Module 1

Understanding and Exploring Genome Assemblies

Kerstin Howe
Wellcome Trust Sanger Institute
zfish-help@sanger.ac.uk

Genome assembly generation

- The Genome Reference Consortium
- Generating Assemblies
- Working With Assemblies

Browsing genome assemblies

- Genome browsing
- The Ensembl gene set
- Guided examples
- [Make your own data visible: BED files]

Genome assembly generation

- The Genome Reference Consortium
- Generating Assemblies
- Working With Assemblies

Genome Reference Consortium
genomereference.org

International consortium looking after
human, mouse and zebrafish reference assemblies

- maintaining reference assemblies
- improving reference assemblies
- adding variation
- All issues documented on website

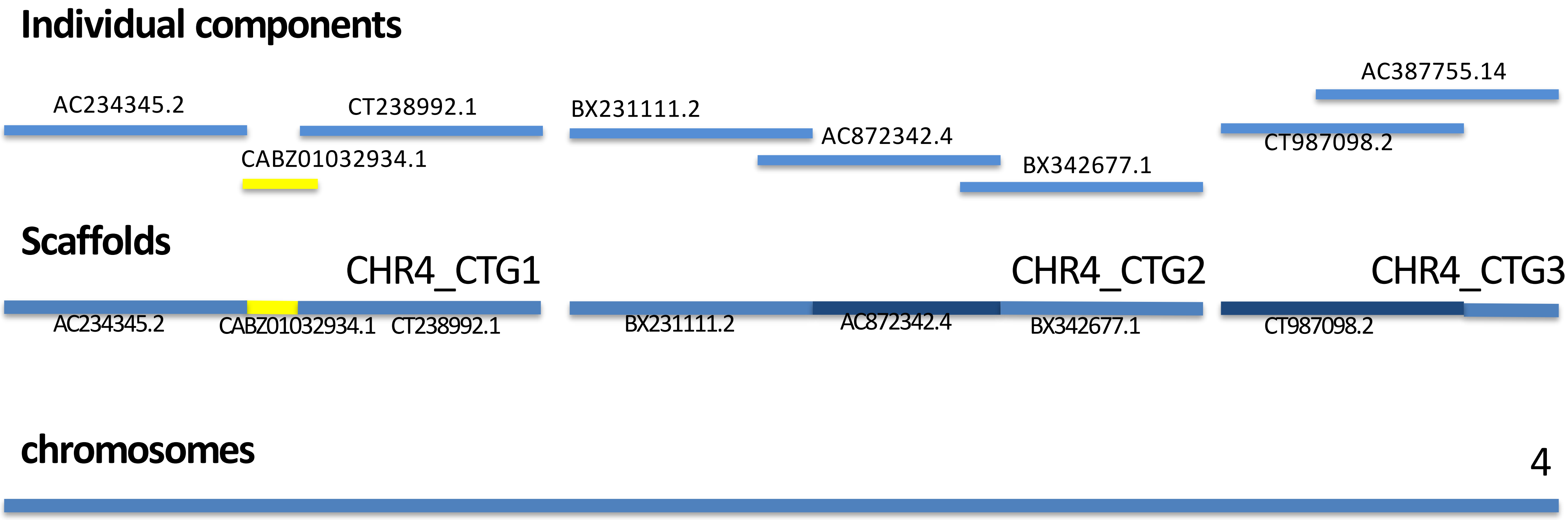
GRCz10 released August 2014
GRCz11 released May 2017

Within GRC:

- Handover from Sanger Institute to ZFIN after GRCz11
- Curation now only in reaction to user enquiries!

