

Module 2: Maps and Genome Sequence

(Sanger)

v - How Do I Find a Zebrafish Gene in the Genome?

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. On the other hand, in Vega the identifiers remain stable since genes are linked to finished clones.

TextView

The simplest way of looking for a gene 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.

Target: Danio rerio

Query: jag2

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

Stable identifier

Features

1. **Ensembl Gene:** [ENSDARG00000021389](#)
 Ensembl gene ENSDARG00000021389 has 2 transcripts: ENSDART00000024922, ENSDART00000049586
 jagged 2 [Source:ZFIN;Acc.ZDB-GENE-011128-3]
 The gene has the following external identifiers mapped to it:
 AFY: Zebrafish: Dr.8287.1.S1_a.at
 EMBL: AF080432, AF229449, AF229450
 EntrezGene: 140422
 protein_id: AAL08214.1, AAL08215.1, AAC98354.1
 RefSeq_dna: NM_131862, NM_131665
 Uniprot/SPTREMBL: Q90Y55, Q90Y56, Q9YHU2
 ZFIN_ID: jag2, ZDB-GENE-011128-3
http://www.ensembl.org/Danio_rerio/geneview?gene=ENSDARG00000021389

Date : Thu Jul 7 18:35:27 2005 [Archive](#) [Permanent page link](#) [Help Desk / Suggestions](#)

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:

The screenshot shows the Vega/Ensembl search interface. At the top, there are logos for Vega, TextView, The Wellcome Trust Sanger Institute, and Ensembl. The search bar contains the text 'jag2'. Below the search bar, there are options for 'Display up to 20 results in standard format'. A yellow box labeled 'Target: all' points to the search input field. Another yellow box labeled 'Danio rerio results' points to the search results section below.

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



[Help Desk / Suggestions](#)

If a text-based search fails to return any meaningful output then we can instead use one of the available alignment algorithms.

SSAHA and BLAST

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

BLASTView

Ensembl Zebrafish MapView

Home Zebrafish What's New Text Search BlastSearch MartSearch Export Data Download Archive sites

Find All [e.g. 15, 12]

Known Genes % GC Repeats SNPs Chromosome 17

BLASTView

Chromosome 17

Length: 42,274,751 bps
Gene Count: 674
Known Gene Count: 140
SNP Count: 66

Change Chromosome

Chromosome: 17

Jump to Contigview

Click anywhere on the chromosome ideogram or one of the feature distribution plots to jump to a contig-level view of features at that point. Alternatively, you can jump to contigview between any two markers on this chromosome:

Between:
and:

Display contig-level view between any two features.

Map your data

Map your own data using KaryoView.

Date: 2005-07-07 20:01:56 [Help Desk / Suggestions](#)

BLASTView

BLASTView

Vega Zebrafish ContigView

Home Dog Human Mouse Zebrafish BLAST Export Data Search Feedback Help

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

Chromosome 20

Chr. 20

Overview

DNA(contigs) BXG14657.9 BXG11626.13 BXG49298.7

Markers 047996 046999

Zfish Genes slsich211-142k18.1 slsich211-142k18.2 ptk2b slsldkey-16in17.2 DREY-16in17.1 slsldkey-16in17.1 slsldkey-16in17.2

Gene Legend ■ KNOWN GENE ■ NOVEL CDS ■ PUTATIVE

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.

Vega BlastSearch (BlastView)

http://vega.sanger.ac.uk/Multi/blastview?species=Danio_rerio

The screenshot shows the Vega BlastView interface. The top navigation bar includes 'Home', 'Dog', 'Human', 'Mouse', 'Zebrafish', 'BLAST', 'Export Data', 'Search', 'Feedback', and 'Help'. The main content area is titled 'BLASTView' and has tabs for 'new', 'SETUP', 'CONFIG', 'RESULTS', and 'DISPLAY'. The 'SETUP' tab is active, showing options to 'Enter the Query Sequence' (paste, upload, or retrieve), 'Select the databases to search against' (species and database), and 'Select the Search Tool' (BLASTN, SSAHA, TBLASTX). A 'RUN' button is visible. On the right, a 'Summary' sidebar shows the status of 'setup', 'configure', 'results', and 'display'. Yellow callout boxes with arrows point to: 'Paste the sequences or...' (text input), 'enter a filename or...' (Browse... button), 'enter accession number' (Retrieve button), 'choose method' (Search Tool dropdown), 'choose target' (Species dropdown), and 'run!' (RUN button).

