Genome Literacy Workshop

Elisabeth Busch-Nentwich
& Pavol Kramár

Queen Mary University of London, UK

Why genome literacy?

- Saves time and effort
  - Ensembl collates a lot of published information
  - Knowing how to find that can (at least partially) replace trawling through papers
- Better experimental design for gene editing experiments
  - Understand gene structure
  - Understand splice variants and expression behaviour
- Download data without coding
  - Sequence, phenotypes, gene ontology annotations
- Customise genome browser
  - Switch on relevant tracks, upload your own data

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

Workshop structure

- Lecture sections
  - Introduce background information
  - Show examples
  - Slides for future reference
- Small exercises
  - In lecture sections
  - I will point out when to switch to browser
- Large exercises
  - Designed to be more challenging
  - Cover all tools and features we introduce
- Your questions
  - Email me so that we can work through them on Day 2

Workshop structure

Day 1

- Part 1
  - Zebrafish genome basics
  - Ensembl basics
  - Nomenclature (Gene names, IDs, etc...)
  - Assembly vs annotation
- Part 2
  - Configuring Ensembl tracks
  - Ensembl “Gene” view
  - Comparative genomics
- Homework
  - Email me your questions so that we can work through them on Day 2
  - Finish exercises

Day 2

- Part 3
  - BioMart
  - Other tools (Variant Effect Predictor, etc...)
  - Custom tracks
- Your questions

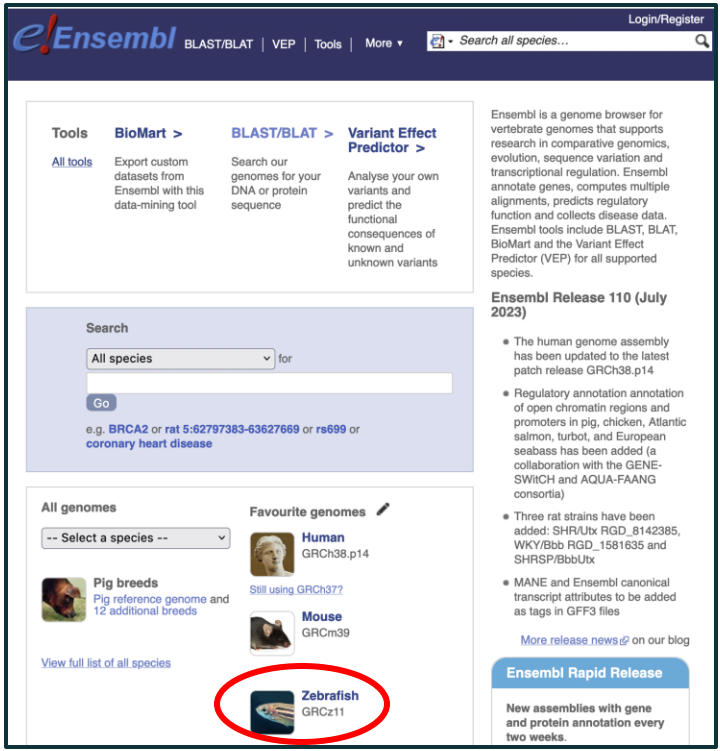
Part 1

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

Ensembl

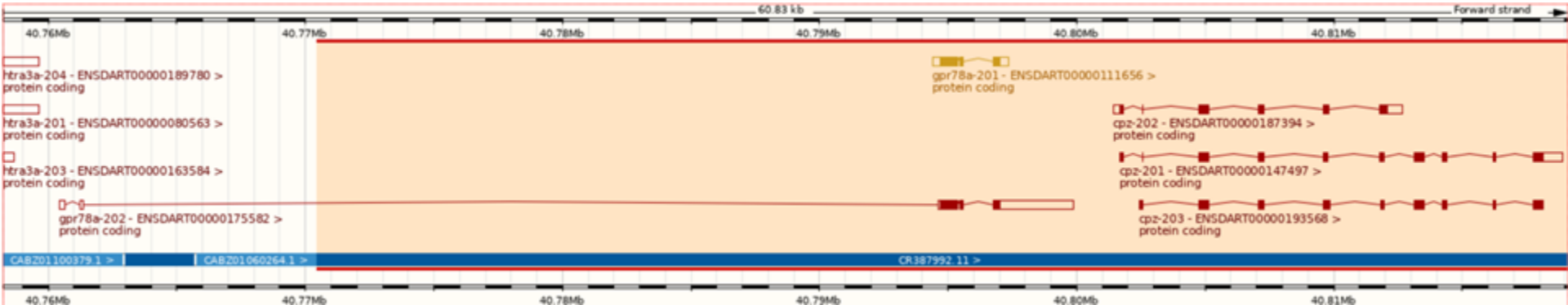
- Most examples from **Ensembl** (we are biased!)
- Probably most widely used genome browser amongst zebrafish researchers
- **Primary source of zebrafish annotation** (UCSC imports Ensembl annotation)

- Currently Ensembl version **110** (July 18th)
- New releases 3 or 4 times / year
- Zebrafish **annotation largely static** between releases
- But **naming and homology** updated (+ new functionality)



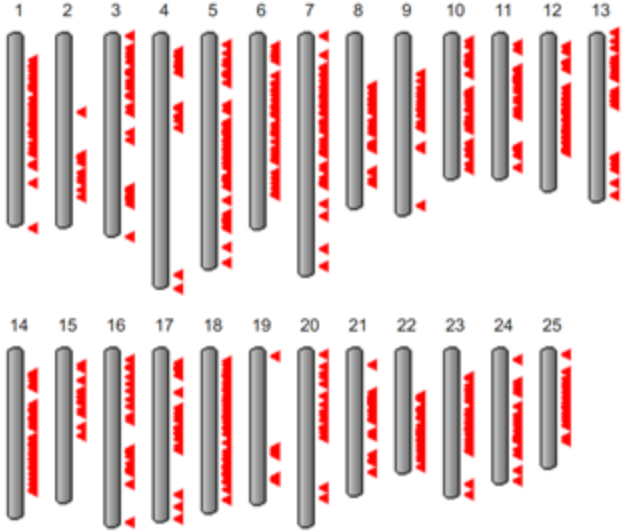
Zebrafish Genome

- **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



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

- When downloading sequence from Ensembl FTP site, "`toplevel`" includes alternative sequence, but "`primary_assembly`" doesn't and is probably what you want

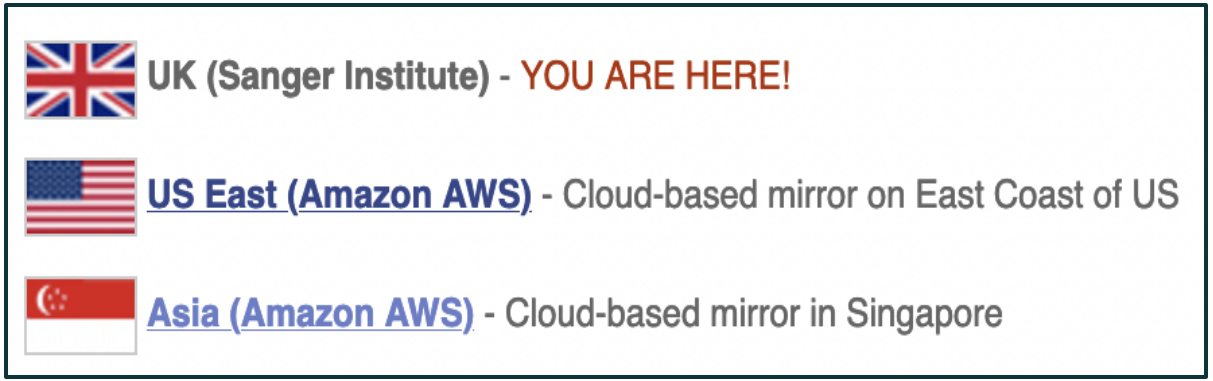
Older Assemblies

- Previous assemblies available in Ensembl **archives**: www.ensembl.org/info/website/archives/assembly.html
  - GRCz10 / danRer10: http://e91.ensembl.org/
  - Zv9 / danRer7: http://e77.ensembl.org/
  - Zv8 / danRer6: http://e54.ensembl.org/

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

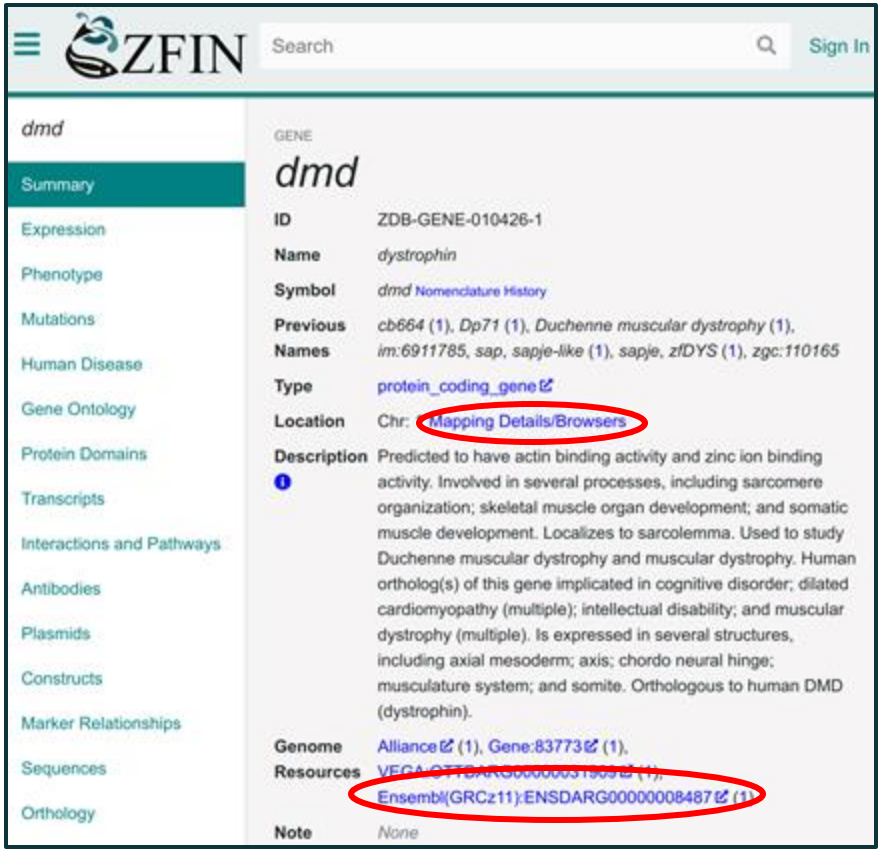
Ensembl Mirrors

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



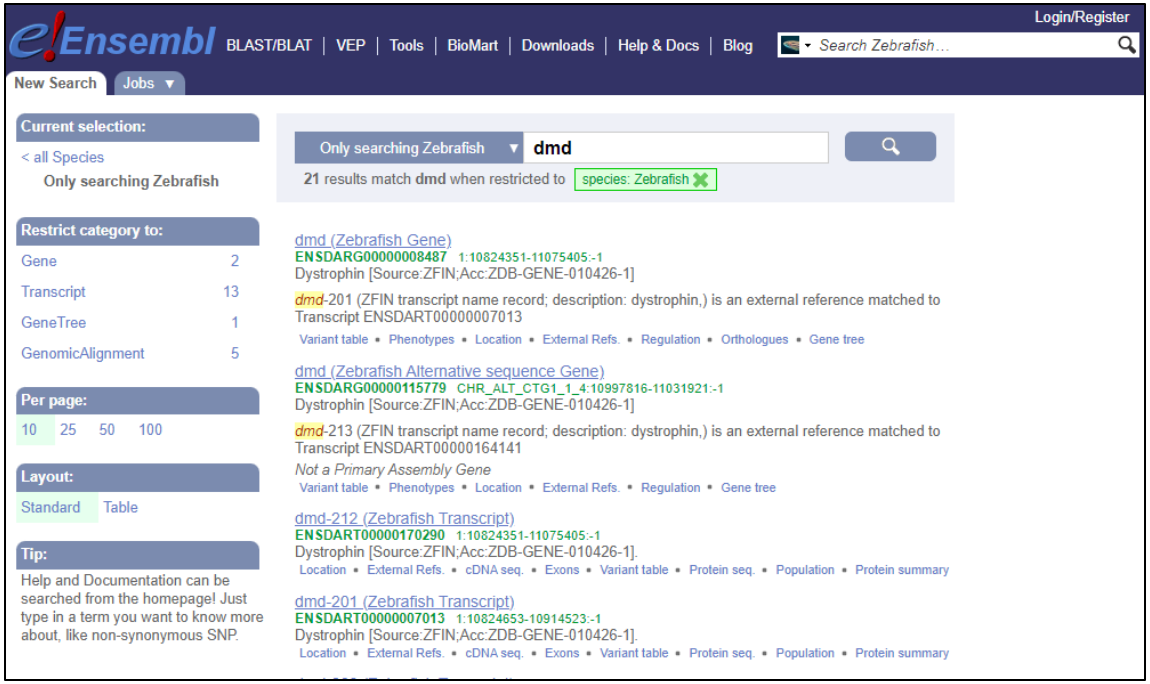
Finding Your Gene

- Follow link from **ZFIN**



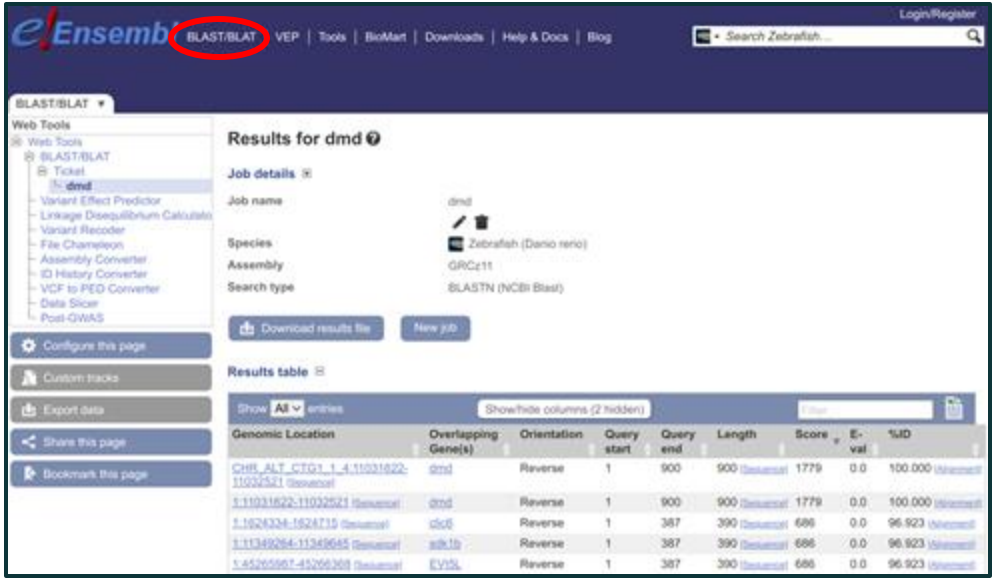
Finding Your Gene

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



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



Finding Your Gene

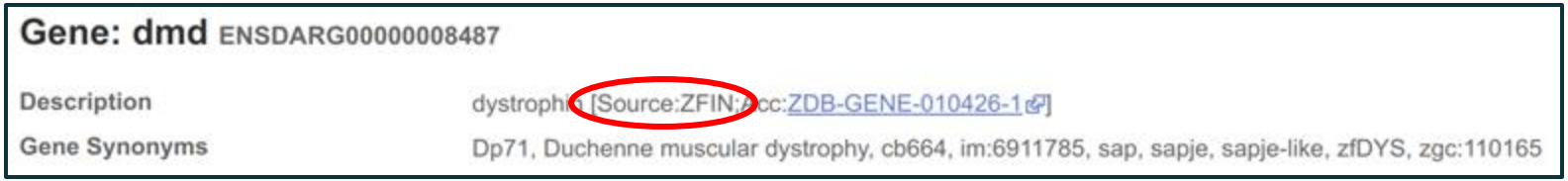
- Follow link from **ZFIN**
- **Search** by gene name on Ensembl (or old name or mutant name)
- Search using **BLAST** or **BLAT** on Ensembl
  - BLAT is faster
  - BLAST finds more distant alignments + alternative scaffolds
  - No BLAST/BLAT on Ensembl archive sites but can use BLAT on UCSC
- Check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

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Japanese medaka HdrR (Oryzias latipes)	1-to-many View Gene Tree	dmd (ENSORLG00000020638) Compare Regions (2:208,119-221,155:1) View Sequence Alignments	84.52 %	88.22 %	0	95.81	Yes
Lumpfish (Cyclopterus lumpus)	1-to-many View Gene Tree	dmd (ENSCLMG00005009931) Compare Regions (2:5,248,684-5,281,983:-1) View Sequence Alignments	82.21 %	82.49 %	0	95.20	Yes
Lyretail cichlid (Neolamprologus brichardi)	1-to-1 View Gene Tree	dmd (ENSNBRG00000015200) Compare Regions (JH422367.1:2,004,027-2,028,054:-1) View Sequence Alignments	87.34 %	89.39 %	0	96.66	Yes
Makobe Island cichlid (Pundamilia nyererei)	1-to-1 View Gene Tree	dmd (ENSPNYG00000022641) Compare Regions (JH419417.1:620,205-712,305:-1) View Sequence Alignments	45.32 %	89.56 %	0	96.75	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, but other databases are also used, e.g. HGNC (HUGO* Gene Nomenclature Committee) and miRBase
- 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)

*HUGO: Human Genome Organization



Stable IDs

- Best to use stable IDs
- e.g. **ENS**DARG00000028213 (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. ENSDAR**G**00000028213 (ttn.2 or ttna)

