

Module 1

Browsing genomes with Ensembl

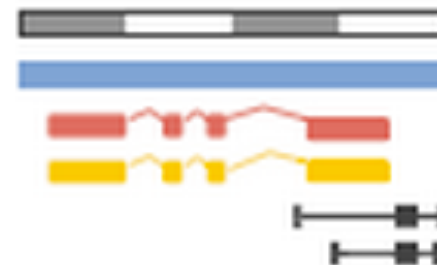
Using the Ensembl Genome Browser to View Annotation of the Human Genome



Emily Perry

Ensembl Outreach Project Leader

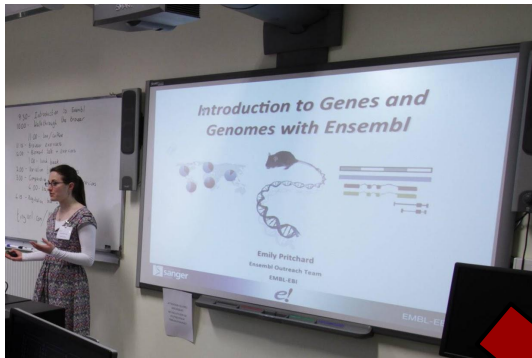
EMBL-EBI



This session

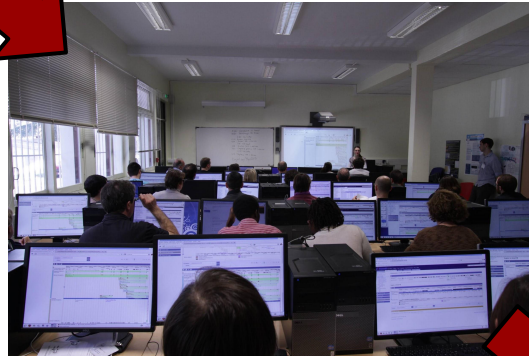
- Introduction to Ensembl
- The Region in detail view walkthrough
- How Ensembl genes and transcripts are produced
- Genes and transcripts walkthrough
- Exercises

Structure



Presentation:
What the data/tool is
How we produce/process the data

Demo:
Getting the data
Using the tool



Exercises:
Trying things out for yourself
Going beyond the demo
Not a test!



Course materials

<http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html>

- Presentations
- Coursebook

- Coursebook page 1-19

Objectives

- What is **Ensembl**?
- What type of data can you get in **Ensembl**?
- How to navigate the **Ensembl** browser **website**.
- Where to go for **help** and **documentation**.



Questions?

Exploring the Ensembl genome browser

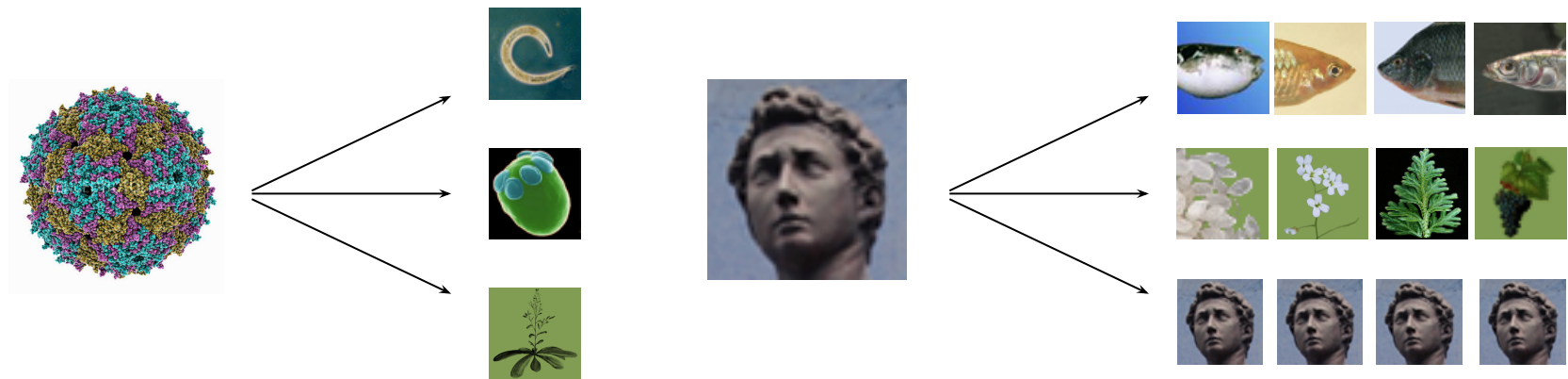
The screenshot shows the Ensembl genome browser homepage. At the top, there is a navigation bar with links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. A search bar is located on the right side of the navigation bar. Below the navigation bar, there is a main search area with a dropdown menu for 'All species' and a search button. Below this, there are several sections: 'Browse a Genome' with a description of the project and a list of popular genomes (Human, Mouse, Zebrafish); 'New to Ensembl?' with a list of tutorials and resources; 'What's New in Release 71 (April 2013)' with a list of new features; and 'Did you know...?' with a link to the YouTube channel. The page is designed with a clean, professional layout and uses a color scheme of blue, white, and yellow.

