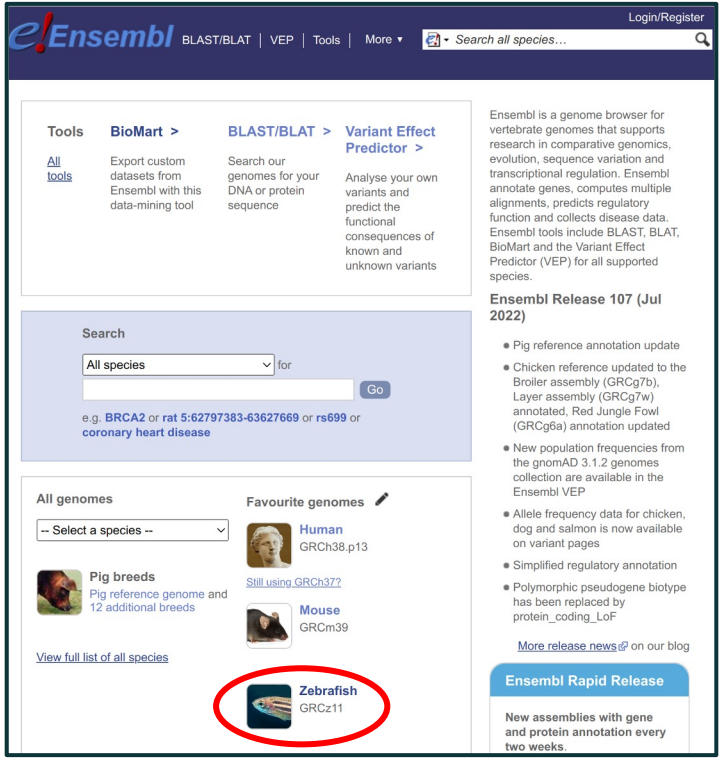
Extra Ensembl

Why Ensembl?

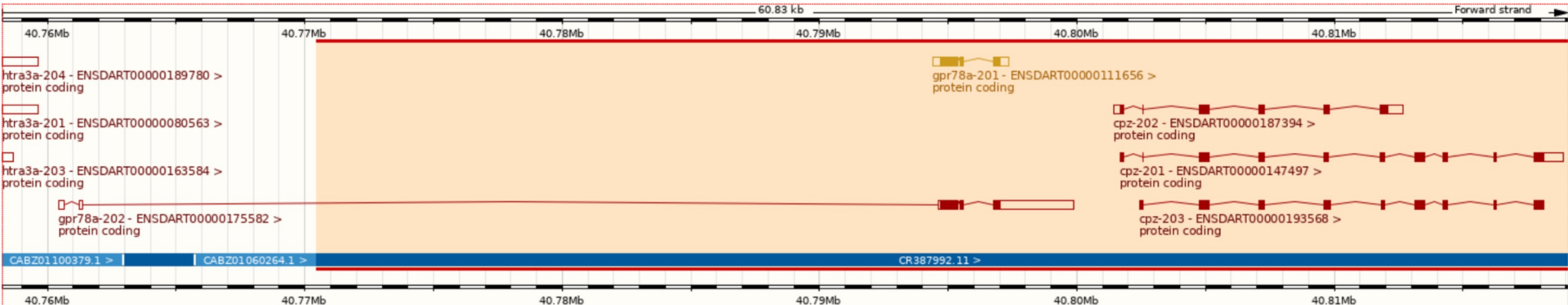
- We are biased!
- But is most widely used genome browser amongst zebrafish researchers
- **Primary source of zebrafish annotation** (UCSC imports Ensembl annotation)

- Zebrafish **annotation largely static** between releases
- But **naming and homology** updated (+ new functionality)



