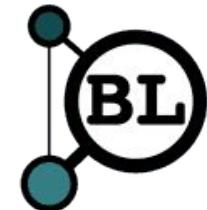


# Genome Literacy Workshop

Elisabeth Busch-Nentwich  
& Ian Sealy



# 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

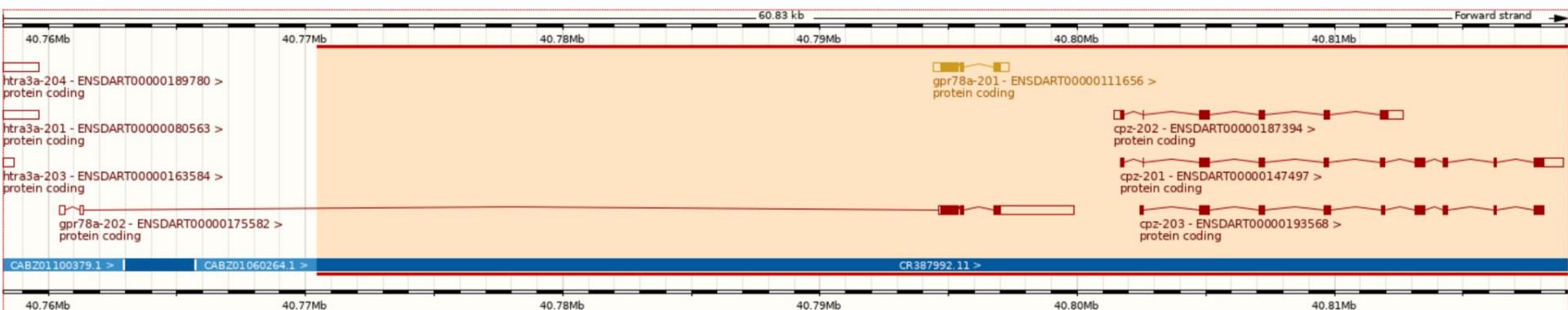
# Ensembl

- 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 **107** (July 12th)
- 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 'BLAST/BLAT', 'VEP', 'Tools', and 'More'. A search bar says 'Search all species...'. Below the navigation, there are sections for 'Tools', 'BioMart >', 'BLAST/BLAT >', and 'Variant Effect Predictor >'. The 'BLAST/BLAT >' section has a link to 'All tools' and a button to 'Export custom datasets from Ensembl with this data-mining tool'. It also includes a search bar for genomes by DNA or protein sequence and a description of the Variant Effect Predictor. To the right, there's a sidebar with information about Ensembl, its tools (BLAST, BLAT, BioMart), and the current release (107). The main content area features a 'Search' section with a dropdown for 'All species' and a 'Go' button. Below it, there's a news section with a list of updates, including one about the Red Jungle Fowl genome. Further down, there are sections for 'All genomes' (with a dropdown to 'Select a species') and 'Favourite genomes' (listing Human, Pig breeds, Mouse, and Zebrafish). The Zebrafish entry is circled in red at the bottom. On the far right, there's a 'Ensembl Rapid Release' section and a link to 'More release news'.

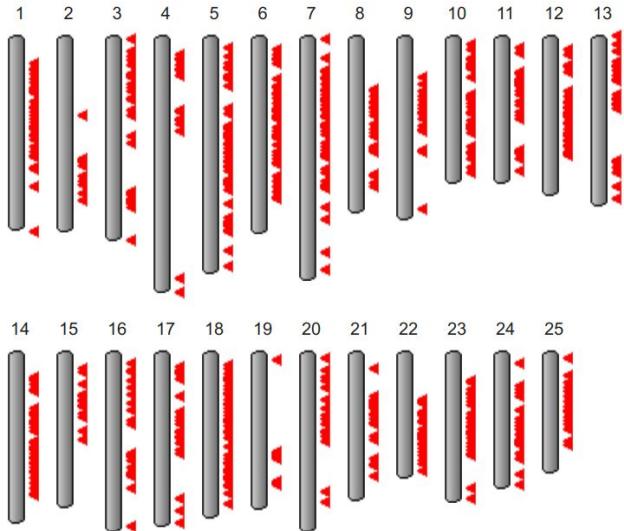
# Zebrafish Genome

- **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](http://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. A search bar is present, along with a 'Login/Register' button. The main content area has a banner for 'Zebrafish (GRCz10)'. Below this, there's a search box for 'Search Zebrafish (Danio rerio)' with dropdowns for 'Search all categories' and 'Search Zebrafish...'. A 'Go' button is next to the search bar. To the right, a yellow box highlights 'What's New in Zebrafish release 91', listing 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'. It also features a 'View karyotype' tool and a 'Example region' for the Zv9 assembly. On the right side, there's a 'Gene annotation' section with a 'Pax6 INS FOXP2 BRCA2 DMD ssh Example gene' diagram and a 'Example transcript' diagram showing a gene structure with exons and a poly-A tail.

# Ensembl Mirrors

- Mirrors: [www.ensembl.org/info/about/mirrors.html](http://www.ensembl.org/info/about/mirrors.html)
- Main site (UK): [www.ensembl.org](http://www.ensembl.org)
- US East mirror: [useast.ensembl.org](http://useast.ensembl.org)
- US West mirror: [uswest.ensembl.org](http://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 circled in red: [Alliance](#) and [Ensembl\(GRCz11\)](#).

GENE	
<b>dmd</b>	
ID	ZDB-GENE-010426-1
Name	<i>dystrophin</i>
Symbol	<i>dmd</i> <a href="#">Nomenclature History</a>
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	<a href="#">protein_coding_gene</a>
Location	Chr: <a href="#">Mapping Details/Browsers</a>
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	<a href="#">Alliance</a> (1), <a href="#">Gene:83773</a> (1),
Resources	<a href="#">VECA</a> (1), <a href="#">DARG000000031909</a> (1), <a href="#">Ensembl(GRCz11)</a> : <a href="#">ENSDARG00000008487</a> (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, name, and a detailed description.

Category	Name	Description
Gene	dmd (Zebrafish Gene) ENSDARG0000008487 1:10824351-11075405-1	Dystrophin [Source:ZFIN;Acc:ZDB-GENE-010426-1]
Transcript	dmd-201 (ZFIN transcript name record; description: dystrophin.)	is an external reference matched to Transcript ENSDART0000007013
GeneTree	dmd-213 (ZFIN transcript name record; description: dystrophin.)	is an external reference matched to Transcript ENSDART00000164141
GenomicAlignment	dmd-211 (Zebrafish Transcript) ENSDART0000148305 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

Other search parameters visible on the left include 'Current selection: Only searching Zebrafish', 'Restrict category to: Gene, Transcript, GeneTree, GenomicAlignment', 'Per page: 10, 25, 50, 100', 'Layout: Standard, Table', and a 'Tip' section.

# 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 BLAST/BLAT search results for the gene **dmd**. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A red circle highlights the "BLAST/BLAT" link in the top menu. The main content area displays "Results for dmd".  
**Job details:**

- Job name: dmd
- Species: Zebrafish (Danio rerio)
- Assembly: GRCz11
- Search type: BLASTN (NCBI Blast)

**Results table:**

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	1779	0.0	100.000 [Alignment]
1:11031622-11032521 [Sequence]	dmd	Reverse	1	900	900	1779	0.0	100.000 [Alignment]
1:11031622-11032521 [Sequence]	dmd	Reverse	1	387	390	686	0.0	96.923 [Alignment]
1:11624334-1624715 [Sequence]	clic6	Reverse	1	387	390	686	0.0	96.923 [Alignment]
1:11349264-11349645 [Sequence]	sdk1b	Reverse	1	387	390	686	0.0	96.923 [Alignment]
1:45265987-45266368 [Sequence]	EVISL	Reverse	1	387	390	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 ( <a href="#">ENSOURLG00000020638</a> )	84.52 %	88.22 %	0	<span style="background-color: green; color: white;">95.81</span>	Yes
		<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (2:208,119-221,155:1) <a href="#">View Sequence Alignments</a>					
Lumpfish ( <i>Cyclopterus lumpus</i> )	1-to-many	dmd ( <a href="#">ENSLCMLG000005009931</a> )	82.21 %	82.49 %	0	<span style="background-color: green; color: white;">95.20</span>	Yes
		<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (2:5,248,684-5,281,983-1) <a href="#">View Sequence Alignments</a>					
Lyretail cichlid ( <i>Neolamprologus brichardi</i> )	1-to-1	dmd ( <a href="#">ENSNBRG00000015200</a> )	87.34 %	89.39 %	0	<span style="background-color: green; color: white;">96.66</span>	Yes
		<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (JH422367.1:2,004,027-2,028,054:-1) <a href="#">View Sequence Alignments</a>					
Makobe Island cichlid ( <i>Pundamilia nyererei</i> )	1-to-1	dmd ( <a href="#">ENSPNYG00000022641</a> )	45.32 %	89.56 %	0	<span style="background-color: green; color: white;">96.75</span>	Yes
		<a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (JH419417.1:620,205-712,305:-1) <a href="#">View Sequence Alignments</a>					

# 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 **CABZ**01052570.1 (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. **ENS**DARG00000028213 (ttn.2 or tt<sub>n</sub>a)
- **ENS** = Ensembl

