# 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 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 with in 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.

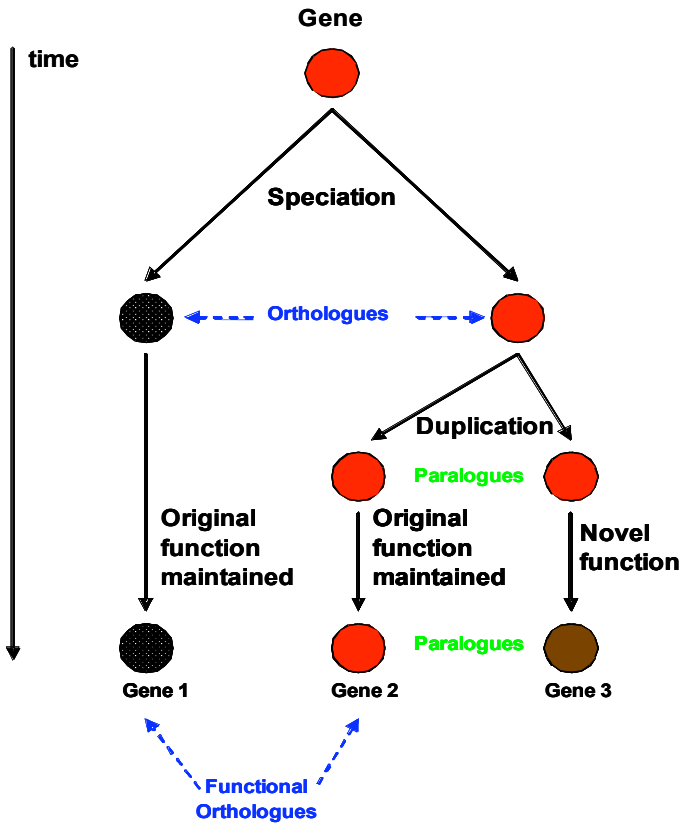



**Figure 6.1 Homologous Gene Sequences**

**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 discontiguous 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).

