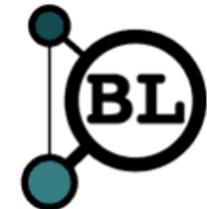


Genome Literacy Workshop

Elisabeth Busch-Nentwich
& Pavol Kramár



Learning Outcomes

- Understand Ensembl as a database
 - basics of default data
 - investigating homology
- Find and switch on optional features
 - find your gene and its associated data
- Download gene and genome data
 - key tools to use
- Upload and display your own data

Part 1

- Zebrafish Genome Project
- Ensembl
- Finding your gene
- Gene name and IDs
- Manual and automatic annotation
- Ensembl “Region” view

Ensembl

- Most examples from **Ensembl** (we are biased!)
- Probably most widely used genome browser amongst zebrafish researchers
- **Primary source of zebrafish annotation** (UCSC imports Ensembl annotation)
- Currently Ensembl version **109** (Feb 8th)
- New releases 3 or 4 times / year
- Zebrafish **annotation largely static** between releases
- But **naming and homology** updated (+ new functionality)

The screenshot shows the Ensembl homepage. At the top, there's a navigation bar with links for 'Login/Register', 'e!Ensembl' logo, 'BLAST/BLAT', 'VEP', 'Tools', and 'More'. Below the navigation is a search bar with the placeholder 'Search all species...' and a magnifying glass icon.

The main content area has several sections:

- Tools**: A link to 'All tools'.
- BioMart >**: A link to 'Export custom datasets from Ensembl with this data-mining tool'.
- BLAST/BLAT >**: A link to 'Search our genomes for your DNA or protein sequence'.
- Variant Effect Predictor >**: A link to 'Analyse your own variants and predict the functional consequences of known and unknown variants'.

A sidebar on the right provides information about Ensembl, including its purpose as a genome browser for vertebrate genomes, its focus on comparative genomics, evolution, and sequence variation, and its tools like BLAST, BLAT, BioMart, and Variant Effect Predictor (VEP). It also mentions the current release (107) and various updates.

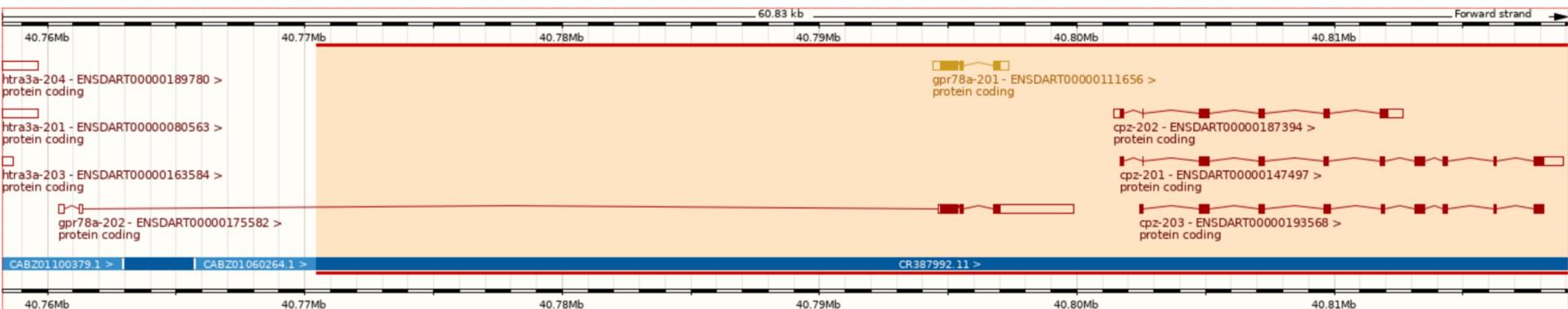
The central part of the page features a search interface with a dropdown menu set to 'All species' and a text input field. Below it is a note: 'e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease'.

Below the search is a section for 'All genomes' and 'Favourite genomes'. It includes a dropdown menu for selecting a species, currently set to 'All species'. Under 'All genomes', there's a section for 'Pig breeds' which includes 'Pig reference genome and 12 additional breeds' and a note 'Still using GRCh37?'. Under 'Favourite genomes', there are entries for 'Human' (GRCh38.p13), 'Mouse' (GRCm39), and 'Zebrafish' (GRCz11), which is highlighted with a red oval.

At the bottom right, there's a blue box for 'Ensembl Rapid Release' and a link to 'More release news'.

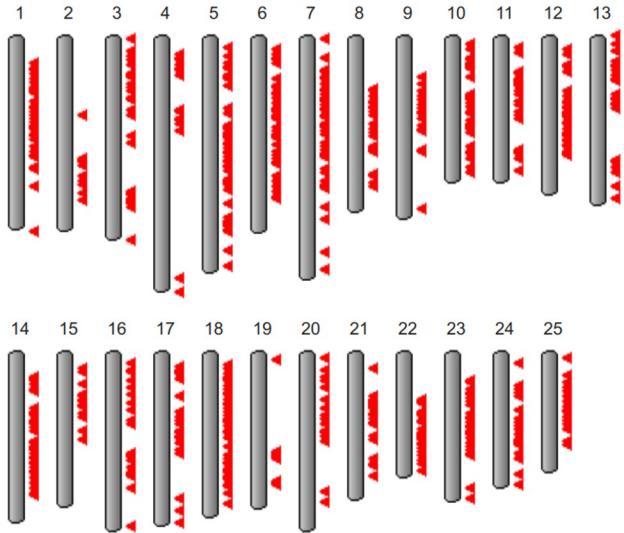
Zebrafish Genome

- **GRCz11** (danRer11) - latest assembly, released in 2017
- Sequencing strategy:
 - 90% clone by clone sequencing
 - **High quality**
 - 10% whole genome shotgun sequencing
 - **Lower quality**
 - Fills gaps between clones
 - Identified by accessions beginning with **CABZ**



Zebrafish Genome History

- Genome project started in **2001** at Sanger Institute
- Initially sequenced pool of **Tübingen** zebrafish
- But zebrafish **very polymorphic** compared to humans
- Too much variation to join clones, so lots of **gaps**
- + same region represented by 2+ clones, leading to **artificial duplication**
- Later used **double haploid** Tübingen fish for some clones and most WGS
- Only **925 gaps** between scaffolds and **N50 > 7 Mbp**
- GRCz11 contains **alternative** scaffolds
- When downloading sequence from Ensembl FTP site, "**toplevel**" includes alternative sequence, but "**primary_assembly**" doesn't and is probably what you want



From <https://www.ncbi.nlm.nih.gov/grc/zebrafish>

Older Assemblies

- Previous assemblies available in Ensembl **archives**:

www.ensembl.org/info/website/archives/assembly.html

- GRCz10 / danRer10: <http://e91.ensembl.org/>
- Zv9 / danRer7: <http://e77.ensembl.org/>
- Zv8 / danRer6: <http://e54.ensembl.org/>

- Even **older** assemblies available in UCSC
- Numbering coordinated when **GRC** (Genome Reference Consortium) took over managing zebrafish assembly from Sanger Institute

The screenshot shows the Ensembl Archive interface for the Zebrafish genome assembly (GRCz10). At the top, there's a navigation bar with links for BioMart, Tools, and More, along with a search bar and a login/register link. The main content area is titled 'Zebrafish (GRCz10)'. It features a search bar with dropdowns for 'Search all categories' and 'Search Zebrafish...', and a 'Go' button. Below the search bar is a 'What's New in Zebrafish release 91' section with a list of updates: 'Structural variants', 'New dbSNP data for zebrafish', and 'Fixing stable ids in the external data database', with a 'More news...' link. The central part of the page displays genome assembly details for GRCz10 (GCA_000002035.3), including links for 'More information and statistics', 'Download DNA sequence (FASTA)', 'Display your data in Ensembl', and 'Other assemblies'. A 'Gene annotation' section on the right provides a summary of what can be found, including protein-coding and non-coding genes, splice variants, cDNA and protein sequences, and non-coding RNAs. It includes links for 'More about this genebuild', 'Download genes, cDNAs, ncRNA, proteins (FASTA)', and examples of Pax6, FOXP2, DMD, and ssh genes with their respective protein and transcript diagrams.

Ensembl Mirrors

- Mirrors: www.ensembl.org/info/about/mirrors.html
- Main site (UK): www.ensembl.org
- US East mirror: useast.ensembl.org
- US West mirror: uswest.ensembl.org
- Most often slow due to chosen tracks though



UK (Sanger Institute) - **YOU ARE HERE!**



[US West \(Amazon AWS\)](#) - Cloud-based mirror on West Coast of US



[US East \(Amazon AWS\)](#) - Cloud-based mirror on East Coast of US



[Asia \(Amazon AWS\)](#) - Cloud-based mirror in Singapore

Finding Your Gene

- Follow link from ZFIN

The screenshot shows the ZFIN gene details page for *dmd*. The left sidebar lists various gene-related categories: Summary, Expression, Phenotype, Mutations, Human Disease, Gene Ontology, Protein Domains, Transcripts, Interactions and Pathways, Antibodies, Plasmids, Constructs, Marker Relationships, Sequences, Orthology, and Note. The 'Summary' category is highlighted with a teal bar. The main content area displays detailed information for the *dmd* gene, including its ID (ZDB-GENE-010426-1), Name (*dystrophin*), Symbol (*dmd*), Previous Names (*cb664*, *Dp71*, *Duchenne muscular dystrophy*, *im:6911785*, *sap*, *sapje-like*, *sapje*, *zfDYS*, *zgc:110165*), Type (*protein_coding_gene*), and Location (Chr: [Mapping Details/Browsers](#)). The 'Description' section provides a detailed biological summary. Two specific resources are highlighted with red circles: 'Alliance' and 'Ensembl(GRCz11):ENSDARG00000008487'. The 'Note' field is listed as 'None'.

GENE	
dmd	dmd
ID	ZDB-GENE-010426-1
Name	<i>dystrophin</i>
Symbol	<i>dmd</i> Nomenclature History
Previous Names	<i>cb664</i> (1), <i>Dp71</i> (1), <i>Duchenne muscular dystrophy</i> (1), <i>im:6911785</i> , <i>sap</i> , <i>sapje-like</i> (1), <i>sapje</i> , <i>zfDYS</i> (1), <i>zgc:110165</i>
Type	protein_coding_gene
Location	Chr: Mapping Details/Browsers
Description	Predicted to have actin binding activity and zinc ion binding activity. Involved in several processes, including sarcomere organization; skeletal muscle organ development; and somatic muscle development. Localizes to sarcolemma. Used to study Duchenne muscular dystrophy and muscular dystrophy. Human ortholog(s) of this gene implicated in cognitive disorder; dilated cardiomyopathy (multiple); intellectual disability; and muscular dystrophy (multiple). Is expressed in several structures, including axial mesoderm; axis; chordo neural hinge; musculature system; and somite. Orthologous to human DMD (<i>dystrophin</i>).
Genome	Alliance (1), Gene:83773 (1),
Resources	VEGAS (1), DARK000000031909 (1), Ensembl(GRCz11):ENSDARG00000008487 (1)
Note	None

Finding Your Gene

- Follow link from ZFIN
- **Search by gene name on Ensembl (or old name or mutant name)**

The screenshot shows the Ensembl search interface for the species Zebrafish. The search term 'dmd' has been entered into the search bar, resulting in 21 matches. The results are displayed in a table with columns for the result type (e.g., Gene, Transcript, GeneTree, GenomicAlignment) and the count of matches (e.g., 2, 13, 1, 5). Below the search bar, there is a 'Did you mean...' dropdown showing alternative search terms like 'dmd (Zebrafish Gene)' and 'dmd-201'. The layout section indicates the results are currently shown in 'Table' format. A tip at the bottom suggests changing the species to see more relevant results.

Current selection:
< all Species
Only searching Zebrafish

Restrict category to:

Gene	2
Transcript	13
GeneTree	1
GenomicAlignment	5

Per page:

Layout:

Tip:

You can choose which results appear near the top of your search by updating your favourite species.

Only searching Zebrafish dmd

21 results match dmd when restricted to species: Zebrafish

Did you mean... ▾

dmd (Zebrafish Gene)
[ENSDARG0000008487](#) 1:10824351-11075405-1
Dystrophin [Source:ZFIN;Acc:ZDB-GENE-010426-1]

dmd-201 (ZFIN transcript name record; description: dystrophin,) is an external reference matched to Transcript ENSDART00000007013

[Variant table](#) • [Phenotypes](#) • [Location](#) • [External Refs.](#) • [Regulation](#) • [Orthologues](#) • [Gene tree](#)

dmd (Zebrafish Alternative sequence Gene)
[ENSDARG00000115779](#) CHR_ALT_CTG1_1_4:10997816-11031921-1
Dystrophin [Source:ZFIN;Acc:ZDB-GENE-010426-1]

dmd-213 (ZFIN transcript name record; description: dystrophin,) is an external reference matched to Transcript ENSDART00000164141

Not a Primary Assembly Gene

[Variant table](#) • [Phenotypes](#) • [Location](#) • [External Refs.](#) • [Regulation](#) • [Gene tree](#)

dmd-211 (Zebrafish Transcript)
[ENSDART00000148305](#) 1:10826296-10841348-1
Dystrophin [Source:ZFIN;Acc:ZDB-GENE-010426-1].

