

Zebrafish Genome Resources Workshop

June 2006, Madison

Contents

- 1 – The Zebrafish Genome Project
- 2 – The Ensembl Genome Browser
- 3 – The Vega Genome Browser
- 4 – How do I find a zebrafish gene?
- 5 – Does my gene have a known orthologue?
- 6 – Data Mining using BioMart

1 - The Zebrafish Genome Project

Aims

- Introduce zebrafish genome project web pages
- Give examples of the services provided
- Show ways to navigate around these pages

Introduction

In spring 2001 the Wellcome Trust Sanger Institute started sequencing the genome of the zebrafish (*Danio rerio*). The strategy being used is clone mapping and sequencing from BAC and PAC libraries complemented with a whole genome shotgun (WGS) assembly. Some of these clones are selected for sequencing based on their location in the tiling path of the physical map. The released zebrafish assembly is based on the integration of the available finished clones with the WGS assembly contigs. The assembly is automatically annotated using the Ensembl pipeline and can be browsed on the Ensembl site. Assemblies are released once or twice a year depending on the available data. The current assembly is Zv6, which was released on March 31st, 2006. The assembly will eventually consist solely of finished clones, with no sequence from the WGS assembly.

Sequences from finished clones come through the sequencing pipeline on a daily basis. They currently cover around 80% (March 2006) of the estimated 1.6 Gb size of the zebrafish genome. In a collaboration with ZFIN, finished clones are manually annotated. The finished clones with manual annotation can be browsed in the Vega database. These data are updated regularly to reflect the changes in the physical map and to make public the annotation. In sections 2 and 3 the structure of the data in Ensembl and Vega is discussed in more detail.

The *Danio rerio* Sequencing Project Page

The main gateway to all the information regarding the zebrafish genome project is:

http://www.sanger.ac.uk/Projects/D_rerio

On the left-hand side of this page there is a quick-access toolbar with links to the services offered, and on the right-hand side there is a report with the jrecent news related to the project. The page is divided in five parts:

- FAQs and contact information
- clone mapping and sequencing
- assembly releases
- other services
- contacts and links

The screenshot shows the Sanger Institute website for the *Danio rerio* Sequencing Project. Several callout boxes highlight specific features:

- Quick-access toolbar**: Points to the top navigation bar containing links for Information, Projects, and Other Services.
- FAQs and contact**: Points to the 'Frequently asked questions' and 'Contact us' links in the left sidebar.
- news**: Points to the 'Vega release' news item dated 8th May 2006.
- clone mapping, sequencing and manual annotation**: Points to the 'Clone mapping and sequencing' section, which includes links to mapping and clone-related pages.
- assembly releases**: Points to the 'Assembly releases' section, which includes links to WGS and assembly-related pages.

The main content area displays the 'Current Sequencing Status' table:

Date	Unfinished	Finished	Total
09-May-2006	654,705,316	1,081,452,147	1,736,157,463

Other sections visible include 'Assembly releases', 'ZF-MODELS project at Sanger', 'Other services', and 'Contacts and links'.

Contacting us

The email address for any enquiry regarding the project is:

zfish-help@sanger.ac.uk

There is also a link to FAQs page where a wide range of questions regarding the project are already answered.

Clone mapping and sequencing

This page lists all the relevant links to the zebrafish clones, from their mapping to the sequence. There are links to the Vega database, the FPC database and to a Blast server for searching all the available sequences from the project.

Assembly releases

This page has information about the current and previous assemblies with links to FTP sites from which the sequences can be downloaded. There is also a link to the trace repository. This is a database that features traces from several projects including all the zebrafish reads used in the whole genome shotgun assembly. These databases can be searched for alignments using SSAHA .

Other services

This section has links to an online RepeatMasker server and tutorials used in several courses and workshops . The repeat analysis is based on the same Repbase database used in the Vega/Ensembl analysis.

The output of this service returns the original sequence where repeats are masked by strings of Ns.

The screenshot shows the Zebrafish RepeatMasker Server web interface. A yellow box labeled "paste sequence or..." points to a large text area under the "SEQUENCE DATA" header. Another yellow box labeled "enter filename" points to a "Browse..." button. The interface includes a left sidebar with navigation links, a main content area with instructions and a "Retrieve BLAST result" section, and a footer with contact information and a "Last Modified" timestamp.

paste sequence or...

enter filename

SEQUENCE DATA

This RepeatMasker service screens DNA sequences in fasta format against a library of repetitive elements identified for zebrafish and returns a masked query sequence. The masking libraries consists of

- all the repeats submitted to repbase
- repeats Dr000001-Dr000409 identified by [Zhirong Bao](#) using his [Recon repeat identification software](#) (Bao and Eddy, submitted). This identifies the majority of repeats present more than 400 times in zebrafish genome.
- repeats dr1410-drr1225 identified by [Rick Waterman](#) as individual examples of repeats (through BLAST analysis of a 2% sample of zebrafish genome against the entire zebrafish WGS database (~4,000,000 traces, data from spring 2002), and includes most repeats present greater than 50 times in the genome. The repeat set is [downloadable](#).

For enquiries please mail zfish-help@sanger.ac.uk.

Select which repeat library to use when masking.
Rebase repeat library (default)

OR upload a sequence file you wish to mask
OR enter a zebrafish clone accession you wish to mask
You may also specify a pipe ("|") separated list of accession's

RESULTS

Mask Reset

Reference for RepeatMasker: A.F.A. Smit & P. Green, unpublished data. RepeatMasker is written and supported by [Arian Smit](#). Queries regarding this RepeatMasker webserver should be sent to webmaster@sanger.ac.uk.

Welcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK Tel: +44 (0)1223 834244 webmaster@sanger.ac.uk
Registered charity number 210183 [Data Release Policy](#) [Conditions of Use](#) [Copyright](#)
Last Modified Tuesday, 19-Oct-2004 15:03:58 BST

2 – The Ensembl Genome Browser

Caveat: At the time of writing this tutorial, Zv6 had not been released with a full gene build yet. All following examples are therefore taken from the Zv5 Ensembl. If you're trying to work through the examples yourself, please be aware of the difference in the scaffold naming ('Zv5_...' versus 'Zv6_...').

Aims

- Explain the source for the data in Ensembl
- Introduce the Ensembl browser
- Show the different Ensembl views with examples

Introduction

Ensembl is a joint project of the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute, funded mainly by the Wellcome Trust, with additional funding from EMBL and NIH-NIAID. Ensembl provides easy access to genomic information with a number of visualisation tools.

The Ensembl site provides automatic baseline annotation of the latest assembly sequence, including gene, transcript and protein predictions. The annotation is integrated with external data sources, such as ZFIN for the zebrafish site. The latest zebrafish assembly is Zv6, which was released on March 31st, 2006.

The key Ensembl web pages are called Views (e.g. GeneView, TextView, MapView, and ContigView). The Ensembl web site gives you the opportunity to directly download data, whether it is a DNA sequence of a genomic contig you are trying to identify novel genes in, or positions of SNPs in a gene you are working on. There is also an FTP site which you can use to download large amounts of data from the Ensembl database, as well as a data mining tool (BioMart, see section 6) which allows flexible and rapid retrieval of information from the databases. There are many ways you can access the data in Ensembl depending on your needs and these are explained here and in other sections.

The Ensembl site is at:

<http://www.ensembl.org>

On this page you will find links to all Ensembl species, documentation, search facilities, downloads and other related links. All Ensembl pages have a tool bar on the left-hand side with quick-access links to several resources and facilities.

The screenshot shows the main Ensembl website interface. A yellow box labeled "quick-access menu" points to the left-hand navigation sidebar. Another yellow box labeled "zebrafish" points to the "zebrafish" link in the "Other chordates" section of the "browse a genome" area. The sidebar includes sections for "Use Ensembl to...", "Docs and downloads", and "Other links". The main content area displays "browse a genome" with categories: Mammals, Other chordates, and Other eukaryotes. The "zebrafish" link is highlighted under "Other chordates".

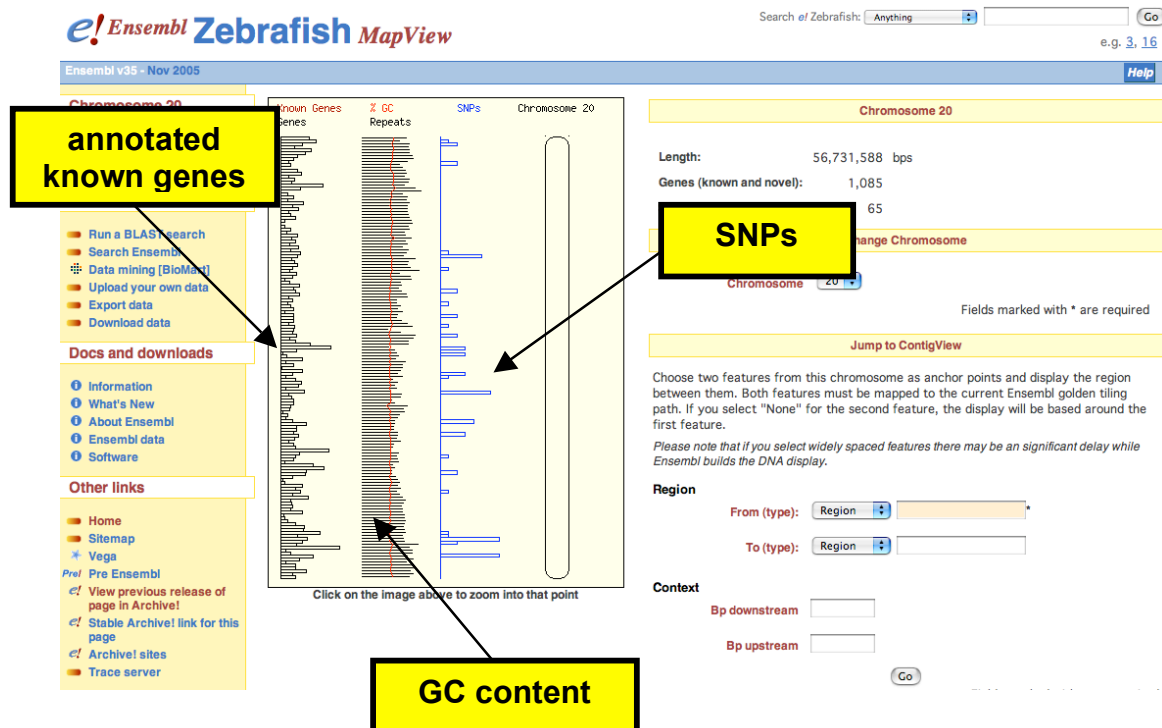
From the main Ensembl site you can access the zebrafish site by clicking on the appropriate species button. As soon a new assembly is released the sequence is made available as a pre-Ensembl site. This includes valuable information such as EST and UniProt alignments and *ab initio* predictions. The main missing data are the Ensembl genes and Ensembl ESTgenes. A full Ensembl dataset for a new assembly is typically made public a couple of months after the assembly release date.

The screenshot shows the Zebrafish Ensembl page. A yellow box labeled "zebrafish chromosomes" points to the "Karyotype" section. The page includes a sidebar with "Use Ensembl to...", "Docs and downloads", and "Select a species". The main content area is titled "Explore the Zebrafish genome" and includes "What's New in Ensembl 35" and a "Karyotype" section. The "Karyotype" section shows a diagram of 25 chromosomes and a search interface for "Chromosome: [] or region []" with "From (bp): []" and "To (bp): []" fields.

MapView and ContigView

This zebrafish Ensembl page provides various access points to the assembly sequence. For example you can browse a particular chromosome. The

chromosomes are linked to the **MapView** pages. The figure below shows the MapView for chromosome 20.



A MapView page plots the gene and SNP density and GC content. From this page you can zoom in to a more detailed display called ContigView by clicking on the schematic figure representing the chromosome.

ContigView can be considered the central view of the Ensembl web site. It shows the fragments (contigs, clones, etc) that make up a genome assembly. It allows you to scroll along entire chromosomes, whilst viewing the annotated features within a selected region in detail.

A ContigView page is divided into four panels: a chromosome overview, a zoomed-in **overview** of the region in the chromosome you are browsing, a **detailed view** showing features and a **basepair view** that goes down to individual bases. In order to continue with this module, jump to the region under the accession BX004766 (in chromosome 20) with start coordinate 1 and end coordinate 200000. (Use the text box provided to enter these coordinates.)

Overview

Features Menu

Detailed view

EST genes

Ensembl genes

Basepair view

The screenshot displays the Ensembl Zebrafish ContigView interface for Chromosome 20. The top navigation bar includes links for Zebrafish, What's New, TextSearch, BlastSearch, MartSearch, Export Data, Download, Archive sites, and Help. A search bar contains the text "[e.g. Zv4_NA18430, Zv4_NA3486]". The main content area is divided into several sections: Overview, Detailed view, and Basepair view. The Overview section shows a genomic map with tracks for DNA(contigs), Markers, Ensembl Genes, and Gene Legend. The Detailed view section shows a zoomed-in view of a specific region (25,272,803 to 25,357,225) with tracks for Length, RefSeq cDNAs, Ensembl mRNAs, gene, Proteins, GenScan, EST trans., Ensembl trans., DNA(contigs), Ensembl mRNAs, Markers, Length, and Gene Legend. The Basepair view section shows a detailed view of the genomic region (25,314,970 to 25,315,110) with tracks for Length, GenScan, Ensembl trans., Amino acids, Restr. Enzymes, and Gene Legend. The interface includes various interactive elements such as zoom controls, window size adjustments, and a features menu.