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 distant 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 by 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 Doc Human Mouse Zebrafish BLAST Export Data Search Feedback Help

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)

Key (%ID): 0 - 20 20 - 40 40 - 60 60 - 80 80 - 100

Alignment Locations vs. Query (click arrow to hide)

Key (%ID): 0 - 20 20 - 40 40 - 60 60 - 80 80 - 100

Alignment Summary (click arrow to hide)

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

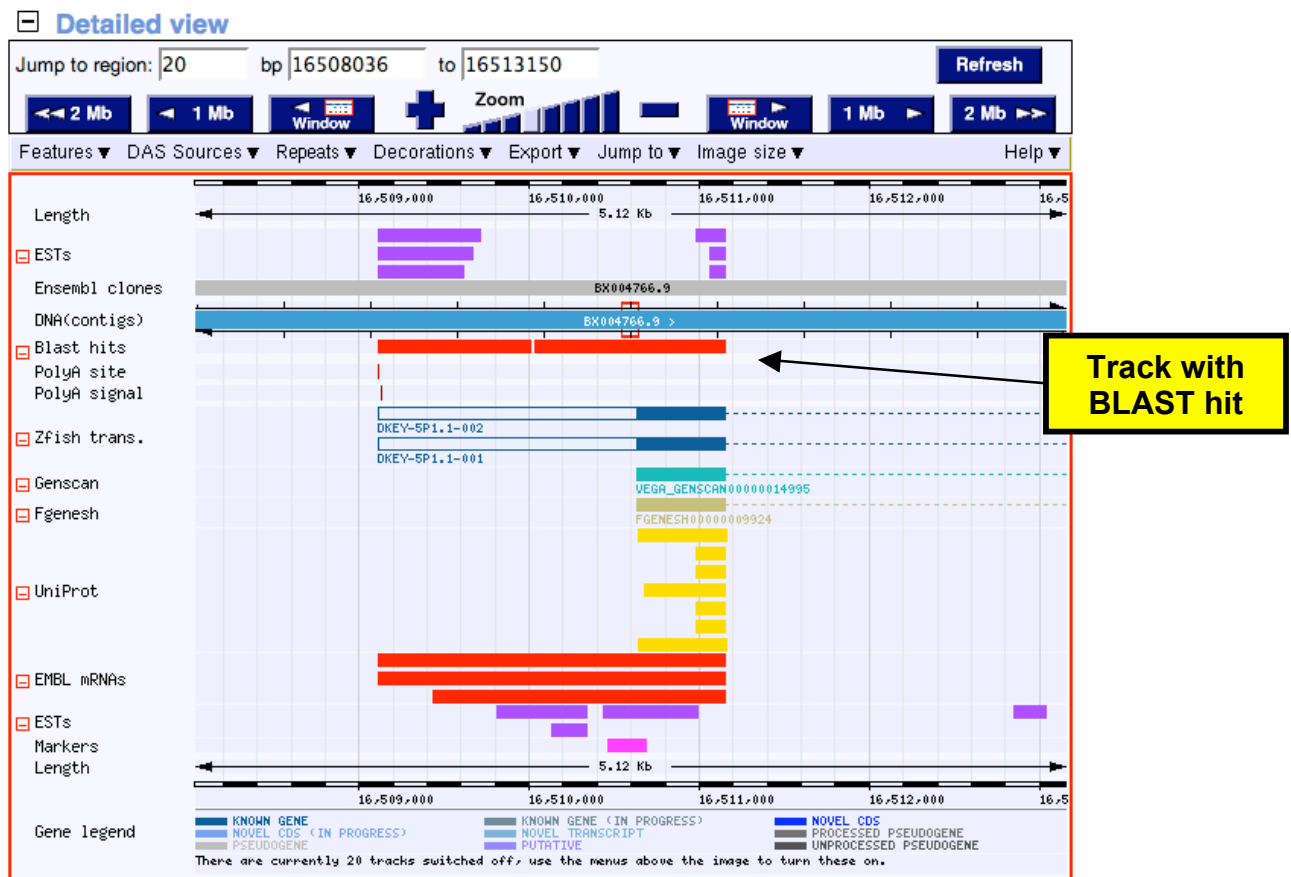
Query	Subject	Chromosome	Clone	Chunk	Stats	Sort By
off	_off_	_off_	_off_	_off_	_off_	>Chunk
Name	Name	Name	Name	Name	Score	<-Score
Start	Start	Start	Start	Start	E-val	>Score
[A] [S] [G] [C]	3397 4513 -	Chr-20	16510036	16511151 +	1091 93.46	1117
[A] [S] [G] [C]	4538 5437 -	Chr-20	16509112	16510011 +	900 100.00	900
[A] [S] [G] [C]	219 530 -	Chr-20	16592548	16592859 +	306 92.31	312
[A] [S] [G] [C]	599 850 -	Chr-20	16557160	16557411 +	249 95.24	252
[A] [S] [G] [C]	2899 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	16536381 +	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	16518936	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	16544832	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

