

Extracting genomic information from Ensembl (Practice)

Rafael Torres-Perez

- Bioinformatics for Genomics and Proteomics (BioinfoGP) -

1. Genome level.

1.1. Guided exercise: Retrieve the human genome sequence and its annotations.

Note: we will be working always with the most updated version of the genomes.

1.1.a. Access the Ensembl initial page. Which is the current Ensembl release number and which is the current human genome assembly code?

Access “*information and statistics*” for the current assembly of *Homo sapiens* and answer:

- How many coding and non-coding genes are annotated in the primary assembly?
- How many transcripts are annotated on average for each coding gene?
- How many variants have been annotated in the human genome?

1.1.b. Locate the “reference” (DNA sequence file in FASTA format) for the human genome (there are several ways to reach this information). Among all the fastas, select the “toplevel” file and download it. (NOTE: the human reference file is very big, you are allowed cancel the download before it ends or delete it after this practice ends ;)

1.1.c. Locate and download the annotations file corresponding to this reference genome version in GFF3 format. How would you be sure of the exact correspondance of both files, the FASTA and the GFF3 file?

1.2. Additional exercise: Retrieve the genome sequence of a particular baker’s yeast and its annotations (Ensembl Genomes).

1.2.1. Access the *information and statistics* for the genome of *Saccharomyces cerevisiae*, strain R64 , version R64-1-1 and answer:

- Who provided the assembly and who the annotations of the genome of this species?
- Which is the size in kilobases of this genome assembly?
- How many transcripts are there on average per coding gene? (compare with the number obtained for *Homo sapiens* in 1.1.a.)

1.2.2. Download the more updated reference genome of *Saccharomyces cerevisiae*, strain R64.

1.2.3. Download the corresponding gff file.

1.2.4. Check if the files from 1.2.2 and 1.2.3 are the same that those you downloaded in a previous session of this series of seminars.

2. Gene level.

Let's access the information about the human gene **ACE2** (official name).

2.1. Disambiguation (*Did you mean...?*) page.

What is its Ensembl stable ID? Learn the logic of this ID... See other suggestions in this page. What "biotype" is deduced for the element identified by ENST00000427411? Can you guess which is the species for the id ENSMUSG00000015405 in the same page? Click the option for the human ACE2 gene.

2.2. Gene Tab

What is the description of this gene? Where is it located in the genome?

Hide the *Transcripts Table* by now (blue button).

Let's visualize the transcripts in the *Overview Browser* at the end of the page. How many contigs cover the gene region? Is ACE2 gene in the forward strand or in the reverse strand? How many transcripts are annotated for ACE2? How many of them code for protein?

Choosing the "best" transcript to be used later: Open the *Transcript Table* (blue button) and determine which transcript is more adequate to set as "more biologically relevant" (Hint: see the *Flags* column).

Lateral panel.

- Functional annotations (Gene Ontology [GO] ontology): Access "*Biological process*" and check if there is a relation of ACE2 with viruses. Choose one of these virus-related GO terms and determine which other human genes share this viral-related function with ACE2 (Hint: "*Search BioMart*" link).

- Determine in the *Variation Table* probably damaging variations located in ACE2. A common criteria to choose this kind of variations is that 3 or more predictors assess a non-benign effect of the mutation. By clicking some right-most column corresponding to the predictors (MutationAssesor, SIFT, POLYPHEN...) you can sort the table by the impact score. Red color values indicates more severe effects. Click on the *dbSNP id* of one of these probably damaging variants and answer: Where is it located? Which is the highest frequency of the minor allele (MAF) of this variant in a population? How many sources of evidence support this variant?

- *Gene Expression*. Filter (button) the data by *Organ* selecting only those with "high" expression and find which the organ with more transcriptomic expression in GTEx platform. [You may want later to search the organ where it is more expressed another major player interacting with SARS-Cov-2 Spike protein, the gene *TMPRSS2*].

2.3. “Region in detail” tab: browsing regulatory elements and the constituents of the transcript.

