
Module 2: Maps and Genome Sequence

(Sanger)

i - The Ensembl Genome Browser

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 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. The latest zebrafish assembly is Zv5, which was released on May 27th. These data are currently present in a pre-Ensembl site at

http://pre.ensembl.org/Danio_rerio

A pre-Ensembl site 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 Zv5 is under preparation. The current zebrafish assembly in Ensembl is Zv4.

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 (EnSMart, see module 3.v) 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 modules.

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.

The screenshot shows the Ensembl Genome Browser homepage. At the top, there are logos for 'project Ensembl', 'wellcome trust sanger institute', and 'EMBL'. Below the logos is the title 'Ensembl Genome Browser' and a search bar. A yellow callout box labeled 'pre-Ensembl zebrafish (Zv5)' points to the 'Zv5' link in the 'Species - Ensembl' table. Another yellow callout box labeled 'Ensembl zebrafish (Zv4)' points to the 'Zv4' link in the same table. The table lists various species with their Ensembl IDs and dates. Below the table, there are sections for 'Help and documentation', 'Data', and 'Have you tried?'. A banner at the bottom promotes the 'CCDS Database' for Homo sapiens.

Species - Ensembl		
Human	NCBI 35	May 05
Mouse	NCBI m33	May 05
Zebrafish	WTSI Zv4	May 05
Rat	RGSC 3.4	May 05
Chicken	WASHUC1	May 05
Mosquito	MOZ 2	May 05
Fugu	Fugu v2.0	May 05
Fruittly	BDGP 3	May 05
Chimp	CHIMP1	May 05
Honeybee	Amel 2.0	May 05
Tetraodon	TETRAODON7	May 05
Dog	BROAD01	May 05
C. elegans	WS 140	May 05
X. tropicalis	JGI3	May 05
S. cerevisiae	SGD	May 05
C. intestinalis	JGI 1.95	May 05
Cow	Btau_1.0	
Opossum	BROAD0.5	

From the main Ensembl site you can access the zebrafish site by clicking on the appropriate species button. Notice that as there is a new zebrafish assembly in preparation there is also a link to the pre-Ensembl site for Zv5.

Ensembl Entry Points

Search for with [Lookup](#)

Show Chr/FPC From To [Lookup](#)

Retrieve a sequence [Export](#) Advanced data [Lookup](#)

Search your sequence [BLAST/SSAHA](#)

Zebrafish Genome Project

The zebrafish genome project is a collaboration between the Sanger Institute and the zebrafish community, announced during the Sanger Institute Zebrafish Workshop 2000 and was started in February 2001.

This Ensembl website features the zebrafish assembly version 4 (Zv4), as released on the 12th July 2004. This assembly was produced by integrating the whole genome shotgun assembly with data from the physical map ([more information](#)).

Datasets used for the analyses that were provided by collaborators are acknowledged [here](#).

The zebrafish sequencing project is funded by the Wellcome Trust.

You may export data from this site. Please see the [Conditions of use](#) for these data.

Annotation

Ensembl zebrafish genes (ENSDAR*) are generated automatically by the Ensembl gene builder. A limited number of Zebrafish clones have also been manually annotated in Vega and these will be imported into Ensembl once the Zebrafish assembly merges the whole genome shotgun and clone sequence data.

Current Release 31.4d

Last Update: 02-09-2004
 Ensembl gene predictions: 23524
 Genscan gene predictions: 57411
 Ensembl gene exons: 214844
 Ensembl gene transcripts: 32062
 Clones: 4100
 Scaffolds: 21333
 Chromosomes: 25
 Base Pairs: 1571018465
 Golden Path Length: 1560425332

[What's New](#)

Browse a Chromosome

1 2 3 4 5 6 7 8 9 10 11 12 13
 14 15 16 17 18 19 20 21 22 23 24 25

Documentation & Help

About Ensembl [Home](#)

For context-sensitive help on any web page click [Help](#)