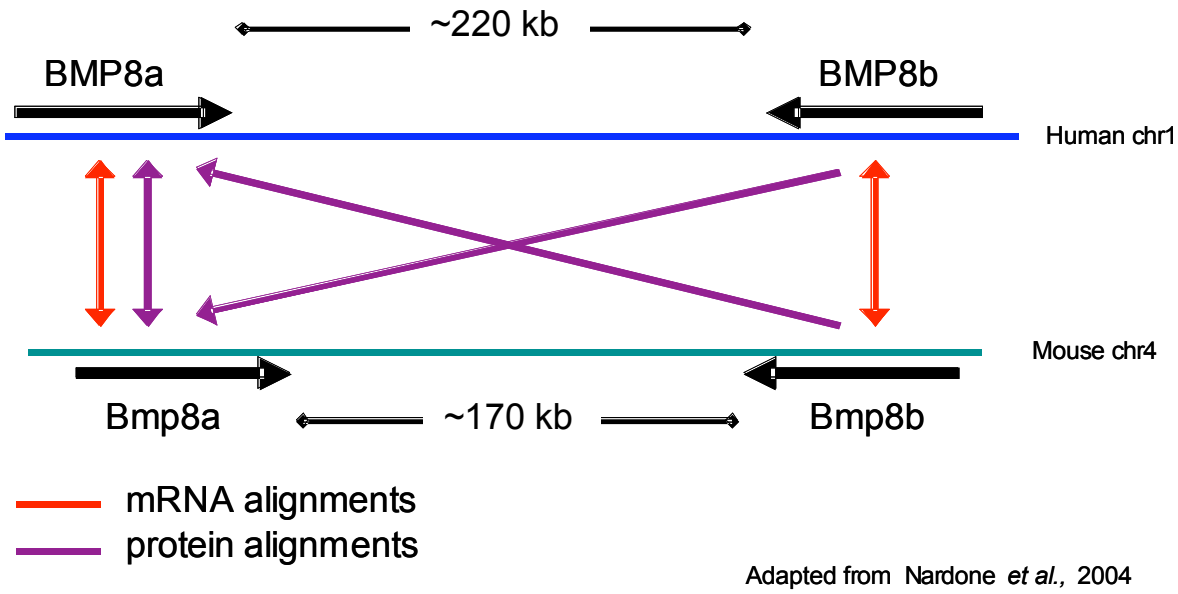



**Figure 4.2 Uncertain *BMP8a* and *BMP8b* orthologues.**

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

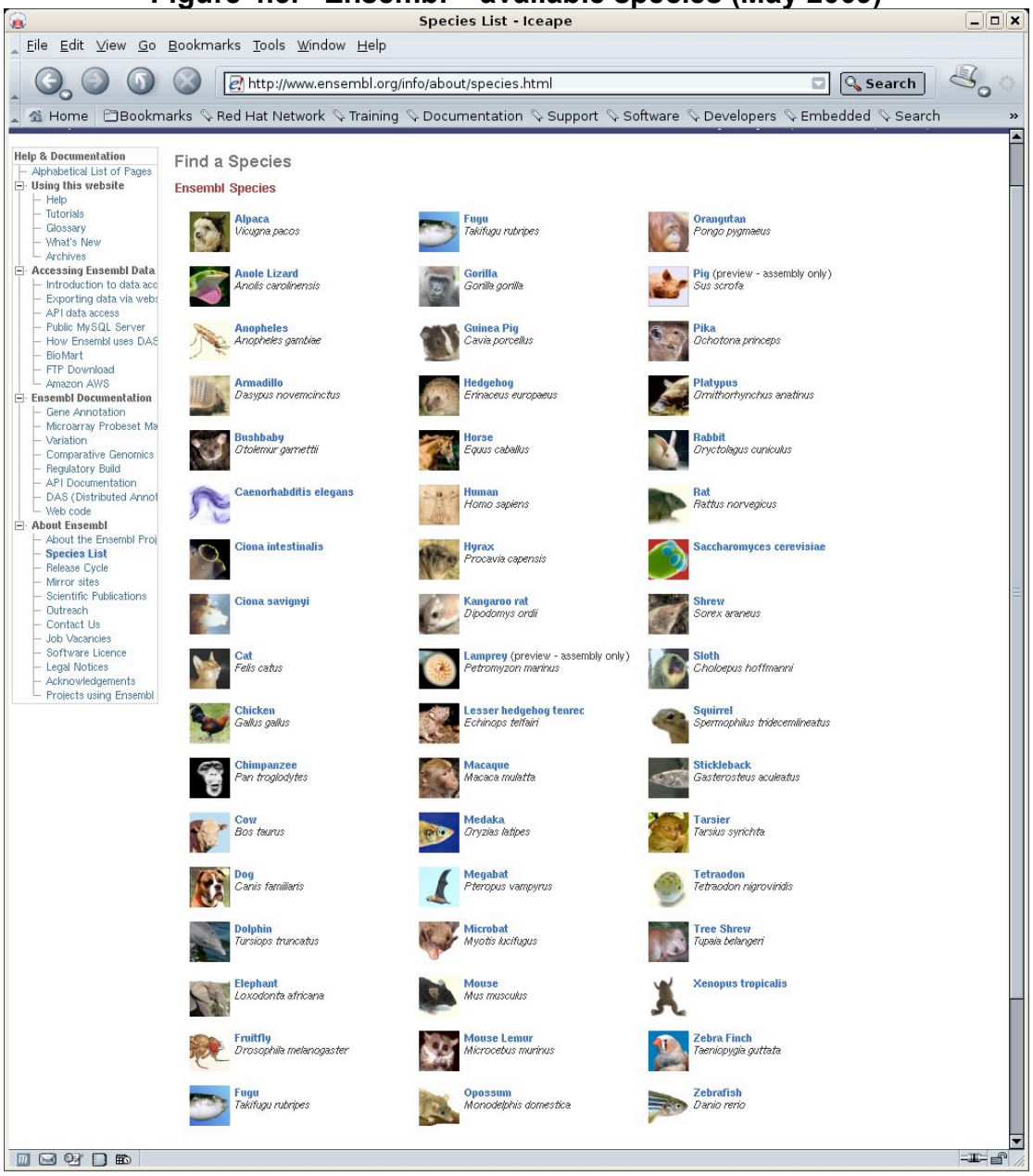
**Comparative Genome Analysis**

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 generated 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 evolutionarily 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<br>Ensembl – Protein family | NCBI - blastn |
| **Identifying orthologues** | Ensembl – Orthologue Prediction<br>UCSC - BLAT | Discontiguous megablast<br>BLAST-n, -p<br>BLink |
| **Confirming true orthology** | Ensembl - MultiContigView<br>- SyntenyView<br>- AlignSliceView | Clustalw |
| **Aligning and identifying conserved sequences** | ECR Browser<br>Genome Vista<br>PhastCons | zPicture<br>Vista<br>LAGAN |

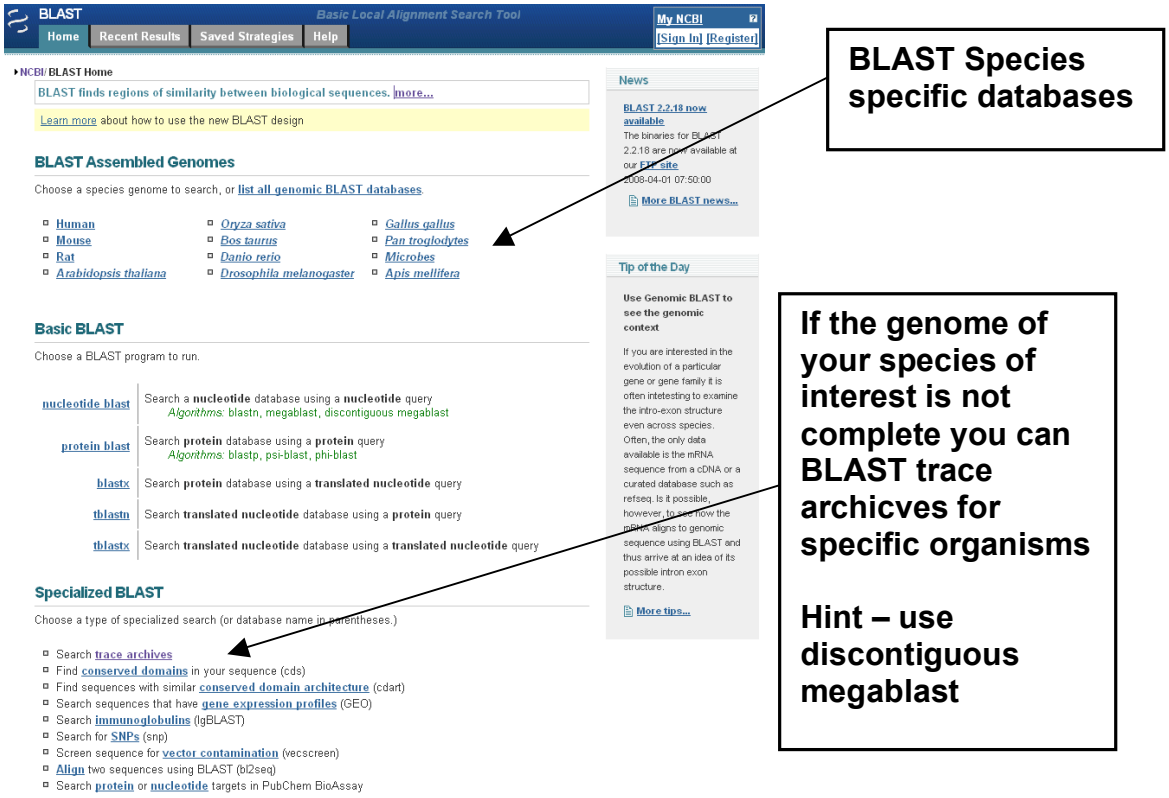
**WORKED EXAMPLES:**

**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.

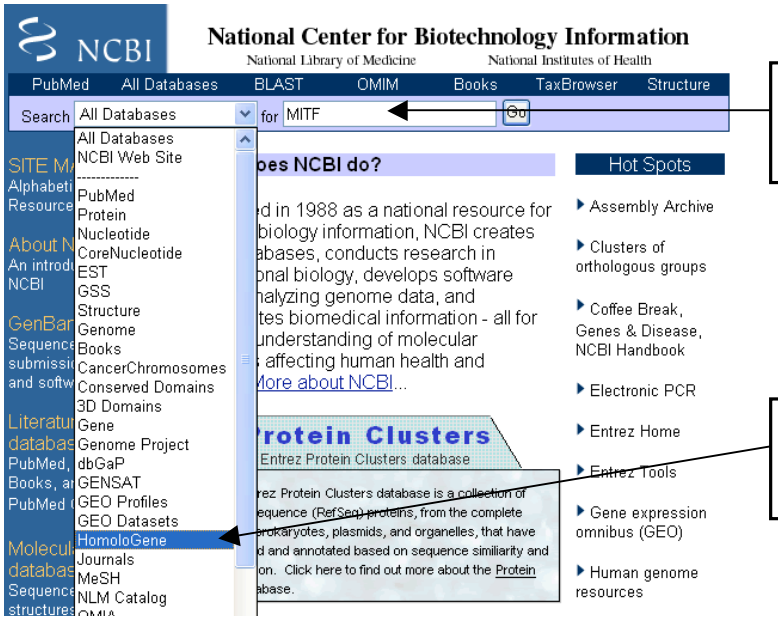
## a) Useful sites at the NCBI

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



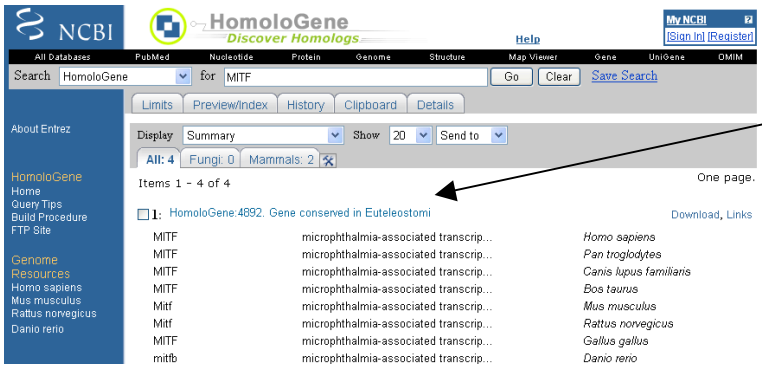

However NCBI has pre-computed orthologues which can be found in the HomoloGene page. However because this page is precomputed this means that it does not contain the most recently assembled genomes.