# Stable IDs

---

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

# Stable IDs

---

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

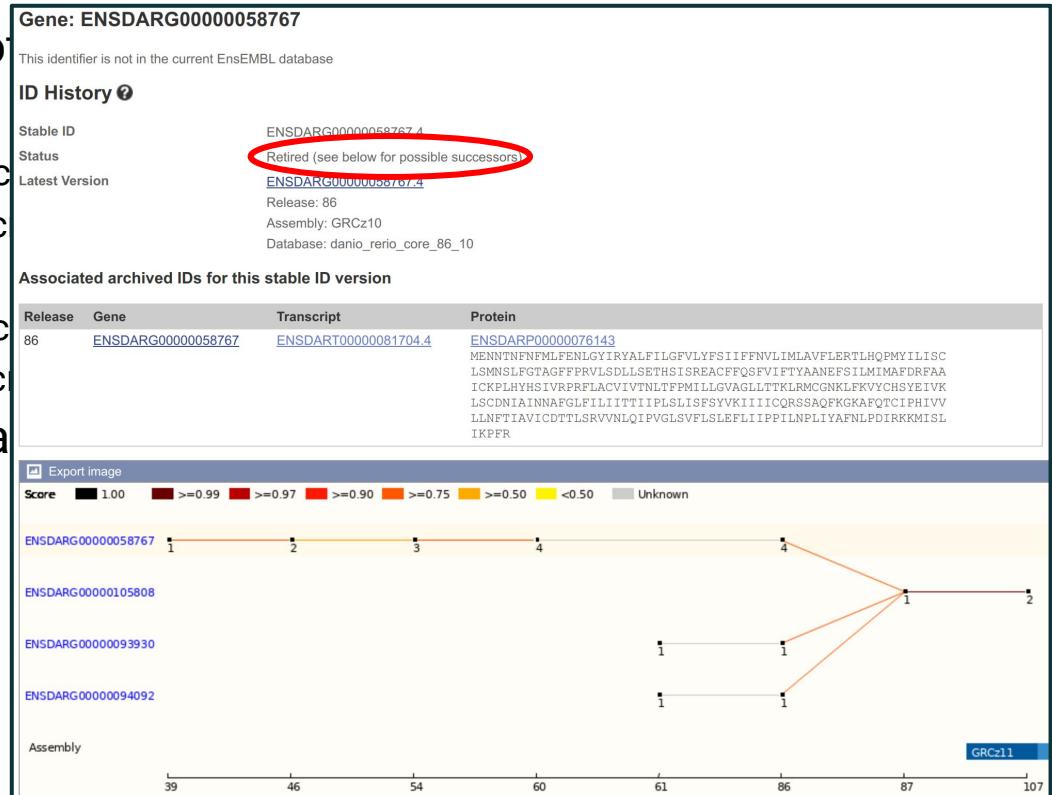
# Stable IDs

---

- 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
  - Version number of **ENSDART** increases
  - Version number of **ENSDARP** increases
  - Version number of **ENSDARE** increases
- Can also be **removed**, e.g. see below



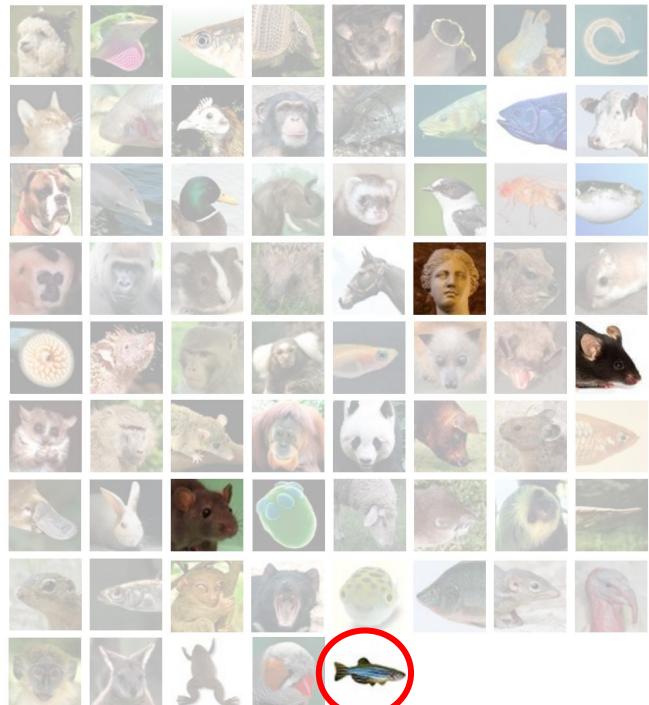
# Stable IDs

---

- 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](http://www.ensembl.org/Danio_rerio/Tools/IDMapper) to convert older IDs to what they **map** to currently in Ensembl

# Gene Annotation

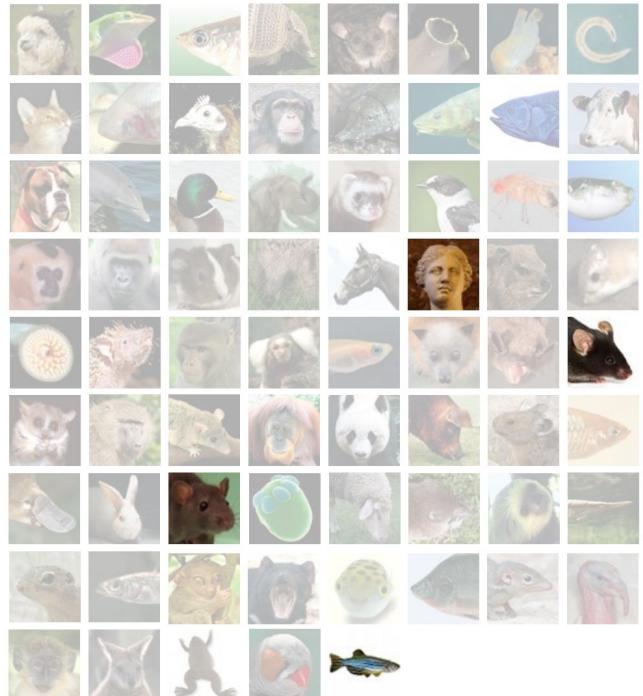
- Zebrafish (+ human, mouse, rat) has **manual** and **automatic** gene annotation
- Other **300+** genomes in Ensembl only have automatic annotation
- [www.ensembl.org/info/about/species.html](http://www.ensembl.org/info/about/species.html)



From Ensembl training materials, CC BY 4.0 license

# Manual Annotation

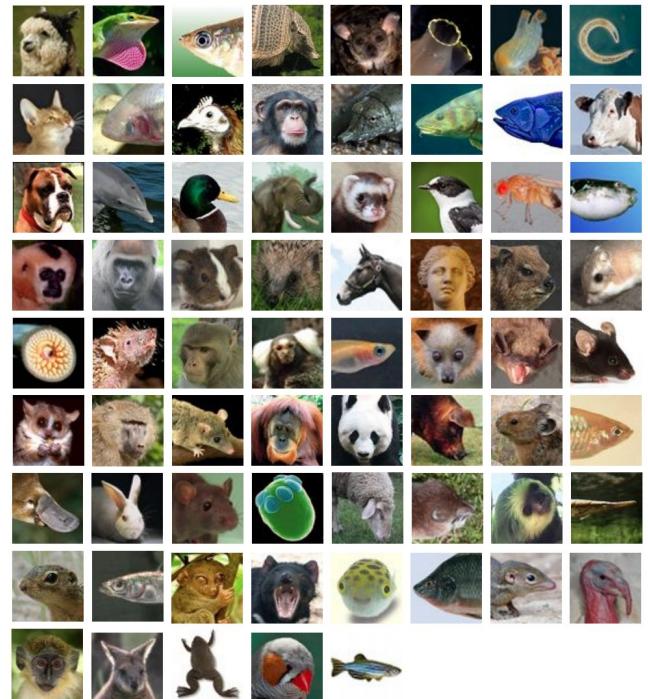
- **Gold** standard
  - Uses information from databases and publications
  - More accurate for tricky areas:
    - e.g. UTRs, splice sites, single exon transcripts
  - **Slower** and more expensive
  - Thorough, but leads to inclusion of transcripts that may not be representative (e.g. low expression)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



From Ensembl training materials, CC BY 4.0 license

# Automatic Annotation

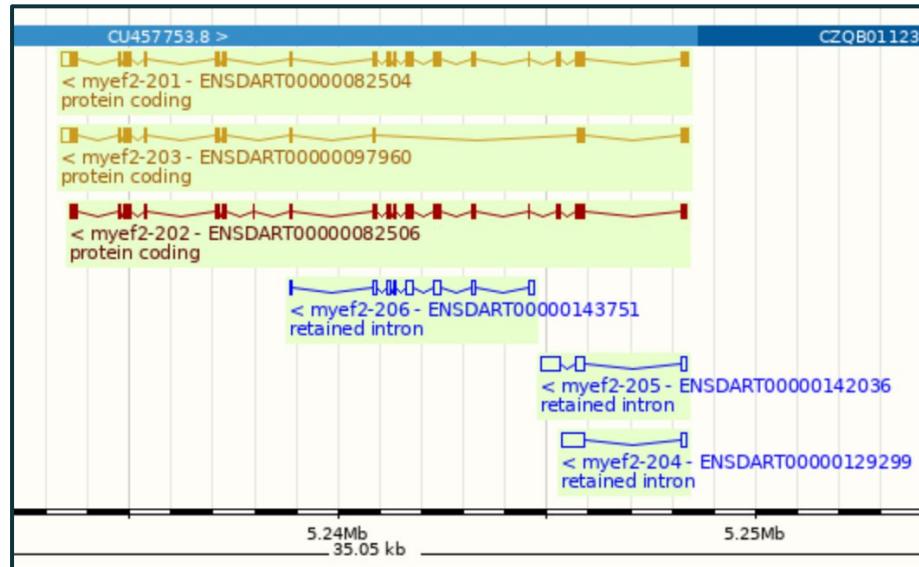
- **Faster**
- Uses evidence from sequences **deposited** in ENA/GenBank/DDBJ and UniProt proteins
- **Overview:**
  - Identify repeats and low complexity sequence with RepeatMasker, Dust and TRF
  - Run GENSCAN to identify *ab initio* gene predictions
  - Align UniProt proteins to GENSCAN predictions, 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

