Module 2 - Maps and Genome Sequence

iv – 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

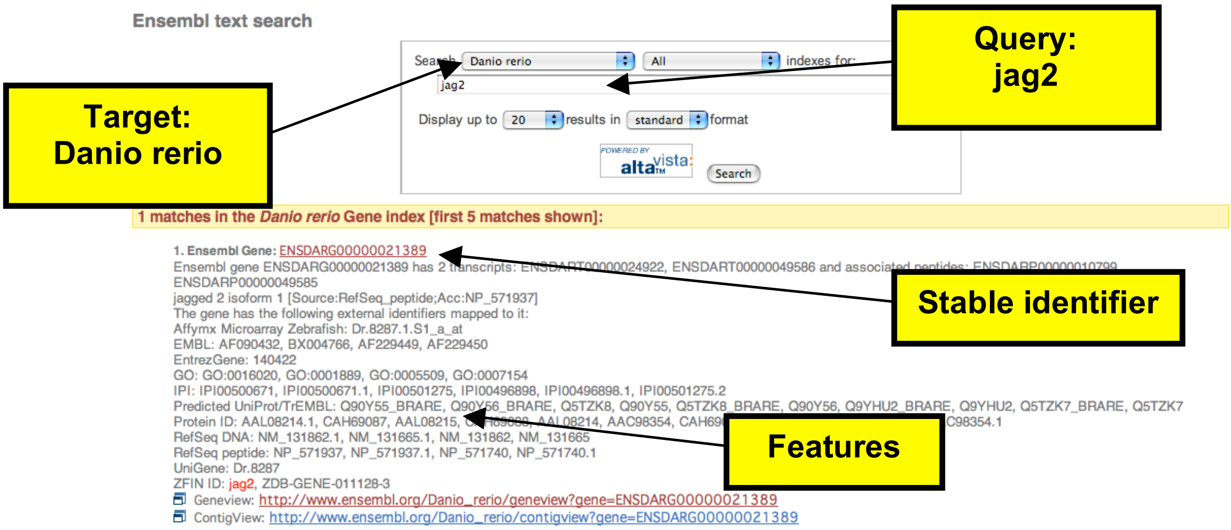
Introduction

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.