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



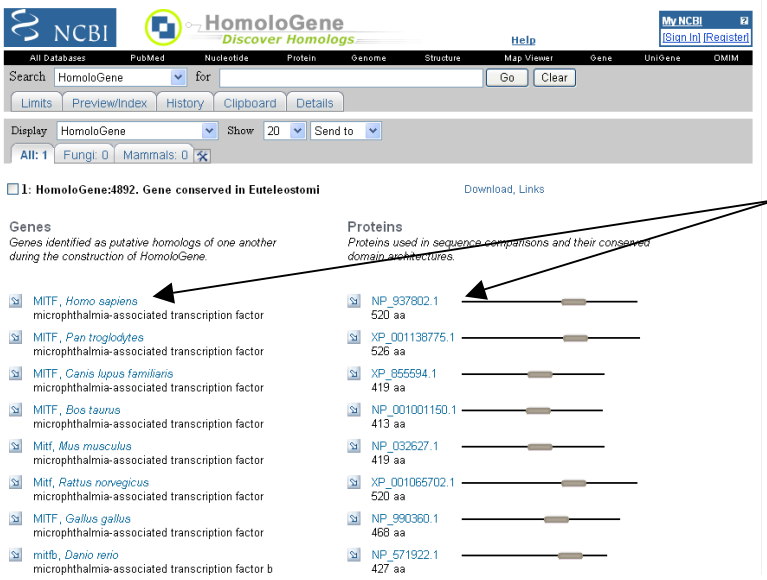

**Type MITF into search box**

**Search HomoloGene**

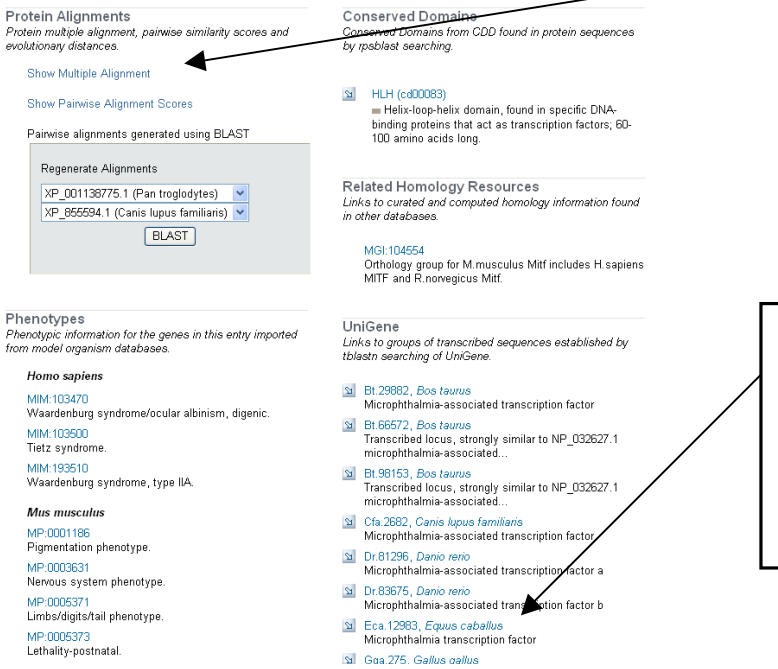



**2: Select HomoloGene:4892**

**The MITF gene family**




**Links to gene and protein information**




**3: Select Show Multiple Alignment**

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

**Multiple Sequence Alignment**

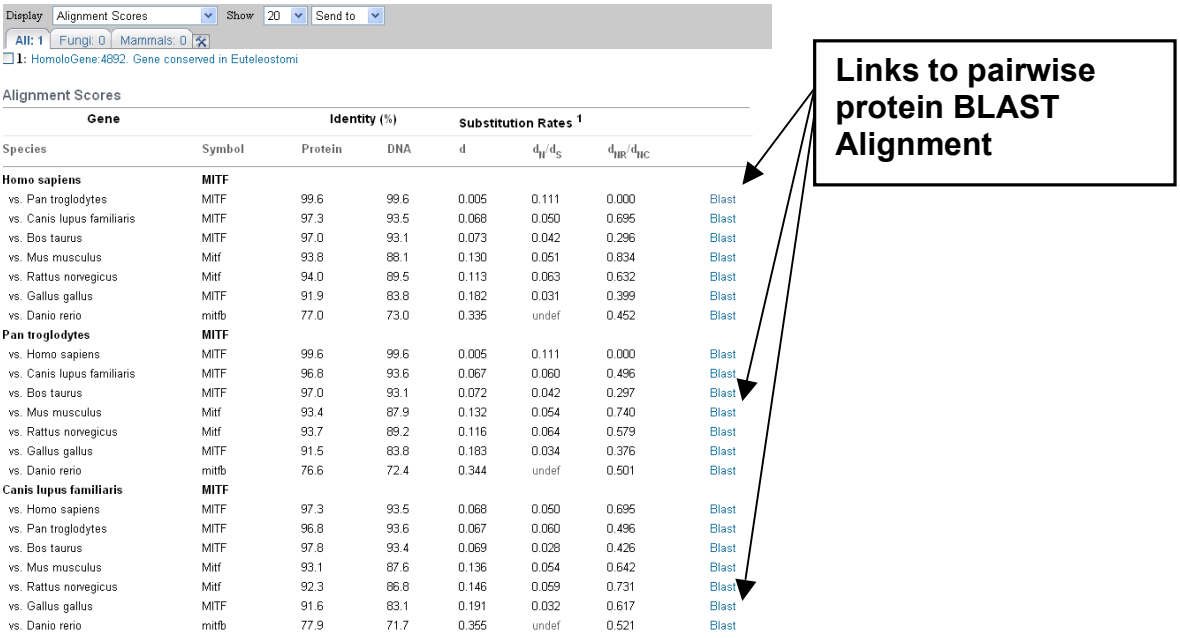
```
NP_937802.1          1   MQSESGIVPDFEVGEEFHEEPKTYYELKSQPLKSSSSAEHPGAS   44
XP_001138775.1       1   MQSESGIVPDFEVGEEFHEEPKTYYELKSQPLKSSSSAEHPGAS   44
XP_855594.1              ---------------------------------------------
NP_001001150.1           ---------------------------------------------
NP_032627.1              ---------------------------------------------
XP_001065702.1       1   MQSESGIVADFEVGEEFHEEPKTYYELKSQPLKSSSSAEHSGAS   44
NP_990360.1              ---------------------------------------------
NP_571922.1          1   MQSESGIVPDFEVGDDFHEEPKTYYELKSQPLQNSNPSEQQ---   41

NP_937802.1         45   KPPISSSSMTSRILLRQQLMREQMQEQERREQQQKLQAAQFMQQ   88
XP_001138775.1      45   KPPISSSSMTSRILLRQQLMREQMQEQERREQQQKLQAAQFMQQ   88
XP_855594.1          1   -------------------MLEMLEYN-----------------   8
NP_001001150.1       1   -------------------MLEMLEYN-----------------   8
NP_032627.1          1   -------------------MLEMLEYS-----------------   8
XP_001065702.1      45   KPPLSSSTMTSRILLRQQLMREQMQEQERREQQQKLQAAQFMQQ   88
NP_990360.1          1   --------MTSRILLRQQLMREQMQEQERREQQQKQQAAQFMQQ   36
NP_571922.1         42   HGSCKPPLGSSRVLLRQQLMREQLQQQERREQQKRQ--------   77

NP_937802.1         89   RVPVSQTPAINVSVPTTLPSATQVPMEVLKVQTHLENPTKYHIQ   132
