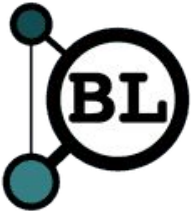


# Genome Literacy Workshop

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

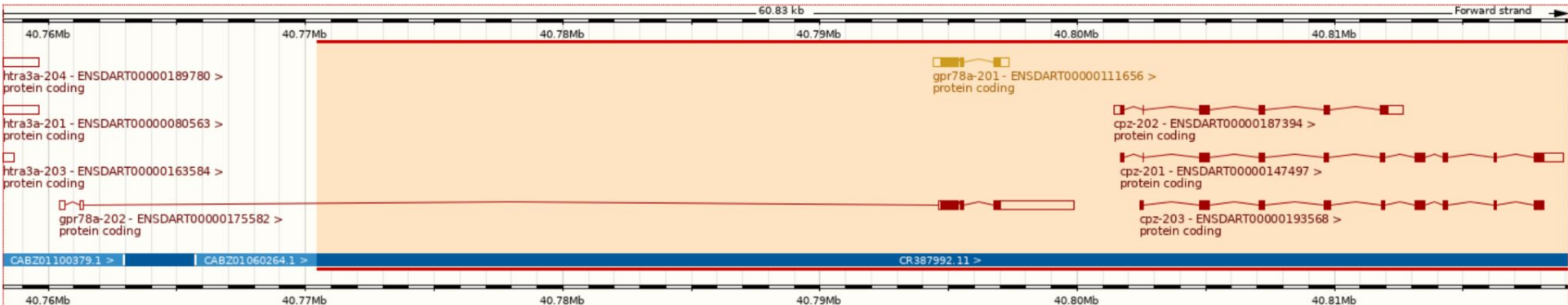
- Most examples from **Ensembl** (we are biased!)
- Probably most widely used genome browser amongst zebrafish researchers
- **Primary source of zebrafish annotation** (UCSC imports Ensembl annotation)
- Currently Ensembl version **107** (July 12th)
- New releases 3 or 4 times / year
- Zebrafish **annotation largely static** between releases
- But **naming and homology** updated (+ new functionality)

The screenshot displays the Ensembl genome browser interface. At the top, there's a navigation bar with links for BLAST/BLAT, VEP, Tools, and More. A search bar is present with the text "Search all species...". Below the navigation bar, there's a section for "Tools" and "BioMart", and a section for "BLAST/BLAT" and "Variant Effect Predictor". The main content area features a "Search" section with a dropdown menu for "All species" and a "Go" button. Below the search section, there's a "All genomes" section with a dropdown menu for "Select a species --". The "Favourite genomes" section lists "Human" (GRCCh38.p13), "Pig breeds" (Pig reference genome and 12 additional breeds), "Mouse" (GRCm39), and "Zebrafish" (GRCz11). The "Zebrafish" entry is circled in red. The right sidebar contains a "News" section titled "Ensembl Release 107 (Jul 2022)" with several bullet points about updates. At the bottom right, there's a "Rapid Release" section with the text "New assemblies with gene and protein annotation every two weeks."



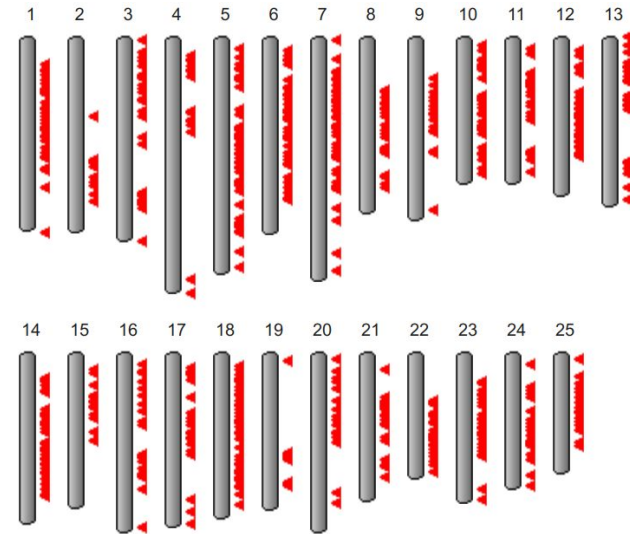
# Zebrafish Genome

- **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, "**toplevel1**" 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

Archive! Ensembl BioMart | Tools | More ▾

Search all species...

Zebrafish (GRCz10) ▾

Search Zebrafish (*Danio rerio*)

Search all categories ▾ Search Zebrafish...

Go

e.g. SLC24A5 or 10:10138322-10349251 or rs3727517 or kinesin

What's New in Zebrafish release 91

- Structural variants
- New dbSNP data for zebrafish
- Fixing stable ids in the external data database

More news...

Genome assembly: GRCz10 (GCA\_000002035.3)

More information and statistics

Download DNA sequence (FASTA)

Display your data in Ensembl

Other assemblies

Zv9 (Ensembl release 79) ▾ Go

View karyotype

Example region

Gene annotation

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

More about this genebuild

Download genes, cDNAs, ncRNA, proteins (FASTA)

Example gene

Example transcript

# 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**

**ZFIN** Search [Sign In](#)

*dmd*

Summary

Expression

Phenotype

Mutations

Human Disease

Gene Ontology

Protein Domains

Transcripts

Interactions and Pathways

Antibodies

Plasmids

Constructs

Marker Relationships

Sequences

Orthology

GENE

*dmd*

**ID** ZDB-GENE-010426-1

**Name** *dystrophin*

**Symbol** *dmd* [Nomenclature History](#)

**Previous Names** *cb664* (1), *Dp71* (1), *Duchenne muscular dystrophy* (1), *im:6911785*, *sap*, *sapje-like* (1), *sapje*, *zfdYS* (1), *zgc:110165*

**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 (dystrophin).

**Genome** [Alliance](#) (1), [Gene:83773](#) (1), [VEGA:OTTDARG0000000319092](#) (1), [Ensembl\(GRCz11\):ENS DARG00000008487 \(1\)](#)

**Resources**

**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 genome browser interface. At the top, the Ensembl logo is on the left, and navigation links for BLAST/BLAT, VEP, Tools, BioMart, and More are in the center. A search bar on the right contains the text "Search Zebrafish...". Below the navigation bar, the "New Search" tab is active. The "Current selection:" section shows "< all Species" and "Only searching Zebrafish". The "Restrict category to:" section lists Gene (2), Transcript (13), GeneTree (1), and GenomicAlignment (5). The "Per page:" section shows options 10, 25, 50, and 100. The "Layout:" section shows "Standard" and "Table". The "Tip:" section provides advice on choosing results and updating species. The search results section shows "Only searching Zebrafish" and "dmd" with a search icon. It states "21 results match dmd when restricted to" and "species: Zebrafish". Below this, a "Did you mean..." section lists three results: "dmd (Zebrafish Gene)", "dmd-201 (ZFIN transcript name record; description: dystrophin.)", and "dmd-213 (ZFIN transcript name record; description: dystrophin.)". Each result includes a link to the gene page and a description of the transcript.

**Ensembl** BLAST/BLAT | VEP | Tools | BioMart | More ▾

New Search

**Current selection:**  
< all Species  
Only searching Zebrafish

**Restrict category to:**  
Gene 2  
Transcript 13  
GeneTree 1  
GenomicAlignment 5

**Per page:**  
10 25 50 100

**Layout:**  
Standard Table

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

Only searching Zebrafish ▾

21 results match **dmd** when restricted to species: Zebrafish ✕

[Did you mean... ▾](#)

[dmd \(Zebrafish Gene\)](#)  
**ENSDARG00000008487** 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 BLAST/BLAT interface. The 'BLAST/BLAT' link in the top navigation bar is circled in red. The left sidebar contains a 'Web Tools' menu with 'BLAST/BLAT' selected, and a sub-menu showing 'dmd' as the selected gene. The main content area displays 'Results for dmd' with job details and a results table.

**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 (Sequence)	1779	0.0	100.000 (Alignment)
1:11031622-11032521 (Sequence)	dmd	Reverse	1	900	900 (Sequence)	1779	0.0	100.000 (Alignment)
1:1624334-1624715 (Sequence)	cltc6	Reverse	1	387	390 (Sequence)	686	0.0	96.923 (Alignment)
1:11349264-11349645 (Sequence)	sdh1b	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 ( <a href="#">ENSORLG00000020638</a> ) <a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (2:208,119-221,155:1) <a href="#">View Sequence Alignments</a>	84.52 %	88.22 %	0	<b>95.81</b>	Yes
Lumpfish ( <i>Cyclopterus lumpus</i> )	1-to-many	dmd ( <a href="#">ENSCLMG00000500931</a> ) <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>	82.21 %	82.49 %	0	<b>95.20</b>	Yes
Lyretail cichlid ( <i>Neolamprologus brichardi</i> )	1-to-1	dmd ( <a href="#">ENSNBRG00000015200</a> ) <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>	87.34 %	89.39 %	0	<b>96.66</b>	Yes
Makobe Island cichlid ( <i>Pundamilia nyererei</i> )	1-to-1	dmd ( <a href="#">ENSPNYG00000022641</a> ) <a href="#">View Gene Tree</a> <a href="#">Compare Regions</a> (JH419417.1:620,205-712,305:-1) <a href="#">View Sequence Alignments</a>	45.32 %	89.56 %	0	<b>96.75</b>	Yes



# 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. ENS**DAR**G00000028213 (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 annotation changes
- Stable IDs have **versions**, e.g.
  - Version number of **ENSDARG** indicates change
  - Version number of **ENSART** indicates change
  - Version number of **ENSDARP** indicates change
  - Version number of **ENSARE** indicates change
- Can also be **removed**, e.g. see below

