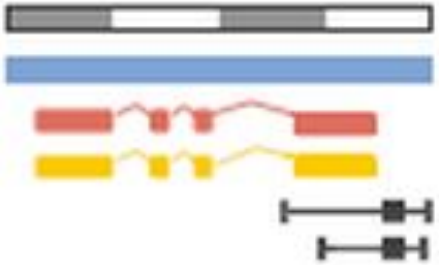
Ensembl Browser Workshop

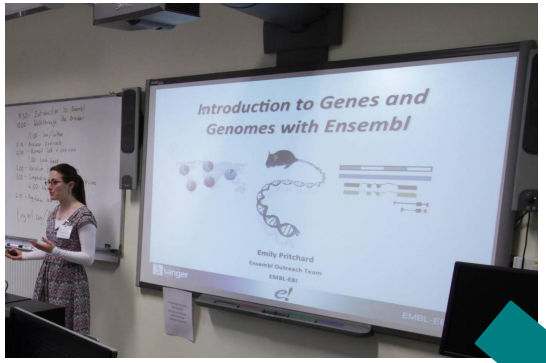


Aleena Mushtaq

Ensembl Outreach

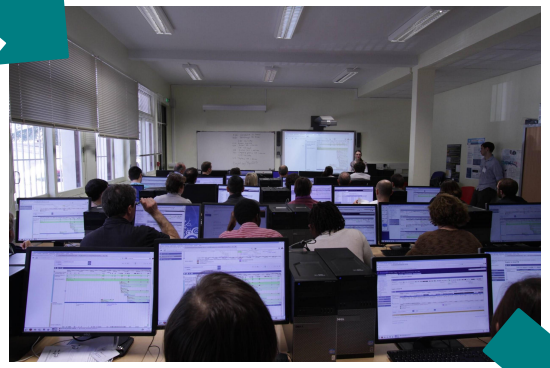
EMBL-EBI

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:
Trying things out for yourself (alone/pairs?)
Going beyond the demo
Not a test!
Pick and choose which ones best suit your use-case

Course material

http://training.ensembl.org/events/

- Presentations
- Demos
- Exercises (with text and video answers)

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.

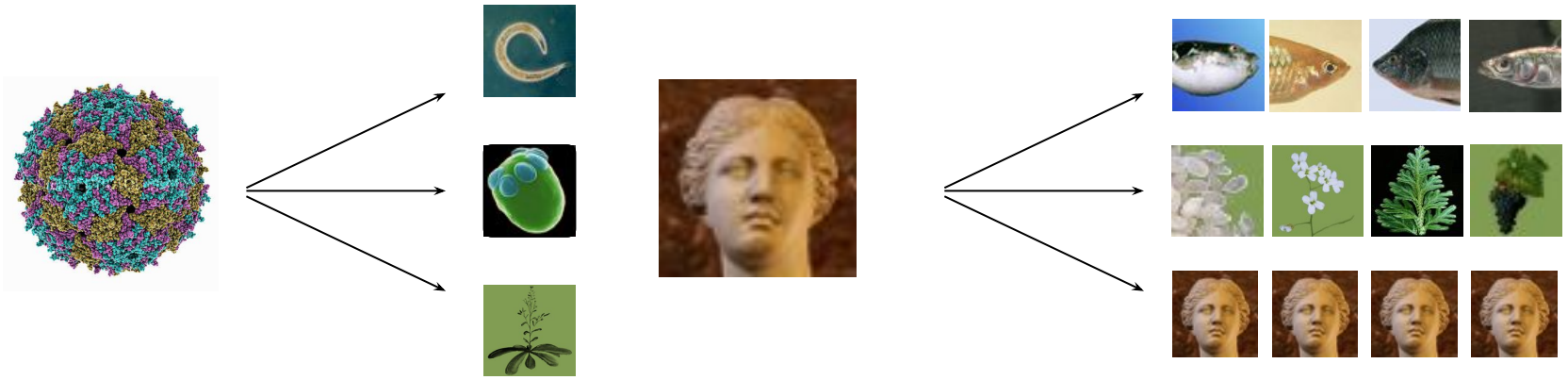
Exploring the *Ensembl genome browser*

Introduction

Why do we need genome browsers?

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

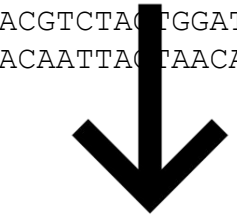
2004: finished human sequence (3 Gb)



CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATCTGAAATTTCTTGGAA
ACACGATCACTTTAACGGAATATTGCTGTTTTGGGGAAGTGTTTTACAGCTGCTGGGCACGCTGTATTTGCCTTACTTAAGC
CCCTGGTAATTGCTGTATTCCGAAGACATGCTGATGGGAATTACCAGGCGGCGTTGGTCTCTAACTGGAGCCCTCTGTCCCC
ACTAGCCACGCGTCACTGGTTAGCGTGATTGAAACTAAATCGTATGAAAATCCTCTTCTCTAGTCGCACTAGCCACGTTTCG
AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAAAATGTTCCCATCCTCACAGTAAGCTGTTACCGTTC
CAGGAGATGGGACTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCTCAGTCACATAATAAGGAATGCAT
CCCTGTGTAAGTGCATTTTGGTCTTCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAAAAATGCCTA
TTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTTTTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAAC
TTTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAAACCCAGTACGTCACAGTGTTGCTTAGAACCATAA
ACTGTTCCTTATGTGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACCACATGATAAATATAGAACGTCT
AGTGGATAAAGAGGAAACTGGCCCCTTGACTAGCAGTAGGAACAATTACTAACAAATCAGAAGCATTAATGTTACTTTATGG
CAGAAGTTGTCCAACTTTTTGGTTTCAGTACTCCTTATACTCTTAAAAATGATCTAGGACCCCCGGAGTGCTTTTGTTTATG
TAGCTTACCATATTAGAAATTTAAAACTAAGAATTTAAGGCTGGGCGTGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGA
GGCCGAGGTGGGCGGATCACTTGAGGCCAGAAGTTTGAGACCAGCCTGGCCAACATGGTGAAACCCTATCTCTACTAAAAAT
ACAAAAAATGTGCTGCGTGTGGTGGTGCGTGCCTGTAATCCCAGCTACACGGGAGGTGGAGGCAGGAGAATCGCTTGAACCC
TGGAGGCAGAGGTTGCAGTGAGCCAAGATCATGCCACTGCACTCTAGCCTGGGCCACATAGCATGACTCTGTCTCAAAACAA
ACAAACAAACAAAAAACTAAGAATTTAAAGTTAATTTACTTAAAAATAATGAAAGCTAACCCATTGCATATTATCACAACAT
TCTTAGGAAAAATAACTTTTTGAAAACAAGTGAGTGGAATAGTTTTTACATTTTTGCAGTTCTCTTTAATGTCTGGCTAAAT
AGAGATAGCTGGATTCACTTATCTGTGTCTAATCTGTTATTTTGGTAGAAGTATGTGAAAAAAAATTAACCTCACGTTGAAA
AAAGGAATATTTTAATAGTTTTCAGTTACTTTTTGGTATTTTTCCTTGTACTTTGCATAGATTTTTCAAAGATCTAATAGAT
ATACCATAGGTCTTTCCCATGTCGCAACATCATGCAGTGATTATTTGGAAGATAGTGGTGTTCTGAATTATACAAAGTTTCC
AAATATTGATAAATTGCATTAAACTATTTTAAAAATCTCATTCATTAATACCACCATGGATGTCAGAAAAGTCTTTTAAGAT
TGGGTAGAAATGAGCCACTGGAAATTCTAATTTTCATTTGAAAGTTCACATTTTGTCATTGACAACAAACTGTTTTCCTTGC
AGCAACAAGATCACTTCATTGATTTGTGAGAAAATGTCTACCAAATTATTTAAGTTGAAATAACTTTGTCAGCTGTTCTTTC
AAGTAAAAATGACTTTTCATTGAAAAAATTGCTTGTTCAGATCACAGCTCAACATGAGTGCTTTTCTAGGCAGTATTGTACT
TCAGTATGCAGAAGTGCTTTATGTATGCTTCCTATTTTGTCAGAGATTATTAAAAGAAGTGCTAAAGCATTGAGCTTCGAAA
TTAATTTTTACTGCTTCATTAGGACATTCTTACATTAAACTGGCATTATTATTACTATTATTTTTAACAAGGACACTCAGTG
GTAAGGAATATAATGGCTACTAGTATTAGTTTGGTGCCACTGCCATAACTCATGCAAATGTGCCAGCAGTTTTACCCAGCAT
CATCTTTGCACTGTTGATACAAATGTCAACATCATGAAAAAGGGTTGAAAAAAGGAATATTTTAATAGTTTTCAGTTACTTT

What is Ensembl?

```
AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAA
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA
CTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCT
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCT
TCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGT
AAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTT
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAACT
TTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAA
ACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACTGTTCCTTATG
TGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACC
ACATGATAAATATAGAACGTCTAGTGGATAAAGAGGAAACTGGCCCC
TTGACTAGCAGTAGGAACAATTAGTAACAAATCAGAAGCATTAATGT
```



Ensembl annotates and maps genomic features from genome sequences

What is Ensembl?

AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAA
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA
CTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCT
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCT
TCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGT
AAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTT
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAACT
TTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAA
ACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACTGTTCCTTATG
TGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACC
ACATGATAAATATAGAACGTCTAC[illegible]GGATAAAGAGGAAACTGGCCCC
TTGACTAGCAGTAGGAACAATTAC[illegible]AACAAATCAGAAGCATTAATGT

Ensembl is an ‘added value resource’ bringing together information from a wide range of other databases in a single site

What is Ensembl?

