

Module 3: Genes and Sequences (Sanger)

ii - Does My Gene Have Known Homologues/Orthologues?

Aims

- Introduce the Compara database
- Explain how Compara data is generated
- Explain how Ensembl predictions are named
- Show how to use orthologues to find a gene in zebrafish
- Introduce MultiContigView

Introduction

Ensembl focuses on metazoan (animal) genomes. The genomes currently available on the Ensembl site are:

- Vertebrates: human, chimpanzee, mouse, rat, dog, chicken, puffer fish, zebrafish, Tetraodon
- Tunicates: *Ciona intestinalis*
- Arthropods: the mosquito *Anopheles gambiae*, *Drosophila melanogaster* and honeybee
- Nematodes: *Caenorhabditis elegans*
- Yeast: *Saccharomyces cerevisiae*

You can reach the home pages for each species via the generic Ensembl home page:

<http://www.ensembl.org>

or by bookmarking a species home page with a URL like:

http://www.ensembl.org/Rattus_norvegicus

For those species for which there is an assembly but not yet any annotation, there is a Preview browser (Pre!). The latest zebrafish assembly Zv5 is currently in a pre-site:

http://pre.ensembl.org/Danio_rerio

For most species, Ensembl runs an automated sequence annotation pipeline and gene build to provide annotation including genome-wide gene and protein sets. There are different challenges associated with building a comprehensive gene set in different organisms. For species where the research community is generating comprehensive manual annotation, Ensembl incorporates those gene and protein sets instead of, or in addition to, its own automated annotation. Thus, manual annotation is displayed for some human chromosomes alongside the Ensembl predictions, and the manually curated genome-wide gene sets for *D. melanogaster*, *S. cerevisiae* and *C. elegans* are used in place of an Ensembl set. Additional types of annotation available will vary to some extent between species. But because annotation is stored and displayed in a consistent way for all species, your experience working

with one species will transfer to a new species. Comparisons of genomic sequence and homologous genes and proteins between species are facilitated.

The *Compara* database is a single multi-species database which stores information on:

- whole genome alignments
- gene orthology/paralogy prediction
- protein clustering
- synteny regions (not available for zebrafish)

In module 2.v we describe how to search for a gene for which you know its sequence (cDNA/protein). In this module we investigate how to use orthology to map a gene known from another species into the zebrafish genome. At the moment the compara database is built only for Ensembl but, with the completion of the genome approaching, Vega will soon add this kind of service.

Whole genome alignments

The alignment of the whole DNA sequence from two organisms is computationally demanding. Such data are of great interest both in studies of the mechanisms of molecular evolution and in attempts to identify conserved functional sequences such as novel genes and regulatory regions. Whole genome alignments become increasingly difficult as the evolutionary distance between two organisms increases. Ensembl is experimenting with different procedures for performing the alignments. Translated BLAT is used to compare, at the amino acid level, genomes from more evolutionarily distant species. Thus regions of similarity will be biased towards those that code for proteins, although highly conserved non-coding regions might be detected as well. You can show a number of tracks displaying the conservation from the 'Compara' menu in ContigView. Links make it easy to navigate back and forth to see details of the region in the two genomes and to download the sequence of regions of interest.

Open the ContigView page showing the jag2 zebrafish gene (see module 2.i) and select from the 'Compara' menu the Fugu translated BLAT track.

Compara menu

Fugu conserved regions

Conserved blocks for the exons of jag2

Detailed view
 Jump to region: 20 bp 25163432 to 25378212
 Features: Compara DAS Sources Repeats Decorations Export Jump to Image size Help

Gene legend
 EST GENE
 ENSEMBL PREDICTED GENES (KNOWN)
 ENSEMBL PREDICTED GENES (NOVEL)
 ENSEMBL PSEUDOGENES

Every conserved block has an associated pop-up window with some options. You can jump to the corresponding Fugu ContigView Page but, more interestingly, you can open a MultiContigView page.

Select the block that corresponds to the first exon of jag2 in zebrafish and jump to MultiContigView.