[Location](#) • [External Refs.](#) • [cDNA seq.](#) • [Exons](#) • [Variant table](#) • [Protein seq.](#) • [Population](#) • [Protein summary](#)

Finding Your Gene

- Follow link from ZFIN
- Search by gene name on Ensembl (or old name or mutant name)
- Search using BLAST or BLAT on Ensembl
 - BLAT is faster
 - BLAST finds more distant alignments + alternative scaffolds
 - No BLAST/BLAT on Ensembl archive sites but can use BLAT on UCSC

The screenshot shows the Ensembl web interface for a BLAST/BLAT search. The URL is https://www.ensembl.org/Blast/BLAT?species=Zebrafish&assembly=GRCz11&search_type=BLASTN&query=dmd. The search term 'dmd' is highlighted with a red circle. The results table displays four hits for the gene dmd:

Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val	%ID	
CHR_ALT_CTG1_1_4:11031622-11032521 [Sequence]	dmd	Reverse	1	900	900	[Sequence]	1779	0.0	100.000 [Alignment]
1:11031622-11032521 [Sequence]	dmd	Reverse	1	900	900	[Sequence]	1779	0.0	100.000 [Alignment]
1:11624334-1624715 [Sequence]	clic6	Reverse	1	387	390	[Sequence]	686	0.0	96.923 [Alignment]
1:11349264-11349645 [Sequence]	sdk1b	Reverse	1	387	390	[Sequence]	686	0.0	96.923 [Alignment]
1:45265987-45266368 [Sequence]	EV15L	Reverse	1	387	390	[Sequence]	686	0.0	96.923 [Alignment]

Finding Your Gene

- Follow link from ZFIN
- Search by gene name on Ensembl (or old name or mutant name)
- Search using BLAST or BLAT on Ensembl
 - BLAT is faster
 - BLAST finds more distant alignments + alternative scaffolds
 - No BLAST/BLAT on Ensembl archive sites but can use BLAT on UCSC
- Check gene correct by checking orthologues and/or synteny

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Japanese medaka HdrR (<i>Oryzias latipes</i>)	1-to-many	dmd (ENSOURLG00000020638)	84.52 %	88.22 %	0	95.81	Yes
		View Gene Tree Compare Regions (2:208,119-221,155:1) View Sequence Alignments					
Lumpfish (<i>Cyclopterus lumpus</i>)	1-to-many	dmd (ENSLCMLG00005009931)	82.21 %	82.49 %	0	95.20	Yes
		View Gene Tree Compare Regions (2:5,248,684-5,281,983:1) View Sequence Alignments					
Lyretail cichlid (<i>Neolamprologus brichardi</i>)	1-to-1	dmd (ENSNBRG00000015200)	87.34 %	89.39 %	0	96.66	Yes
		View Gene Tree Compare Regions (JH422367.1:2,004,027-2,028,054:1) View Sequence Alignments					
Makobe Island cichlid (<i>Pundamilia nyererei</i>)	1-to-1	dmd (ENSPNYG00000022641)	45.32 %	89.56 %	0	96.75	Yes
		View Gene Tree Compare Regions (JH419417.1:620,205-712,305:1) View Sequence Alignments					

Gene Names

- Names assigned to Ensembl genes automatically based on **sequence similarity**
 - Mistakes are possible
 - Names can change
- ZFIN gene symbols** (i.e. the name assigned by ZFIN) are preferred (>23,000 genes), but other databases are also used, e.g. HGNC for ~150 genes, miRBase for ~300 genes
- Description indicates source of name
- Genes without a match are given a name based on the sequence used to identify them, e.g AL645792.1 (clone) or **CABZ01052570.1** (WGS)

Gene: dmd ENSDARG00000008487

Description

dystrophin [Source:ZFIN:Acc:[ZDB-GENE-010426-1](#)]

Gene Synonyms

Dp71, Duchenne muscular dystrophy, cb664, im:6911785, sap, sapje, sapje-like, zfDYS, zgc:110165

Stable IDs

- Best to use stable IDs
- e.g. **ENSDARG00000028213** (ttn.2 or ttna)
- **ENS** = Ensembl

Stable IDs

- Best to use stable IDs
- e.g. **ENSDARG00000028213** (ttn.2 or ttna)
- **ENS** = Ensembl
- **DAR** = *Danio rerio*

Stable IDs

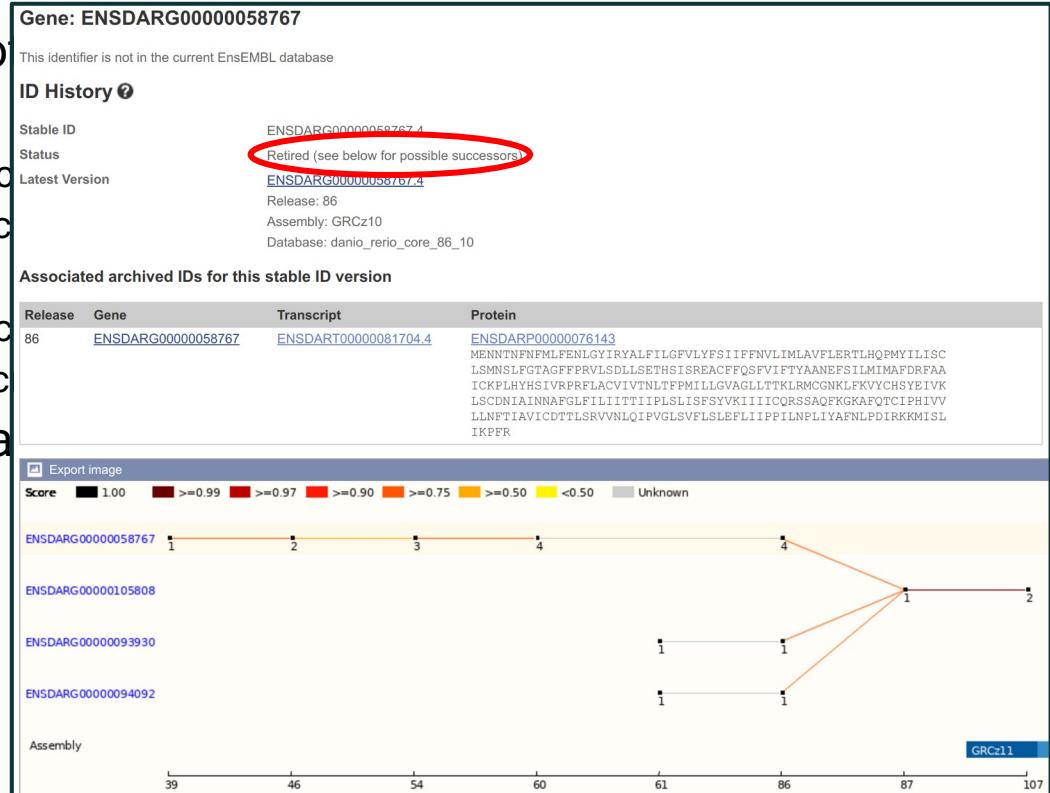
- Best to use stable IDs
- e.g. ENSDARG00000028213 (ttn.2 or ttna)
- ENS = Ensembl
- DAR = Danio rerio
- G = Gene (also T for Transcript, P for Peptide and E for exon)

Stable IDs

- Not completely stable, if annotation or underlying assembly changes
- Stable IDs have **versions**, e.g. ENSDARG00000058767.4
 - Version number of **ENSDARG** increases if transcripts change
 - Version number of **ENSDART** increases if splicing, chromosome or sequence of transcript change
 - Version number of **ENSDARP** increases if peptide's sequence changes
 - Version number of **ENSDARE** increases if exon's sequence changes
- Can also be **removed**, e.g. searching for ENSDARG00000058767

Stable IDs

- Not completely stable, if annotations change
- Stable IDs have **versions**, e.g.
 - Version number of **ENSDARG** increases over time
 - Version number of **ENSDART** increases over time, if annotations change
 - Version number of **ENSDARP** increases over time
 - Version number of **ENSDARE** increases over time
- Can also be **removed**, e.g. see below

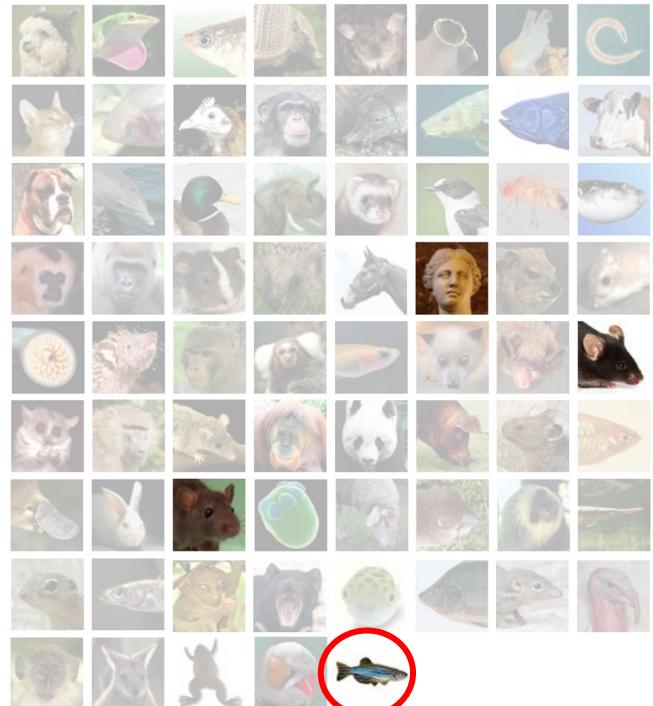


Stable IDs

- Not completely stable, if annotation or underlying assembly changes
- Stable IDs have **versions**, e.g. ENSDARG00000058767.4
 - Version number of **ENSDARG** increases if transcripts change
 - Version number of **ENSDART** increases if splicing, chromosome or sequence of transcript change
 - Version number of **ENSDARP** increases if peptide's sequence changes
 - Version number of **ENSDARE** increases if exon's sequence changes
- Can also be **removed**, e.g. searching for ENSDARG00000058767
- Can use www.ensembl.org/Danio_rerio/Tools/IDMapper to convert older IDs to what they **map** to currently in Ensembl

Gene Annotation

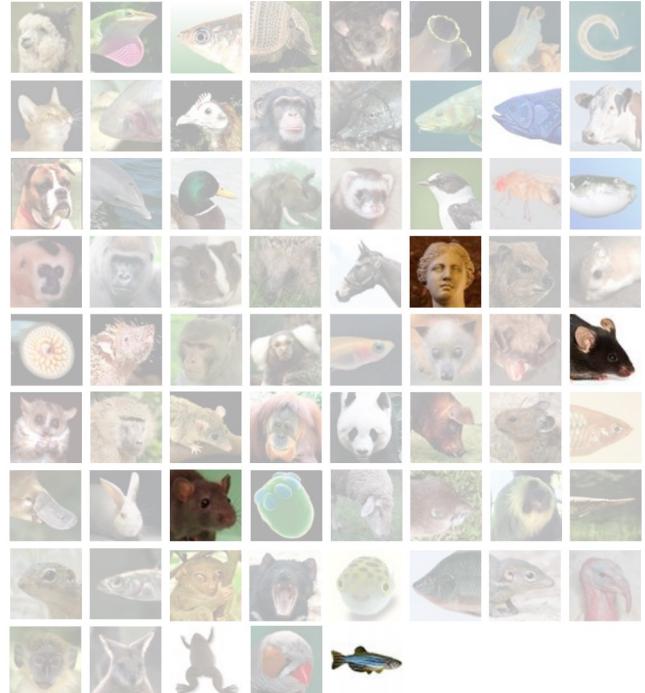
- Zebrafish (+ human, mouse, rat) has **manual** and **automatic** gene annotation
- Other **300+** genomes in Ensembl only have automatic annotation
- www.ensembl.org/info/about/species.html



