

Module 3: Genes and Sequences (NCBI)

ii - Does My Gene Have Known Homologs/Orthologs?

Aims

- Introduce tools to mine homology data
- Suggest ways to identify homologs
- Provide alternative ways to navigate
- Show examples of pre-computed homology comparisons

Introduction

You can mine the pre-computed sequence comparisons identifying putative orthologs (highly similar sequences across genomes) in Homologene. Begin your search for homologs by submitting a search on the Entrez home page (<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>) or by navigating to Homologene (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene>) to view results or submit a text query against the Homologene database. You can also following Links to Homologene from related records in other Entrez databases, such as Entrez Gene or UniGene.

Pre-computed protein comparisons are also available for each protein in Entrez Protein in the BLink (BLAST Link) database. You can view a graphical display of similar proteins by following the BLink link from any Entrez protein record. On the BLink page you can view a blast2 alignment between your protein and each protein identified by BLink as highly similar. Scroll down the list of Protein Descriptions to view the protein names for these proteins.

Homologs can also be identified through cross-species BLAST searches, as described in Module 2_iv.

Since homologs often share similar naming conventions, querying Entrez Gene with a gene name or gene symbol may yield homologous gene records, as seen in the exercise in Module3_i.

Additionally, you can determine if a curated homolog has been identified for a zebrafish gene by following the link to ZFIN found on Entrez Gene and UniGene pages.

Cross-species genome comparisons may also be used to identify homologs. For example, mouse homologs for human genes can be putatively identified based on the placement of mouse genes on the human genome. To see these comparative maps for human, mouse and rat, navigate to the Map Viewer home page (<http://www.ncbi.nlm.nih.gov/mapview/>), choose one of these organisms and select a chromosome to view. The “Maps & Options” button will provide a pop-up window where you can then add maps from human or rat to the mouse Map Viewer page. See Module2_iii to view the zebrafish Map Viewer page. Zebrafish Map Viewer does not currently include comparative maps.

Exercises

1. Homologene: Identify putative homologs based on sequence similarity using pre-computed comparisons in Homologene.

2. BLink: Identify putative homologs in other fish species.

1. Homologene: Identify putative homologs based on sequence similarity using pre-computed comparisons in Homologene.

Starting with an anonymous Entrez Gene record, zgc:86750 (GeneID:415228,

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=415228), follow the link to Homologene.

NCBI Entrez Gene

Search Gene for [] Go Clear current records only

Limits Preview/Index History Clipboard Details

Display Graphics Show 5 Send to

All: 1 gene homologene: 1 gene mapview: 0 gene nucleotide: 1 gene unigene: 0

1: **zgc:86750** **zgc:86750** [*Danio rerio*] MGC cDNA clone, Links
 Genes: 1 updated 31-May-2005
 GeneID: 415228 Locus tag: [ZDB-GENE-040625-145](#)
 Official Symbol: **zgc:86750** and Name: **zgc:86750** provided by [Zebrafish Nomenclature Committee](#)
 Gene type: protein coding
 Gene name: zgc:86750
 Gene description: zgc:86750
 RefSeq status: Predicted
 Organism: [Danio rerio](#)
 Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio*

► **Bibliography:** Gene References into Function (GeneRIF): [Submit](#) [help](#) ?

[PubMed](#) links

► **General gene information** ?

GeneOntology
 Provided by [GO](#)

Component	Evidence
cellular component unknown	ND

Links

- Conserved Domains
- HomoloGene
- Nucleotide
- Full text in PMC
- Protein
- PubMed
- Taxonomy
- Ensembl
- MGC
- UCSC
- ZFIN

Click on Homologene link to view results of pre-computed sequence comparisons

You can also submit a text search against the Homologene database from any Entrez page

View Homologene record links to zgc:86750

NCBI Homologene Discover Homologs

All Databases PubMed Nucleotide Protein Genome Structure Map Viewer

Search HomoloGene for Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Send to

All: 1 Fungi: 0 Mammals: 1

Species	Gene ID	Description
H.sapiens	CRYGD	crystallin, gamma D
P.troglodytes	LOC459906	similar to crystallin, gamma D; gamma crys...
C.familiaris	LOC488495	gamma-D-crystallin
M.musculus	Crygd	crystallin, gamma D
R.norvegicus	Crygd	crystallin, gamma D