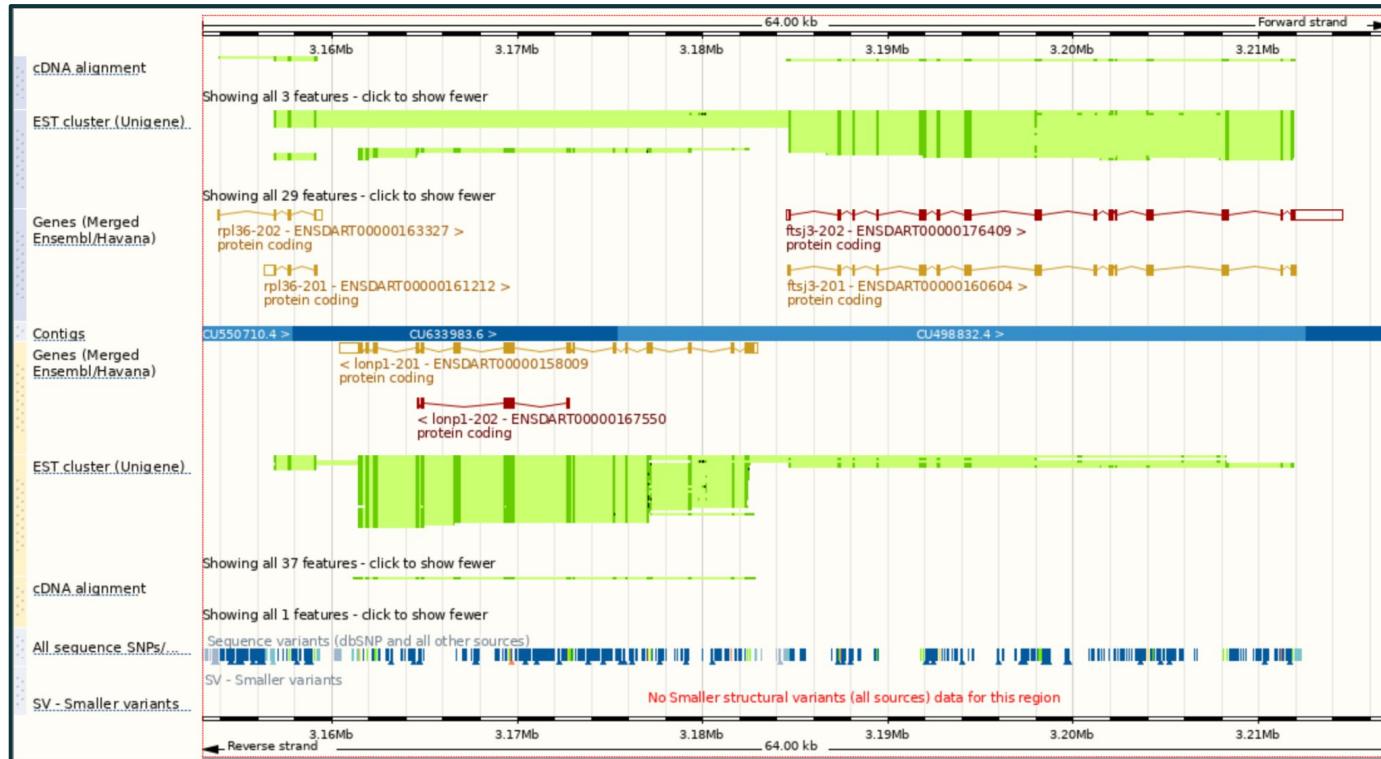
Description: BRISC and BRCA1 A complex member 1 [Source:NCBI gene;Acc:445296]  
Gene Synonyms: zgc:100909  
Location: Chromosome 11: 6,051,287-6,070,192 reverse strand.  
GRCz11:CM002895.2  
About this gene: This gene has 4 transcripts ([splice variants](#)) and [185 orthologues](#).  
Transcripts: [Hide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

Show/hide columns (1 hidden) Filter 

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

# "Region in detail" Demo

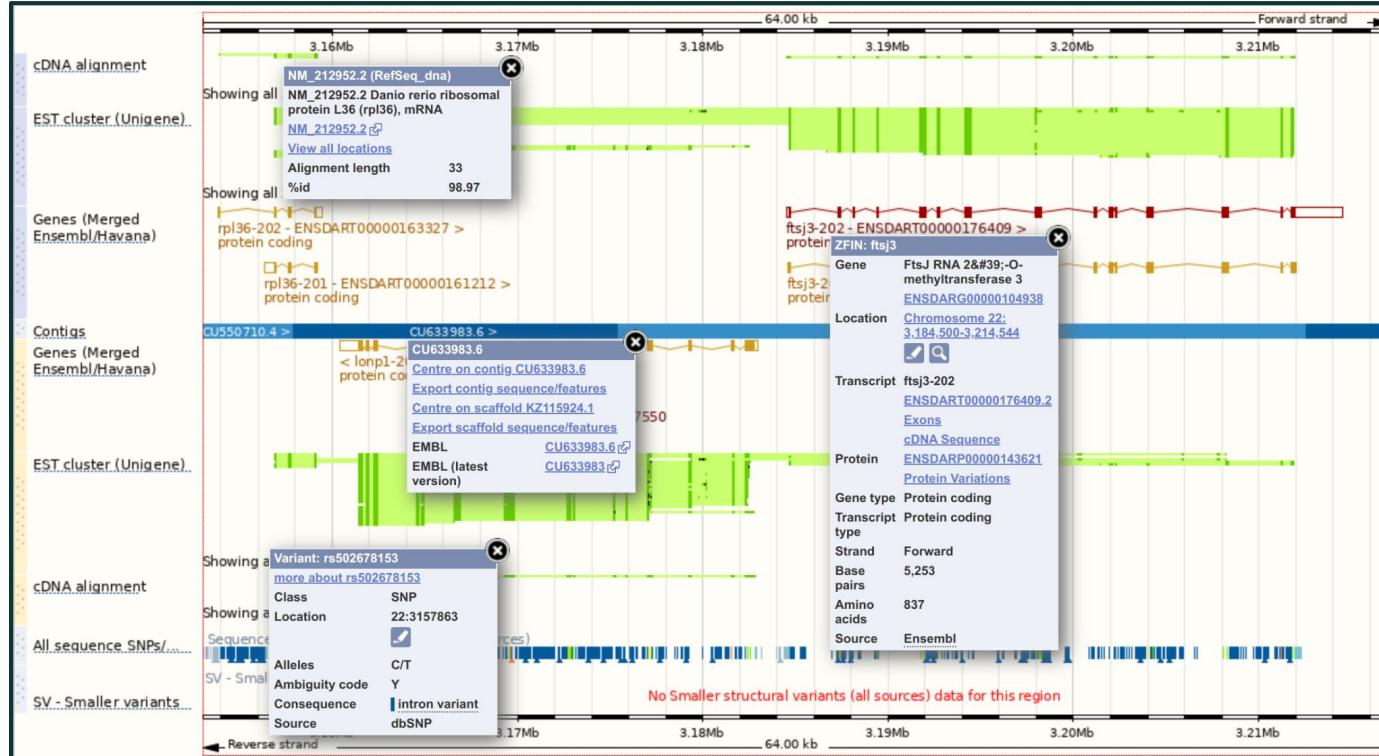
- Go to "22:3153000-3217000"



- 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 [mbl2022.buschlab.org](http://mbl2022.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 top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. On the right, there are "Search Zebrafish..." and "Login/Register" buttons.

The left sidebar contains a tree view of genomic tracks and configurations:

- Location: Whole genome, Chromosome 22, Region 3153000-3217000
- Region: Comparative, Synteny, Alignments, Predictions, RNASeq models (selected), Variations, Resequence, Linkage, Markers, Other genes, UCSC, NCBI, Comparative genomics, Oligo probes, Repeat regions, Information and decorations, Display options.
- Active tracks: Favourite tracks, Track order, Search results, Genome Reference Consortium Issues (0/7), Sequence and assembly (2/7), Sequence (2/4), Simple features (0/3), Genes and transcripts (7/96), Genes (2/2), Prediction transcripts (0/1), RNASeq models (5/93), mRNA and protein alignments (2/4), mRNA alignments (2/3), Protein alignments (0/1), Variation (2/8), Sequence variants (1/2), Failed variants (0/1), Phenotype annotations (0/2), Structural variants (1/3), Comparative genomics (1/68), Multiple alignments (0/2), Conservation regions (1/2), BLASTz/LASTz alignments (0/64), Oligo probes (0/41), Repeat regions (1/23), Information and decorations (12/14), Display options.
- Search for track hubs.

The main content area shows a "Genes and transcripts" section with a "Show tutorial" button. Below it is an "RNASeq models" section with a "Filter by" dropdown set to "All classes" and an "Enter terms to filter by" input field. A legend key indicates:  
Key: Shown (dark blue square), Hidden (light blue square), No Data (grey square).  
Filtered: Shown (green square), Hidden (light green square).

A "Default style:" section allows enabling or disabling various styles, with "ENAs" selected. A table titled "ENAs" lists samples and their expression levels across different time points (1 dpf, 14 dpf, 2 dpf, 3 dpf, 5 dpf) and samples (sample1). The table includes columns for 1 dpf sample1, 14 dpf sample1, 2 dpf sample1, 3 dpf sample1, and 5 dpf sample1, with values ranging from 0 to 1.

