
Module 2: Genome Browsing

Aims

- Explain why it can be useful to look at the whole genome.
- Discuss how genes and other features can be predicted and displayed.
- 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.
- Examples (include location and structure of a known gene and its products; information about a defined chromosomal region; convenient export of selected information).

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 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. cytogenetic bands, markers, SNPs, repeated sequences, regions homologous to other species) and displayed alongside the genomic sequence assembly. Protein,

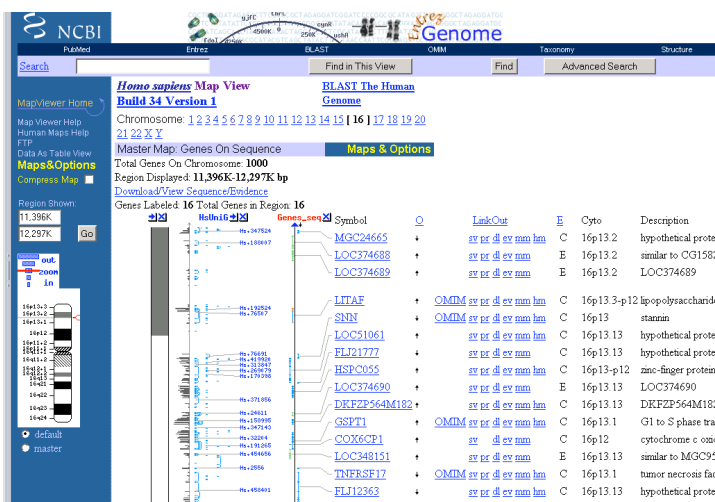
mRNA and EST entries from various sequence databases may also be shown 'mapped' onto the chromosomes.

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

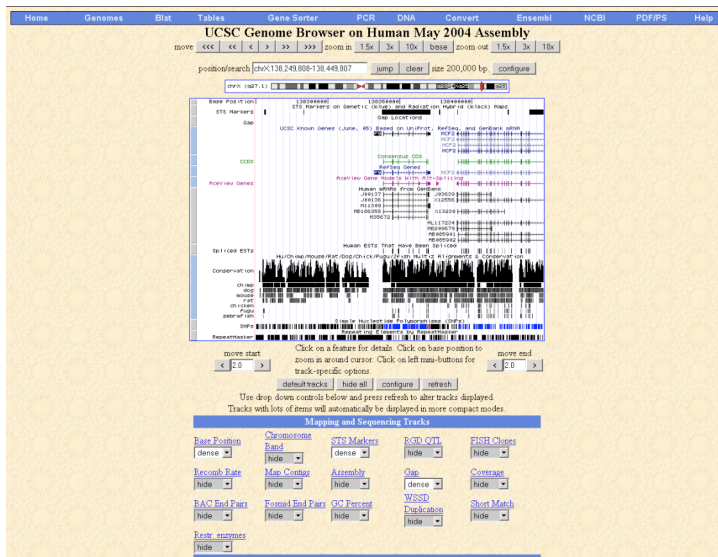
Browsers (and some of their strengths)

- **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>



NCBI Map Viewer

- Good integration with other NCBI resources**



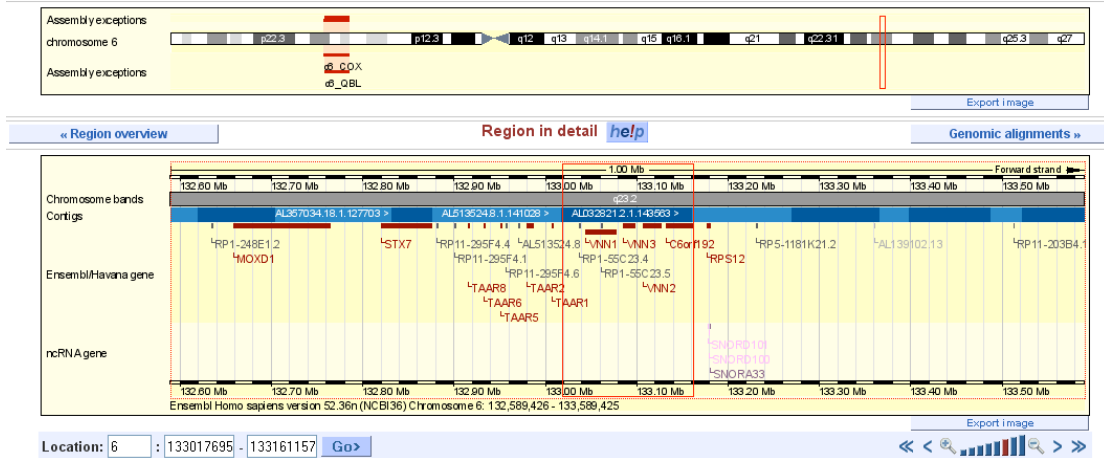
UCSC Genome Browser

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

Ensembl

- Well-supported gene set with evidence
- Range of different views
- Easy retrieval of data sets
- Archive available

Chromosome 6: 133,017,695-133,161,157



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 SNP details, 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

Hubbard, T.J.P. *et al*
Ensembl 2009
Nucleic Acids Res., January 2009; 37: D690 - D697.

Vilella A.J. *et al*
EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.
Genome Res. 2009 Jan 13.

Flicek, P. *et. al*
Ensembl 2008
Nucleic Acids Res. Jan 2008; 36: D707 - D714

Giulietta Spudich, Xosé M. Fernández-Suárez, and Ewan Birney
Genome Browsing with Ensembl: a practical overview
Brief Funct Genomic Proteomic, 2007 Sept; 6: 202-219

Hubbard, T.J.P. *et al.*
Ensembl 2007
Nucl. Acids Res. 2007 **35**: D610-D617

http://nar.oxfordjournals.org/cgi/content/full/35/suppl_1/D610

Xosé M. Fernández-Suárez and Michael K. Schuster

Using the Ensembl Genome Server to Browse Genomic Sequence Data.

UNIT 1.15 in *Current Protocols in Bioinformatics*, Supplement 16, January 2007

<http://mrw.interscience.wiley.com/emrw/9780471250951/cp/cpbi/article/bi0115/current/pdf>

Birney, E. *et al.*

Ensembl 2006

Nucl. Acids Res. 2006 **34**:D556-D561

http://nar.oxfordjournals.org/cgi/content/full/34/suppl_1/D556

Birney, E. *et al.*

An Overview of Ensembl.

Genome Research **14**(5): 925-928 (2004)

<http://www.genome.org/cgi/content/full/14/5/925>

Jekosch, K.

The zebrafish genome project: sequence analysis and annotation.

Methods Cell Biol. **77**:225-39 (2004).

Karolchik, D *et al.*

The UCSC Genome Browser Database.