Introduction

Why do we need genome browsers?

1977: 1st genome to be sequenced (5 kb)

2004: finished human sequence (3 Gb)



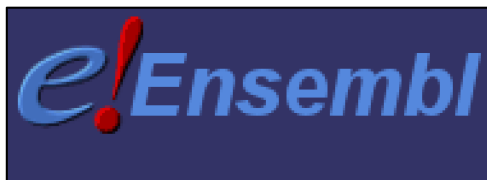
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCAGATGACGCCATCTGAAATTTCTT
GGAAACACGATCACTTTAACGGAATATTGCTGTTTTGGGGAAGTGTTTTACAGCTGCTGGGCACGCTGTATTTGCCTT
ACTTAAGCCCCTGGTAATTGCTGTATTCCGAAGACATGCTGATGGGAATTACCAGGCGGCGTTGGTCTCTAACTGGAG
CCCTCTGTCCCCACTAGCCACGCGTCACTGGTTAGCGTGATTGAAACTAAATCGTATGAAAATCCTCTTCTCTAGTGC
CACTAGCCACGTTTTCGAGTGCTTAATGTGGCTAGTGGCACCGTTTTGGACAGCACAGCTGTAAAATGTTCCCATCCTC
ACAGTAAGCTGTTACCGTTCAGGAGATGGGACTGAATTAGAATTCAAACAAATTTTCCAGCGCTTCTGAGTTTTACC
TCAGTCACATAATAAGGAATGCATCCCTGTGTAAGTGCATTTTGGTCTTCTGTTTTTGCAGACTTATTTACCAAGCATT
GGAGGAATATCGTAGGTA AAAATGCCTATTGGATCCAAAGAGAGGCCAACATTTTTTCAAATTTTAAAGACACGCTGC
AACAAAGCAGGTATTGACAAATTTTATATAACTTTATAAATTACACCGAGAAAGTGTTTTCTAAAAAATGCTTGCTAA
AAACCCAGTACGTCACAGTGTTGCTTAGAACCATAAACTGTTCCCTTATGTGTGTATAAATCCAGTTAACACATAATC
ATCGTTTTGCAGGTTAACCACCTTGACTAGCAGTA
GGAACAATTACTAACAAATCTGGTTTTCAGTACTCC
TTATACTCTTAAAAATGATCAGAAATTTAAACTAA
GAATTTAAGGCTGGGCGTGCAGCGGATCACTTGAGG
CCAGAAGTTTGAGACCAGCCATGTGCTGCGTGTGG
TGGTGCGTGCCTGTAATCCAGCTACACGGGAGGTGGAGGCAGGAGAATCGCTTGAACCTTGGAGGCAGAGGTTGCAG
TGAGCCAAGATCATGCCACTGCACTCTAGCCTGGGCCACATAGCATGACTCTGTCTCAAACAAACAAACAAACAAAA
AACTAAGAATTTAAAGTTAATTTACTTAAAAATAATGAAAGCTAACCCATTGCATATTATCACAACATTTCTTAGGAAA
AATAACTTTTTGAAAACAAGTGAGTGGAATAGTTTTTACATTTTTGCAGTTCTCTTTAATGTCTGGCTAAATAGAGAT
AGCTGGATTCACTTATCTGTGTCTAATCTGTTATTTTTGGTAGAAGTATGTGAAAAAAAATTAACCTCACGTTGAAAAA
AGGAATATTTAATAGTTTTTCAGTTACTTTTTGGTATTTTTCCCTGTACTTTGCATAGATTTTTCAAAGATCTAATAG
ATATACCATAGGTCTTTCCCATGTGCAACATCATGCAGTGATTATTTGGAAGATAGTGGTGTCTGAATTATACAAA
GTTTCCAAATATTGATAAATTGCATTAAACTATTTTTAAAAATCTCATTCAATTAATACCACCATGGATGTCAGAAAAGT
CTTTAAGATTGGGTAGAAATGAGCCACTGGAAATTTCTAATTTTCATTTGAAAGTTCACATTTTGTCATTGACAACAA
ACTGTTTTCCCTTGCAACAACAAGATCACTTCATTGATTTGTGAGAAAATGTCTACCAAATTTAATTAAGTTGAAATAAC
TTTGTCAGCTGTTCTTTCAAGTAAAAATGACTTTTCATTGAAAAAATTGCTTGTTTCAGATCACAGCTCAACATGAGTG
CTTTTCTAGGCAGTATTGTACTTCAGTATGCAGAAGTGCTTTATGTATGCTTCCCTATTTTGTCAGAGATTATTAAAAG
AAGTGCTAAAGCATTGAGCTTCGAAATTAATTTTTACTGCTTCATTAGGACATTCTTACATTAAACTGGCATTATTAT
TACTATTATTTTTAACAAGGACACTCAGTGGTAAGGAATATAATGGCTA