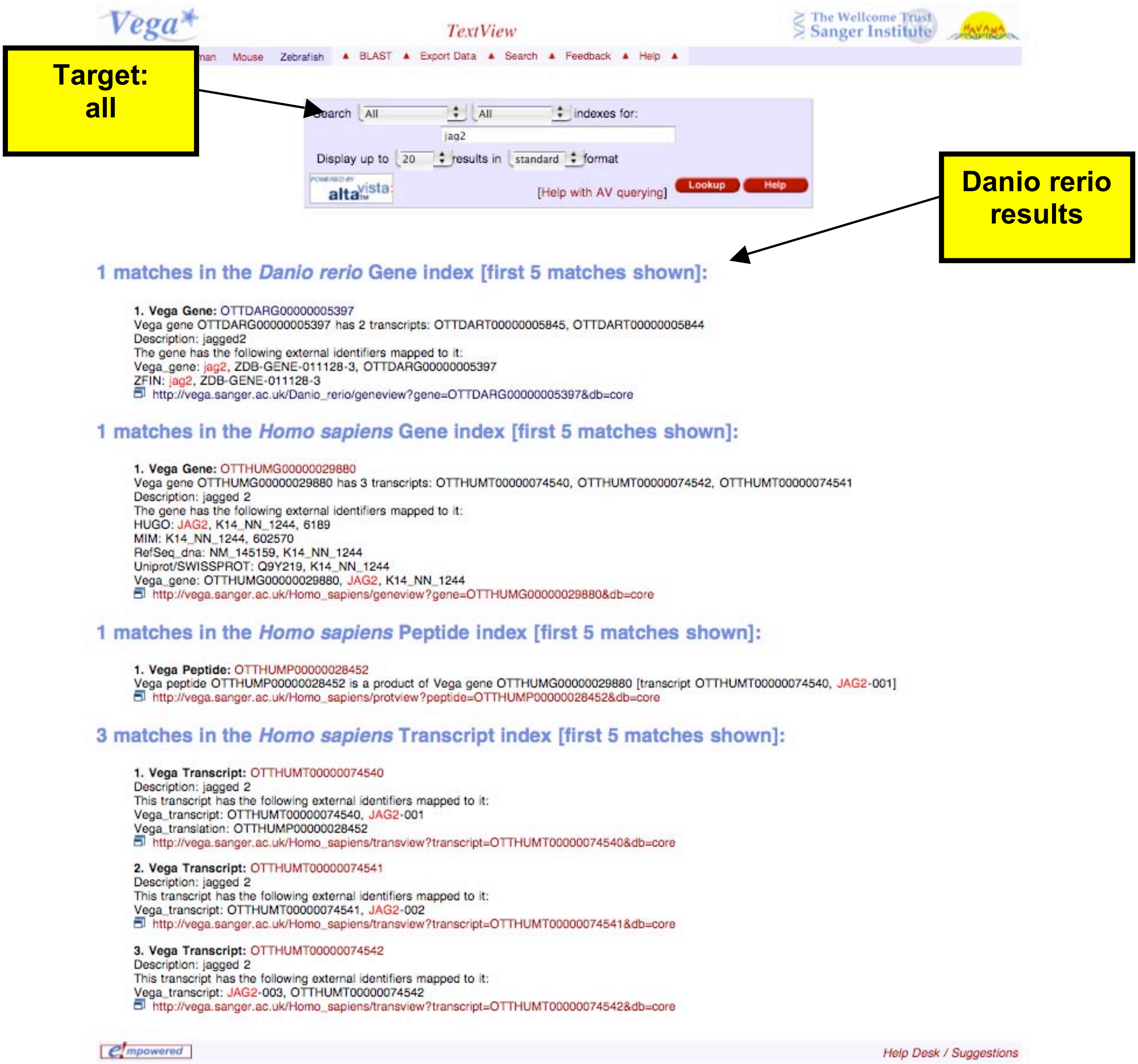
Ensembl text search

Search Danio rerio All indexes for:
jag2
Display up to 20 results in standard format

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: ENSDARP00000010799, 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, CAH69088, AAL08214, AAC98354, CAH69 C98354.1
RefSeq DNA: NM_131862.1, NM_131665.1, NM_131862, NM_131665
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

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:


Vega
TextView
The Wellcome Trust Sanger Institute
HAVANA
man Mouse Zebrafish BLAST Export Data Search Feedback Help
Search All All indexes for:
jag2
Display up to 20 results in standard format
altavista
[Help with AV querying]
Lookup
Help
Target: all
Danio rerio results
empowered
Help Desk / Suggestions

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

1. **Vega Gene:** OTTDARG00000005397
Vega gene OTTDARG00000005397 has 2 transcripts: OTTDART00000005845, OTTDART00000005844
Description: jagged2
The gene has the following external identifiers mapped to it:
Vega_gene: jag2, ZDB-GENE-011128-3, OTTDARG00000005397
ZFIN: jag2, ZDB-GENE-011128-3
http://vega.sanger.ac.uk/Danio_rerio/geneview?gene=OTTDARG00000005397&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

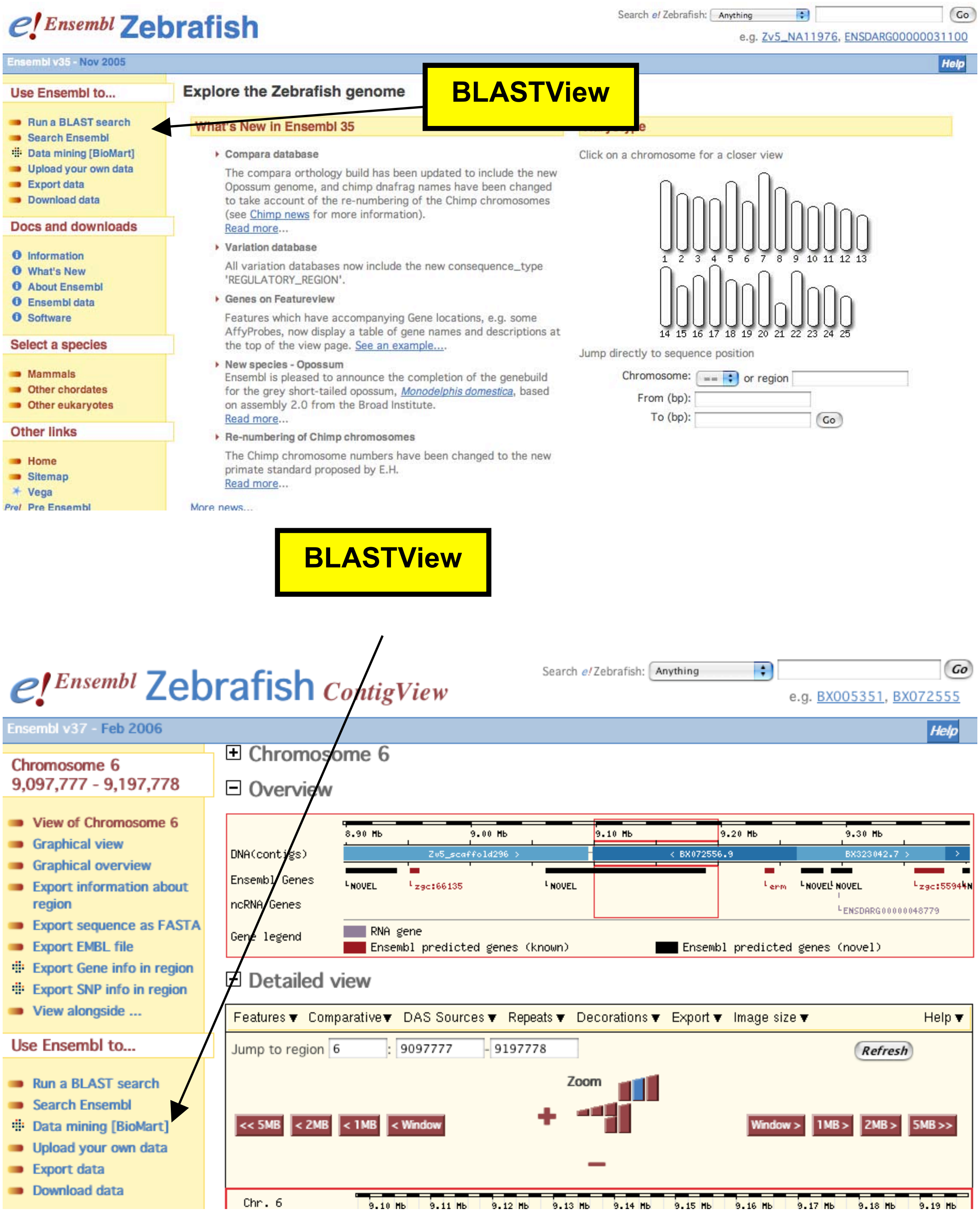
2. **Vega Transcript:** OTTHUMT00000074541
Description: jagged 2
This transcript has the following external identifiers mapped to it:
Vega_transcript: OTTHUMT00000074541, JAG2-002
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000074541&db=core

