Extracting genomic information from Ensembl (Practice and Solutions)

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.
Note: Results are based on Ensembl Release 99 and EnsemblFungi Release 46

1.1.a. Access the Ensembl initial page. Which is the current Ensembl release number and which is the current human genome assembly code?
Ensembl release 99, Human genome assembly GCRh38.p13. In the releases, Ensembl compiles received items (new assemblies, new patches of pre-existent assemblies, variants, regulatory builds) and reprocesses this info to produce new annotations (in human and mouse), new homology comparisons, etc. A genome assembly is a computational representation of the sequence of a genome. This representation is modified from time to time based on patches that modify or add new pieces to the previous assembly version.

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? *20,449 (including 657 readthrough genes. A readthrough gene is a gene that only has transcripts with exons that overlap with exons of other gene [readthrough transcripts]).*
  A note: a participant in the streaming session asked if Ensembl and GENCODE have the same annotation set (in the sense of sharing the sames genes, transcripts). This link from Ensembl clarifies the question (the sort answer is yes…): https://m.ensembl.org/Help/Faq?id=303
- How many transcripts are annotated on average for each coding gene? *11 transcripts aprox.*
- How many variants have been annotated in the human genome? *678,017,608 short variants and 6,073,012 structural variants (find the detailed classification of variants in* https://www.ensembl.org/info/genome/variation/prediction/classification.html*)*
-

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 ;)
You can download the FASTA: a) Selecting the species in the "All genomes" drop-down list or, if "favourite species", clicking its name directly. Next, find the Gene annotation section and click on Download FASTA. Next, in the FTP folder system, select "dna" and finally, select the Homo_sapiens.GRCh38.dna.toplevel.fa.gz (choosing the primary assembly file is also admisible). b)

Main page → Download (at the top) → Download databases or Download data via FTP → Single species data → for Human, click the DNA (fasta) correponding link → dna folder → toplevel fasta file.

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?
Similar to getting the FASTA. For option a), in Gene annotation section click Download GFF3 . For option b) select Gene sets → GFF3. In both options, select at the end the Homo_sapiens.GRCh38.99.gff3.gz file (or the Homo_sapiens.GRCh38.99.chr.gff3.gz file)

1.2. Additional exercise: Retrieve the genome sequence of a particular baker's yeast and its annotations (Ensembl Genomes). http://www.ensemblgenomes.org → https://fungi.ensembl.org/

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? *"The assembly provided on this site is the R64-1-1 assembly, imported from the Saccharomyces Genome Database (SGD)". "The protein-coding and non-coding gene model annotation was imported from the Saccharomyces Genome Database (SGD)" (In human and mouse, annotation is done by Ensembl).*
- Which is the size in kilobases of this genome assembly? 12,157,105 base pairs
- How many transcripts are there on average per coding gene? (compare with the number obtained for *Homo sapiens* in 1.1.a.) *1.08. So, differently from human, you can expect one only transcript represented for each yeast gene.*

1.2.2. Download the more updated reference genome of *Saccharomyces cerevisiae*, strain R64. *Similar to 1.1.b*

1.2.3. Download the corresponding gff file. *Similar to 1.1.c*

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. Saccharomyces_cerevisiae.R64-1-1.dna.toplevel.fa.gz & Saccharomyces_cerevisiae.R64-1-1.99.gff3.gz

2. Gene level.

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

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

Which is its Ensembl stable ID? ENSG00000130234
Learn the logic of this ID… ENS = Ensembl id, G = biotype "**g**ene" 00000130234 = numerical id of the gene.
See other suggestions in this page. What "biotype" is deduced for the element identified by ENST00000427411? Biotype "Transcript" (ENS-**T**-00000427411)

Can you guess which is the species for the id ENSMUSG00000015405 in the same page? *ENS-**MUS**-G-00000015405, It's a gen from MUS musculus.* Click the accession for the human ACE2 gene.