XP_001138775.1      89   RVPVSQTPAINVSVPTTLPSATQVPMEVLKVQTHLENPTKYHIQ   132
XP_855594.1          9   ---------------------------HYQVQTHLENPTKYHIQ   25
NP_001001150.1       9   ---------------------------HYQVQTHLENPTKYHIQ   25
NP_032627.1          9   ---------------------------HYQVQTHLENPTKYHIQ   25
XP_001065702.1      89   RVAVSQTPAINVSVPTTLPSATQVPMEVLKVQTHLENPTKYHIQ   132
NP_990360.1         37   RVPVSQTPAINVSVPASLPPATQVPMEVLKVQTHLENPTKYHIQ   80
NP_571922.1         78   -ISITHSPAINVSHPCGPPSAAQVPMEVLKVQTHLENPTKYHIQ   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

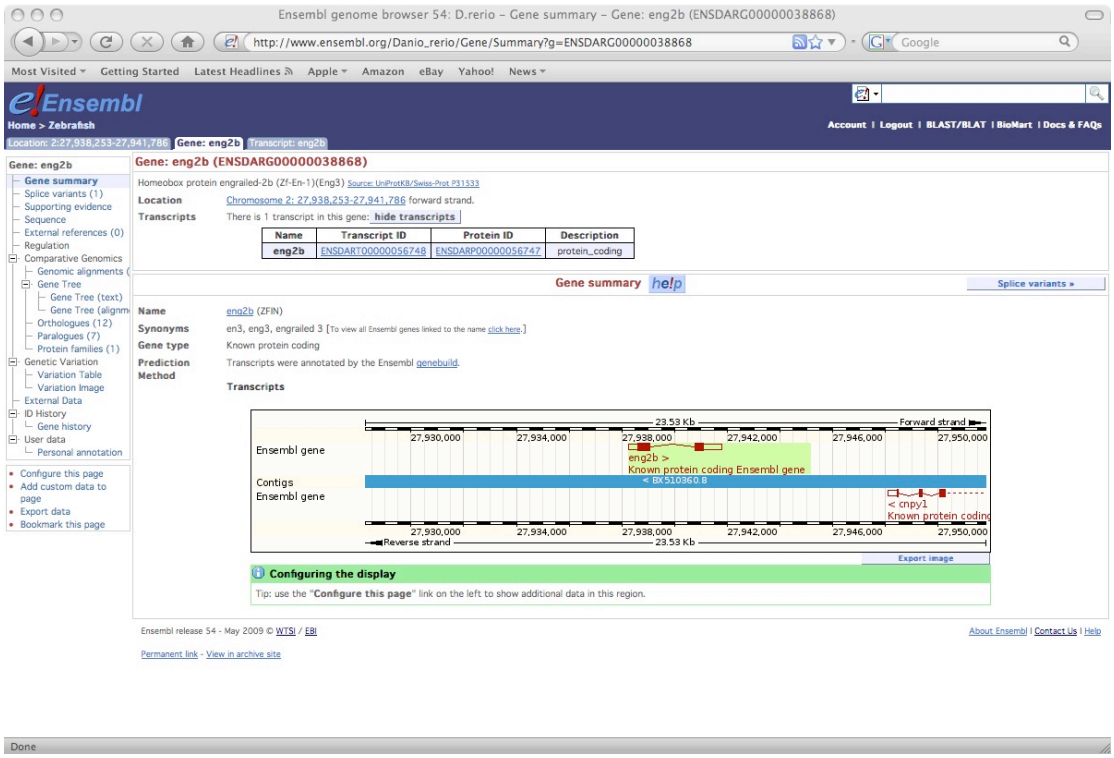
| Gene | | Identity (%) | | Substitution Rates [1] | | | |
|---|---|---|---|---|---|---|---|
| Species | Symbol | Protein | DNA | d | $d_N/d_S$ | $d_{NR}/d_{NC}$ | |
| **Homo sapiens** | **MITF** | | | | | | |
| vs. Pan troglodytes | MITF | 99.6 | 99.6 | 0.005 | 0.111 | 0.000 | Blast |
| vs. Canis lupus 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 | mitfb | 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 | mitfb | 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 | 97.8 | 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 | mitfb | 77.9 | 71.7 | 0.355 | undef | 0.521 | Blast |

- **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.

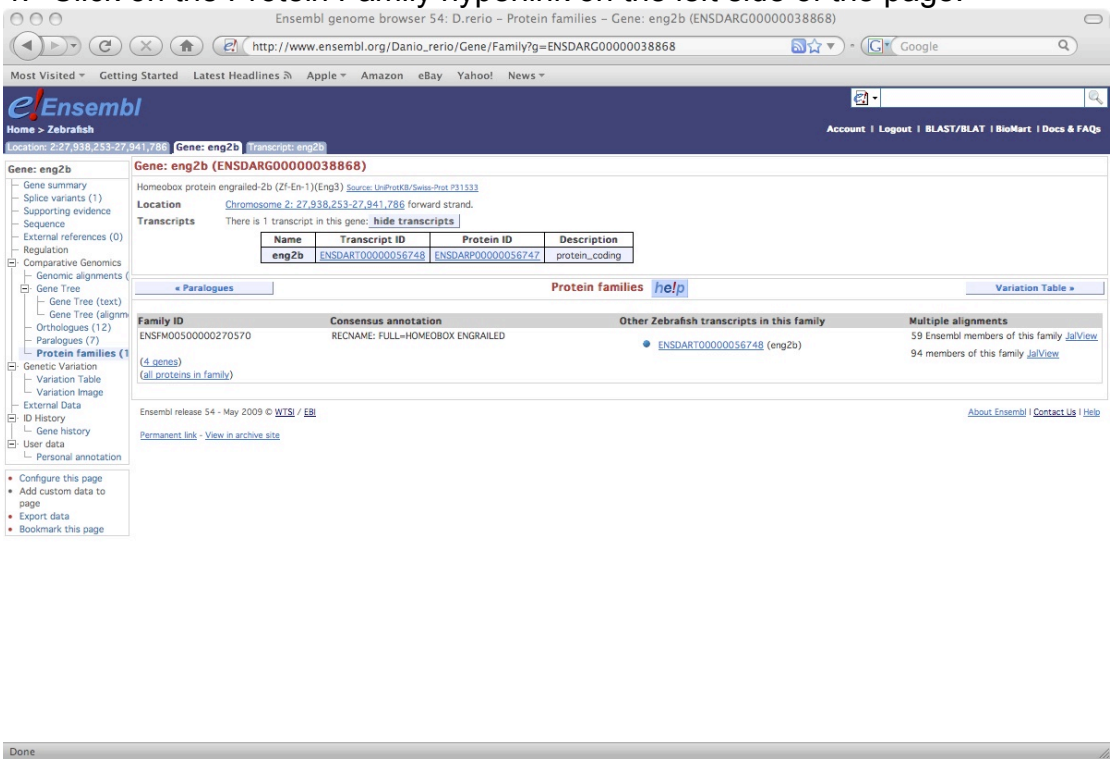
## b) The Ensembl genome browser

1. From the Ensembl homepage select the **Zebrafish** genome browser.
2. Search *e!* zebrafish gene: **eng2b** and press go.
3. Select **Ensembl Gene: ENSDAR00000038868**

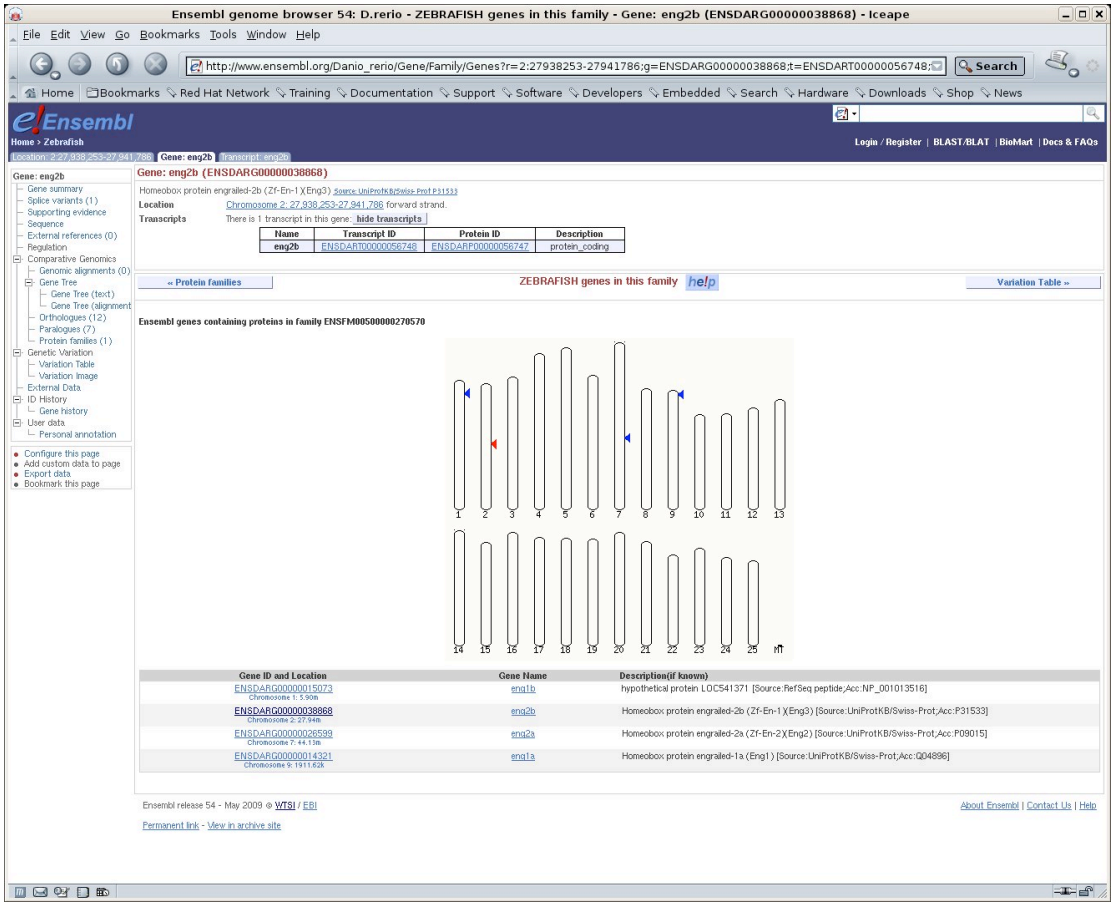
This should bring you to the GeneView page for the eng2b gene:




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




