Browsing Genomes with Ensembl Genomes

www.ensemblgenomes.org

Coursebook

http://www.ebi.ac.uk/~blaise/beca
BECA-ILRI – 16th October 2013

Chat room: http://tinyurl.com/ensembl-nairobi
TABLE OF CONTENTS

Introduction to Ensembl ................................................................. 3
Exploring the Ensembl genome browser ............................................. 5
Demo: Ensembl Genome in context of malaria .................................... 10
Demo EBI Transparent search engine .................................................. 11
Exercise 1 ......................................................................................... 12
Exercise 2 ......................................................................................... 13
Demo Ensembl genome sequence similarity search tool, BLAST ...... 13
Exercise 3 ......................................................................................... 15
Demo Alignment location vs. karyotype .............................................. 16
Exercise 4 ......................................................................................... 17
Demo EBI Alignment location vs. query .............................................. 17
Demo alignment summary ................................................................. 18
Exercise 5 ......................................................................................... 18
Demo The chromosome and region in details view ......................... 19
Exercise 6 ......................................................................................... 20
Demo Customizing the contig view .................................................... 20
Exercise 7 ......................................................................................... 21
Quick guide to databases and projects ............................................. 22
Introduction to Ensembl

Getting started with Ensembl

www.ensembl.org

Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute that annotates chordate genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl ‘compara’ team. Most annotation is updated every two months, leading to increasing Ensembl versions (such as version 70), however the gene sets are determined less frequently. A sister browser at www.ensemblgenomes.org is set up to access non-chordates, namely bacteria, plants, fungi, metazoa, and protists.

Ensembl provides genes and other annotation such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on protein and mRNA evidence in UniProtKB and NCBI RefSeq databases, along with manual annotation from the VEGA/Havana group. All the data are freely available and can be accessed via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interfaces (Perl APIs). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the BioMart web interface, which can extract information from the Ensembl databases without the need for programming knowledge by the user.
Need more help?

- Check Ensembl documentation
- Watch video tutorials on YouTube
- View the FAQs
- Try some exercises
- Read some publications
- Go to our online course

Stay in touch!

- Email the team with comments or questions at helpdesk@ensembl.org
- Follow the Ensembl blog
- Sign up to a mailing list

Further reading

Flicek, P. et al
Ensembl 2013
Nucleic Acids Res. Advanced Access (Database Issue)

Ensembl Methods Series
http://www.biomedcentral.com/series/ENSEMBL2010

Xosé M. Fernández-Suárez and Michael K. Schuster
Using the Ensembl Genome Server to Browse Genomic Sequence Data.

Giulietta M Spudich and Xosé M Fernández-Suárez
Touring Ensembl: A practical guide to genome browsing
Exploring the Ensembl genome browser

Let’s take a look at the Ensembl Genomes homepage at ensemblgenomes.org.

Click on the different taxa to see their homepages. Each one is colour-coded.
You can navigate most of the taxa in the same way as you would with Ensembl, but Ensembl Bacteria has a large number of genomes, so needs slightly different methods. Let’s look at it in more detail.
There’s no full species list for bacteria as it would be hard to navigate with the number of species. To find a species, start to type the species name into the species search box. A drop down list will appear with possible species.

For example, to find a substrain of *Clostridium difficile* type in *Clostridium d*. 

![Image of Ensembl Bacteria interface with search functionalities highlighted](image-url)
The drop down contains various strains of *Clostridium difficile*. Let’s choose *Clostridium difficile 630*. This will take us to another species homepage, where we can explore various features.

Now let’s have a look at the Ensembl Genome Protist. Click on the logo, this returns to EnsemblGenomes main site, click on protist, you should have a screen shot similar to the following.
Click on *Plasmodium Falciparum* under popular genomes. A dedicated page with link to gene annotation, sequence variation genome assembly statistics is found.
Demo: ENSEMBLgenomes in the context of an infectious disease, malaria

An international team of scientists, have identified the first reported inhibitors of a key enzyme namely Glucose 6 phosphate dehydrogenase (pfG6PG), involved in survival of the parasite responsible for malaria. Their findings, which may provide the basis for anti-malarial drug development, are currently published in the online version of the Journal of Medicinal Chemistry. Tropical malaria is responsible for more than 1.2 million deaths annually. Severe forms of the disease are mainly caused by the parasite Plasmodium falciparum, transmitted to humans by female Anopheles mosquitoes.
Reference: (http://www.euroclinix.net/malaria-transmission.html)
We will make use of the EBI services to find out more about the gene mentioned above, its location on the P. falciparum genome and some of its current annotations.

**EBI Transparent search engine**
A) Go to [www.ebi.ac.uk](http://www.ebi.ac.uk)

On the search box, type in PFG6PD and click search. (EBI search engine has been optimize to be biological term aware)
The search should return hits under the nucleotide sequence caption. We are interested in the nucleotide sequence of this gene, as this will be positioned later on the P. falciparum genome.

Click on the first hit (M80655),

**Exercise 1**
1.1 To what EBI service is this identifier linked?
Exercise 2: Some basic information about PFG6PD.

2.1 How long is this sequence in base pairs?
2.2 What is the sequence identifier of this gene?
2.3 To what taxonomic division does *P. falciparum* belong?

Note somewhere the sequence identifier of PFG6PD. This is needed by the Ensembl Genomes to retrieve its sequence from ENA prior the mapping to *P. falciparum* genome.