Nucl. Acids Res. 2003 **31**, 51-54

<http://nar.oupjournals.org/cgi/content/full/31/1/51>

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 website using the **Nuclear respiratory factor 1 (nrf1)** 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
- **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:**
 - What's the conservation track?
 - 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**

The screenshot shows the Ensembl homepage with the following sections:

- Search Ensembl:** A search bar with a dropdown menu set to 'All species' and a 'Go' button. Below it, an example search: 'e.g. human gene BRCA2 or rat X:100000..200000 or insulin'.
- Browse a Genome:** Text explaining the Ensembl project and a link to 'Popular genomes (Log in to customize this list)'. Below this are three links: 'Human (NCBI36)', 'Mouse (NCBI37)', and 'Zebrafish (Zv8)'. The 'Zebrafish' link is circled in red.
- All genomes:** A dropdown menu to 'Select a species' and a link to 'View full list of all Ensembl species'.
- Footer:** Logos for Sanger, Wellcome Trust Sanger Institute, and EMBL-EBI. Text describing the project and funding.
- Right-hand side:**
 - New to Ensembl?:** A list of links: 'Learn how to use Ensembl', 'Add custom tracks', 'Upload your own data', 'Search for a DNA or protein sequence', 'Fetch only the data you want', 'Download our databases via FTP', and 'Mine Ensembl with BioMart'.
 - User Survey:** A yellow box with a clipboard icon and text: 'Almost 6 months have passed since the release of the new website design. If you have a few minutes to spare, we would love to hear what you think of it: Take the survey...'.
 - What's New in Release 54 (5 May 2009):** A list of updates: 'New zebrafish assembly (Zebrafish)', 'Mouse regulatory build (Mouse)', 'New compara views (all species)', 'Variation updates (all species)', and 'Change to default behaviour of TranscriptAdaptor (all species)'.
 - Latest Blog Entries:** A note that no feed is available and a link to 'Go to Ensembl blog'.

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

Search Ensembl Zebrafish

Search for:

Description **Assembly and Genebuild**

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

Assembly

Zv8 is the eighth integrated Whole Genome Shotgun (WGS) assembly of the zebrafish genome at a coverage of 6.5-7x. The project coordination and genome sequencing and assembly is provided by the Wellcome Trust Sanger Institute.

The N50 size is the length such that 50% of the assembled genome lies in blocks of N50 size or longer. The N50 size of the 247,928 contigs is 20,629bp. There are 105,987 supercontigs in the WGS assembly with an N50 size of 687,451bp. ([More information about the assembling process, and further statistics.](#))

Please note: This is still a *preliminary* assembly. The regions of the assembly covered by WGS contigs are of lower quality. The assembly will still contain misjoins, misassemblies and artificial duplications due to retention of haplotypic sequences are likely to occur. During the generation of Zv8, particular attention has been paid to improving the order of the clone path.

Annotation

The zebrafish Zv8 assembly was annotated using a modified Ensembl pipeline. Predictions from zebrafish proteins have been given priority over predictions from other non-mammalian vertebrate species. Aligned zebrafish cDNAs have been used to add UTR regions. Genes are named based on the alignment of their coding regions to known entries in public databases; ZFIN genes have priority in this process.

The final gene-set comprises 24,147 protein-coding genes, 80 genes that have been identified as pseudogenes, and 6 retrotransposed gene predictions. The prediction of ncRNA genes will added for the ensembl 55 release.

[Vega*](#) Additional manual annotation of this genome can be found in [Vega](#)

Ensembl release 54 - May 2009 © [WTSI](#) / [EBI](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)

[Permanent link](#) - [View in archive site](#)

Type 'gene nrf1' into the search bar circled above and click the 'Go' button.

Search Ensembl

Species (1)
 Danio rerio (1)
 Gene (1)
 Feature type (1)
 Gene (1)
 Danio rerio (1)

Ensembl text search

nrf1 corporate/tree:Top/Species/Danio rerio corporate

Your query matched 1 entries in the search database

Ensembl protein_coding Gene: ENSDARG0000000018 (ZFIN: nrf1) [\[Region in detail\]](#)

Ensembl protein_coding gene ENSDARG0000000018 has 1 transcript: ENSDART0000000019, associated peptide: ENSDARP0000000019 and 11 exons: ENSDARE0000000048, ENSDARE0000000049, ENSDARE0000000051, ENSDARE0000000052, ENSDARE0000000055, ENSDARE0000000056, ENSDARE0000000057, ENSDARE0000000058, ENSDARE00000096976, ENSDARE0000155583, ENSDARE00000274045

Nuclear respiratory factor 1 (Nrf-1)(Not really finished protein) [Source:UniProtKB/Swiss-Prot;Acc:Q90X44]

The gene has the following external identifiers mapped to it:
 AfHyMx Microarray Zebrafish: Dr.B179.1.S1_at
 EMBL: AF087671, AL590150
 EntrezGene: SC:323C13.2, ube2h, nrf1, 64604, SC:BZ3C13.1
 GO: GO:006041, GO:0005634, GO:0003677, GO:0006355, GO:0009409, GO:0006950, GO:0016563, GO:0006350, GO:0007275
 IPI: IPI00494020, IPI00494020.2, IPI00775343.1, IPI00775343
 Protein ID: AAC36478, AAC36478.1
 RefSeq DNA: NM_131680.1, NM_131680
 RefSeq peptide: NP_571755, NP_571755.1
 UniGene: Dr.79816, Dr.129833
 UniProtKB/Swiss-Prot: [NRF1](#), DANRE, Q90X44, O93570, Q7ZTS6
 WikiGene: [nrf1](#), 64604
 ZFIN: nuclear respiratory factor, [nrf1](#), ZDB-GENE-001221-1, nrf, not really finished, nrf

Source: e54; Species: Danio rerio; Gene; Feature type: Gene; Danio rerio;

Ensembl release 54 - May 2009 © [WTSI](#) / [EBI](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)

Look through the search results for the appropriate candidates (i.e. *nrf1* is assigned as the gene symbol). In this worked example there is only one search result, but there can be plenty. The following 'Gene' tab will open:

Gene: nrf1

Gene summary [help](#) [Splice variants](#)

Nuclear respiratory factor 1 (Nrf-1)(Not really finished protein) [Source: UniProtKB/Swiss-Prot Q90X44](#)

Location [Chromosome 4: 14,891,849-14,914,055](#) reverse strand.

Transcripts There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
nrf1	ENSDART0000000019	ENSDARP0000000019	protein_coding

Gene summary [help](#)

Name nrf1 (ZFIN)

Synonyms not really finished, nrf, nrf, nuclear respiratory factor [To view all Ensembl genes linked to the name [click here.](#)]

Gene type Known protein coding

Prediction Transcripts were annotated by the Ensembl [genebuild](#).