The bottom of the screen shows a genomic track visualization for the cDNA\_alignment track, spanning from 3.16 Mb to 3.21 Mb. The forward strand is indicated by an arrow pointing to the right.

# 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

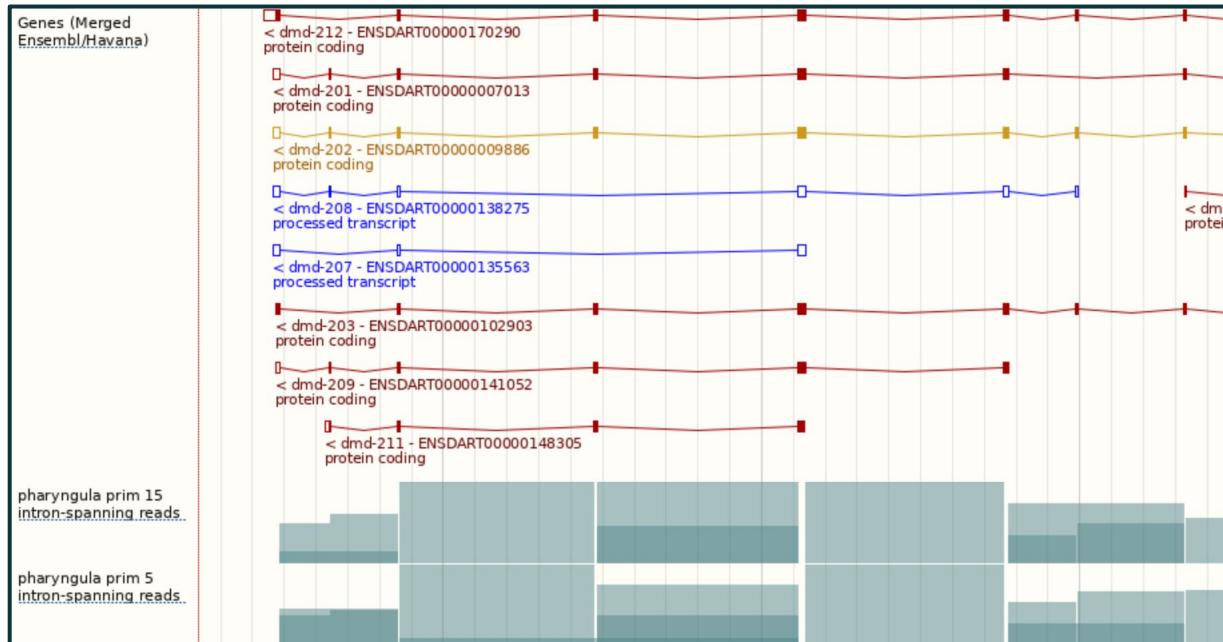
# "Configure this page" Demo

- 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: <a href="#">ZDB-GENE-030131-4006</a> ]
Gene Synonyms	fc64d11, prss15, wu:fc64d11
Location	<a href="#">Chromosome 22: 3,160,447-3,182,965</a> reverse strand. GRCz11:CM002906.2
About this gene	This gene has 2 transcripts ( <a href="#">splice variants</a> ), <a href="#">190 orthologues</a> and <a href="#">1 parologue</a> .
Transcripts	<a href="#">Show transcript table</a>

**Summary**

Name	<a href="#">lonp1</a> (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 <a href="#">article</a> .

# "Gene" Demo - Transcript Table

- Go to ENSDARG00000102765 and click on “Show transcript table”

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
  - 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	
<a href="#">ENSDART00000158009.2</a>	Ionp1-201	4114	<a href="#">966aa</a>	Protein coding	<a href="#">A0A0R4IH79</a>	<a href="#">Ensembl Canonical</a>	<a href="#">APPRIS P1</a>
<a href="#">ENSDART00000167550.2</a>	Ionp1-202	741	<a href="#">247aa</a>	Protein coding	<a href="#">A0A0R4IPW4</a>	<a href="#">CDS 5' and 3' incomplete</a>	

### Summary [?](#)

Name [Ionp1](#) (ZFIN)

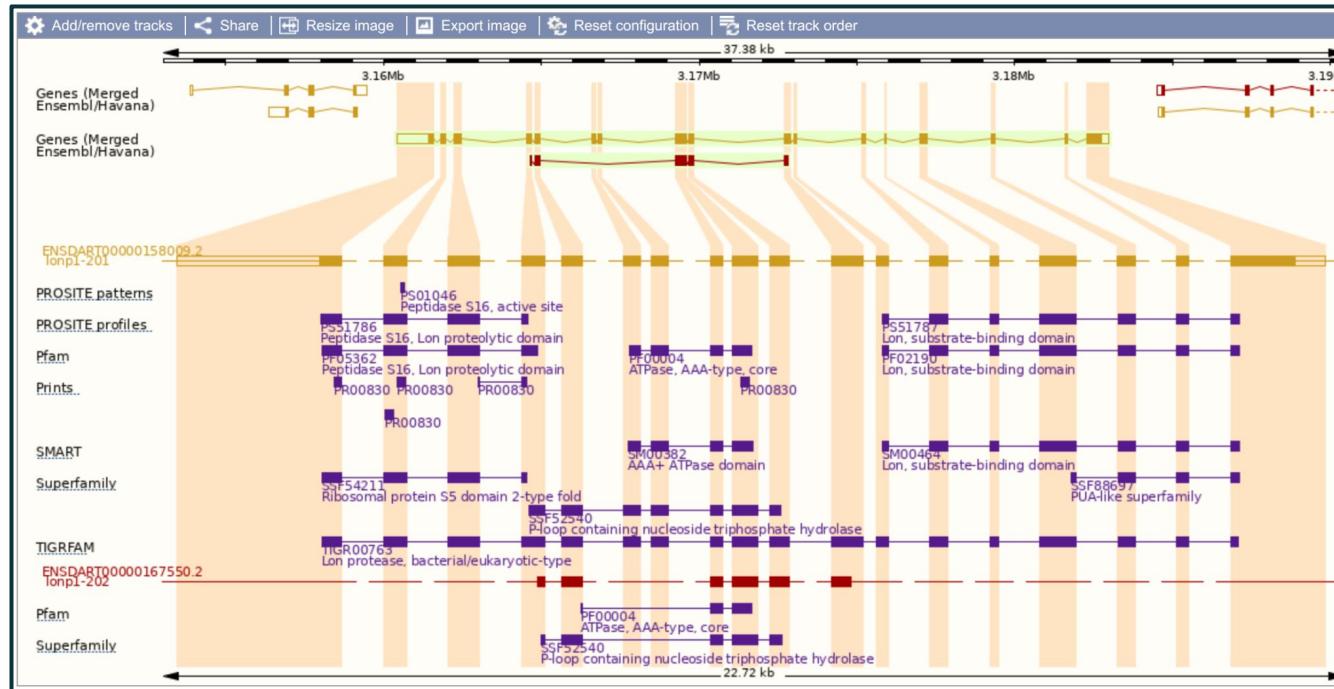
Ensembl version ENSDARG00000102765.2

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 <a href="#">View Gene Tree</a>	LONP1 ( <a href="#">ENSCABG0000010924</a> ) <a href="#">Compare Regions</a> (PKMU01001122.1:170,198-221,187:1) <a href="#">View Sequence Alignments</a>	75.03 %	75.26 %	25	n/a	No
African ostrich ( <i>Struthio camelus australis</i> )	1-to-1 <a href="#">View Gene Tree</a>	LONP1 ( <a href="#">ENSSCUG0000004632</a> ) <a href="#">Compare Regions</a> (KL206174.1:174,870-201,335:1) <a href="#">View Sequence Alignments</a>	80.69 %	70.50 %	50	n/a	Yes
Algerian mouse ( <i>Mus spretus</i> )	1-to-1 <a href="#">View Gene Tree</a>	Lonp1 ( <a href="#">MGP_SPRETEIJ_G0022694</a> ) <a href="#">Compare Regions</a> (17:54,555,161-54,567,779:-1) <a href="#">View Sequence Alignments</a>	75.05 %	74.12 %	0	n/a	No
Alpine marmot ( <i>Marmota marmota marmota</i> )	1-to-1 <a href="#">View Gene Tree</a>	LONP1 ( <a href="#">ENSMMMG00000018859</a> ) <a href="#">Compare Regions</a> (CZRN01000089.1:3,499,006-3,520,539:-1) <a href="#">View Sequence Alignments</a>	62.80 %	62.22 %	0	n/a	No
Amazon Molly ( <i>Poecilia formosa</i> )	1-to-1 <a href="#">View Gene Tree</a>	lonp1 ( <a href="#">ENSPFOG0000001826</a> ) <a href="#">Compare Regions</a> (KI520250.1:178,559-209,184:-1) <a href="#">View Sequence Alignments</a>	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
<a href="#">ENSDART00000158009.2</a>	lonp1-201	4114	<a href="#">966aa</a>	Protein coding	<a href="#">A0A0R4IH79</a>	Ensembl Canonical APPRIS P1
<a href="#">ENSDART00000167550.2</a>	lonp1-202	741	<a href="#">247aa</a>	Protein coding	<a href="#">A0A0R4IPW4</a>	CDS 5' and 3' incomplete

**Paralogues** [?](#)

[Download paralogues](#)

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

# "Gene" Demo - GO Terms

- Go to ENSDARG00000102765 and click on “GO: Molecular function”

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

# "Gene" Demo - External References

- Go to ENSDARG00000102765 and click on “External references”