2.2. *Gene Tab*

Which is the description of this gene? *angiotensin I converting enzyme 2.* Where is it located in the genome? *Chromosome X: 15,561,033-15,602,148 (reverse strand)*

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? *One (accession AC097625.11)* Is ACE2 gene in the forward strand or in the reverse strand? *In the reverse strand: it is below the contig track and its transcripts hold the < sign*. How many transcripts are annotated for ACE2? *Five.* How many of them code for protein? *Two (the red ones, ACE-202 and ACE-201).*

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). *ACE2-201, as it's tagged with MANE indicator (https://www.ensembl.org/info/genome/genebuild/mane.html) in addition to the tag APPRIS P1. Note that ACE2-201 is not the longest transcript in the annotation… Note also that you will find MANE only for human genome. APPRIS is only for human, mouse, rat, zebrafish and pig (sus scrofa).*

Lateral panel.

- Functional annotations (Gene Ontology [GO] ontology): Access "*Biological process*" and check if there is a relation of ACE2 with viruses. There are several GO Terms virus-related: *viral process (evidende: electronic annotation), viral entry into host cell (traceable author statement), receptor-mediated virion attachment to host cell (inferred directly from assay).* 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). *ACE2 shares the GO Term 'viral entry into host cell' with genes CXCR4, CTSB, VAMP8, CDHR3, LAMP1, CD80, ICAM1, CLEC4M and SCARB1.*

- 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? Chosen rs1019324840 dbSNP id variant, located in chromosome X:15572296. Which is the highest frequency of the minor allele (MAF) of this variant in a population? < 0.01 (less of 0.01, the variant is considered "rare"). How many sources of evidence support this variant? Two (frequency and TOPmed).

- *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. Testis [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. *For example, ENS**R**00000902030 located in chromosome X:15,606,401-15,608,200 is a promoter flanking region. This type of regulatory elements are coloured in light red.*

***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? *18* How many residues has the resulting protein? *805*

Click on "*Exons*" in the lateral panel and examine the genomic sequence of this transcript in the browser. Variations information is unneccesary 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? *ENSE00003897519* What does it means the code of colours in this exon sequence? *5´-UTR in orange and coding part of the exon in blue.*

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. *First, he transcript sequence (cDNA); second, the coding sequence underneath it, and third, the protein sequence.*

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

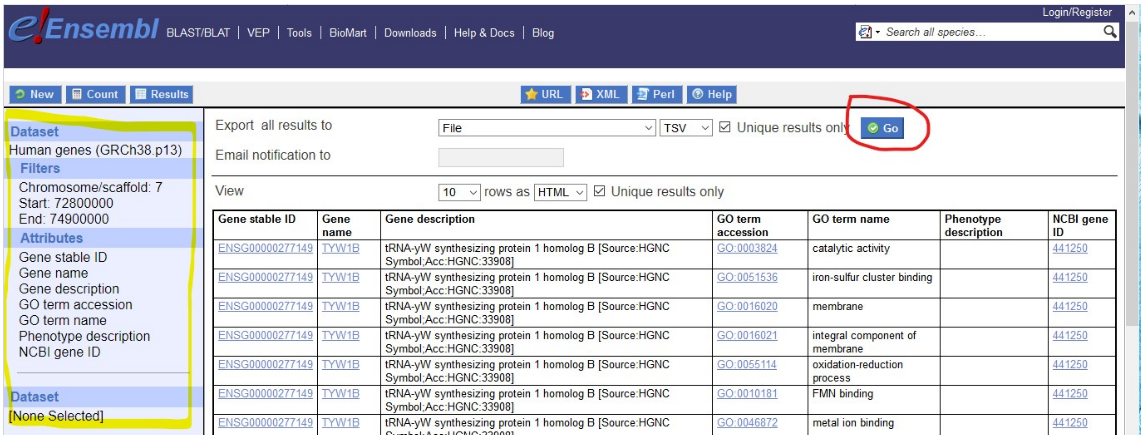
<u>**3.1.** Guided exercise: Retrieve a set of genes and sequences searching a region.</u>

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 <u>chr7:72800000-74900000</u> (GRCh38 assembly)[1].

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

- the Ensembl ID, name, description 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.

