Module 3: Genome Browsing

Aims

- Briefly present the main web-based genome browsers.
- Using Ensembl, demonstrate some of the features and applications of genome browsers.
- Introduce the BioMart data retrieval system.
- Create files with your own data to upload to a genome browser.

Introduction

Web-based genome browsers have been developed to make it easier to access comprehensive information about regions of the human genome and about the whole human gene set. They help you to:

- Explore what is in a chromosomal region
- See features in and around a specific gene
- Search & retrieve data across the whole genome
- Investigate genome organisation
- Compare to other genomes

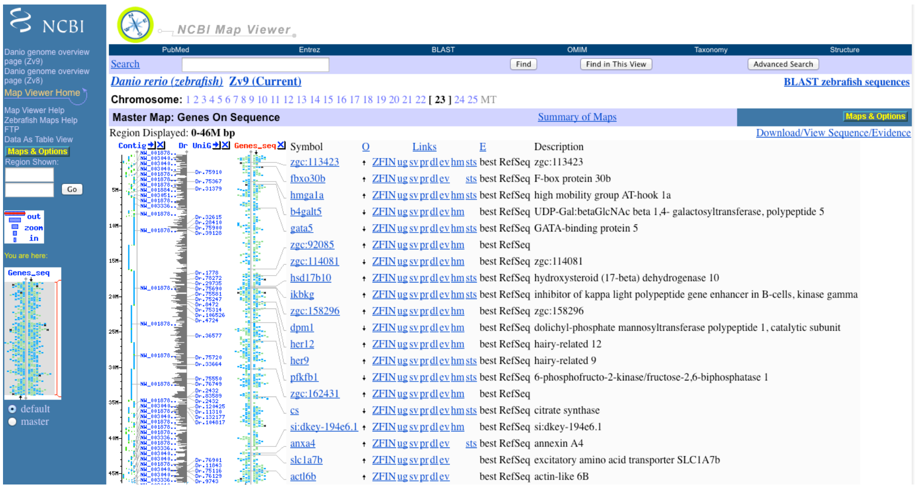
Browsers display the location and structure of known genes and predicted novel genes along with information about the mRNA transcripts and may also include information about protein products. Information about genes is integrated with information about other genomic features (e.g. variation data, markers, repeated sequences, regions homologous to other species) and displayed alongside the genomic sequence assembly. Protein, mRNA and EST entries from various sequence databases may also be shown aligned to the chromosomes.

In addition to providing annotation across the whole genome, browsers provide other resources. The browsers differ in what is provided and how it is presented. Resources that can be found include:

- **Links** to other databases and resources
- **Text Searching**
- **BLAST** and other sequence similarity searching
- **Download** of genomic sequence, gene information and other data
- **Data mining** facilities

Browsers (and some of their strengths)

- **NCBI Map Viewer** – maintained by NCBI
  http://www.ncbi.nlm.nih.gov/mapview/
- **UCSC Genome Browser** – maintained by UCSC
  http://genome.ucsc.edu/cgi-bin/hgGateway
- **Ensembl** – maintained by EBI / Sanger Institute
  http://www.ensembl.org



Danio rerio (zebrafish) Zv9 (Current)

BLAST zebrafish sequences

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

Master Map: Genes On Sequence

Summary of Maps

Maps & Options

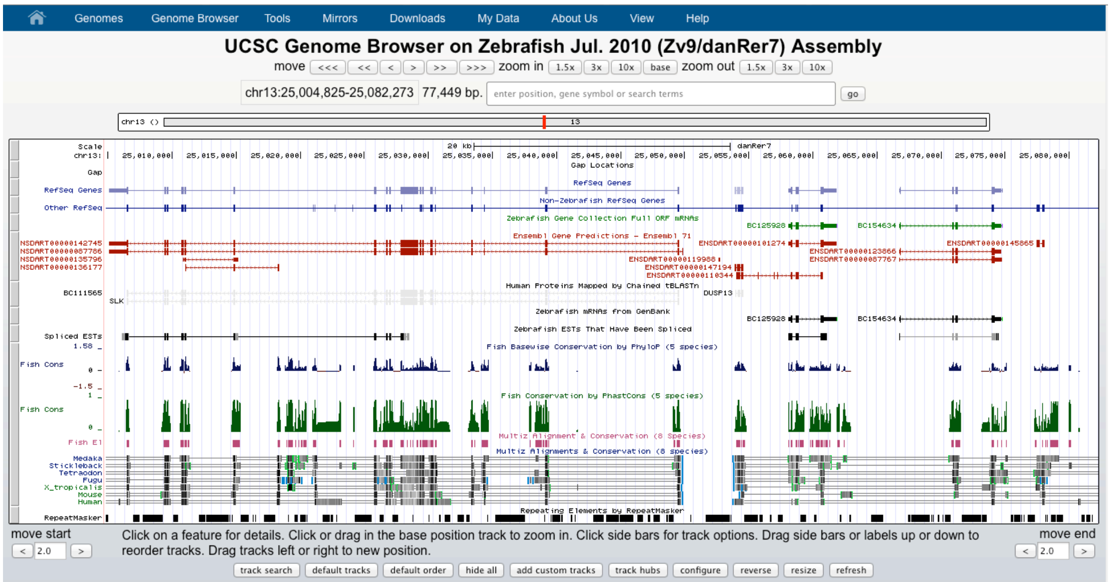
Region Displayed: 0-46M bp

Download/View Sequence/Evidence

Symbol	O	Links	E	Description
zgc:113423	↑	ZFIN ug sv pr dl ev hm sts	best RefSeq	zgc:113423
fbxo30b	↑	ZFIN ug sv pr dl ev sts	best RefSeq	F-box protein 30b
hmga1a	↑	ZFIN ug sv pr dl ev hm sts	best RefSeq	high mobility group AT-hook 1a
b4galt5	↓	ZFIN ug sv pr dl ev hm	best RefSeq	UDP-Gal:betaGlcNAc beta 1,4- galactosyltransferase, polypeptide 5
gata5	↓	ZFIN ug sv pr dl ev hm sts	best RefSeq	GATA-binding protein 5
zgc:92085	↑	ZFIN ug sv pr dl ev hm	best RefSeq	
zgc:114081	↓	ZFIN ug sv pr dl ev hm	best RefSeq	zgc:114081
hsd17b10	↑	ZFIN ug sv pr dl ev hm sts	best RefSeq	hydroxysteroid (17-beta) dehydrogenase 10
ikbkg	↑	ZFIN ug sv pr dl ev hm sts	best RefSeq	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma
zgc:158296	↑	ZFIN ug sv pr dl ev hm	best RefSeq	zgc:158296
dpm1	↓	ZFIN ug sv pr dl ev hm	best RefSeq	dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit
her12	↑	ZFIN ug sv pr dl ev hm	best RefSeq	hairy-related 12
her9	↑	ZFIN ug sv pr dl ev hm sts	best RefSeq	hairy-related 9
pfkfb1	↓	ZFIN ug sv pr dl ev hm sts	best RefSeq	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1
zgc:162431	↑	ZFIN ug sv pr dl ev hm	best RefSeq	
cs	↓	ZFIN ug sv pr dl ev hm sts	best RefSeq	citrate synthase
si:dkey-194e6.1	↑	ZFIN ug sv pr dl ev hm	best RefSeq	si:dkey-194e6.1
anxa4	↑	ZFIN ug sv pr dl ev sts	best RefSeq	annexin A4
slc1a7b	↑	ZFIN ug sv pr dl ev	best RefSeq	excitatory amino acid transporter SLC1A7b
actl6b	↑	ZFIN ug sv pr dl ev	best RefSeq	actin-like 6B

NCBI Map Viewer

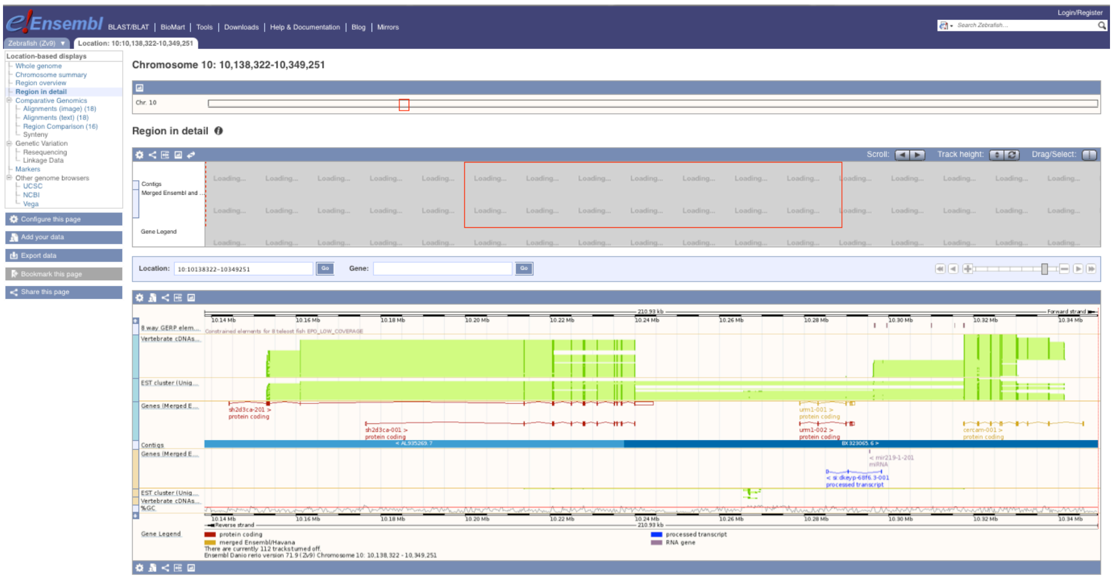
- Good integration with other NCBI resources



Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

UCSC Genome Browser

- Straightforward feature display
- Old assemblies available
- Wide range of tracks supplied by other groups
- Trackhub support



There are currently 112 tracks turned off
Ensembl Danio rerio version 71.9 (Zv9) Chromosome 10: 10,138,322 - 10,349,251

Ensembl

- Well-supported gene set with evidence
- Range of different views
- Archive available
- Trackhub support

While browsers can be very useful tools, they do not provide the definitive answer to every question! Remember, new data and updates make genome browsing a fluid, changing, and improving, process.

Data retrieval and data mining

Genomic annotation data, due to its complexity and volume, does not lend itself to easy access. Presenting it on a web site is important, but so is providing simple but flexible ways to select and retrieve specific sets of data. NCBI has the Entrez query system and UCSC has its Table Browser.

In Ensembl, BioMart facilitates rapid retrieval of richly annotated gene lists, sequences, and variations, among other annotation, integrated with third party data and applications. Genes can be selected by chromosome region, protein domains, associated external identifiers or SNP properties, and these filters can be combined to group and refine biological data, including cross-species analyses, disease links, sequence variations and expression patterns.

BioMart is built upon a query-optimised relational database schema allowing quick and efficient access to voluminous data through a user-friendly, interactive web interface. After selecting the biological object and the species, the results can be refined using a set of pre-defined filters. After each navigation event, the user is provided with immediate feedback on the number of matches found. Output can consist of annotated gene lists, gene structures, SNP details or various kinds of sequence sets. Output can be in HTML, text, Microsoft Excel and compressed formats.

Further reading

Ensembl Help and Documentation http://www.ensembl.org/info/index.html

Cunningham F. *et al*.
Ensembl 2015
Nucleic Acids Res. 2013 Jan;43(Database issue):D662-9

Spudich GM, Fernandez-Suarez XM
Touring Ensembl: a practical guide to genome browsing.
BMC Genomics, 2010 May 11;11:295

Meyer LR, et al.
The UCSC Genome Browser database: extensions and updates 2013.
Nucleic Acids Res. 2012 Nov 15.

Karolchik, D *et al.*
The UCSC Genome Browser Database.
Nucl. Acids Res. 2003 **31**, 51-54

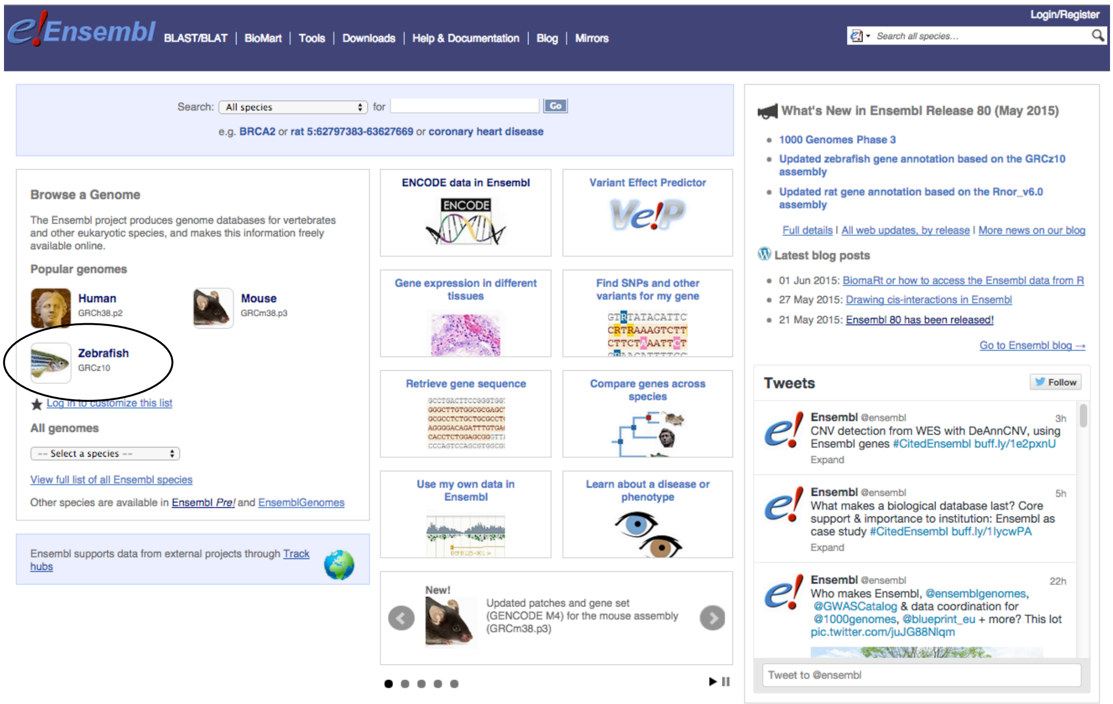
Dombrowski, S M and Maglott, D.
Using the Map Viewer to Explore Genomes
in The NCBI Handbook
http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/handbook/ch20d1.pdf

WALKING THROUGH THE WEBSITE

The instructor will guide you through the release 80 Ensembl website using the **thyrotrophic embryonic factor a (tefa)** gene. The following points will be addressed:

- **The Gene Summary tab and gene-related links:**
  - Are there splice variants?
  - Can I view the genomic sequence with variations?
  - Find orthologues and paralogues, show alignments with other genomes
- **The Transcript tab and related links:**
  - What is the protein sequence?
  - What matching proteins and mRNAs are found in other databases?
  - Gene Ontology
- **The Location tab and related links:**
  - How do I zoom in and change the gene focus.
  - Un-stacking a track (e.g. human cDNAs)
  - Adding a track (i.e. variations)
- **Exporting a sequence and running BLAT/BLAST**

Start by going to **www.ensembl.org**


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species…
Search: All species for Go
ENCODE
Ve!P
Follow
Tweet to @ensembl

e.g. BRCA2 or rat 5:62797383-63627669 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh38.p2

Mouse GRCm38.p3

Zebrafish GRCz10

Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

Ensembl supports data from external projects through Track hubs

ENCODE data in Ensembl

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

Compare genes across species

Use my own data in Ensembl

Learn about a disease or phenotype

New! Updated patches and gene set (GENCODE M4) for the mouse assembly (GRCm38.p3)

What's New in Ensembl Release 80 (May 2015)

- 1000 Genomes Phase 3
- Updated zebrafish gene annotation based on the GRCz10 assembly
- Updated rat gene annotation based on the Rnor_v6.0 assembly

Full details I All web updates, by release I More news on our blog

Latest blog posts

- 01 Jun 2015: BiomaRt or how to access the Ensembl data from R
- 27 May 2015: Drawing cis-interactions in Ensembl
- 21 May 2015: Ensembl 80 has been released!