From Ensembl training materials, CC BY 4.0 license

Manual Annotation

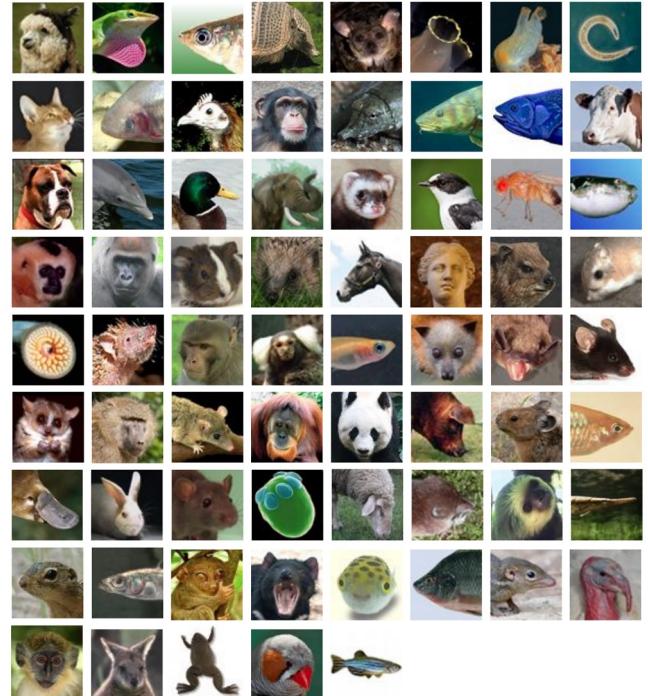
- **Gold** standard
- Uses information from databases and publications
- More accurate for tricky areas:
 - e.g. UTRs, splice sites, single exon transcripts
- **Slower** and more expensive
- Thorough, but leads to inclusion of transcripts that may not be representative (e.g. low expression)
- Only clones manually annotated



From Ensembl training materials, CC BY 4.0 license

Automatic Annotation

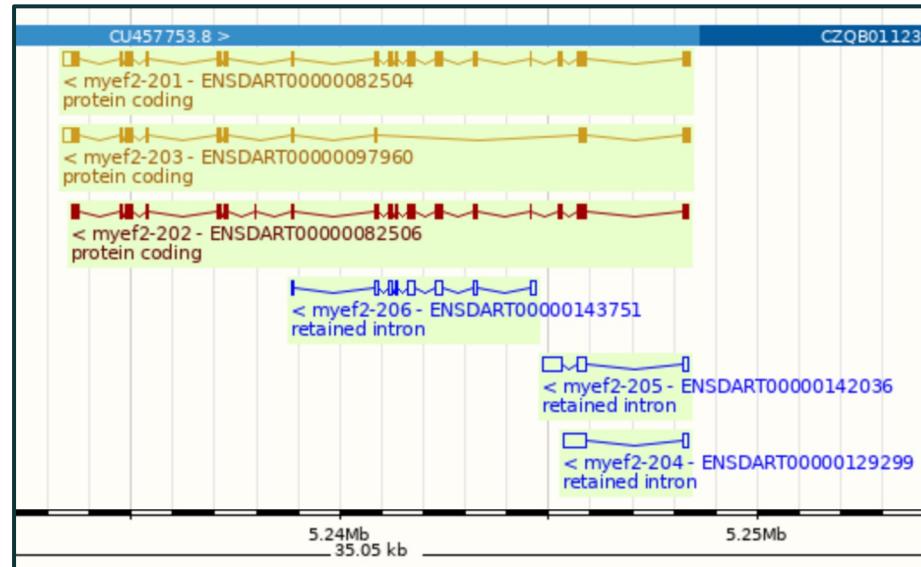
- **Faster**
- Uses evidence from sequences **deposited** in ENA/GenBank/DDBJ and UniProt proteins
- **Overview:**
 - Identify repeats and low complexity sequence with RepeatMasker, Dust and TRF
 - Run GENSCAN to identify *ab initio* gene predictions
 - Align UniProt proteins to GENSCAN predictions, prioritising zebrafish proteins or those from closely related or well annotated species
 - Make gene models using Genewise
 - Align cDNAs, ESTs and RNA-seq to annotate UTRs and make RNA-seq gene models
 - Collapse redundant transcripts and cluster into genes, prioritising manual annotation but including automatic annotation if different splicing
 - Identify pseudogenes by looking for genes with frameshifts / repeats
 - Identify processed pseudogenes by looking for multi-exon equivalent



From Ensembl training materials, CC BY 4.0 license

Merged Annotation

- **Golden:** **Identical** manual and automatic annotation
- **Red:** **Protein-coding** transcript from automatic annotation
- **Blue:** **Non-coding** transcript
- Filled box: **Coding exon**
- Non-filled box: **Non-coding** exon



- In reality, would not trust these retained intron transcripts unless shown to have comparable expression levels

Which Transcript?

- Often **multiple** transcripts
- **Best** transcript for experiments?
- Golden transcript is a good bet
- **Ensembl Canonical** transcript is, on balance, most conserved, most expressed, longest CDS (coding sequence) and in other databases
- APPRIS combines protein structure, important residues and homology to identify a **principal isoform** - APPRIS P1

Gene: **babam1** ENSDARG0000077526

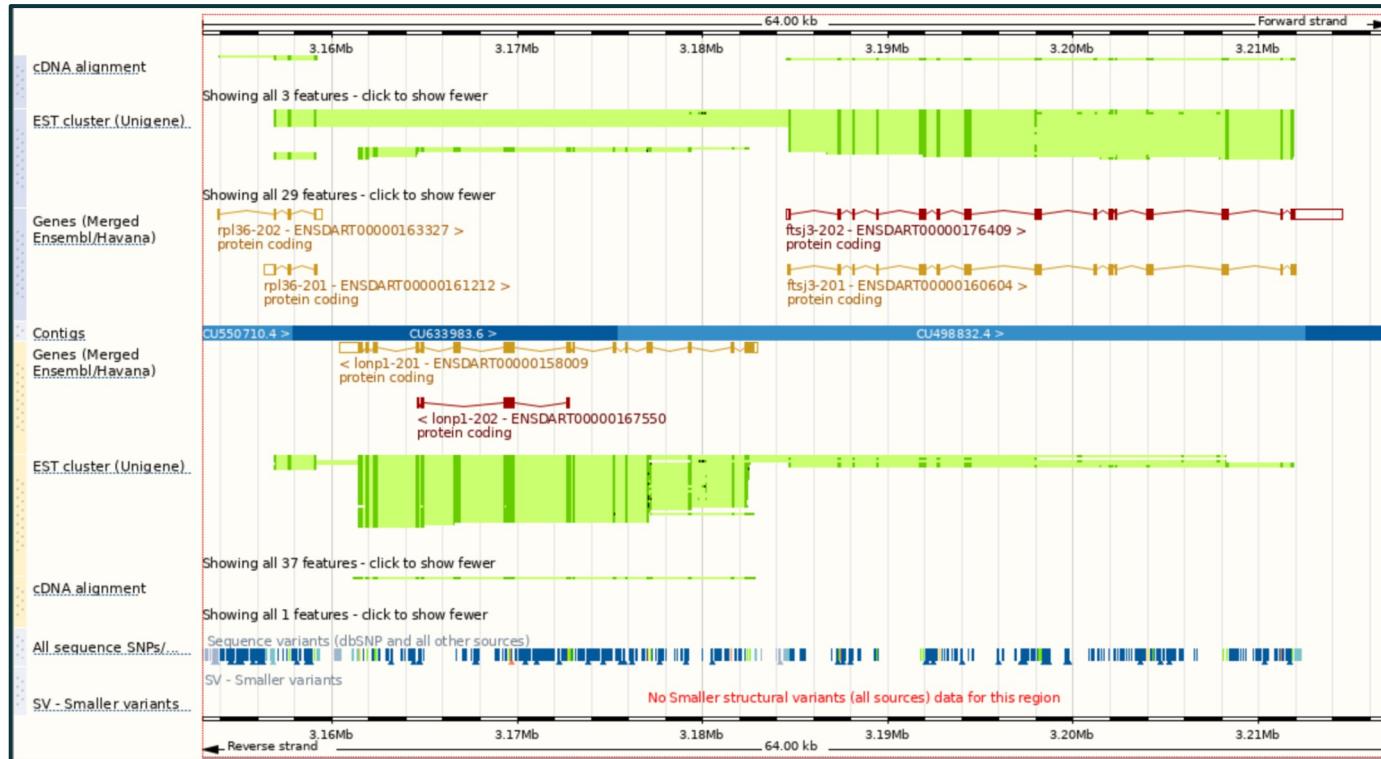
Description: BRISC and BRCA1 A complex member 1 [Source:NCBI gene;Acc:445296]
Gene Synonyms: zgc:100909
Location: Chromosome 11: 6,051,287-6,070,192 reverse strand.
GRCz11:CM002895.2
About this gene: This gene has 4 transcripts ([splice variants](#)) and [185 orthologues](#).
Transcripts: [Hide transcript table](#)

Show/hide columns (1 hidden) Filter

Transcript ID	Name	bp	Protein	Biotype	UniProt Match	Flags
ENSDART00000122262.3	babam1-202	2035	370aa	Protein coding	Q6AXK4	Ensembl Canonical APPRIS P1
ENSDART0000008980.8	babam1-201	1888	370aa	Protein coding	A0A0R4I9A4 Q6AXK4	APPRIS P1
ENSDART00000162776.2	babam1-203	802	197aa	Protein coding	A0A0R4IIK1	CDS 3' incomplete
ENSDART00000167672.2	babam1-204	790	240aa	Protein coding	A0A0R4IMN5	CDS 3' incomplete

"Region in detail" Demo

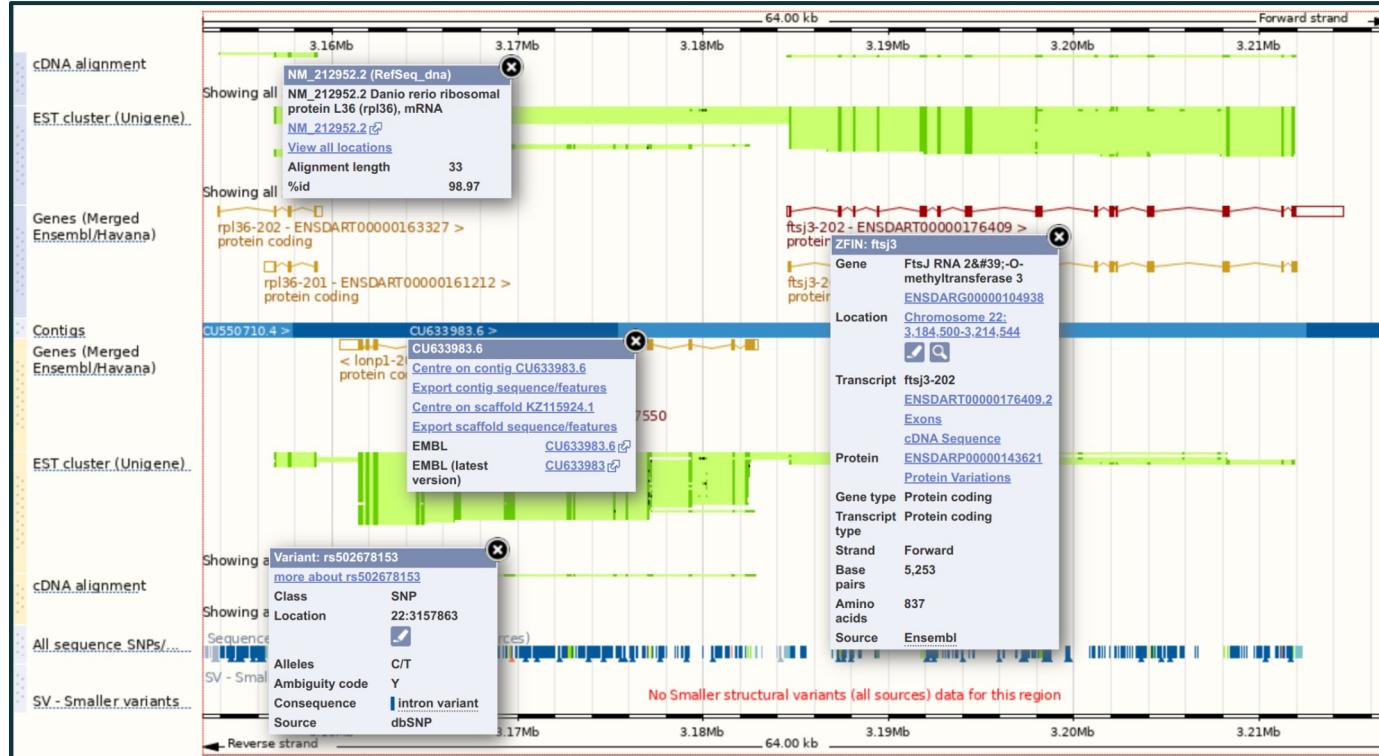
- Go to "22:3153000-3217000"



- 4 clones
- 2 genes on +
- 1 gene on -
- Manual + automatic annotation
- cDNA + EST tracks
- Variant tracks

"Region in detail" Demo

- Go to "22:3153000-3217000"



- 4 clones
- 2 genes on +
- 1 gene on -
- Manual + automatic annotation
- cDNA + EST tracks
- Variant tracks

Exercise 1

- Do Exercise 1 - “exploring the genome”
- Covers:
 - Region view
 - BLAST/BLAT
 - Archive sites
- Go to mbl2023.buschlab.org

Part 2

- Configuring Ensembl tracks
 - Ensembl “Gene” view
 - Comparative genomics
-
- But first, back to the region we were looking at before the exercises:
"22:3153000-3217000"

"Configure this page" Demo

- Go to "22:3153000-3217000" and click "Configure this page"

The screenshot shows the Ensembl genome browser interface for the Zebrafish genome. The URL in the address bar is `22:3153000-3217000`. The main content area displays a genomic track for chromosome 22, spanning from 3.16 Mb to 3.21 Mb. The track shows various genomic features, including exons (green), introns (blue), and gene models (grey). A legend indicates the track types: cDNA alignment, BAM files, Gene models, and Intron-spanning reads.

