# Zebrafish Open Door Workshop – Madison 2018

# Module 2: Using Ensembl to explore zebrafish data

Toby Hunt

Vertebrate Genomics (Ensembl-HAVANA)

EMBL-EBI

e!Ensembl

# Using Ensembl to explore zebrafish data – overview

2a: Comparative Genomics

2b: Sequence Variation and Disease

Ensembl havana

human and vertebrate analysis and annotation

Havana: Human and vertebrate analysis and annotation

# The goal of GENCODE

“Our goal is to identify and classify ~~all gene feature~~s in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation”.

https://www.gencodegenes.org/

# The HAVANA team

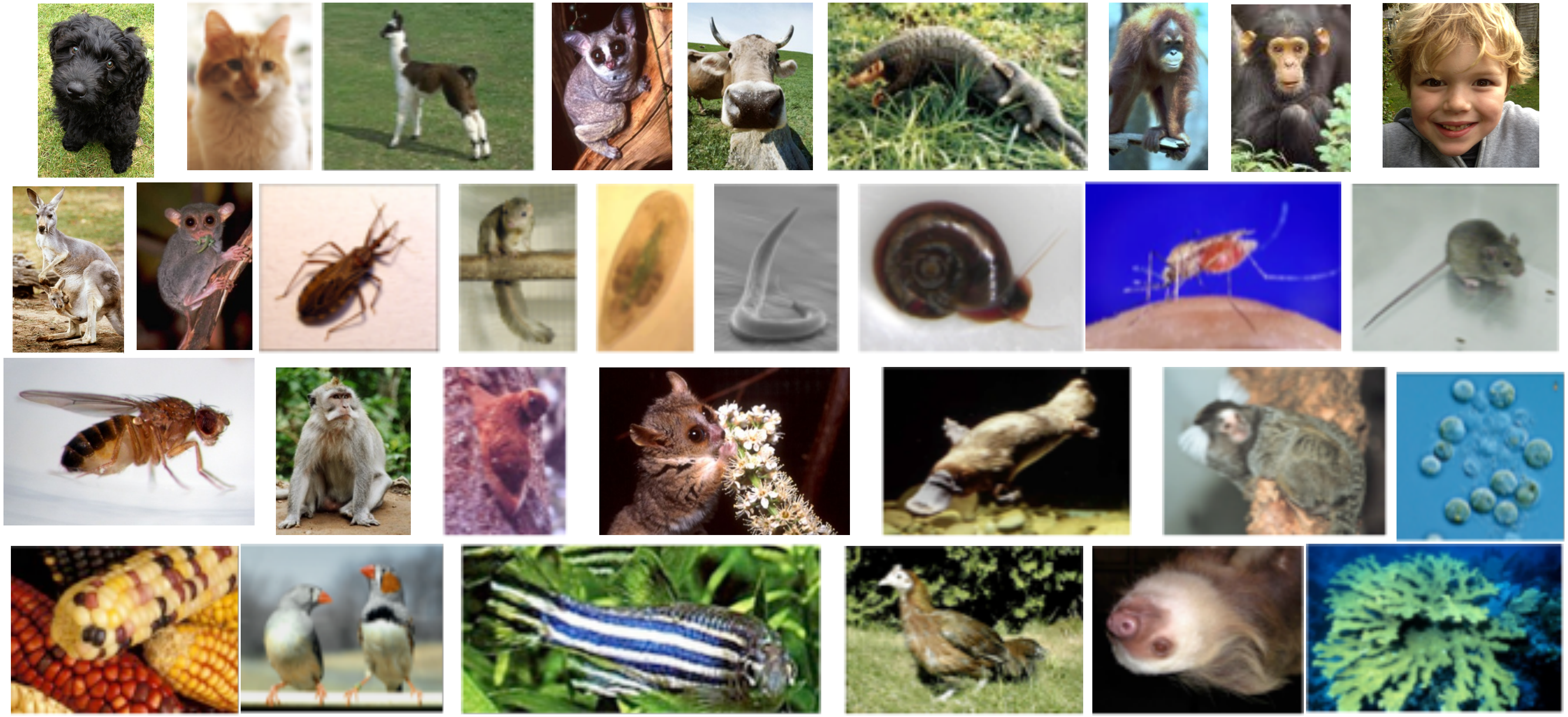
GENCODE

Whole Genome or chromosome

Targeted regions or genes

Community projects

# Comparative Sequence Analysis



# Comparative Sequence Analysis

A tool for decoding genomic information as it is based upon the tenet that:

Functional sequences evolve more slowly than non-functional sequences, therefore sequences that remain conserved throughout evolution *may* perform a biological function.

*"Conservation as a proxy for function"*

# Identify Conserved Regions

## Aligning genome sequences

- Functionally conserved units may be conserved at the sequence level

- Evolutionary Conserved Regions (ECRs)