Go to Ensembl blog →

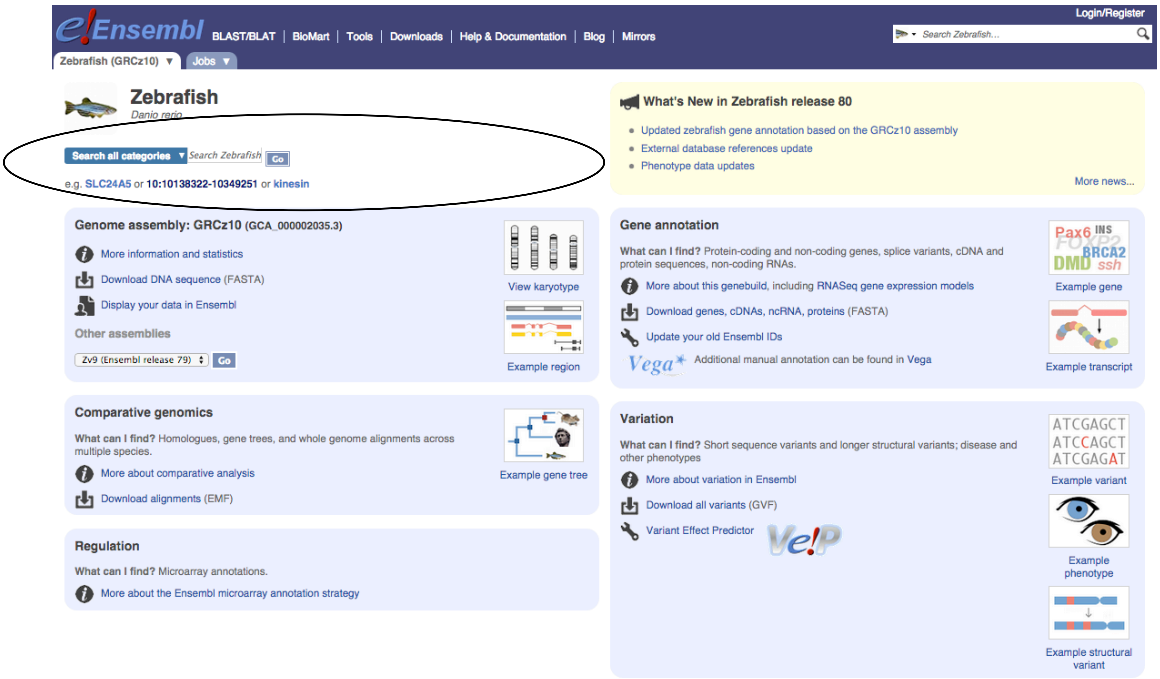
Tweets

Ensembl @ensembl 3h
CNV detection from WES with DeAnnCNV, using Ensembl genes #CitedEnsembl buff.ly/1e2pxnU
Expand

Ensembl @ensembl 5h
What makes a biological database last? Core support & importance to institution: Ensembl as case study #CitedEnsembl buff.ly/1lycwPA
Expand

Ensembl @ensembl 22h
Who makes Ensembl, @ensemblgenomes, @GWASCatalog & data coordination for @1000genomes, @blueprint_eu + more? This lot pic.twitter.com/juJG88Nlqm

Click on ‘Zebrafish’, or the picture circled above, which brings us to the species index page.



Zebrafish
Danio rerio

e.g. SLC24A5 or 10:10138322-10349251 or kinesin

What's New in Zebrafish release 80

- Updated zebrafish gene annotation based on the GRCz10 assembly
- External database references update
- Phenotype data updates

More news...

Genome assembly: GRCz10 (GCA_000002035.3)

More information and statistics

Download DNA sequence (FASTA)

Display your data in Ensembl

Other assemblies

View karyotype

Example region

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild, including RNASeq gene expression models

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Vega Additional manual annotation can be found in Vega

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Example gene tree

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl

Download all variants (GVF)

Variant Effect Predictor

Example variant

Example phenotype

Example structural variant

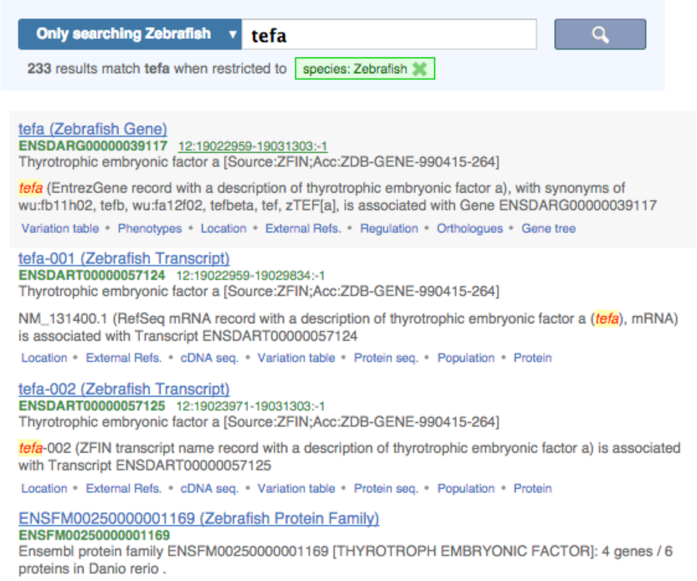
Regulation

What can I find? Microarray annotations.

More about the Ensembl microarray annotation strategy

Type ‘tefa’ into the search bar circled above and click the ‘Go’ button.

The search will return everything that matches tefa, the first result is the *tefa* gene. You can navigate to the 'Gene' page, to a 'Location view' page (coordinates hyperlink) or to any other View' type as listed below the search result.



233 results match **tefa** when restricted to species: Zebrafish ✖

tefa (Zebrafish Gene)
ENSDARG00000039117 12:19022959-19031303:-1
Thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]

tefa (EntrezGene record with a description of thyrotrophic embryonic factor a), with synonyms of wu:fb11h02, tefb, wu:fa12f02, tefbeta, tef, zTEF[a], is associated with Gene ENSDARG00000039117

Variation table • Phenotypes • Location • External Refs. • Regulation • Orthologues • Gene tree

tefa-001 (Zebrafish Transcript)
ENSDART00000057124 12:19022959-19029834:-1
Thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]

NM_131400.1 (RefSeq mRNA record with a description of thyrotrophic embryonic factor a (tefa), mRNA) is associated with Transcript ENSDART00000057124

Location • External Refs. • cDNA seq. • Variation table • Protein seq. • Population • Protein

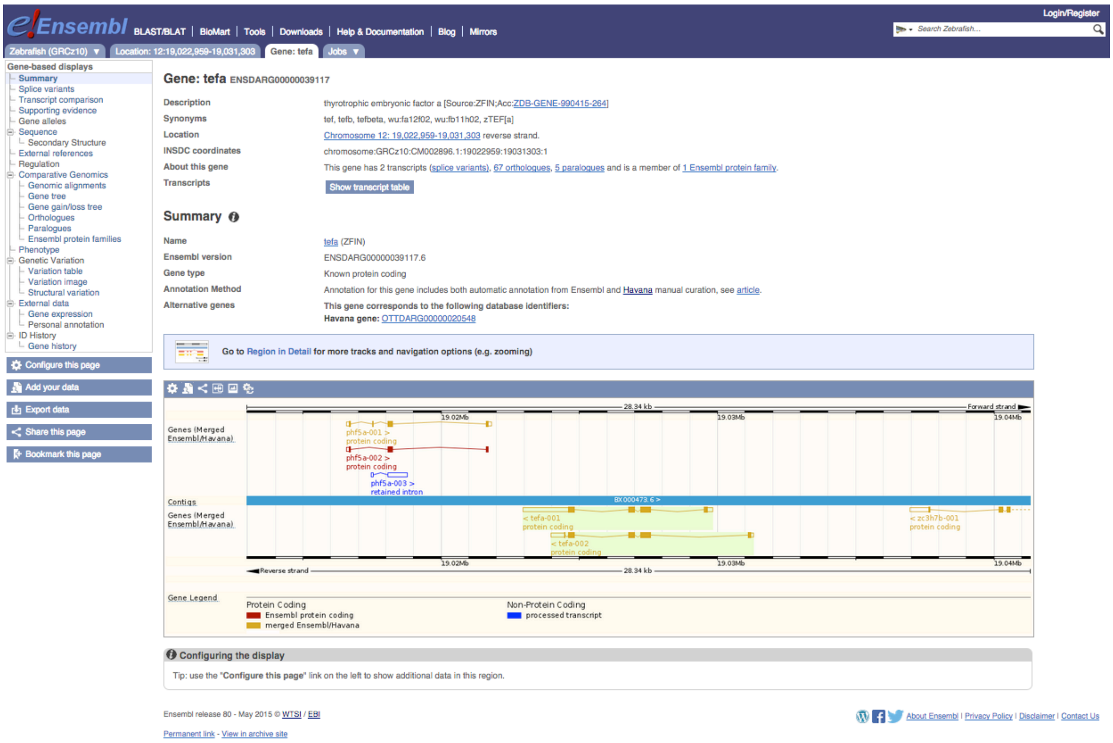
tefa-002 (Zebrafish Transcript)
ENSDART00000057125 12:19023971-19031303:-1
Thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]

tefa-002 (ZFIN transcript name record with a description of thyrotrophic embryonic factor a) is associated with Transcript ENSDART00000057125

Location • External Refs. • cDNA seq. • Variation table • Protein seq. • Population • Protein

ENSFM00250000001169 (Zebrafish Protein Family)
ENSFM00250000001169
Ensembl protein family ENSFM00250000001169 [THYROTROPH EMBRYONIC FACTOR]: 4 genes / 6 proteins in Danio rerio .

Click on 'tefa (Zebrafish Gene)' and the following 'Gene' tab will open:



Gene: tefa ENSDARG00000039117

Description	thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]
Synonyms	tef, tefb, tefbeta, wu:fa12f02, wu:fb11h02, zTEF[a]
Location	Chromosome 12: 19,022,959-19,031,303 reverse strand.
INSDC coordinates	chromosome:GRCz10:CM002896.1:19022959:19031303:1
About this gene	This gene has 2 transcripts (splice variants), 67 orthologues, 5 paralogues and is a member of 1 Ensembl protein family.
Transcripts	Show transcript table

Summary

Name	tefa (ZFIN)
Ensembl version	ENSDARG00000039117.6
Gene type	Known protein coding
Annotation Method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.
Alternative genes	This gene corresponds to the following database identifiers: Havana gene: OTTDARG00000020548

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

Configuring the display

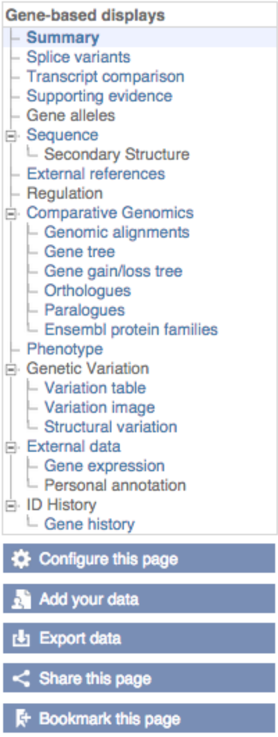
Tip: use the "**Configure this page**" link on the left to show additional data in this region.

Ensembl release 80 - May 2015 © WTSI / EBI

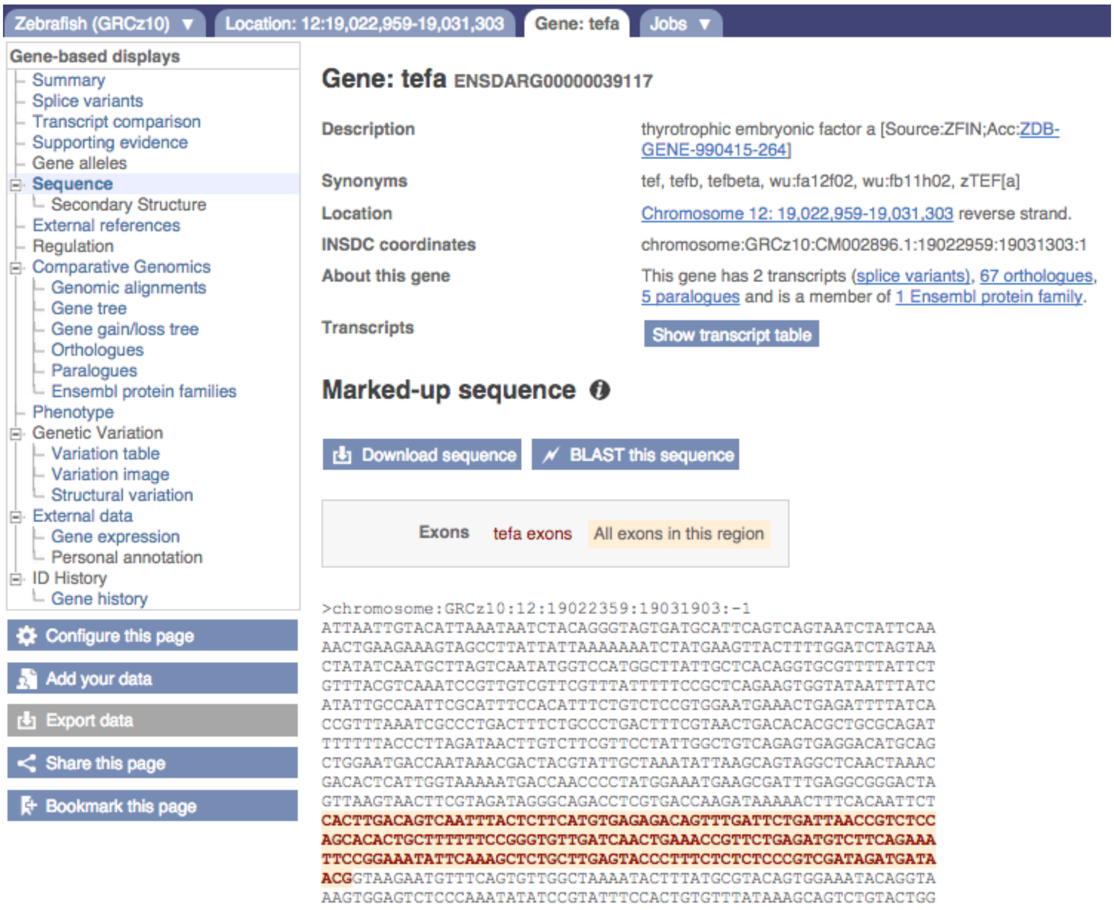
Permanent link - View in archive site

About Ensembl | Privacy Policy | Disclaimer | Contact Us

Let’s walk through some of the links in the left hand navigation column.



How can we view the genomic sequence? Click **‘Sequence’** at the left



Gene: tefa ENSDARG00000039117

Description: thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]

Synonyms: tef, tefb, tefbeta, wu:fa12f02, wu:fb11h02, zTEF[a]

Location: Chromosome 12: 19,022,959-19,031,303 reverse strand.

INSDC coordinates: chromosome:GRCz10:CM002896.1:19022959:19031303:1

About this gene: This gene has 2 transcripts (splice variants), 67 orthologues, 5 paralogues and is a member of 1 Ensembl protein family.

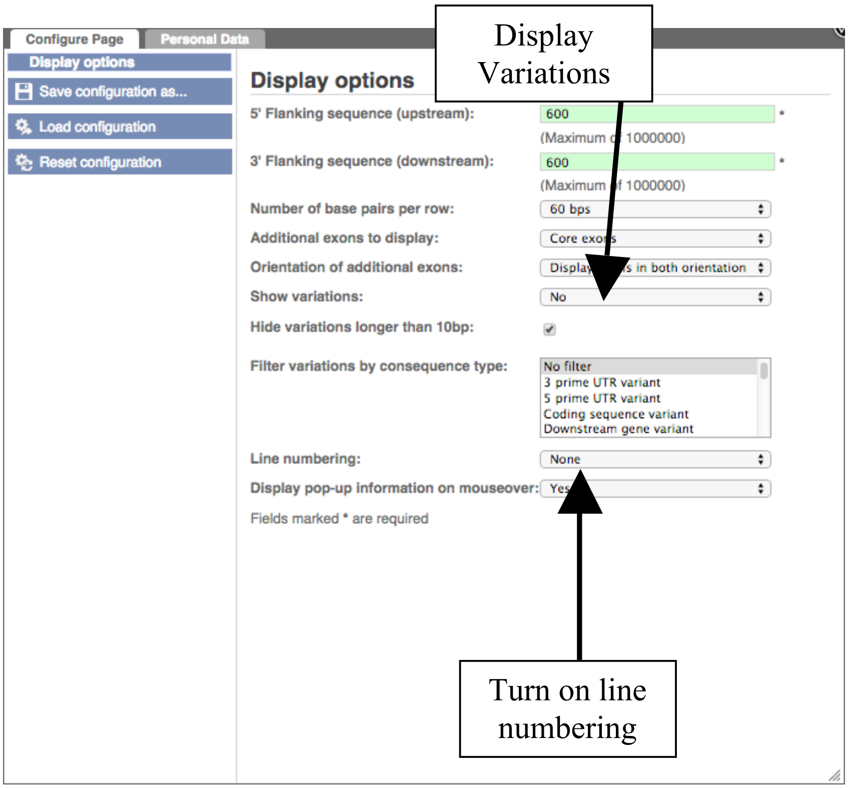
Transcripts

Marked-up sequence

Exons tefa exons All exons in this region

```
>chromosome:GRCz10:12:19022359:19031903:-1
ATTAATTGTACATTAAATAATCTACAGGGTAGTGATGCATTCAGTCAGTAATCTATTCAA
AACTGAAGAAAGTAGCCTTATTATTAAAAAAATCTATGAAGTTACTTTTGGATCTAGTAA
CTATATCAATGCTTAGTCAATATGGTCCATGGCTTATTGCTCACAGGTGCGTTTTATTCT
GTTTACGTCAAATCCGTTGTCGTTCGTTTATTTTTCCGCTCAGAAGTGGTATAATTTATC
ATATTGCCAATTCGCATTTCCACATTTCTGTCTCCGTGGAATGAAACTGAGATTTTATCA
CCGTTTAAATCGCCCTGACTTTCTGCCCTGACTTTCGTAACTGACACACGCTGCGCAGAT
TTTTTTACCCTTAGATAACTTGTCTTCGTTCCTATTGGCTGTCAGAGTGAGGACATGCAG
CTGGAATGACCAATAAACGACTACGTATTGCTAAATATTAAGCAGTAGGCTCAACTAAAC
GACACTCATTGGTAAAAATGACCAACCCCTATGGAAATGAAGCGATTTGAGGCGGGACTA
GTTAAGTAACTTCGTAGATAGGGCAGACCTCGTGACCAAGATAAAAACTTTCACAATTCT
CACTTGACAGTCAATTTACTCTTCATGTGAGAGACAGTTTGATTCTGATTAACCGTCTCC
AGCACACTGCTTTTTTCCGGGTGTTGATCAACTGAAACCGTTCTGAGATGTCTTCAGAAA
TTCCGGAAATATTCAAAGCTCTGCTTGAGTACCCTTTCTCTCTCCCGTCGATAGATGATA
ACGGTAAGAATGTTTCAGTGTTGGCTAAAATACTTTATGCGTACAGTGGAAATACAGGTA
AAGTGGAGTCTCCCAAATATATCCGTATTTCCACTGTGTTTATAAAGCAGTCTGTACTGG
```

‘**Configure this page**’ in the left hand menu allows you to make changes to the display, e.g. add coordinates and make variation visible.