**Large amounts of raw
DNA sequence data**

We need to make the data mean something...

- Explore what is in a chromosomal region
- See features in and around a specific gene
- Search and retrieve data across the whole genome
- Investigate genome organisation
- Compare the human genome to other genomes

We need to make the data mean something...



<http://www.ensembl.org>



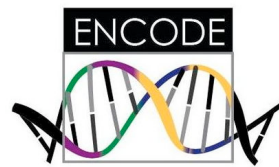
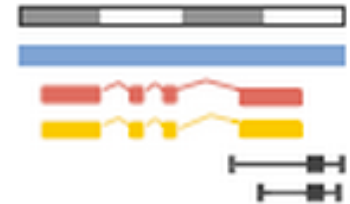
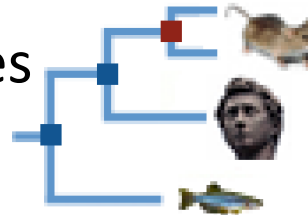
<http://genome.ucsc.edu>



<http://www.ncbi.nlm.nih.gov/mapview>

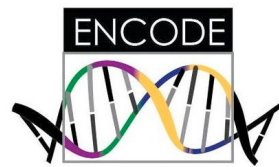
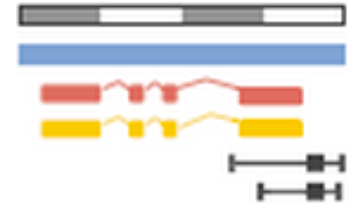
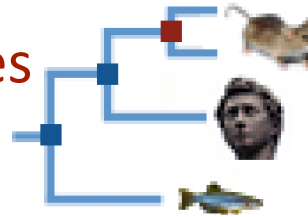
Ensembl Features

- Gene builds for ~70 species
- Gene trees
- Regulatory build (ENCODE)
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source