Gene: ENSDARG00000058767

This identifier is not in the current Ensembl database

**ID History**

Stable ID: ENSDARG00000058767.4  
Status: Retired (see below for possible successors)  
Latest Version: ENSDARG00000058767.4  
Release: 86  
Assembly: GRCh10  
Database: danio\_rerio\_core\_86\_10

Associated archived IDs for this stable ID version

Release	Gene	Transcript	Protein
86	<a href="#">ENSDARG00000058767</a>	<a href="#">ENSART00000081704.4</a>	<a href="#">ENSDARP000000076143</a> MENNTNFMFLFENLGYIRYALFILGFVLYFSIIFNVLMILAVFLERTLHQPMYILISC LSMNSLFGTAGGFFPRVLSDLLSETHSISREACFQSFVIFTYAANESILMIMAFDRFAA ICKPLHYHSIVRPRFLACVIVTNLTFFMILLGVAGLLTTKLRCMGNLKFVYCHSYEIVK LSCDNIANNAGFLFILIITTIPLSLISFSYVKIIICQRSSAQFGKAFQTCIPHIVV LLNFTIAVICDTLSRVVNLQIPVGLSVFLSLEFLIIPILNPLIYAFNLPDIRKRMISL IKPFR

Export image

Score: 1.00 (black), >=0.99 (dark red), >=0.97 (red), >=0.90 (orange-red), >=0.75 (orange), >=0.50 (yellow), <0.50 (light yellow), Unknown (grey)

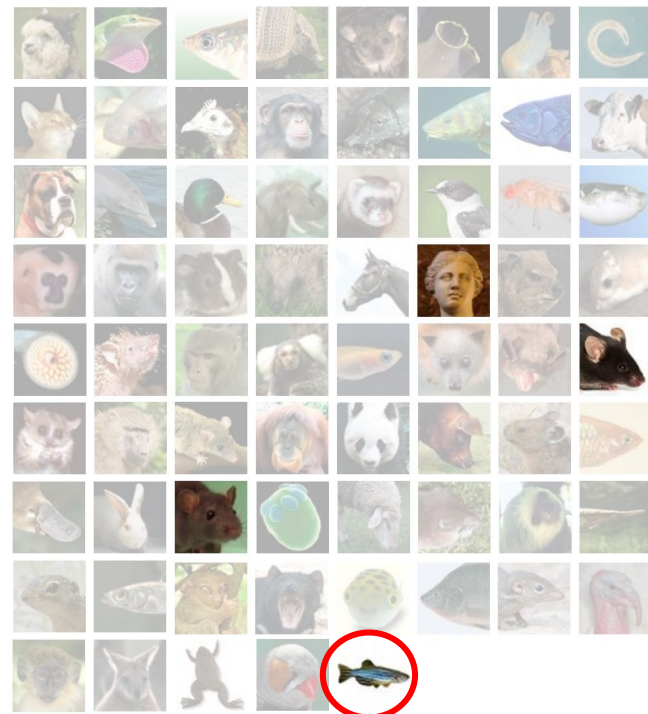
Assembly: GRCh11

# 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)

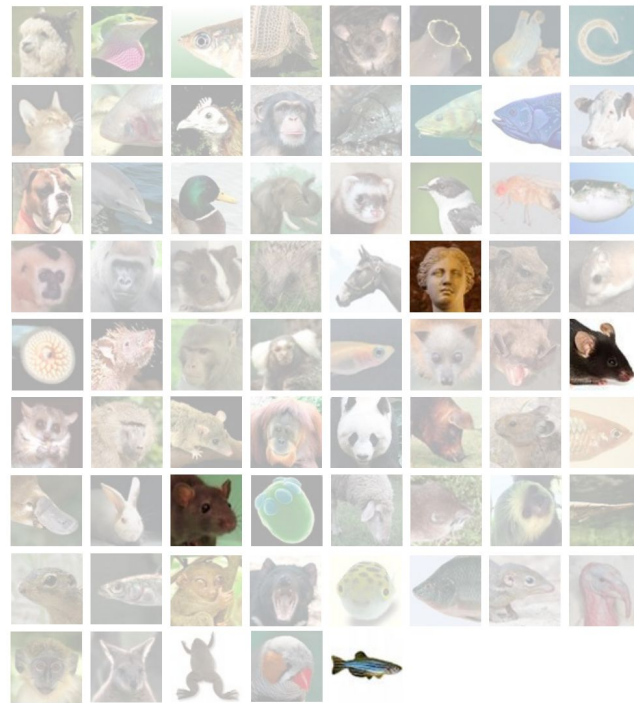


From Ensembl training materials, CC BY 4.0 license



# Manual Annotation

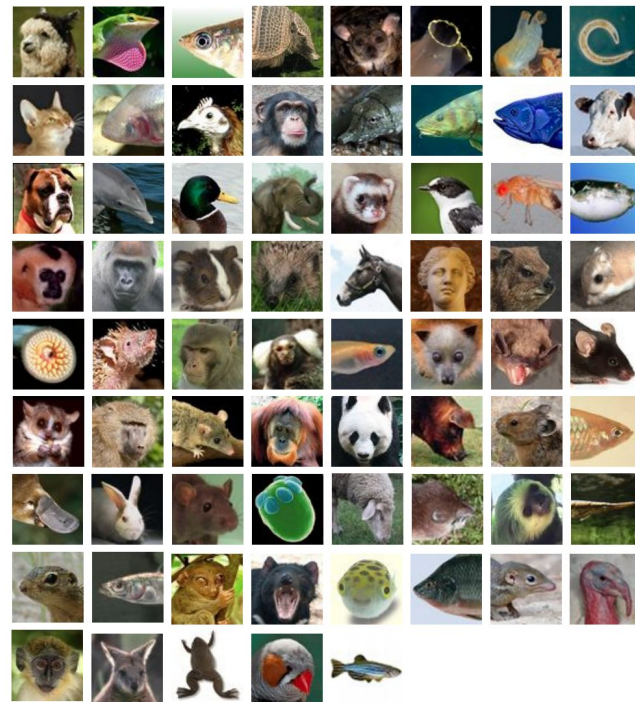
- **Gold** standard
- Uses information from databases and publications
- More accurate for tricky areas:
  - e.g. UTRs, splice sites, single exon transcripts
- **Slower** and more expensive
- Thorough, but leads to inclusion of transcripts that may not be representative (e.g. low expression)
- Only clones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



From Ensembl training materials, CC BY 4.0 license

# Automatic Annotation

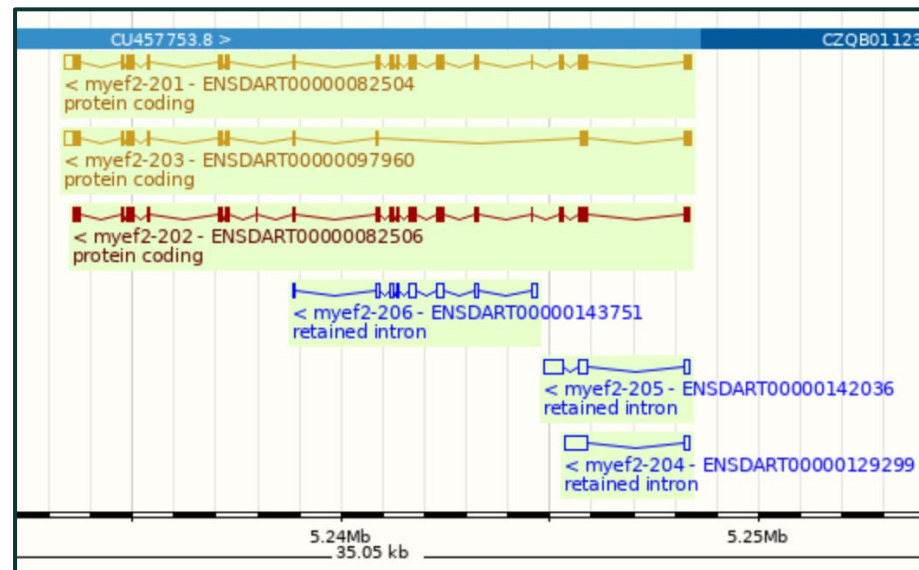
- **Faster**
- Uses evidence from sequences **deposited** in ENA/GenBank/DDBJ and UniProt proteins
- **Overview:**
  - Identify repeats and low complexity sequence with RepeatMasker, Dust and TRF
  - Run GENSCAN to identify *ab initio* gene predictions
  - Align UniProt proteins to GENSCAN predictions, 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** ENSDARG00000077526