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



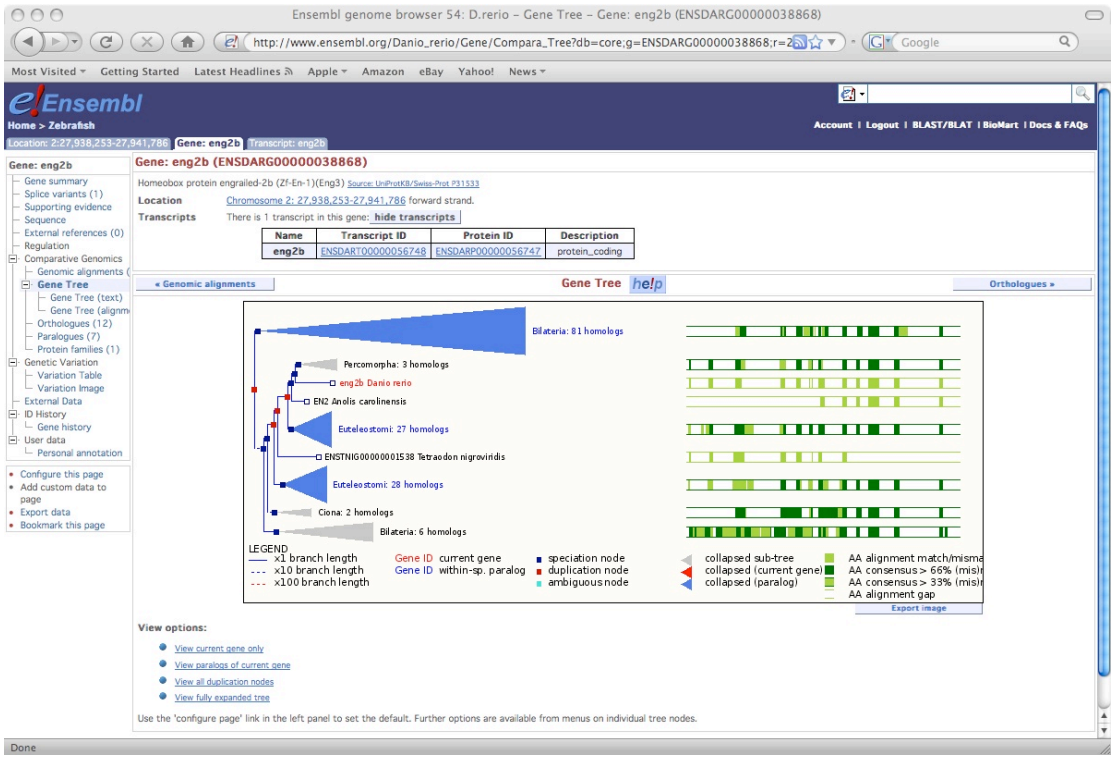

Ensembl '<u>FamilyView</u>' provides a list of closely related Ensembl gene predictions together with a consensus family description and shows the chromosomal location of family members on a karyotype ideogram. It also provides a list of vertebrate <u>UniProt</u> sequences and Ensembl protein predictions 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 – Paralogue prediction.


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.

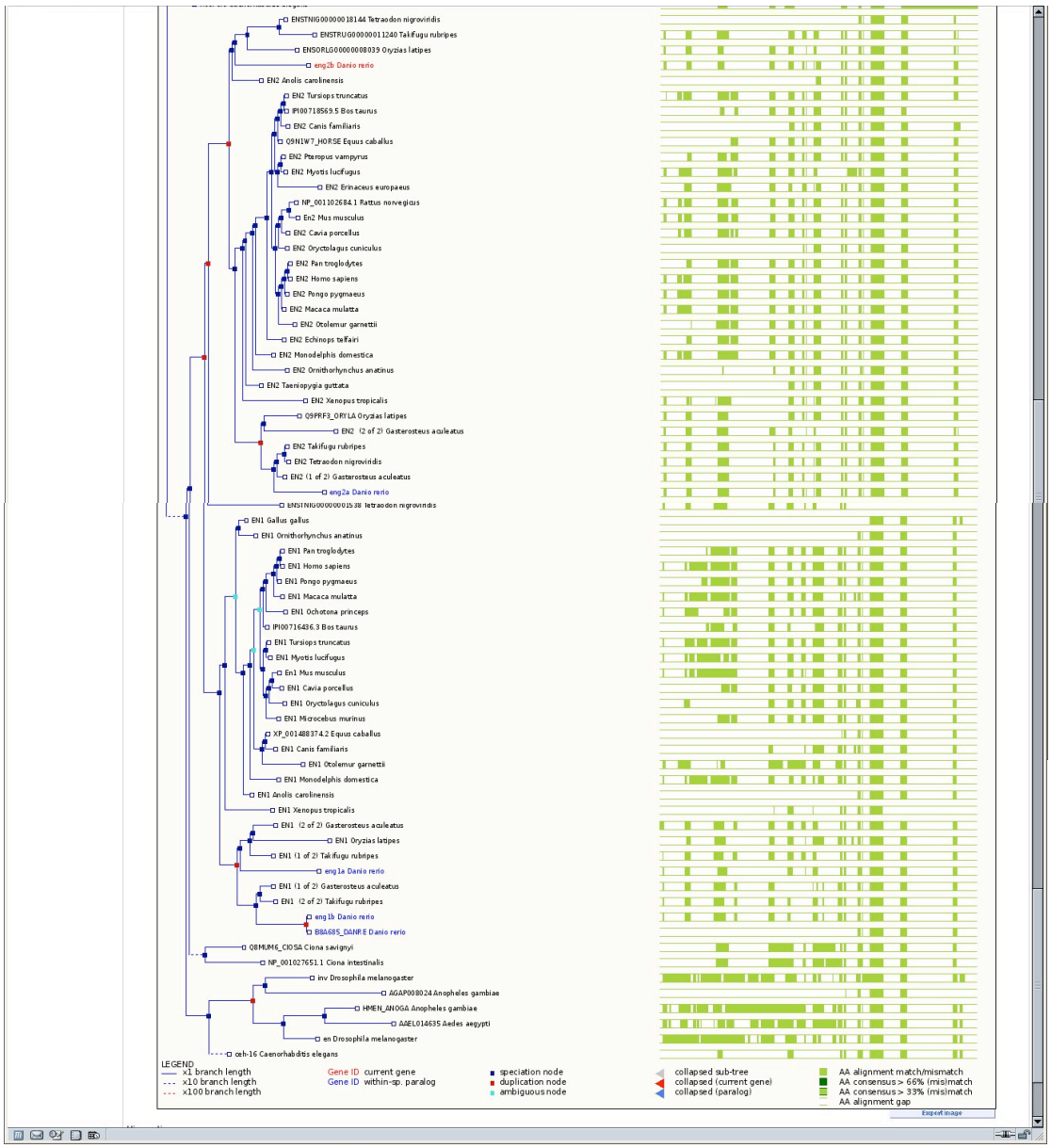
5. Select Gene Tree from the left hand menu:




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

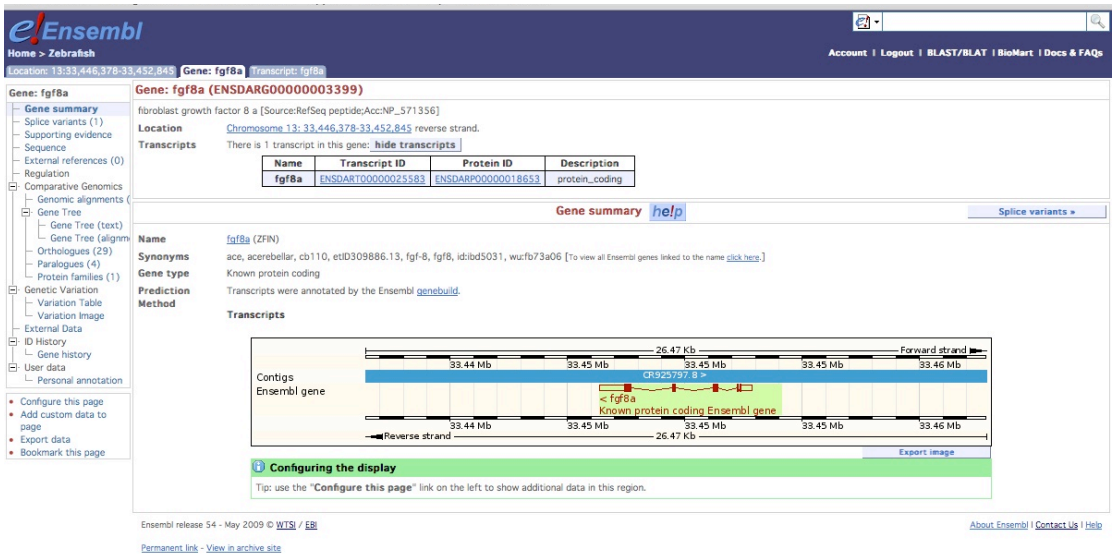
Gene TreeView represents the evolutionary history of gene families.
The Gene Tree displays the maximum likelihood phylogenetic tree (using PHYML) representing the evolutionary history of gene families. Red squares show duplication and blue show speciation. The gereen bars to the right are a representation 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.


Ensembl genome browser 54: D.rerio - Gene Tree - Gene: eng2b (ENSDARG00000038868) - Iceape
File Edit View Go Bookmarks Tools Window Help
http://www.ensembl.org/Danio_rerio/Gene/Compara_Tree?collapse=;db=core;g=ENSDARG00000038868;r=2:27938253-27941786;t=ENSDART
Search
Home Bookmarks Red Hat Network Training Documentation Support Software Developers Embedded Search Hardware Downloads Shop News
Gene Tree (text)
Gene Tree (alignment)
Orthologues (12)
Paralogues (7)
Protein families (1)
Genetic Variation
Variation Table
Variation Image
External Data
ID History
Gene history
User data
Personal annotation
Configure this page