Select a region in the promotor zone of ACE2 in the browser and zoom in. Search in the “*regulatory build*” track a possible regulatory element in this promotor region (a promotor, enhancer, transcription factor binding site...). Click on this element and determine its Ensembl ID and location.

2.4. Gene constituents (Transcript tab): Return to the Gene tab and show the *Transcript table*. Click on the ID of the transcript selected in 2.2. to access the *Transcript tab*. How many exons does it have? How many residues has the resulting protein?

Click on “*Exons*” in the lateral panel and examine the genomic sequence of this transcript in the browser. Variations information is unnecessary by now, so click on “*configure this page*” and remove it (*Show variants: no*). Now, answer: Which is the Ensembl ID of the first exon? What does it mean the code of colours in this exon sequence?

Download the sequence (click on the corresponding button) in FASTA format ignoring the flanking region. Now click on “*Proteins*” (lateral panel) and now download the FASTA aminoacidic sequence of the protein. Idem for the “*cDNA*” sequence removing previously the variants info. Try to understand what does it mean each of the three lines that are displayed.

3. Intermediate (customized) level: Using Ensembl BioMart.

3.1. Guided exercise: Retrieve a set of genes and sequences searching a region.

Williams-Beuren syndrome is a rare genetic disorder that affects an estimated 1 in 7,500 to 10,000 people. The condition is characterized by mild to moderate intellectual disability or learning problems and cardiovascular problems. It is caused by a heterozygous deletion in a region of the chromosome 7 that can cover the range chr7:72800000-74900000 (GRCh38 assembly)¹.

3.1.a. Download a .tsv file containing a table with the following columns, in the mentioned order:

- the name and genomic coordinates of the genes that can be affected by the deletion.
- the Gene Ontology Terms (ids and names) associated with those genes.
- the phenotypes annotated.
- NCBI ID.

3.1.b. Next, obtain also the genomic sequence of these genes in a single multi-FASTA file. Note that it is not possible with BioMart to annex this information to the previous file (you will need to create a new file, but you don’t need to create a strictly new query... :).

1 Adapted from Schubert and Lacone (1996) <https://doi.org/10.3892/ijmm.18.5.799> and from <https://ghr.nlm.nih.gov/condition/williams-syndrome>

3.2. Additional exercise: Retrieve the information in Ensembl Genomes BioMart from a set of genes.

In a previous seminar, it was obtained a list of differentially expressed genes when comparing “p5c5” and “p5c0.04” conditions in *Saccharomyces cerevisiae* R64-1-1. This is a list of IDs of some genes that the analysis showed as having high absolute fold change:

YOL155C
YKR097W
YDR345C
YLR377C
YAR035W
YKL029C
YKL043W
YGL032C
YLR142W
YIL015W
YNR044W
YCL048W-A
YIR016W
YDL218W
YKL163W
YJL153C
YDR461W
YBR068C
YNR002C
YDL214C
YPL187W
YNL036W
YPL058C

3.2.a. Using this gene list as input (you can cut-and-paste it), obtain a table (.tsv file) of the genes that are annotated with any GO Term containing the word “membrane”. The file must hold the following information, in the order of citation:

- Gene ID
- Gene Name
- Genomic coordinates
- Strand
- GO Term ID
- GO Term name
- Transmembrane helices (Protein features): identifier, start and end positions.
- Ensembl Gene ID (Gene stable ID)
- NCBI gene ID
- Uniprot/Swiss-Prot ID

3.2.b. Finally, obtain a list of the *germline variants* of this subset of genes related to “membrane”. How many genes associated to membrane have at least one germline variant annotated in release R64-1-1?

Extract now a tsv file with the variants. It should contain (in your preferred order): *Gen Stable ID*, *Transcript stable ID*, *Variant Name*, *Minor allele frequency*, *Variant alleles*, *Transcript location*, *Polyphen prediction* and *Sift prediction* and *Variant consequence*.

Check the resulting file: Are all the downloaded columns meaningful? Adapt the query and the resulting table according to your impressions.