Method

Transcripts

Configuring the display

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

Ensembl release 54 - May 2009 © [WTSI](#) / [EBI](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)


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

Gene: nrf1

- Gene summary**
- Splice variants (1)
- Supporting evidence
- Sequence
- External references (0)
- Regulation
- Comparative Genomics
 - Genomic alignments (0)
 - Gene Tree
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (44)
 - Paralogues (0)
 - Protein families (1)
- Genetic Variation
 - Variation Table
 - Variation Image
- External Data
- ID History
 - Gene history
- User data
 - Personal annotation

- [Configure this page](#)
- [Add custom data to page](#)
- [Export data](#)
- [Bookmark this page](#)

Click on **'Supporting evidence'** first, which will show which biological sequence records (mRNA and protein) have been used for the annotation of transcripts of a particular gene.


Home > Zebrafish
Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Location: 4:14,891,849-14,914,055
Gene: nrf1
Transcript: nrf1

Gene: nrf1

- Gene summary
- Splice variants (1)
- Supporting evidence**
- Sequence
- External references (0)
- Regulation
- Comparative Genomics
 - Genomic alignments (0)
 - Gene Tree
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (44)
 - Paralogues (0)
 - Protein families (1)
- Genetic Variation
 - Variation Table
 - Variation Image
- External Data
- ID History
 - Gene history
- User data
 - Personal annotation

- [Configure this page](#)
- [Add custom data to page](#)
- [Export data](#)
- [Bookmark this page](#)

Gene: nrf1 (ENSDARG00000000018)

Nuclear respiratory factor 1 (Nrf-1)(Not really finished protein) Source: UniProtKB/Swiss-Prot Q90X44

Location [Chromosome 4: 14,891,849-14,914,055](#) reverse strand.

Transcripts There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
nrf1	ENSDART00000000019	ENSDFARP00000000019	protein_coding

[Splice variants](#)
Supporting evidence [help](#)
[Marked-up sequence](#)

Transcript	CDS support	UTR support	Exon support
ENSDART00000000019 view evidence	align Q90X44.2	align NM_131680.1	6 features

Ensembl release 54 - May 2009 © [WTSI](#) / [EBI](#) [About Ensembl](#) | [Contact Us](#) | [Help](#)

[Permanent link](#) - [View in archive site](#)

How can we view the genomic sequence? Click **'Sequence'** at the left.

THIS STYLE: Location of ENSDARG0000000018 exons
THIS STYLE: Location of Ensembl exons

```
>chromosome: Zv8:4:14891249:14914655:-1
ATTAAGTAAAATTTGAAAAACCTAAACTTGTAAAAGAATAAACGAATGAACATATGAAT
AAAACAATGAAATGCTGTAATGCAGTATGGGCAATTATAATGCAAAGGAAAGATATAA
TAATGTAACCTTATCTAATTTAAATATTTCCAATGGACATTTAGACTAATTGATAATA
TAGTTCCTTAATAAGTGATAAATAGGCCTTTATTCAAACATCTATAACCTTTAAAGCTT
TTTTAAATATACATTTAGCTTAGAGGTGTTTGTATAGTCTGTGGCACTCAAACACT
TTATATATGAAATCTTCTATATCTAATTACATATAAATATCATATTGTTATTAATATA
TTGTTAAAAACAATAAATCGGTGCCATGGAAATGTGCGCATTCCCACACCCATTCCCT
CACTCTCTGTGAGTGTAGTGTGCACCTTGTCTGCTGCATCAGCTAGTAGTGTGTAGCAC
GAGAGGGAGGGGGCGGGTGTCTGTGCACTTCGCTCGTGACCGCGCATCAGTCGCTGA
GCCAGGACCGCGGCACCTGCTGCTCGTTCCTTGTCTCCATTGCAGCTGGTGTTCACA
GCTCCGCGCACCGCGCAGGCCGGAAGTAGCGCAGCCGCTCTGAGGTTAGTGCCCTACCG
CGCAATCTCCAGCCATTTTCAGCCCTGGCGCAGCCGACCAACAACCCGGCAGCGTTTGC
CTTTCACACACTCAGATGCAGATCGCGACGATGCGAAACAGAAGTGCAGATGATTCGAG
CACTTTTGCGCGGATGGCAATTATTATTACCCCGCTCAATTTGGGGCAAAAAAAGAATC
ATGGCGCAGGCGGGGCTTTGGCCTAACGTACTGAATTTAGATGTGCAATCGAATCGCT
TATGTTTTTTCAGATGCTAAATGGGATTTTCAGTGTCTGATAGCTTGTGCTATATGC
TCGATCGCGGAATCGGATTTGAATGCATGTATTATATCGCATTAGCTTAGCAATGTTAT
CGATCAGTCGAGTGTTTATATCGCTTTATTCTCAGTCAACAACGGTGTCTCGCTTACATT
TATCTAAAAGAGCATTTAGATGTTTAGCGCTTATCTCAATGATTGTGATCATTAAGAAG
TTGAATGGGCGACAGTGTTTAATATTTTACCAACACACTGTCCCATTCTTTACAGTAAT
```

← Upstream sequence

← Exon

By default, the exons are highlighted within the genomic sequence.

Variations can be added to this page with the **'Configure this page'** link found at the left. Click on **'Configure this page'** now.

Once you have selected changes (in this example, we display variations and show line numbers) click **'Save and Close'** at the top right (circled in red, above).

THIS STYLE: Location of ENSDARG0000000018 exons

THIS STYLE: Location of Ensembl exons

THIS STYLE: Location of SNPs

```
>chromosome:Zv8:4:14891249:14914655:-1
14914655 ATTAAGTAAAATTTTAAAAACCTAAACTTGTAAGAATAAACGAATGAACATATGAAT 14914596
14914595 AAAACAAATGAAATGCTGTAATGCAAGTATGGGCAATTATAATGCAAAGGAAAGATATAA 14914536
14914535 TAATGTAACTTTATCTAATTTAAATATTTTCCAATGGACATTTAGACTAATTGATAATA 14914476
14914475 TAGTTCCTTAATAAGTGATAAATTAGGCCCTTATTCAAACATCTATAACCTTTAAAGCTT 14914416
14914415 TTTTAAATTATACATTTTAGCTTAGAGGTGTTTGTATAGTCTGTTGCGACTCAAAAACT 14914356
14914355 TTATATATGAAATCTTCTATATCTAATTACATATAAATTATCATATTGTTATTAATATTA 14914296
14914295 TTGXTAAAAACAAAAAAAATCGGTGCCATGGAATGTCGCATTCCCACACCCATCCCT 14914236
14914235 CACTCTCTGTGAGTGTAGTGTGCACCTGCTGCTGCTGCATCACGTAGTAGTGTGTAGCAC 14914176
14914175 GAGAGGAGGGGGCGGGGTTTGTCTGTGCACCTCGCTCGTGACCGCGCATCAGTCGCTGA 14914116
14914115 GCCCAGGACGGCGGCACCTCTGCTGCTCGTCTTGTCTCCATTGCAGCTGGTGTTCACA 14914056
14914055 GCTCCGCGCACCGCGCAGGCCCGGAAGTAGCCGAGCCGCTCTGAGGTTAGTGCCCTACCG 14913996
14913995 CGCAATCTCCAGCCATTTTCAGCCCTGGCGCAGCCGACCAACACCCCGGCAGCGTTTCG 14913936
14913935 CTTTCACACACTCAGATGCAGATCGCGCAGATGCGAAACAGAAGTGCAGATGATTTTCGAG 14913876
14913875 CACTTTTGCCTGGATGGCAATTTATTTACCCCGCTCAATTTGGGGCAAAAAAGAAATC 14913816
14913815 ATGGCGCAGGCGCGGCCCTTTGGCCTAACGTAAGTGAATTTAGATGTGCAATCGAATCGCT 14913756
14913755 TATGTTTTTTGTCCAGATGCTAAATGGGATTTTCAGTCTTGATAGCTTGTGCTCATATGC 14913696
14913695 TCGATCGCGGAATCGGATTTTGAATGCATGTATTATATCGCATTAGCTTAGCAATGTTAT 14913636
```

