

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
Helsinki - 14th June 2016



Structure for this workshop

Introduction to Ensembl

Exploring Ensembl - Genomic regions, Genes and Transcripts

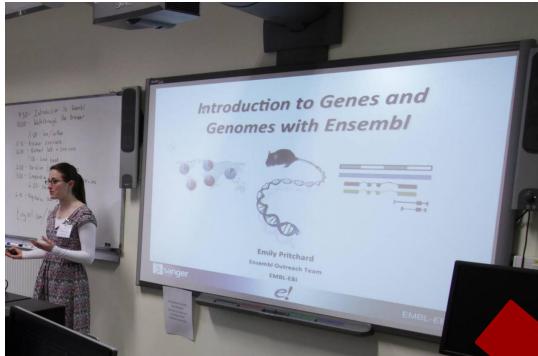
Variation data

The Variant Effect Predictor

- Web-interface
- Perl Script
- REST API

<http://www.ebi.ac.uk/~bmoore/workshops/>

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!

Extra Exercises



Questions?

<http://www.ebi.ac.uk/~bmoore/workshops/>

Course materials

www.ebi.ac.uk/~bmoore/workshops

- Presentations
- Coursebook (demos and exercises)
- Plain Text Files for exercises
- Answerbook (exercise answers)

<http://www.ebi.ac.uk/~bmoore/workshops/>

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**.

<http://www.ebi.ac.uk/~bmoore/workshops/>

Exploring the Ensembl genome browser



e|Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors [Login/Register](#)

Search: for e.g. [BRCA2](#) or [rat X:100000..200000](#) or [coronary heart disease](#)

Browse a Genome
The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

-  **Human** GRC37
-  **Mouse** GRCm38
-  **Zebrafish** Zv9

All genomes
[Select a species](#)

[View full list of all Ensembl species](#)
Other species are available in [Ensembl Pre!](#) and [EnsemblGenomes](#)

ENCODE data in Ensembl

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variations for my gene

Retrieve gene sequence

Compare genes across species

Use my own data in Ensembl

Learn about a disease or phenotype

What's New in Release 73 (September 2013)

- Updated patches for the human assembly (GRCh37.p11)
- New variation citation page and individual genotype search box
- Upload VEP output to the ensembl website

[More release news on our blog](#) →

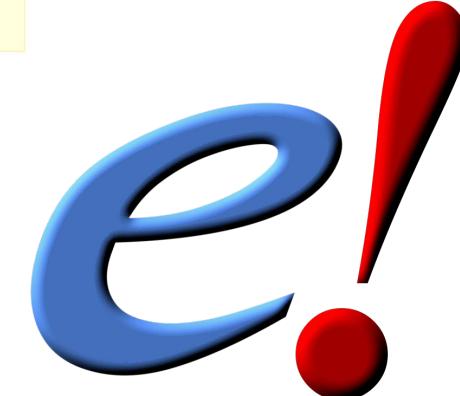
Latest blog posts

- 15 Aug 2013: Ensembl archive 71 downtime, Fri 16th August
- 07 Aug 2013: Announcing Ensembl eCode
- 24 Jul 2013: What's coming in Ensembl release 73

[Go to Ensembl blog](#) →

Did you know...?

FAQs Questions? Check out our [FAQs](#).



Why do we need genome browsers?

1977: 1st genome to be sequenced (5 kb)
2004: finished human sequence (3 Gb)

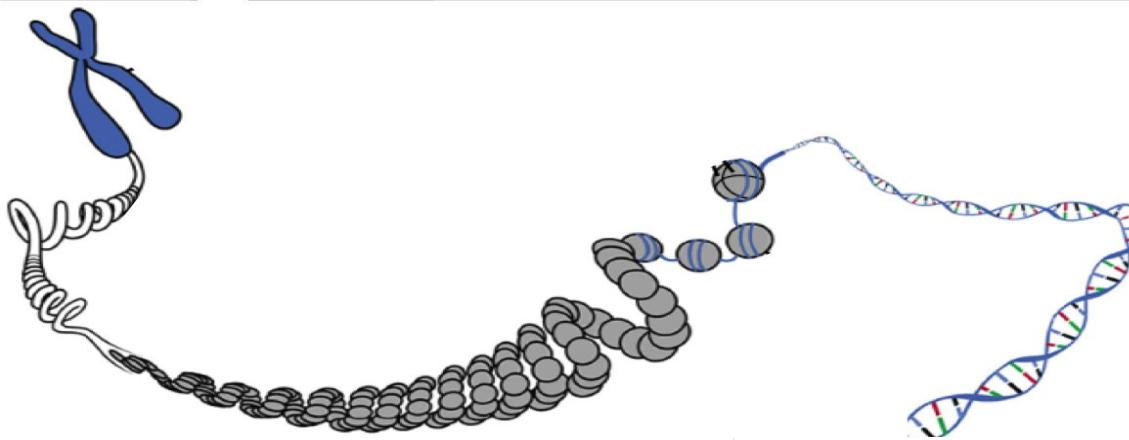


<http://www.ebi.ac.uk/~bmoore/workshops/>

Why do we need genome browsers?

CGGCCTTGGGCTCCGCCTTCAGCTAAGACTTAACCTCCCTCCAGCTGTCCCAGATGACGCCATCTGAAATTCTTGGAAACACGATCAC
TTAACGGAATATTGCTTTGGGAAGTGTACAGCTGCTGGCACGCTGTATTGCCTACTTAAGCCCCTGGTAATTGCTGTATT
CGAAGACATGCTGATGGATTACAGGGCGCGTGGCTCTAACTGGAGCCCTGTCCCCACTAGCCACGCGTCAGTGGTTAGCGTGATT
GAAACTAAATCGTATGAAAATCCTCTCTAGTCGCACTAGGCCACGTTCGAGTGCTTAATGGCTAGTGGCACCGGTTGGACAGCAC
GCTGTAAAATGTTCCCATCCTCACAGTAAGCTGTTACCGTTCCAGGAGATGGGACTGAATTAGAATTCAAACAAATTTCAGCGCTCTGA
GTTTACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTGGCTTCTGTTGCAGACTTACCAAGCATTGGAGGA
ATATCGTAGGAAAAATGCCTATTGGATCCAAAGAGAGGCCAACATTGGAAATTGAAACAGCCTGCAACAAAGCAGGTATTGACAA
ATTTATATAACTTATAAATTACACCGAGAAAGTGTCTAAAAAATGCTGCTAAAACCAGTACGTACAGTGTGCTTAGAACCCAT
AAACTGTTCTTATGTGTGATAAAATCCAGTTAACACATAATCATCGTTGCAGGTTAACCATGATAAATATAGAACGTCTAGTGGATA
AAGAGGAAACTGGCCCTTGACTAGCAGTAGGAACAAATTACTAACAAATCAGAACATTATGTTACTTATGGCAGAAGTGTCCAACCTT
TTGGTTTCAGTACTCCTTACTCTTAAATGATCTAGGACCCCCGGAGTGCTTTGTTATGTAGCTTACCATATTAGAAATTAAAAC
AAGAATTAAAGGCTGGCGTGGCTCACGCTGTAATCCCAGCACTTGGGAGGCCAGGTGGCGGATCATTGAGGCCAGAAGTTGA
GACCAGCCTGGCCAACATGGTGAACACCTATCTACTAAAATACAAAAATGTGCTGCGTGTGGTGGCGCTGTAATCCAGCTAC
ACGGGAGGTGGAGGCAGGAGAACGCTGAACCCCTGGAGGCAGAGGTTGCAGTGAGCCAAGATCATGCCACTGCACTAGCCTGGCCAC
TAGCATGACTCTGTCTAAAACAAACAAACAAAAACTAACAGAATTAAAGTTAACATTAAAGTAAACTTACATTAAATAATGAAAGCTAACCCATTGCA
TATTATCACAAACATTCTAGGAAAAAAACTTTGAAAACAAGTGAATGAGTGGAAATAGTTTACATTGCACTTCTTTAATGTCTGGCT
AAATAGAGATAGCTGGATTCACTTATCTGTGCTAATCTGTTATTGGTAGAAGTATGTGAAAAAAATTACCTCACGTTAAAAAGGA
ATATTAAATAGTTTCAGTTACTTTGGTATTTCCTGTACTTGCATAGATTCAAAGATCTAACAGATATAACCATAGGTCTT
CCATGTCGCAACATCATGCACTGATTATTGAAAGATAGTGGTGTCTGAATTACAAAGTCCAAATATTGATAAATTGCACTAAACAT
TTTAAAAATCTCATTCAATTAAATACCACCATGGATGTCAGAAAAGTCTTAAAGATTGGTAGAAATGAGCCACTGAAATTCTAAC
TTTGAAAGTTCACATTGCAACAAACTGTTCCCTGCAGCAACAGATCACTCATTGATTGTGAGAAAATGTCTACCAAAT
TATTAAAGTTGAAATAACTTGTCACTGTTCAAGTAAAATGACTTTCATTGAAAGATAGTGGTGTCTGAATTACAAAGTCCAAATATT
AGTGTCTTAGGCAGTATTGACTTCAGTATGCAGAAGTGTCTTATGTATGCTTCAAGTAAAATGACTTTCATTGAAAGATAGTGGTGTCTGA
GCATTGAGCTTCGAAATTAAATTTCATTGCTTCATTAGGACATTCTACATTAAACTGGCATTATTACTATTATTTAACAGGACAC
TCAGTGGTAAGGAATATAATGGCTACTAGTATTAGTTGGTGCCTGCCACTGCCATAACTCATGCAAATGTGCCAGCAGTTACCCAGCATC
TTTGCACGTGTGATACAAATGTCAACATCATGAAAAAGGTTGAAAAAGGAATATTAAATAGTTTCAGTTACTTATGACTGTTAGCTA
<http://www.ebi.ac.uk/~bmoore/workshops/>