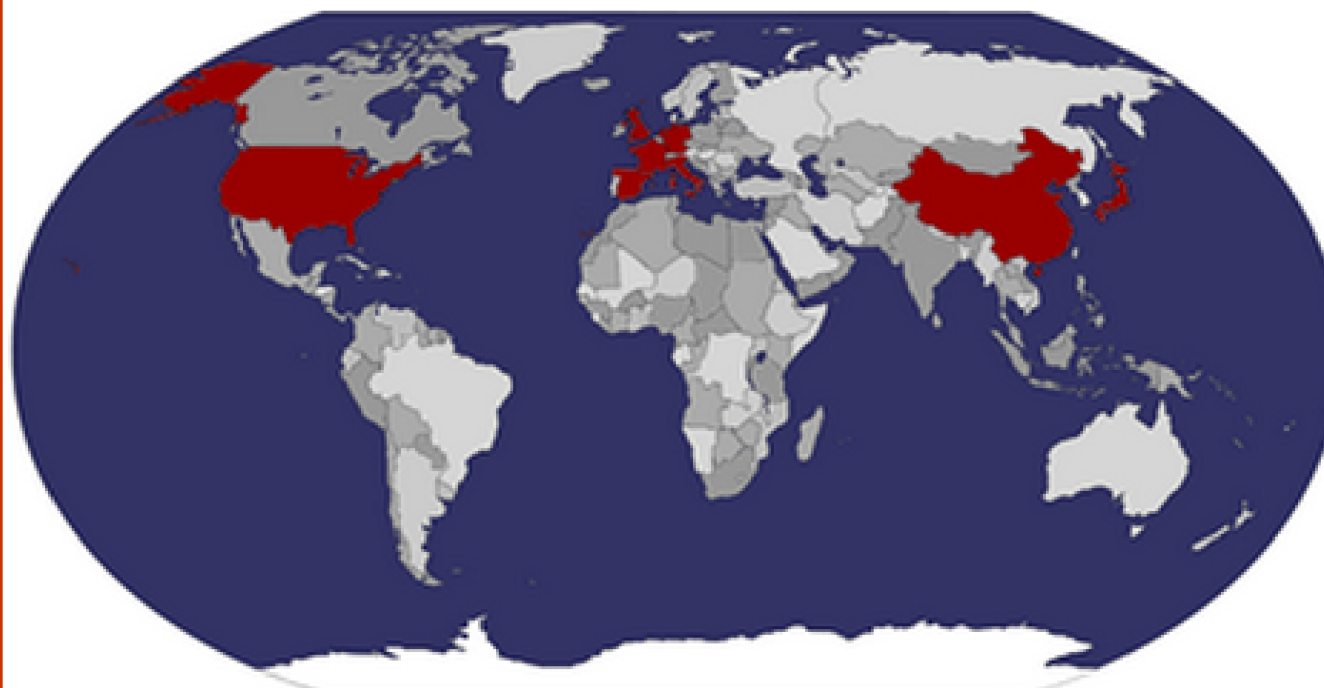


Ensembl Features

- Gene builds for ~70 species
- Gene trees
- Regulatory build (ENCODE)
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source



Ensembl is used worldwide



Top 10 countries

- United Kingdom
- United States
- Spain
- Germany
- China
- France
- Italy
- Japan
- Netherlands
- Singapore

