

Module 4: Comparative Sequence Analysis

Aims

- Review the sequences available from different organisms
- Identify putative homologous genes
- Compare genome sequences from different organisms and identify conserved sequences
- Analyse conserved sequences for potential regulatory function

Comparative sequence analysis is a powerful method for aiding human gene identification, inferring function of a gene's product, and identifying novel functional elements such as those involved in transcriptional regulation. This is because biologically important regions of the genome are, generally, under selective constraint. Comparing the genome sequences from a variety of organisms may facilitate the identification of functionally significant units in the human genome.

The information that can be inferred when comparing sequences is dependent on the evolutionary distance between the two organisms. Organisms that are closely related are more likely to share a higher degree of sequence similarity. Distantly related organisms such as yeast and worm share less sequence similarity and are likely to show sequence conservation in coding regions alone. This may also be true for distantly related vertebrates such as fish. More closely related organisms, such as mouse, are likely to be conserved in coding regions, and other functional elements such as regulatory sequences. However, the closer the evolutionary relationship with human, the more 'sequence noise' is likely to arise where non-functional sequence appears similar because insufficient time has elapsed for the two sequences to diverge.

Evolutionarily Related Gene Sequences

Homologous genes are derived from a common ancestor and may either have a similar sequence or function. In general, homologous genes can be divided into two classes:

Orthologues are genes that often perform the same function in different organisms. They are defined as being homologous genes in different organisms derived from the same gene during speciation. In general, their sequence similarity reflects amount of time since they diverged from a common ancestor i.e., the less time that has elapsed since divergence, the greater the sequence similarity between the two genes. These are genes 1 and 2 in figure 4.1.

Paralogues are genes families that are present within a single species. Often they arise by duplication. These genes are not under the same pressure to maintain their function so that one copy may acquire a novel function. These are genes 2 and 3 in figure 4.1.

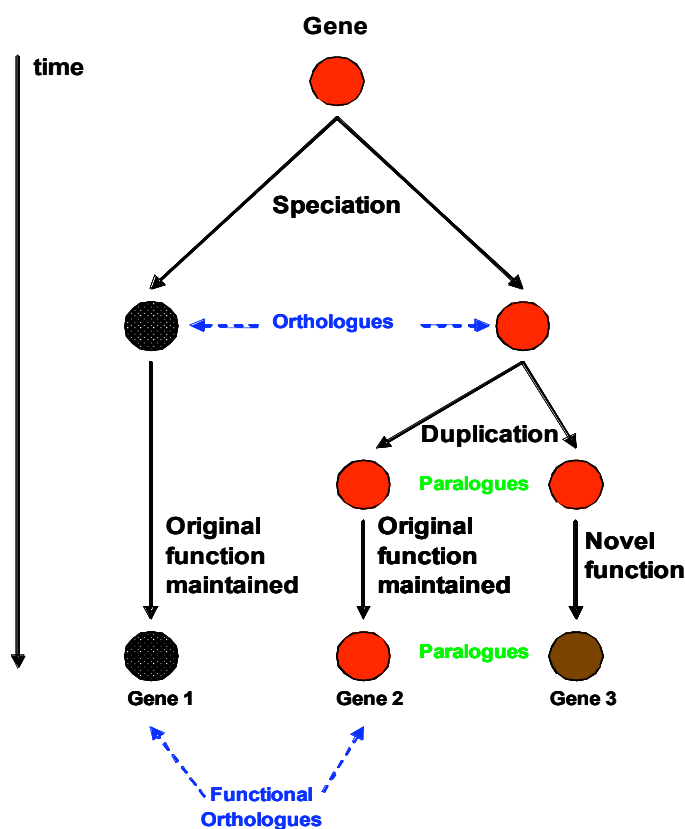


Figure 6.1 Homologous Gene Sequences

Identifying evolutionarily related gene sequences – where to start...

Most searches for orthologous genes begin with blast searches (for more details refer to module 1 of this manual). The type of search that you should perform depends upon what information you have at your disposal. Protein sequence based queries generally find more distantly related matches because of the redundancy in the genetic code (i.e. some amino acids are encoded by more than one codon). Nucleotide searches using the discontinuous megablast parameters are also very useful. We recommend that you survey a number of different databases using a number of different search parameters to obtain the most informative results.

BUT be warned: uncertain Orthologues

You may encounter gene sequences that appear to be orthologous and may be derived from the same ancestor but no longer perform the same biological function (for example genes 1 and 3 in figure 4.1). If you choose to analyse such sequences the sensitivity and specificity of your search will be reduced and it may not yield any informative results.

For example, the gene for bone morphogenetic protein 8 (*BMP8*) was duplicated in a common ancestor of human and mouse giving rise to *BMP8a* and *BMP8b* (see Figure 4.2). BLAST analysis of these four sequences yields quite confusing results. Human and mouse *BMP8a* are reciprocal best alignments using both nucleotide and protein sequences to search. In contrast, both the nucleotide and protein sequences of mouse *Bmp8b* align best to their human *BMP8a* counterparts. Human *BMP8b* mRNA aligns best to mouse *Bmp8b* mRNA, but human *BMP8b* protein aligns best to mouse *Bmp8a* protein, while mouse protein *Bmp8b* aligns best to human protein *BMP8a* (Nardone *et al.*, 2004, figure 4.2).

