Browsing Genes and Genomes with Ensembl

Victoria Newman
Ensembl Outreach Officer
EMBL-EBI
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.
# This webinar course

<table>
<thead>
<tr>
<th>Date</th>
<th>Webinar topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>6th April</td>
<td>Introduction to Ensembl</td>
<td>Helen Sparrow</td>
</tr>
<tr>
<td>13th April</td>
<td>Ensembl genes</td>
<td>Emily Perry</td>
</tr>
<tr>
<td>20th April</td>
<td>Data export with BioMart</td>
<td>Victoria Newman</td>
</tr>
<tr>
<td>27th April</td>
<td>Variation data in Ensembl and the Ensembl VEP</td>
<td>Victoria Newman</td>
</tr>
<tr>
<td>4th May</td>
<td>Comparing genes and genomes with Ensembl Compara</td>
<td>Ben Moore</td>
</tr>
<tr>
<td>11th May</td>
<td>Finding features that regulate genes – the Ensembl Regulatory Build</td>
<td>Ben Moore</td>
</tr>
<tr>
<td>18th May</td>
<td>Uploading your data to Ensembl and advanced ways to access Ensembl data</td>
<td>Emily Perry</td>
</tr>
</tbody>
</table>
Structure

Presentation:
Variation data in Ensembl; introduction to the VEP

Demo:
Viewing variation data in the browser; using the VEP

Exercises:
On the train online course
Questions?

• We’ve muted all the mics
• Ask questions in the Chat box in the webinar interface
• My Ensembl colleagues will respond during the talk
• There’s no threading so please respond with @name

Ben Moore

Emily Perry
Course exercises

http://www.ebi.ac.uk/training/online/course/ensembl-browser-webinar-series-2016

This text will be replaced by a YouTube (link to YouKu too) video of the webinar and a pdf of the slides.

A link to exercises and their solutions will appear in the page hierarchy.

The “next page” will be the exercises.
Get help with the exercises

- Use the exercise solutions in the online course
- Join our Facebook group and discuss the exercises with everybody (see the online course for the link)
- Email us helpdesk@ensembl.org
EMBL-EBI is an Outstation of the European Molecular Biology Laboratory.
Outline

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

1) Small scale: affects one or a few nucleotides (≤50 bp)
   - Small insertions and deletions (DIPs or indels)
   - Single / multi nucleotide polymorphisms (SNPs, MNPs)

   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: affects genomic region (structural variation; >50 bp)
   - Copy number variations (CNVs) and large deletions/duplications
   - Insertions, translocations
Variation consequences
### Consequence terms