The left sidebar contains a navigation tree and a search bar for track hubs. The top navigation bar includes links for BLAST/BLAST+, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right, there are buttons for Login/Register and a search bar labeled "Search Zebrafish...".

The central configuration panel allows users to "Configure this page" for the selected genomic region. It includes sections for Active tracks, Favourite tracks, Track order, Search results, Genome Reference Consortium Issues, Sequence and assembly, Simple features, Genes and transcripts (7/96), RNASeq models (5/93), mRNA and protein alignments, Variation, Comparative genomics, Oligo probes, Repeat regions, Information and decorations, and Display options. A "Configure Region Image" tab is active.

Configuration options include selecting available configurations (Current unsaved), saving the current configuration, and filtering by RNASeq models (All classes). A key legend defines track colors: Shown (dark blue), Hidden (light blue), No Data (grey), and Filtered (green). A "Default style:" section provides options to enable or disable specific styles.

Sample	1 dpf sample1	14 dpf sample1	2 dpf sample1	3 dpf sample1	5 dpf sample1
0	0	0	0	0	0
1	1	1	1	1	1
2	1	1	1	1	1
3	0	1	1	1	1
4	0	1	1	1	1
5	0	1	1	1	1

RefSeq Aside

- NCBI's **annotated** and **curated** database of reference sequences, including transcripts and proteins
- Accessions starting **X** are "Model RefSeq" **predictions** from automatic genome annotation
- Accessions starting **N** are "Known RefSeq" from **manually curated** cDNA and EST data
- Accessions starting **NM & XM** indicate mRNA; **NP & XP** are proteins

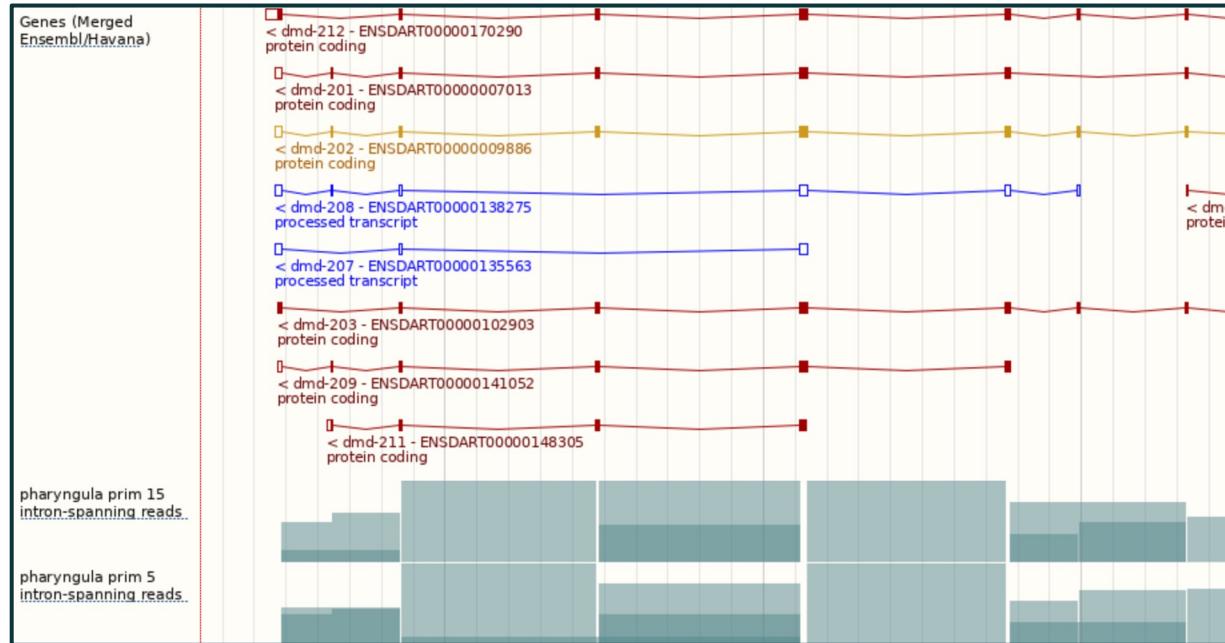
"Configure this page" Demo

- Go to "22:3153000-3217000" and click "Configure this page"



"Configure this page" Demo

- Go to "1:10822281-10882903" and click "Configure this page"
- Under "RNASeq models", turn on "Intron-spanning reads" for "pharyngula prim 5" and "pharyngula prim 15"



"Gene" Demo - Summary

- Go to ENSDARG00000102765

Gene-based displays

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence
 - Secondary Structure
- Comparative Genomics
 - Genomic alignments
 - Gene tree
 - Gene gain/loss tree
 - Orthologues
 - Paralogues
 - Ensembl protein families
- Ontologies
 - GO: Cellular component
 - GO: Biological process
 - GO: Molecular function
- Phenotypes
- Genetic Variation
 - Variant table

Gene: lonp1 ENSDARG00000102765

Description Ion peptidase 1, mitochondrial [Source:ZFIN;Acc:[ZDB-GENE-030131-4006](#)]

Gene Synonyms fc64d11, prss15, wu:fc64d11

Location [Chromosome 22: 3,160,447-3,182,965](#) reverse strand.
GRCz11:CM002906.2

About this gene This gene has 2 transcripts ([splice variants](#)), [190 orthologues](#) and [1 parologue](#).

Transcripts

Show transcript table

Summary

Name [lonp1](#) (ZFIN)

Ensembl version ENSDARG00000102765.2

Gene type Protein coding

Annotation method Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see [article](#).

"Gene" Demo - Transcript Table

- Go to ENSDARG00000102765 and click on "Show transcript table"

Gene-based displays

- Summary** (selected)
- Splice variants
- Transcript comparison
- Gene alleles

- Sequence**
- Secondary Structure

- Comparative Genomics**
- Genomic alignments
- Gene tree
- Gene gain/loss tree
- Orthologues
- Paralogues
- Ensembl protein families

- Ontologies**
- GO: Cellular component
- GO: Biological process
- GO: Molecular function

- Phenotypes**

- Genetic Variation**
- Variant table
- Variant image
- Structural variants

- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence

- ID History

Gene: Ionp1 ENSDARG00000102765

Description Ion peptidase 1, mitochondrial [Source:ZFIN;Acc:[ZDB-GENE-030131-4006](#)]

Gene Synonyms fc64d11, prss15, wu:fc64d11

Location [Chromosome 22: 3,160,447-3,182,965](#) reverse strand.
GRCz11:CM002906.2

About this gene This gene has 2 transcripts ([splice variants](#)), [190 orthologues](#) and [1 parologue](#).

Transcripts [Hide transcript table](#)

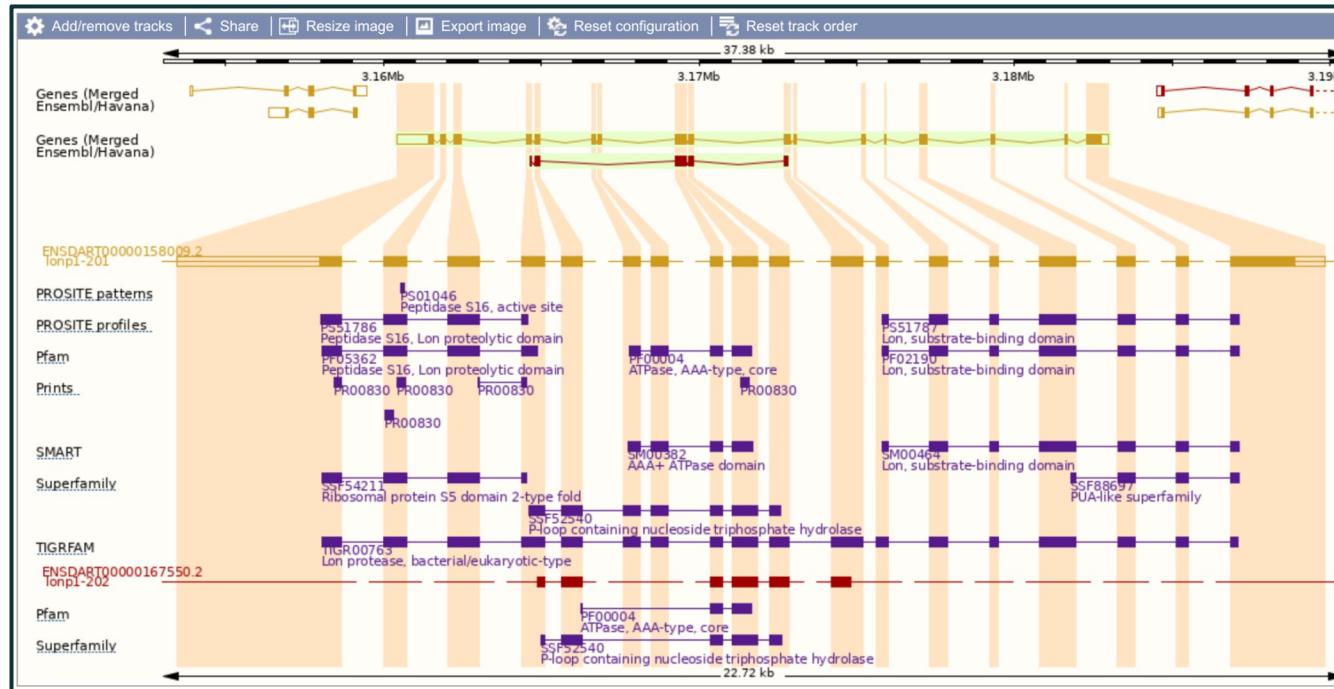
Show/hide columns (1 hidden)							Filter
Transcript ID	Name	bp	Protein	Biotype	UniProt Match	Flags	
ENSDART00000158009.2	Ionp1-201	4114	966aa	Protein coding	A0A0R4IH79	Ensembl Canonical	APPRIS P1
ENSDART00000167550.2	Ionp1-202	741	247aa	Protein coding	A0A0R4IPW4	CDS 5' and 3' incomplete	

Summary [?](#)

Name	Ionp1 (ZFIN)
Ensembl version	ENSDARG00000102765.2
Gene type	Protein coding
Annotation method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article .