Ensembl- unlocking the code

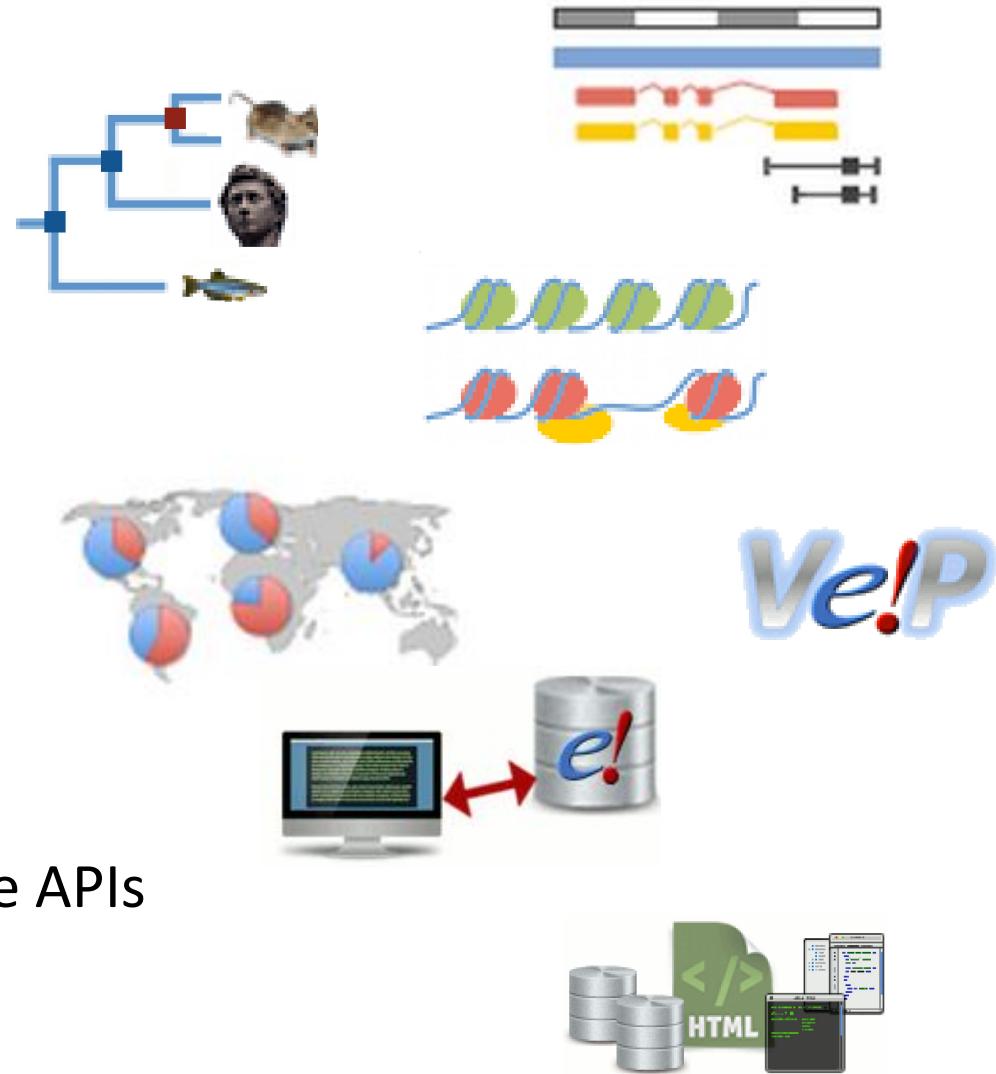


- Genomic assemblies - automated gene annotation
- Variation - Small and large scale sequence variation with phenotype associations
- Comparative Genomics - Whole genome alignments, gene trees
- Regulation - Potential promoters and enhancers, DNA methylation

<http://www.ebi.ac.uk/~bmoore/workshops/>

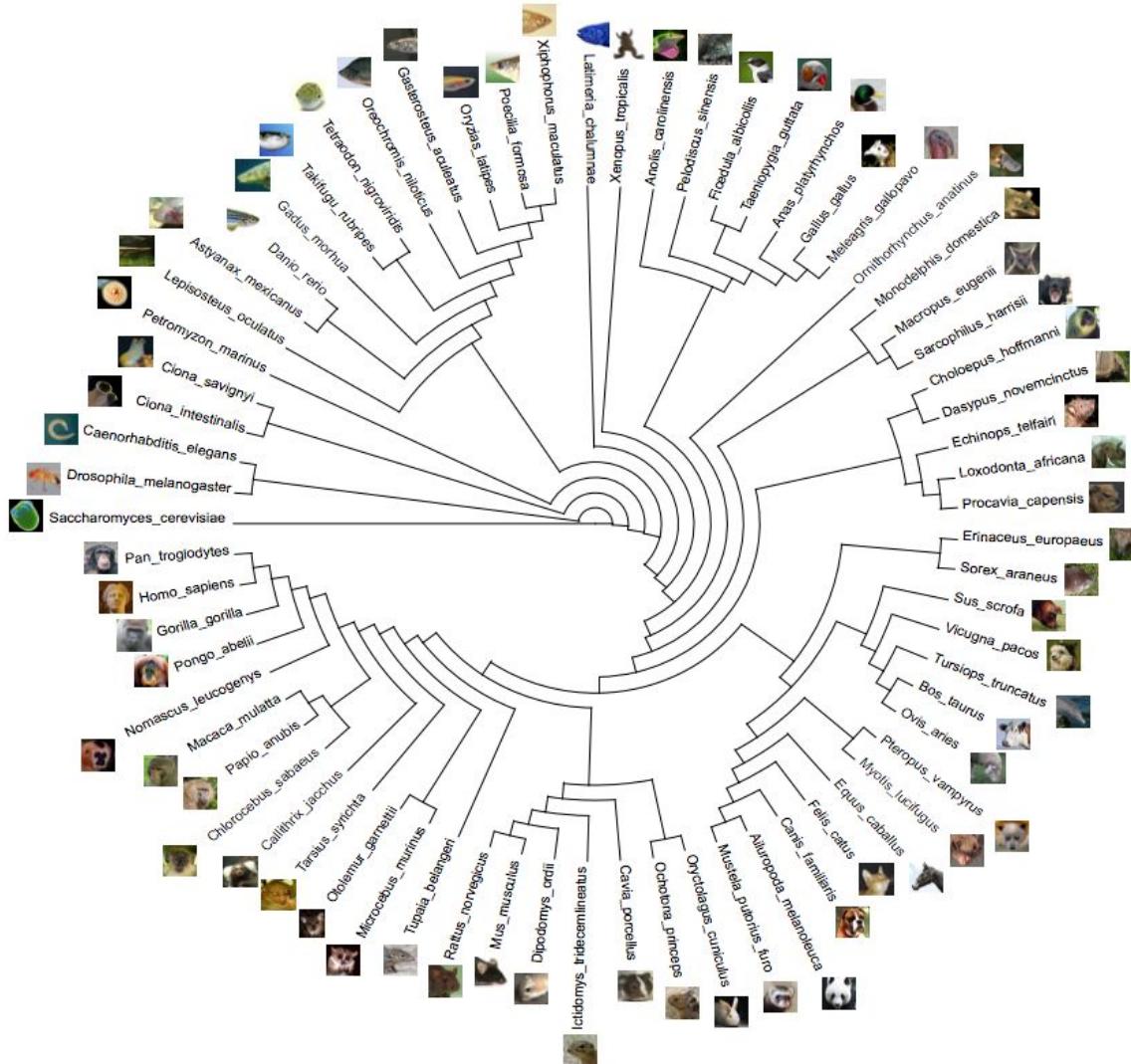
Ensembl Features

- Gene builds for ~70 species
- Gene trees
- Regulatory build
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source



<http://www.ebi.ac.uk/~bmoore/workshops/>

Ensembl- access to 70+ genomes



<http://www.ebi.ac.uk/~bmoore/workshops/>

Ensembl Genomes- expanding Ensembl



www.ensembl.org

- Vertebrates



- Other representative species

<http://www.ebi.ac.uk/~bmoore/workshops/>

Ensembl Genomes- expanding Ensembl



www.ensembl.org



www.ensemblgenomes.org

- Vertebrates



- Other representative species

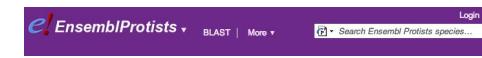
- Bacteria



- Fungi



- Protists



- Metazoa



- Plants



<http://www.ebi.ac.uk/~bmoore/workshops/>

What is a genome assembly?

Sequence reads

CGGCCTTGGCCTCCGCCTTCAGCTCAAGA

CAGCTGTCCCAGATGAC ACTTAACCTCCCTCCCAGCTGTCC

GGGCTCCGCCTTCAGCTC TCCCAGCTGTCCCAGATGACGCCAT

CGGCCTTGGCCTCC AACCTCCCTCCCAGCT
CAGATGACGCC TCCGCCTTCAGCTCAAGACTTAACCTC

Match up overlaps

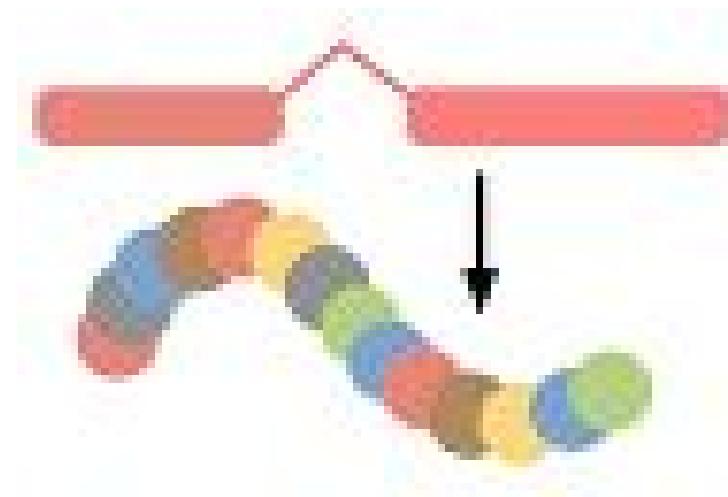
CGGCCTTGGCCTCCGCCTTCAGCTCAAGA AACCTCCCTCCCAGCT CAGATGACGCC
TCCGCCTTCAGCTCAAGACTTAACCTC TCCCAGCTGTCCCAGATGACGCCAT
GGGCTCCGCCTTCAGCTC ACTTAACCTCCCTCCCAGCTGTCC
CGGCCTTGGCCTCC CAGCTGTCCCAGATGAC

Genome assembly

CGGCCTTGGCCTCCGCCTTCAGCTCAAGACTTAACCTCCCTCCCAGCTGTCCCAGATGACGCCAT
<http://www.ebi.ac.uk/~bmoore/workshops/>



