Browsing with Ensembl and Ensembl Genomes

Denise Carvalho-Silva
Ensembl Outreach Team
EMBL-EBI

http://www.ebi.ac.uk/~denise/Norwich_Mar_2012
# This morning

**9:30-12:45**

<table>
<thead>
<tr>
<th>Topic/time</th>
<th>09:30</th>
<th>10:00</th>
<th>10:30</th>
<th>11:00</th>
<th>11:15</th>
<th>11:50</th>
<th>12:40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walkthrough</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coffee Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hands-on (Browser)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biomart + hands on</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrap up (Summary)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- bla bla
- bli blo
- bla blu
Course Objectives

What is Ensembl?

What data can you get in Ensembl? How?

How to mine Ensembl data with BioMart?

Where to go for help and documentation?
What data are you interested in?

Write down 1-2 bullet points on a Post it note and put this up in front of the room.
Why do we need/have genome browsers?

1977: 1\textsuperscript{st} genome to be sequenced (5 kb)
2004: finished human sequence (3 gb)

Large amounts of raw DNA sequence data
Need for comprehensive annotation

http://www.ensembl.org

http://genome.ucsc.edu

Ensembl: an integration point

Ensembl is used worldwide

Top users

United Kingdom
United States
Canada
China
France
Germany
Italy
Japan
Spain
Annotation of vertebrate genomes

61 vertebrates*
D. melanogaster
C. elegans
S. cerevisae


*Release 65 Dec 2011
Annotation of non-vertebrate genomes

Extends the use of Ensembl to other species
Wider taxonomic range (v12, 335 genomes)

www.ensemblgenomes.org
- first release in 2000
- Vertebrates (fly, worm, yeast as outgroups)
- EBI and WTSI
- annotation by the Ensembl project
- www.ensembl.org

- first release in 2009
- non-vertebrates (protists, plants, fungi, metazoa, bacteria)
- EBI
- annotation in collaboration with the scientific communities
- www.ensemblgenomes.org
**Ensembl gene models**

- Automatic annotation
  Genome-wide determination using automated pipeline

- Gene predictions based on ESTs (ESTGenes)

- *Ab initio* predictions (based on the genomic sequence, no experimental evidence)

- RNAseq data

- cDNA and protein alignments (from sequence DBs)
Biological Evidence

• International Nucleotide Sequence databases
  - ENA (European Nucleotide Archive)
  - GenBank
  - DDBJ (DNA Data Bank of Japan)

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

• NCBI RefSeq
  - Manually annotated proteins and mRNAs (NP, NM)
  - Predicted proteins and mRNAs (XP, XM)
Ensembl gene models

- Automatic annotation
  Genome-wide determination using automated pipeline

- Gene predictions based on ESTs (ESTGenes)
- *Ab initio* predictions (based on the genomic sequence, no experimental evidence)
- RNAseq data
- cDNA and protein alignments (from sequence DBs)

- Manual annotation
  Gene determination on a case-by-cases by a curator
Manual Annotation

- **havana** at Wellcome Trust Sanger Institute

- pig, gorilla, dog, wallaby, chimp (gene lists)

- available at the Vega website
  
  http://vega.sanger.ac.uk/index.html

- available at the Ensembl website too!
• automatic annotation
• many species (n=61)
• genome-wide at once

• manual annotation
• fewer species (Hs, Mm, Dr)
• gene by gene

Merged ("gold")

Ensembl (20_)

Havana (00_)
What are the gold transcripts?

- identical annotation from both the Ensembl and Havana teams (merged gene set)
- human, mouse and zebrafish only
- high confidence and quality

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

- annotate a consensus coding DNA sequence set
- EBI, WTSI, UCSC and NCBI
- human and mouse only


Genome Res. 19:1316-23 (2009)
Ensembl genes and transcripts

High Quality

- CCDS transcripts
- Ensembl/Havana merged transcripts
Ensembl stable identifiers

- 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

Available for vertebrates only
Ensembl features

- Gene build for a wider taxonomic range
- Gene trees and regulation build
- Variation display and tools
- BioMart (data export)
- Display of external data (DAS)
- Programmatic access via the Perl API
- Data and code are open source
How is all this information organised?