[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



Dataset
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 7
Start: 72800000
End: 74900000
Attributes
Gene stable ID
Gene name
Gene description
GO term accession
GO term name
Phenotype description
NCBI gene ID
Dataset
[None Selected]

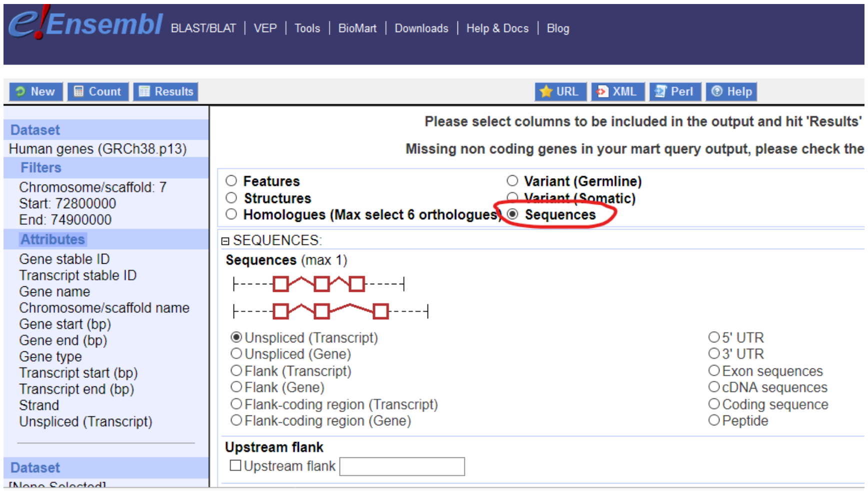
Export all results to: File | TSV | ☑ Unique results only

Email notification to

View 10 rows as HTML ☑ Unique results only

Gene stable ID	Gene name	Gene description	GO term accession	GO term name	Phenotype description	NCBI gene ID
ENSG00000277149	TYW1B	tRNA-yW synthesizing protein 1 homolog B [Source:HGNC Symbol;Acc:HGNC:33908]	GO:0003824	catalytic activity		441250
ENSG00000277149	TYW1B	tRNA-yW synthesizing protein 1 homolog B [Source:HGNC Symbol;Acc:HGNC:33908]	GO:0051536	iron-sulfur cluster binding		441250
ENSG00000277149	TYW1B	tRNA-yW synthesizing protein 1 homolog B [Source:HGNC Symbol;Acc:HGNC:33908]	GO:0016020	membrane		441250
ENSG00000277149	TYW1B	tRNA-yW synthesizing protein 1 homolog B [Source:HGNC Symbol;Acc:HGNC:33908]	GO:0016021	integral component of membrane		441250
ENSG00000277149	TYW1B	tRNA-yW synthesizing protein 1 homolog B [Source:HGNC Symbol;Acc:HGNC:33908]	GO:0055114	oxidation-reduction process		441250
ENSG00000277149	TYW1B	tRNA-yW synthesizing protein 1 homolog B [Source:HGNC Symbol;Acc:HGNC:33908]	GO:0010181	FMN binding		441250
ENSG00000277149	TYW1B	tRNA-yW synthesizing protein 1 homolog B [Source:HGNC	GO:0046872	metal ion binding		441250

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… :).


Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
New
Count
Results
URL
XML
Perl
Help

Dataset
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 7
Start: 72800000
End: 74900000
Attributes
Gene stable ID
Transcript stable ID
Gene name
Chromosome/scaffold name
Gene start (bp)
Gene end (bp)
Gene type
Transcript start (bp)
Transcript end (bp)
Strand
Unspliced (Transcript)
Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results'

