Zebrafish Genome Resources Workshop

June 2006, Madison

Contents

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.

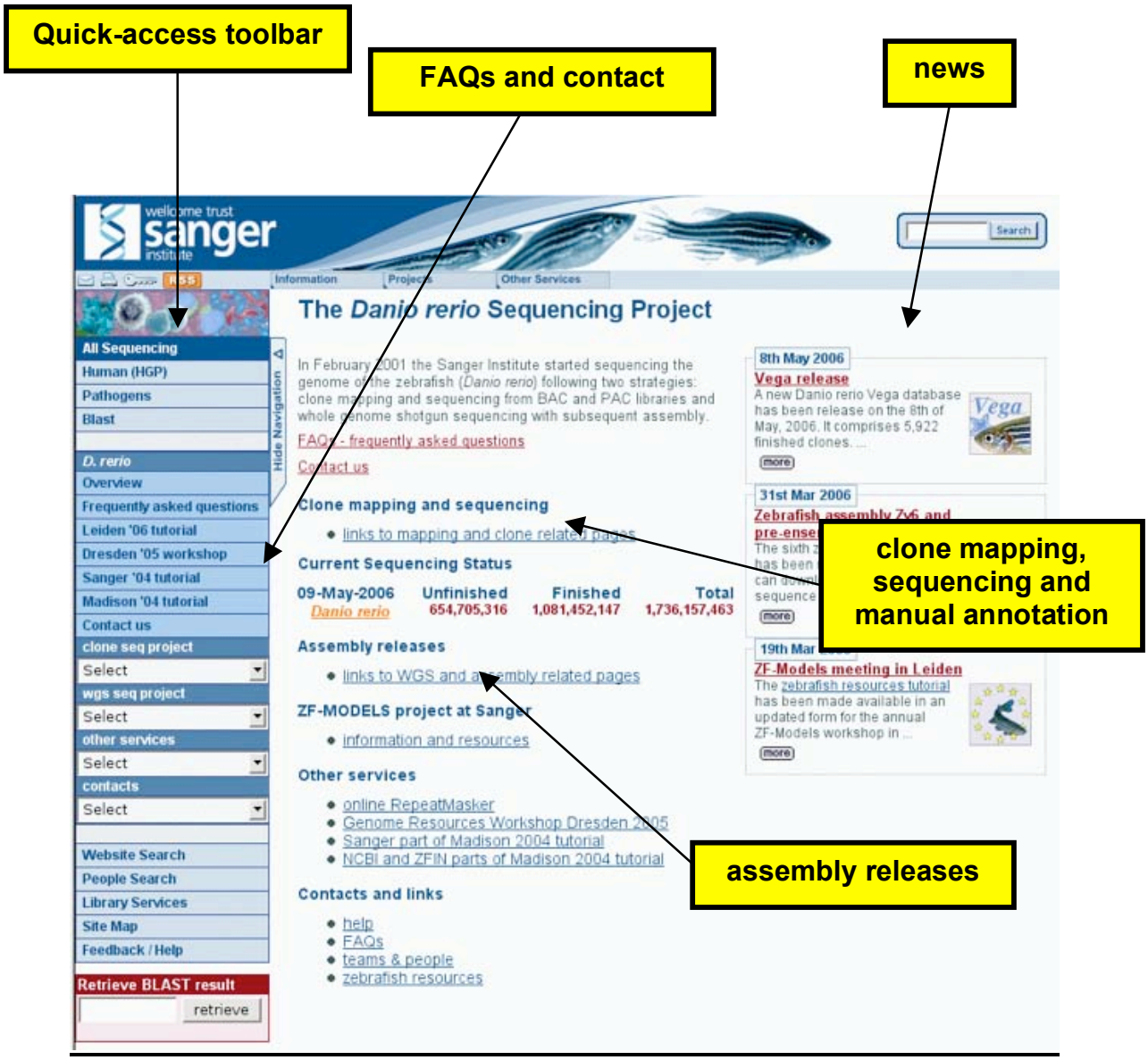
The *Danio rerio* Sequencing Project Page

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



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.

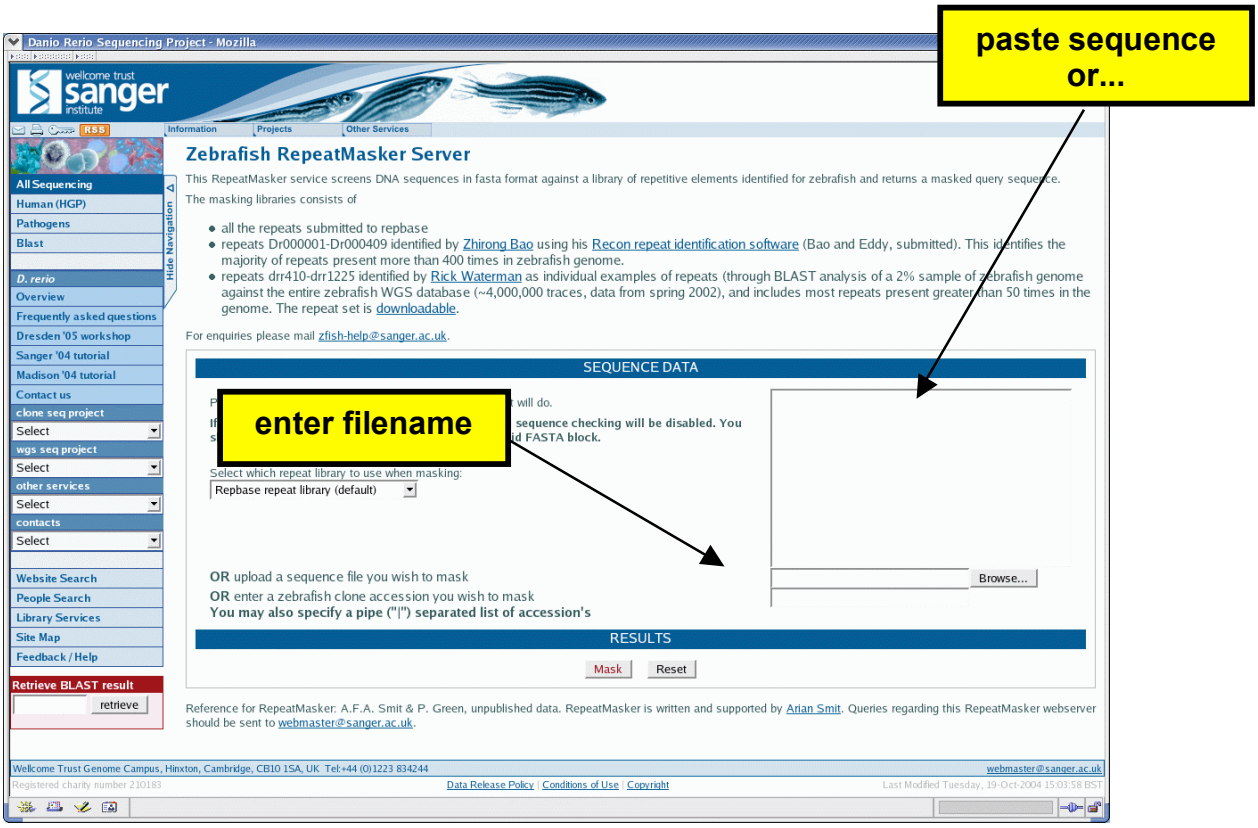
Assembly releases

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.



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.

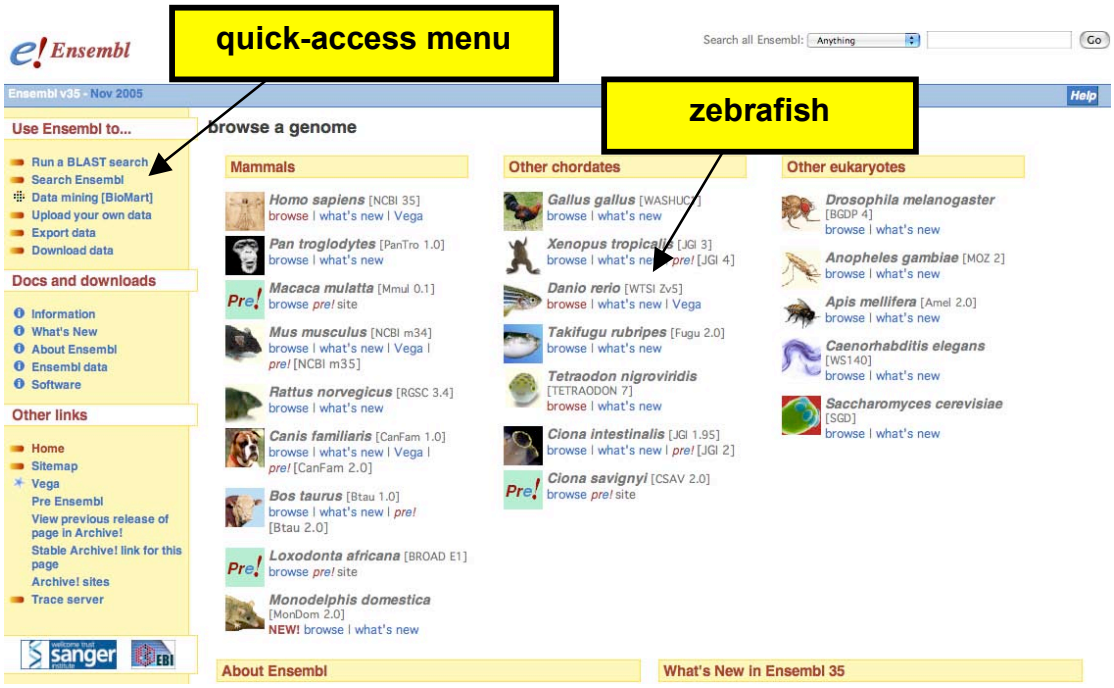
The Ensembl site provides automatic baseline annotation of the latest assembly sequence, including gene, transcript and protein predictions.The annotation is integrated with external data sources, such as ZFIN for the zebrafish site. The latest zebrafish assembly is Zv6, which was released on March 31st,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

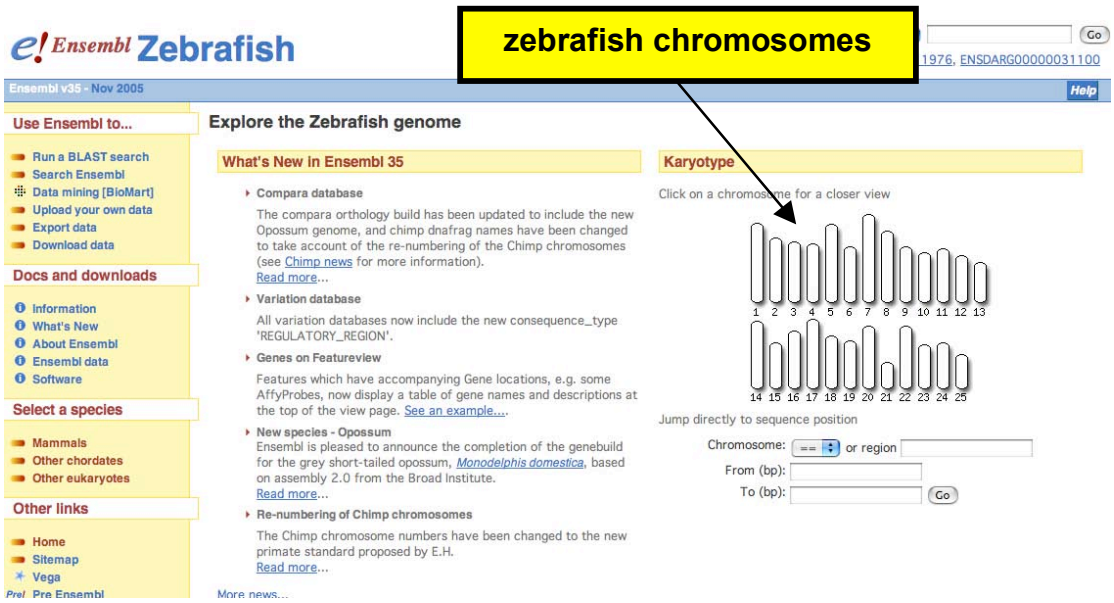
The Ensembl site is at:

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.



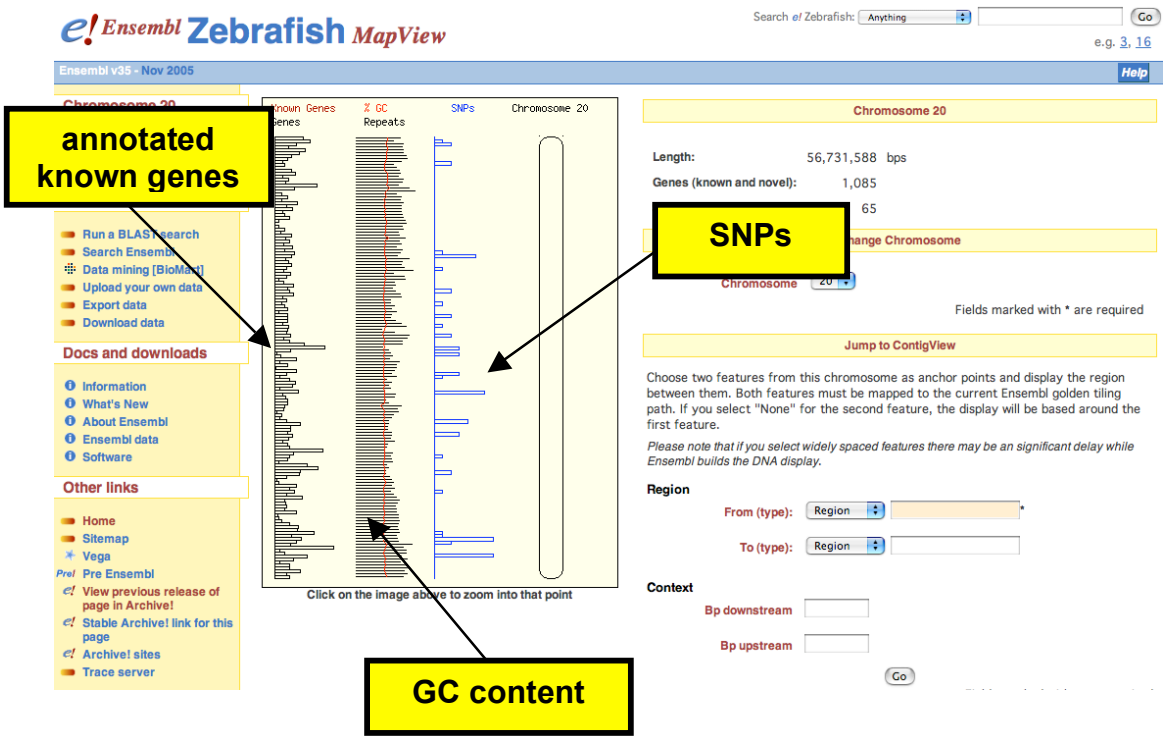
From the main Ensembl site you can access the zebrafish site by clicking on the appropriate species button. As soon a new assembly is released the sequence is made available as a pre-Ensembl site. This includes valuable information such as EST and UniProt alignments and *ab initio* predictions. The main 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.