Description BRIS1 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

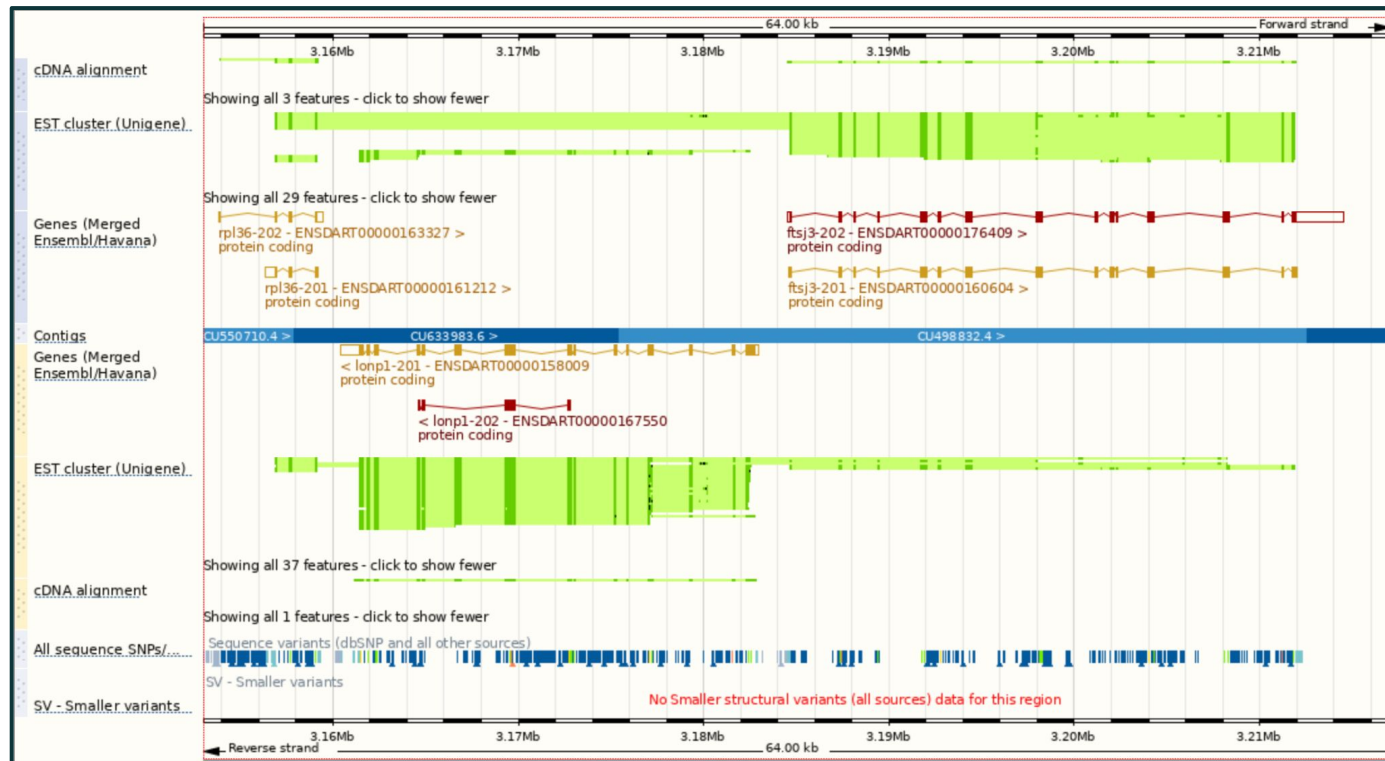
About this gene This gene has 4 transcripts ([splice variants](#)) and [185 orthologues](#).

Transcripts [Hide transcript table](#)

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="#">ENSDART00000008980.8</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="#">A0A0R4I1K1</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

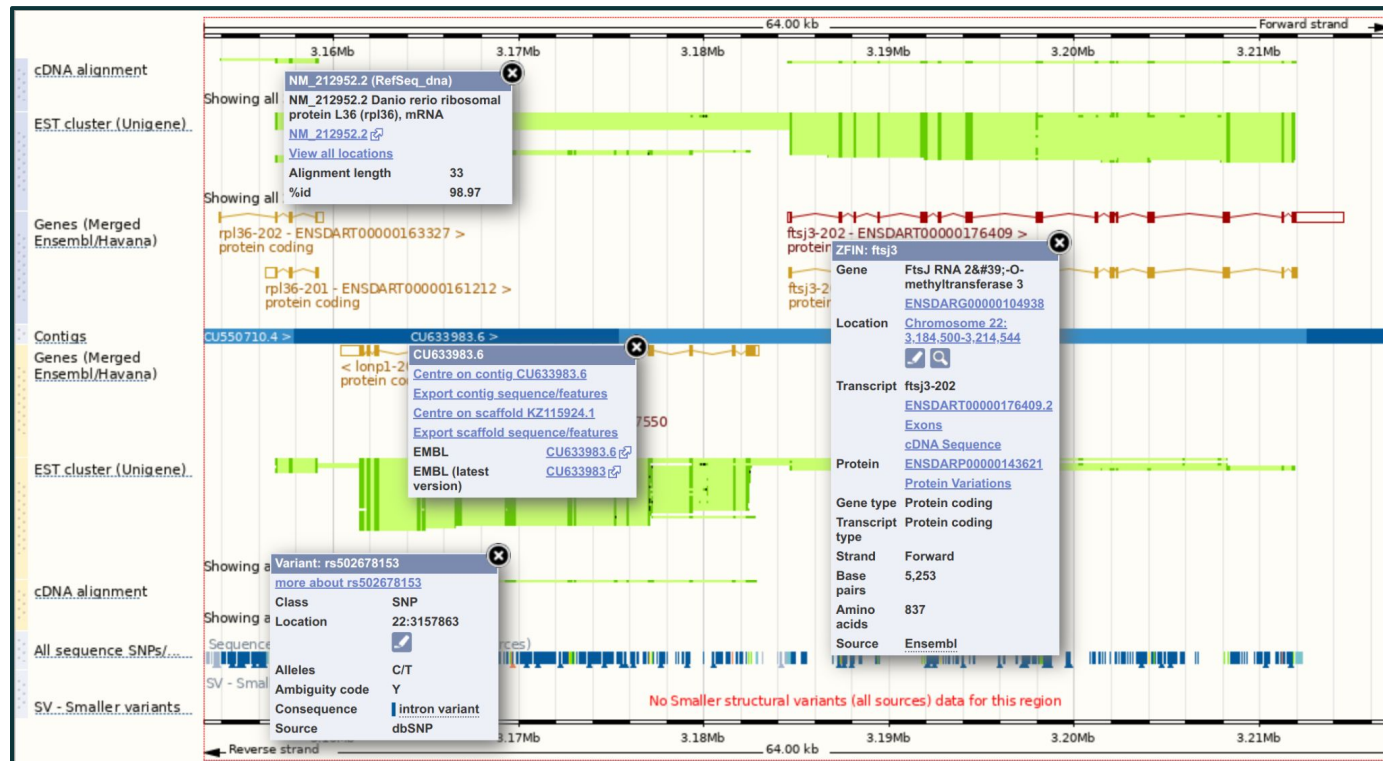
- Go to "22:3153000-3217000"



- 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](https://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 with the 'Configure this page' dialog box open. The dialog box has tabs for 'Configure Region Image', 'Configure Overview Image', 'Configure Chromosome Image', and 'Personal Data'. The 'Configure Region Image' tab is active, showing a list of tracks on the left and configuration options on the right.

**Active tracks:**

- 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

**Configuration options:**

- Select from available configurations: Current unsaved (dropdown) [Save current configuration](#)
- Genes and transcripts [Show tutorial](#)
- RNASeq models
- Filter by: All classes (dropdown)  
Enter terms to filter by (text input)
- Key: Shown (dark blue square), Hidden (light blue square), No Data (grey square), Filtered: Shown (green square), Hidden (light green square)
- Default style: [Enable/disable all](#) (checkboxes for BAM files, Gene models, Intron-spanning reads)

**ENA table:**

ENA	0	1	0	1
1 dpf sample1	0	1	0	1
14 dpf sample1	0	1	1	1
2 dpf sample1	0	1	0	1
3 dpf sample1	0	1	1	1
5 dpf sample1	0	1	1	1

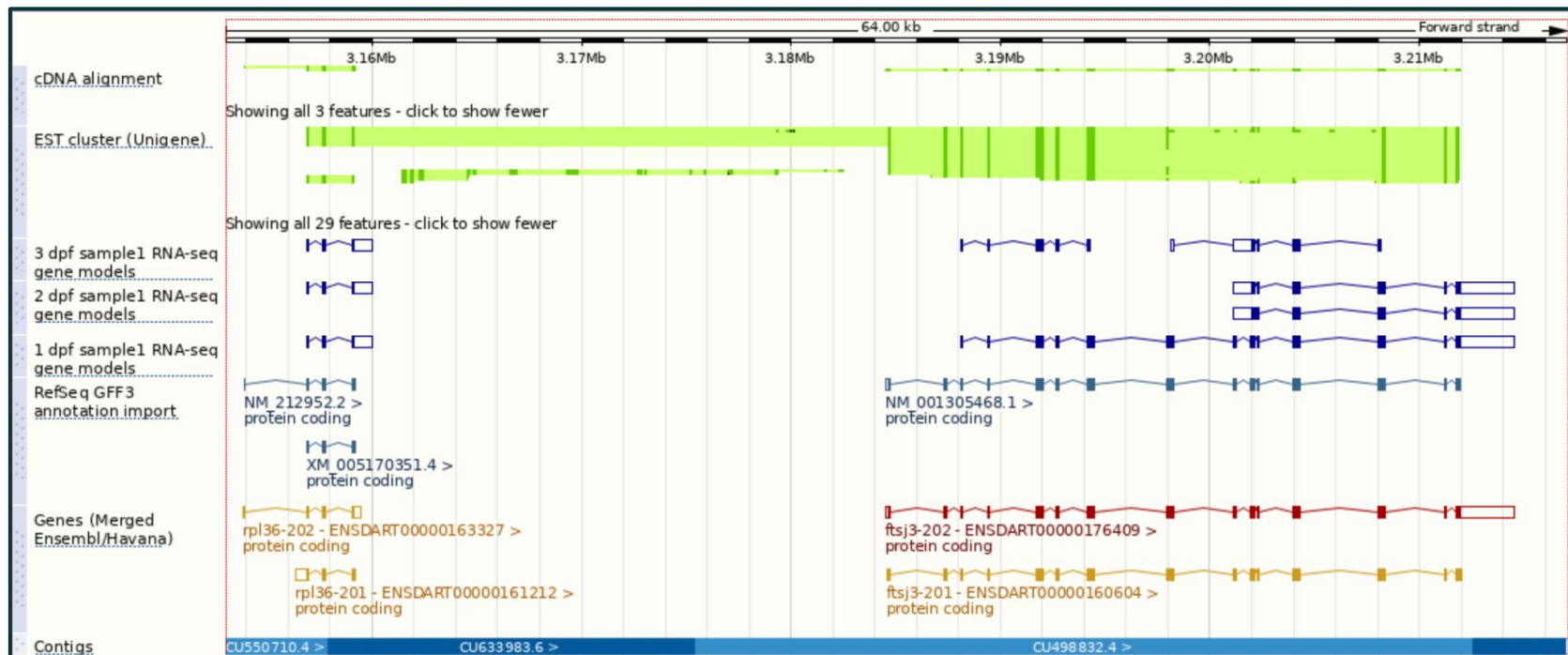
The background shows the Ensembl genome browser interface with the '22:3153000-3217000' region selected. The 'cDNA alignment' track is visible at the bottom.

