Module 1

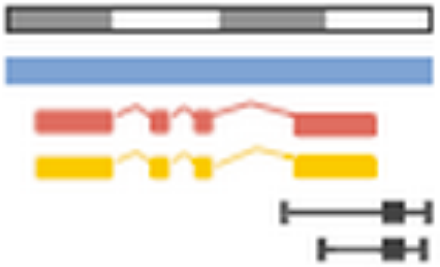
Browsing genomes with Ensembl

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

Emily Perry

Ensembl Outreach Project Leader

EMBL-EBI



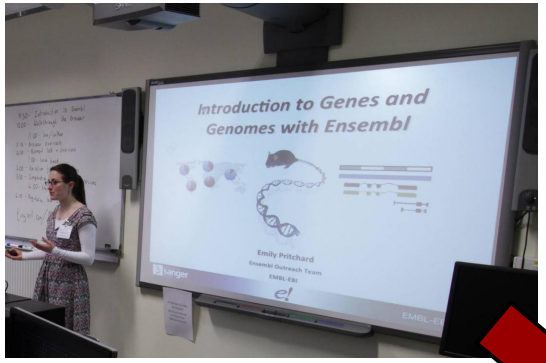
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

This session

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

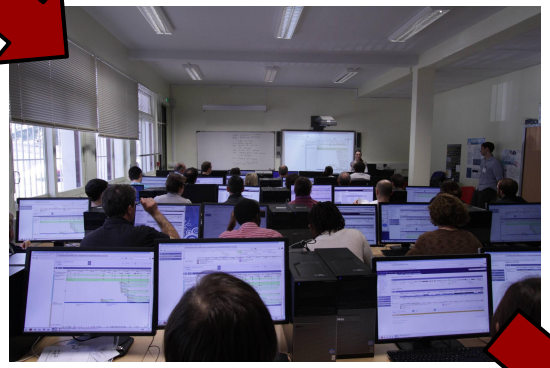
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

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!

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

Course materials

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

- Presentations
- Coursebook

- Coursebook page 1-19

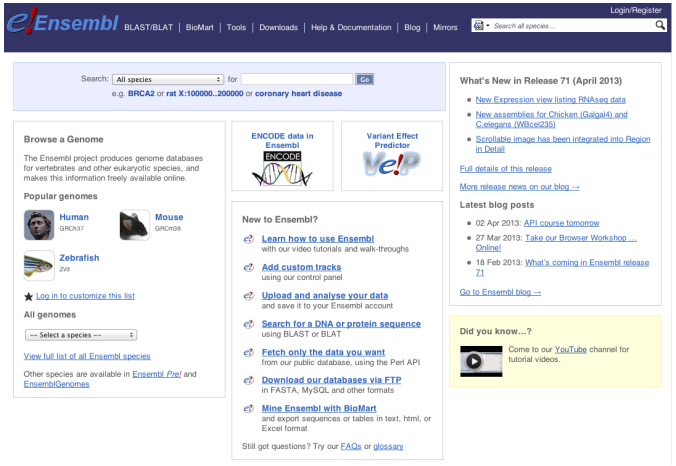
Objectives

- 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.

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

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

Exploring the Ensembl genome browser



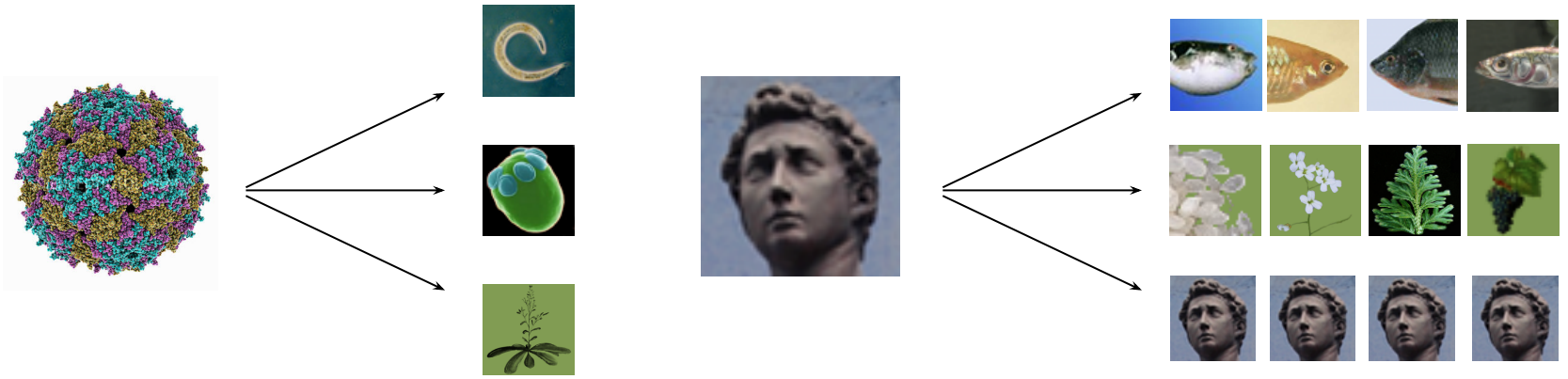
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

Introduction

Why do we need genome browsers?