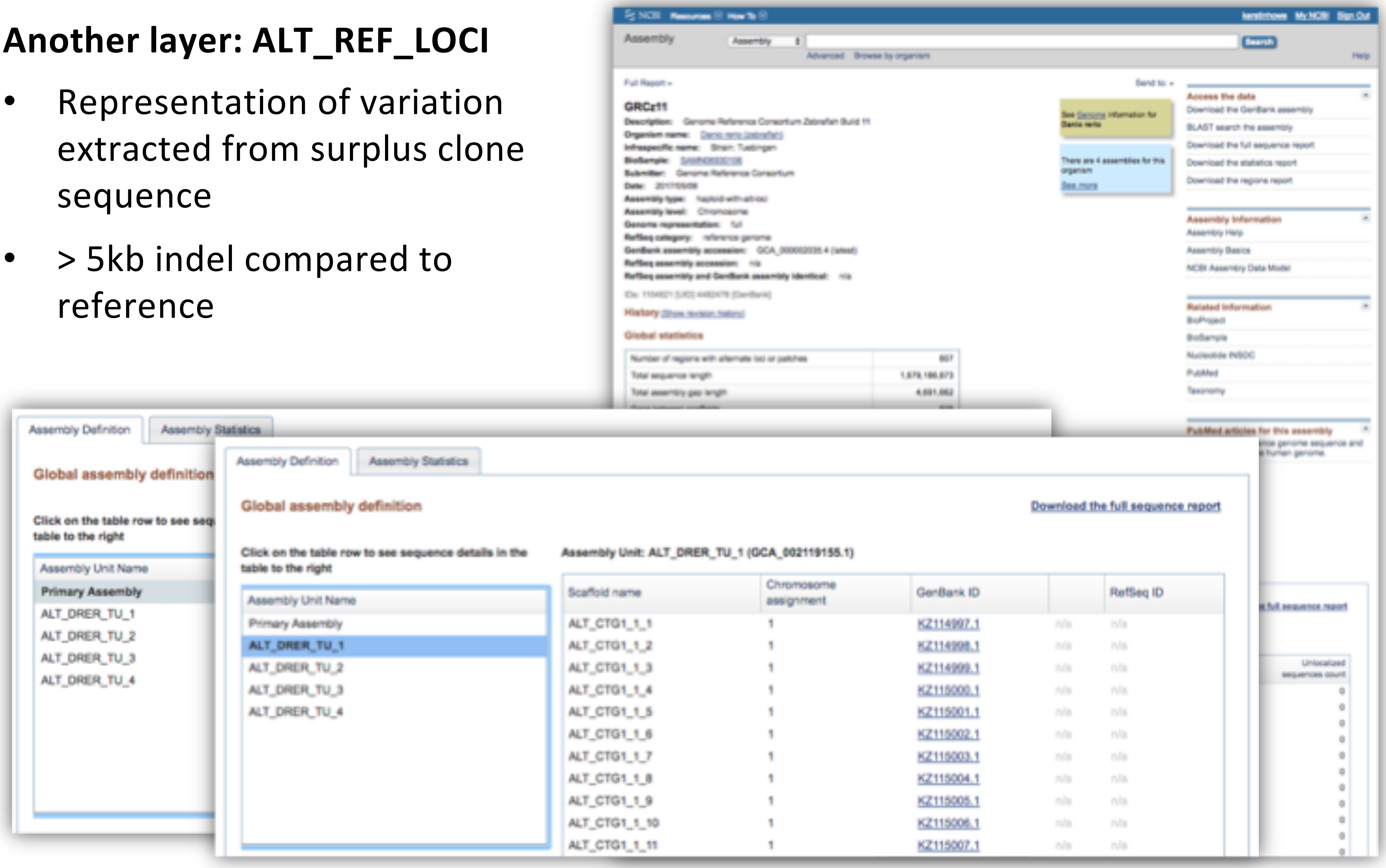
Generating zebrafish assemblies

- Restriction analysis (Fingerprint) and clone contig building (FPC map)
- Meiotic and RH maps
- Fill in WGS contigs
- Check and adjust with additional data (e.g. BioNano maps, Strandseq)



Another layer: ALT_REF_LOCI

- Representation of variation extracted from surplus clone sequence
- > 5kb indel compared to reference



GRCz11

Description: Genome Reference Consortium Zebrafish Build 11
Organism name: Danio rerio (zebrafish)
Infraspecific name: Strain: Tuebingen
BioSample: SAMN06930106
Submitter: Genome Reference Consortium
Date: 2017/05/09
Assembly type: haploid-with-alt-loci
Assembly level: Chromosome
Genome representation: full
RefSeq category: reference genome
GenBank assembly accession: GCA_000002035.4 (latest)
RefSeq assembly accession: n/a
RefSeq assembly and GenBank assembly identical: n/a

IDs: 1104621 [UID] 4482478 [GenBank]

History (Show revision history)

Global statistics

Number of regions with alternate loci or patches	607
Total sequence length	1,679,186,873
Total assembly gap length	4,691,662

See Genome information for Danio rerio

There are 4 assemblies for this organism
See more

Access the data
Download the GenBank assembly
BLAST search the assembly
Download the full sequence report
Download the statistics report
Download the regions report

Assembly Information
Assembly Help
Assembly Basics
NCBI Assembly Data Model

Related Information
BioProject
BioSample
Nucleotide INSDC
PubMed
Taxonomy

PubMed articles for this assembly

Global assembly definition

Download the full sequence report

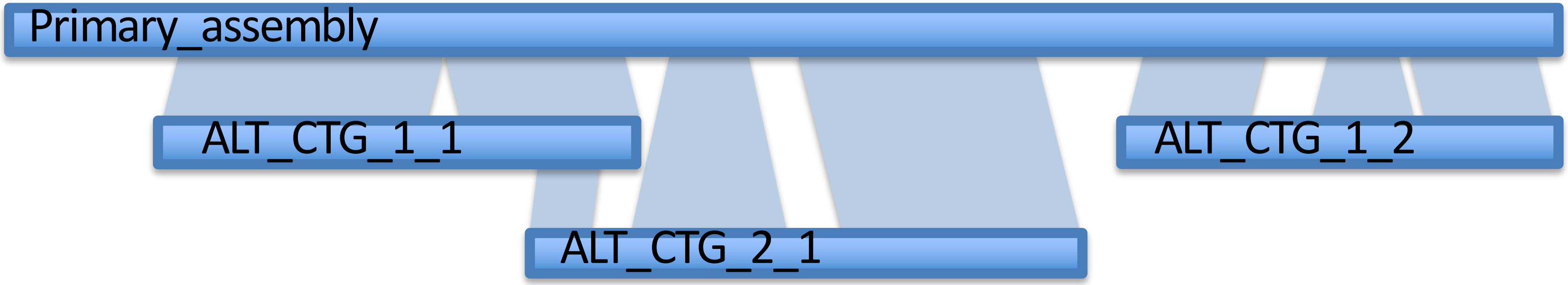
Click on the table row to see sequence details in the table to the right

Assembly Unit Name
Primary Assembly
ALT_DRER_TU_1
ALT_DRER_TU_2
ALT_DRER_TU_3
ALT_DRER_TU_4

Assembly Unit: ALT_DRER_TU_1 (GCA_002119155.1)

Scaffold name	Chromosome assignment	GenBank ID		RefSeq ID
ALT_CTG1_1_1	1	KZ114997.1	n/a	n/a
ALT_CTG1_1_2	1	KZ114998.1	n/a	n/a
ALT_CTG1_1_3	1	KZ114999.1	n/a	n/a
ALT_CTG1_1_4	1	KZ115000.1	n/a	n/a
ALT_CTG1_1_5	1	KZ115001.1	n/a	n/a
ALT_CTG1_1_6	1	KZ115002.1	n/a	n/a
ALT_CTG1_1_7	1	KZ115003.1	n/a	n/a
ALT_CTG1_1_8	1	KZ115004.1	n/a	n/a
ALT_CTG1_1_9	1	KZ115005.1	n/a	n/a
ALT_CTG1_1_10	1	KZ115006.1	n/a	n/a
ALT_CTG1_1_11	1	KZ115007.1	n/a	n/a