AGTGCTTAATGTGGCTAGTGGCACCGGTTTGGACAGCACAGCTGTAA
AATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGA
CTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACCT
CAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCT
TCTGTTTTGCAGACTTATTTACCAAGCATTGGAGGAATATCGTAGGT
AAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTGAAATTT
TTAAGACACGCTGCAACAAAGCAGGTATTGACAAATTTTATATAACT
TTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAAAA
ACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACTGTTCCTTATG
TGTGTATAAATCCAGTTAACAACATAATCATCGTTTGCAGGTTAACC
ACATGATAAATATAGAACGTCTAC[illegible]GGATAAAGAGGAAACTGGCCCC
TTGACTAGCAGTAGGAACAATTAC[illegible]TAACAAATCAGAAGCATTAATGT

NIH U.S. National Library of Medicine NCBI

Genome Data Viewer

www.ncbi.nlm.nih.gov/genome/gdv/

e!Ensembl

www.ensembl.org

UCSC

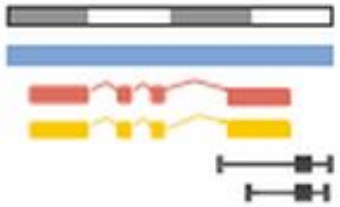
www.genome.ucsc.edu

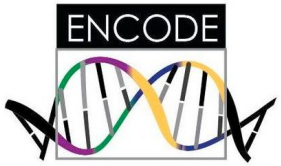
e!EnsemblGenomes

www.ensemblgenomes.org

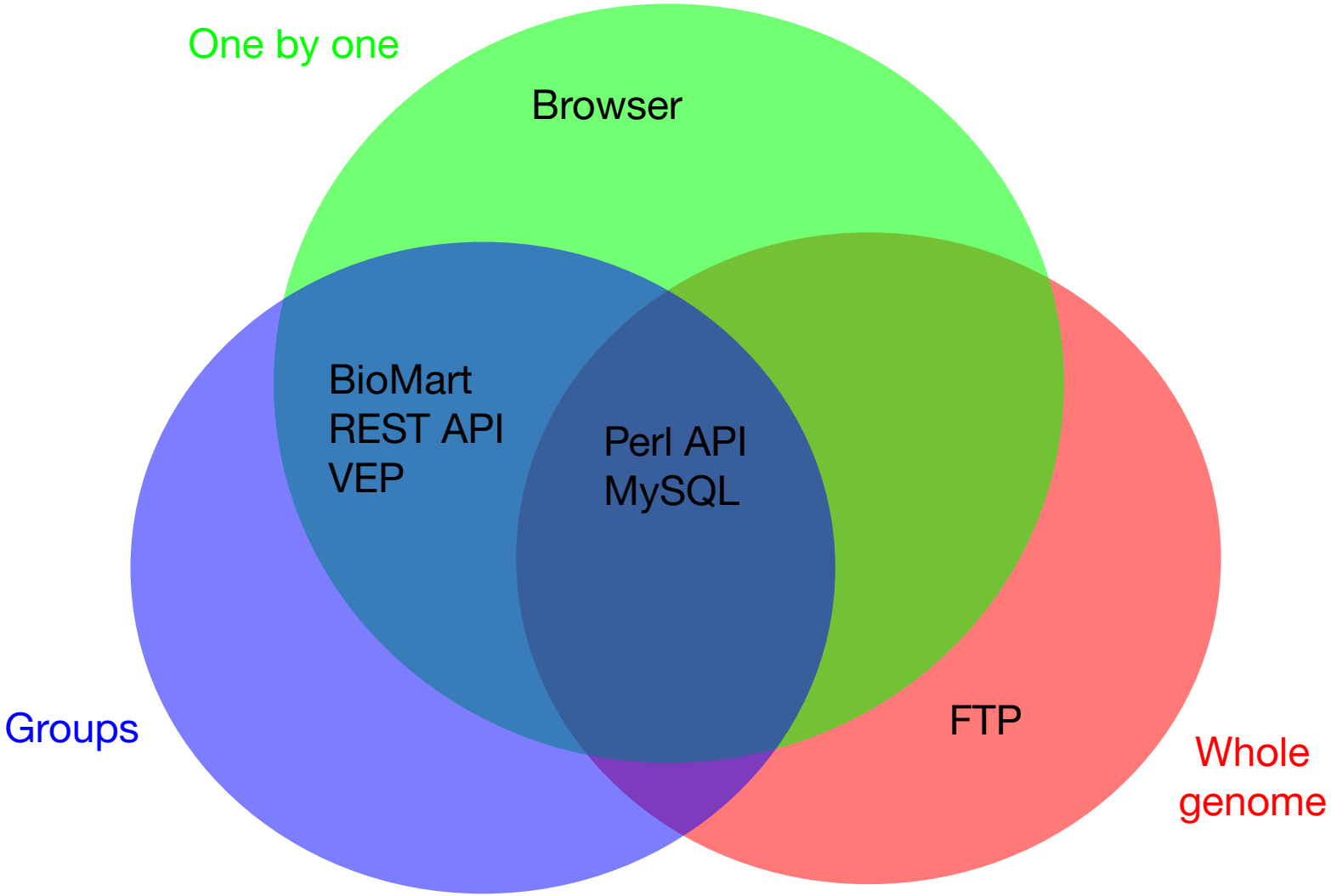
Ensembl features

- Genomes and gene builds for >300 species
- Variation data
- Compara (alignments, gene trees, homologues)
- Regulatory build
- BioMart (data export)
- Tools for data processing, e.g. VEP
- Display your own data
- Programmatic access via APIs
- Completely Open Source (FTP, GitHub)