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

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.



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'.

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

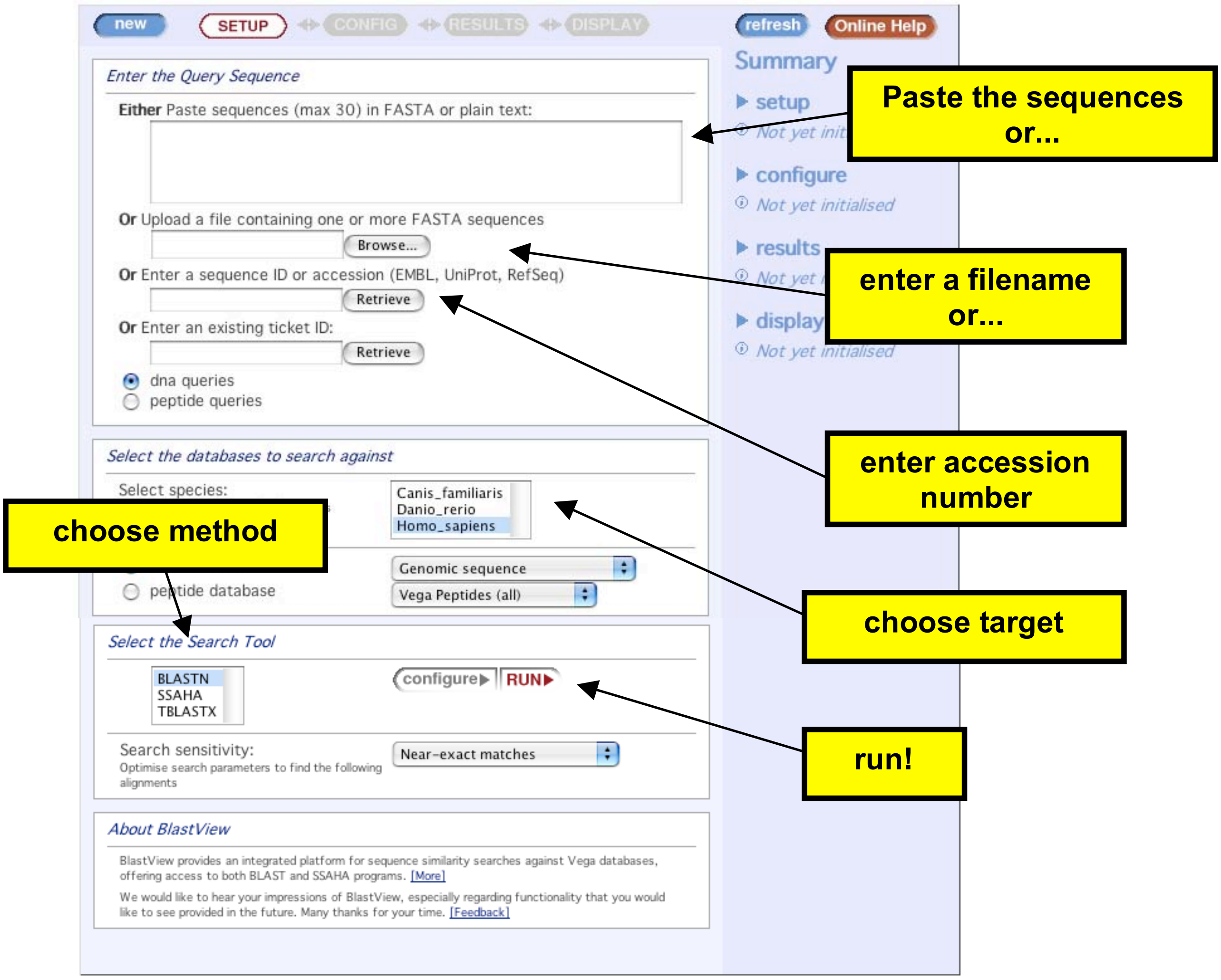
Jump directly to sequence position

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

Jump to region 6 : 9097777 - 9197778

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 Query Sequence

Either Paste sequences (max 30) in FASTA or plain text:

Or Upload a file containing one or more FASTA sequences

Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)

Or Enter an existing ticket ID:

● dna queries
○ peptide queries

Select the databases to search against

Select species: Canis_familiaris, Danio_rerio, Homo_sapiens

Genomic sequence

○ peptide database: Vega Peptides (all)

Select the Search Tool

BLASTN
SSAHA
TBLASTX

Search sensitivity: Near-exact matches

Optimise search parameters to find the following alignments

About BlastView

BlastView provides an integrated platform for sequence similarity searches against Vega databases, offering access to both BLAST and SSAHA programs. [More]

We would like to hear your impressions of BlastView, especially regarding functionality that you would like to see provided in the future. Many thanks for your time. [Feedback]

Paste the sequences or...

enter a filename or...

enter accession number

choose method

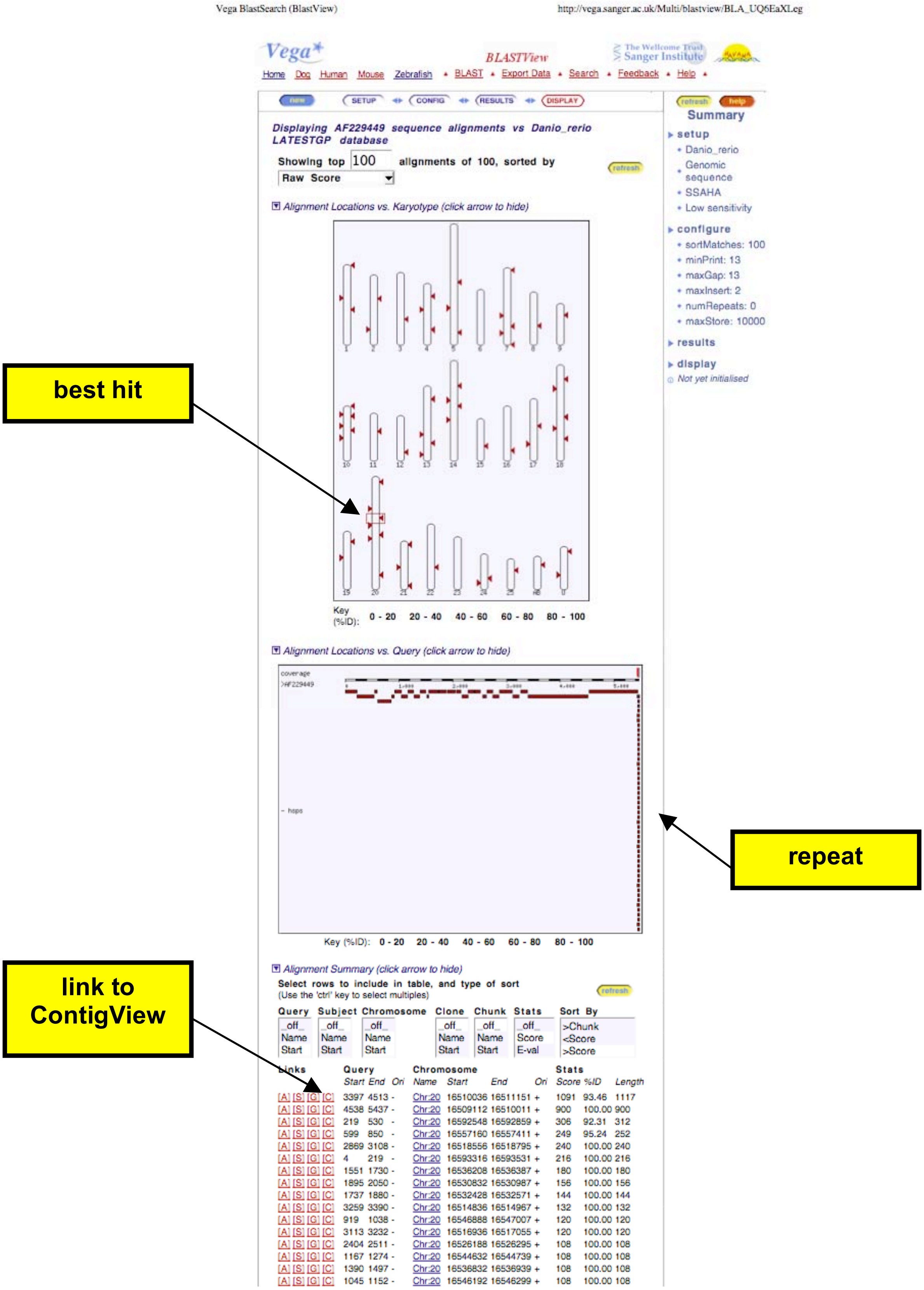
choose target

run!