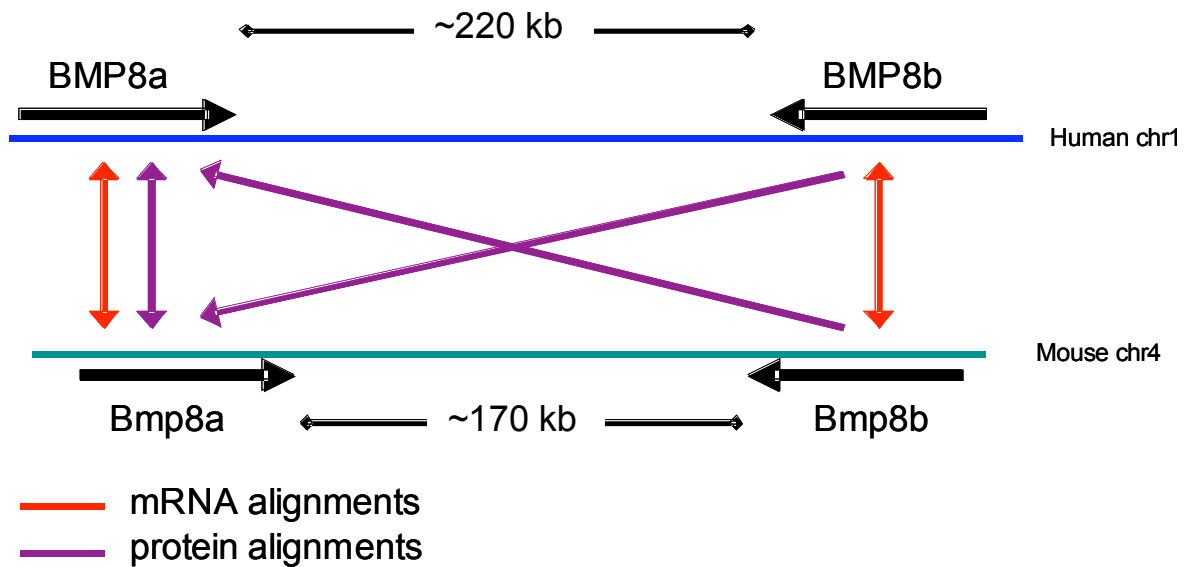


Figure 4.2 Uncertain *BMP8a* and *BMP8b* orthologues.

Therefore, we recommend that you perform the following steps to confirm the true functionality and relatedness of your gene sequences. Much of this information can be obtained from genome browsers such as Ensembl, NBKI NCBI or the UCSC.

- 1) Identify any other paralogues that may affect your analysis. This can be achieved by performing a BLAST search your sequence against its source genome or using the self-chain track at the UCSC genome browser.
- 2) Confirm the percentage identity (similarity) at both the nucleotide and protein level between paralogous and orthologous sequences to ensure that you are analysing the most closely related sequences.
- 3) Perform evolutionary analysis of nucleotide/protein sequence (phylogeny). In contrast to similarity-based methods such as BLAST, phylogenetic methods can better take into account the effects of repeated substitutions at one site and variable rates of evolution

among sequences. Multiple genes are placed in an evolutionary tree representing genealogical relationship

- 4) Compare the exon/intron structure of your orthologous genes.
Evolutionarily related genes often share a similar gene structure

- 5) Examine the chromosomal context of the two orthologous genes.
Closely related species, such as human and mouse often have large conserved segments (for definition see later section Genome sequence analysis) and therefore neighbouring genes are also shared between the two species.

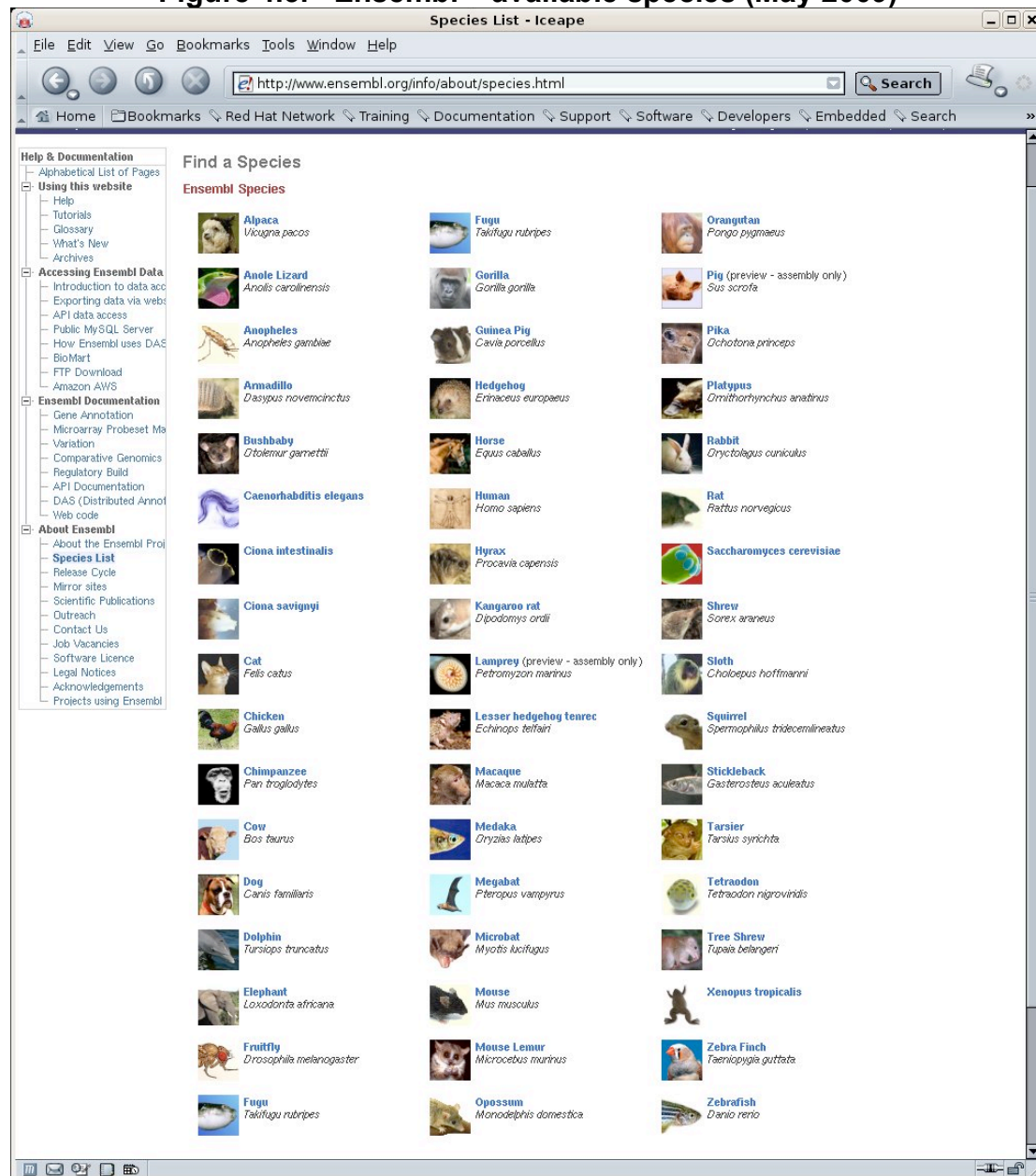
Comparative Genome Analysis

Comparing the DNA sequences of different species is a powerful method for decoding genomic information. This is because functional sequences tend to evolve at a slower rate than non-functional sequences. By comparing the genomic sequences of several species at different evolutionary distances it is possible to identify coding sequences, conserved non-coding sequences and those sequences that are unique to humans.