"Gene" Demo - Splice Variants

- Go to ENSDARG00000102765 and click on "Splice variants"



"Gene" Demo - Orthologues

- Go to ENSDARG00000102765 and click on "Orthologues"

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Abingdon island giant tortoise (<i>Chelonoidis abingdonii</i>)	1-to-1 View Gene Tree	LONP1 (ENSCABG00000010924) Compare Regions (PKMU01001122.1:170,198-221,187:1) View Sequence Alignments	75.03 %	75.26 %	25	n/a	No
African ostrich (<i>Struthio camelus australis</i>)	1-to-1 View Gene Tree	LONP1 (ENSSCUG00000004632) Compare Regions (KL206174.1:174,870-201,335:1) View Sequence Alignments	80.69 %	70.50 %	50	n/a	Yes
Algerian mouse (<i>Mus spretus</i>)	1-to-1 View Gene Tree	Lonp1 (MGP_SPRETEIJ_G0022694) Compare Regions (17:54,555,161-54,567,779:-1) View Sequence Alignments	75.05 %	74.12 %	0	n/a	No
Alpine marmot (<i>Marmota marmota marmota</i>)	1-to-1 View Gene Tree	LONP1 (ENSMMMG00000018859) Compare Regions (CZRN01000089.1:3,499,006-3,520,539:-1) View Sequence Alignments	62.80 %	62.22 %	0	n/a	No
Amazon Molly (<i>Poecilia formosa</i>)	1-to-1 View Gene Tree	lonp1 (ENSPFOG00000001826) Compare Regions (KI520250.1:178,559-209,184:-1) View Sequence Alignments	75.94 %	77.43 %	0	85.71	Yes

"Gene" Demo - Paralogues

- Go to ENSDARG00000102765 and click on "Paralogues"

Gene: lonp1 ENSDARG00000102765

Description Ion peptidase 1, mitochondrial [Source:ZFIN;Acc:[ZDB-GENE-030131-4006](#)] Gene Synonyms fc64d11, prss15, wu:fc64d11 Location Chromosome 22: 3,160,447-3,182,965 reverse strand. GRCz11:CM002906.2 About this gene This gene has 2 transcripts ([splice variants](#)), [190 orthologues](#) and [1 parologue](#). Transcripts [Hide transcript table](#)

Show/hide columns (1 hidden) Filter

Transcript ID	Name	bp	Protein	Biotype	UniProt Match	Flags
ENSDART00000158009.2	lonp1-201	4114	966aa	Protein coding	A0A0R4IH79	Ensembl Canonical APPRI P1
ENSDART00000167550.2	lonp1-202	741	247aa	Protein coding	A0A0R4IPW4	CDS 5' and 3' incomplete

Paralogues [?](#)

[Download paralogues](#)

Show/hide columns Filter

Type	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Paralogues	Bilateral animals (Bilateria)	ENSDARG00000101438 lonp2 Ion peptidase 2, peroxisomal [Source:NCBI gene;Acc:494030]	<ul style="list-style-type: none">Region ComparisonAlignment (protein)Alignment (cDNA)	18:18,475,674-18,524,624:-1	36.31 %	31.57 %

"Gene" Demo - GO Terms

- Go to ENSDARG00000102765 and click on "GO: Molecular function"

GO: Molecular function 					
Show/hide columns (1 hidden) Filter  					
Accession	Term	Evidence	Annotation source	Transcript IDs	
GO:0000166 	nucleotide binding	IEA	UniProt	ENSDART00000158009	<ul style="list-style-type: none">Search BioMartView on karyotype
GO:0003677 	DNA binding	IEA	UniProt	ENSDART00000158009	<ul style="list-style-type: none">Search BioMartView on karyotype
GO:0003697 	single-stranded DNA binding	IBA	GO_Central	ENSDART00000158009	<ul style="list-style-type: none">Search BioMartView on karyotype
GO:0004176 	ATP-dependent peptidase activity	IBA	GO_Central	ENSDART00000167550 ENSDART00000158009	<ul style="list-style-type: none">Search BioMartView on karyotype
GO:0005524 	ATP binding	IEA	UniProt	ENSDART00000158009 ENSDART00000167550	<ul style="list-style-type: none">Search BioMartView on karyotype
GO:0016887 	ATP hydrolysis activity	IEA	UniProt	ENSDART00000158009 ENSDART00000167550	<ul style="list-style-type: none">Search BioMartView on karyotype
GO:0043565 	sequence-specific DNA binding	IEA	UniProt	ENSDART00000158009	<ul style="list-style-type: none">Search BioMartView on karyotype

"Gene" Demo - External References

- Go to ENSDARG00000102765 and click on “External references”

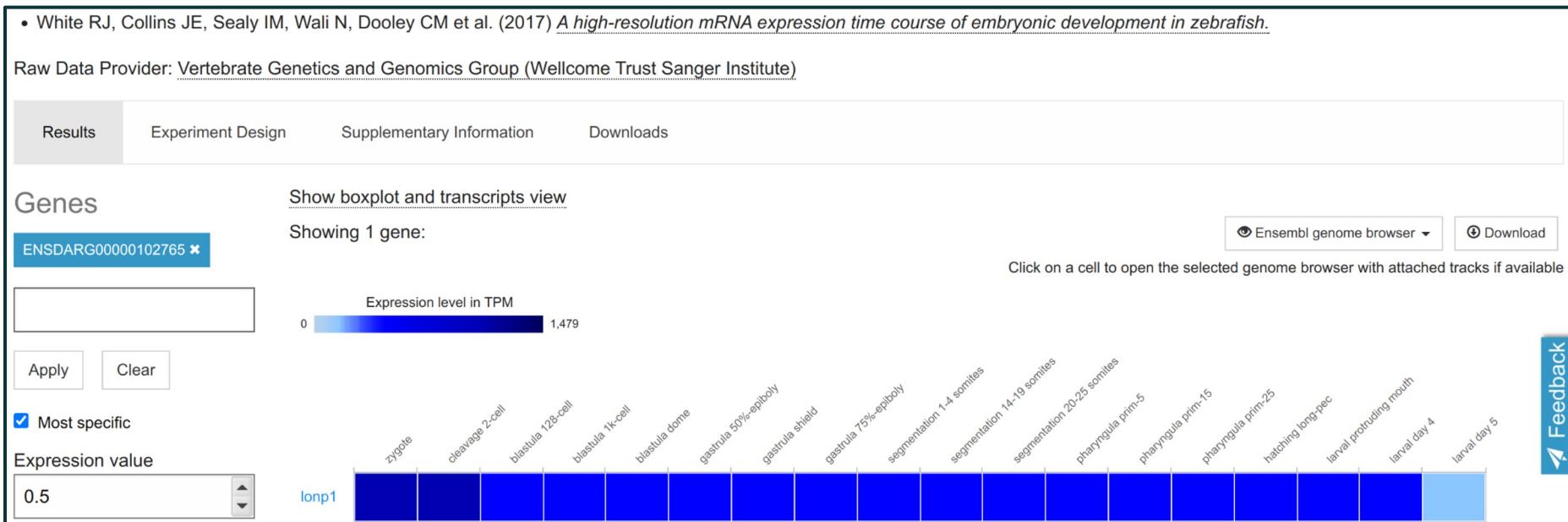
External references

This gene corresponds to the following database identifiers:

External database	Database identifier	Filter 
Expression Atlas	ENSDARG00000102765  [view all locations]	
NCBI gene (formerly Entrezgene)	Iotp1  Ion peptidase 1, mitochondrial [view all locations]	
WikiGene	Iotp1  Ion peptidase 1, mitochondrial [view all locations]	
ZFIN	Iotp1  Ion peptidase 1, mitochondrial [view all locations]	

"Gene" Demo - Expression Atlas

- From “External references” click “Expression Atlas” ID then “18 White et al”

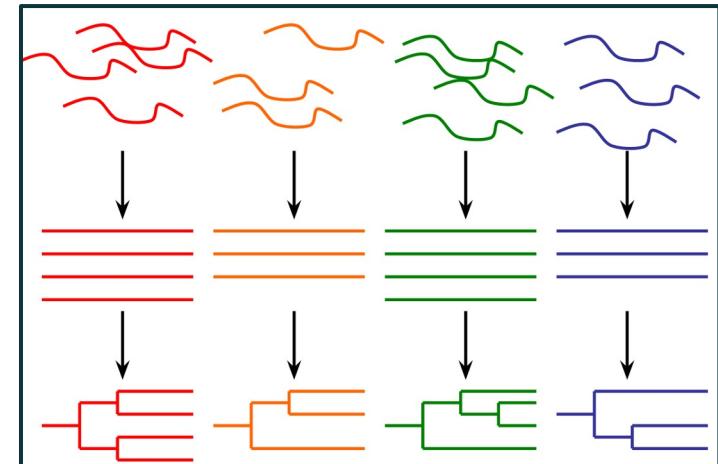


Compara

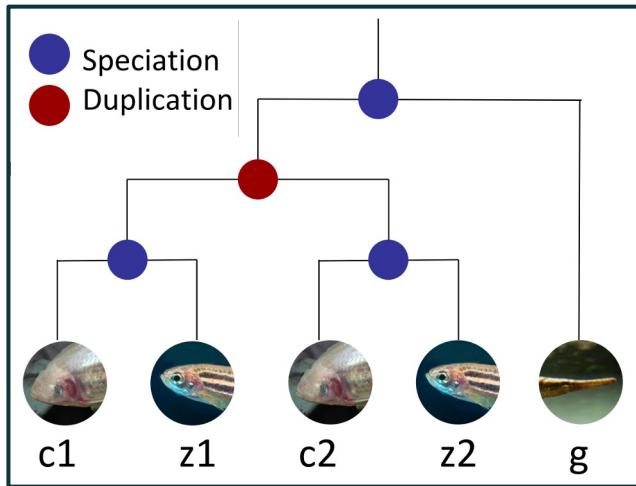
- Compara - produce Ensembl's comparative genomics resources
- Two types of analysis:
 - Gene level comparisons to produce **gene trees**, e.g. infer **homologues** (orthologues & paralogues)
 - **Whole genome alignments** - pairwise and multiple alignments, e.g. **constrained elements** and **synteny**

Compara - Gene Trees

- Separate trees for **proteins** and **ncRNAs** (take secondary structure into account)
- Process:
 - Take **representative** transcripts (e.g. longest CDS) from all genes from all species
 - Classify genes into **clusters** by TreeFam family
 - Build **multiple** alignment
 - Build **gene tree** reconciled with NCBI's taxonomy tree
 - Infer **orthologues** and **paralogues**



Compara - Infer Homologues (Orthologues & Paralogues)



z1 & z2 are **paralogues** (arose from **duplication**), as are **c1 & c2**

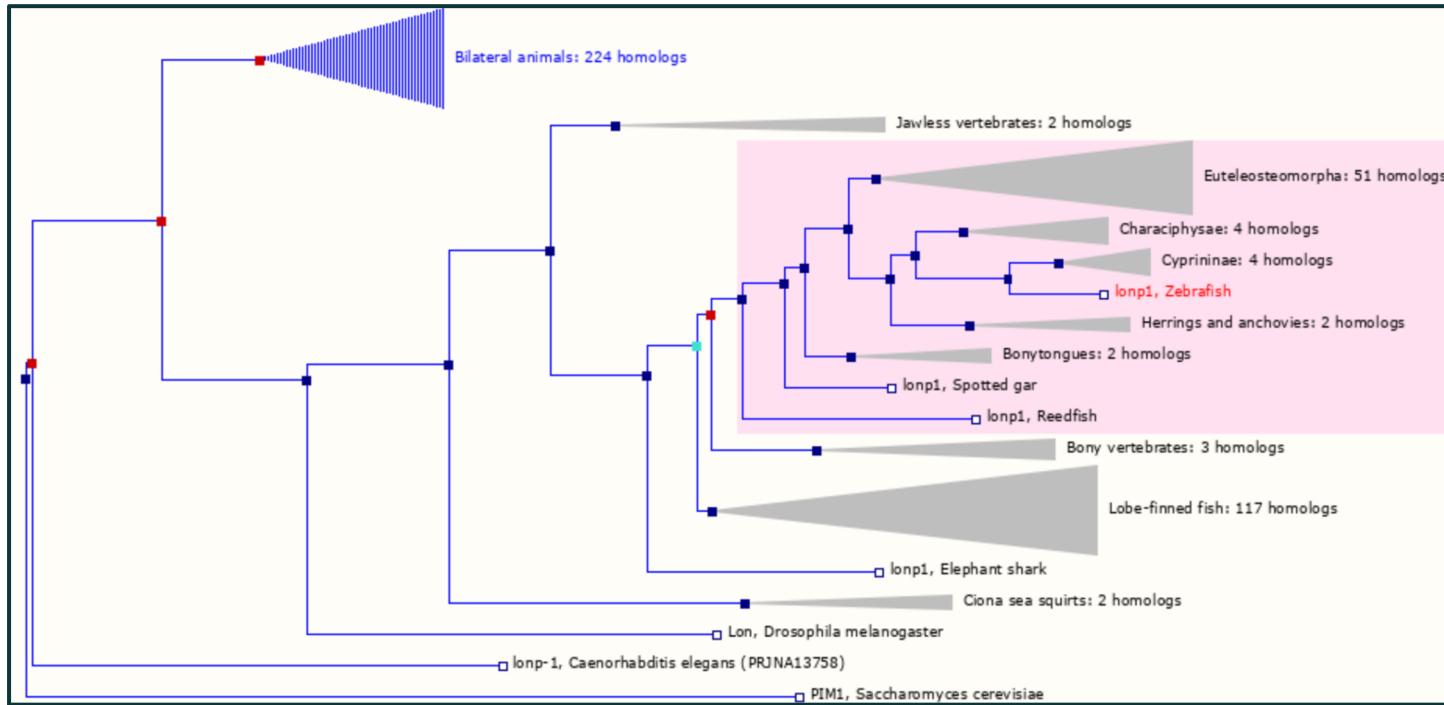
z1 & c1 are **orthologues** (arose from **speciation**), as are **z2 & c2 + z2 & g**, etc...

z1 & c1 have a **one-to-one** relationship

g has a **one-to-many** relationship to e.g. **z1** and **z2**

Homologues labelled "**high confidence**" are supported by conservation of synteny or whole genome alignment blocks

Compara - lonp1 Gene Tree



Compara - Whole Genome Alignments

- **Pairwise whole genome alignments** with LASTZ
- Zebrafish has alignments to **64 species** (plus itself)
- Only human (181) and medaka (65) have more
- Full list at: www.ensembl.org/info/genome/compara/analyses.html
- **Multiple genome alignments** with EPO (Enredo, Pecan, Ortheus)
- Zebrafish is in **2** alignments (out of 11 in Ensembl) - one of **39 fish** and one of **65 fish**
- For lists of species, see:
www.ensembl.org/info/genome/compara/multiple_genome_alignments.html

Synteny Example

- No zebrafish orthologue listed for human RBM20 gene (ENSG00000203867)

 **Species without orthologues**

22 species are not shown in the table above because they don't have any orthologue with ENSG00000203867.