Click on Homologene:36213 to view the detailed HomoloGene record

- View related Genes, their Proteins, related Phenotypes and PubMed entries
- View conserved domains identified in these proteins
- View curated homology data
- View related UniGene clusters

NCBI **Homologene** *Discover Homologs* My NCB
Welcome schriml. [Sign Out](#)

All Databases PubMed Nucleotide Protein Genome Structure Map Viewer LocusLink UniGene OMIM






Search HomoloGene for


Limits Preview/Index History Clipboard Details

Display HomoloGene Show 20 Send to

All: 1 Fungi: 0 Mammals: 1

1: HomoloGene:36213. Gene conserved in Eutheria [Links](#)

<p>Genes <i>Genes identified as putative homologs of one another during the construction of HomoloGene</i></p> <ul style="list-style-type: none"> <input checked="" type="checkbox"/> H.sapiens CRYGD crystallin, gamma D. <input checked="" type="checkbox"/> P.troglodytes LOC459906 similar to crystallin, gamma D; gamma crystallin 4. <input checked="" type="checkbox"/> C.familiaris LOC488495 gamma-D-crystallin. <input checked="" type="checkbox"/> M.musculus Crygd crystallin, gamma D. <input checked="" type="checkbox"/> R.norvegicus Crygd crystallin, gamma D. 	<p>Proteins <i>Proteins used in sequence comparisons and their conserved domain architectures.</i></p> <ul style="list-style-type: none"> <input checked="" type="checkbox"/> NP_008822.2  174 aa <input checked="" type="checkbox"/> XP_516055.1  708 aa <input checked="" type="checkbox"/> XP_545616.1  242 aa <input checked="" type="checkbox"/> NP_031802.1  174 aa <input checked="" type="checkbox"/> NP_149086.1  174 aa
---	--

<p>Alignment Scores <i>Various evolutionary parameters derived from pairwise alignments have been saved.</i></p> <p>Show Table of Pairwise Scores</p>	<p>Conserved Domains <i>Conserved Domains from CDD found in protein sequences by rpsblast searching.</i></p> <ul style="list-style-type: none"> <input checked="" type="checkbox"/> smart00247  XTALbg. Beta/gamma crystallins
--	---



Alignments can be regenerated using BLAST for any selected pair of proteins.

Regenerate Alignments

NP_008822.2(H.sapiens. CRYGD)

XP_516055.1(P.troglodytes. LOC459906)

BLAST

Phenotypes

Phenotypes

Phenotypic information for the genes in this entry imported from model organism databases.

- [H.sapiens MIM:115700](#)
Cataract, crystalline aculeiform. [OMIM]
- [H.sapiens MIM:123690](#)
Cataracts, punctate, progressive juvenile-onset. [OMIM]

PubMed

Articles associated with genes and sequences of this entry plus additional related articles.

- [Pande A, et al.](#)
Decrease in protein solubility and cataract formation caused by the Pro23 to Thr mutation in human gamma D-crystallin. *Biochemistry* 44, 2491-2500 (2005).
The cataract-causing mutation proline23 to threonine does not exhibit any significant structural change relative to the native protein. However, in marked contrast to the native protein, the mutant shows a dramatically lowered solubility.
- [Mackay DS, et al.](#)
A missense mutation in the gammaD crystallin gene (CRYGD) associated with autosomal dominant "coral-like" cataract linked to chromosome 2q. *Mol Vis* 10, 155-162 (2004).

Publications

Related Homology Resources

Links to curated and computed homology information found in other databases.

- [MGI:88524](#)
Orthology group for M.musculus Crygd includes H.sapiens CRYGD and R.norvegicus Crygd.

UniGene

Links to groups of transcribed sequences established by tblastn searching of UniGene.