Add custom data to page
Export data
Bookmark this page
MEOX1 Gasterosteus aculeatus
MEOX1 Oryzias latipes
MEOX1 Takifugu rubripes
MEOX1 Tetraodon nigroviridis
meox1 Danio rerio
ENSTGUG00000002938 Taeniopygia guttata
ENSTGUG00000013769 Taeniopygia guttata
NP_990096.1 Gallus gallus
MEOX1 Anolis carolinensis
MEOX1 Sorex araneus
NP_001102307.1 Rattus norvegicus
MEOX1 Tursiops truncatus
NP_001030453.1 Bos taurus
MEOX1 Myotis lucifugus
XP_001491616.1 Equus caballus
MEOX1 Canis familiaris
MEOX1 Felis catus
MEOX1 Oryctolagus cuniculus
MEOX1 Ochotona princeps
MEOX1 Cavia porcellus
MEOX1 Echinops telfairi
MEOX1 Homo sapiens
MEOX1_PANTR Pan troglodytes
A1YF31_9PRIM Gorilla gorilla
MEOX1_PONPY Pongo pygmaeus
MEOX1 Macaca mulatta
MEOX1 Otolemur garnettii
MEOX1 Spermophilus tridecemlineatus
Meox1 Mus musculus
ENSOCUG00000007860 Oryctolagus cuniculus
MEOX1 Loxodonta africana
MEOX1 Procavia capensis
MEOX1 Pteropus vampyrus
MEOX1 Erinaceus europaeus
MEOX1 Microcebus murinus
MEOX1 Tarsius syrichta
MEOX1 Choloepus hoffmanni
MEOX1 Tupaia belangeri
MEOX1 Monodelphis domestica
MEOX2 Ornithorhynchus anatinus
MEOX2 Anolis carolinensis
MEOX2 Takifugu rubripes
MEOX2 Tetraodon nigroviridis
NP_001038589.1 Danio rerio
MEOX2 Gasterosteus aculeatus
MEOX2 Oryzias latipes
LOC556898 Danio rerio
MEOX2 Tarsius syrichta
MEOX2 Microcebus murinus
MEOX2 Dasypus novemcinctus
MEOX2 Procavia capensis
MEOX2 Tupaia belangeri
NP_001091514.1 Bos taurus
MEOX2 Vicugna pacos
MEOX2 Tupaia belangeri
XP_001915765.1 Equus caballus
MEOX2 Canis familiaris
MEOX2 Pteropus vampyrus
MEOX2 Gorilla gorilla
MEOX2 Pan troglodytes
MEOX2 Homo sapiens
MEOX2 Pongo pygmaeus
MEOX2 Macaca mulatta
MEOX2 Erinaceus europaeus
Meox2 Mus musculus