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

2004: finished human sequence (3 Gb)



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

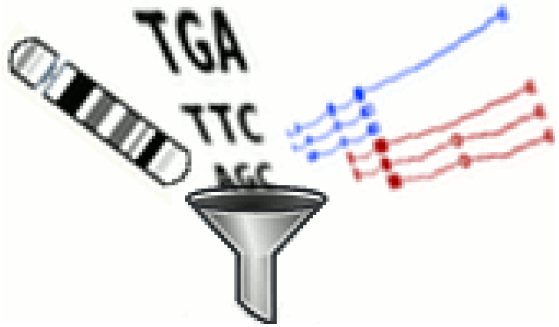
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

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

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

We need to make the data mean something…



Ensembl

http://www.ensembl.org

UCSC Genome

Genomes - Blat - Tables

http://genome.ucsc.edu

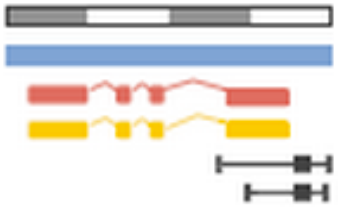
NCBI Home

Map Viewer Home

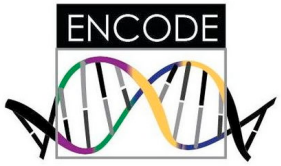
http://www.ncbi.nlm.nih.gov/mapview

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

Ensembl Features

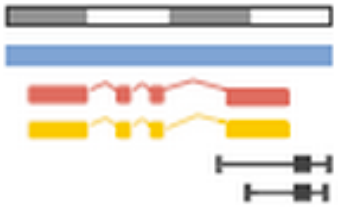


- Gene builds for ~70 species
- Gene trees
- Regulatory build (ENCODE)
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source

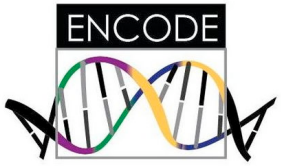


http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

Ensembl Features

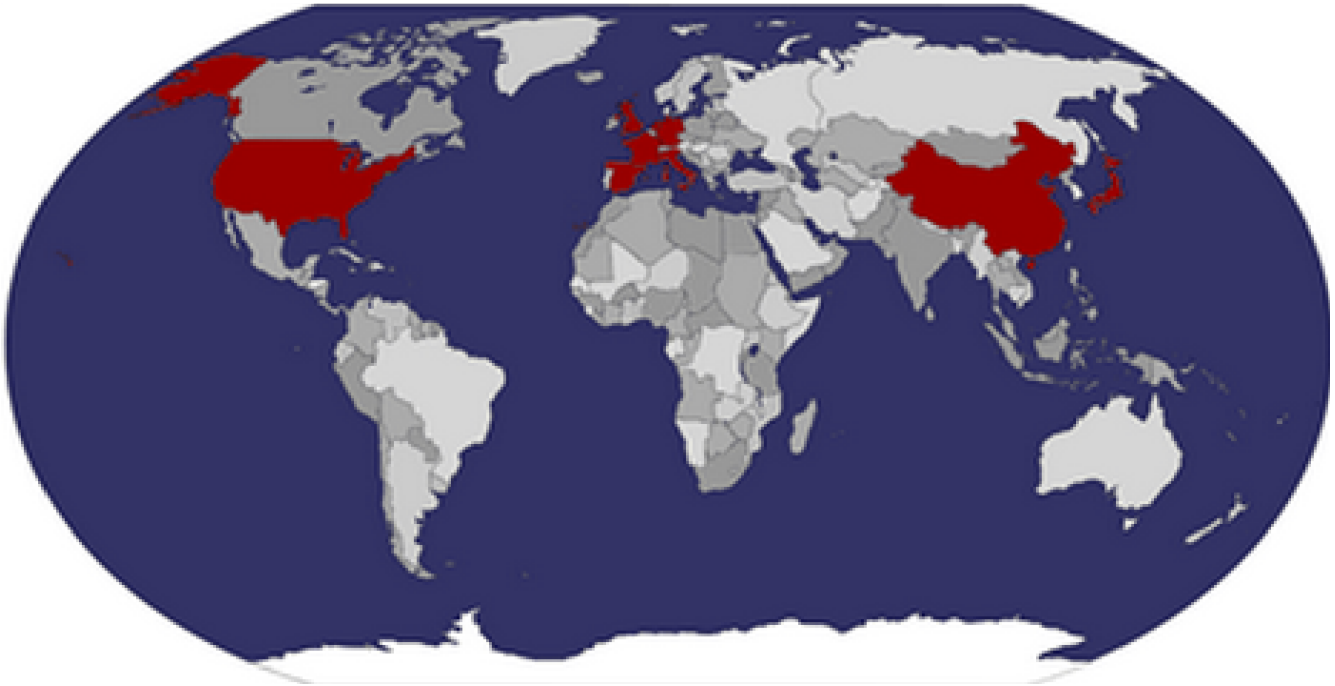


- Gene builds for ~70 species
- Gene trees
- Regulatory build (ENCODE)
- Variation display and VEP
- Display of user data
- BioMart (data export)
- Programmatic access via the APIs
- Completely Open Source