Display options

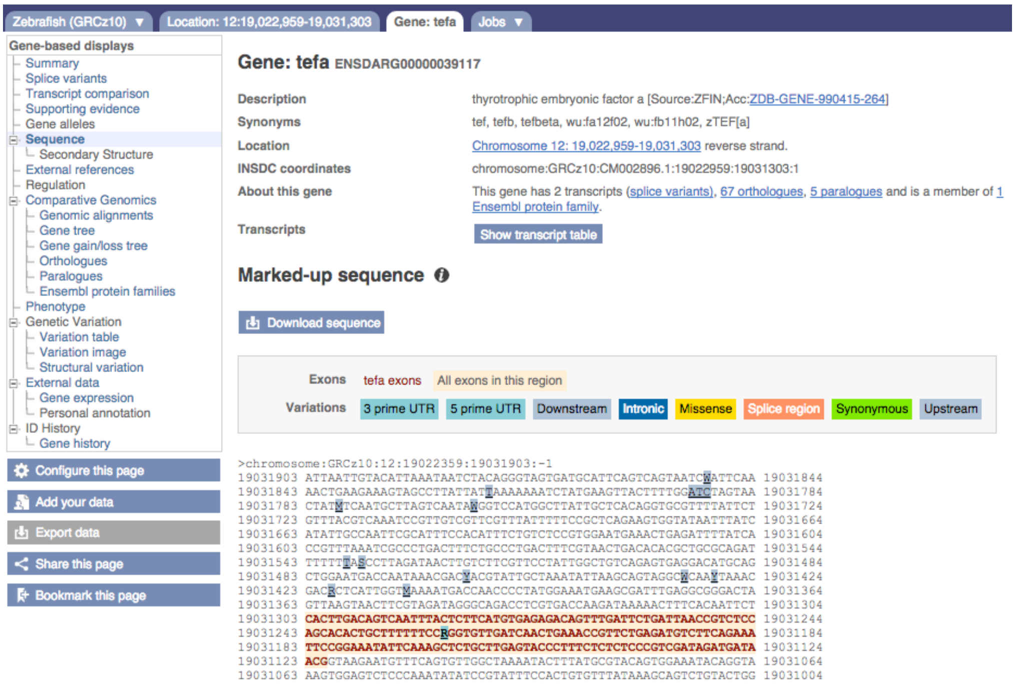
5' Flanking sequence (upstream):	600 * (Maximum of 1000000)
3' Flanking sequence (downstream):	600 * (Maximum of 1000000)
Number of base pairs per row:	60 bps
Additional exons to display:	Core exons
Orientation of additional exons:	Display exons in both orientation
Show variations:	No
Hide variations longer than 10bp:	☑
Filter variations by consequence type:	No filter 3 prime UTR variant 5 prime UTR variant Coding sequence variant Downstream gene variant
Line numbering:	None
Display pop-up information on mouseover:	Yes

Fields marked * are required

Display Variations

Turn on line numbering

Once you have selected changes (in this example, we display variations and show chromosome coordinates) click the **tick** at the top right.



Gene: tefa ENSDARG00000039117

Description	thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]
Synonyms	tef, tefb, tefbeta, wu:fa12f02, wu:fb11h02, zTEF[a]
Location	Chromosome 12: 19,022,959-19,031,303 reverse strand.
INSDC coordinates	chromosome:GRCz10:CM002896.1:19022959:19031303:1
About this gene	This gene has 2 transcripts (splice variants), 67 orthologues, 5 paralogues and is a member of 1 Ensembl protein family.
Transcripts	Show transcript table

Marked-up sequence

Download sequence

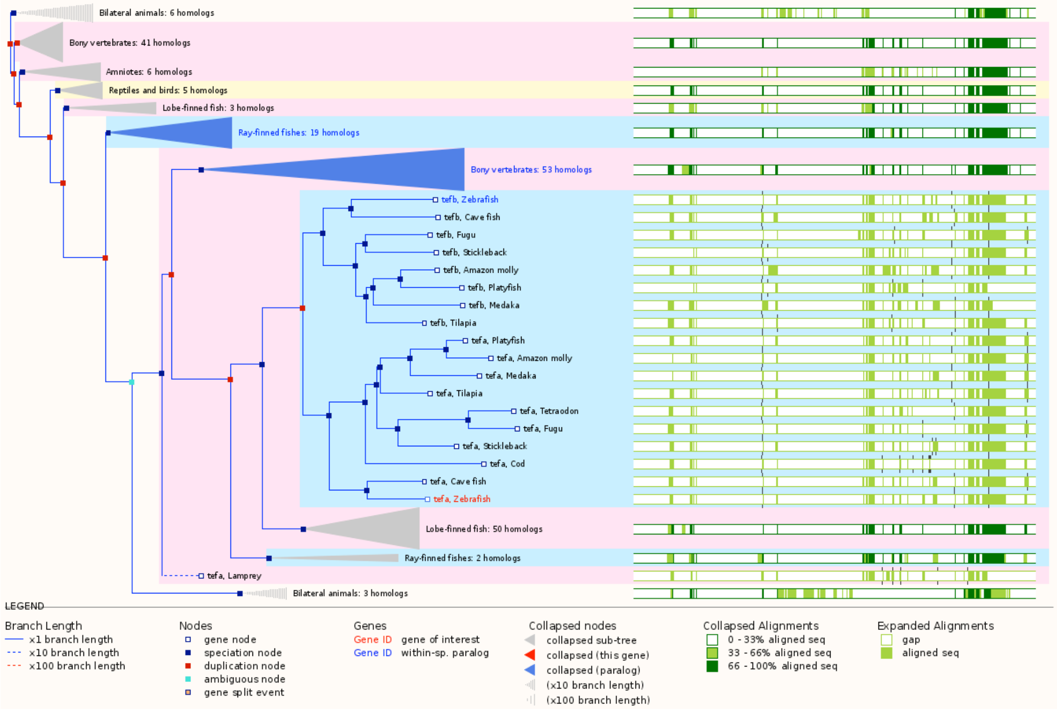
Exons: tefa exons | All exons in this region

Variations: 3 prime UTR | 5 prime UTR | Downstream | Intronic | Missense | Splice region | Synonymous | Upstream

```
>chromosome:GRCz10:12:19022359:19031903:-1
19031903 ATTAATTGTACATTAAATAATCTACAGGGTAGTGATGCATTCAGTCAGTAATCWATTCAA 19031844
19031843 AACTGAAGAAAGTAGCCTTATTATTAAAAAAATCTATGAAGTTACTTTTGGATCTAGTAA 19031784
19031783 CTATMTCAATGCTTAGTCAATAWGGTCCATGGCTTATTGCTCACAGGTGCGTTTTATTCT 19031724
19031723 GTTTACGTCAAATCCGTTGTCGTTCGTTTATTTTTCCGCTCAGAAGTGGTATAATTTATC 19031664
19031663 ATATTGCCAATTCGCATTTCCACATTTCTGTCTCCGTGGAATGAAACTGAGATTTTATCA 19031604
19031603 CCGTTTAAATCGCCCTGACTTTCTGCCCTGACTTTCGTAACTGACACACGCTGCGCAGAT 19031544
19031543 TTTTTTASCCTTAGATAACTTGTCTTCGTTCCTATTGGCTGTCAGAGTGAGGACATGCAG 19031484
19031483 CTGGAATGACCAATAAACGACYACGTATTGCTAAATATTAAGCAGTAGGCWCAAYTAAAC 19031424
19031423 GACRCTCATTGGTMAAAATGACCAACCCCTATGGAAATGAAGCGATTTGAGGCGGGACTA 19031364
19031363 GTTAAGTAACTTCGTAGATAGGGCAGACCTCGTGACCAAGATAAAAACTTTCACAATTCT 19031304
19031303 CACTTGACAGTCAATTTACTCTTCATGTGAGAGACAGTTTGATTCTGATTAACCGTCTCC 19031244
19031243 AGCACACTGCTTTTTTCCRGGTGTTGATCAACTGAAACCGTTCTGAGATGTCTTCAGAAA 19031184
19031183 TTCCGGAAATATTCAAAGCTCTGCTTGAGTACCCTTTCTCTCTCCCGTCGATAGATGATA 19031124
19031123 ACGGTAAGAATGTTTCAGTGTTGGCTAAAATACTTTATGCGTACAGTGGAAATACAGGTA 19031064
19031063 AAGTGGAGTCTCCCAAATATATCCGTATTTCCACTGTGTTTATAAAGCAGTCTGTACTGG 19031004
```

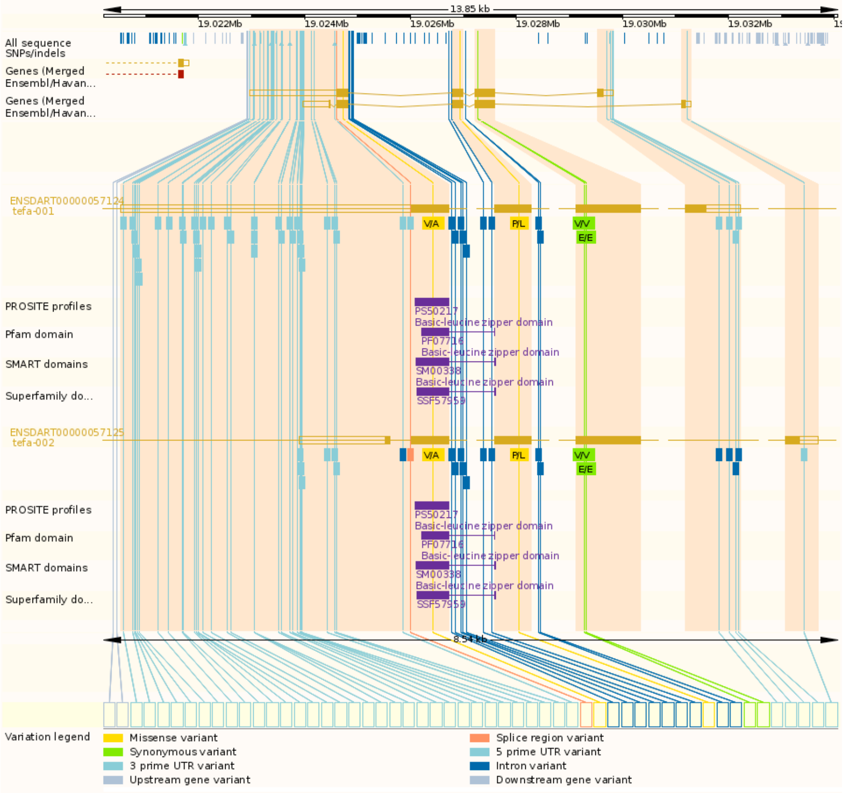
Now variations in the sequence are highlighted. Coordinates have been added.

Now let’s click on **‘Gene tree’**, which will display the current gene in the context of a phylogenetic tree of orthologous and paralogous genes.

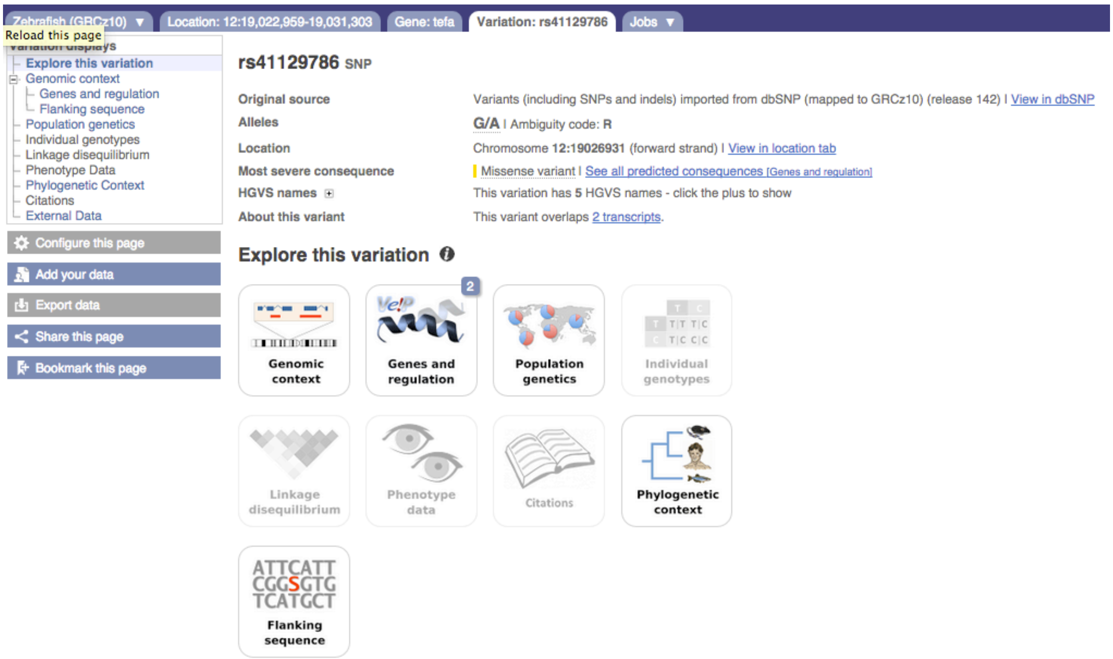


Use the mouse over and ‘expand sub-tree’ to get to the view displayed above. Note that there are two *tef* genes in fish species. In zebrafish these are annotated as *tefa* and *tefb*.

Click on **‘Variation image’** to display genetic variation mapped onto all transcripts of a gene.



Click any variation, then **‘Variation properties’** to learn more about it. A fourth tab will open:

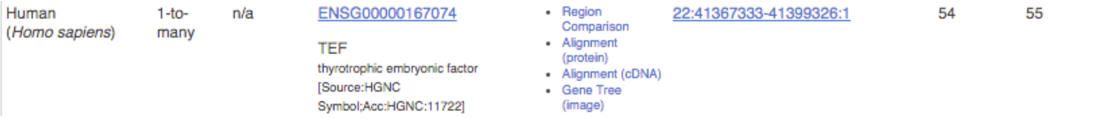


rs41129786 SNP

Original source	Variants (including SNPs and indels) imported from dbSNP (mapped to GRCz10) (release 142) I View in dbSNP
Alleles	G/A I Ambiguity code: R
Location	Chromosome 12:19026931 (forward strand) I View in location tab
Most severe consequence	Missense variant I See all predicted consequences [Genes and regulation]
HGVS names ⊞	This variation has 5 HGVS names - click the plus to show
About this variant	This variant overlaps 2 transcripts.

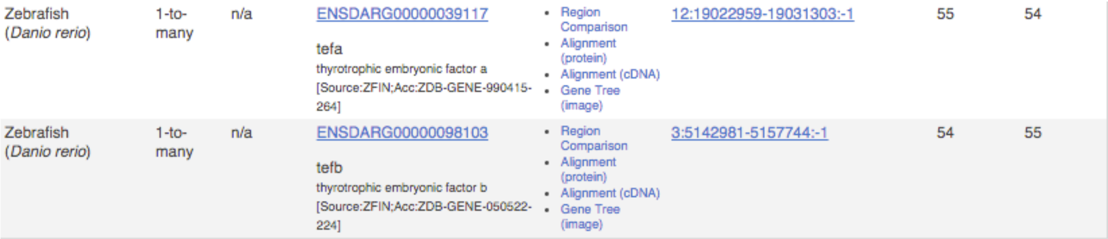
Explore this variation

To find **orthologs** of the *tefa* gene in other species, return to the ‘Gene’ tab and select ‘Orthologues’. You will find that the zebrafish ‘tefa’ gene has one human ortholog, TEF.



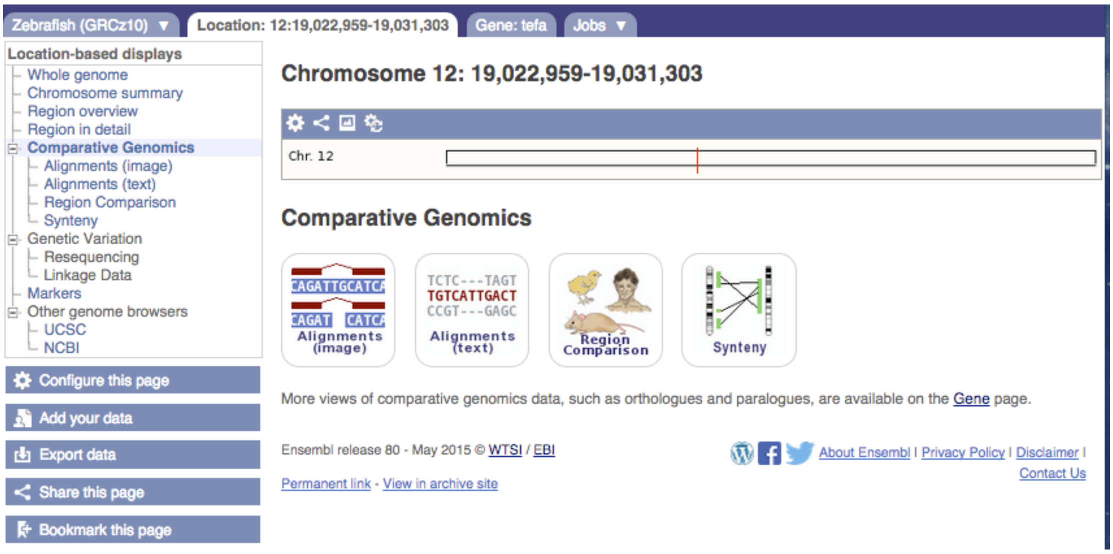
Human (*Homo sapiens*)	1-to-many	n/a	ENSG00000167074 TEF thyrotrophic embryonic factor [Source:HGNC Symbol;Acc:HGNC:11722]	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	22:41367333-41399326:1	54	55

The relationship to the ortholog is ‘1-to-many’, meaning that several zebrafish paralogs share the close relationship with one human TEF gene (compare answer to **Exercise 1**). You can check this by navigating to the human gene and looking at its zebrafish orthologs:



Zebrafish (*Danio rerio*)	1-to-many	n/a	ENSDARG00000039117 tefa thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	12:19022959-19031303:-1	55	54
Zebrafish (*Danio rerio*)	1-to-many	n/a	ENSDARG00000098103 tefb thyrotrophic embryonic factor b [Source:ZFIN;Acc:ZDB-GENE-050522-224]	• Region Comparison • Alignment (protein) • Alignment (cDNA) • Gene Tree (image)	3:5142981-5157744:-1	54	55

In order to find out how the TEF and *tefa* loci compare, go back to the *tefa* location tab and click ‘**Comparative Genomics**’ in the left hand menu.



