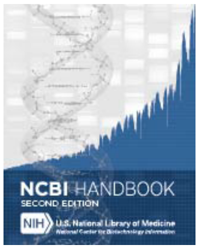
NLM Citation: Maglott DPhD, Pruitt KPhD, Tatusova TPhD, et al. Gene. 2013 Nov 14. In: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.
Bookshelf URL: https://www.ncbi.nlm.nih.gov/books/



Gene

Donna Maglott, PhD, Kim Pruitt, PhD, Tatiana Tatusova, PhD, and Terence Murphy, PhD

Created: November 14, 2013.

Scope

NCBI's Gene database is designed to aggregate gene-specific information from multiple perspectives including sequence, mapping, publications, function of products, expression, evolution, and consequences of variation. Gene makes these data available for diverse scenarios, from occasional interactive access on the Web through computational access to selected or complete data sets.

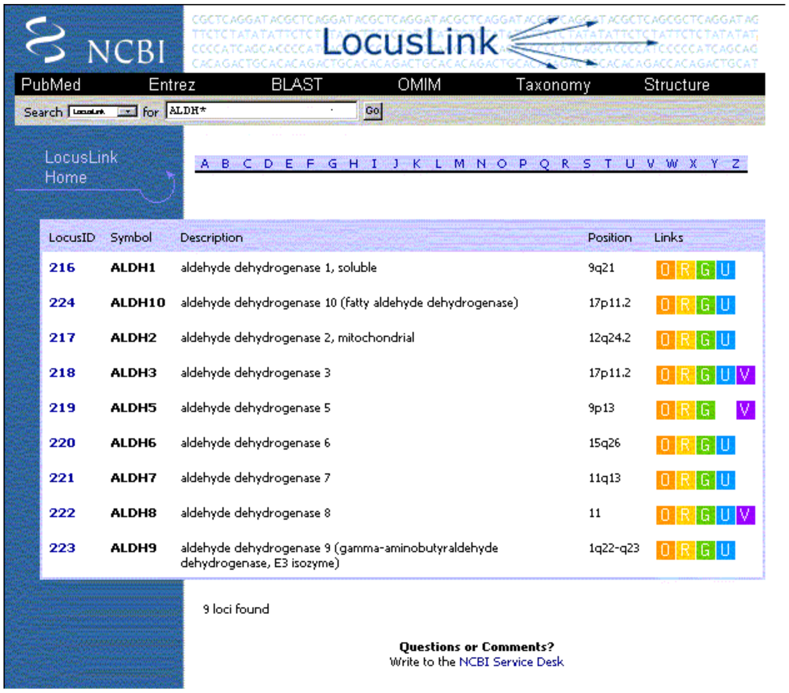
Gene assigns an identifier (the GeneID) for each gene in each taxon either represented in the NCBI Reference Sequence (RefSeq) project, or under consideration by RefSeq. Usually this taxon is defined at the species level, but sometimes will be per isolate, strain or cultivar. Gene is closely coupled with RefSeq, in that genes annotated on RefSeq sequences are assigned GeneIDs for tracking. Not all records in Gene, however, are based on RefSeqs. Gene works closely with multiple groups that may identify a gene before it has been been defined by sequence. In other words, some records in Gene are mapped traits or other phenotypes.

This document does not provide detailed instructions about how to use Gene or comprehensive details about how Gene is built from numerous data sources. For detailed, up- to-date documentation, please refer to Gene's Help document.

History

The database currently known as Gene was first made public in 1999 as LocusLink (1). There was only one species represented (human) and little more than 9000 records. The Web interface supported links only to dbSNP, OMIM, RefSeq, GenBank, and UniGene within NCBI, as well as to the now defunct Genome Database (GDB) and a few other databases externally (Figure 1). By late 2003, when Entrez Gene was released, there were 10 species, almost 195000 records, and links computed to dbSNP, Ensembl, the HUGO Gene Nomenclature Committee (HGNC), GEO, Map Viewer, Mammalian Gene Collection (MGC), Nucleotide, Protein, PubMed, Taxonomy, UCSC, UniSTS, UniGene, and multiple species-specific model organism databases (Figure 2). Now Gene represents more than 11,000 taxa, more than 13,000,000 records, and more than 40 types of links to other NCBI databases.