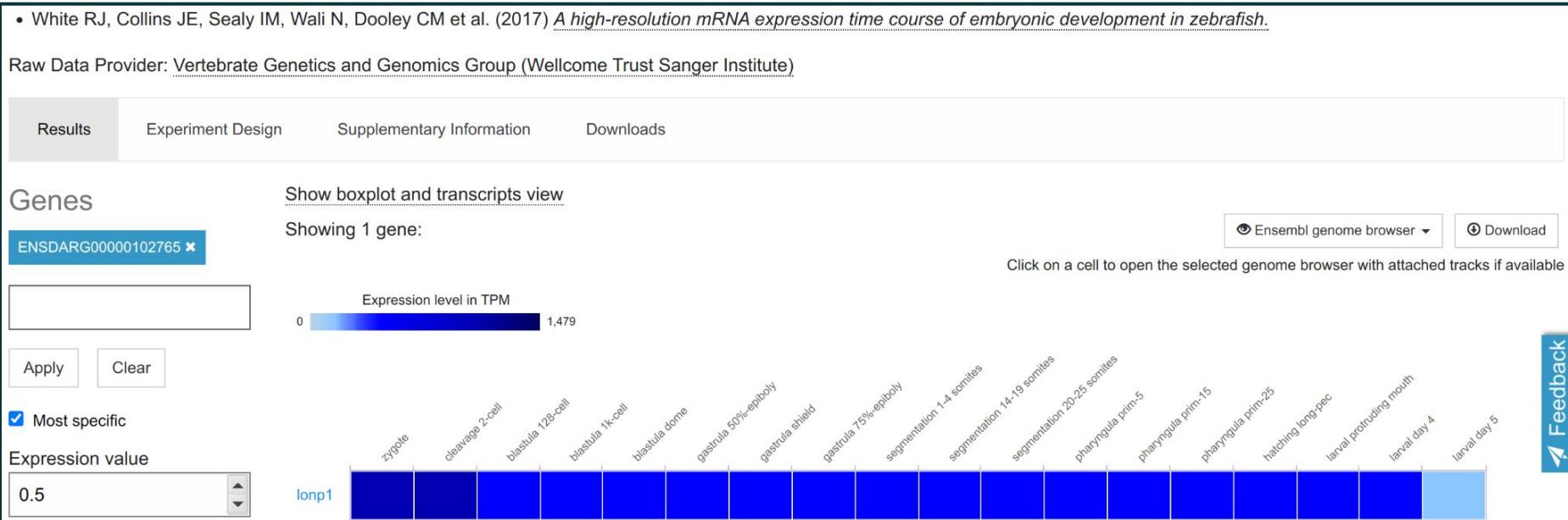
## External references

This gene corresponds to the following database identifiers:

External database	Database identifier	Filter 
Expression Atlas	<a href="#">ENSDARG00000102765</a>  [view all locations]	
NCBI gene (formerly Entrezgene)	<a href="#">lomp1</a>  Ion peptidase 1, mitochondrial <a href="#">[view all locations]</a>	
WikiGene	<a href="#">lomp1</a>  Ion peptidase 1, mitochondrial <a href="#">[view all locations]</a>	
ZFIN	<a href="#">lomp1</a>  Ion peptidase 1, mitochondrial <a href="#">[view all locations]</a>	

# "Gene" Demo - Expression Atlas

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



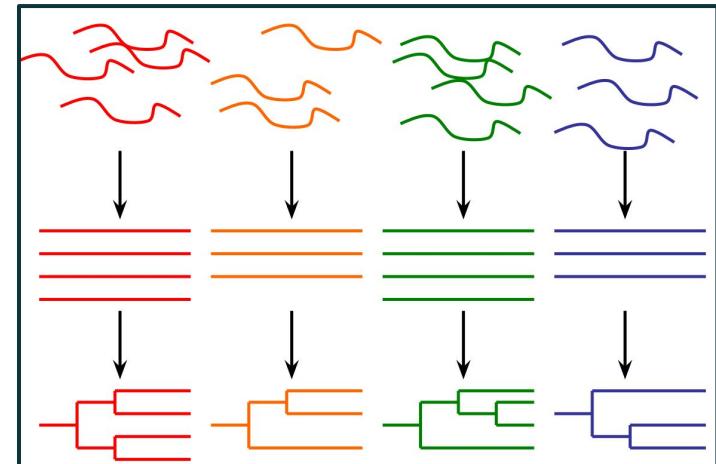
# Compara

---

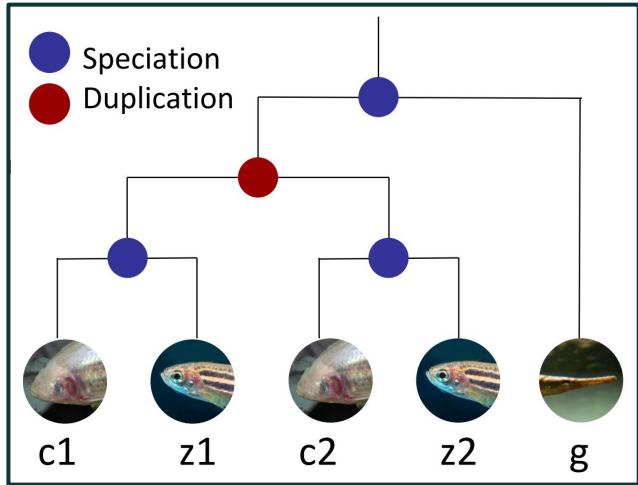
- Compara - produce Ensembl's comparative genomics resources
- Two types of analysis:
  - Gene level comparisons to produce **gene trees**, e.g. infer **homologues** (orthologues & paralogues)
  - **Whole genome alignments** - pairwise and multiple alignments, e.g. **constrained elements** and **synteny**

# 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**

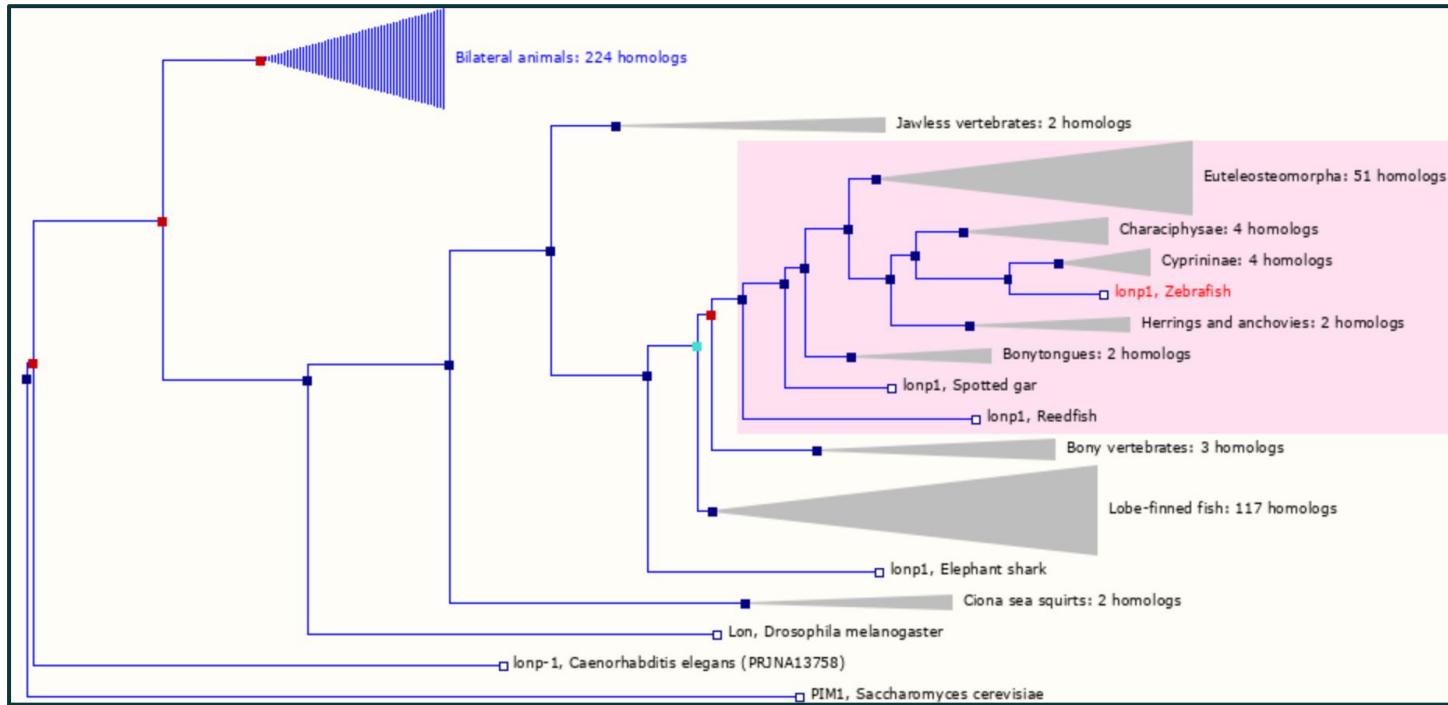
**z1 & c1** are **orthologues** (arose from **speciation**), as are **z2 & c2** + **z2 & g**, etc...