best hit

repeat

link to ContigView

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 in identifying 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 Zv4 but soon to be updated to Zv5). 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 module 2.i 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 half 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 module 2.ii, the Vega database also contains sequenced clones from the AB strain (the AB chromosome). Chromosome U collects all the sequenced clones that have not been assigned to chromosomes.

Searching for all Finished/Unfinished clones

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

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

The screenshot shows the 'D. rerio Blast Server' interface. The page has a blue header with the Sanger Institute logo and navigation links. The main content area is titled 'D. rerio Blast Server' and contains a search form. The form is divided into 'QUERY DATA' and 'OPTIONS' sections. The 'QUERY DATA' section has a text input field for pasting a sequence, a 'Choose File' button, and 'Start Blast' and 'Reset' buttons. The 'OPTIONS' section includes dropdown menus for 'Database' (set to 'D. rerio finished sequences') and 'Executable' (set to 'BLASTN (DNA vs. DNA)'), and checkboxes for 'Filter low complexity regions', 'Mask repetitive sequences using Repeatmasker', and 'Display histogram of score statistics'. The 'Report' field is set to '100' alignments. Three yellow callout boxes with black arrows point to specific elements: 'choose a file' points to the 'Choose File' button, 'Paste your sequence or...' points to the text input field, and 'Select method' points to the 'Executable' dropdown menu.

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_reio/WebFPC/zebrafish/small.shtml

Zebrafish Genome Fingerprinting Project

Search for Fpc Clone. WebFPC [help](#) and [release notes](#) are available.

FPC contigs

Search for a clone name

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		
27	6	2	1		

Search for

Search By Marker

Search By Clone

Type	Name	Ctg

Display
Display
Clear

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

Search for trace name

Ensembl Trace Server

The Ensembl trace repository provides a permanent archive for single pass DNA sequencing reads and associated traces and quality values. These will come from whole genome shotgun projects, EST projects, and other large scale sequencing projects. It is exchanging data regularly with the [NCBI](#) trace archive.

Current services include the ability to examine individual reads by name, to search the whole archive or subsets of the archive with another DNA sequence using [SSAHA](#), a new fast equivalent of BLAST, and to download sets of read sequences in fasta format and associated quality values by [FTP](#). Requests for large data sets of complete trace information to be sent by tape should be made to trace-request@ensembl.org

Trace Repository Statistics

Species	Centre	Trace_type	Count
Acidithiobacillus ferrooxidans ATCC 23270 FTP	Any	Any	39998
Aedes aegypti FTP	Any	Any	15518122
Alligator mississippiensis FTP	Any	Any	23870
Ammonifex degensii FTP	Any	Any	1039
Anabaena variabilis ATCC 29413 FTP	Any	Any	115816
Anaeromyxobacter dehalogenans 2CP-C FTP	Any	Any	59634
Anopheles gambiae FTP	Any	Any	5394864
Aotus nancymae FTP	Any	Any	48112
Apis mellifera FTP	Any	Any	3827668
Archaeoglobus fulgidus DSM 4304 FTP	Any	Any	32
Arthrobacter sp. FB24 FTP	Any	Any	60054
Artibeus jamaicensis FTP	Any	Any	10601
Aspergillus fumigatus FTP	Any	Any	226633
Aspergillus nidulans FTP	Any	Any	623573
Aspergillus terreus FTP	Any	Any	615346
Atelerix albiventris FTP	Any	Any	228311

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.