Advances in bacterial clone mapping and sequencing that evolved during the construction of the sequence-ready maps for particularly the worm and the human genome are now being applied to other organisms. The sequence of many genomes has or is being generated using a combination of the clone-by-clone method (adopted for generating the human genome sequence by the public effort) and whole genome shotgun (WGS - used by Celera to generate the sequence of the fruit fly and their version of the human genome sequence). Finished and unfinished genome assemblies are currently available for 44 vertebrates. The front page of the genome browser Ensembl (<http://www.ensembl.org>) is shown on the following page (figure 4.3). It displays all species whose genomes have been sequenced and assembled. Further information about individual species, its sequence coverage and participating research institutes can be found by following the appropriate

hyperlinks on this page. Similar links can also be found at the UCSC genome browser.

Figure 4.3. Ensembl – available species (May 2009)



When two species diverge from a common ancestor those sequences that maintain their original function are likely to remain conserved in both species throughout their subsequent independent evolution. Therefore comparing sequences in different species is a powerful tool for increasing the confidence

of a predicted functional unit, or identifying novel functional units (e.g. human, mouse and zebrafish).

In general, greater evolutionary distance between the species is reflected by more divergent sequences and fewer shared functional units. Comparing sequences that diverged from a common ancestor approximately 450 million years ago (mya) e.g. human and fish aids the identification of coding sequences. Conserved non-coding regions are generally not identified. If the evolutionary distance between the two species is reduced to approximately 60 mya, e.g. human and mouse, both non-coding and coding units are commonly conserved. A large number of features are conserved between recently evolved species such as human and chimp. The inclusion of a closely related species in a comparative analysis makes it possible to identify coding and non-coding sequences but also those genomic sequences that may be responsible for traits that are unique to the reference species.

Table 4.1: Selection of Species for DNA Comparisons

Human vs.	Chimpanzee	Mouse	Opossum	Fish
Size (Gbp)	3.0	2.5	3.5	0.4
Time since divergence	~5 MYA	~65MYA	~150 MYA	~450 MYA
Sequence conservation (in coding regions)	>99%	~80%	~70%	~65%
Aids identification of	Recently changed sequences and genomic rearrangements	Both coding and non-coding sequences	Both coding and non-coding sequences	Primary coding sequences
Background noise	HIGH	MODERATE	LOW	VERY LOW

Today we will be showing you some of our favourite web-sites where homologous gene sequences can be identified and analysed. These are listed in table 4.2. As with most types of web-based genomic analysis there

are a vast number of sites that can be used for this type of analysis. These sites require different amounts of information from the user; some require an input sequences while others contain pre-calculated information in a database that can be interrogated easily.

Table 4.2 Different steps and web programmes that can be used to identify evolutionarily conserved regions

Step	Using whole genome assemblies	Interrogating and incorporating sequences from databases
Identifying paralogues	UCSC – Self Chain Ensembl – Protein family	NCBI - blastn
Identifying orthologues	Ensembl – Orthologue Prediction UCSC - BLAT	Discontiguous megablast BLAST-n, -p BLink
Confirming true orthology	Ensembl - MultiContigView - SyntenyView - AlignSliceView	Clustalw
Aligning and identifying conserved sequences	ECR Browser Genome Vista PhastCons	zPicture Vista LAGAN

WORKED EXAMPLES:

1) Identifying paralogous and orthologous genes

The easiest and most common method used to identify homologous sequences exploits the sequence conservation between related genes. This can be identified using sequence similarity searches i.e., BLAST searches against nucleotide or protein databases (see module 1). However, in many cases the genome browsers such as NCBI, Ensembl and the UCSC genome browsers have already done the hard work for you. For example orthologous gene predictions from whole genome assemblies can be accessed from either **HomolGene** (NCBI) or **Orthologue Prediction** (Ensembl); while paralogous genes can be identified by following the **Protein Family** link found in Ensembl GeneView.

a) Useful sites at the NCBI

If you have the sequence for your gene of interest in one species and wish to find orthologues BLAST is a good starting point

The screenshot shows the NCBI BLAST homepage. At the top, there are navigation tabs: Home, Recent Results, Saved Strategies, and Help. On the right, there are links for My NCBI, Sign In, and Register. The main content area is divided into several sections:

- NCBI BLAST Home:** A brief description of BLAST and a link to learn more.
- BLAST Assembled Genomes:** A section where users can choose a species genome to search. It lists various species including Human, Mouse, Rat, Arabidopsis thaliana, Oryza sativa, Bos taurus, Danio rerio, Drosophila melanogaster, Gallus gallus, Pan troglodytes, Microbes, and Apis mellifera.
- Basic BLAST:** A section where users can choose a BLAST program to run. It lists options like nucleotide blast, protein blast, blastx, tblastn, and tblastx, each with a brief description and available algorithms.
- Specialized BLAST:** A section where users can choose a type of specialized search. It lists options like trace archives, conserved domains, conserved domain architecture, gene expression profiles, immunoglobulins, SNPs, vector contamination, and align two sequences.

Annotations on the right side of the screenshot include:

- A box titled "BLAST Species specific databases" with an arrow pointing to the "BLAST Assembled Genomes" section.
- A box titled "If the genome of your species of interest is not complete you can BLAST trace archives for specific organisms" with an arrow pointing to the "Specialized BLAST" section.
- A box titled "Hint – use discontinuous megablast" with an arrow pointing to the "Specialized BLAST" section.

However NCBI has pre-computed orthologues which can be found in the HomoloGene page. However because this page is precomputed this means that it does not contain the most recently assembled genomes.