Date : Wed Jul 6 19:57:49 2005

Archive / Permanent page link

Help Desk / Suggestions

The Features menu in the detailed view controls the tracks you can visualise in the panel. Tracks can be turned on and off and the features can be collapsed to simplify the view. Spend some time on this page trying the different menus and studying the displayed features. Observe that there are two tracks for predicted genes: Ensembl transcripts and EST transcripts. (If these features are not visible verify that the corresponding tracks are selected in the menu.)

GeneView, TransView, ExonView and ProtView

Another important view in Ensembl I are the **GeneView** pages with information about the Ensembl predicted genes. In the ContigView page above there is a predicted transcript on the forward strand called **jag2**. Clicking on this transcript displays a pop-up window with several options. Follow the link labelled Ensembl Gene: ENSDARG00000021389. Below we only show the top of the GeneView page for jag2; scroll down to view all the information available.

GeneView provides annotation and supporting evidence for the selected gene. The annotation consists of transcripts, homologues to other species, known and predicted proteins and domains, and links to external documentation. In this example, jag2 is a gene known to ZFIN and so a link to the corresponding external page is provided. The annotation for jag2 is based on 2 transcripts. In the Transcripts sections there are links to the corresponding TransView pages. Click on the link labelled "Transcript info" for the first one with identifier ENSDART00000024922.

Ensembl Gene Report for ENSDARG00000021389

Gene	jag2 (ZFIN: ZFIN:ZDB-GENE-040606-101) click here																						
Ensembl Gene ID	ENSDARG00000021389																						
Genomic Location	This gene can be found on Chromosome 20 at location 22,346,74 . This start of this gene is located in Chunk BX004766.9.2000-212782 .																						
	jagged 2 isoform 1 Source:RefSeq_peptide NP_571937																						
	Genes were annotated by the Ensembl automatic analysis pipeline using GeneWise models from a protein alignment (with priority given to zebrafish proteins). GeneWise models are further combined with available aligned cDNAs and EST clusters to annotate UTRs.																						
Transcripts	ENSDART00000024922 ENSDART00000049586	ENSDARP00000010799 ENSDARP00000049585	jag2 [Transcript info] [Exon info] [Peptide info] jag2 [Transcript info] [Exon info] [Peptide info]																				
	Features Chr. 20 Length DNA(contigs) Ensembl trans. Length 																						
Orthologue Prediction	The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis: <table border="1"> <thead> <tr> <th>Species</th> <th>Type</th> <th>dN/dS</th> <th>Gene identifier</th> </tr> </thead> <tbody> <tr> <td><i>Homo sapiens</i></td> <td>UBRH</td> <td></td> <td>ENSG00000184916 (JAG2) [MultiContigView] [Align]</td> </tr> <tr> <td></td> <td></td> <td></td> <td>Jagged-2 precursor (Jagged2) (HJ2), [Source:Uniprot/SWISSPROT;Acc:Q9Y219]</td> </tr> <tr> <td><i>Mus musculus</i></td> <td>UBRH</td> <td></td> <td>ENSMUSG00000002799 (Jag2) [MultiContigView] [Align]</td> </tr> <tr> <td></td> <td></td> <td></td> <td>jagged 2 [Source:MarkerSymbol;Acc:MGI1098270]</td> </tr> </tbody> </table>			Species	Type	dN/dS	Gene identifier	<i>Homo sapiens</i>	UBRH		ENSG00000184916 (JAG2) [MultiContigView] [Align]				Jagged-2 precursor (Jagged2) (HJ2), [Source:Uniprot/SWISSPROT;Acc:Q9Y219]	<i>Mus musculus</i>	UBRH		ENSMUSG00000002799 (Jag2) [MultiContigView] [Align]				jagged 2 [Source:MarkerSymbol;Acc:MGI1098270]
Species	Type	dN/dS	Gene identifier																				
<i>Homo sapiens</i>	UBRH		ENSG00000184916 (JAG2) [MultiContigView] [Align]																				
			Jagged-2 precursor (Jagged2) (HJ2), [Source:Uniprot/SWISSPROT;Acc:Q9Y219]																				
<i>Mus musculus</i>	UBRH		ENSMUSG00000002799 (Jag2) [MultiContigView] [Align]																				
			jagged 2 [Source:MarkerSymbol;Acc:MGI1098270]																				

Transcript info (points to ENSDART00000024922)

Exon info (points to [Exon info] link)

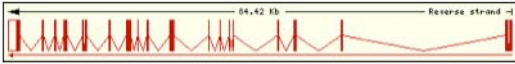
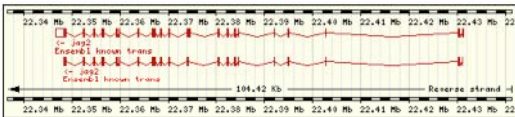
Peptide info (points to [Peptide info] link)

Transcript Structure (points to the Features section)

Ensembl v35 - Nov 2005

e! Ensembl Zebrafish TransView

Ensembl Transcript Report

Transcript	jag2 (ZFIN ID) (to view all Ensembl genes linked to the name click here)
Ensembl Transcript ID	ENSZDART0000024922
Transcript Information	Exons: 26 Transcript length: 5,436 bps Translation length: 1,254 residues This transcript is a product of gene: ENSDARG0000021389
Genomic Location	This transcript can be found on Chromosome 20 at location 22,346,747-22,431,169 . This start of this transcript is located in Chunk BX004766.9.2000-212782 .
Description	jagged 2 isoform 1 Source: RefSeq_peptide NP_571937
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using GeneWise models from a protein alignment (with priority given to zebrafish proteins). GeneWise models are further combined with available aligned cDNAs and EST clusters to annotate UTRs.
Similarity Matches	This Ensembl entry corresponds to the following database identifiers: RefSeq peptide: NP_571740.1 [Target %id: 99; Query %id: 99] [align] NP_571937.1 [Target %id: 99; Query %id: 99] [align] RefSeq DNA: NM_131665.1 [Target %id: 99; Query %id: 99] [align] NM_131862.1 [Target %id: 99; Query %id: 99] [align] Predicted UniProt/TrEMBL: Q5TZK8_BRARE [Target %id: 100; Query %id: 100] [align] Q90Y56_BRARE [Target %id: 99; Query %id: 99] [align] Q9YHU2_BRARE [Target %id: 99; Query %id: 99] [align] EntrezGene: 140422 EMBL: AF090432 [align] AF229449 [align] BX004766 [align] IPI: IPI00500671.1 [Target %id: 99; Query %id: 99] IPI00501275.2 [Target %id: 100; Query %id: 100] Protein ID: AAC98354.1 [align] AAL08214.1 [align] CAH69087.1 [align] UniGene: Dr.8287 [Target %id: 99; Query %id: 99] ZFIN ID: jag2 Affymx Microarray Zebrafish: Dr.8287.1.S1_a_at
GO	The following GO terms have been mapped to this entry via UniProt: GO:0001889 [liver development] GO:0005509 [calcium ion binding] GO:0007154 [cell communication] GO:0016020 [membrane]
InterPro	IPR001438 Type II EGF-like signature - [View other genes with this domain] IPR001881 EGF-like calcium-binding - [View other genes with this domain] IPR001687 ATP/GTP-binding site motif A (P-loop) - [View other genes with this domain] IPR001774 Delta/Serrate/lag-2 (DSL) protein - [View other genes with this domain] IPR000742 EGF-like, subtype 2 - [View other genes with this domain] IPR001093 IMP dehydrogenase/GMP reductase - [View other genes with this domain] IPR000152 Aspartic acid and asparagine hydroxylation site - [View other genes with this domain] IPR006209 EGF-like - [View other genes with this domain] IPR011651 Notch ligand, N-terminal - [View other genes with this domain]
Protein Family	ENSE0000000046 : PRECURSOR This cluster contains 29 Ensembl gene member(s)
Transcript structure	
Transcript neighbourhood	
Transcript sequence	<pre> GTGATCAGACCGAGGAGATCAGACACACCATCACCGCAACACACACGCTCGT GAATTTTGCATGTCAGGAACGAGGATCTGTGCGCGTCATCGGGCTTTTCATCTT CGCTTATAACACATCAATCGCGCATGTGGAAATGTATCAGGATAGGAATGGCTC CAATCGCTGCTGCTGTAAACAGATGTGGACAAAGGTGCCAGGCTCTGCTGCTA GAGCTGCACGTGATTCGTAGAAAATGTAAAGGTGAGTTGTGGAGCGGGAATGT GACACGACCGCAACTCTCAAGACACGCTGCTGCGCGACGAGTGGCATCTACTTT AAAGTGTGCTGAAGGATACCACTGTGAAGTCACCACTGGACAGTGCACCTTCGCG TCCTGATCTACCGACGCTCTGTGGTGAATATAATTTCTTTAAGACCGCAAAACAGC CCACGCAAAAGAGGAGCTGGGAAAGATCATGCTCTTTCATCTGCTGCTGGCGGA TCCTCACACTCATCTTGAAGCTGGGACTGGGATAACTCCATCAGAACATGTGAA GAAAATTTGATCGAAGCGACATTACGCAAGCATGGTAACCCCGCGACCACTGGCAG TCCATCCGCGACCTGTGATCACGCGCCACATTGAATACCGCATCCGTGTCAGGTGTGAT GAGATTAATATGAGGATAGTGCACCAACATGTGCTGCCACAGAGATGACTACTGCT CATACCGATGCTGATCCATCTGGAATATGTGTCTGTTGATGCTGGATGGAGAGGAC TGTGGACAGCGATCTGCAAGCAGGGCTGTAATCTGATTACGGAGGCTGTGGGTGCT GGAGAAATGCAATGCAACTACGCTGCGAGGGCGAGTTCGCGACGAGTGTCTACCTTAT CTGGCTGTTTGGACGCTGCTGTGTATGCTGCTGGCAATGCTGCTGAGAGAACTGG GGCGGCTCTCTGGGATAAAGATCTGAACCTGCGGACGATCATCTCTGTGCTGAA GTGGAACTGATGAATCTGAACCGGATGAATATACTGTGCTGCTCCGAAGGCTAC </pre>

cDNA

TransView provides annotation and supporting evidence for the selected transcript (structure, transcribed proteins, Gene Ontology and InterPro associated entries). The Transcript report panel provides a top-level summary of the transcript, with links to its genomic location, alignments to sequences in external databases, and export options. Underneath the report, the cDNA sequence of the transcript can be shown with codons, peptide sequence and/or SNPs highlighted.

From the GeneView page there are also links to the ExonViews labelled as “Exon info”.

Ensembl Exon Report

Transcript: [jag2](#) (ZFID ID) (to view all Ensembl genes linked to the name [click here](#))

Ensembl Transcript ID: ENSDART0000024922

Transcript Information: Exons: 26 Transcript length: 5,436 bps Translation length: 1,254 residues
This transcript is a product of gene: [ENSDARG00000021389](#)

Genomic Location: This transcript can be found on Chromosome 20 at location [22,346,747-22,431,169](#).
This start of this transcript is located in [Chunk BX004766.9.2000-212782](#).

Description: jagged 2 isoform 1 [Source: RefSeq, peptide NP_571937](#)

Rendering options: Flanking sequence at either end of transcript: 50
Intron base pairs to show at splice sites: 25
☐ Show full intronic sequence
☐ Show exons only

[Go](#)