- [B.taurus Bt.399](#)
Crystallin, gamma B
- [B.taurus Bt.537](#)
Crystallin, gamma D
- [B.taurus Bt.30404](#)
Transcribed locus, moderately similar to NP_149086.1 crystallin, gamma D [Rattus norvegicus]
- [B.taurus Bt.33836](#)
Transcribed locus, moderately similar to XP_516055.1 PREDICTED: similar to crystallin, gamma D; gamma crystallin 4 [Pan troglodytes]
- [B.taurus Bt.37350](#)
Crystallin, gamma C
- [C.familiaris Cfa.23574](#)
Transcribed locus
- [C.familiaris Cfa.23598](#)
Transcribed locus
- [D.erio Dr.15364](#)
GammaM3-crystallin
- [D.erio Dr.18937](#)
Im:7140756
- [D.erio Dr.19621](#)
Crystallin, gamma S4
- [D.erio Dr.29371](#)
Zgc:86723
- [D.erio Dr.29372](#)

Curated Homologs

UniGene clusters

Highly similar zebrafish UniGene clusters

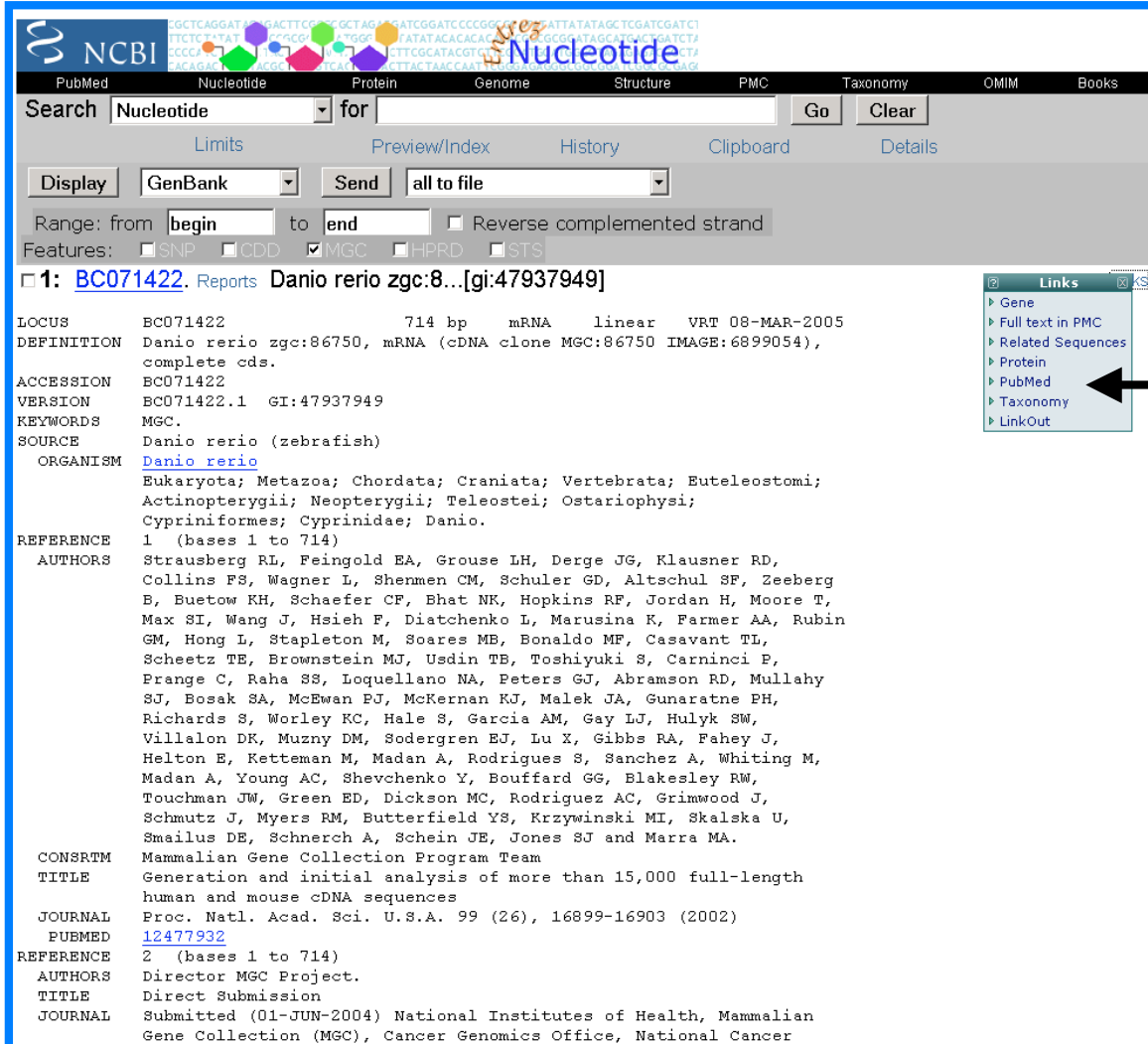
Homologene comparisons have shown that the Entrez Gene zgc:86723 record is highly similar to members of the gamma crystallin gene family.