Ensembl Zebrafish Genome Browser (MultiContigView)

http://www.ensembl.org/Danio_rerio/multicontigview?c=20:25336...

correspondence between conserved regions

Ensembl Zebrafish MultiContigView
 Home Zebrafish What's New TextSearch BlastSearch MapSearch Export Data Downloads Archive sites
 Find: All [Search] [e.g. Zv4_NA9133, ENSDAF]
 Top level
 Overview
 Detailed view
 Jump to region: 20 bp 25311978 to 25361977
 Features: Compara Repeats Decorations Jump to Export Image size Help

Date: Fri Jul 8 10:13:25 2005
 Archived [Permissions page link]
 Help Desk / Suggestions

A **MultiContigView** allows the visualisation of syntenic regions from multiple species. In the example above a region from chromosome 20 in zebrafish is compared to scaffold_3358 in the Fugu assembly. The conserved blocks are connected with green lines. Observe that in this example the exons from jag2 are projected onto a predicted gene in Fugu. The Fugu gene has not been named though. This view gives some evidence that this gene is perhaps the orthologous Fugu jag2 (or a fragment of it). It is also interesting to note the difference in scale between the zebrafish region and the Fugu scaffold (the Fugu genome is almost five times smaller than the zebrafish genome).

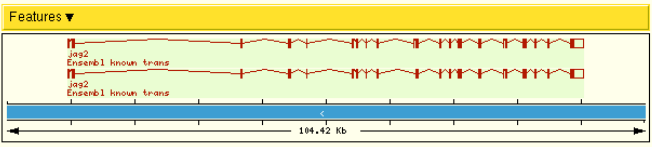
Orthologue predictions

Another kind of comparative analysis focuses on genes and proteins, and attempts to identify orthologues in different genomes. The classic ‘model’ animals are now all represented in Ensembl (*Drosophila*, *C. elegans*, mouse) as well as zebrafish. The automated identification of orthologues is made more difficult by the existence of families of closely related genes. Under such circumstances, Ensembl may show more than one potential orthologue, and the results need to be treated with caution.

Ensembl shows the information about potential orthologues on each GeneView page. The procedure has been applied to all pairs of vertebrates within Ensembl, to the two nematodes, and to the two insects. Open the GeneView page for jag2 in Ensembl and scroll down to the ‘Orthologue Prediction’ entry.

Transcript Structure

1: [jag2](#) (ENSDFART00000024922) [\[Transcript information\]](#) [\[Exon information\]](#) [\[Protein information\]](#)
 2: [jag2](#) (ENSDFART00000049586) [\[Transcript information\]](#) [\[Exon information\]](#) [\[Protein information\]](#)



The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis:

Species	Type	dN/dS	Gene identifier
<i>Drosophila melanogaster</i>	UBRH	-	CG6127 (Ser) [MultiContigView] [Align]
<i>Anopheles gambiae</i>	UBRH	-	ENSANGG00000008746 (Novel Ensembl prediction) [MultiContigView] [Align]
<i>Apis mellifera</i>	UBRH	-	ENSAPMG00000006554 (XM_394560.1) [MultiContigView] [Align] similar to C-Serate-1 protein [Source: RefSeq] [Peptide] [XP_394560]
<i>Ciona intestinalis</i>	UBRH	-	ENSING00000003969 (Novel Ensembl prediction) [MultiContigView] [Align]
<i>Homo sapiens</i>	UBRH	-	ENSG00000184916 (JAG2) [MultiContigView] [Align]
<i>Gallus gallus</i>	UBRH	-	ENSJGALG0000011696 (O12973) [MultiContigView] [Align] C-serate-2 (Fragment) [Source: SPTREMBL] [G42347]
<i>Mus musculus</i>	UBRH	-	ENSMUSG00000002799 (Jag2) [MultiContigView] [Align] Jagged 2 precursor (Jagged2) [Source: Uniprot/SWISSPROT] [Acc:Q9QYE5]
<i>Rattus norvegicus</i>	UBRH	-	ENSRNOG00000013927 (JAG2_RAT) [MultiContigView] [Align] Jagged 2 (Jagged2) (Fragment) [Source: Uniprot/SWISSPROT] [Acc:P97607]
<i>Xenopus tropicalis</i>	UBRH	-	ENSXETG00000006790 (Novel Ensembl prediction) [MultiContigView] [Align]
<i>Tetraodon nigroviridis</i>	UBRH	-	GSTENG00023297001 (GSTENG00023297001) [MultiContigView] [Align]
<i>Fugu rubripes</i>	UBRH	-	SINFRUG00000141145 (Novel Ensembl prediction) [MultiContigView] [Align]

