

Module 3 – Genes and Sequences

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. Some of 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!). 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:

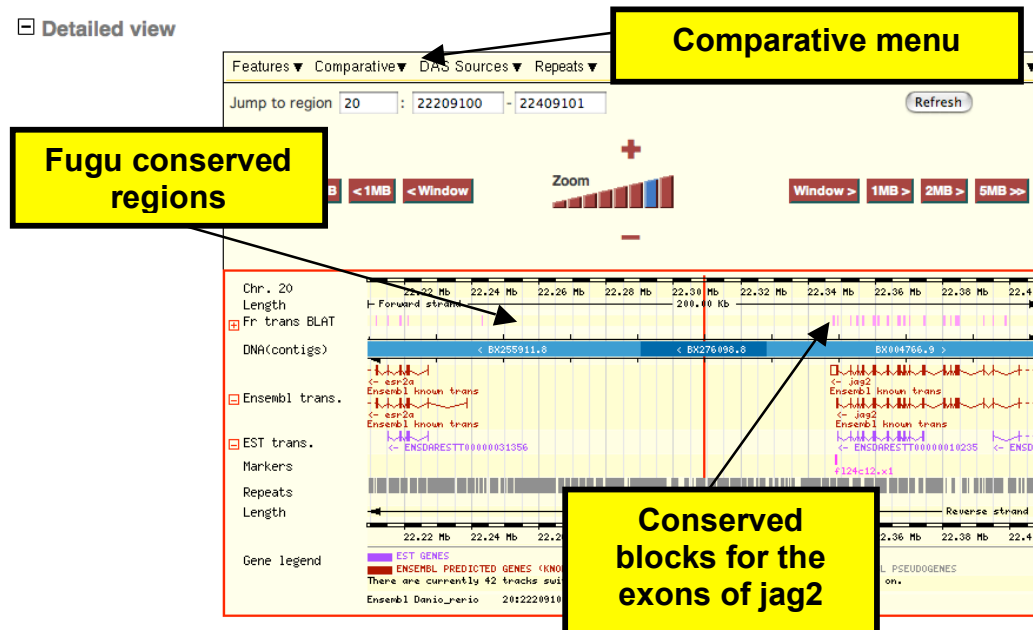
- whole genome alignments
- gene orthology/paralogy prediction
- protein clustering
- syntenic regions (not available for zebrafish)

In the previous section 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

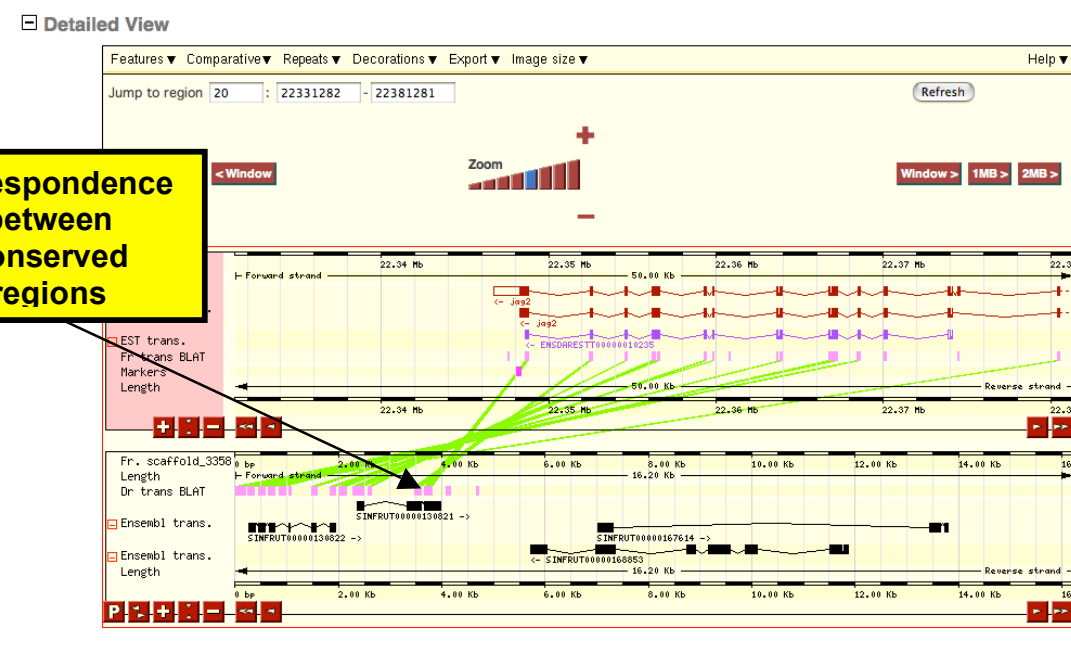
The alignment of the whole DNA sequence from two organisms is computationally demanding. Such data are of great interest both in studies of the mechanisms of molecular evolution and in attempts to identify conserved functional sequences such as novel genes and regulatory regions. Whole genome alignments become more 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 section 2) and select from the 'Comparative' menu the Fugu translated BLAT track.