1) Ensembl browser sites
   Main web site, Pre!, Archive!

2) BioMart (Data Mining tool)

3) Ensembl Database (open source)
   Perl-API, MySQL

4) FTP download site
Help and documentation

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

- Facebook www.facebook.com/Ensembl.org
- Twitter https://twitter.com/Ensembl
- Come visit our blog! www.ensembl.info
Publications

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

• Flicek, P. et. al.
  Ensembl 2012
  http://nar.oxfordjournals.org/content/40/D1/D84.long

• 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
# Ensembl Team

<table>
<thead>
<tr>
<th>Ensembl</th>
<th>Paul Flicek <strong>(EBI)</strong>, Steve Searle <strong>(Sanger Institute)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td>Andy Yates, Stephen Keenan, Monika Komorowska, Rhoda Kinsella, Ian Longden, Thomas Maurel, Kieron Taylor</td>
</tr>
<tr>
<td>Comparative Genomics</td>
<td>Javier Herrero, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Matthieu Muffato, Miguel Pignatelli</td>
</tr>
<tr>
<td>Regulation</td>
<td>Ian Dunham, Nathan Johnson, Daniel Sobral, Steven Wilder</td>
</tr>
<tr>
<td>Variation</td>
<td>Fiona Cunningham, Laurent Gil, Jackie MacArthur, Will McLaren, Graham Ritchie</td>
</tr>
<tr>
<td>Analysis and Annotation</td>
<td>Bronwen Aken, Amonida Zadissa, Dan Barrell, Susan Fairley, Carlos García Girón, Thibaut Hourlier, Andreas Kähäri, Rishi Nag, Magali Ruffier, Simon White</td>
</tr>
<tr>
<td>Web Team</td>
<td>Anne Parker, Ridwan Amode, Simon Brent, Bethan Pritchard, Harpreet Riat, Steve Trevanian (VEGA)</td>
</tr>
<tr>
<td>Outreach</td>
<td>Giulietta M. Spudich, Jeff Almeida-King, Denise Carvalho-Silva, Bert Overduin, Michael Schuster (QC)</td>
</tr>
<tr>
<td>Ensembl Genomes</td>
<td>Paul Kersey, Paul Derwent, Jay Humphrey, Arnaud Kerhornou, Eugene Kulesha, Nick Langridge, Uma Maheswari, Mark McDowall, Michael Nuhn, Helder Pedro, Dan Staines, Iliana Toneva</td>
</tr>
<tr>
<td>Ensembl Strategy</td>
<td>Ewan Birney, Richard Durbin, Paul Flicek, Jen Harrow, Tim Hubbard, Glenn Proctor, Steve Searle</td>
</tr>
</tbody>
</table>
Hands up if you've got a question
What data are you interested in?

Write down 1-2 bullet points on a Post it note and put this up in front of the room.
Ensembl browser

Walk through (pages 10-24)

Hands-on exercises (pages 25-27)
BioMart
Outline

• Definition

• The principle: 4 steps

• An example: simple query in zebrafish

• BioMart: elsewhere

• Where to find it

• Try it out: worked examples and hands on
BioMart

- Data export tool with no programming required
- Quick table generator → Excel
- Web interface to mine Ensembl data
BioMart – 4 step principle

Step 1
DATASET
Choose database and species

Step 2
FILTER
What you know

Step 3
ATTRIBUTES
What you want to know

Step 4
RESULTS
Table/sequences
Step 1

DATASET

Choose your dataset/species
**Step 2**

Select the Filters to limit your data set (what you know).