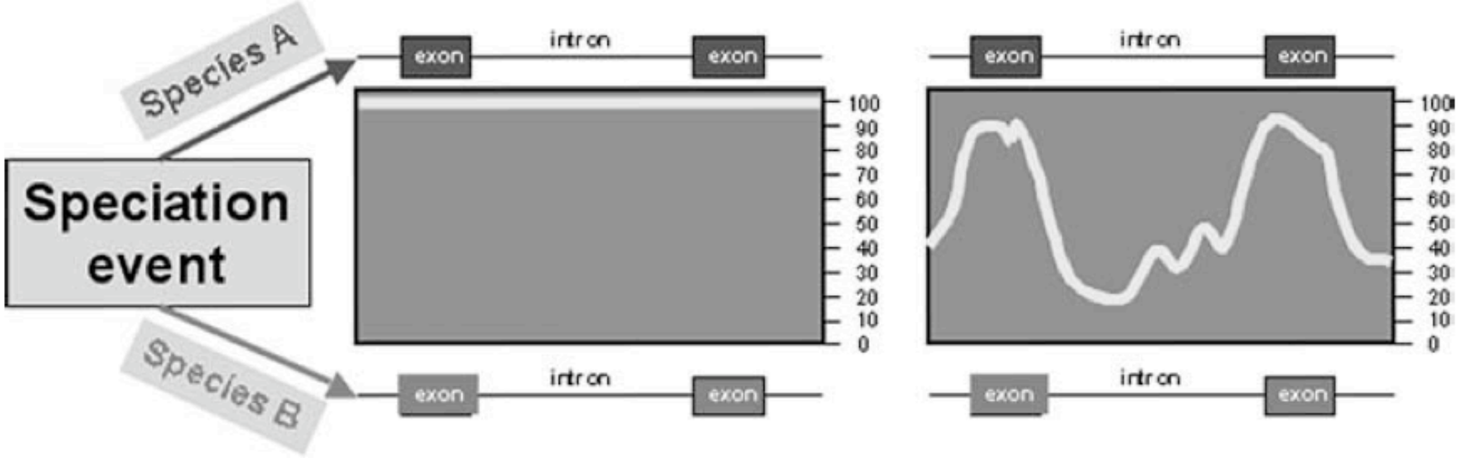


Fig 1. Miller *et al*, 2004. Ann Rev Genomics Hum Gen

# Why Comparative Sequence Analysis?

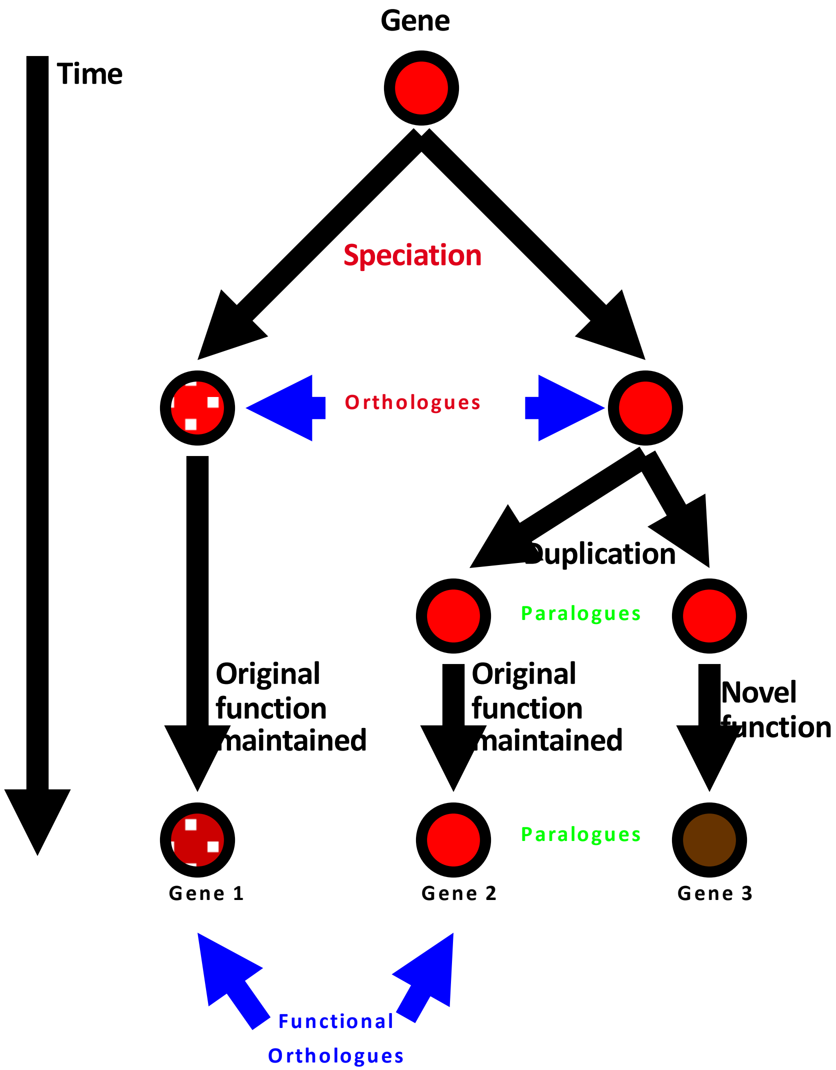
- allows us to achieve a greater understanding of vertebrate evolution

- tells us what is common and what is unique between different species at the genome level

- the function of human genes and other regions may be revealed by studying their counterparts in lower organisms

- helps identify both coding and non-coding genes and regulatory elements

# Homology, Orthology, Paralogy



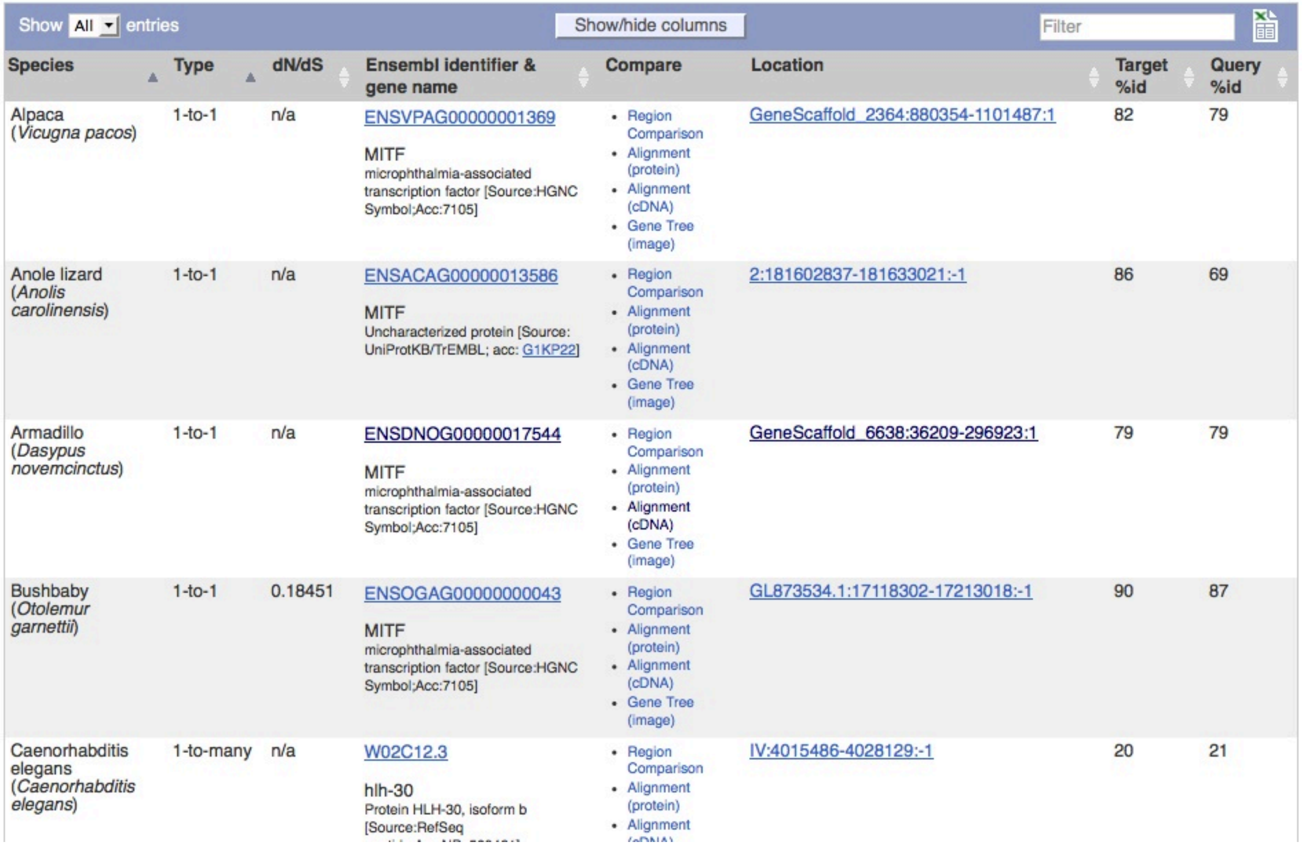
**Homologues** - Genes derived from common ancestral gene

**Orthologues** – Genes in different species that are derived from the same gene in last common ancestor

**Paralogues** – Gene families that have diverged within a single species, often by duplication

# Identifying Orthologous Genes

Orthologue Prediction at Ensembl:http://www.ensembl.org/



| Species | Type | dN/dS | Ensembl identifier & gene name | Compare | Location | Target %id | Query %id |
|---|---|---|---|---|---|---|---|
| Alpaca (*Vicugna pacos*) | 1-to-1 | n/a | ENSVPAG00000001369<br>MITF<br>microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA)<br>• Gene Tree (image) | GeneScaffold_2364:880354-1101487:1 | 82 | 79 |
| Anole lizard (*Anolis carolinensis*) | 1-to-1 | n/a | ENSACAG00000013586<br>MITF<br>Uncharacterized protein [Source: UniProtKB/TrEMBL; acc: G1KP22] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA)<br>• Gene Tree (image) | 2:181602837-181633021:-1 | 86 | 69 |
| Armadillo (*Dasypus novemcinctus*) | 1-to-1 | n/a | ENSDNOG00000017544<br>MITF<br>microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA)<br>• Gene Tree (image) | GeneScaffold_6638:36209-296923:1 | 79 | 79 |
| Bushbaby (*Otolemur garnettii*) | 1-to-1 | 0.18451 | ENSOGAG00000000043<br>MITF<br>microphthalmia-associated transcription factor [Source:HGNC Symbol;Acc:7105] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA)<br>• Gene Tree (image) | GL873534.1:17118302-17213018:-1 | 90 | 87 |
| Caenorhabditis elegans (*Caenorhabditis elegans*) | 1-to-many | n/a | W02C12.3<br>hlh-30<br>Protein HLH-30, isoform b [Source:RefSeq | • Region Comparison<br>• Alignment (protein)<br>• Alignment | IV:4015486-4028129:-1 | 20 | 21 |