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_UQ6EaXLeg
Vega
BLASTView
The Wellcome Trust Sanger Institute
Home Dog Human Mouse Zebrafish • BLAST • Export Data • Search • Feedback • Help •
new SETUP CONFIG RESULTS DISPLAY
refresh help
best hit
repeat
link to ContigView
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 26 U
coverage
>AF229449
hsps
Key (%ID): 0 - 20 20 - 40 40 - 60 60 - 80 80 - 100

Summary
- setup
  - Danio_rerio
  - Genomic sequence
  - SSAHA
  - Low sensitivity
- configure
  - sortMatches: 100
  - minPrint: 13
  - maxGap: 13
  - maxInsert: 2
  - numRepeats: 0
  - maxStore: 10000
- results
- display
  - Not yet initialised

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)

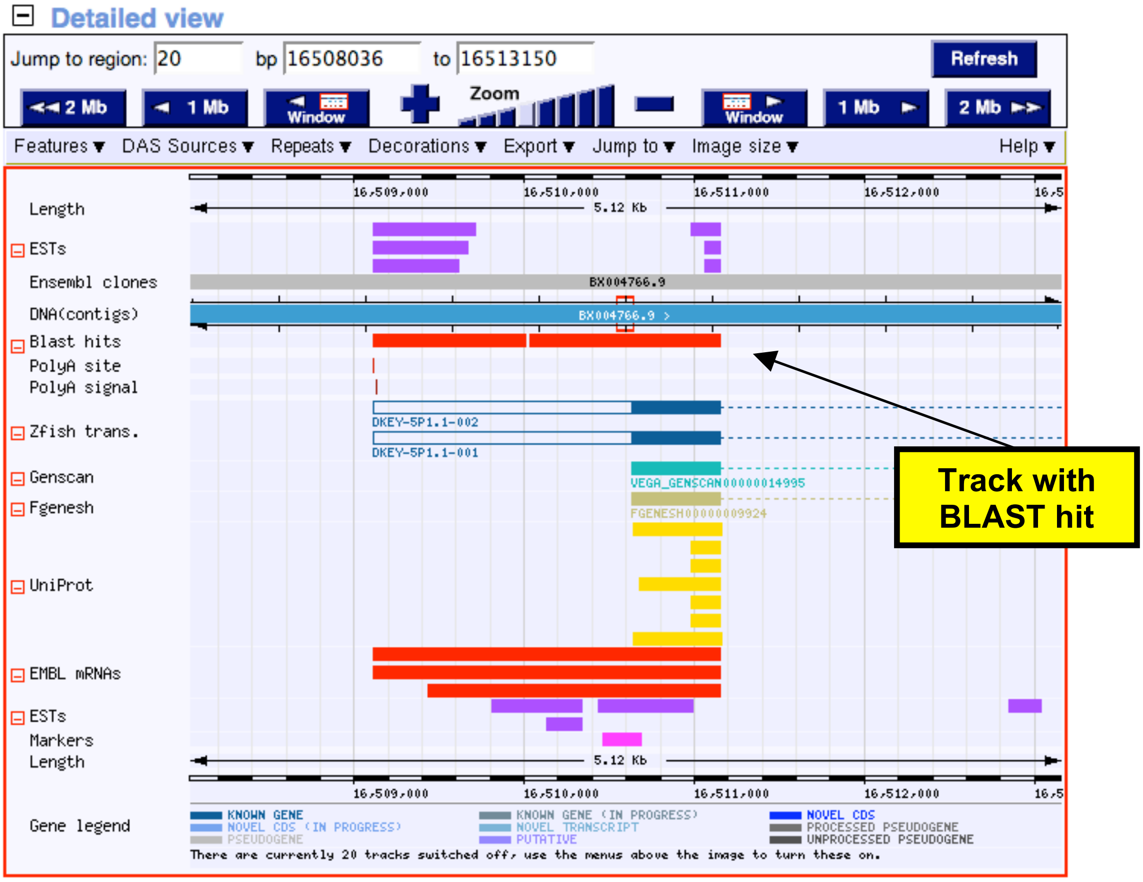
Alignment Summary (click arrow to hide)

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 Name Start	_off_ Name Start	_off_ Name Start	_off_ Name Start	_off_ Name Start	_off_ Score E-val	>Chunk <Score >Score

Links	Query			Chromosome				Stats		
	Start	*End*	*Ori*	*Name*	*Start*	*End*	*Ori*	*Score*	*%ID*	*Length*
[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.



Jump to region: 20 bp 16508036 to 16513150

There are currently 20 tracks switched off, use the menus above the image to turn these on.

