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

EMBL-EBI

Ben Moore

Ensembl Outreach

EMBL-EBI

UNC ClinGen Biocurators - 12th October 2017

Structure for webinar 1

Introduction to Ensembl

Exploring Ensembl

Genes and transcripts

Structure for webinar 2

Variation and phenotype data

Questions?

Course materials

training.ensembl.org

- Presentations
- Coursebook (demos)

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

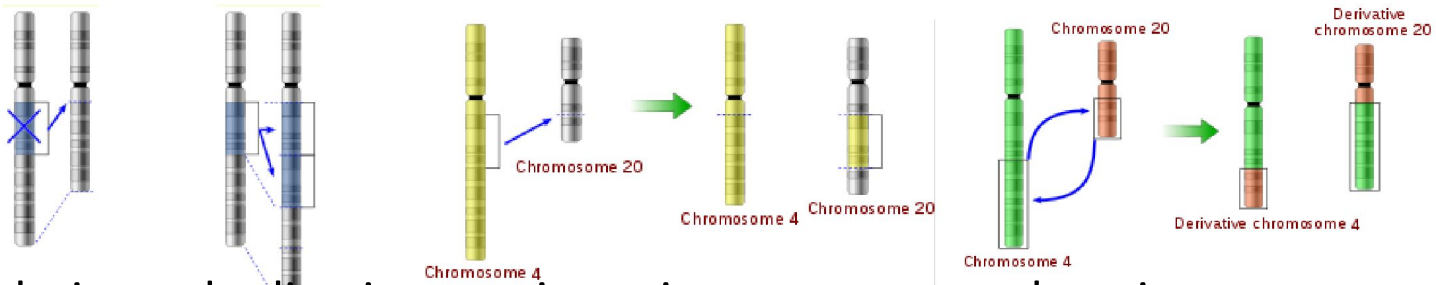
- Small insertions and deletions (DIPs or indels)
- Single nucleotide polymorphism (SNP)

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)

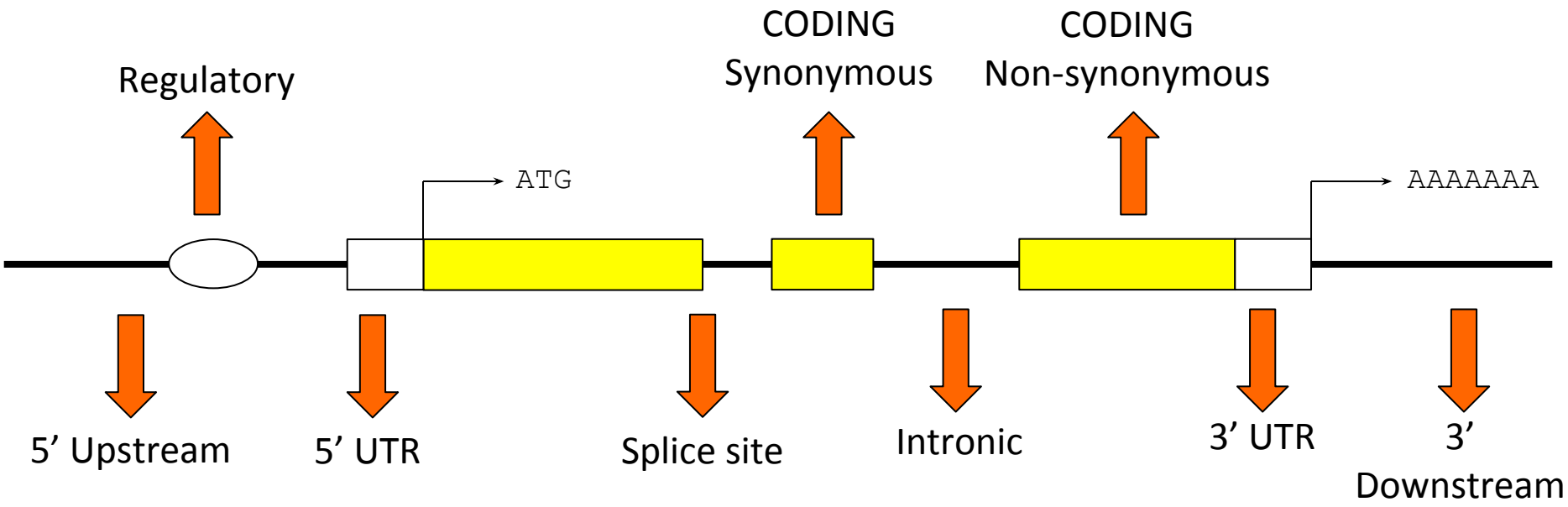
- Copy number variations (CNV)
- Large deletions/duplications, insertions, translocations