# 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

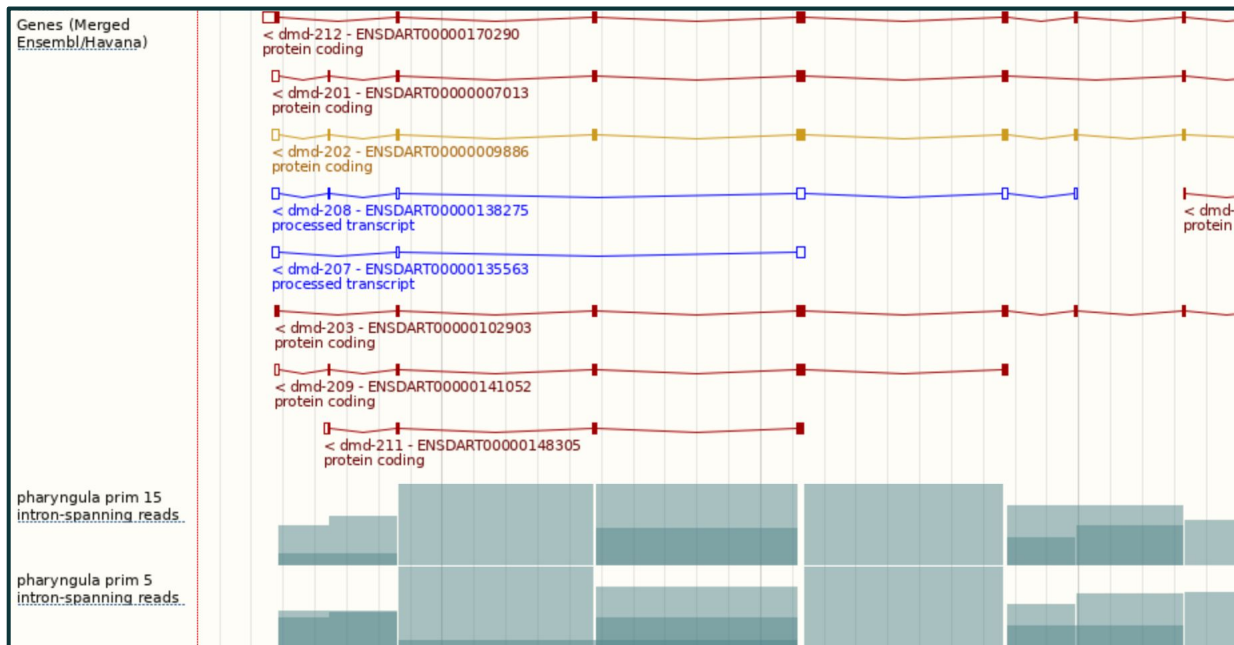
# "Configure this page" Demo

- 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

lon 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 paralogue](#).

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**
  - 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: lonp1** ENSDARG00000102765

**Description** lon 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 paralogue](#).

**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

**Summary**

**Name** [lonp1](#) (ZFIN)

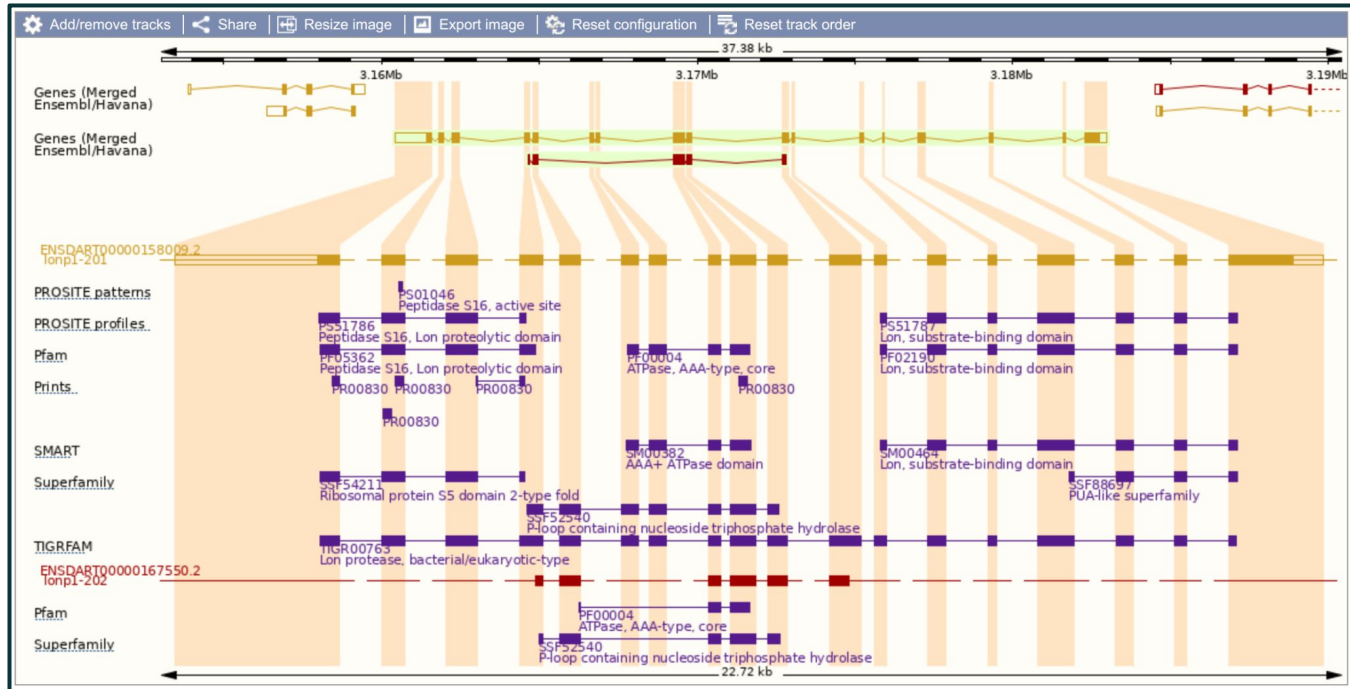
**Ensembl version** ENSDARG00000102765.2

**Gene type** Protein coding

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

# "Gene" Demo - Splice Variants

- Go to ENSDARG00000102765 and click on "Splice variants"



# "Gene" Demo - Orthologues

- Go to ENSDARG00000102765 and click on "Orthologues"

Show <div>All</div> entries		Show/hide columns		Filter				
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="#">(ENSCABG00000010924)</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="#">(ENSSCUG00000004632)</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_SPRETEJ_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="#">(ENSMMSG00000018859)</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="#">(ENSPFOG00000001826)</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

lon 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 paralogue](#).

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

Show/hide columnsFilter

Type	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Paralogues	Bilateral animals (Bilateria)	<a href="#">ENSDARG00000101438</a>  lonp2 lon 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 ?					
Show/hide columns (1 hidden)			Filter		
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:

Filter 	
External database	Database identifier
Expression Atlas	<a href="#">ENSDARG00000102765</a>  <a href="#">[view all locations]</a>
NCBI gene (formerly Entrezgene)	<a href="#">lonp1</a>  lon peptidase 1, mitochondrial <a href="#">[view all locations]</a>
WikiGene	<a href="#">lonp1</a>  lon peptidase 1, mitochondrial <a href="#">[view all locations]</a>
ZFIN	<a href="#">lonp1</a>  lon 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”

• White RJ, Collins JE, Sealy IM, Wali N, Dooley CM et al. (2017) [A high-resolution mRNA expression time course of embryonic development in zebrafish.](#)

Raw Data Provider: [Vertebrate Genetics and Genomics Group \(Wellcome Trust Sanger Institute\)](#)

Results

Experiment Design

Supplementary Information

Downloads

## Genes

[Show boxplot and transcripts view](#)

Showing 1 gene:

ENSDARG00000102765 ✕

[Ensembl genome browser](#) ▼

[Download](#)

Click on a cell to open the selected genome browser with attached tracks if available