- Ancestral sequence
- Siamese fighting fish (*Betta splendens*)
- Sloth (*Choloepus hoffmanni*)
- Channel bull blenny (*Cottoperca gobio*)
- Lumpfish (*Cyclopterus lumpus*)
- Tongue sole (*Cynoglossus semilaevis*)
- Common carp (*Cyprinus carpio carpio*)
- **Zebrafish (*Danio rerio*)**

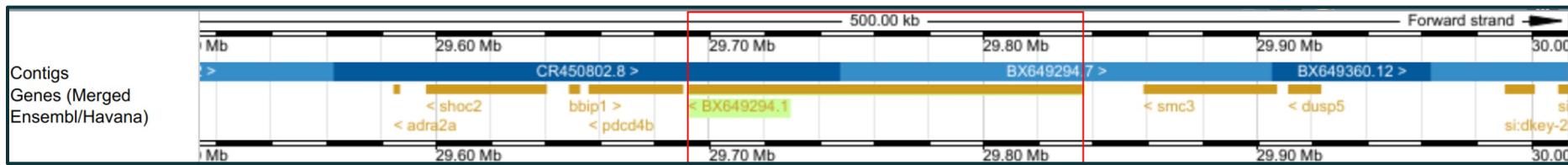
Synteny Example

- If we look at the region around RBM20 in human and then click on **Synteny** we see conservation of synteny with zebrafish chr22

Homo sapiens genes	Location		Danio rerio homologues	Location	
DUSP5 (ENSG00000138166)	10:110497907-110511533	→	dusp5 (ENSDARG00000019307)	22:29911326-29922872	Region Comparison
SMC3 (ENSG00000108055)	10:110567684-110606048	→	smc3 (ENSDARG00000019000)	22:29858535-29906764	Region Comparison
RBM20 (ENSG00000203867)	10:110644336-110839468		No homologues		
PDCD4 (ENSG00000150593)	10:110871795-110900006	→	pdcd4b (ENSDARG00000041022)	22:29655981-29689981	Region Comparison
BBIP1 (ENSG00000214413)	10:110898730-110919201	→	bbip1 (ENSDARG00000071046)	22:29648854-29652356	Region Comparison
SHOC2 (ENSG00000108061)	10:110919367-111017307	→	shoc2 (ENSDARG00000040853)	22:29596646-29640181	Region Comparison
ADRA2A (ENSG00000150594)	10:111077029-111080907	→	adra2a (ENSDARG00000040841)	22:29584800-29586608	Region Comparison

Synteny Example

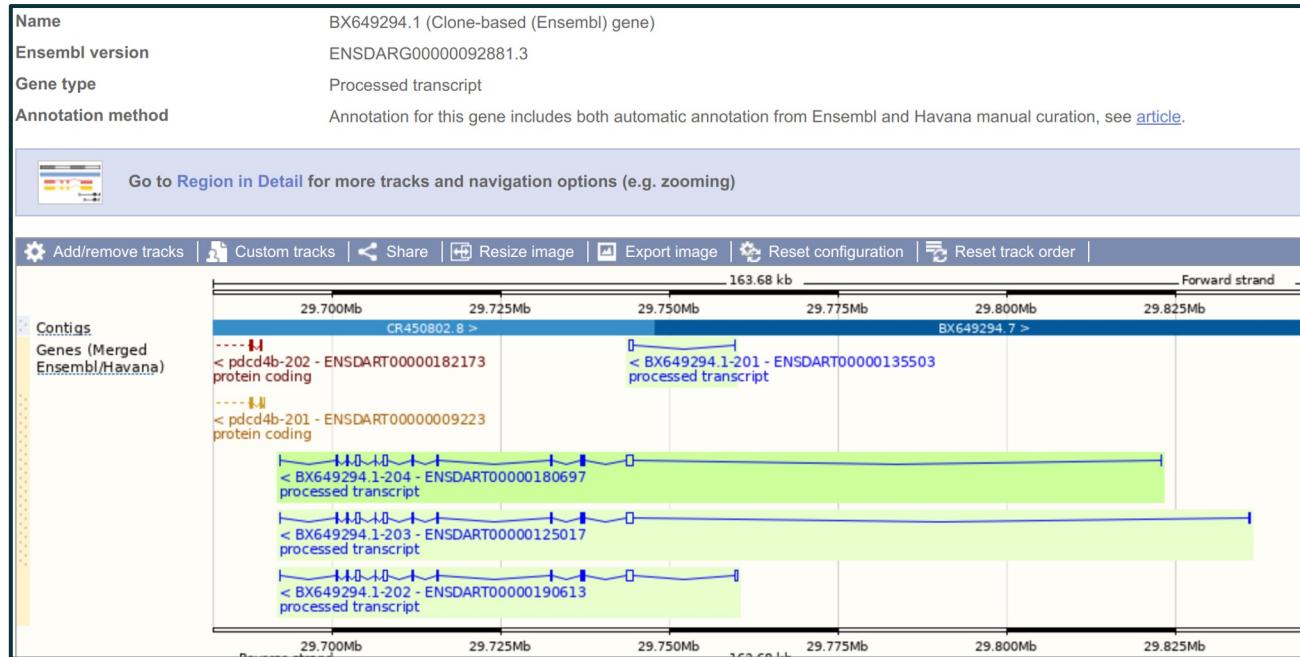
- If we look at the chr22 region in zebrafish then all the surrounding genes are the same and RBM20 is likely to be BX649294.1



Homo sapiens genes	Location	Danio rerio homologues	Location	Region Comparison
DUSP (ENSG00000138166)	10:110497907-110511533	dusp (ENSDARG00000019307)	22:29911326-29922872	Region Comparison
SMC3 (ENSG00000108055)	10:110567684-110606048	smc3 (ENSDARG00000019000)	22:29858535-29906764	Region Comparison
RBM20 (ENSG00000203867)	10:110644336-110839468	No homologues		
PDCD4 (ENSG00000150593)	10:110871795-110900006	pdcd4b (ENSDARG00000041022)	22:29655981-29689981	Region Comparison
BBIP1 (ENSG00000214413)	10:110898730-110919201	bbip1 (ENSDARG00000071046)	22:29648854-29652356	Region Comparison
SHOC2 (ENSG00000108061)	10:110919367-111017307	shoc2 (ENSDARG00000040853)	22:29596646-29640181	Region Comparison
ADRA2A (ENSG00000150594)	10:111077029-111080907	adra2a (ENSDARG00000040841)	22:29584800-29586608	Region Comparison

Synteny Example

- Erroneously labelled as processed transcript and so not in protein gene tree, so not labelled as orthologue or named by orthology



Exercise 2

- Do Exercise 2 - “exploring genes”
- Covers:
 - Gene view
 - Phenotypes
 - Gene Ontology
 - Homologues
 - Gene trees
 - Synteny
- Go to mbl2023.buschlab.org

Part 3

- BioMart
- Other tools
- Custom tracks

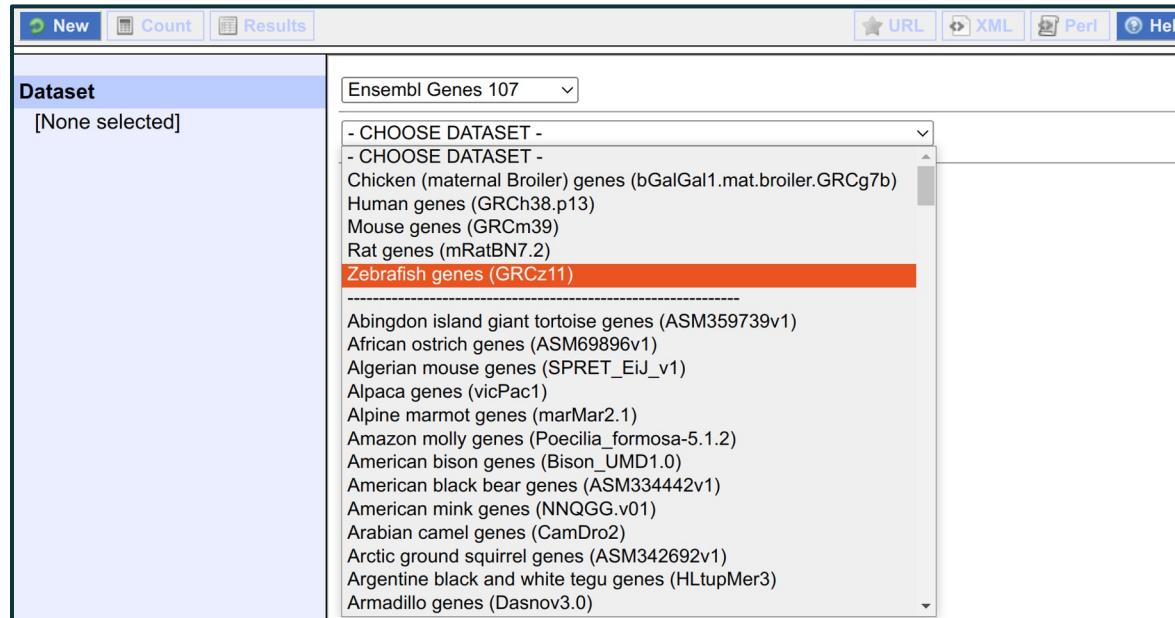
BioMart

- **Export** (large amounts of) Ensembl data without programming
- Completely **customisable**, but **simple** to make complex queries
- Four stages:
 - Dataset
 - Filters
 - Attributes
 - Results

The screenshot shows the Ensembl BioMart interface. At the top, there is a dark blue header bar with the Ensembl logo on the left and a "Login/Register" link on the right. In the center of the header is a search bar with the placeholder "Search all species..." and a magnifying glass icon. Below the header, there is a navigation menu with links: BLAST/BLAT, VEP, Tools, BioMart (which is circled in red), Downloads, Help & Docs, and Blog. At the bottom of the page, there is a light blue footer bar with several buttons: "New", "Count", "Results", "URL", "XML", "Perl", and "Help". On the far left of the footer, there is a "Dataset" section with the text "[None selected]".

BioMart - Dataset

- Choose **database** (e.g. genes or variants) and **species**



BioMart - Filters

- **Filter** to reduce the dataset
- Can select **multiple** filters
- e.g. regions, IDs, GO terms, etc...

The screenshot shows the BioMart interface for filtering Zebrafish genes (GRCz11). The left sidebar displays the dataset selection and basic attributes. The main area shows various filter options:

- Transcript count >=**: Unchecked input field.
- Transcript count <=**: Checked checkbox, with a value of "1" entered in the input field.
- Gene type**: Checked checkbox. A dropdown menu lists several gene types, with "protein_coding" highlighted in orange.
- Transcript type**: Unchecked checkbox. A dropdown menu lists transcript types, with "antisense" highlighted in orange.
- Source (gene)**: Unchecked checkbox. A dropdown menu shows "ensembl" selected.
- Source (transcript)**: Unchecked checkbox. A dropdown menu shows "ensembl" selected.
- APPRIS annotation**: Unchecked checkbox. A radio button group shows "Only" selected.

BioMart - Attributes

- What data to **export**
- e.g. IDs, genomic locations, sequences, homologues, etc...

The screenshot shows the BioMart Attributes interface. At the top, there are buttons for New, Count, Results, URL, XML, Perl, and Help. On the left, a sidebar shows 'Dataset 6 / 37241 Genes' (Zebrafish genes (GRCz11)) and 'Filters' (Chromosome/scaffold: 22, Start: 3000000, End: 4000000, Transcript count <= 1, Gene type: protein_coding). Below that is an 'Attributes' section with checkboxes for: Gene stable ID, Gene name, Source of gene name, APPRIS annotation, Chromosome/scaffold name, Gene start (bp), Gene end (bp), Strand, Chromosome/scaffold name, Gene start (bp), Gene end (bp), Strand, and Karyotype band. The 'Dataset' section at the bottom says [None Selected]. The main area has instructions: 'Please select columns to be included in the output and hit 'Results' when ready' and 'Missing non coding genes in your mart query output, please check the following FAQ'. It contains radio buttons for 'Features' (selected), 'Structures', 'Homologues (Max select 6 orthologues)', 'Variant (Germline)', and 'Sequences'. A large list of checkboxes for columns includes: APPRIS annotation, Ensembl Canonical, Readthrough, Gene name (checked), Source of gene name (checked), Transcript name, Source of transcript name, Transcript count, Gene % GC content, Gene type, Transcript type, Source (gene), and Source (transcript).