ALT_REF_LOCI

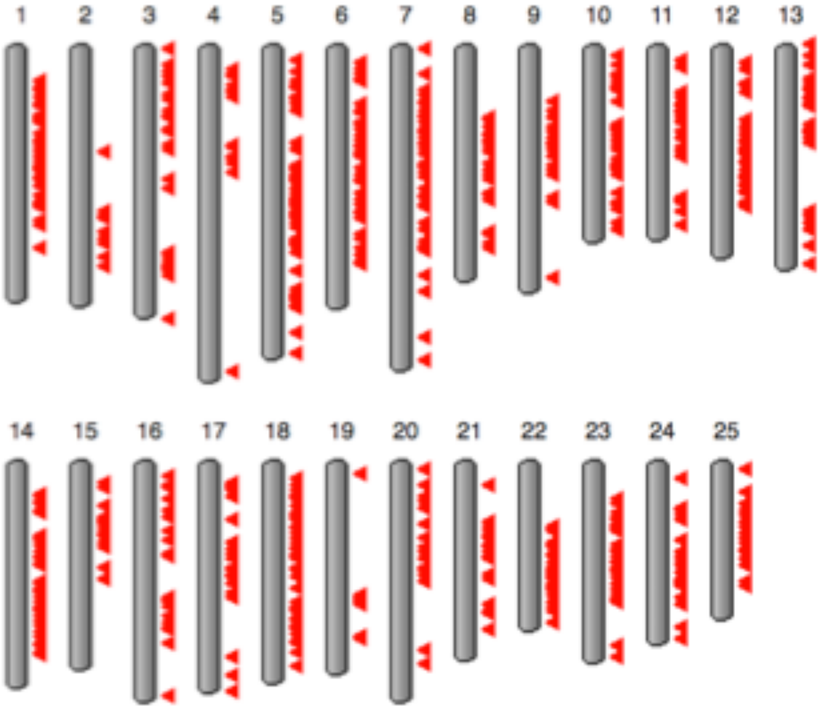


Primary_assembly

ALT_CTG_1_1

ALT_CTG_1_2

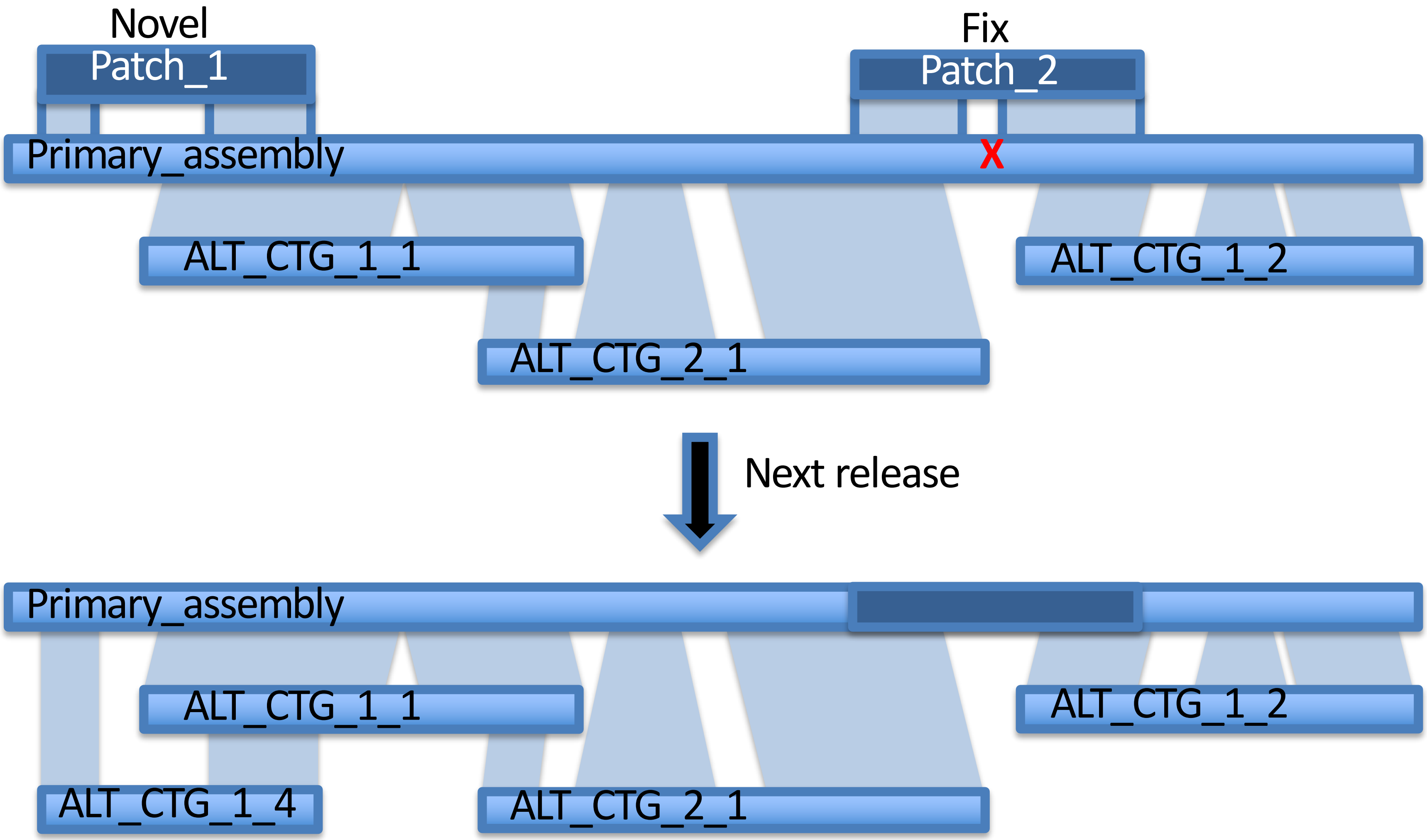
ALT_CTG_2_1



930 ALT_REF_LOCI for GRCz11

Generating Assemblies

…and another layer: Patches



Working With Assemblies

You might want to

- Find a clone that covers a feature
- Confirm that a region is correctly positioned
- Find out whether a gap could be closed
- Check whether a gene is really duplicated