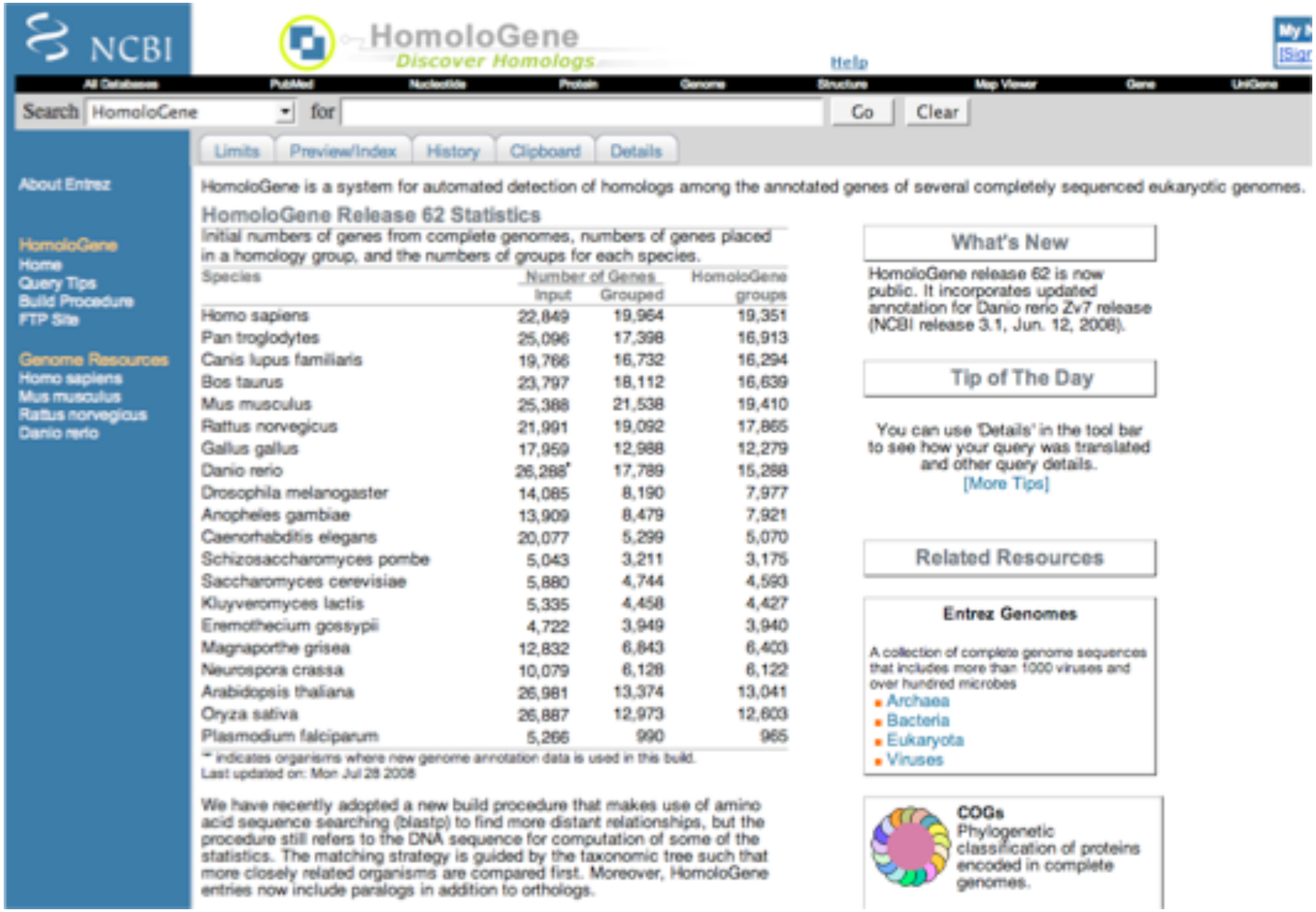
Links to the closest putative orthologous genes in other species

Hyperlinks to view alignments & positional information

# Identifying Orthologous Genes

**NCBI Homologene**

http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene&cmd



HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes.

### HomoloGene Release 62 Statistics

Initial numbers of genes from complete genomes, numbers of genes placed in a homology group, and the numbers of groups for each species.

| Species | Number of Genes | | HomoloGene |
|---|---|---|---|
| | Input | Grouped | groups |
| Homo sapiens | 22,849 | 19,964 | 19,351 |
| Pan troglodytes | 25,096 | 17,398 | 16,913 |
| Canis lupus familiaris | 19,766 | 16,732 | 16,294 |
| Bos taurus | 23,797 | 18,112 | 16,639 |
| Mus musculus | 25,388 | 21,538 | 19,410 |
| Rattus norvegicus | 21,991 | 19,092 | 17,865 |
| Gallus gallus | 17,959 | 12,988 | 12,279 |
| Danio rerio | 26,288* | 17,789 | 15,288 |
| Drosophila melanogaster | 14,085 | 8,190 | 7,977 |
| Anopheles gambiae | 13,909 | 8,479 | 7,921 |
| Caenorhabditis elegans | 20,077 | 5,299 | 5,070 |
| Schizosaccharomyces pombe | 5,043 | 3,211 | 3,175 |
| Saccharomyces cerevisiae | 5,880 | 4,744 | 4,593 |
| Kluyveromyces lactis | 5,335 | 4,458 | 4,427 |
| Eremothecium gossypii | 4,722 | 3,949 | 3,940 |
| Magnaporthe grisea | 12,832 | 6,843 | 6,403 |
| Neurospora crassa | 10,079 | 6,128 | 6,122 |
| Arabidopsis thaliana | 26,981 | 13,374 | 13,041 |
| Oryza sativa | 26,887 | 12,973 | 12,603 |
| Plasmodium falciparum | 5,266 | 990 | 965 |

* indicates organisms where new genome annotation data is used in this build.
Last updated on: Mon Jul 28 2008

We have recently adopted a new build procedure that makes use of amino acid sequence searching (blastp) to find more distant relationships, but the procedure still refers to the DNA sequence for computation of some of the statistics. The matching strategy is guided by the taxonomic tree such that more closely related organisms are compared first. Moreover, HomoloGene entries now include paralogs in addition to orthologs.

**What's New**

HomoloGene release 62 is now public. It incorporates updated annotation for Danio rerio Zv7 release (NCBI release 3.1, Jun. 12, 2008).

**Tip of The Day**

You can use 'Details' in the tool bar to see how your query was translated and other query details.
[More Tips]

**Related Resources**

**Entrez Genomes**

A collection of complete genome sequences that includes more than 1000 viruses and over hundred microbes

- Archaea
- Bacteria
- Eukaryota
- Viruses

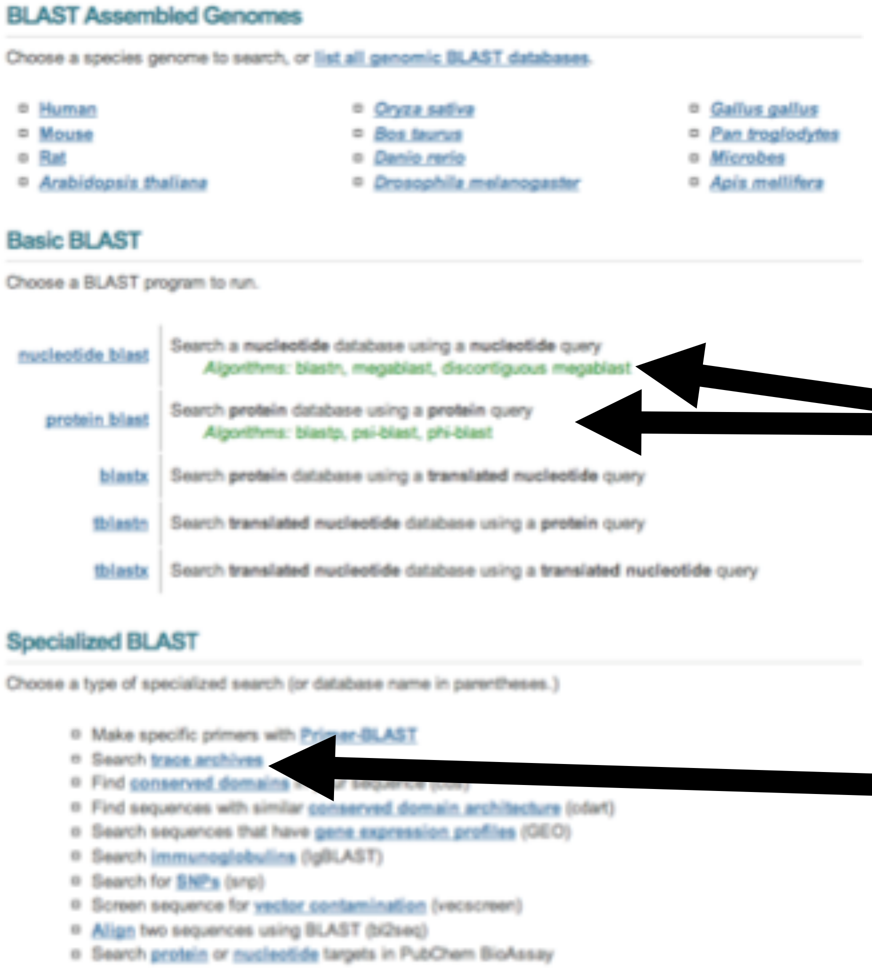
**COGs**
Phylogenetic classification of proteins encoded in complete genomes.