BioMart - Results

- Access your selected data in multiple formats
- e.g. HTML, TSV, CSV, XLS

New Count Results

Dataset 6 / 37241 Genes
Zebrafish genes (GRCz11)

Filters
Chromosome/scaffold: 22
Start: 3000000
End: 4000000
Transcript count <= 1
Gene type: protein_coding

Attributes
Gene stable ID
Gene name
Source of gene name
APPRIS annotation
Chromosome/scaffold name
Gene start (bp)
Gene end (bp)
Strand

Export all results to File

Unique results only

Email notification to

View rows as Unique results only

TSV HTML CSV XLS

Gene stable ID	Gene name	Source of gene name	APPRIS annotation	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Strand
ENSDARG00000103139	LO017843.1	Clone-based (Ensembl) gene	principal1	22	3045495	3078347	1
ENSDARG00000100132	CU929402.1	Clone-based (Ensembl) gene	principal1	22	3232925	3234494	1
ENSDARG00000100533	sicch1073-178p5.3	ZFIN	principal1	22	3238474	3239834	1
ENSDARG00000110077	CU929402.2	Clone-based (Ensembl) gene	principal1	22	3244950	3271707	1
ENSDARG00000053074	gipc3	ZFIN	principal1	22	3303671	3328241	1
ENSDARG00000104717	ttxa2r	ZFIN	principal1	22	3336723	3344613	-1

More Tools

The screenshot shows the Ensembl Tools page. At the top, there's a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right, there's a "Login/Register" button and a search bar. Below the navigation, a sidebar titled "In this section" lists "Ensembl Variant Effect Predictor" with sub-links for VEP web interface, VEP command line, Data formats, Variant Recoder, Haploviewer, VEP FAQ, Variant Simulator, and VCF to PED Converter. A "Search documentation" input field with a "Go" button is also present. The main content area is titled "Ensembl Tools" and contains a sub-section "Processing your data". It lists several tools with their descriptions, online status, upload limit, download script availability, and documentation links:

Name	Description	Online tool	Upload limit	Download script	Documentation
Variant Effect Predictor (VeP)	Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool.	Available	50MB*	Available	Available
Variant Recoder	Translate a variant identifier, HGVS notation or genomic SPDI notation to all possible variant IDs, HGVS, VCF format and genomic SPDI.	Available	Maximum 1000 variants recommended	Available	Available
BLAST/BLAT	Search our genomes for your DNA or protein sequence.	Available	50MB	Available	Available
File Chameleon	Convert Ensembl files for use with other analysis tools	Available		Available	Available
Assembly Converter	Map (liftover) your data's coordinates to the current assembly.	Available	50MB	Available	Available
ID History Converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	Available	50MB	Available	Available
Linkage Disequilibrium Calculator	Calculate LD between variants using genotypes from a selected population.	Available		Available	Available
VCF to PED converter	Parse a vcf file to create a linkage pedigree file (ped) and a marker information file, which together may be loaded into ld visualization tools like Haploview.	Available		Available	Available
Data Slicer	Get a subset of data from a BAM or VCF file.	Available		Available	Available
Post-GWAS	Upload GWAS summary statistics and highlight likely causal gene candidates.	Available		Available	Available

- Results from all tools can be stored indefinitely if create an **Ensembl account**

Variant Effect Predictor

- VEP predicts **consequences** of variants
- www.ensembl.org/Danio_rerio/Tools/VEP
- Example:

22 3169475 3169475 G/T 1

22 3169514 3169514 A/T 1

22 3166910 3166910 C/A 1

(Chr, Start, End, REF/ALT, Strand)

- Custom Ensembl format, but standard formats like **VCF** can be used

Variant Effect Predictor

Category	Count
Variants processed	3
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	1
Overlapped transcripts	2
Overlapped regulatory features	-

Consequences (all)

stop_gained: 33%
missense_variant: 33%
intron_variant: 17%
splice_acceptor_variant: 17%

Coding consequences

stop_gained: 50%
missense_variant: 50%

Results preview

Navigation (per variant) | Filters | Download | New job

Page: 1 of 1 | Show: All variants

Uploaded variant is defined

Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature	Biotype	cDNA position	CDS position	Protein position	Amino acids	Codons	APPRIS	SIFT
22_3166910_C/A	22:3166910-3166910	A	splice_acceptor_variant	lonp1	ENSDARG00000102765	ENSDART00000158009.2	protein_coding	-	-	-	-	-	P1	-
22_3166910_C/A	22:3166910-3166910	A	intron_variant	lonp1	ENSDARG00000102765	ENSDART00000167550.2	protein_coding	-	-	-	-	-	-	-
22_3169475_G/T	22:3169475-3169475	T	stop_gained	lonp1	ENSDARG00000102765	ENSDART00000158009.2	protein_coding	1885	1674	558	Y/*	TAC/TAA	P1	-
22_3169475_G/T	22:3169475-3169475	T	stop_gained	lonp1	ENSDARG00000102765	ENSDART00000167550.2	protein_coding	405	405	135	Y/*	TAC/TAA	-	-
22_3169514_A/T	22:3169514-3169514	T	missense_variant	lonp1	ENSDARG00000102765	ENSDART00000158009.2	protein_coding	1846	1635	545	S/R	AGT/AGA	P1	0
22_3169514_A/T	22:3169514-3169514	T	missense_variant	lonp1	ENSDARG00000102765	ENSDART00000167550.2	protein_coding	366	366	122	S/R	AGT/AGA	-	0

Assembly Converter

- Assembly Converter allows converting coordinates from one assembly to another
- Also known as **LiftOver**
- e.g. used for converting coordinates found in old papers
- www.ensembl.org/Danio_rerio/Tools/AssemblyConverter
- Example:

```
22 3144711 3144711 sa39354
22 3145013 3145013 sa43743
(Chr, Start, End, Name)
```
- **BED format:** www.ensembl.org/info/website/upload/bed.html
- (Only first three fields are essential)

Assembly Converter

Assembly Converter ?

New job Clear form

This online tool currently uses [CrossMap](#), which supports a limited number of formats (see our online documentation for [details of the individual data formats](#) listed below). CrossMap also discards metadata in files, so track definitions, etc, will be lost on conversion.

Species:

Assembly mapping:

Name for this job (optional):

Input file format:

Either paste data:

```
22 3144711 3144711 sa39354  
22 3145013 3145013 sa43743
```

Or upload file: No file chosen

Or provide file URL:

Run >

Assembly Converter

Assembly Converter ?

New job Clear form

This online tool currently uses metadata in files, so track de

Species:
22 3144711 3144711 sa39354

Assembly mapping:
22 3145013 3145013 sa43743

Name for this job (optional):

Input file format:

Either paste data:

22 3161984 3161984 sa39354

22 3162286 3162286 sa43743

CrossMap also discards

Or upload file: No file chosen

Or provide file URL:

Run >

UCSC In-Silico PCR

- Fast search for possible products from a pair of **PCR** primers
- genome.ucsc.edu/cgi-bin/hgPcr

UCSC In-Silico PCR

- Fast search
- [genome.ucsc.edu](http://genome.ucsc.edu/cgi-bin/hgPCR/PCR.cgi)

The screenshot shows the UCSC In-Silico PCR search interface. At the top, there's a navigation bar with links to Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, and About Us. Below the navigation bar is the main search form titled "UCSC In-Silico PCR". The form includes dropdown menus for "Genome" (set to Zebrafish) and "Assembly" (set to May 2017 (GRCz11/danRer11)), and text input fields for "Forward Primer" (CCGGGGAGCAGTTGA) and "Reverse Primer" (TGGGTGGAGTAGGTCTG). There are also input fields for "Max Product Size" (4000), "Min Perfect Match" (15), "Min Good Match" (15), and a checkbox for "Flip Reverse Primer". Below the search form is a section titled "About In-Silico PCR" which provides a brief description of the service and a link to a YouTube video. The "Configuration Options" section details the parameters: Genome and Assembly, Target, Forward Primer, Reverse Primer, Max Product Size, Min Perfect Match, Min Good Match, and Flip Reverse Primer. The "Output" section explains the format of the search results, mentioning FASTA output with capitalized headers where primers match. A sample sequence is shown at the bottom.

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

UCSC In-Silico PCR

Genome: Assembly: Forward Primer: Reverse Primer:

Zebrafish May 2017 (GRCz11/danRer11) CCGGGGAGCAGTTGA TGGGTGGAGTAGGTCTG submit

Max Product Size: 4000 Min Perfect Match: 15 Min Good Match: 15 Flip Reverse Primer:

About In-Silico PCR

In-Silico PCR searches a sequence database with a pair of PCR primers, using an indexing strategy for fast performance. See an example [video](#) on our YouTube channel.

Configuration Options

Genome and Assembly - The sequence database to search.
Target - If available, choose to query transcribed sequences.
Forward Primer - Must be at least 15 bases in length.
Reverse Primer - On the opposite strand from the forward primer. Minimum length of 15 bases.
Max Product Size - Maximum size of amplified region.
Min Perfect Match - Number of bases that match exactly on 3' end of primers. Minimum match size is 15.
Min Good Match - Number of bases on 3' end of primers where at least 2 out of 3 bases match.
Flip Reverse Primer - Invert the sequence order of the reverse primer and complement it.

Output

When successful, the search returns a sequence output file in fasta format containing all sequence in the database that lie between and include the primer pair. The fasta header describes the region in the database and the primers. The fasta body is capitalized in areas where the primer sequence matches the database sequence and in lower-case elsewhere. Here is an example from human:

```
>chr22:31000551+31001000 TAACAGATTGATGATGCATGAAATGGG CCCATGAGTGGCTCTAAAGCAGCTGC  
TtACAGATTGATGATGCATGAAATGGGgggtggccagggggtgggggtga
```

UCSC In-Silico PCR

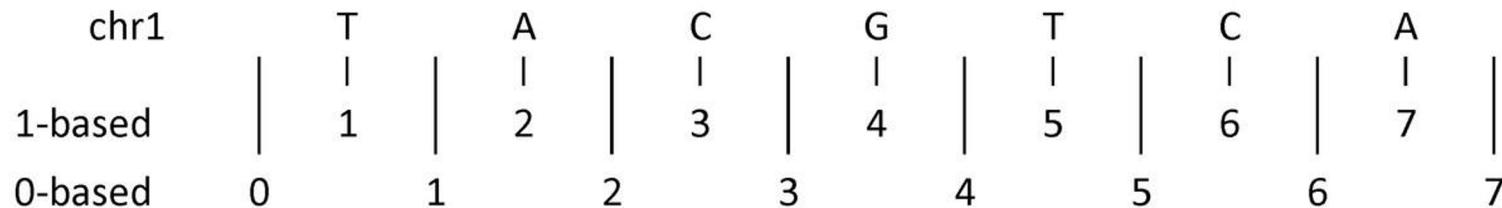
- Fast search for primers
- genome.ucsc.edu