MapView and ContigView

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

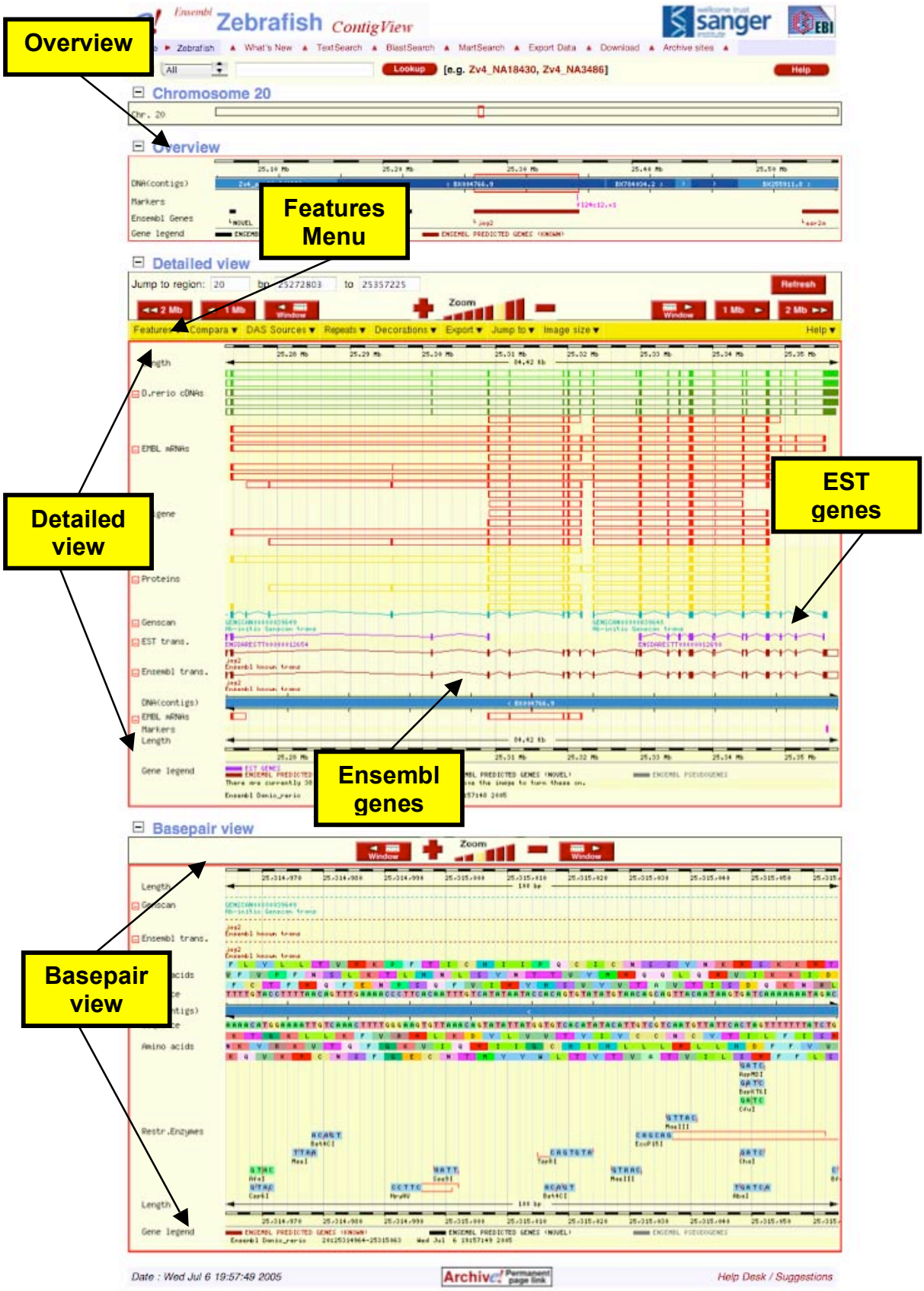
chromosomes are linked to the **MapView** pages. The figure below shows the MapView for chromosome 20.



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.)

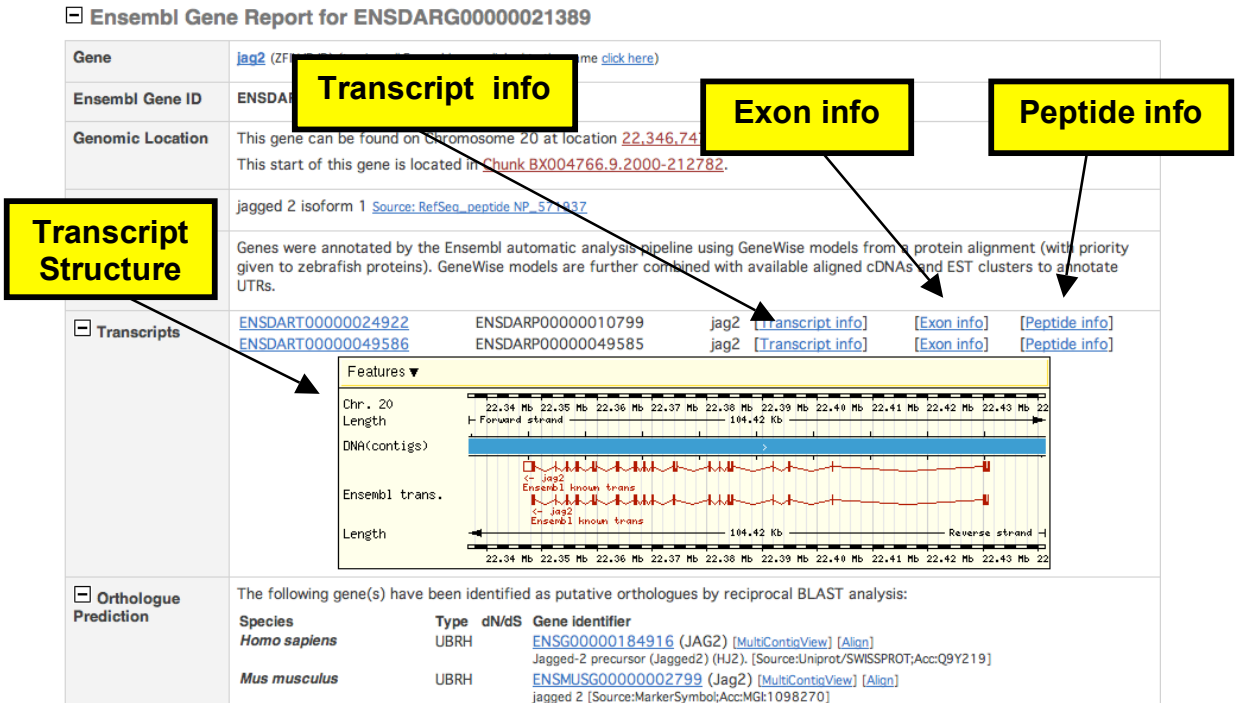

Overview
Features Menu
Detailed view
EST genes
Ensembl genes
Basepair view
Zebrafish ContigView
Chromosome 20
Overview
Detailed view
Basepair view
Date : Wed Jul 6 19:57:49 2005
Help Desk / Suggestions

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 ink labelled Ensembl Gene: ENSDARG00000021389. Below we only show the top of the GeneView page for jag2; scroll down to view all the information available.

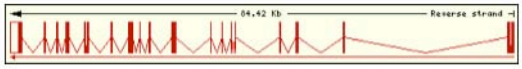
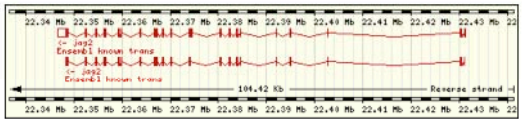
GeneView provides annotation and supporting evidence for the selected gene. The annotation consists of transcripts, homologies to other species, known and predicted proteins and domains, and links to external documentation. In this example, jag2 is a gene known to ZFIN and so a link to the corresponding external page is provided. The annotation for jag2 is based on 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 v35 - Nov 2005

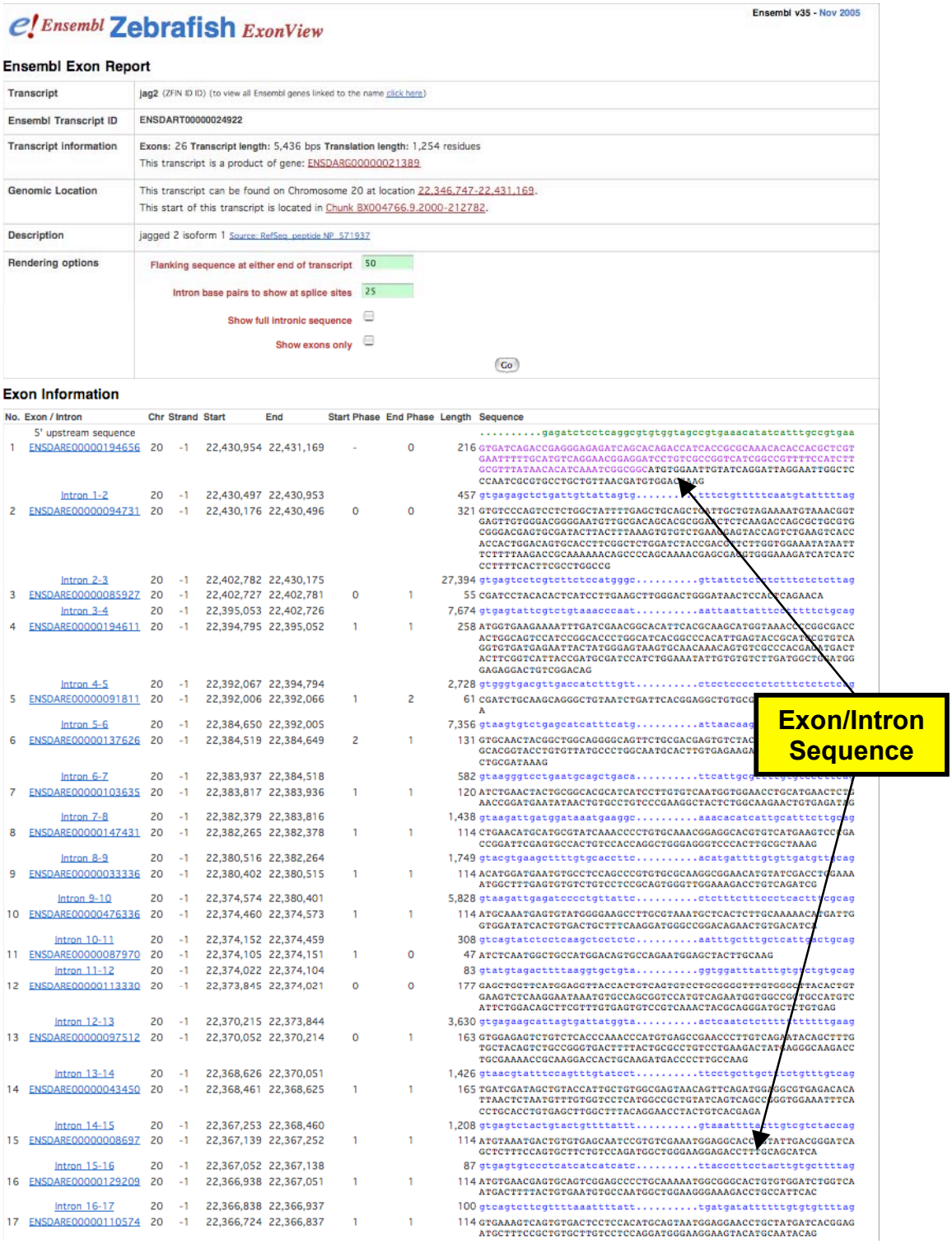
Ensembl Zebrafish TransView

Ensembl Transcript Report

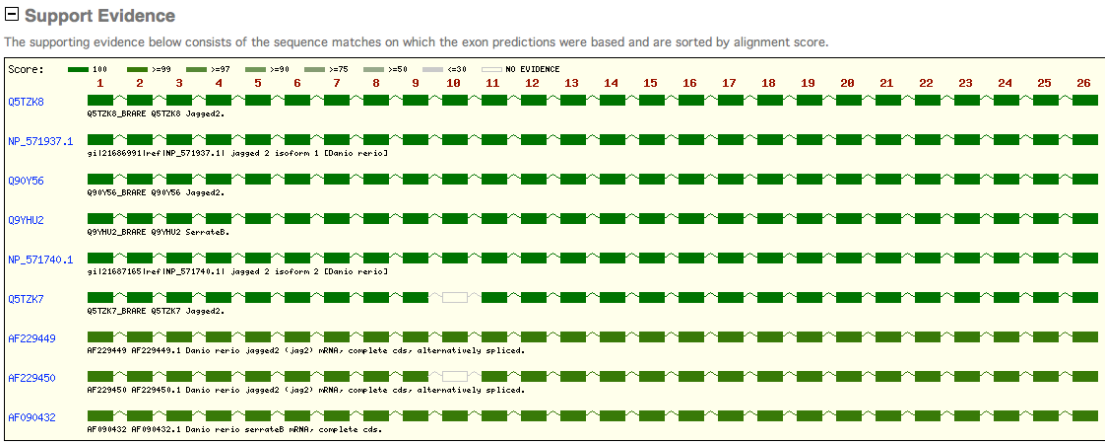
Transcript	**jag2** (ZFIN ID ID) (to view all Ensembl genes linked to the name click here)
Ensembl Transcript ID	**ENSDART00000024922**
Transcript information	**Exons:** 26 **Transcript length:** 5,436 bps **Translation length:** 1,254 residues This transcript is a product of gene: ENSDARG00000021389
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
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: 93; Query %id: 98] [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	ENSF00000000046 : PRECURSOR This cluster contains 29 Ensembl gene member(s)
Transcript structure	
Transcript neigbourhood	
Transcript sequence	GTGATCAGACCGAGGGAGAGATCAGCACAGACCATCACCGCGCAAACACACCACGCTCGT GAATTTTTGCATGTCAGGAACGGAGGATCCTGTCGCCGGTCATCGGCCGTTTTCCATCTT GCGTTTATAACACATCAAATCGGCGGCATGTGGAATTGTATCAGGATTAGGAATTGGCTC CCAATCGCGTGCCTGCTGTTAACGATGTGGACGAAGGTGTCCCAGTCCTCTGGCTAT GAGCTGCAGCTGATTGCTGTAGAAAATGTAAACGGTGAGTTGTGGGACGGGGAATGTT GACAGCACGCGGAACTCTCAAGACCAGCGCTGCGTGCGGGACGAGTGCGATACTTACTTT AAAGTGTGTCTGAAGGAGTACCAGTCTGAAGTCACCACCACTGGACAGTGCACCTTCGGC TCTGGATCTACCGACGTTCTTGGTGGAAATATAATTTCTTTTAAGACCGCAAAAAACAGC CCCAGCAAAACGAGCGACGTGGGAAAGATCATCATCCCTTTTCACTTCGCCTGGCCGCGA TCCTACACACTCATCCTTGAAGCTTGGGACTGGGATAACTCCACTCAGAACAATGGTGAA GAAAATTTGATCGAACGGCACATTCACGCAAGCATGGTAAACCCCGGCGACCACTGGCAG TCCATCCGGCACCCTGGCATCACGGCCCACATTGAGTACCGCATCCGTGTCAGGTGTGAT GAGAATTACTATGGGAGTAAGTGCAACAAACAGTGTCGCCCACGAGATGACTACTTCGGT CATTACCGATGCGATCCATCTGGAAATATTGTGTGTCTTGATGGCTGGATGGGAGAGGAC TGTCGGACAGCGATCTGCAAGCAGGGCTGTAATCTGATTCACGGAGGCTGTGCGGTGCCT GGAGAATGCAACGTGCAACTACGGCTGGCAGGGGCAGTTCTGCGACGAGTGTCTACCTTAT CCTGGCTGTTTGCACGGTACCTGTGTTATGCCCTGGCAATGCACTTGTGAGAAGAACTGG GGCGGCCTCCTCTGCGATAAAGATCTGAACTACTGCGGCACGCATCATCCTTGTGTCAAT GGTGGAACCTGCATGAACTCTGAACCGGATGAATATAACTGTGCCTGTCCGAAGGCTAC **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.