In addition to the taxonomic scope, the number of records, and connectivity, the look and feel of Gene has changed over the years. The current database implementation provides a hierarchical Table of Contents to facilitate navigation, integration with MyNCBI to support personalized display of sections of the record, an embedded viewer of NCBI's annotation of any gene on one or more genomic RefSeqs, a page dedicated to the display of GeneRIFs, and, especially for human, enhanced access to gene-specific variation and phenotype reports (Figure 3).



LocusID	Symbol	Description	Position	Links
216	ALDH1	aldehyde dehydrogenase 1, soluble	9q21	O R G U
224	ALDH10	aldehyde dehydrogenase 10 (fatty aldehyde dehydrogenase)	17p11.2	O R G U
217	ALDH2	aldehyde dehydrogenase 2, mitochondrial	12q24.2	O R G U
218	ALDH3	aldehyde dehydrogenase 3	17p11.2	O R G U V
219	ALDH5	aldehyde dehydrogenase 5	9p13	O R G V
220	ALDH6	aldehyde dehydrogenase 6	15q26	O R G U
221	ALDH7	aldehyde dehydrogenase 7	11q13	O R G U
222	ALDH8	aldehyde dehydrogenase 8	11	O R G U V
223	ALDH9	aldehyde dehydrogenase 9 (gamma-aminobutyraldehyde dehydrogenase, E3 isozyme)	1q22-q23	O R G U

9 loci found

Questions or Comments?
Write to the NCBI Service Desk

Figure 1. Representation of gene-specific information in LocusLink.