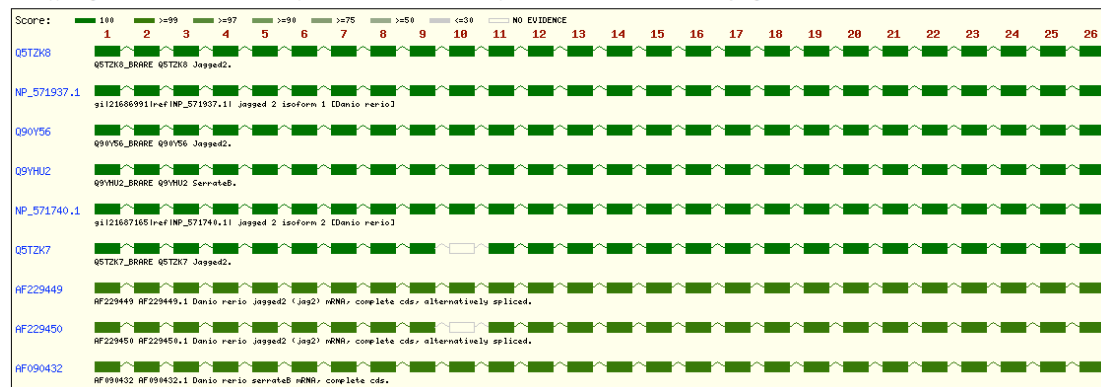
Exon Information

No. Exon / Intron	Chr	Strand	Start	End	Start Phase	End Phase	Length	Sequence
5' upstream sequence							gagatctctcctcagggtgtgtgtagccgtgaacacatcatcttgcggtgaa
1 ENSDARE00000194656	20	-1	22,430,954	22,431,169	-	0	216	GTGATCAGACGAGGGAGAGATCAGCAGACACCATCACCGCGCAACACACCACGCTCGT GAATTTTGCATGTGAGAACGGAGGATCCTGTGCGCGGTATCGGCCGTTTTCATCTT GCTTTATACACATCAATCGGCGGATGTGGAAATGTATCAGGATTAGGAATTGGCTC CCAATCGCGGCTGCTGTAAACGATGTGGAG
Intron 1-2	20	-1	22,430,957	22,430,953			457	gtgagagctctgattgtattatgtg.....attctgttttcaatgtatttttag
2 ENSDARE00000094731	20	-1	22,430,176	22,430,496	0	0	321	GTGTCCAGTCTCTGGCTATTTTGAAGCTGACGCTGCTGTGAGAAAATGTAAACGGT GATTTGTGGGACGGGAATGTGGGACAGCAGCGGAACTTCAAGACACGCGCTGGG CGGACGAGTGGGATCTACTTTAAAGTGTGTCTGAAGAGTACAGTCTGAAGTCACC ACCAGTGGACAGTGCACCTTCGGCTCTGGATCTACCGACCTTCTGGTGGAAATATAAT TCTTTAAGACCGCAAAACAGCCCAAGCAAAACAGCGAATGTGGAAAGATCATCATC CCTTTCACTTCGCTGGCGG
Intron 2-3	20	-1	22,402,782	22,430,175			27,394	gtgagctctctgtcttctccatggg.....gtttattctctctctctctctcttag
3 ENSDARE00000085927	20	-1	22,402,727	22,402,781	0	1	55	CGATCCTACACACATCATCTTGAAGCTTGGGACTGGGATAAATCCATCAGAACA
Intron 3-4	20	-1	22,395,053	22,402,726			7,674	gtgagttatctgtctgtgaacccaat.....atttattattctctctctctctcttag
4 ENSDARE00000194611	20	-1	22,394,795	22,395,052	1	1	258	ATGTGTGAAGAAATTTGATCGAACGGCACATTTCACGCAAGCATGTGAACGCGCGGACC ACTGGCAGTCCATCCGGCACCTTGGCATCACGCGCCACATTTAGTACCGCATCTGTCTCA GGTGTGATGAGAATTACTATGGGAGTAAGTCAACAAACAGTGTGCGCCACGAGTGTACT ACTTCGGTCATTACCGATCGCATCCATCTGGAATAATTTGTGTCTGTATGGCTCATGG GAGAGGACTGTGGGACAG
Intron 4-5	20	-1	22,392,067	22,394,794			2,728	gtgggtgagcttgaccattcttgggt.....ctctctctctctctctctctctctcttag
5 ENSDARE00000091811	20	-1	22,392,006	22,392,066	1	2	61	CGATCTCGAACGAGGCGTGAATCTGATTCCAGGAGCTGTGGCA
Intron 5-6	20	-1	22,384,650	22,392,005			7,356	gttaagtgtctgagcatcttctcatg.....atttaacaaag
6 ENSDARE00000137626	20	-1	22,384,519	22,384,649	2	1	131	GTGCAACTACGCTGGCAGGGGACGTTCTGCGACGAGTGTCTAC GCACGGTACTGTGTATGCGCTGGCAATGCACCTGTGAGAAGA CTGGGATAAG
Intron 6-7	20	-1	22,383,937	22,384,518			582	gtaaaggtgtctgaatgcagctgaca.....tctattgtctctctctctctctcttag
7 ENSDARE00000103635	20	-1	22,383,817	22,383,936	1	1	120	ATCTGAACACTCGGCGACGATCATCTCTGTGTCAATGGTGAACCTGCATGAACCTC AACCGGATGAATAAATCTGTGCTGTCCGAGGCTACTCTGCGAAGACTGTGAGATG
Intron 7-8	20	-1	22,382,379	22,383,816			1,438	gttaagtgtgtggaatgaagga.....aaacacatctctctctctctctcttag
8 ENSDARE00000147431	20	-1	22,382,265	22,382,378	1	1	114	CTGAACATGATGCTATCAACCCCTGTGCAACGAGGACGACGTGTGATGAAGTCCGA CCGGATTCGAGTCCCATGTGCCACAGGCTGGGAGGCTCCCACTTGGCGTAAG
Intron 8-9	20	-1	22,380,516	22,382,264			1,749	gtacgtgaaagttttgtgcaacttc.....acatgattttgtgtgtgtgtgttag
9 ENSDARE00000033336	20	-1	22,380,402	22,380,515	1	1	114	ACATGGATGAATGTGCTCCAGCCCGTGTGCGCAGGCGGGAACATGTATGACCTGGA ATGGCTTTGAGTGTGTCTGTCTCCGAGTGGGTGTGAAGAAGCTGTGAGATG
Intron 9-10	20	-1	22,374,574	22,380,401			5,828	gttaagtgtgagatccctgttattc.....ctctctctctctctctctctctctcttag
10 ENSDARE000000476336	20	-1	22,374,460	22,374,573	1	1	114	ATGCAAAATGAGTGTATGGGAAGCCTTGGCTAAATGCTCACTCTTGCAAAAACAGGATT GTGGATATCACTGTGACGCTTTCAAGGATGGGCGGACAGACATGTGACATCA
Intron 10-11	20	-1	22,374,152	22,374,459			308	gtcagttatctctcgaactctctc.....atttgtttgtctctctctctctcttag
11 ENSDARE00000087970	20	-1	22,374,105	22,374,151	1	0	47	ATCTCAATGGCTGCCATGGACAGTGGCAGAAATGGAGCTACTTGCAG
Intron 11-12	20	-1	22,374,022	22,374,104			83	gtatgtagacttttaaggtgtgta.....gttggaattttgtgtgtgtgttag
12 ENSDARE00000113330	20	-1	22,373,845	22,374,021	0	0	177	GACCTGCTTATGGAGGCTTACCACTGTGAGTGTGCTGCGCGGCTTGTGGGCTACACTGT GAAAGCTCAAGGAATAAATGTGCGCGGCTCAATGTGAGATGTGTGCGCTGTGCATGTCT ATTCTGGACAGCTTCCTGTGTGAGTGTCCGTCAACTACCGAGGAGTCTGTGTAG
Intron 12-13	20	-1	22,370,215	22,373,844			3,630	gtgagaagcattagtattatgta.....actcaactctctctctctctcttag
13 ENSDARE00000097512	20	-1	22,370,052	22,370,214	0	1	163	GTGAGAGCTGTCTTCAACCAACCCATGTGAGCGGAACCTTGTGAGTATACACTTTG TGCTACAGTCTGCGGGGTGACTTTTACTGCGCTGTCTGAAGACTATAGGCGAAGACC TGCAGAAACCGCAAGACCATGCAAGATGACCCCTTGGCAAG
Intron 13-14	20	-1	22,368,626	22,370,051			1,426	gttaagctatttccagtttgtatctc.....tctctgtgtgtgtgtgtgtgttag
14 ENSDARE00000043450	20	-1	22,368,461	22,368,625	1	1	165	TGATCGATAGCTGATCACTTGTGTGCGGAGTAAACAGTTCAGATGGGCGCTGAGACACA TTAACTCTAATGTTTGTGGCTCTACATGGCGCTGATCAGTACAGCCAGGTGGAATTTCA CCTGCACCTGTGAGCTTGGCTTACAGGAACCTACTGTACAGAGA
Intron 14-15	20	-1	22,367,253	22,368,460			1,208	gtgagttctactgtactgttttattt.....gttaattttgtgtgtgtgtgttag
15 ENSDARE00000008697	20	-1	22,367,139	22,367,252	1	1	114	ATGTAATGACTGTGTGAGCAATCCGTGTGCAAAATGGAGGACCTTATTGACGGGATCA GCTCTTTCCAGTGTCTGTGTCAGATGGCTGGGAAGGAGACCTTTCAGCATCA
Intron 15-16	20	-1	22,367,052	22,367,138			87	gtgagttgtctctctctctctctc.....taccctctctctctctctcttag
16 ENSDARE00000129209	20	-1	22,366,938	22,367,051	1	1	114	ATGTGAACGAGTGTGAGTGGAGCCCTGCAAAAATGGCGGCACTGTGTGATCTGTGTCA ATGACTTTTACGTGAATGTGCCAATGGCTGGGAAGGAAAGCACTGTGCATTAC
Intron 16-17	20	-1	22,366,838	22,366,937			100	gtcagttctgttttaattttattt.....tgtattttttgtgtgtgtgttag
17 ENSDARE00000110574	20	-1	22,366,724	22,366,837	1	1	114	GTGAAGTCAAGTGTGACTCTCCACATGCAGTAATGGAGGAACCTGTATATCAGGAG ATGCTTTCCGCTGTGCTGTCTCCAGATGGGAAGGAATACATGCAATACAG

ExonView provides annotation and supporting evidence for the exons of a selected transcript. Ensembl gene predictions are based on aligned evidence from external databases like UniProt and RefSeq. At the bottom of an ExonView page you can find the evidence linked to this prediction.

Support Evidence

The supporting evidence below consists of the sequence matches on which the exon predictions were based and are sorted by alignment score.



Finally from the links labelled “Peptide info” in the GeneView page we can visit the ProtView page for the associated translation.

Ensembl Zebrafish ProtView Ensembl v35 - Nov 2005

Ensembl Protein Report

Peptide	jag2 (ZFIN ID) (to view all Ensembl genes linked to the name click here)
Ensembl Peptide ID	ENS DARP00000010799
Translation Information	This protein is a translation of transcript ENS DART00000024922 , which is a product of gene ENS DARG00000021389 .
Genomic Location	This peptide can be found on Chromosome 20 at location 22,348,271-22,431,022 . This start of this peptide is located in Chunk BX004766.9.2000-212782 .
Description	jagged 2 isoform 1 Source: RefSeq_peptide NP_571937
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using GeneWise models from a protein alignment (with priority given to zebrafish proteins). GeneWise models are further combined with available aligned cDNAs and EST clusters to annotate UTRs.
InterPro	IPR001438 Type II EGF-like signature - [View other genes with this domain] IPR001881 EGF-like calcium-binding - [View other genes with this domain] IPR001687 ATP/GTP-binding site motif A (P-loop) - [View other genes with this domain] IPR001774 Delta/Serrate/jag-2 (DSL) protein - [View other genes with this domain] IPR000742 EGF-like, subtype 2 - [View other genes with this domain] IPR001093 IMP dehydrogenase/GMP reductase - [View other genes with this domain] IPR000152 Aspartic acid and asparagine hydroxylation site - [View other genes with this domain] IPR006209 EGF-like - [View other genes with this domain] IPR011651 Notch ligand, N-terminal - [View other genes with this domain]
Protein Family	ENSE00000000046 : PRECURSOR This cluster contains 29 Ensembl gene member(s)
Protein Features	
Protein Sequence	<pre> MWNCIRIRNWLPIACLLTMTKVSQSGYFELQLIAVENVNGELMDGECDCSTRNSQDQ RCVREDCDTYFKVCLKYSQSVTTTSGCTFSGSTDLVIGNIISFKTANSPTKSTQDVGK IIPFPFAPRPSYLLIAWQWQSTQNSGENIIRIIRASNVSPQWQWQIRPQSTTA HISYRIKRVCDENYYSKCKKQCRPRDPTFGHYRCDFPSONIVCLDGNMGECCTAICKQG CNLINGCCAVPGECCKNYGQWQGFCDCELPYPCGLSGTCVHPWQCTCERNMGLLQDKDL NYCOTSHPCVNGQTCMNSPDEYKACFEGYSKNCIEIRACVSNPCANGGTCHEVPTG FECRCPPWQEGTCAKMDCEASSPCAQGGCTCIDLNGFECVCPQWVGKTCQIDANECM GKPCVNAISCKRMIGYSCDCFGKAGQRCCLINGSGGCGMGATXELVGGYRCCCP </pre>

ProtView shows information about the structure and function of the encoded protein in the transcript's report with external links to various databases like Pfam, Prosite, etc...

ExportView

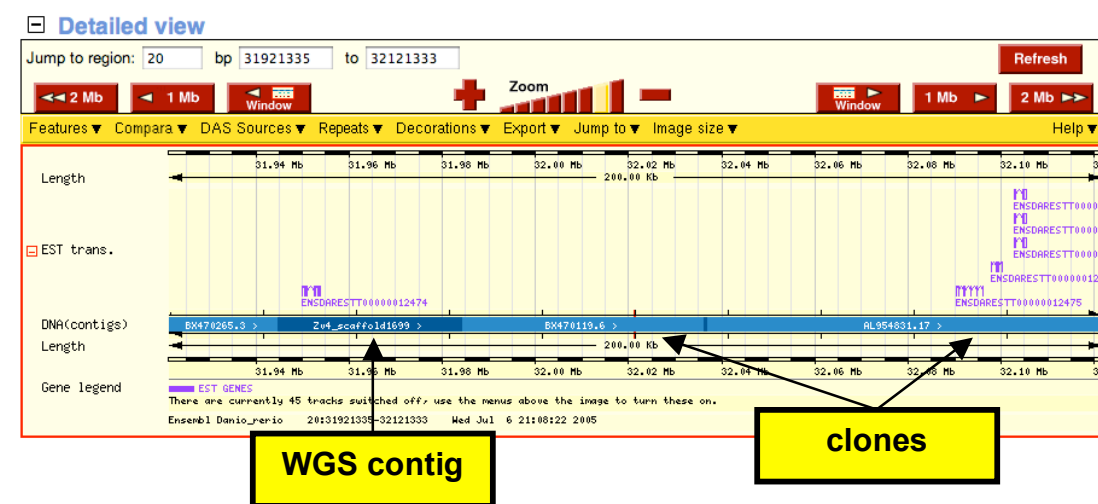
ExportView lets you download/dump data. All the features for a genomic region may be downloaded or exported to several formats (for example, FASTA, GenBank or EMBL-style flat file, as a feature list or an image). The ExportView pages are accessible from the link 'Export data' in the left-hand side menu from any of the pages above.