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

Ensembl is used worldwide

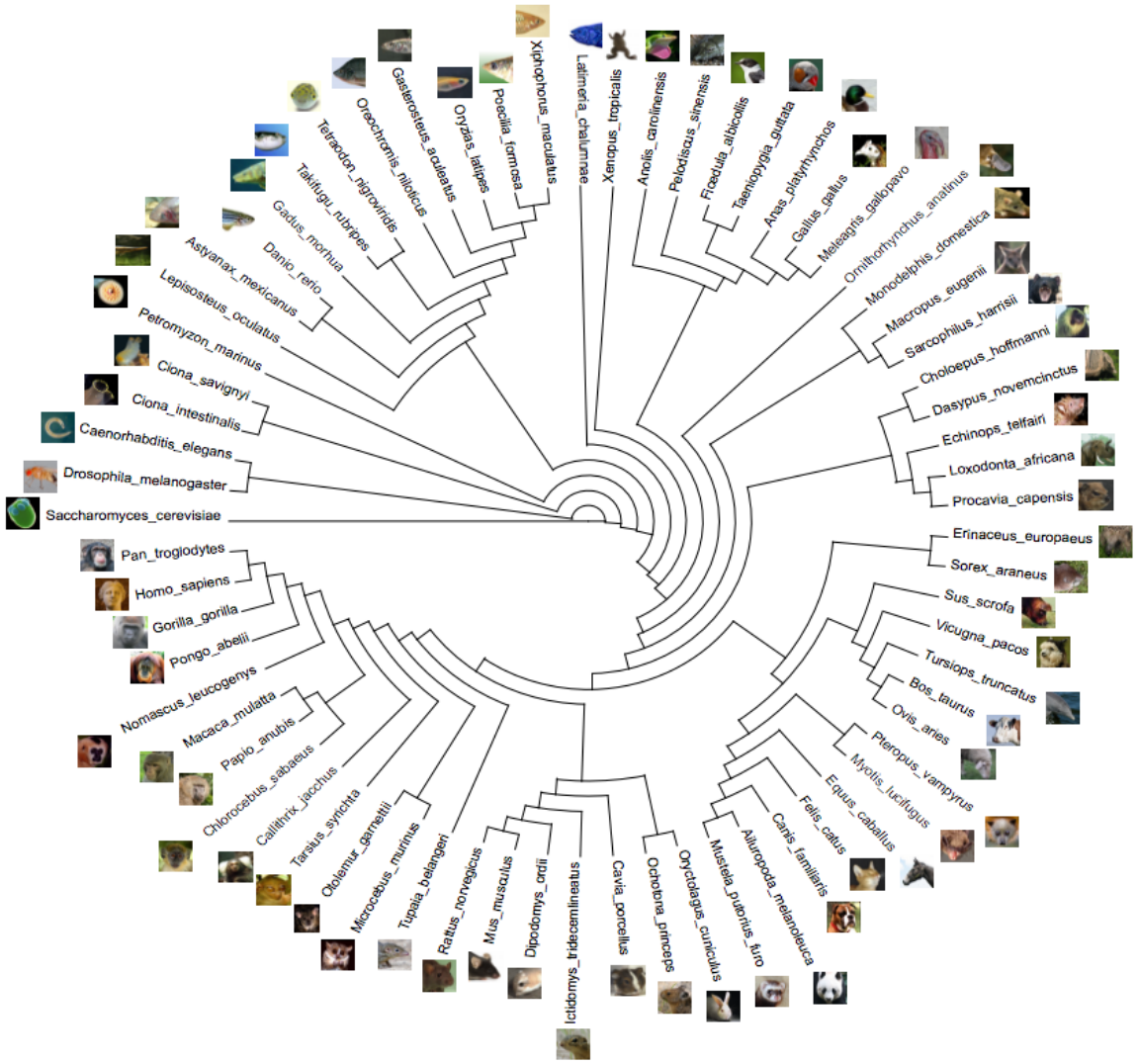


Top 10 countries

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

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

Vertebrate species on Ensembl



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

What is a genome assembly?

Sequence reads

```
                    CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA
  CAGCTGTCCCAGATGAC     ACTTAACTTCCCTCCCAGCTGTCC
 GGGCTCCGCCTTCAGCTC                          TCCCAGCTGTCCCAGATGACG
                     AACTTCCCTCCCAGCT
    CGGCCTTTGGGCTCC                    TCCGCCTTCAGCTCAAGACTTAACTTC
                      CAGATGACG
```

Match up overlaps

```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGA   AACTTCCCTCCCAGCT    CAGATGACG
            TCCGCCTTCAGCTCAAGACTTAACTTC  TCCCAGCTGTCCCAGATGACG
        GGGCTCCGCCTTCAGCTC   ACTTAACTTCCCTCCCAGCTGTCC
CGGCCTTTGGGCTCC                                CAGCTGTCCCAGATGAC
```

Genome assembly

```
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGACG
```