2. **BLink:** Identify putative homologs in other fish species. Begin your search from mRNA accession.

For example: BC071422

(<http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=BC071422.1>
)

Click on the Links menu and select 'Protein' to go to the corresponding Entrez Protein record



NCBI Nucleotide

Search for

Limits Preview/Index History Clipboard Details

Display

Range: from to Reverse complemented strand

Features: SNP CDD MGC HPRD STS

1: [BC071422](#). Reports [Danio rerio zgc:8...\[gi:47937949\]](#)

LOCUS BC071422 714 bp mRNA linear VRT 08-MAR-2005

DEFINITION [Danio rerio zgc:86750](#), mRNA (cDNA clone MGC:86750 IMAGE:6899054), complete cds.

ACCESSION BC071422

VERSION BC071422.1 GI:47937949

KEYWORDS MGC.

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 714)
AUTHORS Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF, Zeeberg B, Bustow KH, Schaefer CF, Bhat NK, Hopkins RF, Jordan H, Moore T, Max SI, Wang J, Hsieh F, Diatchenko L, Marusina K, Farmer AA, Rubin GM, Hong L, Stapleton M, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Brownstein MJ, Usdin TB, Toshiyuki S, Carninci P, Prange C, Raha SS, Loquellano NA, Peters GJ, Abramson RD, Mullahy SJ, Bosak SA, McEwan PJ, McKernan KJ, Malek JA, Gunaratne PH, Richards S, Worley KC, Hale S, Garcia AM, Gay LJ, Hulyk SW, Villalón DK, Muzny DM, Sodergren EJ, Lu X, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madan A, Young AC, Shevchenko Y, Bouffard GG, Blakesley RW, Touchman JW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Krzywinski MI, Skalska U, Smailus DE, Schnerch A, Schein JE, Jones SJ and Marra MA.
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](#)

REFERENCE 2 (bases 1 to 714)
AUTHORS Director MGC Project.
TITLE Direct Submission
JOURNAL Submitted (01-JUN-2004) National Institutes of Health, Mammalian Gene Collection (MGC), Cancer Genomics Office, National Cancer

Links

- Gene
- Full text in PMC
- Related Sequences
- Protein
- PubMed
- Taxonomy
- LinkOut

To navigate to the corresponding Entrez Protein record: first scroll down the page to the Protein Link



```

Clone distribution: MGC clone distribution information can be found
through the I.M.A.G.E. Consortium/LLNL at: http://image.llnl.gov
Series: IRAL Plate: 56 Row: g Column: 20
This clone was selected for full length sequencing because it
passed the following selection criteria: matched mRNA gi: 47937949.
FEATURES             Location/Qualifiers
     source            1..714
                        /organism="Danio rerio"
                        /mol_type="mRNA"
                        /db_xref="taxon:7955"
                        /clone="MGC:86750 IMAGE:6899054"
                        /tissue_type="Embryo, 7 different stages (from just
                        fertilized embryos to 72 hours just hatched baby fish)"
                        /clone_lib="GISZF001"
                        /lab_host="DH10B"
                        /note="Vector: pDNR-LIB"
     gene              1..714
                        /gene="zgc:86750"
                        /db_xref="GeneID:415228"
                        /db_xref="ZFIN:ZDB-GENE-040625-145"
     CDS                67..591
                        /gene="zgc:86750"
                        /codon_start=1
                        /product="zgc:86750"
                        /protein_id="AAH71422.1"
                        /db_xref="GI:47937950"
                        /db_xref="GeneID:415228"
                        /db_xref="ZFIN:ZDB-GENE-040625-145"
                        /translation="MGKVIFYEDRNFQGRSYECMGDCGDFSSYMNRCHSCRVESGCWM
MYDQTNMYMGSYGFFRRGEYADYMSMFGMNNCIRSCRMIIPMYRGSYRMRIYERDNFMGQ
MYEMDDCDNIMNRYRMSHCQSCHVMDGHWLFYDQPNYRGRMWHFPGQYRNFSNYGG
MRFMSMRRIMDSWY"
ORIGIN
1 caacacagaa aatcagtttc agcttctcct ttgtgcaatc accaagggtc agctaaagta
61 accatgatgg gcaaggatcat cttctacgag gacagaaact tccagggtcg ctcttatgag
121 tgtatgggag actgtgggtga cttctcctcc tacatgaatc gctgtcactc ttgcagagtg
181 gagagcggct gctggatgat gtacgaccaa accaactaca tgggaagtgg atattcttc
241 aggaggggag agtatgctga ttacatgtct atgtttggaa tgaacaactg catcaggtcc
301 tgccgtatga tccccatgta caggggatcc tacagaatga ggatctacga gagggacaac
361 ttcattgggc agatgtacga gatgatggat gactgtgaca acatcatgaa ccgttaccgc