Zebrafish assembly in Ensembl

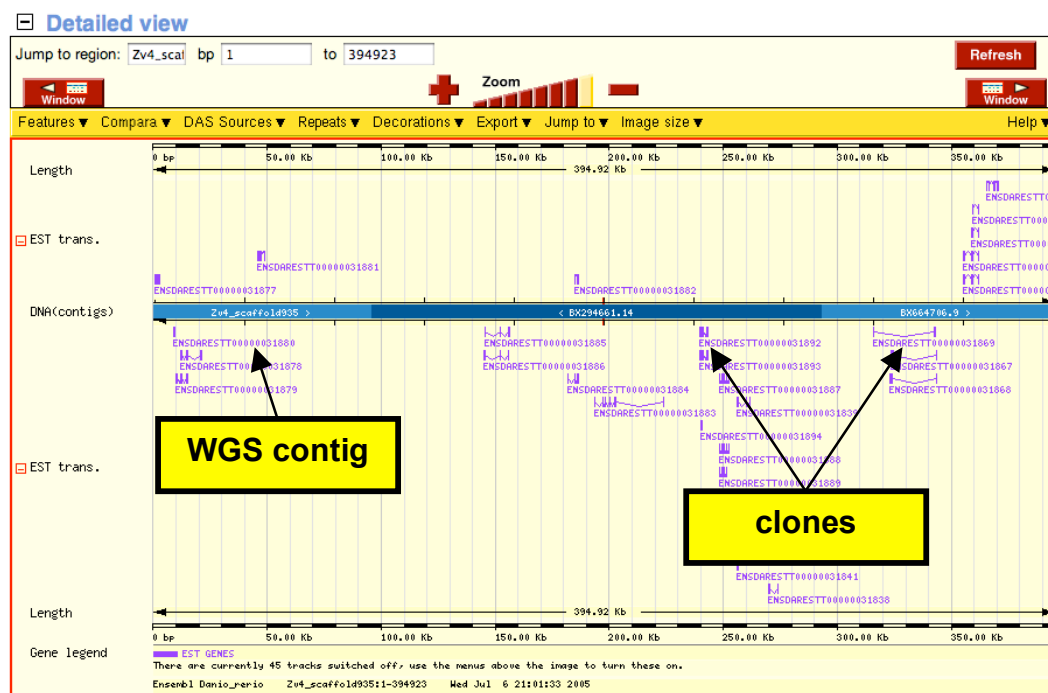
The sequence in the *Danio rerio* Ensembl database is the latest assembly release with automatic annotation. The genomic sequence released is based on all the sequenced clones with remaining gaps covered by contigs from a whole genome shotgun (WGS) assembly. The WGS fragments are placed in those gaps using a mixed strategy that looks at sequence similarity and other anchors as BAC-ends and markers. This placement is hard to perform without errors - mainly due to the presence of mis-joins in the WGS assembly and duplicates. It is even more difficult to place sequence where there is no sequenced clone or marker to use as an anchor.

In this context the user has to evaluate the data with a critical eye. In particular when the sequence of interest is known to the community but it is wrong in the assembly. There are three kinds of scaffolds and these are, in order of quality from best to worst:

1. scaffolds that have been attached to chromosomes (they may contain sequenced clones),
2. scaffolds that can be aligned to clones but the physical map cannot assign a chromosome yet (they may contain sequenced clones), and
3. NA (non-attached) scaffolds that corresponds to WGS contigs that could not be placed in the map (they don not contain sequenced clones).



Zv5_scaffold935 is an example of a region that is part of the map but, when the assembly was built, did not have a placement in a chromosome (category 2). This example shows that the region contains some sequenced clones as shown by the presence of their accession numbers.



Finally a scaffold from category 3 is *Zv5_NA10*. This region does not contain any finished clones.

Exercises

This section introduces the Ensembl browser and some of its basic views. In other section we will study more advanced features like the compara database and Blast/SSAHA search facilities. The user is encouraged to navigate the site and experiment with the different views discussed above.

1. Find the GeneView page for jag2 (Ensembl gene), and scroll down to the first 'Transcript/Translation Summary'. As jag2 has been identified in Zv6 you can use this gene name in a text search box.
2. Examine the genomic context. From GeneView, follow the link 'View gene in genomic location' to ContigView.
3. Customise the display of ContigView selecting different tracks and comparing the data from different tracks.
4. In ContigView zoom in to examine the data in more detail.
5. Export a file containing the cDNA of one of the predicted transcripts for jag2.
6. One of the Ensembl tracks displays probes for which ZFIN has a expression pattern page. Search for the mapping of the EST with accession CK685476 and open the corresponding ContigView page. Make sure that the 'expression pattern' track is selected in the 'features' menu. The ContigView page displays a link to the expression pattern page in ZFIN, try it.

3 – The Vega Genome Browser

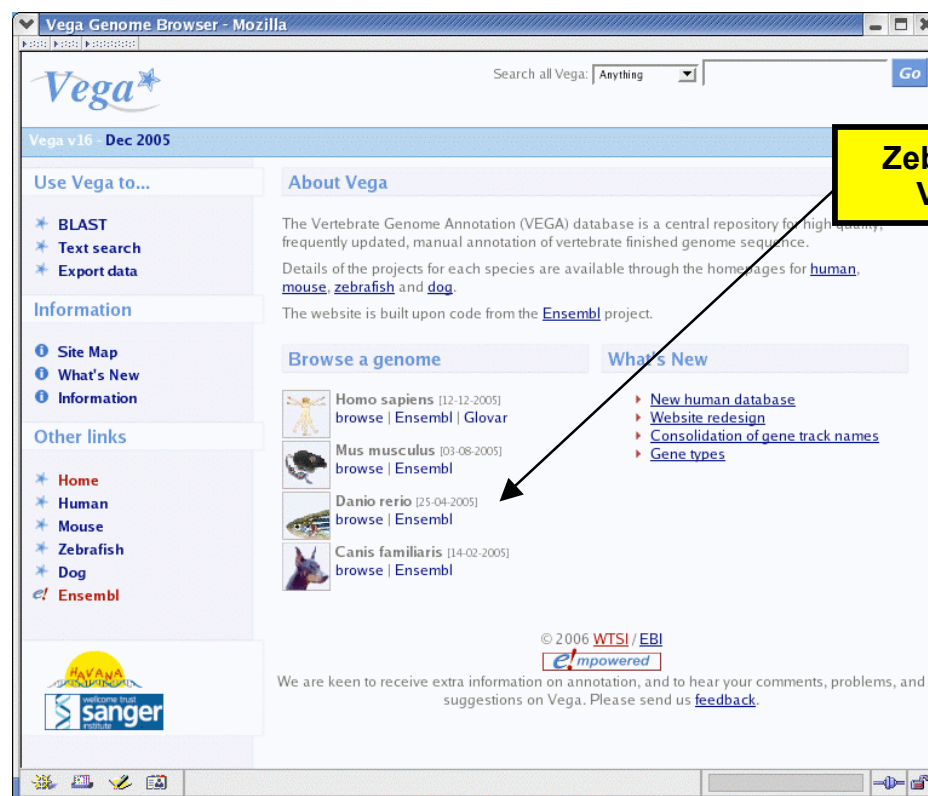
Aims

- Introduce the Vega genome browser
- Explain the source of the data in Vega
- Show the different Vega views stressing the differences to the Ensembl views

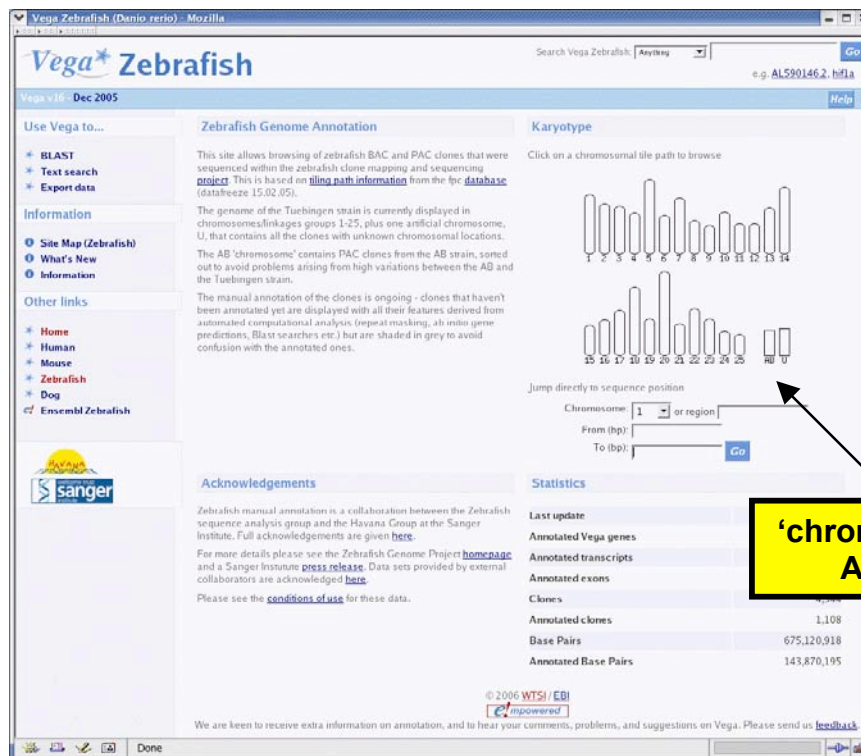
Introduction

The Vertebrate Genome Annotation (Vega) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence. The *Danio rerio* Vega database contains all the finished clones. Unlike the *Danio rerio* Ensembl database, the Vega database only contains high-quality sequence with high-quality manual annotation. The annotation is undertaken in collaboration and synchronisation with the central zebrafish database ZFIN. The implementation of the Vega browser is based on the Ensembl code and so they share many features and functionality. This section gives a brief introduction to the Vega views emphasising the differences with Ensembl. Refer to section 2 for more details on the Ensembl views.

The main Vega page is <http://vega.sanger.ac.uk>. One obvious difference between Ensembl and Vega is the background colour. In Ensembl it is yellow whereas in Vega is blue.



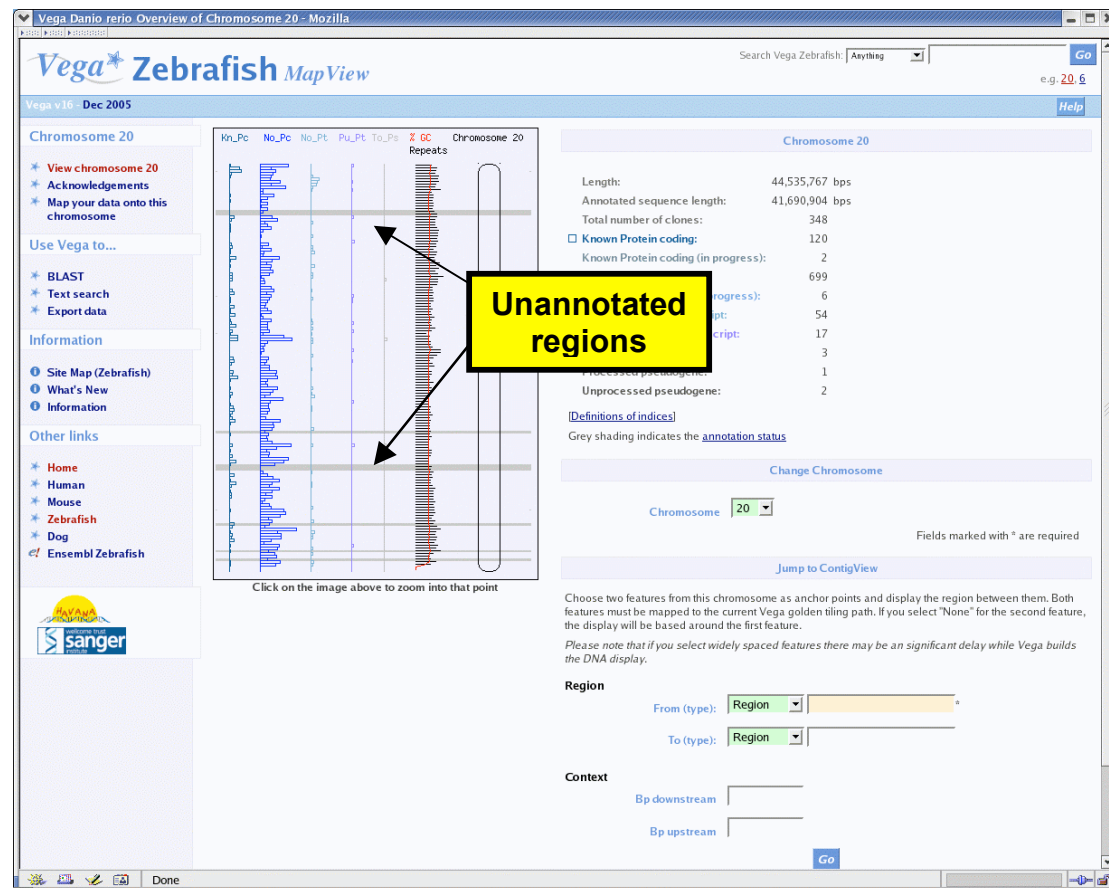
Follow the link to the zebrafish database.



This zebrafish page displays all the chromosomes plus two 'artificial' ones called AB and U. Chromosome AB groups all the sequenced clones for a PAC library made from the AB strain. Some of these clones have been manually annotated. Chromosome U contains finished clones which have not been placed in the physical map. The lengths do not represent an estimation of the size of the real chromosomes, but the amount of the current finished sequence. Chromosome 20 is top priority and that is reflected by the fact that it is the longest, ie the one with the most finished sequence.