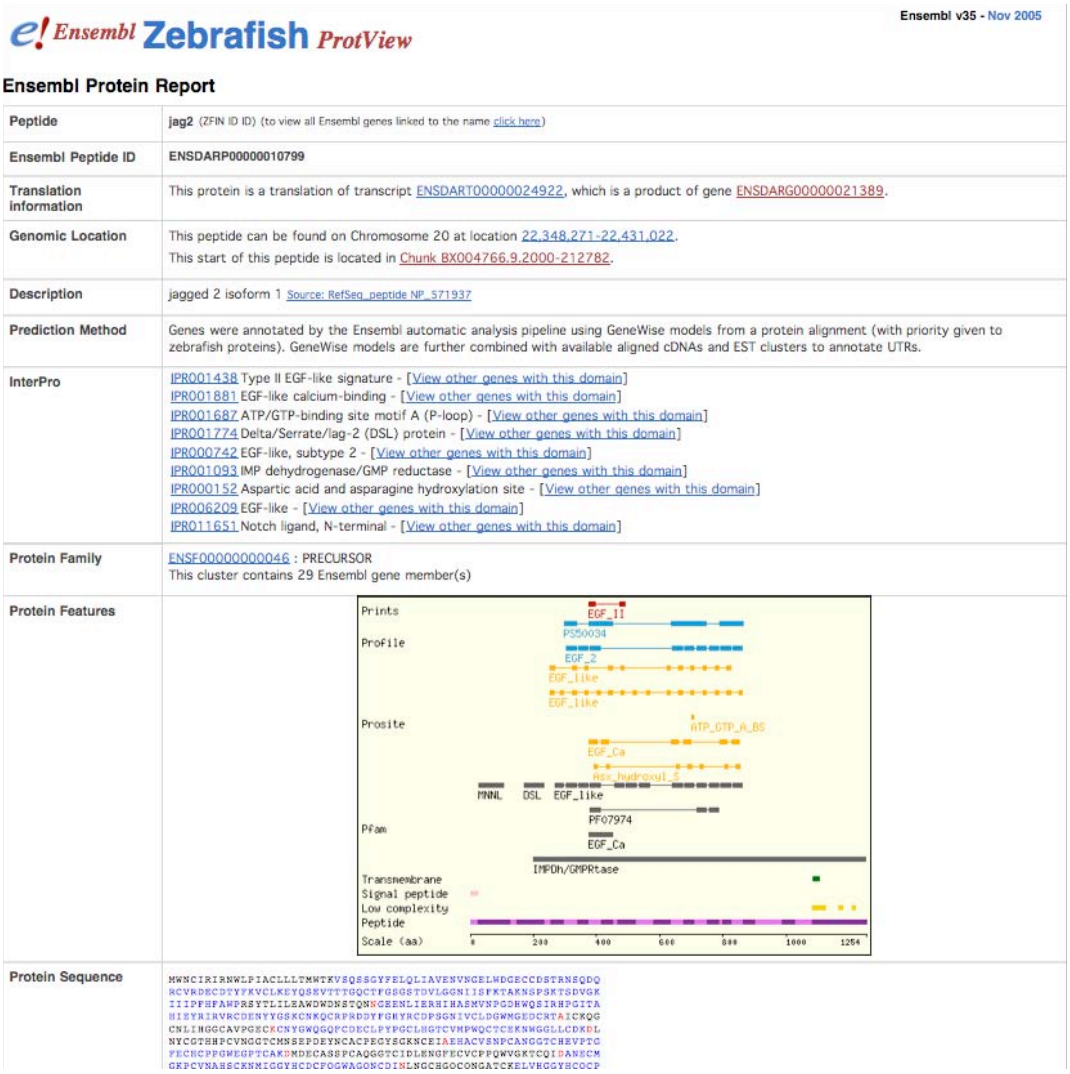
From the GeneView page there are also links to the ExonViews labelled as “Exon info”.



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



Ensembl Zebrafish ProtView

Ensembl v35 - Nov 2005

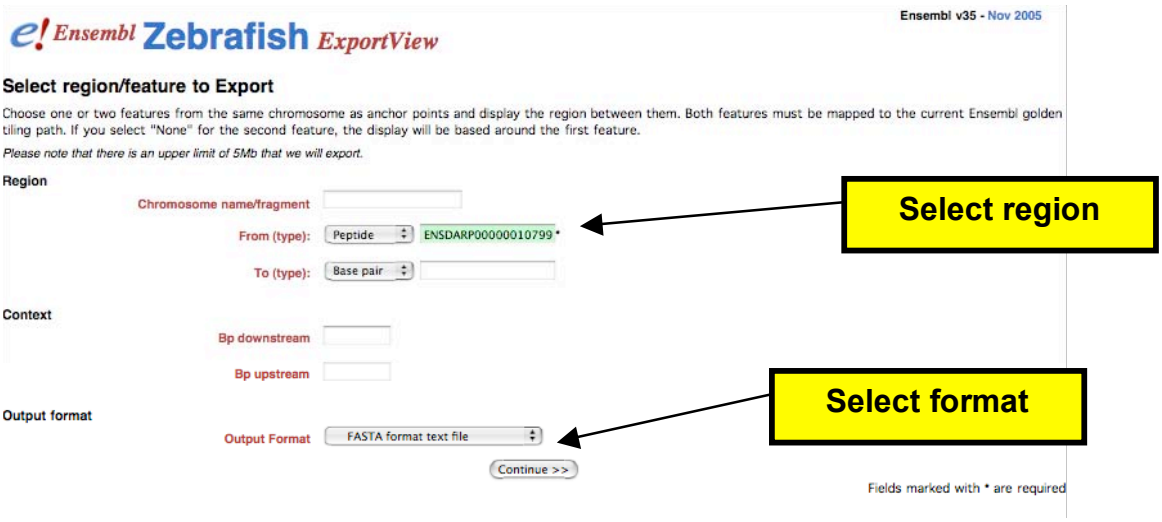
Ensembl Protein Report

Peptide	**jag2** (ZFIN ID ID) (to view all Ensembl genes linked to the name click here)
Ensembl Peptide ID	**ENSDARP00000010799**
Translation information	This protein is a translation of transcript ENSDART00000024922, which is a product of gene ENSDARG00000021389.
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/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	ENSF00000000046 : PRECURSOR This cluster contains 29 Ensembl gene member(s)
Protein Features	
Protein Sequence	

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.



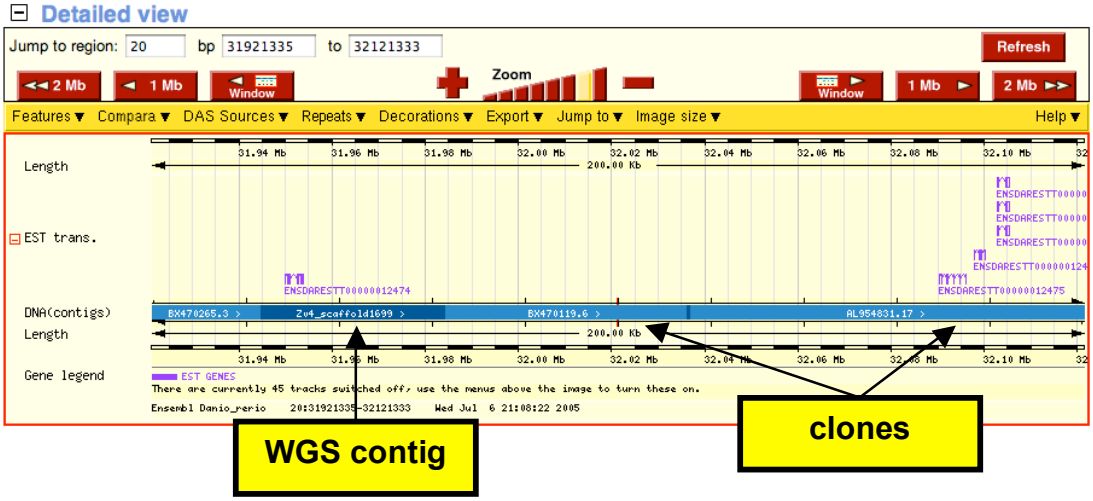
Zebrafish assembly in Ensembl

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:

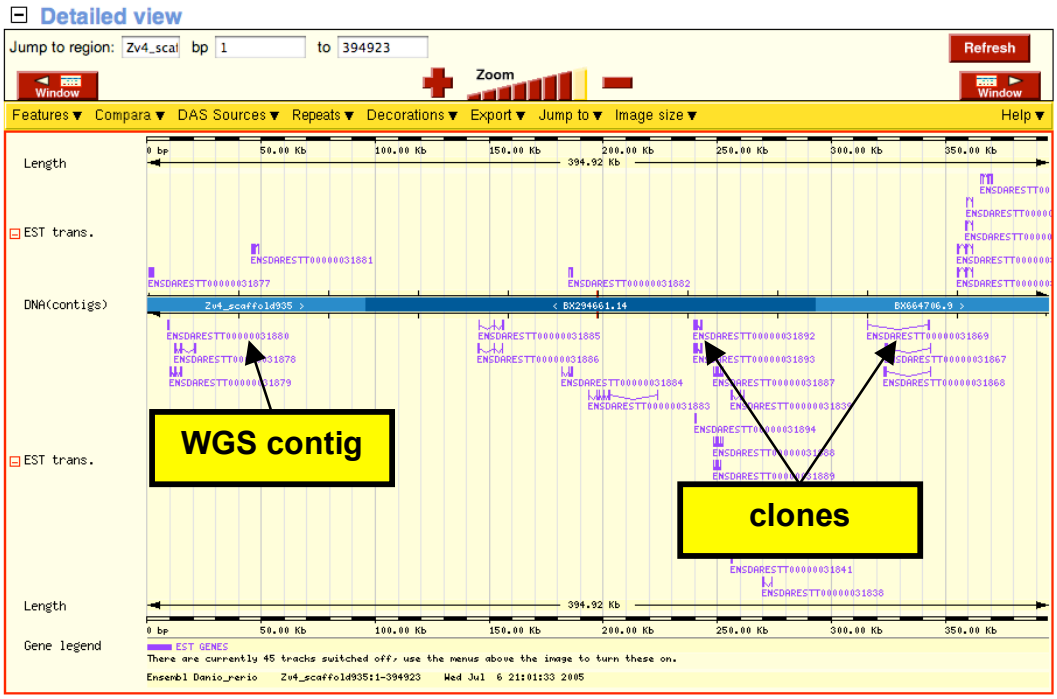
1. scaffolds that have been attached to chromosomes (they may contain sequenced clones),
2. scaffolds that can be aligned to clones but the physical map cannot assign a chromosome yet (they may contain sequenced clones), and
3. NA (non-attached) scaffolds that corresponds to WGS contigs that could not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

Zv5_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 may contain 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 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

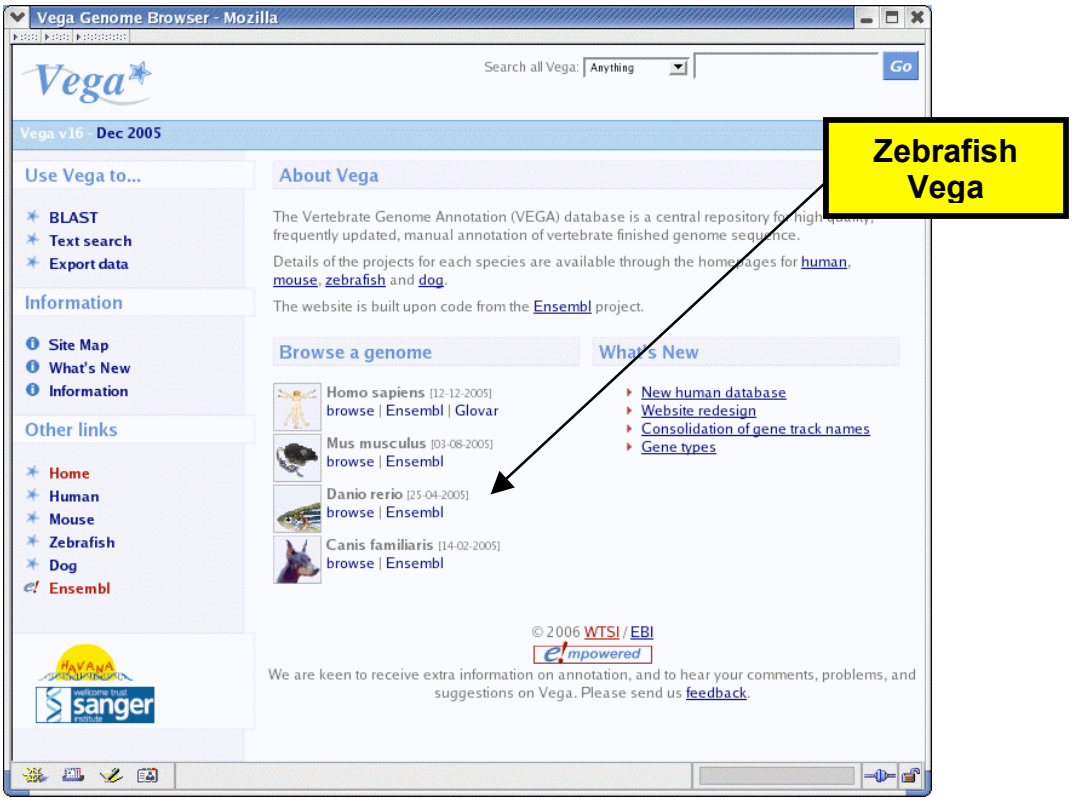
Aims

- 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

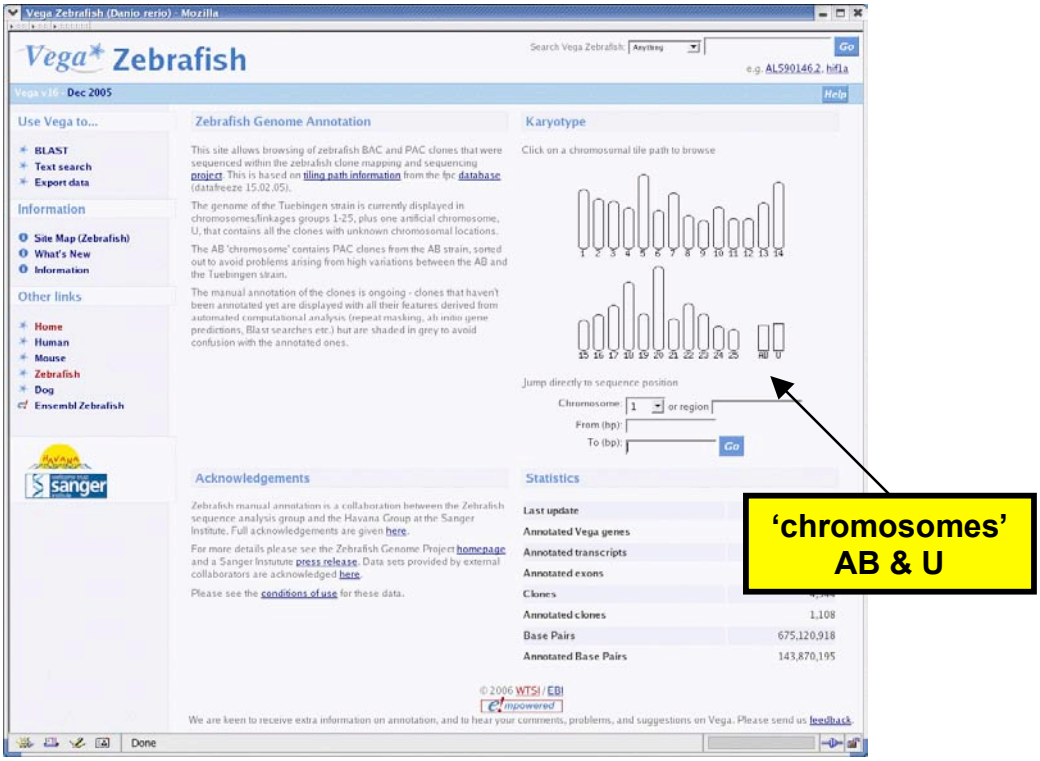
Introduction

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.