Help is at hand

- zfish-help@sanger.ac.uk
- Ensembl
- gEVAL geval.sanger.ac.uk

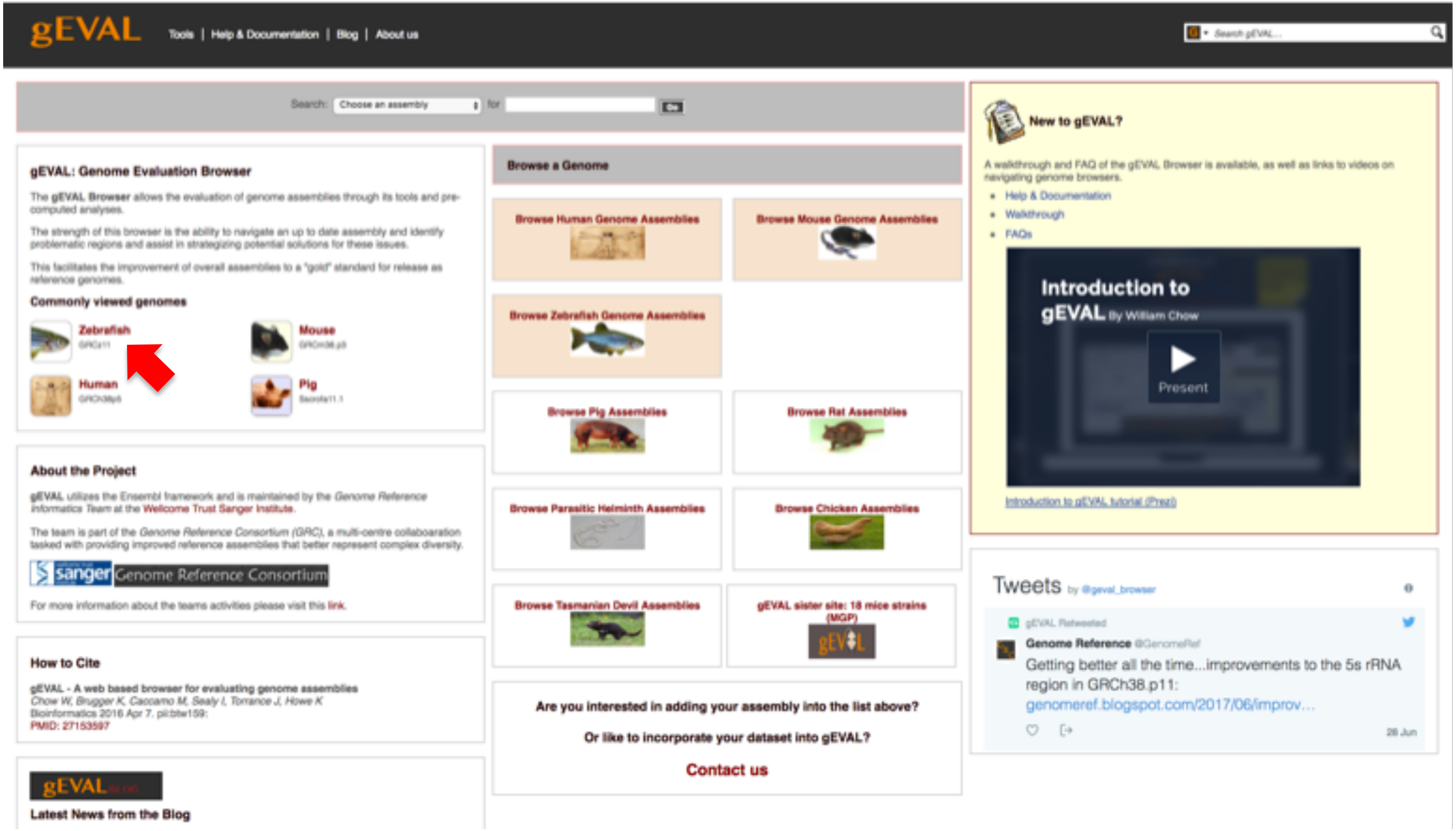
gEVAL

geval.sanger.ac.uk

- genome evaluation browser
- Numerous assembly versions
- alignments of BAC/FOS ends, markers, optical maps, cDNAs, other genome assemblies, etc. to check consistency in the assembly
- reports GRC investigations
- offers ‘punchlists’ denoting issues with an assembly
- extensive documentation
- featured GRCz11 from release date

Module 1: Understanding and Exploring Genome Assemblies

geval.sanger.ac.uk


gEVAL
Tools | Help & Documentation | Blog | About us
Search gEVAL...
Search: Choose an assembly for
Go
Present

gEVAL: Genome Evaluation Browser

The **gEVAL Browser** allows the evaluation of genome assemblies through its tools and pre-computed analyses.

The strength of this browser is the ability to navigate an up to date assembly and identify problematic regions and assist in strategizing potential solutions for these issues.

This facilitates the improvement of overall assemblies to a "gold" standard for release as reference genomes.

Commonly viewed genomes

Zebrafish GRCz11	Mouse GRCm38.p5
Human GRCh38p5	Pig Sscrofa11.1

About the Project

gEVAL utilizes the Ensembl framework and is maintained by the *Genome Reference Informatics Team* at the Wellcome Trust Sanger Institute.

The team is part of the *Genome Reference Consortium (GRC)*, a multi-centre collaboaration tasked with providing improved reference assemblies that better represent complex diversity.

sanger Genome Reference Consortium

For more information about the teams activities please visit this link.

How to Cite

gEVAL - A web based browser for evaluating genome assemblies
Chow W, Brugger K, Caccamo M, Sealy I, Torrance J, Howe K
Bioinformatics 2016 Apr 7. pii:btw159:
PMID: 27153597

gEVAL BLOG

Latest News from the Blog

Browse a Genome

- Browse Human Genome Assemblies
- Browse Mouse Genome Assemblies
- Browse Zebrafish Genome Assemblies
- Browse Pig Assemblies
- Browse Rat Assemblies
- Browse Parasitic Helminth Assemblies
- Browse Chicken Assemblies
- Browse Tasmanian Devil Assemblies
- gEVAL sister site: 18 mice strains (MGP)

Are you interested in adding your assembly into the list above?

Or like to incorporate your dataset into gEVAL?

Contact us

New to gEVAL?

A walkthrough and FAQ of the gEVAL Browser is available, as well as links to videos on navigating genome browsers.

- Help & Documentation
- Walkthrough
- FAQs

Introduction to gEVAL By William Chow

Introduction to gEVAL tutorial (Prezi)

Tweets by @geval_browser

gEVAL Retweeted

Genome Reference @GenomeRef

Getting better all the time...improvements to the 5s rRNA region in GRCh38.p11:
genomeref.blogspot.com/2017/06/improv...