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

Genome contigs


BL


AL


CM


IM

BL102
AL476
CM553
IM768

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

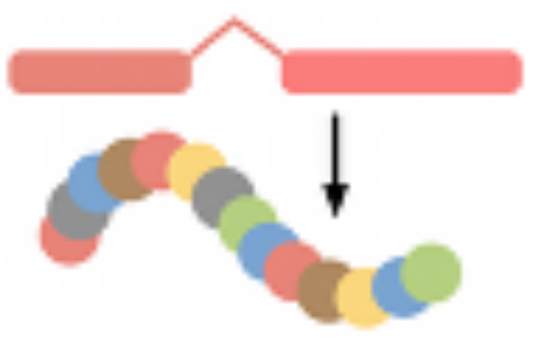
Hands on

- 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

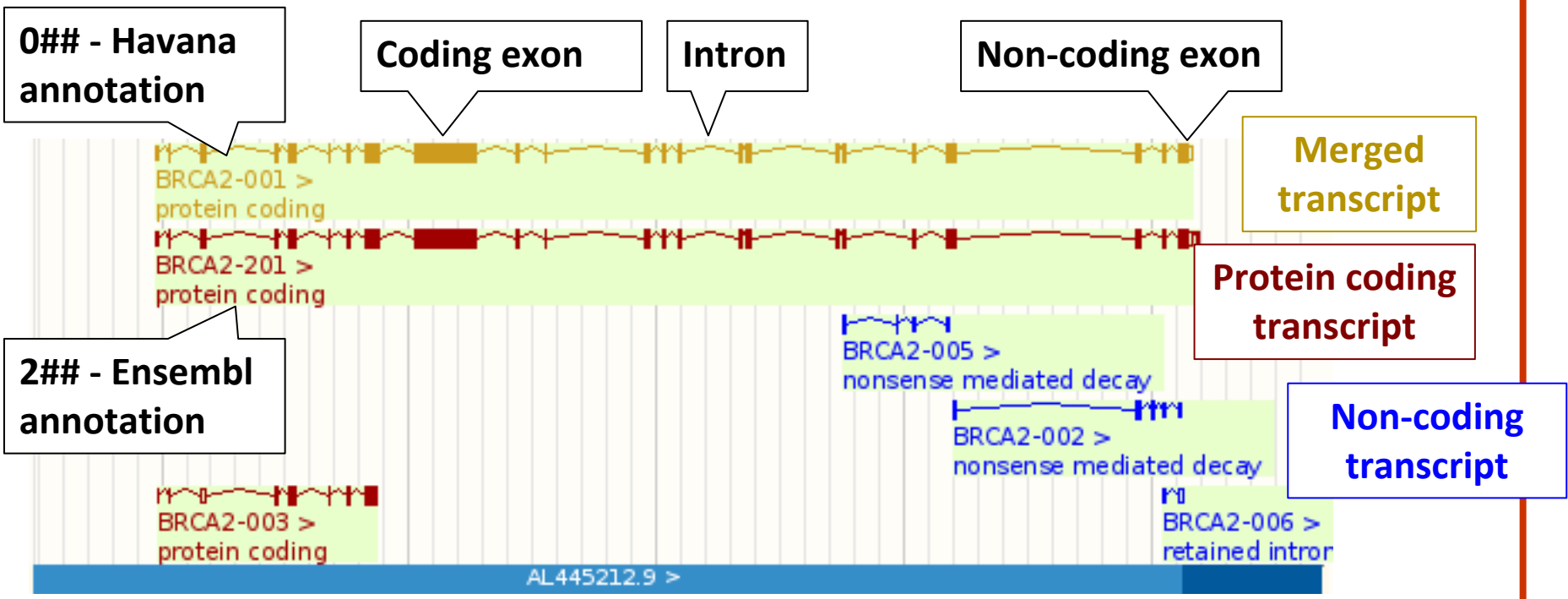
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