Missing non coding genes in your mart query output, please check the

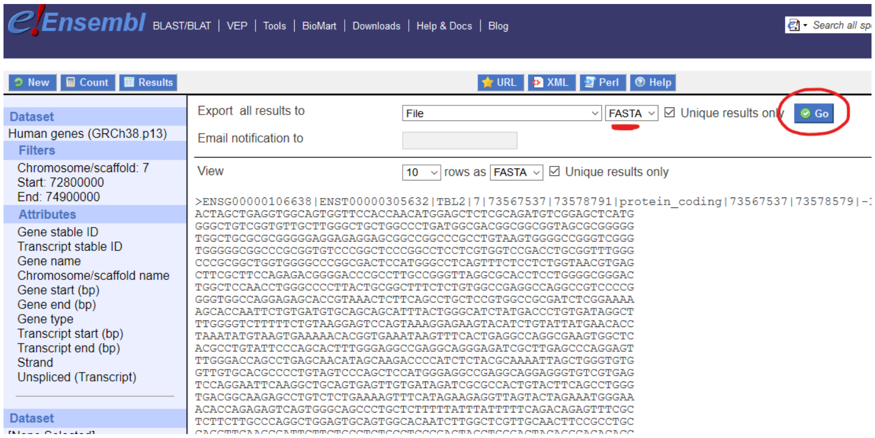
○ Features ○ Variant (Germline)
○ Structures ○ Variant (Somatic)
○ Homologues (Max select 6 orthologues) ● Sequences

⊟ SEQUENCES:

Sequences (max 1)

● Unspliced (Transcript) ○ 5' UTR
○ Unspliced (Gene) ○ 3' UTR
○ Flank (Transcript) ○ Exon sequences
○ Flank (Gene) ○ cDNA sequences
○ Flank-coding region (Transcript) ○ Coding sequence
○ Flank-coding region (Gene) ○ Peptide

Upstream flank
□ Upstream flank


Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Search all sp
New
Count
Results
URL
XML
Perl
Help
Go

Dataset
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 7
Start: 72800000
End: 74900000
Attributes
Gene stable ID
Transcript stable ID
Gene name
Chromosome/scaffold name
Gene start (bp)
Gene end (bp)
Gene type
Transcript start (bp)
Transcript end (bp)
Strand
Unspliced (Transcript)
Dataset

Export all results to: File | FASTA | ☑ Unique results only

Email notification to

View 10 rows as FASTA ☑ Unique results only

```
>ENSG00000106638|ENST00000305632|TBL2|7|73567537|73578791|protein_coding|73567537|73578579|-1
ACTAGCTGAGGTGGCAGTGGTTCCACCAACATGGAGCTCTCGCAGATGTCGGAGCTCATG
GGGCTGTCGGTGTTGCTTGGGCTGCTGGCCCTGATGGCGACGGCGGCGGTAGCGCGGGGG
TGGCTGCGCGCGGGGGAGGAGAGGAGCGGCCGGCCCGCCTGTAAGTGGGGCCGGGTCGGG
TGGGGGCGGCCCGCGGTGTCCCGGCTCCCGGCCTCCTCGTGGTCCGACCTGCGGTTTGGG
CCCGCGGCTGGTGGGGCCCGGCGACTCCATGGGCCTCAGTTTCTCCTCTGGTAACGTGAG
CTTCGCTTCCAGAGACGGGGACCCGCCTTGCCGGGTTAGGCGCACCTCCTGGGGCGGGAC
TGGCTCCAACCTGGGCCCCTTACTGCGGCTTTCTCTGTGGCCGAGGCCAGGCCGTCCCCG
GGGTGGCCAGGAGAGCACCGTAAACTCTTCAGCCTGCTCCGTGGCCGCGATCTCGGAAAA
AGCACCAATTCTGTGATGTGCAGCAGCATTTACTGGGCATCTATGACCCTGTGATAGGCT
TTGGGGTCTTTTTCTGTAAGGAGTCCAGTAAAGGAGAAGTACATCTGTATTATGAACACC
TAAATATGTAAGTGAAAAACACGGTGAAATAAGTTTCACTGAGGCCAGGCGAAGTGGCTC
ACGCCTGTATTCCCAGCACTTTGGGAGGCCGAGGCAGGGAGATCGCTTGAGCCCAGGAGT
TTGGGACCAGCCTGAGCAACATAGCAAGACCCCATCTCTACGCAAAATTAGCTGGGTGTG
GTTGTGCACGCCCCTGTAGTCCCAGCTCCATGGGAGGCCGAGGCAGGAGGGTGTCGTGAG
TCCAGGAATTCAAGGCTGCAGTGAGTTGTGATAGATCGCGCCACTGTACTTCAGCCTGGG
TGACGGCAAGAGCCTGTCTCTGAAAAGTTTCATAGAAGAGGTTAGTACTAGAAATGGGAA
ACACCAGAGAGTCAGTGGGCAGCCCTGCTCTTTTTATTTATTTTTCAGACAGAGTTTCGC
TCTTCTTGCCCAGGCTGGAGTGCAGTGGCACAATCTTGGCTCGTTGCAACTTCCGCCTGC
```

