Training materials

- Ensembl training materials are protected by a CC BY license
- http://creativecommons.org/licenses/by/4.0/
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers
- http://www.ensembl.org/info/about/publications.html

Browsing Genes and Genomes with Ensembl

Ben Moore

Ensembl Outreach

EMBL-EBI

UC Irvine - 31st October 2022

EMBL-EBI

Structure

Presentation:

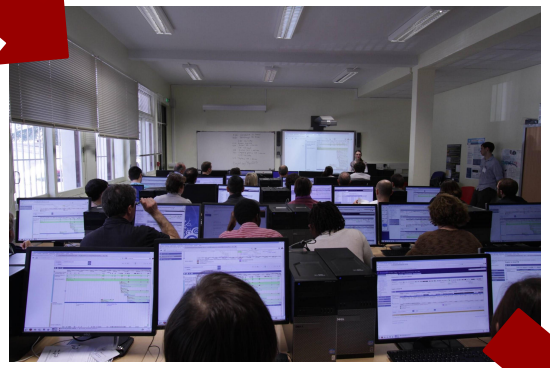
What the data/tool is

How we produce/process the data

Demo:

Getting the data

Using the tool



Follow along if you want to

Exercises:

Available on the EBI Train Online website

Trying things out for yourself

Going beyond the demo

http://training.ensembl.org/events

Course materials

training.ensembl.org

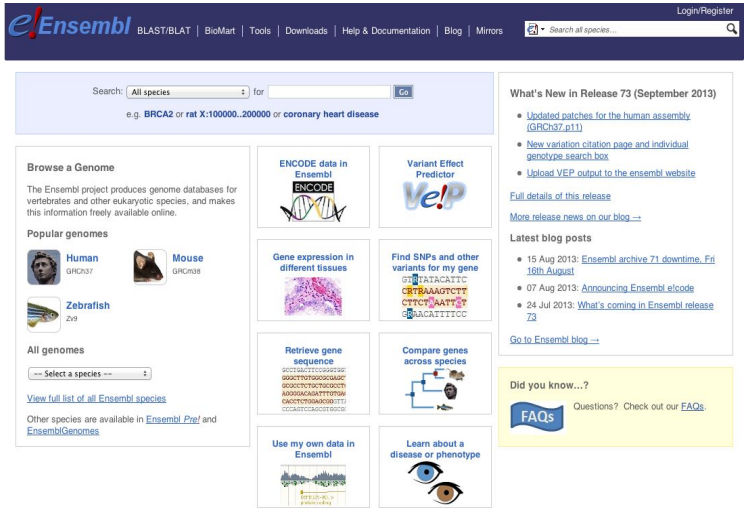
- Presentations
- Demonstrations
- Exercises and answers
- Living Document

http://training.ensembl.org/events

Objectives

- What is Ensembl?
- What type of data can you get in Ensembl?
- How to navigate the Ensembl browser website.
- Where to go for help and documentation.

Exploring the Ensembl genome browser

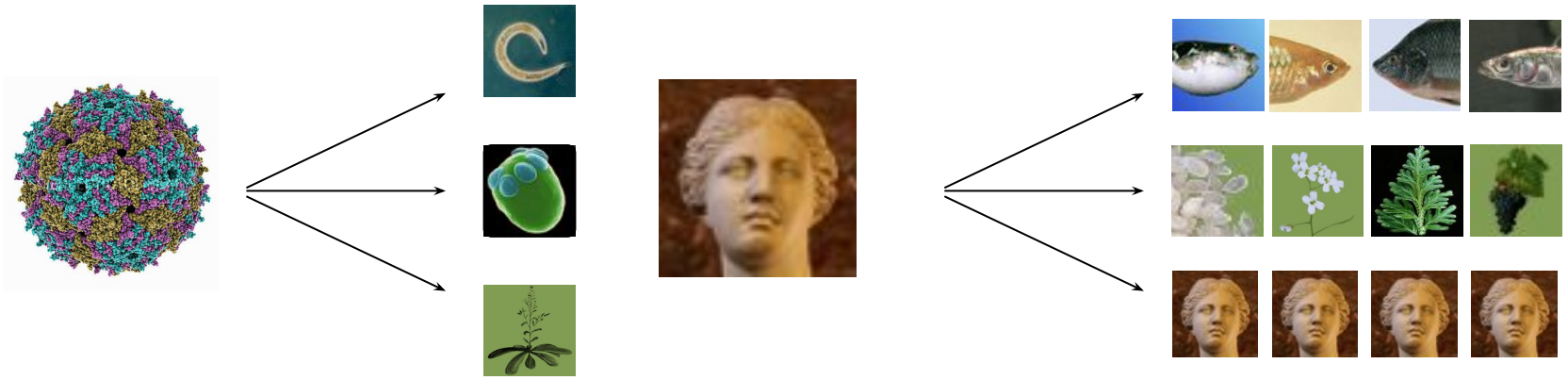


EMBL-EBI

Why do we need genome browsers?

1977: 1st genome to be sequenced (5 kb)

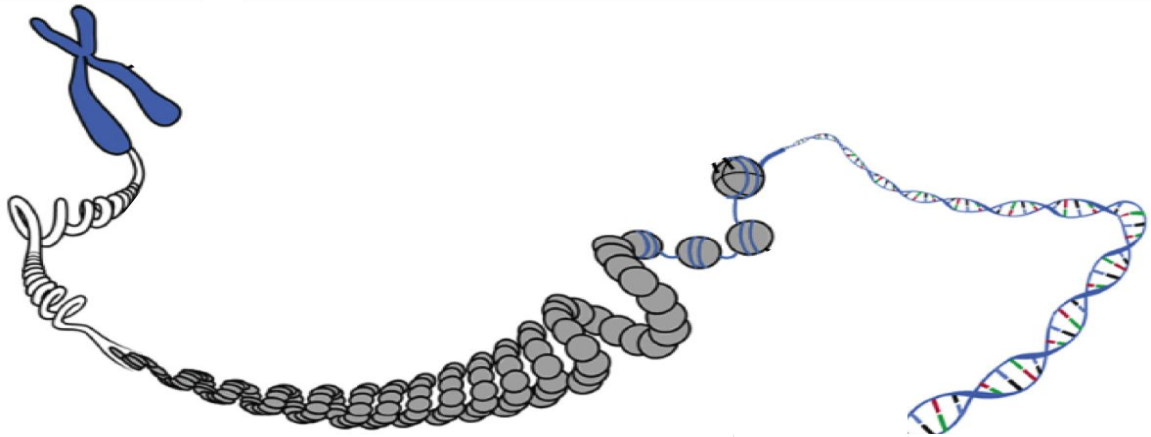
2004: finished human sequence (3 Gb)



Why do we need genome browsers?

http://training.ensembl.org/events

Ensembl- unlocking the code

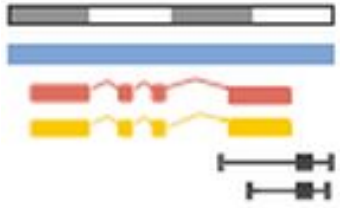


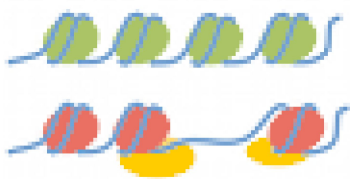
- *Genomic assemblies* - automated gene annotation
- *Variation* - Small and large scale sequence variation with phenotype associations
- *Comparative Genomics* - Whole genome alignments, gene trees
- *Regulation* - Potential promoters and enhancers, DNA methylation

http://training.ensembl.org/events

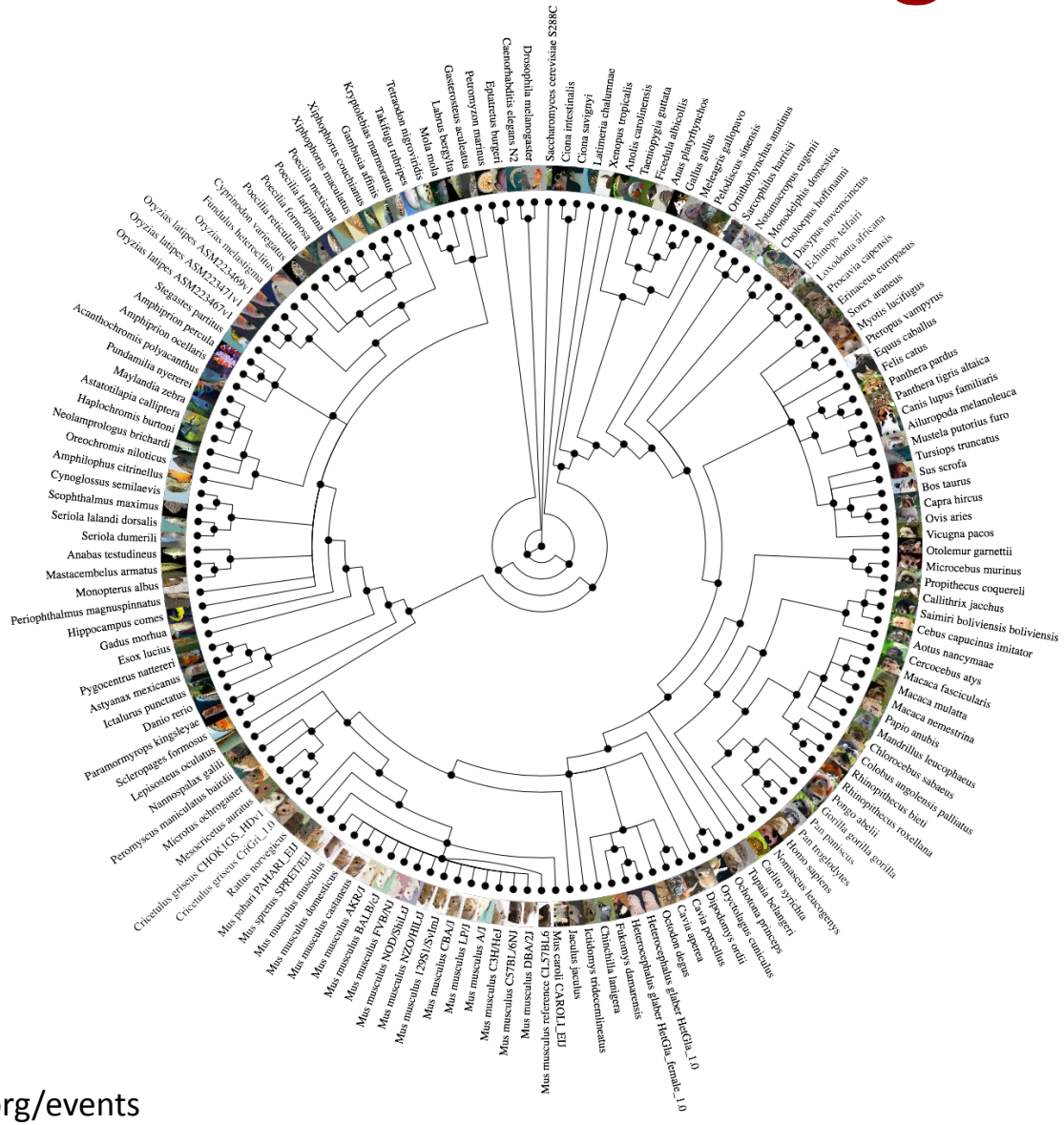
Ensembl Features

- Gene builds for ~300 species
- Variation, comparative genomics and regulatory data display
- Display of user data
- Tools for data processing, e.g VEP
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source






Ensembl- access to 300+ genomes



http://training.ensembl.org/events

Ensembl Genomes- expanding Ensembl

www.ensembl.org

- Vertebrates

- Other representative species

http://training.ensembl.org/events

Ensembl Genomes- expanding Ensembl

www.ensembl.org

www.ensemblgenomes.org

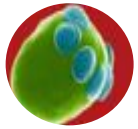
- Vertebrates



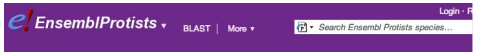
- Other representative species

- Bacteria


- Fungi

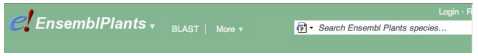



- Protists

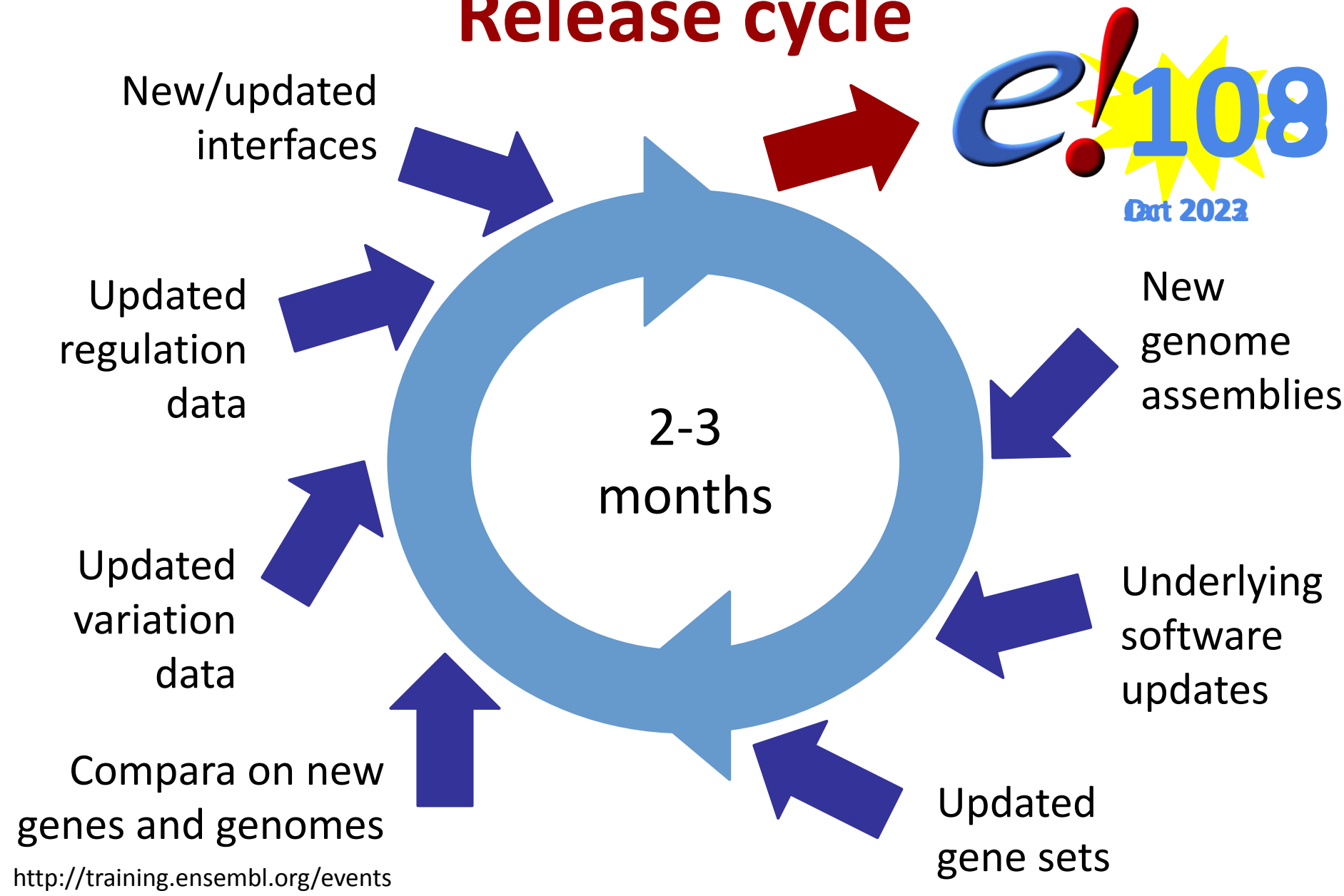
- Metazoa


- Plants

http://training.ensembl.org/events

Release cycle



http://training.ensembl.org/events

What is a genome assembly?

Sequence reads

```
                    CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA
  CAGCTGTCCCAGATGAC     ACTTAACTTCCCTCCCAGCTGTCC
GGGCTCCGCCTTCAGCTC                    TCCCAGCTGTCCCAGATGACGCCAT
                  AACTTCCCTCCCAGCT
   CGGCCTTTGGGCTCC                    TCCGCCTTCAGCTCAAGACTTAACTTC
                   CAGATGACGCC
```

Match up overlaps

```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA   AACTTCCCTCCCAGCT    CAGATGACGCC
           TCCGCCTTCAGCTCAAGACTTAACTTC  TCCCAGCTGTCCCAGATGACGCCAT
       GGGCTCCGCCTTCAGCTC   ACTTAACTTCCCTCCCAGCTGTCC
CGGCCTTTGGGCTCC                          CAGCTGTCCCAGATGAC
```

Genome assembly

```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCAT
```

Genome contigs


BL


AL


CM


IM

BL102
AL476
CM553
IM768

http://training.ensembl.org/events

Human genome assemblies

- GRCh38 (aka hg38)
  - www.ensembl.org
  - Most up-to-date and supported

- GRCh37 (aka hg19)
  - Greater gap length than GRCh38
  - grch37.ensembl.org
  - Limited data and software updates

- NCBI36 (aka hg18)
  - 150,000 gaps
  - ncbi36.ensembl.org
  - No longer updated

http://training.ensembl.org/events

Hands on

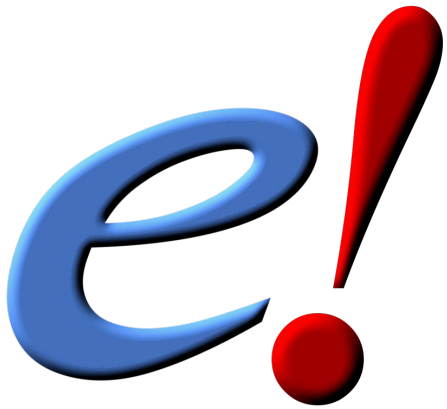
- We're going to look at the Ensembl homepage and how to find information about the species and genome assemblies in Ensembl.

- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

http://training.ensembl.org/events

Exploring genomic locations: the Region in Detail view

EMBL-EBI

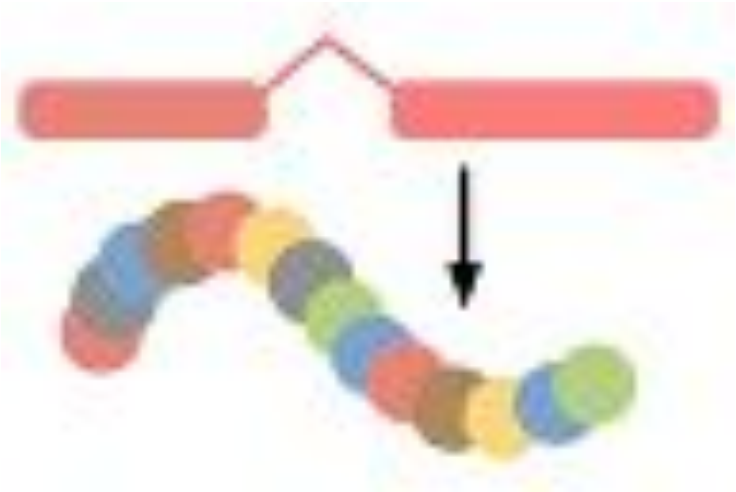


Hands on

- We're going to look at a region of the human genome, 4:122868000-122946000, and manipulate the view to see the data we're interested in.

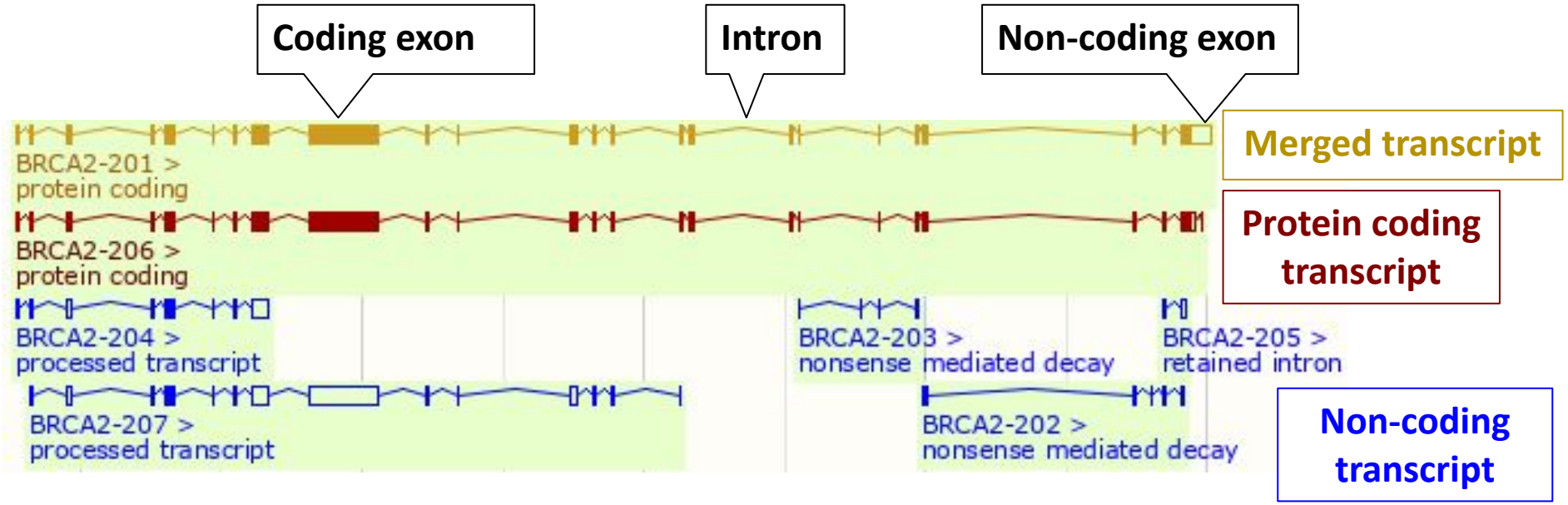
- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

Genes and Transcripts



EMBL-EBI

Gene views


Coding exon
Intron
Non-coding exon
BRCA2-201 >
protein coding
BRCA2-206 >
protein coding
BRCA2-204 >
processed transcript
BRCA2-203 >
nonsense mediated decay
BRCA2-205 >
retained intron
BRCA2-207 >
processed transcript
BRCA2-202 >
nonsense mediated decay
Merged transcript
Protein coding transcript
Non-coding transcript

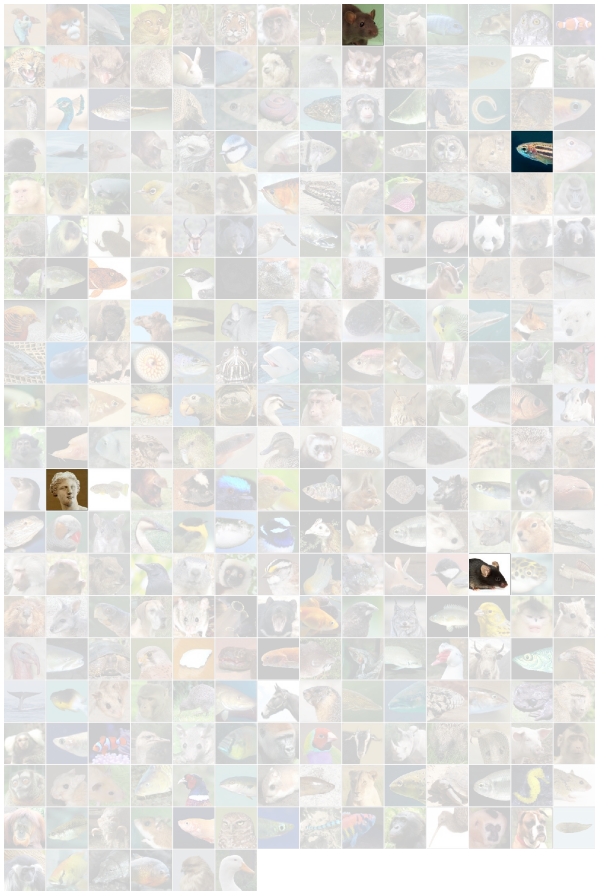
Ensembl and Havana annotation

e! Ensembl

havana

Automatic annotation

Manual annotation



http://training.ensembl.org/events

Automatic gene annotation

- Genome-wide determination using the Ensembl automated pipeline

- Predictions based on experimental (biological) data

http://training.ensembl.org/events

Biological Evidence

- **International Nucleotide Sequence databases**
  - cDNAs
  - ESTs
  - RNAseq

ENA European Nucleotide Archive

GenBank

DDBJ DNA Data Bank of Japan

- **Protein sequence databases**
  - Swiss-Prot: manually curated
  - TrEMBL: unreviewed translations

UniProt

- **Homologous sequences**

  Infer genes from homology to other species
  - predict genes in  by mapping cDNAs/proteins from  to the  genome.

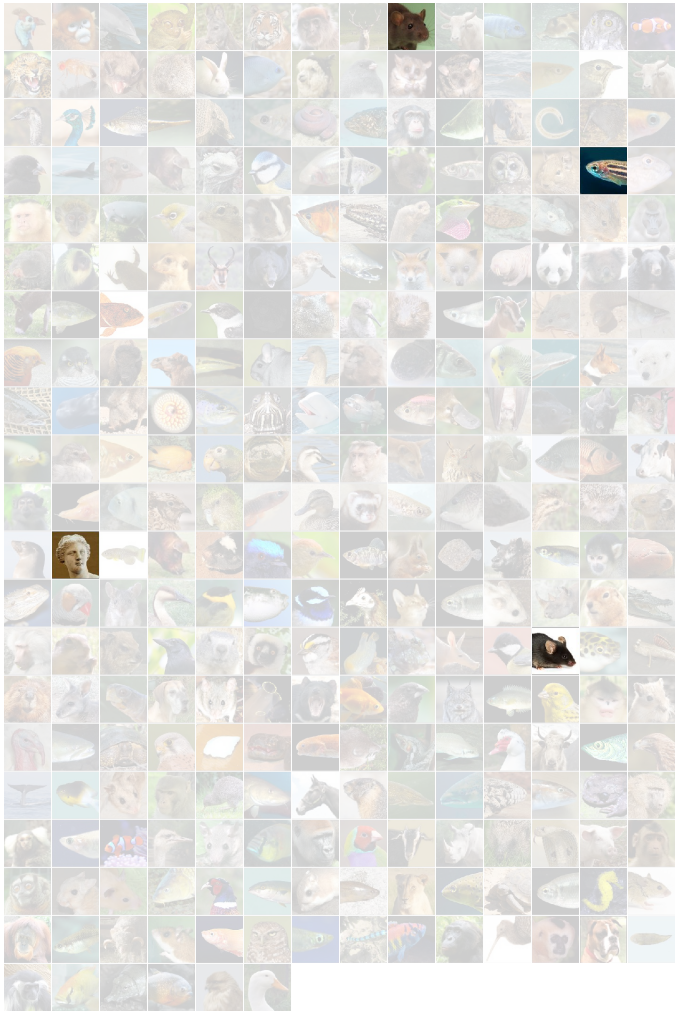
http://training.ensembl.org/events

Manual gene annotation

- Gene determination on a case-by-case basis by a person

- Uses data from databases and papers:
  - INSDC databases
  - RNAseq
  - long read transcriptomic data
  - intron data
  - CAGE
  - PolyA-Seq
  - Mass Spec
  - ONT
  - publications

havana
human and vertebrate analysis and annotation



http://training.ensembl.org/events

Gene annotation: Manual

Benefit	Disadvantages

- More comprehensive
- More genes and transcripts overall
- Require less evidence (quality > thresholds)
- More accurate for difficult regions:
  - UTRs
  - Splice sites
  - Single exon transcripts
  - Exceptions (i.e. immunoglobins)
	- Slower
	- Small scale

Golden transcripts

- Identical annotation

- Higher confidence and quality

GENCODE

- The GENCODE gene set is made up of:
  - The merged set of **Ensembl automatically annotated genes** and **Havana manually annotated genes**

- GENCODE is the default gene set used by gnomAD/ExAC, ENCODE, 1000 Genomes and other major projects.

MANE transcripts

- **M**atched **A**nnotation from the **N**CBI and **E**BI (MANE) project
- Biologically relevant transcript set with 100% identity between resources (including non-coding regions)
-  ~100% of human protein coding transcripts

Which transcript should I use?

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags
ESRRA-201	ENST00000000442.11	2274	423aa	Protein coding	CCDS41667	P11474-1	NM_004451.5	TSL:1 GENCODE basic APPRIS P2 MANE Select v0.91
ESRRA-202	ENST00000405666.5	2283	423aa	Protein coding	CCDS41667	P11474-1	-	TSL:1 GENCODE basic APPRIS P2
ESRRA-203	ENST00000406310.6	2293	506aa	Protein coding	-	P11474-2	-	TSL:1 GENCODE basic
ESRRA-204	ENST00000467987.1	610	No protein	Retained intron	-	-	-	TSL:3
ESRRA-205	ENST00000468670.2	738	108aa	Protein coding	-	F5GWT5	-	CDS 3' incomplete TSL:2
ESRRA-206	ENST00000539594.5	747	162aa	Protein coding	-	F5H0E9	-	CDS 3' incomplete TSL:3
ESRRA-207	ENST00000545035.1	677	198aa	Protein coding	-	H0YGT3	-	CDS 5' incomplete TSL:2
ESRRA-208	ENST00000677967.1	2266	422aa	Protein coding	CCDS60830	-	-	GENCODE basic APPRIS ALT1

MANE Select v0.5	'Matched Annotation NCBI and Ensembl' Transcript with 100% identical annotation with RefSeq
APPRIS P1	APPRIS principal isoform: The major isoform(s) from combining protein structural information, functionally important residues and evidence from cross-species alignments
GENCODE basic	"Complete" transcripts (where a gene has complete transcripts)
TSL:1	Transcript support level: Scored 1-5 for quality (1= best)

http://training.ensembl.org/events

http://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html

Ensembl stable IDs

- ENSG##########.# Ensembl Gene ID
- ENST##########.# Ensembl Transcript ID
- ENSP##########.# Ensembl Peptide ID
- ENSE##########.# Ensembl Exon ID
- ENSR##########.# Ensembl Regulatory region ID

- For non-human species a suffix is added:

  MUS (*Mus musculus*) for mouse ENSMUSG###

  DAR (*Danio rerio*) for zebrafish: ENSDARG###

http://www.ensembl.org/info/genome/stable_ids/index.html

Hands on

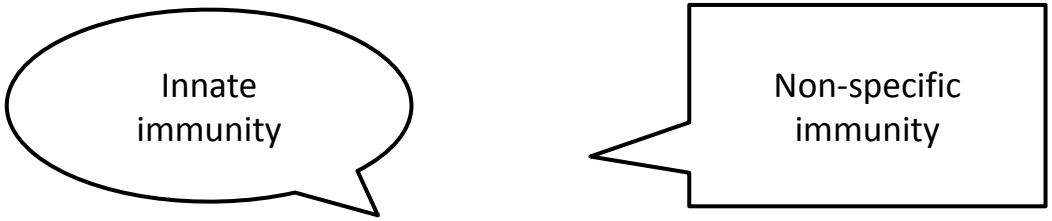
- We're going to look at an Ensembl gene, *UQCRQ*, and find out information about it and its transcripts.

- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

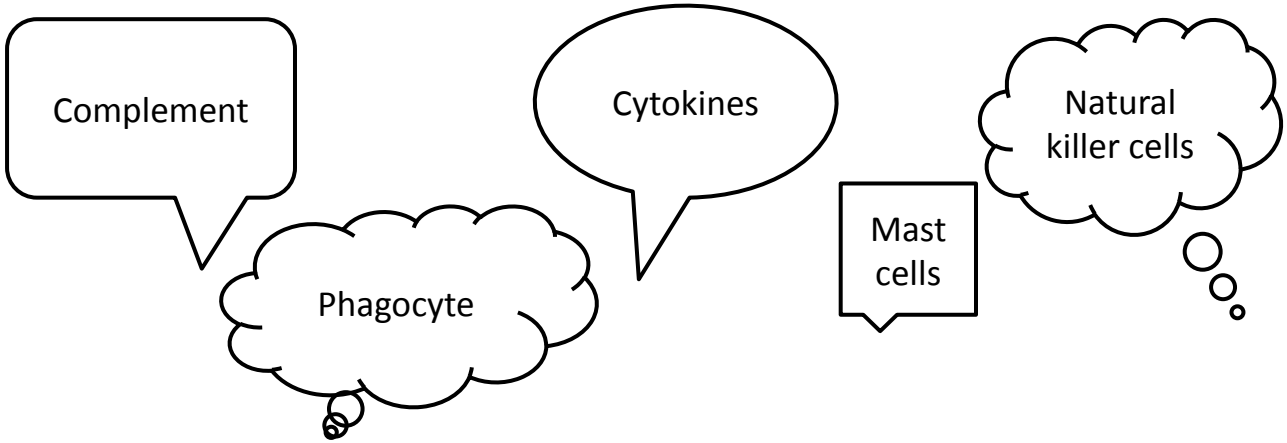
http://training.ensembl.org/events

Why Gene Ontology (GO)?

Multiple terms for the same thing



Gene descriptions too specific



GO terms form a controlled vocabulary

GO:0045087 - innate immune response

Innate immune responses are defense responses mediated by germline encoded components that directly recognise components of potential pathogens.

Hands on

- We're going to look at an Ensembl gene, *UQCRQ*, and find out information about it and its transcripts.

- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

http://training.ensembl.org/events

Variation

EMBL-EBI

Outline

- Classification of variants
- Species and sources of variation
- Browsing variation data
  - Gene tab
  - Location tab
  - Variation tab
- Variant Effect Predictor 

Variation types

1) Small scale in one or few nucleotides of a gene

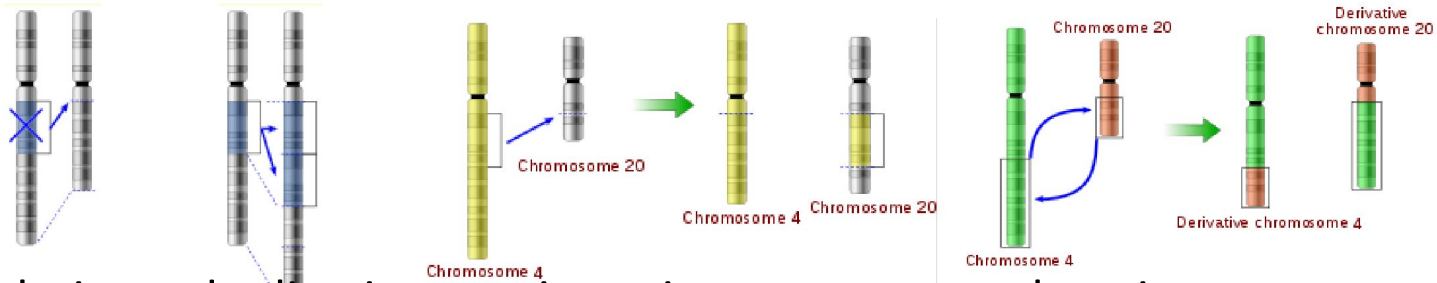
- Small insertions and deletions (DIPs or indels)
- Single nucleotide polymorphism (SNP)

A G A C T T G A C C T G T C T – A A C T G G A

T G A C T T G A C – T G T C T G A A C G G G A

2) Large scale in chromosomal structure (structural variation)

- Copy number variations (CNV)
- Large deletions/duplications, insertions, translocations



deletion duplication insertion translocation

24 species with variation data

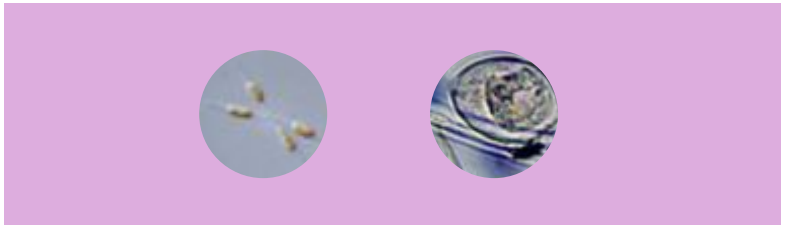
	**Cat** *Felis catus*	3.6 M
	**Chicken** *Gallus gallus*	24 M
	**Chimpanzee** *Pan troglodytes*	1.6 M
	**Cow** *Bos taurus*	104 M
	**Dog** *Canis lupus familiaris*	5.9 M
	**Gibbon** *Nomascus leucogenys*	1.1 M
	**Goat** *Capra hircus*	37 M
	**Horse** *Equus caballus*	21 M
	**Human** *Homo sapiens*	679 M
	**Macaque** *Macaca mulatta*	53 M
	**Mouse** *Mus musculus*	84 M

	**Opossum** *Monodelphis domestica*	1.1 M
	**Orangutan** *Pongo abelii*	10 M
	**Pig** *Sus scrofa*	67 M
	**Platypus** *Ornithorhynchus anatinus*	1.3 M
	**Rat** *Rattus norvegicus*	5 M
	**Sheep** *Ovis aries rambouillet*	61 M
	**Sheep (texel)** *Ovis aries*	61 M
	**Tetraodon** *Tetraodon nigroviridis*	902 K
	**Turkey** *Meleagris gallopavo*	9 K
	**Zebra finch** *Taeniopygia guttata*	1.7 M
	**Zebrafish** *Danio rerio*	17 M

http://www.ensembl.org/info/genome/variation/species/sources_documentation.html

http://training.ensembl.org/events

Species with variation data in Ensembl Genomes



http://training.ensembl.org/events

Where does the data come from?