Genes and Transcripts



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

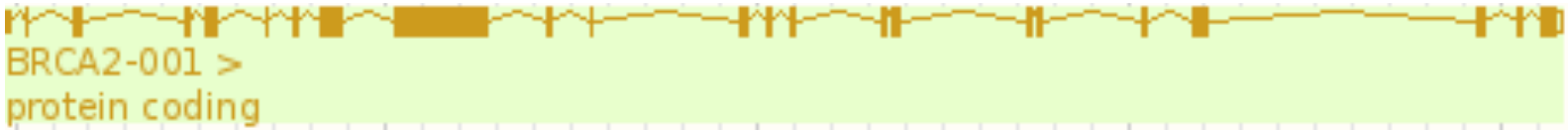
Gene views



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

Golden transcripts

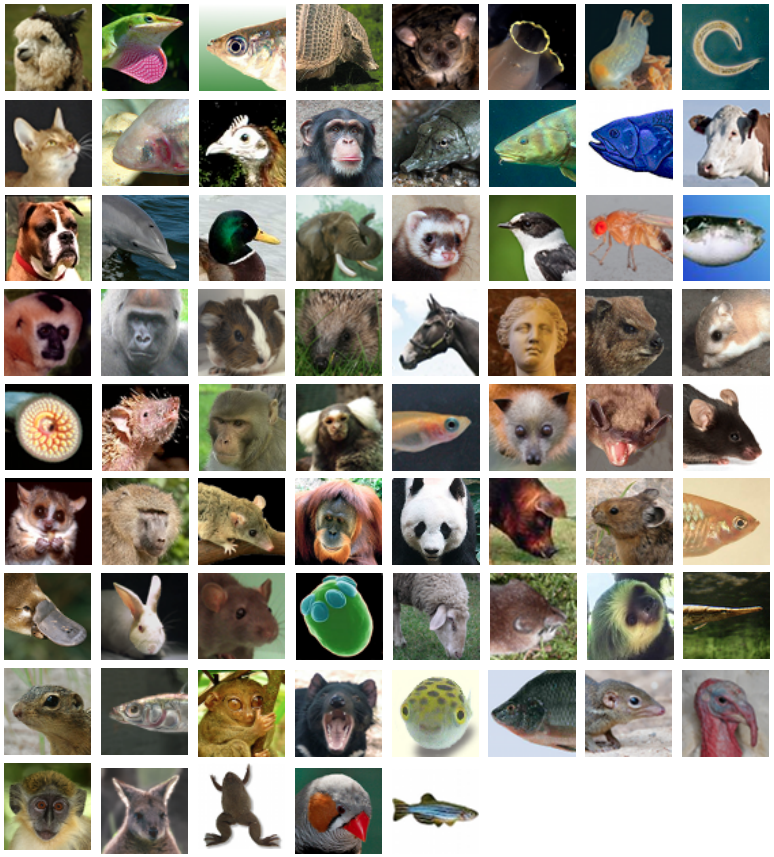
- Identical annotation



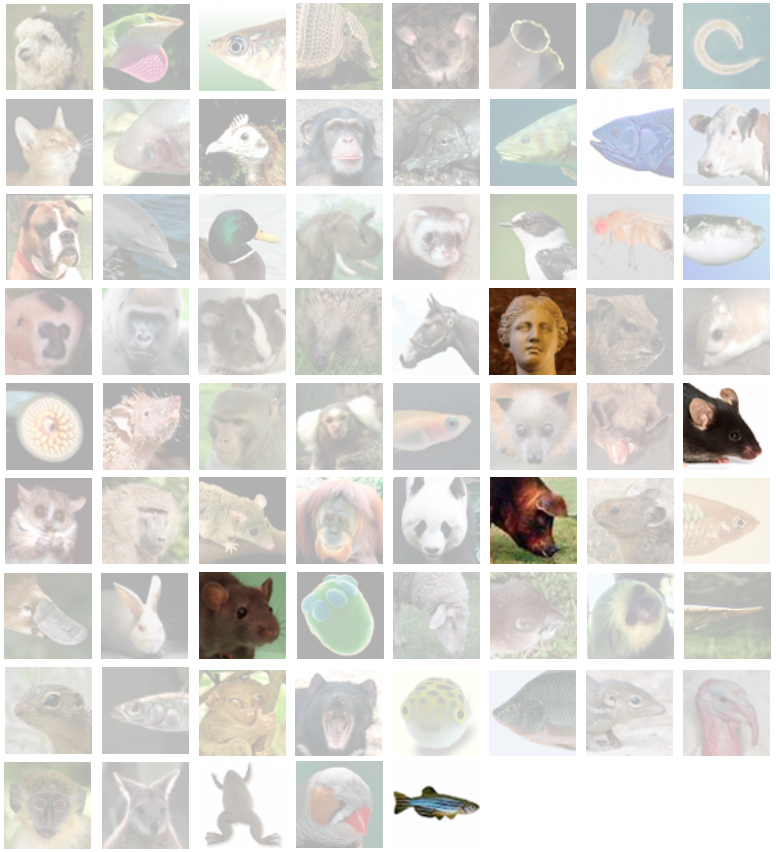
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

Ensembl and Havana annotation

Automatic annotation



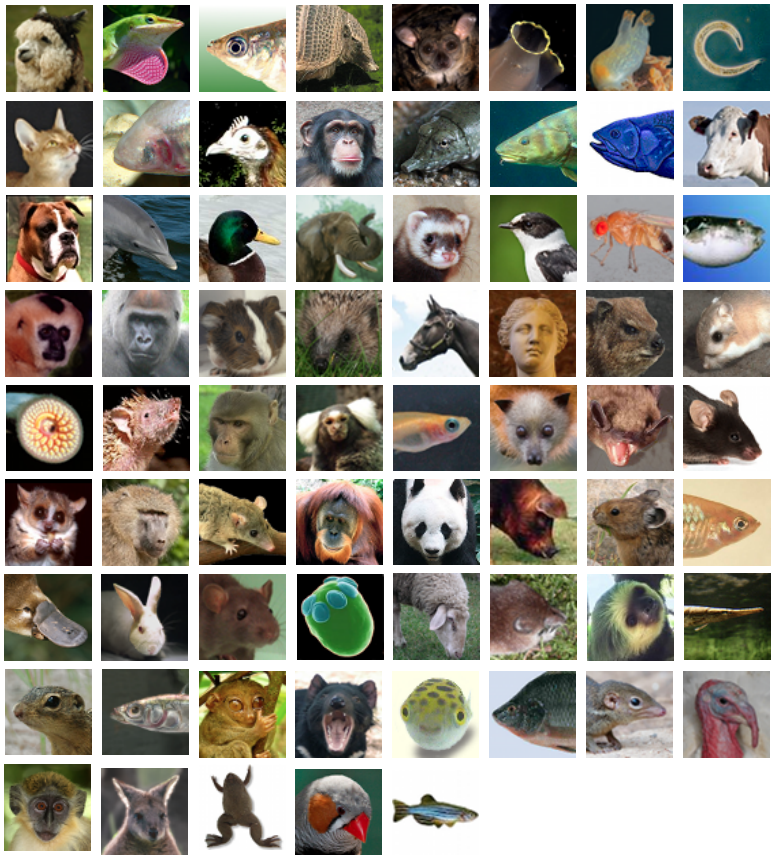
Manual annotation



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

Automatic gene annotation

- Genome-wide determination using the Ensembl automated pipeline
- Predictions based on experimental (biological) data
- Predictions based on the genomic sequence (*ab initio*)



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

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

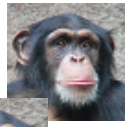
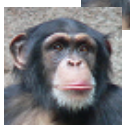
UniProt

- NCBI RefSeq

RefSeq genomes·transcripts·proteins

  - Manually annotated proteins and mRNAs (NP, NM)