Questions or suggestions? Try [Help Desk](#)

Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site) [Documentation](#)

Ensembl Links and Site Map

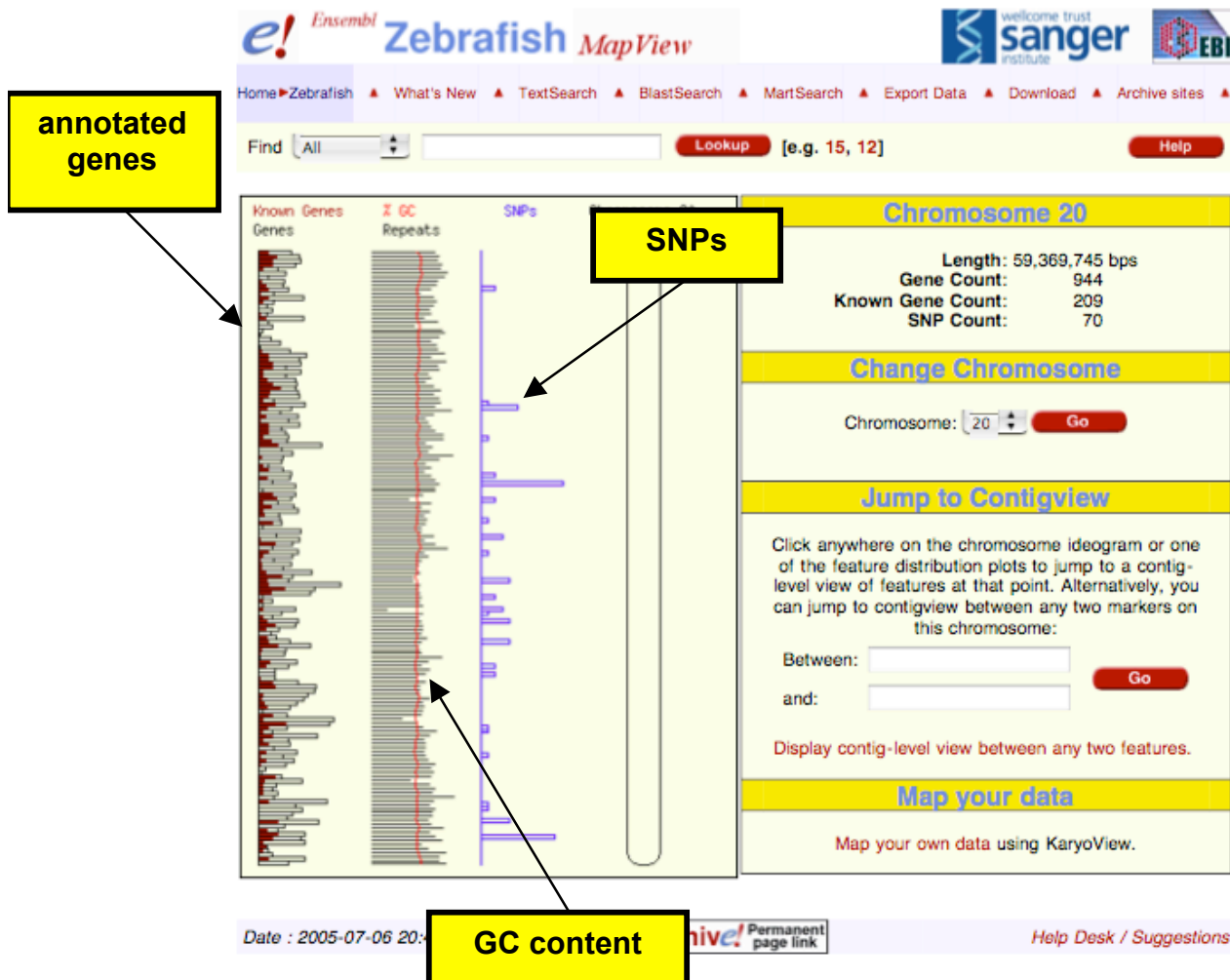
[Download](#)
[Export](#)
[EnSMart](#)
[BLAST/SSAHA](#)