<table>
<thead>
<tr>
<th>SO term</th>
<th>SO description</th>
<th>SO accession</th>
<th>Old Ensembl term</th>
</tr>
</thead>
<tbody>
<tr>
<td>transcript_ablation</td>
<td>A feature alteration whereby the deleted region includes a transcript feature</td>
<td>SO:0001908</td>
<td>Transcript ablation</td>
</tr>
<tr>
<td>splice_donor_variant</td>
<td>A splice variant that changes the 2 base region at the 5' end of an intron</td>
<td>SO:0001572</td>
<td>Essential splice site</td>
</tr>
<tr>
<td>splice_acceptor_variant</td>
<td>A splice variant that changes the 2 base region at the 3' end of an intron</td>
<td>SO:0001573</td>
<td></td>
</tr>
<tr>
<td>stop_gained</td>
<td>A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript</td>
<td>SO:0001907</td>
<td>Stop gained</td>
</tr>
<tr>
<td>frameshift_variant</td>
<td>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</td>
<td>SO:0001904</td>
<td>Frameshift coding</td>
</tr>
<tr>
<td>stop_lost</td>
<td>A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript</td>
<td>SO:0001903</td>
<td>Stop lost</td>
</tr>
<tr>
<td>inframe_insertion</td>
<td>An inframe non-synonymous variant that inserts bases into the coding sequence</td>
<td>SO:0001902</td>
<td>Non-synonymous coding</td>
</tr>
<tr>
<td>inframe_deletion</td>
<td>An inframe non-synonymous variant that deletes bases from the coding sequence</td>
<td>SO:0001903</td>
<td></td>
</tr>
<tr>
<td>missense_variant</td>
<td>A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved</td>
<td>SO:0001900</td>
<td>Partial coding</td>
</tr>
<tr>
<td>transcript_amplification</td>
<td>A feature amplification of a region containing a transcript</td>
<td>SO:0001902</td>
<td>Transcript amplification</td>
</tr>
<tr>
<td>splice_region_variant</td>
<td>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 2-6 bases of the intron</td>
<td>SO:0001900</td>
<td>Splice site</td>
</tr>
<tr>
<td>incomplete_terminal_codon_variant</td>
<td>A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed</td>
<td>SO:0001905</td>
<td>Incomplete coding</td>
</tr>
<tr>
<td>synonymous_variant</td>
<td>A sequence variant where there is no resulting change to the encoded amino acid</td>
<td>SO:0001908</td>
<td>Synonymous coding</td>
</tr>
<tr>
<td>stop_related_variant</td>
<td>A sequence variant where at least one base in the terminator codon is changed, but the terminator remains</td>
<td>SO:0001907</td>
<td></td>
</tr>
<tr>
<td>coding_sequence_variant</td>
<td>A sequence variant that changes the coding sequence</td>
<td>SO:0001908</td>
<td>Coding unknown</td>
</tr>
<tr>
<td>mature_mRNA_variant</td>
<td>A transcript variant located with the sequence of the mature mRNA</td>
<td>SO:0001902</td>
<td>Within mature mRNA</td>
</tr>
<tr>
<td>3_prime_UTR_variant</td>
<td>A UTR variant of the 3' UTR</td>
<td>SO:0001904</td>
<td>3'prime UTR</td>
</tr>
<tr>
<td>3_prime_UTR_variant</td>
<td>A UTR variant of the 3' UTR</td>
<td>SO:0001902</td>
<td>3'prime UTR</td>
</tr>
<tr>
<td>interon_variant</td>
<td>A transcript variant occurring within an intron</td>
<td>SO:0001907</td>
<td>Intron</td>
</tr>
<tr>
<td>NMD_transcript_variant</td>
<td>A variant in a transcript that is the target of NMD</td>
<td>SO:0001902</td>
<td>NMD transcript</td>
</tr>
<tr>
<td>non_coding_exon_variant</td>
<td>A sequence variant that changes non-coding exon sequence</td>
<td>SO:0001902</td>
<td>Within non-coding gene</td>
</tr>
<tr>
<td>nc_transcript_variant</td>
<td>A transcript variant of a non-coding RNA</td>
<td>SO:0001905</td>
<td></td>
</tr>
<tr>
<td>upstream_gene_variant</td>
<td>A sequence variant located 5' of a gene</td>
<td>SO:0001901</td>
<td>Upstream</td>
</tr>
<tr>
<td>downsteam_gene_variant</td>
<td>A sequence variant located 3' of a gene</td>
<td>SO:0001902</td>
<td>Downstream</td>
</tr>
<tr>
<td>TFBS_ablation</td>
<td>A feature alteration whereby the deleted region includes a transcription factor binding site</td>
<td>SO:0001903</td>
<td></td>
</tr>
<tr>
<td>TFBS_amplification</td>
<td>A feature amplification of a region containing a transcription factor binding site</td>
<td>SO:0001902</td>
<td></td>
</tr>
<tr>
<td>regulatory_region_variant</td>
<td>A sequence variant located within a regulatory region</td>
<td>SO:0001902</td>
<td>Regulatory region</td>
</tr>
<tr>
<td>regulatory_region_ablation</td>
<td>A feature alteration whereby the deleted region includes a regulatory region</td>
<td>SO:0001902</td>
<td>Regulatory region deletion</td>
</tr>
<tr>
<td>regulatory_region_amplification</td>
<td>A feature amplification of a region containing a regulatory region</td>
<td>SO:0001902</td>
<td>Regulatory region amplification</td>
</tr>
<tr>
<td>feature_elongation</td>
<td>A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence</td>
<td>SO:0001907</td>
<td>Feature elongation</td>
</tr>
<tr>
<td>feature_translocation</td>
<td>A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence</td>
<td>SO:0001900</td>
<td>Feature translocation</td>
</tr>
<tr>
<td>intergenic_variant</td>
<td>A sequence variant located in the intergenic region, between genes</td>
<td>SO:0001908</td>
<td>Intergenic</td>
</tr>
</tbody>
</table>

