EBI is an Outstation of the European Molecular Biology Laboratory.

Genes and Transcripts
# This webinar course

<table>
<thead>
<tr>
<th>Date</th>
<th>Webinar topic</th>
<th>Instructor</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4th Sept</td>
<td>Introduction to Ensembl</td>
<td>Astrid Gall</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ensembl genes</td>
<td>Emily Perry</td>
<td></td>
</tr>
<tr>
<td>6th Sept</td>
<td>Variation data in Ensembl and the Ensembl VEP</td>
<td>Erin Haskell</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comparing genes and genomes with Ensembl Compara</td>
<td>Astrid Gall</td>
<td></td>
</tr>
<tr>
<td>11th Sept</td>
<td>Finding features that regulate genes – the Ensembl Regulatory Build</td>
<td>Emily Perry</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data export with BioMart</td>
<td>Erin Haskell</td>
<td></td>
</tr>
<tr>
<td>13th Sept</td>
<td>Uploading your data to Ensembl</td>
<td>Astrid Gall</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Introduction to the Ensembl REST APIs</td>
<td>Emily Perry</td>
<td></td>
</tr>
</tbody>
</table>
Structure

Presentation: Where Ensembl genes come from

Demo: Getting gene data

Exercises: On the train online course
Gene views

- Coding exon
- Intron
- Non-coding exon
- Merged transcript
- Protein coding transcript
- Non-coding transcript
Golden transcripts

- Identical annotation
- Higher confidence and quality
Ensembl and Havana annotation

Automatic annotation

Manual annotation
Automatic gene annotation

- Genome-wide determination using the Ensembl automated pipeline
- Predictions based on experimental (biological) data
- Known proteins/cDNAs plotted onto the genome using sequence matching
- Ab initio prediction of genes based on ORFs
Biological Evidence

- International Nucleotide Sequence databases
  - cDNAs
  - ESTs
  - RNA-seq

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

- Infer genes from homology to other species
  - Eg predict genes in [species] by mapping cDNAs/proteins from [other species] to the [species] genome
Manual gene annotation

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

• \textbf{havana}

• Genome-wide

• Uses data from databases and papers
Manual vs automatic

• Manual annotation is more comprehensive
  • More transcripts per gene, especially non-coding transcripts
  • More genes overall, especially non-coding genes
• Manually annotated transcripts may have less evidence supporting them
  • Evidence is evaluated individually by quality not just thresholds
• Manual annotation can be more accurate for difficult to annotate features such as:
  • UTRs
  • Splice sites
  • Single exon transcripts
  • Exceptions, such as immunoglobulins
GENCODE

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

- GENCODE is the default gene set used by ENCODE, 1000 genomes and other major projects.
CCDS transcripts

- Consensus **coding** DNA sequence set
- Agreement between EBI, UCSC and NCBI
Imported annotation

- We have imported annotation for a small number of species from external sources
- Quality ensured by:
  - Our rigorous QC
  - Trusted sources

Chinese hamster ovary

Ryukyu mouse

Shrew mouse
Which transcript to use?

- GENCODE Basic: Only the “complete” transcripts (where a gene has complete transcripts) (http://www.ensembl.org/Help/Glossary?id=500)

- Transcript support level: Scored 1-5 for quality, where 1 is the best (http://www.ensembl.org/Help/Glossary?id=492)

- APPRIS principal isoform: The major isoform(s) from combining protein structural information, functionally important residues and evidence from cross-species alignments. (http://www.ensembl.org/Help/Glossary?id=521)

- + CCDS, + Golden transcripts
Ensembl stable IDs

- ENSG############ Ensembl Gene ID
- ENST############ Ensembl Transcript ID
- ENSP############ Ensembl Peptide ID
- ENSE############ Ensembl Exon ID

- For non-human species a suffix is added:
  MUS (*Mus musculus*) for mouse ENSMUSG###
  DAR (*Danio rerio*) for zebrafish: ENSDARG###

http://www.ensembl.org/info/genome/stable_ids/index.html
Why Gene Ontology (GO)?