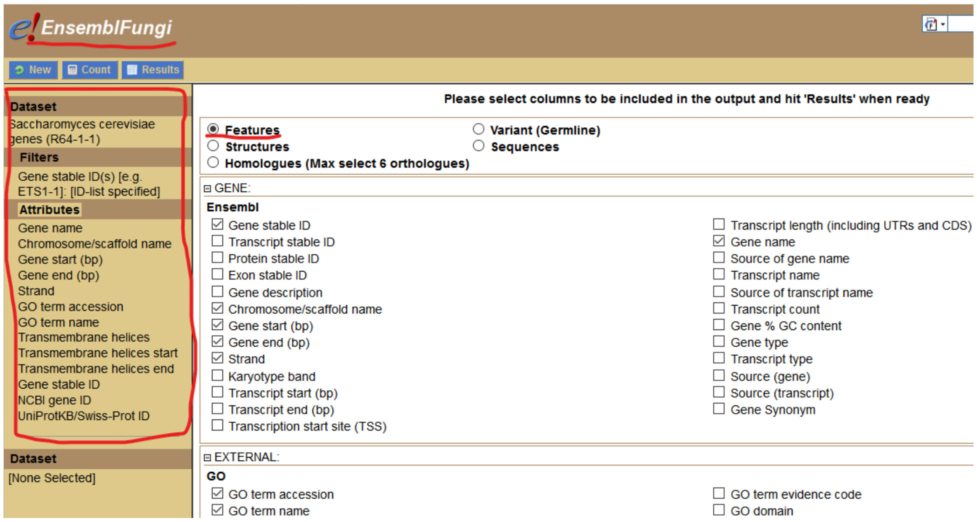
3.2. Additional exercise: Retrieve the information in Ensembl Genomes BioMart from a set of genes.

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



Dataset
Saccharomyces cerevisiae genes (R64-1-1)

Filters
Gene stable ID(s) [e.g. ETS1-1]: [ID-list specified]

Attributes
Gene name
Chromosome/scaffold name
Gene start (bp)
Gene end (bp)
Strand
GO term accession
GO term name
Transmembrane helices
Transmembrane helices start
Transmembrane helices end
Gene stable ID
NCBI gene ID
UniProtKB/Swiss-Prot ID

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

● **Features**
○ **Structures**
○ **Homologues (Max select 6 orthologues)**
○ **Variant (Germline)**
○ **Sequences**

⊟ GENE:

Ensembl

☑ Gene stable ID
☐ Transcript stable ID
☐ Protein stable ID
☐ Exon stable ID
☐ Gene description
☑ Chromosome/scaffold name
☑ Gene start (bp)
☑ Gene end (bp)
☑ Strand
☐ Karyotype band
☐ Transcript start (bp)
☐ Transcript end (bp)
☐ Transcription start site (TSS)
☐ Transcript length (including UTRs and CDS)
☑ Gene name
☐ Source of gene name
☐ Transcript name
☐ Source of transcript name
☐ Transcript count
☐ Gene % GC content
☐ Gene type
☐ Transcript type
☐ Source (gene)
☐ Source (transcript)
☐ Gene Synonym