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

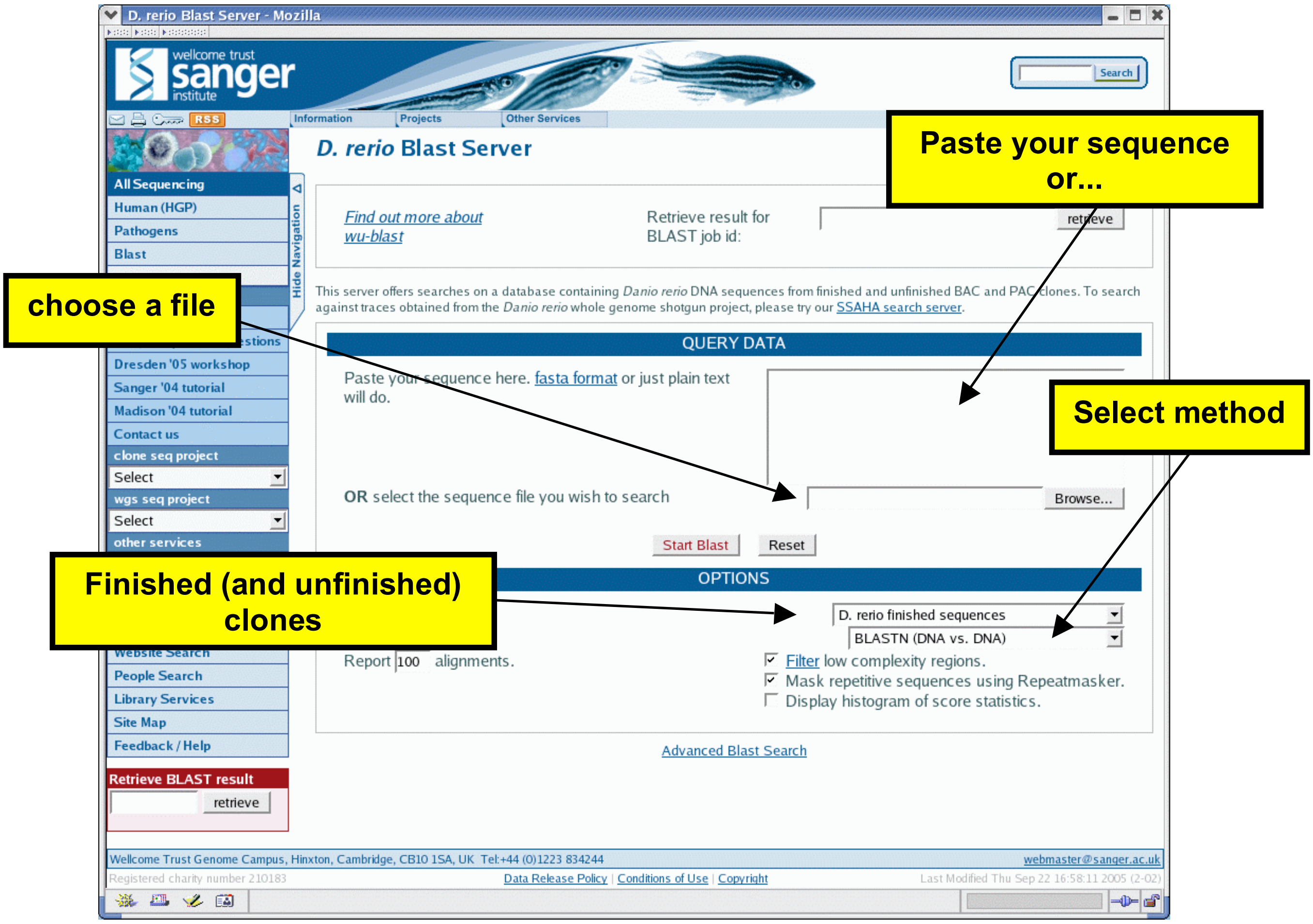
The Ensembl database contains the latest zebrafish assembly with automatic annotation (currently version Zv6 – March 2006). The zebrafish assembly is obtained by integrating all the available sequenced clones with a whole genome shotgun assembly. When reading and interpreting the outcome of a search 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



D. rerio Blast Server

Retrieve result for BLAST job id:

This server offers searches on a database containing *Danio rerio* DNA sequences from finished and unfinished BAC and PAC clones. To search against traces obtained from the *Danio rerio* whole genome shotgun project, please try our SSAHA search server.

Paste your sequence here. fasta format or just plain text will do.

OR select the sequence file you wish to search

D. rerio finished sequences

BLASTN (DNA vs. DNA)

Report 100 alignments.

☑ Filter low complexity regions.
☑ Mask repetitive sequences using Repeatmasker.
☐ Display histogram of score statistics.

Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK Tel:+44 (0)1223 834244

Registered charity number 210183

Last Modified Thu Sep 22 16:58:11 2005 (2-02)

Paste your sequence or...