Now variations in the sequence are highlighted in green. Line numbers 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.

The screenshot shows the Ensembl genome browser interface for the *nrf1* gene. The main content area displays a phylogenetic tree titled "Gene Tree" with a legend. The legend includes symbols for branch lengths (x1, x10, x100), duplication nodes, and alignment quality (AA alignment match/mismatch, AA consensus > 65%, AA consensus > 33%). The tree shows the current gene (*nrf1*) in zebrafish and its orthologs in other species, including *Gasterosteus aculeatus*, *Drosophila*, and *Ciona*. A table below the tree lists the transcripts and protein IDs for the current gene and its orthologs.

Name	Transcript ID	Protein ID	Description
<i>nrf1</i>	ENSDART0000000019	ENSDARP0000000019	protein_coding

Use the mouse over and 'expand sub-tree' to get to the view displayed above. Click **'View fully expanded tree'** at the bottom.

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

The screenshot shows the Ensembl interface for the *nrf1* gene. The 'Variation Image' tab is active, displaying a genomic track from 14.88 Mb to 14.92 Mb. The track includes the Ensembl gene model, protein domains (GIA, TIC, A/G, G/C, G/T, G/A), and various genetic variations. A legend at the bottom explains the color coding for different types of variations: 3' UTR (blue), G/C (green), T/C (yellow), A/G (orange), T/C (red), G/A (purple), G/T (brown), and G/A (pink).

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

The screenshot shows the Ensembl interface for the variation **rs40788869**. The 'Variation summary' tab is active, displaying detailed information about the variation, including its class (SNP), synonyms, alleles (G/A), and location (4:14909816). A callout box points to the 'Variation of interest' (rs40788869) and another callout box points to the 'Links to population frequency, if available' link.

Now, we would like to work with the transcript of this gene. Select the **transcript** from the header section by clicking on the **Transcript tab** for nrf1. This will lead to the Transcript-Summary display.

Ensembl
Home > Zebrafish
Location: 414,891,849-14,914,055 Gene: nrf1 Transcript: nrf1 Variation: rs40788869

Transcript-based displays

- Transcript summary
- Supporting evidence (8)
- Sequence
 - Exons (11)
 - cDNA
 - Protein
- External References
 - General identifiers (13)
 - Oligo probes (1)
 - Gene ontology (9)
- Genetic Variation
 - Population comparison
 - Comparison image
- Protein Information
 - Protein summary
 - Domains & features (10)
 - Variations (5)
- External Data
- ID History
 - Transcript history
 - Protein history
- User data
 - Personal annotation

Transcript: nrf1 (ENSDART00000000019)
Nuclear respiratory factor 1 (Nrf-1)(Not really finished protein) [Source:UniProtKB/Swiss-Prot;Acc:Q90X44]
Location: [Chromosome 4: 14,891,849-14,914,055](#) reverse strand.
Gene: This transcript is a product of gene [ENSDARG00000000018](#) - There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
nrf1	ENSDART00000000019	ENSDARPO00000000019	protein_coding

Transcript summary [help](#) [Supporting evidence >](#)

Statistics Exons: 11 Transcript length: 2,908 bps Translation length: 514 residues
Type Known protein coding
Prediction Method Transcripts were annotated by the Ensembl [genebuild](#).

Ensembl release 54 - May 2009 © WTSI / EBI [About Ensembl](#) | [Contact Us](#) | [Help](#)
[Permanent link](#) - [View in archive site](#)

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

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

Transcript-based displays

- Transcript summary
- Supporting evidence (8)
- Sequence
 - Exons (11)
 - cDNA
 - Protein
- External References
 - General identifiers (13)
 - Oligo probes (1)
 - Gene ontology (9)
- Genetic Variation
 - Population comparison
 - Comparison image
- Protein Information
 - Protein summary
 - Domains & features (10)
 - Variations (5)
- External Data
- ID History
 - Transcript history
 - Protein history
- User data
 - Personal annotation

- Configure this page
- Add custom data to page
- Export data
- Bookmark this page

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).

Green: Flanking sequence

Purple: UTR

Black: Coding sequence

Blue: Introns

Have you forgotten what the colours mean? No worries- click on the ‘Help’ button (circled in red) and read the help for this page. A link to the **glossary** is also provided.

Next, follow the ‘Supporting Evidence’ link. The following data display is quite an important one, as it shows which biological evidence has been used for the annotation of this transcript.

Red boxes: exons in the Ensembl transcript (mRNA)

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.

The screenshot shows the Ensembl genome browser interface for the *nrf1* gene on chromosome 4. The main display area shows tracks for Contigs, Marker, Ensembl gene, and EMBL vertebrate cDNA. A 'Configure the display' dialog box is open at the bottom, showing options to turn on/off various tracks. The dialog box indicates that there are currently 54 tracks turned off and provides instructions on how to change the tracks.

Ensembl **'Location'** displays are also highly configurable. You can switch on additional tracks displaying the various feature types that Ensembl annotates in the genome. Again, 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 **'cDNAs'** track in **'normal'** expanded form.