1: We will start by searching for our gene of interest **MITF** at NCBI

Type MITF into search box

Search HomoloGene

2: Select HomoloGene:4892

The MITF gene family

Links to gene and protein information

Genes	Proteins
<input checked="" type="checkbox"/> MITF, <i>Homo sapiens</i> microphthalmia-associated transcription factor	<input checked="" type="checkbox"/> NP_937802.1 520 aa
<input checked="" type="checkbox"/> MITF, <i>Pan troglodytes</i> microphthalmia-associated transcription factor	<input checked="" type="checkbox"/> XP_001138775.1 526 aa
<input checked="" type="checkbox"/> MITF, <i>Canis lupus familiaris</i> microphthalmia-associated transcription factor	<input checked="" type="checkbox"/> XP_855594.1 419 aa
<input checked="" type="checkbox"/> MITF, <i>Bos taurus</i> microphthalmia-associated transcription factor	<input checked="" type="checkbox"/> NP_001001150.1 413 aa
<input checked="" type="checkbox"/> Mitf, <i>Mus musculus</i> microphthalmia-associated transcription factor	<input checked="" type="checkbox"/> NP_032627.1 419 aa
<input checked="" type="checkbox"/> Mitf, <i>Rattus norvegicus</i> microphthalmia-associated transcription factor	<input checked="" type="checkbox"/> XP_001065702.1 520 aa
<input checked="" type="checkbox"/> MITF, <i>Gallus gallus</i> microphthalmia-associated transcription factor	<input checked="" type="checkbox"/> NP_990360.1 468 aa
<input checked="" type="checkbox"/> mitfb, <i>Danio rerio</i> microphthalmia-associated transcription factor b	<input checked="" type="checkbox"/> NP_571922.1 427 aa

3: Select Show Multiple Alignment

Protein Alignments
Protein multiple alignment, pairwise similarity scores and evolutionary distances.

Show Multiple Alignment

Show Pairwise Alignment Scores

Pairwise alignments generated using BLAST

Regenerate Alignments

XP_001138775.1 (Pan troglodytes) [v]
XP_855594.1 (Canis lupus familiaris) [v]

BLAST

Conserved Domains
Conserved Domains from CDD found in protein sequences by psblast searching.

HLH (cd00083)
Helix-loop-helix domain, found in specific DNA-binding proteins that act as transcription factors; 60-100 amino acids long.

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

MGI:104554
Orthology group for M.musculus Mitf includes H.sapiens MITF and R.novegicus Mitf.

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

Homo sapiens

MIM:103470
Waardenburg syndrome/ocular albinism, digenic.

MIM:103500
Tietz syndrome.

MIM:193510
Waardenburg syndrome, type IIA.

Mus musculus

MP:0001186
Pigmentation phenotype.

MP:0003631
Nervous system phenotype.

MP:0005371
Limbs/digits/tail phenotype.

MP:0005373
Lethality-postnatal.

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

Bt.29882, *Bos taurus*
Microphthalmia-associated transcription factor

Bt.66572, *Bos taurus*
Transcribed locus, strongly similar to NP_032627.1 microphthalmia-associated...

Bt.98153, *Bos taurus*
Transcribed locus, strongly similar to NP_032627.1 microphthalmia-associated...

Cfa.2682, *Canis lupus familiaris*
Microphthalmia-associated transcription factor

Dr.81296, *Danio rerio*
Microphthalmia-associated transcription factor a

Dr.83675, *Danio rerio*
Microphthalmia-associated transcription factor b

Eca.12993, *Equus caballus*
Microphthalmia transcription factor

Gaa.275, *Gallus gallus*

Further orthologues identified by searching UniGene with tblastn. Here there is the horse orthologue

Multiple Sequence Alignment

NP_937802.1	1	MQSESGIVDPFVVGEEFHEEPKTYEYELKSSQPLKSSSSAEHPGAS	44
XP_001138775.1	1	MQSESGIVDPFVVGEEFHEEPKTYEYELKSSQPLKSSSSAEHPGAS	44
XP_855594.1		-----	
NP_001001150.1		-----	
NP_032627.1		-----	
XP_001065702.1	1	MQSESGIVADPFVVGEEFHEEPKTYEYELKSSQPLKSSSSAEHSGAS	44
NP_990360.1		-----	
NP_571922.1	1	MQSESGIVDPFVVGDDPFHEEPKTYEYELKSSQPLQNSNPSEQQ---	41
NP_937802.1	45	KPPISSSSMTRILLRQQLMREQMQRERREQQQLQAAQPMQQ	88
XP_001138775.1	45	KPPISSSSMTRILLRQQLMREQMQRERREQQQLQAAQPMQQ	88
XP_855594.1	1	-----MLEMLEYN-----	8
NP_001001150.1	1	-----MLEMLEYN-----	8
NP_032627.1	1	-----MLEMLEYS-----	8
XP_001065702.1	45	KPPISSSTMTSRILLRQQLMREQMQRERREQQQLQAAQPMQQ	88
NP_990360.1	1	-----MTRILLRQQLMREQMQRERREQQQLQAAQPMQQ	36
NP_571922.1	42	HGSCKPPPLGSSRVLLRQQLMREQLQQQRERREQQKRC-----	77
NP_937802.1	89	RVPVSQTPAINVSVPTTLP SATQVPMVVLKVQTHLENPTKYHIQ	132
XP_001138775.1	89	RVPVSQTPAINVSVPTTLP SATQVPMVVLKVQTHLENPTKYHIQ	132
XP_855594.1	9	-----HYQVQTHLENPTKYHIQ	25
NP_001001150.1	9	-----HYQVQTHLENPTKYHIQ	25
NP_032627.1	9	-----HYQVQTHLENPTKYHIQ	25
XP_001065702.1	89	RVAVSQTPAINVSVPTTLP SATQVPMVVLKVQTHLENPTKYHIQ	132
NP_990360.1	37	RVPVSQTPAINVSVPA SLPPATQVPMVVLKVQTHLENPTKYHIQ	80
NP_571922.1	78	-ISITHSPAINVSHPCGPPSAAQVPMVVLKVQTHLENPTKYHIQ	120

Multiple Alignment of orthologous protein sequences

4: return to HomoloGene page and select Show Pairwise Alignment Scores.