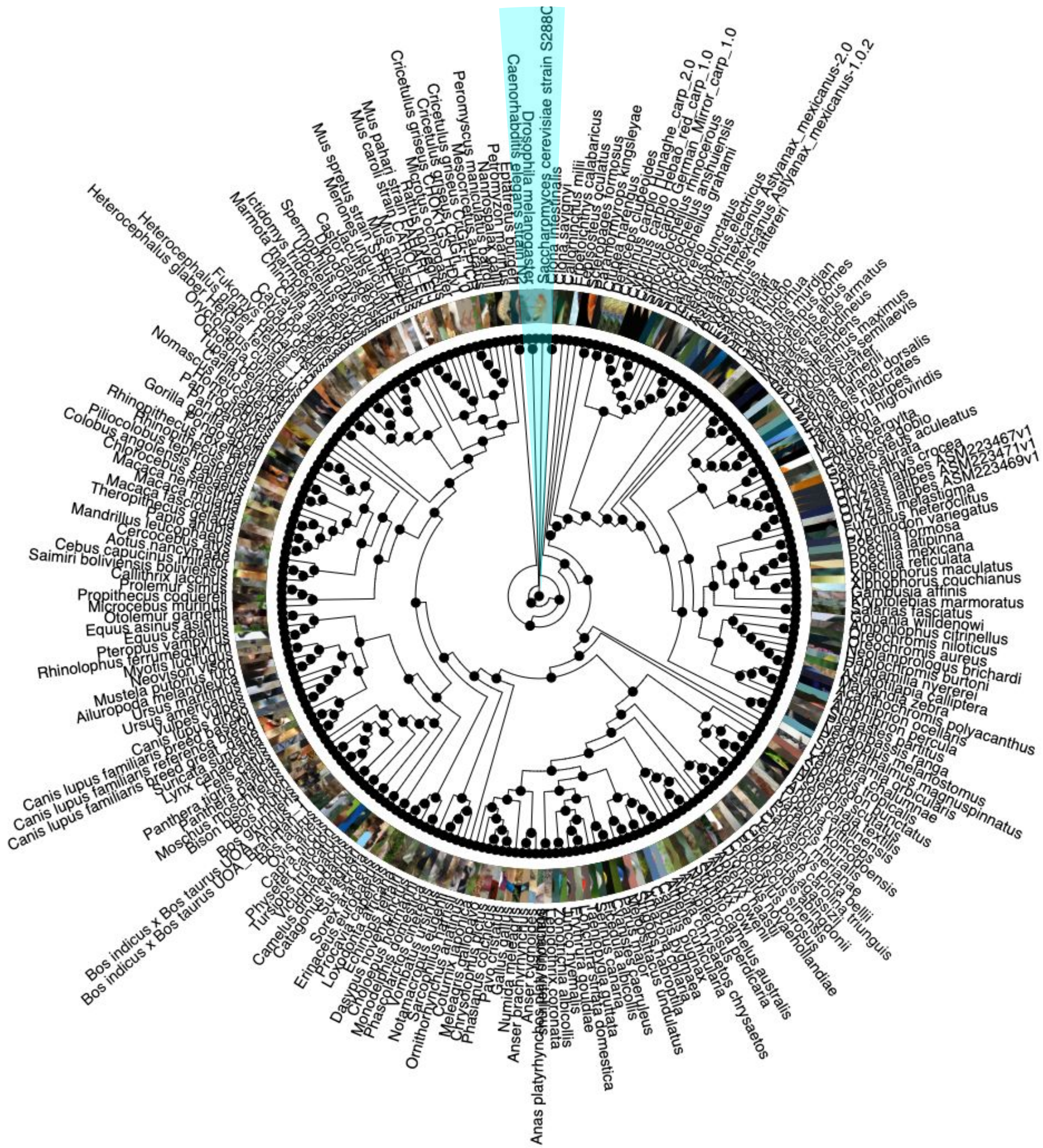




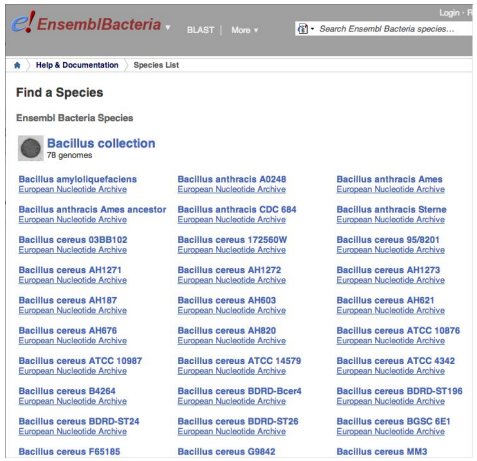
Access scales



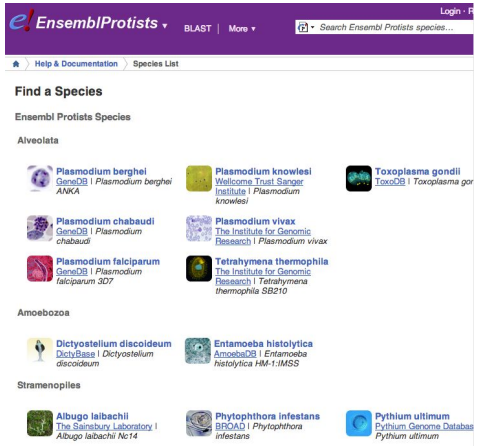
Vertebrate species in Ensembl



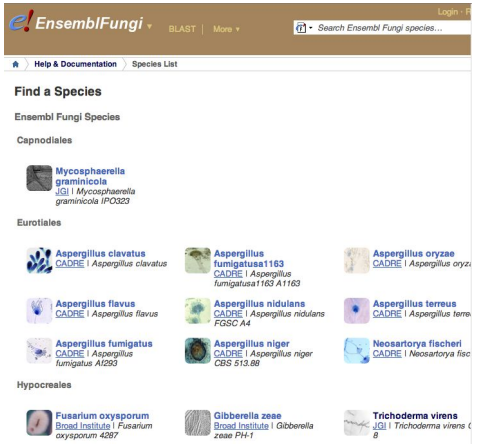
Non-vertebrates on Ensembl genomes



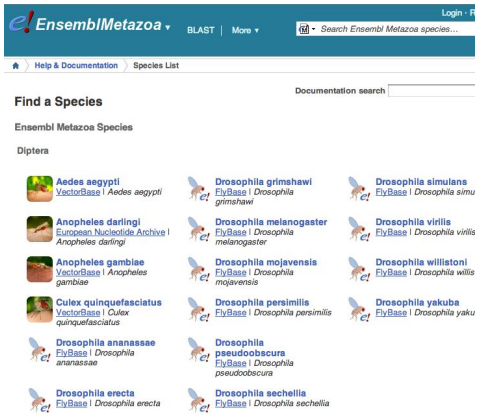
Bacteria



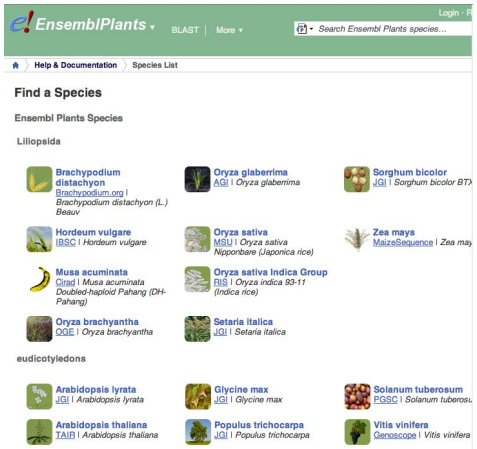
Protists



Fungi



Metazoa



Plants

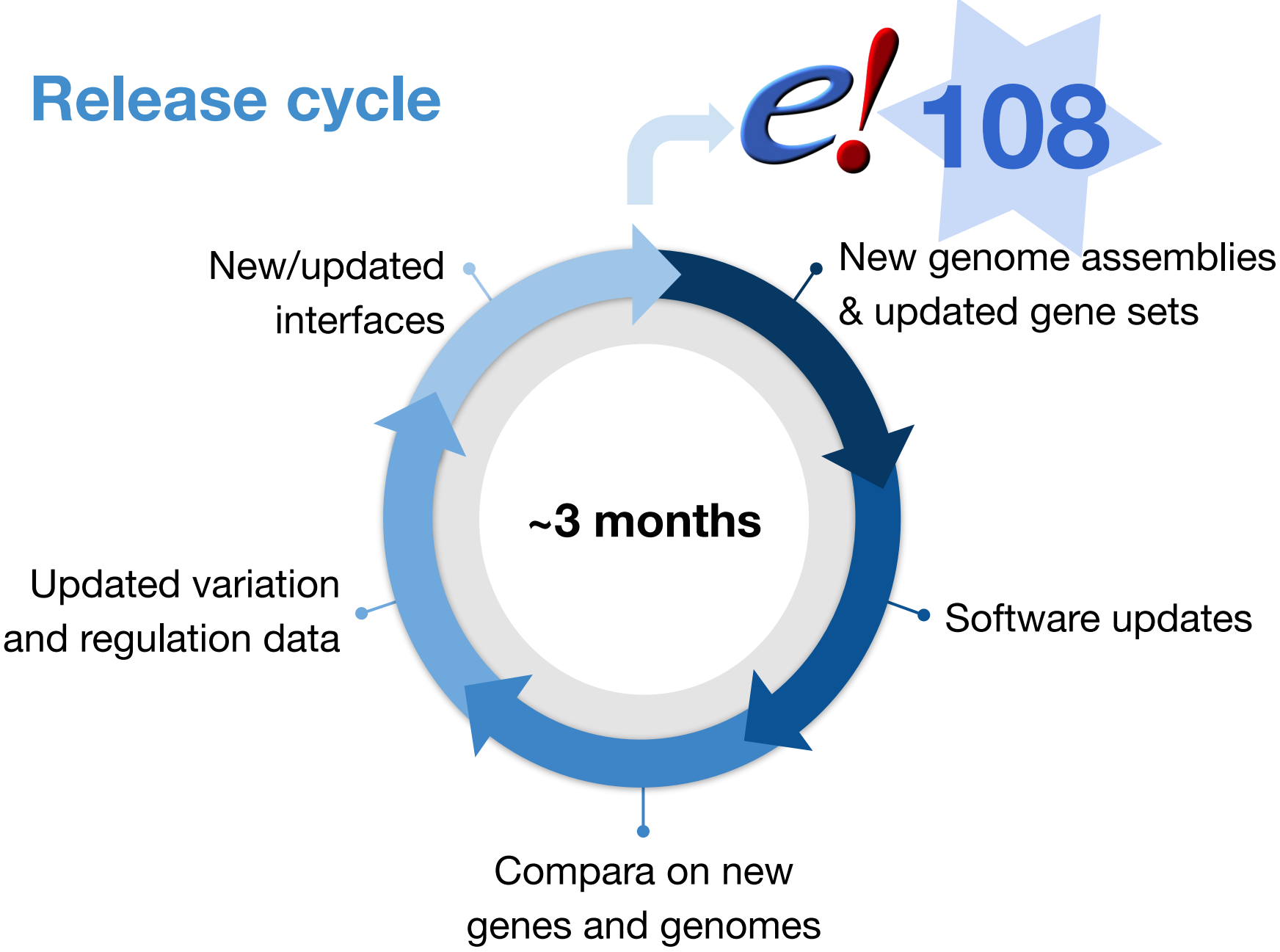
www.ensemblgenomes.org

EMBL-EBI

Ensembl and Ensembl Genomes

	Ensembl	Ensembl Genomes
Released	2000	2009
Species	Vertebrates (fly, worm and yeast as outgroups)	Non-vertebrates (protists, plants, fungi, metazoa, bacteria)
Annotation	by Ensembl	in collaboration with the scientific communities
URL	www.ensembl.org	www.ensemblgenomes.org

Release cycle



Ensembl Rapid Release

- Released every two weeks ✅
- Genome with gene annotation only ✅
- BLAST ✅
- Homology predictions ✅
- No BioMart ❌
- No variation ❌

Cloning into BACs

BL

Cut into chunks and clone into BACs

Grow up in bacterial cultures

Making a contig

Sequence reads

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

Match up overlaps

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

Contig

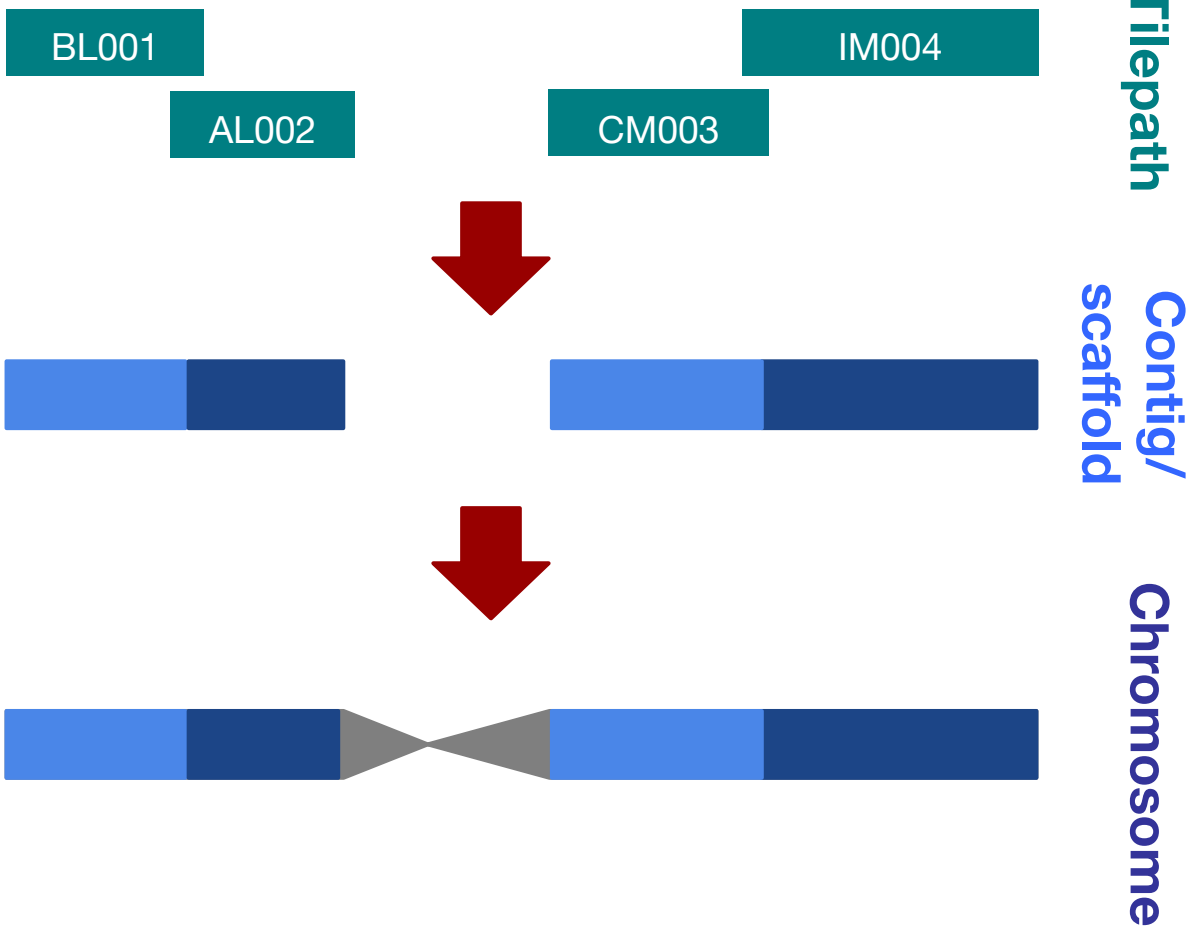
```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCATC
```