Expression level in TPM  
0 1,479

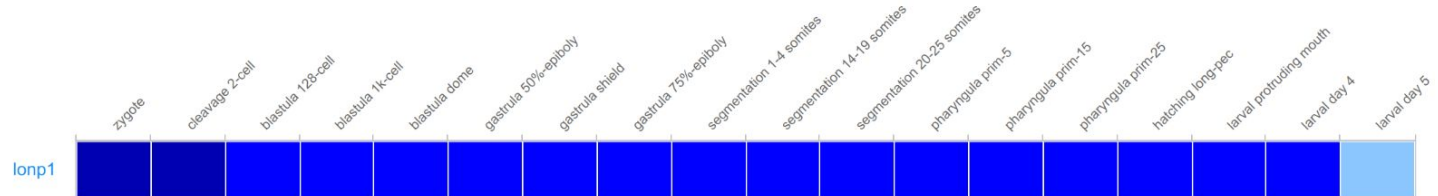
Apply

Clear

☒ Most specific

Expression value

0.5



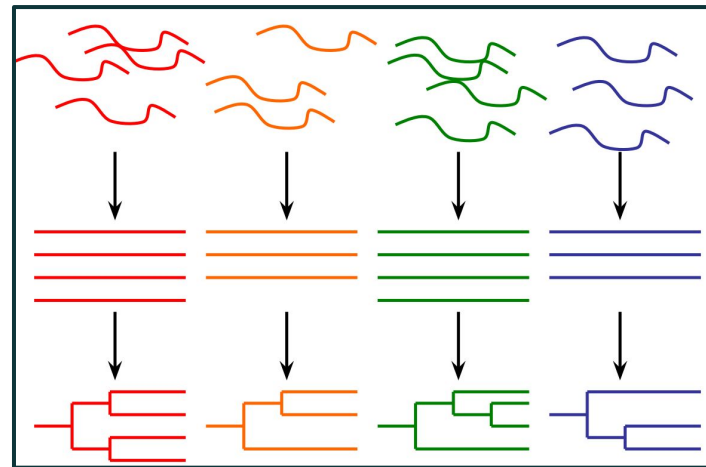
Feedback

# Compara

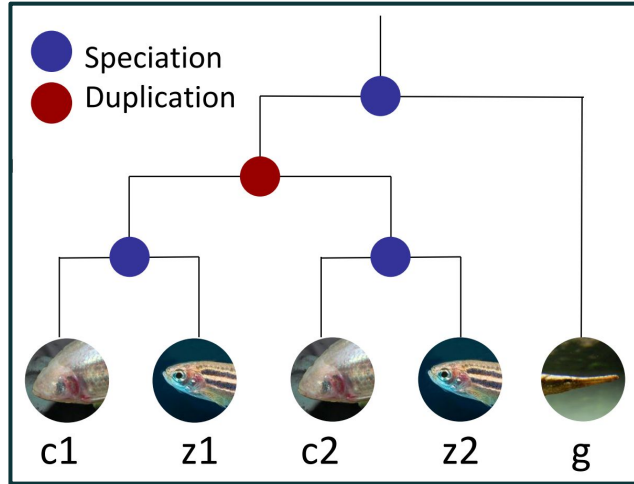
- Compara - produce Ensembl's comparative genomics resources
- Two types of analysis:
  - Gene level comparisons to produce **gene trees**, e.g. infer **homologues** (orthologues & paralogues)
  - **Whole genome alignments** - pairwise and multiple alignments, e.g. **constrained elements** and **synteny**

# 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**



# Compare - Infer Homologues (Orthologues & Paralogues)



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

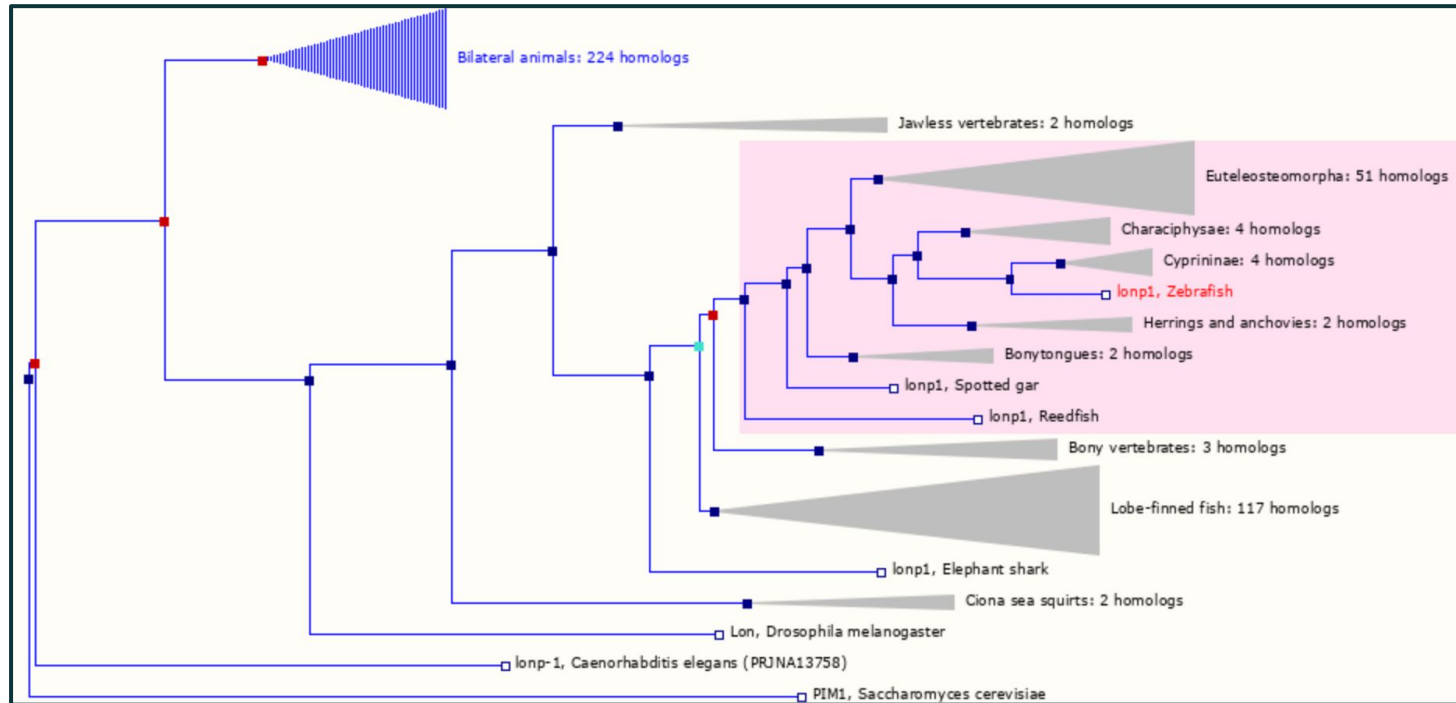
**z1 & c1** are **orthologues** (arose from **speciation**), as are **z2 & c2** + **z2 & g**, etc...

**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*)

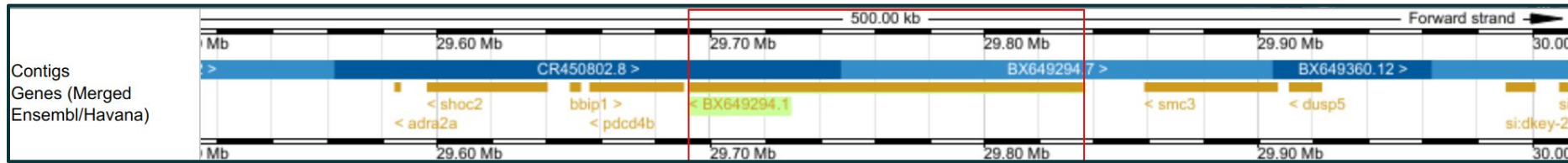
# Synten Example

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

<i>Homo sapiens</i> genes	Location		<i>Danio rerio</i> 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> (ENSDARG000000041022)	<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> (ENSDARG000000071046)	<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> (ENSDARG000000040853)	<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> (ENSDARG000000040841)	<a href="#">22:29584800-29586608</a>	<a href="#">Region Comparison</a>

# Synteny Example

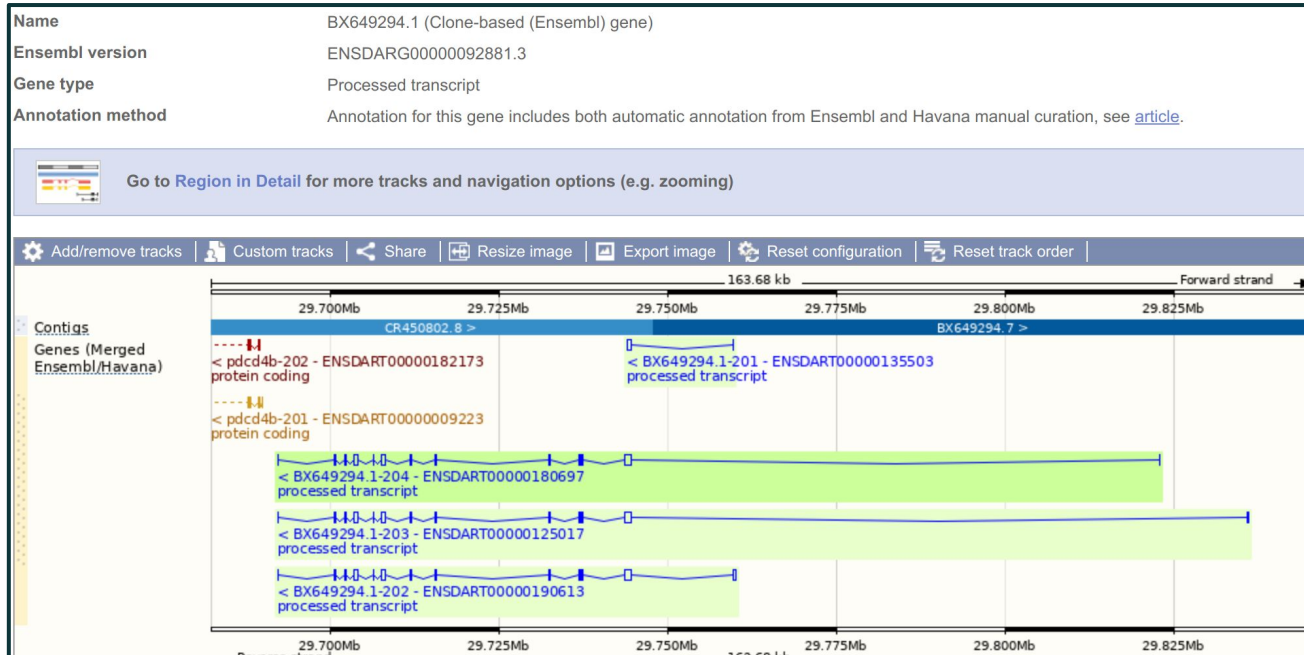
- If we look at the chr22 region in zebrafish then all the surrounding genes are the same and RBM20 is likely to be BX649294.1