**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](http://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](http://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*)**

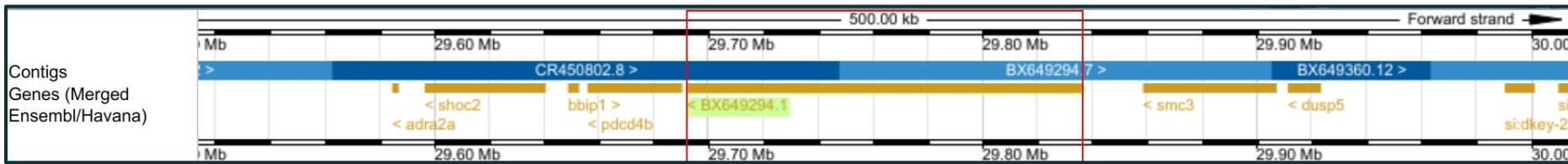
# Synteny Example

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

# Synteny Example

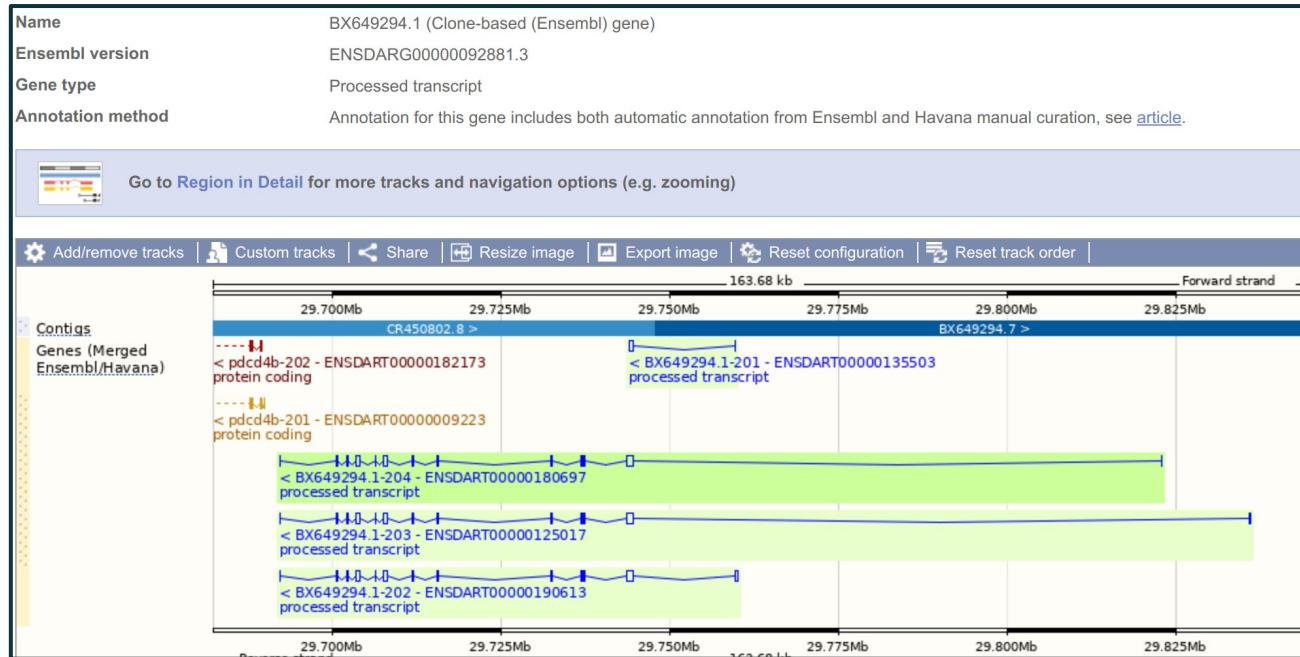
- If we look at the chr22 region in zebrafish then all the surrounding genes are the same and RBM20 is likely to be BX649294.1



Homo sapiens genes	Location	Danio rerio homologues	Location	Region Comparison
<a href="#">DUSP5</a> (ENSG00000138166)	<a href="#">10:110497907-110511533</a>	<a href="#">dusp5</a> (ENSDARG00000019307)	<a href="#">22:29911326-29922872</a>	<a href="#">Region Comparison</a>
<a href="#">SMC3</a> (ENSG00000108055)	<a href="#">10:110567684-110606048</a>	<a href="#">smc3</a> (ENSDARG00000019000)	<a href="#">22:29858535-29906764</a>	<a href="#">Region Comparison</a>
<a href="#">RBM20</a> (ENSG00000203867)	<a href="#">10:110644336-110839468</a>	No homologues		
<a href="#">PDCD4</a> (ENSG00000150593)	<a href="#">10:110871795-110900006</a>	<a href="#">pdc4b</a> (ENSDARG00000041022)	<a href="#">22:29655981-29689981</a>	<a href="#">Region Comparison</a>
<a href="#">BBIP1</a> (ENSG00000214413)	<a href="#">10:110898730-110919201</a>	<a href="#">bbip1</a> (ENSDARG00000071046)	<a href="#">22:29648854-29652356</a>	<a href="#">Region Comparison</a>
<a href="#">SHOC2</a> (ENSG00000108061)	<a href="#">10:110919367-111017307</a>	<a href="#">shoc2</a> (ENSDARG00000040853)	<a href="#">22:29596646-29640181</a>	<a href="#">Region Comparison</a>
<a href="#">ADRA2A</a> (ENSG00000150594)	<a href="#">10:111077029-111080907</a>	<a href="#">adra2a</a> (ENSDARG00000040841)	<a href="#">22:29584800-29586608</a>	<a href="#">Region Comparison</a>

# 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 [mbl2022.buschlab.org](http://mbl2022.buschlab.org)

# Part 3

---

- BioMart
- Other tools
- Custom tracks

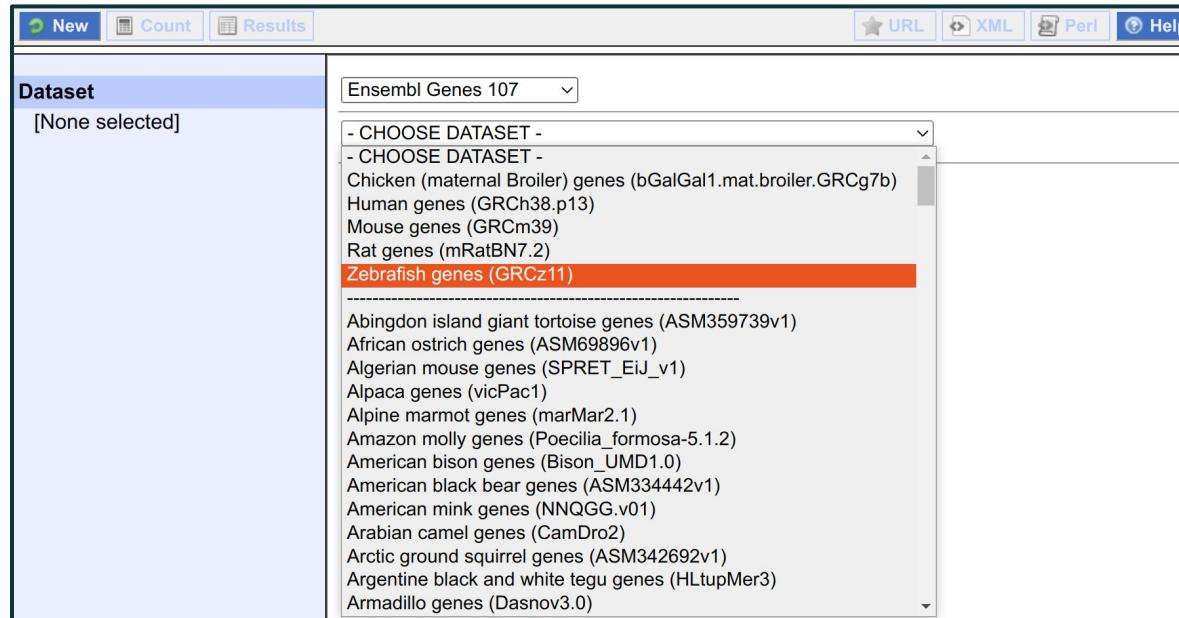
# BioMart

- **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. Below the header, there is a toolbar with buttons for "New", "Count", and "Results". To the right of the toolbar are links for "URL", "XML", "Perl", and "Help". At the bottom left, there is a "Dataset" section with a dropdown menu labeled "- CHOOSE DATABASE -". The main content area below the toolbar is currently empty.

# 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 the Zebrafish genes dataset (GRCz11). The left sidebar displays the dataset details and available filters. The main area shows various filtering options:

- Transcript count <=:** A dropdown menu shows options: polymorphic\_pseudogene, processed\_pseudogene, processed\_transcript, protein\_coding (which is selected), and pseudogene.
- Gene type:** A dropdown menu shows options: antisense, IG\_C\_gene, IG\_C\_pseudogene, IG\_J\_pseudogene, IG\_pseudogene, and IG\_V\_pseudogene.
- Source (gene):** A dropdown menu shows options: ensembl (selected).
- Source (transcript):** A dropdown menu shows options: ensembl (selected).
- APPRI annotation:** Radio buttons for "Only" (selected) and "Excluded".

# 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 navigation links: New, Count, Results, URL, XML, Perl, and Help. On the left, a sidebar displays the selected dataset: "Dataset 6 / 37241 Genes" (Zebrafish genes (GRCz11)). It includes sections for "Filters" (Chromosome/scaffold: 22, Start: 3000000, End: 4000000, Transcript count <= 1, Gene type: protein\_coding) and "Attributes" (Gene stable ID, Gene name, Source of gene name, APPRIS annotation, Chromosome/scaffold name, Gene start (bp), Gene end (bp), Strand). A "Dataset" section at the bottom indicates "[None Selected]". The main panel has a header: "Please select columns to be included in the output and hit 'Results' when ready" and a link "Missing non coding genes in your mart query output, please check the following FAQ". It contains two groups of radio buttons: "Features" (selected), "Structures", and "Homologues (Max select 6 orthologues)". Below these are two sections: "GENE:" and "Ensembl". The "GENE:" section lists attributes with checkboxes: Gene stable ID (checked), Gene stable ID version, Transcript stable ID, Transcript stable ID version, Protein stable ID, Protein stable ID version, Exon stable ID, Gene description, Chromosome/scaffold name (checked), Gene start (bp) (checked), Gene end (bp) (checked), Strand (checked), and Karyotype band. The "Ensembl" section lists attributes with checkboxes: APPRIS annotation (checked), 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). Most checkboxes in the "Ensembl" section are checked.