Contigs to scaffolds

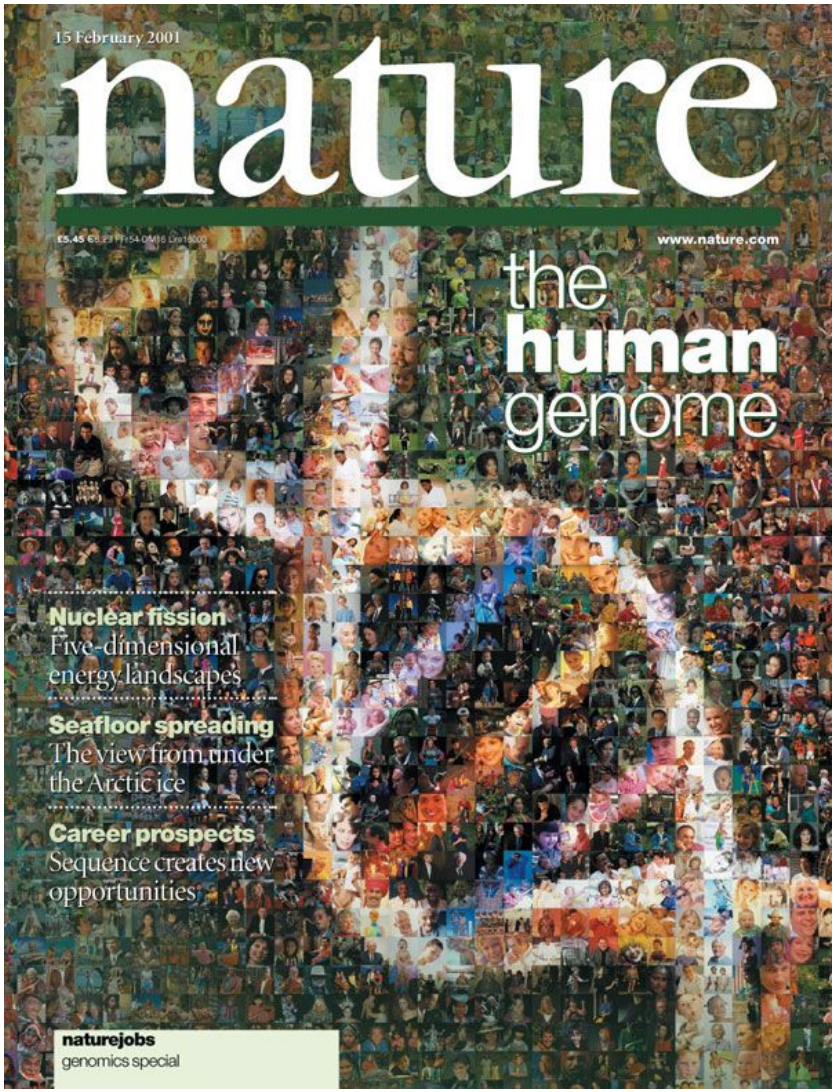
BACs from different individuals assembled together with overlaps Tilepath

Overlaps trimmed to give contigs. A run of contigs with no gaps is a scaffold.

Genetic maps are used to assemble scaffolds into a chromosome



Genome assemblies



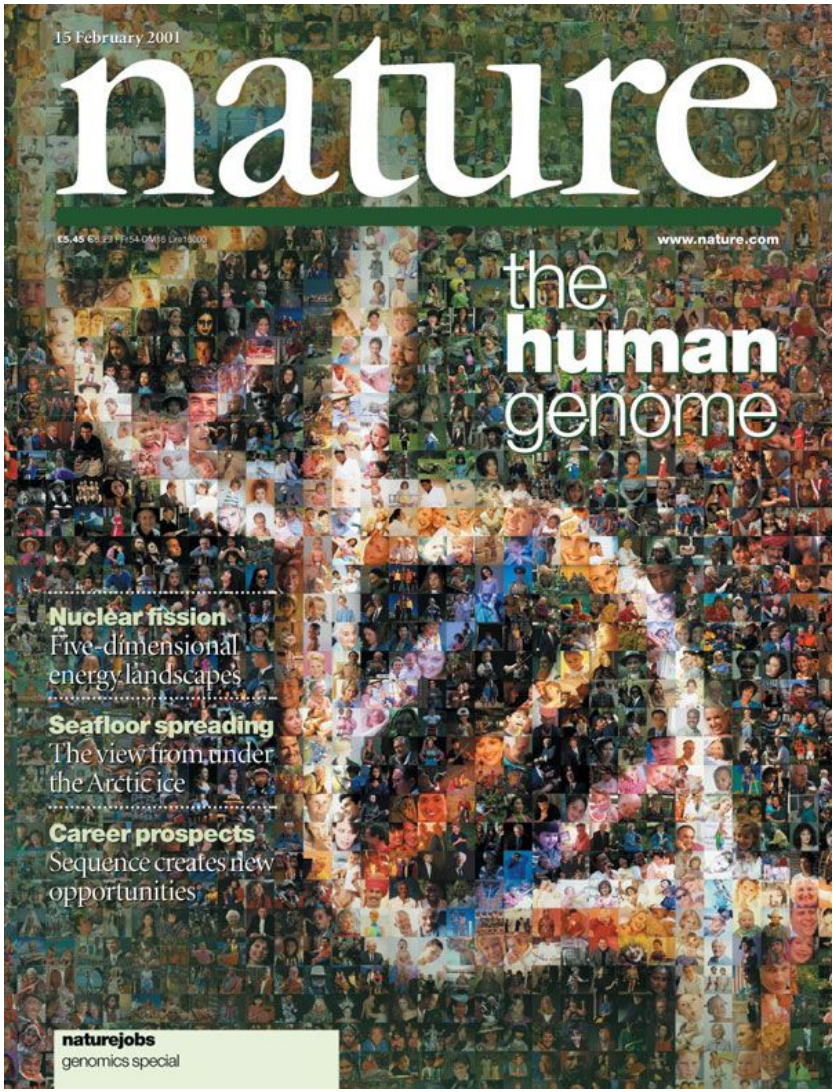
BL

AL

CM

IM

Genome contigs



BL

AL

CM

IM

BL001

AL002

CM003

IM004

Genome assemblies

Genome
"DNA within a cell"

Genome *assembly*
Representation of a genome
Contains errors and gaps
Coordinate system

Human genome assemblies

- GRCh38 (aka hg38) 

  - Many rare/private alleles replaced.
  - www.ensembl.org
  - Most up-to-date and supported
- GRCh37 (aka hg19) 

  - Some large gaps
  - grch37.ensembl.org
  - Limited data and software updates
  - Still the preferred genome of the clinical community
- NCBI36 (aka hg18) 

  - Many gaps
  - ncbi36.ensembl.org
  - No longer updated

Hands on

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

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.

Ensembl Browser Workshop

Aleena Mushtaq

12 December 2022

Ensembl Outreach

EMBL-EBI

Reference Variation Datasets

Variation types

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

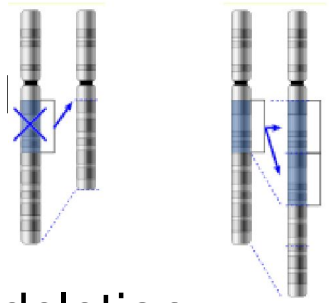
- Small insertions and deletions (DIPs or indels)
- Single nucleotide polymorphism (SNP)

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

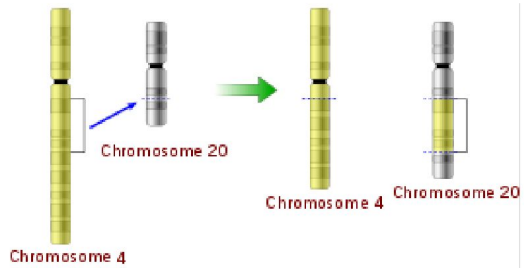
2) Large scale in chromosomal structure (structural variation)

- Copy number variations (CNV)