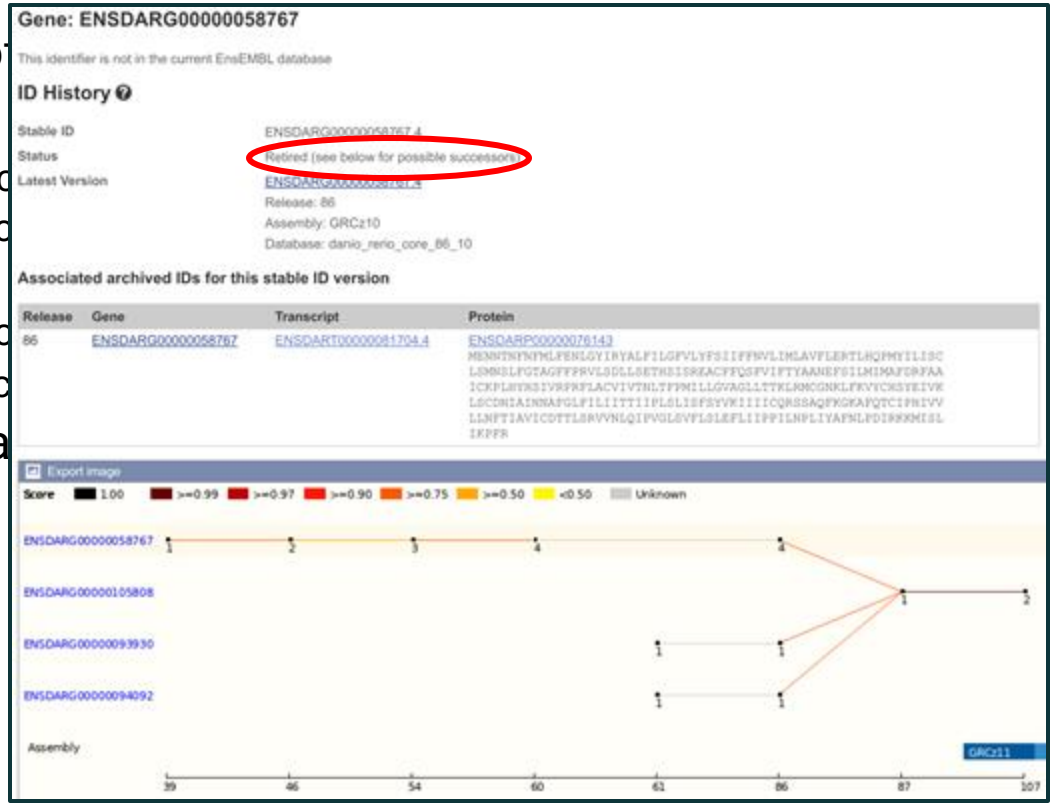
- **ENS** = Ensembl
- **DAR** = Danio rerio
- **G** = Gene (also T for Transcript, P for Peptide and E for exon)

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 anno
- Stable IDs have **versions**, e.g.
  - Version number of **ENSDARG** inc
  - Version number of **ENSDART** inc change
  - Version number of **ENSDARP** inc
  - Version number of **ENSDARE** inc
- Can also be **removed**, e.g. sea



Stable IDs

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

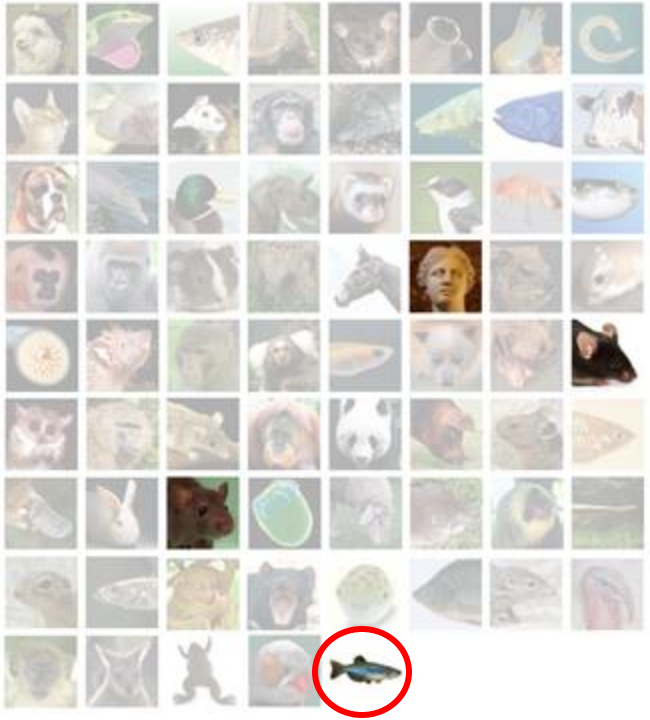
Stable IDs

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

Mini-exercise: Check whether your favourite genes have had stable ID changes. What's the ENSDARG version number? (Hint: look for your gene's ID history)

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



Manual Annotation

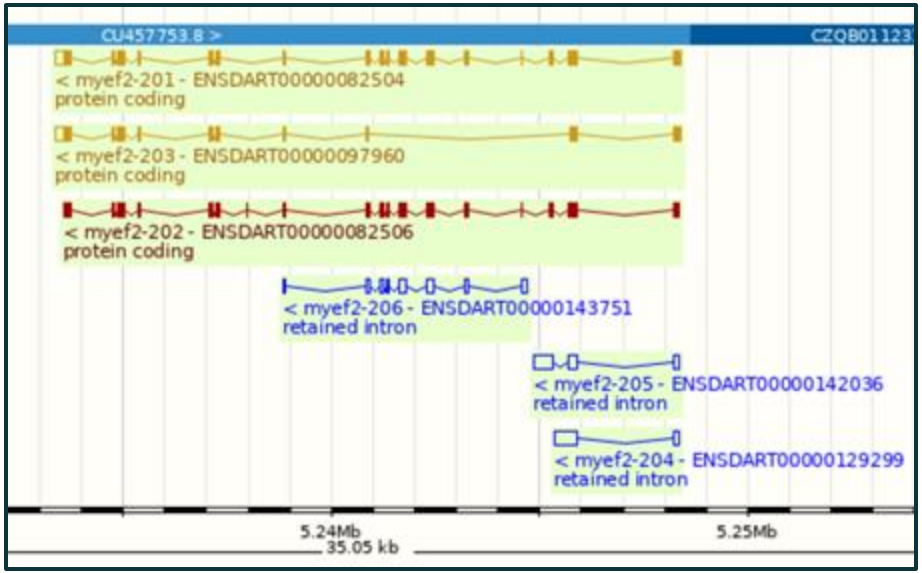
- **Gold** standard
- Uses information from databases and publications
- More accurate for tricky areas:
  - e.g. UTRs, splice sites, single exon transcripts
- **Slower** and more expensive
- Thorough, but leads to inclusion of transcripts that may not be representative (e.g. low expression)
- Only clones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

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

Merged Annotation

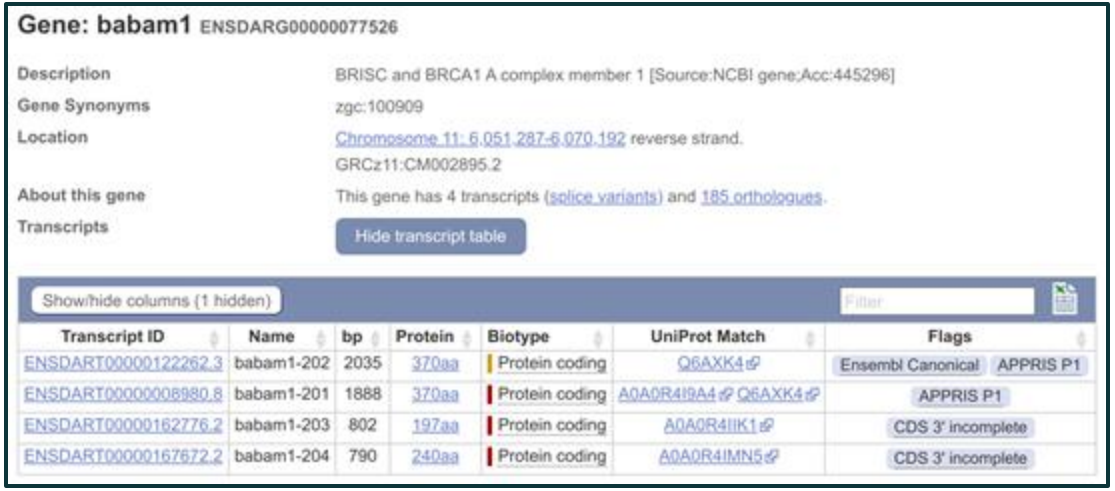
- Golden: **Identical** manual and automatic annotation
- Red: **Protein-coding** transcript from automatic annotation
- Blue: **Non-coding** transcript
- Filled box: **Coding exon**
- Non-filled box: **Non-coding** exon



- 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

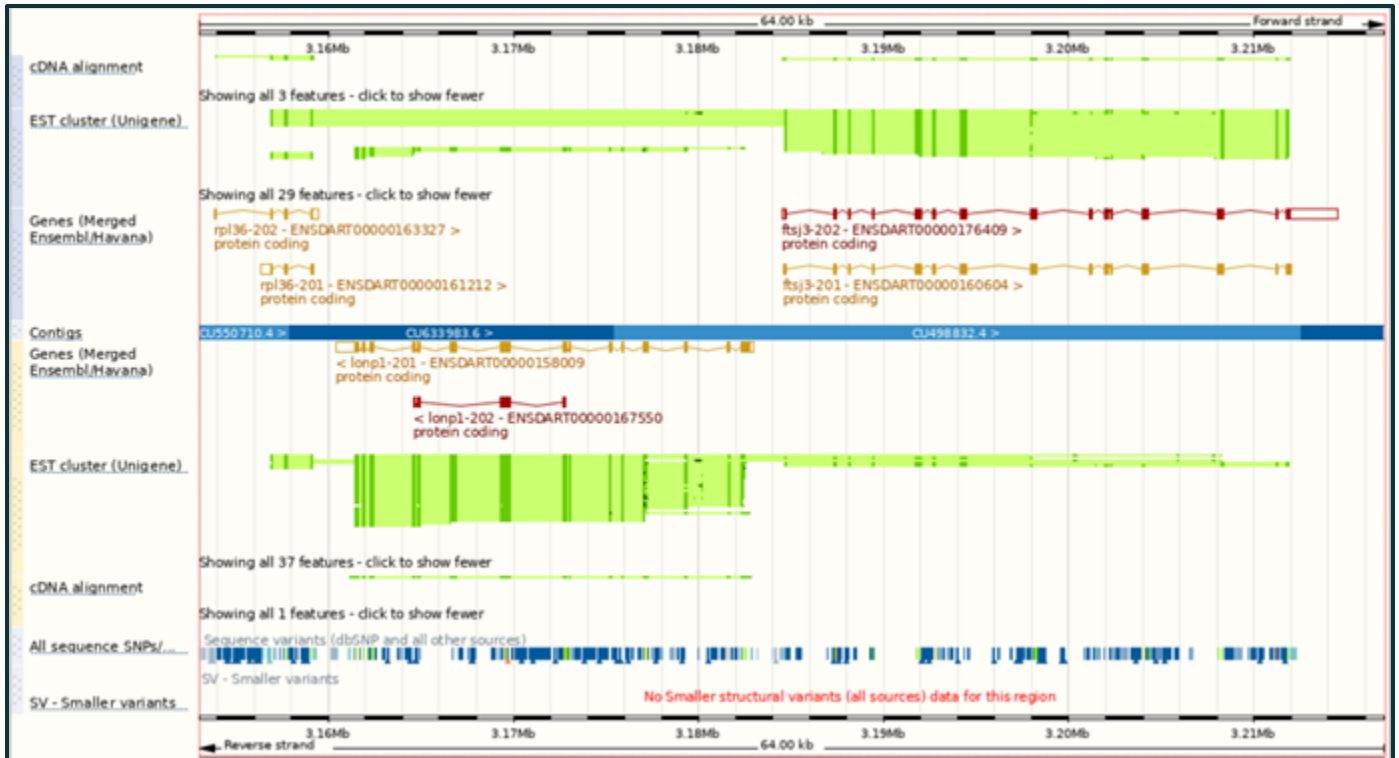
Description	BRISC and BRCA1 A complex member 1 [Source:NCBI gene;Acc:445296]
Gene Synonyms	zgc:100909
Location	Chromosome 11: 6,051,287-6,070,192 reverse strand. GRCz11:CM002895.2
About this gene	This gene has 4 transcripts (splice variants) and 185 orthologues.
Transcripts	Hide transcript table

Show/hide columns (1 hidden) Filter

Transcript ID	Name	bp	Protein	Biotype	UniProt Match	Flags
ENSDART00000122262.3	babam1-202	2035	370aa	Protein coding	Q6AXK4	Ensembl Canonical APPRIS P1
ENSDART00000008980.8	babam1-201	1888	370aa	Protein coding	A0A0R4I9A4 Q6AXK4	APPRIS P1
ENSDART00000162776.2	babam1-203	802	197aa	Protein coding	A0A0R4IIK1	CDS 3' incomplete
ENSDART00000167672.2	babam1-204	790	240aa	Protein coding	A0A0R4IMN5	CDS 3' incomplete

"Region in detail" Demo

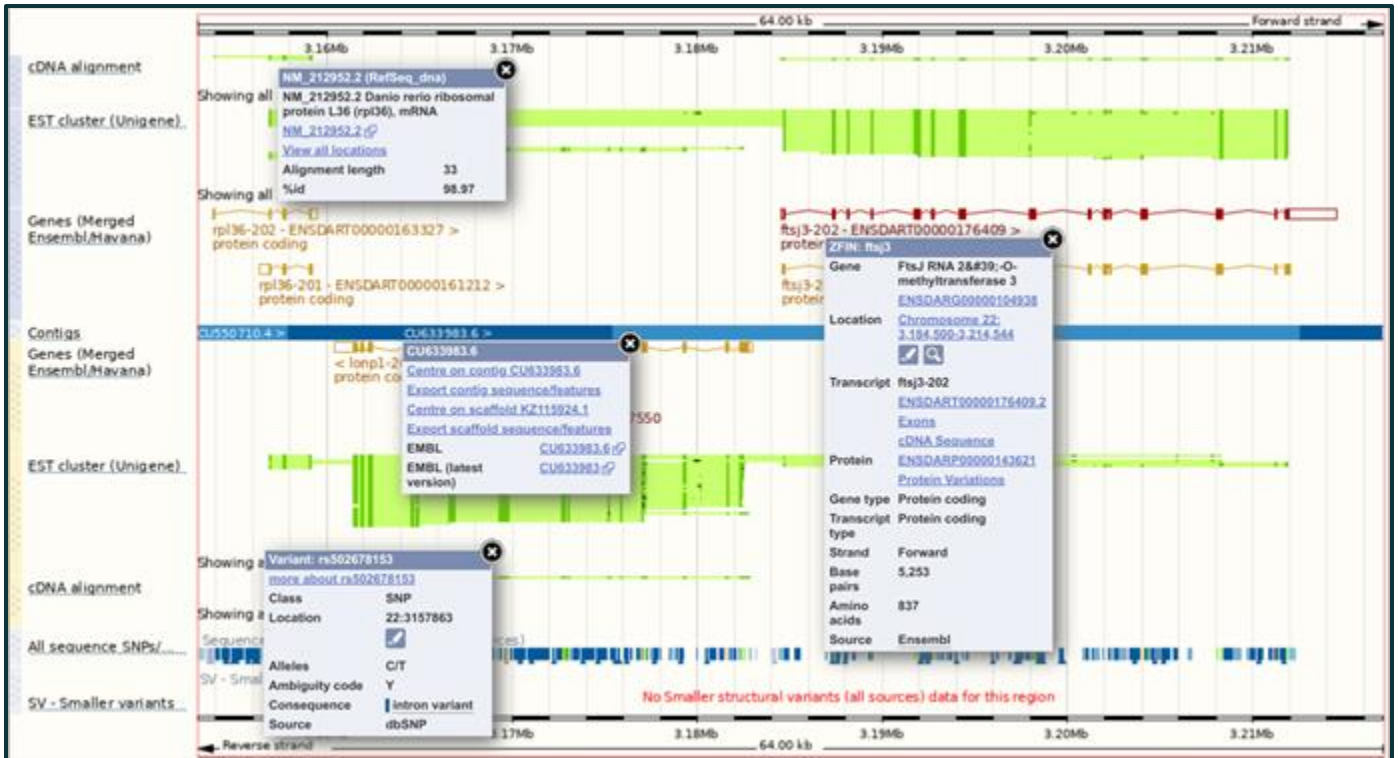
- Go to "22:3153000-3217000"



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

"Region in detail" Demo

- Go to "22:3153000-3217000"



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

Quiz time!

Exercise 1

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

Part 1 summary

- Zebrafish genome basics
- Ensembl basics
- Nomenclature (Gene names, IDs, etc...)
- Assembly vs annotation
- Learning outcomes:
  - Understanding gene features for experiments, e.g. CRISPR, ISHs
  - Distinguish between annotation and assembly, gene versions
  - Being able to switch between releases
  - Judging which transcripts/ gene features are reliable
- Any questions?

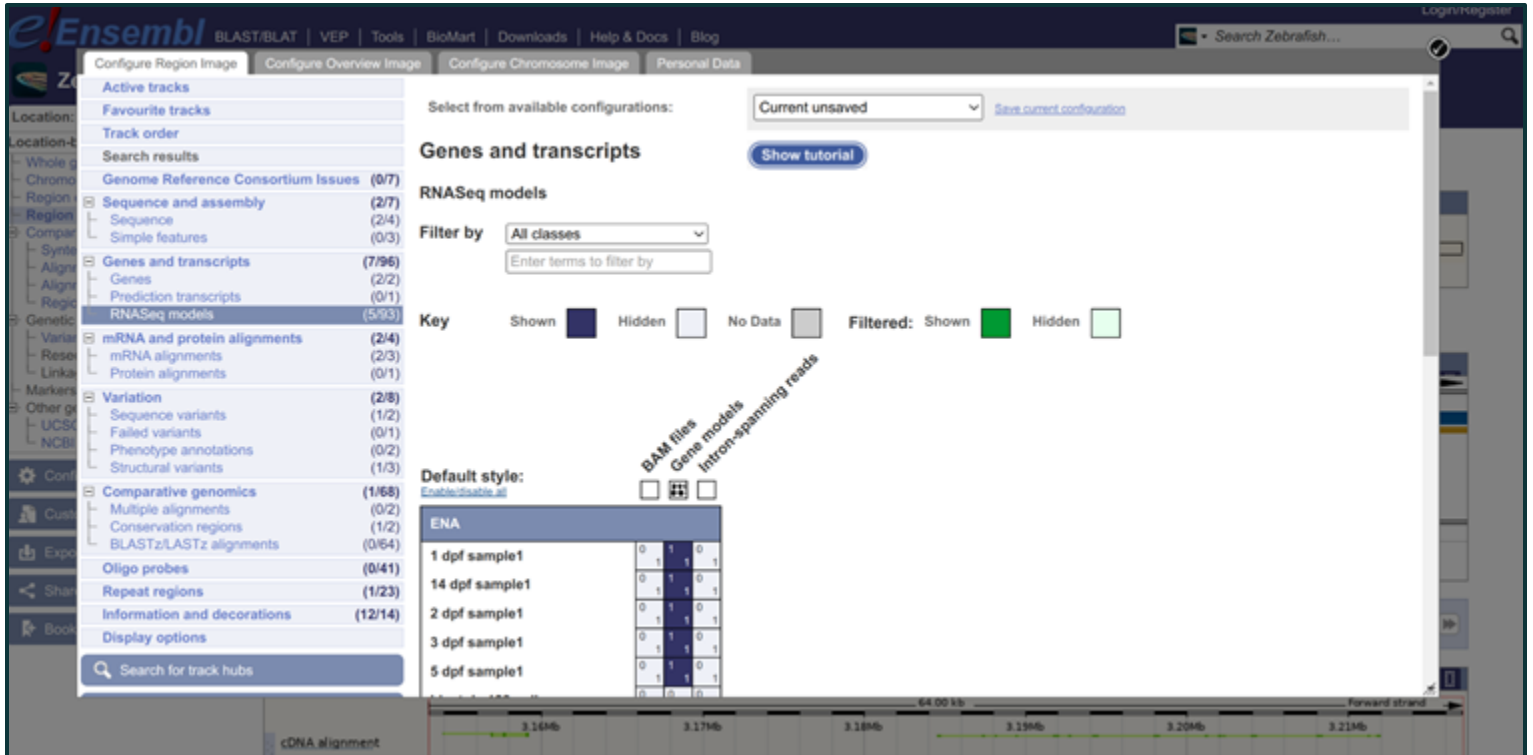
Part 2

- 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”