deletion duplication insertion translocation

http://training.ensembl.org/events

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

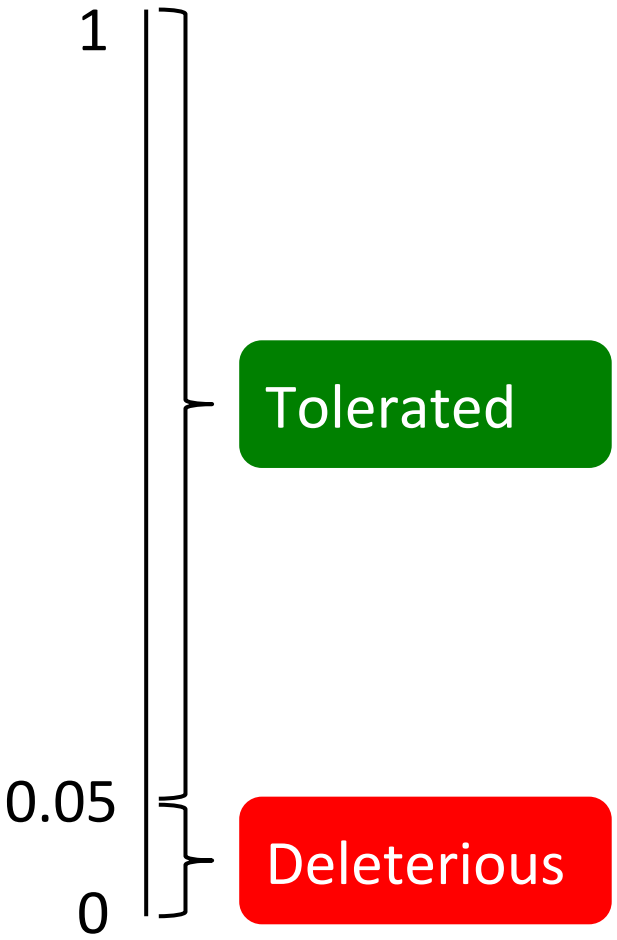
Missense variants- pathogenicity

SIFT and PolyPhen score changes in amino acid sequence based on:

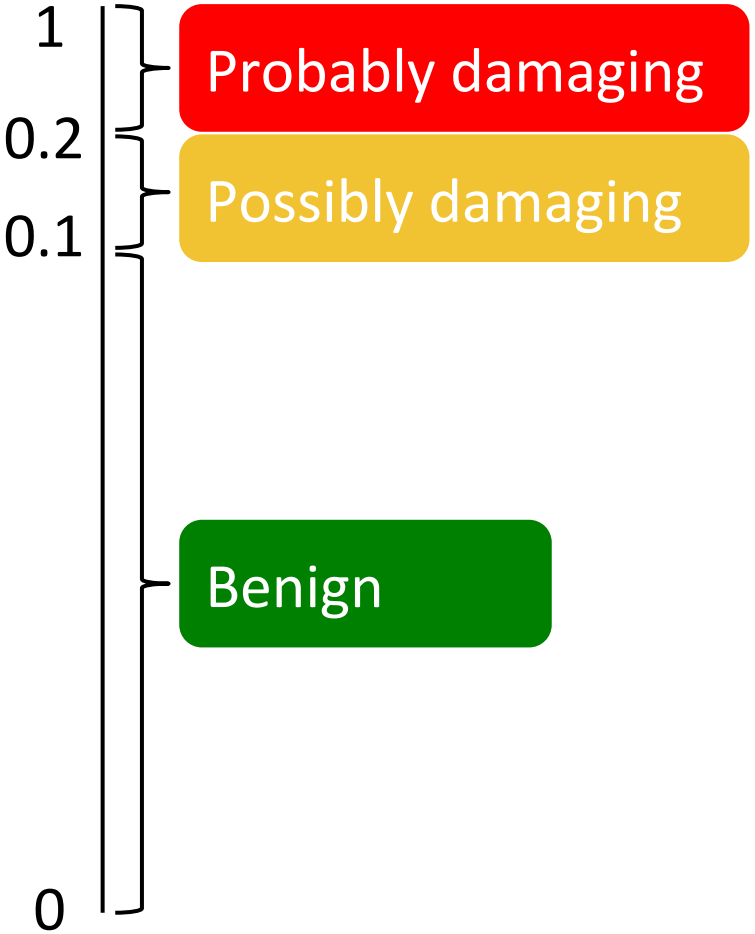
- How well conserved the amino acid is
- The chemical change in the amino acid