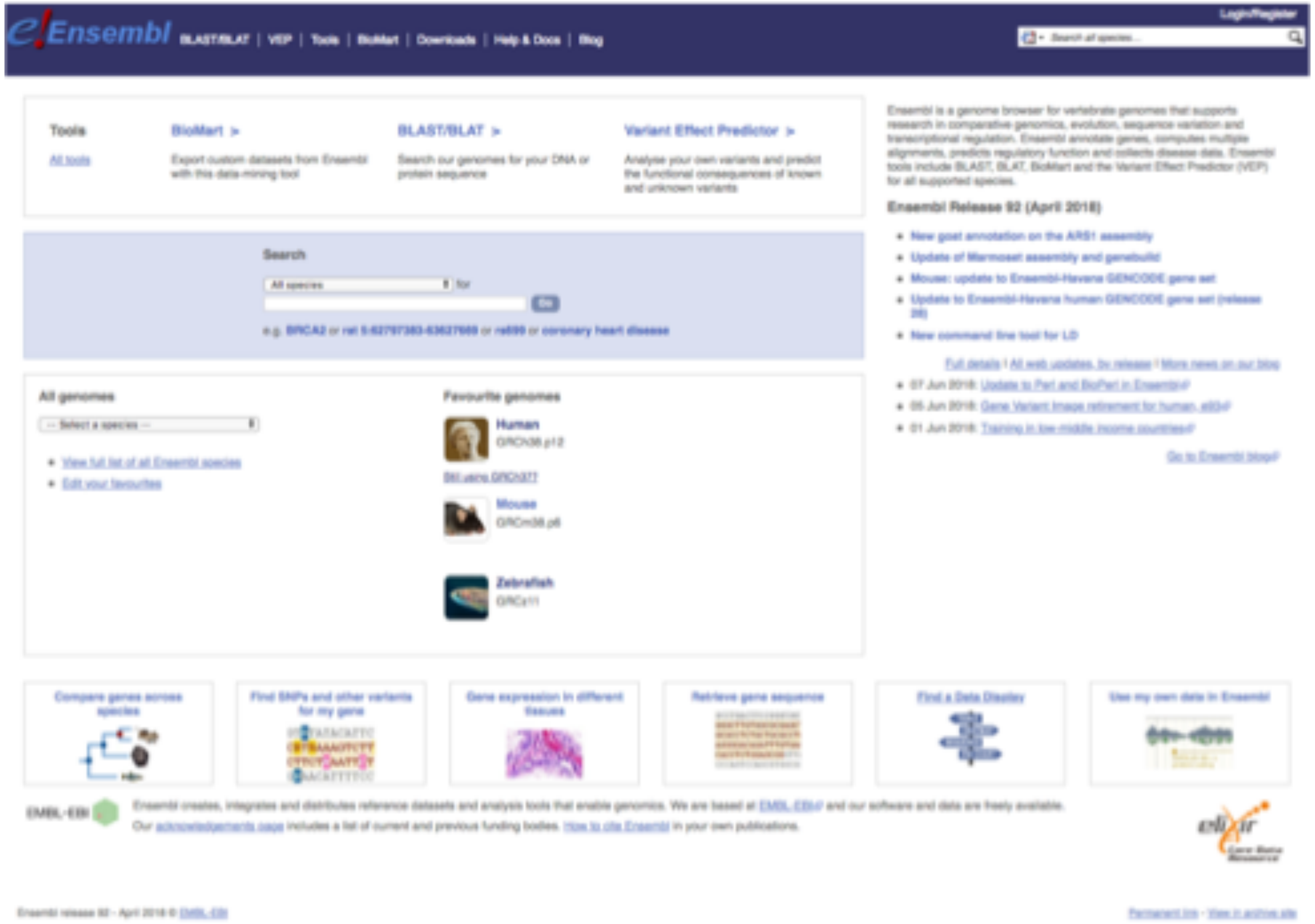
28 Jun

Browsing genome assemblies

- Genome browsing
- The Ensembl gene set
- Guided examples
- [Make your own data visible: BED files]

Genome Browsers

- NCBI Map Viewer http://www.ncbi.nlm.nih.gov/mapview/
- UCSC Genome Browser http://genome.ucsc.edu/
- Ensembl Genome Browser http://www.ensembl.org/



e!Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
Search all species...

Tools
All tools

BioMart >
Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >
Search our genomes for your DNA or protein sequence

Variant Effect Predictor >
Analyse your own variants and predict the functional consequences of known and unknown variants

Search
All species for
Go
e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes
-- Select a species --
- View full list of all Ensembl species
- Edit your favourites

Favourite genomes
Human GRCh38.p12
Still using GRCh37?
Mouse GRCm38.p6
Zebrafish GRCz11

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 92 (April 2018)

- New goat annotation on the ARS1 assembly
- Update of Marmoset assembly and genebuild
- Mouse: update to Ensembl-Havana GENCODE gene set
- Update to Ensembl-Havana human GENCODE gene set (release 28)
- New command line tool for LD

Full details | All web updates, by release | More news on our blog

- 07 Jun 2018: Update to Perl and BioPerl in Ensembl
- 05 Jun 2018: Gene Variant Image retirement for human, e93
- 01 Jun 2018: Training in low-middle income countries

Go to Ensembl blog

Compare genes across species | Find SNPs and other variants for my gene | Gene expression in different tissues | Retrieve gene sequence | Find a Data Display | Use my own data in Ensembl

EMBL-EBI Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available. Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

Ensembl release 92 - April 2018 © EMBL-EBI

Permanent link - View in archive site

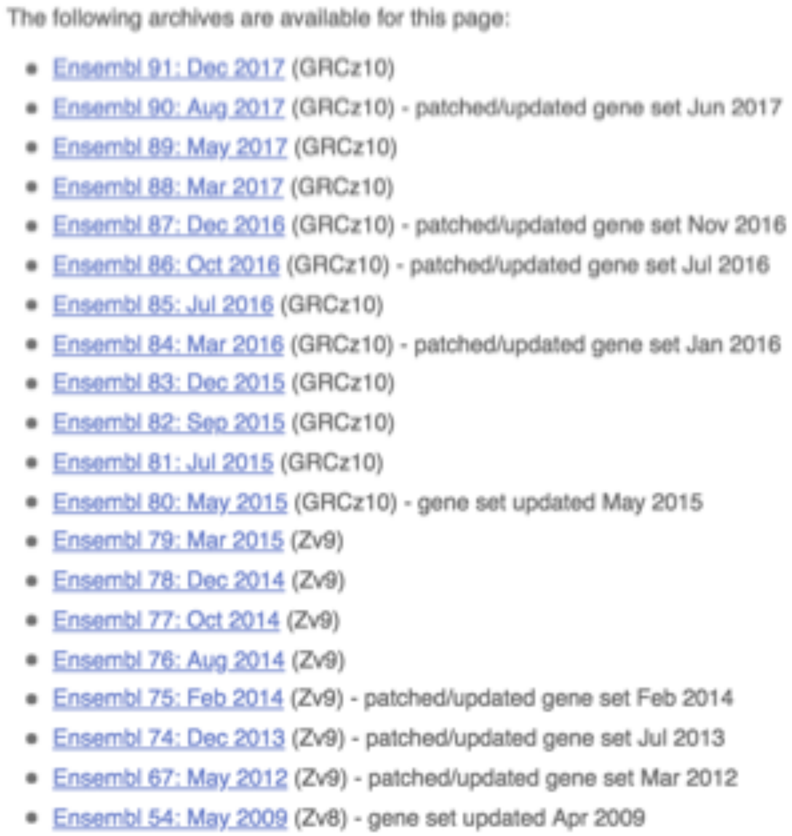
Caveat: check the assembly version!