Display Alignment Scores Show 20 Send to

All: 1 Fungi: 0 Mammals: 0

1: HomoloGene:4892, Gene conserved in Euteleostomi

Alignment Scores

Species	Gene	Symbol	Identity (%)		Substitution Rates ¹			Blast
			Protein	DNA	d	d_N/d_S	d_{NR}/d_{NC}	
Homo sapiens								
	MITF							
vs. Pan troglodytes	MITF		99.6	99.6	0.005	0.111	0.000	Blast
vs. Canis lupus familiaris	MITF		97.3	93.5	0.068	0.050	0.695	Blast
vs. Bos taurus	MITF		97.0	93.1	0.073	0.042	0.296	Blast
vs. Mus musculus	Mitf		93.8	88.1	0.130	0.051	0.834	Blast
vs. Rattus norvegicus	Mitf		94.0	89.5	0.113	0.063	0.632	Blast
vs. Gallus gallus	MITF		91.9	83.8	0.182	0.031	0.399	Blast
vs. Danio rerio	mitf		77.0	73.0	0.335	undef	0.452	Blast
Pan troglodytes								
	MITF							
vs. Homo sapiens	MITF		99.6	99.6	0.005	0.111	0.000	Blast
vs. Canis lupus familiaris	MITF		96.8	93.6	0.067	0.060	0.496	Blast
vs. Bos taurus	MITF		97.0	93.1	0.072	0.042	0.297	Blast
vs. Mus musculus	Mitf		93.4	87.9	0.132	0.054	0.740	Blast
vs. Rattus norvegicus	Mitf		93.7	89.2	0.116	0.064	0.579	Blast
vs. Gallus gallus	MITF		91.5	83.8	0.183	0.034	0.376	Blast
vs. Danio rerio	mitf		76.6	72.4	0.344	undef	0.501	Blast
Canis lupus familiaris								
	MITF							
vs. Homo sapiens	MITF		97.3	93.5	0.068	0.050	0.695	Blast
vs. Pan troglodytes	MITF		96.8	93.6	0.067	0.060	0.496	Blast
vs. Bos taurus	MITF		97.0	93.4	0.069	0.028	0.426	Blast
vs. Mus musculus	Mitf		93.1	87.6	0.136	0.054	0.642	Blast
vs. Rattus norvegicus	Mitf		92.3	86.8	0.146	0.059	0.731	Blast
vs. Gallus gallus	MITF		91.6	83.1	0.191	0.032	0.617	Blast
vs. Danio rerio	mitf		77.9	71.7	0.355	undef	0.521	Blast

Links to pairwise protein BLAST Alignment

- **d**: the number of nucleotide substitutions per site, corrected for multiple substitutions using the method of Jukes and Cantor (1969).
- **d_N/d_S** : the ratio of the rate of nonsynonymous substitutions (d_N) to the rate of synonymous substitutions (d_S), calculated using the method of [Nei and Gojobori \(1986\)](#). A high value of this metric indicates adaptive selection, whereas a low value indicates purifying selection.
- **d_{NR}/d_{NC}** : the ratio of radical nonsynonymous substitutions (d_{NR}) to conservative nonsynonymous substitutions (d_{NC}), calculated using the method of [Hughes et al. \(1990\)](#). This metric is analogous to d_N/d_S , but it has the advantage of being useful for studying the evolution of sequences that diverged in the distant past.

b) The Ensembl genome browser

1. From the Ensembl homepage select the **Zebrafish** genome browser.
2. Search *e!* zebrafish gene: **eng2b** and press go.
3. Select **Ensembl Gene: ENSDAR0000038868**

This should bring you to the GeneView page for the eng2b gene:

Ensembl genome browser 54: D.rerio - Gene summary - Gene: eng2b (ENSDARG00000038868)

Location: 2:27,938,253-27,941,786

Gene: eng2b (ENSDARG00000038868)

Homeobox protein engrailed-2b (ZF-En-1)(Eng3) Source: UniProtKB/Swiss-Prot:P91333

Location: Chromosome 2: 27,938,253-27,941,786 forward strand.

Transcripts: There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
eng2b	ENSDART00000056748	ENSDARFP00000056747	protein_coding

Gene summary [help](#) [Splice variants >](#)

Name: [eng2b](#) (ZFN)

Synonyms: en3, eng3, engrailed 3 [To view all Ensembl genes linked to the name [click here](#).]

Gene type: Known protein coding

Prediction: Transcripts were annotated by the Ensembl [genebuild](#).

Method: Transcripts

Configuring the display

Tip: use the "Configure this page" link on the left to show additional data in this region.

Ensembl release 54 - May 2009 © WTSI / EBI

Permanent link - View in archive site

4. Click on the Protein Family hyperlink on the left side of the page:

Ensembl genome browser 54: D.rerio - Protein families - Gene: eng2b (ENSDARG00000038868)

Location: 2:27,938,253-27,941,786

Gene: eng2b (ENSDARG00000038868)

Homeobox protein engrailed-2b (ZF-En-1)(Eng3) Source: UniProtKB/Swiss-Prot:P91333

Location: Chromosome 2: 27,938,253-27,941,786 forward strand.

Transcripts: There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
eng2b	ENSDART00000056748	ENSDARFP00000056747	protein_coding

Protein families [help](#) [Variation Table >](#)

Family ID: ENSFM0050000270570

Consensus annotation: RECNAME: FULL-HOMEOBOX ENGRAILED

Other Zebrafish transcripts in this family: [ENSDART00000056748 \(eng2b\)](#)

Multiple alignments: 59 Ensembl members of this family [JaView](#)
94 members of this family [JaView](#)

(4 genes)
[\(all proteins in family\)](#)

Ensembl release 54 - May 2009 © WTSI / EBI

Permanent link - View in archive site

Click on the 4 genes link to bring up family view.

The screenshot shows the Ensembl genome browser interface for the gene **eng2b** (ENSDARG00000038868). The main content area features a karyotype ideogram of Zebrafish chromosomes, with vertical bars representing genes. The **eng** family genes are highlighted: **eng1a** (chromosome 1), **eng1b** (chromosome 1), **eng2a** (chromosome 7), **eng2b** (chromosome 7), and **eng13** (chromosome 13). Below the ideogram, a table lists the gene IDs and locations for these family members.

Gene ID and Location	Gene Name	Description (if known)
ENSDARG0000015073 Chromosome 1: 5,908	eng1b	hypothetical protein LOC541371 [Source:RefSeq peptide;Acc:NP_001013516]
ENSDARG00000038868 Chromosome 2: 21,198	eng2b	Homeobox protein engrailed-2b (Zf-En-1)(Eng3) [Source:UniProtKB/Swiss-Prot;Acc:P31533]
ENSDARG00000026599 Chromosome 7: 44,136	eng2a	Homeobox protein engrailed-2a (Zf-En-2)(Eng2) [Source:UniProtKB/Swiss-Prot;Acc:P09015]
ENSDARG00000014321 Chromosome 13: 1911,626	eng1a	Homeobox protein engrailed-1a (Eng1) [Source:UniProtKB/Swiss-Prot;Acc:Q04696]

Ensembl 'FamilyView' provides a list of closely related Ensembl gene predictions together with a consensus family description and shows the chromosomal location of family members on a karyotype ideogram. It also provides a list of vertebrate UniProt sequences and Ensembl protein predictions from other species that have been used to define the family. It therefore provides a way of exploring orthologues and closely related homologues across a range of animal species.

