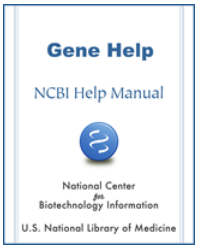
U.S. National Library of Medicine
National Center for Biotechnology Information

NLM Citation: Gene Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. Gene Frequently Asked Questions. 2008 Apr 21 [Updated 2018 Apr 9].
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



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.

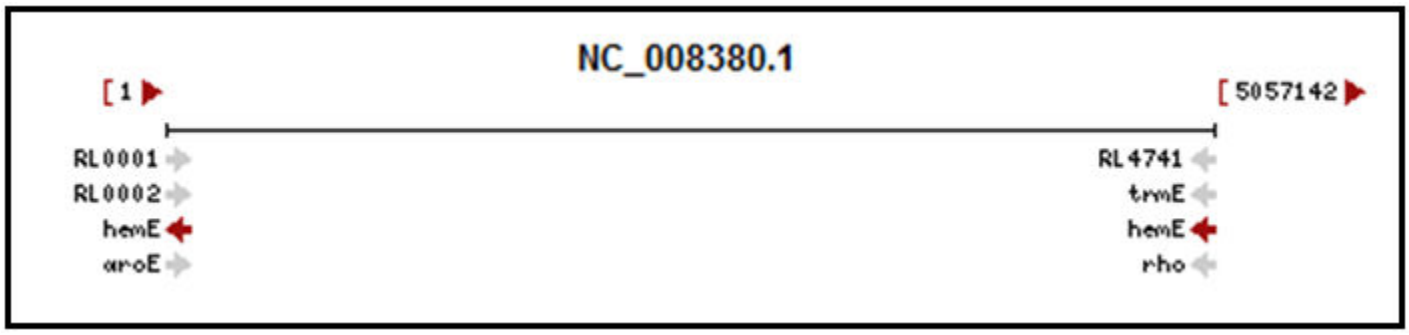


Figure 1. The Genomic Context section for hemE of Rhizobium leguminosarum bv. viciae 3841 (GeneID 4402322), a gene that spans the origin of replication of a circular genome. The complete genomic RefSeq accession is shown as a linearized molecule opened at the origin. The spanning gene is colored in maroon.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

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]

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.