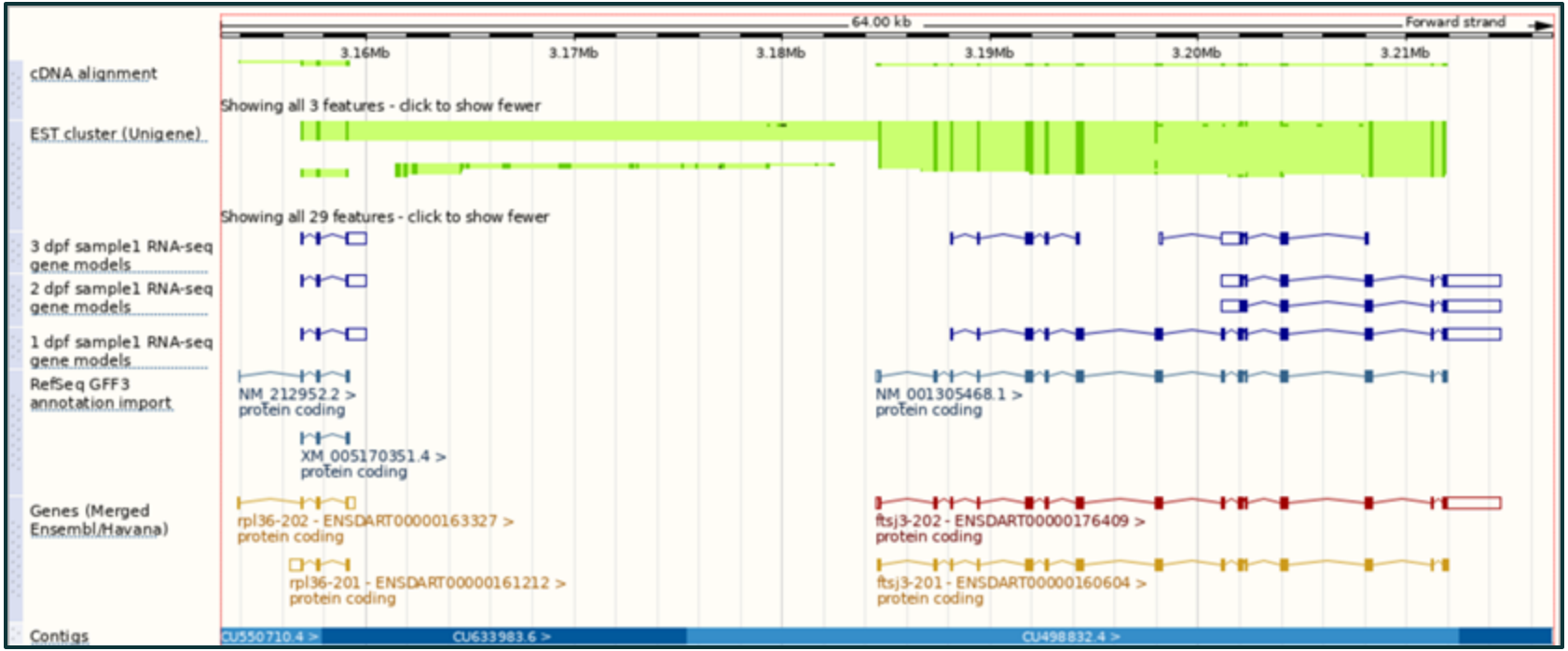


RefSeq Aside

- 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

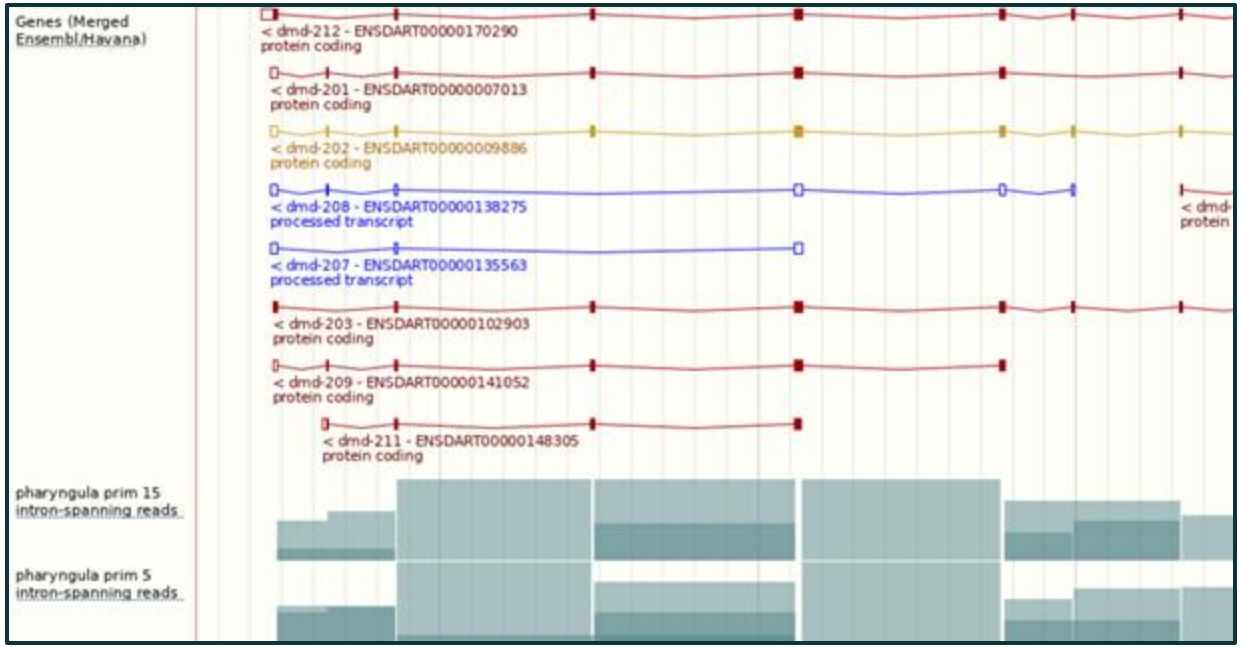
"Configure this page" Demo

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



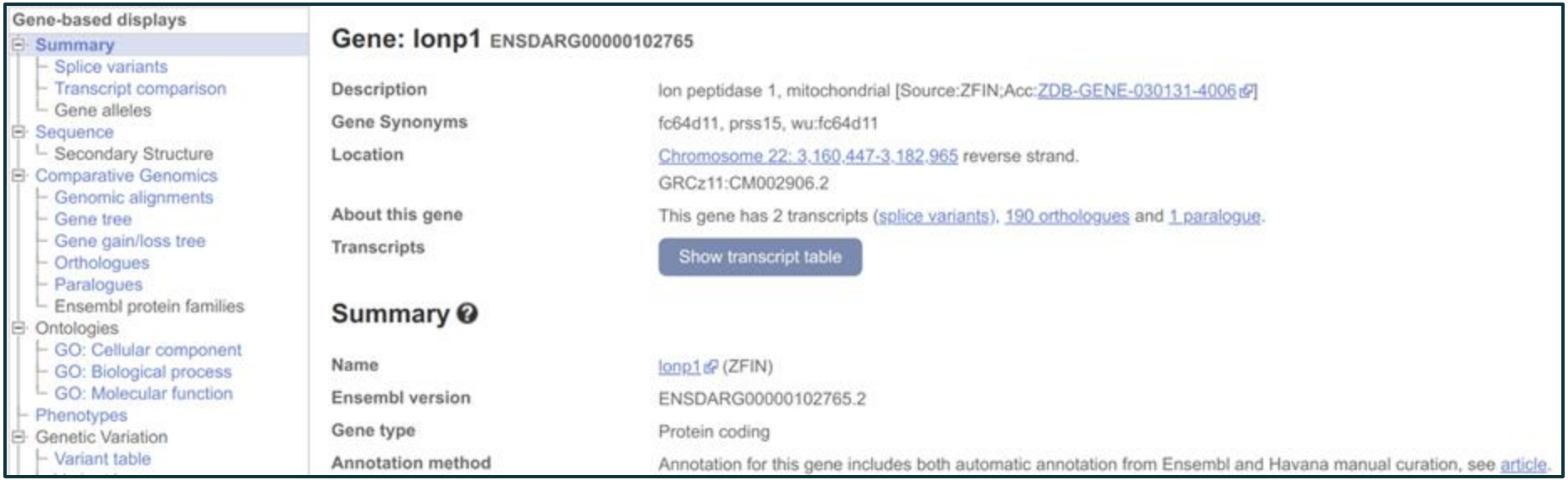
"Configure this page" Demo

- 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”



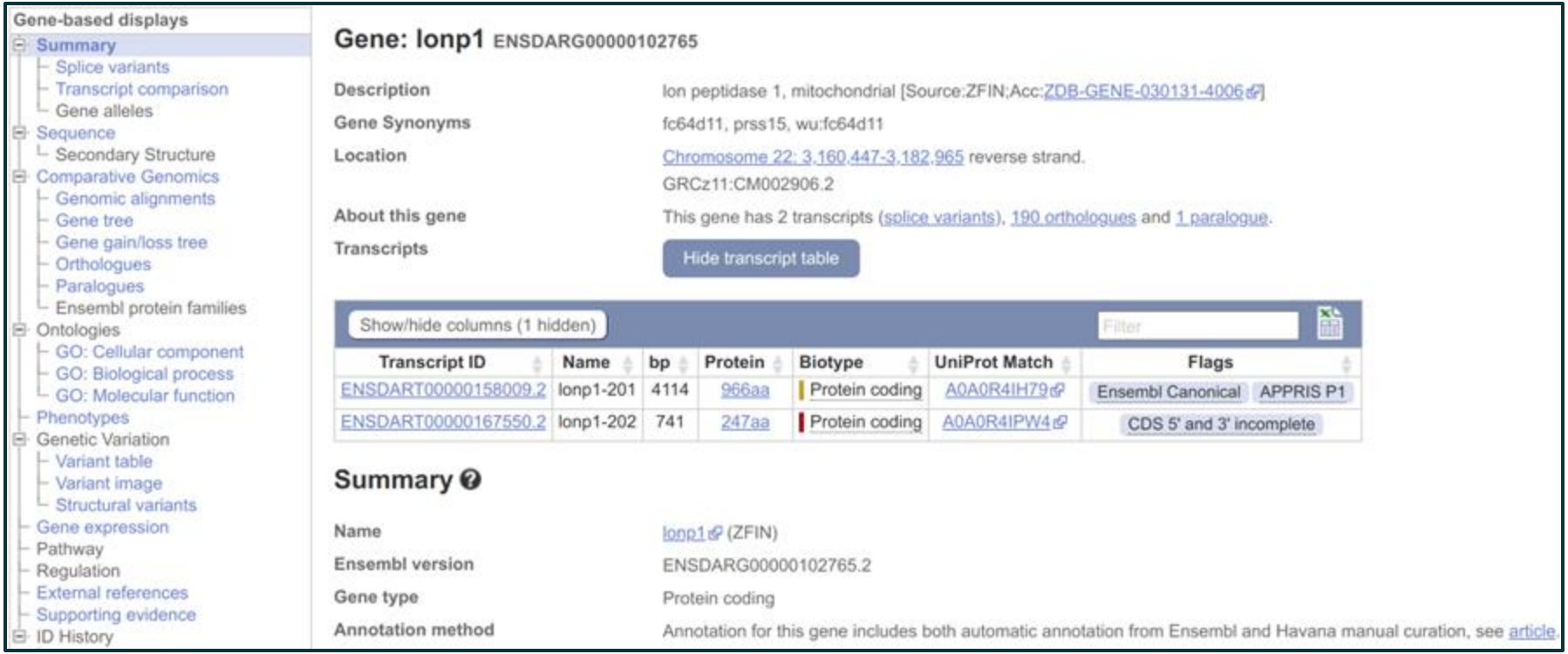
"Gene" Demo - Summary

- Go to ENSDARG00000102765



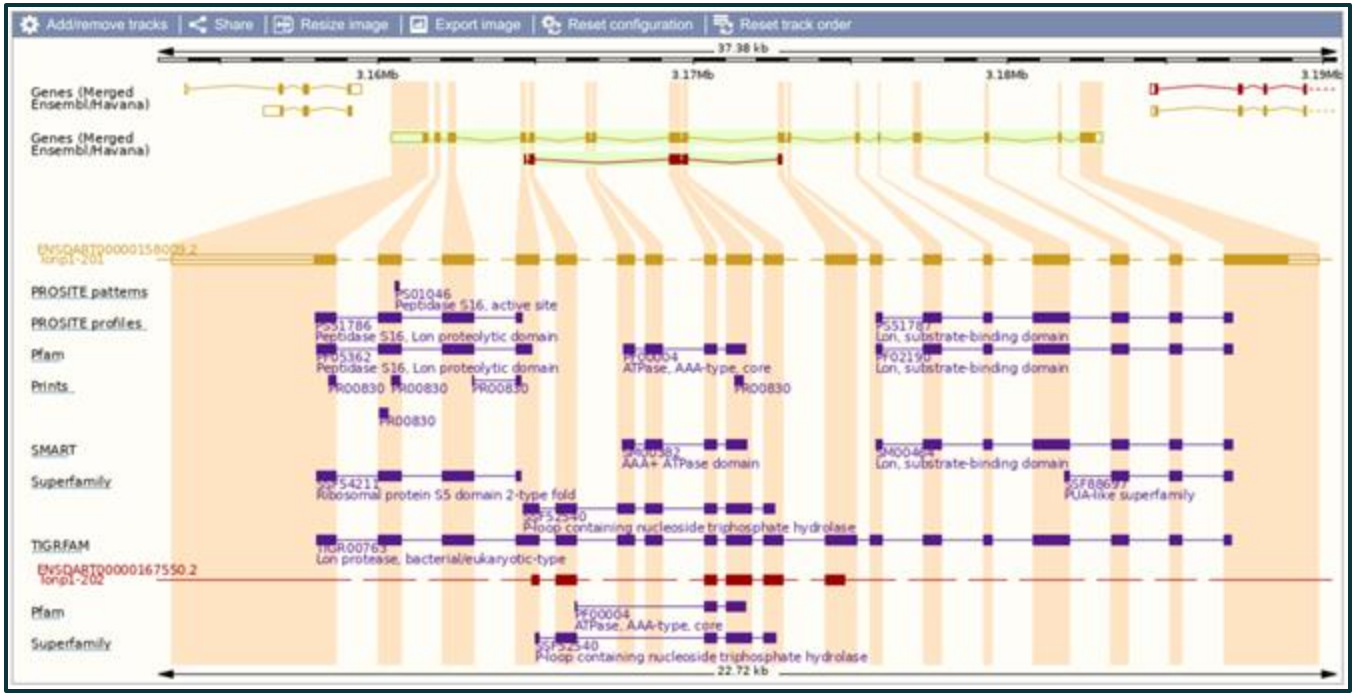
"Gene" Demo - Transcript Table

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



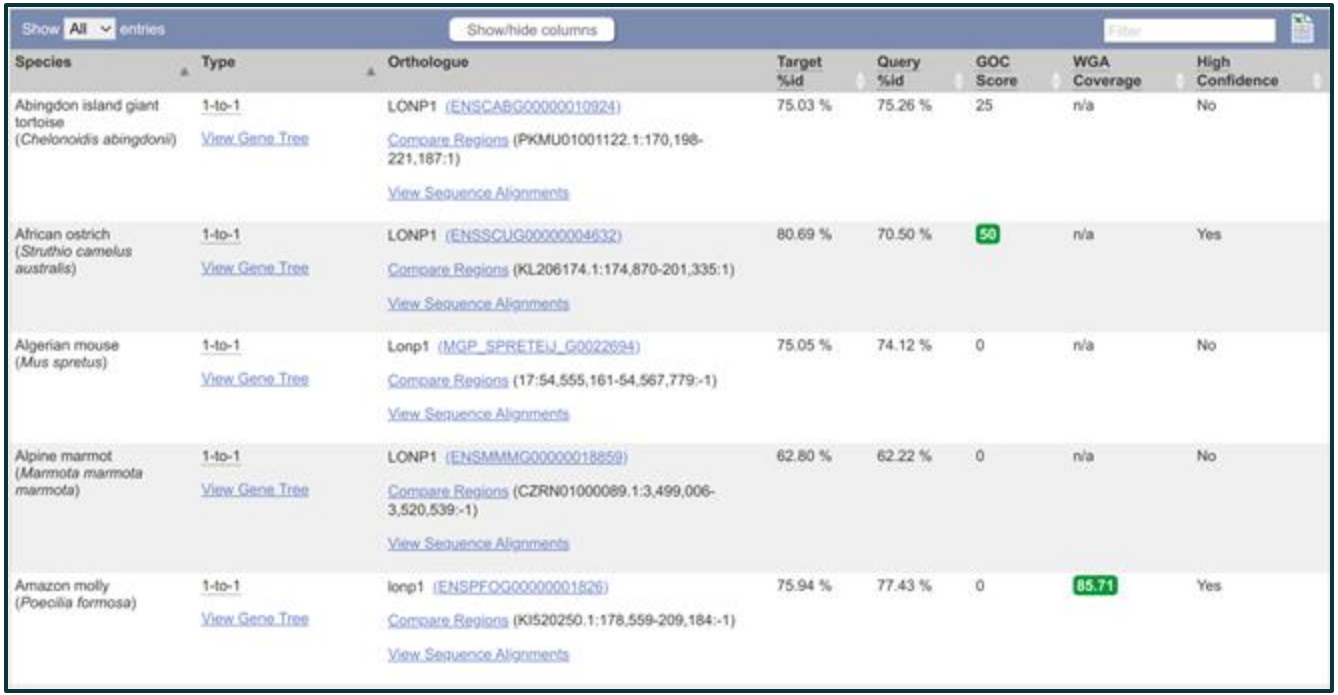
"Gene" Demo - Splice Variants

- Go to ENSDARG00000102765 and click on “Splice variants”



"Gene" Demo - Orthologues

- Go to ENSDARG00000102765 and click on “Orthologues”