Contains a wealth of information about homologous genes and links to other resources

# Identifying Orthologous Genes

## BLAST searches

http://www.ncbi.nlm.nih.gov/BLAST/



### BLAST Assembled Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- Arabidopsis thaliana
- Oryza sativa
- Bos taurus
- Danio rerio
- Drosophila melanogaster
- Gallus gallus
- Pan troglodytes
- Microbes
- Apis mellifera

### Basic BLAST

Choose a BLAST program to run.

| | |
|---|---|
| nucleotide blast | Search a **nucleotide** database using a **nucleotide** query<br>Algorithms: blastn, megablast, discontiguous megablast |
| protein blast | Search **protein** database using a **protein** query<br>Algorithms: blastp, psi-blast, phi-blast |
| blastx | Search **protein** database using a **translated nucleotide** query |
| tblastn | Search **translated nucleotide** database using a **protein** query |
| tblastx | Search **translated nucleotide** database using a **translated nucleotide** query |

### Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

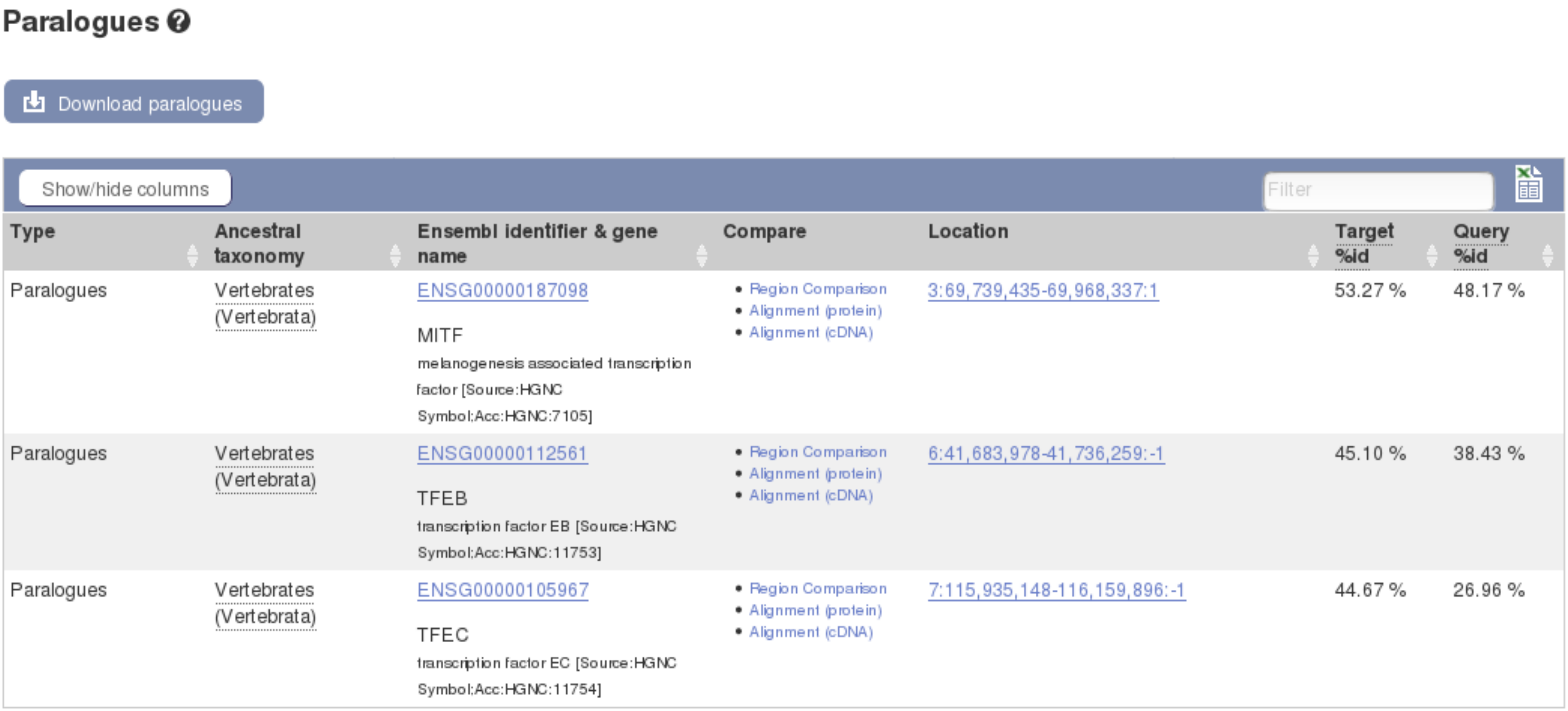
- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search for SNPs (snp)
- Screen sequence for vector contamination (vecscreen)
- Align two sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay

Species specific searches

Nucleotide or protein searches

Trace archives:
A good place to look
If you species of interest
doesn't have a browser

# Paralogues in Ensembl:



## Paralogues

| Type | Ancestral taxonomy | Ensembl identifier & gene name | Compare | Location | Target %id | Query %id |
|---|---|---|---|---|---|---|
| Paralogues | Vertebrates (Vertebrata) | ENSG00000187098<br>MITF<br>melanogenesis associated transcription factor [Source:HGNC Symbol;Acc:HGNC:7105] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA) | 3:69,739,435-69,968,337:1 | 53.27 % | 48.17 % |
| Paralogues | Vertebrates (Vertebrata) | ENSG00000112561<br>TFEB<br>transcription factor EB [Source:HGNC Symbol;Acc:HGNC:11753] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA) | 6:41,683,978-41,736,259:-1 | 45.10 % | 38.43 % |
| Paralogues | Vertebrates (Vertebrata) | ENSG00000105967<br>TFEC<br>transcription factor EC [Source:HGNC Symbol;Acc:HGNC:11754] | • Region Comparison<br>• Alignment (protein)<br>• Alignment (cDNA) | 7:115,935,148-116,159,896:-1 | 44.67 % | 26.96 % |

# How best to ensure that you have identified an orthologous gene

- Percentage identity (protein and nucleotide)

  (e.g. ClustalOmega, MUSCLE, sometimes Homologene)

- Compare the size and number of exons in orthologous genes

  (EST/cDNA to genomes – Splign , Ensembl ExonView)

- Positional information  - neighbouring genes

  (Ensembl– SyntenyView, UCSC, Genomicus )