<i>Homo sapiens</i> genes	Location		<i>Danio rerio</i> homologues	Location	
<a href="#">DUSP5</a> (ENSG00000138166)	<a href="#">10:110497907-110511533</a>	→	<a href="#">dusp5</a> (ENSARG00000019307)	<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> (ENSARG00000019000)	<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> (ENSARG00000041022)	<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> (ENSARG00000071046)	<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> (ENSARG00000040853)	<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> (ENSARG00000040841)	<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

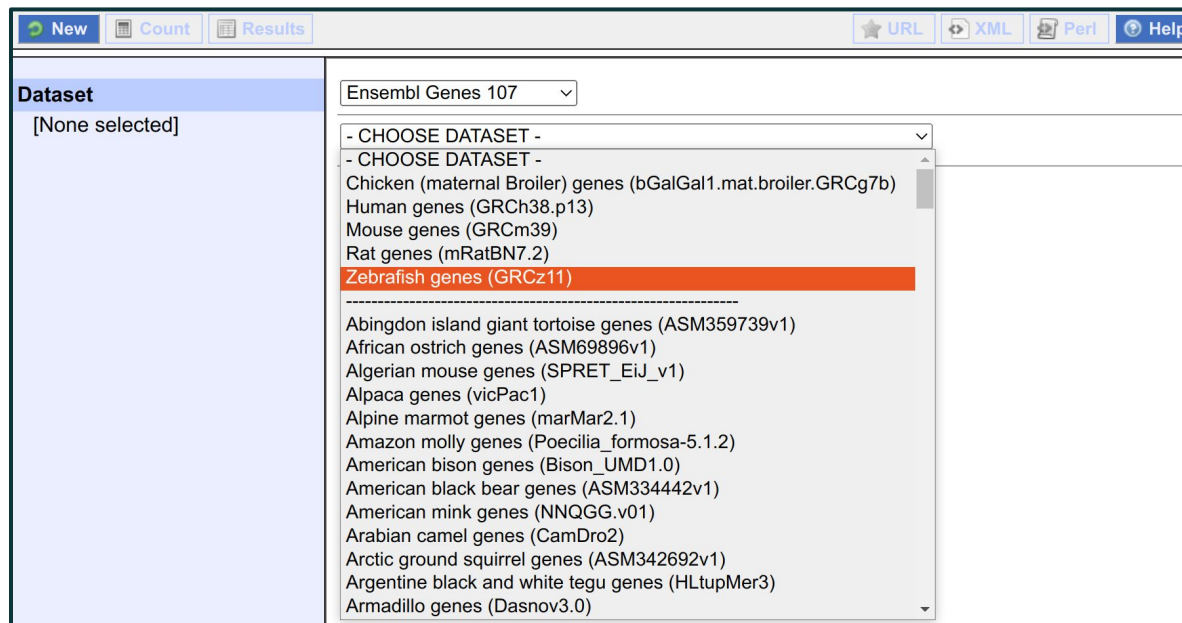
- **Export** (large amounts of) Ensembl data without programming
- Completely **customisable**, but **simple** to make complex queries
- Four stages:
  - Dataset
  - Filters
  - Attributes
  - Results

The screenshot shows the Ensembl BioMart web interface. The top navigation bar is dark blue with the Ensembl logo and links for BLAST/BLAT, VEP, Tools, **BioMart** (circled in red), Downloads, Help & Docs, and Blog. A search bar on the right says "Search all species...". Below the navigation bar is a light blue bar with buttons for New, Count, Results, URL, XML, Perl, and Help. The main content area has a "Dataset" section on the left with "[None selected]" and a dropdown menu on the right set to "- CHOOSE DATABASE -".



# 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 displays the BioMart web interface for filtering Zebrafish genes (GRCz11). The interface is divided into several sections:

- Dataset:** Zebrafish genes (GRCz11)
- Filters:** A list of filters is shown, including "Chromosome/scaffold: 22", "Start: 3000000", "End: 4000000", "Transcript count <=: 1", and "Gene type: protein\_coding".
- Attributes:** A list of attributes is shown, including "Gene stable ID", "Gene stable ID version", "Transcript stable ID", and "Transcript stable ID version".
- Filter Selection:** The "Gene type" filter is selected, and a dropdown menu shows the following options: "antisense", "IG\_C\_gene", "IG\_C\_pseudogene", "IG\_J\_pseudogene", "IG\_pseudogene", "IG\_V\_pseudogene", "polymorphic\_pseudogene", "processed\_pseudogene", "processed\_transcript", "protein\_coding" (highlighted in orange), and "pseudogene".
- Source:** The "Source (gene)" and "Source (transcript)" filters are set to "ensembl".
- APPRIS annotation:** The "APPRIS annotation" filter is set to "Only".

The interface also includes a top navigation bar with tabs for "New", "Count", and "Results", and a right sidebar with links for "URL", "XML", "Perl", and "Help".

# BioMart - Attributes

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

NewCountResults

URLXMLPerlHelp

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  
Dataset  
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready  
Missing non coding genes in your mart query output, please check the following [FAQ](#)

☒ Features  
☐ Structures  
☐ Homologues (Max select 6 orthologues)

☐ Variant (Germline)  
☐ Sequences

GENE:  
Ensembl

☒ Gene stable ID  
☐ 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  
☒ Gene start (bp)  
☒ Gene end (bp)  
☒ Strand  
☐ Karyotype band

☒ APPRIS annotation  
☐ Ensembl Canonical  
☐ Readthrough  
☒ Gene name  
☒ Source of gene name  
☐ Transcript name  
☐ Source of transcript name  
☐ Transcript count  
☐ Gene % GC content  
☐ Gene type  
☐ Transcript type  
☐ Source (gene)  
☐ Source (transcript)

# BioMart - Results

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

New

Count

Results

URLXMLPerlHelp

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

TSV

HTML

CSV

TSV

XLS

☐ Unique results only

Go

Email notification to

View

10

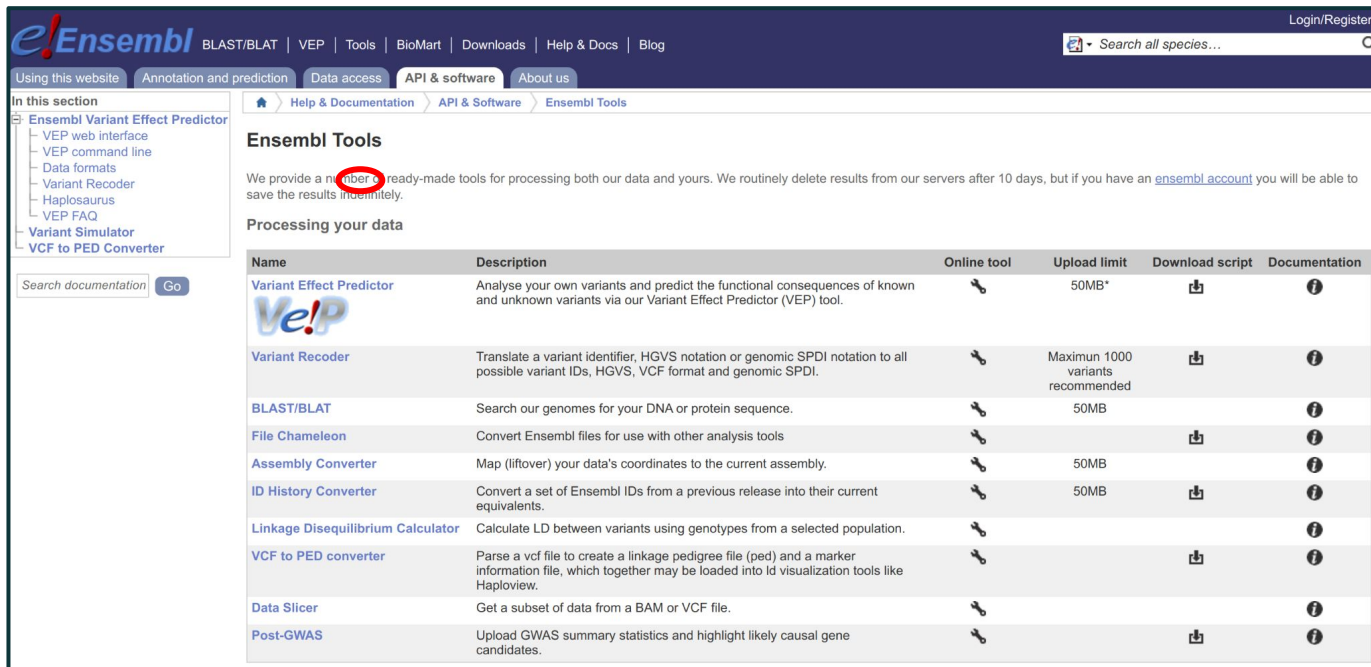
rows as

HTML

☐ 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="#">si:ch1073-178p5.3</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



**e!Ensembl** BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Using this website | Annotation and prediction | Data access | **API & software** | About us

Search all species...

**In this section**




























- Ensembl Variant Effect Predictor
  - VEP web interface
  - VEP command line
  - Data formats
  - Variant Recoder
  - HaploSaurus
  - VEP FAQ
- Variant Simulator
- VCF to PED Converter

Search documentation Go

## Ensembl Tools

We provide a **number** of ready-made tools for processing both our data and yours. We routinely delete results from our servers after 10 days, but if you have an [ensembl account](#) you will be able to save the results indefinitely.