The Ensembl variation process

http://www.ensembl.org/info/genome/variation/index.html

Ensembl variation process: Import

Variant import → Quality control → Linked data → Ensembl analysis

Import variant data from publicly available archives and data repositories

dbSNP
Short Genetic Variations

DGVarchive

EVA

COSMIC
The core of COSMIC, an expert-curated database of somatic mutations

Cancer Gene Census
A catalogue of genes with mutations that are causally implicated in cancer

http://www.ensembl.org/info/genome/variation/species/sources_documentation.html

Ensembl variation process: QC

- Mapping to reference assembly
  - GRCh37 → GRCh38
- Checks on alleles
- Checks for IUPAC ambiguity codes

Excluding 'suspect' variants

http://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#quality_control

Ensembl variation process: Linked data

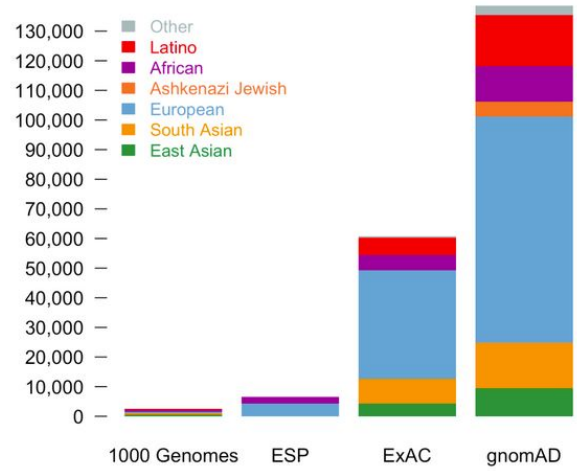
Variant import → Quality control → **Linked data** → Ensembl analysis

Import 'value added' data

- Allele frequencies
- Phenotype/disease
- Publication data

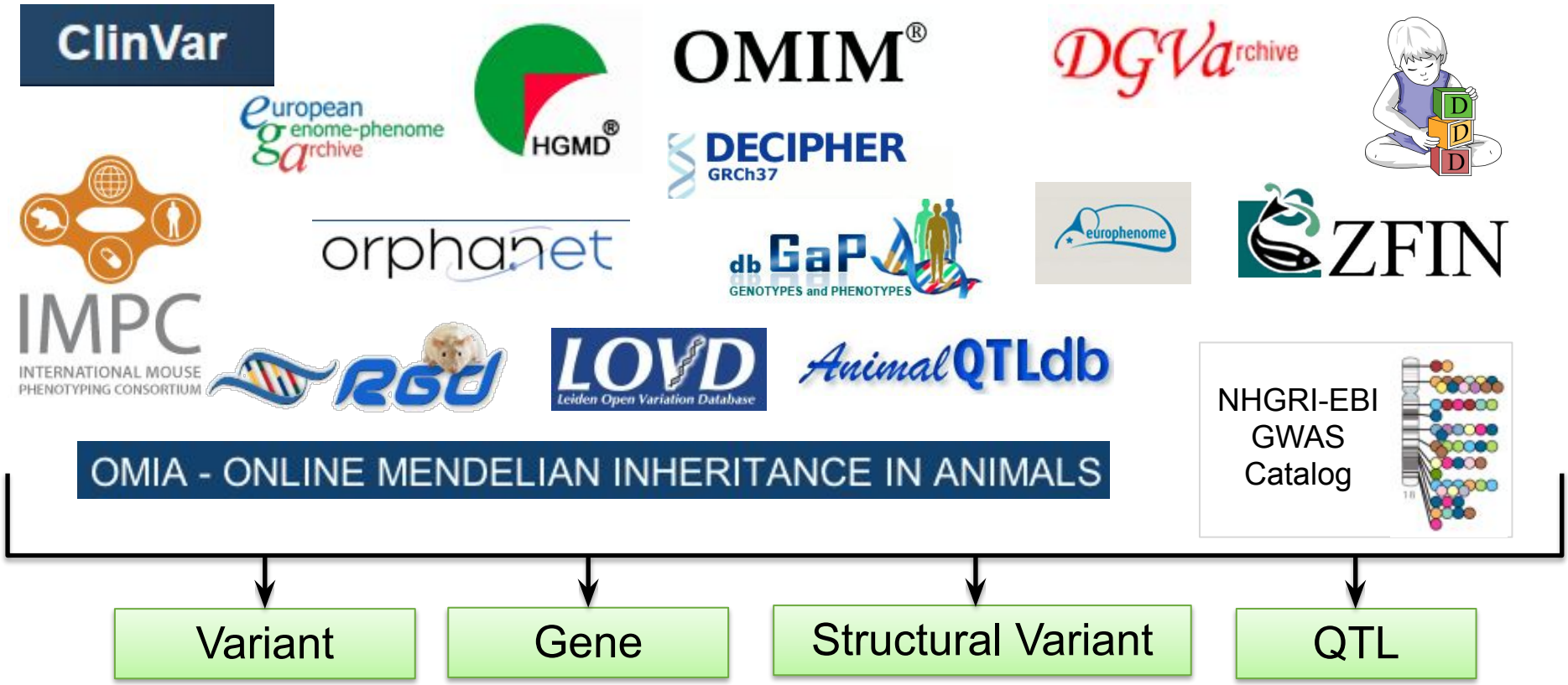
Linked data: Allele Frequencies

- 1000 Genomes: worldwide healthy WGS
- gnomAD: mixed disease/healthy WGS and exomes, skewed to most studied populations (included ExAC)
- UK10K: UK-wide disease exomes
- TOPMed: disease WGS
- NCBI Allele Frequency Aggregator (ALFA)



http://training.ensembl.org/events

Phenotype and Disease Data



Attributes types held for phenotype features include:

- clinical significance reported by ClinVar
- inheritance type
- reported genes /variants
- risk allele
- p-value
- odds ratio
- beta coefficient

http://training.ensembl.org/events

Ensembl variation process: Analysis

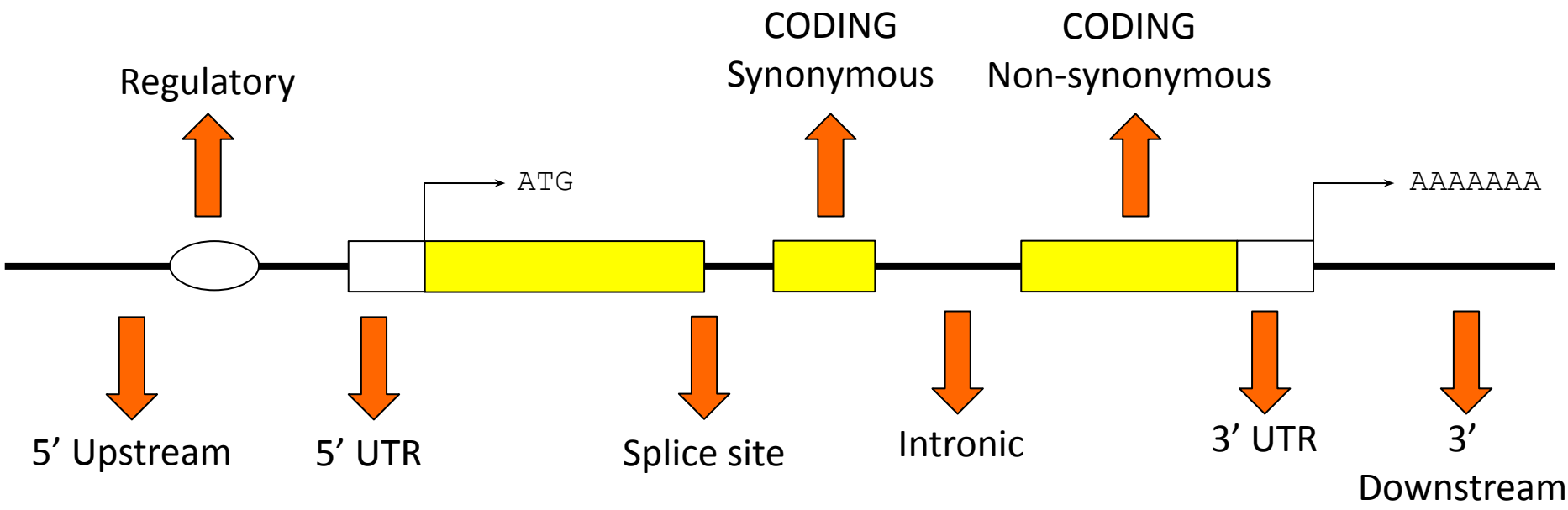
Variant import → Quality control → Linked data → **Ensembl analysis**

Ensembl predicts:

- Variant consequences
- Protein function prediction
- Linkage disequilibrium data
- Variant conservation across species

http://www.ensembl.org/info/genome/variation/prediction/index.html

Variation consequences



Consequence terms

* SO term	SO description	SO accession	Old Ensembl term
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Essential splice site
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	
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 three	SO:0001589	Frameshift coding
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
initiator_codon_variant	A codon variant that changes at least one base of the first codon of a transcript	SO:0001582	Non synonymous coding
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	
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	
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification
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 site
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	Partial codon
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous coding
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	
coding_sequence_variant	A sequence variant that changes the coding sequence	SO:0001580	Coding unknown
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	Within mature miRNA
5_prime_UTR_variant	A UTR variant of the 5' UTR	SO:0001623	5prime UTR
3_prime_UTR_variant	A UTR variant of the 3' UTR	SO:0001624	3prime UTR
intron_variant	A transcript variant occurring within an intron	SO:0001627	Intronic
NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO:0001621	NMD transcript
non_coding_exon_variant	A sequence variant that changes non-coding exon sequence	SO:0001792	Within non coding gene
nc_transcript_variant	A transcript variant of a non coding RNA	SO:0001619	
upstream_gene_variant	A sequence variant located 5' of a gene	SO:0001631	Upstream
downstream_gene_variant	A sequence variant located 3' of a gene	SO:0001632	Downstream
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	Regulatory region
regulatory_region_variant	A sequence variant located within a regulatory region	SO:0001566	
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
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

http://www.ensembl.org/info/docs/variation/predicted_data.html

http://training.ensembl.org/events

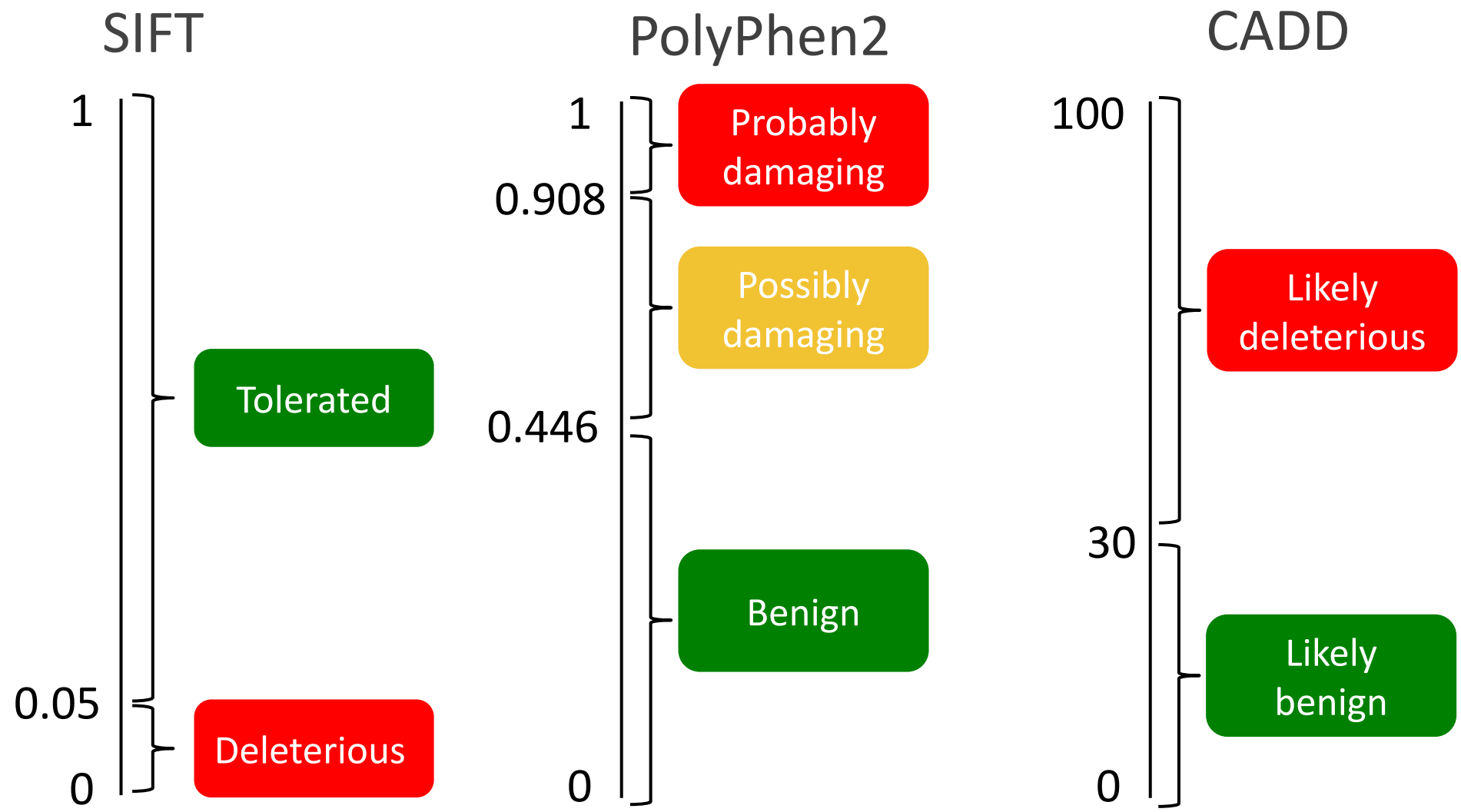
Missense variants- pathogenicity

Various algorithms score or rank changes in amino acid sequence based on:

- How well conserved the protein is
- The chemical change in the amino acid position
- 3D structure and domains

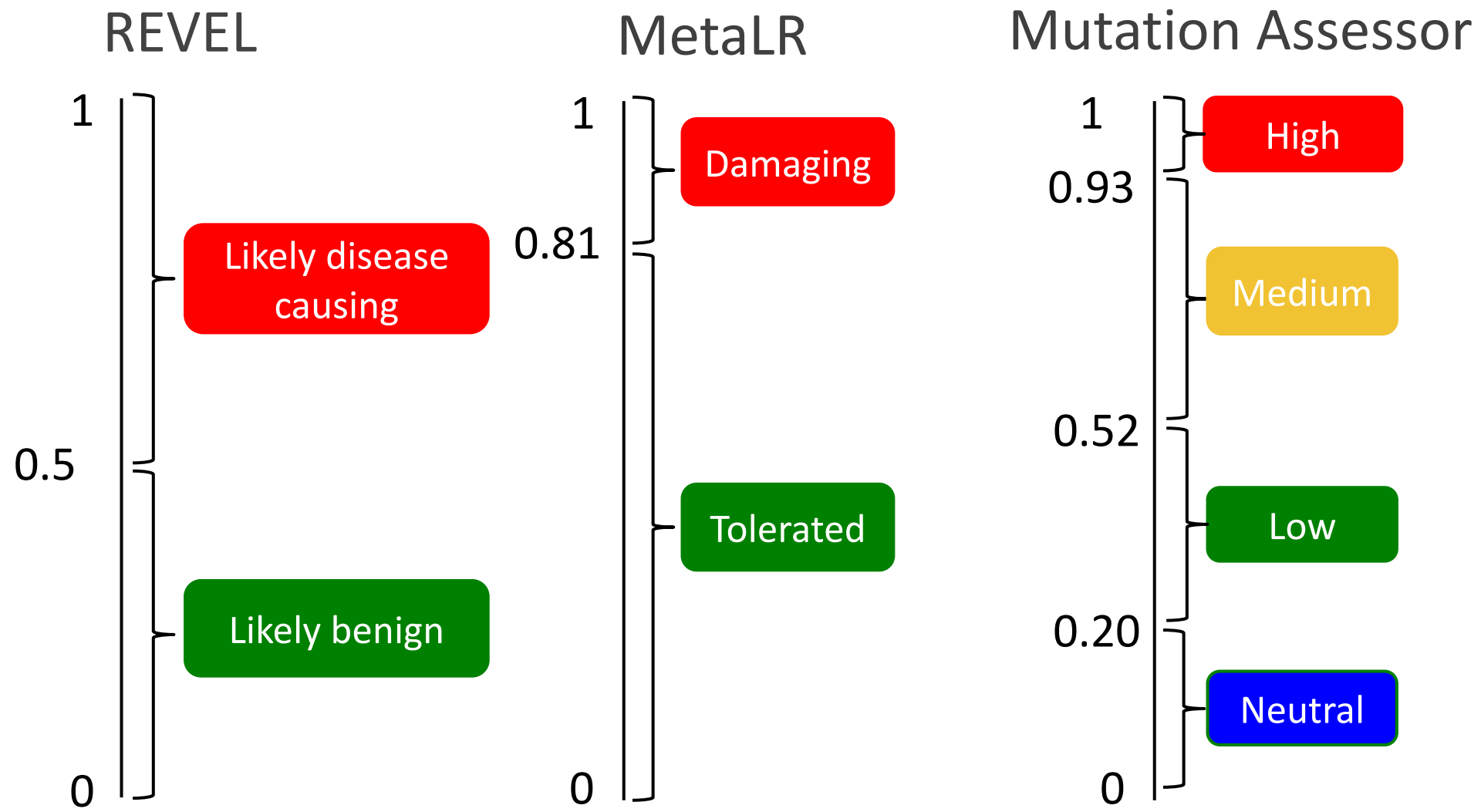
- These are predictions, not facts
- A prediction will never be as good as experimental validation

Missense variants- pathogenicity



http://training.ensembl.org/events

Missense variants- pathogenicity



Reference alleles


BL


AL


CM


IM

BL102
AL476
CM553
IM768

BL102

AGTCGTAGCTAGC**T**AGGCCATAGGCGA

Frequency T = 0.05, frequency G = 0.95
G is the allele in all primates
T causes disease susceptibility