# BioMart - Results

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

New | Count | Results | URL | XML | Perl | Help

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

Gene stable ID	Gene name	Source of gene name	APPRIS annotation	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Strand
<a href="#">ENSDARG00000103139</a>	<a href="#">LO017843.1</a>	Clone-based (Ensembl) gene	<a href="#">principal1</a>	<a href="#">22</a>	<a href="#">3045495</a>	<a href="#">3078347</a>	1
<a href="#">ENSDARG00000100132</a>	<a href="#">CU929402.1</a>	Clone-based (Ensembl) gene	<a href="#">principal1</a>	<a href="#">22</a>	<a href="#">3232925</a>	<a href="#">3234494</a>	1
<a href="#">ENSDARG00000100533</a>	<a href="#">sich1073-178p5.3</a>	ZFIN	<a href="#">principal1</a>	<a href="#">22</a>	<a href="#">3238474</a>	<a href="#">3239834</a>	1
<a href="#">ENSDARG00000110077</a>	<a href="#">CU929402.2</a>	Clone-based (Ensembl) gene	<a href="#">principal1</a>	<a href="#">22</a>	<a href="#">3244950</a>	<a href="#">3271707</a>	1
<a href="#">ENSDARG00000053074</a>	<a href="#">gipc3</a>	ZFIN	<a href="#">principal1</a>	<a href="#">22</a>	<a href="#">3303671</a>	<a href="#">3328241</a>	1
<a href="#">ENSDARG00000104717</a>	<a href="#">tbxa2r</a>	ZFIN	<a href="#">principal1</a>	<a href="#">22</a>	<a href="#">3336723</a>	<a href="#">3344613</a>	-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 "Using this website" link is also present. The main content area is titled "Ensembl Tools" and contains a note about ready-made tools for processing data. Below this, a section titled "Processing your data" lists several tools:

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.	🔗	50MB*	🔗	🔗
Variant Recoder	Translate a variant identifier, HGVS notation or genomic SPDI notation to all possible variant IDs, HGVS, VCF format and genomic SPDI.	🔗	Maximum 1000 variants recommended	🔗	🔗
BLAST/BLAT	Search our genomes for your DNA or protein sequence.	🔗	50MB	🔗	🔗
File Chameleon	Convert Ensembl files for use with other analysis tools	🔗		🔗	🔗
Assembly Converter	Map (liftover) your data's coordinates to the current assembly.	🔗	50MB	🔗	🔗
ID History Converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	🔗	50MB	🔗	🔗
Linkage Disequilibrium Calculator	Calculate LD between variants using genotypes from a selected population.	🔗		🔗	🔗
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.	🔗		🔗	🔗
Data Slicer	Get a subset of data from a BAM or VCF file.	🔗		🔗	🔗
Post-GWAS	Upload GWAS summary statistics and highlight likely causal gene candidates.	🔗		🔗	🔗

- 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](http://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)**  
Page: 1 of 1 | Show: All variants

**Filters**  
Uploaded variant is defined Add

**Download**  
All: VCF VEP TXT  
BioMart: Variants Genes

Show/hide columns (16 hidden)

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](http://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](http://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

**Input:**

Species: 22 3144711 3144711 sa39354

Assembly mapping: 22 3145013 3145013 sa43743

Name for this job (optional)

Input file format:

Either paste data:

**Output:**

22 3161984 3161984 sa39354

22 3162286 3162286 sa43743

Or upload file:  Choose file Two 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](http://genome.ucsc.edu/cgi-bin/hgPcr)

# UCSC In-Silico PCR

- Fast search
- [genome.ucsc.edu](http://genome.ucsc.edu/cgi-bin/hgPCR/PCRSearchFormV2?db=Zebrafish&species=Danio+rerio)