### Processing your data

Name	Description	Online tool	Upload limit	Download script	Documentation
<a href="#">Variant Effect Predictor</a> 	Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool.		50MB*		
<a href="#">Variant Recoder</a>	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		
<a href="#">BLAST/BLAT</a>	Search our genomes for your DNA or protein sequence.		50MB		
<a href="#">File Chameleon</a>	Convert Ensembl files for use with other analysis tools				
<a href="#">Assembly Converter</a>	Map (liftover) your data's coordinates to the current assembly.		50MB		
<a href="#">ID History Converter</a>	Convert a set of Ensembl IDs from a previous release into their current equivalents.		50MB		
<a href="#">Linkage Disequilibrium Calculator</a>	Calculate LD between variants using genotypes from a selected population.				
<a href="#">VCF to PED converter</a>	Parse a vcf file to create a linkage pedigree file (ped) and a marker information file, which together may be loaded into Id visualization tools like Haploview.				
<a href="#">Data Slicer</a>	Get a subset of data from a BAM or VCF file.				
<a href="#">Post-GWAS</a>	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

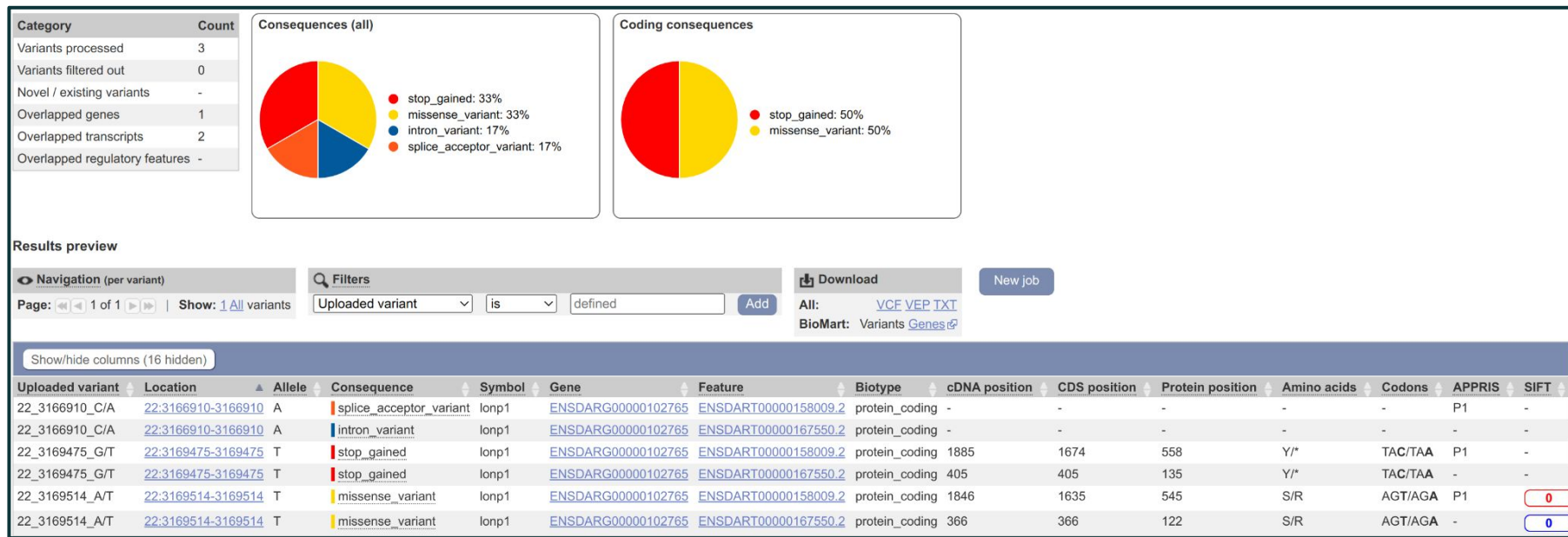
22 3169514 3169514 A/T 1

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



# 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:

Zebrafish (Danio rerio) ▼

Assembly mapping:

GRCz10 -> GRCz11 ▼

Name for this job (optional):

Input file format:

BED ▼

Either paste data:

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

Or upload file:

Choose file No file chosen

Or provide file URL:

Run ›

# Assembly Converter

**Assembly Converter** ?

New jobClear form

This online tool currently uses CrossMap to convert between different assembly mapping formats. CrossMap also discards metadata in files, so track details are lost.

Species:

Assembly mapping:

Name for this job (optional):

Input file format:

Either paste data:

**Input:**  
22 3144711 3144711 sa39354  
22 3145013 3145013 sa43743

**Output:**  
22 3161984 3161984 sa39354  
22 3162286 3162286 sa43743

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

# UCSC In-Silico PCR

- Fast search
- [genome.ucsc.edu/](https://genome.ucsc.edu/)

[Home](#) [Genomes](#) [Genome Browser](#) [Tools](#) [Mirrors](#) [Downloads](#) [My Data](#) [Projects](#) [Help](#) [About Us](#)

## UCSC In-Silico PCR

Genome:  
Zebrafish

Assembly:  
May 2017 (GRCz11/danRer11)