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

```
>4 dna:chromosome chromosome:Zv8:4:14891849:14914055:1
TTGCCAATGTCACCATCATTTTGCACCAGCAACACGGGAGGGTTTGGTTTCATCAAACCG
TATATATTTTACAATTTACATAAAAAAGCCTCTCATTGTGTGAGCTATAAAAAATTCGA
ACCGATGAAAGGAGCTGAAACTCGACAGAGTGGATCACGGGTGGAAGAAACAGTAACAAA
ACATCCAAAATGCACTGTAGAAACAATACAAGAGCTTGGGAGATAAAGAAAGAGATATG
AATTTATGGGTATTTCACTTTGCGATAACATTAGAATTCGAATACATCCTACATCATATC
TAAACCTTTTCCCATGTGTCTTATTCAAAAGATATGAGCGCACATCTAAACAAACCAAAA
TAAATAAAAAATAAATAAAACATTTTCCCTAGCAGTTTCATAACAGGATCAGTTTCCAAA
ATGCGGACATACCATGAATATGTAAGCTCTCTGTTTTCGAAGTATAAAGCTGAACGA
TATGGAGTCTACATCTTTCATCCTTGTCTACTGACATTTATAGCAATCGATTAATAGTA
CAATTTAGTGAATAGTCTATAGATAGCACGTGAGAGTTCTGAAAGTTCAAATCTGTACCGG
CAGGAAACGGTGGTCTCATTCGATATACCACATCATTCGAGATCATACAATTTTTTTTTT
TACAATGGGAATATGTGAAATGCATACAAGAAACTACATTTGTACAGACTTACAAAAA
AAAAAAAAAAAAACCATGCATATCTCAGTAGTCAATGGGATCTAACACTAAAATTTGTAATG
CAATATACACAATATATTAATGAAATATTCAAAATATGGCACAAGATGGTTCACAATATA
TATAATATGTACATAAAAACTGAGGAGAAACGCATTAAGAATGACCAACATAATGCCAT
ATACCGGGGAGAAATGAACAAATTCGCGCCAGTTTGAAGGCAGAAAAATTTGGACTGAG
GTTTTTCTTTTCTTGTAACTCGTACAGAAAGAAATAATTGGTAAATTCATAGTATCTCTGC
ATTACATTTGTCAAACACCAGCTAAAAAAGAGAAACGAACTAACGTTAATGAA
...
```

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'.

Query	Subject	Chromosome	Scaffold	Contig	Stats	Sort By			
off Name Start	off Name Start	off Name Start	off Name Start	off Name Start	off Score E-val	>Contig <Score >Score			
Links	Query	Chromosome	Start	End	Ori	Score	E-val	%ID	Length
[A] [S] [G] [C]	1	720 +	Chr:4	14891849	14892568 +	3566	0.0e+00	100.00	720

Finally, follow links to an alignment **[A]**, the query sequence **[S]**, the genome sequence **[G]** and the corresponding Location-View **[C]** (for its former name ContigView... or to C (see) the BLAST hit!

The screenshot shows the Ensembl genome browser interface for Zebrafish. The main display area is titled "Region in detail" and shows a genomic region on Chromosome 4. The tracks include:

- Contigs:** Shows genomic contigs with coordinates and sizes.
- Marker:** Shows various markers such as Bx6493473, Bx571796110, and Bx5373585.
- Ensembl gene:** Shows the gene structure for *nrfl* and *pkna4*.
- BLAT/BLAST hits:** Shows a red bar representing a BLAT hit, with a black arrow pointing to it from a box labeled "BLAT hit".
- Known protein coding Ensembl gene:** Shows the protein-coding region of the *nrfl* gene.

At the bottom of the main display area, there is a section titled "Configuring the display" with the following text:

You currently have 0 tracks in the overview panel and 54 tracks in the main panel turned off. To change the tracks you are displaying, use the "Configure this page" link on the left.

Export the image using the link at the bottom.

EXERCISES and ANSWERS

Note: The answers to these exercises correspond to version 54 of Ensembl. If you use a newer version and your answer doesn't correspond with the given answer, please consult the instructors. Alternatively, you can use version 54 from the Ensembl Archive site.

Exercise 1 – Exploring a gene

(a) Search for the zebrafish pax6a 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?

(c) Have a look at the General identifiers and the Gene ontology terms for one of the pax6a transcripts (ENSDART00000066224/ENSDART00000066225). Click on some of the links. What is the function of pax6a?

(d) Which PFAM domains does the protein encoded by pax6a contain?

(e) Is there a human ortholog predicted for the zebrafish pax6a 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) How does the manual annotation in Vega compare to the Ensembl annotation? Why are there differences?

(2) What does ZFIN say about pax6a?

Answers

(a)

- Go to <http://www.ensembl.org>.
- Under 'Search Ensembl' type 'zebrafish pax6a gene'.
- Click [Go].
- On the page with search results click on 'Ensembl protein_coding Gene: ENSDARG00000045045 (ZFIN: pax6a)'.

The zebrafish pax6a gene is located on linkage group 25. Ensembl has two transcripts annotated for this gene. The transcripts are transcribed from the forward strand of the genome assembly.

(b)

- Click on the Ensembl Transcript IDs (ENSDART00000066225).

The longest transcript is ENSDART00000066225. The length of this transcript is 2790 base pairs and the length of the encoded protein 451 amino acids.

- Click on 'ENSDART00000066225' in case you are not already on the 'Transcript: pax6a' tab.
- Click on 'Exons' in the side menu.

ENSDART00000066225 has 13 exons, of which the first one is completely untranslated and the second and last one are partially untranslated.

(c)

- Click on 'General identifiers' in the side menu.
- Explore some of the links (good places to start are 'ZFIN' and 'UniProtKB/Swiss-Prot').
- Do the same for 'Gene ontology'.

Pax6a encodes a nuclear transcription factor involved in pattern formation and brain development.

(d)

- Click on 'Domains & features' in the side menu.

The pax6a product contains 2 PFAM domains: Homeobox and Paired_box_N.

(e)

- Click on the 'Gene: pax6a' tab.

- Click on 'Orthologues' in the side menu.

There is one human ortholog predicted for zebrafish pax6a, PAX6 (ENSG00000007372). It has the type 1-to-many.

- Click on the 'Orthologues' next to the 'help' link.
- Click on the link for the detailed description, read through 'homology types'.
- Click on the human orthologue ENSDARG00000007372, follow the links to its zebrafish orthologues

The type 1-to-many is set because one human gene (PAX6) is the ortholog to two zebrafish genes (pax6a and pax6b)

Exercise 2 – Exploring a region

(a) Go to the region from bp 33300000 to 33500000 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, contigs from unfinished clones and whole genome shotgun contigs)?

(b) Do the tilepath clones (i.e. the BAC clones that were sequenced to generate the sequence for the human genome assembly) correspond with the contigs? Note that these clones are not shown by default! Which clone library does the clone containing the fgf8 gene come from?

(c) Zoom in on the fgf8 transcript, including a bit of flanking sequence on both sides. Which marker is located close by? Does this marker appear anywhere else in the genome?

(d) CpG islands are genomic regions that contain a high frequency of CG dinucleotides and are often located near the promoter of mammalian genes. Is there a CpG island associated with the fgf8 transcript? And did the Eponine program (<http://www.sanger.ac.uk/Software/analysis/eponine/>) predict a transcription start site for the fgf8 transcript?