Other Species

[Mosquito](#) [Honeybee](#) [C. elegans](#)
[Dog](#) [C.intestinalis](#) [Fruitfly](#)
[Fugu](#) [Chicken](#) [Human](#)
[Mouse](#) [Chimp](#) [Rat](#)
[S. cerevisiae](#) [Tetraodon](#) [X. tropicalis](#)

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 following figure 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 point of the Ensembl web site. It shows the fragments (contigs and clones) 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 more local 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 in chromosome 20 with start coordinate 25272803 and end coordinate 25357225. (Use the text box provided to enter these coordinates.)

The screenshot displays the Ensembl Zebrafish ContigView interface for Chromosome 20. The top navigation bar includes the Ensembl logo, 'Zebrafish ContigView', and logos for Sanger and EBI. A search bar contains the text '[e.g. Zv4_NA18430, Zv4_NA3486]'. Below the search bar, the 'Overview' section shows a genomic map with contigs and genes. A 'Features Menu' box highlights the 'Overview' and 'Detailed view' tabs. The 'Detailed view' section is expanded, showing a zoomed-in view of a region from 25,272,803 to 25,357,225 bp. It includes a 'Features Menu' with options like 'Compare', 'DAS Sources', 'Repeats', 'Decorations', 'Export', 'Jump to', and 'Image size'. The main content area shows various genomic features: Length, 0.refseq cDNAs, ENBL mRNAs, Unigene, Proteins, Genscan, EST trans., Ensembl trans., and DNH(contigs). A 'Features Menu' box highlights the 'Ensembl genes' section. The 'Basepair view' section at the bottom shows the DNA sequence, amino acid translations, and restriction enzyme sites. A 'Date: Wed Jul 6 19:57:49 2005' and 'Archiv' logo are visible at the bottom.