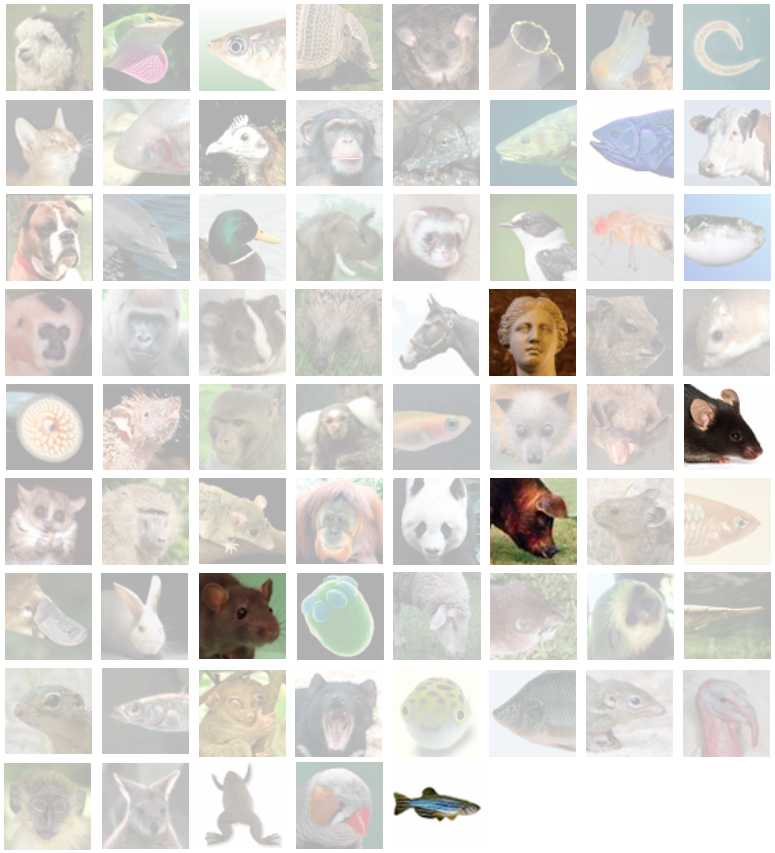
Other species

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

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

Manual gene annotation

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



- havana
- Genome-wide
- Genes list
- vega.sanger.ac.uk Vega

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

GENCODE

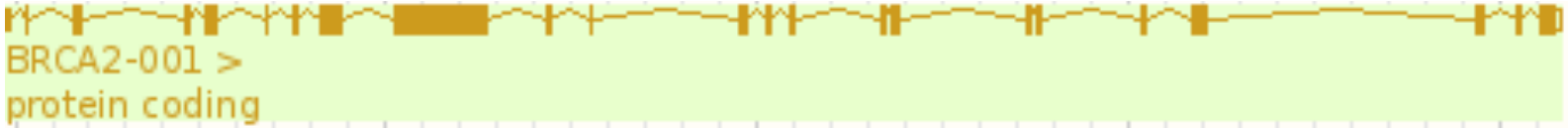
- The GENCODE gene set is made up of:
  - Ensembl automatically annotated genes
  - Havana manually annotated genes
  - The merged gene set

- GENCODE is the default gene set used by ENCODE, 1000 genomes and other major projects.