Ensembl : What annotation is available?

Genes	protein coding genes gene/transcript/peptide models (coding and non-coding)
Comparative data	orthologues and paralogues, gene trees protein families whole genome alignments syntenic regions
Variation data	Single Nucleotide Polymorphisms (SNPs), indels, phenotypes, population frequencies, variant effect prediction (VEP), etc.
Regulatory data	e.g. regulatory elements from ENCODE
IDs	crossreferences to other databases
aligned data	cDNAs, RNAseq, peptides, micro array probes, BAC clones, etc.
Cytogenetic bands, markers, repeats etc.	
External resources	e.g. GRC trackhub, mapped next-gen reads

The Ensembl gene set

- **Bimonthly releases, updated gene set ~ every 6 months**
- **New 'Genebuild' with every new assembly**

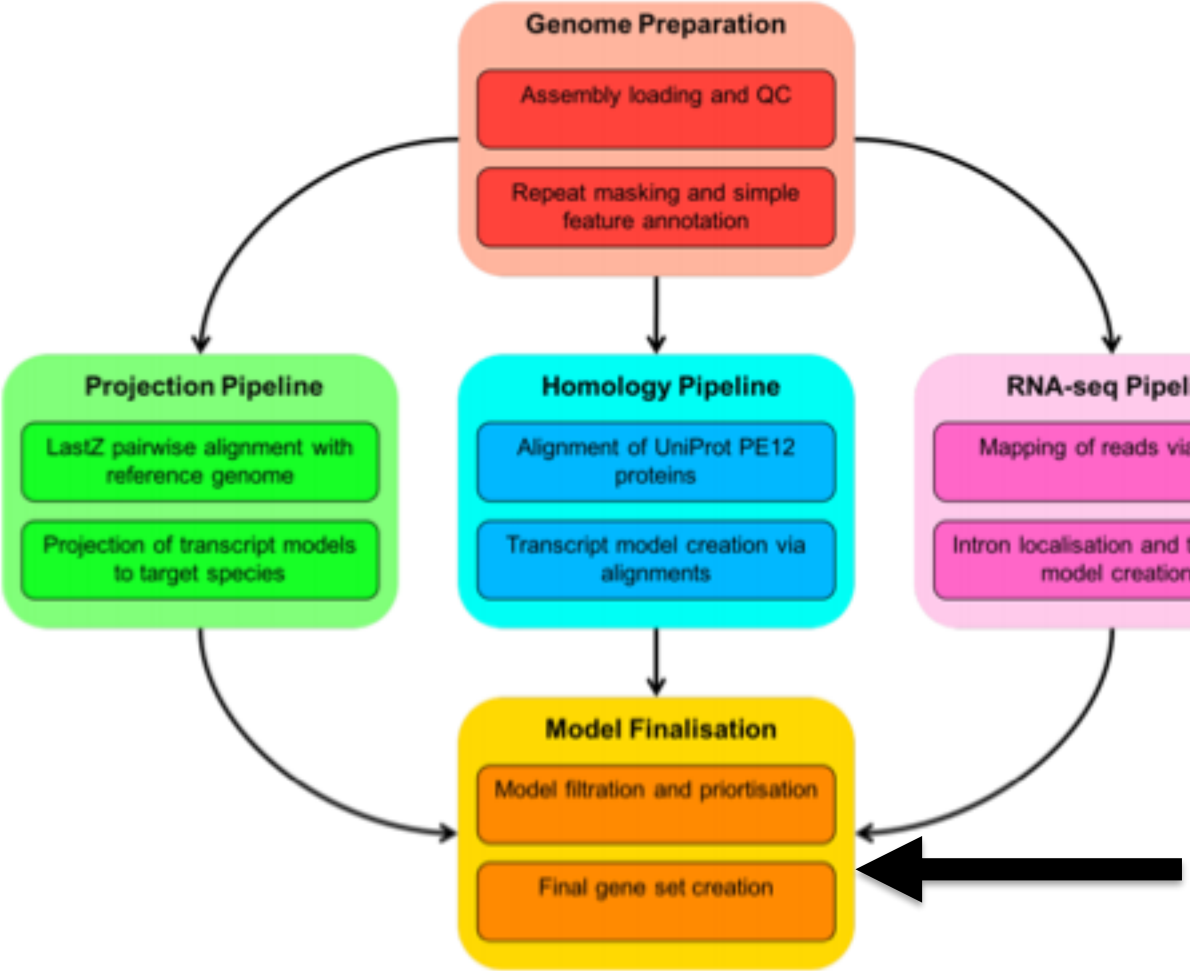


The following archives are available for this page:

- Ensembl 91: Dec 2017 (GRCz10)
- Ensembl 90: Aug 2017 (GRCz10) - patched/updated gene set Jun 2017
- Ensembl 89: May 2017 (GRCz10)
- Ensembl 88: Mar 2017 (GRCz10)
- Ensembl 87: Dec 2016 (GRCz10) - patched/updated gene set Nov 2016
- Ensembl 86: Oct 2016 (GRCz10) - patched/updated gene set Jul 2016
- Ensembl 85: Jul 2016 (GRCz10)
- Ensembl 84: Mar 2016 (GRCz10) - patched/updated gene set Jan 2016
- Ensembl 83: Dec 2015 (GRCz10)
- Ensembl 82: Sep 2015 (GRCz10)
- Ensembl 81: Jul 2015 (GRCz10)
- Ensembl 80: May 2015 (GRCz10) - gene set updated May 2015
- Ensembl 79: Mar 2015 (Zv9)
- Ensembl 78: Dec 2014 (Zv9)
- Ensembl 77: Oct 2014 (Zv9)
- Ensembl 76: Aug 2014 (Zv9)
- Ensembl 75: Feb 2014 (Zv9) - patched/updated gene set Feb 2014
- Ensembl 74: Dec 2013 (Zv9) - patched/updated gene set Jul 2013
- Ensembl 67: May 2012 (Zv9) - patched/updated gene set Mar 2012
- Ensembl 54: May 2009 (Zv8) - gene set updated Apr 2009