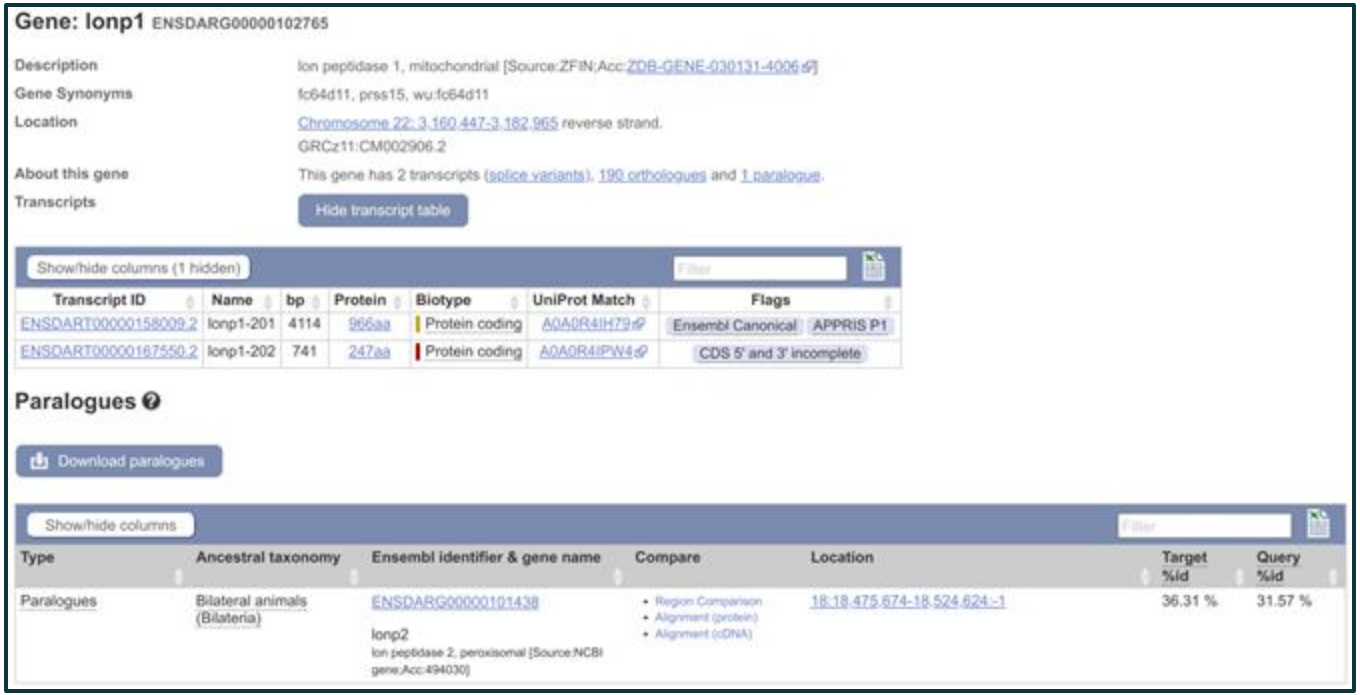


Show All entries | Show/hide columns | Filter

Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Abingdon island giant tortoise (*Chelonoidis abingdonii*)	1-to-1 View Gene Tree	LONP1 (ENSCABG00000010924) Compare Regions (PKMU01001122.1:170,198-221,187:1) View Sequence Alignments	75.03 %	75.26 %	25	n/a	No
African ostrich (*Struthio camelus australis*)	1-to-1 View Gene Tree	LONP1 (ENSSCUG00000004632) Compare Regions (KL206174.1:174,870-201,335:1) View Sequence Alignments	80.69 %	70.50 %	50	n/a	Yes
Algerian mouse (*Mus spretus*)	1-to-1 View Gene Tree	Lonp1 (MGP_SPRETEiJ_G0022694) Compare Regions (17:54,555,161-54,567,779:-1) View Sequence Alignments	75.05 %	74.12 %	0	n/a	No
Alpine marmot (*Marmota marmota marmota*)	1-to-1 View Gene Tree	LONP1 (ENSMMMG00000018859) Compare Regions (CZRN01000089.1:3,499,006-3,520,539:-1) View Sequence Alignments	62.80 %	62.22 %	0	n/a	No
Amazon molly (*Poecilia formosa*)	1-to-1 View Gene Tree	lonp1 (ENSPFOG00000001826) Compare Regions (KI520250.1:178,559-209,184:-1) View Sequence Alignments	75.94 %	77.43 %	0	85.71	Yes

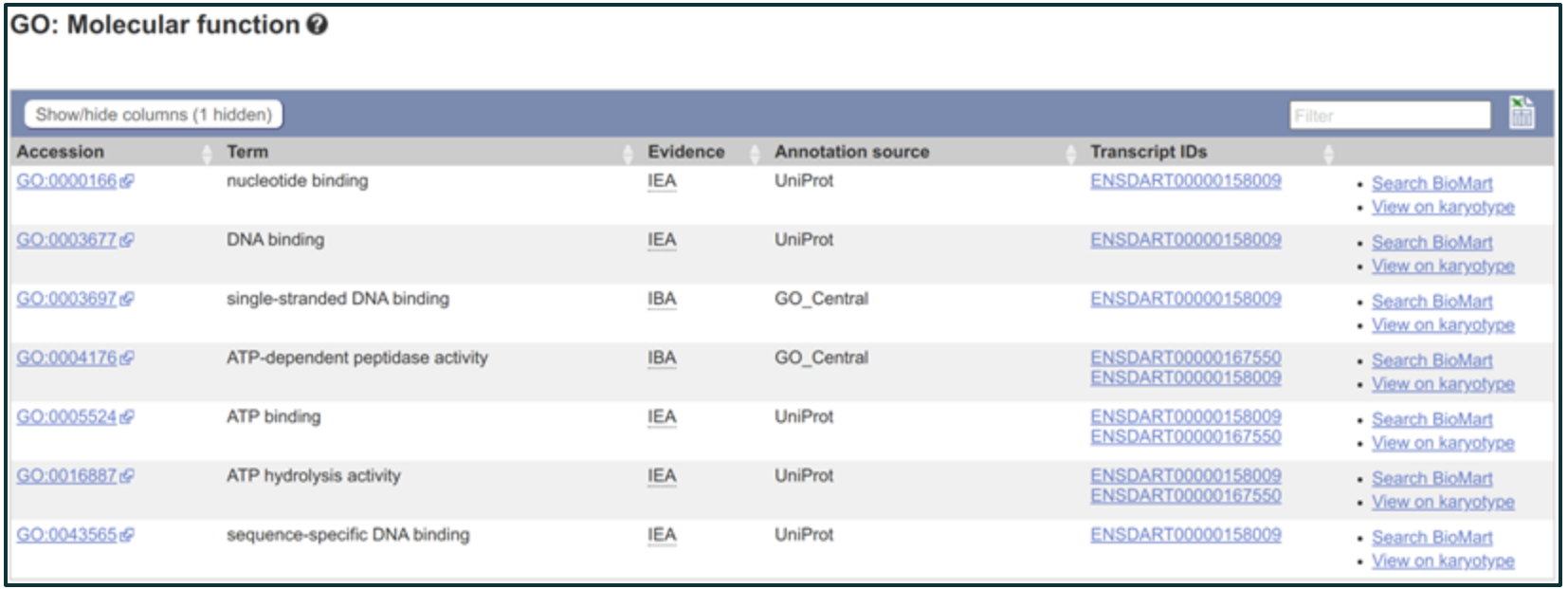
"Gene" Demo - Paralogues

- Go to ENSDARG00000102765 and click on “Paralogues”



"Gene" Demo - GO Terms

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



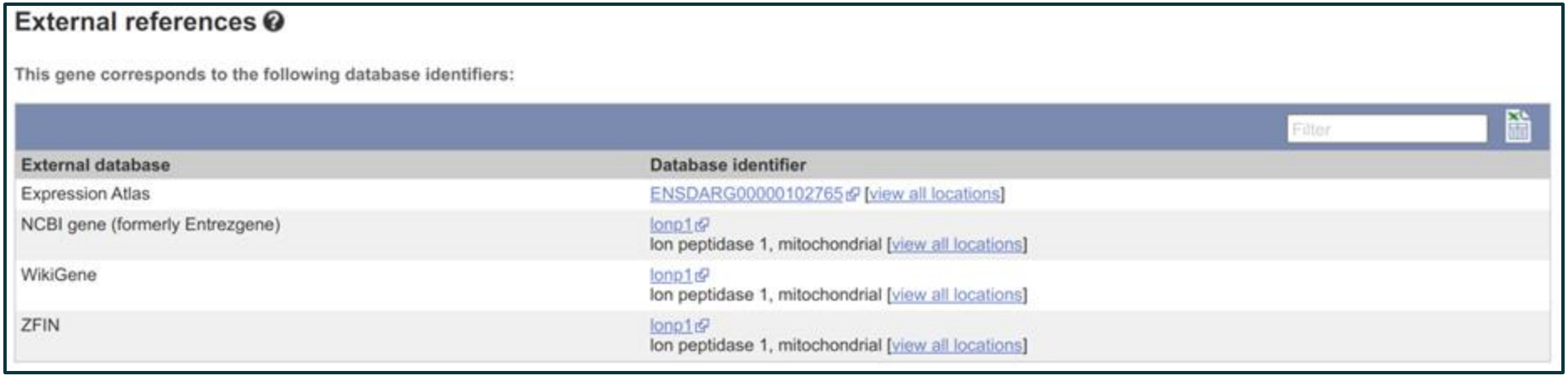
GO: Molecular function

Show/hide columns (1 hidden) Filter

Accession	Term	Evidence	Annotation source	Transcript IDs	
GO:0000166	nucleotide binding	IEA	UniProt	ENSDART00000158009	• Search BioMart • View on karyotype
GO:0003677	DNA binding	IEA	UniProt	ENSDART00000158009	• Search BioMart • View on karyotype
GO:0003697	single-stranded DNA binding	IBA	GO_Central	ENSDART00000158009	• Search BioMart • View on karyotype
GO:0004176	ATP-dependent peptidase activity	IBA	GO_Central	ENSDART00000167550 ENSDART00000158009	• Search BioMart • View on karyotype
GO:0005524	ATP binding	IEA	UniProt	ENSDART00000158009 ENSDART00000167550	• Search BioMart • View on karyotype
GO:0016887	ATP hydrolysis activity	IEA	UniProt	ENSDART00000158009 ENSDART00000167550	• Search BioMart • View on karyotype
GO:0043565	sequence-specific DNA binding	IEA	UniProt	ENSDART00000158009	• Search BioMart • View on karyotype

"Gene" Demo - External References

- Go to ENSDARG00000102765 and click on “External references”



External references

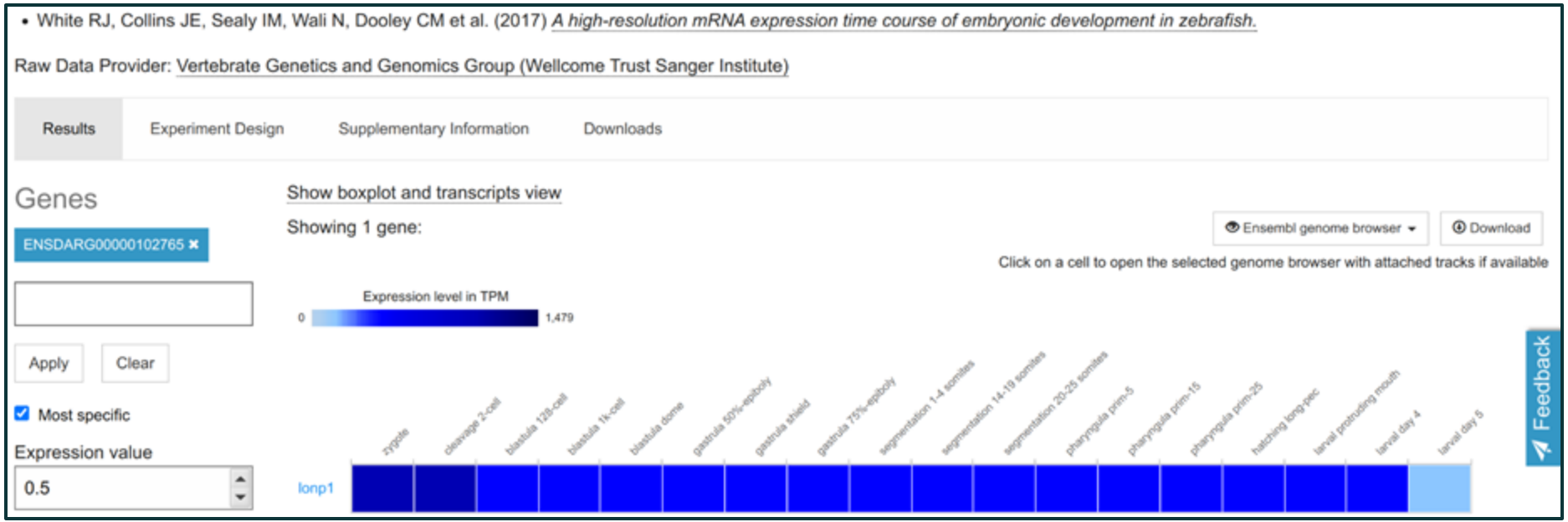
This gene corresponds to the following database identifiers:

Filter

External database	Database identifier
Expression Atlas	ENSDARG00000102765 [view all locations]
NCBI gene (formerly Entrezgene)	lonp1 lon peptidase 1, mitochondrial [view all locations]
WikiGene	lonp1 lon peptidase 1, mitochondrial [view all locations]
ZFIN	lonp1 lon peptidase 1, mitochondrial [view all locations]

"Gene" Demo - Expression Atlas

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

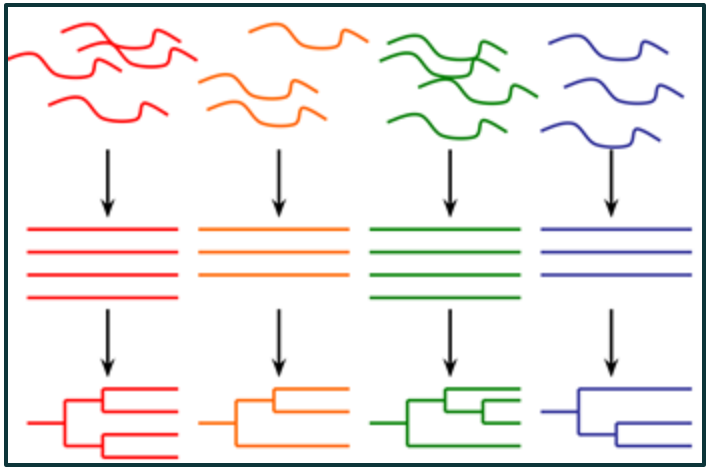


Compara

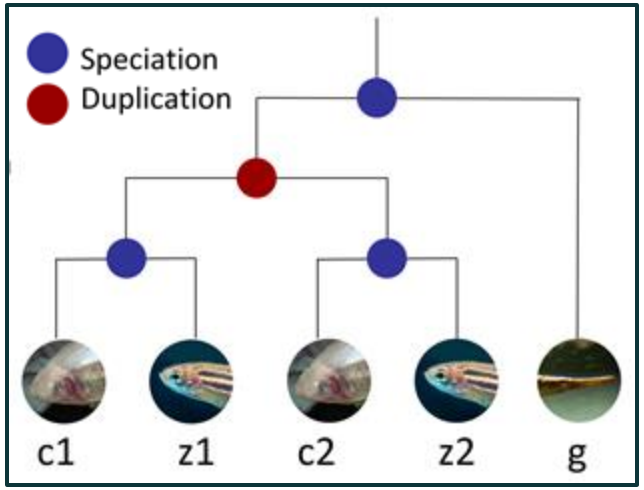
- Compara - produce Ensembl's comparative genomics resources
- Two types of analysis:
  - Gene level comparisons to produce **gene trees**, e.g. infer **homologues** (orthologues & paralogues)
  - **Whole genome alignments** - pairwise and multiple alignments, e.g. **constrained elements** and **synteny**

Compara - Gene Trees

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



Compara - Infer Homologues (Orthologues & Paralogues)



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

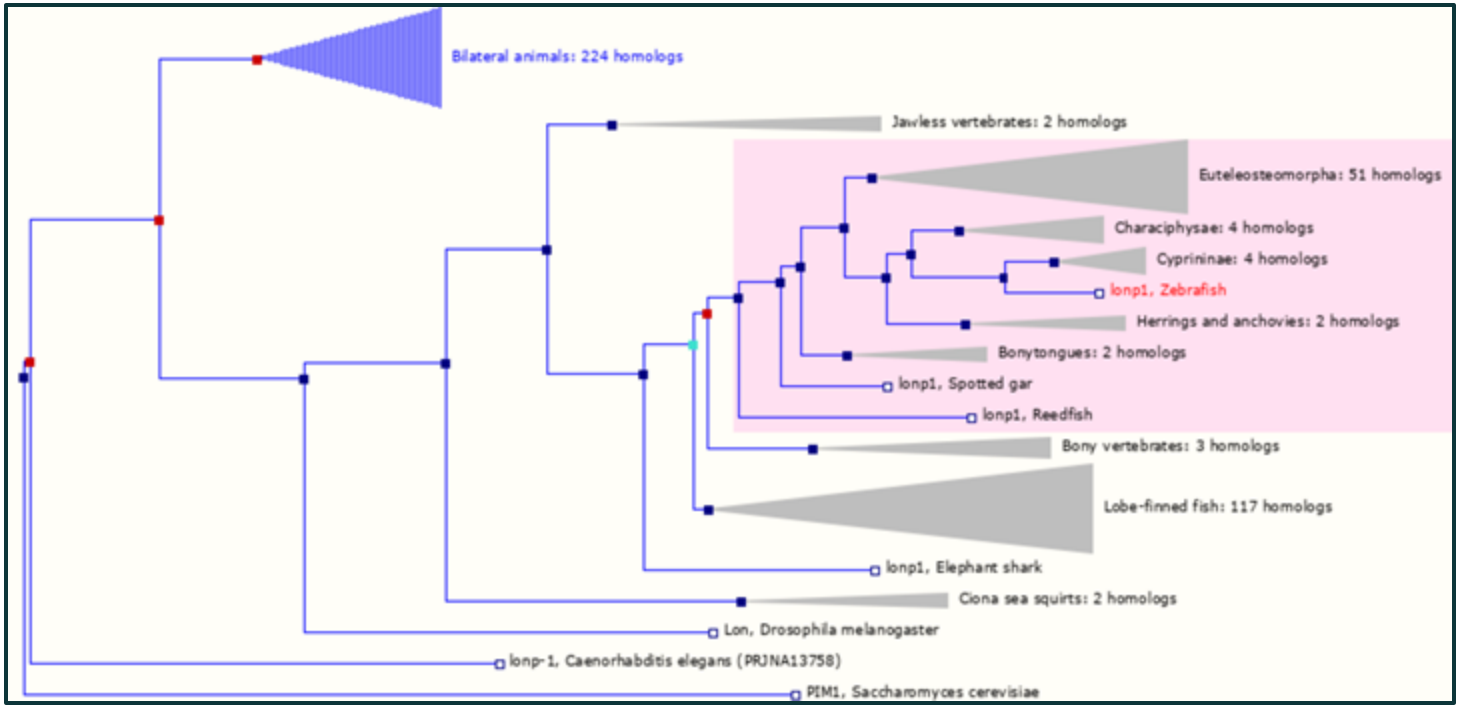
z1 & c1 are **orthologues** (arose from **speciation**), as are **z2 & c2** + **z2 & g**, etc…