Chromosome 12: 19,022,959-19,031,303

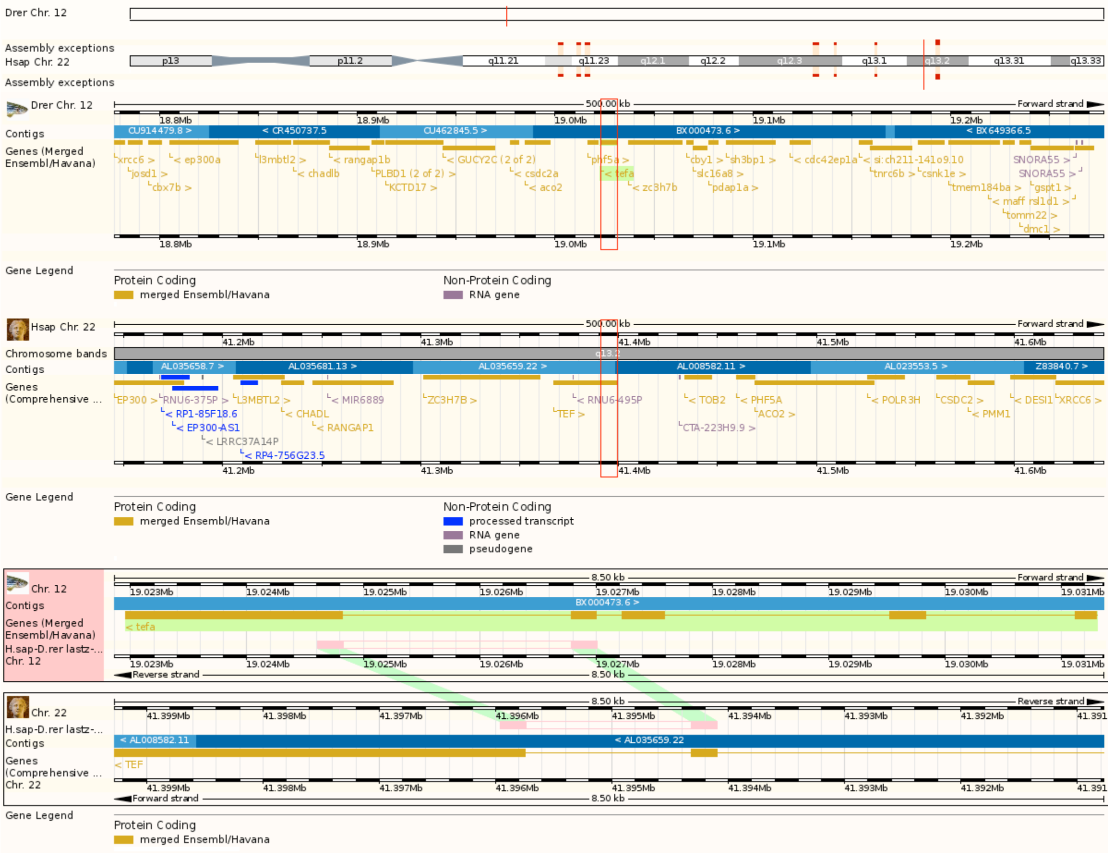
Comparative Genomics

More views of comparative genomics data, such as orthologues and paralogues, are available on the Gene page.

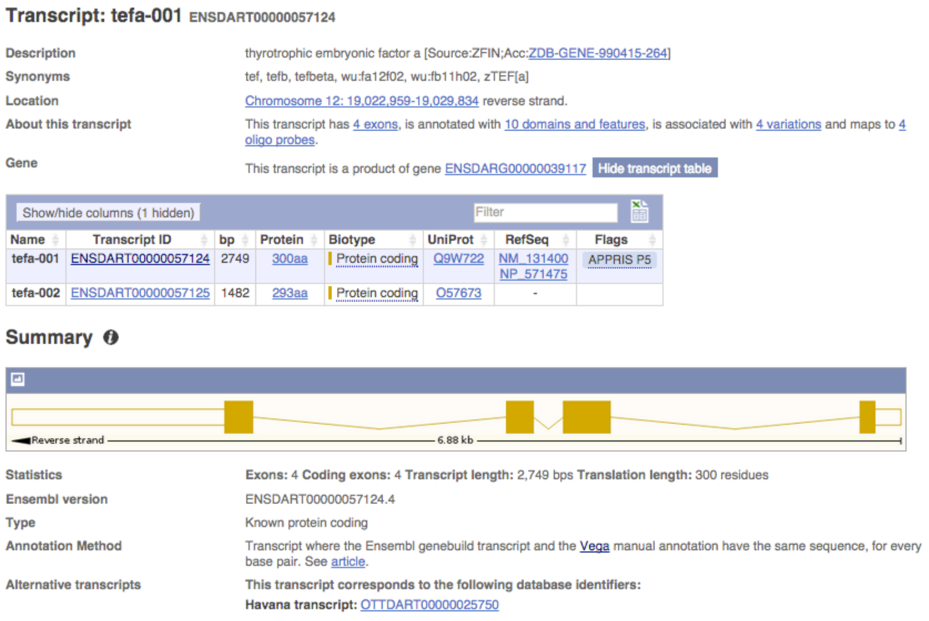
Ensembl release 80 - May 2015 © WTSI / EBI

Permanent link - View in archive site

‘Region Comparisons’ provides pre-computed alignments between species. Select ‘Human’ to see the below.



Now, we would like to work with a **transcript** of this gene. Return to the *tefa* ‘Gene‘ page, select ‘Show transcript table’ and and click tefa-001. This will lead to the transcript summary display.



Transcript: tefa-001 ENSDART00000057124

Description	thyrotrophic embryonic factor a [Source:ZFIN;Acc:ZDB-GENE-990415-264]
Synonyms	tef, tefb, tefbeta, wu:fa12f02, wu:fb11h02, zTEF[a]
Location	Chromosome 12: 19,022,959-19,029,834 reverse strand.
About this transcript	This transcript has 4 exons, is annotated with 10 domains and features, is associated with 4 variations and maps to 4 oligo probes.
Gene	This transcript is a product of gene ENSDARG00000039117 Hide transcript table

Show/hide columns (1 hidden) Filter

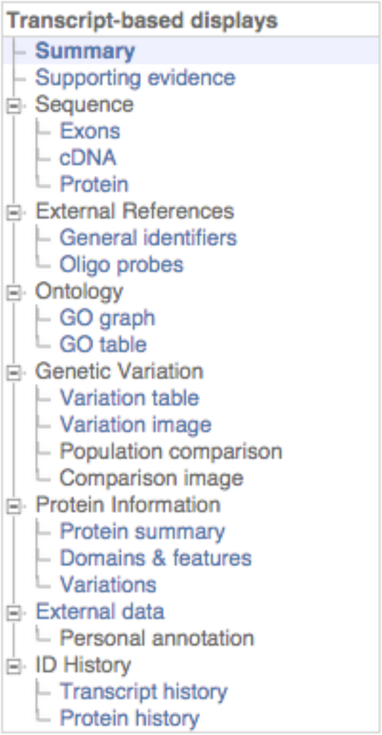
Name	Transcript ID	bp	Protein	Biotype	UniProt	RefSeq	Flags
tefa-001	ENSDART00000057124	2749	300aa	Protein coding	Q9W722	NM_131400 NP_571475	APPRIS P5
tefa-002	ENSDART00000057125	1482	293aa	Protein coding	O57673	-	

Summary

Statistics	**Exons:** 4 **Coding exons:** 4 **Transcript length:** 2,749 bps **Translation length:** 300 residues
Ensembl version	ENSDART00000057124.4
Type	Known protein coding
Annotation Method	Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See article.
Alternative transcripts	**This transcript corresponds to the following database identifiers:** **Havana transcript:** OTTDART00000025750

Note that this is a ‘merged’ or ‘golden’ transcript, i.e. the automated (Ensembl) and manual (Vega, vega.sanger.ac.uk) annotation are identical.

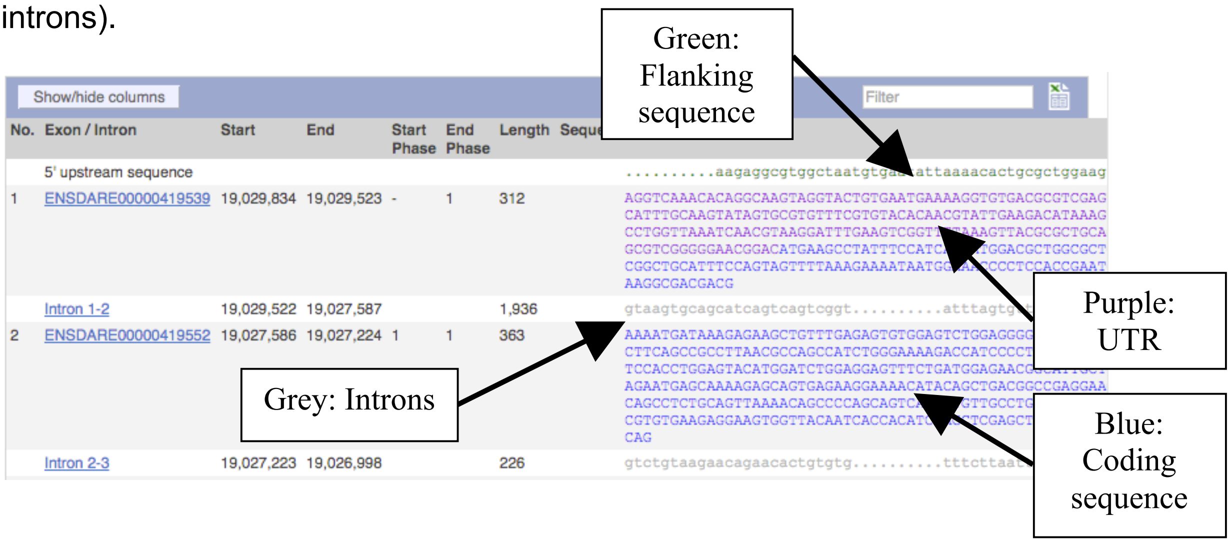
Again, the left hand navigation column provides several options for this particular transcript.



Transcript-based displays

- Summary
- Supporting evidence
- Sequence
  - Exons
  - cDNA
  - Protein
- External References
  - General identifiers
  - Oligo probes
- Ontology
  - GO graph
  - GO table
- Genetic Variation
  - Variation table
  - Variation image
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features
  - Variations
- External data
  - Personal annotation
- ID History
  - Transcript history
  - Protein history

Choose the **'Exons'** option first, which displays exon sequences in full and introns in a configurable context. Use the **'Configure this page'** link to change the display (for example, show more flanking sequence, show full introns).



Show/hide columns

Filter

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Seque
	5' upstream sequence						..........aagaggcgtggctaatgtga[illegible]attaaaacactgcgctggaag
1	ENSDARE00000419539	19,029,834	19,029,523	-	1	312	AGGTCAAACACAGGCAAGTAGGTACTGTGAATGAAAAGGTGTGACGCGTCGAG CATTTGCAAGTATAGTGCGTGTTTCGTGTACACAACGTATTGAAGACATAAAG CCTGGTTAAATCAACGTAAGGATTTGAAGTCGGTT[illegible]TAAAGTTACGCGCTGCA GCGTCGGGGAACGGACATGAAGCCTATTTCCATCA[illegible]TGGACGCTGGCGCT CGGCTGCATTTCCAGTAGTTTTAAAGAAAATAATGG[illegible]CCCCTCCACCGAAT AAGGCGACGACG
	Intron 1-2	19,029,522	19,027,587			1,936	gtaagtgcagcatcagtcagtcggt..........atttagtg
2	ENSDARE00000419552	19,027,586	19,027,224	1	1	363	AAAATGATAAAGAGAAGCTGTTTGAGAGTGTGGAGTCTGGAGGGG CTTCAGCCGCCTTAACGCCAGCCATCTGGGAAAAGACCATCCCCT TCCACCTGGAGTACATGGATCTGGAGGAGTTTCTGATGGAGAACG[illegible] AGAATGAGCAAAAGAGCAGTGAGAAGGAAAACATACAGCTGACGGCCGAGGAA CAGCCTCTGCAGTTAAAACAGCCCCAGCAGTCA[illegible]GTTGCCTG CGTGTGAAGAGGAAGTGGTTACAATCACCACATC[illegible]CTCGAGCT CAG
	Intron 2-3	19,027,223	19,026,998			226	gtctgtaagaacagaacactgtgtg..........tttcttaat

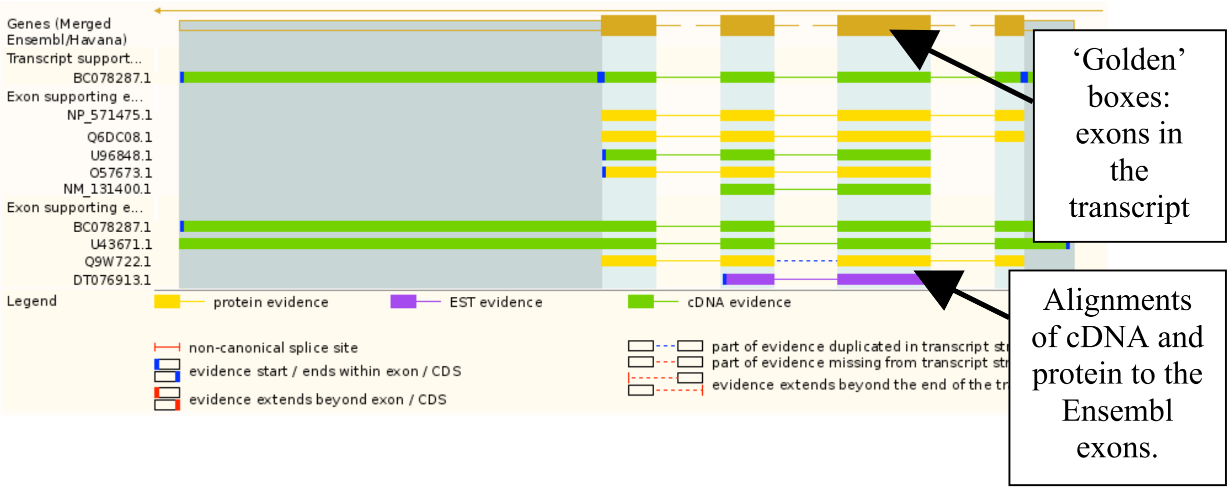
Green: Flanking sequence

Purple: UTR

Grey: Introns

Blue: Coding sequence

Next, follow the **'Supporting Evidence'** link, which shows which biological evidence has been used for the annotation of this transcript.

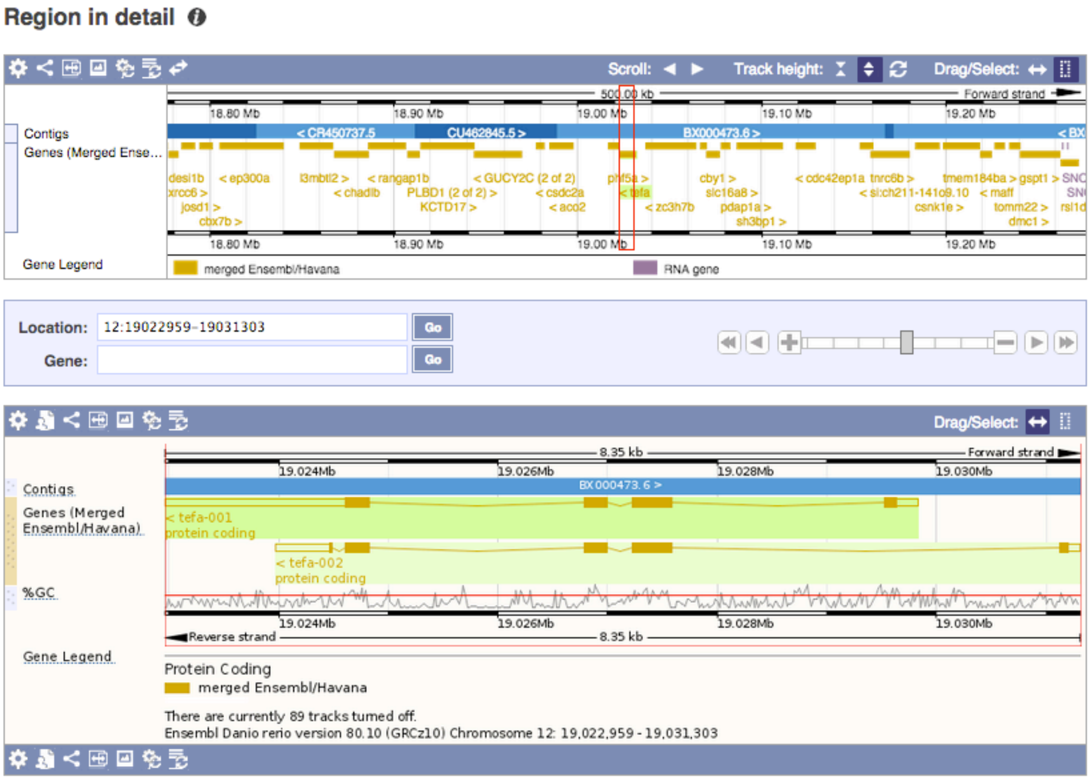

Genes (Merged Ensembl/Havana)
Transcript support...
BC078287.1
Exon supporting e...
NP_571475.1
Q6DC08.1
U96848.1
O57673.1
NM_131400.1
Exon supporting e...
BC078287.1
U43671.1
Q9W722.1
DT076913.1
Legend
protein evidence
EST evidence
cDNA evidence
non-canonical splice site
evidence start / ends within exon / CDS
evidence extends beyond exon / CDS
part of evidence duplicated in transcript str
part of evidence missing from transcript str
evidence extends beyond the end of the tr

'Golden' boxes: exons in the transcript

Alignments of cDNA and protein to the Ensembl exons.