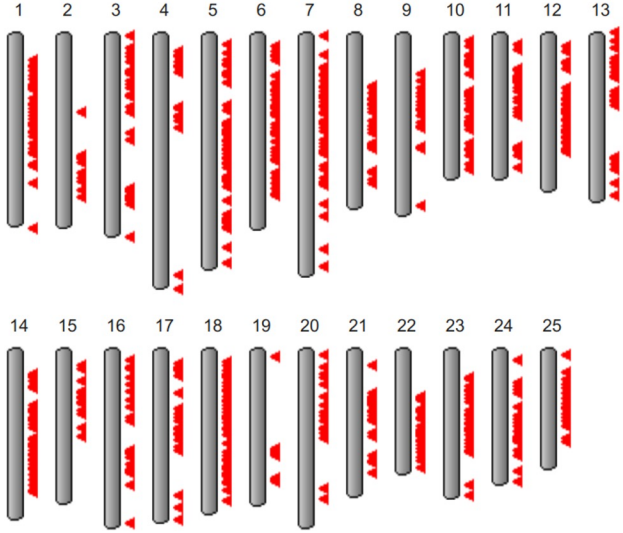
Zebrafish Genome

- **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

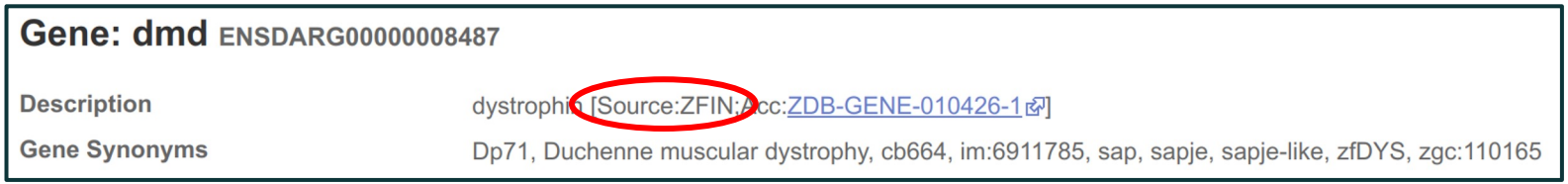
Older Assemblies

- 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

Gene Names

- Names assigned to Ensembl genes automatically based on **sequence similarity**
  - Mistakes are possible
  - Names can change
- **ZFIN gene symbols** (i.e. the name assigned by ZFIN) are preferred (>23,000 genes), but other databases are also used, e.g. HGNC for ~150 genes, miRBase for ~300 genes
- Description indicates source of name
- Genes without a match are given a name based on the sequence used to identify them, e.g AL645792.1 (clone) or **CABZ**01052570.1 (WGS)

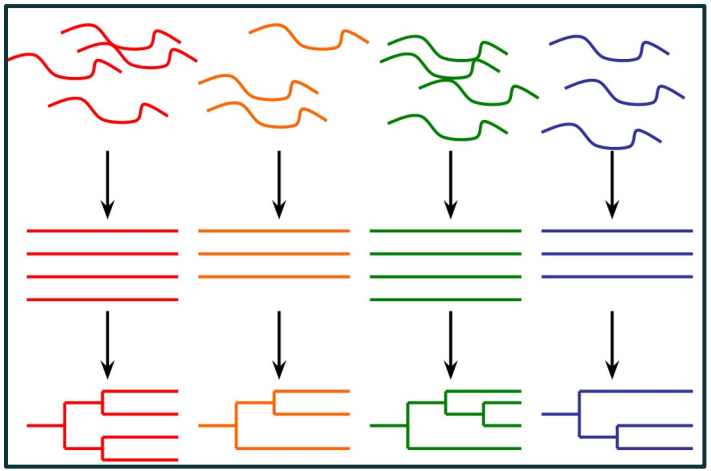


Compara

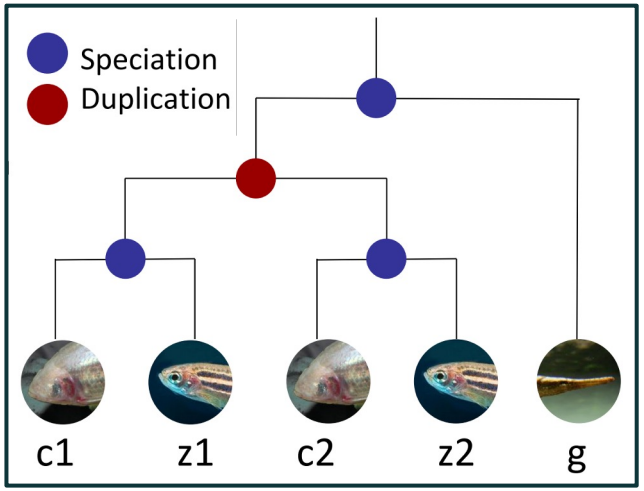
- Compara - produce Ensembl's comparative genomics resources
- Two types of analysis:
  - Gene level comparisons to produce **gene trees**, e.g. infer **homologues** (orthologues & paralogues)
  - **Whole genome alignments** - pairwise and multiple alignments, e.g. **constrained elements** and **synteny**