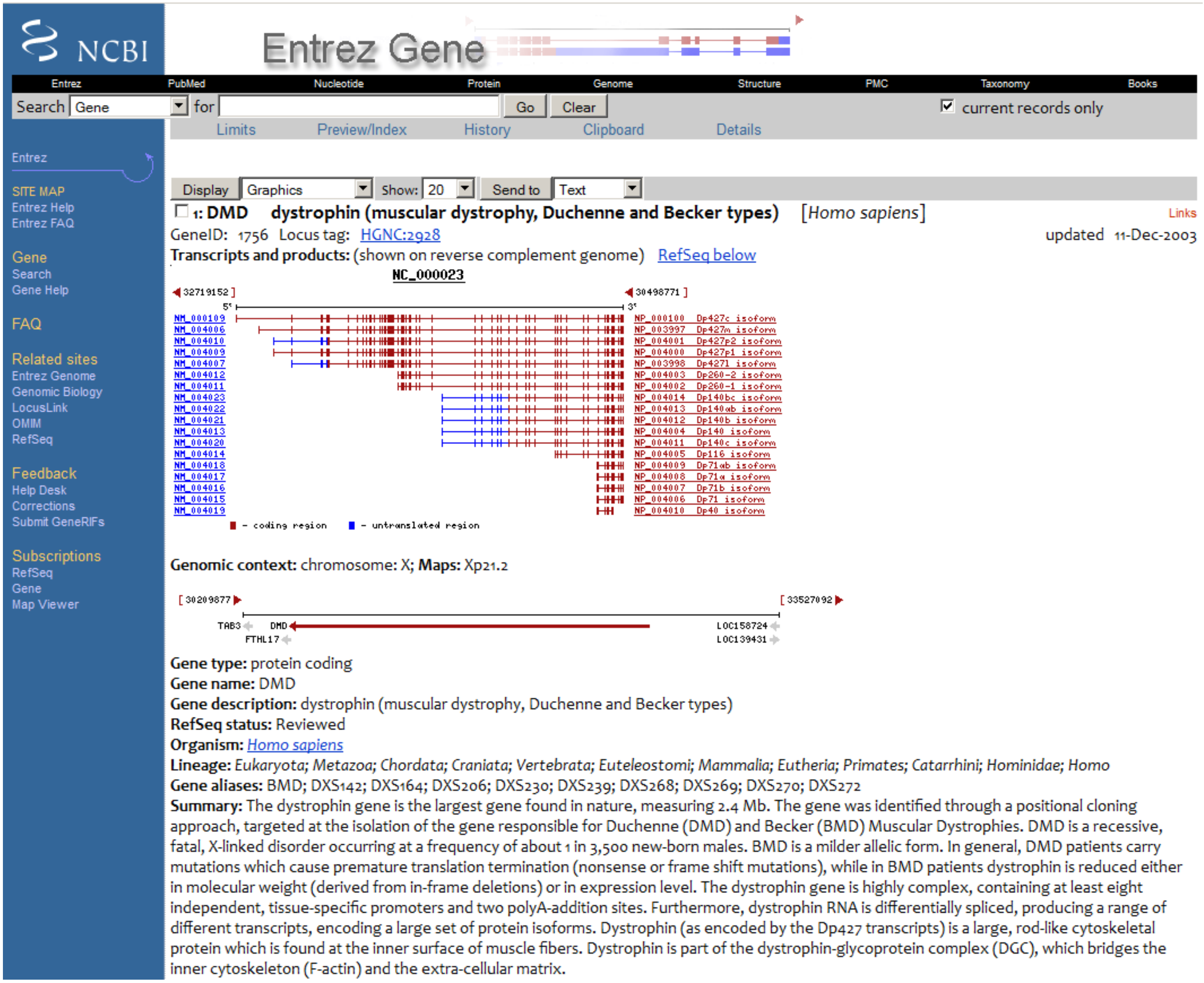
Data Model

Gene has a simple data model. Once the concept of a gene is defined by sequence or mapped location, it is assigned a unique integer identifier or GeneID. Then data of particular types are connected to that identifier. These types include sequence accessions, names, summary descriptions, genomic locations, terms from the Gene Ontology Consortium (2), interactions, related phenotypes, and summaries of orthology. For some of the commonly requested elements, and because of the simplicity of the data model, Gene provides tab-delimited files of content anchored on the GeneID.

The full extraction of Gene is exported as binary ASN.1 (ftp://ftp.ncbi.nih.gov/gene/DATA/ASN_BINARY/) with a tool provided to convert to XML (ftp://ftp.ncbi.nlm.nih.gov/asn1-converters/by_program/gene2xml). The ASN.1 representation of a Gene record (http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene.asn) incorporates several objects used by other resources (Gene-ref, BioSource, RNA-ref, etc.), but also has several several objects specific to Gene to represent the type of gene, map location, and properties. A major component of Gene's ASN.1 representation is the generic Gene-commentary that is used to represent content defined by type, heading, label, text, and source.

Dataflow

Gene is updated daily and incrementally. In other words, on any given day a record may be changed but not all records will be changed on the same day. The FTP site is refreshed comprehensively each day, except for special reports and documentation files.


NCBI
Entrez Gene
Entrez
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Search
Gene
for
Go
Clear
☑ current records only
Limits
Preview/Index
History
Clipboard
Details
Entrez
SITE MAP
Entrez Help
Entrez FAQ
Gene
Search
Gene Help
FAQ
Related sites
Entrez Genome
Genomic Biology
LocusLink
OMIM
RefSeq
Feedback
Help Desk
Corrections
Submit GeneRIFs
Subscriptions
RefSeq
Gene
Map Viewer
Display
Graphics
Show:
20
Send to
Text
Links
NC_000023
32719152
30498771
5'
3'
NM_000109 NP_000100 Dp427c isoform
NM_004006 NP_003997 Dp427m isoform
NM_004010 NP_004001 Dp427p2 isoform
NM_004009 NP_004000 Dp427p1 isoform
NM_004007 NP_003998 Dp427l isoform
NM_004012 NP_004003 Dp260-2 isoform
NM_004011 NP_004002 Dp260-1 isoform
NM_004023 NP_004014 Dp140bc isoform
NM_004022 NP_004013 Dp140ab isoform
NM_004021 NP_004012 Dp140b isoform
NM_004013 NP_004004 Dp140 isoform
NM_004020 NP_004011 Dp140c isoform
NM_004014 NP_004005 Dp116 isoform
NM_004018 NP_004009 Dp71ab isoform
NM_004017 NP_004008 Dp71a isoform
NM_004016 NP_004007 Dp71b isoform
NM_004015 NP_004006 Dp71 isoform
NM_004019 NP_004010 Dp40 isoform
🟥 - coding region
🟦 - untranslated region
30209877
33527092
TAB3
DMD
FTHL17
LOC158724
LOC139431

□ 1: DMD dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]