Forward Primer:  
CCCGGGGAGCAGTTGA

Reverse Primer:  
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  CCCATGAGTGGCTCCTAAAGCAGCTGC  
TtACAGATTGATGATGCATGAAATGGGgggtggccaggggtggggggtga
```

# UCSC In-Silico PCR

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

[illegible]

# UCSC & Ensembl Differences

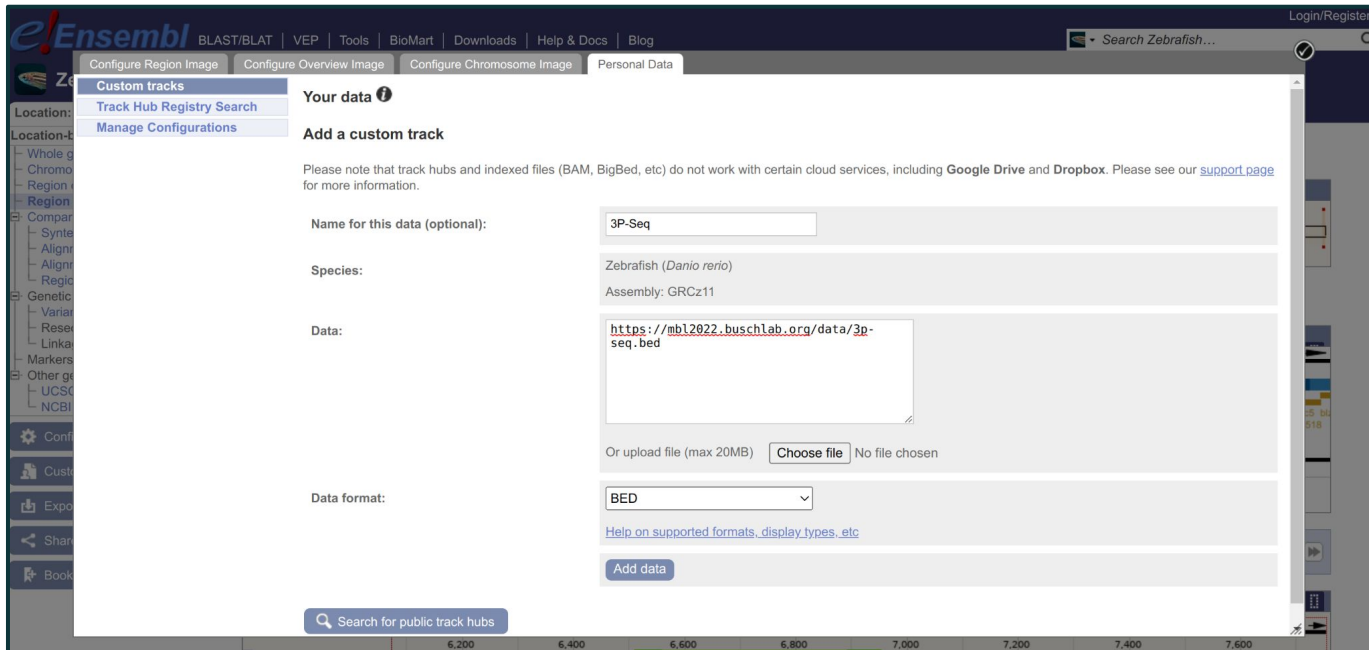
- **Ensembl:** 1  
**UCSC:** chr1
- **Ensembl:** 1-based coordinates (bases numbered)  
**UCSC:** 0-based coordinates (numbers between bases)

chr1		T		A		C		G		T		C		A	
1-based		1		2		3		4		5		6		7	
0-based	0		1		2		3		4		5		6		7

- 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. On the left, a sidebar contains a tree view with categories like 'Location', 'Genetic', and 'Markers'. The 'Custom tracks' section is highlighted in the sidebar. The main content area is titled 'Your data' and contains a form for adding a custom track. The form includes fields for 'Name for this data (optional):' (filled with '3P-Seq'), 'Species:' (filled with 'Zebrafish (Danio rerio)'), 'Assembly:' (filled with 'GRCz11'), and 'Data:' (filled with the URL 'https://mbl2022.buschlab.org/data/3p-seq.bed'). Below the 'Data' field, there is an option to 'Or upload file (max 20MB)' with a 'Choose file' button and the text 'No file chosen'. The 'Data format:' dropdown is set to 'BED'. A link 'Help on supported formats, display types, etc' is provided. At the bottom of the form is an 'Add data' button. The Ensembl logo and navigation links are visible at the top of the page.

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search Zebrafish...

Configure Region Image | Configure Overview Image | Configure Chromosome Image | Personal Data

**Custom tracks**

Track Hub Registry Search

Manage Configurations

**Your data**

**Add a custom track**

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.

Name for this data (optional): 3P-Seq

Species: Zebrafish (*Danio rerio*)

Assembly: GRCz11

Data: <https://mbl2022.buschlab.org/data/3p-seq.bed>

Or upload file (max 20MB) Choose file No file chosen

Data format: BED

[Help on supported formats, display types, etc](#)

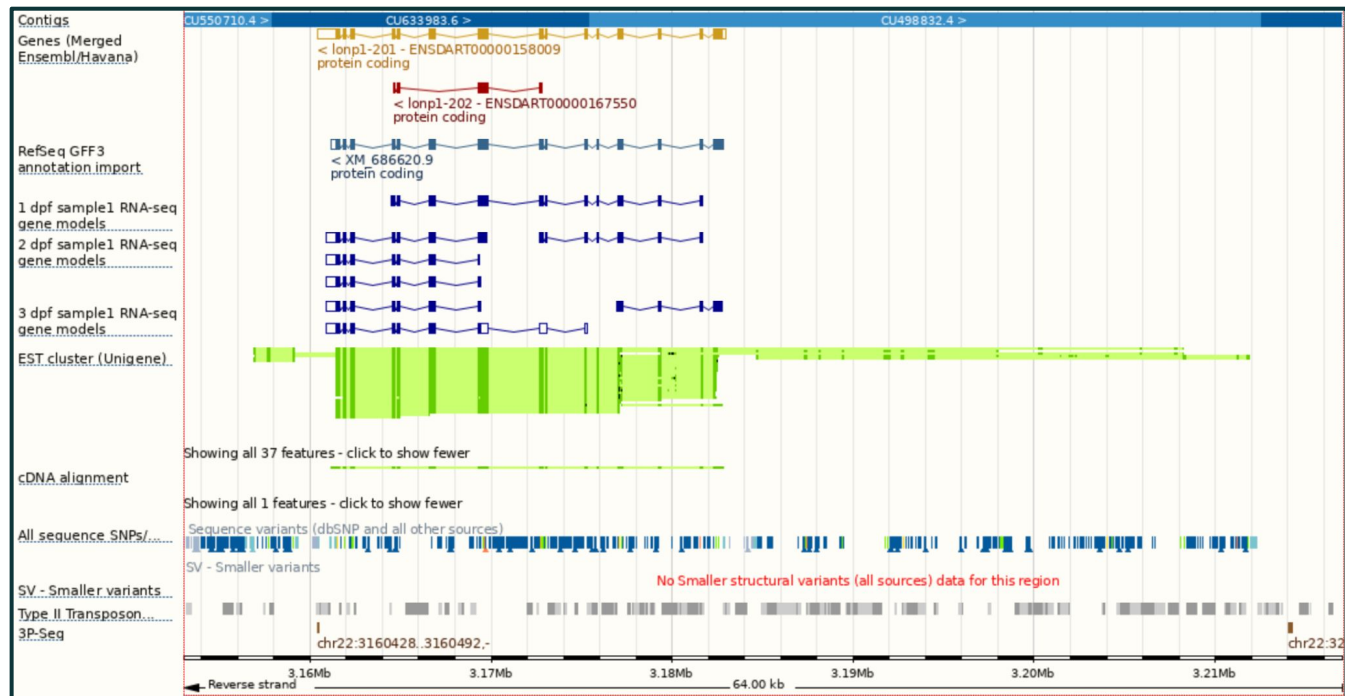
Add data

Search for public track hubs

- 24 hpf 3P-Seq data from Bartel lab

# Custom Tracks

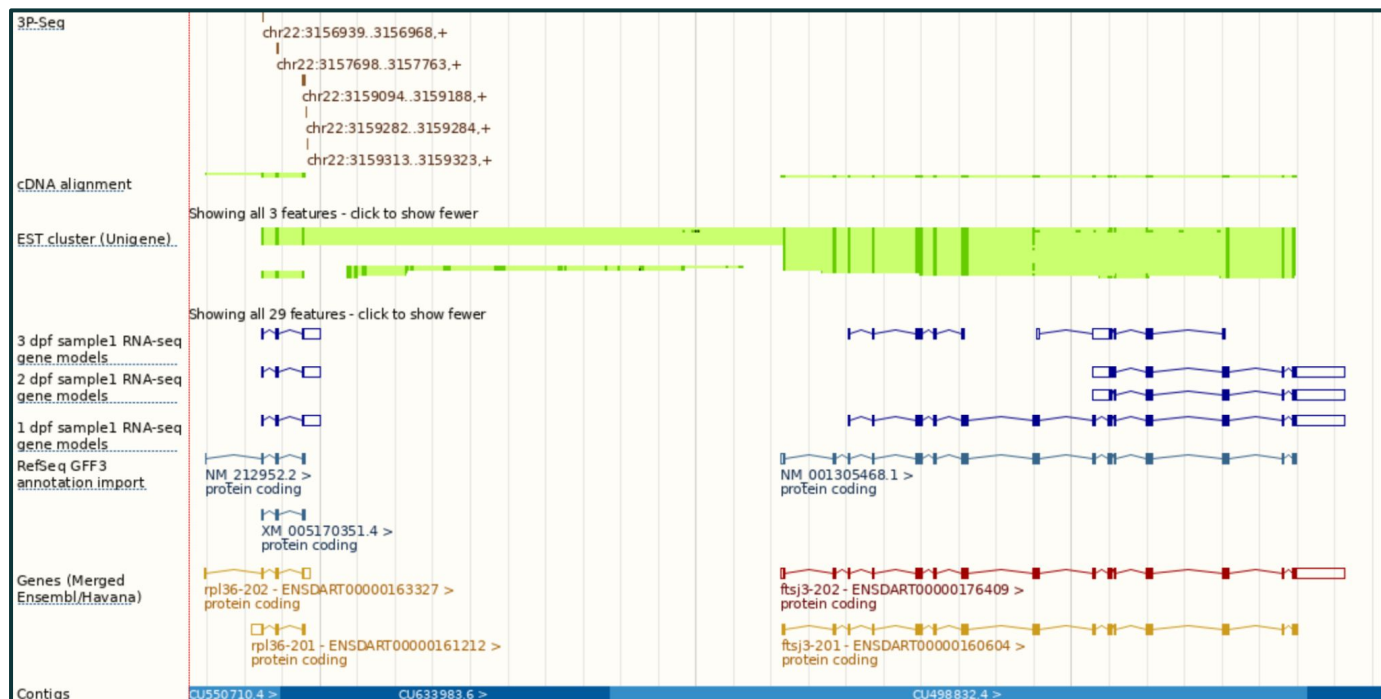
- Go to "22:3153000-3217000" (reverse strand)





# Custom Tracks

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

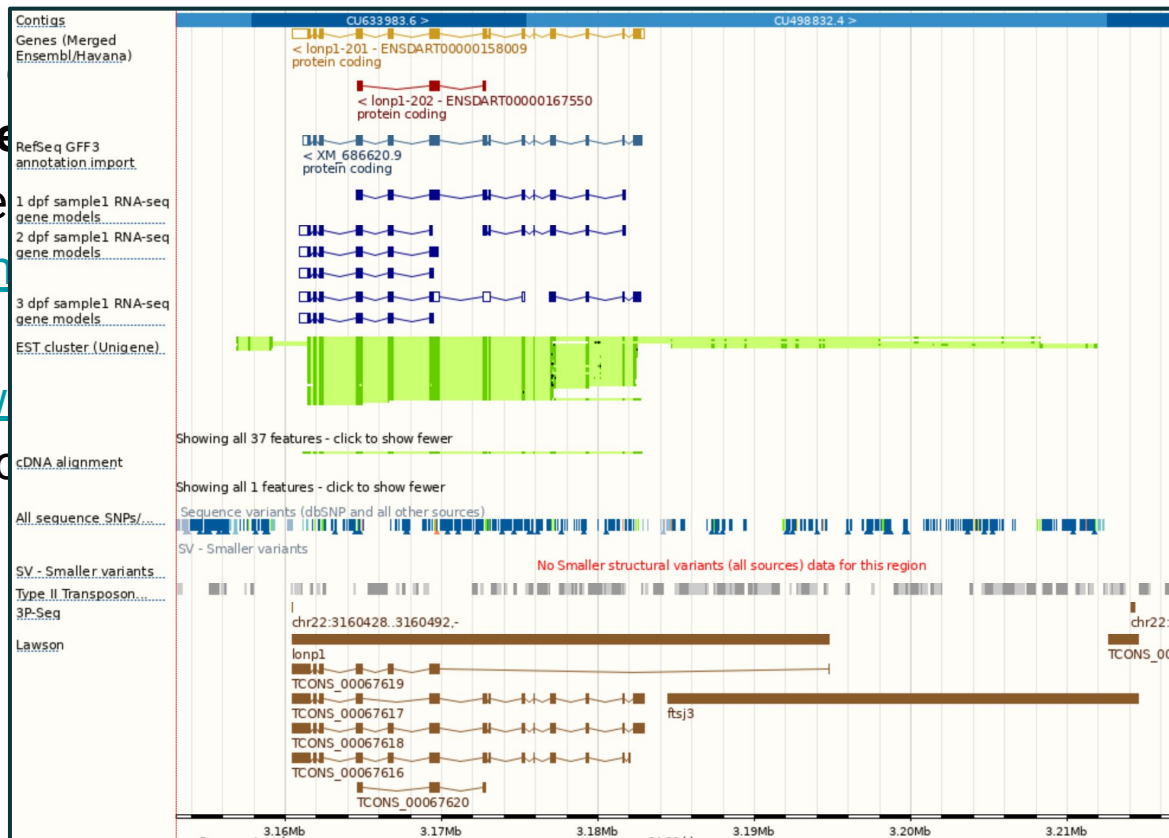


# Custom Tracks - Lawson Lab Annotation

- 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.umich.edu](http://www.umich.edu)
- Add:
- <https://www.ncbi.nlm.nih.gov/assembly/chr22.3160428.3160492-1>
- Large, so



otation for  
nes"

ome/

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](https://mbl2022.buschlab.org)

# Thank You!

Any questions?