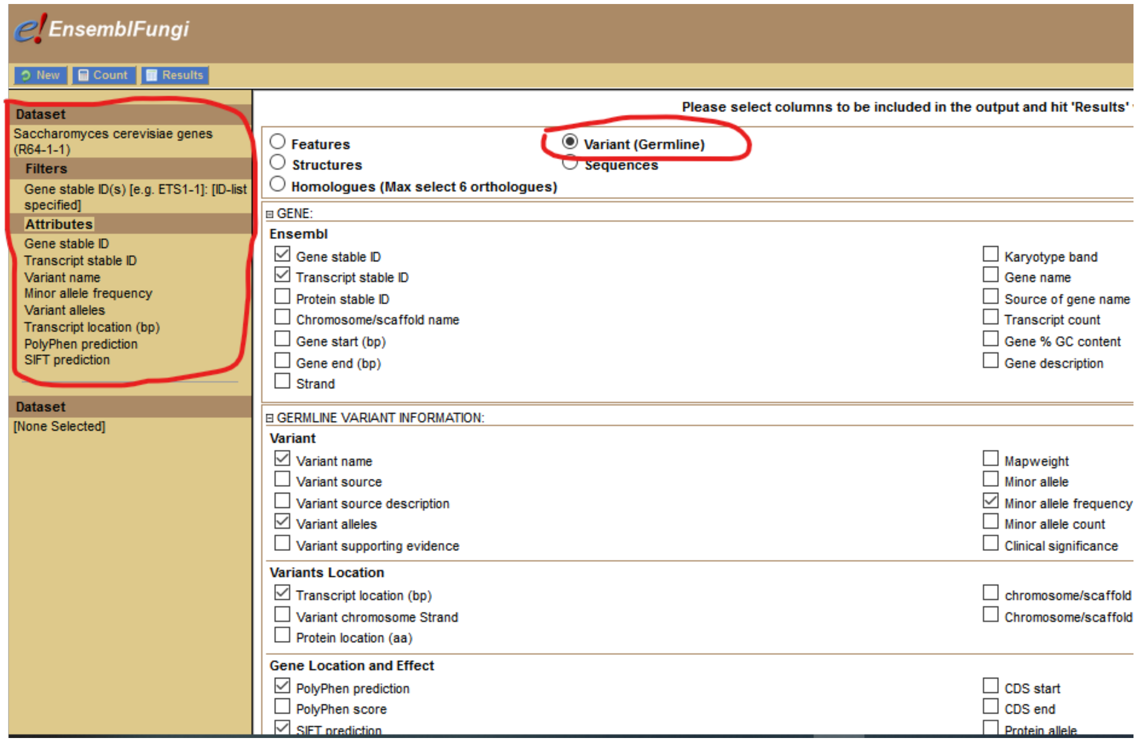
⊟ EXTERNAL:

GO

☑ GO term accession
☑ GO term name
☐ GO term evidence code
☐ GO domain

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*.



Dataset
Saccharomyces cerevisiae genes (R64-1-1)
Filters
Gene stable ID(s) [e.g. ETS1-1]: [ID-list specified]
Attributes
Gene stable ID
Transcript stable ID
Variant name
Minor allele frequency
Variant alleles
Transcript location (bp)
PolyPhen prediction
SIFT prediction

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results'

○ Features ● Variant (Germline)
○ Structures ○ Sequences
○ Homologues (Max select 6 orthologues)

⊟ GENE:

Ensembl

☑ Gene stable ID ☐ Karyotype band
☑ Transcript stable ID ☐ Gene name
☐ Protein stable ID ☐ Source of gene name
☐ Chromosome/scaffold name ☐ Transcript count
☐ Gene start (bp) ☐ Gene % GC content
☐ Gene end (bp) ☐ Gene description
☐ Strand

⊟ GERMLINE VARIANT INFORMATION:

Variant

☑ Variant name ☐ Mapweight
☐ Variant source ☐ Minor allele
☐ Variant source description ☑ Minor allele frequency
☑ Variant alleles ☐ Minor allele count
☐ Variant supporting evidence ☐ Clinical significance

Variants Location

☑ Transcript location (bp) ☐ chromosome/scaffold
☐ Variant chromosome Strand ☐ Chromosome/scaffold
☐ Protein location (aa)

Gene Location and Effect

☑ PolyPhen prediction ☐ CDS start
☐ PolyPhen score ☐ CDS end
☑ SIFT prediction ☐ Protein allele

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

The table doesn't contain any value in the column for Minor allele frequency, so probably it should be removed from the table design.