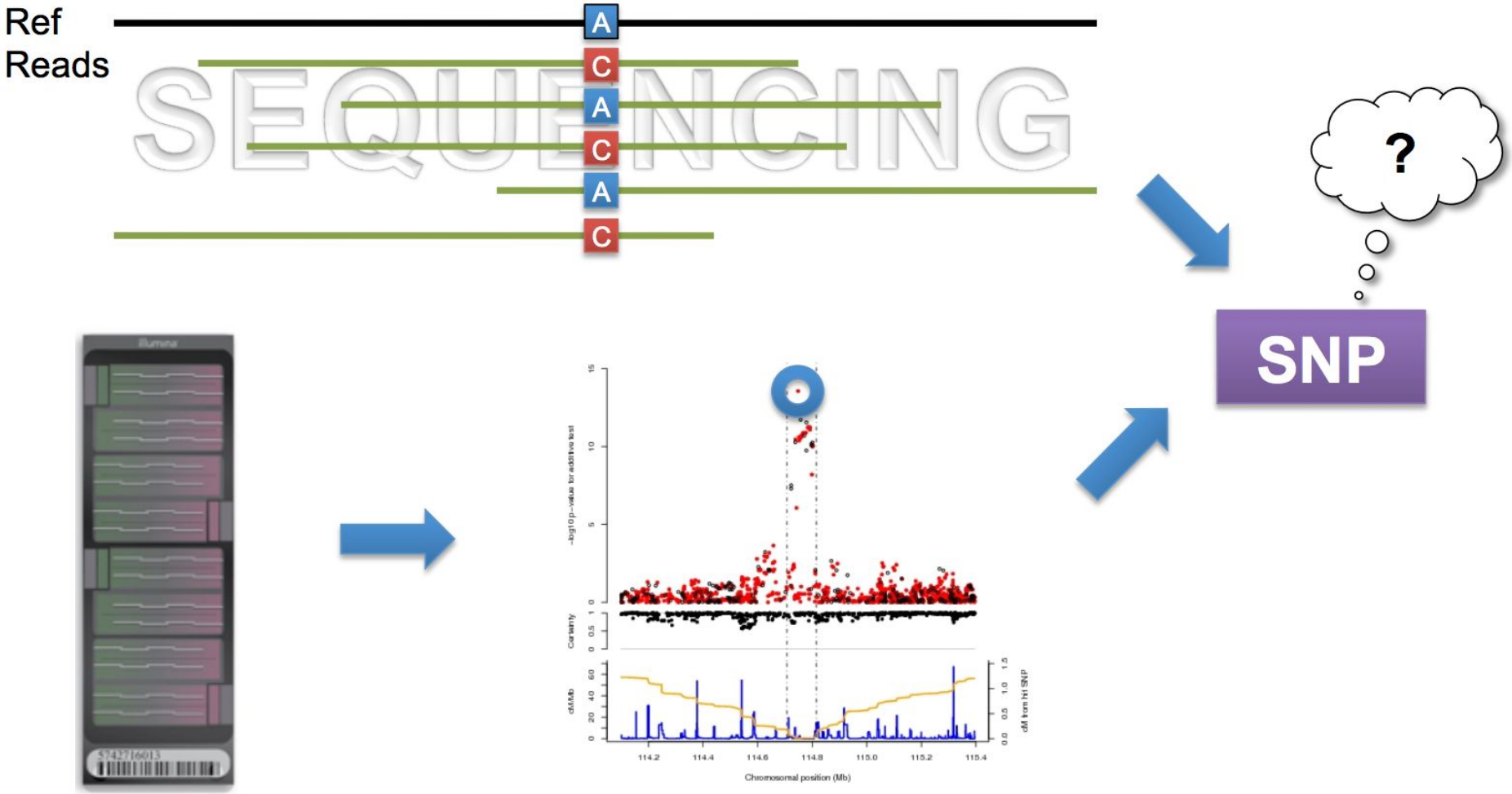
T is allele in the contig used
∴ T is the reference allele
∴ G is the alternate allele
∴ Alleles are T/G

Hands on

- We're going to look at a gene *MCM6* to find variants in the gene.
- We will look at the region of *MCM6* to find variants in the region.
- We will look at a variant rs4988235 to find more information about it.

- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

What is the VEP for?



What can I do with the VEP?

A tool to predict and annotate the **functional consequences of variants**

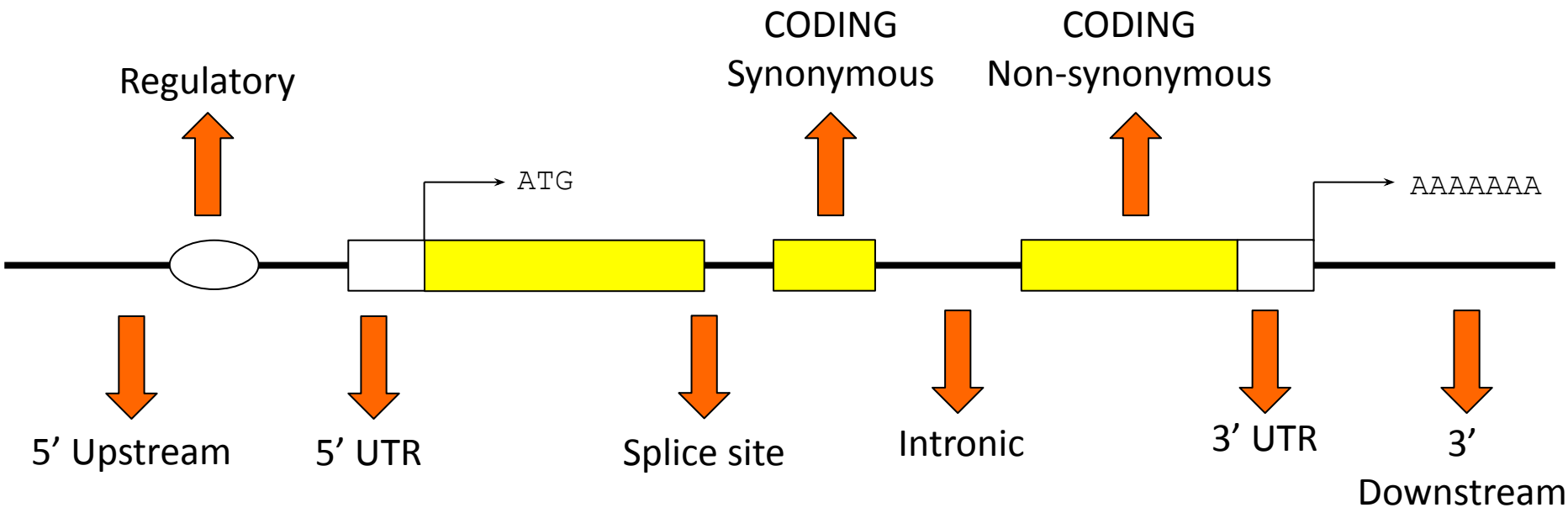
(SNPs, insertions, deletions, CNVs or structural variants)

Data input

- Variant coordinates
- VCF
- HGVS
- Variant IDs
- SPDI

- Affected gene, transcript and protein sequence
- Pathogenicity
- Frequency data
- Regulatory consequences
- Splicing consequences
- Literature citations

Variation consequences



SO term	SO description	SO accession	Old Ensembl term
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Essential splice site
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	
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 three	SO:0001589	Frameshift coding
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
initiator_codon_variant	A codon variant that changes at least one base of the first codon of a transcript	SO:0001582	Non synonymous coding
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	
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	
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification
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 site
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	Partial codon
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous coding
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	
coding_sequence_variant	A sequence variant that changes the coding sequence	SO:0001580	Coding unknown
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	Within mature miRNA
5_prime_UTR_variant	A UTR variant of the 5' UTR	SO:0001623	5prime UTR
3_prime_UTR_variant	A UTR variant of the 3' UTR	SO:0001624	3prime UTR
intron_variant	A transcript variant occurring within an intron	SO:0001627	Intronic
NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO:0001621	NMD transcript
non_coding_exon_variant	A sequence variant that changes non-coding exon sequence	SO:0001792	Within non coding gene
nc_transcript_variant	A transcript variant of a non coding RNA	SO:0001619	
upstream_gene_variant	A sequence variant located 5' of a gene	SO:0001631	Upstream
downstream_gene_variant	A sequence variant located 3' of a gene	SO:0001632	Downstream

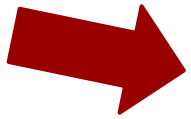
http://www.ensembl.org/info/docs/variation/predicted_data.html

http://training.ensembl.org/events

Use the VEP

Variant Effect Predictor

VeP



Web interface

- Point-and-click interface
- Suits smaller volumes of data

Documentation
Launch the web interface

Standalone perl script

- More options, more flexibility
- For large volumes of data

Documentation
Download latest version

REST API

- Language-independent API
- Simple URL-based queries
- GET single variants, POST many

Documentation

http://www.ensembl.org/info/docs/tools/vep/index.html

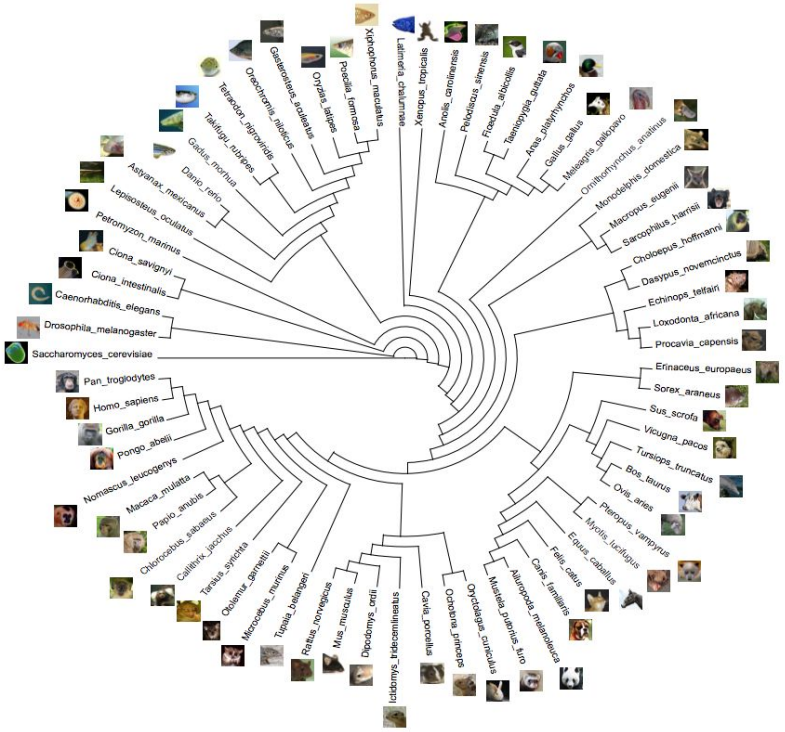
Hands on

We have identified five variants on human chromosome nine, an A deletion at 128328461, C->A at 128322349, C->G at 128323079, G->A at 128322917 and C->A at 128203516.

We will use the **Ensembl VEP** to determine:

- Whether my variants have already been annotated in Ensembl
- What genes are affected by my variants?
- Do any of my variants affect gene regulation?

Comparative Genomics



EMBL-EBI

Overview of the talk

- Comparative genomics: applications and species
- Gene trees
- Homology predictions
- Whole genome alignments
  - pairwise
  - multiple

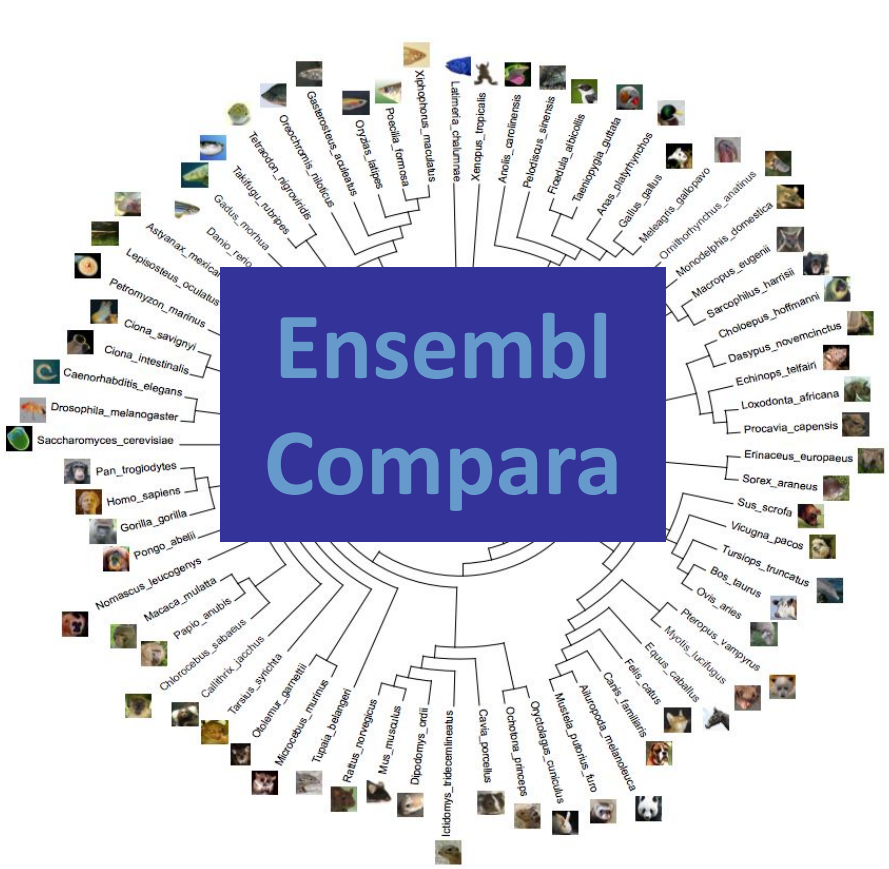
- Shared synteny

Applications of Comparative Genomics

Comparative genomics allows us to understand:

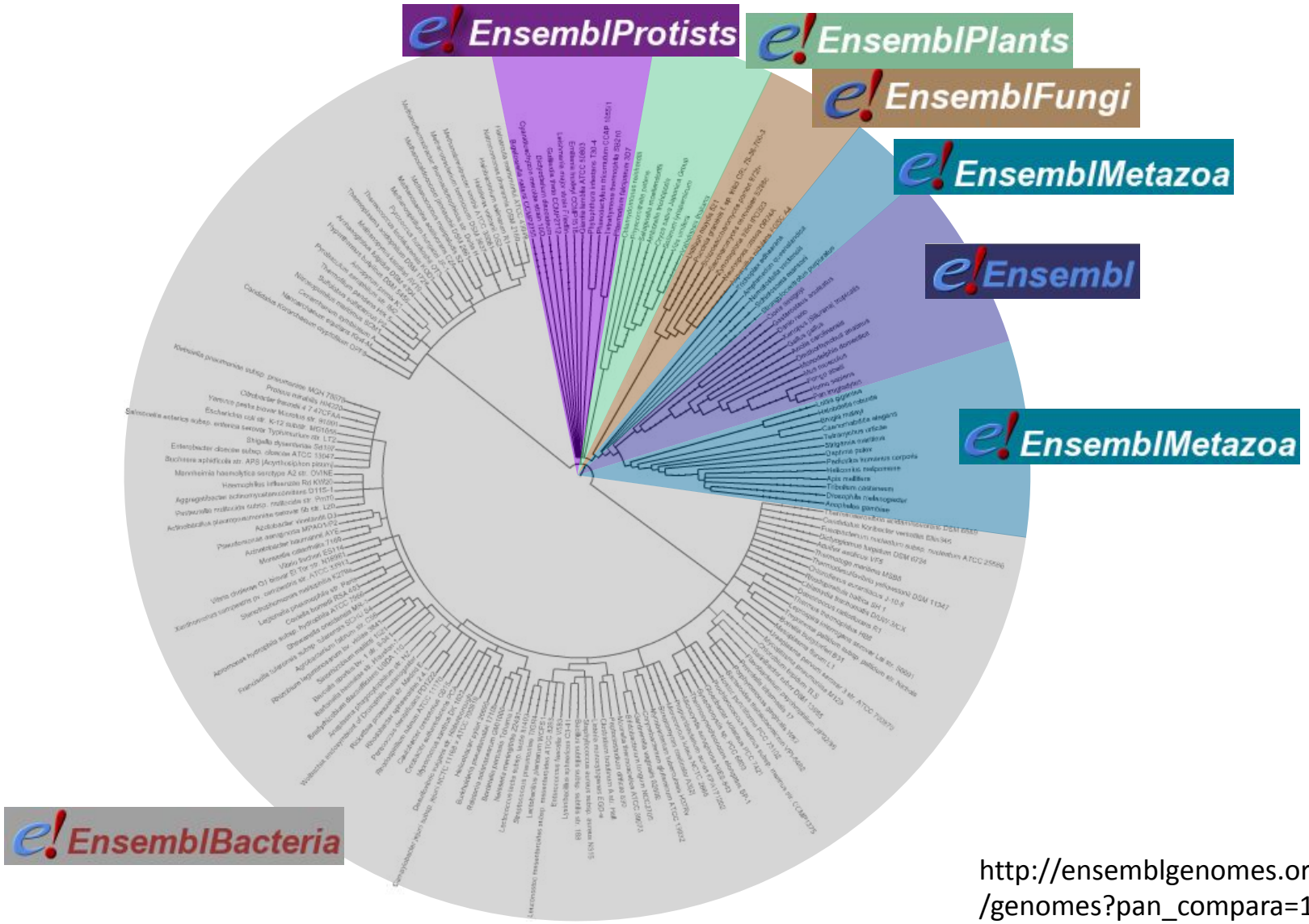
- vertebrate evolution
- differences between species at the genome level
- gene function based on homology
- the distribution of highly conserved regions

