Computer lab

Ed Tobias

1. Browsing Ensembl Genome Browser for a particular gene
2. Find where gene is located on the chromosome
3. Looking at the Single Nucleotide Variants (SNVs) identified by the 1000 Genomes Project
4. How to find summary information about a particular gene in Ensembl
5. Get summary statistics for the gene
6. Using VEP (Variant Effect Predictor) to evaluate NGS-detected missense mutation
7. Interpreting SIFT and PolyPhen scores
8. 3D-visualisation of proteins

1. Browsing Ensembl Genome Browser for a particular gene

Go to Ensembl

Click on Ensembl Genome Browser

Change to GRCh37 genome build by clicking on “Still using GRCh37?”

to get:

Change “All Species” to Human
Enter specific gene name SOX9

Select first gene offered.

Scroll down then click on Region in Detail
2. Find where gene is located on the chromosome

The gene is shown in lower window. The top window shows where it is in relation to whole chromosome.
Middle window shows a 1 Mbp region containing it.

Try zooming out (Zoom out with the – button ) and also using Scroll controls to move left or right.

3. Have a look at the Single Nucleotide Variants (SNVs) identified by the 1000 Genomes Project

by scrolling down until you see something like this:

On the left hand side you should see 1KG/All SNPs/Indels.
This track represents the variants from the 1000 Genomes Project
4. How to find information about a particular gene in Ensembl

e.g. To look up human CFTR gene, find how many transcripts it has and find its genomic location (start and end position)

Choose GRCh37 build
Select human.
Enter gene name. CFTR.
Click on CFTR gene.
Click on Show Transcript Table.

5. Get summary statistics for the gene (e.g. size of gene in genome including introns, no. of coding exons, transcript length, no. of amino acids) by clicking on the Transcript ID of the first transcript (or the one with the NM and NP RefSeq numbers and may have a small golden rectangle in the Biotype column) and then scrolling down.

Click on transcript ID ENST 00000003084.6 – the one with the golden colour bar beneath Biotype: (CFTR-001)

<table>
<thead>
<tr>
<th>Transcript: CFTR-001 ENST00000003084.6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>Synonyms</strong></td>
</tr>
<tr>
<td><strong>Location</strong></td>
</tr>
</tbody>
</table>
| **About this transcript** | This transcript has 27 exons, is annotated with 19 domains and features, is associated with 1017 variants and maps to 19 splice sites.  
This transcript is a product of gene ENST00000003084.8. |  
| **Gene** |  |

![Transcript Table]

Choose transcript 001. Scroll down to see Summary.
See detailed diagram showing where the gene is located on the chromosome
Near the top of the page, click on the link to a genomic region that is given on the Location line (e.g. **Chromosome 7: 117,120,017-117,308,715**)

![Genomic Diagram](image-url)
Patient with distal arthrogryposis and a mutation detected by next generation DNA sequencing of: FBN2 gene p.C1323F and c.3968G>T ? pathogenic

6. USING VEP (Variant Effect Predictor) – very powerful & useful for e.g. analysing whether an amino acid substitution type of mutation is likely to damage the protein’s function or not.

To enter VEP, click on the Variant Effect Predictor box from the Ensembl home page:

Then launch VEP
Click on New Job

If not already in the “GRCh37” genome build, change it to GRCh 37 by clicking on: Go to GRCh37 website
Enter a name for the query. This is optional but helps when you have more than one query running at the same time – which is quite possible!

In the Paste Data box, enter the ENST identifier \textbf{(ENST00000262464)} followed by the coding sequence description for the mutation. For the FBN2 gene \textit{p.C1323F} mutation that we’re going to analyse, the coding sequence mutation description is actually \textit{c.3968G>T}

Therefore, in the box labelled as “Either paste data” enter this:
\textbf{ENST00000262464:c.3968G>T}

making sure not to leave any space before this or after this in the box, as it upsets the program!

Now, if offered the opportunity, click on “Instant results” for first variant to see if it works OK. You should get:
Close that results box.

Now, to get the full (non-instant) analysis, tick various frequency data boxes. Then click on Run, lower down the page.

Once the job is “done”, click on View results and get this output:
and scroll to the right to see the SIFT and Polyphen results:

The top line is for the transcript of interest.

You can see

- which exon number contains the mutation (i.e. exon 30)
- the total number of exons (65 exons).
- the amino acid change - from C (cysteine) to F (phenylalanine)
- the SIFT score (0 in this case). [SIFT is the online tool for evaluating a missense mutation that is based on sequence alignments and is run on the Craig Venter Institute’s server].
- the PolyPhen Score (0.854 in this case). [PolyPhen is the Polymorphism Phenotyping online software that is freely accessible, running on a server at Harvard Medical School. It is based on sequence alignments and also on the physical properties of amino acids.]
7. SIFT and PolyPhen score interpretation:

- from ThermoFisher at

SIFT and PolyPhen scores predict whether an amino acid substitution is likely to affect protein function. NB PolyPhen-2 and SIFT scores both use the same range, 0.0 to 1.0, but with opposite meanings. A variant with a PolyPhen score of 0.0 is predicted to be benign but a variant with a SIFT score of 1.0 is predicted to be benign.

The SIFT score ranges from 0.0 (deleterious) to 1.0 (tolerated). The score can be interpreted as follows:
- 0.0 to 0.05 -- Variants with scores in this range are considered deleterious. Variants with scores closer to 0.0 are more confidently predicted to be deleterious.
- 0.05 to 1.0 -- Variants with scores in this range are predicted to be tolerated (benign). Variants with scores very close to 1.0 are more confidently predicted to be tolerated.

The PolyPhen-2 score ranges from 0.0 (tolerated) to 1.0 (deleterious). Variants with scores of 0.0 are predicted to be benign. Values closer to 1.0 are more confidently predicted to be deleterious. The score can be interpreted as follows:
- 0.0 to 0.15 -- Variants with scores in this range are predicted to be benign.
- 0.15 to 1.0 -- Variants with scores in this range are possibly damaging.
- 0.85 to 1.0 -- Variants with scores in this range are more confidently predicted to be damaging.
8. 3D visualisation of proteins

Type PDB into Google and go to [www.rcsb.org](http://www.rcsb.org)

Once at the database search for CFTR NBD
Then select Homo sapiens from Filter or Refinements on the left

Look for structure with reference 2PZE
i.e. the one that is named:

**Minimal human CFTR first nucleotide binding domain as a head-to-tail dimer**

Then click on the very small box labelled “3D view”, just beneath the molecule picture for 2PZE.
Scroll down and select Jsmol viewer (or, if that fails, Jmol or NGL).
If you're asked if you want to run it (in a message at the bottom of the screen), choose Run This Time.
Try to find the phenylalanine amino acid (F508) that is often deleted in mutant CF alleles.

In NGL Viewer: amino acid & number is shown in top left of screen:

Now try looking at the structure for the same molecule but containing the F508 mutation (ie 2PZF) and see if you can notice any major difference in the structure.