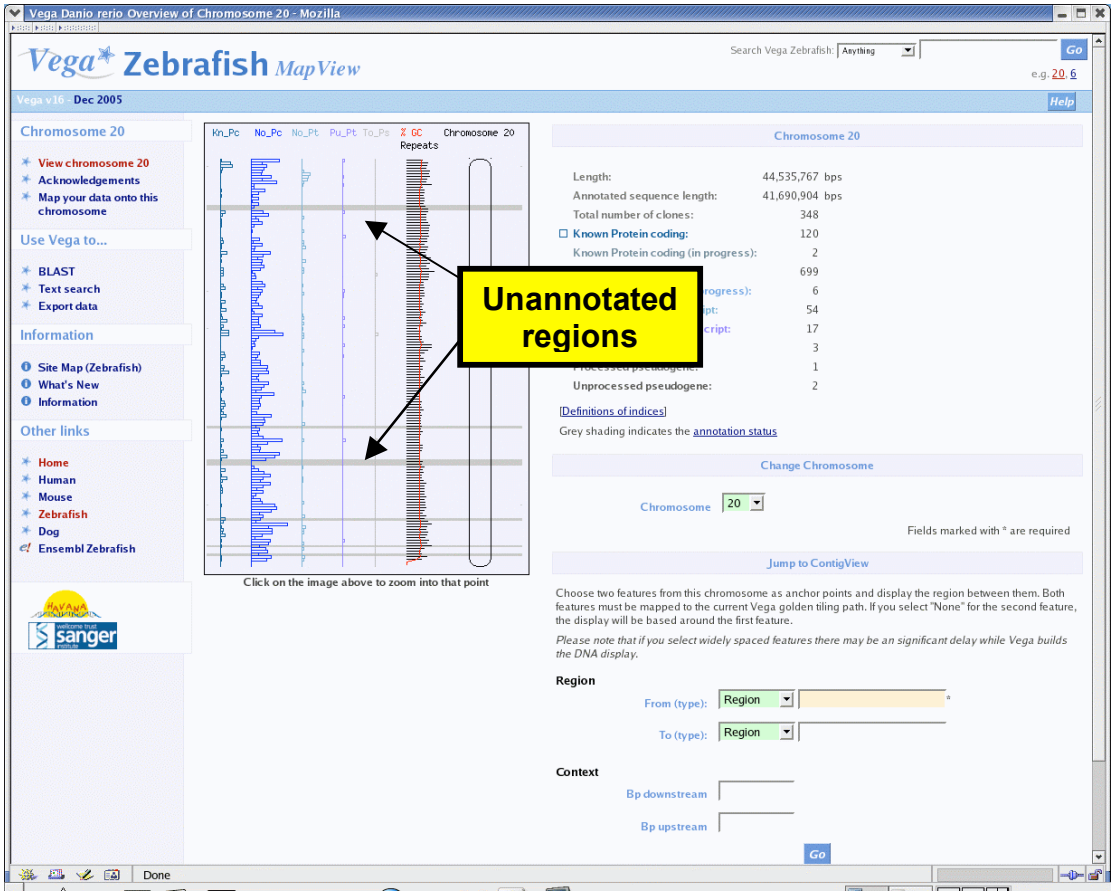
Follow the link to the zebrafish database.



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

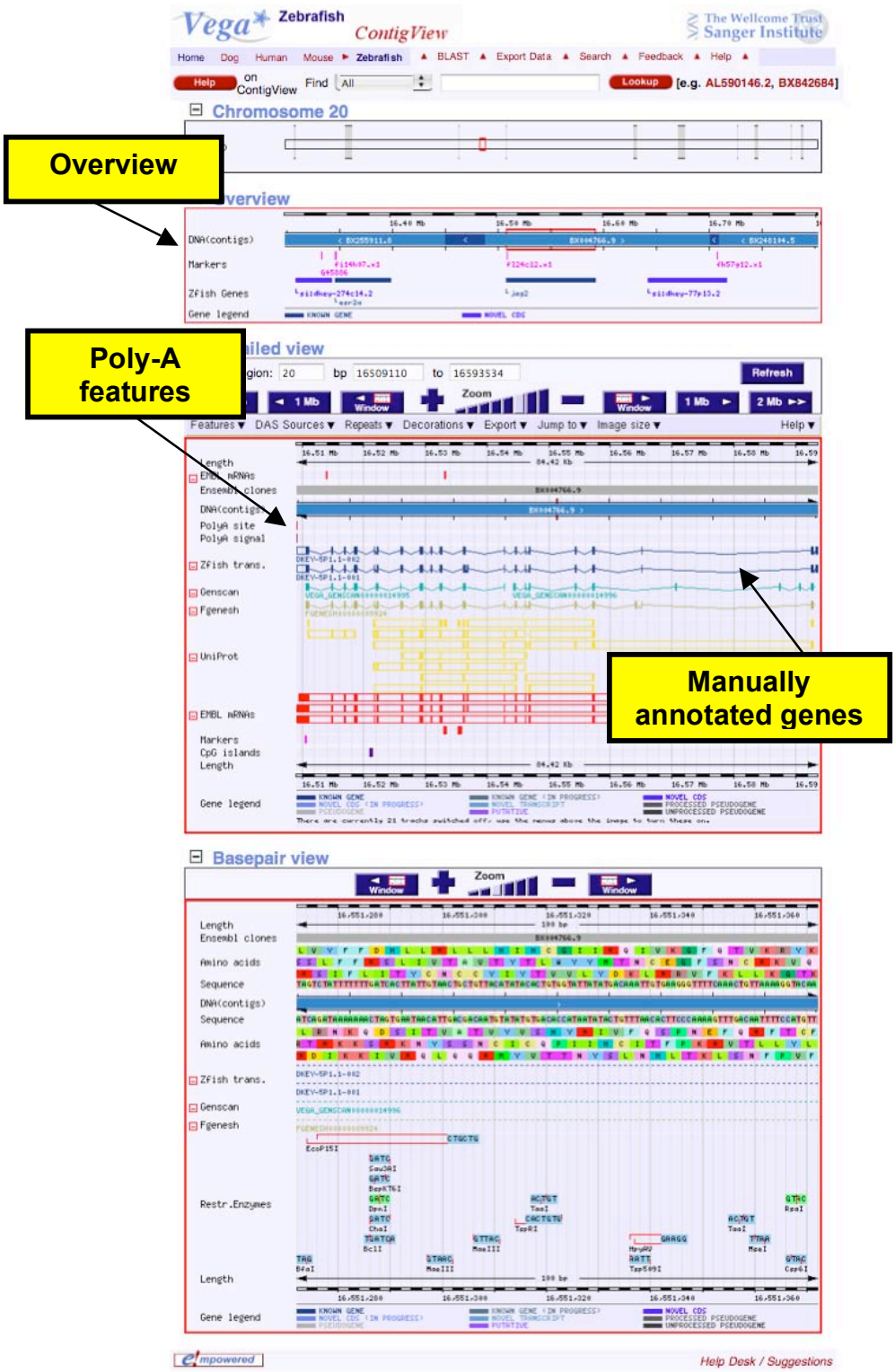
Clicking on a chromosome links to the corresponding MapView page.



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.

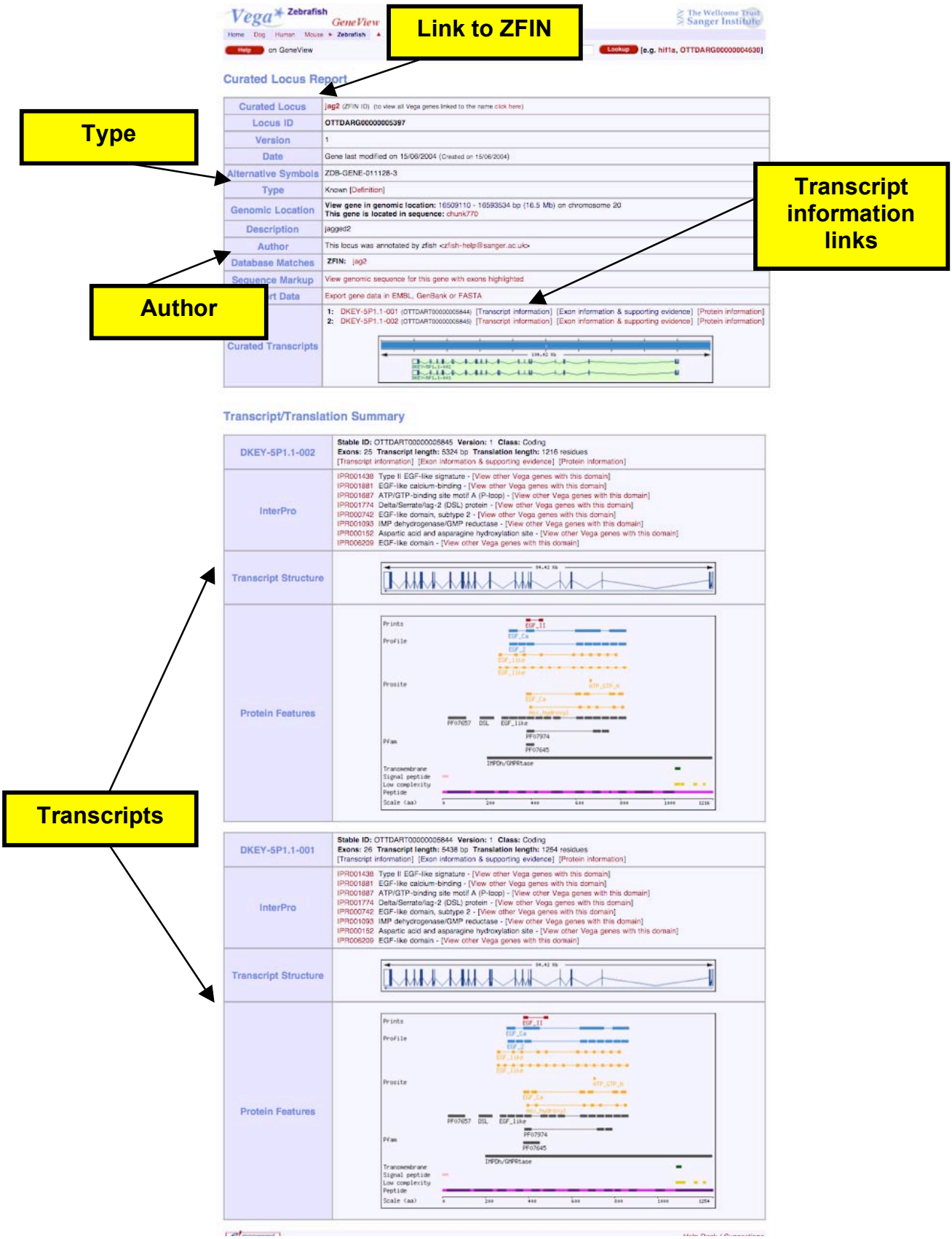
Jump to the ContigView for the region in chromosome 20 from 17925000 to 18135782.



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.



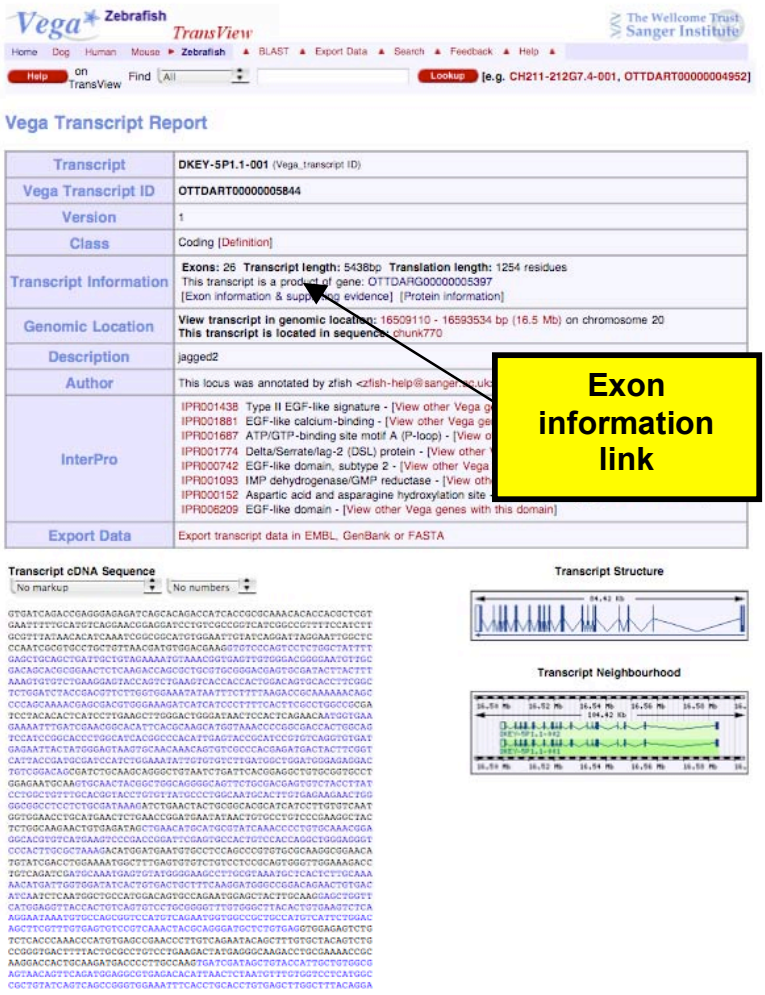
Every annotated gene in Vega has a ZFIN gene entry. Follow the link in the example by clicking on the name of the gene (jag2). Another special feature of the Vega GeneView page is the fields for the authors of the annotation, and

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 OTTDART00000005844 to open the **TransView** page.



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 — The Wellcome Trust Sanger Institute

Home Dog Human Mouse ▸ Zebrafish ▲ BLAST ▲ Export Data ▲ Search ▲ Feedback ▲ Help ▲

Help on ExonView Find All Lookup [e.g. CH211-212G7.4-001, OTTDART00000004952]

Vega Exon Report

Transcript	**DKEY-5P1.1-001** (Vega_transcript ID)
Vega Transcript ID	**OTTDART00000005844**
Version	1
Class	Coding [Definition]
Transcript Information	This transcript is a product of Ensembl gene OTTDARG00000005397 [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

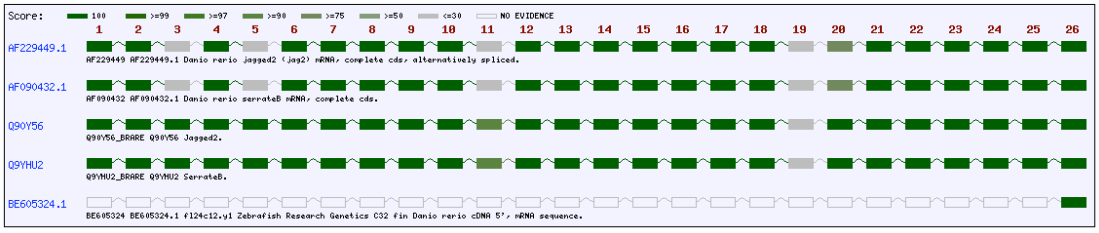
Exon Information

No.	Exon / Intron	Chr	Strand	Start	End	Length	Sequence
	5' upstream sequence						.....gagatctcctcaggcgtgtggtagccgtgaaacatatcatttgccgtgaa
1	OTTDARE00000033254	20	-1	16593319	16593534	216 bp	GTGATCAGACCGAGGGAGAGATCAGCACAGACCATCACCGCGCAAACACACCACGCTCGT GAATTTTTGCATGTCAGGAACGGAGGATCCTGTCGCCGGTCATCGGCCGTTTTCCATCTT GCGTTTATAACACATCAAATCGGCGGCATGTGGAATTGTATCAGGATTAGGAATTGGCTC CCAATCGCGTGCCTGCTGTTAACGATGTGGACGAAG
	Intron 1-2	20	-1	16592862	16593318	457 bp	gtgagagctctgattgttattagtg.....tttctgtttttcaatgtatttttag
2	OTTDARE00000033260	20	-1	16592541	16592861	321 bp	GTGTCCCAGTCCTCTGGCTATTTTGAGCTGCAGCTGATTGCTGTAGAAAATGTAAACGGT GAGTTGTGGGACGGGGAATGTTGCGACAGCACGCGGAACTCTCAAGACCAGCGCTGCGTG CGGGACGAGTGCGATACTTACTTTAAAGTGTGTCTGAAGGAGTACCAGTCTGAAGTCACC ACCACTGGACAGTGCACCTTCGGCTCTGGATCTACCGACGTTCTTGGTGGAAATATAATT TCTTTTAAGACCGCAAAAAACAGCCCCAGCAAAACGAGCGACGTGGGAAAGATCATCATC CCTTTTCACTTCGCCTGGCCG
	Intron 2-3	20	-1	16565147	16592540	27394 bp	gtgagtcctcgtcttctccatgggc.....gttattctctctctttctctcttag
3	OTTDARE00000033271	20	-1	16565092	16565146	55 bp	CGATCCTACACACTCATCCTTGAAGCTTGGGACTGGGATAACTCCACTCAGAACA

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

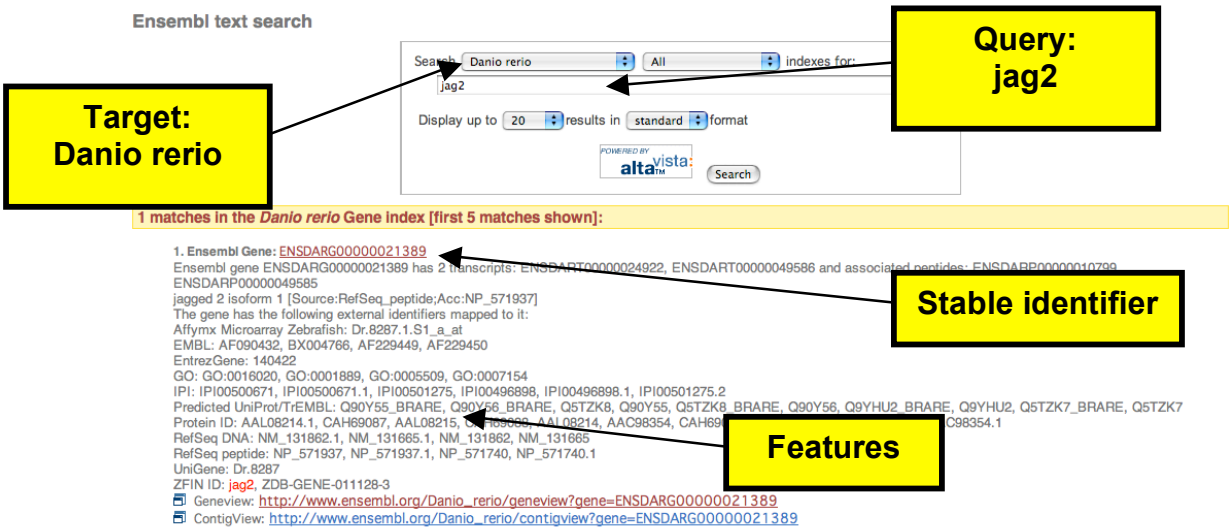
Introduction

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.