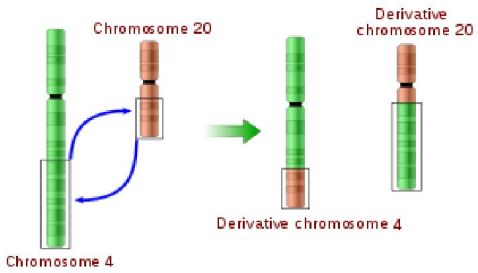


deletion

duplication



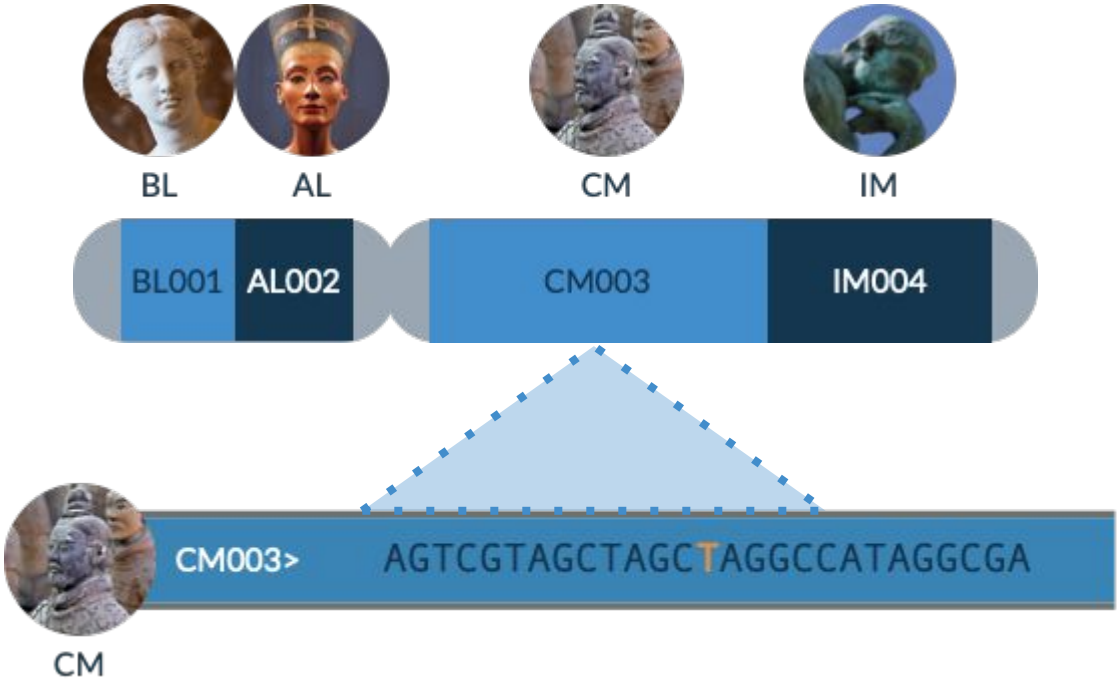
insertion



translocation

http://training.ensembl.org/events

Reference alleles



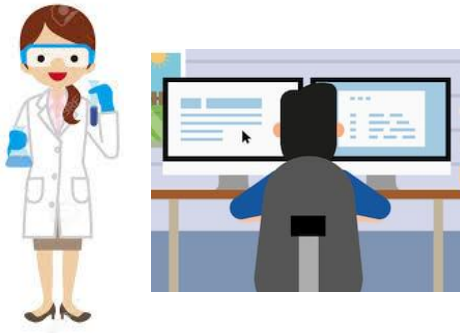
T is the allele in the sequenced contig, therefore:
T is the reference allele
G is the alternative allele
Alleles are **T**/**G**

Reference Variation Datasets

- Provide list of known variants across populations
  - Use for filtering in variant prioritisation (low MAF or novel)
  - Use to identify known genotype-phenotype links

Reference Variation Datasets

Data Generation → **Archives** → **Annotation in genome browsers**



- **Genotyping arrays**
- **Whole genome sequencing (WGS)**
- **Whole exome sequencing (WES)**

dbSNP
Short Genetic Variations

DGVarchive

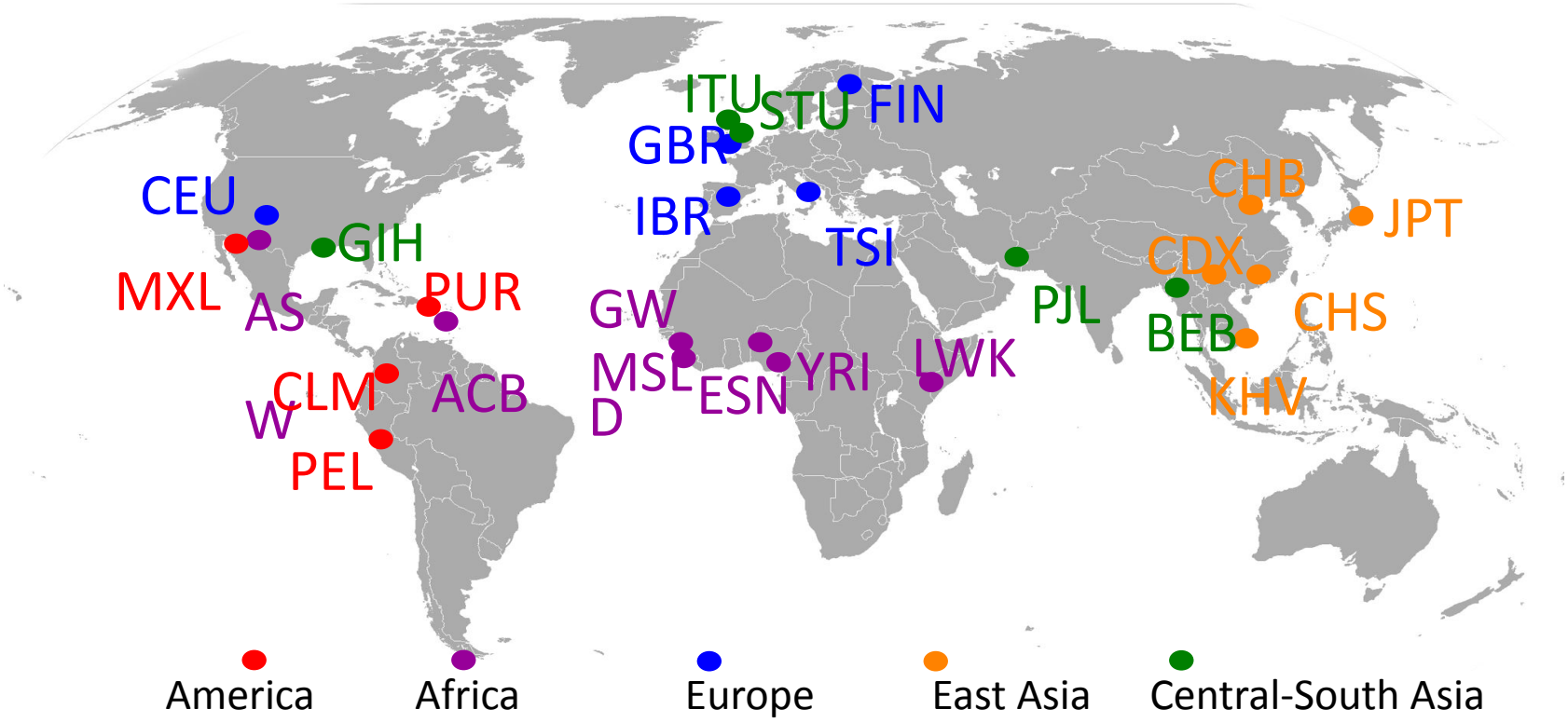
EVA

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://training.ensembl.org/events

Allele Frequencies- the 1000 Genomes Project



The 1000 Genomes Phase 3 data provides genotype data for 2504 individuals from 26 different narrowly defined populations in 5 groupings.

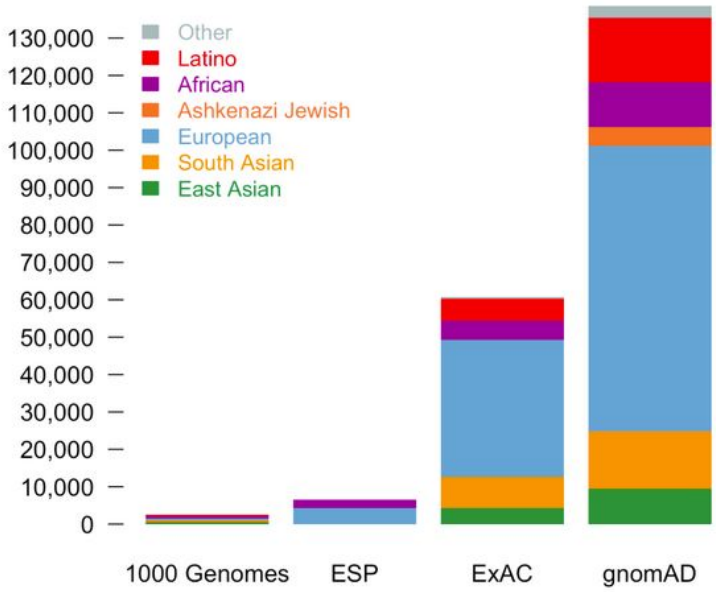
http://training.ensembl.org/events

Allele Frequencies- GnomAD

The Genome Aggregation Database provides allele frequency data for over 130,000 samples from 8 different populations

POPULATION	DESCRIPTION	GENOMES	EXOMES	TOTAL
AFR	African/African American	4,368	7,652	12,020
AMR	Admixed American	419	16,791	17,210
ASJ	Ashkenazi Jewish	151	4,925	5,076
EAS	East Asian	811	8,624	9,435
FIN	Finnish	1,747	11,150	12,897
NFE	Non-Finnish European	7,509	55,860	63,369
SAS	South Asian	0	15,391	15,391
OTH	Other (population not assigned)	491	2,743	3,234
	Total	15,496	123,136	138,632

Sample numbers



From: https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/

Human Population Genetics

Allele Frequencies

- **HapMap:** Worldwide, genotyping
- **1000 Genomes:** Worldwide, WGS, healthy individuals
- **ExAC:** Exomes, diseased/healthy individuals, skewed to most studied populations
- **gnomAD:** WGS and exomes, diseased/healthy individuals, skewed to most studied populations
- **UK10K:** UK-wide, exomes, diseased individuals
- **TOPMed:** WGS, diseased individuals

Phenotype and Disease Data

ClinVar

European Genome-phenome Archive

HGMD®

OMIM®

DGVarchive

DECIPHER GRCh37

orphanet

dbGaP GENOTYPES and PHENOTYPES

europhenome

ZFIN

IMPC INTERNATIONAL MOUSE PHENOTYPING CONSORTIUM

RGD

LOVD Leiden Open Variation Database

AnimalQTLdb

NHGRI-EBI GWAS Catalog

OMIA - ONLINE MENDELIAN INHERITANCE IN ANIMALS

Variant

- clinical significance
- inheritance type
- reported genes /variants
- risk allele