Overview

Features Menu

Detailed view

EST genes

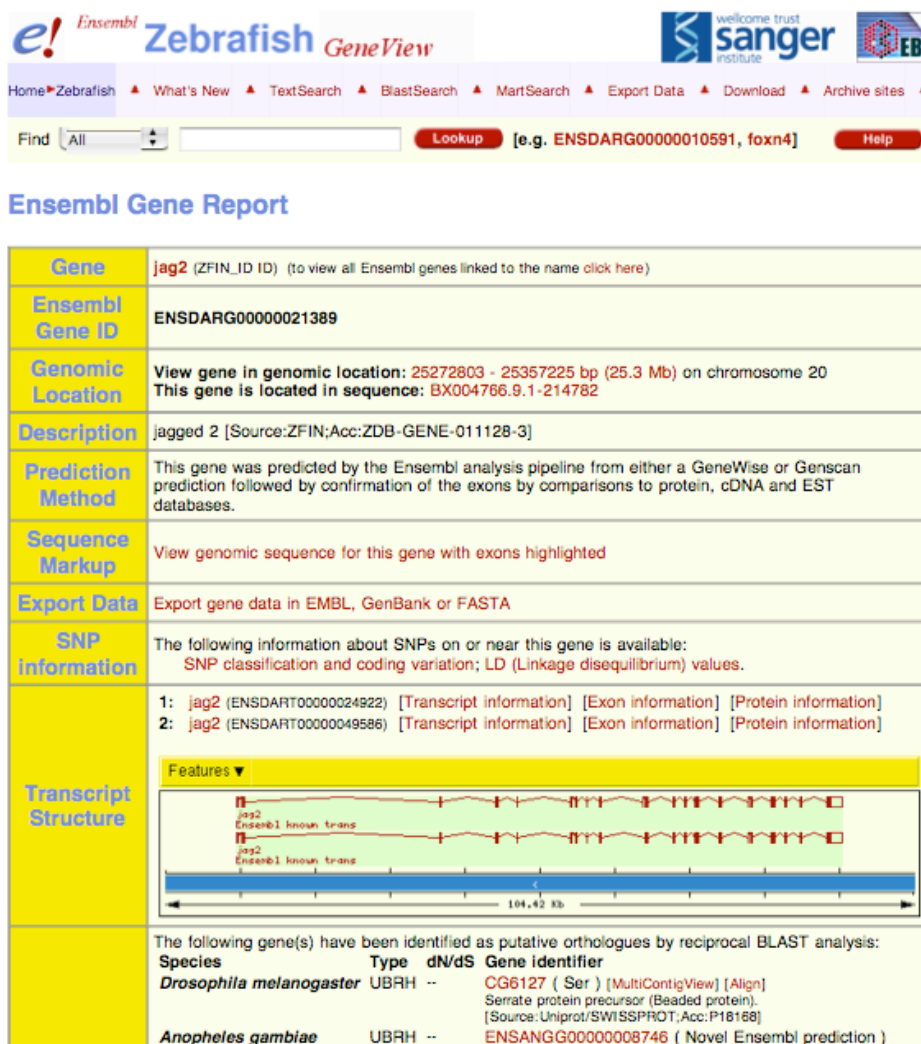
Ensembl genes

Basepair view

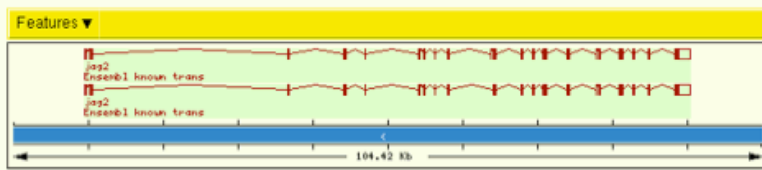
The Features menu in the detailed view controls the tracks you can visualise in the panel. Tracks can be turned on and off and the features can be collapsed to simplify the view. Spend some time on this page trying the different menus and studying the displayed features. Observe that there are two tracks 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 ProteinView

Another important view in Ensembl is the **GeneView** page with information about 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 first link labelled Gene:ENSDARG00000021389. Below we only show the top of the GeneView page for jag2; scroll down to view all the information available.



Ensembl Gene Report

Gene	jag2 (ZFIN_ID ID) (to view all Ensembl genes linked to the name click here)
Ensembl Gene ID	ENSDARG00000021389
Genomic Location	View gene in genomic location: 25272803 - 25357225 bp (25.3 Mb) on chromosome 20 This gene is located in sequence: BX004766.9.1-214782
Description	jagged 2 [Source:ZFIN;Acc:ZDB-GENE-011128-3]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases.
Sequence Markup	View genomic sequence for this gene with exons highlighted
Export Data	Export gene data in EMBL, GenBank or FASTA
SNP information	The following information about SNPs on or near this gene is available: SNP classification and coding variation ; LD (Linkage disequilibrium) values .
Transcript Structure	<p>1: jag2 (ENSDART0000024922) [Transcript information] [Exon information] [Protein information]</p> <p>2: jag2 (ENSDART0000049586) [Transcript information] [Exon information] [Protein information]</p> <p>Features</p>  <p>104.42 kb</p>
	The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis:
Species	Type dN/dS Gene identifier
<i>Drosophila melanogaster</i>	UBRH -- CG6127 (Ser) [MultiContigView] [Align] Serrate protein precursor (Beaded protein). [Source:Uniprot/SWI SSFPROT;Acc:P18168]
<i>Anopheles gambiae</i>	UBRH -- ENSANGG00000008746 (Novel Ensembl prediction)

Transcript Structure

GeneView provides annotation and supporting evidence for the selected gene. The annotation consists of transcripts, homologies 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 two transcripts. In the Transcript Structure sections there are links to the corresponding TransView pages. Click on the link labelled “Transcript information” for the first one with identifier ENSDART00000024922.