MapView and ContigView

Clicking on a chromosome links to the corresponding MapView page.



The regions shaded in grey in the **MapView** pages indicate segments of the chromosome that have not been annotated yet. Check the difference between chromosomes 20 and 7 in terms of how much sequence has been annotated. As chromosome 20 is a priority one most of the sequence is not greyed, for chromosome 7 the situation is just the opposite.

The **ContigView** is, like in Ensembl, one of the main pages in Vega. The contents of the ContigView include some data specific to the manual annotation, for example there is a track for polyA signals. In order to facilitate the task of the annotators, the alignments of protein and ESTs is done more aggressively over finished clones than over the assembly. The most important track in the Vega ContigView page is the 'Zfish transcripts', the manually annotated transcripts. Observe that shaded regions do not contain this kind of transcript since they have not been annotated yet.

Jump to the ContigView for the region in chromosome 20 from 17925000 to 18135782.

Vega Zebrafish ContigView

Home Dog Human Mouse **Zebrafish** BLAST Export Data Search Feedback Help

Help on ContigView Find All [e.g. AL590146.2, BX842684]

Chromosome 20

Overview

DNA(contigs) 16.48 Mb 16.58 Mb 16.68 Mb 16.78 Mb

Markers 16.48 Mb 16.58 Mb 16.68 Mb 16.78 Mb

Zfish Genes 16.48 Mb 16.58 Mb 16.68 Mb 16.78 Mb

Gene legend 16.48 Mb 16.58 Mb 16.68 Mb 16.78 Mb

Poly-A features

Region: 20 bp 16509110 to 16535334

Features DAS Sources Repeats Decorations Export Jump to Image size Help

Length ☐ ENSEMBL clones ☐ DNA(contigs) ☐ PolyA site ☐ PolyA signal ☐ Zfish trans. ☐ GenScan ☐ Fgenesh ☐ UniProt ☐ EMBL mRNAs ☐ Markers ☐ CpG islands ☐ Length ☐ Gene legend

Manually annotated genes

Basepair view

Length ☐ ENSEMBL clones ☐ Amino acids ☐ Sequence ☐ DNA(contigs) ☐ Amino acids ☐ Zfish trans. ☐ GenScan ☐ Fgenesh ☐ Restr. Enzymes ☐ Length ☐ Gene legend

Help Desk / Suggestions

This region contains a transcript labelled jag2. Look for this name in the 'Overview' panel.

GeneView, TransView, ExonView and ProteinView

Follow the link to the GeneView page from the jag2 transcript jag2-001.

Vega Zebrafish GeneView The Wellcome Trust Sanger Institute

Home Dog Human Mouse Zebrafish [e.g. hif1a, OTTDARG0000004630]

Link to ZFIN

Type

Author

Transcript information links

Transcripts

Curated Locus Report

Curated Locus	jag2 (ZFIN ID) (to view all Vega genes linked to the name click here)
Locus ID	OTTDARG00000005397
Version	1
Date	Gene last modified on 15/06/2004 (Created on 15/06/2004)
Alternative Symbols	ZDB-GENE-011128-3
Type	Known [Definition]
Genomic Location	View gene in genomic location: 16509110 - 16593534 bp (16.5 Mb) on chromosome 20 This gene is located in sequence: chunk770
Description	jagged2
Author	This locus was annotated by zfish <zfish-help@sanger.ac.uk>
Database Matches	ZFIN: jag2
Sequence Markup	View genomic sequence for this gene with exons highlighted
Export Data	Export gene data in EMBL, GenBank or FASTA
Curated Transcripts	<p>1: DKEY-5P1.1-001 (OTTDART0000005844) [Transcript information] [Exon information & supporting evidence] [Protein information]</p> <p>2: DKEY-5P1.1-002 (OTTDART0000005845) [Transcript information] [Exon information & supporting evidence] [Protein information]</p>

Transcript/Translation Summary

DKEY-5P1.1-002	<p>Stable ID: OTTDART0000005845 Version: 1 Class: Coding</p> <p>Exons: 25 Transcript length: 5324 bp Translation length: 1216 residues</p> <p>[Transcript information] [Exon information & supporting evidence] [Protein information]</p>
InterPro	<p>IPR001438 Type II EGF-like signature - [View other Vega genes with this domain]</p> <p>IPR001881 EGF-like calcium-binding - [View other Vega genes with this domain]</p> <p>IPR001687 ATP/GTP-binding site motif A (P-loop) - [View other Vega genes with this domain]</p> <p>IPR001774 Delta/Serrate/lag-2 (DSL) protein - [View other Vega genes with this domain]</p> <p>IPR007042 EGF-like domain, subtype 2 - [View other Vega genes with this domain]</p> <p>IPR001093 IMP dehydrogenase/GMP reductase - [View other Vega genes with this domain]</p> <p>IPR001052 Aspartic acid and asparagine hydroxylation site - [View other Vega genes with this domain]</p> <p>IPR006209 EGF-like domain - [View other Vega genes with this domain]</p>
Transcript Structure	
Protein Features	<p>Prints</p> <p>Profile</p> <p>Prosite</p> <p>Pfam</p> <p>Transmembrane</p> <p>Signal peptide</p> <p>Low complexity</p> <p>Peptide</p> <p>Scale (aa)</p>

DKEY-5P1.1-001	<p>Stable ID: OTTDART0000005844 Version: 1 Class: Coding</p> <p>Exons: 26 Transcript length: 5438 bp Translation length: 1264 residues</p> <p>[Transcript information] [Exon information & supporting evidence] [Protein information]</p>
InterPro	<p>IPR001438 Type II EGF-like signature - [View other Vega genes with this domain]</p> <p>IPR001881 EGF-like calcium-binding - [View other Vega genes with this domain]</p> <p>IPR001687 ATP/GTP-binding site motif A (P-loop) - [View other Vega genes with this domain]</p> <p>IPR001774 Delta/Serrate/lag-2 (DSL) protein - [View other Vega genes with this domain]</p> <p>IPR007042 EGF-like domain, subtype 2 - [View other Vega genes with this domain]</p> <p>IPR001093 IMP dehydrogenase/GMP reductase - [View other Vega genes with this domain]</p> <p>IPR001052 Aspartic acid and asparagine hydroxylation site - [View other Vega genes with this domain]</p> <p>IPR006209 EGF-like domain - [View other Vega genes with this domain]</p>
Transcript Structure	
Protein Features	<p>Prints</p> <p>Profile</p> <p>Prosite</p> <p>Pfam</p> <p>Transmembrane</p> <p>Signal peptide</p> <p>Low complexity</p> <p>Peptide</p> <p>Scale (aa)</p>

Every annotated gene in Vega has a ZFIN gene entry. Follow the link in the example by clicking on the name of the gene (jag2). Another special feature of the Vega GeneView page is the fields for the authors of the annotation, and

the type of gene. The gene type gives an indication of the confidence of the annotation based on the available evidence, for example:

- a gene has type **known** if it was listed by ZFIN at the moment of the annotation (eventually every annotated gene will have an entry in ZFIN), and
- a gene has type **novel CDS** if its product was similar to, but not identical to, a known protein from zebrafish or another organism.

Transcripts are also classified in several categories as well.

The gene *jag2* has been annotated with two transcripts. Follow the link labelled 'Transcript information' for the transcript OTTDART0000005844 to open the **TransView** page.

Vega Transcript Report

Transcript	DKEY-SP1.1-001 (Vega transcript ID)
Vega Transcript ID	OTTDART0000005844
Version	1
Class	Coding [Definition]
Transcript Information	Exons: 26 Transcript length: 5438bp Translation length: 1254 residues This transcript is a product of gene: OTTDART0000005837 [Exon information & supporting evidence] [Protein information]
Genomic Location	View transcript in genomic location: 16509110 - 16593534 bp (16.5 Mb) on chromosome 20 This transcript is located in sequence chunk770
Description	jagged2
Author	This locus was annotated by zfish-zfish-help@sanger.ac.uk
InterPro	IPR001438 Type II EGF-like signature - [View other Vega genes with this domain] IPR001881 EGF-like calcium-binding - [View other Vega genes with this domain] IPR001687 ATP/GTP-binding site motif A (P-loop) - [View other Vega genes with this domain] IPR001774 Delta/Serratelag-2 (DSL) protein - [View other Vega genes with this domain] IPR000742 EGF-like domain, subtype 2 - [View other Vega genes with this domain] IPR001093 IMP dehydrogenase/GMP reductase - [View other Vega genes with this domain] IPR000152 Aspartic acid and asparagine hydroxylation site - [View other Vega genes with this domain] IPR008209 EGF-like domain - [View other Vega genes with this domain]
Export Data	Export transcript data in EMBL, GenBank or FASTA

Transcript cDNA Sequence
[No markup] [No numbers]

Transcript Structure
[Diagram showing exons and introns]

Transcript Neighbourhood
[Diagram showing the transcript's position relative to other transcripts]

The **ExonView** page for this transcript gives more information about the sequence of the exons and introns and the supporting evidence used in the annotation. Follow the link labelled 'Exon information & supporting evidence' to open the ExonView page.

Vega Zebrafish *ExonView* 

Home Dog Human Mouse Zebrafish BLAST Export Data Search Feedback Help

Help on ExonView Find [e.g. CH211-212G7.4-001, OTTDART0000004952]

Vega Exon Report

Transcript	DKEY-SP1.1-001 (Vega_transcript ID)
Vega Transcript ID	OTTDART0000005844
Version	1
Class	Coding [Definition]
Transcript Information	This transcript is a product of Ensembl gene OTTDARG0000005397 [Transcript Information] [Supporting Evidence] [Peptide Information]
Genomic Location	View transcript in genomic location: 16509110 - 16593534 bp (16.5 Mb) on chromosome 20 This transcript is located in sequence: chunk770
Description	jagged2

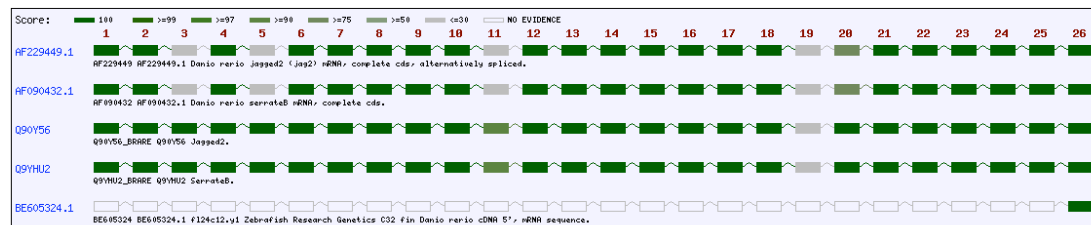
Exon Information

No.	Exon / Intron	Chr	Strand	Start	End	Length	Sequence
	5' upstream sequence					gagatctcctcaggcgtgtggtagccgtgaacatcatcattgccgtgaa
1	OTTDARE00000033254	20	-1	16593319	16593534	216 bp	GTGATCAGACCGAGGGAGAGATCAGCAGACACCATCACCGCGCAACACACCCACGCTCGT GAATTTTTCATGTCAGGAACGGAGATCCTGTCGCGGTCATCGGCCGTTTCCATCTT GCGTTTATACACATCAATCGCGGCATGTGGAAATTGTATCAGGATTAGGAATTGGCTC CCAATCGCGTGCCTGCTTTAAGCATGTGACGAAG gtgagagctctgattgttattagtg.....ttttgttttcaatgtatttttag
	Intron 1-2	20	-1	16592862	16593318	457 bp	
2	OTTDARE00000033260	20	-1	16592541	16592861	321 bp	GTGTCACAGTCTCTGGCTATTTGAGCTGCAGCTGATTGCTGTAGAAAATGTAAACGGT GAGTTGTGGGACGGGGAATGTTGCGACAGCAGCGGAACCTCAAGACACGCGCTGCGTG CGGACGAGTCCGATACCTACTTTAAAGTGTGCTGAAGGATACCACTCTGAAGTCACC ACCACTGGACAGTGCACCTTCGGCTCTGGATCTACGACGTTCTTGGTGGAAATATAAT TCTTTTAAGACCGCAAAAAACAGCCCAAGCAAAACGAGCGACGTGGGAAAGATCATCATC CCTTTTCACCTCGCCTGGCCG gtgagtcctcgtcttctccatgggc.....gttattctctctctctctctcttag
	Intron 2-3	20	-1	16565147	16592540	2794 bp	
3	OTTDARE00000033271	20	-1	16565092	16565146	55 bp	CGATCTACACACTCATCTTGAAGCTGGGACTGGGATACTCCACTCAGAACA

At the bottom of this page there is diagram showing the supporting evidence for this annotation.

Supporting evidence

The supporting evidence below consists of the sequence matches on which the exon predictions were based and are sorted by alignment score.



In this example there is no evidence for exon 19, indicating that the annotator has 'built' this exon from other evidences such as splice sites, codon bias and ORFs. Compare this situation and what you see in the Ensembl predictions.

The link labelled 'Peptide Information' opens the **ProteinView** page. This data is generated automatically using the predicted translation in very much the same fashion as done for the Ensembl annotation.

Other views that we discussed in the module for the Ensembl browser are also present in Vega, for example, **ExportView** to download the data in files.

Exercises

1. Open the GeneView page for jag2 and visit the associated links. In particular open the ContigView page showing this gene.