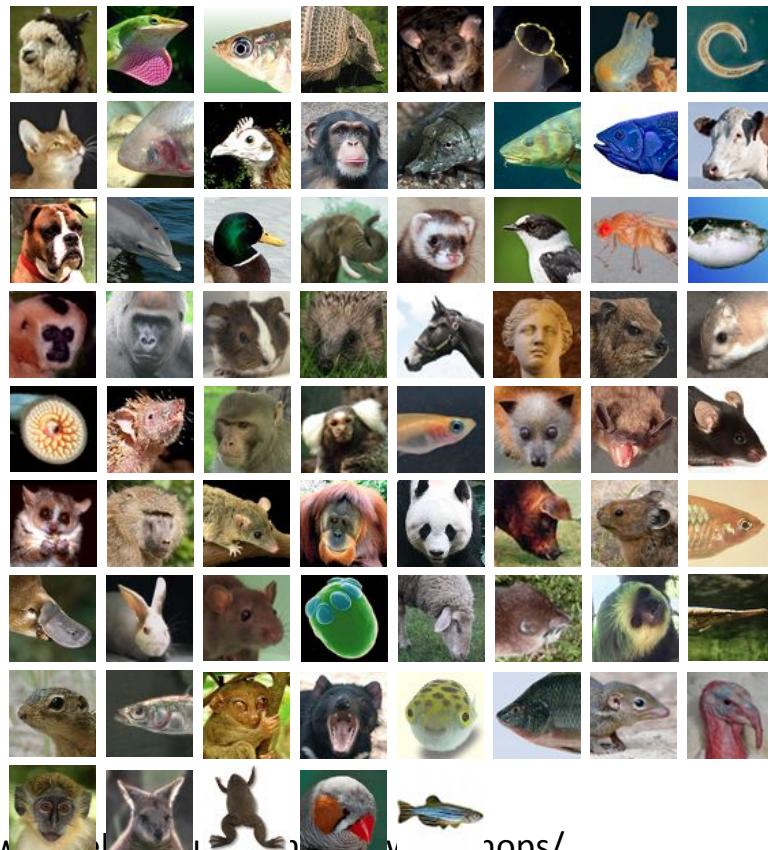
Genes and Transcripts



Ensembl and Havana annotation



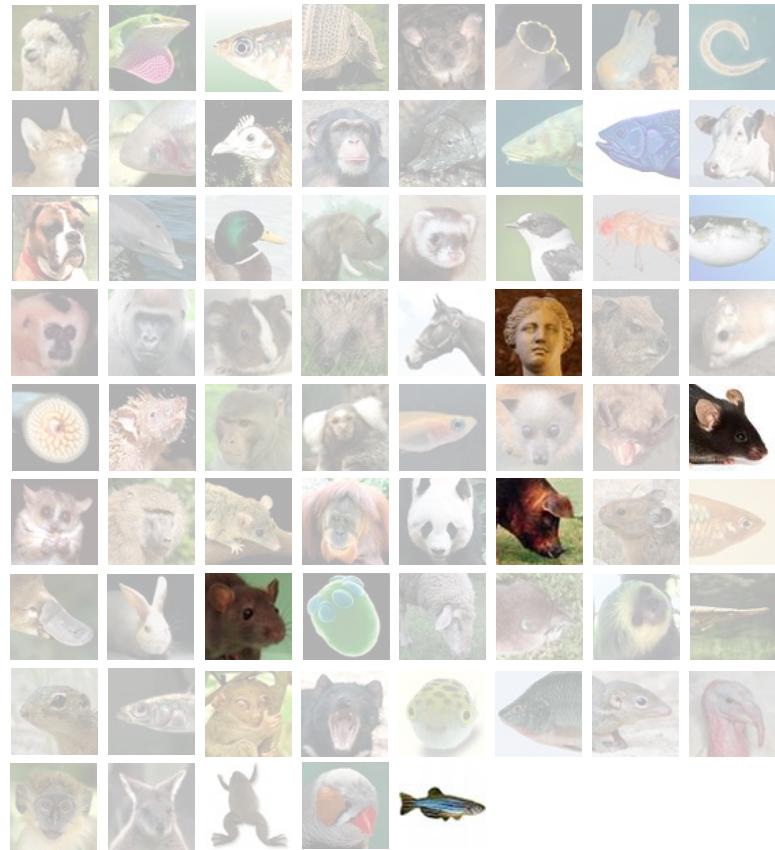
Automatic annotation



<http://www.ensembl.org/info/docs/>

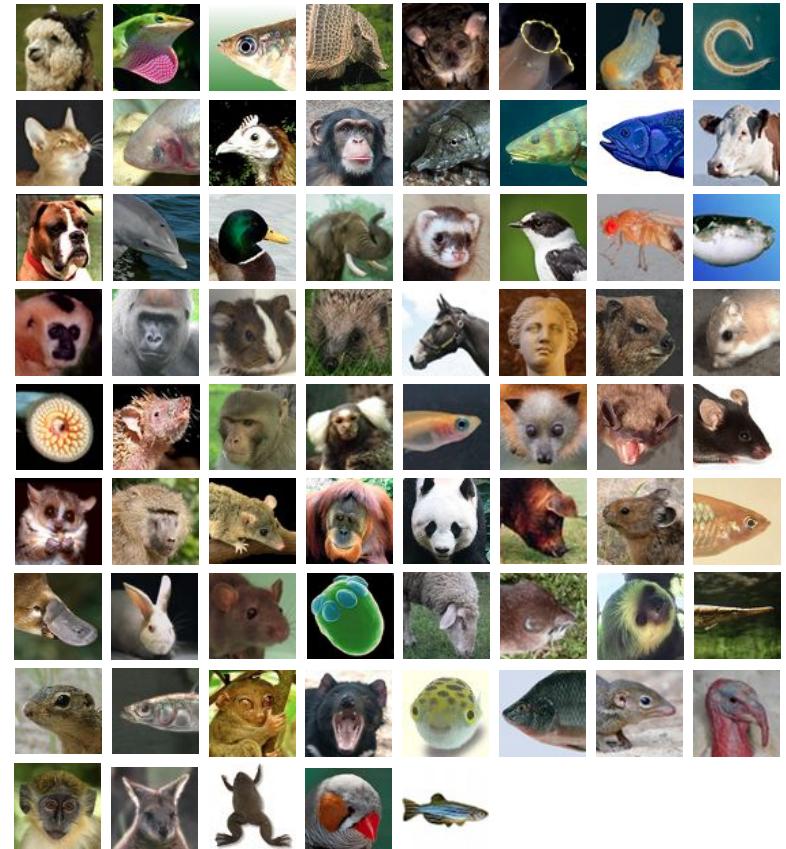


Manual annotation



Automatic gene annotation

- Genome-wide determination using the Ensembl automated pipeline
- Predictions based on experimental (biological) data



<http://www.ebi.ac.uk/~bmoore/workshops/>

Biological Evidence

- International Nucleotide Sequence databases



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



- NCBI RefSeq
 - Manually annotated proteins and mRNAs (NP, NM)

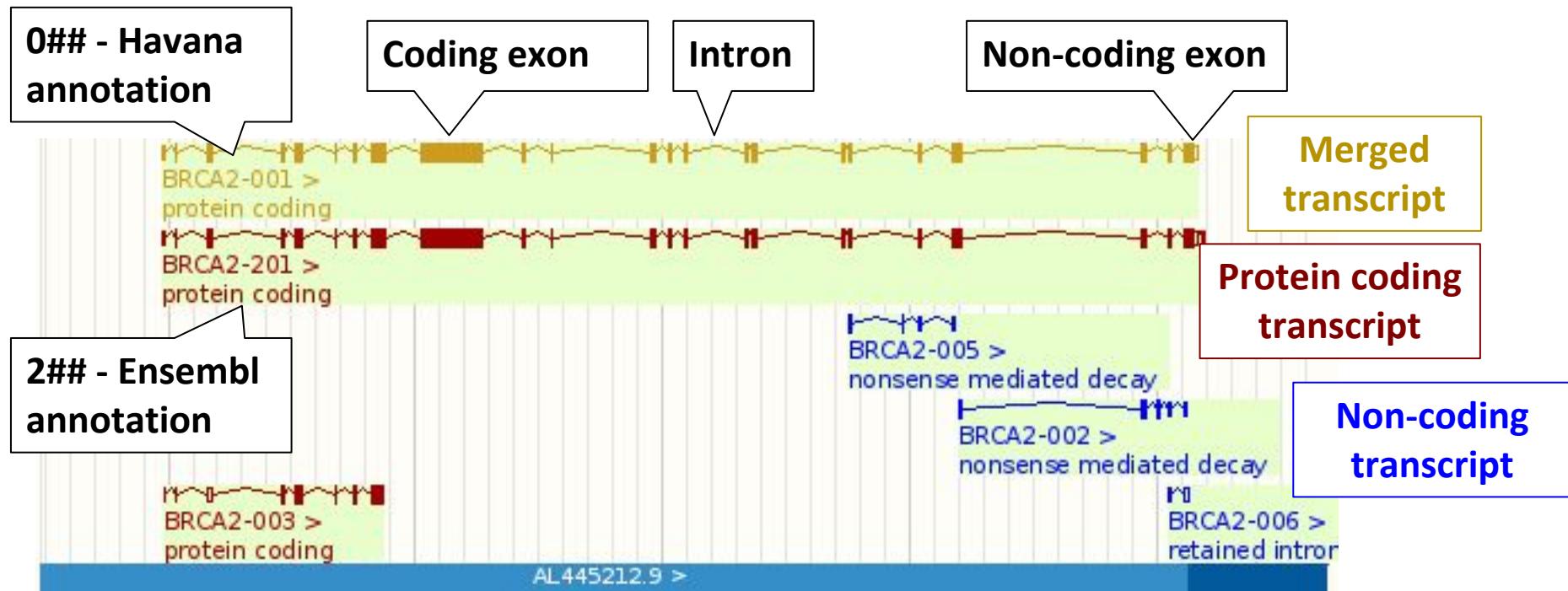


Other species

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



Gene views



Golden transcripts

- Identical annotation

*e!*Ensembl

havana
Human and mouse analysis



- Higher confidence and quality



<http://www.ebi.ac.uk/~bmoore/workshops/>

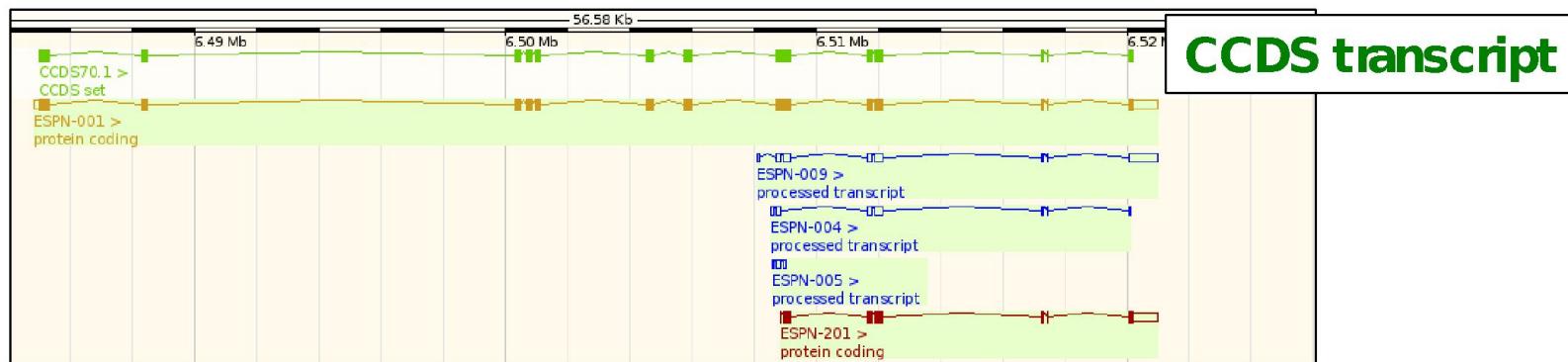
CCDS transcripts



- Consensus coding DNA sequence set
- Agreement between EBI, WTSI, UCSC and NCBI



<http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi>



<http://www.ebi.ac.uk/~bmoore/workshops/>

Ensembl stable IDs

- ENS**G**##### Ensembl Gene ID
- ENS**T**##### Ensembl Transcript ID
- ENS**P**##### Ensembl Peptide ID
- ENS**E**##### Ensembl Exon ID
- ENS**R**##### Ensembl Regulatory region ID
- For non-human species a suffix is added:
MUS (*Mus musculus*) for mouse ENS**MUS**G###
DAR (*Danio rerio*) for zebrafish: ENS**DAR**G###