Ensembl Transcript Report

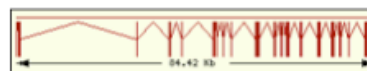
Exon Information

Transcript	jag2 (ZFIN_ID ID)
Ensembl Transcript ID	ENSDART00000024922 Ensembl
Ensembl Transcript	Exons: 26 Transcript length: 5436bp Translation length: 1254 residues This transcript is a product of gene: ENSDARG00000021389 [Exon information] [Protein information]
Genomic Location	View transcript in genomic location: 25272803 - 25357225 bp (25.3 Mb) on chromosome 20 This transcript is located in sequence: BX004766.9.1-214782
Description	jagged 2 [Source:ZFIN;Acc:ZDB-GENE-011128-3]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases.
Similarity Matches	This Ensembl entry corresponds to the following database identifiers: AFY Zebrafish: Dr.8287.1.S1_a_at EMBL: AF090432 [align] AF229449 [align] EntrezGene: 140422 Protein ID: AAC98354.1 [align] AAL08214.1 [align] RefSeq dna: NM_131665 [Target %id: 99; Query %id: 99] [align] NM_131862 [Target %id: 99; Query %id: 99] [align] UniProt/TrEMBL: Q90Y56 [Target %id: 99; Query %id: 99] [align] Q9YHU2 [Target %id: 99; Query %id: 99] [align] ZFIN ID: jag2
InterPro	IPR001438 Type II EGF-like signature - [View other EnsEMBL genes with this domain] IPR001881 EGF-like calcium-binding - [View other EnsEMBL genes with this domain] IPR001687 ATP/GTP-binding site motif A (P-loop) - [View other EnsEMBL genes with this domain] IPR001774 Delta/Serrate/lag-2 (DSL) protein - [View other EnsEMBL genes with this domain] IPR000742 EGF-like domain, subtype 2 - [View other EnsEMBL genes with this domain] IPR000152 Aspartic acid and asparagine hydroxylation site - [View other EnsEMBL genes with this domain] IPR006209 EGF-like domain - [View other EnsEMBL genes with this domain] IPR001007 von Willebrand factor, type C - [View other EnsEMBL genes with this domain]
Export Data	Export transcript data in EMBL, GenBank or FASTA