Missense variants- pathogenicity

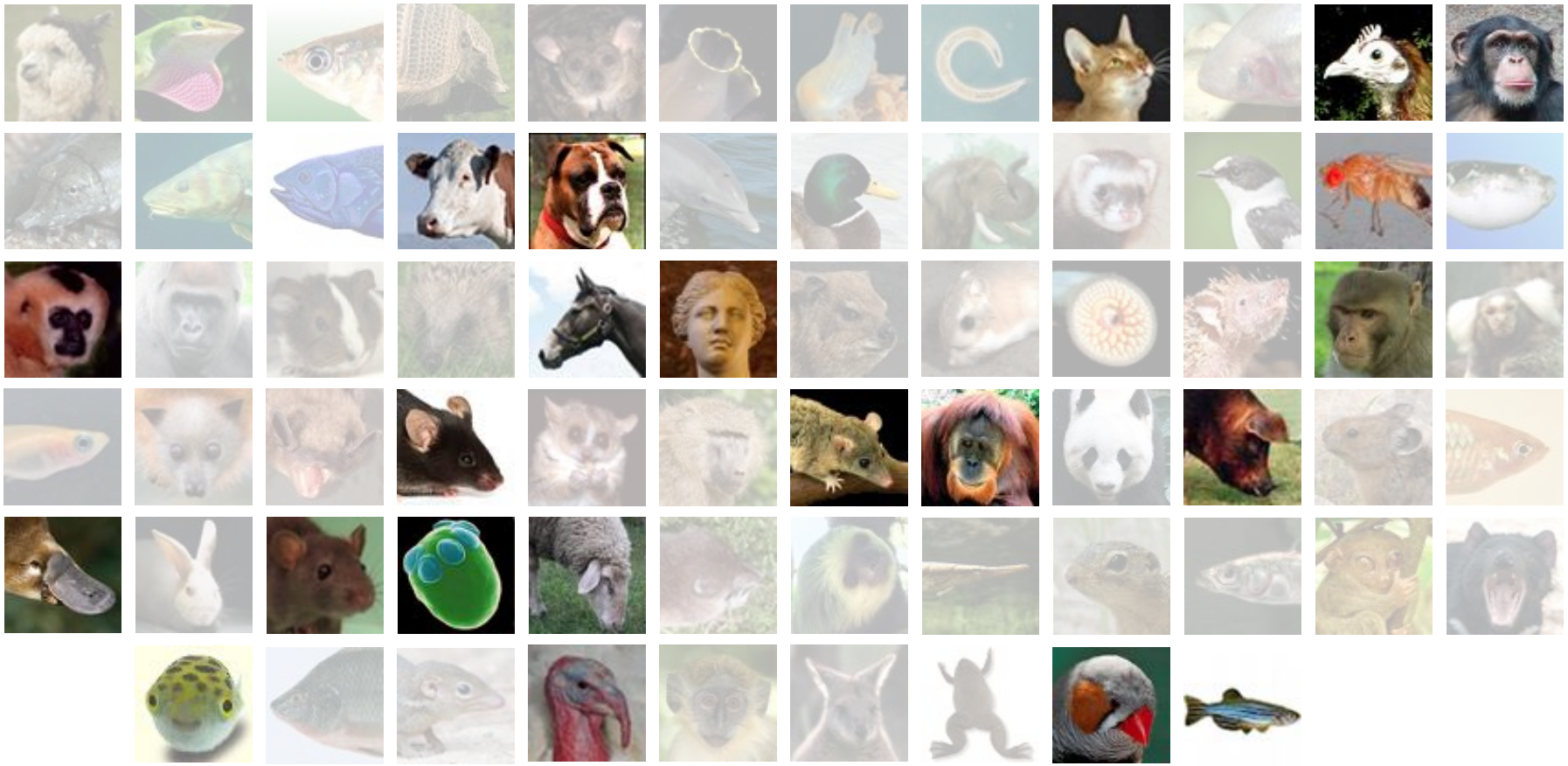
SIFT



PolyPhen



Species with variation data



+ Ensembl Plants, Fungi, Protists and Metazoa

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

Variation sources

dbSNP
Short Genetic Variations

orphanet

OMIM®

DGVarchive

DECIPHER v5.1
GRCh37

PubMed

HGMD®
The Human Gene Mutation Database
Cardiff

UniProt

european genome-phenome archive

GIANT

Meta-Analyses of Glucose and Insulin-related traits Consortium
MAGIC

GEFOS

Genetics

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

http://training.ensembl.org/events

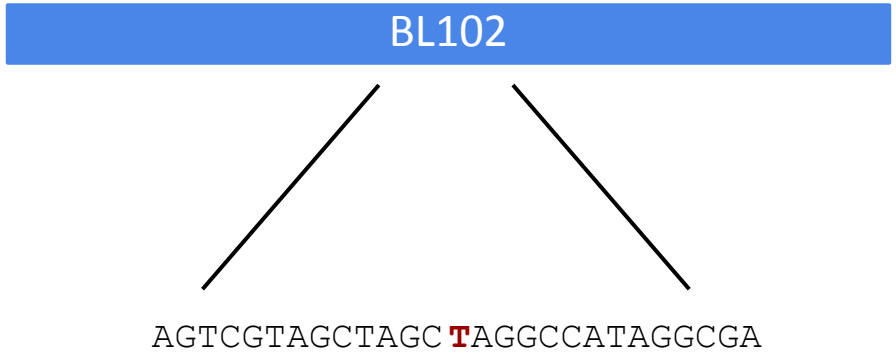
Reference alleles

BL

AL

CM

IM



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

Allele strand

AGTCGTAGCTAGC**T/G**AGGCCATAGGCGA

TCGCCTATGGCCT**A/C**GCTAGCTACGACT

Exon sequence:

TATGGCCT**A/C**GCTAGC

Ensembl	Alleles in database = T/G Alleles in gene = A/C
dbSNP Short Genetic Variations	Alleles = A/C -ve strand or T/G +ve strand
PubMed	Alleles = A/C or T/G Often lack further info

Hands on

- We're going to look at a gene *MCM6* to find variants in the gene.
- We will look at a variant rs4988235 to find more information about it.
- Demo: coursebook page 29-35

Feedback

Survey:

https://goo.gl/forms/52NqsEhD1My52yds1

Find it on the training pages:

http://training.ensembl.org/events/2017/2017-09-14-UNC_BioCurators

Wrap-up

Ensembl is a genome browser which integrates:

- gene annotation
- variation
- regulation
- comparative genomics

How is all this data organised?

- Ensembl browser sites
  - Main website, *Pre!*, *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
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

http://training.ensembl.org/events

Publications

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

Aken B. *et al*

Ensembl 2017

Nucleic Acids Research