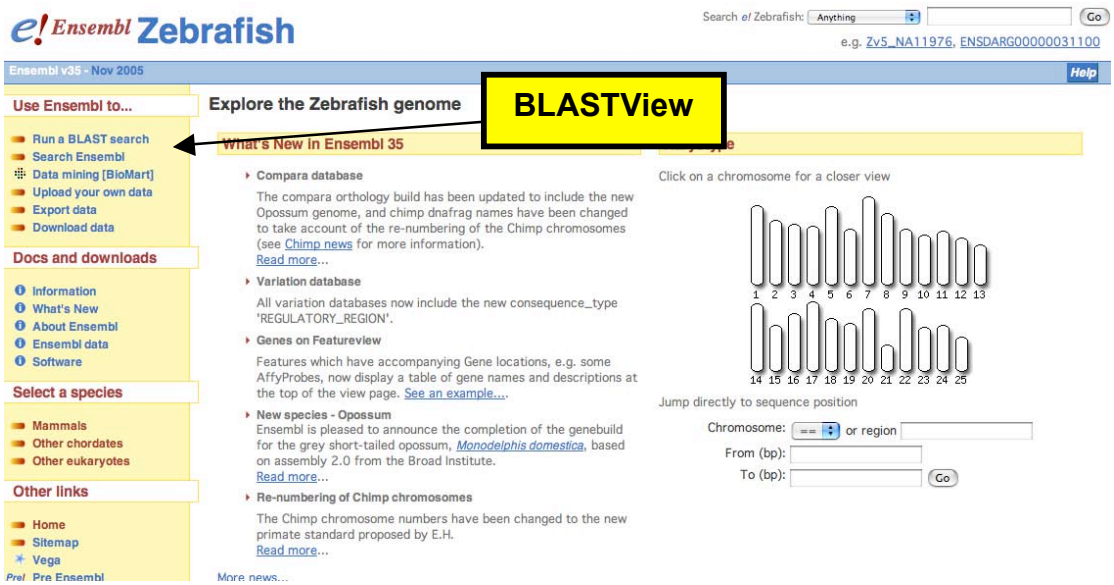


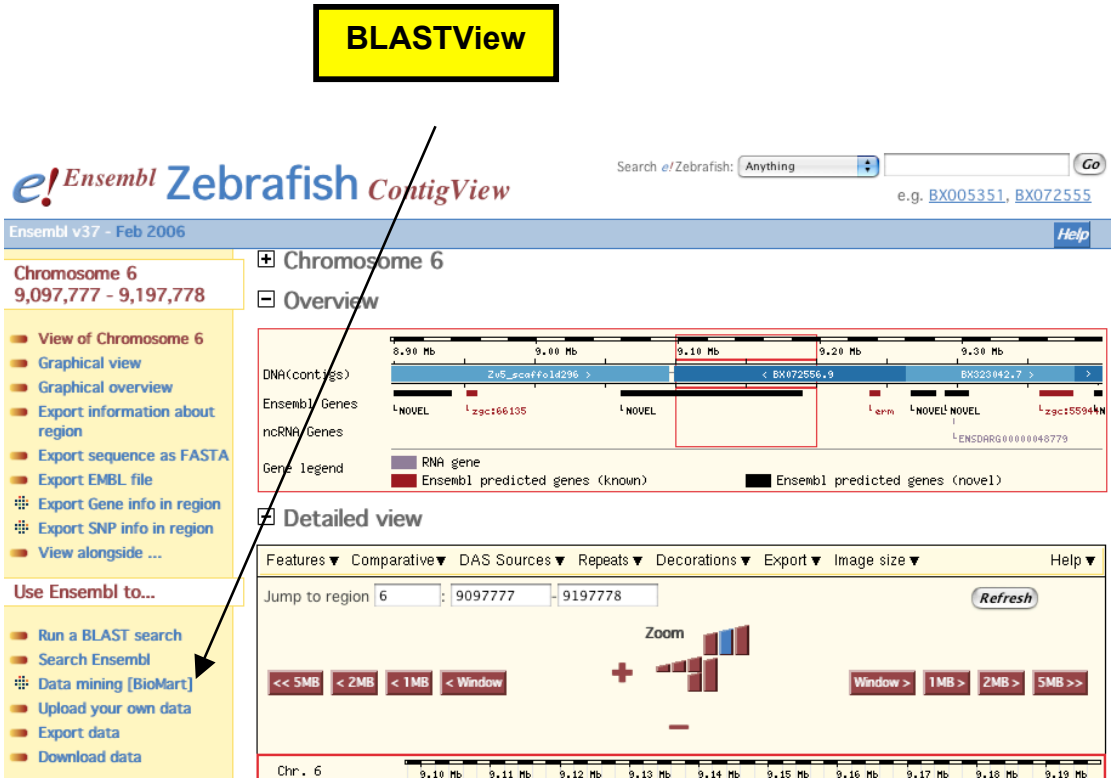
A **TextView** page summarises the result of a text-based search. If the query appears under different indices then the TextView page organises the results in categories. For example the page below corresponds to the result of searching for jag2 in Vega:

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 searched 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

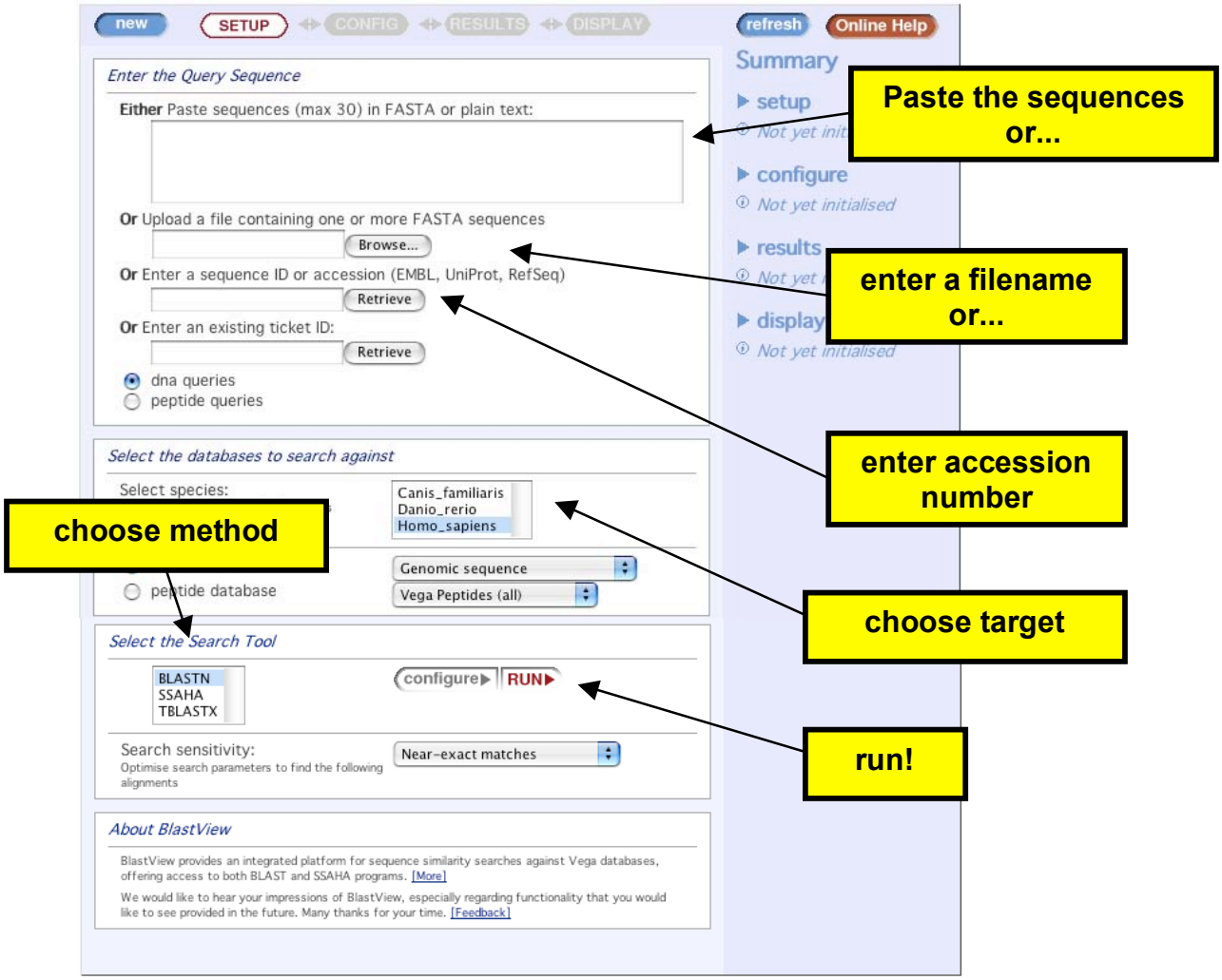
The Vega/Ensembl browsers provide a page where you can search using different sequences to query the databases. You can access the BLASTView page from any of the Vega/Ensembl views through the link labelled 'run a BLAST search' in Ensembl or 'BLAST' in Vega.





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.

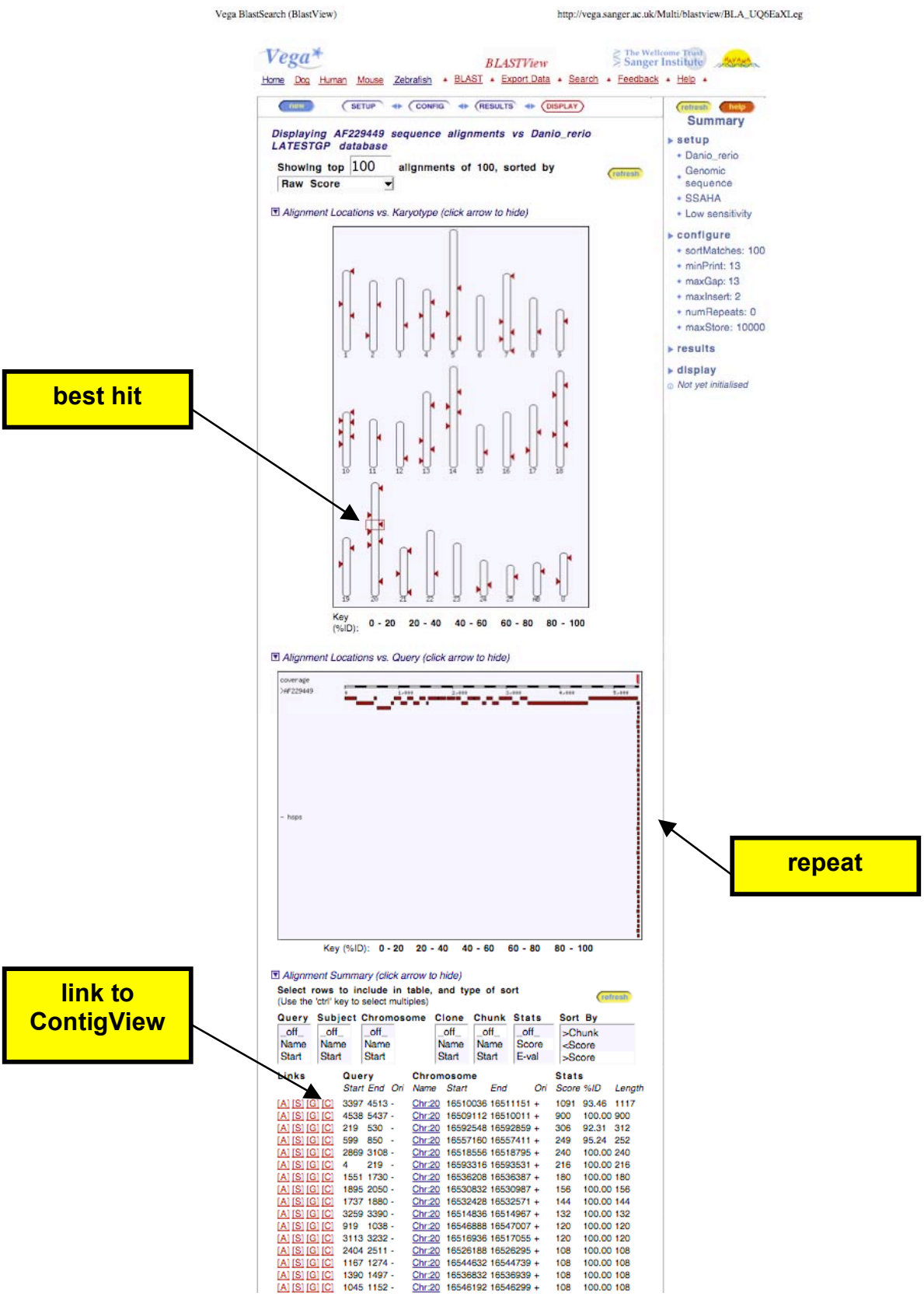
Open a BLASTView page in the Vega browser. In order to specify the query for the search you have the option of either using the sequence(s) or using an EMBL/GenBank accession number.