choose a file

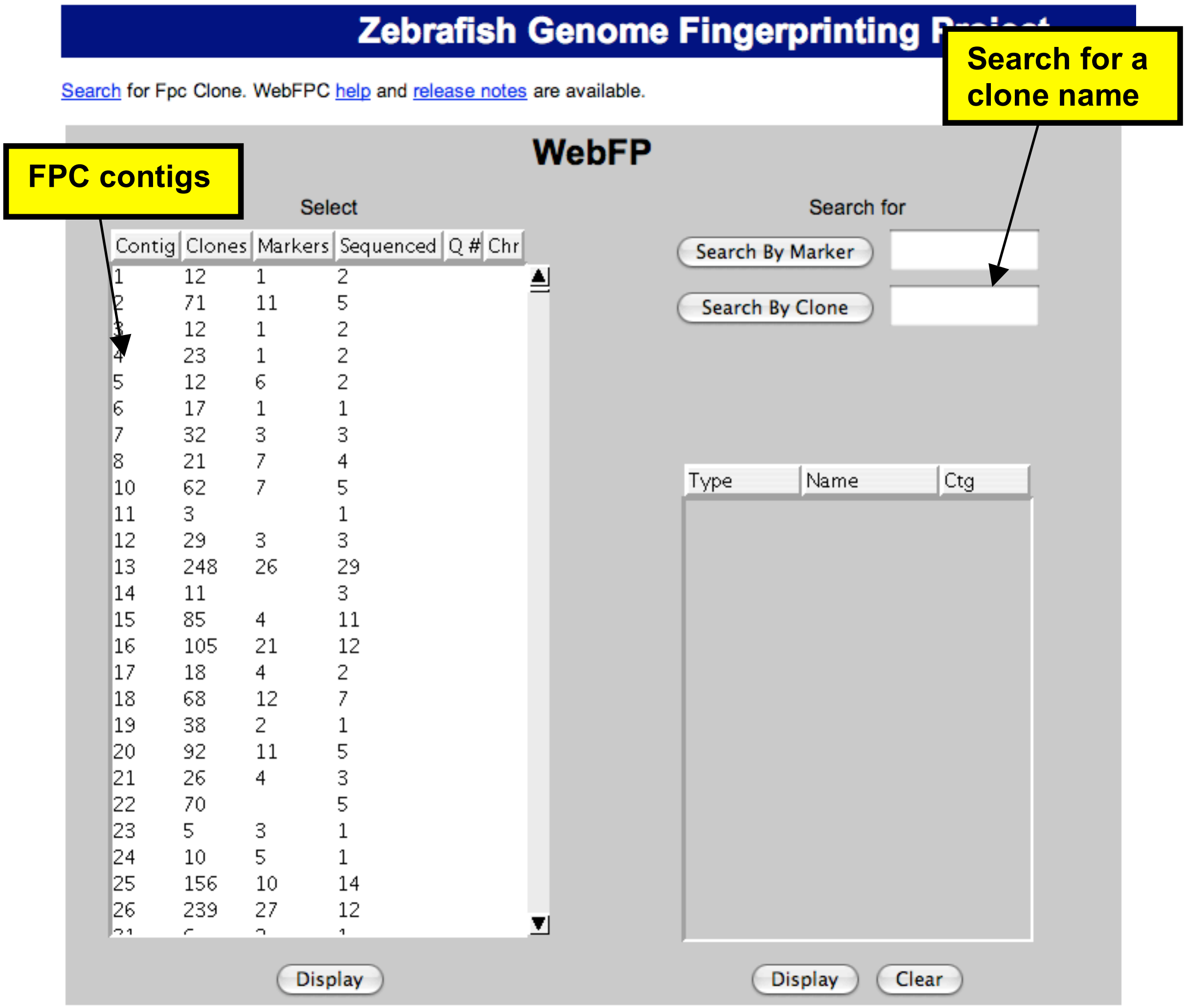
Select method

Finished (and unfinished) clones

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



Zebrafish Genome Fingerprinting

Search for Fpc Clone. WebFPC help and release notes are available.

WebFP

Select

Contig	Clones	Markers	Sequenced	Q #	Chr
1	12	1	2		
2	71	11	5		
3	12	1	2		
4	23	1	2		
5	12	6	2		
6	17	1	1		
7	32	3	3		
8	21	7	4		
10	62	7	5		
11	3		1		
12	29	3	3		
13	248	26	29		
14	11		3		
15	85	4	11		
16	105	21	12		
17	18	4	2		
18	68	12	7		
19	38	2	1		
20	92	11	5		
21	26	4	3		
22	70		5		
23	5	3	1		
24	10	5	1		
25	156	10	14		
26	239	27	12		