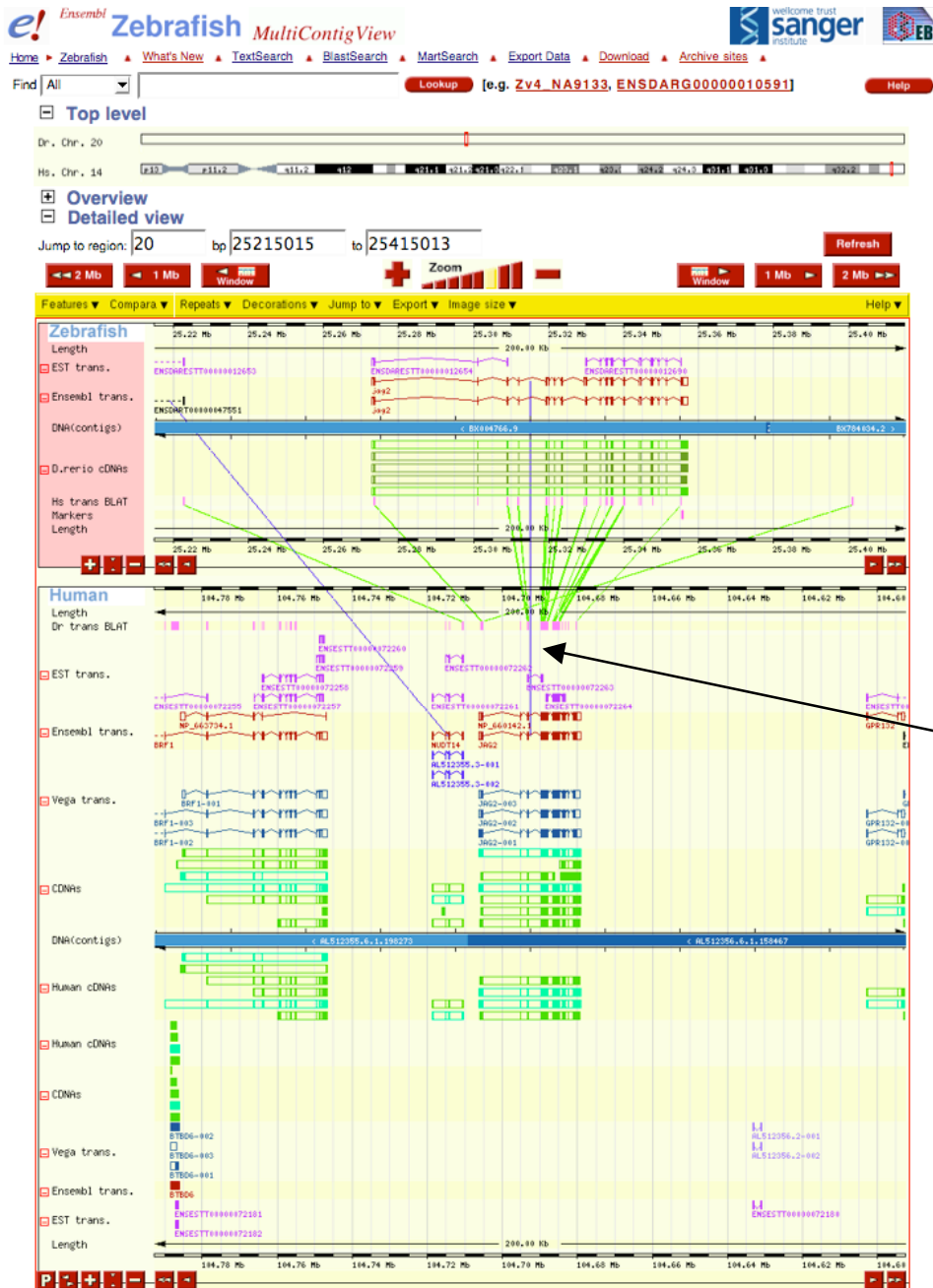
[View alignments of homologies.](#)