Putative paralogues are also listed in the hyperlinks on the left side of the page – Parologue prediction.

The screenshot shows the 'Orthologues' section for the **eng2b** gene family. It lists putative paralogues within species (Euteleostomi and Bilateria) and across species. The list includes details such as taxonomic level, gene identifiers, UniProt accession numbers, and alignment links.

Species	Gene ID	Gene Name	Description
Euteleostomi	ENSDARG00000026599	eng2a	Homeobox protein engrailed-2a (Zf-En-2)(Eng2) [Source:UniProtKB/Swiss-Prot (P09015)]
	ENSDARG00000015073	eng1b	hypothetical protein LOC541371 [Source:RefSeq peptide;Acc:NP_001013516]
	ENSDARG00000074916	BBA685_DANRE	Novel protein similar to vertebrate engrailed homolog 1 (Zgc:109892, EN1) Fragment [Source:UniProtKB/TrEMBL (BBA685)]
	ENSDARG00000014321	eng1a	Homeobox protein engrailed-1a (Eng1) [Source:UniProtKB/Swiss-Prot (Q04696)]
	ENSDARG00000007891	meox1	mesenchyme homeobox 1 [Source:RefSeq peptide;Acc:NP_001002450]
Bilateria	ENSDARG00000061818	(NP_001038589.1)	hypothetical protein LOC566969 [Source:RefSeq peptide;Acc:NP_001038589]
	ENSDARG00000040911	(LOC566898)	No description
	ENSDARG00000007891	(meox1)	mesenchyme homeobox 1 [Source:RefSeq peptide;Acc:NP_001002450]

The gene orthology and paralogy predictions are generated by a pipeline where maximum likelihood phylogenetic gene trees (generated by PHYML) play a central role. They aim to represent the evolutionary history of gene families, i.e. genes that diverged from a common ancestor. These gene

trees reconciled with their species tree (using RAP) have their internal nodes annotated to distinguish duplication or speciation events.

5. Select Gene Tree from the left hand menu:

Ensembl genome browser 54: D.rerio - Gene Tree - Gene: eng2b (ENSDARG00000038868)

Location: 2:27,938,253-27,941,786 Gene: eng2b

Gene: eng2b (ENSDARG00000038868)

Homeobox protein engrailed-2b (Zf-En-1)(Eng3) Source: UniProtKB/Swiss-Prot:P31133

Location: Chromosome 2: 27,938,253-27,941,786 forward strand.

Transcripts: There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
eng2b	ENSDART00000056748	ENSDARP00000056747	protein_coding

Gene Tree [help](#)

View options:

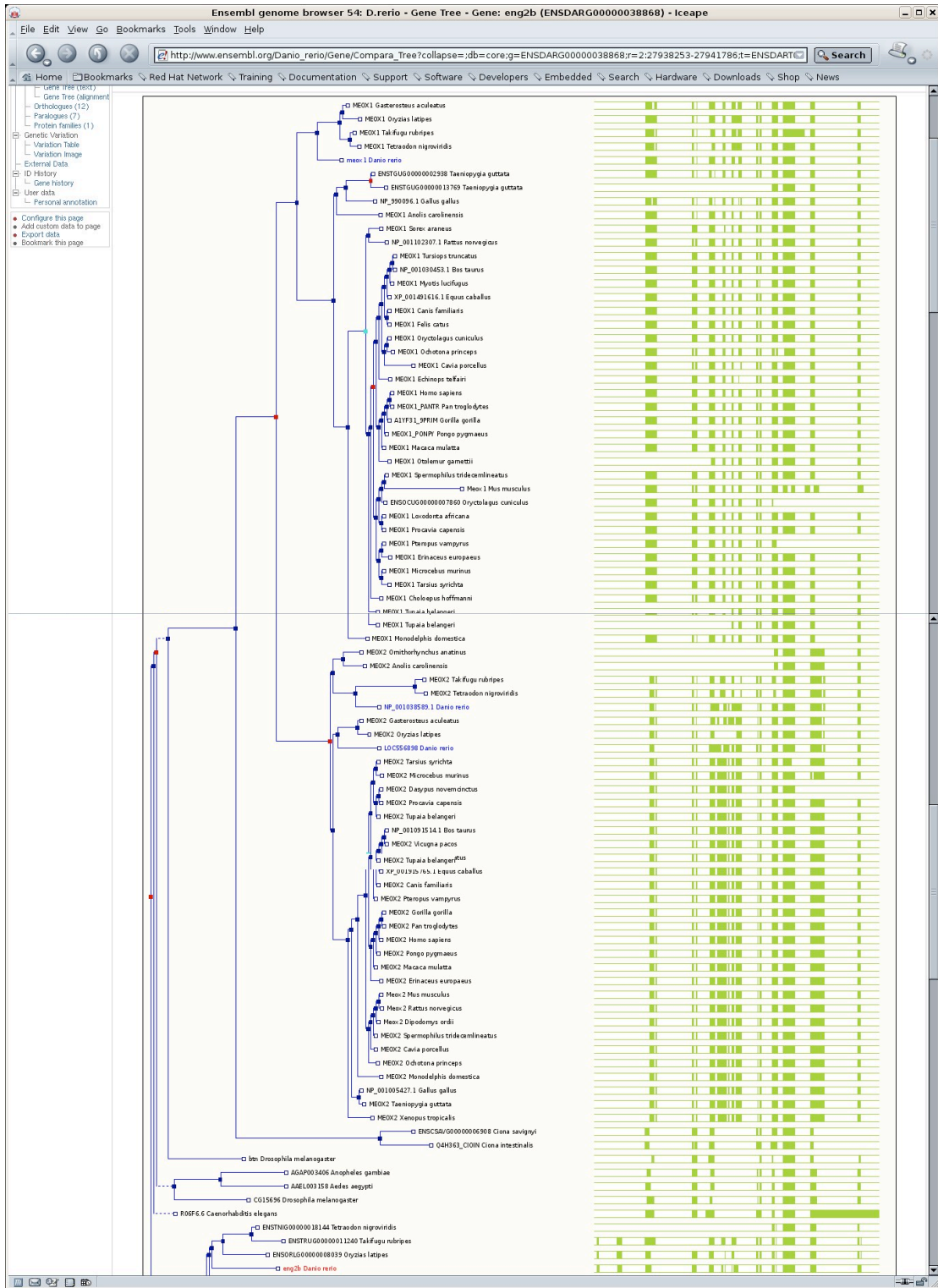
- View current gene only
- View paralogs of current gene
- View all duplication nodes
- View fully expanded tree