Genome contigs



BL



AL



CM



IM

BL102

AL476

CM553

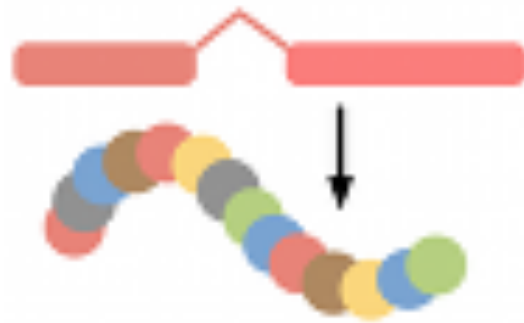
IM768

Hands on

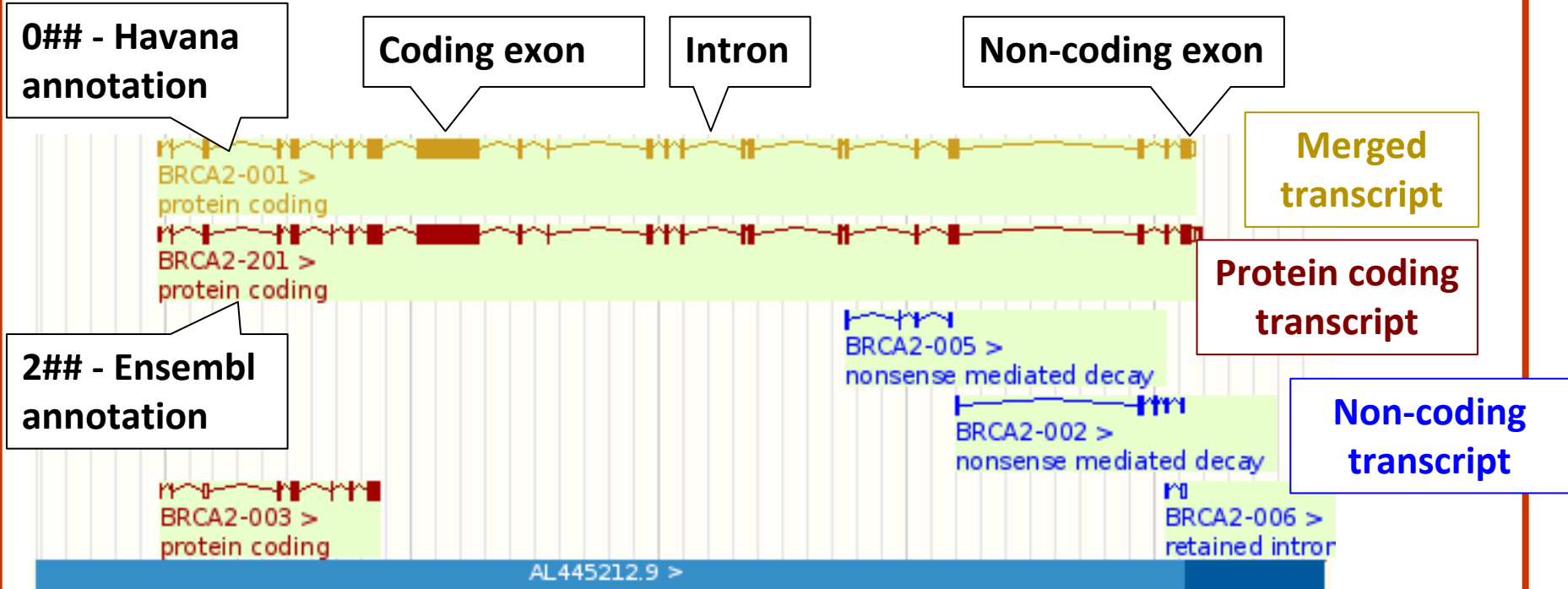
- We're going to look at the Ensembl homepage and how to find information about the species and genome assemblies in Ensembl.
- Demo: page 2-4

- We're going to look at a region of the human genome, **4: 122868000-122946000**, and manipulate the view to see the data we're interested in.
- Demo: page 4-9