The screenshot shows the UCSC In-Silico PCR search interface. At the top, there is a navigation bar with links to Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, and About Us. Below the navigation bar, the title "UCSC In-Silico PCR" is displayed. The main search form includes fields for Genome (Zebrafish), Assembly (May 2017 (GRCz11/danRer11)), Forward Primer (CCGGGGAGCAGTTGA), Reverse Primer (TGGGTGGAGTAGGTCTG), Max Product Size (4000), Min Perfect Match (15), Min Good Match (15), and a Flip Reverse Primer checkbox. Below the search form is a section titled "About In-Silico PCR" which provides a brief description of the tool and a link to a video example. 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 for primer matches. A sample sequence output 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
```

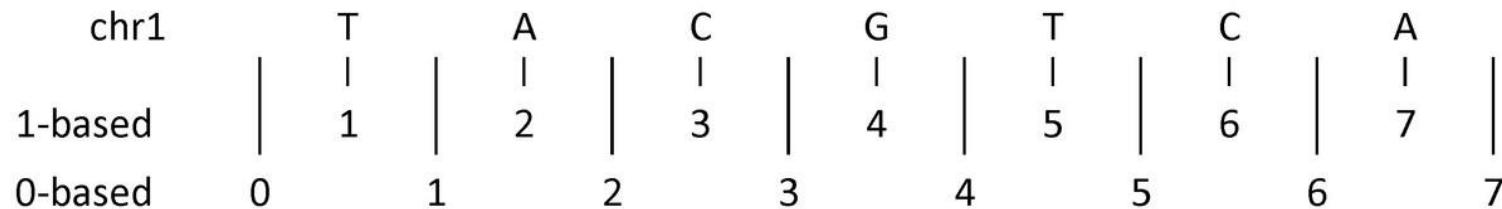
# UCSC In-Silico PCR

- Fast search for people
  - [genome.ucsc.edu](http://genome.ucsc.edu)

Genomes	Genome Browser	Tools	Mirrors	Downloads	My Data
<h2>UCSC In-Silico PCR</h2>					
>chr14_KZ115440v1_alt:182433-183230	798bp CCCGGGGAGCAGTTGATcccttgcggggtaagactaaatcccct tgcattaaatgtcatgtcgtttccataactaggatttgcagaaggttgcatg tgcataaaaatgtcgacattnaaataaataaataatgttgttatgataaat gtgtatattttgcgtcccttcattttaaataaataaaggttatacatgtgg gttattttgtataatgtgtataataatattgcatgttgaaagtaaaaaa tatcattttaaataaaggctgacaggcgttataatggagttactgtttaaa ctaaggaaatgtgaaggatcttgcggggataaattttaaaggctgtgc cttttcgttt tggaaaaggatcaggatgtggtaaggataaaggactgtcgagagaaaa tgcacatgttagatgttttttttttttttttttttttttttttttttt ccggcaaatgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt ccctttgcgtccatcgcacccgtctgtctgtgtgtgtgtgtgtgt cagagtctcgacgcgcgcgcgcgcgcgcgcgcgcgcgcgcgc attttatgtactaaaaataactgttttttttttttttttttttttt accacactcttttttttttttttttttttttttttttttttttttt aattttccatggagaaacctctgtcagatCAGACCTACTCCACCAACG				
>chr14:21334984-21335781	798bp CCCGGGGAGCAGTTGAT CGTTGGGTGGAGTAGGTCTG CCCGGGGGAGCAGTTGATcccttgcggggtaagactaaatcccct tgcattaaatgtcatgtcgtttccataactaggatttgcagaaggttgcatg tgcataaaaatgtcgacattnaaataaataaataatgttgttatgataaat gtgtatattttgcgtcccttcattttaaataaataaaggttatacatgtgg gttattttaaataatgtgtataataatattgcatgttgaaagtaaaaaa tatcattttaaataaaggctgacaggcgttataatggagttactgtttaaa ctaaggaaatgtgaaggatcttgcggggataaattttaaaggctgtgc cttttcgttttttttttttttttttttttttttttttttttttttt tggaaaaggatcaggatgtggtaaggataaaggactgtcgagagaaaa tgcacatgttagatgttttttttttttttttttttttttttttt ccggcaaatgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgtgt ccctttgcgtccatcgcacccgtctgtctgtgtgtgtgtgtgt cagagtctcgacgcgcgcgcgcgcgcgcgcgcgcgcgc attttatgtactaaaaataactgttttttttttttttttttttt accacactcttttttttttttttttttttttttttttttttt aattttccatggagaaacctctgtcagatCAGACCTACTCCACCAACG				

# 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

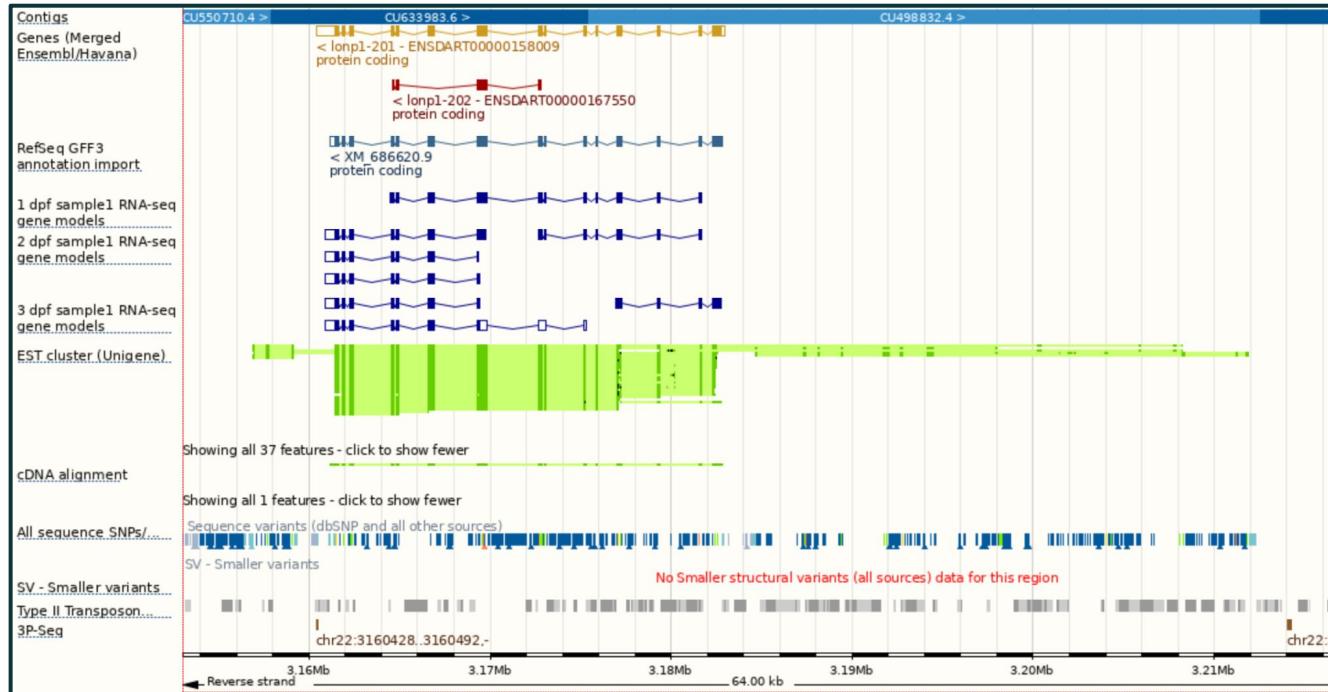
- Click “Custom tracks” and add <https://mbl2022.buschlab.org/data/3p-seq.bed>

The screenshot shows the Ensembl genome browser interface for Zebrafish (Danio rerio) GRCz11 assembly. The left sidebar shows various genomic tracks like Whole Genome, Chromosome, and UCSC. The main panel is titled "Custom tracks" under "Your data". It includes 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 that track hubs and indexed files (BAM, BigBed, etc) do not work with certain cloud services like Google Drive and Dropbox. A "Help on supported formats, display types, etc" link is also present. On the right, a genomic track viewer displays several tracks across a genomic region from 6,200 to 7,600.

- 24 hpf 3P-Seq data from Bartel lab

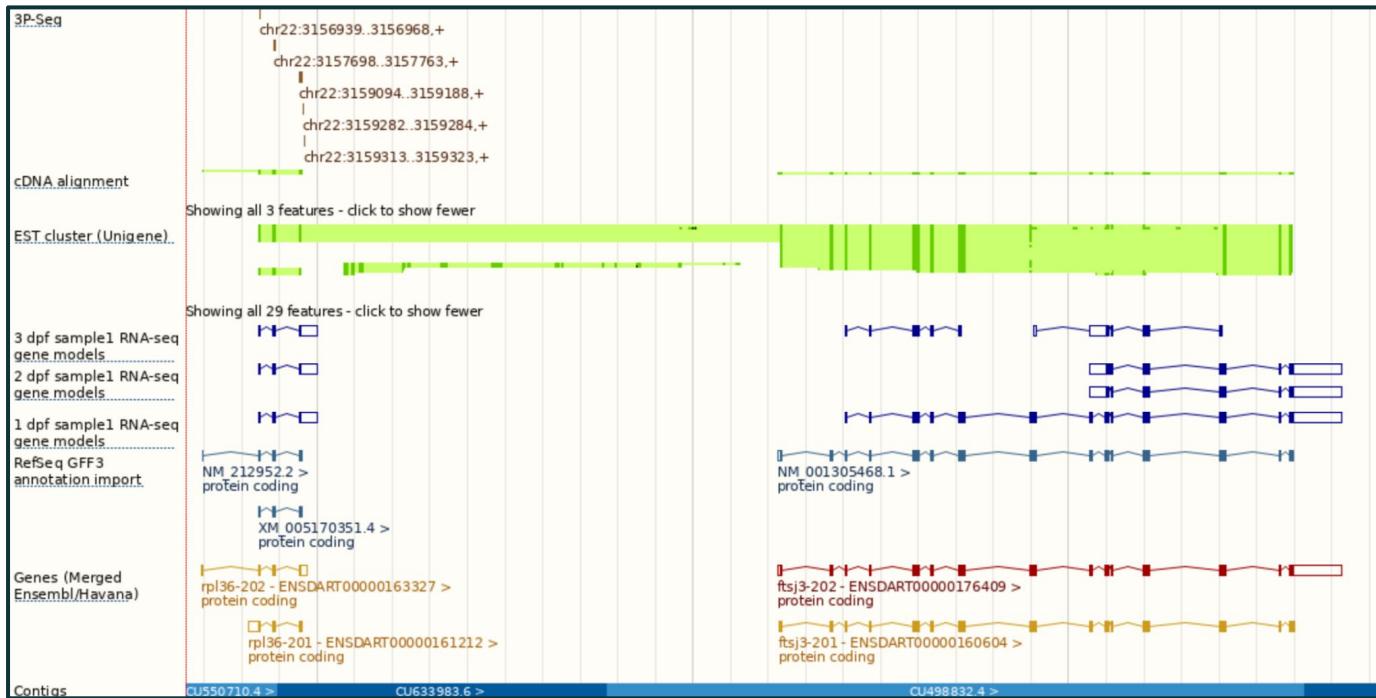
# Custom Tracks

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



# Custom Tracks

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



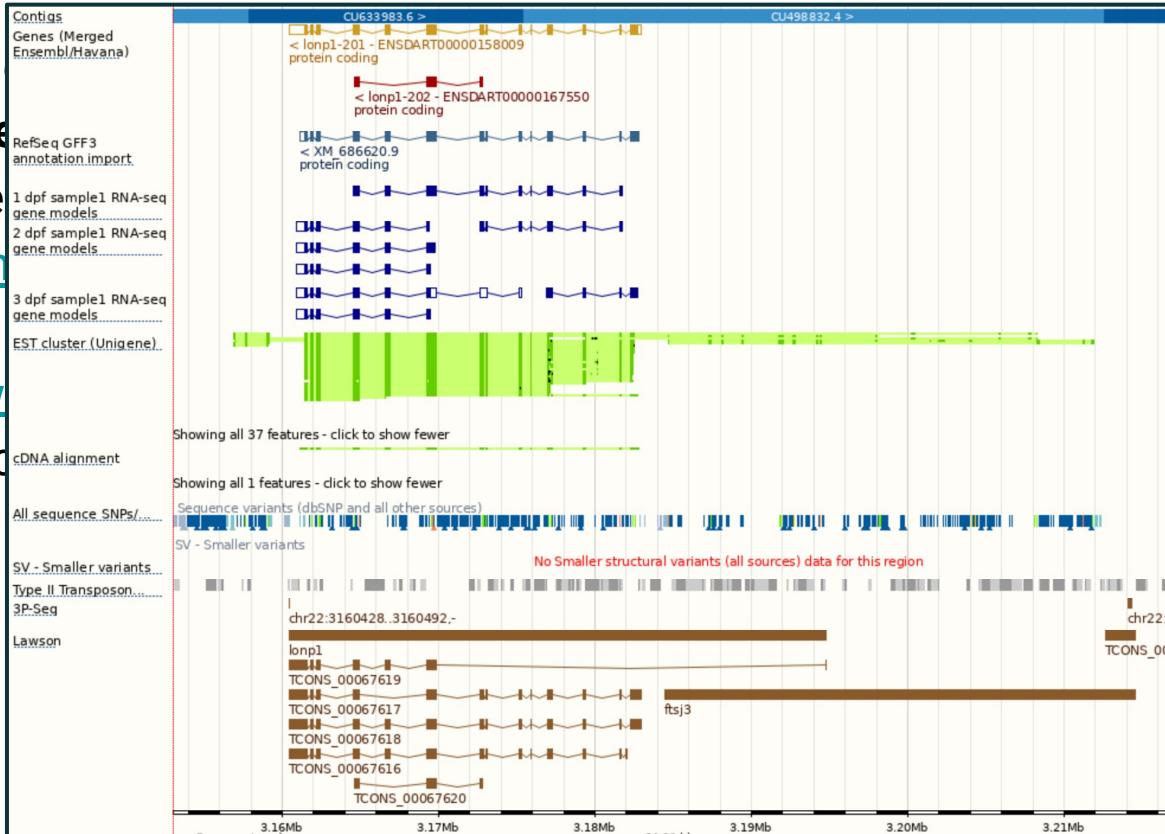
# Custom Tracks - Lawson Lab Annotation

---

- 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/](http://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:e [www.umass.edu](http://www.umass.edu)
- Add: <https://www.umass.edu/biology/lawson-lab>
- Large, so



otation for  
nes"  
ome/  
<https://www.umass.edu/biology/lawson-lab>  
s/v4.3.2.gtf

# Exercise 3

---

- Do Exercise 3 - “exploring data”
- Covers:
  - BioMart
  - Making BED files
  - Finding candidate genes
  - Finding orthologues
- Go to [mbl2022.buschlab.org](http://mbl2022.buschlab.org)

# Thank You!

Any questions?