<table>
<thead>
<tr>
<th>Filters</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGION:</td>
<td>Ensembl Gene ID</td>
</tr>
<tr>
<td>GENE:</td>
<td>Ensembl Transcript ID</td>
</tr>
<tr>
<td>EVENT:</td>
<td>Dataset</td>
</tr>
</tbody>
</table>

Please restrict your query using criteria below.
Step 3

**ATTRIBUTES**

Determine output columns (What you want to know)
**Step 4**

**RESULTS**

Tables/sequences

Export/email notification
Troponin required to regulate Ca\textsuperscript{2+} dependent contraction

Mutant forms of troponins associated with muscle diseases

Troponin genes expressed in muscle during fish development: \textit{tnnt3a}, \textit{tnnt3b} and \textit{tnnt2c}
Simple query

For the zebrafish *tnnt3a*, *tnnt3b* and *tnnt2c* genes, retrieve a list of Ensembl gene and transcript IDs, and their anatomical system and development stage (ZFIN)
Step 1

DATASET
Ensembl Genes
Danio rerio

Step 2
FILTERS
- tnnt3a
- tnnt3b
- tnnt2c

Step 3
ATTRIBUTES
- Ensembl gene ID
- Ensembl transcript ID
- Anatomical system
- Development stage

Step 4
RESULTS
Step 1 – Database and Dataset

Current Ensembl

- Choose Database -

Archives also available

zebrafish genes dataset
Step 2 - Filters
Enter the known genes

Expand the GENE panel to enter in the gene symbols (ZFIN)
Click “Count” to see if BioMart is reading 3 out of 32,313 zebrafish genes.

In “ID list limit” paste in the gene symbols (ZFIN)
Step 3 – Determine the Attributes

Expand panels to enter various attributes
### Step 4 - The Results

#### Table Preview

For the full table: click **“ALL” rows** or **“Go”**

- Click **“Unique results only”**

**Dataset 3 / 32313 Genes**

- **Danio rerio genes (Zv9)**

**Filters**

- **ZFIN symbol(s) [e.g. zgc:153184]:** [ID-list specified]

**Attributes**

- **Ensembl Gene ID**
- **Ensembl Transcript ID**
- **Anatomical System (ZFIN)**
- **Development Stage (ZFIN)**

<table>
<thead>
<tr>
<th>Ensembl Gene ID</th>
<th>Ensembl Transcript ID</th>
<th>Anatomical System</th>
<th>Development Stage (ZFIN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSDARG00000032232</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000017</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000018</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000019</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000020</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000021</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000022</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000023</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000024</td>
</tr>
<tr>
<td>ENSDARG00000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin muscle</td>
<td>ZFS:0000025</td>
</tr>
<tr>
<td>Ensembl Gene ID</td>
<td>Ensembl Transcript ID</td>
<td>Anatomical System (ZFIN) Development Stage (ZFIN)</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000032242</td>
<td>ENSDART00000044154</td>
<td>pectoral fin musculature ZFS:0000017</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000018</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000019</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000020</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000021</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000022</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000023</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000024</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000025</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000026</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000027</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000028</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000029</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000030</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000031</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000032</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>pectoral fin musculature ZFS:0000033</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000017</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000018</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000019</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000020</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000021</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000022</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000023</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000024</td>
<td></td>
</tr>
<tr>
<td>ENSDARG0000030270</td>
<td>ENSDART00000125901</td>
<td>unspecified ZFS:0000025</td>
<td></td>
</tr>
</tbody>
</table>
BioMart started at Ensembl
Where to find BioMart

central.biomart.org
www.ensembl.org/biomart/martview
Production Team

Rhoda Kinsella  Thomas Maurel
Hands up if you’ve got a question
BioMart

Worked examples (page 32)

*Hands-on* exercises (pages 37-40)
Wrap up (Summary)

Ensembl is a genome browser where gene annotation, variation, regulation and comparative genomic data are all integrated. It can be accessed through the website, BioMart or the Perl APIs.
Your take home message
Help and documentation

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