- **Genes are built on evidence, no gene is predicted on sequence alone!**

The Ensembl gene build



Genome Preparation

Assembly loading and QC

Repeat masking and simple feature annotation

Projection Pipeline

LastZ pairwise alignment with reference genome

Projection of transcript models to target species

Homology Pipeline

Alignment of UniProt PE12 proteins

Transcript model creation via alignments

RNA-seq Pipeline

Mapping of reads via BWA

Intron localisation and transcript model creation

Model Finalisation

Model filtration and priortisation

Final gene set creation

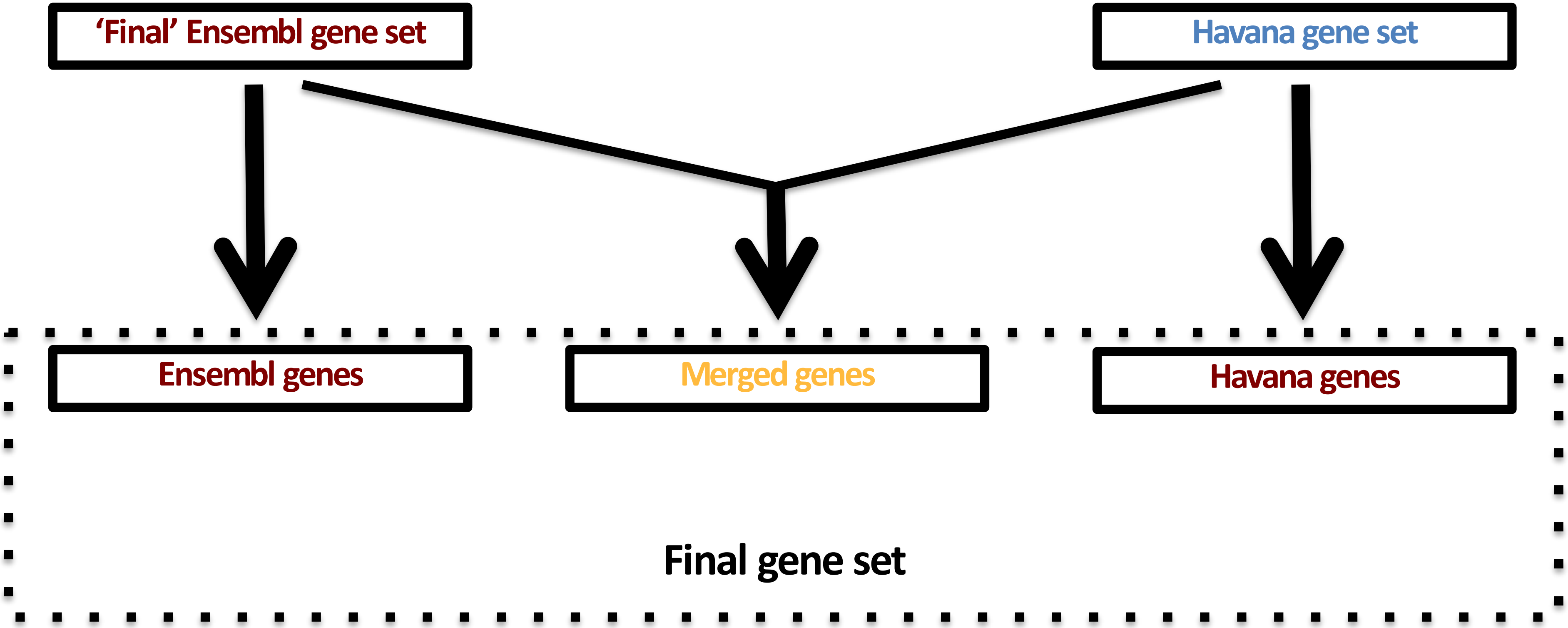
Addtion of

- UTRs
- alternative transcripts
- Pseudogenes
- Immune genes

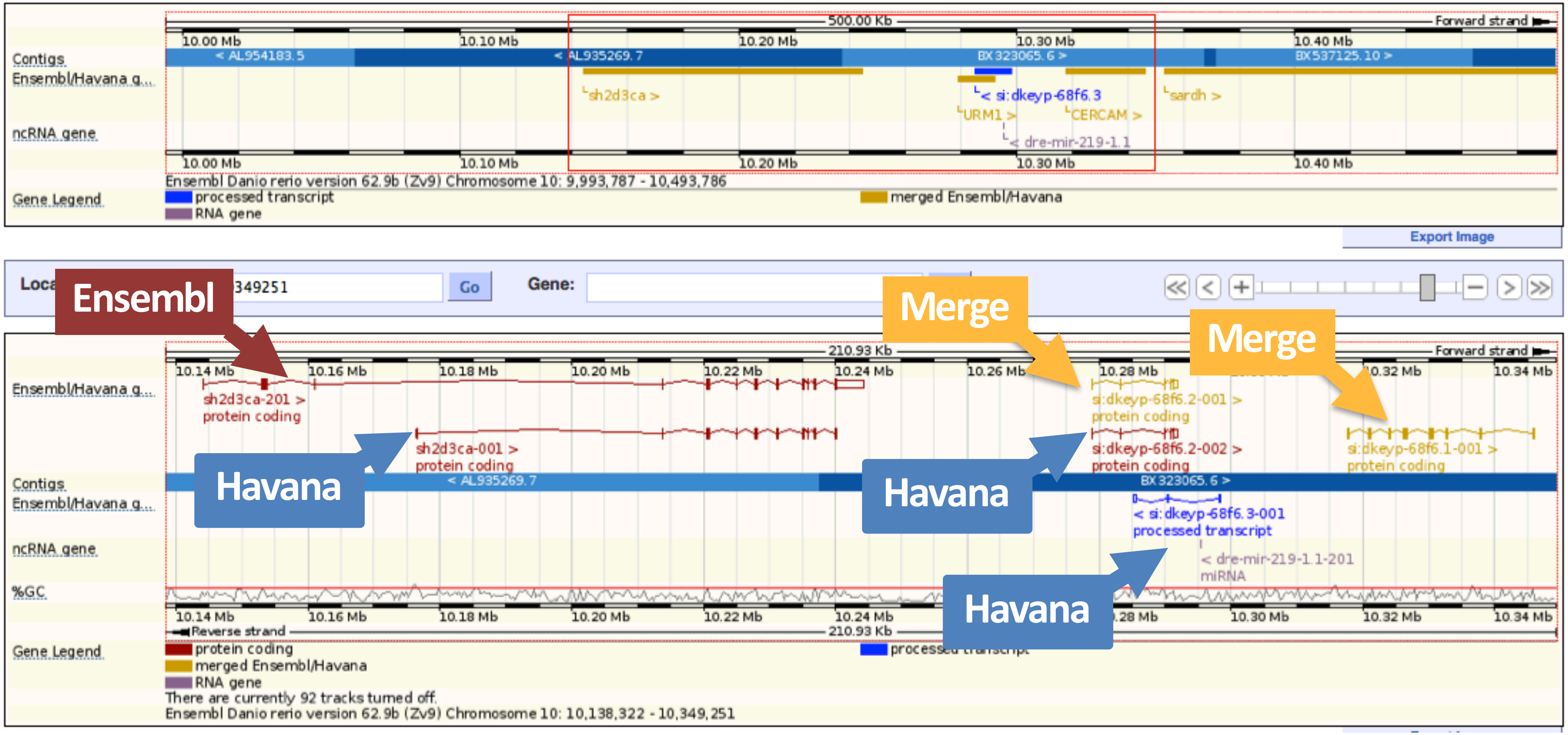
From: http://www.ensembl.org/info/genome/genebuild/2018_03_zebrafish_genebuild.pdf