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

Golden transcripts

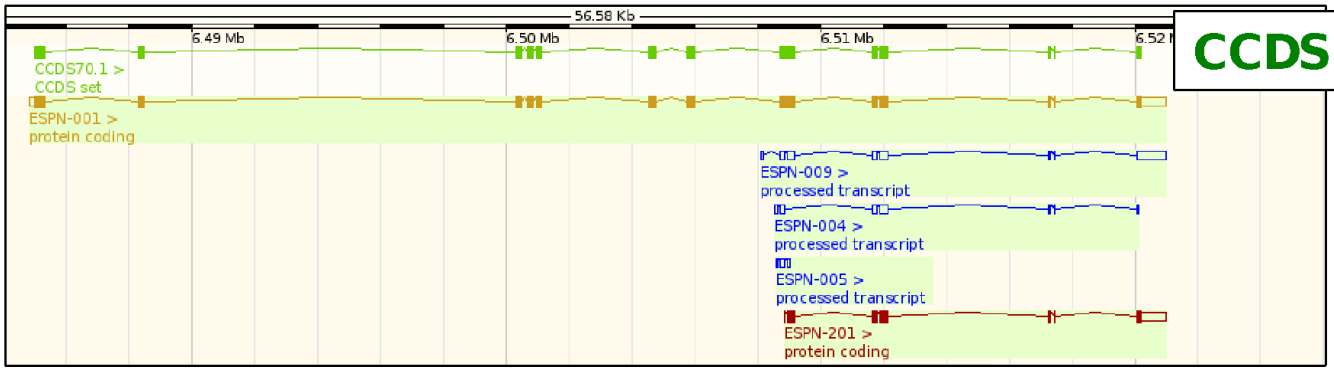
- Identical annotation



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

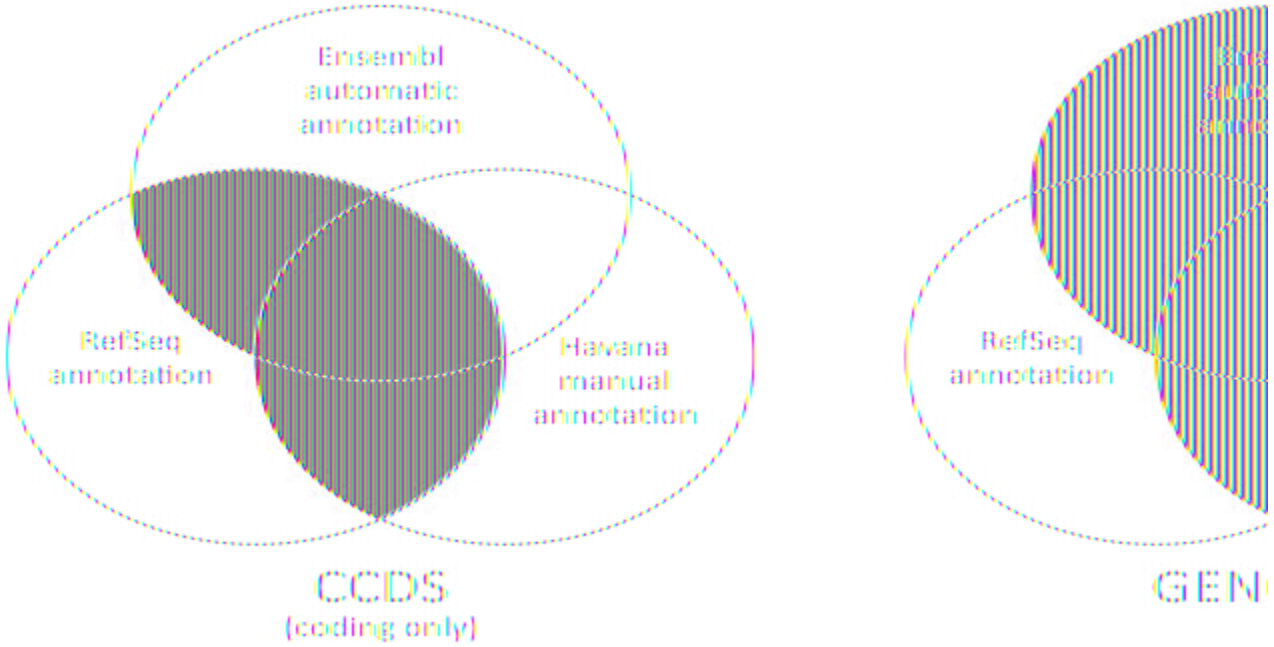
CCDS transcripts

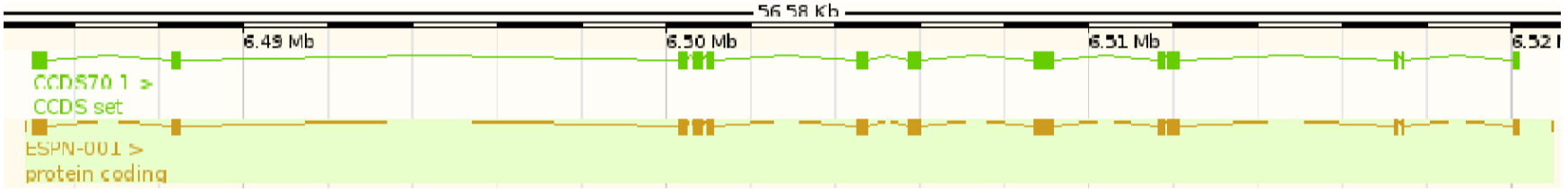
- Consensus coding DNA sequence set
- Agreement between EBI, WTSI, UCSC and NCBI
- v  
- http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi



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

Higher quality transcripts



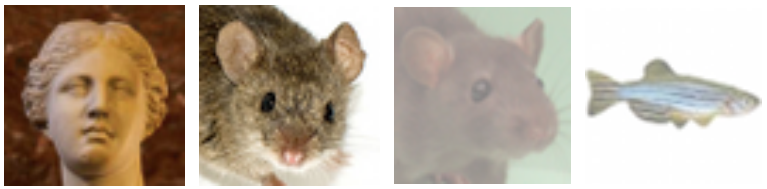


http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

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)
- Transcript support level: Scored 1-5 for quality, where 1 is the be TSL:1 (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)