Meox2 Rattus norvegicus
Meox2 Dipodomys ordii
MEOX2 Spermophilus tridecemlineatus
MEOX2 Cavia porcellus
MEOX2 Ochotona princeps
MEOX2 Monodelphis domestica
NP_001005427.1 Gallus gallus
MEOX2 Taeniopygia guttata
MEOX2 Xenopus tropicalis
ENSCSAVG00000006908 Ciona savignyi
Q4H363_CIOIN Ciona intestinalis
btn Drosophila melanogaster
AGAP003406 Anopheles gambiae
AAEL003158 Aedes aegypti
CG15696 Drosophila melanogaster
R06F6.6 Caenorhabditis elegans
ENSTNIG00000018144 Tetraodon nigroviridis
ENSTRUG00000011240 Takifugu rubripes
ENSORLG00000008039 Oryzias latipes
eng2b Danio rerio

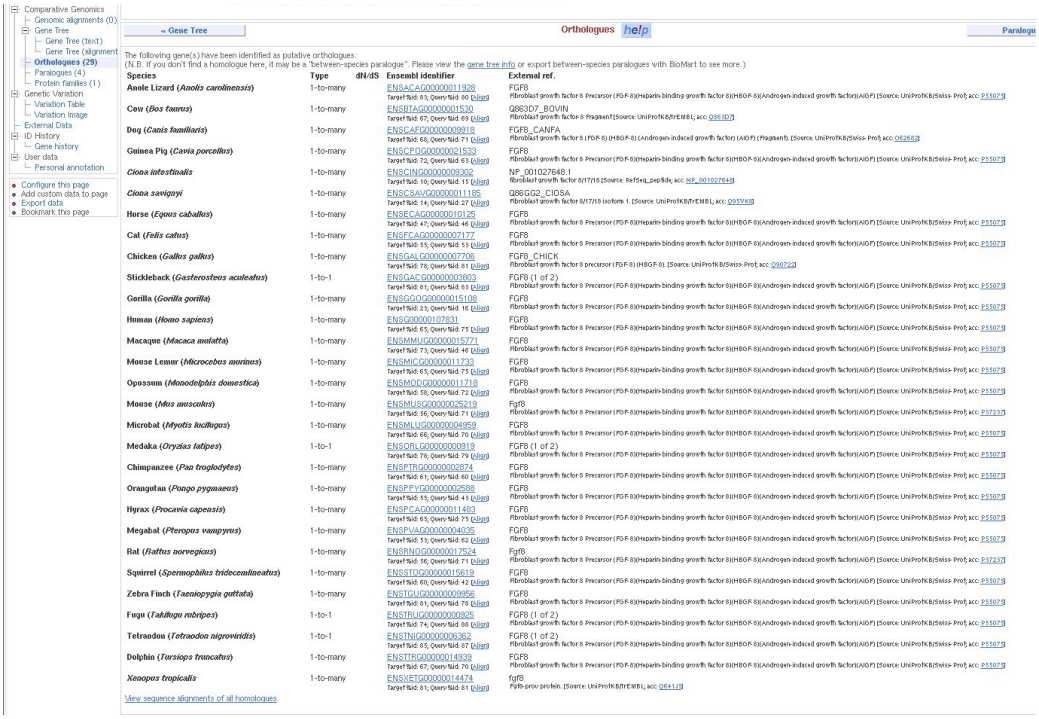

ENSTNIG00000018144 Tetraodon nigroviridis
ENSTRUG00000011240 Takifugu rubripes
ENSORLG00000008039 Oryzias latipes
eng2b Danio rerio
EN2 Anolis carolinensis
EN2 Tursiops truncatus
IPI00718569.5 Bos taurus
EN2 Canis familiaris
Q9N1W7_HORSE Equus caballus
EN2 Pteropus vampyrus
EN2 Myotis lucifugus
EN2 Erinaceus europaeus
NP_001102684.1 Rattus norvegicus
En2 Mus musculus
EN2 Cavia porcellus
EN2 Oryctolagus cuniculus
EN2 Pan troglodytes
EN2 Homo sapiens
EN2 Pongo pygmaeus
EN2 Macaca mulatta
EN2 Otolemur garnettii
EN2 Echinops telfairi
EN2 Monodelphis domestica
EN2 Ornithorhynchus anatinus
EN2 Taeniopygia guttata
EN2 Xenopus tropicalis
Q9PRF3_ORYLA Oryzias latipes
EN2 (2 of 2) Gasterosteus aculeatus
EN2 Takifugu rubripes
EN2 Tetraodon nigroviridis
EN2 (1 of 2) Gasterosteus aculeatus
eng2a Danio rerio
ENSTNIG00000015138 Tetraodon nigrovindis
EN1 Gallus gallus
EN1 Ornithorhynchus anatinus
EN1 Pan troglodytes
EN1 Homo sapiens
EN1 Pongo pygmaeus
EN1 Macaca mulatta
EN1 Ochotona princeps
IPI00716436.3 Bos taurus
EN1 Tursiops truncatus
EN1 Myotis lucifugus
En1 Mus musculus
EN1 Cavia porcellus
