Ensembl and ENA
High level overview and use cases

Denise Carvalho-Silva
Ensembl Outreach Team

On behalf of Ensembl and ENA teams
European Molecular Biology Laboratories
European Bioinformatics Institute

SME Bioinformatics Forum
Barcelona 8-9 October 2012
Outline

Ensembl project: background and goals

Data available

Data access and Ensembl tools

Use cases: Ensembl and ENA

Ensembl Outreach and Support

Acknowledgements
Ensembl project

Launched in 1999: before the release of the draft of the human genome

Joint project between the EBI and WTSI

www.ensembl.org

launched in March 2000
Goals

Provide comprehensive annotation of genomes

Integrate the annotation with other biological data

Make them all publicly available
Annotation of non-vertebrate genomes

Extends the use of Ensembl to other species
Wider taxonomic range (v15, 354 genomes)

www.ensemblgenomes.org

launched in 2009
Data available in Ensembl 68

- Gene annotation for 66 vertebrate species
- Variation data for 19 species
- Comparative Genomics data for 69 species
- Regulation data for 16 species
Data access: browser sites

www.ensembl.org
pre.ensembl.org
archive.ensembl.org
Data access: BioMart

- web interface to export Ensembl data
- no programming skills required

www.ensembl.org/biomart/martview
**BioMart results**

### Dataset 1 / 54345 Genes
Homo sapiens genes (GRCh37.p5)

#### Filters
- HGNC symbol(s) [e.g. ZFY]: [ID-list specified]

#### Attributes
- Ensembl Gene ID
- Ensembl Transcript ID
- Ensembl Protein ID
- Chromosome Name
- Gene Start (bp)
- MIM Morbid Description

#### Export all results to
- File
- TSV
- [Unique results only]

#### Email notification to

#### View
- 10 rows as HTML
- [Unique results only]

<table>
<thead>
<tr>
<th>Ensembl Gene ID</th>
<th>Ensembl Transcript ID</th>
<th>Ensembl Protein ID</th>
<th>Chromosome</th>
<th>Gene Start (bp)</th>
<th>MIM Morbid Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENST00000001628</td>
<td>ENST000000046805</td>
<td></td>
<td>7</td>
<td>117109383</td>
<td>VAS DEFERRIS, CONGENITAL BILATERAL APLASIA OF; CBAVD</td>
</tr>
<tr>
<td>ENST00000001628</td>
<td>ENST000000046805</td>
<td></td>
<td>7</td>
<td>117109383</td>
<td>CYSTIC FIBROSIS; CF</td>
</tr>
<tr>
<td>ENST00000001628</td>
<td>ENST000000046805</td>
<td></td>
<td>7</td>
<td>117109383</td>
<td>BRONCHECTASIS WITH OR WITHOUT ELEVATED SWEAT CHLORIDE 1; BESC1</td>
</tr>
<tr>
<td>ENST00000001628</td>
<td>ENST000000046805</td>
<td></td>
<td>7</td>
<td>117109383</td>
<td>PANCREATITIS, HEREDITARY, PCTT</td>
</tr>
</tbody>
</table>

**Tables/sequences Export/email**
Data access: APIs and FTP

• Ensembl Database (open source): Perl-API, MySQL

http://www.ensembl.org/info/docs/api/index.html

• FTP download site

http://www.ensembl.org/info/data/ftp/index.html
Ensembl Tools

We provide a number of ready-made tools for processing your data. At the moment, small datasets can be uploaded to our servers and processed online; for larger datasets, we provide an API script that can be downloaded (you will also need to install our Perl API to use these).

In the near future we aim to offer an intermediate service, whereby medium-to-large data sets can be submitted to a queue, similar to BLAST.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Online tool</th>
<th>Download code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly converter</td>
<td>Map your data to the current assembly. Accepted file formats: GFF, GTF, BED, PSL. N.B. Export is currently in GFF only</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID History converter</td>
<td>Convert a set of Ensembl IDs from a previous release into their current equivalents (max 30 IDs).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region Report</td>
<td>Export standard data sets (genes, sequence, variations, etc) from one or more regions, in either GFF or simple text format (max 5MB total, or 1MB for variation/regulation data).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variant Effect Predictor</td>
<td>(Formerly SNP Effect Predictor). Analyse your own variants, the functional consequences of known and unknown variants.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual Machine</td>
<td>VirtualBox virtual Machine with Ubuntu desktop and pre-compiled the latest Ensembl API plus Variant Effect Predictor (VEP). Note, download is &gt;1 GB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assembly converter
ID history converter
Virtual Machine
Region Report
Variant Effect Predictor

http://www.ensembl.org/tools.html
Gene annotation

• Automatic pipeline
  Genome-wide determination

• Manual curation
  Gene determination on a case-by-case basis by an annotator

+ 63 species

+ gene lists 5 species
Gene annotation on the browser

Exons are drawn as boxes. Filled boxes are translated (coding) exons, empty boxes are untranslated regions (UTRs).

- Merged ("gold") gene set: identical annotation from Ensembl and Havana for human, mouse, zebrafish
- high confidence and quality
Biological Evidence

- International Nucleotide Sequence databases
  - ENA (European Nucleotide Archive)
  - GenBank
  - DDBJ (DNA Data Bank of Japan)
- Protein sequence databases
  - UniProt
- NCBI RefSeq
- RNAseq (transcriptomic) data
European Nucleotide Archive

ENA provides a comprehensive, accessible and publicly available repository for nucleotide sequence data

Data submission

Data search/download

http://www.ebi.ac.uk/ena/
Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary.

Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany.

We also obtained the complete mtDNA from the extinct European cave bear using a 44,000 year old bone found in Gamssulzen Cave, Austria. Again, we used a 2-step multiplex approach, but in this case, all PCR products were cloned and multiple clones were sequenced (EMBL:FM177760). Moreover, to ensure sequence accu-

Retrieve and browse the mitochondrial genome of the cave bear (*Ursus spelaeus*).
I have submitted a DNA sequence to ENA and got the ID AF489725. Can I view this ID in Ensembl?

- Which gene is associated with?
- Which chromosome is the gene found on?
- What are the neighbouring genes?
- Is there a homologue to this gene in dog?
- Find the cDNA alignment between the two genes
- Can I jump to ENA from Ensembl?
Our sequencing results identified a known SNP (rs4988235) in one of our samples in individuals from Barcelona (Spain).

- What is the major allele for this SNP? Is it the same in all 1000 Genomes super-populations?
- What is the ancestral allele? Is it conserved in vertebrates?
- Are there any phenotypes associated with this SNP?
- How many variants are associated with this phenotype?
- Which gene is associated to this phenotype?
Ensembl Outreach and Support

- Course online www.ensembl.info/ecourse
- Tutorials www.ensembl.org/info/website/tutorials
- YouTube channel www.youtube.com/user/EnsemblHelpdesk
- Mailing lists announce@ensembl.org, dev@ensembl.org
- Comments and questions? helpdesk@ensembl.org
Acknowledgements

Funded by the Wellcome Trust, NIH-NHGRI, EU and EMBL

Funded by EMBL, EU, Wellcome Trust, BBSRC

datasubs@ebi.ac.uk
update@ebi.ac.uk