GeneID: 1756 Locus tag: HGNC:2928 updated 11-Dec-2003

Transcripts and products: (shown on reverse complement genome) RefSeq below

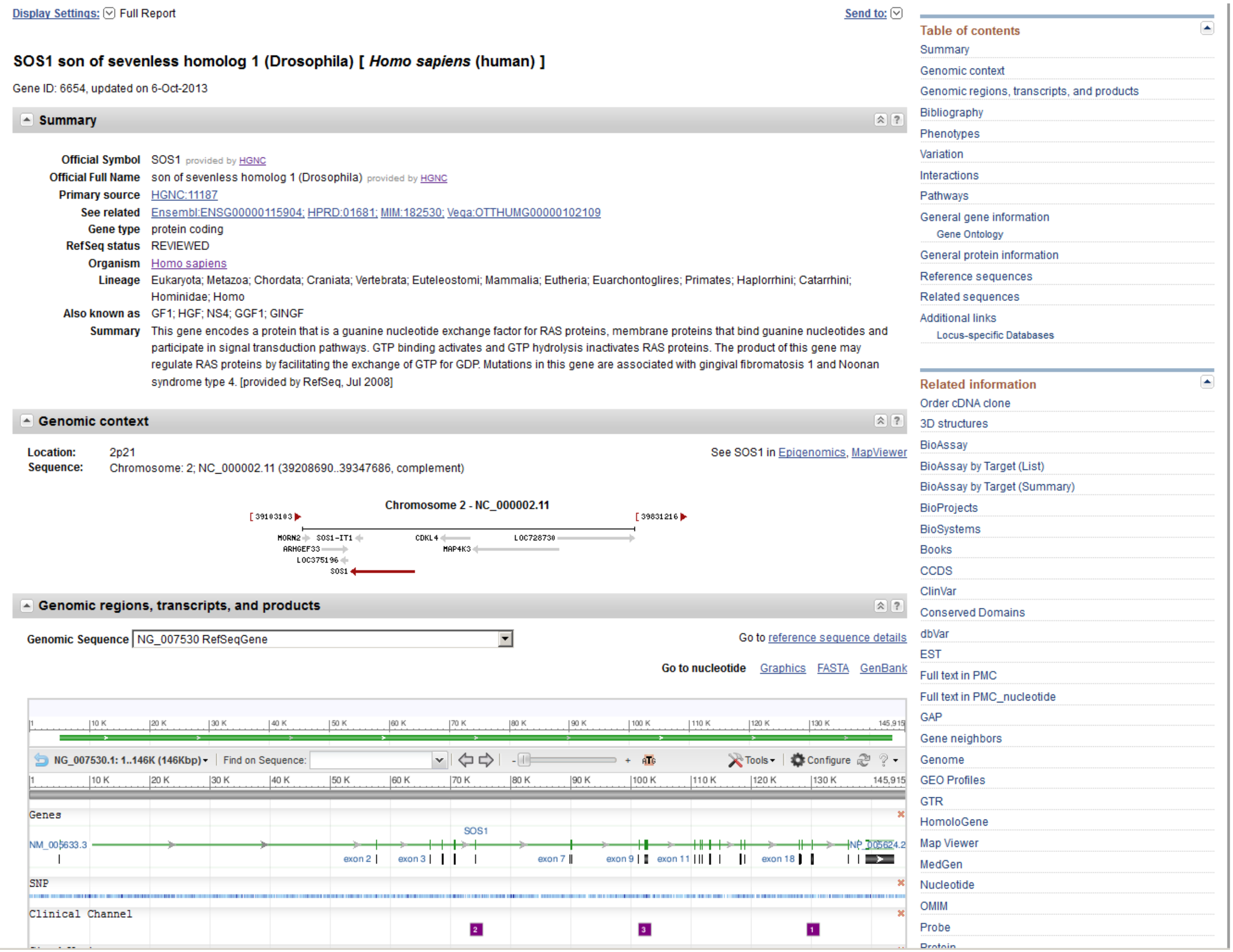
Genomic context: chromosome: X; **Maps:** Xp21.2

