ENSEMBL

Nociones Básicas de Bioinformática y Genómica

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2015-03-30

Ensembl

http://www.ensembl.org/

Joint project:

- EMBL - EBI and the
- Wellcome Trust Sanger Institute
- started in 1999 (Human Genome Project)

General repository for annotated genomes
(mainly **vertebrates** / eukaryotic)

- imports genome sequences/assemblies from diverse consortia
- **automatically** filter and annotate genome sequences
- integrate these data with other biological information
- make the results freely available to the community

Basic genome annotations

- Genes
- Genomic location
- Gene model structures
- Exons
- Introns
- UTRs
- Transcript(s)
- Pseudogenes
- Non-coding RNA
- Protein(s)
- Links to other sources of information: NCBI, Gene Ontology, ...

Ensembl Own (stable) IDs

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

Human

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

Other species

A suffix is added

- MUS (Mus musculus) : ENSMUSG###
- DAR (Danio rerio) : ENSDARG###

Species

A list of the currently available species:

- http://www.ensembl.org/info/about/species.html

Each species has its descriptive page

- http://www.ensembl.org/Homo_sapiens/Info/Index
- http://www.ensembl.org/Mus_musculus/Info/Index
- http://www.ensembl.org/Danio_rerio/Info/Index

Information about the Genome build and annotation in the links:

- More information and statistics

Other Species

Pre-Ensembl

The Ensembl pre-build site: genomes that are in the process of being annotated.

http://pre.ensembl.org/index.html

Ensembl Genomes

Species other than vertebrates:

- Ensembl Bacteria
- Ensembl Fungi
- Ensembl Metazoa
- Ensembl Plants

http://ensemblgenomes.org/

Ensembl Databases Internal Organization

- Genes: main database
  - transcripts
  - exons
  - ...
- Variations:
  - germinal: heritable (wikipedia)
  - somatic: acquired mutation (wikipedia)
- Regulation:
- VEGA
- PRIDE

Ensembl Databases: Ensembl Variation

www.ensembl.org/info/genome/variation/index.html
All possible variations here

Sequence variants (short)

- SNP
- Insertion (short)
- Deletion (short)
- InDel (short): an insertion **and** a deletion
- Substitution

Structural variants (long)

- CNV: Copy Number Variation
- Inversion
- Translocation

Ensembl Databases: Regulation

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

Regulation in different cell types for:

- human
- mouse
- drosohpila

Ensembl Databases: Vega

http://vega.sanger.ac.uk

- A repository for high-quality gene models produced by the manual annotation of vertebrate genomes.
- Sanger dependent

Havana

- The Havana team is a subset of Vega
- provides the **manual** annotation of human, mouse, zebrafish and other vertebrate genomes

see http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/

Ensembl Databases: PRIDE

PRIDE: PRoteomics IDEntifications

PRIDE Archive: proteomics data repository

http://www.ebi.ac.uk/pride/archive/

- public data repository for proteomics data
- EBI
- protein and peptide identifications
- post-translational modifications

Ensembl Utilities

Form the web site menu:

- BLAST/BLAT: alignment
- **BioMart**: data-mining tool
- *Tools*
- **Downloads**: FTP
- Help & Documentation: nice glossary *(see ambiguity code)*
- *Blog*
- *Mirrors*

Ensembl FTP

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

Annotation File Formats

- BED
- GFF/GTF : General Feature Format / General Transfer Format

Checksum

Checksum files are available in most FTP directories. See:

- http://en.wikipedia.org/wiki/Checksum
- http://en.wikipedia.org/wiki/Md5sum

Ensembl Biomart

http://www.ensembl.org/biomart/

What is BioMart

http://www.biomart.org/

- A federated database system that provides unified access to disparate, geographically distributed data sources.
- Any existing databases can easily be incorporated into the BioMart framework.
- It is designed to be data platform independent.
- Efficient. Ej. parallel query processing ...
- Unified technology.
- No need of programming: graphical user interfaces.

The BioMart project provides:

- software: server ... client
- **services**: already set up servers that provide access to a data

BioMart: Standardized Access to data

- Choose a Database. The repository.
- Dataset. For instance, Species.
- Select Attributes: IDs, descriptions, sequences
  Attributes are what we want to know about the genes.
- Set Filters: Indicate where our search should be restricted.

Examples:

- BioMart Central Portal
- EnsMart
- HapMap
- WormBase
- some Spanish

Links

http://www.ensembl.org/info/website/glossary.html