Gene

Structural Variant

- p-value
- odds ratio
- beta coefficient

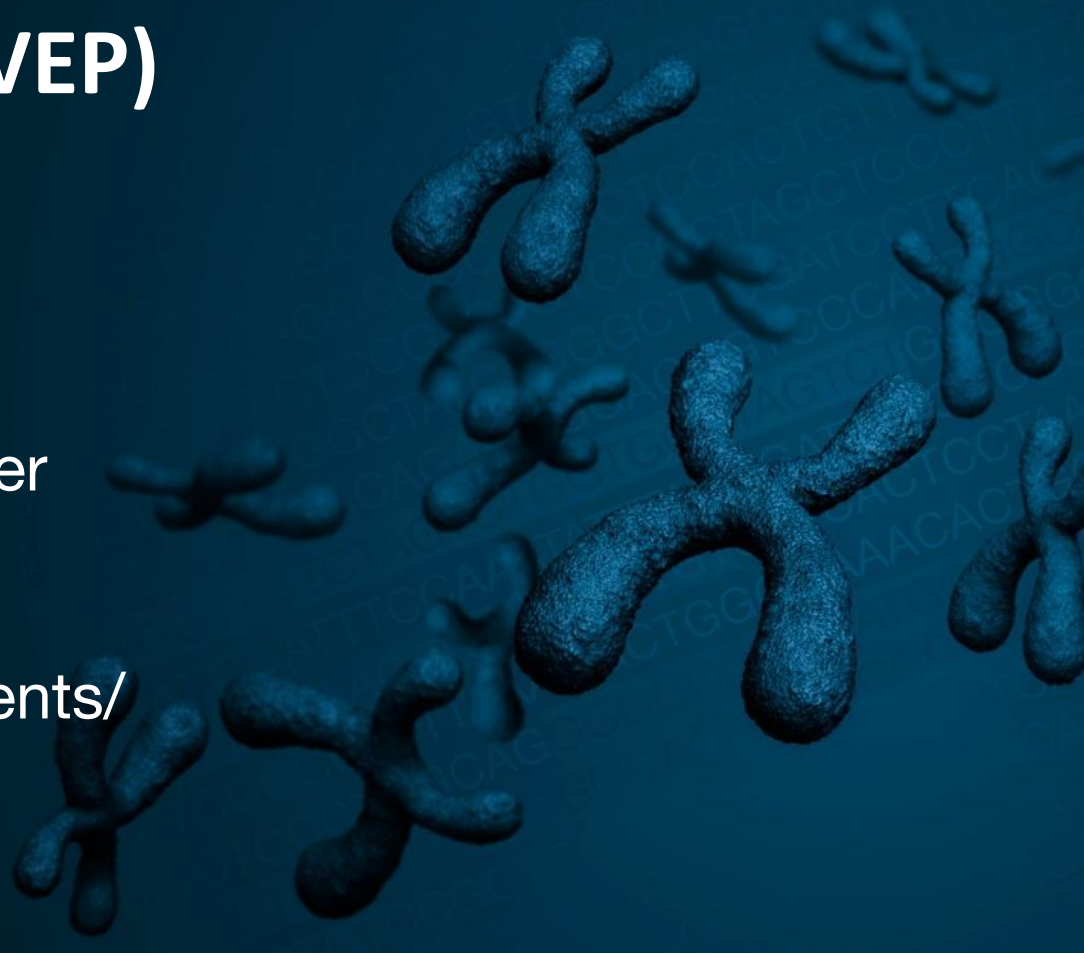
QTL

http://training.ensembl.org/events

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.

Variant Annotation: Variant Effect Predictor (VEP)



Aleena Mushtaq
Ensembl Outreach Officer

Slides available from
training.ensembl.org/events/

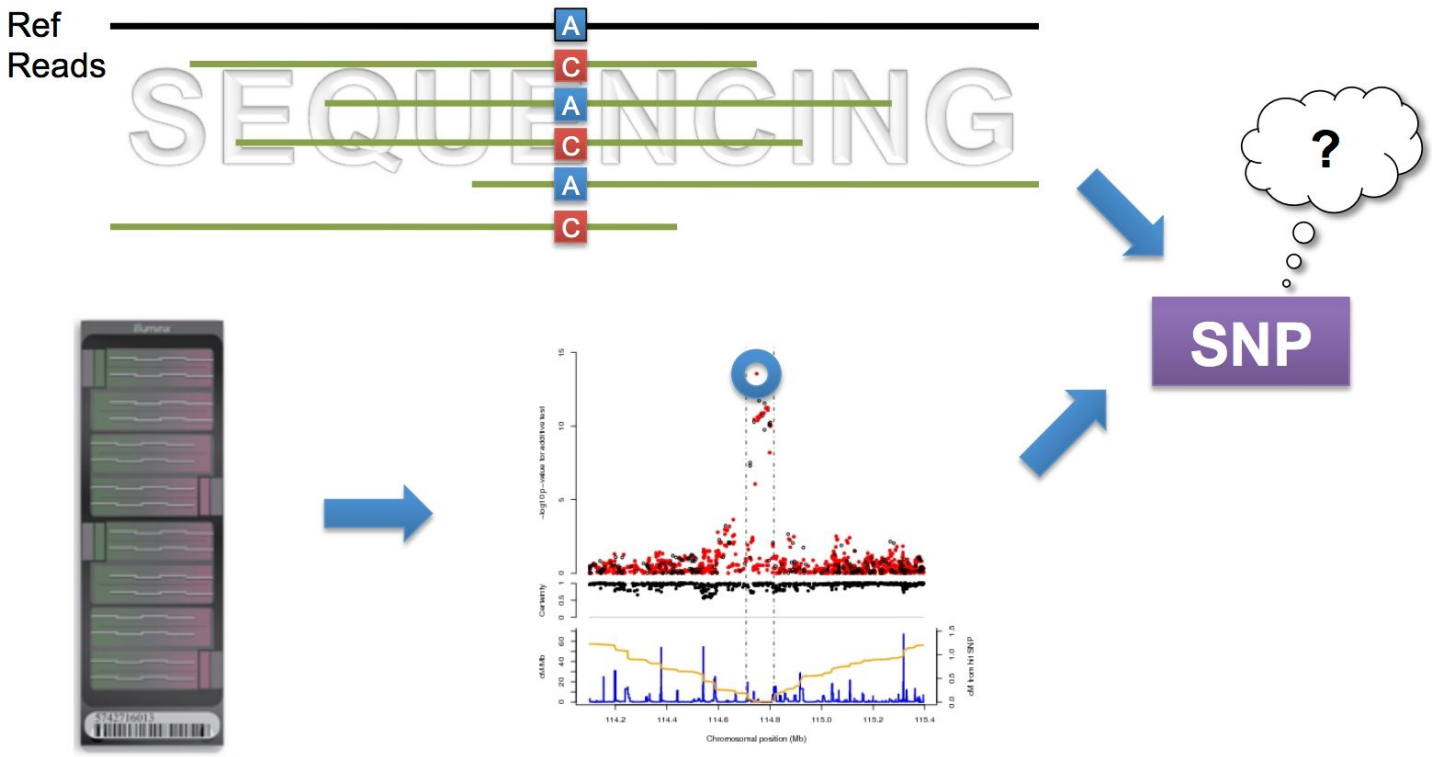
EMBL-EBI

Ensembl

What is the VEP for?

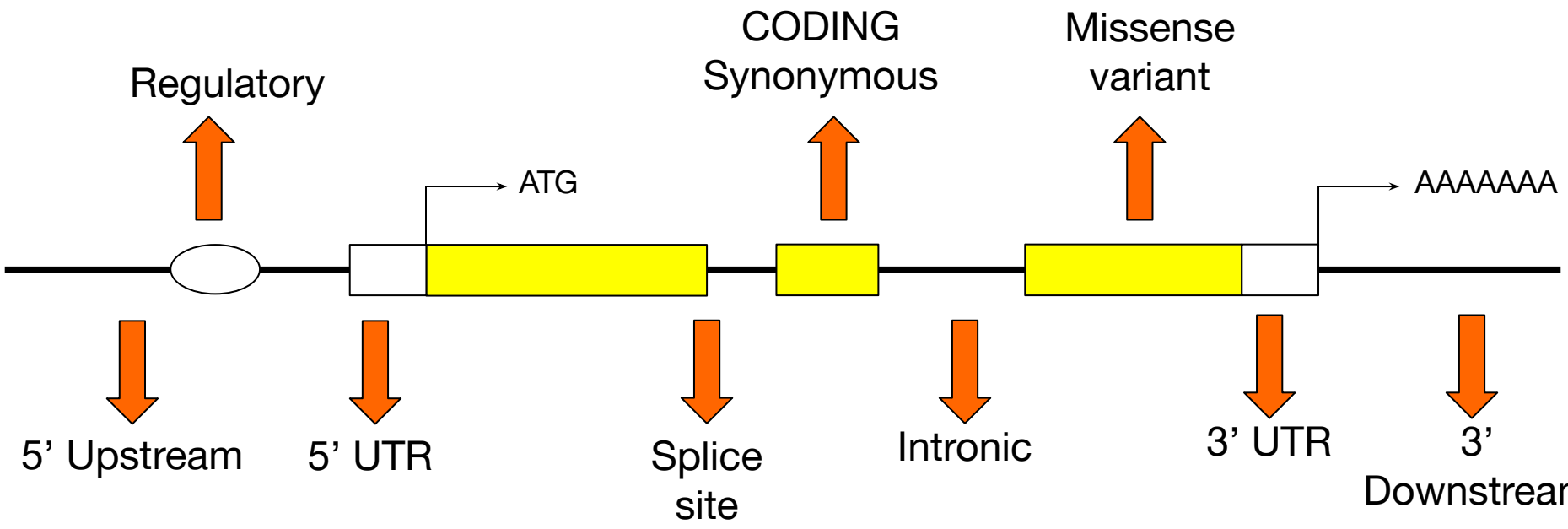
A tool to predict and annotate the **functional consequences of variants**