Gene type: protein coding
Gene name: DMD
Gene description: dystrophin (muscular dystrophy, Duchenne and Becker types)
RefSeq status: Reviewed
Organism: *Homo sapiens*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo*
Gene aliases: BMD; DXS142; DXS164; DXS206; DXS230; DXS239; DXS268; DXS269; DXS270; DXS272
Summary: The dystrophin gene is the largest gene found in nature, measuring 2.4 Mb. The gene was identified through a positional cloning approach, targeted at the isolation of the gene responsible for Duchenne (DMD) and Becker (BMD) Muscular Dystrophies. DMD is a recessive, fatal, X-linked disorder occurring at a frequency of about 1 in 3,500 new-born males. BMD is a milder allelic form. In general, DMD patients carry mutations which cause premature translation termination (nonsense or frame shift mutations), while in BMD patients dystrophin is reduced either in molecular weight (derived from in-frame deletions) or in expression level. The dystrophin gene is highly complex, containing at least eight independent, tissue-specific promoters and two polyA-addition sites. Furthermore, dystrophin RNA is differentially spliced, producing a range of different transcripts, encoding a large set of protein isoforms. Dystrophin (as encoded by the Dp427 transcripts) is a large, rod-like cytoskeletal protein which is found at the inner surface of muscle fibers. Dystrophin is part of the dystrophin-glycoprotein complex (DGC), which bridges the inner cytoskeleton (F-actin) and the extra-cellular matrix.

Figure 2. Gene in 2003. The diagram of the gene structure was idiosyncratic to Gene; the organization of the page followed the NCBI conventions of the time by using a blue sidebar at the left to provide general information about Gene and other resources. Links to related data in other databases was accessed by clicking on the Links menu at the upper right.

Data are added to Gene by integrating automated and curatorial flows. For some taxa, primarily genomes submitted to NCBI with annotation of genes, data are loaded to Gene by extracting information annotated on the gene feature of the genomic sequence that was submitted. Those data may be supplemented by data from Gene Ontology (GO) based on identifiers in the sequence, according to rules reported by Gene in ftp://ftp.ncbi.nih.gov/gene/DATA/go_process.xml. The content of the Gene record for these species is thus updated only when a new annotation of the genome is supplied, or when supplementary data such as GeneRIFs, GO terms, or UniGene clusters are updated.

For the taxa included in RefSeq's curated set (see the RefSeq chapter for more information), updates may happen daily, and independently of a re-annotation of a genome. There are automated flows to reconcile official gene symbols and full names, protein names, and database identifiers. Curators may modify summaries, add or redefine transcript or RefSeqGene RefSeqs, or add citations to the record. When this happens the Gene record is updated. Genes that are in scope for more frequent updates can usually be detected because the NCBI Reference



Display Settings: Full Report

Send to:

SOS1 son of sevenless homolog 1 (Drosophila) [*Homo sapiens* (human)]

Gene ID: 6654, updated on 6-Oct-2013

Summary

Official Symbol	SOS1 provided by HGNC
Official Full Name	son of sevenless homolog 1 (Drosophila) provided by HGNC
Primary source	HGNC:11187
See related	Ensembl:ENSG00000115904; HPRD:01681; MIM:182530; Vega:OTTHUMG00000102109
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	GF1; HGF; NS4; GGF1; GINGF
Summary	This gene encodes a protein that is a guanine nucleotide exchange factor for RAS proteins, membrane proteins that bind guanine nucleotides and participate in signal transduction pathways. GTP binding activates and GTP hydrolysis inactivates RAS proteins. The product of this gene may regulate RAS proteins by facilitating the exchange of GTP for GDP. Mutations in this gene are associated with gingival fibromatosis 1 and Noonan syndrome type 4. [provided by RefSeq, Jul 2008]

Genomic context

Location: 2p21

Sequence: Chromosome: 2; NC_000002.11 (39208690..39347686, complement)

See SOS1 in Epigenomics, MapViewer

Genomic regions, transcripts, and products

Genomic Sequence NG_007530 RefSeqGene

Go to reference sequence details

Go to nucleotide Graphics FASTA GenBank

Related information

Order cDNA clone
3D structures
BioAssay
BioAssay by Target (List)
BioAssay by Target (Summary)
BioProjects
BioSystems
Books
CCDS
ClinVar
Conserved Domains
dbVar
EST
Full text in PMC
Full text in PMC_nucleotide
GAP
Gene neighbors
Genome
GEO Profiles
GTR
HomoloGene
Map Viewer
MedGen
Nucleotide
OMIM
Probe
Protein

Figure 3. Gene in 2013. Partial display of a record in Gene showing content comparable to that in Figure 2, namely the summary section, the genomic context, and part of the embedded view of the annotation of the gene on a selected genomic sequence. In this example the genomic sequence is a RefSeqGene, and thus shows a more limited set of alternative transcipts, and report the exon numbering system defined by the RefSeqGene.

