



The Kalmanovitz Library and  
The Center for Knowledge Management

# Entrez Hands On

Gilberto da Gente  
Bioinformatics Specialist  
June 12<sup>th</sup> 2008

University of California, San Francisco



# Filter

The term **filter** is used to describe categories of records grouped based on their relationship either to other Entrez databases or to external resources that have submitted LinkOut connections.

<u>Filter name</u>	<u>Definition</u>
all	total records, current or not
gene all	all current records
gene books	Gene records with explicit links to Entrez Books
gene gensat	Gene records with explicit links to Entrez GenSAT
gene geo	Gene records with explicit links to Entrez GEO
gene homologene	Gene records with explicit links to Entrez HomoloGene
gene nucleotide	Gene records with explicit links to Entrez nucleotide, excluding RefSeq chromosome or contig accessions
gene nucleotide pos	Gene records with explicit links to Entrez nucleotide, limited to those of RefSeq chromosome or contig accessions, and thus including position data

# Filter

<u>Filter name</u>	<u>Definition</u>
gene omim	Gene records with explicit links to Entrez OMIM, and thus includes links to both disease and 'gene' records in OMIM
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 genes variation information available in dbSNP
gene taxonomy	Gene records with explicit links to Entrez Taxonomy
gene unigene	Gene records with explicit links to Entrez UniGene
gene unists	Gene records with explicit links to Entrez UniSTS (marker data)

# Properties

**Properties** are assigned to gene records based on content, rather than relationship to other database records

## Property categories

1. type of gene
2. source of the gene
3. type of RefSeq provided for the gene
4. other

# Properties: gene type

1. **type of gene**: property named as *genetype name\_of\_type*

<u>Property name</u>	<u>Explanation</u>
genetype miscrna	gene encodes an RNA not in any of the specifics below
genetype other	of know type, but not any of the specific known categories
genetype protein coding	encodes a protein
genetype pseudo	pseudogene
genetype rna	encodes ribosomal RNA
genetype scrna	encodes small cytoplasmic RNA
genetype snorna	encodes small nucleolar RNA
genetype srna	encodes small nuclear RNA
genetype trna	encodes transfer RNA
genetype unknown	the type of gene is not known

# Properties: gene source

2. **source of the gene**: property named as *source name\_of\_source*

<u>Property name</u>	<u>Explanation</u>
source extrachromosomal	located extrachromosomally
source genomic	located on a chromosome
source mitochondrion	located in the mitochondrion
source other	location not included in other specifics
source organelle	located in an organelle (includes mitochondrion and plastid)
source plasmid	located in a plasmid
source plastid	located in a plastid
source proviral	located in a provirus
source virion	located in a virion

# Properties: refseq comment

LOCUS NM\_001114313 3693 bp mRNA linear VRT 06-FEB-2008  
DEFINITION Danio rerio HEAT repeat containing 6 (heatr6), mRNA.  
ACCESSION NM\_001114313 XM\_691143  
VERSION NM\_001114313.1 GI:166851849  
KEYWORDS .  
SOURCE Danio rerio (zebrafish)  
ORGANISM [Danio rerio](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
Actinopterygii; Neopterygii; Teleostei; Ostariophysi;  
Cypriniformes; Cyprinidae; Danio.  
REFERENCE 1 (bases 1 to 3693)  
AUTHORS Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G.,  
Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D.,  
Altschul,S.F., Zeeberg,B., Buetow,K.H., Schaefer,C.F., Bhat,N.K.,  
Hopkins,R.F., Jordan,H., Moore,T., Max,S.I., Wang,J., Hsieh,F.,  
Diatchenko,L., Marusina,K., Farmer,A.A., Rubin,G.M., Hong,L.,  
Stapleton,M., Soares,M.B., Bonaldo,M.F., Casavant,T.L.,  
Scheetz,T.E., Brownstein,M.J., Usdin,T.B., Toshiyuki,S.,  
Carninci,P., Prange,C., Raha,S.S., Loquellano,N.A., Peters,G.J.,  
Abramson,R.D., Mullahy,S.J., Bosak,S.A., McEwan,P.J.,  
McKernan,K.J., Malek,J.A., Gunaratne,P.H., Richards,S.,  
Worley,K.C., Hale,S., Garcia,A.M., Gay,L.J., Hulyk,S.W.,  
Villalon,D.K., Muzny,D.M., Sodergren,E.J., Lu,X., Gibbs,R.A.,  
Fahey,J., Helton,E., Kettman,M., Madan,A., Rodrigues,S.,  
Sanchez,A., Whiting,M., Madan,A., Young,A.C., Shevchenko,Y.,  
Bouffard,G.G., Blakesley,R.W., Touchman,J.W., Green,E.D.,  
Dickson,M.C., Rodriguez,A.C., Grimwood,J., Schmutz,J., Myers,R.M.,  
Butterfield,Y.S., Krzywinski,M.I., Skalska,U., Smailus,D.E.,  
Schnerch,A., Schein,J.E., Jones,S.J. and Marra,M.A.  
CONSRM Mammalian Gene Collection Program Team  
TITLE Generation and initial analysis of more than 15,000 full-length  
human and mouse cDNA sequences  
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 99 (26), 16899-16903 (2002)  
PUBMED [12477932](#)  
COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to final  
NCBI review. The reference sequence was derived from [BC155667.1](#).  
On Feb 6, 2008 this sequence version replaced gi:[125848749](#).