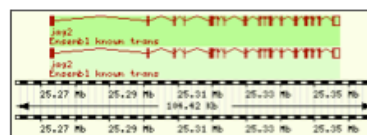
Transcript cDNA Sequence

```
GTGATCAGACCGAGGGAGAGATCAGCACAGACCATCACCGGCAAAACACCCACGCTCGT
GAATTTTGCATGTCAGGAACGGAGGATCCCTGTCCGGGCTCATCGGCCGTTTTCATCTTT
CCGTTTTAAACACATCAAATCGCGGCATGTGGAAATGTATCAGGATTAGGAATTGGCTC
CCAAATCGCGTGCCTGCTTTAACGATGTGGACGAAGGTGTCCAGTCCCTTGGCTATTTT
GAGCTGCACGTGATTGCTGTAGAAAATGTAAACGGTGAATTTGGGACGGGAAATGTTGC
GACAGCACCGGGAACCTCTCAAGACCAAGCGTTCGCTGCGGGACGAGTGCATACCTTACTTT
AAGTGTGCTGAAGGATACCACTGCTGAAGTCAACCACTGGACAGTGCACCTTCGGC
CTCGGATACCGACCTCTCTGGGAAATTAATTTCTTTAGACCGCAAAACACAGC
CCACGCAAAACGAGCGAGCTGGGAAAGATCATCATCCCTTTCACCTCCGCTCGCGGCA
TCCTACACACTCACTCTGAACTGAGCTTGGGACTGGGAACTCCACTCAGAACAAAGGTGAA
GAAAATTTGATCGAACCGCATTCACGCAACCATGGTAAACCCCGGACCATCGGACG
TCCATCCGGCACCCCTGGCATCACGGCCACATTSAGTACCGCATCCCTGTGAGGTGTGAT
GAGAATTAATGAGGAGTAAGTGAACAACAGTGTCCGCCACGAGATGACTACTTCGTT
CATTACCGATCGCATCCATCTGGAAATATCTGTGCTTTGATGGCTGGATGGGAGGAC
TGTCCGACAGCGATCTGCAAGCAGGGCTGTAAATCTGATTCACGGAGGCTGTCCGGTGCCT
GGAAATGCAAGTGCACACTACGGCTGGCAGGGCAAGTTCGCGACGAGTCTACCTTAT
CCTGGCTGTTTGCACGCTACCTGTGTTATGCCCTGGCAATGCACCTTGTGAGAAAGACTGG
GGCGCCCTCCCTTGGCATAAAGATCTGAACTACTGGCCGACGATCATCTTGTGTCAAT
GGTGGAACTGCAATGAATCTGAACCGGATGAATATAACTGTGCTGTCCCGAAGGCTAC
TCTGGCAAGAACTGTGAGATAGCTGAACATGCAATGCGTATCAAAACCCCTGTGCAACCGA
GGCACCTGTGATGAAGTCCCGACCGGATTCGAGTCCGCACTGTCCACCGCTGGGGGGGT
CCCACTTGGGCTAAAGCATGGATGAATGTGGCTCCAGCCCGTGTGGCCAAAGGCGGAACA
TGTATCGACCTGGAAATGGCTTGGATGTGCTGTCCCTCCCGAGTGGGTTGGAAAGACCT
TGTGATCGGATGCAAAATGAGTGTATGGGAGGCTTGGTAAATGCTCACTTGTGCAAA
AACATGATGGTGGATATCACGTGACCTTTCAGGATGGCCGACGAACTGTGAC
ATCAATCTCAATGGTGTCCATGGACGTGCCAGAATGGAGCTACTGCAAGGAGCTGGT
CATGGAGGTTACCACCTGTCAGTGTCTCCGCGGGTTTGGGGCTTACACTGTGAACTCTCA
AGGNATAAATGTGCCAGCGGTCATGTGAGAATGGTGGCCGCTGCCATGTCAATCTGGAC
AGCTTCGTTTGTGAGTGTCCGTCAAACTACCGAGGATGCTCTGTGAGGTGGAGAGTCTG
TCTCACCCAAACCCATGTGAGCCGAACCTTGTGCAATACAGCTTGTGCTACAGTCTG
```

Transcript Structure

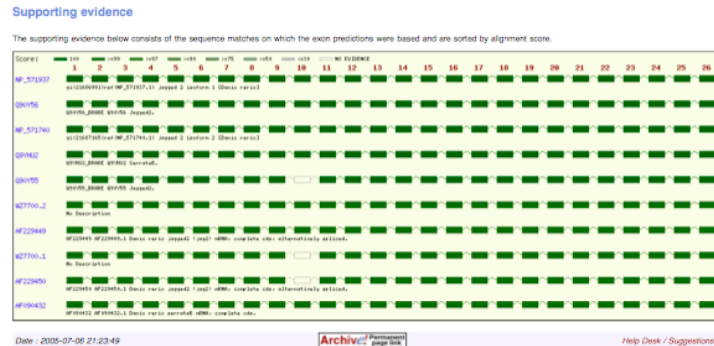


Transcript Neighbourhood



cDNA

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.



Finally from the links labelled “Protein/Peptide information” we can visit the ProteinView page for the associated translation.