(SNPs, insertions, deletions or structural variants)



http://training.ensembl.org/events

Sequence Ontology: Variant consequences



Consequence terms

*	SO term	SO description	SO accession	Display term	IMPACT
	transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
	splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
	splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
	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	HIGH
	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 variant	HIGH
	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	HIGH
	start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
	transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
	inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
	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	MODERATE
	protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
	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	LOW
	splice_donor_5th_base_variant	A sequence variant that causes a change at the 5th base pair after the start of the intron in the orientation of the transcript	SO:0001787	Splice donor 5th base variant	LOW
	splice_donor_region_variant	A sequence variant that falls in the region between the 3rd and 6th base after splice junction (5' end of intron)	SO:0002170	Splice donor region variant	LOW
	splice_polypyrimidine_tract_variant	A sequence variant that falls in the polypyrimidine tract at 3' end of intron between 17 and 3 bases from the end (acceptor -3 to acceptor -17)	SO:0002169	Splice polypyrimidine tract variant	LOW
	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	LOW
	start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
	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	LOW
	synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW
	coding_sequence_variant	A sequence variant that changes the coding sequence	SO:0001580	Coding sequence variant	MODIFIER
	mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	Mature miRNA variant	MODIFIER
	5_prime_UTR_variant	A UTR variant of the 5' UTR	SO:0001623	5 prime UTR variant	MODIFIER
	3_prime_UTR_variant	A UTR variant of the 3' UTR	SO:0001624	3 prime UTR variant	MODIFIER
	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	MODIFIER
	intron_variant	A transcript variant occurring within an intron	SO:0001627	Intron variant	MODIFIER
	NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO:0001621	NMD transcript variant	MODIFIER
	non_coding_transcript_variant	A transcript variant of a non coding RNA gene	SO:0001619	Non coding transcript variant	MODIFIER
	upstream_gene_variant	A sequence variant located 5' of a gene	SO:0001631	Upstream gene variant	MODIFIER
	downstream_gene_variant	A sequence variant located 3' of a gene	SO:0001632	Downstream gene variant	MODIFIER
	TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	SO:0001895	TFBS ablation	MODIFIER
	TFBS_amplification	A feature amplification of a region containing a transcription factor binding site	SO:0001892	TFBS amplification	MODIFIER
	TF_binding_site_variant	A sequence variant located within a transcription factor binding site	SO:0001782	TF binding site variant	MODIFIER
	regulatory_region_ablation	A feature ablation whereby the deleted region includes a regulatory region	SO:0001894	Regulatory region ablation	MODERATE
	regulatory_region_amplification	A feature amplification of a region containing a regulatory region	SO:0001891	Regulatory region amplification	MODIFIER
	feature_elongation	A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence	SO:0001907	Feature elongation	MODIFIER
	regulatory_region_variant	A sequence variant located within a regulatory region	SO:0001566	Regulatory region variant	MODIFIER
	feature_truncation	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence	SO:0001906	Feature truncation	MODIFIER
	intergenic_variant	A sequence variant located in the intergenic region, between genes	SO:0001628	Intergenic variant	MODIFIER

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

http://training.ensembl.org/events

Pathogenicity predictions

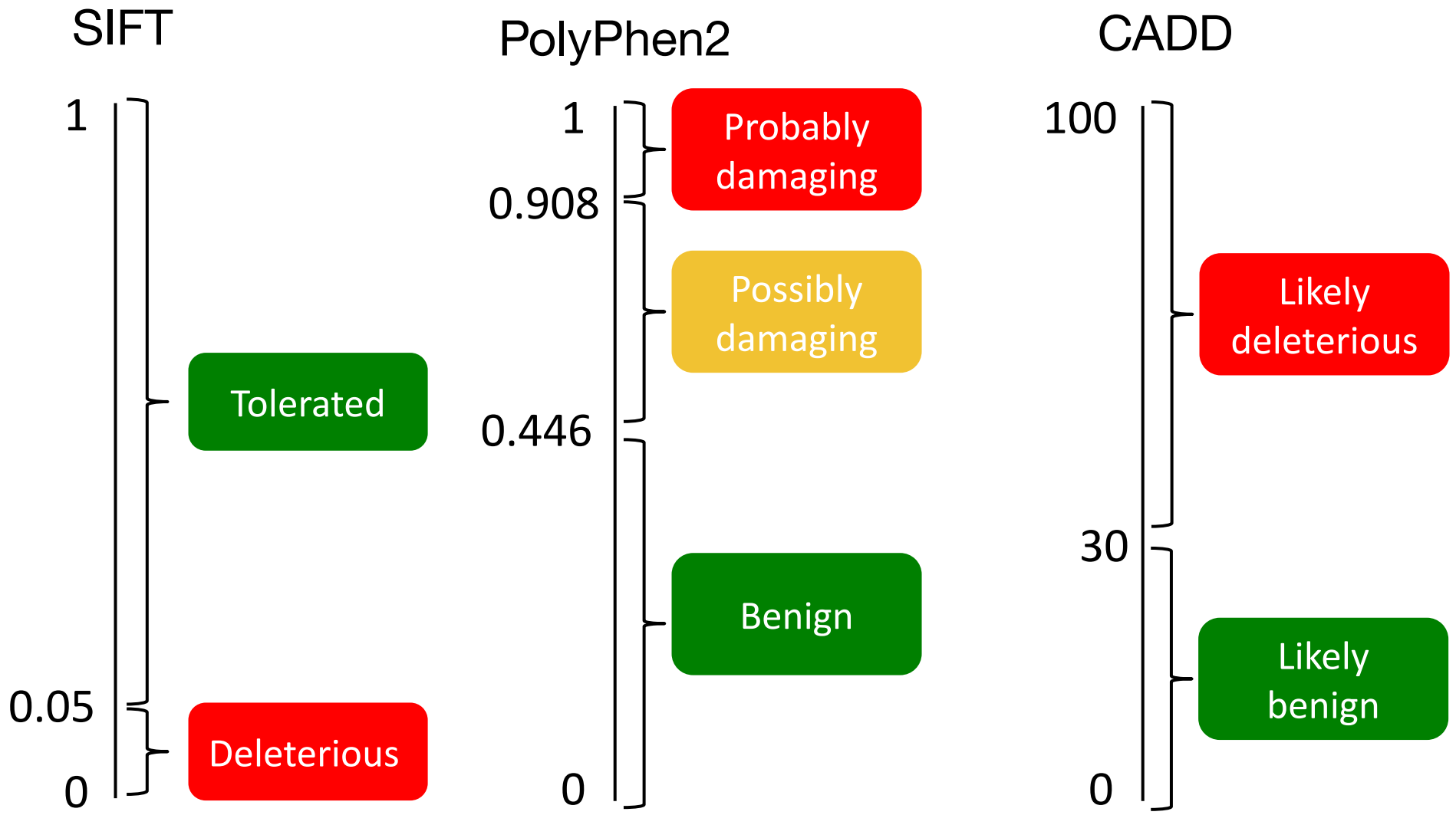
Various algorithms score or rank changes in amino acid sequence based on:

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

These are predictions, not facts

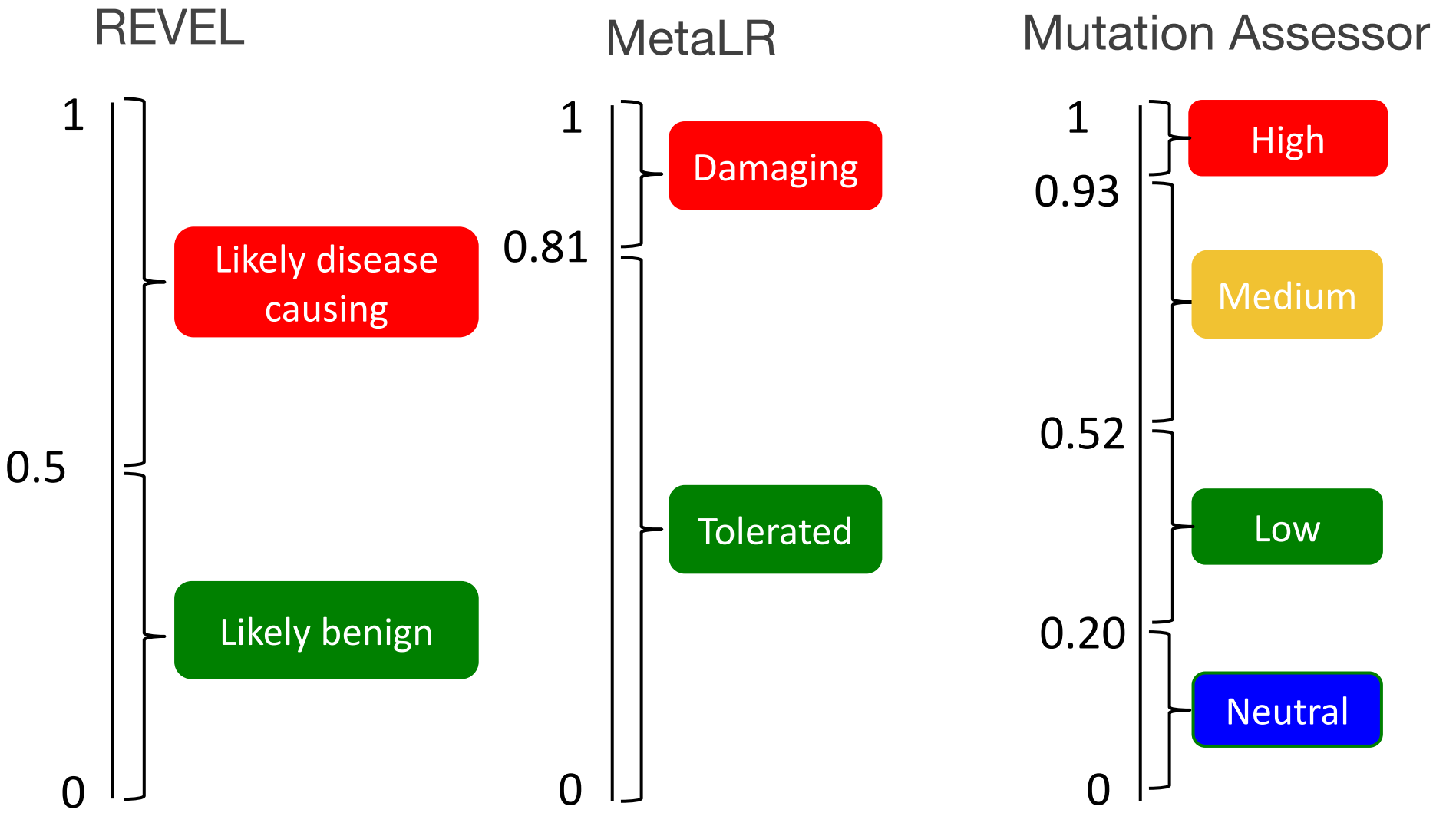
A prediction will never be as good as experimental validation

Missense variants: Pathogenicity



http://training.ensembl.org/events

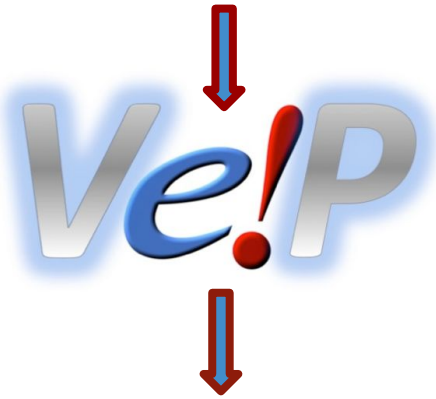
Missense variants- pathogenicity



What can you do with the VEP?