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

<http://www.ebi.ac.uk/~bmoore/workshops/>

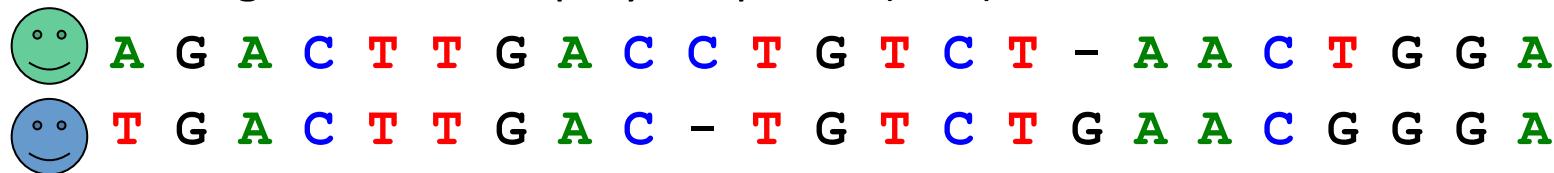
Variation



Variation types

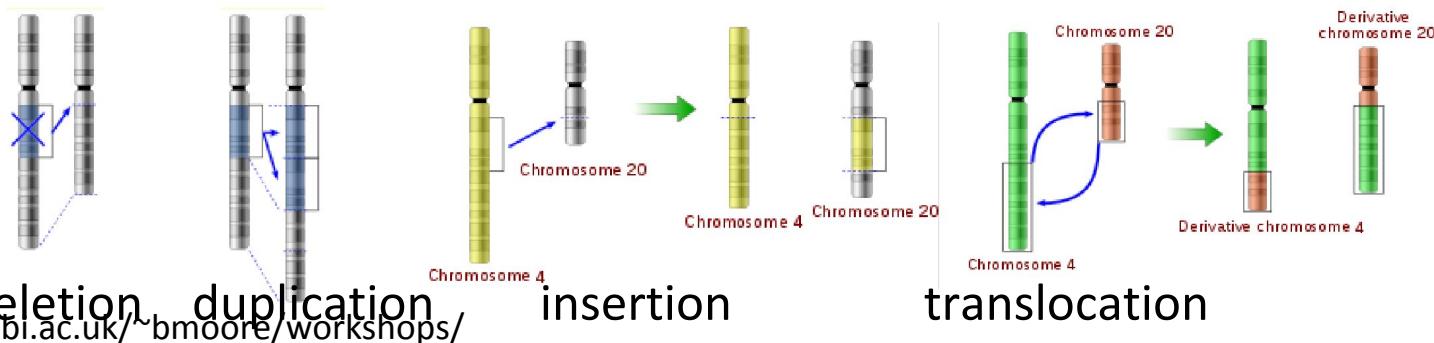
1) Small scale in one or few nucleotides of a gene

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



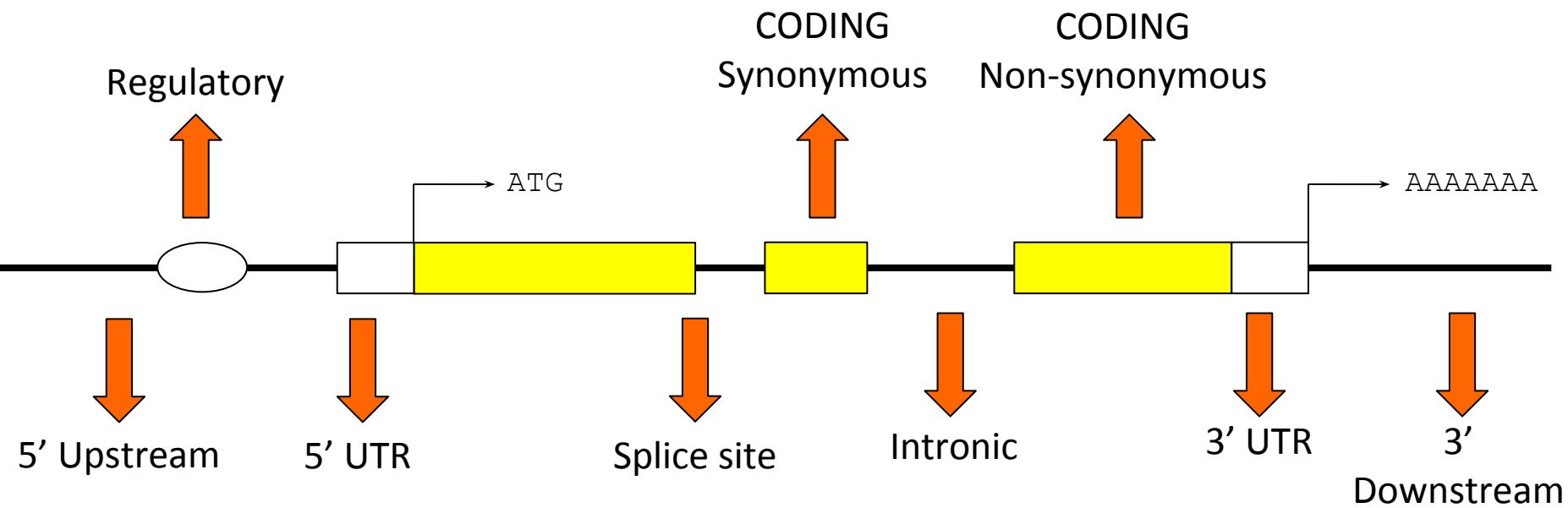
2) Large scale in chromosomal structure (structural variation)

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



<http://www.ebi.ac.uk/~bmoore/workshops/>

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 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

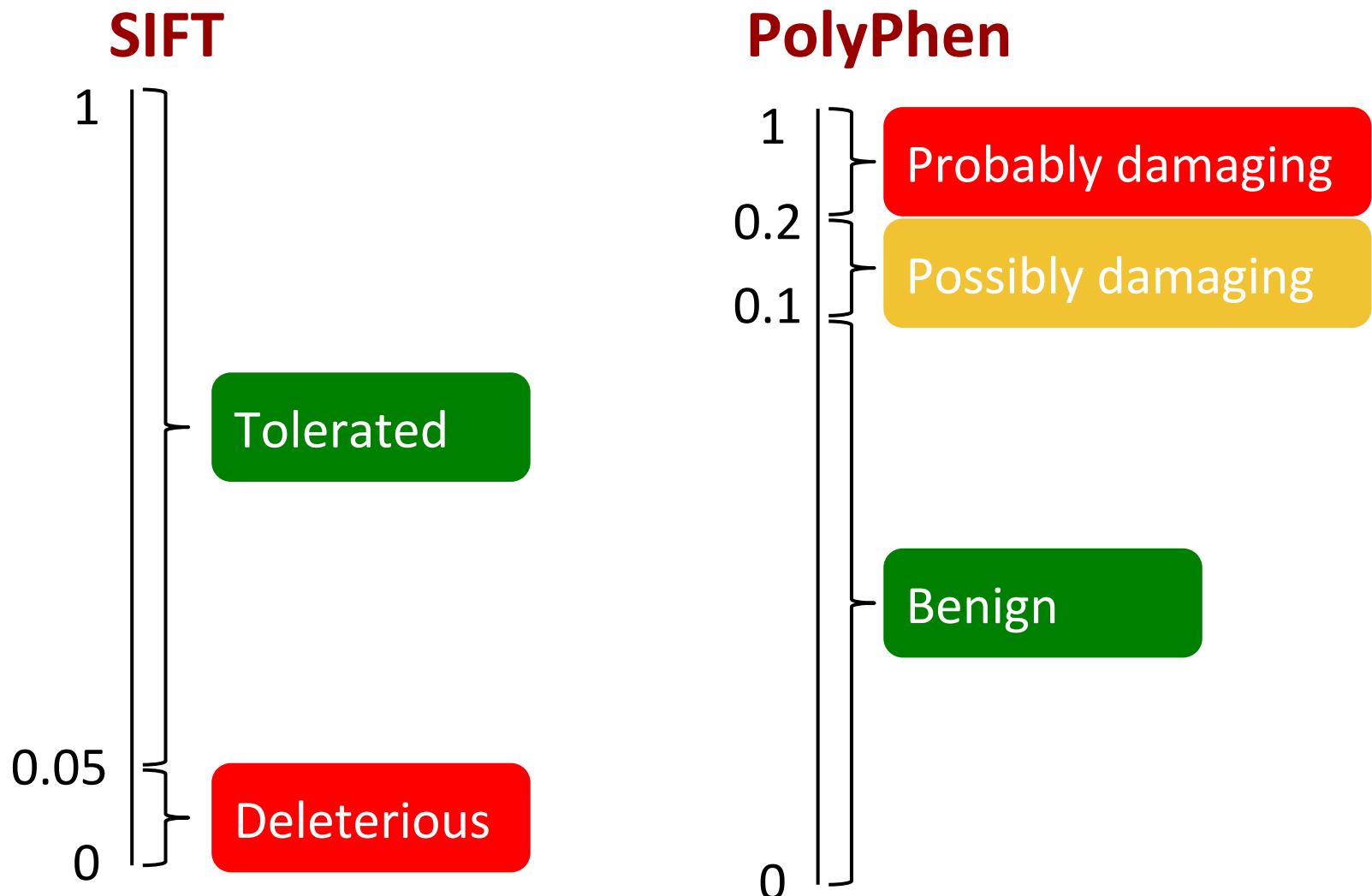
<http://www.ebi.ac.uk/~bmoore/workshops/>