Comparative analysis by taxa



http://training.ensembl.org/events

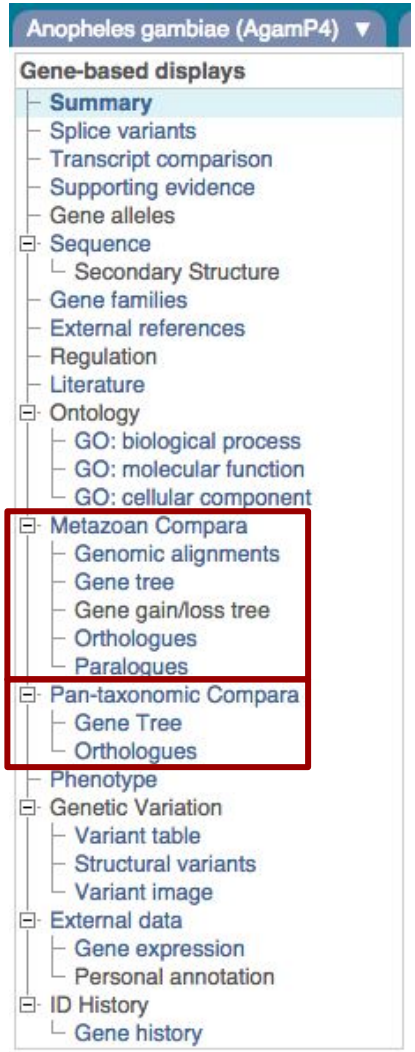
Pan-taxonomic compara

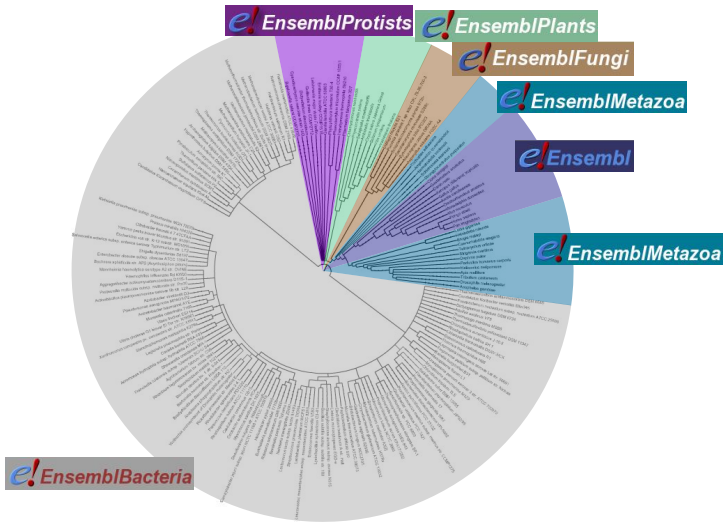


http://ensemblgenomes.org/info/genomes?pan_compara=1

http://training.ensembl.org/events

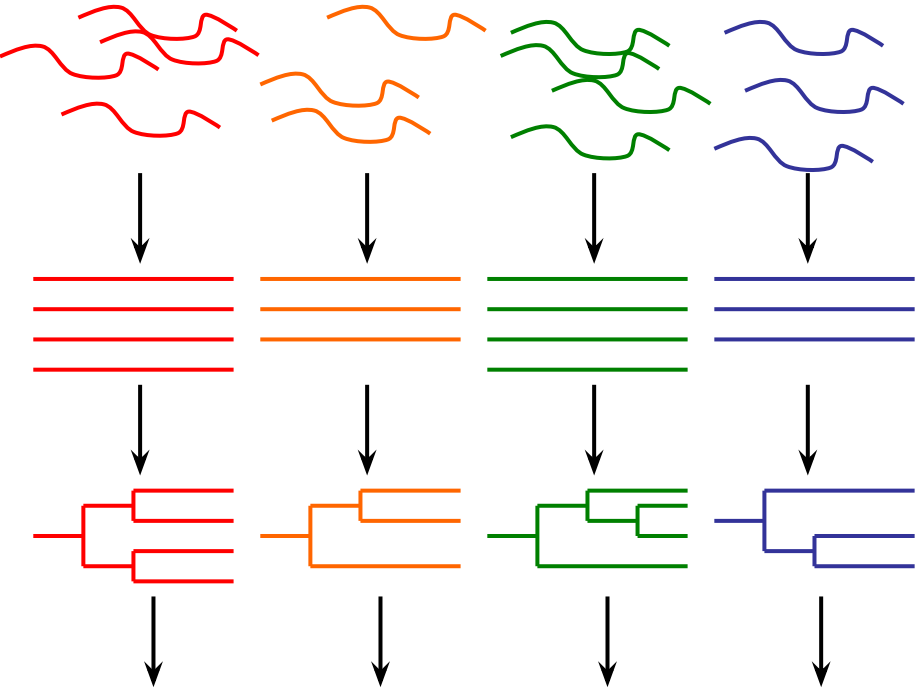
Pan-taxonomic compara





http://training.ensembl.org/events

Gene trees



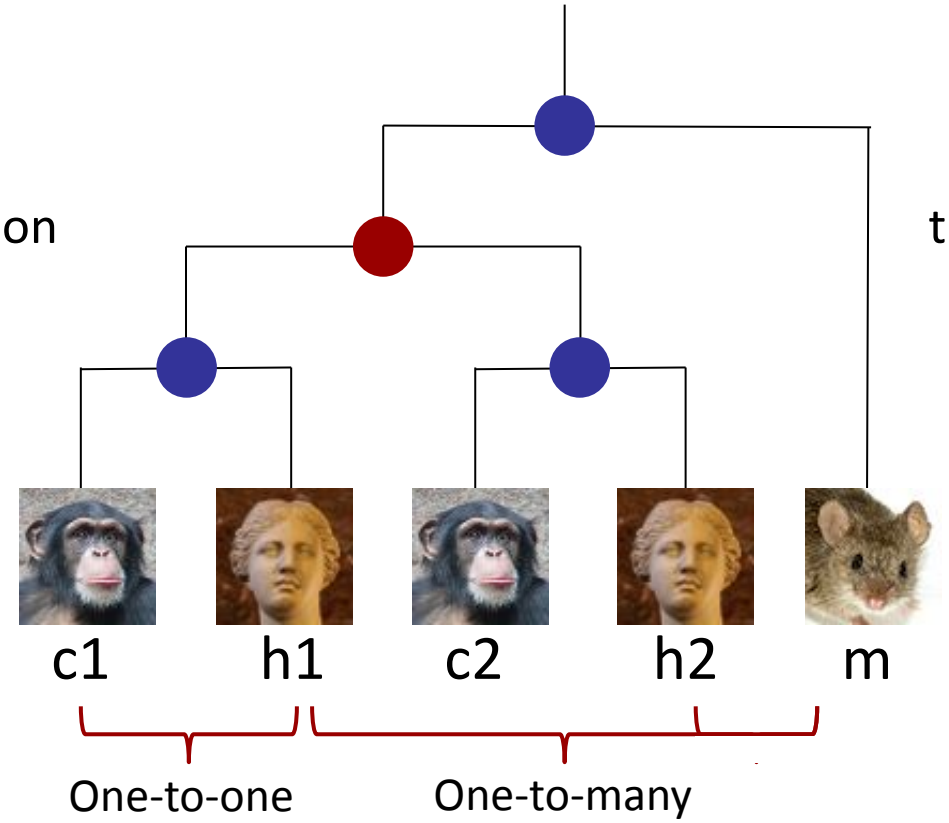
Orthologues and Paralogues

- Based on protein alignments
- Representative protein of each Ensembl gene
- Clustering, Blast, multiple alignments
- Reconciliation with species tree
- Orthologue/Paralogue inference

http://www.ensembl.org/info/docs/compara/homology_method.html

Homology relationships

Paralogues
Genes emerged through a duplication event
eg
c1 and c2
h1 and h2



Orthologues
Genes emerged through a speciation event
eg
c1 and h1
h2 and m
c2 and m

Speciation
Duplication

http://training.ensembl.org/events

Hands on

- We're going to look at a gene *BRCA2* to find homologues.

Whole genome alignments

- To identify highly conserved regions
  - sequences that evolve slowly
  - regions likely to be functional
  - both coding and non-coding sequences

- To spot trouble gene predictions
- To define syntenic regions

- Types: pairwise *versus* multiple

Pairwise alignments

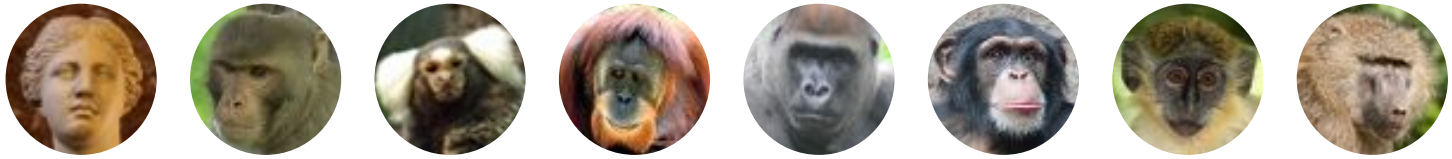
- Pairwise alignments with BLASTZ (older) LASTZ-net (newer)

- Human: everything
- Model organisms: related species
- Agricultural mammals: each other

http://www.ensembl.org/info/genome/compara/analyses.html

Multiple alignments

- EPO (Enredo-Pecan-Ortheus) analysis
  - fish, sauropsids, eutherian mammals, primates



- EPO-extended analysis (allows fragmented assemblies)
    - fish, sauropsids, eutherian, primates, pig breeds (+ other agricultural mammals)
- Mercator-Pecan analysis
    - amniota vertebrates (mammals+birds)
- Cactus
    - Murinae

http://www.ensembl.org/info/genome/compara/multiple_genome_alignments.html

Hands on

- We will look at a human genomic region 2:176087000-176202000 which contains the *HoxD* cluster to find alignments and conservation regions.

- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

Regulation of gene expression

EMBL-EBI

Regulation

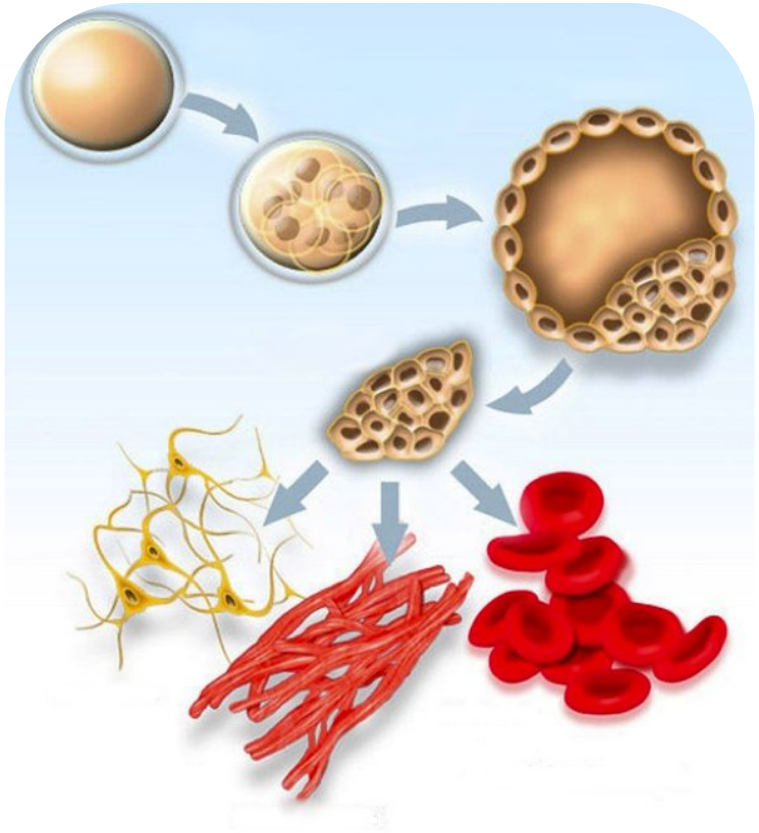
- Annotation of the genome with functional regulatory elements; promoters, enhancers, repressors

- Epigenetic marks
  - Histone modifications
  - DNA methylation

- Transcription Factor binding
- RNA Pol binding

- Predicted open/closed chromatin
  - DNase I sensitivity

One genome - many cell types



Essentially all cells of an individual share

- the same genome
- the same genes

But…

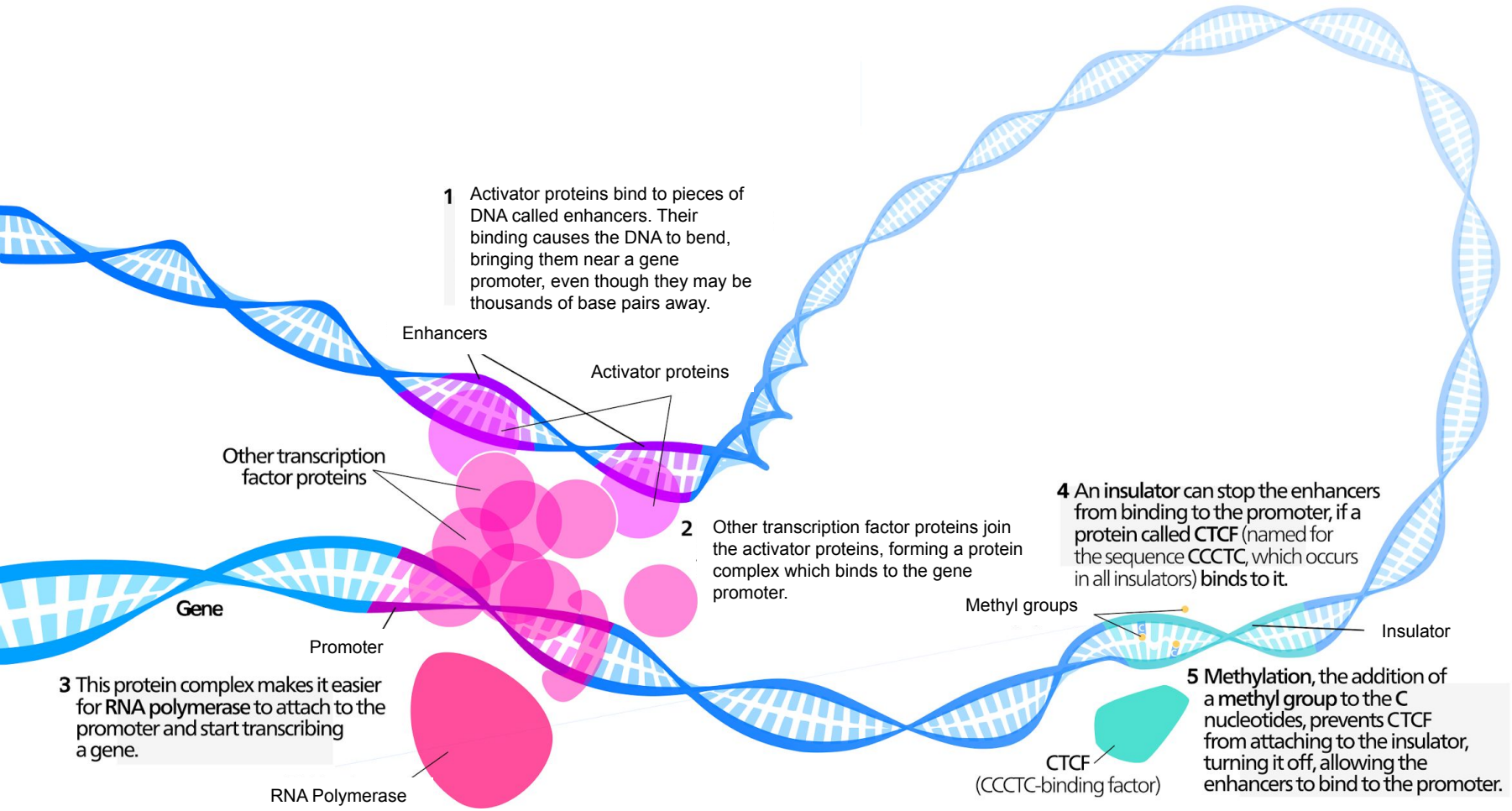
there are hundreds of **different cell types** with a clearly **distinct phenotype**

Difference?

Different gene expression profiles = different epigenomes*

*epigenome = cell type (sometimes cell line) = tissue type

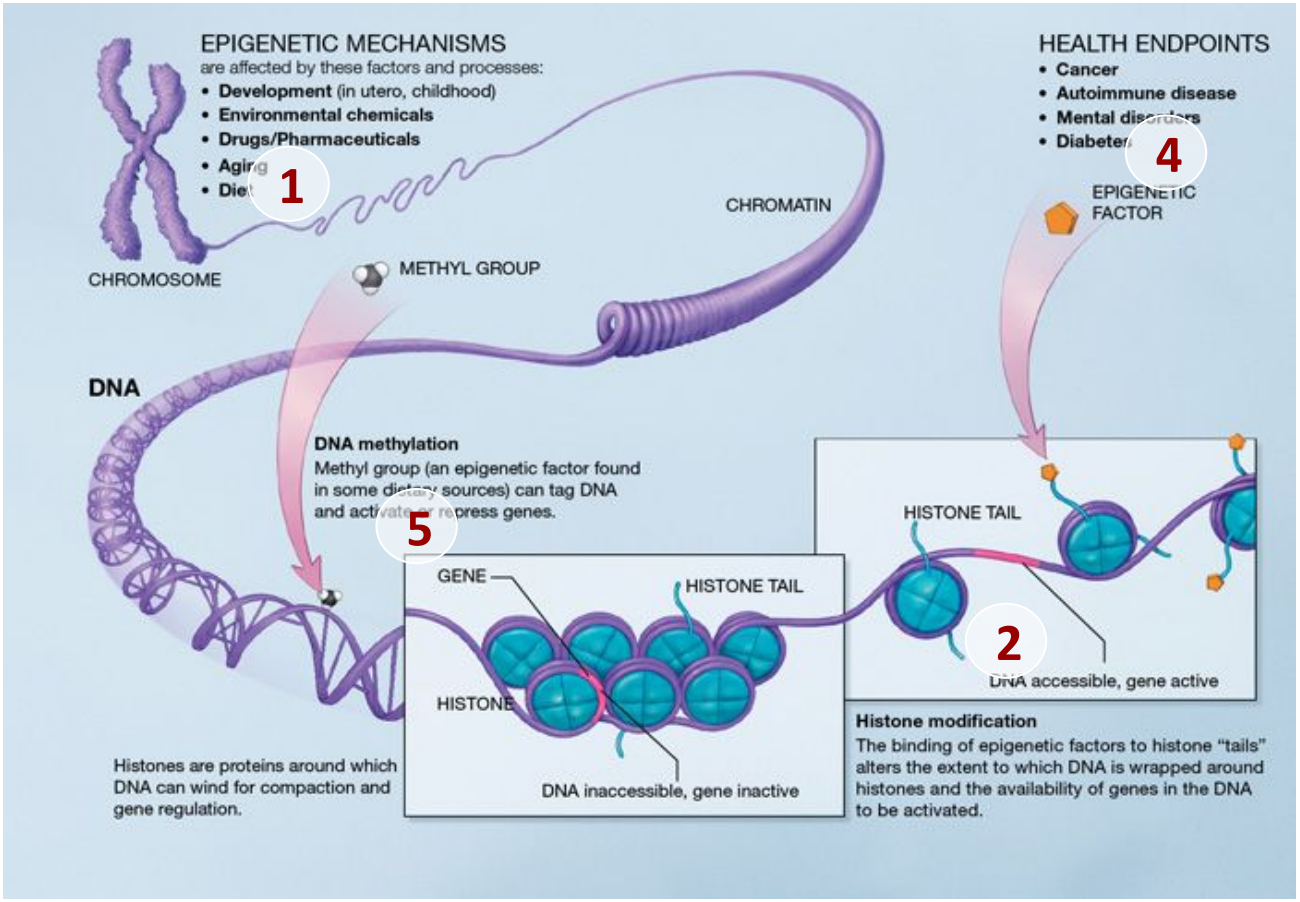
Layers of transcriptional regulation



"Transcription Factors" by Kelvinsong - Own work
https://commons.wikimedia.org/wiki/File:Transcription_Factors.svg#/media/File:Transcription_Factors.svg