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

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

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

Compara - lonp1 Gene Tree



Mini-exercise: Go to lonp1 gene tree and expand/collapse subtrees to show both paralogs

Problem with latest release

- Due to an apparent production issue, many fewer genes in release 110 have **ZFIN** names
- **23,328** genes in release 109 have ZFIN names, but only **15,924** now
- Meanwhile, only **164** genes in 109 had HGNC names, but now **3,113** do
- (308 genes have names from miRBase in both 109 and 110)
- Expect fix, but not until next release at the earliest (i.e. 3 or 4 months)
- Affects all types of protein-coding genes, but a particular problem for **paralogs...**

Problem with latest release

Teleost-specific paralogs are normally called *a* and *b*, e.g. *kdm2aa* and *kdm2ab*
Paralogs for *kdm2aa* in Ensembl release 109:

Show/hide columns | Filter

Type	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Paralogues	Osteoglossocephalai	ENSDARG00000078133 kdm2ab lysine (K)-specific demethylase 2Ab [Source:ZFIN;Acc:ZDB-GENE-101007-5]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	14:21,754,521-21,782,227:1	52.40 %	53.62 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000046010 kdm2bb lysine (K)-specific demethylase 2Bb [Source:ZFIN;Acc:ZDB-GENE-080225-13]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	10:41,826,331-41,907,213:-1	42.70 %	43.32 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000036593 kdm2ba lysine (K)-specific demethylase 2Ba [Source:ZFIN;Acc:ZDB-GENE-040426-2195]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	8:40,238,647-40,276,193:-1	29.45 %	20.93 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000018559 kdm7ab lysine (K)-specific demethylase 7Ab [Source:NCBI gene;Acc:503902]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	25:17,244,532-17,310,301:1	22.20 %	16.43 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000006584 phf8 PHD finger protein 8 [Source:NCBI gene;Acc:566534]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	23:27,591,366-27,608,257:-1	20.74 %	17.23 %

Problem with latest release

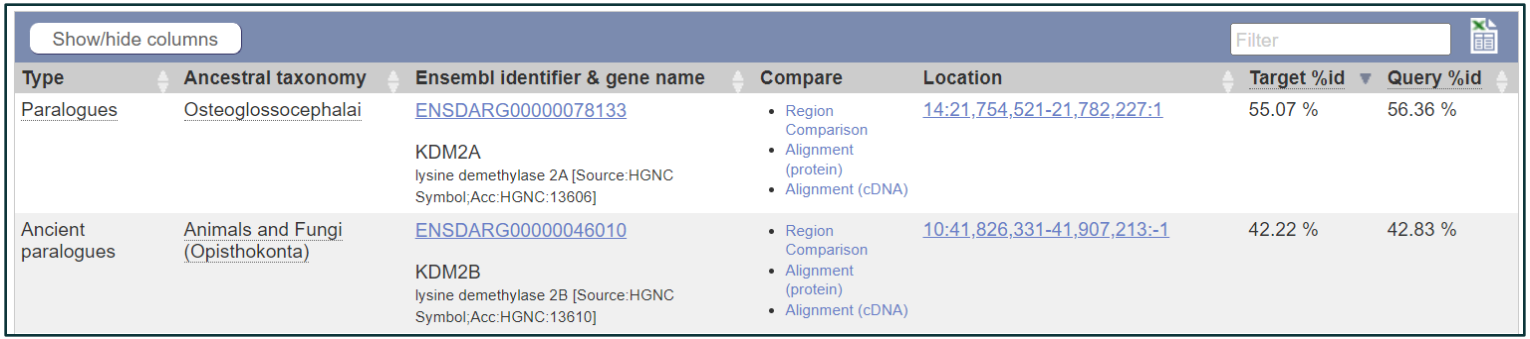
Teleost-specific paralogs are normally called a and b, e.g. *kdm2aa* and *kdm2ab*
Paralogs for *kdm2aa* in Ensembl release 110:

Show/hide columns | Filter

Type	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Paralogues	Osteoglossocephalai	ENSDARG00000078133 KDM2A lysine demethylase 2A [Source:HGNC Symbol;Acc:HGNC:13606]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	14:21,754,521-21,782,227:1	55.07 %	56.36 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000046010 KDM2B lysine demethylase 2B [Source:HGNC Symbol;Acc:HGNC:13610]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	10:41,826,331-41,907,213:-1	42.22 %	42.83 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000036593 kdm2ba lysine (K)-specific demethylase 2Ba [Source:ZFIN;Acc:ZDB-GENE-040426-2195]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	8:40,238,647-40,276,193:-1	29.33 %	20.85 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000018559 kdm7ab lysine (K)-specific demethylase 7Ab [Source:ZFIN;Acc:ZDB-GENE-050309-32]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	25:17,244,532-17,310,301:1	21.65 %	16.02 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000006584 phf8 PHD finger protein 8 [Source:NCBI gene (formerly Entrezgene);Acc:566534]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	23:27,591,366-27,608,257:-1	20.64 %	17.15 %

Problem with latest release

Teleost-specific paralogs are normally called a and b, e.g. *kdm2aa* and *kdm2ab*
Paralogs for *kdm2aa* in Ensembl release 110:


Show/hide columns

Filter

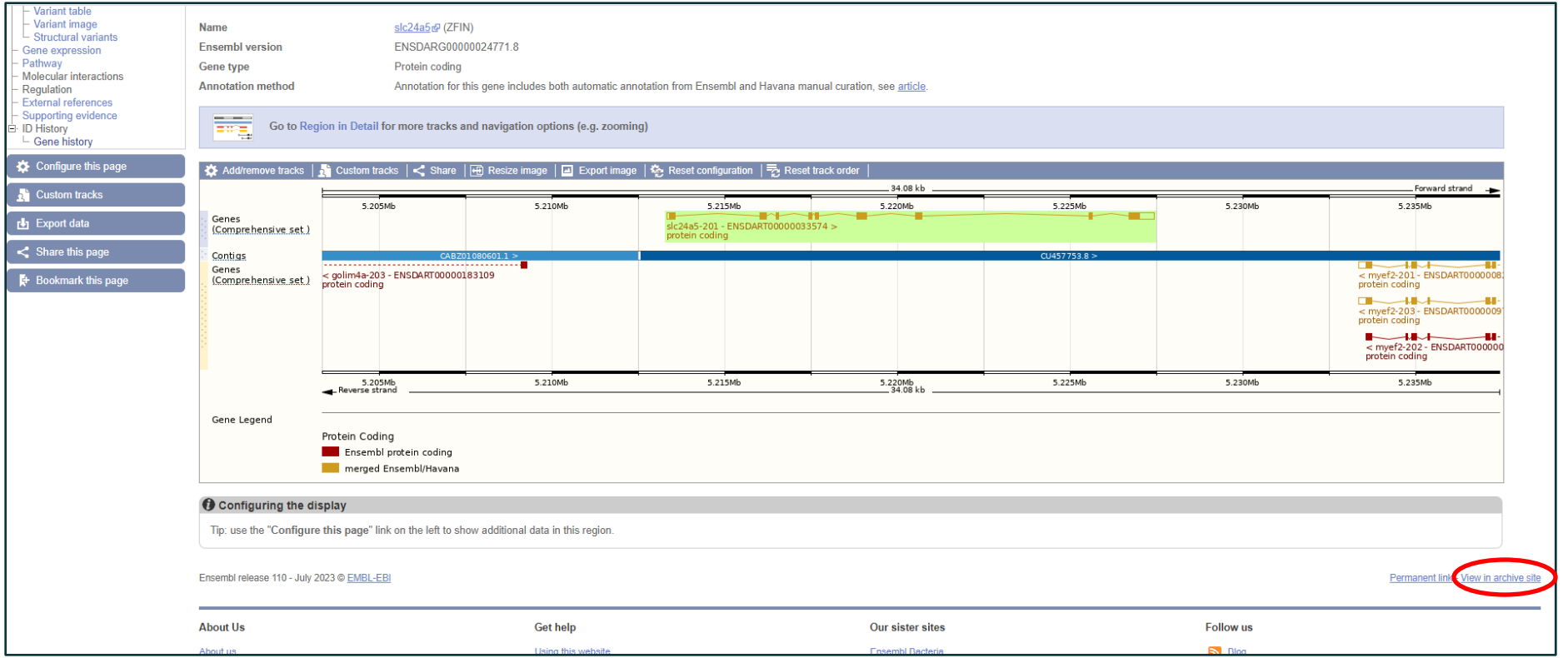
Type	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id	Query %id
Paralogues	Osteoglossocephalai	ENSDARG00000078133 KDM2A lysine demethylase 2A [Source:HGNC Symbol;Acc:HGNC:13606]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	14:21,754,521-21,782,227:1	55.07 %	56.36 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000046010 KDM2B lysine demethylase 2B [Source:HGNC Symbol;Acc:HGNC:13610]	• Region Comparison • Alignment (protein) • Alignment (cDNA)	10:41,826,331-41,907,213:-1	42.22 %	42.83 %

Three problems:
- Makes it unclear that a gene is a paralog
- Violates nomenclature rules
- Breaks link to ZFIN genes

Species	Gene	Protein
Zebrafish	*shha*	Shha
Human	*SHH*	SHH
Mouse	*Shh*	SHH

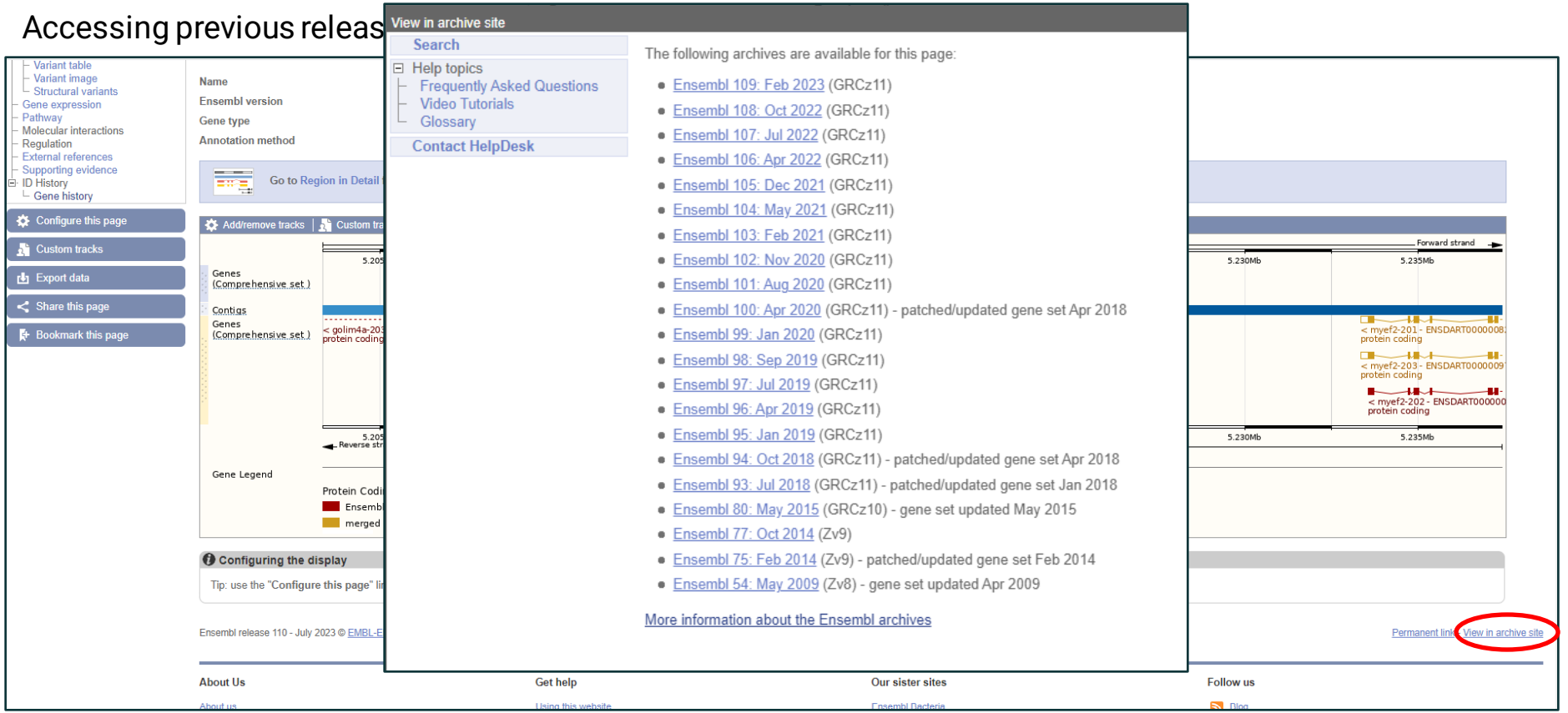
Problem with latest release

Accessing previous release:



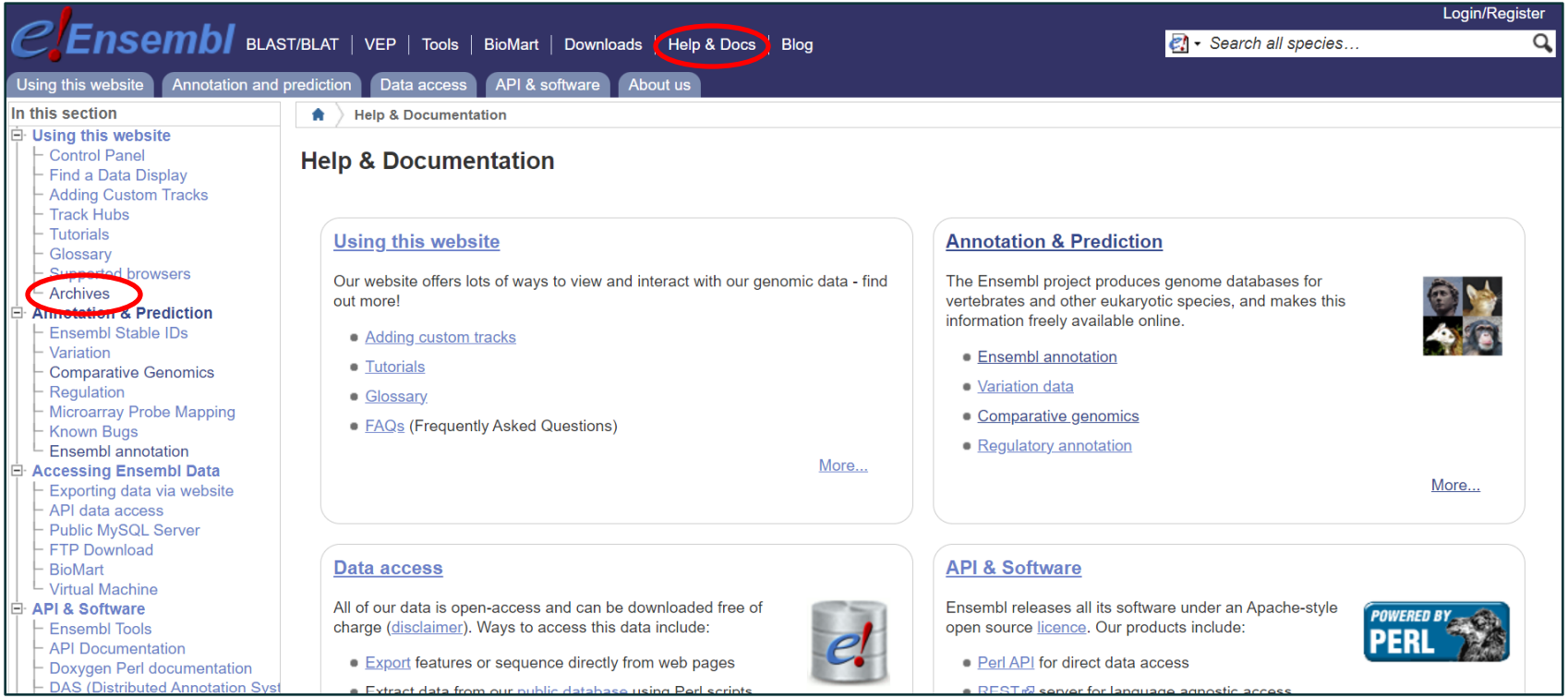
Problem with latest release

Accessing previous releas



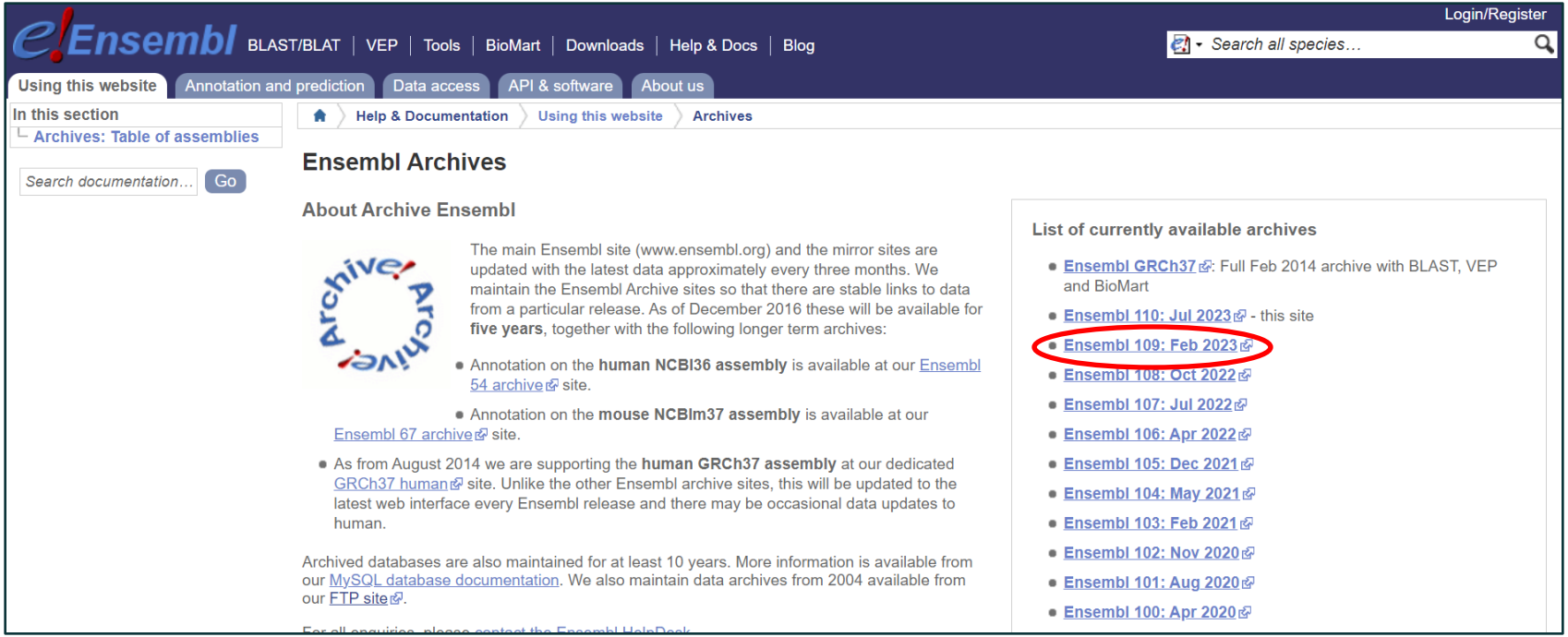
Problem with latest release

Accessing previous release:



Problem with latest release

Accessing previous release:

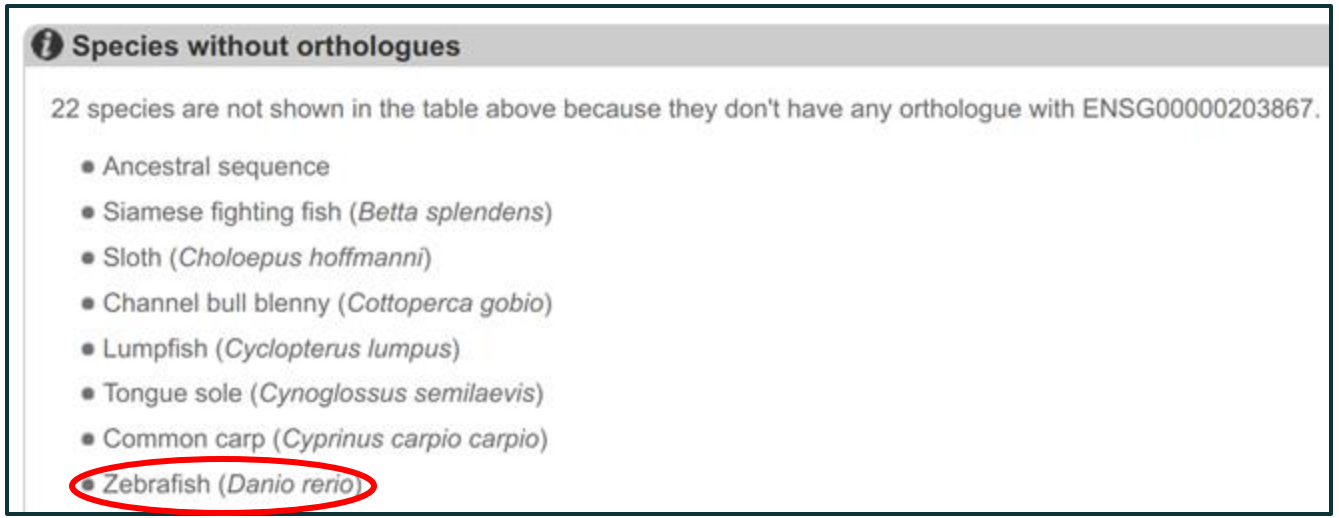


Compara - Whole Genome Alignments

- **Pairwise whole genome alignments** with LASTZ
- Zebrafish has alignments to **64 species** (plus itself)
- Only human (180) and medaka (65) have more
- Full list at: www.ensembl.org/info/genome/compara/analyses.html
- **Multiple genome alignments** with EPO (Enredo, Pecan, Ortheus)
- Zebrafish is in **2** alignments (out of 11 in Ensembl) - one of **39 fish** and one of **65 fish**
- For lists of species, see: www.ensembl.org/info/genome/compara/multiple_genome_alignments.html

Synteny Example

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



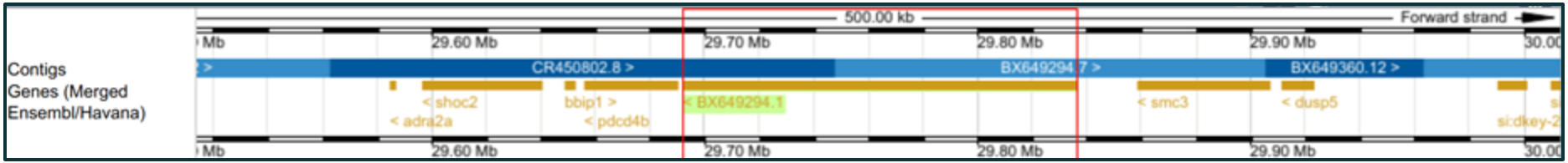
Synteny Example

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

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

Synteny Example

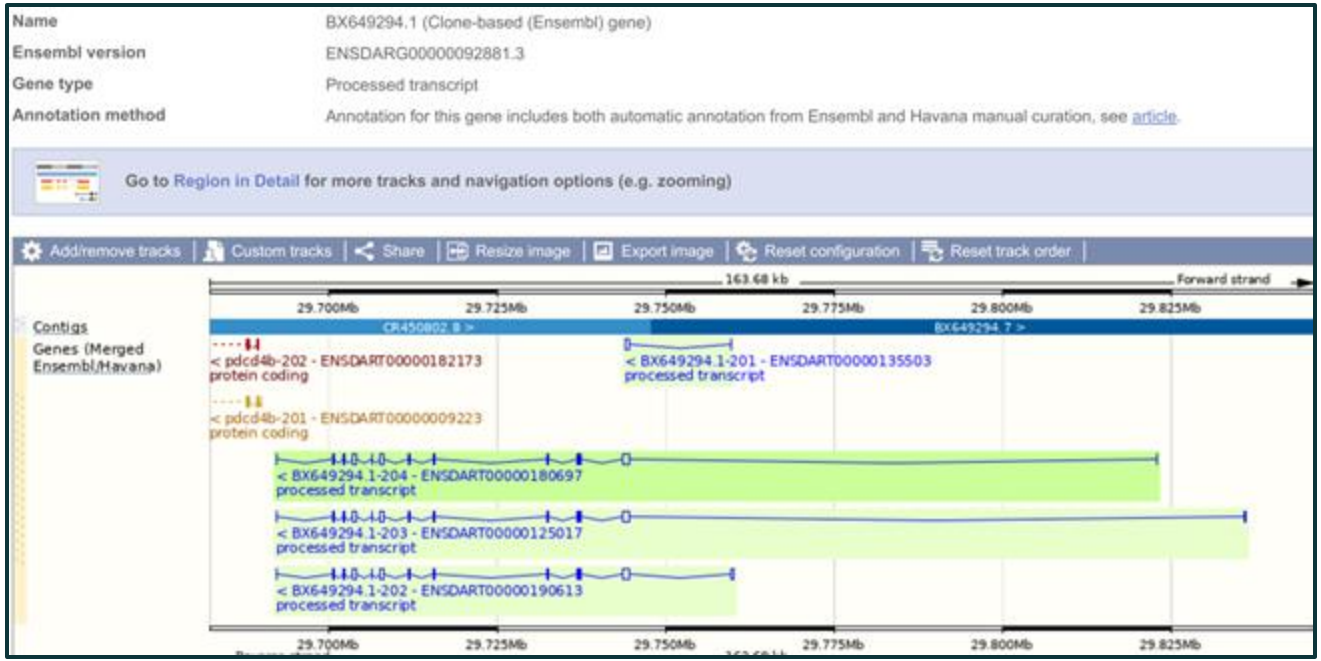
- If we look at the chr22 region in zebrafish then all the surrounding genes are the same and RBM20 is likely to be BX649294.1



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

Synteny Example

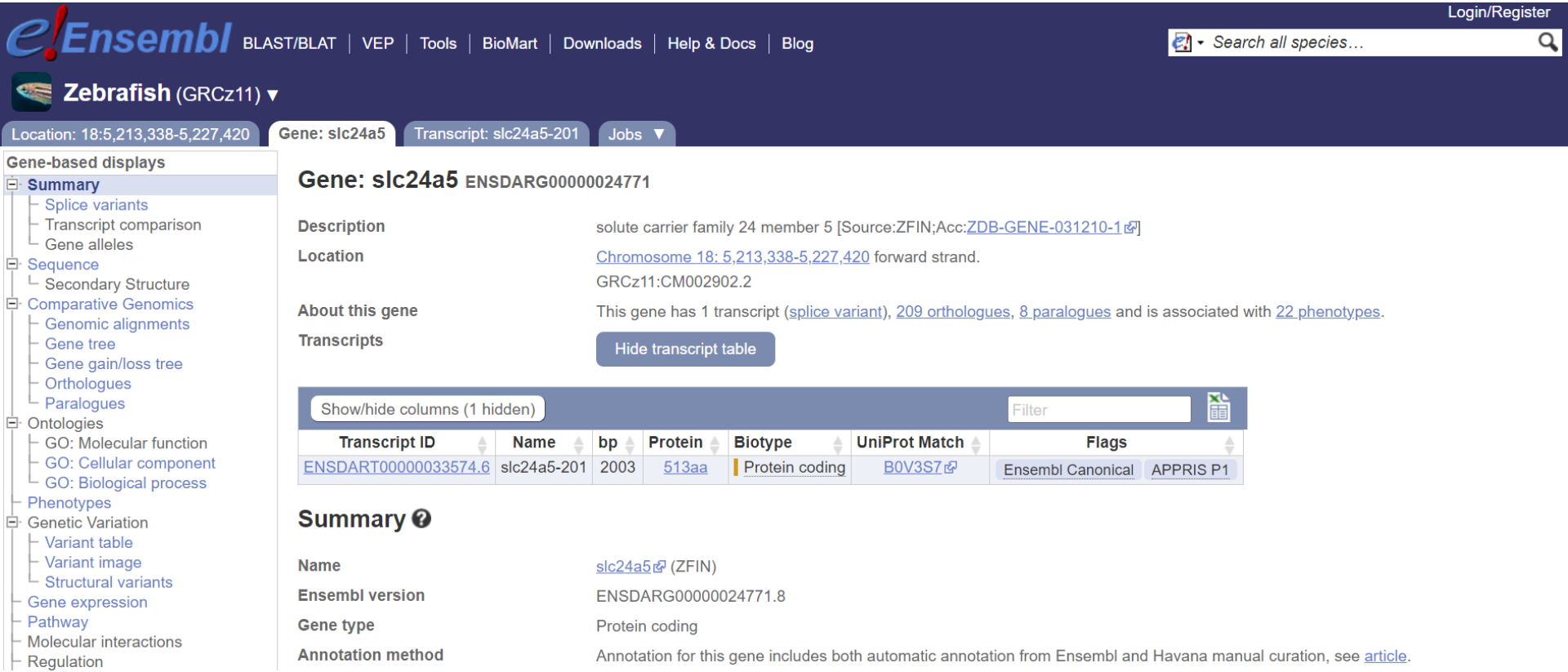
- Erroneously labelled as processed transcript and so not in protein gene tree, so not labelled as orthologue or named by orthology



Exercise 2

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

Summary overview of Ensembl Browser



Part 2 summary

- Switching on additional tracks
- Gene view
- Comparative genomics
- Learning outcomes:
  - Customise Ensembl tracks for your needs
  - Navigate gene view to find splice variants, exon tables, external data, etc
  - Identify orthologs and paralogs
- Any questions?
- Email me any questions that you think of later: e.busch-nentwich@qmul.ac.uk

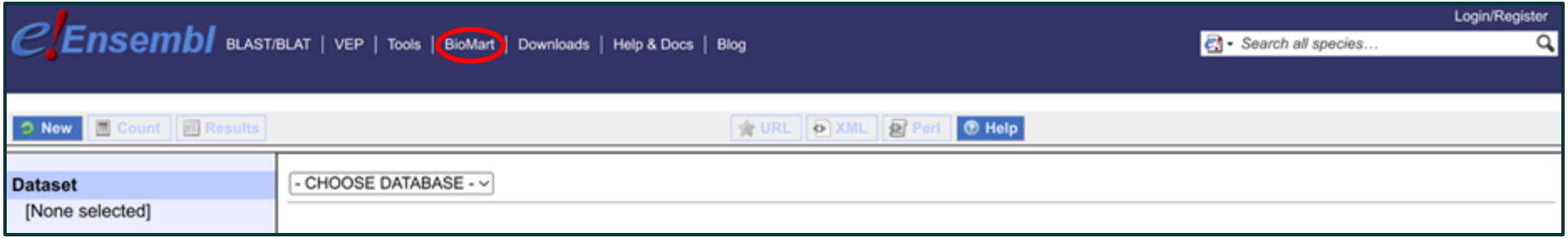
Part 3 (Day 2)

- BioMart
- Other tools
- Custom tracks

- Before we start: Any questions from yesterday?