Missense variants- pathogenicity

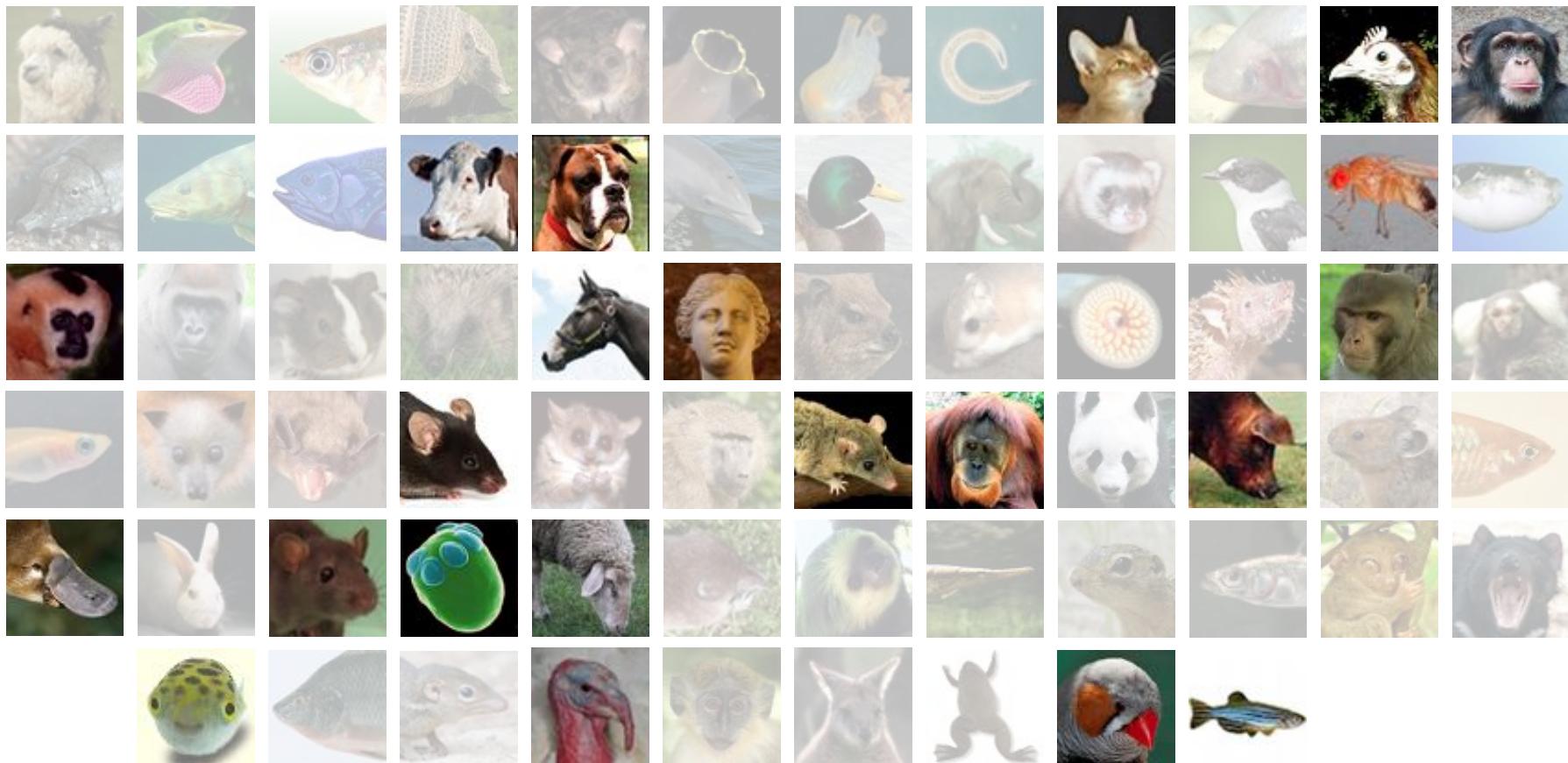
SIFT and PolyPhen score changes in amino acid sequence based on:

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

Missense variants- pathogenicity



Species with variation data



+ Ensembl Plants, Fungi, Protists and Metazoa

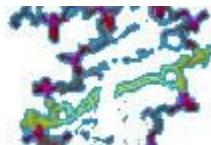
http://www.ensembl.org/info/genome/variation/sources_documentation.html

<http://www.ebi.ac.uk/~bmoore/workshops/>

Variation sources

dbSNP

Short Genetic Variations



orphanet

 **DECIPHER** v5.1
GRCh37

OMIM®

*DGVa*rchive



PublMed

European
*g*enome-phenome
*g*archive


HGMD®
The Human Gene Mutation Database
Cardiff



G
I
ANT

Meta-Analyses of Glucose and Insulin-related traits Consortium

MAGIC

GEF  **S**



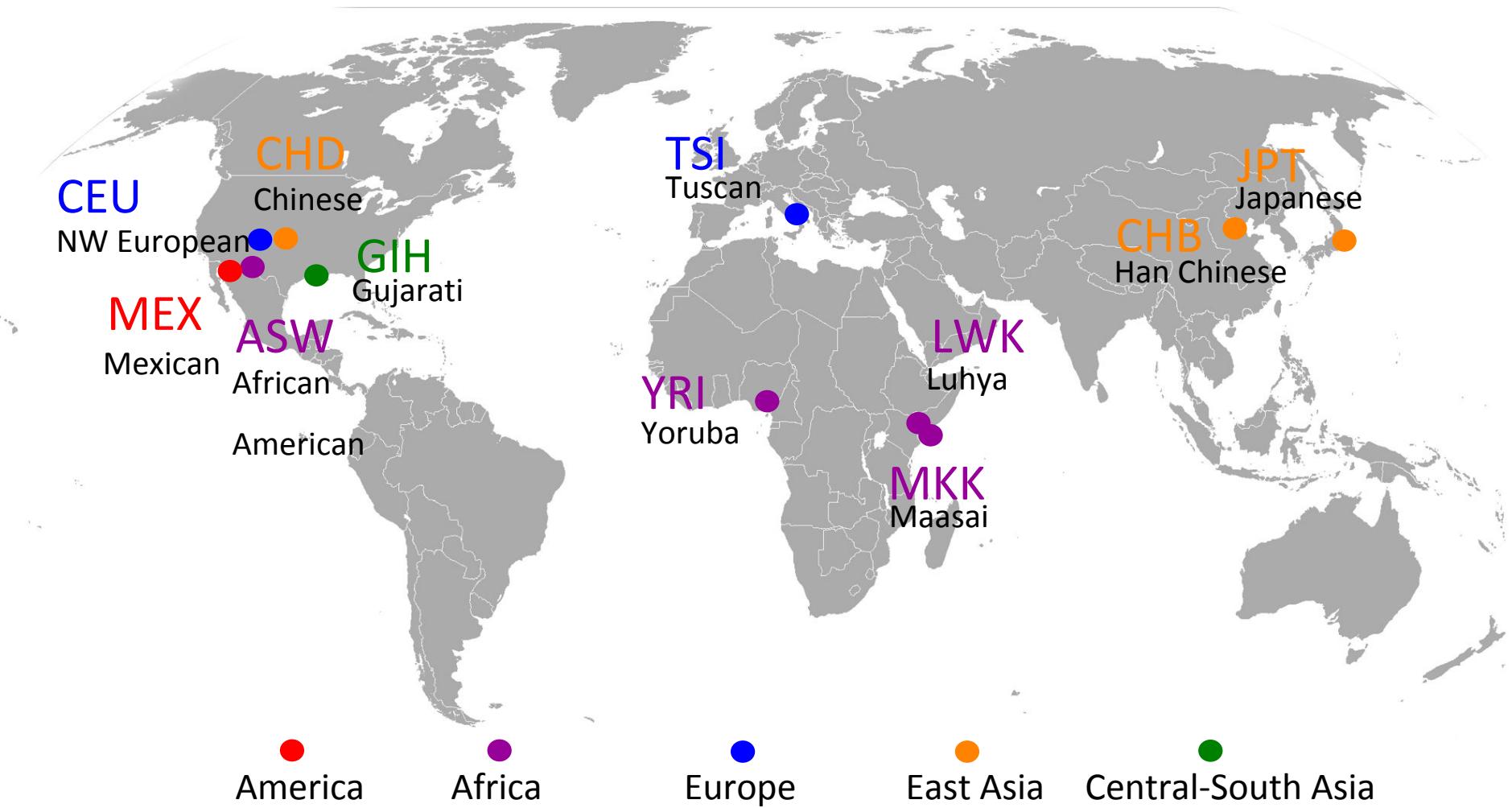
 **B**3 Genetics

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

<http://www.ebi.ac.uk/~bmoore/workshops/>

HapMap project

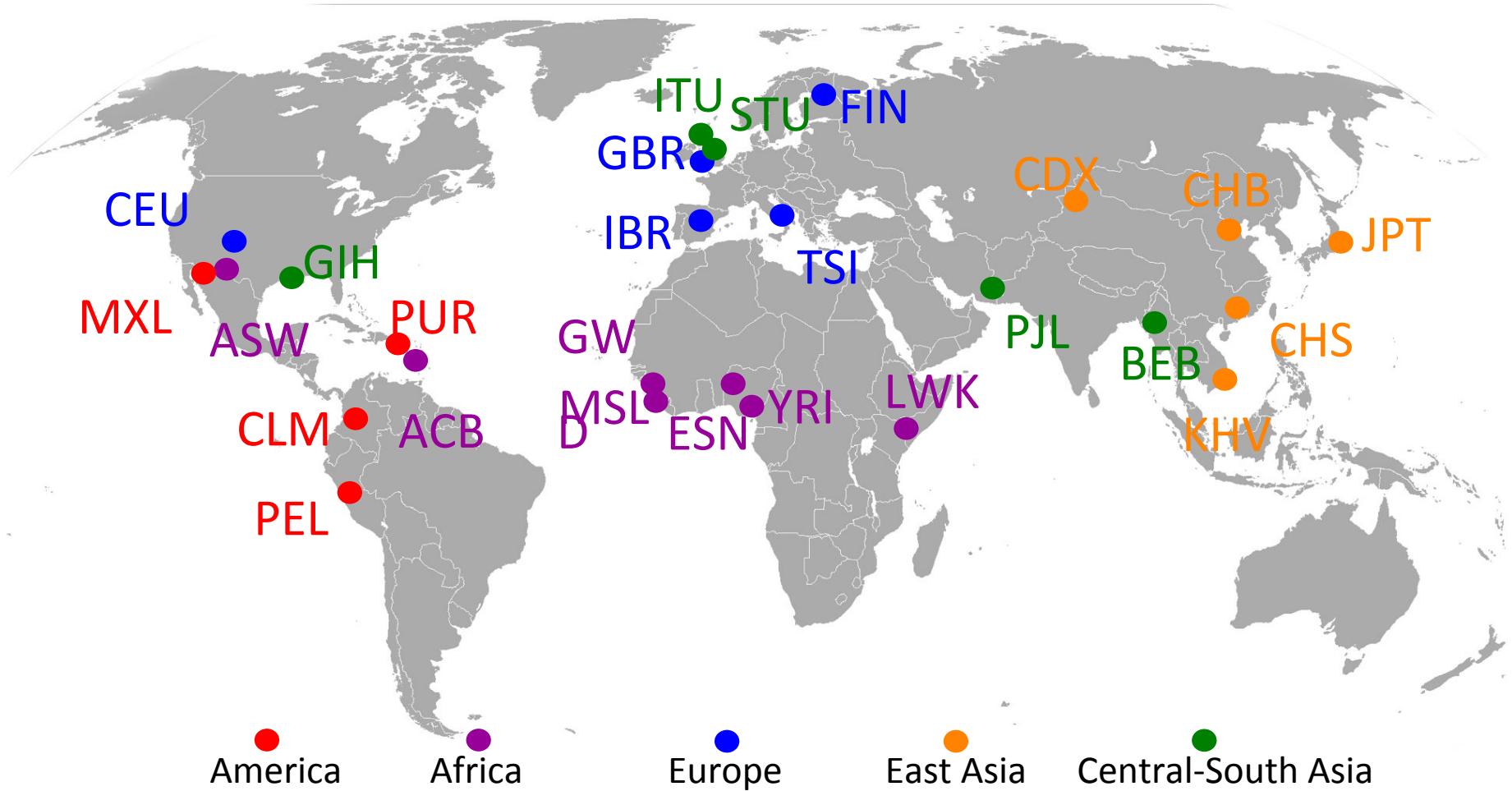
Genotyping 1,301 individuals from 11 populations.