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**

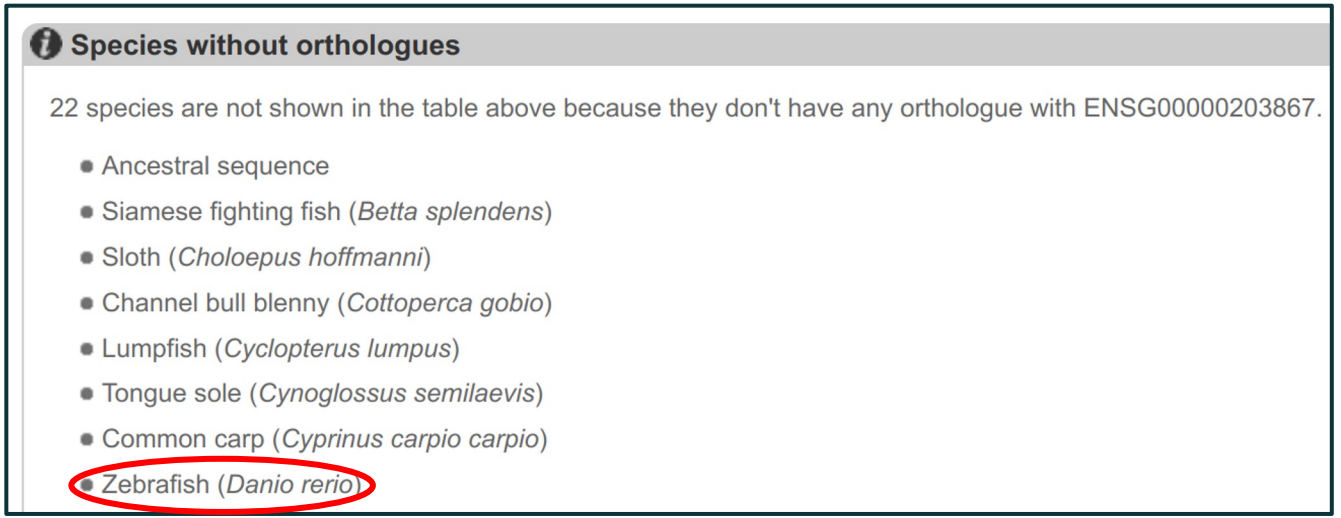
Homologues labelled “**high confidence**” are supported by conservation of synteny or whole genome alignment blocks