**ENSEMBL Genomes Sequence similarity search tools: BLAST**

Go to [http://ensemblgenomes.org/](http://ensemblgenomes.org/) and click on Protists (*Plasmodium* belongs to this taxa). Ensembl genome extends ENSEMBL across the taxonomic space.
Select Plasmodium falciparum from the search pull down menu.

We will make use of the sequence similarity search tool BLAST to locate PFG6PD on the plasmodium falciparum genome, click on BLAST.

Now we will need the PFG6PD sequence identifier saved earlier. Copy and paste the identifier into the Sequence identifier retrieval box. Ensembl will connect to ENA and fetch the sequences for us.
Exercise 3

3.1 What type of sequence is retrieved by ENSEMBL genome?

NB: Since we have nucleotide sequences make sure DNA database – LATESTGP and the search tool BLASTN are selected. You could click on RUN, however we will skip this as the result here: http://protists.ensembl.org/Multi/blastview/BLA_cD2EZmNh9
Click or copy and paste the above link into your browser. We are presented with the following screens. Note down the result name of our search, this is kept on EBI server for a maximum of 7 days.

Click on retrieve to retrieved a page with link(s) to our result

Click on view to see the results

The view result page is made up of three main blocks:
Alignment location vs. karyotype
The following blue arrow shows the direction of BLAST processing stages. The red colour of the result caption indicates we are at the result stage.

This shows that partial and/or full query sequence had match on 14 chromosomes with the best shown in red.

Exercise 4:
4.1 On what chromosome do we find the best match?

Alignment location vs. Query
The alignment location shows the high scoring pairs (HSP) on the forward strand, above grey-black bar (chromosome) as well as on the reverse strand (below the chromosome).

**Alignment summary**

Hover over the links to find out more information about them.

**Exercise 5:**

5.1 What is the length of the longest subject sequence match and what is its percentage
identity to the query sequence. (Hint: select >Length in Sort by and click refresh).

Click on A, to view the alignment between the query sequence and the subject sequences. 
Click on S to view the query sequence highlighted with matching bases for selected HSP and matching bases for other HSPs in selected hit.
Click on G to view the genome sequence with, location of selected alignment, location of other alignments and location of Exons.
Click on C to view the contig view that is subdivided into main 3 graphs.

The chromosome view

Scroll along the chromosome by clicking and dragging within the image. As you do this you’ll see the image below grey out and two blue buttons appear. Clicking on Update this image would jump the lower image to the region central to the scrollable image. We want to go back to where we started, so we’ll click on Reset scrollable image.
More on region in details

Click on the protein coding, a popup similar to this will show up, follow the link there in to find out more about the gene product function.

Exercise 6: PFG6PD annotation

6.1 What is the length of its protein sequence hit, its transcript sequence hit?
6.2 Is this description in accord with the initial gene name we used to search EBI at the beginning of this tutorial?
6.3 What is the source of the latter annotation?
6.4 What is the molecular function of this gene?
6.5 How was this gene molecular function assigned?

Hints (follow the source links on the description line from above link)

Customizing the region in details view
We customize next the contig view by adding known polymorphism track on the PFG6PD

Close the popup page by clicking on the tick mark on the top right.

The zoomed-in region in detail should contain SNP pooled out of dbSNP and look as follows

Exercise 7: Variation (Single nucleotide polymorphism, SNP)

7. What type of variant do you find (hint: legend)?
Quick Guide to Databases and Projects
Here is a list of databases and projects you will come across in these exercises. Google any of these to learn more. Projects include many species, unless otherwise noted.

Other help: The Ensembl Glossary: http://www.ensembl.org/Help/Glossary
FAQs: http://www.ensembl.org/Help/Faq

SEQUENCES EMBL---Bank, NCBI GenBank, DDBJ – Contain nucleic acid sequences deposited by submitters such as wet----lab biologists and gene sequencing projects. These three databases are synchronised with each other every day, so the same sequences should be found in each.

UniProtKB – the “Protein knowledgebase”, a comprehensive set of protein sequences. Divided into two parts: Swiss---Prot and TrEMBL

UniProt Swiss---Prot – the manually annotated, reviewed protein sequences in the UniProtKB. High quality.

UniProt TrEMBL – the automatically annotated, unreviewed set of proteins (EMBL---Bank translated). Varying quality.

PROTEIN SIGNATURES InterPro – A collection of domains, motifs, and other protein signatures. Protein signature records are extensive, and combine information from individual projects such as UniProt, along with other databases such as SMART, PFAM and PROSITE (explained below).

PFAM – A collection of protein families

PROSITE – A collection of protein domains, families, and functional sites.

SMART – A collection of evolutionarily conserved protein domains.

OTHER PROJECTS NCBI dbSNP – A collection of sequence polymorphisms; mainly single nucleotide polymorphisms, along with insertion---deletions.
Acknowledgement:

James Allen, Dan Bolser, Mikkel Christensen, Paul Davis, Paul Derwent, Lee Falin, Christoph Grabmueller, Malcolm Hinsley, Kevin Howe, Jay Humphrey, Arnaud Kerhornou, Julia Khobova, Eugene Kulesha, Nick Langridge, Dan Lawson, Ernesto Lowy, Mark McDowall, Uma Maheswari, Gareth Maslen, Michael Nuhn, Chuang Kee Ong, Michael Paulini, Helder Pedro, Anton Petrov, Dan Staines, Brandon Walts, Derek Wilson, Gary Williams