- Confirm that no other paralogous genes are present in your species of interest

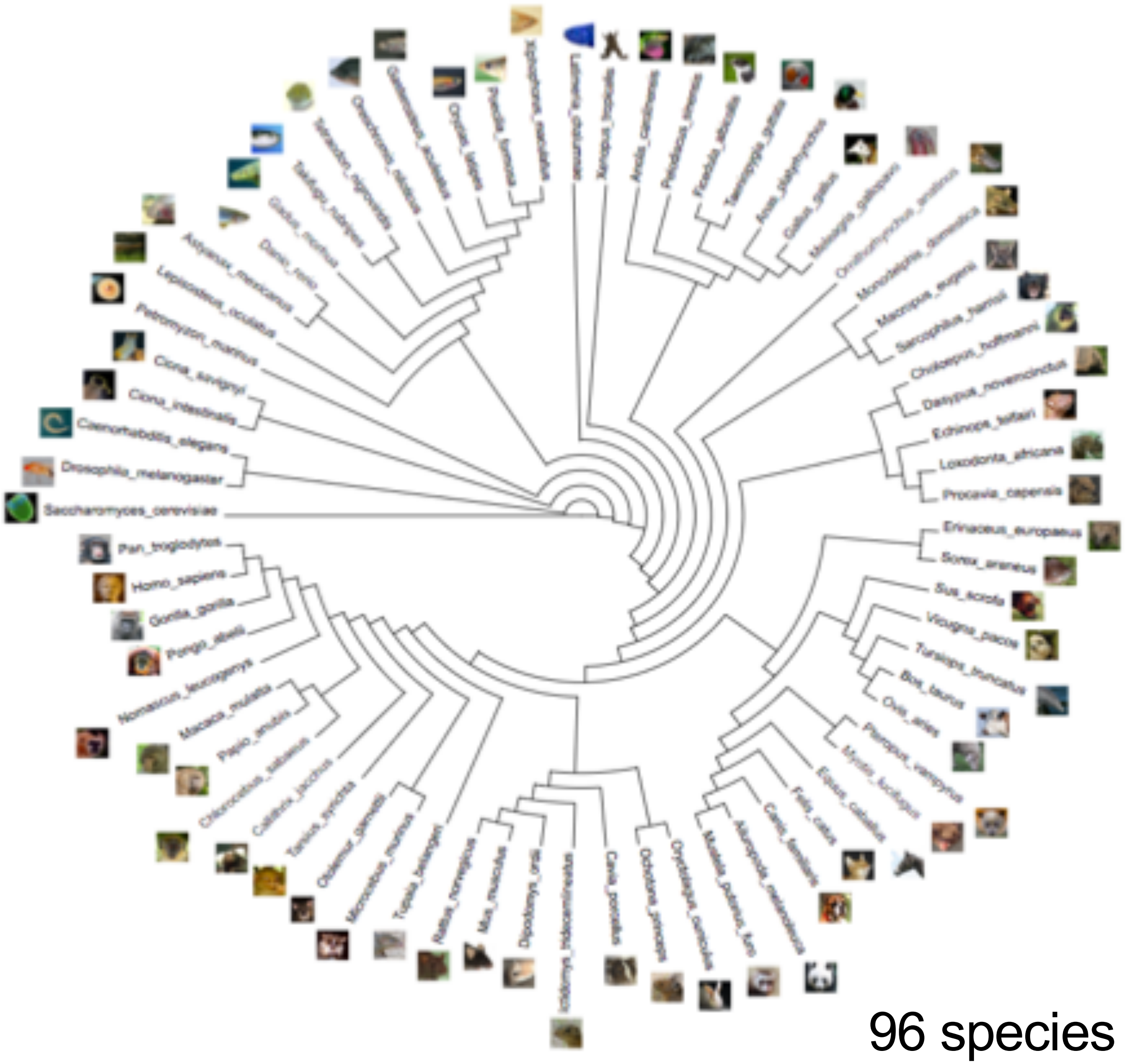
  (BLAST, self-chain @UCSC, paralogues Ensembl)

# Comparative Genome Analysis: Where to Start?

To identify conserved regions, you must:

- Decide which species you would like to compare
- Identify and extract the relevant genome sequences
- Annotate genes and other features found in the genome sequences
- Ensure that repetitive sequences are masked

# How many vertebrate genomes are available?



96 species + mouse strains

86 species in e! v89

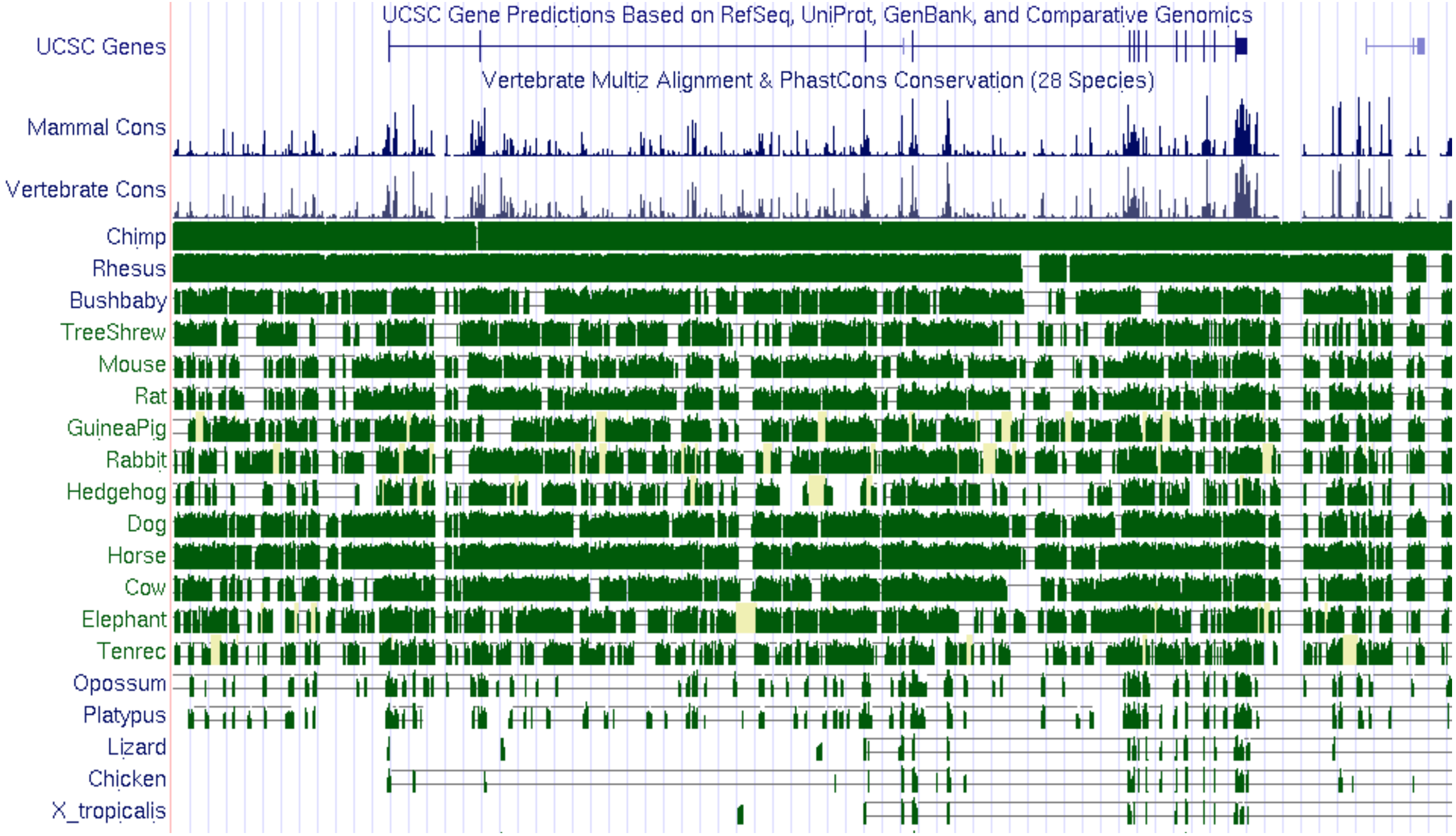
# Selection of Species for DNA comparisons

| **Human** vs. | **Chimpanzee** | **Mouse** | **Opossum** | **Pufferfish** |
|---|---|---|---|---|
| Size (Gbp) | 3.0 | 2.5 | 4.2 | 0.4 |
| Time since divergence | ~6 MYA | ~ 90 MYA | ~150 MYA | ~450 MYA |
| Sequence conservation (in coding regions) | >99% | ~80% | ~70-75% | ~65% |
| Aids identification of... | Recently changed sequences and genomic rearrangements | Both coding and non-coding sequences | Both coding and non-coding sequences | Primarily coding sequences |
| Background noise | High | Moderate | Low | Lower |