```

Navigate to BLink by the provided link

NCBI Entrez Protein

Welcome schrim Sign Out

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

All: 1 bacteria: 0 RefSeq: 0

1: [AAH71422](#) Reports BLink, Domains, MGC cDNA clone, Links

Zgc:86750 [Danio rerio]
gi|47937950|gb|AAH71422.1|[47937950]

- View the graphical display of protein hits
- Click on the BLAST score to see the alignment
- Click on the Accession to go to the Entrez Protein record
- Scroll down the Protein Descriptions to protein names
- Click on “Show identical” to include all identical hits in the display
- Click on “Best hits” to view the hits grouped by organism
- Click on “Conserved Domain Database hits” to view domain details

Query Protein

NCBI

BLAST Protein Structure PubMed Taxonomy
Genome Nucleotide 3D-Domains Books Help

Query: gi|47937950 Zgc:86750 [Danio rerio]
Matching gi: [50344932](#)

Show identical Best hits Common Tree Taxonomy Report 3D structures CDD-Search GI list

198 BLAST hits to 21 unique species [Sort by taxonomy proximity](#)

0 Archaea 0 Bacteria 198 Metazoa 0 Fungi 0 Plants 0 Viruses 0 Other Eukaryotae

Keep only Cut-Off 100 Select Reset

174 aa

	SCORE	E	ACCESSION	GI	PROTEIN DESCRIPTION
Conserved Domain Database hits					
991	31		AAH95033	63100520	Gamma crystallin M2 [Danio rerio]
988	31		NP_001...	66472384	Gamma crystallin M2 [Danio rerio]
957	31		NP_001...	50344940	hypothetical protein LOC415232 [Danio rerio]
955	31		NP_001...	50540230	hypothetical protein LOC436855 [Danio rerio]
923	31		NP_001...	50540228	hypothetical protein LOC436854 [Danio rerio]
879	28		P10044	117451	Gamma crystallin M2 (Gamma-M2)
848	31		NP_001...	50539876	hypothetical protein LOC436681 [Danio rerio]
810	31		NP_001...	66472812	crystallin, gamma M2b [Danio rerio]
809	31		NP_001...	56090503	crystallin, gamma M2c [Danio rerio]
779	31		NP_001...	66392178	gammaM2a-crystallin [Danio rerio]
755	24		AAP07205	6319202	gamma crystallin M [Astyanax mexicanus]
745	24		I50142	2147391	gamma-crystallin M2-3 - Clarias fuscus
731	21		JC2354	1083899	gamma-crystallin M2-1 - Petenia splendida x Cichlasoma synspilum
719	24		JEU323	7441339	gamma3-crystallin 1 - catfish
718	21		CAG08130	47221468	unnamed protein product [Tetraodon nigroviridis]
715	21		JC2356	632011	gamma-crystallin M2-2 - Petenia splendida x Cichlasoma synspilum

BLink has identified highly similar proteins in several fish including: zebrafish, Mexican tetra, whitespotted clarias, catfish and freshwater pufferfish