UBRH - (U)nique (B)est (R)eciprocal (H)it
 MBRH - one of (M)any (B)est (R)eciprocal (H)its
 RHS = Reciprocal Hit based on Synteny around BRH
 DWGA = Derived from Whole Genome Alignment

Orthologue predictions

In Ensembl, orthologues are identified starting with comparisons at the protein level. ‘All-versus-all’ BLASTP+SW (Smith-Waterman algorithm) is first used to identify those protein pairs that are best reciprocal hits (BRH) between two sets of proteins that represent every gene in the two organisms. Additional putative orthologues are then sought using synteny and these are known as

RHS (Reciprocal Hit supported by Synteny). Where two homologous proteins are encoded by genes each located within 1 Mb of a pair of BRH, they are good candidates for being an additional orthologous pair. Currently we divide these BRH into UBRH (Unique Best Reciprocal Hit) and MBRH (Multiple Best Reciprocal Hit). The latter have multiple but identical best hits, which can happen if there is perfect protein sequence duplication of translated genes within a species. The same approach permits the identification of adjacent family members that may be recently duplicated lineage-specific paralogues. For every orthologue prediction there is a link to a MultiContigView page. Follow the link labelled 'MultiContigView' for the human prediction:

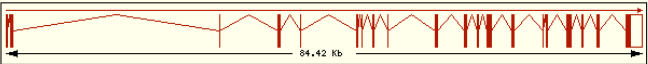


Jag2 in zebrafish corresponds to JAG2 in human. These genes are used to anchor the region. Orthologous genes are indicated by a blue line and conserved blocks are connected by green lines. Observe that there is also a link between two putative orthologous genes on the left.

Protein Families

Another option is to look for proteins that share particular domains. Ensembl runs domain prediction programs on all its protein sets, and provides access to this information in ProteinView (for individual proteins) and in **DomainView** (showing all the genes in a species that share a particular InterPro domain). The family database is generated by running the Tribe-MCL sequence clustering algorithm on a set of peptides consisting of the Ensembl predictions for each species, together with all metazoan sequences from UniProt/Swiss-Prot and UniProt/TrEMBL. On this set of peptides, an all-against-all BLASTP is run to establish similarities. Using these similarities, clusters can be established using the MCL algorithm.

Scroll down in the GeneView page for jag2 until you find the protein family entry:

jag2	Stable ID: ENSDART00000049586 Exons: 25 Transcript length: 5322 bp Translation length: 1216 residues [Transcript information] [Exon information] [Protein information]
Similarity Matches	This Ensembl entry corresponds to the following database identifiers: AFY Zebrafish: Dr.8287.1.S1.a.at EMBL: AF229450 [align] Protein ID: AAL08215.1 [align] UniProt/TrEMBL: Q90Y55 [Target %id: 99; Query %id: 99] [align] ZFIN ID: jag2
InterPro	IPR001438 Type II EGF-like signature - [View other EnsEMBL genes with this domain] IPR001881 EGF-like calcium-binding - [View other EnsEMBL genes with this domain] IPR001687 ATP/GTP-binding site motif A (P-loop) - [View other EnsEMBL genes with this domain] IPR001774 Delta/Serrate/lag-2 (DSL) protein - [View other EnsEMBL genes with this domain] IPR000742 EGF-like domain, subtype 2 - [View other EnsEMBL genes with this domain] IPR000152 Aspartic acid and asparagine hydroxylation site - [View other EnsEMBL genes with this domain] IPR006209 EGF-like domain - [View other EnsEMBL genes with this domain] IPR001007 von Willebrand factor, type C - [View other EnsEMBL genes with this domain]
Protein Family	ENSF00000000049 : PRECURSOR This cluster contains 30 Ensembl gene member(s)
Transcript Structure	

This links to the **FamilyView** page with a summary of all the genes in the family and their orthologues in other species.