https://academic.oup.com/nar/article/45/D1/D635/2605734/Ensembl-2017

Xosé M. Fernández-Suárez and Michael K. Schuster
Using the Ensembl Genome Server to Browse Genomic Sequence Data.
Current Protocols in Bioinformatics 1.15.1-1.15.48 (2010)
www.ncbi.nlm.nih.gov/pubmed/20521244

Giulietta M Spudich and Xosé M Fernández-Suárez
Touring Ensembl: A practical guide to genome browsing
BMC Genomics **11**:295 (2010)
www.biomedcentral.com/1471-2164/11/295

Follow us

www.facebook.com/Ensembl.org

@Ensembl

@BENsembl

www.ensembl.info

http://training.ensembl.org/events

Acknowledgements

The Entire Ensembl Team

Bronwen L. Aken[1], Premanand Achuthan[1], Wasiu Akanni[1], M. Ridwan Amode[1], Friederike Bernsdorff[1], Jyothish Bhai[1], Konstantinos Billis[1], Denise Carvalho-Silva[1], Carla Cummins[1], Peter Clapham[2], Laurent Gil[1], Carlos García Girón[1], Leo Gordon[1], Thibaut Hourlier[1], Sarah E. Hunt[1], Sophie H. Janacek[1], Thomas Juettemann[1], Stephen Keenan[1], Matthew R. Laird[1], Ilias Lavidas[1], Thomas Maurel[1], William McLaren[1], Benjamin Moore[1], Daniel N. Murphy[1], Rishi Nag[1], Victoria Newman[1], Michael Nuhn[1], Chuang Kee Ong[1], Anne Parker[1], Mateus Patricio[1], Harpreet Singh Riat[1], Daniel Sheppard[1], Helen Sparrow[1], Kieron Taylor[1], Anja Thormann[1], Alessandro Vullo[1], Brandon Walts[1], Steven P. Wilder[1], Amonida Zadissa[1], Myrto Kostadima[1], Fergal J. Martin[1], Matthieu Muffato[1], Emily Perry[1], Magali Ruffier[1], Daniel M. Staines[1], Stephen J. Trevanion[1], Fiona Cunningham[1], Andrew Yates[1], Daniel R. Zerbino[1] and Paul Flicek[1,2,*]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [2]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

Open Targets

MedBioinformatics

Co-funded by the European Union

http://training.ensembl.org/events

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