structure. If you are using an older version of gene2xml with current data, you may encounter errors, in which case you should check the version of gene2xml that you are using and see if it is the latest version.

Gene

Entrez supports reporting any record or set of records in XML format. After you have retrieved record(s) of interest, select XML from **Display Settings** and the result will be displayed according to Gene's DTD. You can then send that result to a file.

Note: to convert multiple records to XML via the Entrez interface, check the boxes to left of the gene symbol in the query result view.

Unzipping Compressed ASN.1 Binary Format FTP Files

The ftp files in the ASN_BINARY subdirectory of Gene's ftp site are binary concatenated gzip files. This type of content is defined in the specification RFC-1952:

"2.2. File format

A gzip file consists of a series of "members" (compressed data sets). The format of each member is specified in the following section. The members simply appear one after another in the file, with no additional information before, between, or after them."

This specification can be found at the Internet Engineering Task Force web site at http://www.ietf.org/rfc/rfc1952.txt.

If you are developing applications to decompress Gene's ASN.1 binary format ftp files, be sure that any compression library that you are using supports this standard. For example, there is a known issue with the compression library in Microsoft ® .NET Framework 3.5 which does not support decompressing this type of content. For further information about this issue, see http://connect.microsoft.com/VisualStudio/feedback/ViewFeedback.aspx?FeedbackID=357758

How to extract the Summary text from records in Gene

One of the following methods could be used:

1 Use **geneDocSum.pl**, a Perl program freely available for download from ftp://ftp.ncbi.nih.gov/gene/tools/. Instructions and options are provided in the accompanying README file. Test what you want to retrieve via the web site, and then use that query as input to the program. To illustrate its use, all current (alive) human records that include a Summary can be retrieved by running:

geneDocSum.pl -q "has_summary[prop] AND human[orgn]" -o tab -t Name -t Summary

2. Use the ESearch and EFetch functions of E-utilities.

Using ESearch, identify the GeneIDs of interest that include a Summary. For example, to retrieve all current (alive) human records with a Summary:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=has_summary[prop]+AND+human[orgn]+AND+alive[prop]

By default, ESearch returns only 20 records. You can increase that count by redefining the maximum:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=gene&term=has_summary[prop]+AND+human[orgn]+AND+alive[prop]&retmax=5000

Use EFetch to retrieve the full records corresponding to the GeneIDs retrieved by ESearch. Input to EFetch may be either a single or comma-delimited list of UIDs, or come from the **Entrez History Server**.

For more details about how to use E-utility functions, please refer to Entrez Programming Utilities Help. You will note there are sections on Downloading Full Records as well as sample applications.