Other transcript-specific displays include the cDNA sequence, general identifiers and gene ontology terms from the GO consortium (www.geneontology.org).

Let's now view the genomic region in which this gene and its transcript have been annotated by clicking onto the **'Location'** tab.



Region in detail

Location: 12:19022959-19031303

There are currently 89 tracks turned off.
Ensembl Danio rerio version 80.10 (GRCz10) Chromosome 12: 19,022,959 - 19,031,303

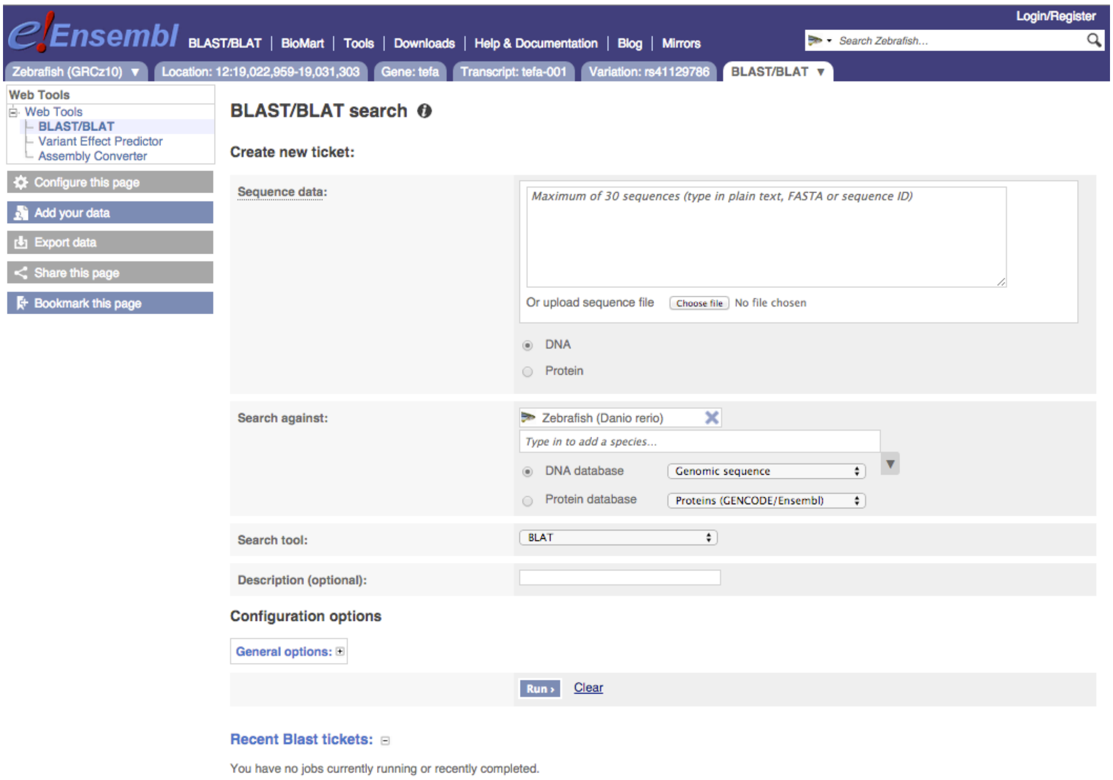
Ensembl **'Location'** displays are also highly configurable. To enter the configuration dialogue, use the **'Configure this page'** link. As an exercise, add **all variations** to the **'Region in detail'** display and view the **zebrafish cDNA** track in **'normal'** expanded form.

Rather than reconfiguring your preferred view with every new visit to the site, you can preserve your configuration by registering and logging into **your account** (upper right corner).

After investigating the **'Location'** display, we would like to export genomic sequence. Click the **'Export data'** option and select the **'FASTA'** sequence format.

```
>12 dna:chromosome chromosome:GRCz10:12:19022959:19031303:1
TAAAACAAATCAGAGTTTTTTAATAATATGTACAACATGAAACAGCCAAGGCATTCAACA
TTTACAAGTCAAAAACATCTAAAAGACAATCCAAAAAGAAAAACTTTTCAAAACGGACAA
TTAGAAAAAGACTTTTTGTACAGGAACAAATTACATATTCAACCAAGATTCATGGAGTTA
GTCGCTGTTCCTCTCAAATCAGCCTGCATTGGTTTGCTTTCTCGTCTTACACTTAGGAAG
CTTGAAGAGAAACGTCAAACTGAAGGTGCGCTTGTAAAGAGATTCTAGTACAGACCGCAA
ATATGCACATCGTAAAACCTGATTTAAGTGAAGTCATTGTGTACTAGCCACAAAAACTAA
GCTTCAGAGTTTCAATCAATTAGCTTTGGTCAAGCTCACATTACAGCAACTGCCATTCTG
AAAAAAAAACACATGAAAAAAATACAAACAGGCCTCACAATGAGTACTGCAACATTAGAC
TGCTAGCCTACTTCACAACAATACACAGAGCAAAATACACGACATATAGTGAATCTTAAG
AACTGACCCTTTATCTTCTTTCCCTGATGATAAAAAAGGACAGAGACTAAATATGTGCCT
ACAATAAAAAAAAAAGCAGGAAGGTGGCAAGCAGAGGTTTGAATTCAGTTTCTGATATGT
AACACACTAGTGCTTTTATAAACAGTCACGCTGCTCAAGACACTTATGGATCACTGCAAA
ACCACTTTTGTTAAACATTTGATGGTTATCTGAAAAGATAAATGATGCATACGAAAACGA
```

Select the header and a few lines of sequence and then follow the 'BLAST/BLAT' link in the blue header bar. Paste the sequence into the appropriate box and select 'BLAT' as the search algorithm and 'Danio_rerio' as species. Finally, click 'Run'.



BLAST/BLAT search

Create new ticket:

Sequence data: Maximum of 30 sequences (type in plain text, FASTA or sequence ID)

Or upload sequence file [Choose file] No file chosen

● DNA
○ Protein

Search against: Zebrafish (Danio rerio)
Type in to add a species…
● DNA database: Genomic sequence
○ Protein database: Proteins (GENCODE/Ensembl)

Search tool: BLAT

Description (optional):

Configuration options

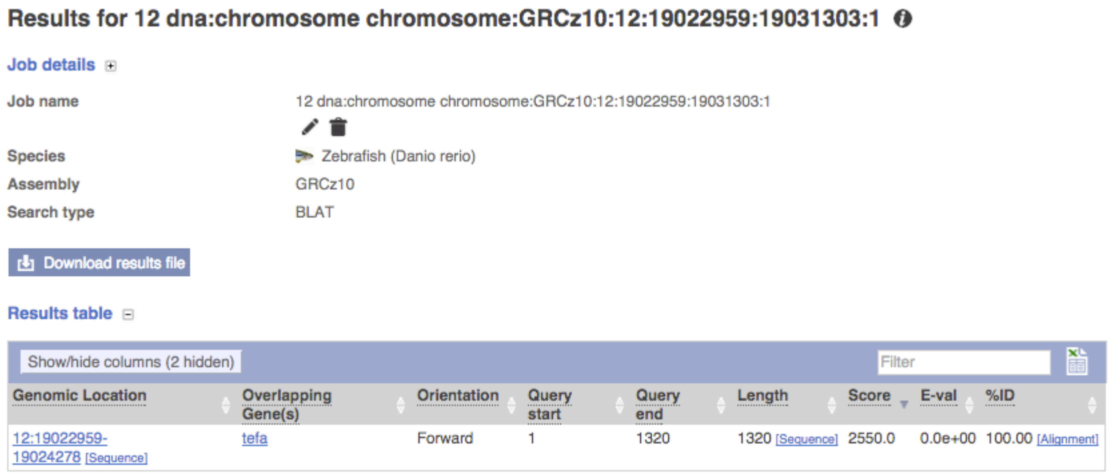
General options:

Run › Clear

Recent Blast tickets:

You have no jobs currently running or recently completed.

Wait a bit until:



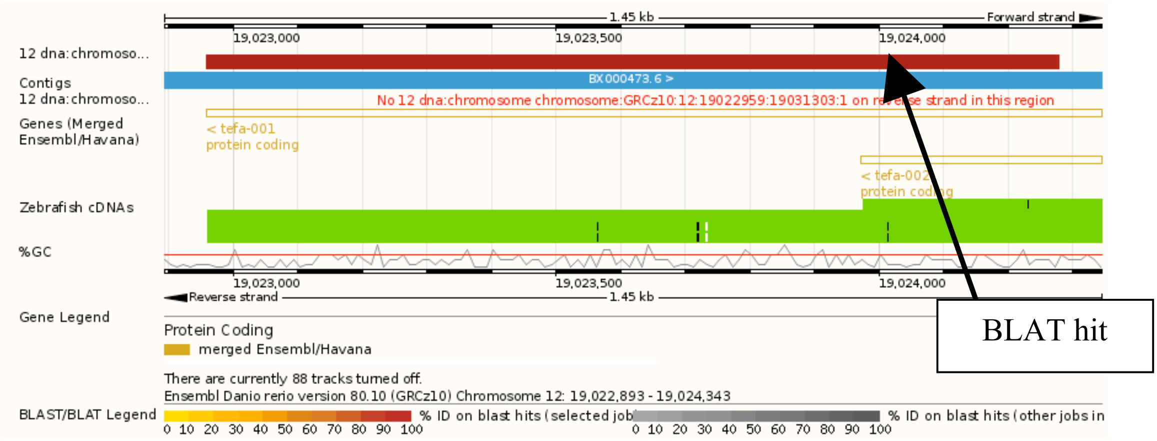
Results for 12 dna:chromosome chromosome:GRCz10:12:19022959:19031303:1

Job details

Job name: 12 dna:chromosome chromosome:GRCz10:12:19022959:19031303:1
Species: Zebrafish (Danio rerio)
Assembly: GRCz10
Search type: BLAT

Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val	%ID
12:19022959-19024278 [Sequence]	tefa	Forward	1	1320	1320 [Sequence]	2550.0	0.0e+00	100.00 [Alignment]

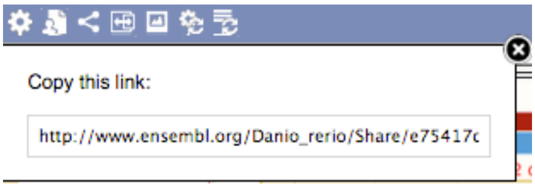
Note that the Results Table can be configured/reordered to display the desired data. Follow links to the Location View ('Genomic bp')



There are currently 88 tracks turned off.
Ensembl Danio rerio version 80.10 (GRCz10) Chromosome 12: 19,022,893 - 19,024,343

BLAT hit

Export or share the image using the links at the bottom. The ‘share’ option will preserve your configuration and might be more helpful than sharing a ‘picture’.



Copy this link:

http://www.ensembl.org/Danio_rerio/Share/e75417c

EXERCISES and ANSWERS

Note: The answers to these exercises correspond to version 80 of Ensembl featuring the GRCz10 assembly. If you use a different version and your answer doesn't correspond with the given answer, please consult the instructors. Note that certain versions are preserved on the Ensembl Archive site.

Exercise 1 – Exploring a gene

(a) Search for the zebrafish *tead1a* gene. On which chromosome is this gene located? How many transcripts (splice variants) has Ensembl annotated for it? Are these transcribed from the forward or from the reverse strand of the genome assembly?

(b) What is the longest transcript? How long is the protein it encodes? How many exons does it have? Are any of the exons completely or partially untranslated? How do the transcripts differ?

(c) Have a look at the General identifiers for one of the *tead1a* transcripts. Click on some of the links. What is the function of *tead1a*?

(d) Which PFAM domains do the proteins encoded by *tead1a* contain?

(e) Is there a human ortholog predicted for the zebrafish *tead1a* gene? What 'type' does it have? Why?

(f) If you have yourself a gene of interest, explore what information Ensembl displays about it!

Advanced questions drawing in other modules:

(1) What does ZFIN say about *tead1a*?

(2) What are the paralogs of *tead1a*?

Answers

(a)

🖱 Go to http://www.ensembl.org.
🖱 Under 'Search' select 'Zebrafish' and type 'tead1a'.
🖱 Click [Go].
🖱 On the page with search results follow 'gene' -> 'Zebrafish' and click the gene ID of *tead1a*.

The zebrafish *tead1a* gene is located on linkage group 25. Ensembl has 3 transcripts annotated for this gene. The transcripts are transcribed from the forward strand of the genome assembly.

(b)

🖱 Have a look at the transcript table at the top of the page.

The longest transcript is ENSDART00000125925. The length of this transcript is 1744 base pairs and the length of the encoded protein 422 amino acids.

🖱 Click on 'ENSDART00000125925'.
🖱 Click on 'Exons' in the side menu.

ENSDART00000125925 has 13 exons, of which the first two are untranslated and the third and the last one are partially translated.

🖱 Click on the 'Location' tab and zoom in on different areas of the transcripts
🖱 Click on the transcripts and in in the pop-up menu check the 'Analysis' entry.

tead1a-001 is longer than the other transcripts. *tead1a*-001 was annotated both by the automated Ensembl pipeline as well by manual annotation (Havana), resulting in the same structure, and was therefore merged. *tead1a*-002 was manually annotated, and found to be protein-coding. *tead1a*-003 was manually annotated and found to contain a retained intron; it is non-coding.

(c)

🖱 Click on 'General identifiers' in the side menu of a Transcript tab or 'External References' from a Gene tab.
🖱 Explore some of the links (a good place to start is 'ZFIN').

tead1a encodes an DNA-binding transcription factor involved in the hippo signalling cascade.

(d)

🖱 Select a protein.

E.g. the *tead1a*-001 protein contains a TEA/ATTS domain.

(e)

🖱 Click on the 'Gene: *tead1a*' tab.
🖱 Click on 'Orthologues' in the side menu.

There is one human ortholog predicted for zebrafish *tead1a*, TEAD1 (ENSG000000187079). It has the type '1-to-many'.

🖱 Explore the 'Help & Documentation' pages, a definition of homology types can be found at http://www.ensembl.org/info/genome/compara/homology_method.html

Human TEAD1 is the ortholog of the zebrafish genes *tead1a* and *tead1b*.

Exercise 2 – Exploring a region

(a) Go to the region from bp 33100000 to 33350000 on zebrafish chromosome 13. How many contigs make up this portion of the assembly (contigs are contiguous stretches of DNA sequence that have been assembled solely based on direct sequencing information, in the zebrafish assembly there are finished clones and whole genome shotgun contigs)?

(b) Make the tilepath clones (i.e. the BAC clones that were sequenced to generate the sequence for the human genome assembly) visible, what are the clone names in this region? Note that these clones are not shown by default! Which clone library does the clone containing the *btbd6a* gene come from?

(c) Zoom in on the *btbd6a* transcript, including a bit of flanking sequence on both sides. Which markers are located close by? Do the markers appear anywhere else in the genome?

(d) Export the genomic sequence of the region you are looking at in FASTA format.

(e) Is this region being worked on by the Genome Reference Consortium?

(f) If you have yourself a genomic region of interest, explore what information Ensembl displays about it!

Answer

(a)

🖱 Go to the Ensembl homepage.
🖱 Under 'Search Ensembl' type 'zebrafish 13: 33100000-33350000'.
🖱 Click [Go].

This genomic region is made up of 8 contigs, indicated by the alternatingly light and dark blue coloured bars in the 'Contigs' track.

(b)

🖱 Click on 'Configure this page' in the side menu.
🖱 Click on the individual contigs to see more details.
🖱 Follow the EMBL link to the submission record to find out about the name.

The tilepath clones in this region are DKEY-71P21, CH1073-380H14, CH1073-224M6 and CH1073-127N13. There are also 4 whole genome shotgun contigs, their accessions start with CABZ.

🖱 Click on 'CU855940.5', then follow the EMBL link
🖱 Read the last lines of the comments.

CU855940.5 (CH1073-380H14) is from the CHORI-1073 Zebrafish double haploid fosmid library.

(c)

🖱 Switch on the 'Markers' track under 'Sequence and Assembly'.
🖱 Draw a box around the transcript.
🖱 Click on 'Jump to region' in the pop-up menu.
🖱 Click on the markers and 'Marker info'

Gene *btbd6a* is e.g. close to the fc21e08.y1 marker. This marker is only placed in this location.