Layers of transcriptional regulation

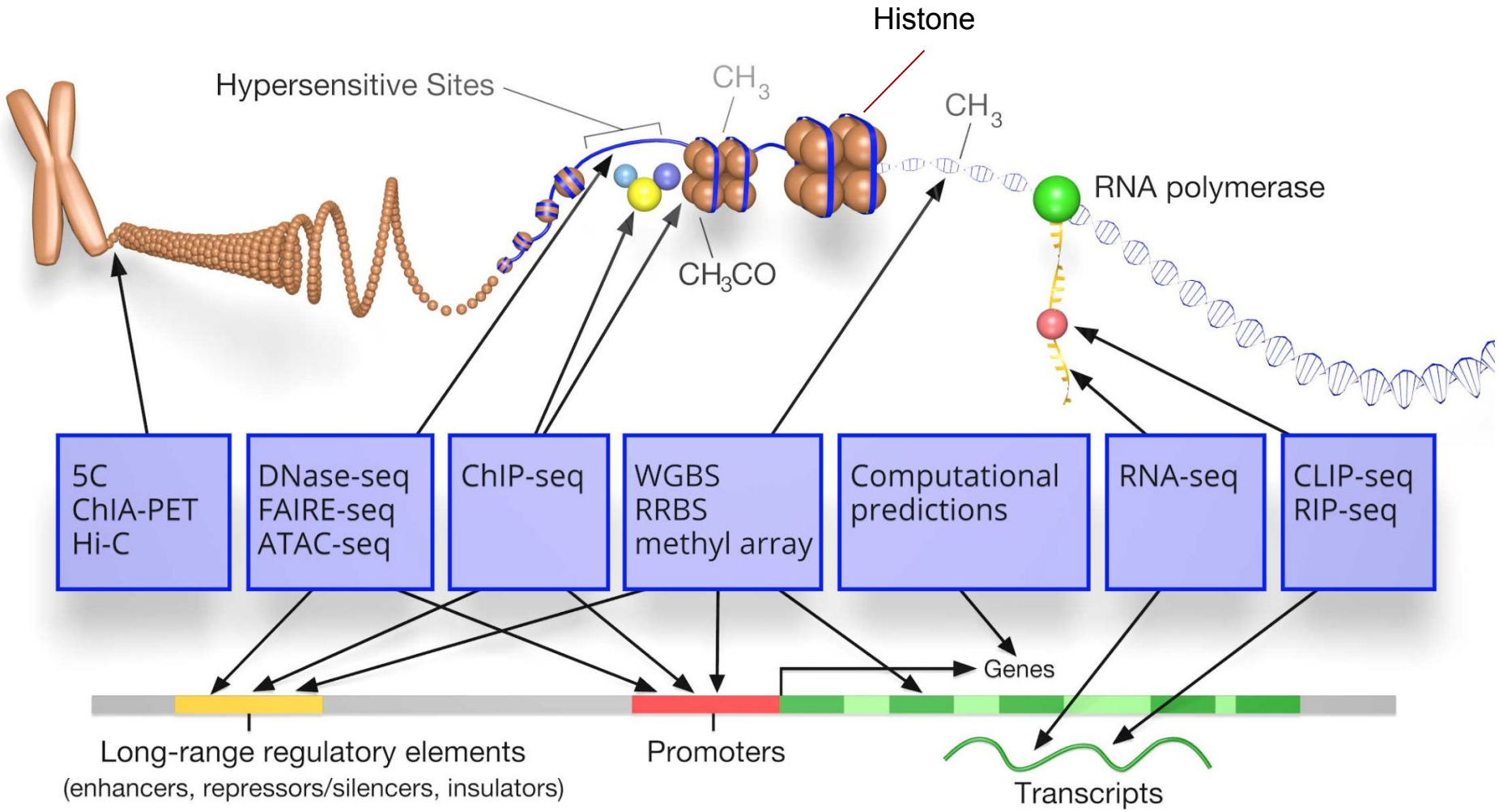
DNA quaternary structure: Histones and chromatin



1. DNA methylation
2. Open chromatin
3. Nucleosome positioning
4. Modification of histone tails
5. 3D conformation

Image taken from: http://commonfund.nih.gov/epigenomics/figure

Experimental data



source: https://www.encodeproject.org/

The Ensembl Regulatory Build

EMBL-EBI

What data does Ensembl generate?

Step 1: The processed **data** is imported from the various sources.

Step 2: Data processed to **predict the positions** of regulatory **features**

(i.e. promoters, promoter flanking regions, enhancers, CTCF binding sites, transcription factor binding sites & open chromatin)

Step 3: The **activity** of these features is predicted in different cell types

(i.e. active, poised, repressed, inactive, unknown)

Step 4: Display in the genome browser.

ensembl.org/info/genome/funcgen/regulatory_build.html

http://training.ensembl.org/events

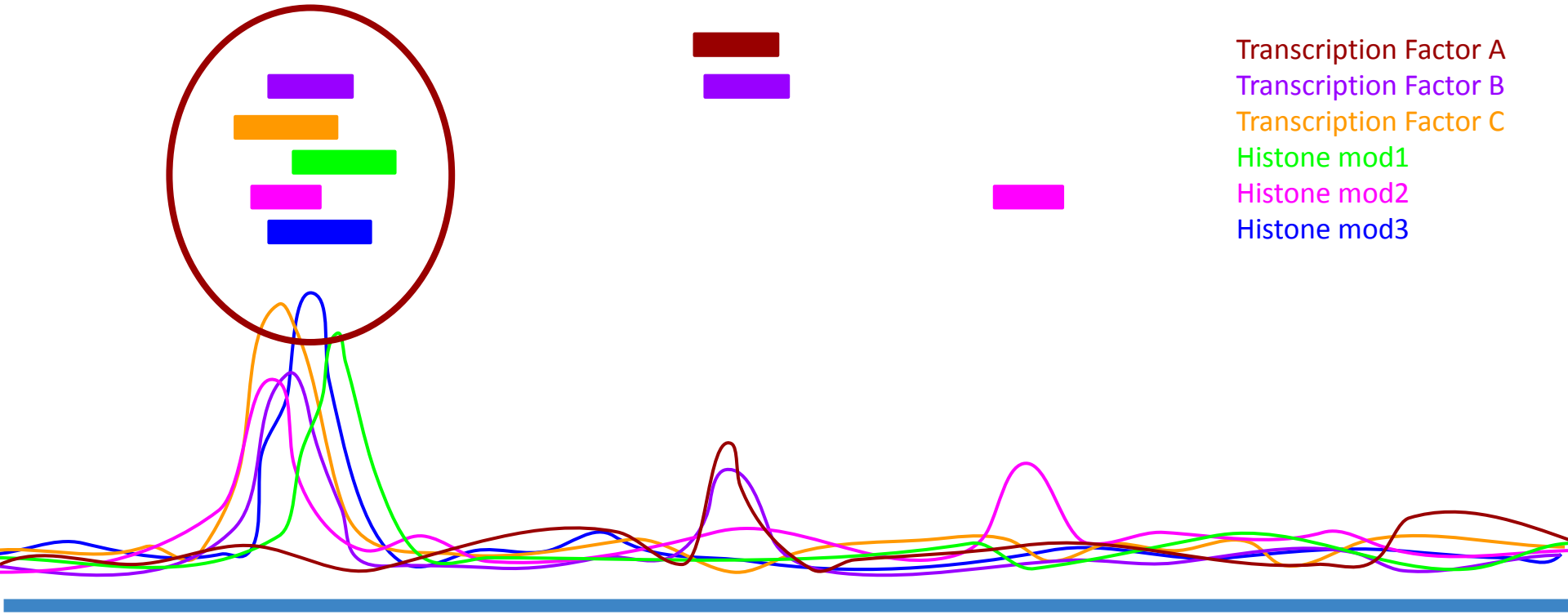
Where does the data come from?

Species	Data source	Assay types	Epigenomes
	ENCODE ROADMAP epigenomics PROJECT	- ChIP-seq (histone mods) - TF binding sites - RNApol - DNase sensitivity (open chromatin)	48 cultured cell lines
	ENCODE	- ChIP-seq (histone mods) - TF binding sites - RNApol - DNase sensitivity (open chromatin)	8 cultured cell lines
	BLUEPRINT epigenome	- ChIP-seq (histone mods) - DNase sensitivity (open chromatin)	20 primary cells from haematopoietic cell lineage (direct from human cells)

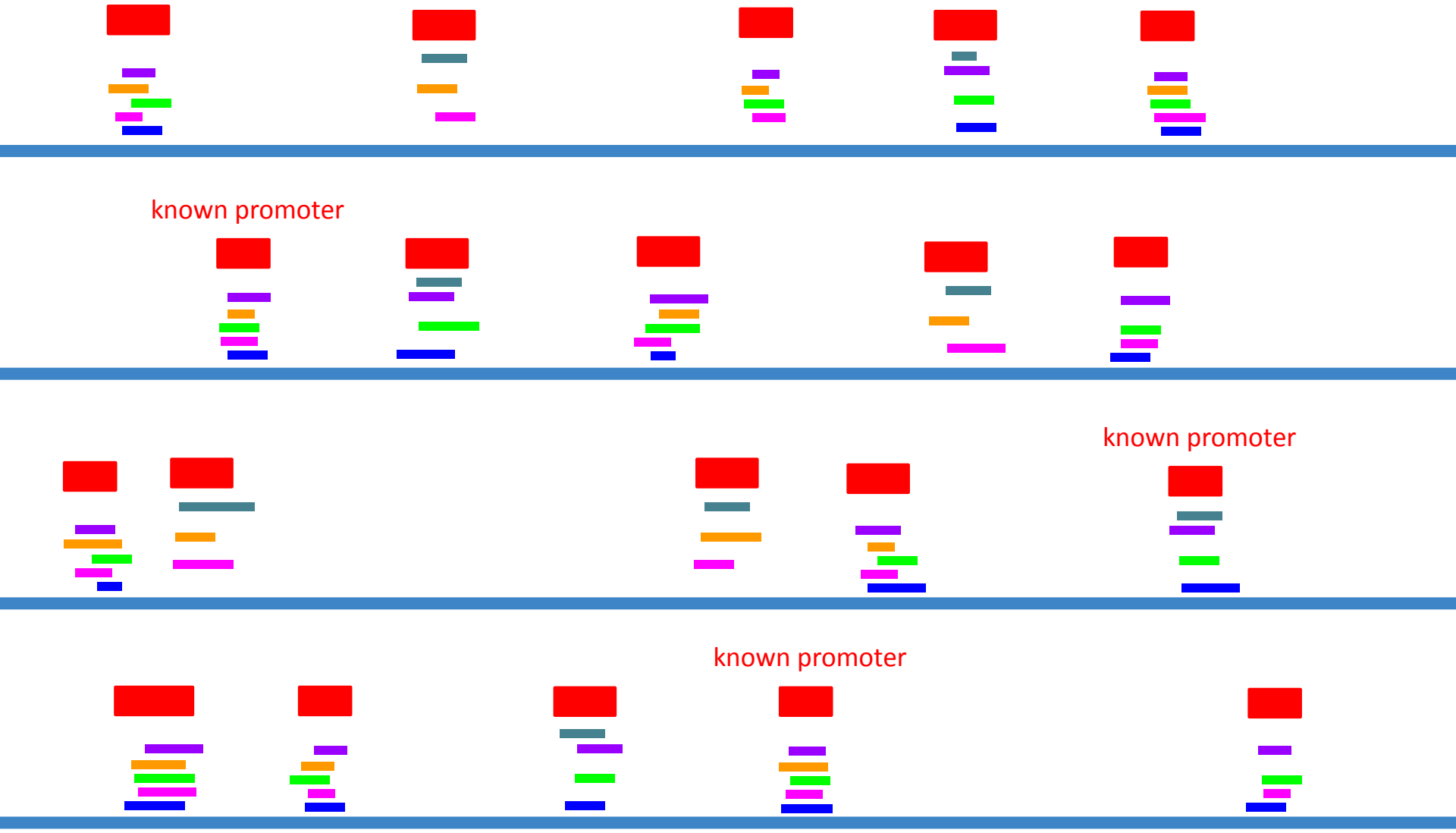
ensembl.org/info/genome/funcgen/regulation_sources.html