Ensembl Protein Report

ExportView Link

Peptide	jag2 (ZFIN_ID ID)
Ensembl Translation ID	ENSDARP00000010799
Ensembl Translation	This peptide is a product of gene ENSDARG00000021389 [Transcript Information] [Exon Information]
Description	jagged 2 [Source:ZFIN;Acc:ZDB-GENE-011128-3]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases.
InterPro	<ul style="list-style-type: none"> IPR001438 Type II EGF-like signature - [View other genes with this domain] IPR001881 EGF-like calcium-binding - [View other genes with this domain] IPR01687 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 domain, subtype 2 - [View other genes with this domain] IPR000152 Aspartic acid and asparagine hydroxylation site - [View other genes with this domain] IPR006209 EGF-like domain - [View other genes with this domain] IPR001007 von Willebrand factor, type C - [View other genes with this domain]
Protein Family	ENSF00000000048 : PRECURSOR This cluster contains 30 Ensembl gene member(s)
DAS Sources	<ul style="list-style-type: none"> UniProt (Protein knowledgebase) Manage Sources
Protein Features	

ProteinView 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 several locations in the Ensembl web site.

The screenshot shows the Ensembl Zebrafish ExportView interface. The page is titled "Ensembl Zebrafish ExportView" and includes logos for the Wellcome Trust Sanger Institute and EBI. The navigation bar includes links for Home, Zebrafish, What's New, TextSearch, BlastSearch, MartSearch, Export Data, Download, and Archive sites. A search bar contains the query "[e.g. Zv4_NA9133, ENSDARG00000010591]" and a "Lookup" button. Below the search bar are tabs for "Flat File", "FASTA", "Feature List", and "Pip".

The main content area is divided into several sections:

- Select data to export:** Includes a "Feature" section with a dropdown menu set to "Gene" and an "ID:" field. Below it is a "Region" section with a "Chromosome:" dropdown set to "1" and "Markers from:" and "Bases from:" fields.
- Select export options:** Includes "Export as" radio buttons for "EMBL" (selected) and "GenBank". Below this is a list of features to export with checkboxes: Similarity features, Repeat features, Prediction features (genscan), Contig Information, Variation features, Marker Features, Gene Information, and EST Gene Information.
- Select output format:** Includes radio buttons for "Text" (selected), "HTML", and "Zip". Below these are "Export" and "Reset" buttons.

Four yellow callout boxes with black borders and arrows point to specific elements:

- "Select feature or..." points to the "Gene" dropdown menu.
- "region" points to the "Markers from:" field.
- "Select format for data" points to the "EMBL" radio button.
- "Select format file" points to the "Text" radio button.

At the bottom of the page, there is a footer with the date "Date : Wed Jul 6 20:38:48 2005", an "Archive! Permanent page link" button, and a "Help Desk / Suggestions" link.

Zebrafish assembly in Ensembl

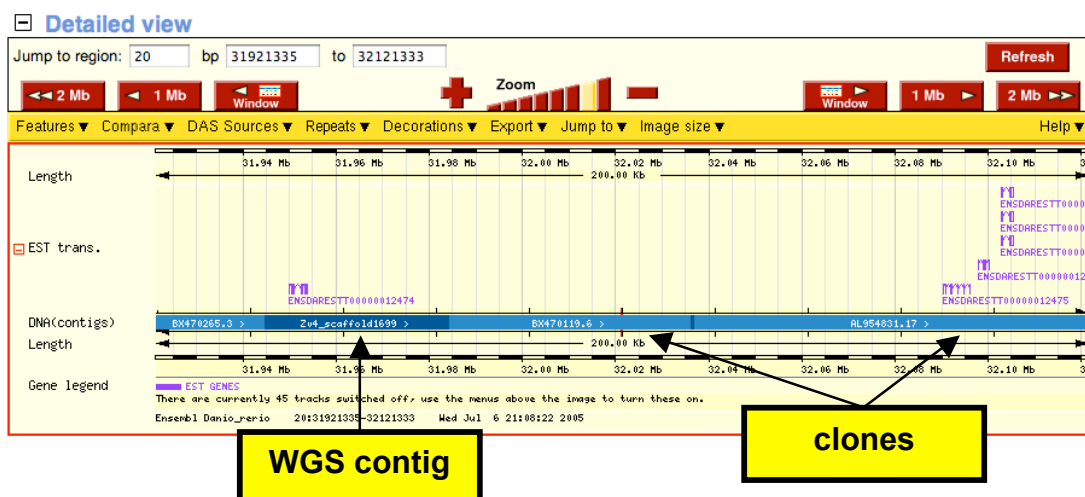
The sequence in the *Danio rerio* Ensembl database is the latest assembly release with automatic annotation. The genomic sequence released is based on all the sequenced clones with remaining gaps covered by contigs from a whole genome shotgun (WGS) assembly. The WGS fragments are placed in those gaps using a mixed strategy that looks at sequence similarity and