Sequences section will include a subsection entitled RefSeqs maintained independently of Annotated Genomes (Figure 4).

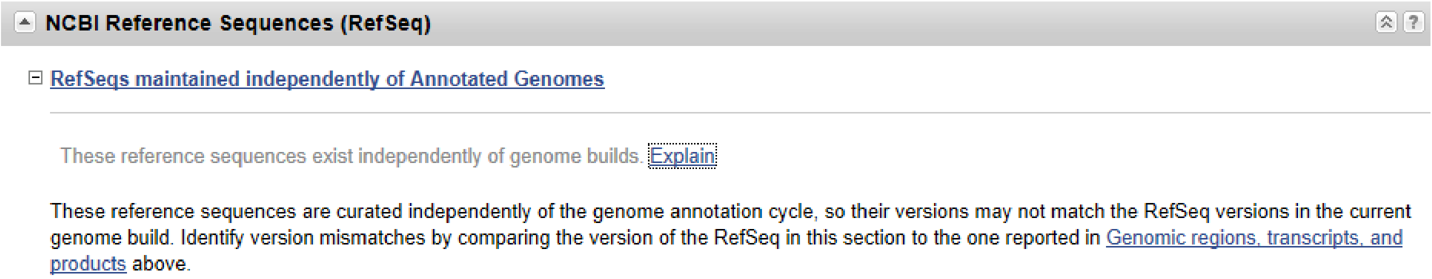
More detailed information about the maintenance of information in Gene is provided in Gene Help.

Access

Web

Gene

Gene is accessed on the Web via http://www.ncbi.nlm.nih.gov/gene/. If the GeneID is known, the path to a specfic record is generated based on the root path plus the GeneID, e.g., http://www.ncbi.nlm.nih.gov/gene/672 for human BRCA1, for which the GeneID is 672.



NCBI Reference Sequences (RefSeq)

RefSeqs maintained independently of Annotated Genomes

These reference sequences exist independently of genome builds. Explain

These reference sequences are curated independently of the genome annotation cycle, so their versions may not match the RefSeq versions in the current genome build. Identify version mismatches by comparing the version of the RefSeq in this section to the one reported in Genomic regions, transcripts, and products above.

Figure 4. A record maintained independently of annotation releases. If this information is included in the Reference Sequences portion of the Gene record, other content of the record is also likely to change more often.

Gene's website is searched via NCBI's Entrez system. The fields, filters, and properties that support effective queries are documented in Gene's Help book. Among those that are used most often are the gene symbol ([gene]) and a sequence accession.

Other NCBI databases

Gene is also accessed from other databases at NCBI. For example, a query to sequence databases, ClinVar, MedGen, or PubMed will detect what looks like a Gene symbol, and provide a display summarizing what is available in Gene (Figure 5). Records in Gene related to other database entries can be identified by following the links to Gene displayed in the panel at the right.

FTP

Information about genes is accessible from any FTP site of NCBI that includes GeneIDs as part of the content. These will not be enumerated in this document; just be aware that if a record reports a GeneID or gene_id, that is the identifier from NCBI's Gene database.