Compara - Whole Genome Alignments

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

Synteny Example

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



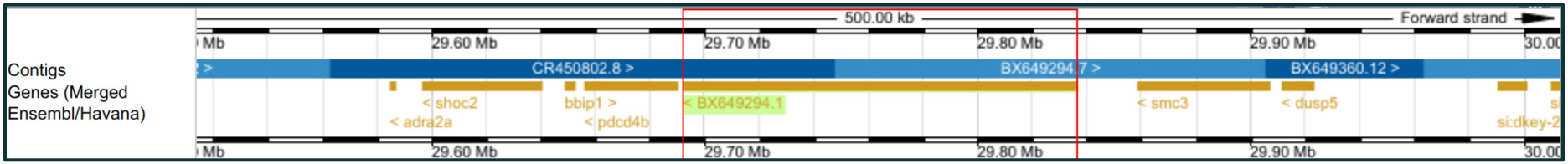
Synteny Example

- 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

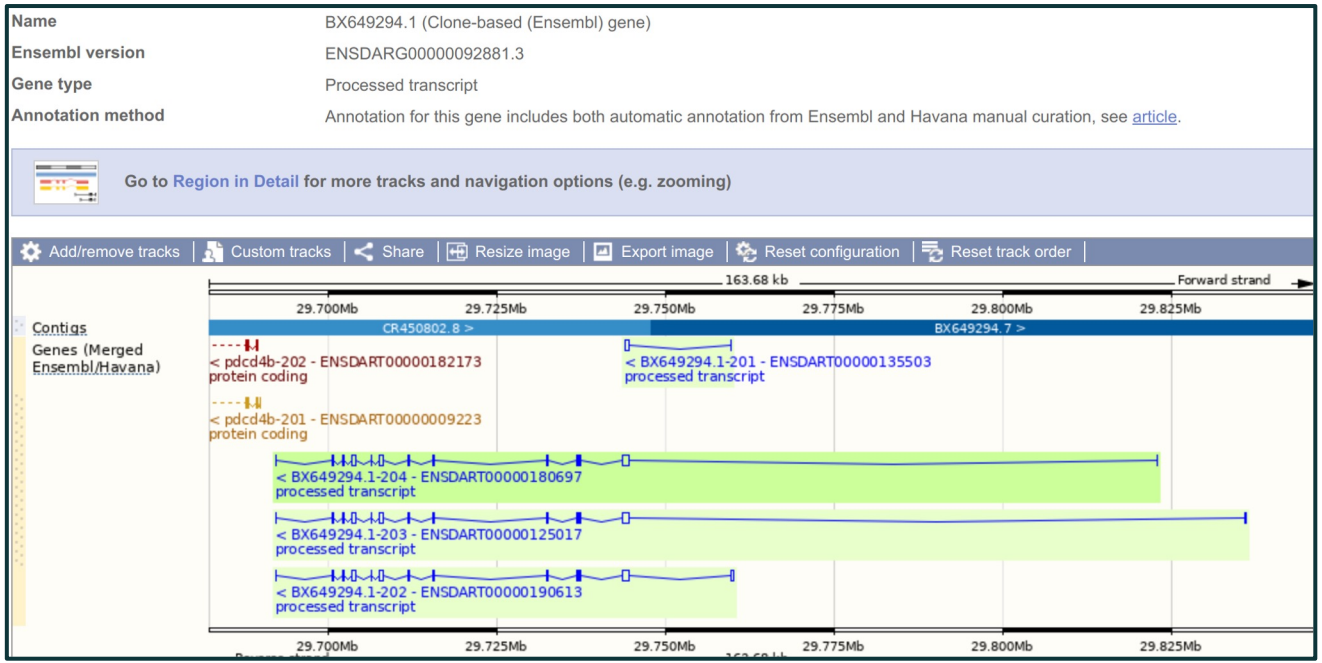
- If we look at the chr22 region in zebrafish then all the surrounding genes are the same and RBM20 is likely to be BX649294.1



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

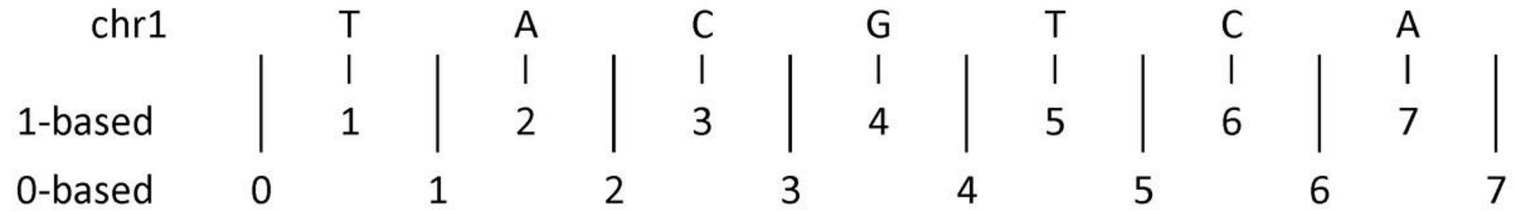
Synteny Example

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



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

Thank You!

Any questions?