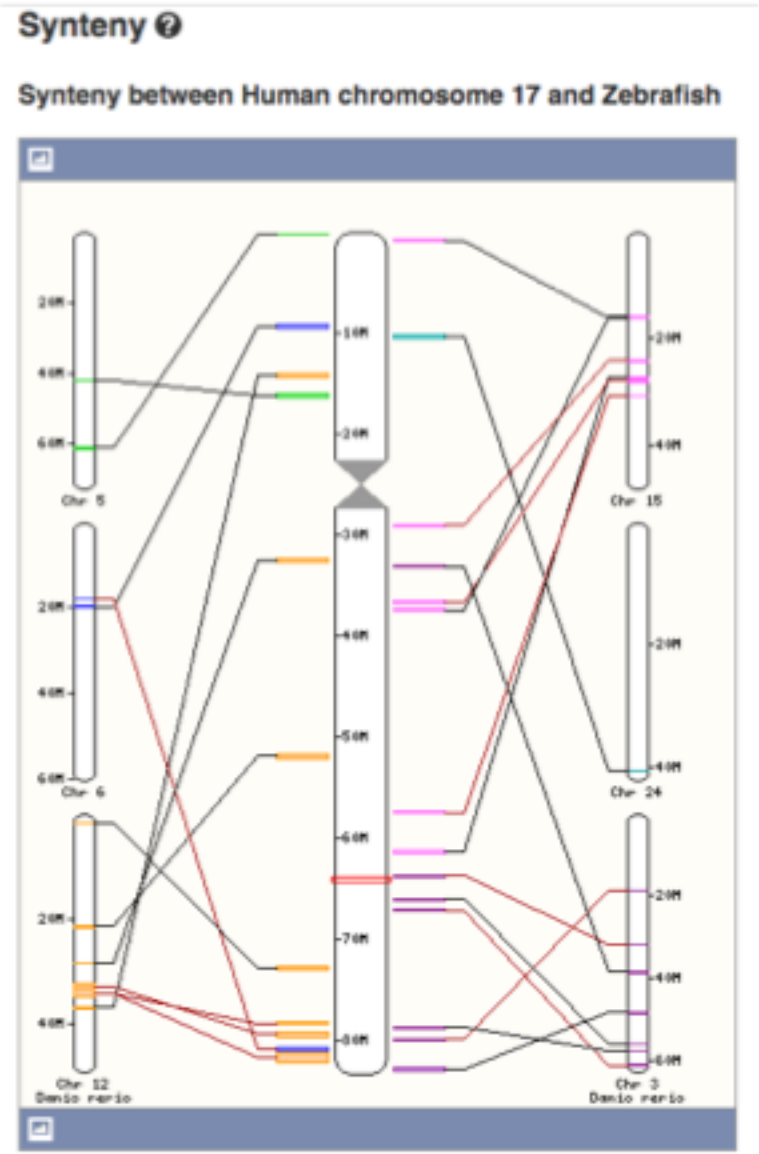
# Aligning genomic sequence

- Pair-wise genome sequence alignments combined with additional phylogenetic information

(eg PhastCons@UCSC, RankVista,)



# Aligning genome sequences - synteny



Synteny

Synteny between Human chromosome 17 and Zebrafish

- Syntenic regions are calculated where possible from pairwise (two-species) whole genome alignments.(e.g. Compara@Ensembl)

- The centre chromosome represents the species of interest, and the smaller chromosomes show syntenic regions with a second species. Blocks are coloured according to the chromosome number on the second species.

# Worked Demos and Exercises

# Exploring sequence variation and disease

# Human v Fish

- Disease resources are very human-centric
- More variation information is available for humans

# Gene Expression Databases

**GEO profiles** (NCBI)

- Gene expression profiles
- Derived from GEO (Gene Expression Omnibus)

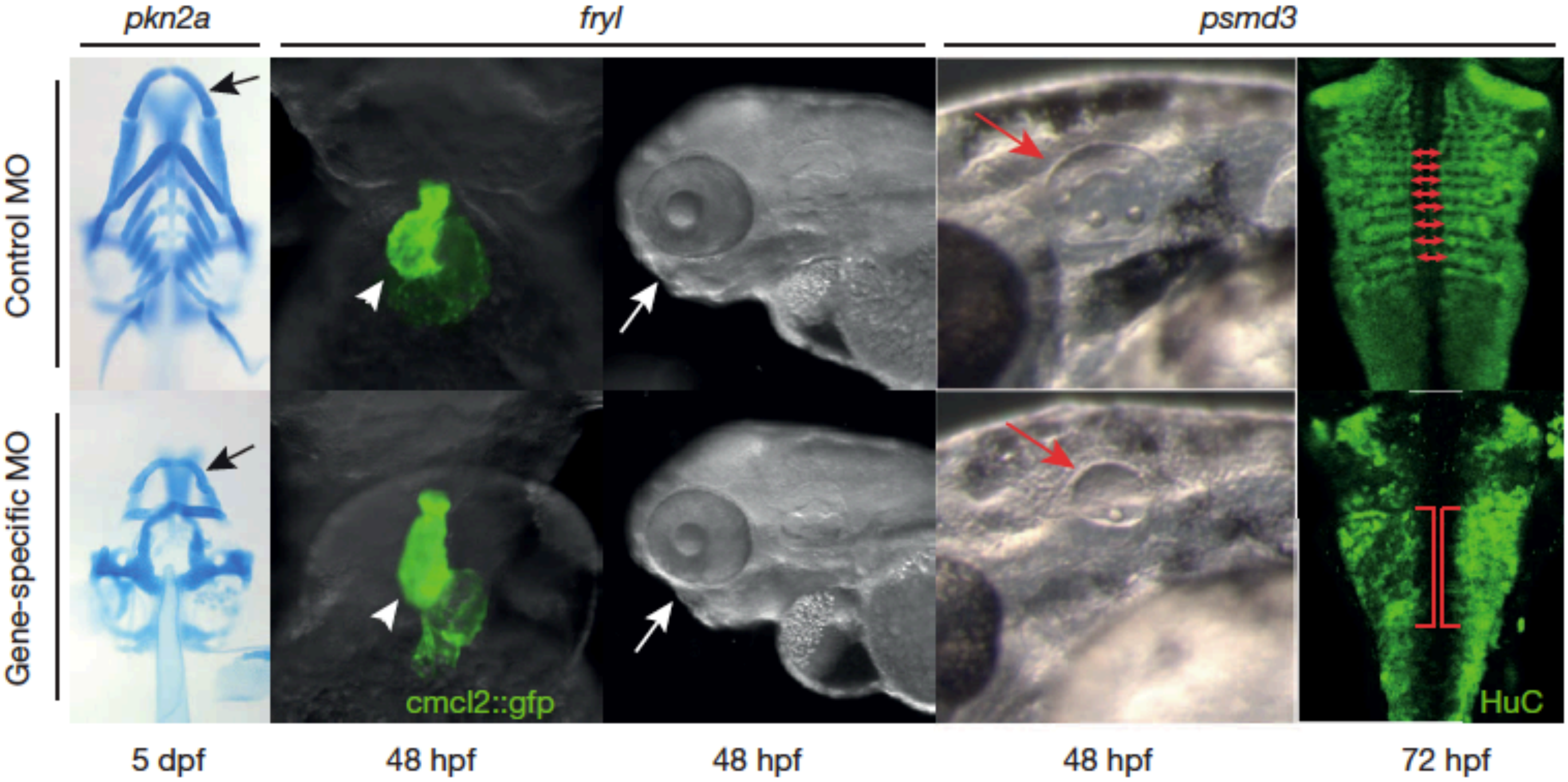
**Expression Atlas** (EBI)

- Baseline Atlas: which gene products and their abundance in “normal” conditions
- Differential Atlas: genes that are up or down regulated in a variety of different experimental Conditions
- Derived from Array Express

# Examining phenotypic effect of Mutations

- OMIM
  - Online Mendelian Inheritance in Man
  - Catalogue of all known diseases with a genetic component
- COSMIC
  - Catalogue Of Somatic Mutations In Cancer
- DECIPHER
  - DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
  - Database of genomic variation data from analysis of patient DNA