http://training.ensembl.org/events

Step 1: Raw data & peak calling



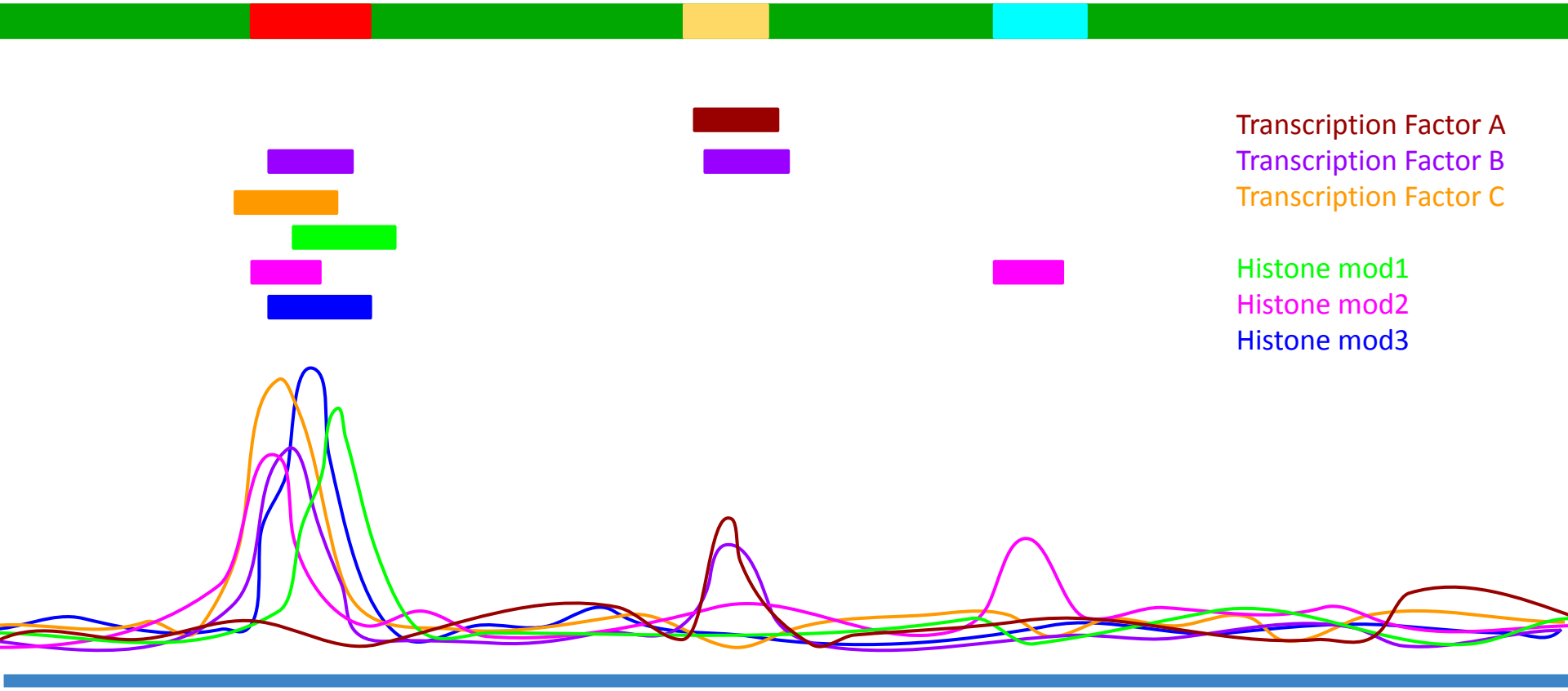
Step 2: Predicting feature positions



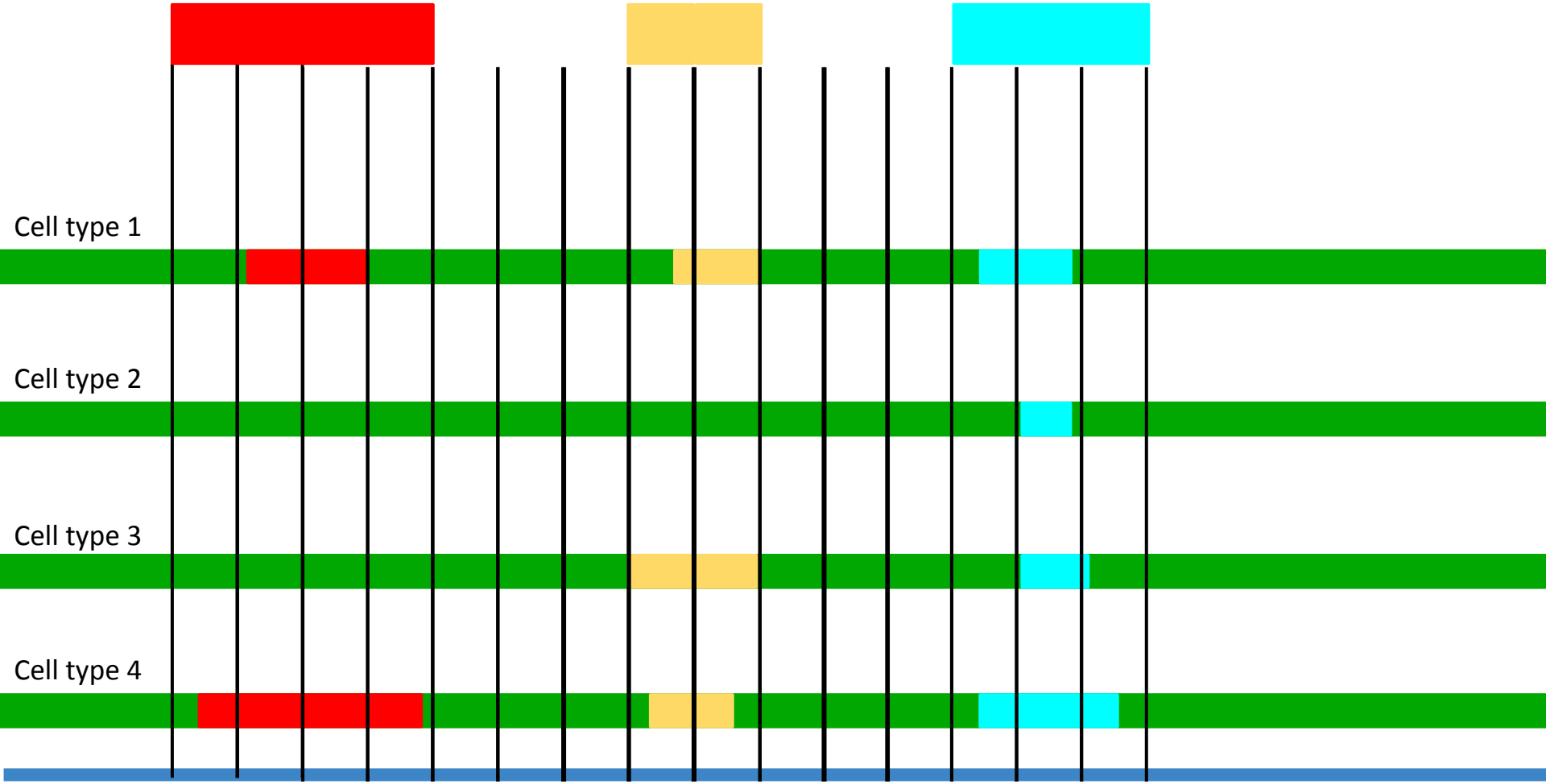
http://training.ensembl.org/events

Step 2: Predicting feature positions

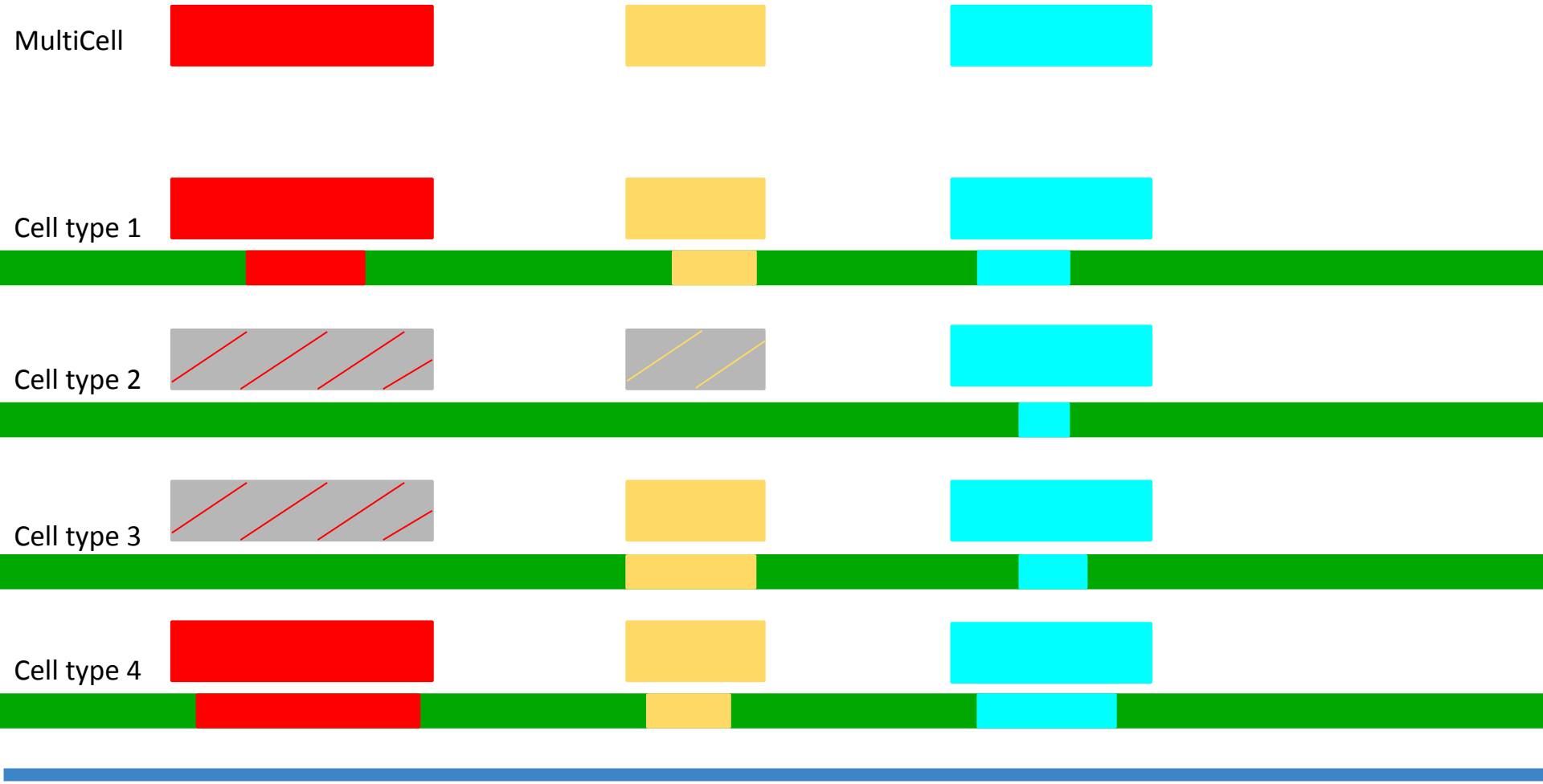
Segmentation



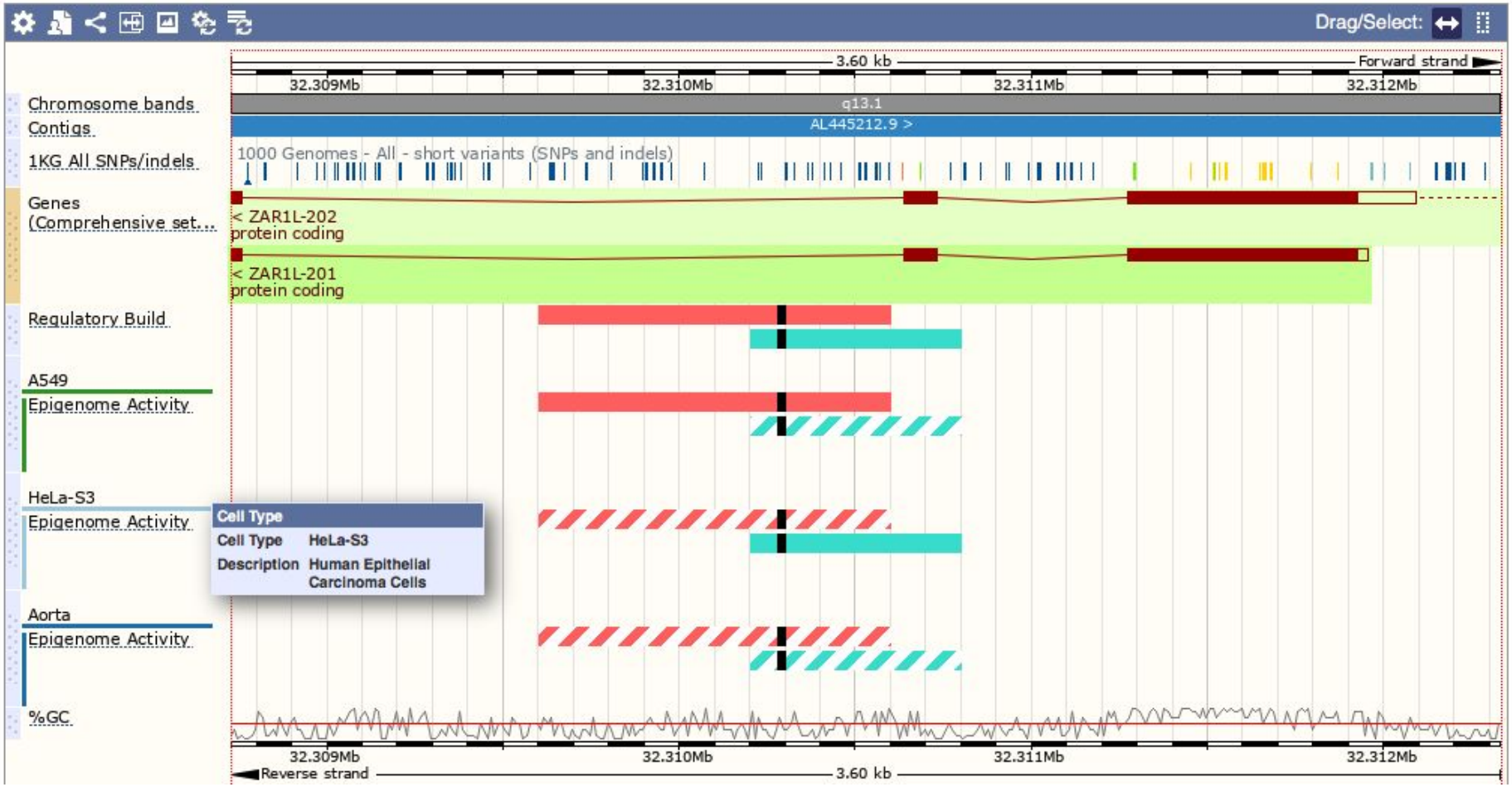
Step 3: Predicting cell type activity

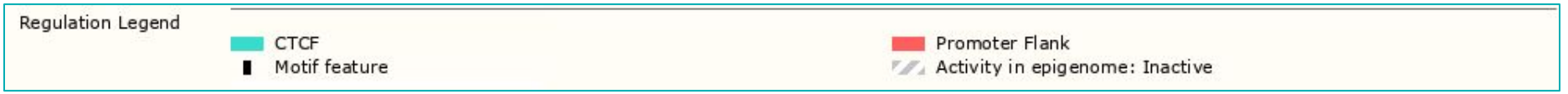


Step 3: Predicting cell type activity



Step 4: View in genome browser





Link to example page
http://training.ensembl.org/events

What does Ensembl not do?

We **do not**:

- Link regulatory features to genes
  - We allow you see the location of features.
- Link regulatory features to gene expression.
  - We have cell-line specific regulation data and tissue specific expression data.

You are required to make your own inferences about this data.

Regulatory data is incredibly complex and still in relative infancy.

There is no comprehensive database of regulation data...yet!

Other regulatory resources in Ensembl

- Human and mouse  
  - VISTA elements (enhancers)
  - TarBase miRNA targets
- Human only 
  - Fantom TSS & enhancers
  - GTEX eQTLs
- Methylation (WGBS/RRBS) data from ENCODE
- Microarray probe mapping to transcripts

http://www.ensembl.org/Help/View?id=142

Future directions…

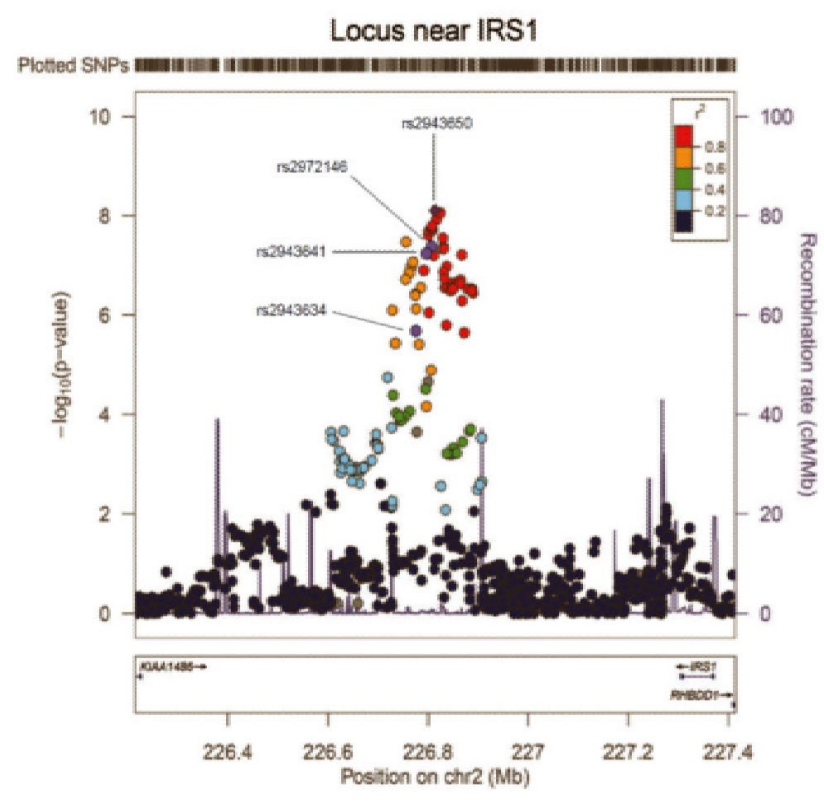
FAANG
Functional Annotation of Animal Genomes

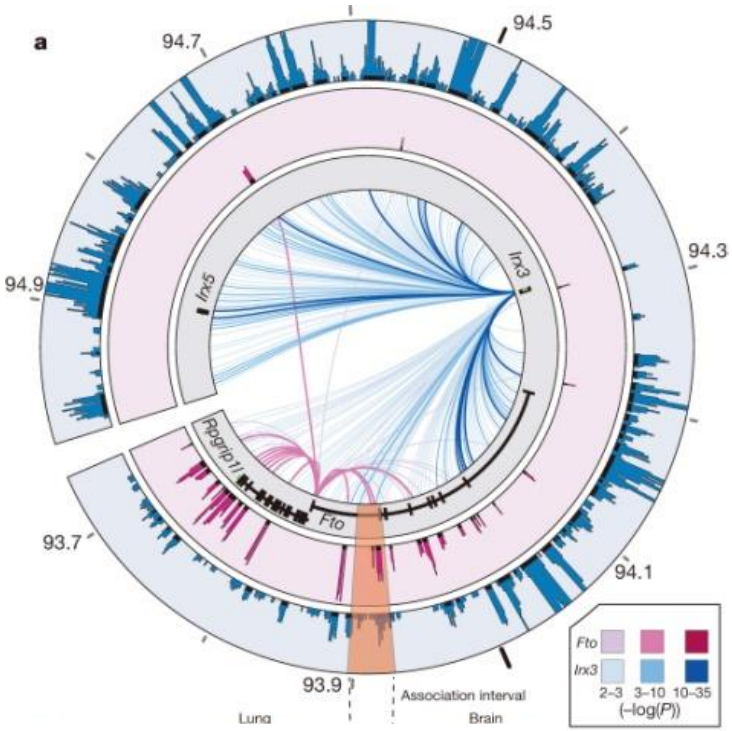
Zebrafish **Genome Biology**
CSIR Institute of Genomics & Integrative Biology

http://training.ensembl.org/events

Future directions…

Using eQTL & Hi-C to link regulatory regions to genes





Kilpeläinen TO et al., Nat Genet. 2011 Jun 26;43(8):753-60.
Smemo S et al. Nature. 2014 Mar 20;507(7492):371-5.

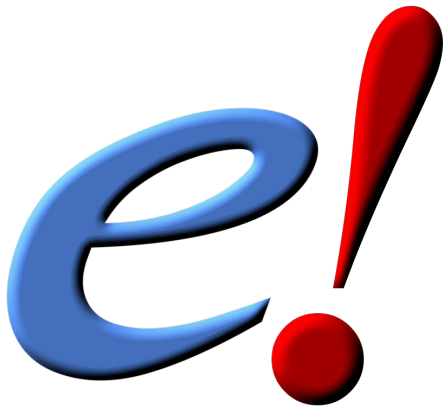
Hands on

- We're going to look at the region of a gene *KPNA2* to find regulatory features and explore what cells types they are active in and what evidence there is to show this.

- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

Data Mining with BioMart

EMBL-EBI



Outline of this session

- What is BioMart?
- The principle: 4 steps
- Demo and Exercises

What is BioMart?

- A tool in your browser:
  - Export Ensembl data with no programming required
  - Build queries with a few mouse clicks
  - Generates customisable datatables and files

Why use BioMart?