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.



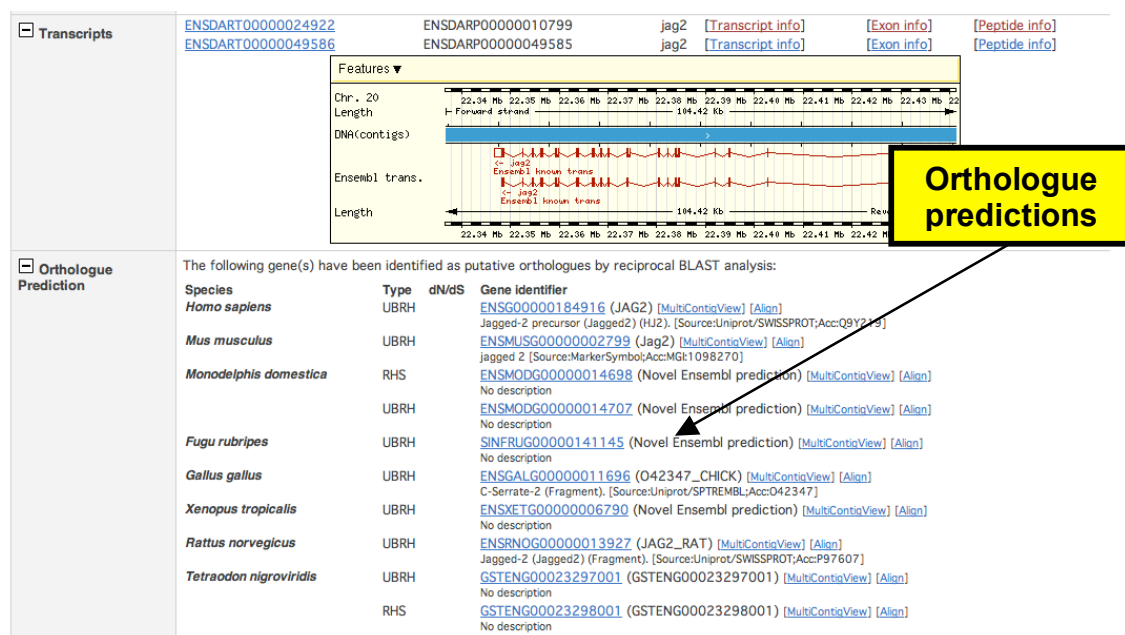
A **MultiContigView** allows the visualisation of syntenic regions from multiple species. In the example above a region from chromosome 20 in zebrafish is compared to scaffold_3358 in the Fugu assembly. The conserved blocks are connected with green lines. Observe that in this example the exons from jag2 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. This is particularly relevant for zebrafish, it is now widely accepted that there was an extra whole genome duplication in the teleost lineage posterior to the split with tetrapods. A duplicated orthologues prediction might be due to an assembly mistake rather than an actual duplicate.

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.