(d)

🖱 Click on 'Export data' in the side menu.
🖱 Click on [Next>].
🖱 Click on 'HTML'.

Note that the sequence has a header that provides information about the genome assembly (GRCz10), the nature of the sequence (dna), the coordinate system (chromosome), the coordinate system descriptor (13), the start and end coordinates (e.g. 33249320:33297062) and the strand (1):

>13 dna:chromosome chromosome:GRCz10:13:33249320:33297062:1

(e)

🖰 Go to 'Configure this page' and switch on 'Sequence and assembly' -> 'GRC alignments' and switch on 'Genome curation'

This reveals that there is indeed an issue with the assembly here, registered as ZG-6933. The report states that a gap exists between clones components BX284673.9 and CU855940.5. The GRC is currently awaiting sequence data to close the gap that is filled with WGS contigs.

Data mining

On top of visualising genome data, all browsers offer data to be extracted and stored. This ranges from a simple download of sequence or features for a certain region to genome-wide pre-prepared data collections. However, if you are only interested in data that passes certain filters it becomes trickier. Sometimes, this can be extracted from the offered downloads, but this might require advanced bioinformatics skills and additional data.

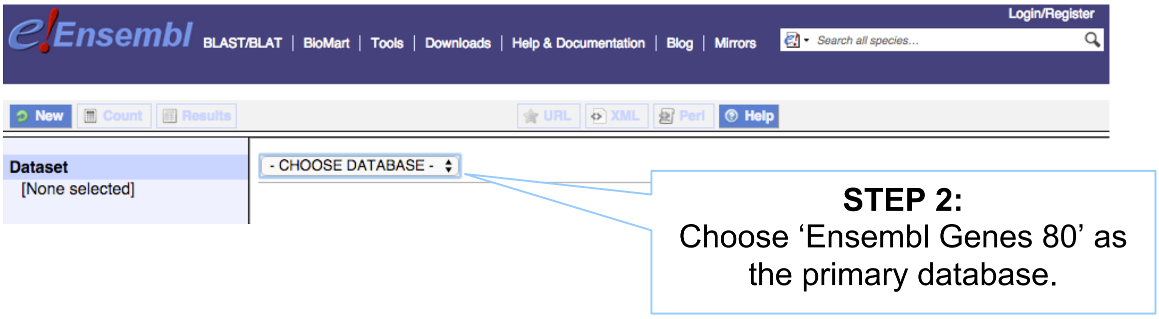
UCSC and Ensembl offer a service to filter and extract data according to your needs. UCSC provides the **Table Browser** (https://genome.ucsc.edu/goldenPath/help/hgTablesHelp.html) and Ensembl provides **Biomart** which we explain in detail below.

Mining data using BioMart - worked example

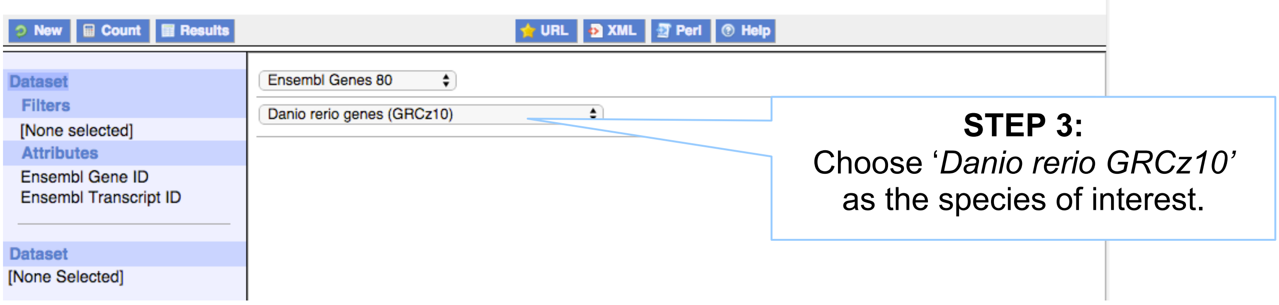
- Find all protein-coding zebrafish genes on linkage group 1 that have a human orthologue.
- Display the Ensembl IDs of the zebrafish and human genes plus the chromosomal location of the human gene.
- Download the sequence of all available 5’ UTRs of these genes.

Note that the below example was created on Ensembl version 80. Since the gene set gets adapted to the ongoing manual gene annotation with every other release, the results might differ with a different release, even on the same genome assembly.

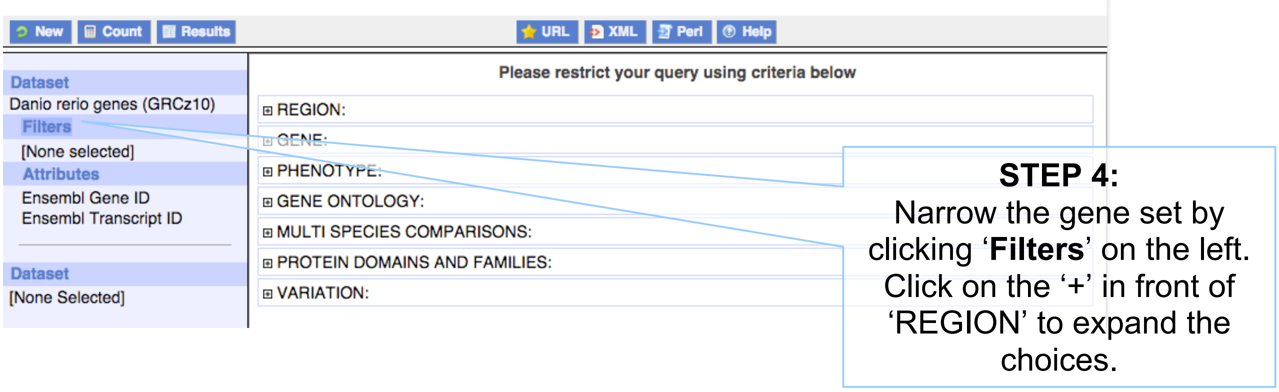
STEP 1: Click on ’BioMart’ in the top header bar of the Ensembl home page.



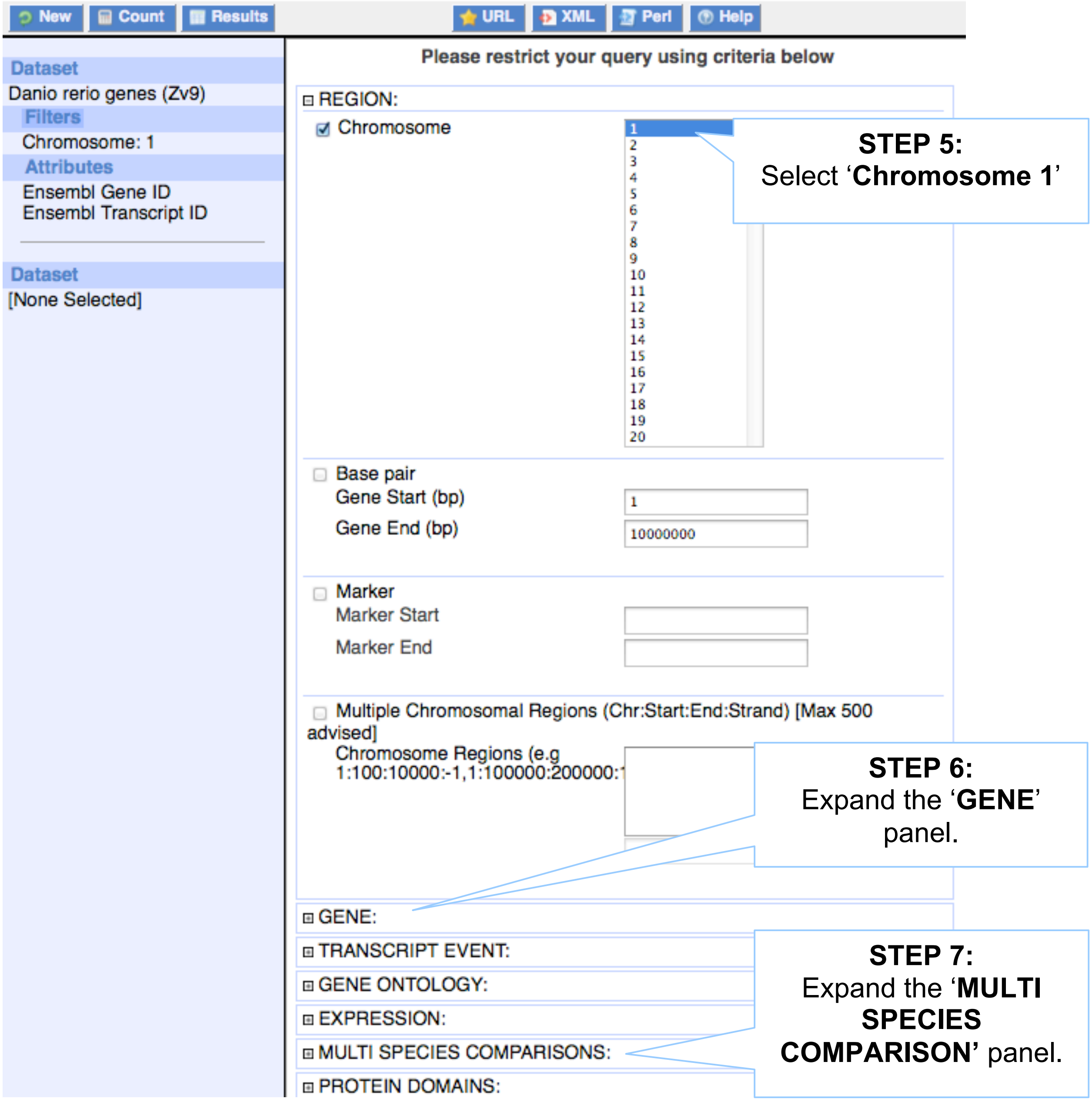
STEP 2:
Choose ‘Ensembl Genes 80’ as the primary database.



STEP 3:
Choose ‘*Danio rerio GRCz10’* as the species of interest.



STEP 4:
Narrow the gene set by clicking ‘**Filters**’ on the left. Click on the ‘+’ in front of ‘REGION’ to expand the choices.



Dataset
Danio rerio genes (Zv9)
Filters
Chromosome: 1
Attributes
Ensembl Gene ID
Ensembl Transcript ID

Dataset
[None Selected]

Please restrict your query using criteria below

⊟ REGION:

☑ Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

☐ Base pair
Gene Start (bp): 1
Gene End (bp): 10000000

☐ Marker
Marker Start
Marker End

☐ Multiple Chromosomal Regions (Chr:Start:End:Strand) [Max 500 advised]
Chromosome Regions (e.g 1:100:10000:-1,1:100000:200000:1

⊞ GENE:

⊞ TRANSCRIPT EVENT:

⊞ GENE ONTOLOGY:

⊞ EXPRESSION:

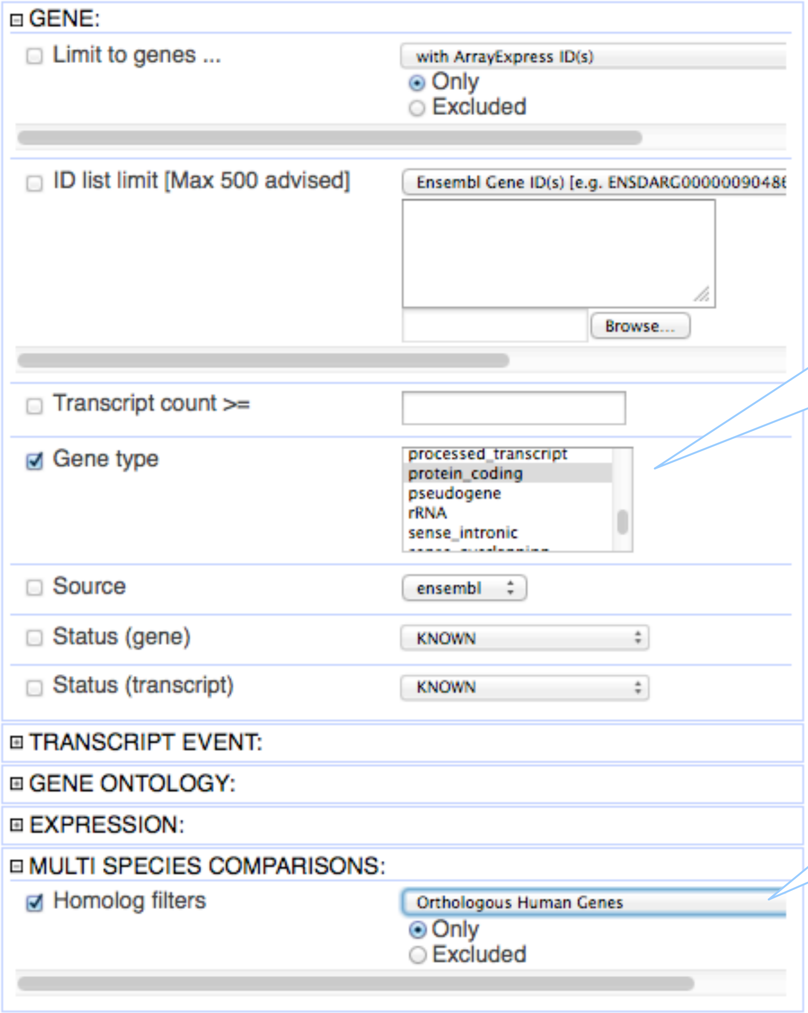
⊞ MULTI SPECIES COMPARISONS:

⊞ PROTEIN DOMAINS:

STEP 5:
Select ‘**Chromosome 1**’

STEP 6:
Expand the ‘**GENE**’ panel.

STEP 7:
Expand the ‘**MULTI SPECIES COMPARISON’** panel.



⊟ GENE:

□ Limit to genes ... — with ArrayExpress ID(s) ● Only ○ Excluded

□ ID list limit [Max 500 advised] — Ensembl Gene ID(s) [e.g. ENSDARG0000009048 — Browse...

□ Transcript count >=

☑ Gene type — processed_transcript, protein_coding, pseudogene, rRNA, sense_intronic

□ Source — ensembl

□ Status (gene) — KNOWN

□ Status (transcript) — KNOWN

⊞ TRANSCRIPT EVENT:

⊞ GENE ONTOLOGY:

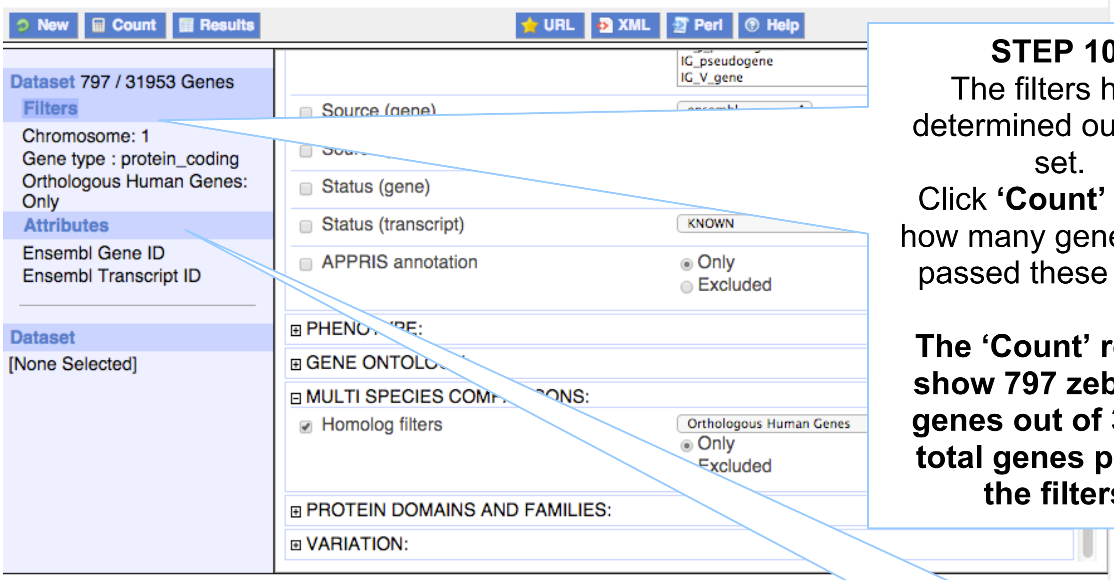
⊞ EXPRESSION:

⊟ MULTI SPECIES COMPARISONS:

☑ Homolog filters — Orthologous Human Genes ● Only ○ Excluded

STEP 8:
Limit to genes of type **'protein coding'**

STEP 9:
Limit to **'Orthologous Human Genes Only'**



Dataset 797 / 31953 Genes

Filters

Chromosome: 1
Gene type : protein_coding
Orthologous Human Genes: Only

Attributes

Ensembl Gene ID
Ensembl Transcript ID

Dataset
[None Selected]

IG_pseudogene
IG_V_gene

□ Source (gene)

□ Status (gene)

□ Status (transcript) — KNOWN

□ APPRIS annotation — ● Only ○ Excluded

⊞ PHENO RE:

⊞ GENE ONTOLO

⊟ MULTI SPECIES COM ONS:

☑ Homolog filters — Orthologous Human Genes ● Only Excluded

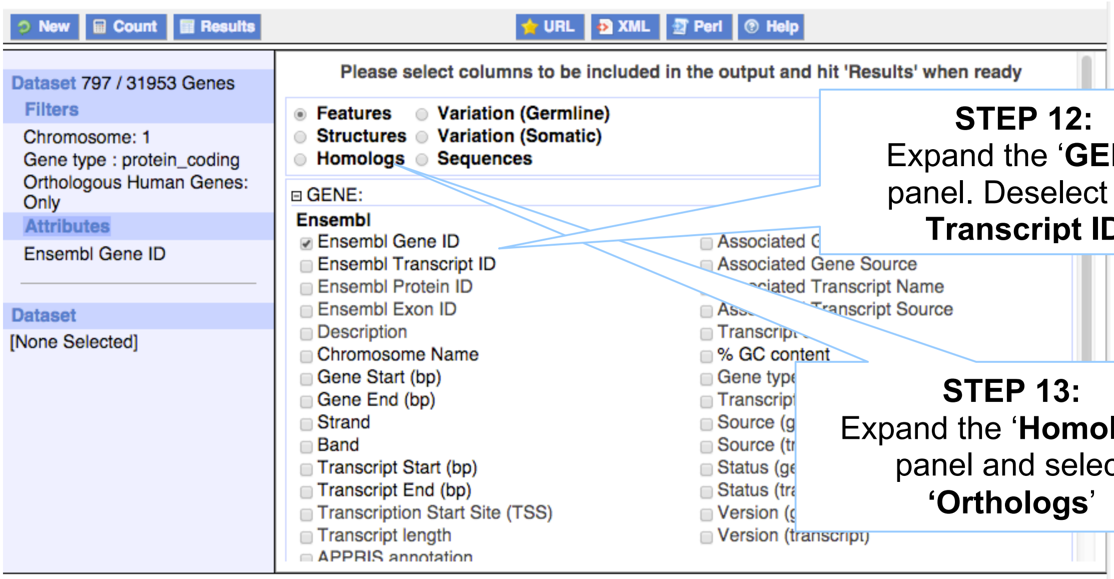
⊞ PROTEIN DOMAINS AND FAMILIES:

⊞ VARIATION:

STEP 10:
The filters have determined our gene set.
Click **'Count'** to see how many genes have passed these filters.