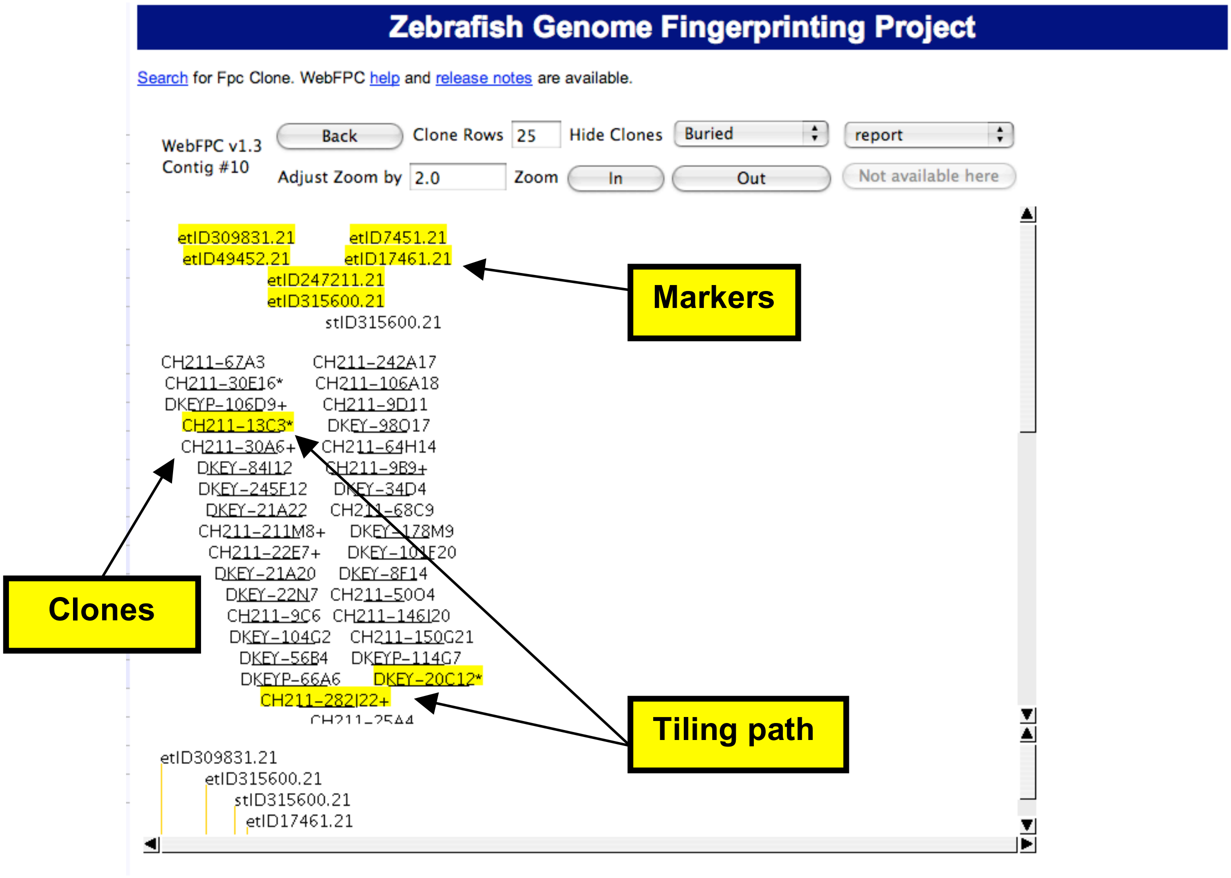
Search for

Type	Name	Ctg

FPC contigs

Search for a clone name

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



Zebrafish Genome Fingerprinting Project

Search for Fpc Clone. WebFPC help and release notes are available.

WebFPC v1.3
Contig #10

Clone Rows 25 · Hide Clones Buried · report · Adjust Zoom by 2.0

Trace repository

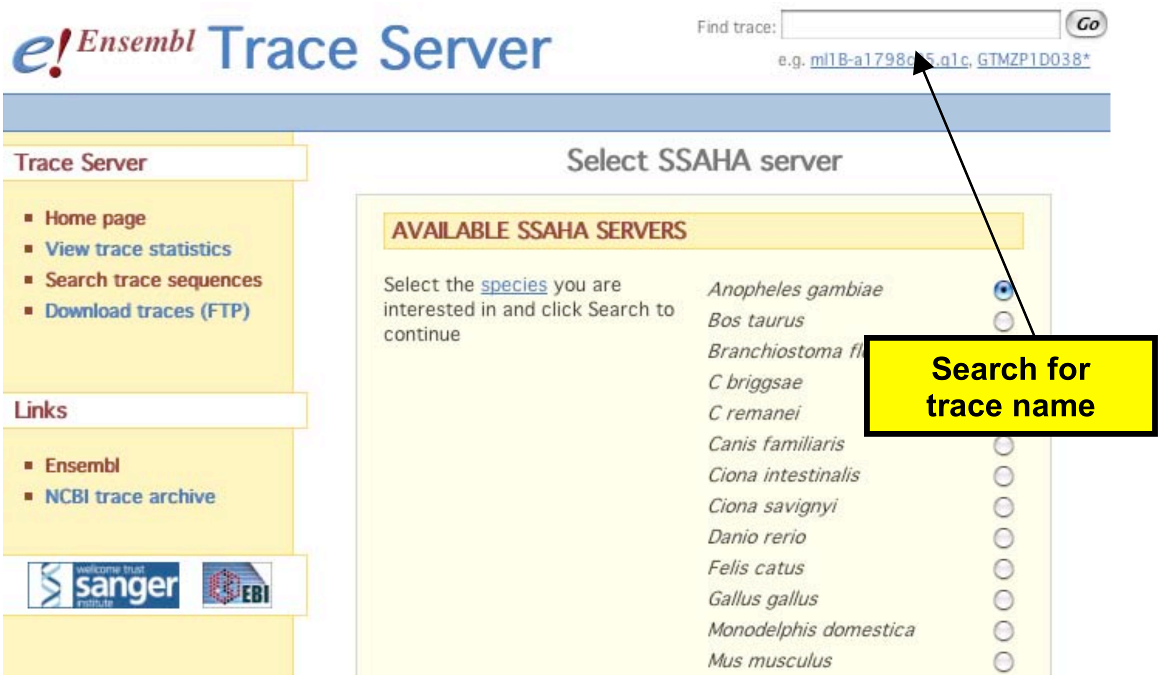
The whole genome shotgun assembly used to build the zebrafish assembly is based on a collection of reads from cosmids, fosmids and BAC ends. This collection currently gives around 7x coverage of the genome. All these reads are deposited into the trace repository at

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.



Select SSAHA server

AVAILABLE SSAHA SERVERS

Select the species you are interested in and click Search to continue

- *Anopheles gambiae* ●
- *Bos taurus* ○
- *Branchiostoma fl*
- *C briggsae*
- *C remanei*
- *Canis familiaris* ○
- *Ciona intestinalis* ○
- *Ciona savignyi* ○
- *Danio rerio* ○
- *Felis catus* ○
- *Gallus gallus* ○
- *Monodelphis domestica* ○
- *Mus musculus* ○

Search for trace name

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.