The Ensembl gene build : Gene merge

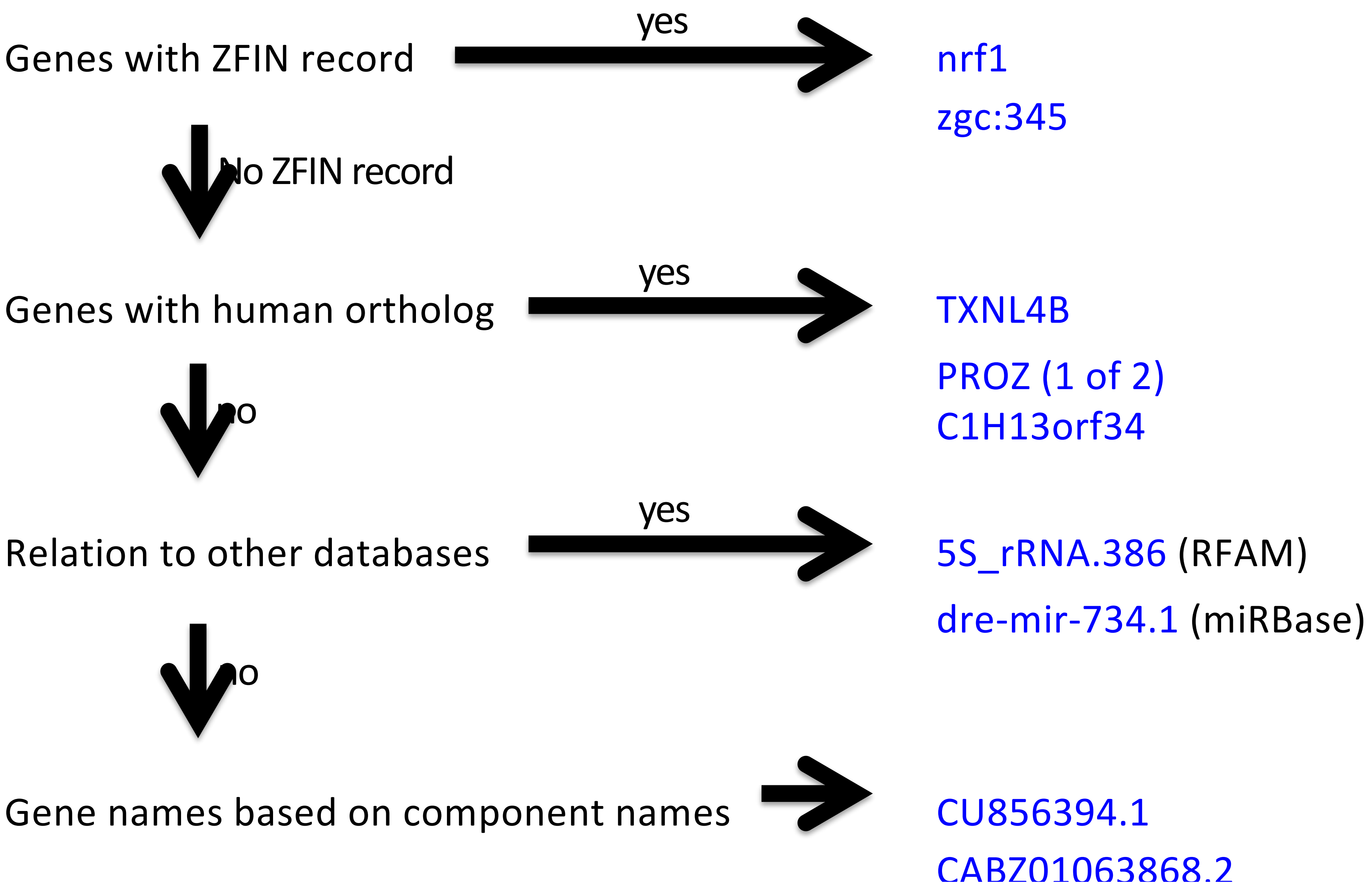
Getting the optimum gene set by combining automated and manual annotation



The Ensembl gene build : Gene merge



Ensembl gene names



Genes with ZFIN record

yes

nrf1

zgc:345

No ZFIN record

Genes with human ortholog

yes

TXNL4B

PROZ (1 of 2)

C1H13orf34

no

Relation to other databases

yes

5S_rRNA.386 (RFAM)

dre-mir-734.1 (miRBase)

no

Gene names based on component names

CU856394.1

CABZ01063868.2

Access to Genome Annotation

• Release web site	http://www.ensembl.org/
• Archive	http://archive.ensembl.org
• BioMart	http://www.ensembl.org/biomart/martview
• Downloads	http://www.ensembl.org/info/data/ftp/index.html
• Perl API	http://www.ensembl.org/info/docs/index.html
• REST API	

Help and Information

- Zebrafish specific help zfish-help@sanger.ac.uk / zfinadmin@zfin.org
- Zebrafish genome project at Sanger Institute www.sanger.ac.uk/science/data/zebrafish-genome-project
- Zebrafish at the GRC genomereference.org
- Zebrafish genome assembly evaluation geval.sanger.ac.uk
- Ensembl helpdesk helpdesk@ensembl.org
- View animated tutorials www.ensembl.org/info/website/tutorials/index.html
- Mailing lists: announce@ensembl.org
  ensembl-dev@ensembl.org

Guided examples

- A stroll through Ensembl - abbreviated
- Introduction to BioMart - abbreviated
- Making your own data visible (in your own time)