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

Kumar, P, et al., 2009, *Nature Protocols*
Adzhubei, I, et al., 2010, *Nature Methods*
Missense variants – pathogenicity

**SIFT**
- 1: Tolerated
- 0.05: Deleterious
- 0: Deleterious

**PolyPhen**
- 1: Probably damaging
- 0.2: Possibly damaging
- 0.1: Benign
- 0: Benign
Species with variation data

+ Ensembl Plants, Fungi, Protists, and Metazoa
http://www.ensembl.org/info/genome/variation/sources_documentation.html
Variation sources

http://www.ensembl.org/info/docs/variation/sources_documentation.html
http://www.ensembl.org/info/genome/variation/sources_phenotype_documentation.html
HapMap project

Genotyping arrays to assess variant frequency in 1,301 individuals from 11 populations
1000 Genomes Project

Sequencing to assess variant frequency in 2,504 individuals from 26 populations at 4X coverage

http://www.ensembl.org/Help/Faq?id=328
Reference alleles

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

AGTCGTA^T^GCGA
Allele strand

AGTCGTAGCTAGC T/G GAGGCCATAGGCGA
TGGCGCATAGGCGAT
AGGCCCATAGGCGAT

Exon sequence:
TATGGCCT A/C CGCTAGC

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

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

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

- We’re going to look at the gene *MCM6* and explore its variants.
- We will look at the region surrounding *MCM6* to find variants as well.
- We will look at the variant *rs4988235* to find more information about it.
What is the VEP?

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

- Variant coordinates
- VCF
- HGVS
- Variant IDs

Affected gene, transcript, and protein sequence

Pathogenicity

Frequency data

Regulatory consequences

Splicing consequences

Literature citations

http://www.ensembl.org/info/docs/tools/vep/script/index.html
Species that work with the VEP
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
Use the VEP

http://www.ensembl.org/info/docs/tools/vep/index.html
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
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
- The genes affected by my variants
- Whether any of my variants affect gene regulation
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/hmoore/workshops/
Publications

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

Aken, B. et al.
Ensembl 2017
Nucleic Acids Research
http://europepmc.org/abstract/med/27899575

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
www.biomedcentral.com/1471-2164/11/295
Acknowledgements

The Entire Ensembl Team

Bronwen L. Aken¹, Premanand Achuthan¹, Wasiu Akanni¹, M. Ridwan Amode¹, Friederike Bernsdorf¹, Jyothish Bhai¹, Konstantinos Billis¹, Denise Carvalho-Silva¹, Carla Cummins¹, Peter Clapham², Laurent Gil¹, Carlos García Girón¹, Leo Gordon¹, Thibaut Hourlier¹, Sarah E. Hunt¹, Sophie H. Janacek¹, Thomas Juettemann¹, Stephen Keenan¹, Matthew R. Laird¹, Ilias Lavidas¹, Thomas Maurel¹, William McLaren¹, Benjamin Moore¹, Daniel N. Murphy¹, Rishi Nag¹, Victoria Newman¹, Michael Nuhn¹, Chuang Kee Ong¹, Anne Parker¹, Mateus Patricio¹, Harpreet Singh Riat¹, Daniel Sheppard¹, Helen Sparrow¹, Kieron Taylor¹, Anja Thormann¹, Alessandro Vullo¹, Brandon Walts¹, Steven P. Wilder¹, Amonida Zadissa¹, Myrto Kostadima¹, Fergal J. Martin¹, Matthieu Muffato¹, Emily Perry¹, Magali Ruffier¹, Daniel M. Staines¹, Stephen J. Trevanion¹, Fiona Cunningham¹, Andrew Yates¹, Daniel R. Zerbino¹ and Paul Flicek¹,²,*

http://www.ebi.ac.uk/bmoore/workshops/
Training materials

- Ensembl training materials are protected by a CC BY license
  ![CC BY]

- 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