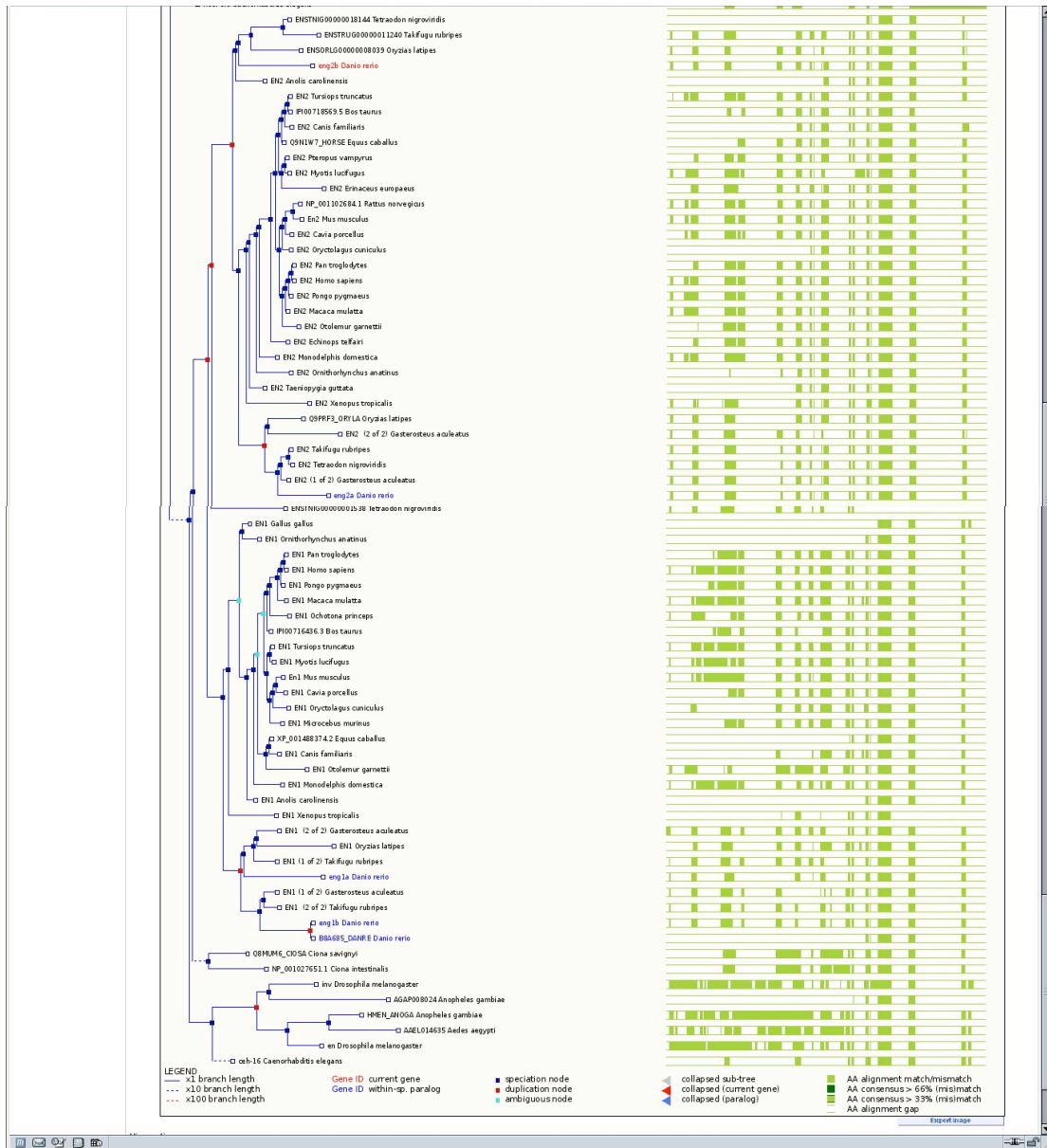
Use the 'configure page' link in the left panel to set the default. Further options are available from menus on individual tree nodes.

Click on view fully expanded tree, under view options to see the full image.

Gene TreeView represents the evolutionary history of gene families.

The Gene Tree displays the maximum likelihood phylogenetic tree (using PHYL) representing the evolutionary history of gene families. Red squares show duplication and blue show speciation. The green bars to the right are a representation of the multiple alignment of peptides made using MUSCLE. Full boxes indicate matches/mismatches, open boxes indicate gaps in the alignment. Both the image and the alignment can be dumped.





Orthologue prediction at Ensembl

This is one of the easiest and most accurate sites for identifying orthologous genes. Information is only available for species that have a sequenced and assembled genome, so if you want to identify sequences from other species you will need to use Blast.

Search for fgf8a:

The screenshot shows the Ensembl genome browser interface for the gene *fgf8a* in Zebrafish. The main content area displays the following information:

- Gene summary:** fibroblast growth factor 8 a [Source:RefSeq peptide;Acc:NP_571356]
- Location:** Chromosome 13: 33,446,378-33,452,845 reverse strand.
- Transcripts:** There is 1 transcript in this gene: [hide transcripts](#)
- Table of transcripts:**

Name	Transcript ID	Protein ID	Description
fgf8a	ENSDFART00000025583	ENSDFARPO0000018653	protein_coding
- Gene summary:** [help](#)
- Name:** *fgf8a* (ZFN)
- Synonyms:** ace, acerebellar, cb110, etD309886.13, fgf-8, fgf8, id:ibid5031, wurfb73a06 [To view all Ensembl genes linked to the name [click here](#).]
- Gene type:** Known protein coding
- Prediction Method:** Transcripts were annotated by the Ensembl [genebuild](#).
- Transcripts:** (Section for transcript details)
- Genomic map:** A visual representation of the gene's location on chromosome 13, showing contigs, the gene structure, and the forward and reverse strands. The map is centered on a 26.47 Kb region from 33,446 Mb to 33,452 Mb.

Click on the orthologues link in the left hand menu.

The screenshot shows the Orthologues page for the gene *fgf8a* in Zebrafish. The page displays a list of orthologous genes across various species. The table includes the following columns:

- Species:** Anole Lizard (*Anolis carolinensis*), Cow (*Bos taurus*), Dog (*Canis familiaris*), Guinea Pig (*Cavia porcellus*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Human (*Homo sapiens*), and others.
- dN/dS:** Ratio of non-synonymous to synonymous substitutions.
- Ensembl identifier:** Ensembl gene ID for the orthologous gene.
- External ref.:** External reference link for the orthologous gene.