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

(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 13:33300000-33500000'.
- Click [Go].

This genomic region is made up of 3 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.
- Search for 'Misc regions'.
- Select 'WGS/Clones assembly'.
- Click [SAVE and close].

The tilepath clones correspond neatly to the contigs and it is easy to see from which BAC clones which contig sequence in the assembly is derived. The tilepath clones overlap.

- Click on 'CR925797' in the 'WGS/Clones assembly' track
- In the new menu, click on the EMBL accession. Read the last lines of the comments.

CR925797 is clone CH211-194I8 from the CHORI-211 BAC library.

(c)

- Draw a box around the *fgf8* transcript.
- Click on 'Jump to region' in the pop-up menu.

Gene *fgf8* overlaps the *fgf8* marker

- Click on the marker and 'Marker info'

The *fgf8* marker is uniquely placed into this position.

(d)

- Click on 'Configure this page' in the side menu.
- Search for 'cpg'.
- Select 'CpG islands'.
- Search for 'eponine'.
- Select 'TSS (Eponine)'.

- Click [SAVE and close].

There is indeed a CpG island located at the 5' end of the fgf8 transcript. Eponine TSS-finder predicts a transcription start site quite a bit away from the assumed Ensembl fgf8 transcript start site.

(e)

- 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 (Zv8), the chromosome (13), the start and end coordinates (33420763 and 33536545) and the strand (1):

```
>13 dna:chromosome chromosome:Zv8:13:33420763:33536545:1
```

BioMart

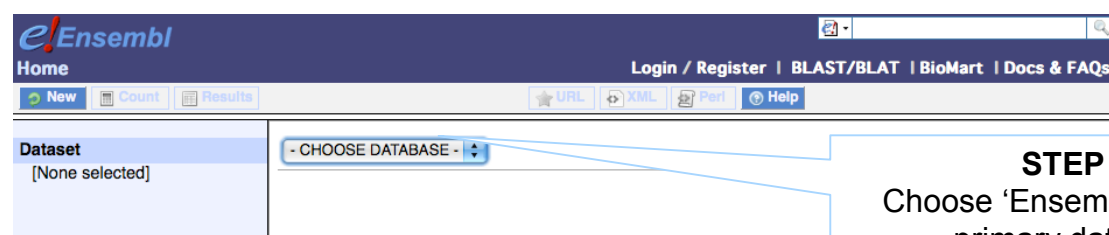
Mining data- worked example

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.

Where and when are these zebrafish genes expressed?

Download the sequence of all available 5' UTRs of these genes.

STEP 1: Either click on 'BioMart' in the top right header bar of the Ensembl home page, or go to <http://www.biomart.org/> and click on the 'MartView' tab.



The screenshot shows the Ensembl BioMart interface. The top navigation bar includes 'Home', 'Login / Register', 'BLAST/BLAT', 'BioMart', and 'Docs & FAQs'. Below the navigation bar, there are buttons for 'New', 'Count', and 'Results'. A 'Dataset' section shows '[None selected]'. A dropdown menu labeled '- CHOOSE DATABASE -' is highlighted with a blue box. A callout box points to this dropdown menu with the text: 'STEP 2: Choose 'Ensembl 54' as the primary database.'

The screenshot shows the Ensembl homepage with the 'Dataset' dropdown menu open, displaying 'Ensembl 54' and a '- CHOOSE DATASET -' option below it. A callout box points to the dropdown menu.

STEP 3:
Choose '*Danio rerio*' as the species of interest.

The screenshot shows the 'Filters' section on the left sidebar expanded. The 'REGION' checkbox is checked, and the 'Please restrict your query using criteria below' section is visible. A callout box points to the 'Filters' section.

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

The screenshot shows the 'REGION' section expanded. The 'Chromosome' checkbox is checked, and a dropdown menu shows '1' selected. The 'Gene Start (bp)' and 'Gene End (bp)' fields are also visible. A callout box points to the 'Chromosome' dropdown.

STEP 5:
Select '**Chromosome 1**'

STEP 6:
Expand the '**GENE**' panel.

STEP 7:
Expand the '**MULTI SPECIES COMPARISON**' panel.

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

STEP 9:
Limit to 'Orthologous Human Genes Only'

STEP 9:
The filters have determined our gene set. Click 'Count' to see how many genes have passed these filters.

The 'Count' results show 800 zebrafish genes out of 24,233 total genes passed the filters.

STEP 10:
Click on 'Attributes' to select output options (i.e. what we would like to know about our gene set).

The filters have determined our gene set. Click 'Count' (at the top) to see how many genes have passed these filters at any time during your search.

The screenshot shows the Ensembl genome browser interface. The left sidebar contains navigation options like 'New', 'Count', and 'Results'. The main content area is titled 'Please select columns to be included in the output and hit 'Results' when ready'. Under the 'Features' radio button, the 'GENE' section is expanded, showing a list of checkboxes for various attributes such as 'Ensembl Gene ID', 'Ensembl Transcript ID', 'Associated Gene Name', etc. A callout box labeled 'STEP 11' points to the 'GENE' section.

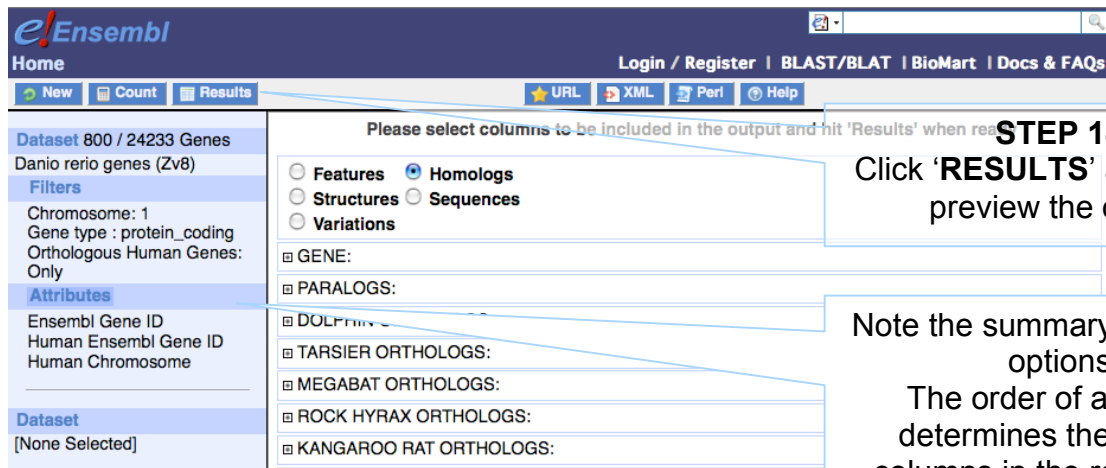
STEP 11:
Expand the 'GENE' panel. Deselect the Transcript ID

The screenshot shows the Ensembl genome browser interface with the 'Homologs' radio button selected. The 'HUMAN ORTHOLOGS' section is expanded, showing a list of checkboxes for attributes like 'Human Ensembl Gene ID', 'Human Chromosome', 'Human Chromosome End (bp)', etc. A callout box labeled 'STEP 12' points to the 'HUMAN ORTHOLOGS' section.