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:

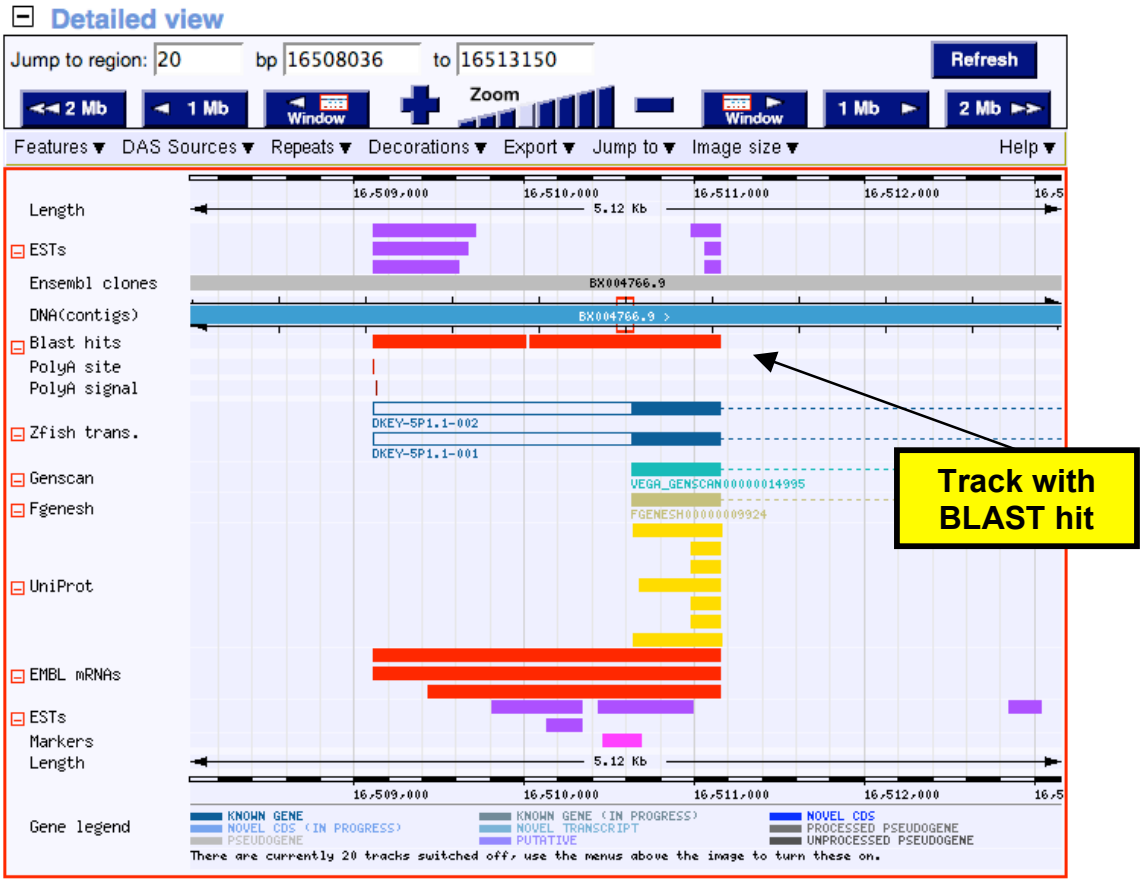
- BLASTN - the well-known BLAST algorithm performing a DNA-DNA search.
- SSAHA - this tool runs a hash-based algorithm. It is very fast since most of the needed data structures are pre-loaded. It works very well when searching for near-exact matches. It only performs searches where the query is DNA.
- 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:



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

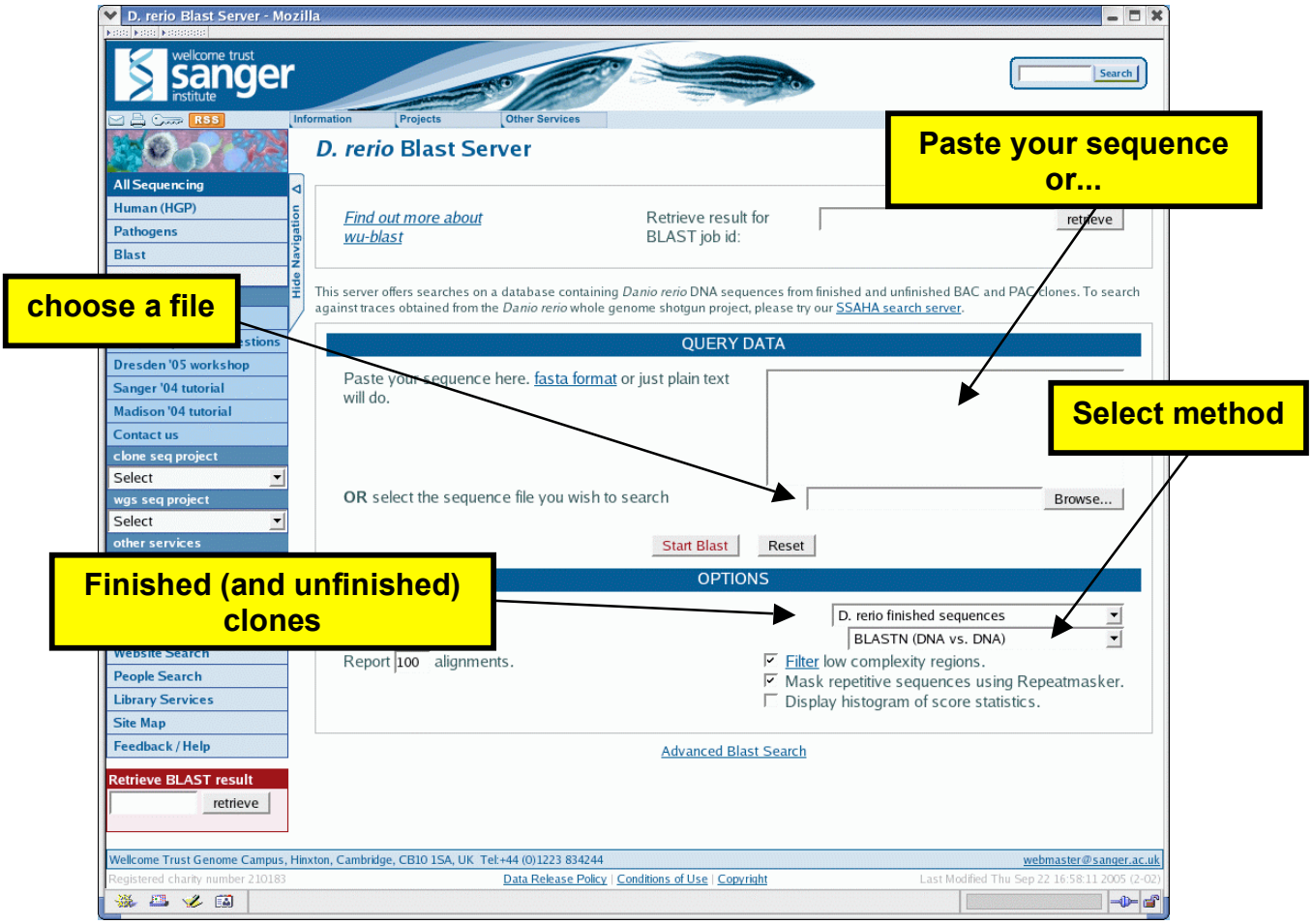
The Ensembl database contains the latest zebrafish assembly with automatic annotation (currently version Zv6 – March 2006). The zebrafish assembly is obtained by integrating all the available sequenced clones with a whole genome shotgun assembly. When reading and interpreting the outcome of a search 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 though the *Danio rerio* project page or directly at:

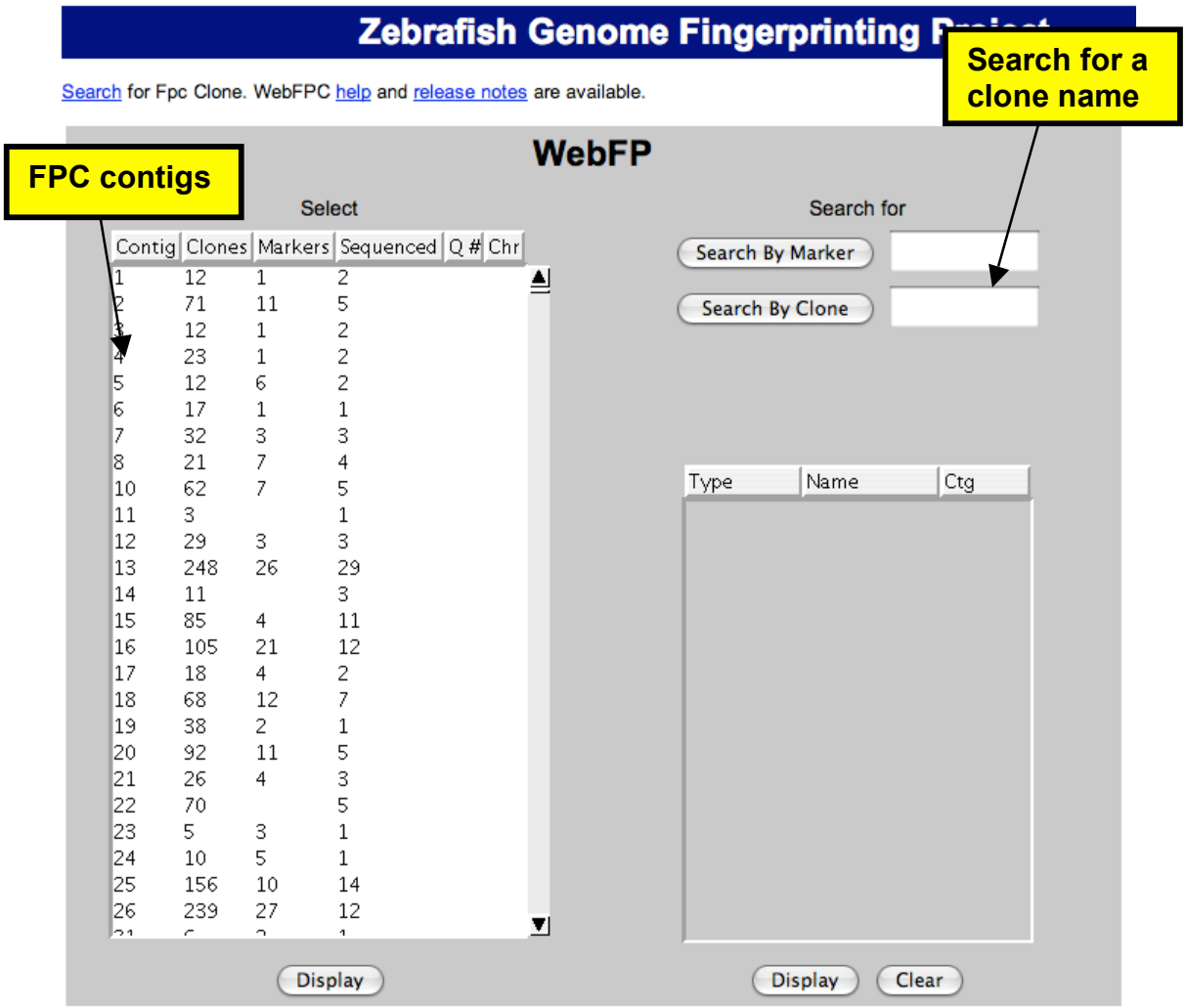
http://www.sanger.ac.uk/cgi-bin/blast/submitblast/d_rerio



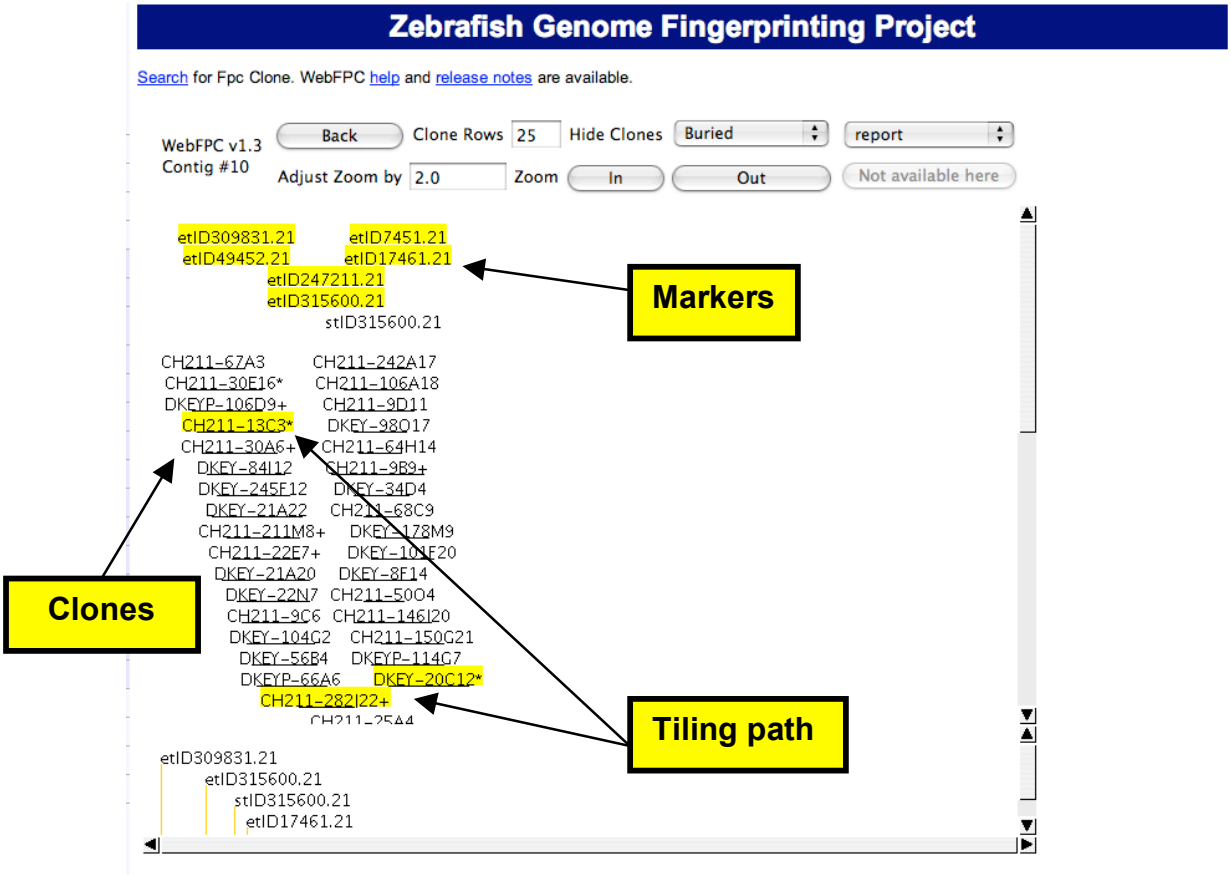
This search can be used to find out whether a clone containing your region of interest is covered by a sequenced clone. An unfinished clone might be submitted in several contigs with artificial gaps between them. Contigs from unfinished clones can be long enough to contain a gene but they will not present in Vega until they are properly finished.

This collection of all sequenced clones does not replace the assembly since it is incomplete and also lacks all the extra features like alignments and automatic gene predictions. Moreover the sequences in this collection are isolated without a tiling path or contextual information. If you want to learn more about a clone and its flanking regions you can query the FPC database at:

http://www.sanger.ac.uk/Projects/D_rerio/WebFPC/zebrafish/small.shtml



FPC contig 10, for example, contains 62 clones and 5 have been sequenced. Click on this FPC contig to see more information:



Trace repository

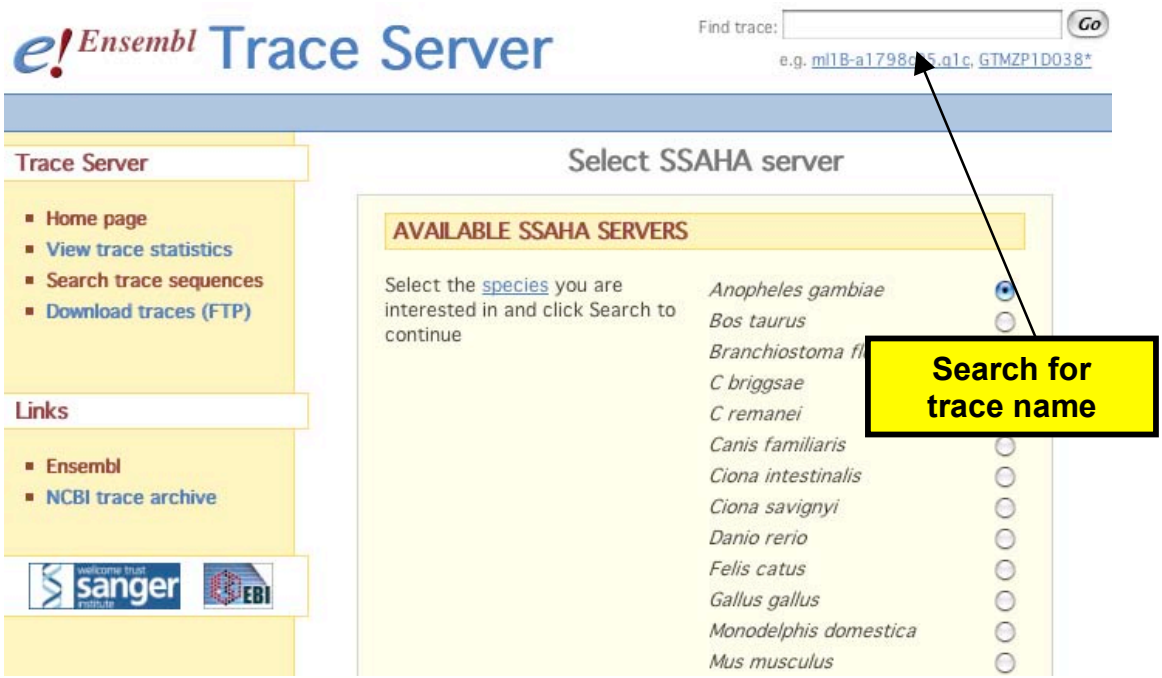
The whole genome shotgun assembly used to build the zebrafish assembly is based on a collection of reads from cosmids, fosmids and BAC ends. This collection currently gives around 7x coverage of the genome. All these reads are deposited into the trace repository at