The 'Count' results show 797 zebrafish genes out of 31,953 total genes passed the filters.

STEP 11:
Click on **'Attributes'** to select output options (i.e. what we would like to know about our gene set).



Dataset 797 / 31953 Genes

Filters

Chromosome: 1
Gene type : protein_coding
Orthologous Human Genes: Only

Attributes

Ensembl Gene ID

Dataset

[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

● Features ○ Variation (Germline)
○ Structures ○ Variation (Somatic)
○ Homologs ○ Sequences

⊟ GENE:

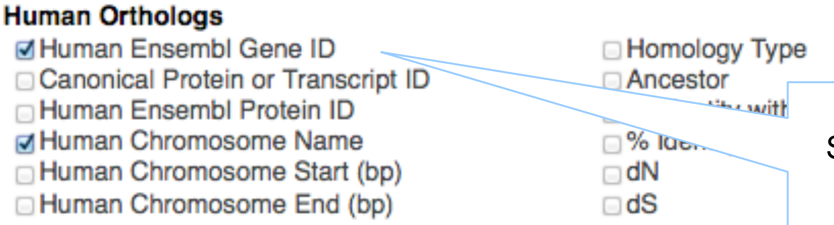
Ensembl

☑ Ensembl Gene ID
□ Ensembl Transcript ID
□ Ensembl Protein ID
□ Ensembl Exon ID
□ Description
□ Chromosome Name
□ Gene Start (bp)
□ Gene End (bp)
□ Strand
□ Band
□ Transcript Start (bp)
□ Transcript End (bp)
□ Transcription Start Site (TSS)
□ Transcript length
□ APPRIS annotation

□ Associated G
□ Associated Gene Source
□ ciated Transcript Name
□ Ass Transcript Source
□ Transcript
□ % GC content
□ Gene type
□ Transcript
□ Source (g
□ Source (tr
□ Status (ge
□ Status (tra
□ Version (g
□ Version (transcript)

STEP 12:
Expand the '**GENE**' panel. Deselect the **Transcript ID**

STEP 13:
Expand the '**Homologs**' panel and select '**Orthologs**'

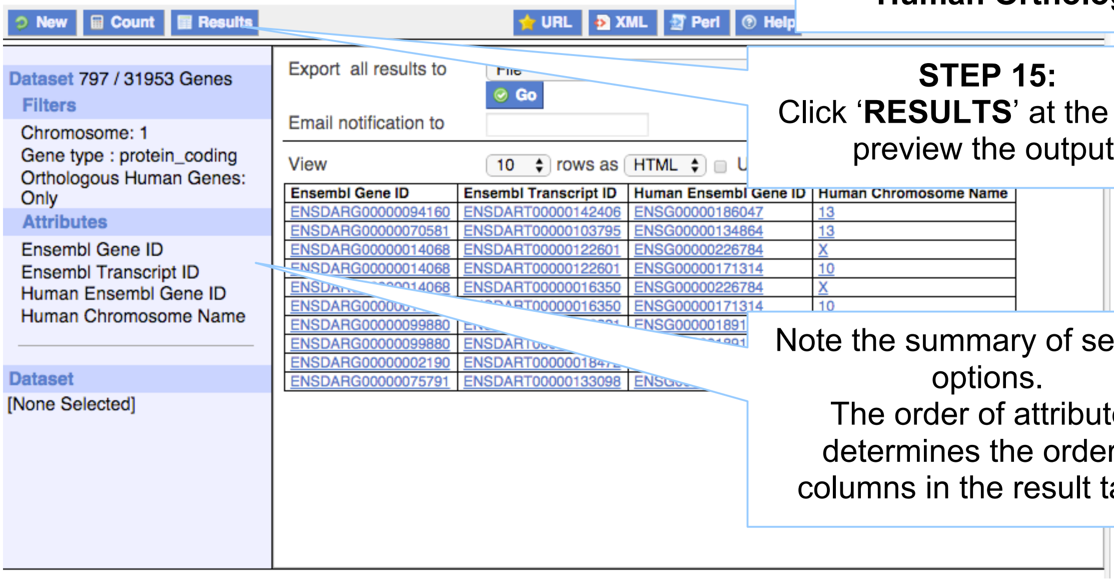


Human Orthologs

☑ Human Ensembl Gene ID
□ Canonical Protein or Transcript ID
□ Human Ensembl Protein ID
☑ Human Chromosome Name
□ Human Chromosome Start (bp)
□ Human Chromosome End (bp)

□ Homology Type
□ Ancestor
□ % Ider
□ dN
□ dS

STEP 14:
Select '**Human Ensembl Gene ID**' and '**Human Chromosome' from 'Human Orthologs'**



Dataset 797 / 31953 Genes

Filters

Chromosome: 1
Gene type : protein_coding
Orthologous Human Genes: Only

Attributes

Ensembl Gene ID
Ensembl Transcript ID
Human Ensembl Gene ID
Human Chromosome Name

Dataset

[None Selected]

Export all results to File
Go

Email notification to

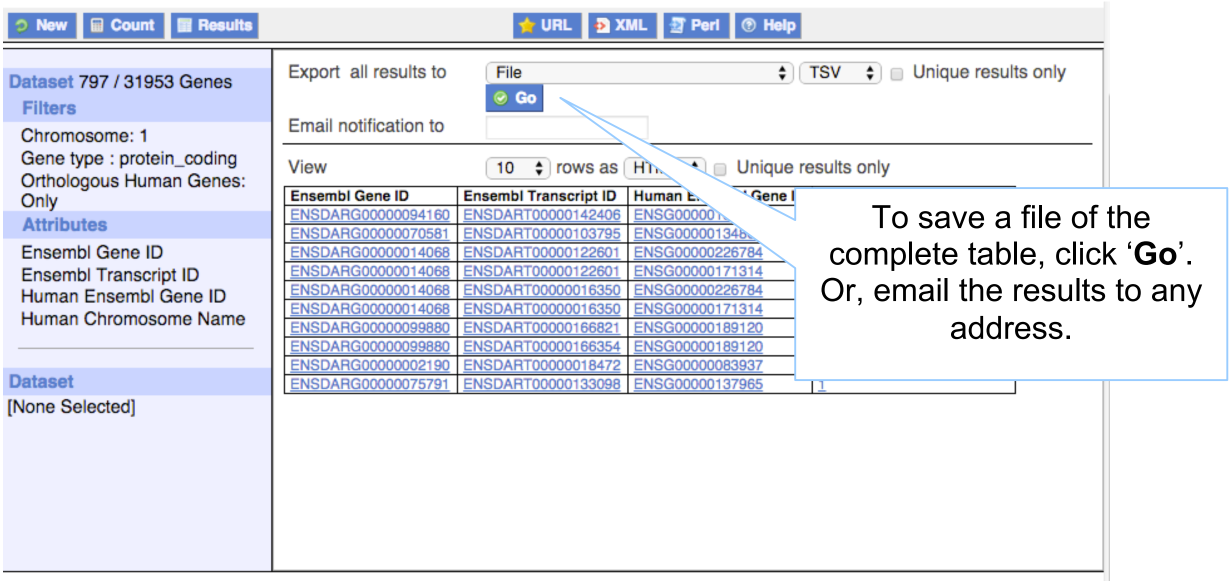
View 10 rows as HTML □ U

Ensembl Gene ID	Ensembl Transcript ID	Human Ensembl Gene ID	Human Chromosome Name
ENSDARG00000094160	ENSDART00000142406	ENSG00000186047	13
ENSDARG00000070581	ENSDART00000103795	ENSG00000134864	13
ENSDARG00000014068	ENSDART00000122601	ENSG00000226784	X
ENSDARG00000014068	ENSDART00000122601	ENSG00000171314	10
ENSDA[illegible]14068	ENSDART00000016350	ENSG00000226784	X
ENSDARG000000[illegible]	[illegible]ART00000016350	ENSG00000171314	10
ENSDARG00000099880	[illegible]	ENSG000001891[illegible]	[illegible]
ENSDARG00000099880	ENSDART[illegible]	[illegible]	[illegible]
ENSDARG00000002190	ENSDART0000001847[illegible]	[illegible]	[illegible]
ENSDARG00000075791	ENSDART00000133098	ENSG[illegible]	[illegible]

STEP 15:
Click '**RESULTS**' at the top to preview the output.

Note the summary of selected options.
The order of attributes determines the order of columns in the result table.

And here you have the first 10 results; you can change the number of displayed results in the drop down menu. Expanded to 'all' this gives you a nice overview of possible syntenic regions in the two genomes.



Dataset 797 / 31953 Genes
Filters
Chromosome: 1
Gene type : protein_coding
Orthologous Human Genes: Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Human Ensembl Gene ID
Human Chromosome Name
Dataset
[None Selected]

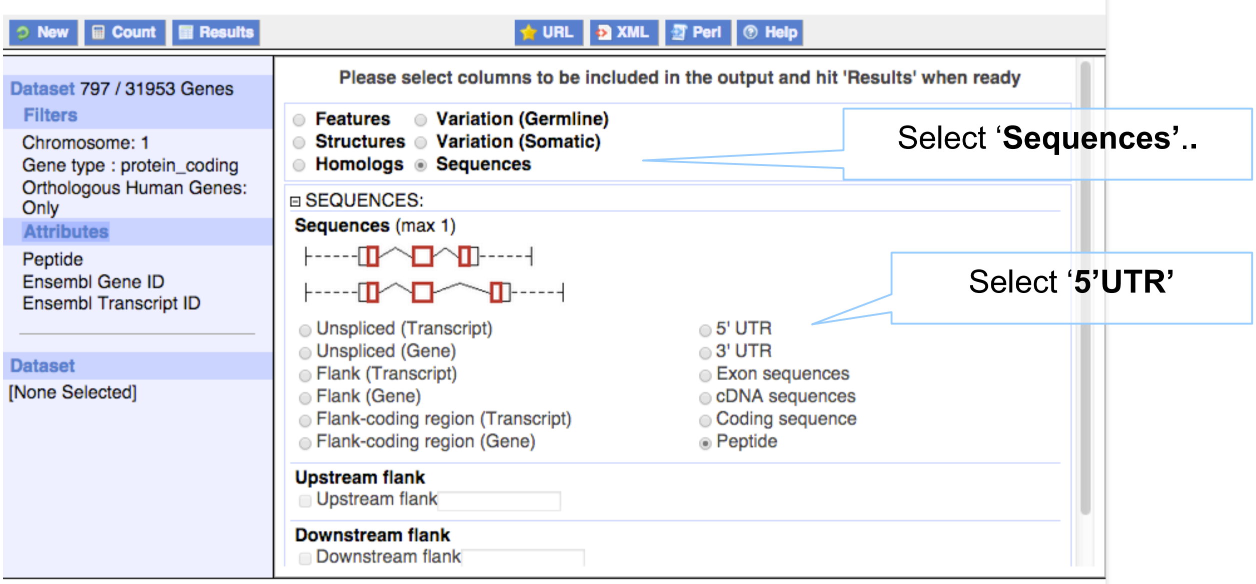
Export all results to File TSV □ Unique results only
Go
Email notification to

View 10 rows as HT[ML] □ Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Human E[nsembl] Gene I[D]	
ENSDARG00000094160	ENSDART00000142406	ENSG00000	
ENSDARG00000070581	ENSDART00000103795	ENSG00001348	
ENSDARG00000014068	ENSDART00000122601	ENSG00000226784	
ENSDARG00000014068	ENSDART00000122601	ENSG00000171314	
ENSDARG00000014068	ENSDART00000016350	ENSG00000226784	
ENSDARG00000014068	ENSDART00000016350	ENSG00000171314	
ENSDARG00000099880	ENSDART00000166821	ENSG00000189120	
ENSDARG00000099880	ENSDART00000166354	ENSG00000189120	
ENSDARG00000002190	ENSDART00000018472	ENSG00000083937	
ENSDARG00000075791	ENSDART00000133098	ENSG00000137965	1

To save a file of the complete table, click '**Go**'. Or, email the results to any address.

In order to obtain all 5'UTRs of these genes, go back to the '**Attributes**'.



Dataset 797 / 31953 Genes
Filters
Chromosome: 1
Gene type : protein_coding
Orthologous Human Genes: Only
Attributes
Peptide
Ensembl Gene ID
Ensembl Transcript ID
Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

○ Features ○ Variation (Germline)
○ Structures ○ Variation (Somatic)
○ Homologs ● Sequences

⊟ SEQUENCES:
Sequences (max 1)

○ Unspliced (Transcript) ○ 5' UTR
○ Unspliced (Gene) ○ 3' UTR
○ Flank (Transcript) ○ Exon sequences
○ Flank (Gene) ○ cDNA sequences
○ Flank-coding region (Transcript) ○ Coding sequence
○ Flank-coding region (Gene) ● Peptide

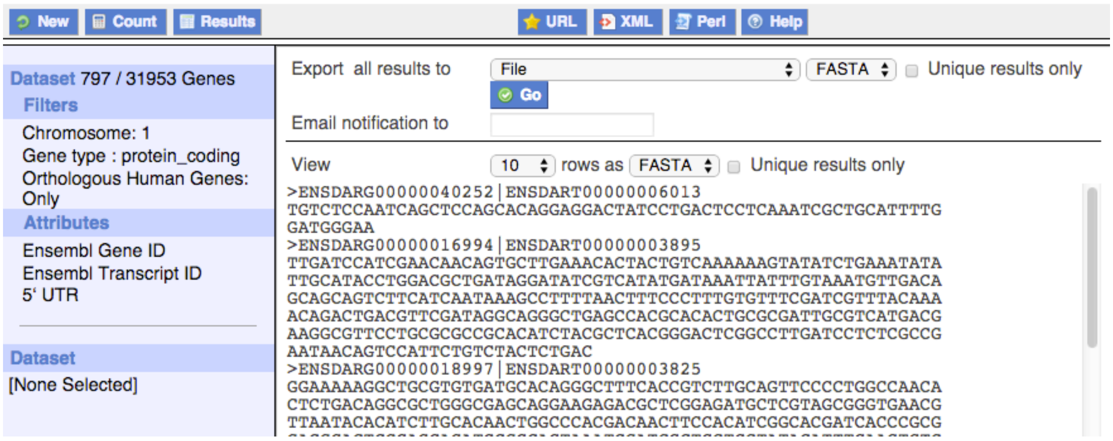
Upstream flank
□ Upstream flank

Downstream flank
□ Downstream flank

Select '**Sequences**'..

Select '**5'UTR**'

Click '**Results**' and you will get the required list. Note that not all genes have 5' UTRs annotated.



Dataset 797 / 31953 Genes
Filters
Chromosome: 1
Gene type : protein_coding
Orthologous Human Genes: Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
5' UTR
Dataset
[None Selected]

Export all results to File FASTA □ Unique results only
Go
Email notification to

View 10 rows as FASTA □ Unique results only

```
>ENSDARG00000040252|ENSDART00000006013
TGTCTCCAATCAGCTCCAGCACAGGAGGACTATCCTGACTCCTCAAATCGCTGCATTTTG
GATGGGAA
>ENSDARG00000016994|ENSDART00000003895
TTGATCCATCGAACAACAGTGCTTGAAACACTACTGTCAAAAAAGTATATCTGAAATATA
TTGCATACCTGGACGCTGATAGGATATCGTCATATGATAAATTATTTGTAAATGTTGACA
GCAGCAGTCTTCATCAATAAAGCCTTTTAACTTTCCCTTTGTGTTTCGATCGTTTACAAA
ACAGACTGACGTTCGATAGGCAGGGCTGAGCCACGCACACTGCGCGATTGCGTCATGACG
AAGGCGTTCCTGCGCGCCGCACATCTACGCTCACGGGACTCGGCCTTGATCCTCTCGCCG
AATAACAGTCCATTCTGTCTACTCTGAC
>ENSDARG00000018997|ENSDART00000003825
GGAAAAAGGCTGCGTGTGATGCACAGGGCTTTCACCGTCTTGCAGTTCCCCTGGCCAACA
CTCTGACAGGCGCTGGGCGAGCAGGAAGAGACGCTCGGAGATGCTCGTAGCGGGTGAACG
TTAATACACATCTTGCACAACTGGCCCACGACAACTTCCACATCGGCACGATCACCCGCG
```

EXERCISES and ANSWERS

Note: The answers to these exercises correspond to version 80 of Ensembl.