# Zebrafish as a model for DDD



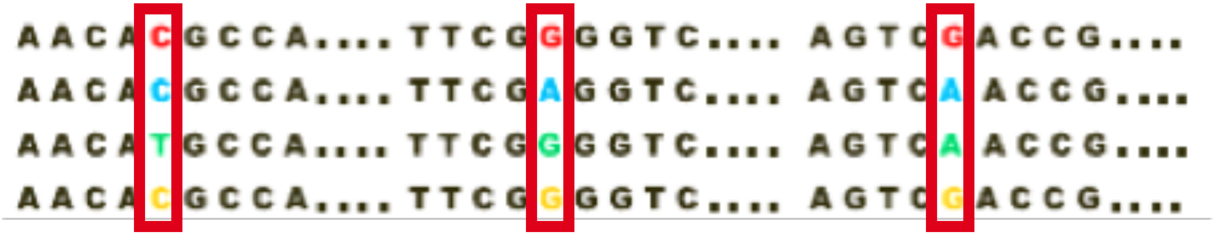
The Deciphering Developmental Disorders Study (2014) Nature

# Variation is useful

- Determine disease risk
- Predict reactions to environmental triggers
- Predict responsiveness to drug treatments
- Forensics
- Genetic and physical mapping
- Evolution

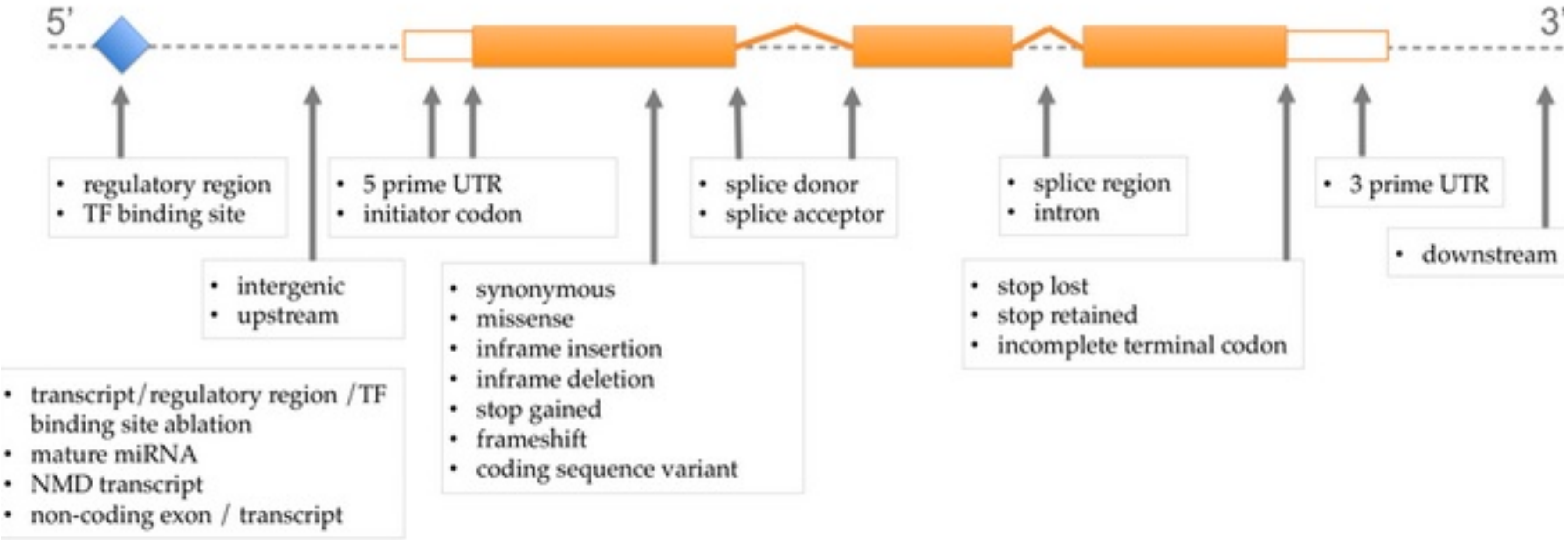
# Variation Types

- Cytological level:
  - Chromosome numbers
  - Segmental duplications, rearrangements, and deletions
- Molecular level:
  - Transposable Elements
  - Short Deletions/Insertions, Tandem Repeats
- Sequence level:
  - Single Nucleotide Polymorphisms (SNPs)
  - Small Nucleotide Insertions and Deletions (Indels)



```
AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
AACACGCCA.... TTCGAGGTC.... AGTCAACCG....
AACATGCCA.... TTCGGGGTC.... AGTCAACCG....
AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
```

# Types of SNPs



- regulatory region
- TF binding site

- intergenic
- upstream

- transcript/regulatory region /TF binding site ablation
- mature miRNA
- NMD transcript
- non-coding exon / transcript

- 5 prime UTR
- initiator codon

- synonymous
- missense
- inframe insertion
- inframe deletion
- stop gained
- frameshift
- coding sequence variant

- splice donor
- splice acceptor

- splice region
- intron

- stop lost
- stop retained
- incomplete terminal codon

- 3 prime UTR

- downstream

Genic, coding SNPs; Genic, non-coding SNPs; Intergenic, regulatory

# Variant predictor programs

- PolyPhen and SIFT
  - Provides a scoring for a SNP/Mutation and effect on phenotype
- Variant effect predictor – VEP (Ensembl)
- Variant Annotation Integrator (UCSC)

# Variant Effect Predictor (VEP)

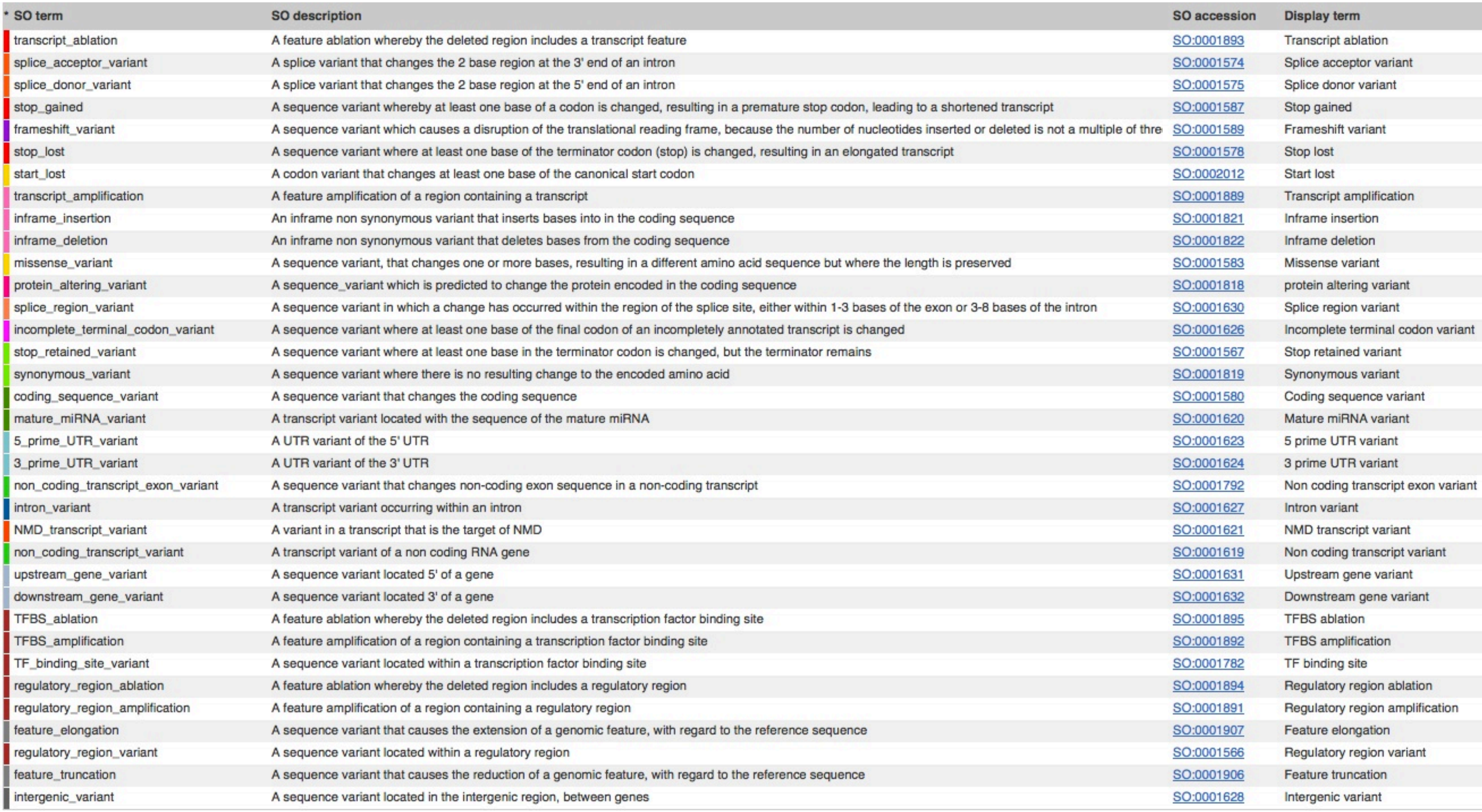
Predicts:

- Functional consequences of known and unknown variants
- Substitutions, insertions, deletions and structural variants

Output:

- Affected genes / transcripts / regulatory features / motifs
- Gene symbols
- IDs from Ensembl, CCDS, UniProt, HGVS
- Consequence (missense, stop gained etc)
- Location of variant
- Co-located known variant (s)
- Minor allele frequencies from 1000 Genomes Project
- PolyPhen and SIFT scores

# Colour-coding in Ensembl



| * | SO term | SO description | SO accession | Display term |
|---|---|---|---|---|
| 🟥 | transcript_ablation | A feature ablation whereby the deleted region includes a transcript feature | SO:0001893 | Transcript ablation |
| 🟧 | splice_acceptor_variant | A splice variant that changes the 2 base region at the 3' end of an intron | SO:0001574 | Splice acceptor variant |
| 🟧 | splice_donor_variant | A splice variant that changes the 2 base region at the 5' end of an intron | SO:0001575 | Splice donor variant |
| 🟥 | stop_gained | A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript | SO:0001587 | Stop gained |
| 🟪 | frameshift_variant | A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of thre | SO:0001589 | Frameshift variant |
| 🟥 | stop_lost | A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript | SO:0001578 | Stop lost |
| 🟨 | start_lost | A codon variant that changes at least one base of the canonical start codon | SO:0002012 | Start lost |
| 🟪 | transcript_amplification | A feature amplification of a region containing a transcript | SO:0001889 | Transcript amplification |
| 🟪 | inframe_insertion | An inframe non synonymous variant that inserts bases into in the coding sequence | SO:0001821 | Inframe insertion |
| 🟪 | inframe_deletion | An inframe non synonymous variant that deletes bases from the coding sequence | SO:0001822 | Inframe deletion |
| 🟨 | missense_variant | A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved | SO:0001583 | Missense variant |
| 🟪 | protein_altering_variant | A sequence_variant which is predicted to change the protein encoded in the coding sequence | SO:0001818 | protein altering variant |
| 🟧 | splice_region_variant | A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron | SO:0001630 | Splice region variant |
| 🟪 | incomplete_terminal_codon_variant | A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed | SO:0001626 | Incomplete terminal codon variant |
| 🟩 | stop_retained_variant | A sequence variant where at least one base in the terminator codon is changed, but the terminator remains | SO:0001567 | Stop retained variant |
| 🟩 | synonymous_variant | A sequence variant where there is no resulting change to the encoded amino acid | SO:0001819 | Synonymous variant |
| 🟩 | coding_sequence_variant | A sequence variant that changes the coding sequence | SO:0001580 | Coding sequence variant |
| 🟩 | mature_miRNA_variant | A transcript variant located with the sequence of the mature miRNA | SO:0001620 | Mature miRNA variant |
| 🟦 | 5_prime_UTR_variant | A UTR variant of the 5' UTR | SO:0001623 | 5 prime UTR variant |
| 🟦 | 3_prime_UTR_variant | A UTR variant of the 3' UTR | SO:0001624 | 3 prime UTR variant |
| 🟩 | non_coding_transcript_exon_variant | A sequence variant that changes non-coding exon sequence in a non-coding transcript | SO:0001792 | Non coding transcript exon variant |
| 🟦 | intron_variant | A transcript variant occurring within an intron | SO:0001627 | Intron variant |
| 🟧 | NMD_transcript_variant | A variant in a transcript that is the target of NMD | SO:0001621 | NMD transcript variant |
| 🟩 | non_coding_transcript_variant | A transcript variant of a non coding RNA gene | SO:0001619 | Non coding transcript variant |
| ⬜ | upstream_gene_variant | A sequence variant located 5' of a gene | SO:0001631 | Upstream gene variant |
| ⬜ | downstream_gene_variant | A sequence variant located 3' of a gene | SO:0001632 | Downstream gene variant |
| 🟫 | TFBS_ablation | A feature ablation whereby the deleted region includes a transcription factor binding site | SO:0001895 | TFBS ablation |
| 🟫 | TFBS_amplification | A feature amplification of a region containing a transcription factor binding site | SO:0001892 | TFBS amplification |
| 🟫 | TF_binding_site_variant | A sequence variant located within a transcription factor binding site | SO:0001782 | TF binding site |
| 🟫 | regulatory_region_ablation | A feature ablation whereby the deleted region includes a regulatory region | SO:0001894 | Regulatory region ablation |
| 🟫 | regulatory_region_amplification | A feature amplification of a region containing a regulatory region | SO:0001891 | Regulatory region amplification |
| ⬛ | feature_elongation | A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence | SO:0001907 | Feature elongation |
| 🟫 | regulatory_region_variant | A sequence variant located within a regulatory region | SO:0001566 | Regulatory region variant |
| ⬛ | feature_truncation | A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence | SO:0001906 | Feature truncation |
| ⬛ | intergenic_variant | A sequence variant located in the intergenic region, between genes | SO:0001628 | Intergenic variant |

# Warning!

- All these tools make **predictions**

- Findings should **always** be confirmed experimentally

# Worked Demos and Exercises

# Ensembl Acknowledgements

## The Entire Ensembl Team

Daniel R. Zerbino[1], Premanand Achuthan[1], Wasiu Akanni[1], M. Ridwan Amode[1], Daniel Barrell[1,2], Jyothish Bhai[1], Konstantinos Billis[1], Carla Cummins[1], Astrid Gall[1], Carlos García Giroń[1], Laurent Gil[1], Leo Gordon[1], Leanne Haggerty[1], Erin Haskell[1], Thibaut Hourlier[1], Osagie G. Izuogu[1], Sophie H. Janacek[1], Thomas Juettemann[1], Jimmy Kiang To[1], Matthew R. Laird[1], Ilias Lavidas[1], Zhicheng Liu[1], Jane E. Loveland[1], Thomas Maurel[1], William McLaren[1], Benjamin Moore[1], Jonathan Mudge[1], Daniel N. Murphy[1], Victoria Newman[1], Michael Nuhn[1], Denye Ogeh[1], Chuang Kee Ong[1], Anne Parker[1], Mateus Patricio[1], Harpreet Singh Riat[1], Helen Schuilenburg[1], Dan Sheppard[1], Helen Sparrow[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Brandon Walts[1], Amonida Zadissa[1], Adam Frankish[1], Sarah E. Hunt[1], Myrto Kostadima[1], Nicholas Langridge[1], Fergal J. Martin[1], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Dan M. Staines[1], Stephen J. Trevanion[1], Bronwen L. Aken[1], Fiona Cunningham[1], Andrew Yates[1] and Paul Flicek[1,3]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Eagle Genomics Ltd., Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK and [3]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

## Funding

wellcome

EMBL

BBSRC bioscience for the future

National Human Genome Research Institute

TRANSFORMING GENETIC MEDICINE INITIATIVE

SAVE THE TASMANIAN DEVIL PROGRAM APPEAL

Open Targets

MedBioinformatics

Co-funded by the European Union