<http://www.ebi.ac.uk/~bmoore/workshops/>

1000 genomes project

Sequencing 2,500 individuals at 4X coverage



<http://www.ebi.ac.uk/~bmoore/workshops/>

<http://www.ensembl.org/Help/Faq?id=328>

Reference alleles



BL



AL



CM



IM

BL102 AL476
CM553 IM768



BL102

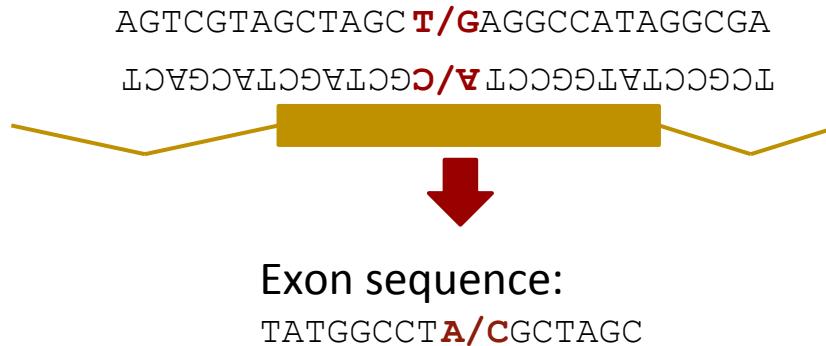
AGTCGTAGCTAGC **T**AGGCCATAGGCGA

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

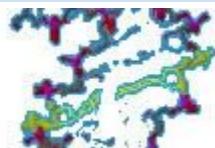
- T is the reference allele
- G is the alternate allele
- Alleles are T/G

Allele strand



Alleles in database = T/G
Alleles in gene = A/C

dbSNP
Short Genetic Variations



Alleles = A/C -ve strand or
T/G +ve strand



Alleles = A/C or T/G
Often lack further info

<http://www.ebi.ac.uk/~bmoore/workshops/>

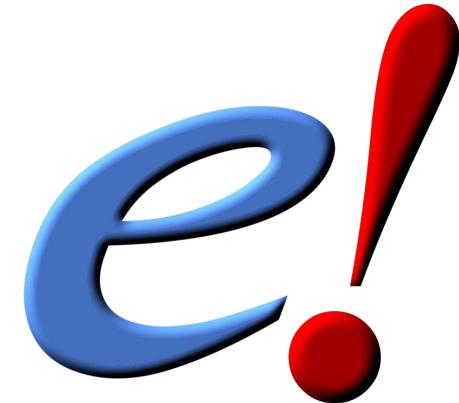


Hands on

- 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.
- Demo: coursebook page 6-21
- Exercises: coursebook page 21-23
 - Answers: answer book page 3-5



Data Mining with BioMart



Outline of this session

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

<http://www.ebi.ac.uk/~bmoore/workshops/>

What is BioMart?

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

<http://www.ebi.ac.uk/~bmoore/workshops/>

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://www.ebi.ac.uk/~bmoore/workshops/>

Availability

Ensembl

Ensembl Plants

Ensembl Fungi (some exceptions)

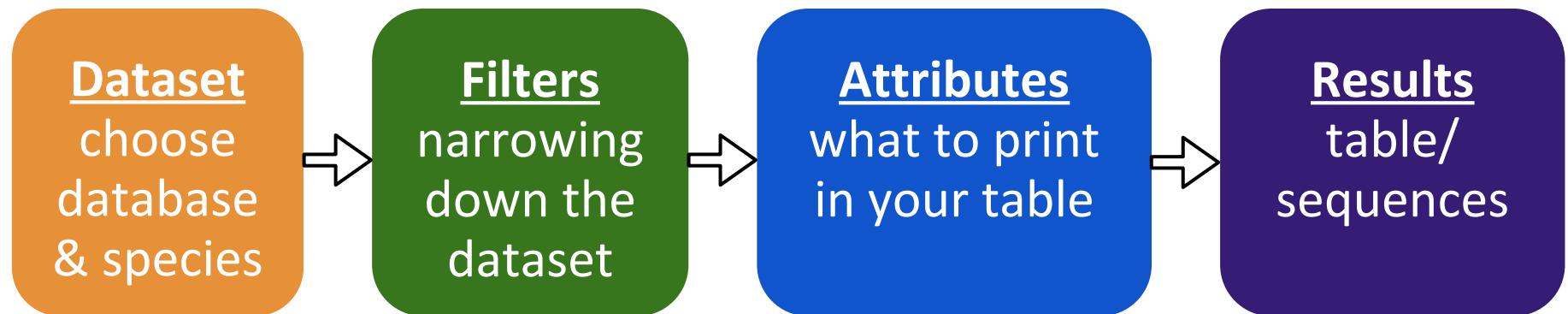
Ensembl Metazoa

Ensembl Protists (some exceptions)

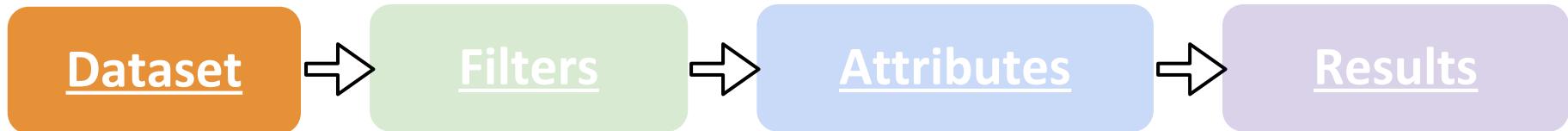
<http://www.ebi.ac.uk/~bmoore/workshops/>

How do I use BioMart?

The 4 steps

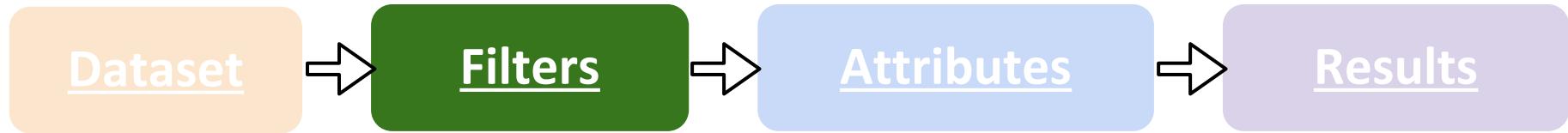


Step 1: Dataset



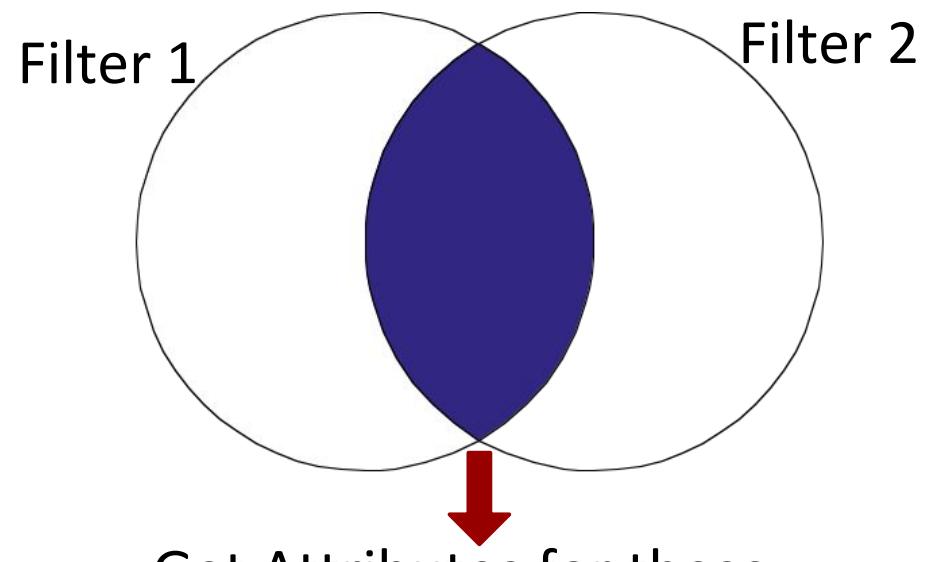
- Define the database that you want to search with your filters
 - Genes, Variation, Regulation
- Define the species

Step 2: Filters

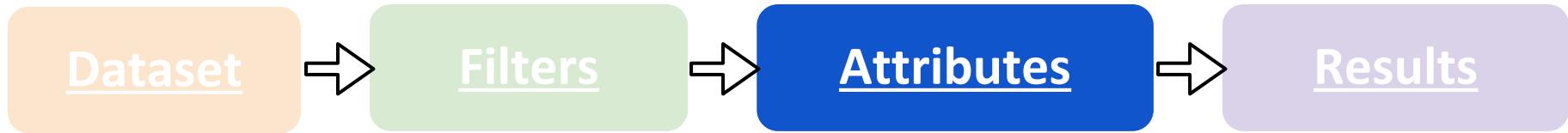


Define a (large) set of genes/variants by combinations of parameters, eg:

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



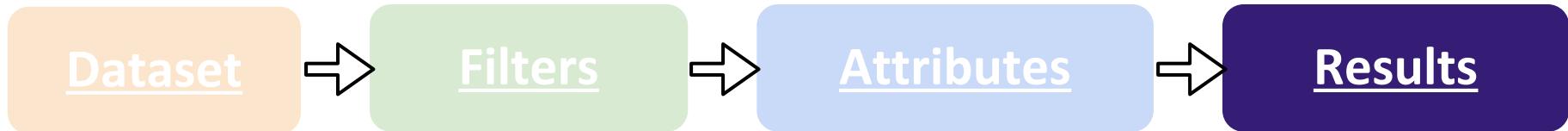
Step 3: Attributes



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

- IDs
- Features
- Sequences
- Orthologues/Paralogues

Step 4: Results

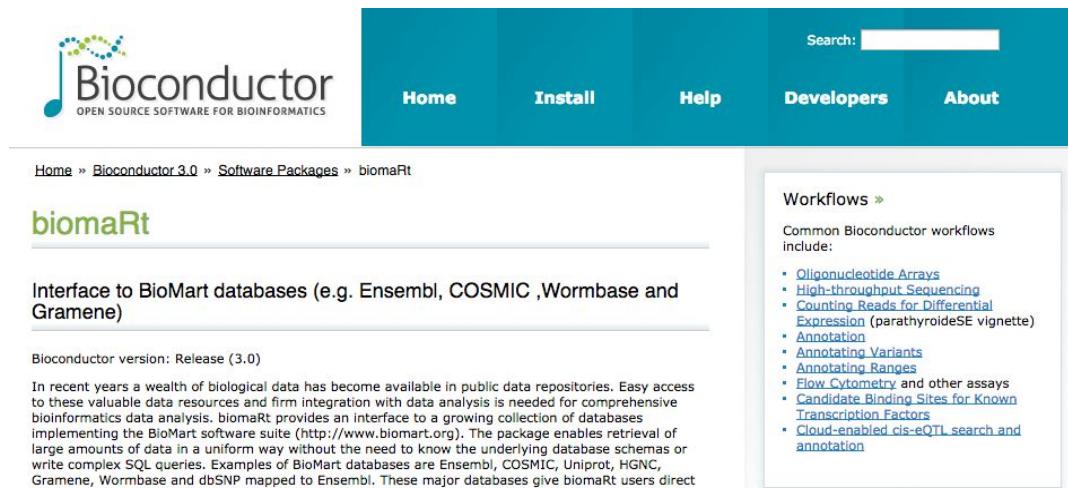


View and download the datatable in a number of formats:

- html
- tsv
- CSV
- xls
- fasta

<http://www.ebi.ac.uk/~bmoore/workshops/>

biomaRt



The screenshot shows the Bioconductor website with a teal header. The header includes the Bioconductor logo (a stylized musical note composed of colored dots), the text 'Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS', and navigation links for 'Home', 'Install', 'Help', 'Developers', and 'About'. A search bar is also present. Below the header, the page title is 'biomaRt' and the subtitle is 'Interface to BioMart databases (e.g. Ensembl, COSMIC, Wormbase and Gramene)'. The content area contains a paragraph about Bioconductor version 3.0, a detailed description of the biomaRt package, and a sidebar titled 'Workflows' listing various Bioconductor workflows.

Workflows »

Common Bioconductor workflows include:

- [Oligonucleotide Arrays](#)
- [High-throughput Sequencing](#)
- [Counting Reads for Differential Expression](#) (parathyroideSE vignette)
- [Annotation](#)
- [Annotating Variants](#)
- [Annotating Ranges](#)
- [Flow Cytometry](#) and other assays
- [Candidate Binding Sites for Known Transcription Factors](#)
- [Cloud-enabled cis-eQTL search and annotation](#)

- 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>

<http://www.ebi.ac.uk/~bmoore/workshops/>

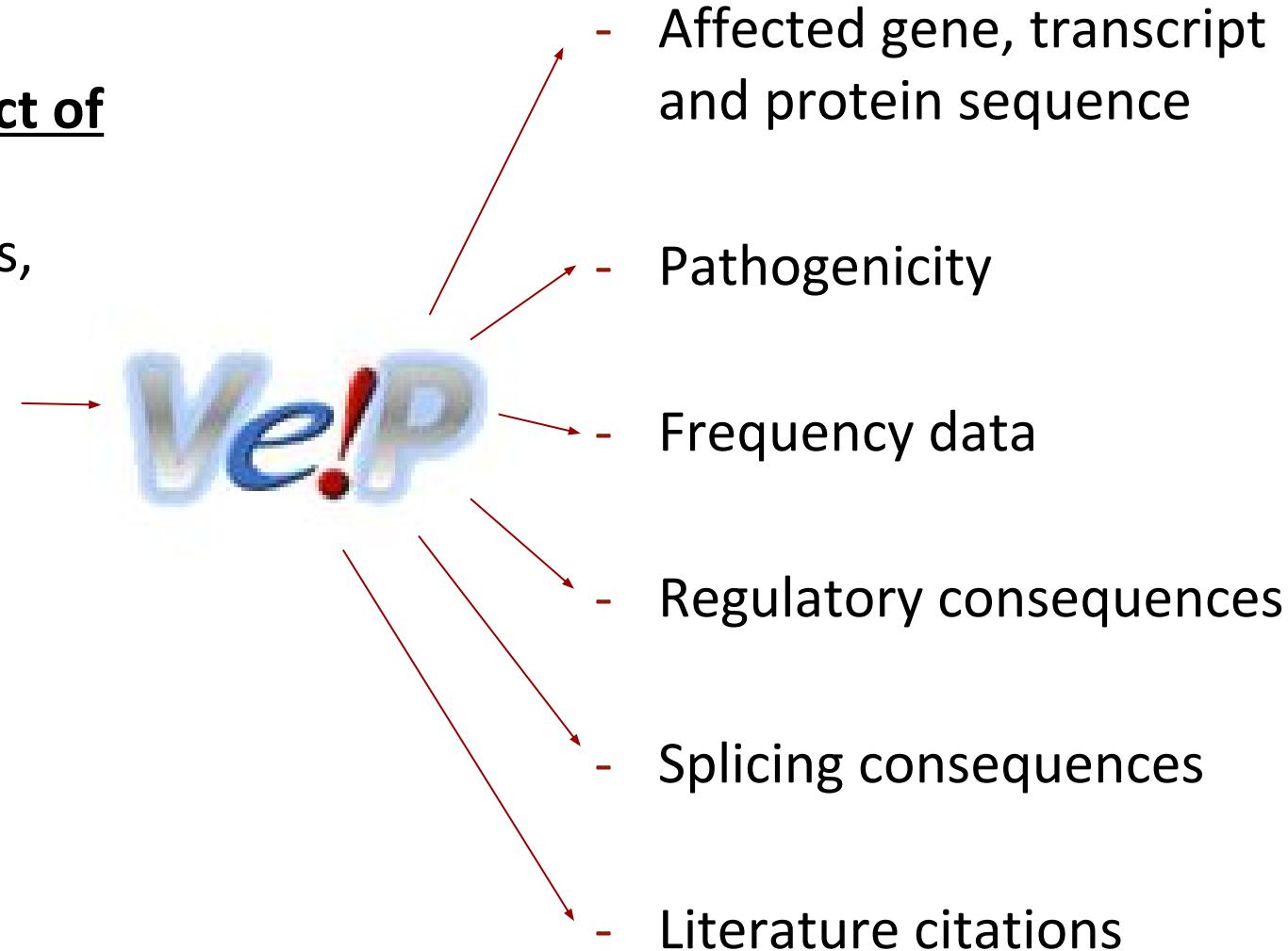
Hands on

- We're going to look at a set of six *Homo sapiens* variants *rs333*, *rs334*, *rs344*, *rs1800413*, *rs74653330* and *rs137854567* and find out:
 - Their location
 - Their alleles
 - Their MAF
 - Their phenotype associations
 - Their flanking sequences
- Demo: coursebook page 24-28

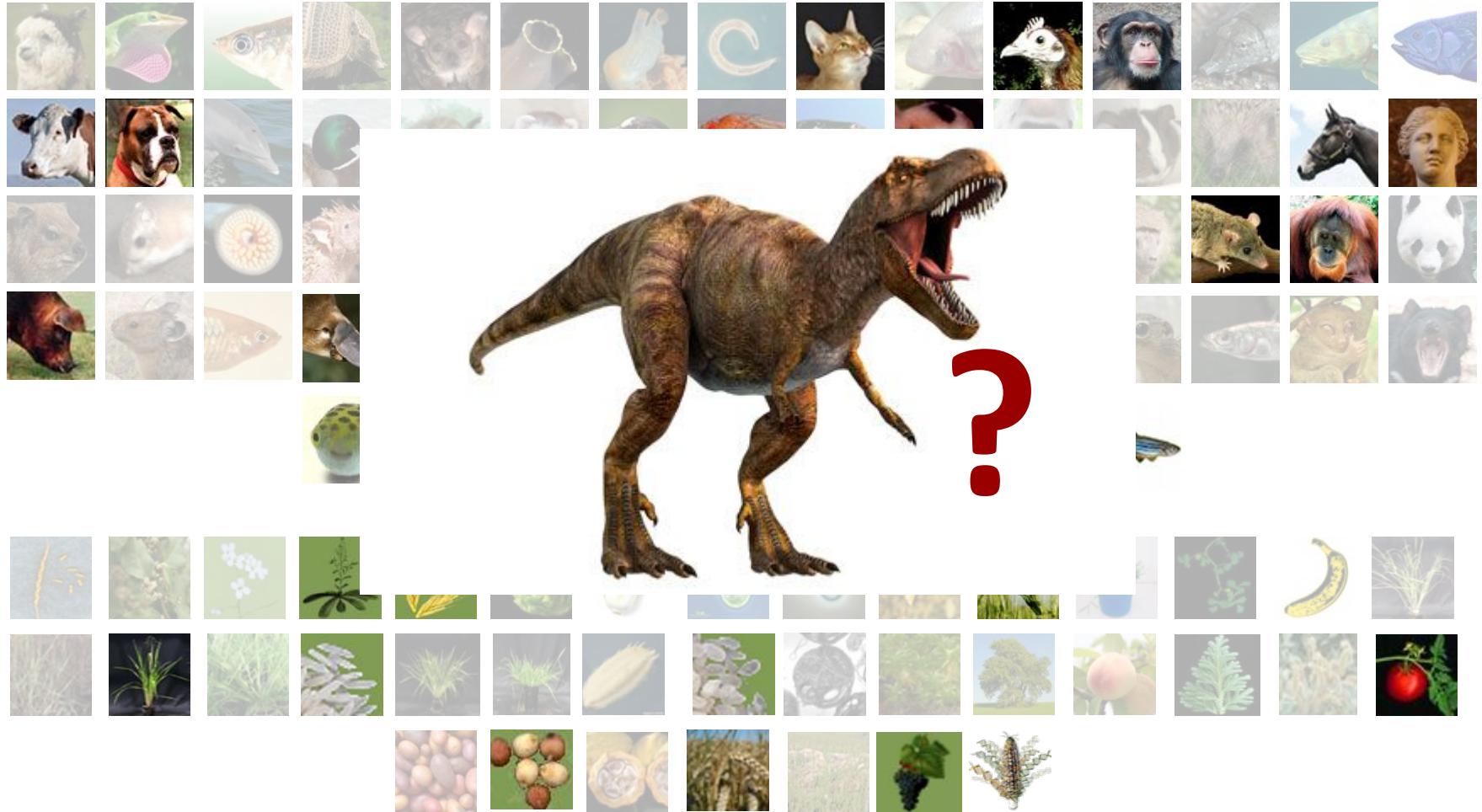
What is the VEP?

