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
Exploring Biological Sequences with Ensembl

Ben Moore
Ensembl Outreach
EMBL-EBI
9th October 2019
Structure for today

- Introduction to Ensembl
- Retrieving genomic sequences using the Ensembl browser
  
  coffee/tea

- Exporting genomic sequences using BioMart
- Exporting genomic sequences using the REST API
- Exporting genomic sequences from the FTP site
Course materials

training.ensembl.org

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

- What type of sequence data can you get in Ensembl?
- How to retrieve sequence data using the Ensembl browser website.
- How to retrieve sequence data from Ensembl on a larger scale.
- Where to go for help and documentation.
Questions?
The Ensembl genome browser

EMBL-EBI
Why do we need genome browsers?

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

http://training.ensembl.org/events
Why do we need genome browsers?

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

- Gene builds for ~200 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://training.ensembl.org/events
Ensembl - access to 200+ genomes

http://training.ensembl.org/events
Ensembl Genomes- expanding Ensembl

www.ensembl.org

- Vertebrates

- Other representative species

http://training.ensembl.org/events
What is a genome assembly?

Genome
“DNA within a cell”

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

http://training.ensembl.org/events
What is a genome assembly?

Sequence reads

CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA
CAGCTGTCCCAGATGAC
ACTTAACTTCCCTCCCAGCTGTCC
GGGCTCCGCCTTCAGCTCT
AACTTCCCTCCCAGCT
TCCCAGCTGTCCCAGATGACGCCT
TCCGCCTTTCAGCTCAAGACTTAACTTC
CAGATGACGCC
TCCCAGCTGTCCCAGATGACGCCT
CAGATGACGCC

Match up overlaps

CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA
AACTTCCCTCCCAGCT
CAGATGACGCC
TCCCAGCTGTCCCAGATGACGCCCT
TCCCCAGCTGTCCCAGATGACGCCCT
GGGCTCCGCCTTCAGCTCT
AACTTCCCTCCCAGCT
TCCCAGCTGTCCCAGATGACGCCCT
AACTTCCCTCCCAGCT
CAGCTGTCCCAGATGACGCCCT
TCCCAGCTGTCCCAGATGACGCCCT
CAGATGACGCC

Genome assembly

CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACGCCCT
http://training.ensembl.org/events
Ensembl gene annotation

Automatic annotation

Manual annotation

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

Coding exon | Intron | Non-coding exon

Merged transcript
Protein coding transcript
Non-coding transcript

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

- We’re going to look at a region of the human genome, 4:122868000-122946000, and export the genomic sequence of this region.
- We will also search for an Ensembl gene, UQCRQ, and find out sequence data about it and its transcripts.

- Demo: Coursebook page 6-22
- Exercises: Coursebook page 23-24
  - Answers: Answer book page 3-6

http://training.ensembl.org/events
Data Mining with BioMart
What is BioMart?

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

Ensembl
Ensembl Plants
Ensembl Fungi (some exceptions)
Ensembl Metazoa
Ensembl Protists (some exceptions)
How do I use BioMart?
The 4 steps

Dataset
choose database & species

Filters
narrowing down the dataset

Attributes
what to print in your table

Results
table/sequences

http://training.ensembl.org/events
Step 1: Dataset

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

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

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

Step 2: Filters

Dataset ➔ Filters ➔ Attributes ➔ Results

Get Attributes for these

http://training.ensembl.org/events
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://training.ensembl.org/events
Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data using R statistical programming language.

- Package for Biomart called BiomaRt:

- Easy to install in R:
  - biocLite("biomaRt")

Hands on

- We’re going to look at a set of six *Homo sapiens* genes *ESPN, MYH9, USH1C, CISD2, THR5* and *BRCA2* and find out:
  - Their Ensembl stable IDs
  - Their cDNA sequences
Hands on

- We’re going to look at a set of six *Homo sapiens* genes *ESPN, MYH9, USH1C, CISD2, THRBD, and BRCA2* and find out:
  - Their Ensembl stable IDs
  - Their cDNA sequences

- Demo: Coursebook page 25-28
- Exercises: Coursebook page 28-29
  - Answers: Answer book page 7-8
Hands on

- We’re going to export sequences programmatically using the Ensembl REST API.
  - Demo: Coursebook page 30-32

- We will also export reference sequence files from the Ensembl FTP site and using the File chameleon tool.
  - Demo: Coursebook page 33-36
How is all this data organised?

- Ensembl browser sites
  - Main website, Archive!
- BioMart ‘DataMining tool’
- Ensembl Database (open source)
  - Perl-API, REST API, MySQL
- FTP download site
Help and documentation

Course online
https://www.ebi.ac.uk/training/online/course-list
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

Cunningham, F. et al
**Ensembl 2019**
Nucleic Acids Research
https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky1113/5165265

Kersey, PJ et al.
**Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species**
Nucleic Acids Research (Database Issue)
https://doi.org/10.1093/nar/gkx1011

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

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

The Entire Ensembl Team


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

Funding

Co-funded by

the European Union

EBI is an Outstation of the European Molecular Biology Laboratory.
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