2. Study the differences between the manual annotation for jag2 in Vega and the automatic annotation in Ensembl (see the Ensembl section for an example of how to open the GeneView page for jag2 in Ensembl).
3. One of the differences between the automatic prediction and the manually annotated jag2 is the number of exons and the UTR. Why do you think these data are different?
4. Customise the ContigView page to turn on the track for poly-A signals.
5. Another special track in Vega ContigView is 'Assembly tags'. This features information on special regions of the clones. These data are entered by the person in charge of finishing the sequence. A region that contains one of these tags is the finished clone 'AL928990'. Open a ContigView page for this clone in Vega and check the text for the assembly tag.
6. Many clones present in Vega are also placed in the assembly and therefore can be browsed in Ensembl. As Vega is updated more often there might be a difference in the versions. A sequenced clone may go through several updates from its first to its final submission, these are recorded via the version numbers. Check for clone present in the assembly and compare it to one in Vega.

4 – How do I find a zebrafish gene?

Aims

- Introduce the different search facilities in Vega/Ensembl
- Discuss strategies for locating genes in the zebrafish genome
- Present other resources like the trace repository

Introduction

The genome sequence would not be of much use without annotation. The interfaces of the Vega/Ensembl browsers are designed to efficiently present users with relevant information, but the interpretation of much of the data is still in the user's domain. Searching for a region of interest can be a difficult task in its own right.

Every gene, transcript, exon and translation in Vega and Ensembl have an identifier, for example, ENSDARG00000021389 in Ensembl or OTTDARG00000005397 in Vega. These identifiers can be used as external references. In every new Ensembl release the set of identifiers from an old version are carried over wherever possible. In some cases, as the assembly is not finished yet, the identifiers cannot be mapped and they might vanish when moving to the latest release. Since October 2004 the old Ensembl releases are available from the Ensembl Archive site so old data can be checked and compared to the latest assembly. Links to the Ensembl Archive site can be found in the left-hand side menu bar in the Ensembl pages. In Vega the identifiers remain stable since genes are linked to finished clones.

TextView

The simplest way of searching for a gene (or indeed any term or accession) is using the text-based searches. The Vega/Ensembl pages have text boxes where the user can enter a keyword to perform a search over a collection of pre-indexed items. If you know the name of a gene or a keyword that might be present in its description then you can use it in this kind of search. Try searching with the name jag2. The search result is displayed in a TextView page.

Ensembl text search

Target: Danio rerio

Query: jag2

Search: All indexes for:

Display up to results in format

POWERED BY **alta** vista

1 matches in the Danio rerio Gene Index [first 5 matches shown]:

1. Ensembl Gene: [ENSDARG00000021389](#)
 Ensembl gene ENSDARG00000021389 has 2 transcripts: ENSDART00000024922, ENSDART00000049586 and associated peptides: ENSDABP00000010799, ENSDARP00000049585
 jagged 2 isoform 1 [Source:RefSeq_peptide;Acc:NP_571937]
 The gene has the following external identifiers mapped to it:
 Affymx Microarray Zebrafish: Dr.8287.1.S1_a_at
 EMBL: AF090432, BX004766, AF229449, AF229450
 EntrezGene: 140422
 GO: GO:0016020, GO:0001889, GO:0005509, GO:0007154
 IPI: IPI00500671, IPI00500671.1, IPI00501275, IPI00496898, IPI00496898.1, IPI00501275.2
 Predicted UniProt/TREMBL: Q90Y55_BRARE, Q90Y56_BRARE, Q5TZK8, Q90Y55, Q5TZK8_BRARE, Q90Y56, Q9YHU2_BRARE, Q9YHU2, Q5TZK7_BRARE, Q5TZK7
 Protein ID: AAL08214.1, CAH69087, AAL08215, C98354, AAL08214, AAC98354, CAH69087, C98354.1
 RefSeq DNA: NM_131862.1, NM_131865.1, NM_131862, NM_131865
 RefSeq peptide: NP_571937, NP_571937.1, NP_571740, NP_571740.1
 UniGene: Dr.8287
 ZFIN ID: [jag2](#), ZDB-GENE-011128-3
 Geneview: http://www.ensembl.org/Danio_rerio/geneview?gene=ENSDARG00000021389
 ContigView: http://www.ensembl.org/Danio_rerio/contigview?gene=ENSDARG00000021389

Stable identifier

Features

A **TextView** page summarises the result of a text-based search. If the query appears under different indices then the TextView page organises the results in categories. For example the page below corresponds to the result of searching for jag2 in Vega:

Target: all

Danio rerio results

1 matches in the *Danio rerio* Gene index [first 5 matches shown]:

1. **Vega Gene:** OTTDARG0000005397
 Vega gene OTTDARG0000005397 has 2 transcripts: OTTDART0000005845, OTTDART0000005844
 Description: jagged2
 The gene has the following external identifiers mapped to it:
 Vega_gene: jag2, ZDB-GENE-011128-3, OTTDARG0000005397
 ZFIN: jag2, ZDB-GENE-011128-3
http://vega.sanger.ac.uk/Danio_rerio/geneview?gene=OTTDARG0000005397&db=core

1 matches in the *Homo sapiens* Gene index [first 5 matches shown]:

1. **Vega Gene:** OTTHUMG00000029880
 Vega gene OTTHUMG00000029880 has 3 transcripts: OTTHUMT00000074540, OTTHUMT00000074542, OTTHUMT00000074541
 Description: jagged 2
 The gene has the following external identifiers mapped to it:
 HUGO: JAG2, K14_NN_1244, 6189
 MIM: K14_NN_1244, 602570
 RefSeq_dna: NM_145159, K14_NN_1244
 Uniprot/SWISSPROT: Q9Y219, K14_NN_1244
 Vega_gene: OTTHUMG00000029880, JAG2, K14_NN_1244
http://vega.sanger.ac.uk/Homo_sapiens/geneview?gene=OTTHUMG00000029880&db=core

1 matches in the *Homo sapiens* Peptide index [first 5 matches shown]:

1. **Vega Peptide:** OTTHUMP00000028452
 Vega peptide OTTHUMP00000028452 is a product of Vega gene OTTHUMG00000029880 [transcript OTTHUMT00000074540, JAG2-001]
http://vega.sanger.ac.uk/Homo_sapiens/protview?peptide=OTTHUMP00000028452&db=core

3 matches in the *Homo sapiens* Transcript index [first 5 matches shown]:

1. **Vega Transcript:** OTTHUMT00000074540
 Description: jagged 2
 This transcript has the following external identifiers mapped to it:
 Vega_transcript: OTTHUMT00000074540, JAG2-001
 Vega_translation: OTTHUMP00000028452
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000074540&db=core

2. **Vega Transcript:** OTTHUMT00000074541
 Description: jagged 2
 This transcript has the following external identifiers mapped to it:
 Vega_transcript: OTTHUMT00000074541, JAG2-002
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000074541&db=core

3. **Vega Transcript:** OTTHUMT00000074542
 Description: jagged 2
 This transcript has the following external identifiers mapped to it:
 Vega_transcript: JAG2-003, OTTHUMT00000074542
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000074542&db=core

empowered Help Desk / Suggestions

If a text-based search fails to return any meaningful output then we can instead use one of the available alignment algorithms. If a term does not return any output in a text-based search does not mean that associated feature is missing. It might be that due to a difference in the annotation (like a missing exon) or because the term is not present in ZFIN it was not feasible to link it to an annotated feature.

SSAHA and BLAST

The Vega/Ensembl browsers provide a page where you can search using different sequences to query the databases. You can access the BLASTView page from any of the Vega/Ensembl views through the link labelled 'run a BLAST search' in Ensembl or 'BLAST' in Vega.

e!Ensembl Zebrafish Search e! Zebrafish: e.g. [Zv5_NA11976](#), [ENSDARG0000031100](#)

Ensembl v35 - Nov 2005 [Help](#)

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Select a species

- Mammals
- Other chordates
- Other eukaryotes

Other links

- Home
- Sitemap
- Vega

Explore the Zebrafish genome

What's New in Ensembl 35

- Compara database**
The compara orthology build has been updated to include the new Opossum genome, and chimp dnafrag names have been changed to take account of the re-numbering of the Chimp chromosomes (see [Chimp news](#) for more information). [Read more...](#)
- Variation database**
All variation databases now include the new consequence_type 'REGULATORY_REGION'. [Read more...](#)
- Genes on Featureview**
Features which have accompanying Gene locations, e.g. some AffyProbes, now display a table of gene names and descriptions at the top of the view page. [See an example...](#)
- New species - Opossum**
Ensembl is pleased to announce the completion of the genebuild for the grey short-tailed opossum, *Monodelphis domestica*, based on assembly 2.0 from the Broad Institute. [Read more...](#)
- Re-numbering of Chimp chromosomes**
The Chimp chromosome numbers have been changed to the new primate standard proposed by E.H. [Read more...](#)

[More news...](#)

Click on a chromosome for a closer view

Jump directly to sequence position

Chromosome: or region

From (bp):

To (bp):

BLASTView

e!Ensembl Zebrafish ContigView Search e! Zebrafish: e.g. [BX005351](#), [BX072555](#)

Ensembl v37 - Feb 2006 [Help](#)

Chromosome 6
9,097,777 - 9,197,778

- View of Chromosome 6
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- View alongside ...

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Chromosome 6

Overview

Detailed view

Features ▾ Comparative ▾ DAS Sources ▾ Repeats ▾ Decorations ▾ Export ▾ Image size ▾ [Help](#) ▾

Jump to region : -

Zoom

Chr. 6

The **BLASTView** page is an interface to set up, run and visualise the output of a sequence-based search. It is designed to work in clear steps where the user can first enter the query and select the target for the search, configure the parameters for the algorithm and finally customise the format of the output.

Open a BLASTView page in the Vega browser. In order to specify the query for the search you have the option of either using the sequence(s) or using an EMBL/GenBank accession number.

The screenshot shows the BlastView web interface with several yellow callout boxes pointing to specific features:

- Paste the sequences or...**: Points to the 'Enter the Query Sequence' section, specifically the text input field for pasting sequences.
- enter a filename or...**: Points to the 'Browse...' button for uploading FASTA sequences.
- enter accession number**: Points to the 'Retrieve' button for entering a sequence ID or accession.
- choose method**: Points to the 'Select the Search Tool' section, highlighting the BLASTN, SSAHA, and TBLASTX options.
- choose target**: Points to the 'Select the databases to search against' section, highlighting the 'Genomic sequence' and 'Vega Peptides (all)' options.
- run!**: Points to the 'RUN' button in the 'Select the Search Tool' section.

Enter the accession number AF229449 and click on the button 'Retrieve'. Verify that *Danio rerio* is selected as the target database. You can choose the method to be used for the search. If your query is DNA you can choose from:

- BLASTN - the well-known BLAST algorithm performing a DNA-DNA search.
- SSAHA - this tool runs a hash-based algorithm. It is very fast since most of the needed data structures are pre-loaded. It works very well when searching for near-exact matches. It only performs searches where the query is DNA.
- TBLASTX - the BLAST algorithm where query and target are translated to the six possible reading frames. This is recommended when the query sequence is from a different organism.

Select the SSAHA algorithm and run the search by clicking on the 'Run' button. After some seconds you will be presented with the result. Every search is identified with a ticket that can be used later to retrieve a result (the results are stored for a couple of days). A Blast search can take a few minutes and your job may have to queue until it gets executed. The result of searching the *Danio rerio* Vega database with AF229449 is the following page:

Vega BlastSearch (BlastView)

http://vega.sanger.ac.uk/Multi/blastview/BLA_UQ6EaXLg

best hit

repeat

link to ContigView

Displaying AF229449 sequence alignments vs Danio_rerio LATESTGP database

Showing top 100 alignments of 100, sorted by Raw Score

☒ Alignment Locations vs. Karyotype (click arrow to hide)

☒ Alignment Locations vs. Query (click arrow to hide)

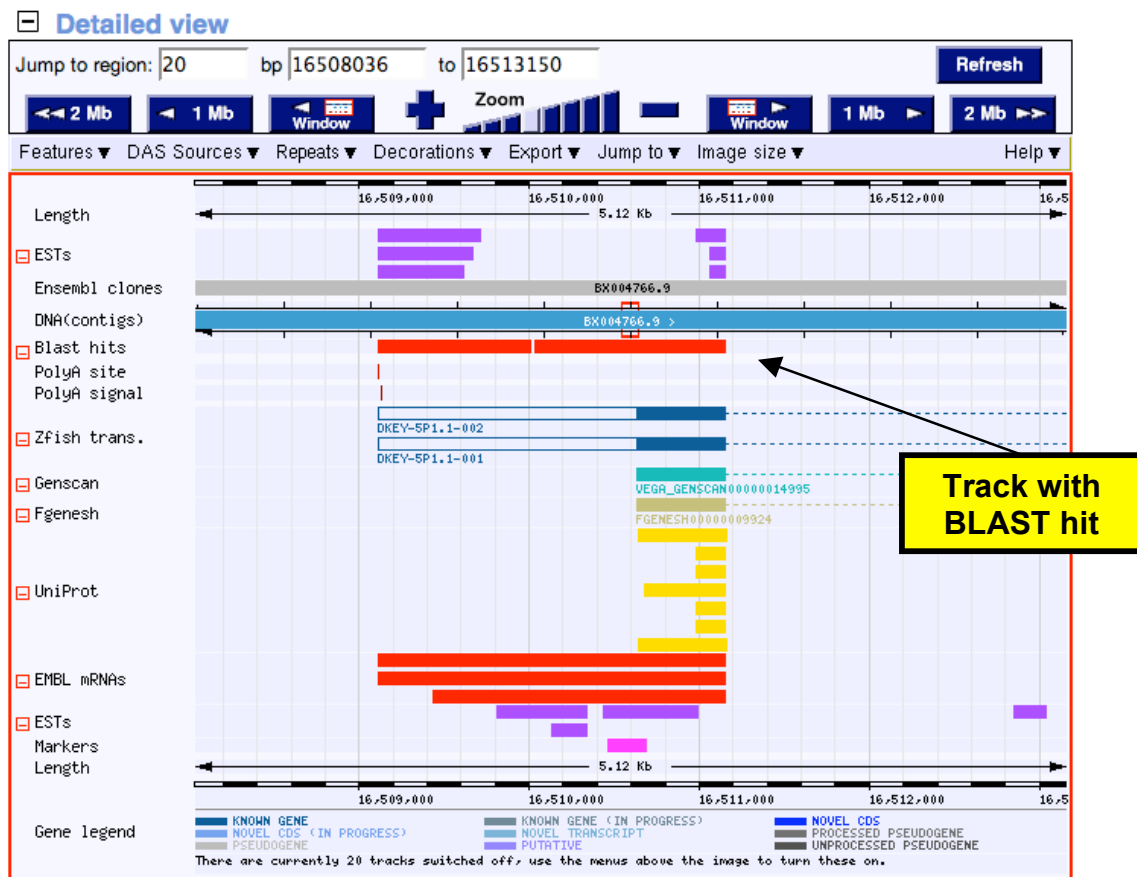
☒ Alignment Summary (click arrow to hide)