Multiple terms for the same thing

- Innate immunity
- Non-specific immunity

Gene descriptions too specific

- Complement
- Cytokines
- Phagocyte
- Mast cells
- Natural killer cells
GO terms form a controlled vocabulary

**GO:0045087** - innate immune response

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

- GO:0006955 immune response
- GO:0045087 innate immune response
- GO:0035420 MAPK cascade involved in innate immune response
  - GO:0001867 complement activation, lectin pathway
  - GO:0035006 melanisation defence response
  - GO:0009626 positive reg of innate immune response
  - GO:0009626 plant-type hypersensitive response
  - GO:0009626 negative reg of innate immune response
  - GO:0002228 natural killer cell mediated immunity
  - GO:0002228 response to type I interferon
  - GO:0002228 response to type II interferon
  - GO:0009616 virus induced gene silencing
  - GO:0009616 response to interferon-gamma
  - GO:0034341 regulation of innate immune response
  - GO:0009682 induced systemic resistance
  - GO:0009814 defence response, incompatible interaction

- GO:0006957 complement activation, alternative pathway
- GO:0042381 hemolymph coagulation
- GO:0034340 response to type I interferon
- GO:0045089 positive reg of innate immune response
More information

The Ensembl gene annotation system.
Aken et al. 2016
http://europepmc.org/articles/PMC4919035
Hands on

- We’re going to look at an Ensembl gene, *ESPN*, and find out information about it and its transcripts.
- Demo: coursebook page 6-11
- Exercises: coursebook page 11-12
  - Answers: answer book page
Questions

- We’ve muted all the mics
- Ask questions in the Slack workspace

Astrid Gall
Erin Haskell
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 Slack workspace and discuss the exercises with everybody in dedicated channels (register to get sent a link)
• Email us helpdesk@ensembl.org
<table>
<thead>
<tr>
<th>Date</th>
<th>Webinar topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>4th Sept</td>
<td>Introduction to Ensembl</td>
<td>Astrid Gall</td>
</tr>
<tr>
<td></td>
<td>Ensembl genes</td>
<td>Emily Perry</td>
</tr>
<tr>
<td>6th Sept</td>
<td>Variation data in Ensembl and the Ensembl VEP</td>
<td>Erin Haskell</td>
</tr>
<tr>
<td></td>
<td>Comparing genes and genomes with Ensembl Compara</td>
<td>Astrid Gall</td>
</tr>
<tr>
<td>11th</td>
<td>Finding features that regulate genes – the Ensembl Regulatory Build</td>
<td>Emily Perry</td>
</tr>
<tr>
<td>Sept</td>
<td></td>
<td>Erin Haskell</td>
</tr>
<tr>
<td></td>
<td>Data export with BioMart</td>
<td></td>
</tr>
<tr>
<td>13th</td>
<td>Uploading your data to Ensembl</td>
<td>Astrid Gall</td>
</tr>
<tr>
<td>Sept</td>
<td>Introduction to the Ensembl REST APIs</td>
<td>Emily Perry</td>
</tr>
</tbody>
</table>
Next webinar – Variation data in Ensembl and the Ensembl VEP

Ensembl imports variation and phenotype data from a number of sources. Learn how to find variants in genes and regions, and access additional information, including population frequencies.

We will then introduce the Variant Effect Predictor (VEP), a tool which allows you to analyse your own variation data for potential effects on genes.

Thursday 6th September 3pm BST

Erin Haskell
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 2018
Zerbino et al
http://europepmc.org/abstract/MED/29155950

Topic-specific publications mentioned throughout workshop
Ensembl 2018
Ensembl Acknowledgements

The Entire Ensembl Team

Daniel R. Zerbino¹, Premanand Achuthan¹, Wasiu Akanni¹, M. Ridwan Amode¹, Daniel Barrell¹,², Jyothish Bhai¹, Konstantinos Billis¹, Carla Cummins¹, Astrid Gall¹, Carlos García Giroń¹, Laurent Gil¹, Leo Gordon¹, Leanne Haggerty¹, Erin Haskell¹, Thibaut Hourlier¹, Osagie G. Izuogu¹, Sophie H. Janacek¹, Thomas Juettemann¹, Jimmy Kiang To¹, Matthew R. Laird¹, Ilia Lavidas¹, Zhicheng Liu¹, Jane E. Loveland¹, Thomas Maurel¹, William McLaren¹, Benjamin Moore¹, Jonathan Mudge¹, Daniel N. Murphy¹, Victoria Newman¹, Michael Nuhn¹, Denye Ogeh¹, Chuang Kee Ong¹, Anne Parker¹, Mateus Patricio¹, Harpreet Singh Riat¹, Helen Schuilenburg¹, Dan Sheppard¹, Helen Sparrow¹, Kieron Taylor¹, Anja Thomann¹, Alessandro Vullo¹, Brandon Walts¹, Amonida Zadissa¹, Adam Frankish¹, Sarah E. Hunt¹, Myrto Kostadima¹, Nicholas Langridge¹, Fergal J. Martin¹, Matthieu Muffato¹, Emily Perry¹, Magali Ruffier¹, Dan M. Staines¹, Stephen J. Trevanion¹, Bronwen L. Aken¹, Fiona Cunningham¹, Andrew Yates¹ and Paul Flicek¹,³

¹European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, ²Eagle Genomics Ltd., Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, UK and ³Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Funding

Co-funded by the European Union