Ensembl calculates the closest putative orthologues for species pairs. The stable identifiers of putative orthologous genes lead to the corresponding 'GeneView' display of that gene within the web site for the other species. The homologous genes present the best reciprocal BLAST hits for the two species with additional pairs obtained by a combination of BLAST and location information for more closely related species. These homologues may therefore potentially represent orthologues. Types of orthologous gene pairs are described on the webpage; dN/dS ratios (non-synonymous to synonymous substitutions – an indicator of selective pressure) are also displayed.

Click on align under the human orthologue to show the alignment:

The screenshot shows the Ensembl genome browser interface for the gene **fgf8a**. The main content area displays the gene's location on chromosome 13, its transcripts, and a table of orthologues. The orthologue table shows two species: *Danio rerio* and *Homo sapiens*. Below the table, a multiple sequence alignment is shown using CLUSTAL W (1.61).

Gene: fgf8a (ENSDARG00000003399)
 fibroblast growth factor 8 a [Source:RefSeq peptide;Acc:NP_571356]
Location [Chromosome 13: 33,446,378-33,452,845 reverse strand.](#)
Transcripts There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
fgf8a	ENSDART00000025583	ENSDARP00000018653	protein_coding

Orthologues **Ortholog Alignme**

Ortholog type: 1 to many orthologue

Species	Gene ID	Peptide ID
Danio rerio	ENSDARG00000003399	ENSDARP00000018653
Homo sapiens	ENSG00000107831	ENSP00000321797

CLUSTAL W(1.61) multiple sequence alignment

```

ENSDARP00000018653/1-210  MRLIPSELSYLFHLFAFCYYAQ-----VTIQSPPH
ENSP00000321797/1-244    MGSPRSALSCLLHLVLCQAQEGPGRGPALGRELASLFRAGEPQGVSQVTVQSSPI
*   * * * * . : * * * * *

ENSDARP00000018653/1-210  FTQHVSEQSKVTVDSRRLIRTVQLYSRTSGGHVQVLANIKINAMAEDGDVAKLIUETD
ENSP00000321797/1-244    FTQHVREQSLVTDQLSERLIRTVQLYSRTSGGHVQVLANIKINAMAEDGDFAKLIUETD
*****

ENSDARP00000018653/1-210  TFGSEVRKGAETGFYICMNRGKLGKIKGLGKDCIFTEIVLEHNYTALQKVKYEGWYM
ENSP00000321797/1-244    TFGSEVRVRGAETGLYICMNRGKLGKIAKSNKGGKDCVPTIVLEHNYTALQKVKYEGWYM
*****

ENSDARP00000018653/1-210  APTKRGPRKGSKTRQHQREVRHMKRLPKGHQ--IAEHKPFDFINYPFHR-----TKRTK
ENSP00000321797/1-244    APTKRGPRKGSKTRQHQREVRHMKRLPRGHRTT-EQSLRFELINYPFHR-----TKRTK
*****

ENSDARP00000018653/1-210  YSG-ER
ENSP00000321797/1-244    APEP-R
*
    
```

Ensembl release 54 - May 2009 © [WTSI](#) / [EBI](#)
[Permanent link](#) - [View in archive site](#)

N.B. Orthologue prediction only allows two aligned sequences to be viewed. Aligned eutherian or amniota genome sequences can be viewed within Ensembl using the MLAGAN parameters. Multiple sequence alignment programmes such as ClustalW should be used to align more than two orthologous nucleotide (mRNA) or protein sequences (covered in modules 2 and 3).

Confirming true orthology:

Other features of orthologue prediction will now be used to assess the exon/intron structures and chromosomal context of putative orthologues. Orthologous gene pairs often have similar transcript structures. These can be compared visually using AlignSliceView at Ensembl. This is not available in the current Ensembl release, so we will have to use an archive site.

Click back to the Gene summary page for *fgf8a*, and click on the view in archive site link at the bottom of the page and chose ensembl 47.

The screenshot displays the Ensembl GeneView interface for the gene *fgf8a* (ENSDARG0000003399) in *Danio rerio*. The page is titled "Ensembl Gene Report for ENSDARG0000003399 - Iceape". The main content area shows the gene's location on Chromosome 13 and a description of fibroblast growth factor 8. Below this, there is a section for "Orthologue Prediction" which lists several putative orthologues from various species. The table below summarizes the key information from this section.

Species	Type	Gene identifier
<i>Bos taurus</i>	1-to-many	ENSBTAG0000001530 (QB63D7_BOVIN) [multicontigview] [align]
<i>Canis familiaris</i>	1-to-many	ENSCAF00000008818 (FGF8_CANIFA) [multicontigview] [align]
<i>Ciona intestinalis</i>	1-to-many	ENSCIC00000000393 (NP_001027648.1) [multicontigview] [align]
<i>Ciona savignyi</i>	1-to-many	ENSICSA00000001185 (QB6G2_CIOSA) [multicontigview] [align]
<i>Felis catus</i>	1-to-many	ENSFCAG00000007172 (FGF8) [multicontigview] [align]
<i>Gallus gallus</i>	1-to-many	ENSNGAL00000007708 (FGF8_CHICK) [multicontigview] [align]
<i>Gasterosteus aculeatus</i>	1-to-1	ENSNGAC00000003803 (FGF8 (1 of 2)) [multicontigview] [align]
<i>Homo sapiens</i>	1-to-many	ENSOG00000007831 (FGF8) [multicontigview] [align]
<i>Macaca mulatta</i>	1-to-many	ENSMUL00000001571 (FGF8) [multicontigview] [align]
<i>Monodelphis domestica</i>	1-to-many	ENSMDOD00000001718 (FGF8) [multicontigview] [align]
<i>Mus musculus</i>	1-to-many	ENSMUS000000008219 (Fgf8) [multicontigview] [align]
<i>Myotis bairdii</i>	1-to-many	ENSMYB00000000499 (FGF8) [multicontigview] [align]
<i>Oryzias latipes</i>	1-to-1	ENSOROL00000009819 (FGF8 (1 of 2)) [multicontigview] [align]
<i>Pan troglodytes</i>	1-to-many	ENSPTBG00000002874 (FGF8) [multicontigview] [align]

From within orthologue prediction find the human entry and click on multicontigview to view long-range sequence conservation between zebrafish and human.