Genes and Transcripts



Gene views



Golden transcripts

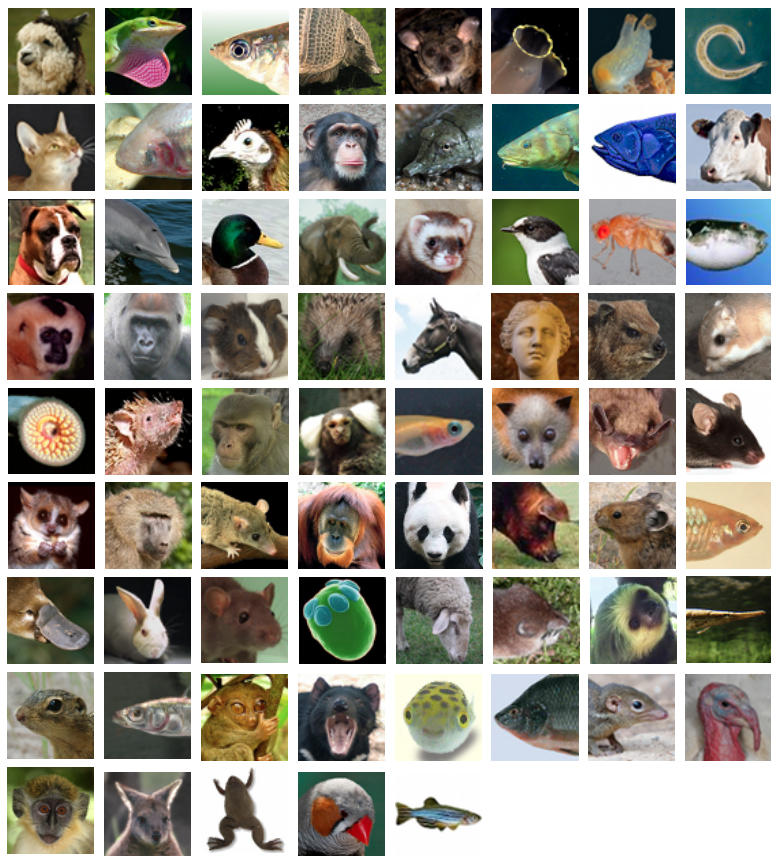
- Identical annotation



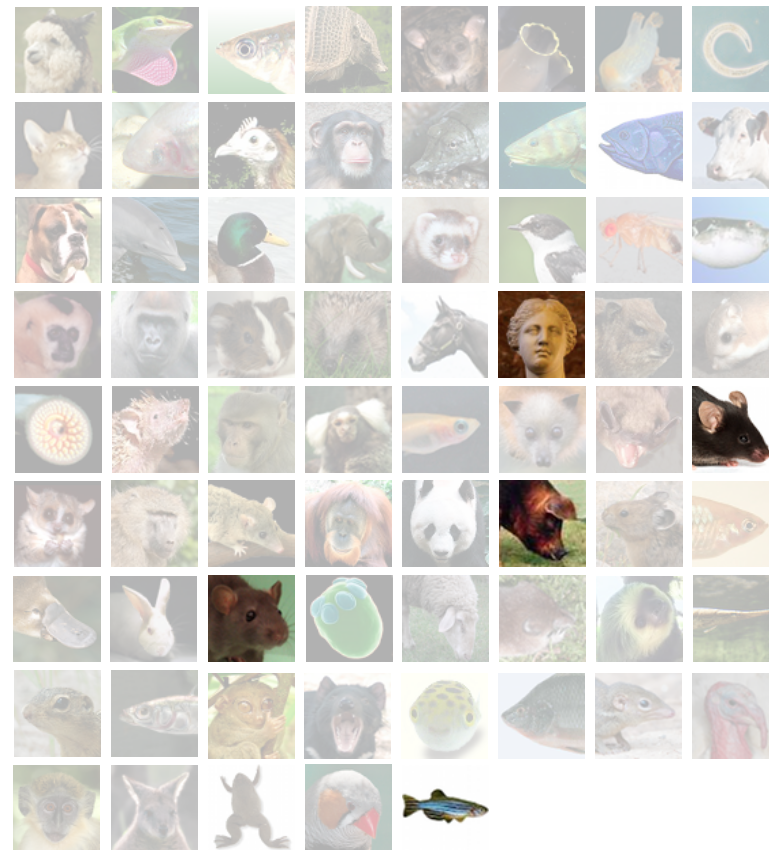
Ensembl and Havana annotation



Automatic annotation



Manual annotation



Biological Evidence

- International Nucleotide Sequence databases



GenBank



- Protein sequence databases

- Swiss-Prot: manually curated
- TrEMBL: unreviewed translations






- NCBI RefSeq



- Manually annotated proteins and mRNAs (NP, NM)



Other species

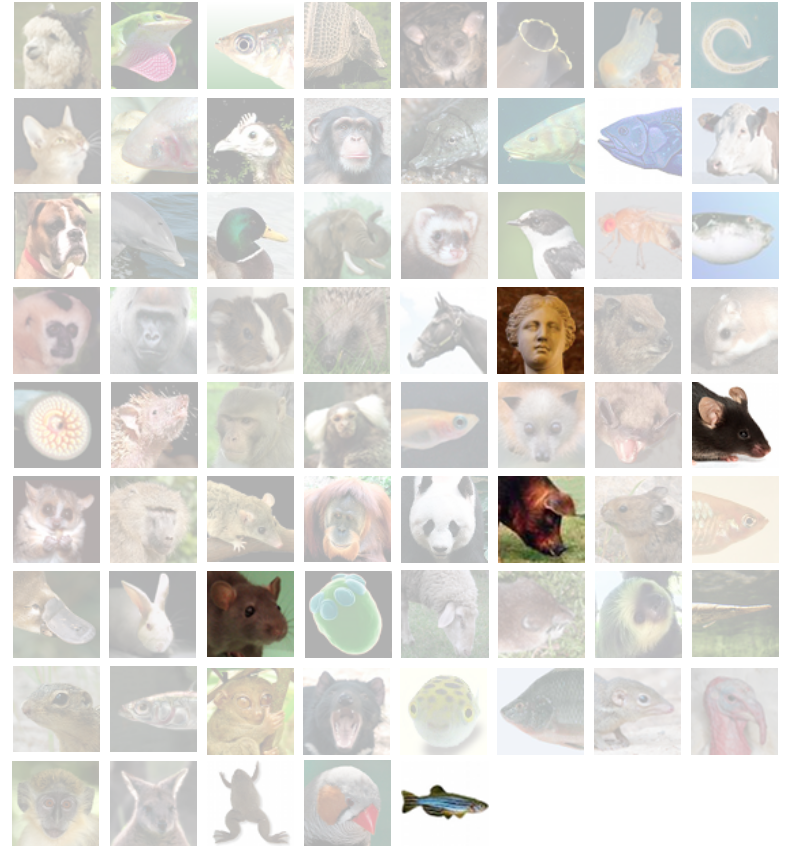
- Infer genes from homology to other species
 - Eg predict genes in  by mapping cDNAs/proteins from  to the  genome
- RNAseq data