Determine the effect of variants (SNPs, insertions, deletions, CNVs or structural variants):

- Variant Co-ordinates
- VCF
- HGVS
- Variant IDs



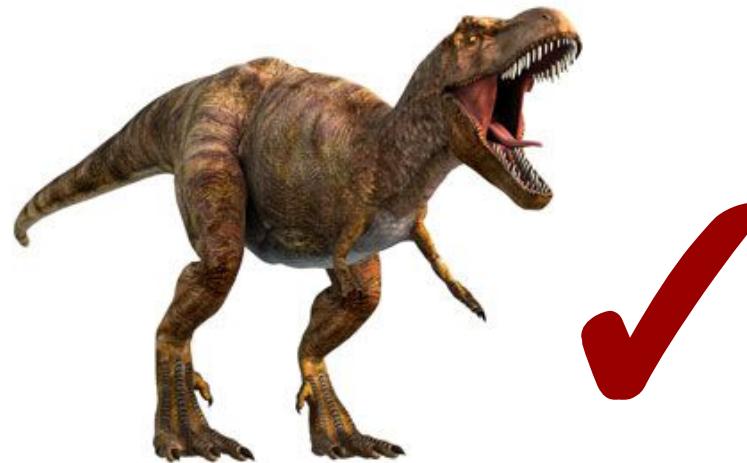
Species that work with the VEP



<http://www.ebi.ac.uk/~bmoore/workshops/>

Set up a cache

- Speed up your VEP script with an offline cache.
- Use prebuilt caches for Ensembl species.
- Or make your own from GTF and FASTA files - even for genomes not in Ensembl.



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

<http://www.ebi.ac.uk/~bmoore/workshops/>

Use the VEP

Variant Effect Predictor



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](#)

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

<http://www.ebi.ac.uk/~bmoore/workshops/>

VEP plugins

- Plugins add extra functionality to the VEP
- They may extend, filter or manipulate the output of the VEP
- Plugins may make use of external data or code

<http://www.ebi.ac.uk/~bmoore/workshops/>

Hands on

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

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: coursebook page 29-33

Ensembl data through the Perl API

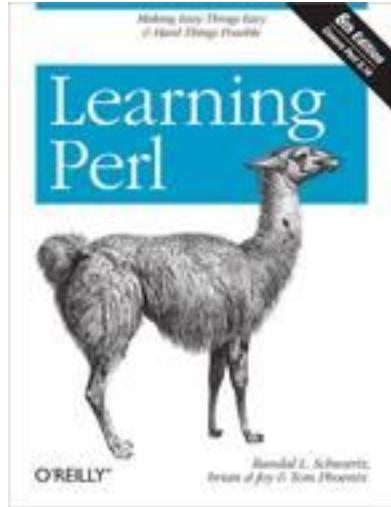
- Database querying using Perl scripts
- We use object-oriented Perl



```
my $gene_adaptor = $registry->  
    get_adaptor( 'human', 'core', 'gene' );  
  
my $gene = $gene_adaptor->  
    fetch_by_display_label( 'brca2' );  
  
print $gene->stable_id, "\n";
```

<http://www.ensembl.org/info/data/api.html>

Perl API



Learn Perl

Get out all possible
Ensembl data. ←
Output in any
format you like.

```
#!/usr/bin/perl
use strict;
use warnings;
use Bio::EnsEMBL::Registry;

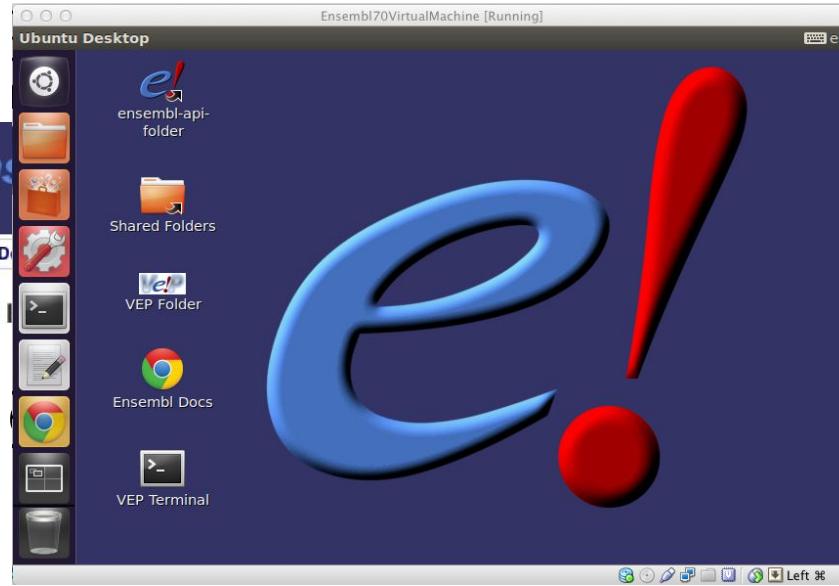
my $registry = "Bio::EnsEMBL::Registry";
$registry->load_registry_from_db(
    -host => 'ensemldb.ensembl.org',
    -user => 'anonymous'
);

my $transcript_adaptor = $registry -> get_adaptor('Mouse', 'Core', 'Transcript');
my @transcripts = @{$transcript_adaptor->fetch_all};

while (my $transcript = shift @transcripts) {
    my $refseqs = @{$transcript->get_all_object_xrefs('RefSeq')};

    if ($refseqs) {
        print $transcript->stable_id;
        while (my $refseq = shift @refseqs) {
            print " ", $refseq->display_id;
        }
        print "\n";
    }
}
```

Write scripts



AN
load more
dules)

public Listref Bio::EnsEMBL::Gene::get_all_Exons ()
Example :
my @exons = @{\$gene->get_all_Exons };
Description: Returns a set of all the exons associated with this gene.
Retruntype : Listref of Bio::EnsEMBL::Exon objects
Exceptions : none
Caller : general
Status : Stable
Code:
click to view

Learn Ensembl API

Running the VEP through the Perl API

- I want a script that gets a gene name from the command line and prints its sequence.
- We've already learnt how to use the API and know our way around the documentation
- We need to write a script.

Hands on

We have identified a number of human variants, which are contained in the VCF available at: www.ebi.ac.uk/~bmoore/workshops

We will use the **Standalone Perl script for VEP** to determine:

- What genes are affected by my variants?
- Do any of the variants affect protein structure/function?
 - Demo: coursebook page 34-37

Data access via REST

- We've had a Perl API for a long time ...
- ... but not everybody works in Perl
- Our RESTful service allows language agnostic access to our data.
- Visit rest.ensembl.org for installation, documentation and examples

What is REST?

- REST allows you to query the database using simple URLs giving output in plain text format

eg `http://rest.ensembl.org/xrefs/symbol/homo_sapiens/BRCA2?content-type=application/json`

gives

```
[ {"type":"gene", "id":"ENSG00000139618"}, {"type":"gene", "id":"LRG_293"} ]
```

- This means you can write scripts in any language to construct these URLs and read their output

Hands on

We have identified a single common variant in a cohort of patients with hypertension

Use grch37.rest.ensembl.org
for GRCh37

We will use the **REST API** to determine:

- What genes are affected by my variant?
- Does the variant affect protein structure/function?
 - Demo: coursebook page 38-40
 - Exercises: coursebook page 41
 - Answers: answer book page 6-7

Feedback survey

<https://www.surveymonkey.co.uk/r/Helsinki2016>

<http://www.ebi.ac.uk/bmoore/workshops/>

Wrap-up

Ensembl is a genome browser which integrates:

- gene annotation
- variation
 - The VEP
- comparative genomics
- regulation

<http://www.ebi.ac.uk/bmoore/workshops/>

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>

<http://www.ebi.ac.uk/bmoore/workshops/>

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

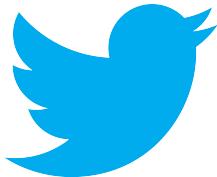
Ensembl public mailing lists dev@ensembl.org,
announce@ensembl.org

<http://www.ebi.ac.uk/bmoore/workshops/>

Follow us



www.facebook.com/Ensembl.org



@Ensembl



www.ensembl.info

<http://www.ebi.ac.uk/bmoore/workshops/>

Publications

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

Yates, A. *et al*

Ensembl 2016

Nucleic Acids Research

<http://nar.oxfordjournals.org/content/early/2015/12/19/nar.gkv1157.full>

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F

Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor

BMC Bioinformatics 26(16):2069-70(2010)

<http://bioinformatics.oxfordjournals.org/content/26/16/2069>

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

<http://www.ebi.ac.uk/bmoore/workshops/>

Ensembl 2016



<http://www.ebi.ac.uk/bmoore/workshops/>

Acknowledgements

The Entire Ensembl Team

Andrew Yates¹, Wasiu Akanni¹, M. Ridwan Amode¹, Daniel Barrell^{1,2}, Konstantinos Billis¹, Denise Carvalho-Silva¹, Carla Cummins¹, Peter Clapham², Stephen Fitzgerald¹, Laurent Gil¹, Carlos García Girón¹, Leo Gordon¹, Thibaut Hourlier¹, Sarah E. Hunt¹, Sophie H. Janacek¹, Nathan Johnson¹, Thomas Juettemann¹, Stephen Keenan¹, Ilias Lavidas¹, Fergal J. Martin¹, Thomas Maurel¹, William McLaren¹, Daniel N. Murphy¹, Rishi Nag¹, Michael Nuhn¹, Anne Parker¹, Mateus Patrício¹, Miguel Pignatelli¹, Matthew Rahtz², Harpreet Singh Riat¹, Daniel Sheppard¹, Kieron Taylor¹, Anja Thormann¹, Alessandro Vullo¹, Steven P. Wilder¹, Amonida Zadissa¹, Ewan Birney¹, Jennifer Harrow², Matthieu Muffato¹, Emily Perry¹, Magali Ruffier¹, Giulietta Spudich¹, Stephen J. Trevanion¹, Fiona Cunningham¹, Bronwen L. Aken¹, Daniel R. Zerbino¹ and Paul Flicek^{1,2,*}

¹European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK and ²Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Funding



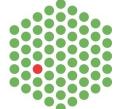
National
Human Genome
Research Institute



Centre for Therapeutic
Target Validation



epigenome



BioMedBridges



MedBioinformatics

Co-funded by the
European Union

<http://www.ebi.ac.uk/bmoore/workshops/>



EMBL-EBI



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>

<http://www.ebi.ac.uk/bmoore/workshops/>