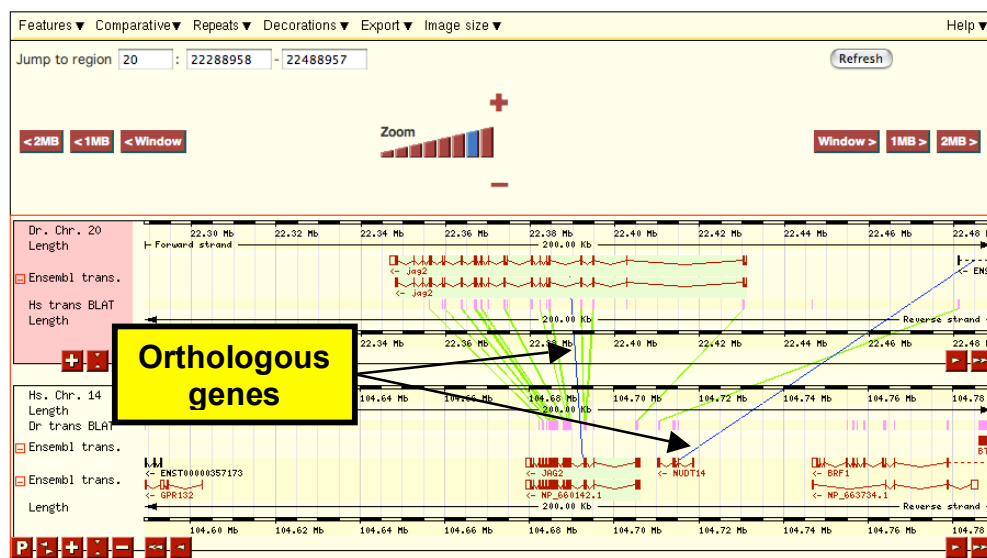
The screenshot displays the Ensembl GeneView page for the *jag2* gene. The top section shows transcript information for *ENSDART00000024922* and *ENSDART00000049586*. Below this, a genomic track shows the gene structure on chromosome 20, including exons and introns. A yellow box highlights the 'Orthologue predictions' section, which lists putative orthologues identified by reciprocal BLAST analysis.

Species	Type	dN/dS	Gene identifier
<i>Homo sapiens</i>	UBRH		ENSG00000184916 (JAG2) [MultiContigView] [Align] Jagged-2 precursor (Jagged2) (HJ2). [Source:Uniprot/SWISSPROT;Acc:Q9Y213]
<i>Mus musculus</i>	UBRH		ENSMUSG00000002799 (Jag2) [MultiContigView] [Align] jagged 2 [Source:MarkerSymbol;Acc:MG1098270]
<i>Monodelphis domestica</i>	RHS		ENSMODG00000014698 (Novel Ensembl prediction) [MultiContigView] [Align] No description
	UBRH		ENSMODG00000014707 (Novel Ensembl prediction) [MultiContigView] [Align] No description
<i>Fugu rubripes</i>	UBRH		SINFRUG000000141145 (Novel Ensembl prediction) [MultiContigView] [Align] No description
<i>Gallus gallus</i>	UBRH		ENSGALG00000011696 (O42347_CHICK) [MultiContigView] [Align] C-Serrate-2 (Fragment). [Source:Uniprot/SPTREMBL;Acc:O42347]
<i>Xenopus tropicalis</i>	UBRH		ENSXETG00000006790 (Novel Ensembl prediction) [MultiContigView] [Align] No description
<i>Rattus norvegicus</i>	UBRH		ENSRNOG00000013927 (JAG2_RAT) [MultiContigView] [Align] Jagged-2 (Jagged2) (Fragment). [Source:Uniprot/SWISSPROT;Acc:P97607]
<i>Tetraodon nigroviridis</i>	UBRH		GSTENG00023297001 (GSTENG00023297001) [MultiContigView] [Align] No description
	RHS		GSTENG00023298001 (GSTENG00023298001) [MultiContigView] [Align] No description

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.

☒ Detailed View



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:

Ensembl v36 - Dec 2005

e!Ensembl Zebrafish FamilyView

Ensembl Family ENSF00000000048

Family ID	ENSF00000000048
Consensus annotation	PRECURSOR
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.
Multiple alignments	<p>Click to view multiple alignments of the 695 Ensembl members of this family. JalView</p> <p>Click to view multiple alignments of the 1077 members of this family. JalView</p>
Ensembl genes containing peptides in family ENSF00000000048	

Location of Ensembl genes containing family ENSF00000000048

Gene ID	Gene Name	Genome Location	Description(if known)
ENSDARG00000010791	dla	Chromosome 1:	deltaA [Source:RefSeq; peptide:Acc:NP_571029]

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

BioMart (next section) 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.