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:

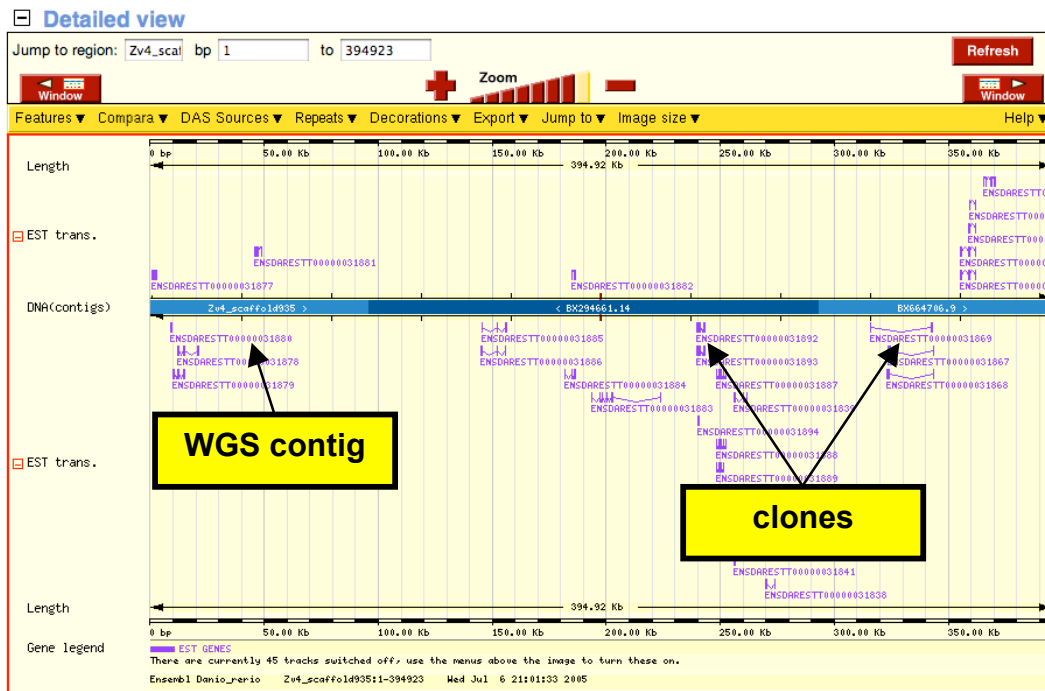
1. scaffolds that have been attached to chromosomes (they contain sequenced clones),
2. scaffolds that can be aligned to clones but the physical map cannot assign a chromosome yet, and
3. NA (non-attached) scaffolds that corresponds to WGS contigs that could not be placed in the map.

Zv5_scaffold1699 is an example of category 1 above. This scaffold is placed in chromosome 20.



In the detailed view for this page there is a genomic region labelled BX470265.3. This is the accession number of a sequenced clone. The region labelled Zv5_scaffold1699 is a WGS supercontig. This is of lower quality than the sequenced clone (and can contain some small gaps represented by a sequence of Ns).

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 Zv4_NA10. This region does not contain any finished clones.

Exercises

This module introduces the Ensembl browser and some of the basic views. In other modules 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 Zv4 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.
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.