EN1 Oryctolagus cuniculus
EN1 Microcebus murinus
XP_001488374.2 Equus caballus
EN1 Canis familiaris
EN1 Otolemur garnettii
EN1 Monodelphis domestica
EN1 Anolis carolinensis
EN1 Xenopus tropicalis
EN1 (2 of 2) Gasterosteus aculeatus
EN1 Oryzias latipes
EN1 (1 of 2) Takifugu rubripes
eng1a Danio rerio
EN1 (1 of 2) Gasterosteus aculeatus
EN1 (2 of 2) Takifugu rubripes
eng1b Danio rerio
BBA68S_DANRE Danio rerio
Q8MUM6_CIOSA Ciona savignyi
NP_001027651.1 Ciona intestinalis
inv Drosophila melanogaster
AGAP008024 Anopheles gambiae
HMEN_ANOGA Anopheles gambiae
AAEL014635 Aedes aegypti
en Drosophila melanogaster
ceh-16 Caenorhabditis elegans
LEGEND
x1 branch length
x10 branch length
x100 branch length
Gene ID current gene
Gene ID within-sp. paralog
speciation node
duplication node
ambiguous node
collapsed sub-tree
collapsed (current gene)
collapsed (paralog)
AA alignment match/mismatch
AA consensus > 66% (mis)match
AA consensus > 33% (mis)match
AA alignment gap
Export image

## 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:

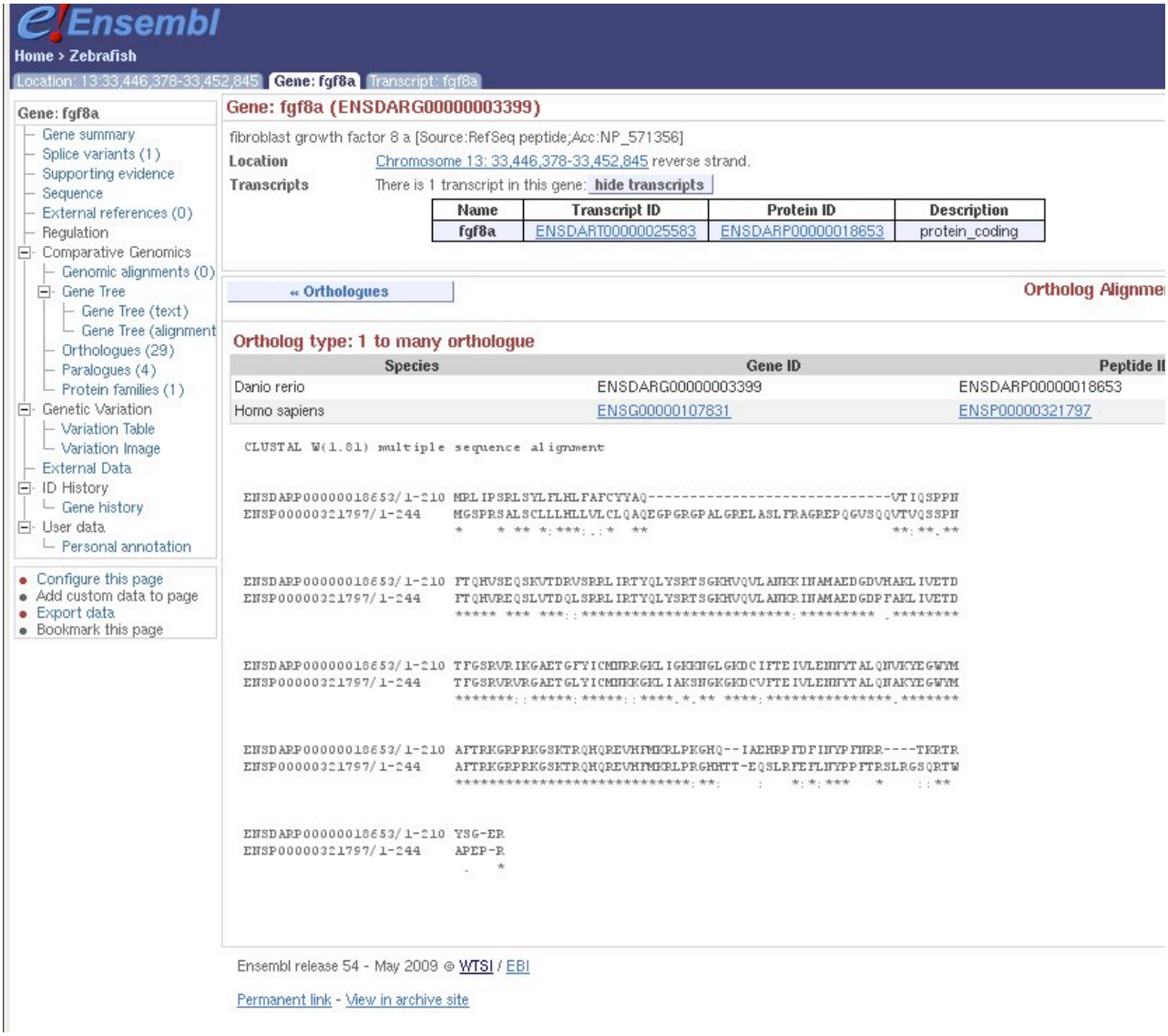



Click on the orthologues link in the left hand menu.




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 represent 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:

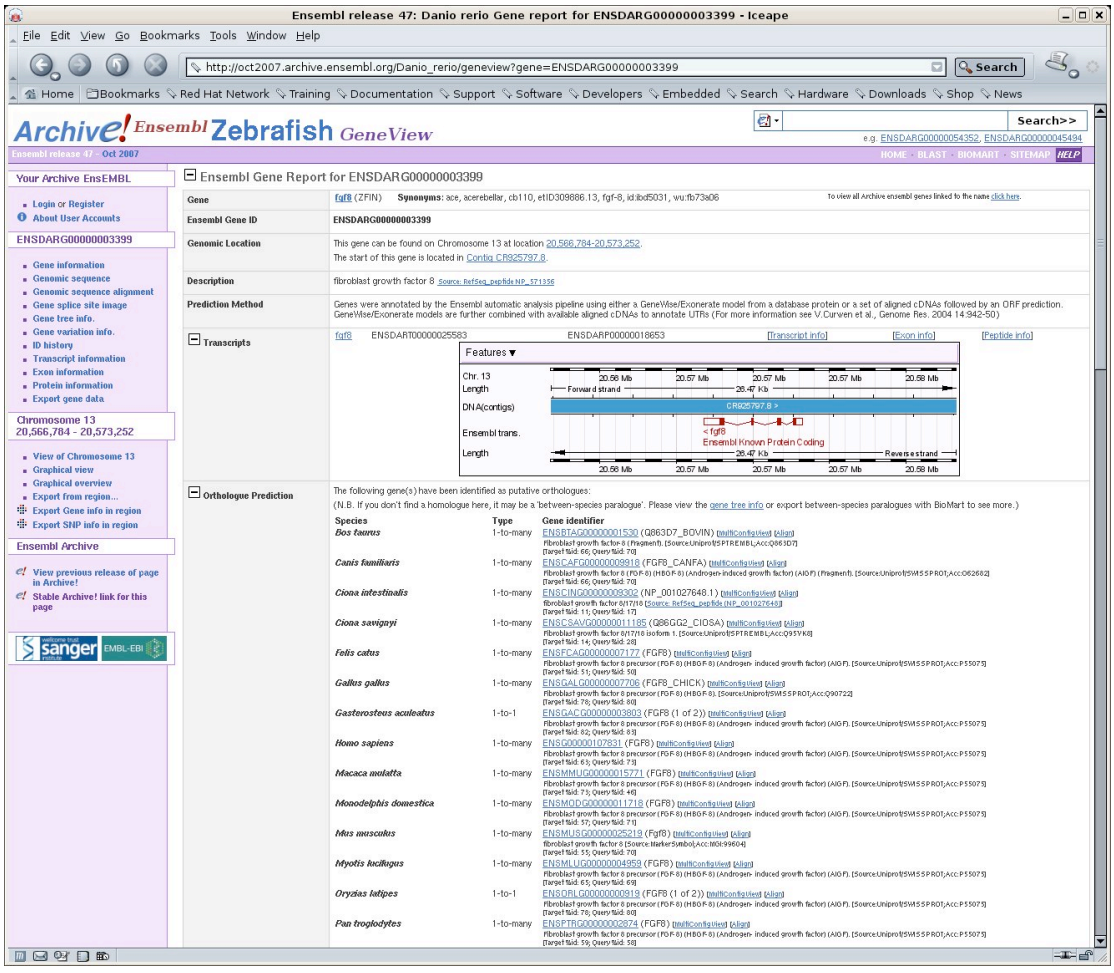



N.B. Orthologue prediction only allows two aligned sequences to be viewed. Aligned eutherian or amniota genome sequences can be viewed within Ensembl using the MLAGAN parameters. Multiple sequence alignment programmes such as ClustalW should be used to align more than two orthologous nucleotide (mRNA)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




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