STEP 12:
Expand the 'Homologs' panel and select 'Human Orthologs'

This is a close-up of the 'HUMAN ORTHOLOGS' section. The 'Orthologs' sub-section is expanded, showing checkboxes for 'Human Ensembl Gene ID', 'Human Chromosome', 'Human Chromosome End (bp)', 'Orthology Type', 'dN', 'dS', 'Representative Protein ID', 'Human Ensembl Protein ID', and 'Human Chromosome Start (bp)'. A callout box labeled 'STEP 13' points to the 'Human Ensembl Gene ID' and 'Human Chromosome' checkboxes.

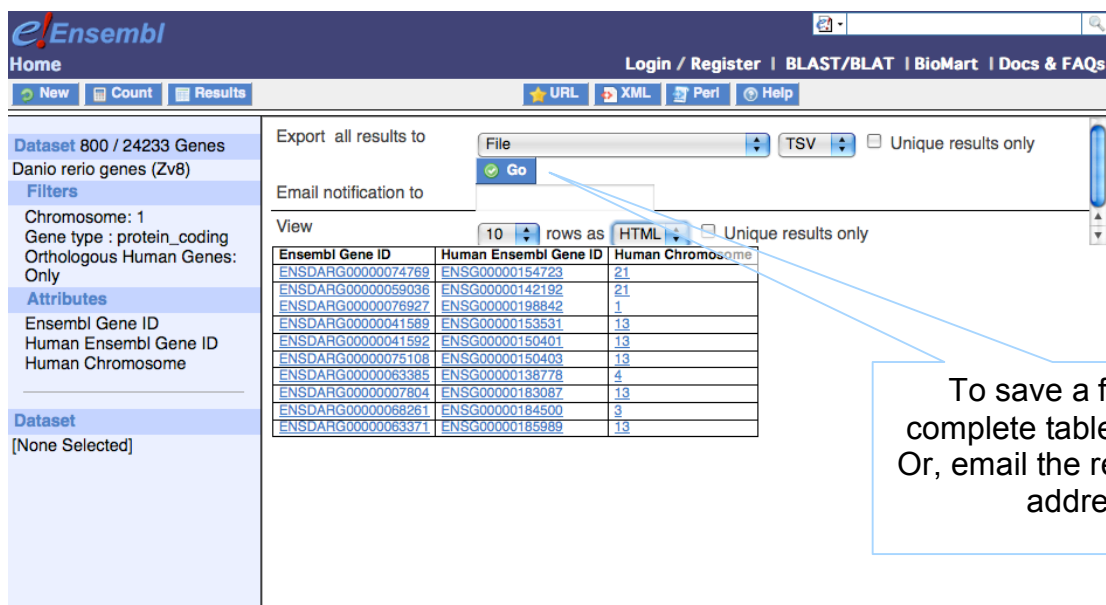
STEP 13:
Select 'Human Ensembl Gene ID' and 'Human Chromosome'



STEP 14:
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.



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

Ensembl Gene ID	Human Ensembl Gene ID	Human Chromosome
ENSDARG00000074769	ENSG00000154723	21
ENSDARG00000059036	ENSG00000142192	21
ENSDARG00000076927	ENSG00000198842	1
ENSDARG00000041589	ENSG00000153531	13
ENSDARG00000041592	ENSG00000150401	13
ENSDARG00000075108	ENSG00000150403	13
ENSDARG00000063385	ENSG00000138778	4
ENSDARG00000007804	ENSG00000183087	13
ENSDARG00000068261	ENSG00000184500	3
ENSDARG00000063371	ENSG00000165989	13

Continue to find out about the expression of these 800 zebrafish genes.

The screenshot shows the Ensembl interface with the 'Features' section selected. Callouts highlight the 'Features' radio button and the checked options under 'Source of Expression Data'.

Clicking 'Results' will show you a list of the genes with associated expression stages/anatomical locations, if any.

The screenshot shows the 'Results' table with the following data:

Ensembl Gene ID	Anatomical System (ZFIN)	Development Stage (ZFIN)
ENSDARG00000074769		
ENSDARG00000074769		
ENSDARG00000059036		
ENSDARG00000076927		
ENSDARG00000041589		
ENSDARG00000041592	unspecified	Zygote:1-cell
ENSDARG00000041592	unspecified	Cleavage:2-cell
ENSDARG00000041592	unspecified	Cleavage:4-cell
ENSDARG00000041592	unspecified	Cleavage:8-cell
ENSDARG00000041592	unspecified	Cleavage:16-cell
ENSDARG00000041592	unspecified	Cleavage:32-cell
ENSDARG00000041592	unspecified	Cleavage:64-cell
ENSDARG00000041592	unspecified	Blastula:128-cell
ENSDARG00000041592	unspecified	Blastula:256-cell
ENSDARG00000041592	unspecified	Blastula:512-cell
ENSDARG00000041592	unspecified	Blastula:1k-cell
ENSDARG00000041592	unspecified	Blastula:High
ENSDARG00000041592	unspecified	Blastula:Oblong
ENSDARG00000041592	unspecified	Blastula:Sphere
ENSDARG00000041592	unspecified	Blastula:Dome

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

The screenshot shows the Ensembl web interface. On the left sidebar, the 'Attributes' section is expanded, showing 'Ensembl Gene ID', 'Ensembl Transcript ID', and '5' UTR'. The main content area is titled 'Please select columns to be included in the output and hit 'Results' when ready'. Under the 'SEQUENCES' section, the 'Sequences (max 1)' radio button is selected. Below it, the '5' UTR' radio button is selected. Other options include 'Unspliced (Transcript)', 'Unspliced (Gene)', 'Flank (Transcript)', 'Flank (Gene)', 'Flank-coding region (Transcript)', 'Flank-coding region (Gene)', '3' UTR', 'Exon sequences', 'cDNA sequences', 'Coding sequence', and 'Peptide'. There are also input fields for 'Upstream flank' and 'Downstream flank'. A callout box points to the 'Sequences' radio button with the text 'Select 'Sequences'..' and another callout points to the '5' UTR' radio button with the text 'Select '5'UTR''.

Click **'Results'** and you will get the required list.