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

GCCCTTAACTGATCGGCTTCTTCAGCAAGAGAGTTGCAAAGTACAG

to search in Ensembl using SSAHA. Where did you find the best match? Try using the same sequence with BLAST. Why do 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://ww.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:

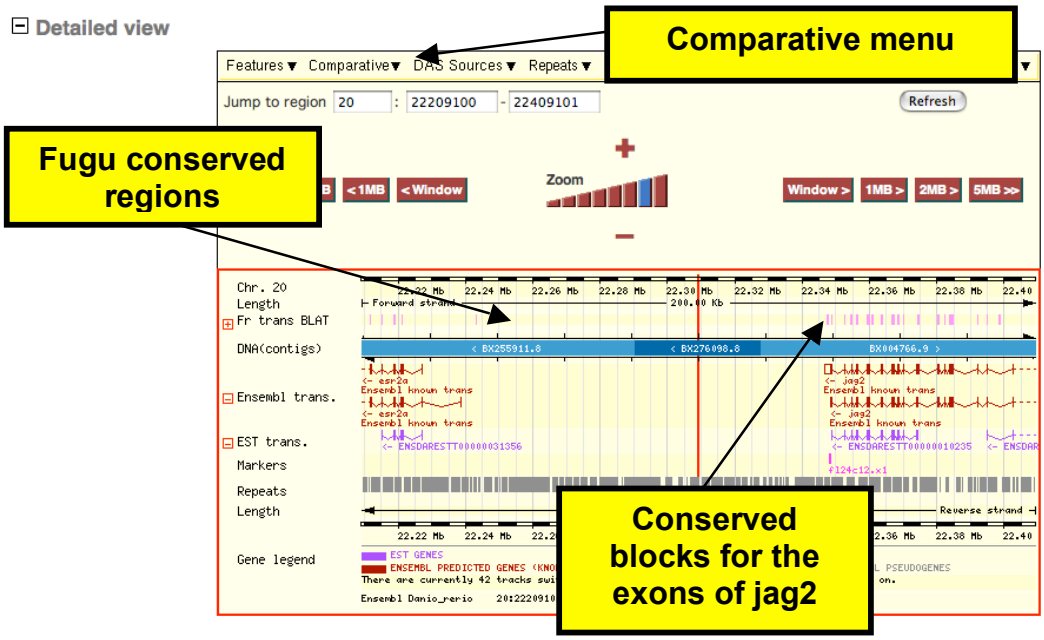
- whole genome alignments
- gene orthology/paralogy prediction
- protein clustering
- synteny regions (not available for zebrafish)

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

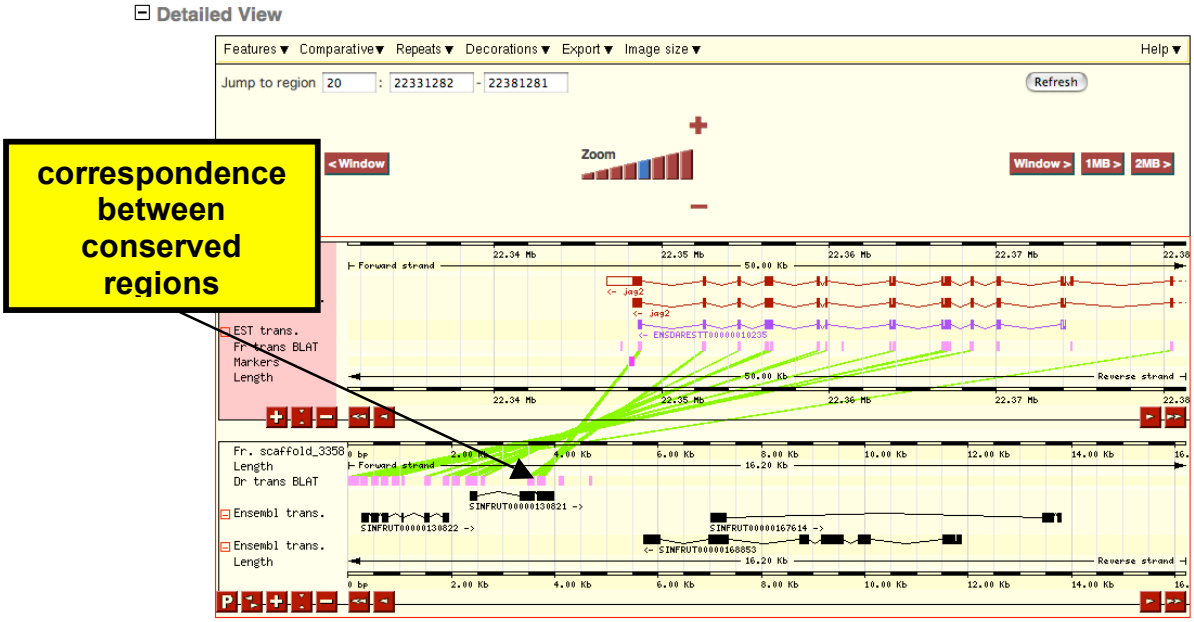
The alignment of the whole DNA sequence from two organisms is computationally demanding. Such data are of great interest both in studies of the mechanisms of molecular evolution and in attempts to identify conserved functional sequences such as novel genes and regulatory regions. Whole genome alignments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



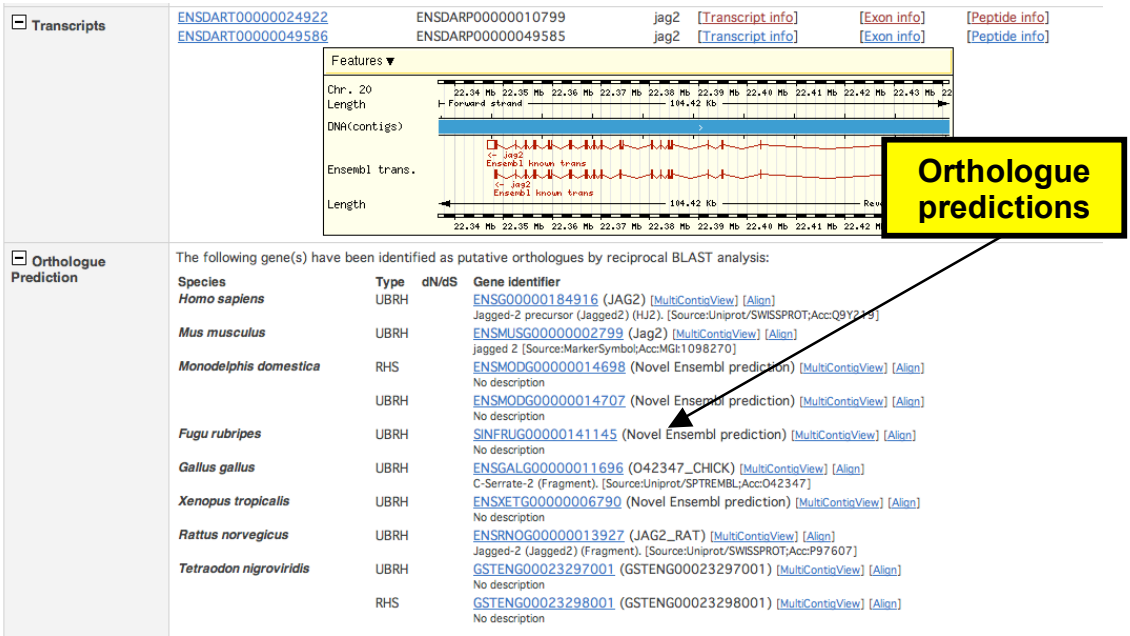
A **MultiContigView** allows the visualisation of syntenic regions from multiple species. In the example above a region from chromosome 20 in zebrafish is compared to scaffold_3358 in the Fugu assembly. The conserved blocks are connected with green lines. Observe that in this example the exons from jag2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

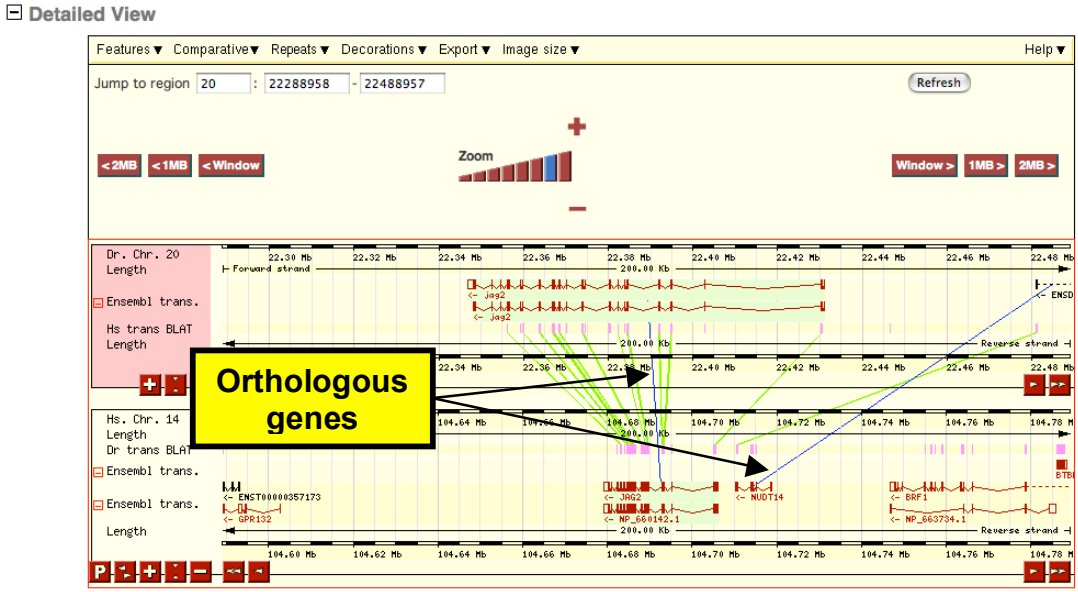
Another kind of comparative analysis focuses on genes and proteins, and attempts to identify orthologues in different genomes. The classic 'model' animals are now all represented in Ensembl (*Drosophila, C. elegans*, mouse) as well as zebrafish. The automated identification of orthologues is made more 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 linage 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.



In Ensembl, orthologues are identified starting with comparisons at the protein level. 'All-versus-all' BLASTP+SW (Smith-Waterman algorithm) is first used to identify those protein pairs that are best reciprocal hits (BRH) between two sets of proteins that represent every gene in the two organisms. Additional putative orthologues are then sought using synteny and these are known as RHS (Reciprocal Hit supported by Synteny). Where two homologous proteins are encoded by genes each located within 1 Mb of a pair of BRH, they are

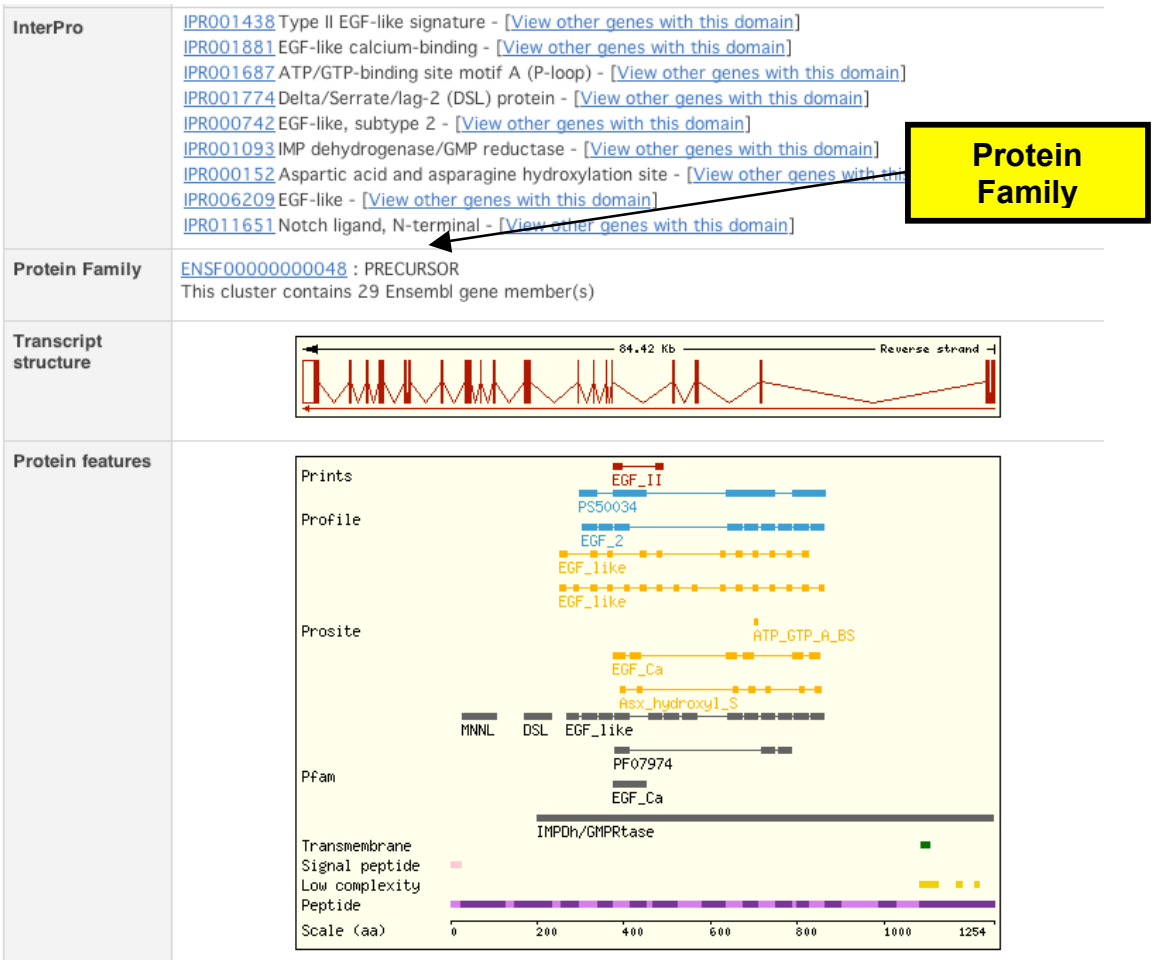
good candidates for being an additional orthologous pair. Currently we divide these BRH into UBRH (Unique Best Reciprocal Hit) and MBRH (Multiple Best Reciprocal Hit). The latter have multiple but identical best hits, which can happen if there is perfect protein sequence duplication of translated genes within a species. The same approach permits the identification of adjacent family members that may be recently duplicated lineage-specific paralogues. For every orthologue prediction there is a link to a MultiContigView page. Follow the link labelled 'MultiContigView' for the human prediction:
Jag2 in zebrafish corresponds to JAG2 in human. These genes are used to anchor the region. Orthologous genes are indicated by a blue line and conserved blocks are connected by green lines. Observe that there is also a link between two putative orthologous genes on the left.