Ensembl Zebrafish FamilyView

Home Zebrafish What's New TextSearch BlastSearch MartSearch Export Data Download Archive sites Help

Find [All] [e.g. ENSF0000000084, ENSF00000000823]

Ensembl protein family report

Family ID	ENSF0000000048	
Consensus annotation	PRECURSOR The annotation confidence score of this family is 97	
Prediction method	Protein families were generated using the MCL (Markov CLustering) package available at http://micans.org/mcl/ . The application of MCL to biological graphs was initially proposed by Enright A.J., Van Dongen S. and Ouzounis C.A. (2002) "An efficient algorithm for large-scale detection of protein families." Nucl. Acids. Res. 30, 1575-1584.	
Export data	Export a list of genes containing this family Dump protein sequences for family members in FASTA format Export multiple alignments of all members of this family in format: <input type="text" value="FASTA"/> <input type="button" value="Go"/>	
Multiple alignments	Click to view multiple alignments of the 654 Ensembl members of this family. <input type="button" value="JalView"/> Click to view multiple alignments of all 1005 members of this family. <input type="button" value="JalView"/>	
Ensembl genes containing peptides in family ENSF0000000048	<p>Please click on the gene identifier arrow to go to graphical gene view</p>	
Scaffold	Gene	Description
Zv4_NA13677	ENSDARG00000033981	
Zv4_NA15389	ENSDARG00000005097	Notch homolog 2 [Source:ZFIN;Acc:ZDB-GENE-000329-4]
Zv4_NA1937	ENSDARG00000020219	
Zv4_scaffold1523	ENSDARG00000010522	Notch homolog 3 [Source:ZFIN;Acc:ZDB-GENE-000329-5]
Zv4_scaffold2280	ENSDARG00000017806	
Zv4_scaffold2285	ENSDARG00000028216	Notch homolog 3 [Source:ZFIN;Acc:ZDB-GENE-000329-5]
1	ENSDARG00000010791	DeltaD [Source:ZFIN;Acc:ZDB-GENE-990415-47]
4	ENSDARG00000020680	
5	ENSDARG00000008881	
5	ENSDARG00000004232	DeltaB [Source:ZFIN;Acc:ZDB-GENE-980526-114]
5	ENSDARG00000022080	
5	ENSDARG00000013811	Notch homolog 1b [Source:ZFIN;Acc:ZDB-GENE-990415-183]
7	ENSDARG00000013526	
11	ENSDARG00000021259	
12	ENSDARG00000030946	
12	ENSDARG00000019724	
13	ENSDARG00000011089	
13	ENSDARG00000010677	
13	ENSDARG00000014246	
13	ENSDARG00000029999	
13	ENSDARG00000034035	
15	ENSDARG00000004595	DeltaC [Source:ZFIN;Acc:ZDB-GENE-000125-4]
15	ENSDARG00000002336	DeltaC [Source:ZFIN;Acc:ZDB-GENE-000125-4]
15	ENSDARG00000001374	
16	ENSDARG00000003362	
19	ENSDARG00000013168	
20	ENSDARG00000021389	Jagged 2 [Source:ZFIN;Acc:ZDB-GENE-011128-3]
20	ENSDARG00000012737	
22	ENSDARG000000011213	
24	ENSDARG000000030289	

JalView is an external tool that allows the visualisation and evaluation of multiple alignments between the translations involved.

Ensmart (module 3.v) provides the means to rapidly and easily download sets of transcript or protein sequences with particular domains or from particular families, which can be very useful as starting points for alignment and phylogenetic analysis.

Exercises

1. Blocks of conserved regions can be visualised as dotplot diagrams. Turn on a Compara track and open **DotterView**.
2. Follow the link to the associated protein family of jag2. How many genes produce proteins in this family? Are they all 'known' genes? Are there members of the same family in other species? How many? Have a look at the section Orthologue Predictions. Follow the link to human JAG2.
3. Look for the mouse JAG2 and verify whether it aligns to the zebrafish prediction.
4. Find the zebrafish hoxb1b gene and identify its orthologue in Fugu. Compare the two genes with respect to length and number of exons. Visualise both in MultiContigView. Open the 'homeobox' FamilyView page.