Manual gene annotation

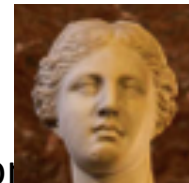
- Gene determination on a case-by-case basis by a person

-  human and vertebrate analysis and
- Genome-wide     
- Genes list    
- vega.sanger.ac.uk 



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.





Golden transcripts

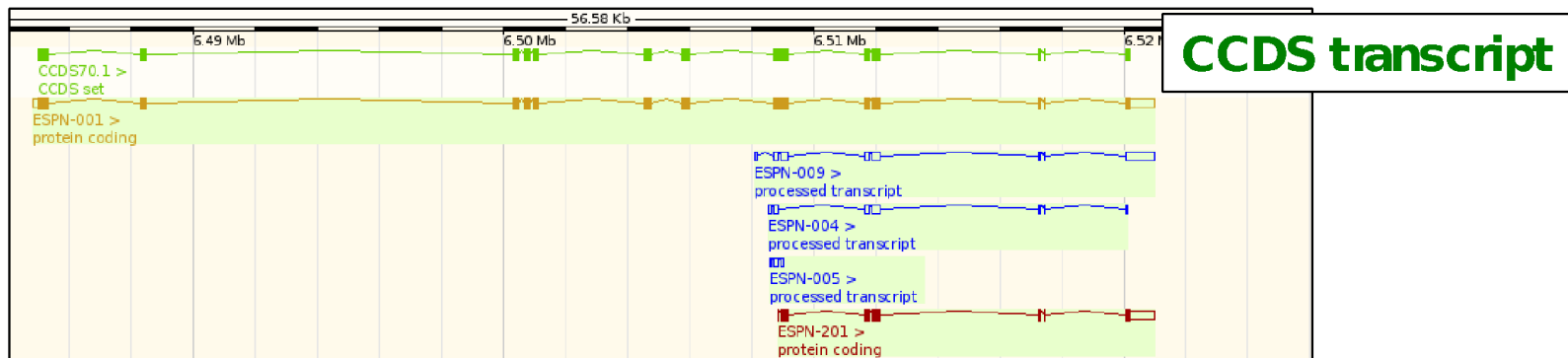
- Identical annotation



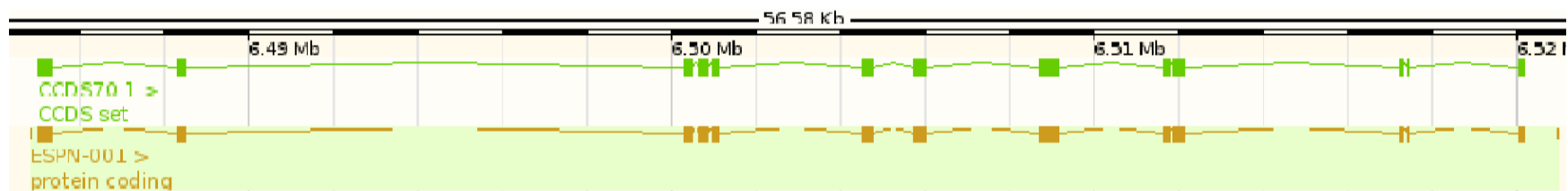
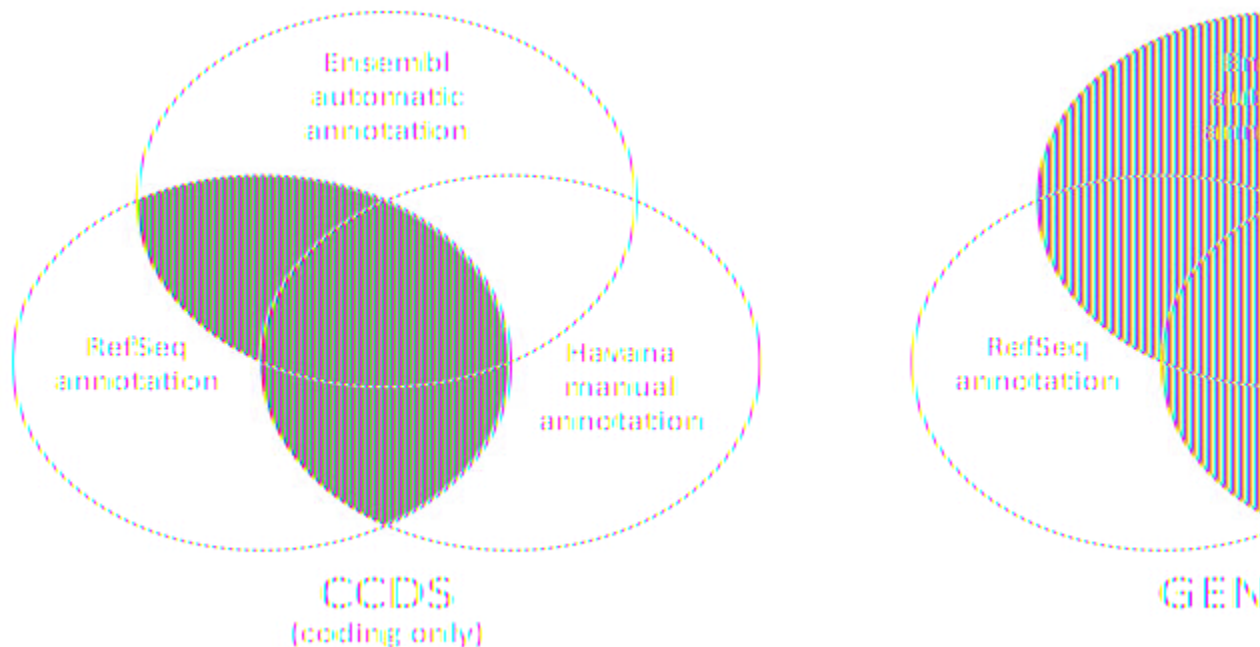
CCDS transcripts



- Consensus coding DNA sequence set
- Agreement between EBI, WTSI, UCSC and NCBI
- v8  
- <http://www.ncbi.nlm.nih.gov/CCDS/CcidsBrowse.cgi>



Higher quality transcripts



Which transcript to use?

- GENCODE Basic: Only the “complete” transcripts (where a gene has complete transcripts) [GENCODE basic](http://www.ensembl.org/Help/Glossary?id=500) (<http://www.ensembl.org/Help/Glossary?id=500>)
- Transcript support level: Scored 1-5 for quality, where 1 is the best [TSL:1](http://www.ensembl.org/Help/Glossary?id=492) (<http://www.ensembl.org/Help/Glossary?id=492>)
- APPRIS principal isoform: The most expressed isoform from proteomics analysis [APPRIS PI](http://www.ensembl.org/Help/Glossary?id=493) (<http://www.ensembl.org/Help/Glossary?id=493>)
- + 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

<http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html>

Hands on

- We're going to look at an Ensembl gene, *NUDT6*, and find out information about it and its transcripts.
- Demo:
 - Gene tab page 9-14
 - Transcript tab page 14-17
- Exercises: page 18-19
 - Answers: page 20-22

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:0009682
induced systemic resistance

GO:0009814
defence response, incompatible interaction

GO:0009616
virus induced gene silencing

GO:0034341
response to interferon-gamma

GO:0045088
regulation of innate immune response

GO:0002227
innate immune response in mucosa

GO:0002228
natural killer cell mediated immunity

GO:0045824
negative reg of innate immune response

GO:0001867
complement activation, lectin pathway

GO:0035006
melanisation defence response

GO:0009626
plant-type hypersensitive response

GO:0006957
complement activation, alternative pathway

GO:0042381
hemolymph coagulation

GO:0034342
response to type II interferon

GO:0034340
response to type I interferon

GO:0045089
positive reg of innate immune response

Hands on

- We're going to look at an Ensembl gene, *NUDT6*, and find out information about it and its transcripts.
- Demo:
 - Gene tab page 9-14
 - Transcript tab page 14-17
- Exercises: page 18-19
 - Answers: page 20-22

Help and documentation



Course 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