Gene

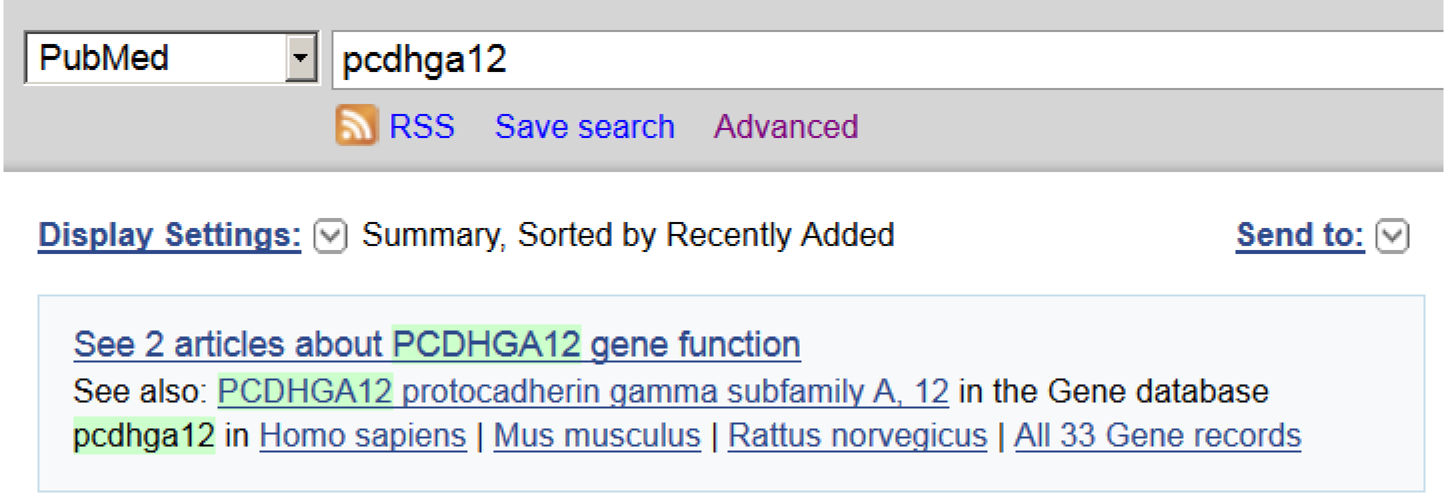
Gene's FTP site (ftp://ftp.ncbi.nlm.nih.gov/gene/) is divided into DATA, GeneRIF, and tools sections. The ftp://ftp.ncbi.nlm.nih.gov/gene/README file describes all sections, reports maintenance details, and provides detailed information about files available from Gene, as well as the annotation-specific files provided from ftp://ftp.ncbi.nih.gov/genomes. The DATA subdirectory provides several comprehensive files, but also includes subdirectories for the full extractions (ASN_BINARY) and tab-delimited reports (GENE_INFO) that provide subsets of data divided by major taxonomic groups.

GFF

For those interested in the location of genes and exons in a genomic context, the genomes path provides a GFF directory for many species. The README_CURRENT_RELEASE file indicates the NCBI Annotation Release being reported and the dates on which data were frozen to support the annotation. In the GFF file, GeneID is reported as a cross-reference, e.g., Dbxref=GeneID:1080. NCBI uses the GFF3 standard.

E-Utilities

Gene is fully accessible programmatically using NCBI's E-Utilities. The tools section on Gene's FTP site provides some sample perl scripts to extract information from Gene based on esummary and efetch and elink.



Summary, Sorted by Recently Added

See 2 articles about PCDHGA12 gene function

See also: PCDHGA12 protocadherin gamma subfamily A, 12 in the Gene database

pcdhga12 in Homo sapiens | Mus musculus | Rattus norvegicus | All 33 Gene records

Figure 5. Gene sensor in PubMed. A query that matches a gene symbol provides the user with link to more information in Gene, as well as the listing of citations in PubMed that satisfy the query (http://www.ncbi.nlm.nih.gov/pubmed/?term=pcdhga12)

Related Tools

In addition to the scripts available from the tools directory of Gene's FTP site (ftp://ftp.ncbi.nlm.nih.gov/gene/tools/README), gene2xml (ftp://ftp.ncbi.nlm.nih.gov/asn1-converters/by_program/gene2xml/) supports conversion of Gene's ASN.1 representation to XML. Gene-related programming tips are included in Gene Help and FAQ.

References

1. Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. Nucleic Acids Res. 2000;Jan 128(1):126–8. PubMed PMID: 10592200.
2. Gene Ontology Consortium.Blake JA, Dolan M, Drabkin H, et al Gene Ontology annotations and resources. Nucleic Acids Res. 2013;Jan41(Database issue):D530–5.doiEpub 2012 Nov 17 doi: 10.1093/nar/gks1050. PubMed PMID: 23161678.