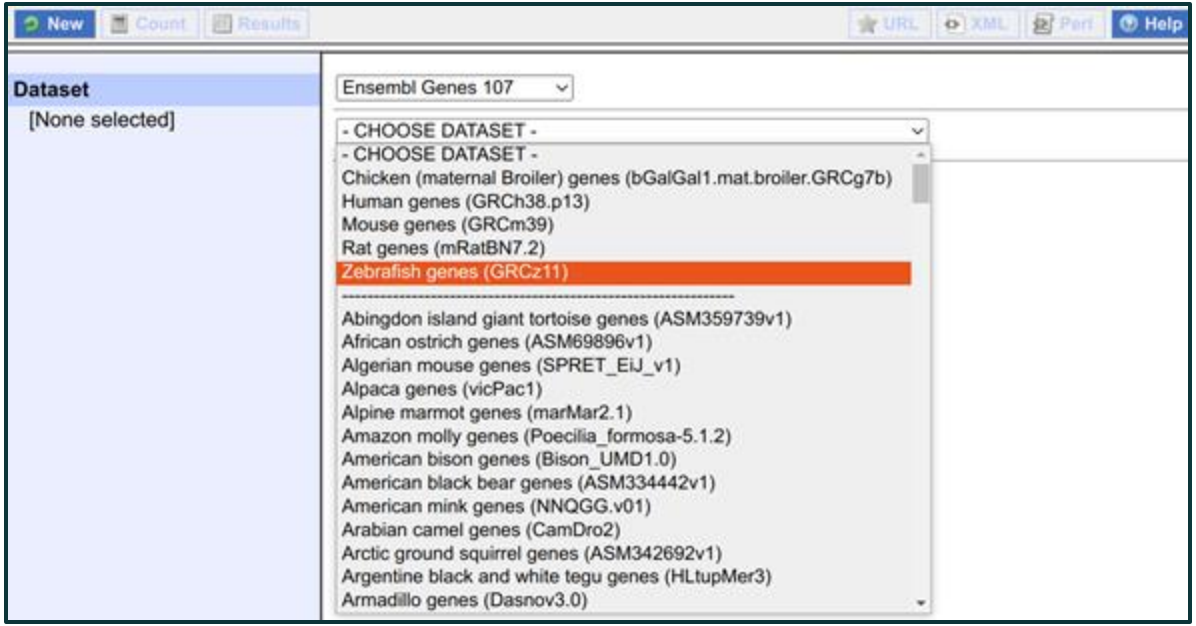
BioMart

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



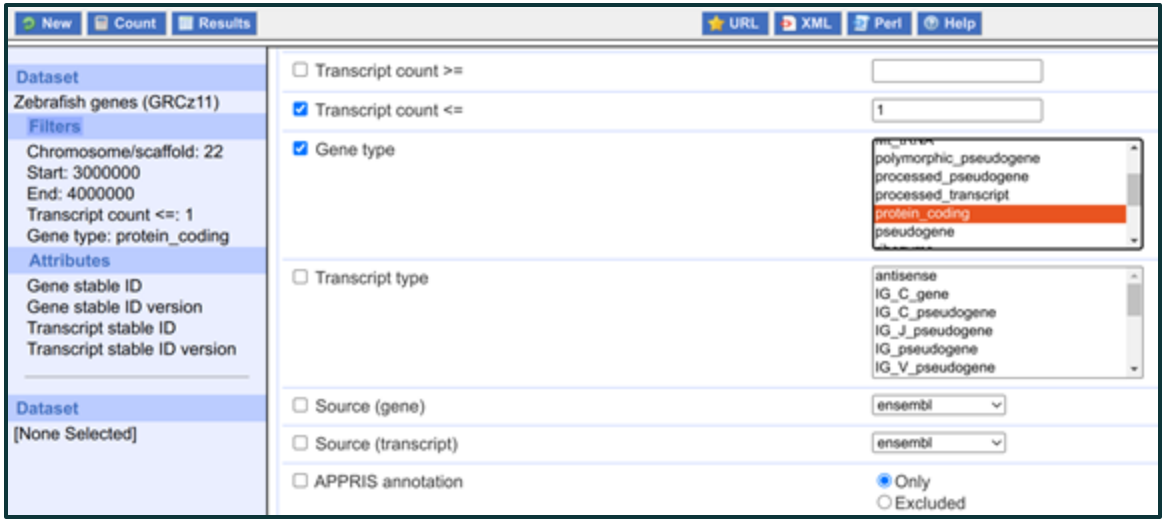
BioMart - Dataset

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



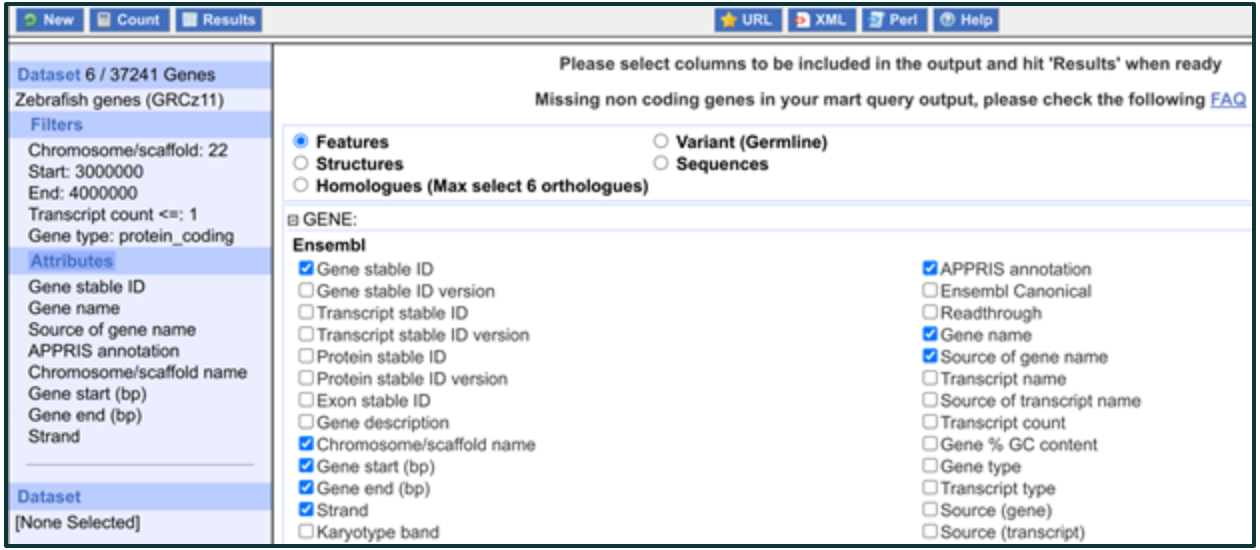
BioMart - Filters

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



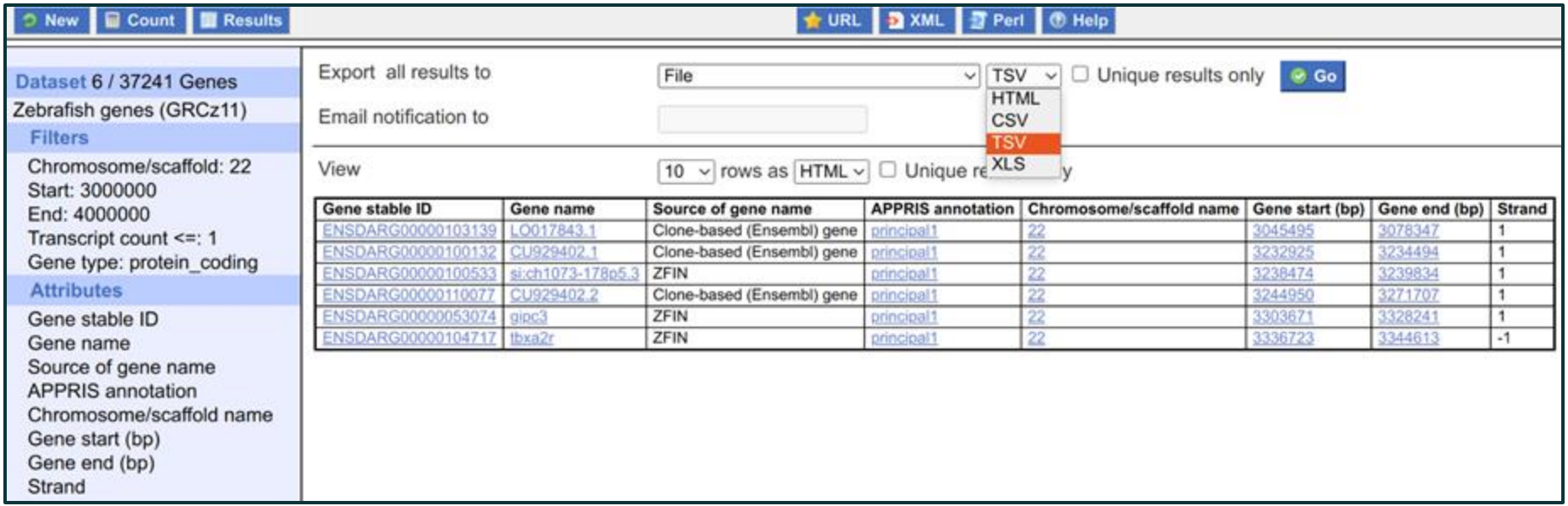
BioMart - Attributes

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



BioMart - Results

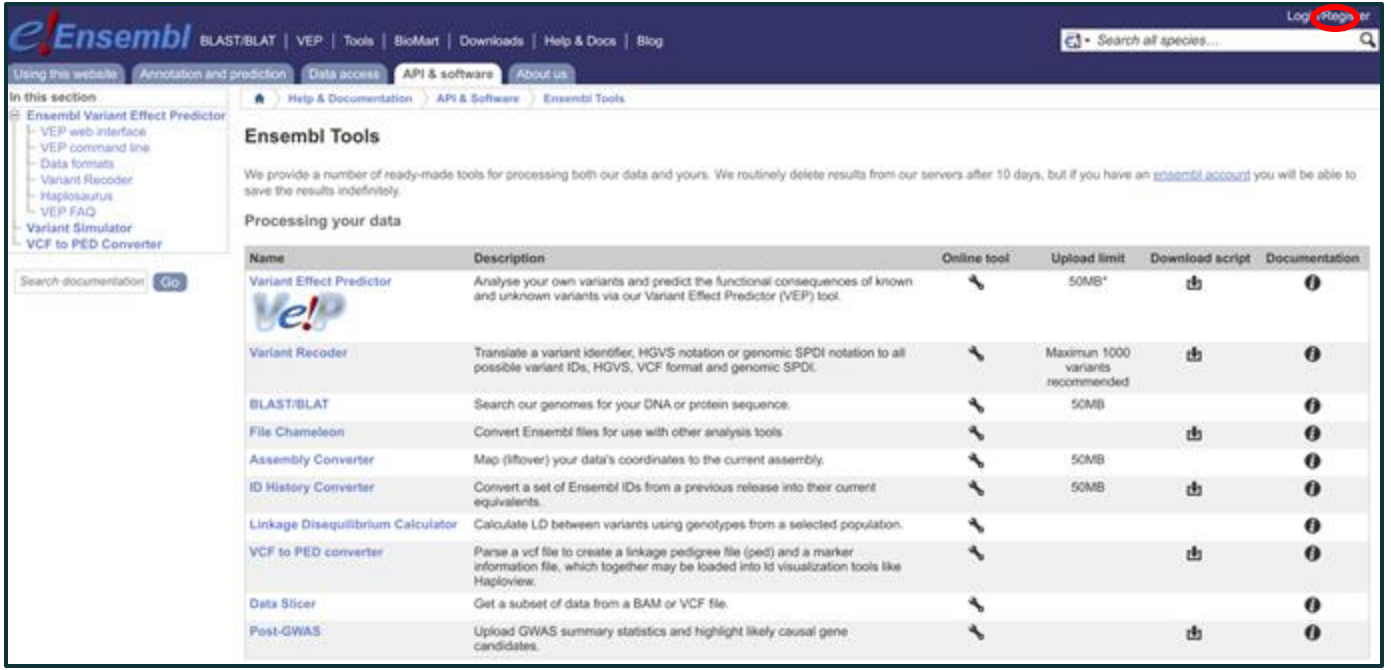
- Access your selected data in multiple formats
- e.g. HTML, TSV, CSV, XLS



Exercise 3

- Do Exercise 3 - “CRISPR design”
- Covers:
  - Using BioMart to download sequence
  - Designing and checking guideRNAs
- Go to mbl2023.buschlab.org

More Tools



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

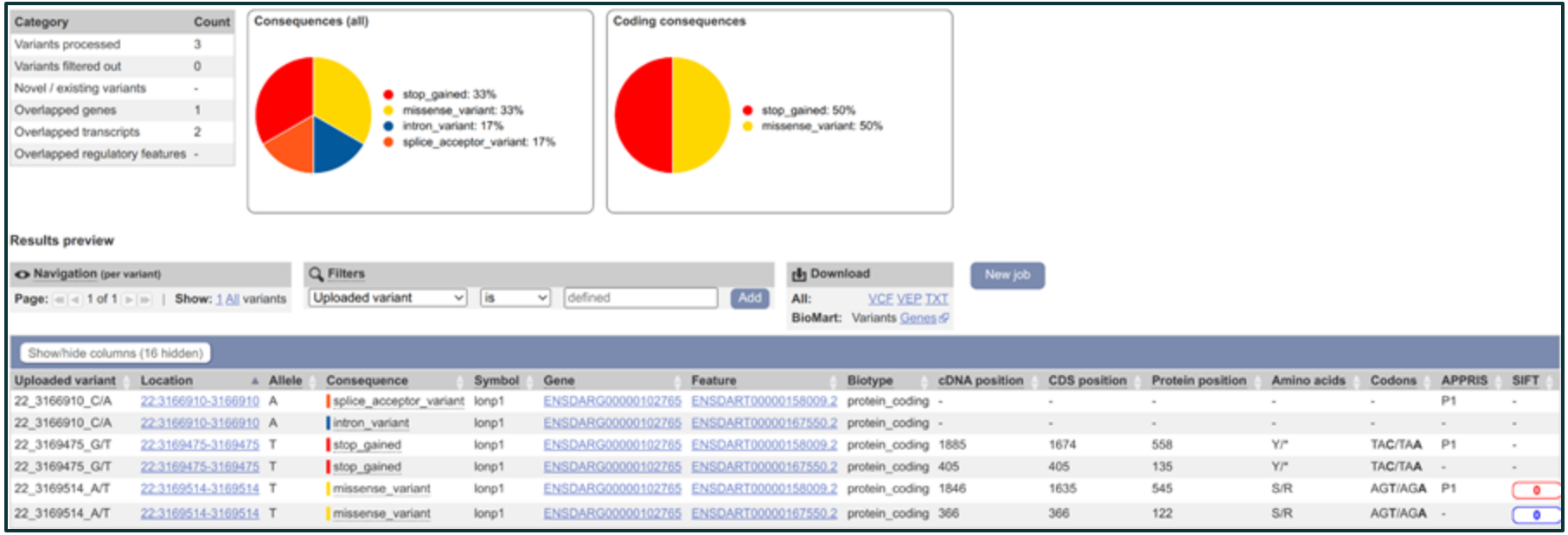
Variant Effect Predictor

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

```
22 3169475 3169475 G/T 1
22 3169514 3169514 A/T 1
22 3166910 3166910 C/A 1
```

  (Chr, Start, End, REF/ALT, Strand)
- Custom Ensembl format, but standard formats like **VCF** can be used

Variant Effect Predictor



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

Results preview

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

Page: 1 of 1 | Show: 1 All variants

Uploaded variant | is | defined | Add

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

Exercise 4

- Do Exercise 4 - “exploring variation”
- Covers:
  - Variant nomenclature
  - Variant Effect Predictor
  - Variant tables
- Go to mbl2023.buschlab.org

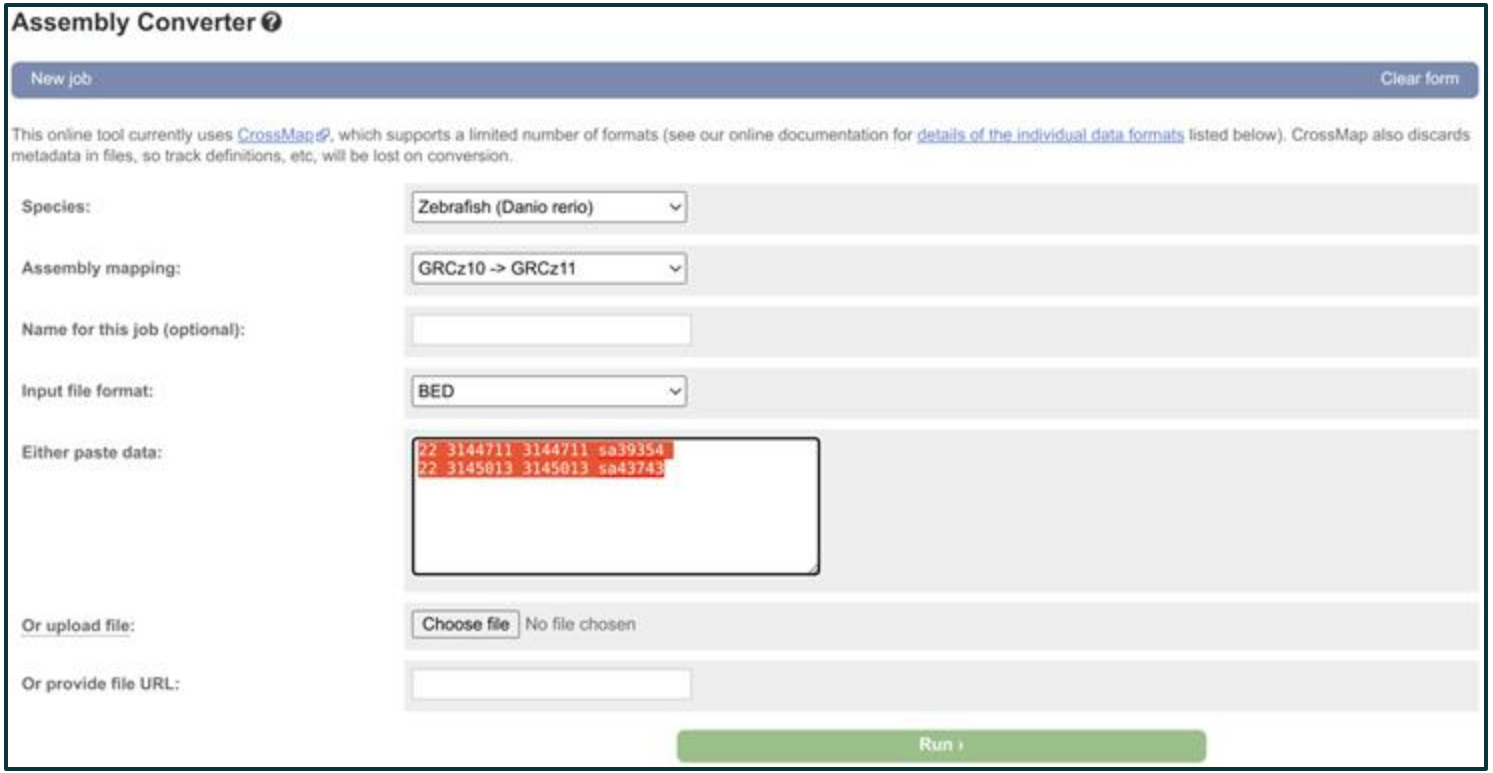
Assembly Converter

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

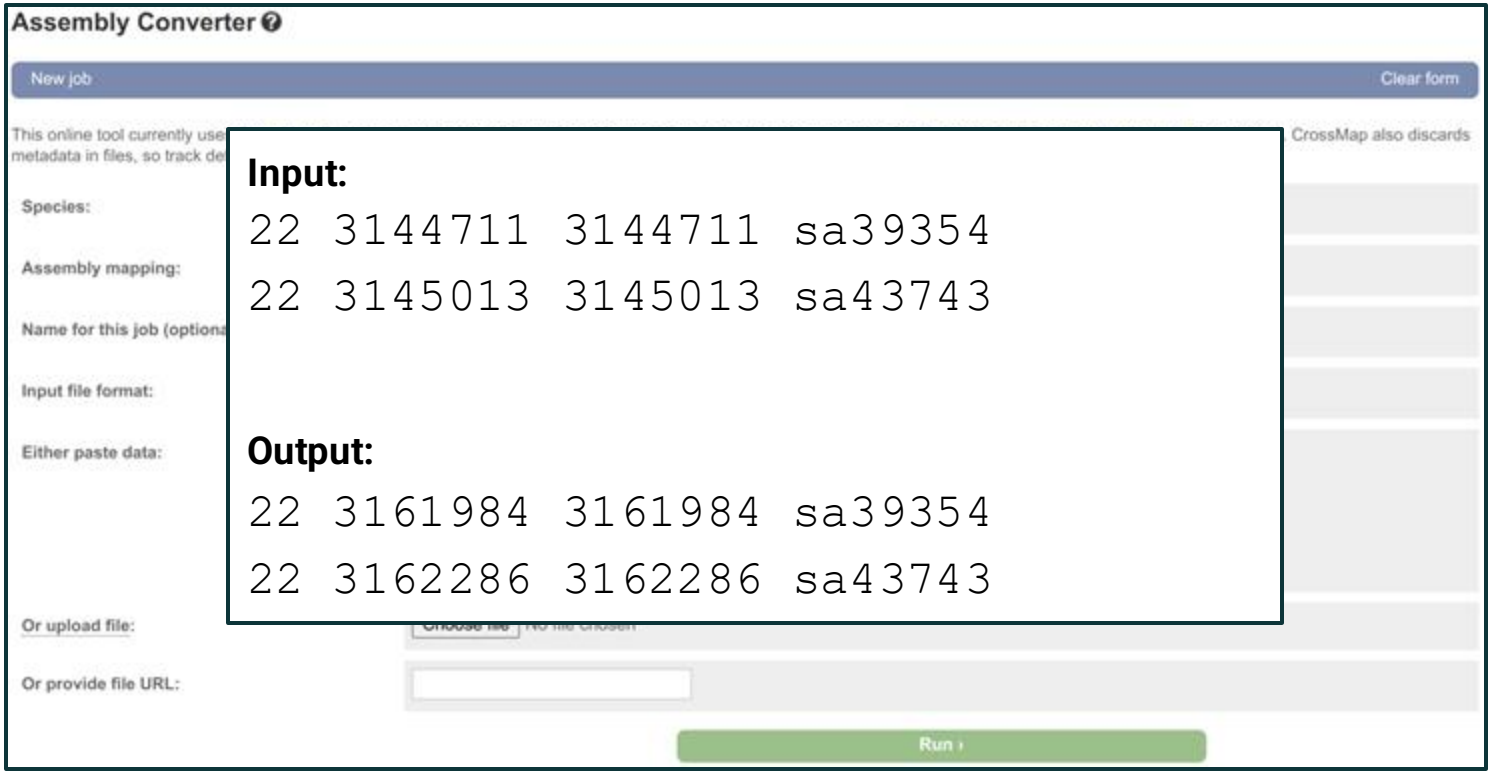
```
22 3144711 3144711 sa39354
22 3145013 3145013 sa43743
```