Exercise 1

Generate a list of all zebrafish protein coding genes on chr1 with a ZFIN ID that have more than one splice variants and that are causing the caudal fin to be absent when mutated. Download the peptide sequences and make sure the header states the Ensembl ID, a description, the associated gene name and the associated gene DB.

Answer

🖱 Go to the Ensembl homepage.
🖱 Click the BioMart link on the toolbar.

Start with all the zebrafish Ensembl genes:

🖱 Choose the 'Ensembl 80' database.
🖱 Choose the 'Danio rerio genes GRCz10' dataset.

Now filter for the genes on chromosome 1:

🖱 Click on 'Filters' in the left panel.
🖱 Expand the 'REGION' section by clicking on the + box.
🖱 Select 'Chromosome – 1'. Make sure the check box in front of the filter is ticked, otherwise the filter won't work.
🖱 Click the [Count] button on the toolbar.

This should give you 1,386 / 31,953 Genes. Now for genes with a ZFIN ID

🖱 Click 'Limit to genes…' and choose 'with ZFIN IDs'
🖱 Click the [Count] button on the toolbar.

1,186 genes have ZFIN IDs.

Now filter further for genes that are protein coding:

🖱 Expand the 'GENE' section by clicking on the + box.
🖱 Select 'Gene type – protein_coding'.
🖱 Click the [Count] button on the toolbar.

This should give you 1,081 / 31,953 Genes.

Now only select those genes with at least 2 alternative splice variants.

🖱 Expand the 'GENE' section again by clicking on the + box.
🖱 Select "Transcript count >=' and enter '2'
🖱 Click the [Count] button on the toolbar.

661 genes left. Let's see what disrupts the caudal fin development:

🖱 Expand the 'Phenotype' filter.
🖱 Select 'caudal fin absent' under 'Phenotype'.
🖱 Select 'ZFIN' under 'Phenotype source'.

One gene left!

Now download the cDNA sequences with the Ensembl gene and transcript IDs, the associated gene name, gene DB and a description.

🖱 Click on 'Attributes' in the left panel.
🖱 Select the 'Sequences' attributes page.
🖱 Select 'cDNA'.
🖱 Expand the 'Header Information' section and select 'Ensembl Gene ID', 'Description', 'Associated Gene Name' and 'Associated Gene Source' plus 'Ensembl Transcript ID'
🖱 Click the [Results] button on the toolbar.

```
>ENSDARG00000031894|ENSDART00000047876|lymphoid enhancer-binding factor 1
[Source:ZFIN;Acc:ZDB-GENE-990714-26]|lef1|ZFIN
GGAGCACGACACAGACCTGATGCACATGAAACCTCAGCACGAGCAGAGAAAGGAGCAGGA
GCCCAAAAGACCTCACATCAAGAAACCTCTAAACGCTTTCATGCTGTATATGAAAGAGAT
GCGCGCCAATGTGGTGGCCGAATGCACGCTGAAGGAGAGCGCCGCTATCAATCAGATCCT
CGGCCGGAGGTGGCATGCTTTATCTCGGGAAGAGCAAGCTAAGTATTACGAATTAGCCCG
CAAGGAACGGCAGCTCCATATGCAGCTTTACCCAGGATGGTCTGCCAGAGACAATTATGG
AAAGAAAAAAAAGCGGAAGAGGGAAAAGATCCAGGAACCTGCTTCAGATGGAAATGGCTT
TTTCTTTTATGGAACACAAAAGGTACAGGCCAGAGAATGAAAACGGCGTACATCTGAACA ATGGTAAGAG
>ENSDARG00000031894|ENSDART00000132405|lymphoid enhancer-binding factor 1
[Source:ZFIN;Acc:ZDB-GENE-990714-26]|lef1|ZFIN
GTAGTCAGTCAGAGATCAGGGGGAGGAGTACAGCACTACACTCTCTCCAGCCCAACATTA
CTCTCAGTCTCTGCTGAGCTCATTTCTGAAGAGGGACACCTTTTTTACCCAACAAACCAA
ACGGGAATGACACACACCATCTGAACTCCAACATTTCTTTTTTTGTTGTTGTTGCTTTTA
TTTTGAAACAAGTGAAACTGTCCTTTTCTGAACTTTAAGTTCCAACTTTTTCCTTCCACC
AAAGGATTCGTATTTTAACTTTTTCCCCAAACCCGCTATTTTTCTTCCTCTGGATTCCCG
AGAGTTTTTCCACCGGACGCGCGCGCTCTTGTTACCGTAAACCAAACACACTCACGCGCG
CGTGTTCGGAGTGCGCGAGCTGACCAGAAACAAAACAACTATACGGGGGGTTTAATTTCA
ATTGCACGCGTTTGCGTCCCTGGCGTTTGTAGGGTGAGGAGGACTTTCATTCACCCGAGA

…
```

Exercise 2

BioMart is a very handy tool when you want to map IDs between different databases. The following is a list of 29 IDs of human proteins from the RefSeq database of NCBI (http://www.ncbi.nlm.nih.gov/projects/RefSeq/):

NP_001218, NP_203125, NP_203124, NP_203126, NP_001007233, NP_150636, NP_150635, NP_001214, NP_150637, NP_150634, NP_150649, NP_001216, NP_116787, NP_001217, NP_127463, NP_001220, NP_004338, NP_004337, NP_116786, NP_036246, NP_116756, NP_116759, NP_001221, NP_203519, NP_001073594, NP_001219, NP_001073593, NP_203520, NP_203522

Generate a list that shows to which Ensembl Gene IDs and to which HGNC symbols these RefSeq IDs correspond. Which of these genes have a zebrafish ortholog?

Answer

🖱 Click [New].
🖱 Choose the 'Ensembl 80' database.
🖱 Choose the 'Homo sapiens genes (GRCh38.p2)' dataset.

🖱 Click on 'Filters' in the left panel.
🖱 Expand the 'GENE' section by clicking on the + box.
🖱 Select 'Input external references ID list - Refseq protein ID(s)'
🖱 Enter the list of IDs in the text box below (either comma separated or as a list).

🖱 Click on 'Attributes' in the left panel.
🖱 Select the 'Features' attributes page.
🖱 Expand the 'GENE' section by clicking on the + box.
🖱 Deselect 'Ensembl Transcript ID'.
🖱 Expand the 'External' section by clicking on the + box.
🖱 Select 'HGNC symbol' and 'RefSeq Protein ID'.

🖱 Click the [Results] button on the toolbar.
🖱 Select 'View All rows as HTML' or export all results to a file. Tick the box 'Unique results only'.

Note: BioMart is 'transcript-centric', which means that it will give a separate row of output for each transcript of a gene, even if you don't include the Ensembl Transcript ID in your output. When you don't want this, use the 'Unique results only' option.

Your results should show 11 genes, most of them Caspase (CASP) genes. Several RefSeq IDs map to the same Ensembl Gene ID and HGNC symbol.

Now narrow down to genes with zebrafish orthologs.

🖰 Click on 'Filters' in the left panel.
🖰 Expand the 'MULTI SPECIES COMPARISONS' section by clicking on the + box.
🖰 Select 'Homolog filters' and select 'Orthologous Zebrafish Genes - Only'
🖰 Click the [Count] button on the toolbar.

You will be left with 8 genes.

Exercise 3

Generate a list of all zebrafish genes on chr 1 that have an human ortholog on human chr 13. Display the gene names, are they the same? Note: This requires you to select an additional data set.

🖰 Choose database 'Ensembl 80' and dataset 'Danio rerio genes (GRCz10).
🖰 Narrow down by filtering for 'REGION' 'Chromosome - 1' and 'MULTI SPECIES COMPARISONS' selecting 'Homolog Filters' 'Orthologuos Human genes -Only'
🖰 Click on "Attributes', then 'Features' , deselect 'Ensembl Transcript ID', select 'Associated Gene Name'

🖰 Click on 'Dataset' (bottom left) and select '[Ensembl 80] Homo sapiens genes (GRCh38.p2)'
🖰 Narrow down by filtering for 'REGION' 'Chromosome - 13' and 'MULTI SPECIES COMPARISONS' selecting 'Homolog Filters' 'Orthologuos Zebrafish genes -Only'
🖰 Click on "Attributes', then 'Features' , deselect 'Ensembl Transcript ID', select 'Associated Gene Name'

You will end up with a list where quite a lot of names are identical in zebrafish and human. Note that the uppercase zebrafish gene names are projected from the most likely human ortholog whereas lowercase names are given by ZFIN.

Exercise 4

Design your own query!

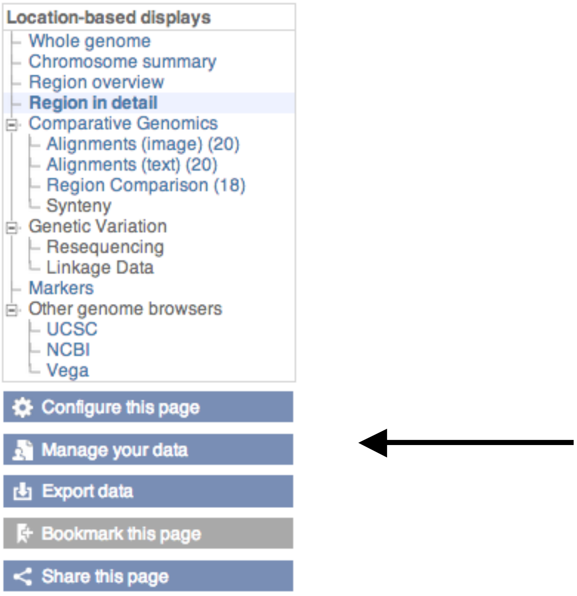
Make your own data visible with BED files

BED files have a very simple format (tab delimited, only chr, start and stop required) and can be quickly created to provide a list of features you want to make visible in a genome browser like e.g. Ensembl or the UCSC browser. The BED format also allows for optional extension, which can make these data collections very powerful tools, even providing whole trackhubs. The format is described in detail at http://genome.ucsc.edu/FAQ/FAQformat.html#format1

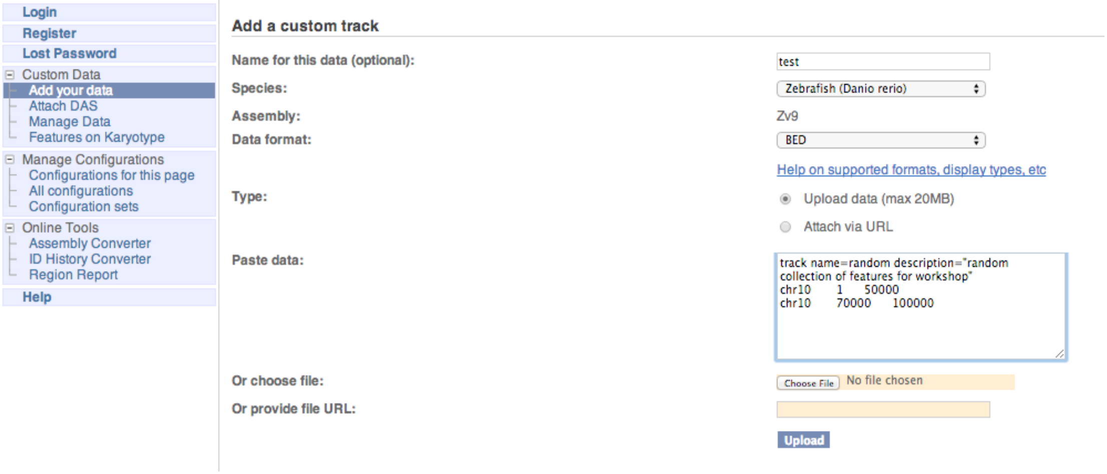
Here is an example for a bed file content (the header is optional):

```
track name=random description="random collection of features for
workshop"
chr10 0      50000 gene1
chr10 69999 100000      gene2
chr10 109999      120000      gene2
```

Go to the Ensembl browser, bring up a zebrafish location view and choose ‘Add your data’ and select ‘Data format – BED’



Paste the bed file data and upload.



Add a custom track

Name for this data (optional): test

Species: Zebrafish (Danio rerio)

Assembly: Zv9

Data format: BED

Type: ● Upload data (max 20MB) ○ Attach via URL

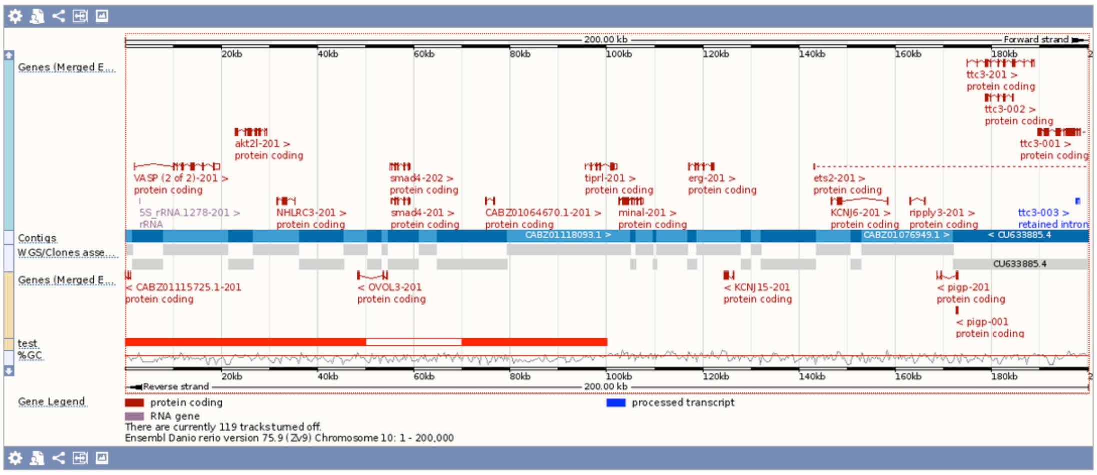
Paste data:

```
track name=random description="random
collection of features for workshop"
chr10     1     50000
chr10     70000     100000
```

Or choose file: No file chosen

Or provide file URL:

Navigate to the start of chromosome 10 and you should see the two features added.

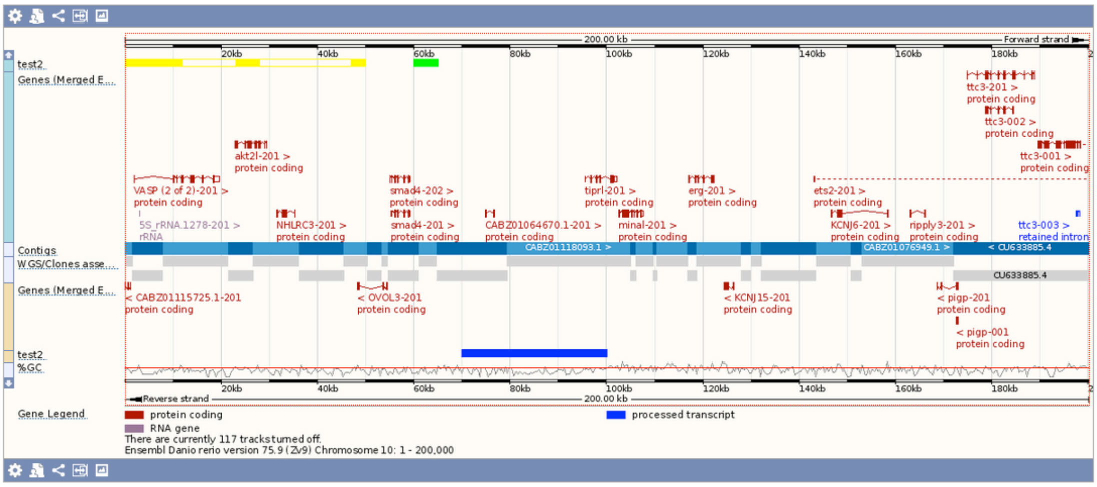


There are currently 119 tracks turned off.
Ensembl Danio rerio version 75.9 (Zv9) Chromosome 10: 1 - 200,000

Here is a more advanced example:

```
track name=random description="random collection of features for
workshop" useScore=1 itemRgb="On"
chr10 0      50000 feature1     1      +      0      50000 255,255,0    3
      12000,5000,3000    0,22999,46999
chr10 59999 65000 feature2     0.5    +      59999 65000 0,255,0
chr10 69999 100000      feature3     1      -      69999 100000
      0,0,255
```

resulting in



🟥 protein coding

🟪 RNA gene

🟦 processed transcript

There are currently 117 tracks turned off.
Ensembl Danio rerio version 75.9 (Zv9) Chromosome 10: 1 - 200,000

Exercises:

1. Play around with the data. What do you need to do to display separate items and what to display exon/intron like structures?
2. Create your own data collection in a bed file and make it visible in Ensembl.
3. Try to query biomart for a certain range of features and adapt the output to bed format. A text editor should help with e.g. the ‘chr’ prefix. Make the features visible in Ensembl.
4. Where exactly does gene1 in the first example start and end? Why?

Answers:

1. Column 4 accepts a name for the feature. If you give the same name for more than one feature, those with the same name will be drawn in exon/intron style.

4. The bed format is requiring a 0-based start and a 1-based end, so using the range of 0-50,000 in our example bed file specifies to use the bp starting after 0 (i.e. bp 1), ending at 50,000. A good explanation of coordinate systems can be found here: https://www.biostars.org/p/84686/