Select rows to include in table, and type of sort (Use the 'ctrl' key to select multiples)

Query	Subject	Chromosome	Clone	Chunk	Stats	Sort By			
off.	off.	off.	off.	off.	off.				
Name	Name	Name	Name	Name	Score	>Chunk			
Start	Start	Start	Start	Start	E-val	<Score			
[A] [S] [G] [C]	3397	4513	-	Chr:20	16510036	16511151	1091	93.46	1117
[A] [S] [G] [C]	4538	5437	-	Chr:20	16509112	16510011	900	100.00	900
[A] [S] [G] [C]	219	530	-	Chr:20	16592548	16592859	306	92.31	312
[A] [S] [G] [C]	599	850	-	Chr:20	16557160	16557411	249	95.24	252
[A] [S] [G] [C]	2869	3108	-	Chr:20	16518556	16518795	240	100.00	240
[A] [S] [G] [C]	4	219	-	Chr:20	16593316	16593531	216	100.00	216
[A] [S] [G] [C]	1551	1730	-	Chr:20	16536208	16536387	180	100.00	180
[A] [S] [G] [C]	1895	2050	-	Chr:20	16530832	16530987	156	100.00	156
[A] [S] [G] [C]	1737	1880	-	Chr:20	16532428	16532571	144	100.00	144
[A] [S] [G] [C]	3259	3390	-	Chr:20	16514836	16514967	132	100.00	132
[A] [S] [G] [C]	919	1038	-	Chr:20	16546888	16547007	120	100.00	120
[A] [S] [G] [C]	3113	3232	-	Chr:20	16516936	16517055	120	100.00	120
[A] [S] [G] [C]	2404	2511	-	Chr:20	16526188	16526295	108	100.00	108
[A] [S] [G] [C]	1167	1274	-	Chr:20	16544632	16544739	108	100.00	108
[A] [S] [G] [C]	1390	1497	-	Chr:20	16536832	16536939	108	100.00	108
[A] [S] [G] [C]	1045	1152	-	Chr:20	16546192	16546299	108	100.00	108

The result page can be customised and the relevant hits sorted in different ways. The most relevant match is framed in the diagrammatic view of the chromosomes. There is also a diagram indicating the coverage of the query,

which can be relevant to identify a repetitive subsequence in the query. At the bottom of the page there is a list of all hits. You can change the order of this list using the toolbar provided. If you are looking for a gene it is helpful to order the hits by, for example, chromosome coordinates. The matches can also be displayed in a ContigView page by selecting the [C] link.



This ContigView page adds a new track with the relevant hits. In this example the alignment coincides with an exon of a manually annotated gene (perhaps jag2!).

Important note

The Ensembl database contains the latest zebrafish assembly with automatic annotation (currently version Zv6 – March 2006). The zebrafish assembly is obtained by integrating all the available sequenced clones with a whole genome shotgun assembly. When reading and interpreting the outcome of a search it is important to understand the quality of the underlying sequence. In particular remember that the current assembly still includes sequences that do not have a chromosome assigned. If the best hit of your search matches one of these 'floating' fragments it will not appear in a framed box (since this only covers chromosomes 1 to 25). Refer to the exercises for an example. In section 2 the structure of the zebrafish assemblies is explained in more detail.

The Vega database contains all the finished clones featuring high-quality manual annotation. This database is updated more often than Ensembl in order to incorporate new annotation and reflect changes in the map. When searching Vega you should bear in mind that it currently covers part of the zebrafish genome. An unsuccessful search in Vega is not sufficient evidence to conclude that the query sequence is not present in zebrafish. Despite not being complete, Vega features the best sequence with the best annotation and should be your starting point when searching the zebrafish genome. As explained in section 3, the Vega database also contains sequenced clones from the AB strain (the AB chromosome). Chromosome U collects all the sequenced clones that have not been assigned to chromosomes.

Searching for all Finished/Unfinished clones

New sequenced clones come through the pipeline on a daily basis. These sequences are submitted to EMBL/GenBank. Although sequenced clones are made public as soon as possible it takes time until they appear in Vega or in a new assembly. The Sanger Institute offers a Blast search page whose target is all the available sequenced clones for zebrafish. This service can be accessed through the *Danio rerio* project page or directly at:

http://www.sanger.ac.uk/cgi-bin/blast/submitblast/d_rerio

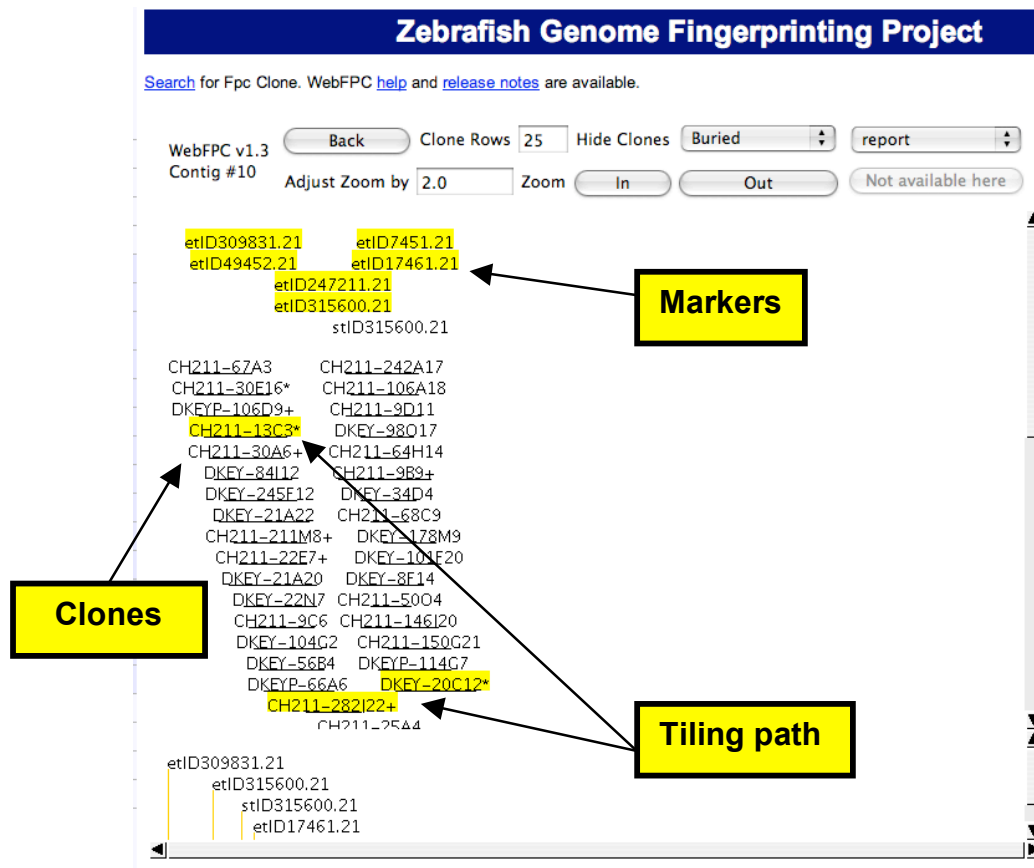
The screenshot shows the 'D. rerio Blast Server' web interface. The page has a header with the Sanger Institute logo and navigation links. A left sidebar contains a 'Hide Navigation' menu with categories like 'All Sequencing', 'Human (HGP)', 'Pathogens', and 'Blast'. The main content area is titled 'D. rerio Blast Server' and includes a 'Find out more about wu-blast' link. Below this is a 'QUERY DATA' section with a text input for 'Paste your sequence here' and a 'Browse...' button for selecting a file. An 'OPTIONS' section below that allows selecting the database ('D. rerio finished sequences') and the search method ('BLASTN (DNA vs. DNA)'). Checkboxes for 'Filter low complexity regions' and 'Mask repetitive sequences using Repeatmasker' are also present. A 'Retrieve BLAST result' section at the bottom has a 'retrieve' button. Four yellow callout boxes with arrows point to specific features: 'choose a file' points to the 'Browse...' button; 'Paste your sequence or...' points to the text input field; 'Finished (and unfinished) clones' points to the database selection dropdown; and 'Select method' points to the search method dropdown.

choose a file

Paste your sequence or...

Finished (and unfinished) clones

Select method



Trace repository

The whole genome shotgun assembly used to build the zebrafish assembly is based on a collection of reads from cosmids, fosmids and BAC ends. This collection currently gives around 7x coverage of the genome. All these reads are deposited into the trace repository at

<http://trace.ensembl.org>

This page contains a long list. Look for *Danio rerio* and then expand the list. The reads are sorted by type and the sequencing centre of origin. All these reads can be downloaded but you can also perform a SSAHA search using the server at:

<http://trace.ensembl.org/perl/ssahaview>

If you know the name of a read or the prefix from the plasmid, fosmid or BAC you can use the provided text boxes to search for them.

Exercises

1. A TextView search looks for data in different indices and includes stable ids and other text-based information. Try for example using the text “activator of transcription” (including the quotes) in Vega or Ensembl.
2. Perform a SSAHA search with AF229449 but using the *Danio rerio* Ensembl database as the target (in the example above the search was done with Vega as the target). Do you obtain the same results using SSAHA and BLAST?
3. Search the Ensembl/Vega database with a human/mouse cDNA. What is the best approach?
4. Search the Ensembl/Vega database with a human protein. Why is SSAHA not in the list of possible tools?
5. Use the sequence

GCCCTTAAGTATCGGCTTCTTCAGCAAGAGAGTTGCAAAGTACAG

to search in Ensembl using SSAHA. Where did you find the best match? Try using the same sequence with BLAST. Why do you think you get more hits with BLAST? Try using the same sequence in Vega.

5 – Does my gene have a known orthologue?

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:

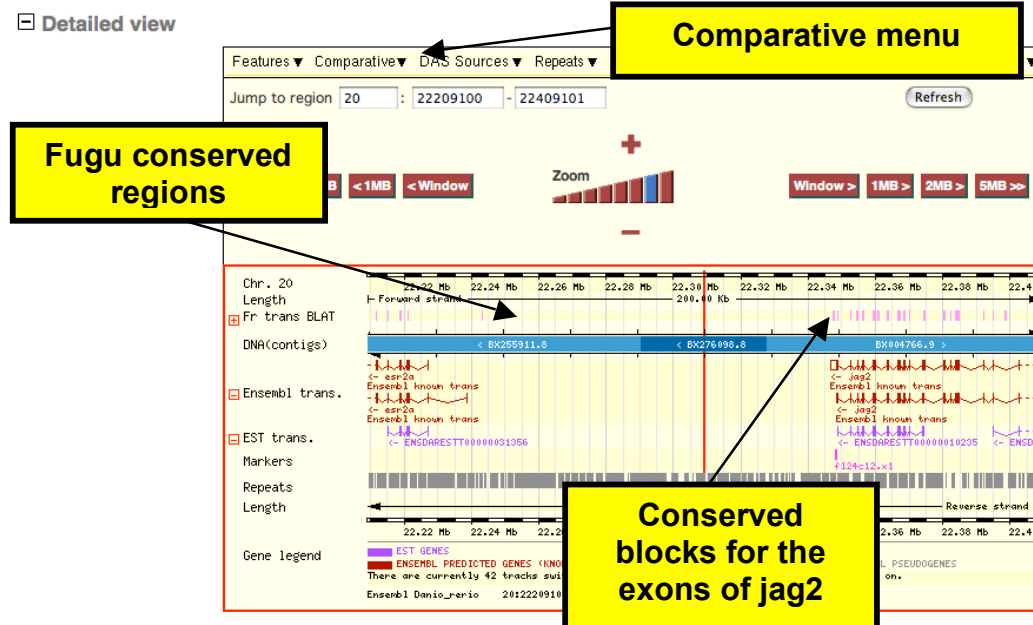
- whole genome alignments
- gene orthology/paralogy prediction
- protein clustering
- syntenic regions (not available for zebrafish)