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

Assembly Converter



Assembly Converter

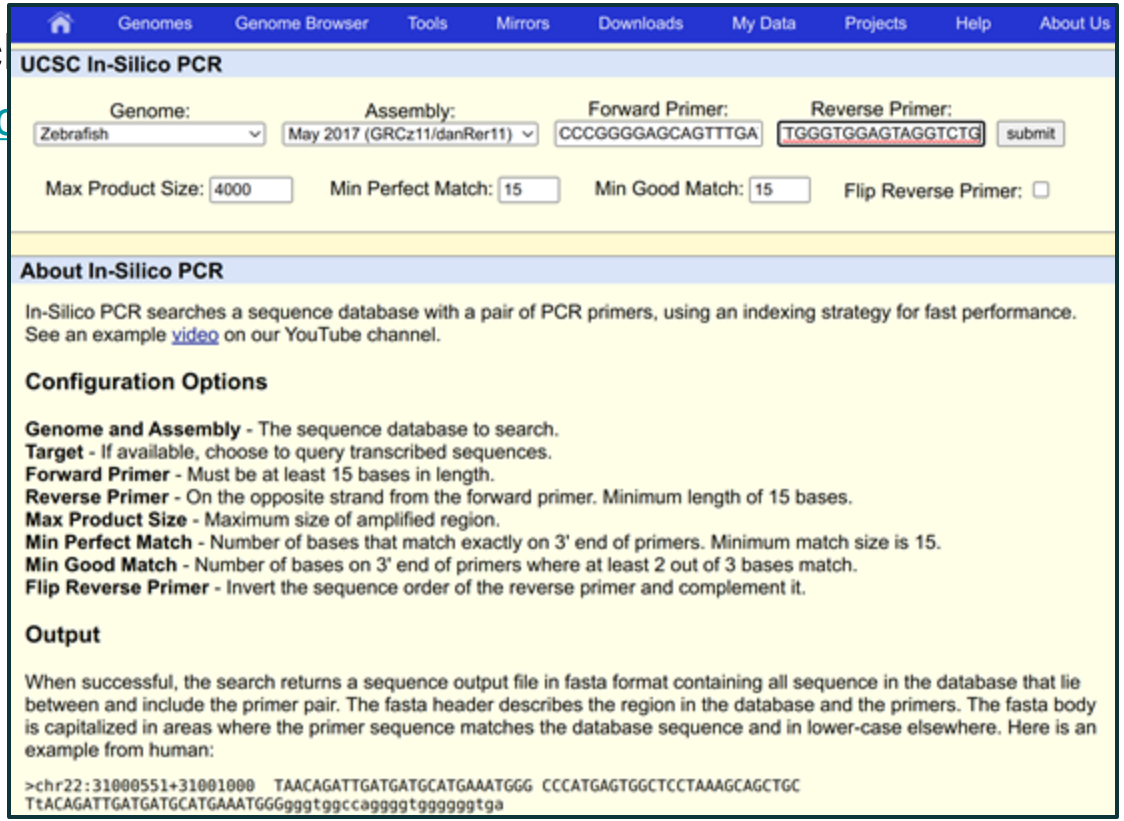


UCSC In-Silico PCR

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

UCSC In-Silico PCR

- Fast searc
- genome.uc

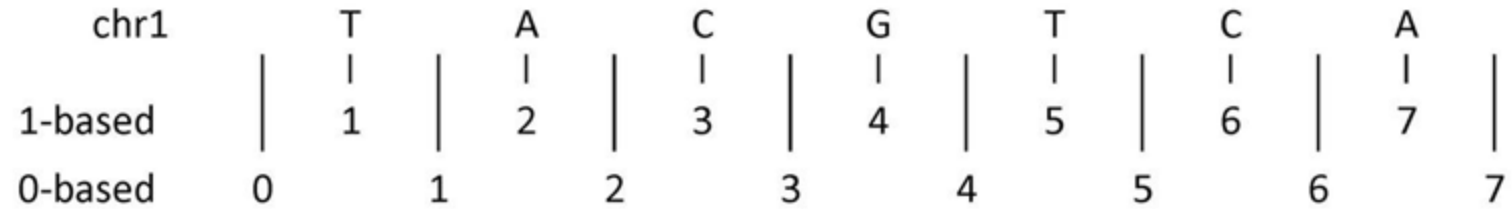


UCSC In-Silico PCR

- Fast search for p mers
- genome.ucsc.edu

UCSC & Ensembl Differences

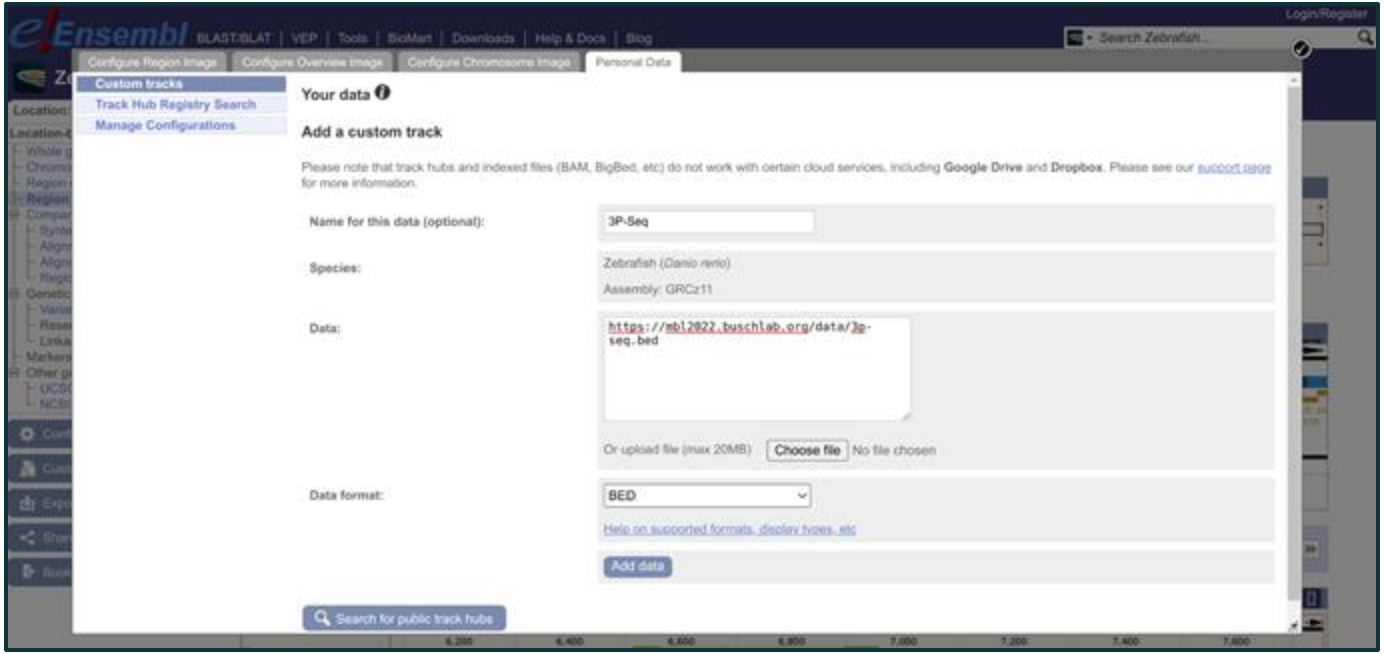
- **Ensembl:** 1
  UCSC: chr1
- **Ensembl:** 1-based coordinates (bases numbered)
  UCSC: 0-based coordinates (numbers between bases)



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

Custom Tracks

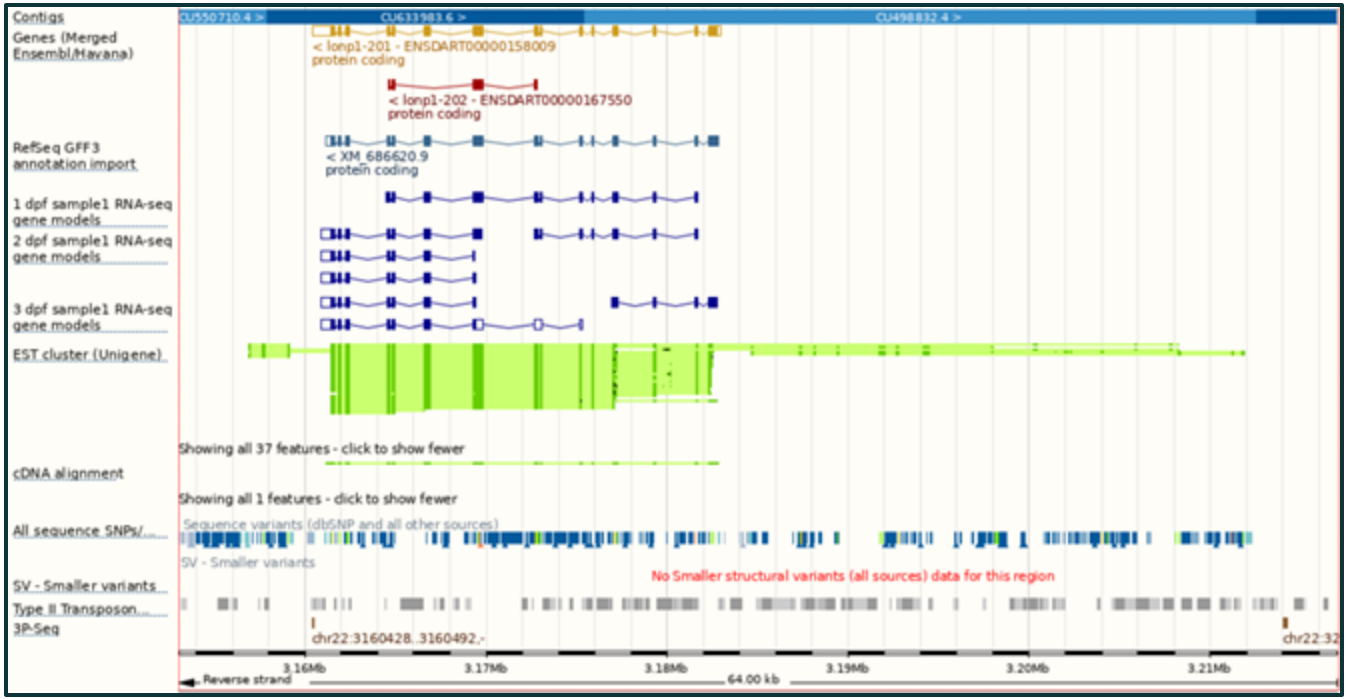
- Click “Custom tracks” and add https://mbl2023.buschlab.org/data/3p-seq.bed



- 24 hpf 3P-Seq data from Bartel lab

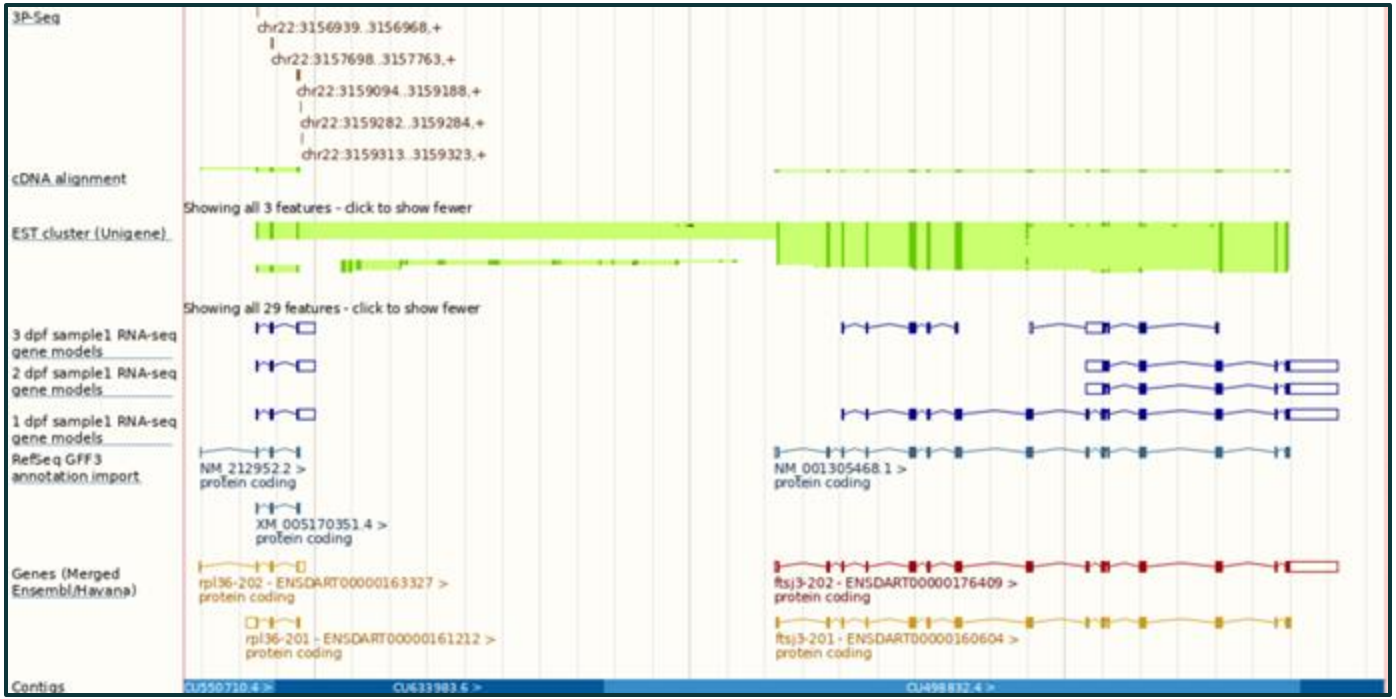
Custom Tracks

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



Custom Tracks

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

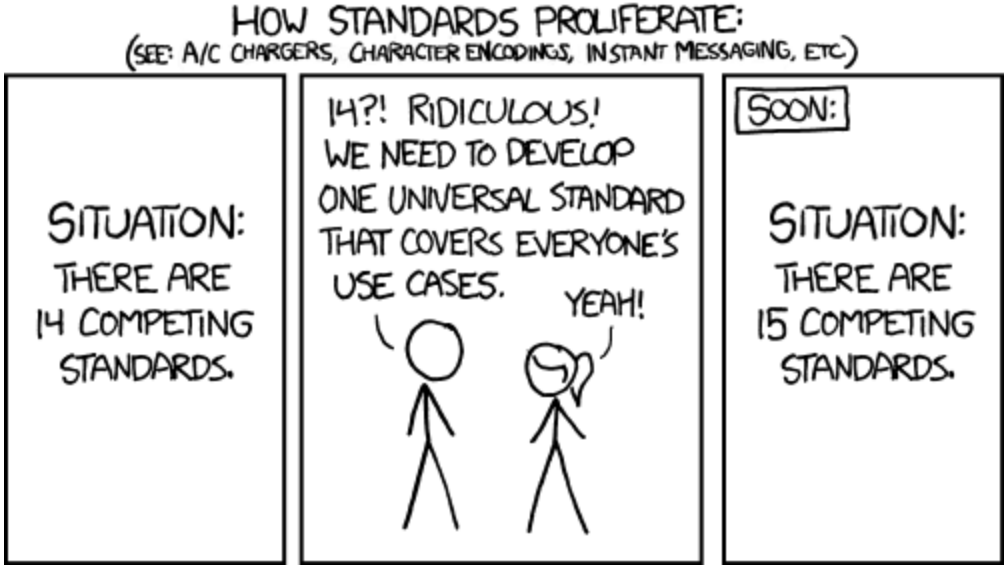


Custom Tracks - Lawson Lab Annotation

- Lawson *et al.* (2020) “**An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes**” eLife 9:e55792
- Aim: To unify discrepancies between Ensembl and RefSeq annotations

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



Custom Tracks - Lawson Lab Annotation

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

Custom Tracks - Lawson Lab Annotation

- Lawson **otation for sensitive nes”** eLife 9:e
- www.um ome/
- Add: https://w es/v4.3.2.gtf
- Large, sc

Regulatory information

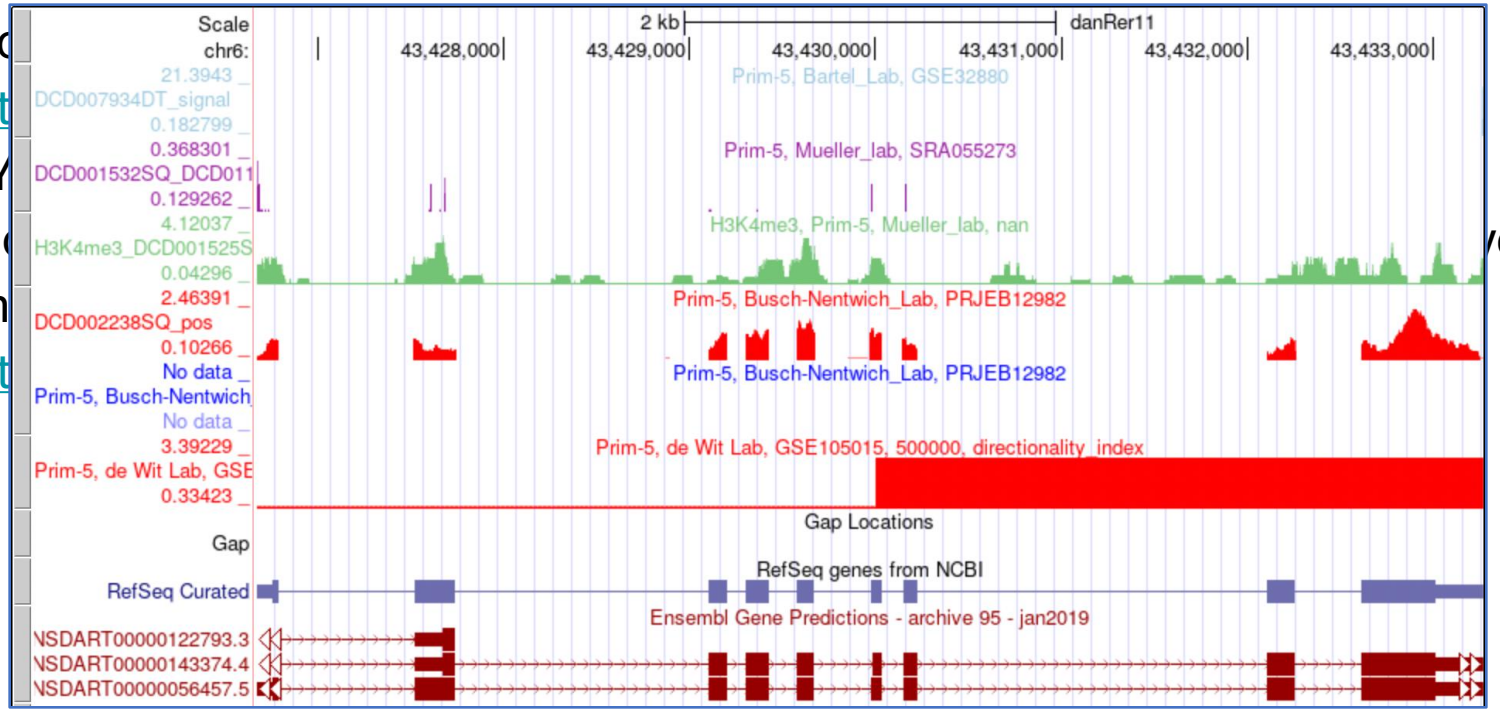
- Ensembl integrates **regulatory information** from a variety of genome-wide assays (e.g. ChIP-seq and DNase-seq)
- But only has annotation for a limited set of species (mainly human and mouse, but also pig, chicken and some fish)
- **Not** zebrafish
- Instead have **DANIO-CODE**, a compendium of functional genomics datasets: https://danio-code.zfin.org/
- Datasets submitted to DANIO-CODE can be visualised in Ensembl (and UCSC)

Regulatory information

- For Ensembl, add a **custom track** using the DANIO-CODE track hub: https://danio-code.zfin.org/trackhub/DANIO-CODE.hub.txt
- (You will need to select "**Track Hub**" for the data format.)
- Does make Ensembl quite slow, even with just the default tracks displayed
- Instead, could use UCSC by following instructions at: https://danio-code.zfin.org/help/

Regulatory information

- Fc
  ht
- (Y
- Dc
- In
  ht



Exercise 5

- Do Exercise 5 - “exploring data”
- Covers:
  - BioMart
  - Making BED files
  - Finding candidate genes
  - Finding orthologues
- Go to mbl2023.buschlab.org

Part 3 Summary

- BioMart
- Other tools
- Custom tracks
- Learning outcomes:
  - Use BioMart to download data
  - Analyse variants
  - Making BED files
  - Upload custom tracks

Thank You!

Any questions?