The RefSeq COMMENT block indicates the Status of the record and the GenBank sequence data that was used to provide the record. In addition, the COMMENT may identify a collaboration that supplied the defining sequence information for the genome, gene, or protein. The level of curation may differ between different collaborating groups.

COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to final  
NCBI review. The reference sequence was derived from [BC155667.1](#).  
On Feb 6, 2008 this sequence version replaced gi:[125848749](#).

# Properties: refseq type

3. type of RefSeq provided for the gene:  
property named as *srcdb refseq*  
*type\_of\_refseq*

<u>Query Restriction</u>	<u>Description</u>
<code>srcdb_refseq[prop]</code>	All NCBI RefSeq records
<code>srcdb_refseq_reviewed[prop]</code>	reviewed records
<code>srcdb_refseq_provisional[prop]</code>	provisional records
<code>srcdb_refseq_predicted[prop]</code>	predicted records
<code>srcdb_refseq_validated[prop]</code>	validated records
<code>srcdb_refseq_inferred[prop]</code>	inferred records; annotation inferred based on alignments from other genes or organisms
<code>srcdb_refseq_known[prop]</code>	reviewed, validated, provisional, predicted, inferred nucleotide or protein; excludes RefSeq records that are provided by the NCBI genome annotation
<code>srcdb_refseq_model[prop]</code>	RefSeq records generated by the NCBI gene annotation pipeline; model records

# Properties: refseq comment

GENOME ANNOTATION	This identifies RefSeq records provided by the NCBI Genome Annotation process. These records are provided via automated processing and are not subject to individual review or revision between builds (see description of the <a href="#">assembly and annotation process</a> ). The mRNA records are identified based on alignments of other mRNAs to the genomic sequence and the proteins are conceptual translations of these mRNAs. These model transcripts and proteins may differ from pre-existing curated RefSeq (accession prefix NM, NR, NP) or GenBank records because they correspond to the genomic sequence.
INFERRED	Not curated. Inferred by genome sequence analysis with no direct same-species support for the product. Support for the record may include a combination of orthologous or paralogous protein homology and alignments of transcripts from related genes. A portion of the sequence may be defined by ab initio prediction.
MODEL	Not curated. The RefSeq record is predicted by a whole-genome computational genome annotation pipeline. The record may represent an ab initio prediction, or may have some level of transcript or protein homology support.
PREDICTED	Not curated. Automatically provided based on GenBank sequence data; limited or partial support for the transcript or protein. A portion of the transcript or protein may reflect an ab initio annotation prediction that was submitted to GenBank.
PROVISIONAL	Not curated. Automatically provided based on GenBank sequence data; there is support for the transcript and protein. This is the default status code applied to some genomes for which there is no clear information about the method used to define the sequence.
REVIEWED	Curated. The RefSeq record has been reviewed to provide the preferred sequence standard and to add additional functional descriptive information and feature annotation, as relevant.
VALIDATED	Curated. The RefSeq record has undergone an initial review to provide the preferred sequence standard.
WGS	Not curated. The RefSeq record represents a collection of whole genome shotgun (WGS) sequences. This status code is applied to genomic records.

# Properties: other

## 4. other

<u>Property name</u>	<u>Explanation</u>
alive	a current, primary record (i.e., not secondary or discontinued). The term secondary means a record that has been merged into another.
GeneRIF	a record having one or more GeneRIF annotations attached
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.
phenotype	has an associated phenotype
phenotype only	only method of defining this gene is by phenotype
has ccds	a gene that encodes a protein sequence that is a member of a Consensus CDS (CCDS). See <a href="http://www.ncbi.nlm.nih.gov/projects/CCDS/">http://www.ncbi.nlm.nih.gov/projects/CCDS/</a>

# Examples

---

**Example 1:** I have an unknown band in a protein gel which weighs 2.4 kilodaltons

**23000:25000[molwt]**

**Example 2:** View a complete list of genes present on human chromosome 3

**Human[organism] AND 3[chr]**

**Example 3:** Retrieve a list of genes from chromosome 3, which are also associated with schizophrenia

**3[chr] AND schizophrenia [dis]**

# Examples

**Example 4:** Find all Gene records from fungi that have expression data in UniGene or GEO.

**fungi[organism] AND ( "gene unigene"[filter] OR "gene geo"[filter])**

**Example 5:** Find human and mouse genes not annotated on the genome but having reviewed RefSeq records.

**(Human[organism] OR mouse[organism]) AND "srcdb refseq reviewed"[Properties] NOT "gene nucleotide pos"[Filter]**

# Examples

**Example 6:** Find genes mapped to *Arabidopsis thaliana* chromosome 3 that have orthologs reported in HomoloGene

**"Arabidopsis thaliana"[Organism] AND 3[chr]  
AND "gene homologene"[filter]**

**Example 7:** Find mouse genes on chromosome 11 that has snp and transcript variants

**mouse[orgn] AND 11[chromosome] AND  
"gene snp"[filter] AND "has transcript  
variants"[prop] AND alive[prop]**