Data input

- Variant Coordinates
- VCF
- HGVS
- Variant IDs
- SPDI

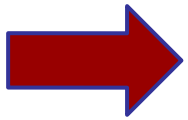


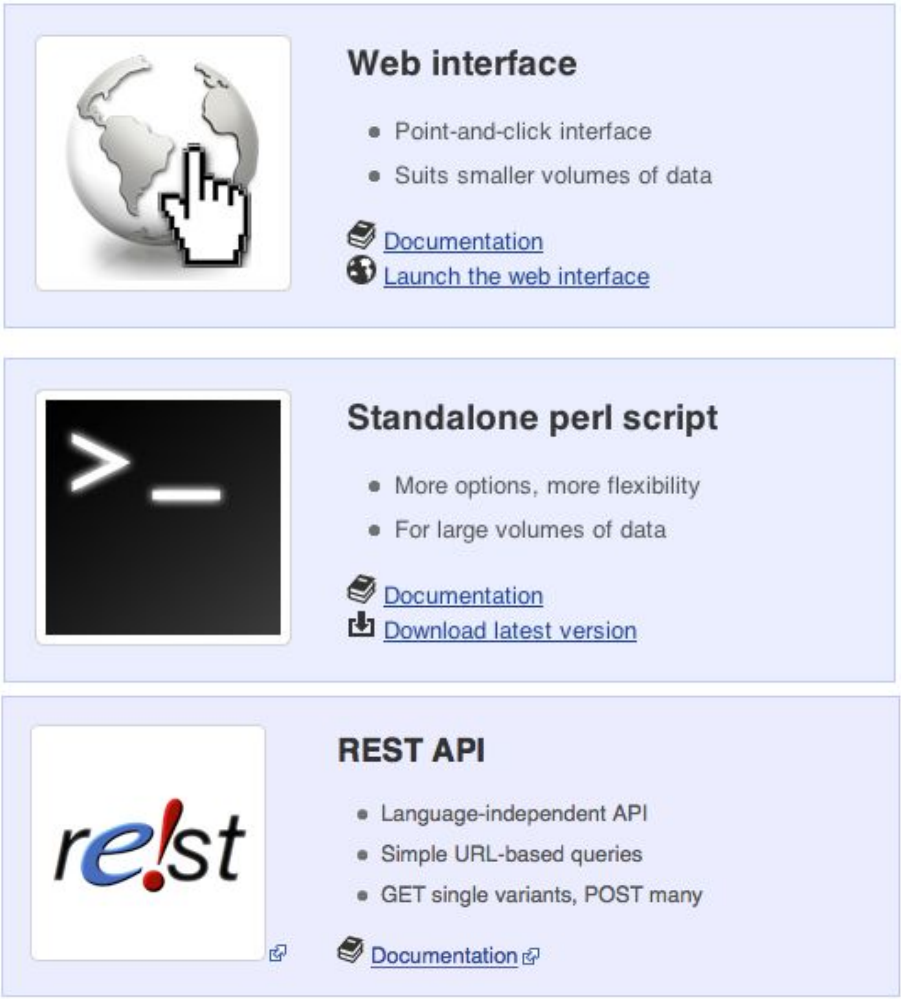
VEP Output

- Genes, transcripts affected
- Pathogenicity
- Frequency data
- Regulatory and splicing consequences
- Pubmed citations

http://training.ensembl.org/events

Use the VEP





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

VEP Features

Transcript consequence predictions:	Overlap with regulatory regions:	Known variants:	Allele frequencies:	Pathogenicity predictions:	Phenotype, disease, clinical significance:
Ensembl / GENCODE	ENCODE	dbSNP	1000 Genomes	SIFT	OMIM
GENCODE basic	BLUEPRINT	COSMIC	ESP	PolyPhen	Orphanet
RefSeq	NIH Epigenomics Roadmap	ClinVar	gnomAD	dbNSFP	GWAS catalog
Merged	Can be limited to specific cell types	ESP		CADD	ClinVar
		HGMD Public		Condel	
		PhenCode		LoFTool	
				FATHMM	
				MutationTaster	
				Many more!	

Hands on

We have identified five variants on human chromosome nine.

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

➔ Demo and exercises on https://training.ensembl.org/events/2022/2022-10-13-EBI_structural_bioinformatics_course_VEP_October_108

Wrap-up

Ensembl is a genome browser which integrates:

- Gene annotation
- Variation
- Comparative genomics
- Regulation

How is all this data organised?

- Ensembl browser sites

Main website, GRCh37, Ensembl Genomes, Archive!

- BioMart ‘DataMining tool’
- Ensembl Database (open source)

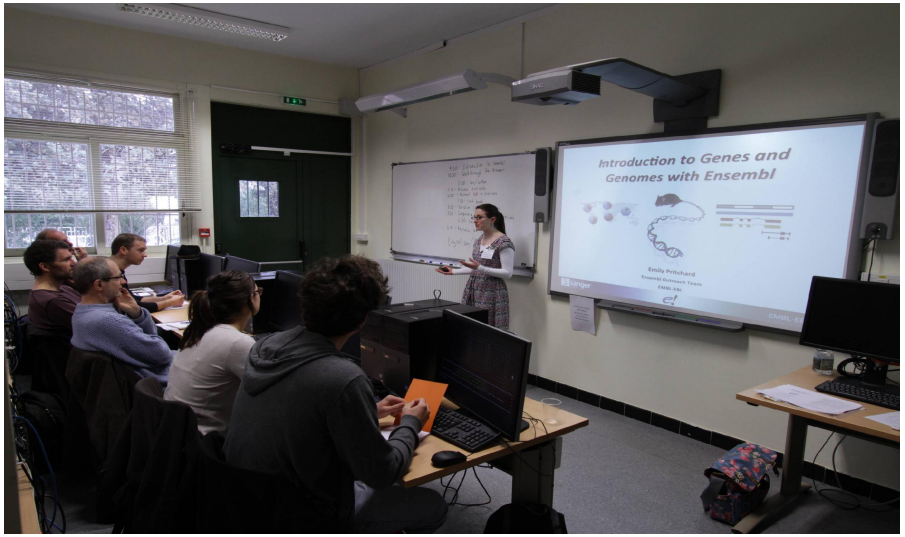
Perl-API, REST API, MySQL

- FTP download site
  http://www.ensembl.org/info/data/ftp/index.html

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

Help and documentation

Courses online
http://www.ebi.ac.uk/training/online/subjects/11

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

Flash animations

www.youtube.com/user/EnsemblHelpdesk

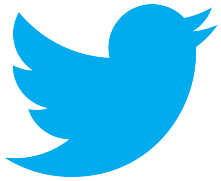
http://u.youku.com/Ensemblhelpdesk

Email us helpdesk@ensembl.org

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

Follow us!

www.facebook.com/Ensembl.org



@Ensembl

www.ensembl.info

Publications

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

Ensembl 2022
Cunningham *et al*
https://europepmc.org/article/med/34791404

Topic-specific publications mentioned throughout workshop

Ensembl 2019

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, Benjamin Moore, Bethany Flint, Brandon Walts, Bruno Contreras-Moreira, Carla Cummins, Carlos Garcia Giron, Claire Davidson, Cristina Guijarro, Dan Sheppard, Daniel Zerbino, David Alejandro Urbina Gomez, David Thybert, Denye Ogeh, Diana Lemos, Dimitrios Paraschas, Elizabeth Lewis, Emily Perry, Fergal Martin, Fiona Cunningham, Francesca Tricomi, Gareth Maslen, Gareth Williams, Garth Ilsley, Guy Naamati, Helen Schuilenburg, IF Barnes, Ilias Lavidas, Irina Armean, James Allen, Jamie Allen, Jane Loveland, John Tate, 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, Magdalena Zarowiecki, Manoj Sakthivel, Manuel Carbajo Martinez, Marc Chakiachvili, Mark Quinton-Tulloch, Marie-Marthe Suner, Matthew Hardy, Matthew Russell, Matthieu Barba, Matthieu Muffato, Mehrnaz Charkhchi, Michael Paulini, Michal Szpak, Mike Kay, Mikkel Christensen, Mira Sycheva, Nishadi De Silva, Osagie Izuogu, Paul Davis, Paul Flicek, Reham Fatima, Ridwan Amode, Ruth Bennett, Sanjay Boddu, Sarah Donaldson, Sarah Hunt, Shamika Mohanan, Stephen Trevanion, Thiago Genez, Thibaut Hourlier, Thomas Juettemann, Toby Hunt, Tuan Le, Vasili Sitnik, Vinay Kaikala, Yasmin Fathy, Zoe Hollis

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