- + CCDS, + Golden transcripts



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

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

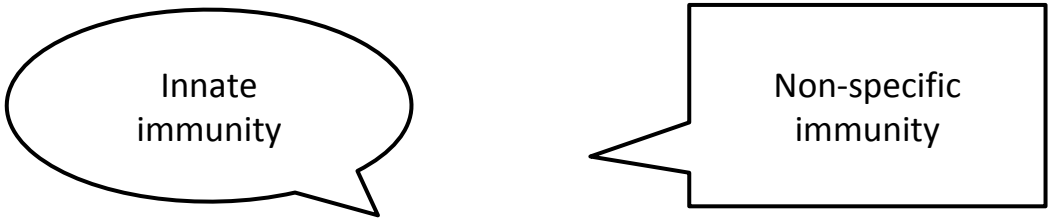
Hands on

- 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

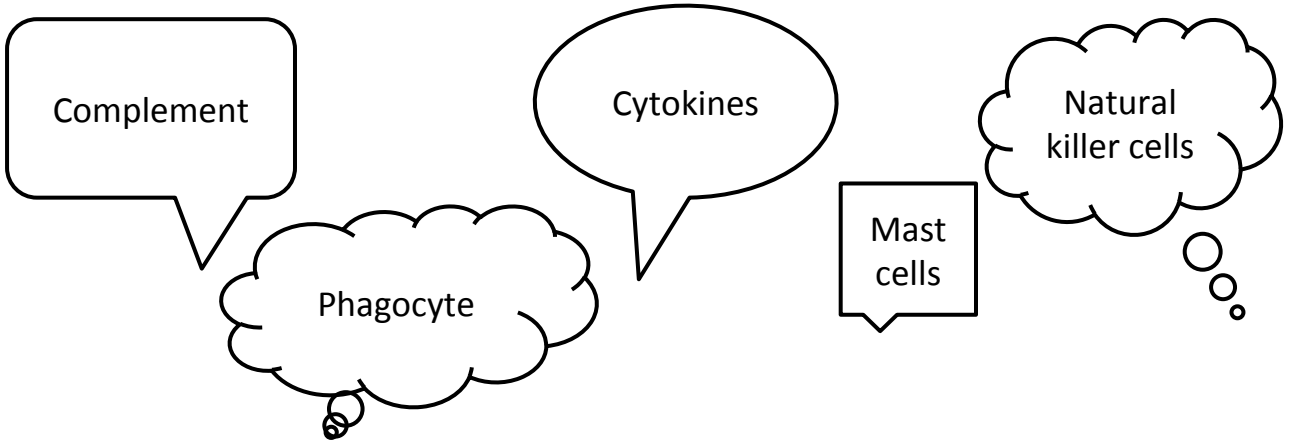
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

Why Gene Ontology (GO)?

Multiple terms for the same thing



Gene descriptions too specific



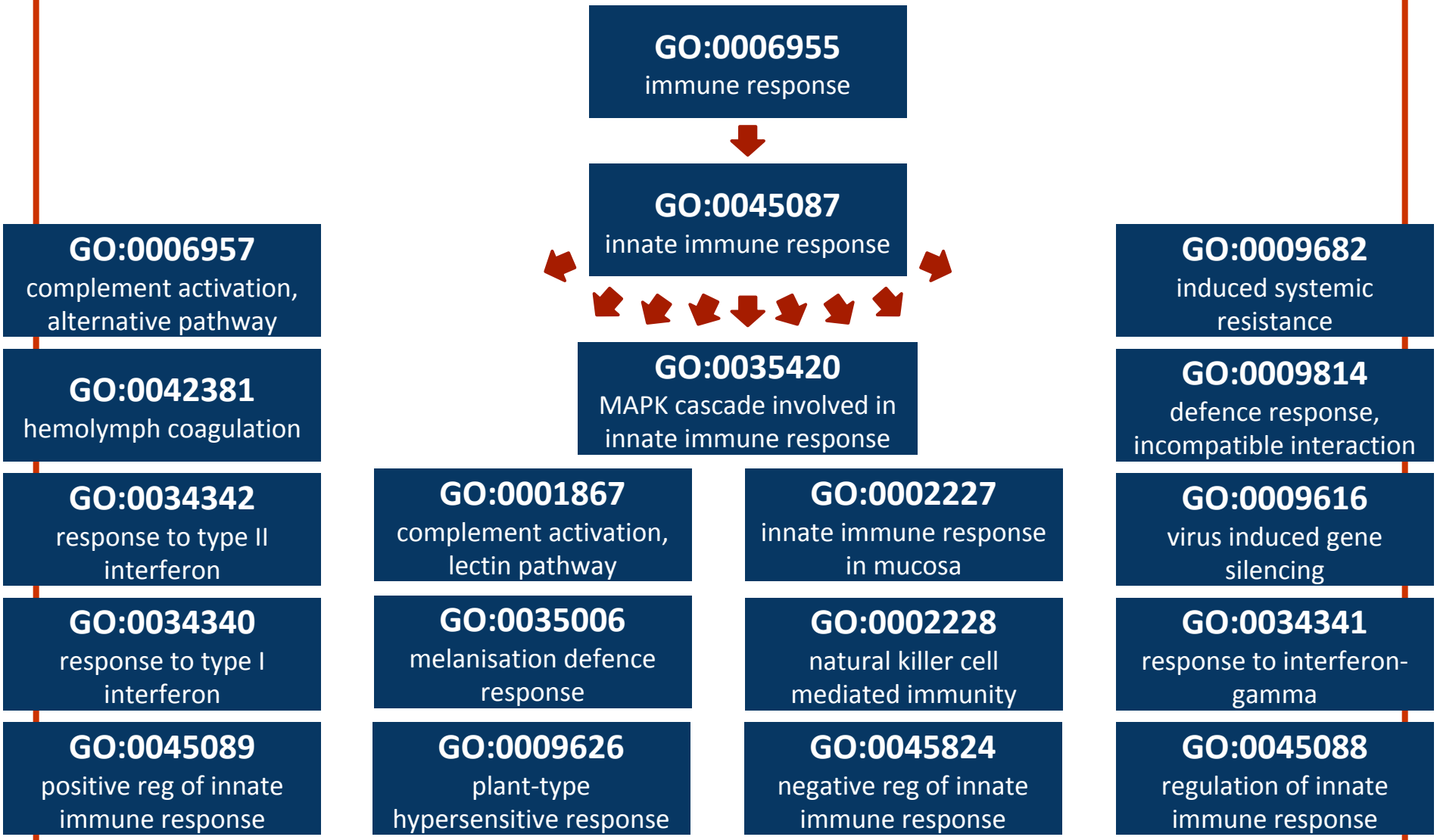
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/hinxton.html

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



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

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

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

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