The screenshot shows the Ensembl web interface after clicking 'Results'. The 'Export all results to' dropdown is set to 'FASTA'. The 'View' section shows '20 rows as FASTA'. The main content area displays a list of gene sequences in FASTA format, including Ensembl IDs and transcript IDs. The sequences are as follows:

```

>ENSDARG0000000086 | ENSDART0000000087
TGCGGCTGGACGGCAGCTGACGGACTGAACC
>ENSDARG00000013695 | ENSDART00000005116
CGCGTGCCGTGCTCCTCGCGGCTCCCGCTGCTGATTGAGGCTCCGGTTCCTCCGGTCA
GTCTCCGGTCCGGC
>ENSDARG00000015506 | ENSDART00000004361
Sequence unavailable
>ENSDARG00000005525 | ENSDART00000004233
CTGAGAGTTGCTAGGGCGTCAAACATCATGTGTGAAAGATGATCCAAGGGTGGCGGCTGA
GGCAAGGCAGAAAAGGAGGAGGCATCCCTGCAGTGCCGTGGCCCTGCAGCTAGAAT
AGGCTGCACCTGGCTGAATGTGTGTCACACGCATGCCCTGTACACCTCACCGTATGTG
TCCAAAGTCTCATATCCCTTGTGCAACTTCCCTCAGTAAGACGATTAGAGGGACAGCGT
TGGGAAGAAGGAGCAATCAACACTTCTGTCTTAAATATAACG
>ENSDARG00000005626 | ENSDART00000003097
Sequence unavailable
>ENSDARG00000003054 | ENSDART00000003278
Sequence unavailable
>ENSDARG00000009472 | ENSDART00000003022
AGACAGAAGCTTACGGACGTTAAGTAATGTCAACCGAATACAAACGTCACATAAATTA
ACCGTCGACGCTCGCAGCGTAGTTTTTGTATC
>ENSDARG000000075827 | ENSDART00000003463
AGA
>ENSDARG00000016994 | ENSDART00000003895
GCACATCTACGCTCACGGGACTCGGCCTTGATCCTCTCGCCGAATAACACTTTTAAACGT
CACAGTCCATCTGTCTACTCTGAC
>ENSDARG00000014313 | ENSDART00000004062
CGGGTGTCTGTGTCGGTCTGCTGGTGGTGTCAACACAAGAAGC
>ENSDARG00000006392 | ENSDART00000002886
TGAAGTGCATCAACACCGCGCTGCTTACTCTGAATGGAGCATGTGAAATGATACAT
TGCAAAATCGAGTTGTTTATGTGAAATAGGCTATTCAAGCTGTCAA
    
```

EXERCISES and ANSWERS

Note: The answers to these exercises correspond to version 54 of Ensembl. If you use a newer version and your answer doesn't correspond with the given answer, please consult the instructors. Alternatively, you can use version 54 from the Ensembl Archive site.

Exercise 1

Generate a list of all zebrafish genes on chr4 with a ZFIN ID that are expressed in the gastrula shield and have a transmembrane domain. Narrow this down to genes without a human ortholog. Narrow it down again to genes with at least two splice variants.

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 54' database.
- Choose the 'Danio rerio genes (Zv8)' dataset.

Now filter for the genes on chromosome 4:

- Click on 'Filters' in the left panel.
- Expand the 'REGION' section by clicking on the + box.
- Select 'Chromosome - 4'. 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 446 / 37435 Genes. 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 1079 / 24233 Genes. Now for genes with a ZFIN ID

- Click 'ID list limit' and choose 'ZFIN ID'
- Click the [Count] button on the toolbar.

All 1079 have ZFIN IDs

Now select only those genes that are expressed in the shield.

- Expand the 'EXPRESSION' section by clicking on the + box.
- Click 'ZFIN developmental stage data' and browse to 'Gastrula', then select 'Gastrula:Shield'
- Click the [Count] button on the toolbar.

You should be down to 136 / 24233 genes now. Select those with a transmembrane domain:

- Expand the 'PROTEIN DOMAINS' section by clicking on the + box.
- Select 'Transmembrane domains' and also 'Only'
- Click the [Count] button on the toolbar.

This leaves 25 genes. Narrow down further to those without a Human ortholog.

- Expand the 'MULTI SPECIES COMPARISONS' section by clicking on the + box.
- Select 'Homolog filters' and select 'Orthologous Human Genes - Excluded'
- Click the [Count] button on the toolbar.

Down to 3 / 24233. Now only select those genes with 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.

One gene left. Now download the cDNA sequence with the Ensembl ID, the associated gene name, gene DB and a description.

- Click on 'Attributes' in the left panel.
- Select the 'Sequences' attributes page.
- Select 'Peptide'.
- Expand the 'Header Information' section and select 'Ensembl Gene ID', 'Description', 'Associated Gene Name' and 'Associated Gene DB'
- Click the [Results] button on the toolbar.

If you are happy with how the results look in the preview, output all the results:

- Select 'View All rows as HTML' or export all results to a file.

Note: When you select 'View All rows as HTML', your results will be shown under a new tab in your internet browser.

Although you have filtered for only one gene, your results will contain more than one row. This is because the gene has more than one transcript and the results contain a separate row for each transcript.

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 53' database.
- Choose the 'Homo sapiens genes (NCBI36)' dataset.

- Click on 'Filters' in the left panel.
- Expand the 'GENE' section by clicking on the + box.
- Select 'ID list limit - Refseq protein ID(s)'
- Enter the list of IDs in the text box (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.

- 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

List all genes between the markers Z17393 and Z65461. Where did they get their names from?

Answer

- Click [New].
- Choose the 'Ensembl 54' database.
- Choose the 'Danio rerio genes (Zv8)' dataset.

- Click on 'Filters' in the left panel.
- Expand the 'REGION' section by clicking on the + box.
- Enter 'Marker Start: Z17393' and 'Marker End: Z65461'.

- 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'.
- Select 'Associated Gene Name' and 'Associated Gene DB'.

- 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'.

Your results should show 29 genes. Among these there should be genes with a ZFIN record, genes without a ZFIN record but an entry in Entrez, genes without a ZFIN record but closely related to a human gene (HGNC, check this is really the case by following up *vav2* in human and also in ZFIN!) and genes where no good relation could be found, hence the DB column is empty.

Exercise 4

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 54' and dataset 'Danio rerio genes (Zv8).
- Narrow down by filtering for 'REGION' 'Chromosome - 1' and 'MULTI SPECIES COMPARISONS' selecting 'Homolog Filters' 'Orthologous Human genes -Only'
- Click on "Attributes', then 'Features' , deselect 'Ensembl Transcript ID', select 'Associated Gene Name'

- Click on 'Dataset' (bottom left) and select '[Ensembl 54] Homo sapiens genes (NCBI36)'
- Narrow down by filtering for 'REGION' 'Chromosome - 13' and 'MULTI SPECIES COMPARISONS' selecting 'Homolog Filters' 'Orthologous 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, whereas a few zebrafish genes seem to be in need of renaming.

Exercise 5

Design your own query!