The screenshot shows the UCSC In-Silico PCR interface with a blue header bar containing links for Genomes, Genome Browser, Tools, Mirrors, Downloads, and My Data. Below the header, the title "UCSC In-Silico PCR" is displayed in a light blue box. The main content area contains two sequence entries:

```
>chr14_KZ115440v1_alt:182433-183230 798bp CCCGGGGAGCAGTTGAT CGTTGGTGGAGTAGGTCTG  
CCCGGGGAGCAGTTGATcaccttgcggaggtaagcactaaatccctct  
tgattaaatttgcacatgtgcgtttcataacttagatttgcaggatgttcatg  
tgcataaaaatgtgcacatttaataaataatgtttatgataaaat  
gttatattttatgcgtccttcattttatataatgtttatgcgttgc  
gttatttgcataatgtgtataatataatattgcacatgtgtttatgc  
tatcatatataatgtgcacatgtttatgcgttgc  
ctaaggaaatgaatgtttatgcgttgc  
cttttcagttttatgcgtttatgcgttgc  
tgaaaaaaggatcaggatgtgc  
tgcatatgcgttgc  
cgccaaatgtgtgcgtgtgtgtatgtgtgttttgtgggggt  
cccttgcgtccatcagcacttgcgttgc  
cagatgtgcacaggcaggcccccaataatcacatgc  
atttatatgcataatataatgcgttgc  
accacttcgtttgcgtttgc  
aatttccatggagaacccatgc  
aatttccatggagaacccatgc  
>chr14:21334984-21335781 798bp CCCGGGGAGCAGTTGAT CGTTGGTGGAGTAGGTCTG  
CCCGGGGAGCAGTTGATcaccttgcggaggtaagcactaaatccctct  
tgattaaatttgcacatgtgcgtttcataacttagatttgcaggatgttcatg  
tgcataaaaatgtgcacatttaataaataatgtttatgataaaat  
gttatatttgcgtccttcattttatataatgtttatgcgttgc  
gttatttgcataatgtgtataatataatattgcacatgtgtttatgc  
tatcatatataatgtgcacatgtttatgcgttgc  
ctaaggaaatgaatgtttatgcgttgc  
cttttcagttttatgcgtttatgcgttgc  
tgaaaaaaggatcaggatgtgc  
tgcatatgcgttgc  
cgccaaatgtgtgcgtgtgtatgtgtgttttgtgggggt  
cccttgcgtccatcagcacttgcgttgc  
cagatgtgcacaggcaggcccccaataatcacatgc  
atttatatgcataatataatgcgttgc  
accacttcgtttgcgtttgc  
aatttccatggagaacccatgc
```

UCSC & Ensembl Differences

- **Ensembl:** 1
UCSC: chr1
- **Ensembl:** 1-based coordinates (bases numbered)
UCSC: 0-based coordinates (numbers between bases)



- The **G** is **1:4-4** in Ensembl coordinates but **1:3-4** in UCSC

Custom Tracks

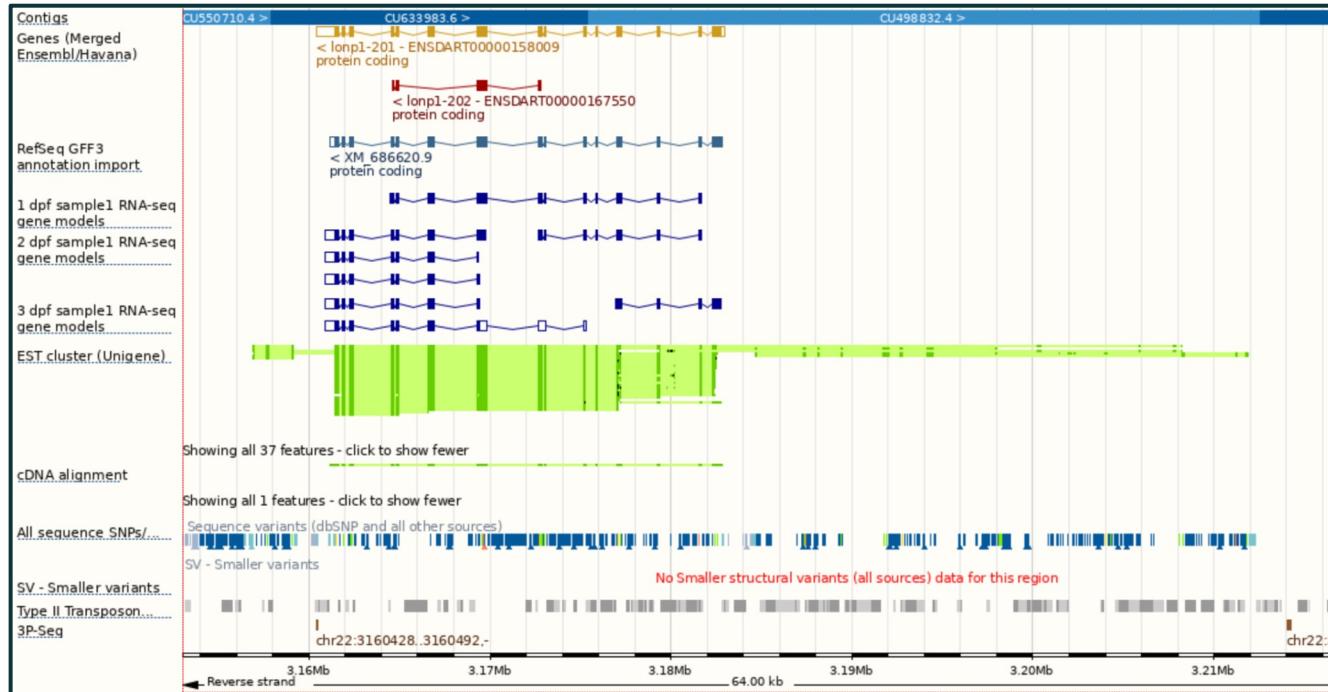
- Click “Custom tracks” and add <https://mbl2023.buschlab.org/data/3p-seq.bed>

The screenshot shows the Ensembl genome browser interface for Zebrafish (Danio rerio). The left sidebar shows various genomic tracks like Whole genome, Chromosome, Region, and UCSC. The main panel is titled "Custom tracks" and contains fields for "Name for this data (optional)" (3P-Seq), "Species" (Zebrafish (Danio rerio)), "Assembly" (GRCz11), "Data" (URL: <https://mbl2022.buschlab.org/data/3p-seq.bed>), "Data format" (BED), and a "Choose file" button. A note at the top states: "Please note that track hubs and indexed files (BAM, BigBed, etc) do not work with certain cloud services, including Google Drive and Dropbox. Please see our [support page](#) for more information." On the right, a chromosome browser track for chromosome 1 is visible, showing genomic data across the 6,200 to 7,600 region.

- 24 hpf 3P-Seq data from Bartel lab

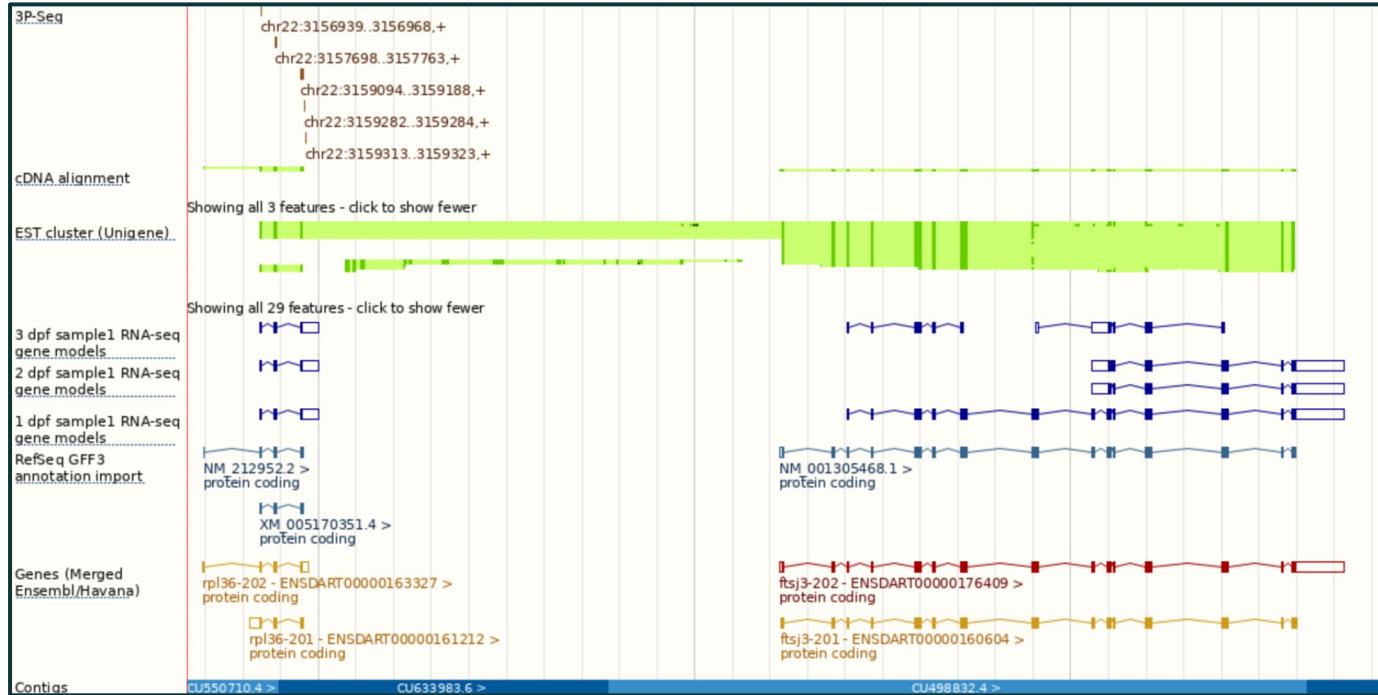
Custom Tracks

- Go to "22:3153000-3217000" (reverse strand)



Custom Tracks

- Go to "22:3153000-3217000" (forward strand)

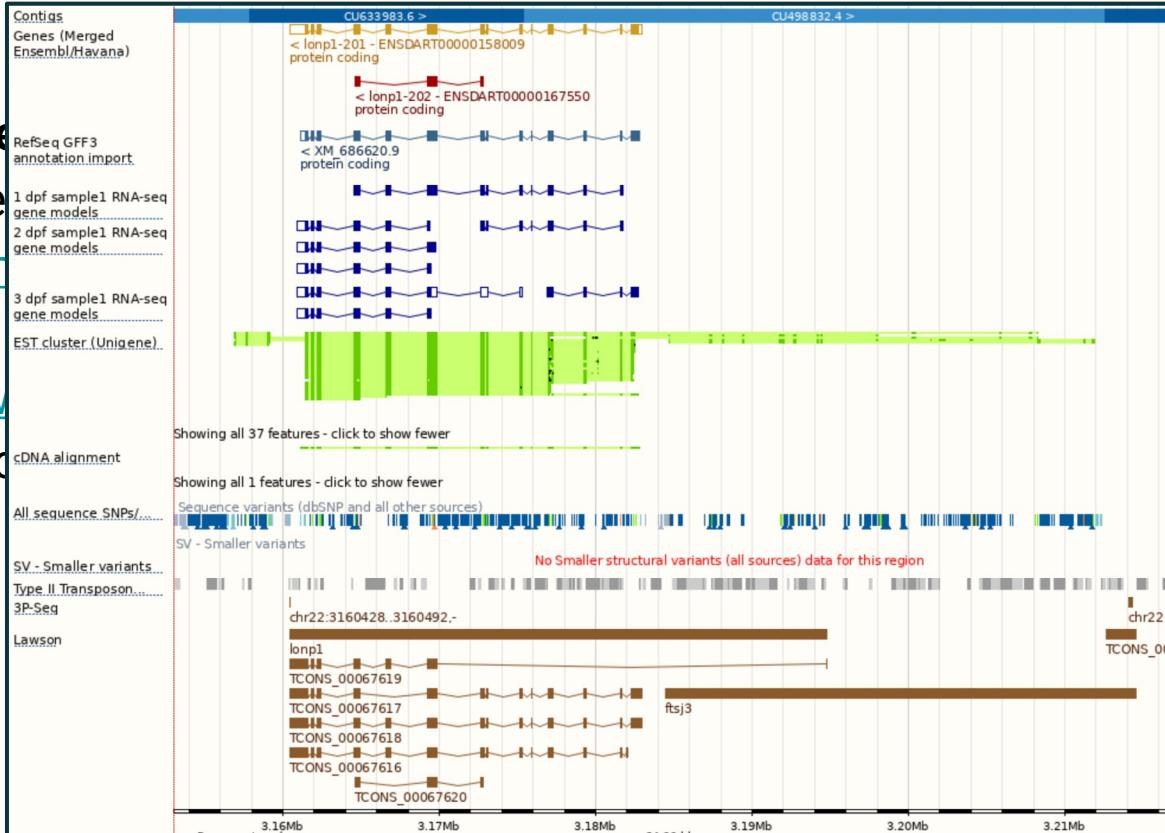


Custom Tracks - Lawson Lab Annotation

- Lawson et al. (2020) “**An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes**”
eLife 9:e55792
- www.umassmed.edu/lawson-lab/reagents/zebrafish-transcriptome/
- Add:
<https://www.umassmed.edu/globalassets/lawson-lab/downloadfiles/v4.3.2.gtf>
- Large, so Ensembl will be slow - disable or delete when done

Custom Tracks - Lawson Lab Annotation

- Lawson sensitive eLife 9: www.umass.edu
- Add: <https://www.ncbi.nlm.nih.gov/gene/1000000000000000000>
- Large, so



Annotation for genes"
ome/
es/v4.3.2.gtf

Exercise 3

- Do Exercise 3 - “exploring data”
- Covers:
 - BioMart
 - Making BED files
 - Finding candidate genes
 - Finding orthologues
- Go to mbl2023.buschlab.org

Thank You!

Any questions?