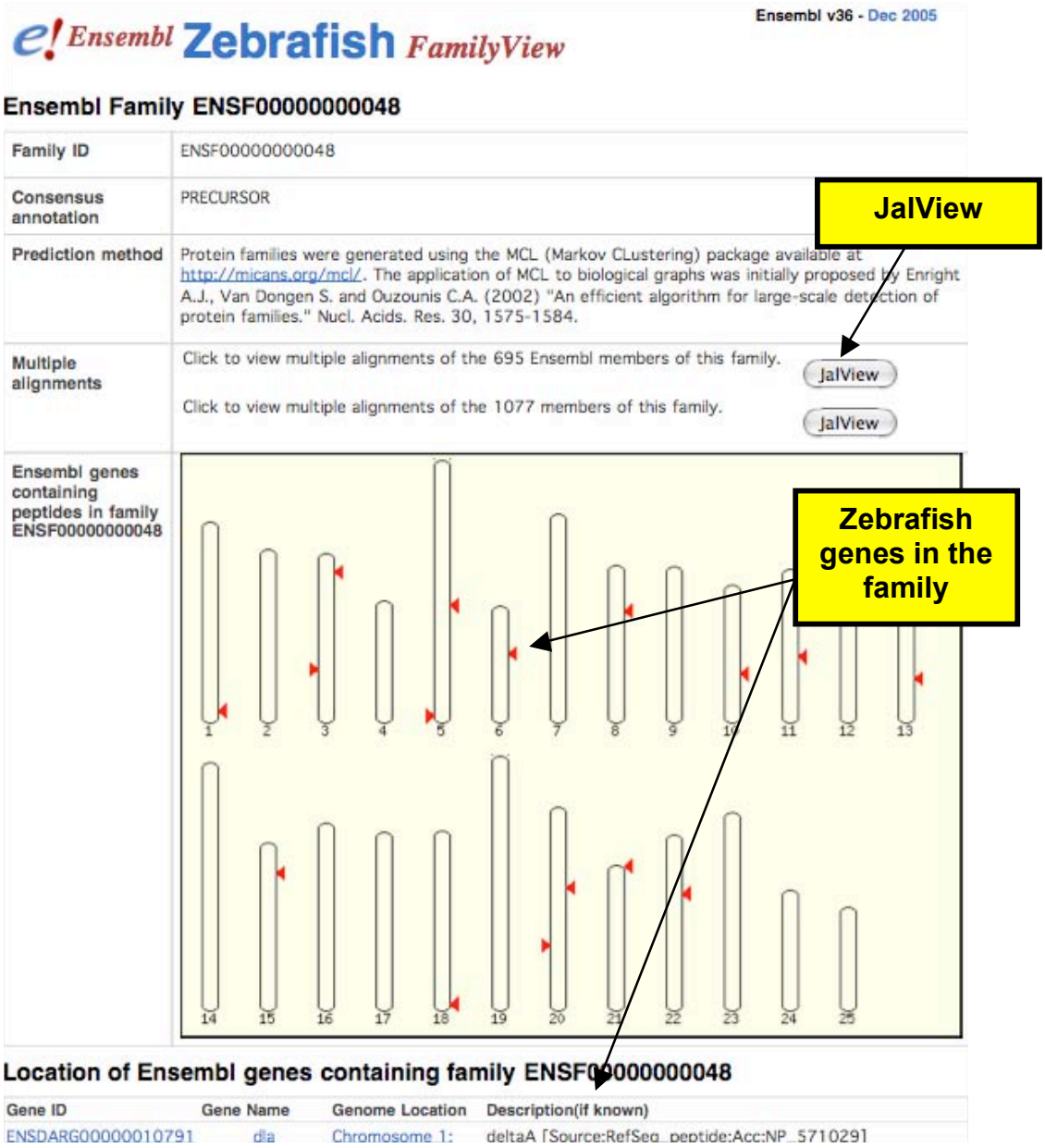
Protein Families

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

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



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



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

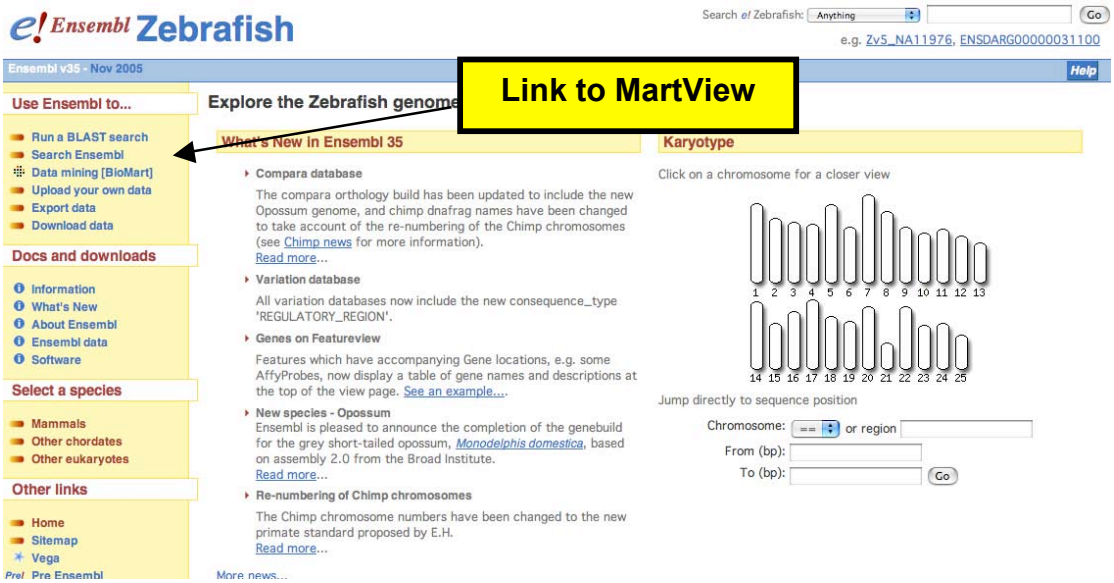
- Introduce BioMart, a data mining system for large datasets

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. EnsMart 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:

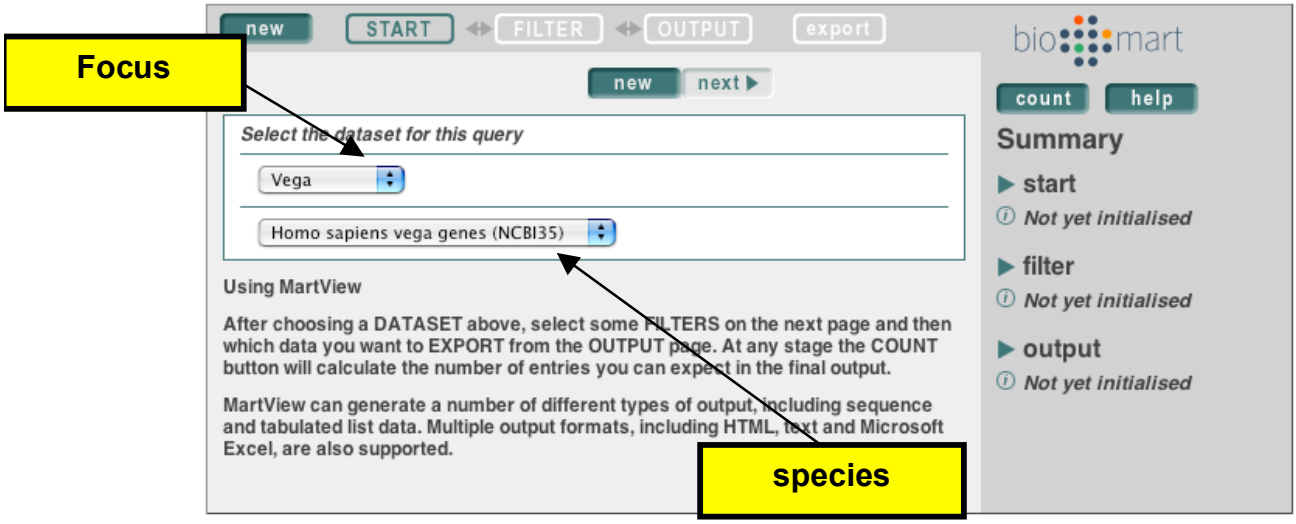


Queries in BioMart are organised into three steps: **start**, **filter** and **output**. The user can navigate between these three stages using the 'Back' and 'Next'

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'.

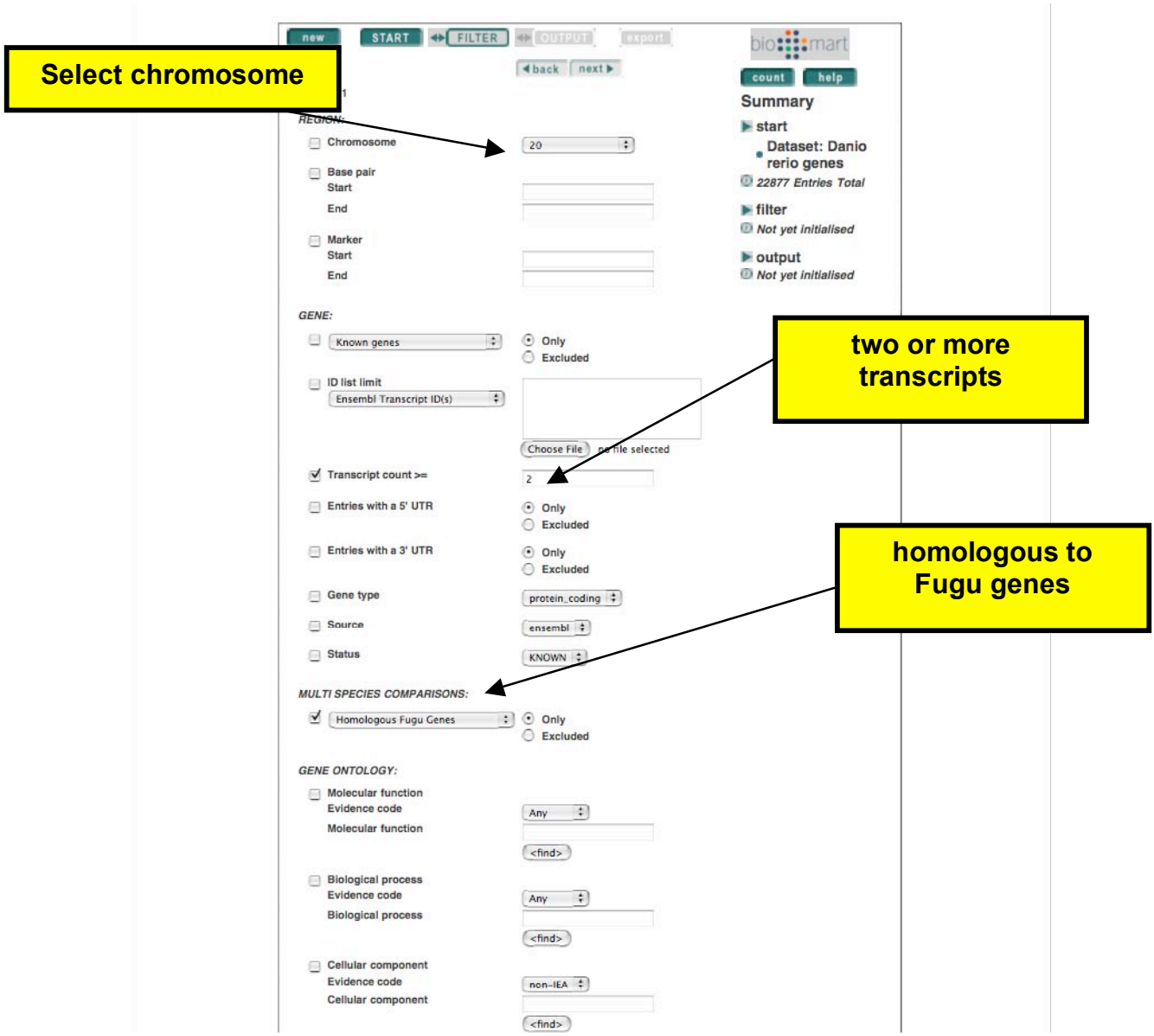


Filter

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

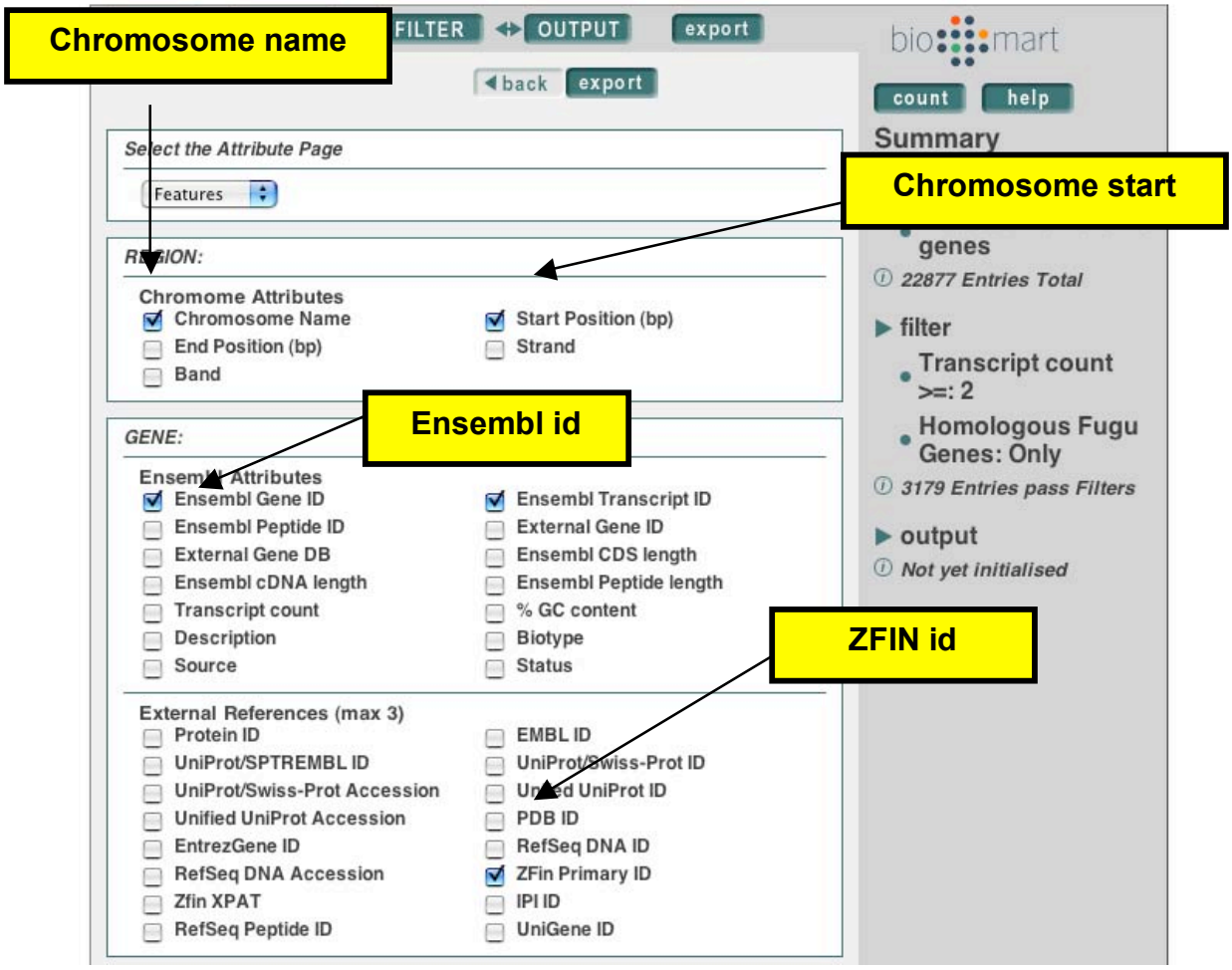


For each filter a MartView user can define whether the criteria should be satisfied or not. Click next to advance to the next stage.

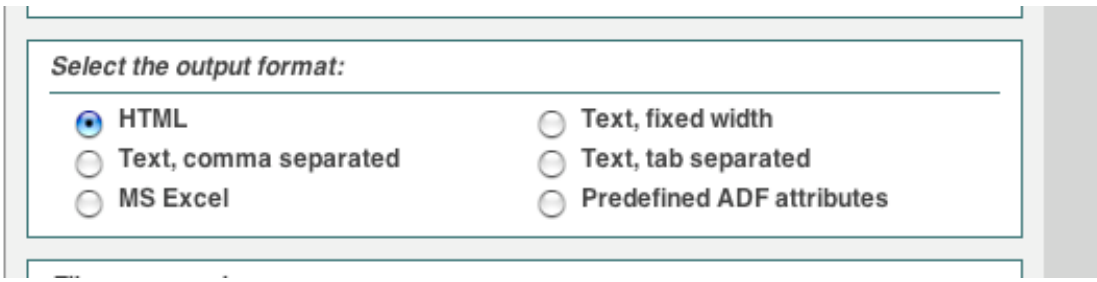
Output

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



- and finally the output format is HTML



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.