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

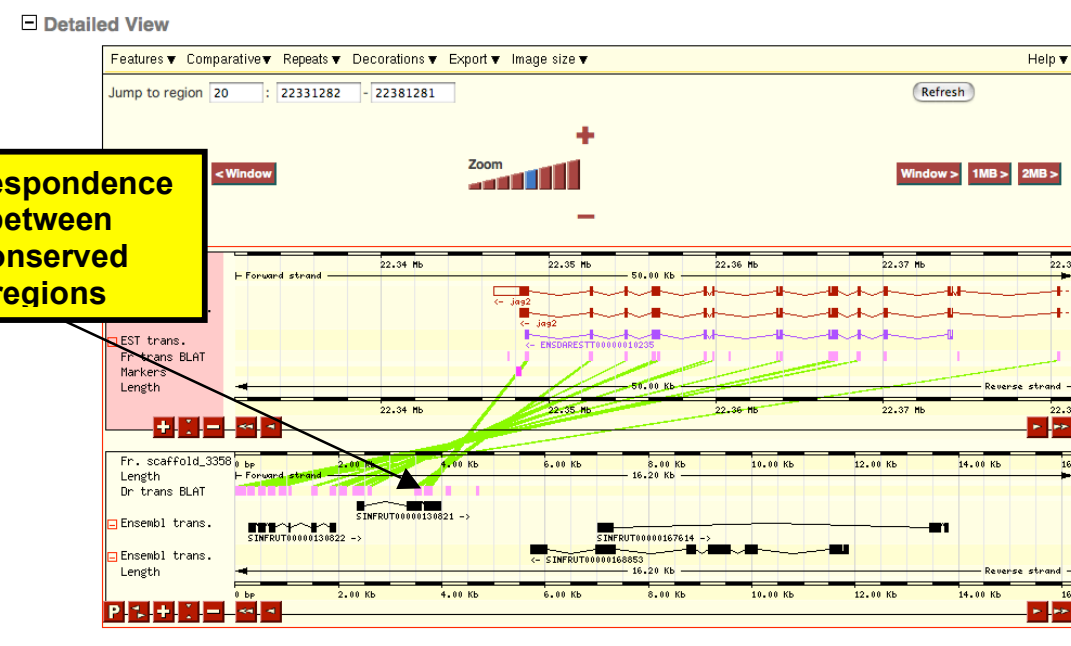
The alignment of the whole DNA sequence from two organisms is computationally demanding. Such data are of great interest both in studies of the mechanisms of molecular evolution and in attempts to identify conserved functional sequences such as novel genes and regulatory regions. Whole genome alignments 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.



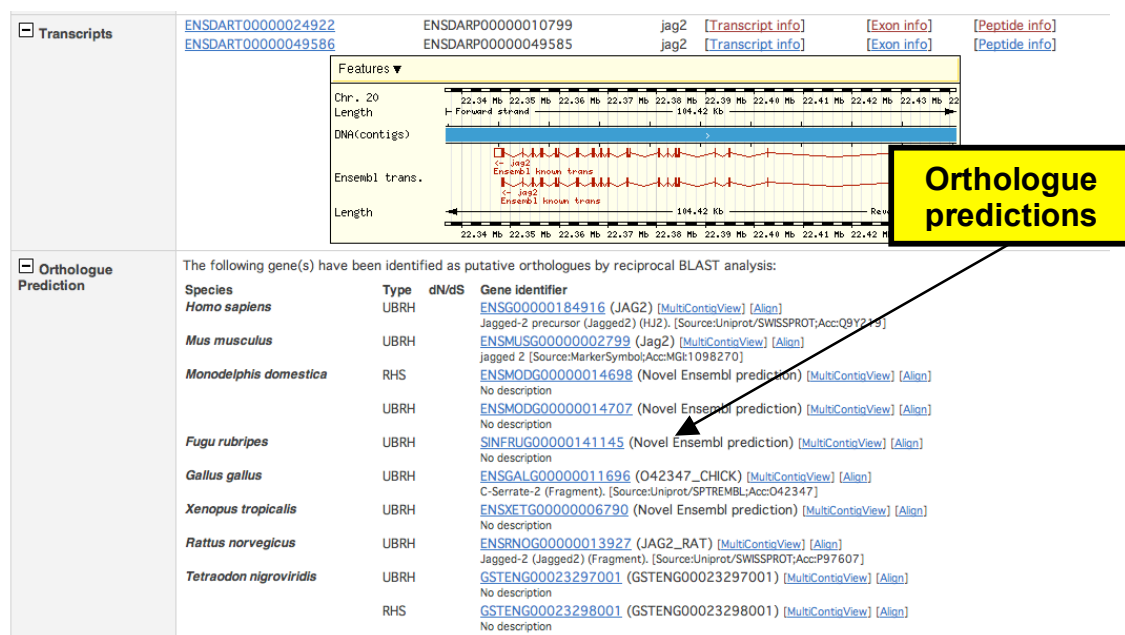
A **MultiContigView** allows the visualisation of syntenic regions from multiple species. In the example above a region from chromosome 20 in zebrafish is compared to scaffold_3358 in the Fugu assembly. The conserved blocks are connected with green lines. Observe that in this example the exons from jag2 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.



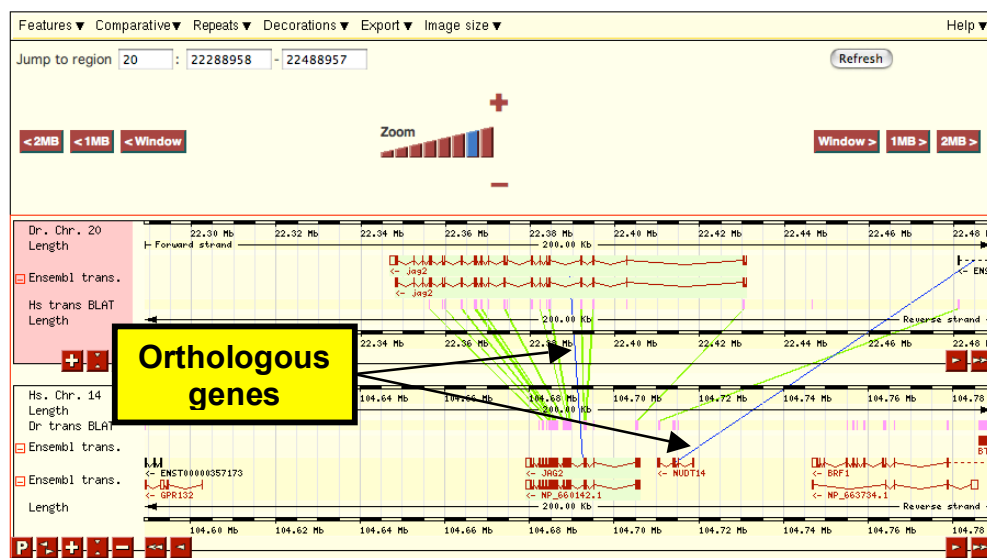
The screenshot displays the Ensembl GeneView page for the *jag2* gene. The top section shows the gene's location on chromosome 20, with a scale bar indicating a 194,42 kb region. Below this, the 'Orthologue Prediction' section lists putative orthologues identified by reciprocal BLAST analysis. A yellow box labeled 'Orthologue predictions' points to this section.

Species	Type	dN/dS	Gene identifier
<i>Homo sapiens</i>	UBRH		ENSG00000184916 (JAG2) [MultiContigView] [Align] Jagged-2 precursor (Jagged2) (HJ2). [Source:Uniprot/SWISSPROT;Acc:Q9Y219]
<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.

6 – Data Mining using BioMart

Aims

- Introduce BioMart, a data mining system for large datasets

Introduction

The BioMart system extends the Ensembl genome browser's capabilities, facilitating rapid retrieval of customised datasets. A wide variety of complex queries are supported, on various types of annotations, for numerous species. These can be applied to many research problems, ranging from SNP selection for candidate gene screening, through cross-species evolutionary comparisons, to microarray annotation. Users can group and refine biological data according to many criteria, including cross-species analyses, disease links, sequence variations, and expression patterns. Both tabulated list data, and biological sequence output can be generated on the fly, in HTML, text, Microsoft Excel and compressed formats. A wide range of sequence types, such as cDNA, peptides, coding regions, UTRs and exons, with additional upstream and downstream regions, can be retrieved. Ensembl can be accessed via a public web site or through a Java application suite.

MartView

MartView implements the user interfaces to the system. Follow the link to BioMart from the left-hand menu toolbar:

The screenshot shows the Ensembl Zebrafish website interface. On the left-hand side, there is a menu titled 'Use Ensembl to...'. Under this menu, the option 'Data mining (BioMart)' is highlighted with a yellow box. A yellow callout box with the text 'Link to MartView' points to this option. The main content area displays 'Explore the Zebrafish genome' with various links and a karyotype visualization. The top of the page includes a search bar and a 'Go' button.

Queries in BioMart are organised into three steps: **start**, **filter** and **output**. The user can navigate between these three stages using the 'Back' and 'Next'

buttons provided. Below is a detailed description of each step using MartView as an example.

Start

The start stage includes the initial selection of the species and focus for the query. Each species is designated with its genome assembly version. There are three possible foci: Ensembl, SNP and Vega. Select the dataset for Ensembl and *Danio rerio* in the species box and click 'next'.

Focus

Select the dataset for this query

Vega

Homo sapiens vega genes (NCBI35)

Using MartView

After choosing a DATASET above, select some FILTERS on the next page and then which data you want to EXPORT from the OUTPUT page. At any stage the COUNT button will calculate the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.

species

bio.mart

count help

Summary

▶ start
① Not yet initialised

▶ filter
① Not yet initialised

▶ output
① Not yet initialised

Filter

This stage allows the user to limit the initial search to a subset satisfying particular criteria. A wide range of filter types can be applied, in any combination. The system supports batch querying and a set of external identifiers can be uploaded directly from a file. The region filter allows a search to be carried out on the full genome, on a single chromosome, or on a portion of a chromosome (as determined by markers, bands or base pair coordinates). The availability of other filter options depends on the data content for a particular species and focus. For gene foci, multi-species filters can limit the selection of genes to those associated with homologues in other species, or with an upstream region that is conserved between species. Further filters allow restriction to a particular gene type or to genes that have been mapped to a particular external id set (for example, Affymetrix, EMBL, Gene Ontology or ZFIN identifiers). Searches can also be limited to genes with protein products possessing particular features, such as the presence of a transmembrane domain, signal sequence, or other domain specified using identifiers from domain databases. Access to expression data stored in BioMart is provided via the eVOC controlled expression vocabulary. Currently two datasets can be accessed in this way: the GNF microarray dataset and EST-derived expression data. Finally, one can restrict searches to genes with SNPs in particular regions (for example, coding or UTR), or to genes that have non-synonymous SNPs.

For example, the following configuration of filters selects genes that satisfy the following criteria:

- placed in chromosome 20
- have at least two transcripts
- have identified orthologous genes in Fugu

The screenshot shows the bioMart web interface with the following configurations:

- REGION:**
 - ☐ Chromosome: 20
 - ☐ Base pair: Start, End
 - ☐ Marker: Start, End
- GENE:**
 - ☐ Known genes: Only
 - ☐ ID list limit: Ensembl Transcript ID(s)
 - ☒ Transcript count >= 2
 - ☐ Entries with a 5' UTR: Only
 - ☐ Entries with a 3' UTR: Only
 - ☐ Gene type: protein_coding
 - ☐ Source: ensembl
 - ☐ Status: KNOWN
- MULTI SPECIES COMPARISONS:**
 - ☒ Homologous Fugu Genes: Only
- GENE ONTOLOGY:**
 - ☐ Molecular function: Evidence code: Any, Molecular function: <find>
 - ☐ Biological process: Evidence code: Any, Biological process: <find>
 - ☐ Cellular component: Evidence code: non-IEA, Cellular component: <find>

Summary on the right:

- start: Dataset: Danio rerio genes, 22877 Entries Total
- filter: Not yet initialised
- output: Not yet initialised

For each filter a MartView user can define whether the criteria should be satisfied or not. Click next to advance to the next stage.

Output

In this stage we can select the format for the output, but first it might be of interest to check how many genes passed our criteria. We can find this information on the right-hand side of the page. For the output we can for example require the following data:

- chromosome name (in this case, all should be on chromosome 20) and chromosome start
- Ensembl id for the gene
- ZFIN id if available

The screenshot shows the bioMart interface for selecting attributes. The 'Region' section has 'Chromosome Name' and 'Start Position (bp)' selected. The 'Gene' section has 'Ensembl Gene ID' and 'ZFin Primary ID' selected. The right sidebar shows a summary of 22877 entries and filter settings.

- and finally the output format is HTML

The screenshot shows the 'Select the output format:' section. The 'HTML' radio button is selected. Other options include 'Text, fixed width', 'Text, comma separated', 'Text, tab separated', 'MS Excel', and 'Predefined ADF attributes'.

In order to get the output, click on the Export button. The output for the genes appears in a table with links to the Ensembl database. Click on one of these genes and verify that all the selected criteria have been satisfied.

Exercises

1. Try your own queries. Experiment with different filters and outputs. In particular try with the “sequence” option for output. You can export cDNA, genomic sequences and so on.
2. Dump all predicted coding regions that contain a tubulin domain. How would you approach this query?
3. In the filter stage an extra dataset can be added. In which situation is it useful to query two datasets?
4. In the Filter step the user can specify terms from the ZFIN control vocabularies for anatomical and developmental stages data. For example get a list of all the genes known to be expressed in the blastula developmental stage. Click on ‘find’ to get the full list of terms.