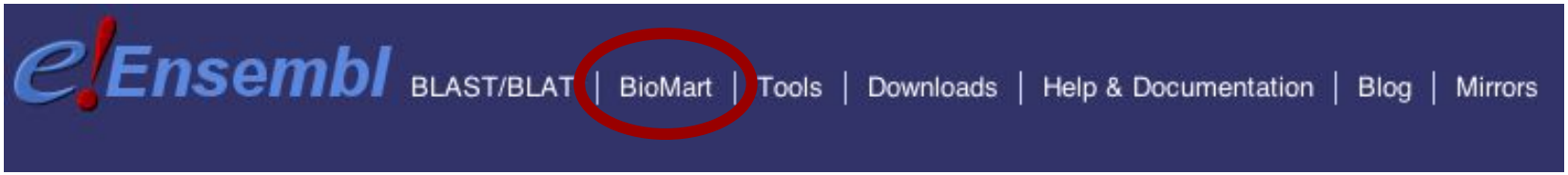
For things that would be time consuming/ difficult with the Ensembl browser

- Query multiple things (gene/ variants) at once:
  - ID conversions
  - Gene locations
  - Download sequences

- Export large amounts of data

Where to find BioMart

- www.ensembl.org/biomart/martview



- metazoa.ensembl.org/biomart/martview

http://training.ensembl.org/events

Availability

Ensembl

Ensembl Plants

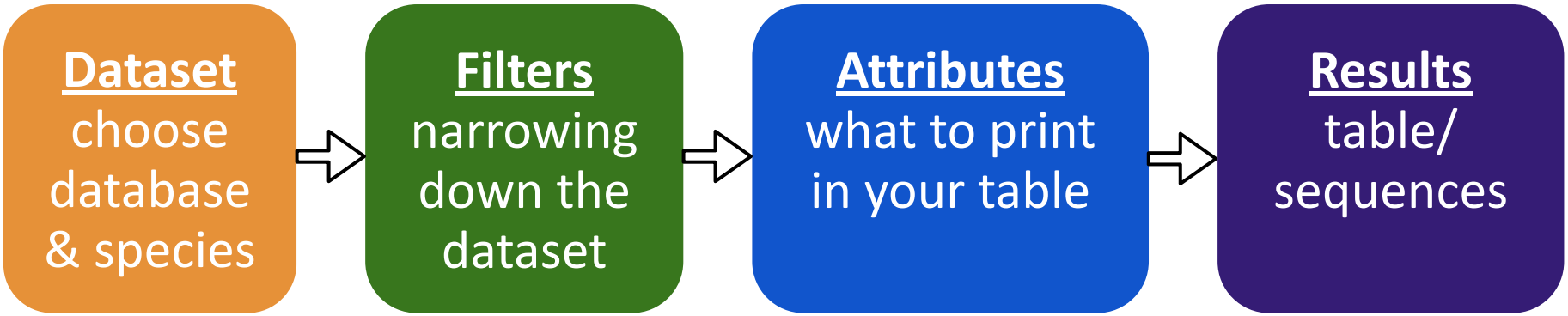
Ensembl Fungi (some exceptions)

Ensembl Metazoa

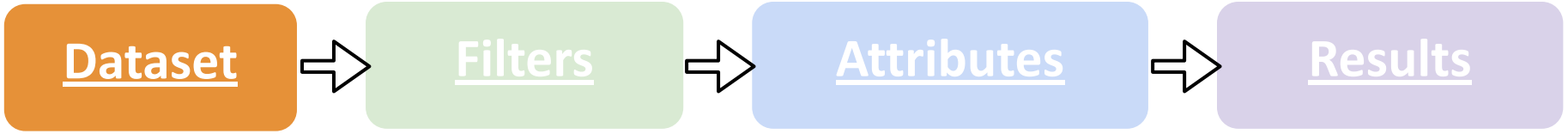
Ensembl Protists (some exceptions)

http://training.ensembl.org/events

How do I use BioMart?
The 4 steps



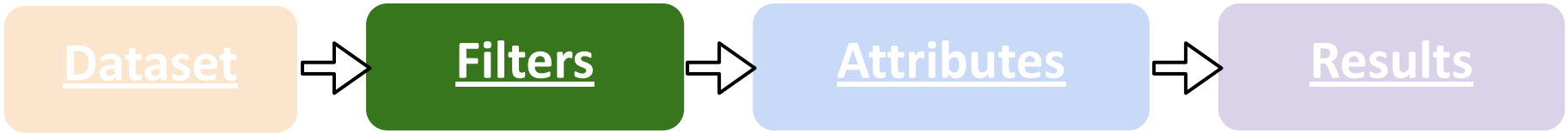
Step 1: Dataset



- Define the database that you want to search with your filters
    - Genes, Variation, Regulation

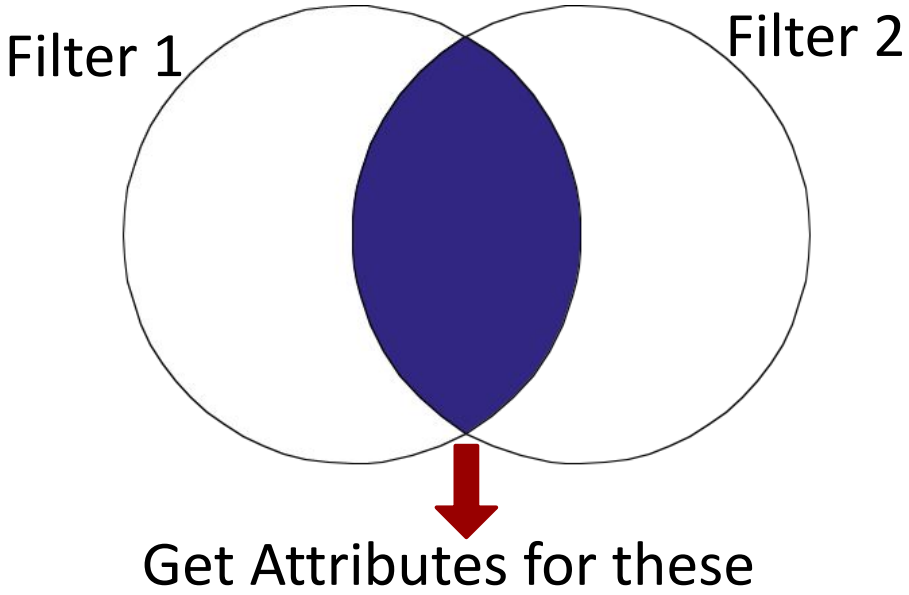
- Define the species

Step 2: Filters

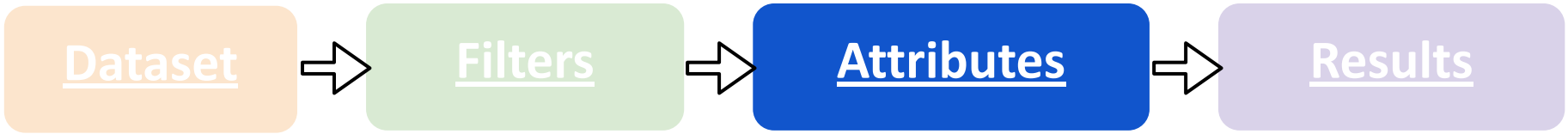


Define a (large) set of genes/variants by combinations of parameters, eg:

- A region
- A list of IDs
- Function (GO term)
- Phenotypes



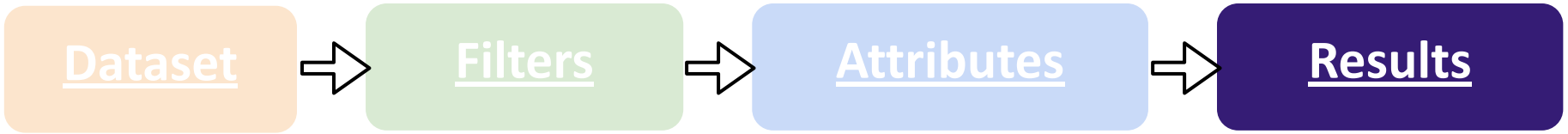
Step 3: Attributes



Define the data you want for that set, e.g:

- IDs
- Features
- Sequences
- Orthologues/Paralogues

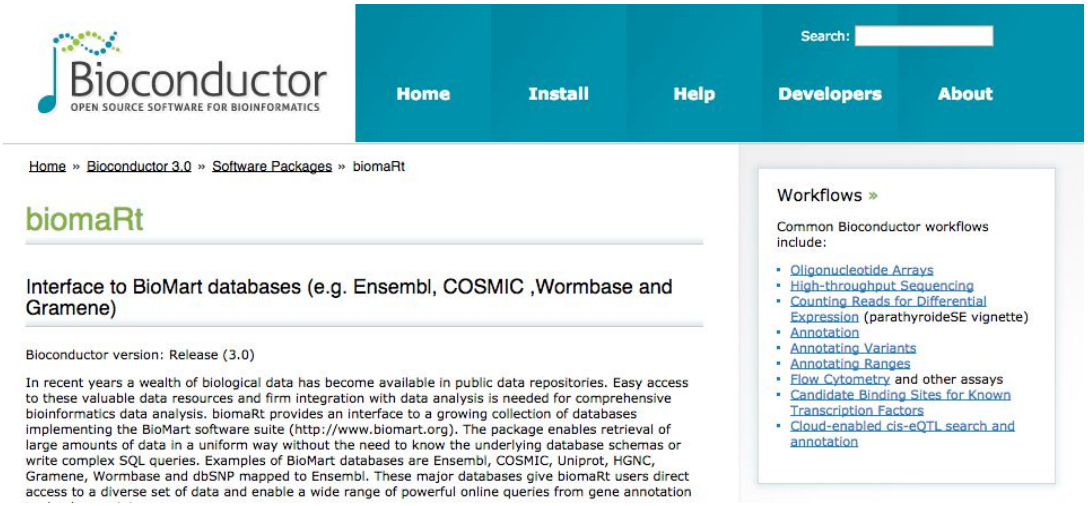
Step 4: Results



View and download the datatable in a number of formats:

- html
- tsv
- csv
- xls
- fasta

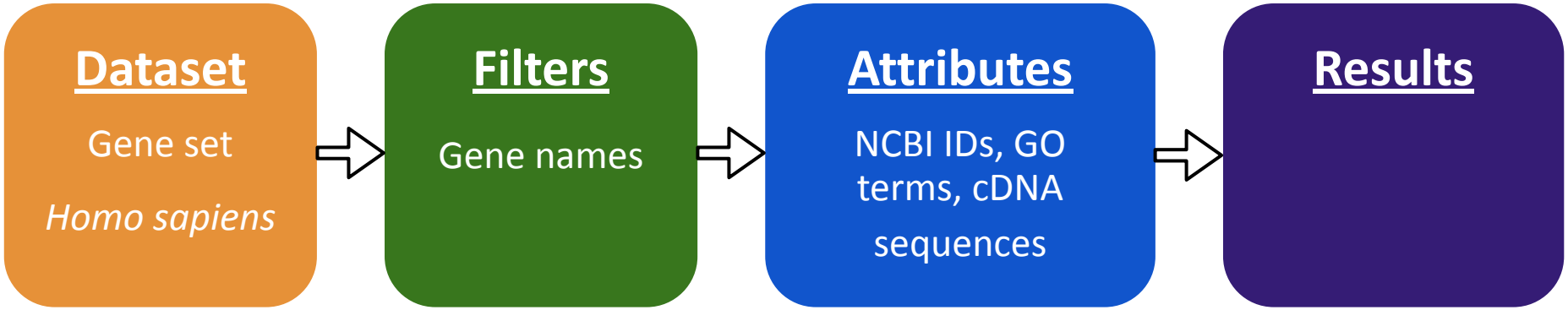
biomaRt



- Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data using R statistical programming language.
- Package for Biomart called BiomaRt : http://www.bioconductor.org/packages/release/bioc/html/biomaRt.html
- Easy to install in R :
  - source("http://bioconductor.org/biocLite.R")
  - biocLite("biomaRt")
- Documentation:http://www.bioconductor.org/packages/release/bioc/vignettes/biomaRt/inst/doc/biomaRt.pdf

Hands on

- We're going to look at a set of six *Homo sapiens* genes *ESPN, MYH9, USH1C, CISD2, THRB* and *BRCA2* and find out:
  - Their NCBI IDs
  - Their function via GO terms
  - Their cDNA sequences

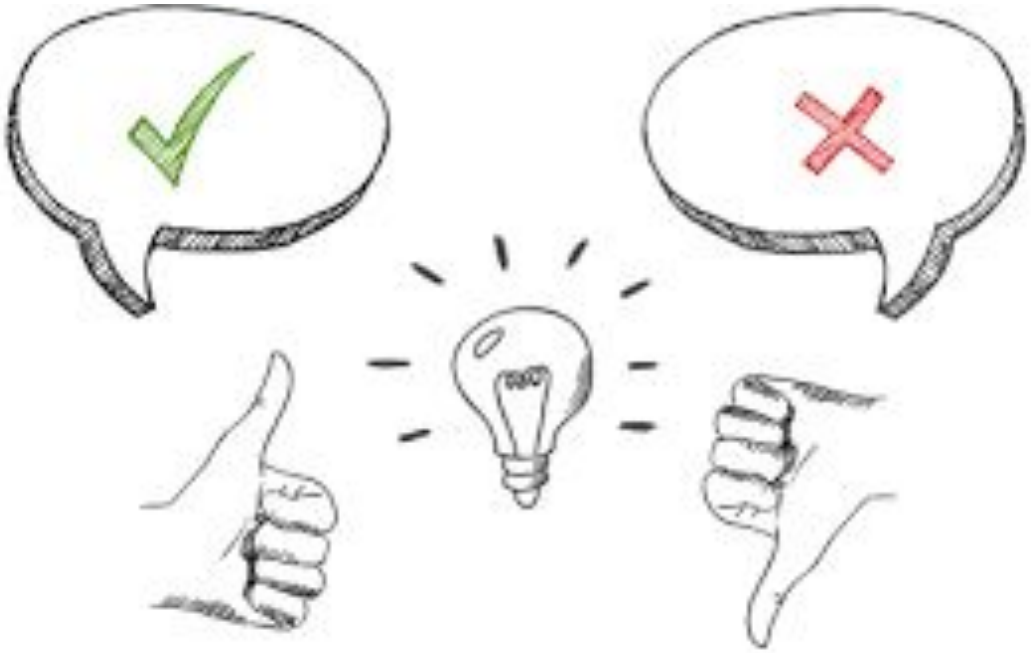


Hands on

- We're going to look at a set of six *Homo sapiens* genes *ESPN, MYH9, USH1C, CISD2, THRB* and *BRCA2* and find out:
  - Their NCBI IDs
  - Their function via GO terms
  - Their cDNA sequences

- There are more exercises than we have time for: pick and choose the ones most relevant to your work and you're welcome to finish them in your own time.

Feedback



training.ensembl.org/events

Wrap-up

Ensembl is a genome browser which integrates:

- gene annotation
- variation
- regulation
- comparative genomics

http://training.ensembl.org/events

How is all this data organised?

- Ensembl browser sites
  - Main website, Ensembl Genomes, GRCh37, *Archive!*
- BioMart 'DataMining tool'
- Ensembl Database (open source)
  - Perl-API, REST API, MySQL
- FTP download site
  - http://www.ensembl.org/info/data/ftp/index.html

Help and documentation

Course online
https://www.ebi.ac.uk/training/online/course-list

Tutorials www.ensembl.org/info/website/tutorials

Flash animations

www.youtube.com/user/EnsemblHelpdesk

http://u.youku.com/Ensemblhelpdesk

Email us helpdesk@ensembl.org

Ensembl public mailing lists dev@ensembl.org, announce@ensembl.org

http://training.ensembl.org/events

Publications

http://www.ensembl.org/info/about/publications.html

Cunningham, F *et al.*

Ensembl 2022

Nucleic Acids Research (Database Issue)

https://doi.org/10.1093/nar/gkab1049

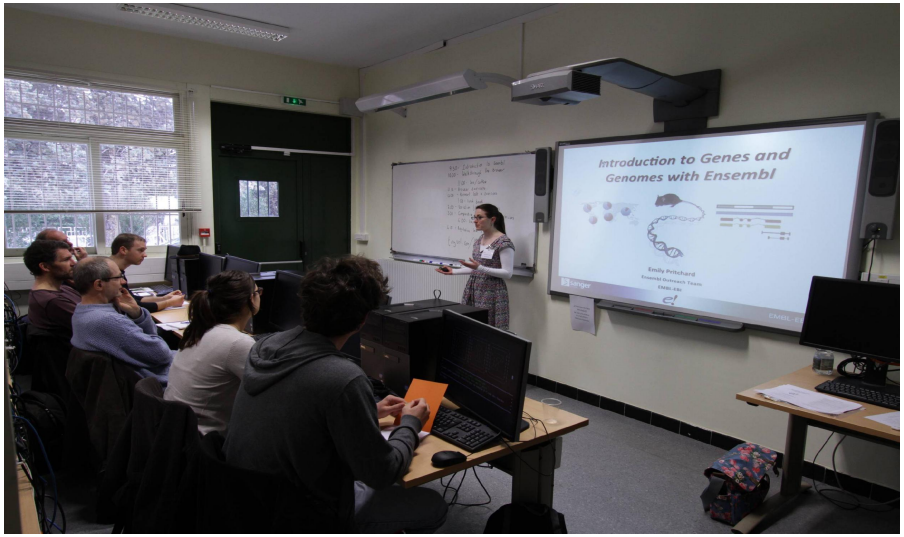
Andrew D Yates *et al.*
Ensembl Genomes 2022: an expanding genome resource for non-vertebrates
Nucleic Acids Research (Database Issue)
https://doi.org/10.1093/nar/gkab1007

Newman, V. *et al.*
The Ensembl Genome Browser: Strategies for Accessing Eukaryotic Genome Data
Methods in Molecular Biology (Clifton, N.J.), 01 Jan 2018, 1757:115-139
https://europepmc.org/article/MED/29761458

Recommend us to your friends

We can teach an Ensembl course at any institute for free (plus trainers' expenses in high income countries).

Email us: helpdesk@ensembl.org



Browser course

One day course on the Ensembl browser, aimed at wet-lab scientists.

REST API course

Half day course on the Ensembl REST API, aimed at bioinformaticians.

Train the Trainer course

One day course on delivering the Ensembl browser course.

training.ensembl.org/hosting

http://training.ensembl.org/events

Follow us

www.facebook.com/Ensembl.org

@Ensembl

@BENsembl

www.ensembl.info

http://training.ensembl.org/events

Ensembl Acknowledgements

The Entire Ensembl Team

Adam Frankish, Ahamed Imran Abdul Salam, Alexandra Bignell, Ameya Chaubal, Andrea Winterbottom, Andrew Berry, Andrew Parton, Andrey Azov, Andy Yates, Anja Thormann, Anmol Jaywant Hemrom, Anne Lyle, Astrid Gall, Benjamin Moore, Bethany Flint, Brandon Walts, Bruno Contreras-Moreira, Carla Cummins, Carlos Garcia Giron, Claire Davidson, Cristina Guijarro, Dan Sheppard, Daniel Zerbino, David Thybert, Denye Ogeh, Diana Lemos, Elizabeth Lewis, Emily Perry, Fergal Martin, Fiona Cunningham, Gareth Maslen, Gareth Williams, Garth Ilsley, Guy Naamati, Helen Schuilenburg, IF Barnes, Ilias Lavidas, Irina Armean, James Allen, Jamie Allen, Jane Loveland, Jonathan Mudge, Jorge Alvarez-Jarreta, Jose Carlos Marugan, Jose Manuel Gonzalez Martinez, Jyothish Bhai, Kamalkumar Jayantilal Dodiya, Kevin Howe, Kieron Taylor, Kostas Billis, Lahcen Campbell, Leanne Haggerty, Luca Da Rin Fioretto, Magali Ruffier, Manoj Sakthivel, Manuel Carbajo Martinez, Marc Chakiachvili, Marek Szuba, Marie-Marthe Suner, Matthew Hardy, Matthew Russell, Matthieu Barba, Matthieu Muffato, Michael Paulini, Michal Szpak, Mike Kay, Mikkel Christensen, Mira Sycheva, Nick Langridge, Nishadi De Silva, Osagie Izuogu, Paul Davis, Paul Flicek, Premanand Achuthan, Reham Fatima, Ridwan Amode, Ruth Bennett, Sanjay Boddu, Sarah Donaldson, Sarah Hunt, Shamika Mohanan, Stephen Trevanion, Thibaut Hourlier, Thomas Juettemann, Thomas Maurel, Tiago Grego, Toby Hunt, Tuan Le, Vasili Sitnik

Funding

wellcome

EMBL

National Human Genome Research Institute

Open Targets

SAVE THE TASMANIAN DEVIL PROGRAM APPEAL

BBSRC bioscience for the future

TRANSFORMING GENETIC MEDICINE INITIATIVE

Co-funded by the European Union

Training materials

- Ensembl training materials are protected by a CC BY license
- http://creativecommons.org/licenses/by/4.0/
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers
- http://www.ensembl.org/info/about/publications.html