Data mining with Ensembl Biomart (answers to questions)

Stéphanie Le Gras
(slegras@igbmc.fr)

Exercise 1: get annotations of a gene

- 1.
  - Click on Filters (left panel),
  - Expand the “GENE” section
  - Select “Input external references ID list”, select “Gene Name(s)” in the drop down list and enter IDH1.
  - Click on Count in the top left panel. You should get 1/68005 Genes
  - Click on Attributes (left menu)
  - Select “Features” (selected by default)
  - Select Gene stable ID, Transcript stable ID and Gene Name
  - Click on Results (top left menu)

Gene stable ID	Transcript stable ID	Gene name
ENSG00000138413	ENST00000345146	IDH1
ENSG00000138413	ENST00000446179	IDH1
ENSG00000138413	ENST00000415913	IDH1
ENSG00000138413	ENST00000484575	IDH1
ENSG00000138413	ENST00000415282	IDH1
ENSG00000138413	ENST00000462386	IDH1
ENSG00000138413	ENST00000417583	IDH1
ENSG00000138413	ENST00000451391	IDH1
ENSG00000138413	ENST00000481557	IDH1

  - 9 transcripts are found

Exercise 1: get annotations of a gene

- 2.
  - You can leave the Dataset and Filters the same, and go directly to the Attributes section
  - Click on Attributes (left panel)
  - Select “Sequences”
  - Expand the SEQUENCES section
  - Select Exon sequences
  - Expand “Header Information”
  - Unselect “Gene stable ID” (Gene Information)
  - Select Gene name (Gene Information), transcript stable IDs (Transcript Information) and Exon stable IDs (Exon Information).
  - Click on Results
- 3.
  - You can leave the Dataset and Filters the same, and go directly to the Attributes section
  - Click on Attributes (left panel)
  - In the SEQUENCES section
  - select Coding sequence
  - “Header Information”: unselect Gene name (Gene Information) and select transcript stable ID (Transcript Information) and Exon stable IDs (Exon Information).
  - Click on Results

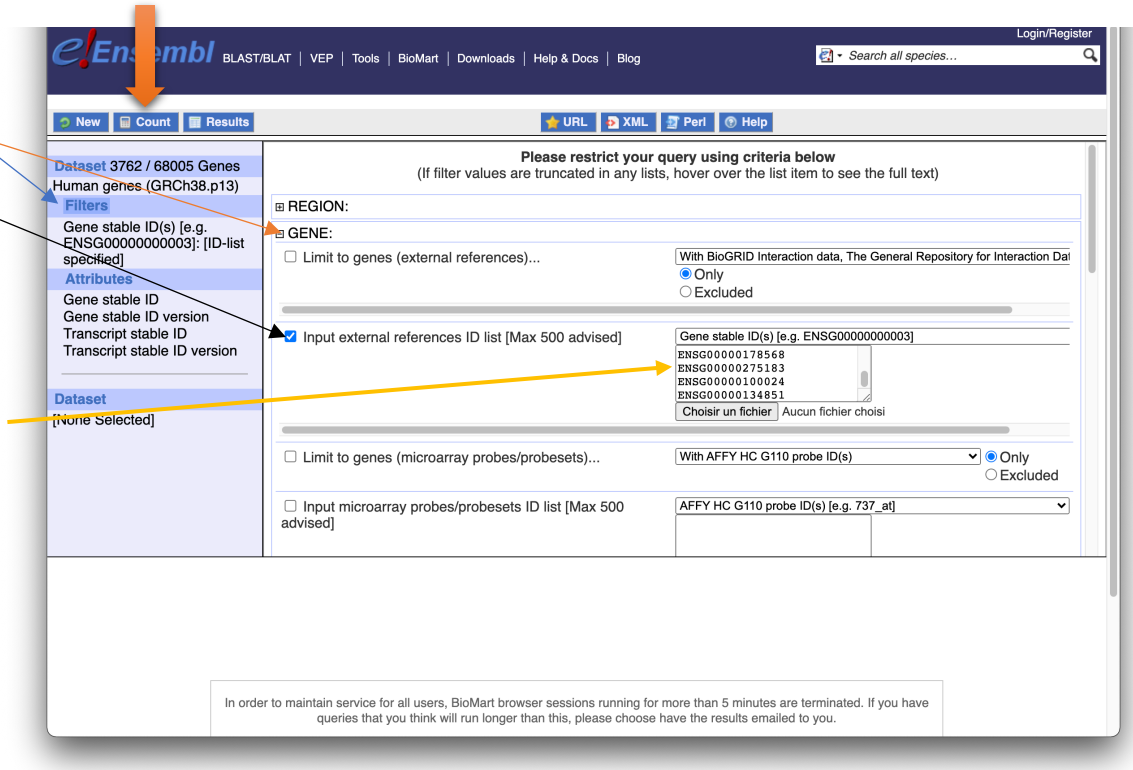
Exercise 1: get annotations of a gene

- 4.
  - You can leave the Dataset and Filters the same, and go directly to the Attributes section
  - Click on Attributes (left panel)
  - Select “Features” (selected by default)
  - In the GENE section: Gene stable ID, Transcript stable ID and Gene Name should be selected
  - Expand the EXTERNAL section
  - Select GO Term Name, GO domain and GO Term Accession
  - Click on Results
- 5.
  - You can leave the Dataset and Filters the same, and go directly to the Attributes section
  - Click on Attributes (left panel)
  - Select “Variant (Germline)”
  - In the GENE section: Gene stable ID, Transcript stable ID and Gene Name should be selected
  - Expand the GERMLINE VARIANT INFORMATION section
  - Select Variant Name, Variant Alleles, Minor allele frequency, Chromosome/scaffold name, Chromosome /scaffold position start (bp), Chromosome/scaffold position end (bp), Variant Consequence
  - Click on Results

Exercise 2: get annotations for a set of genes

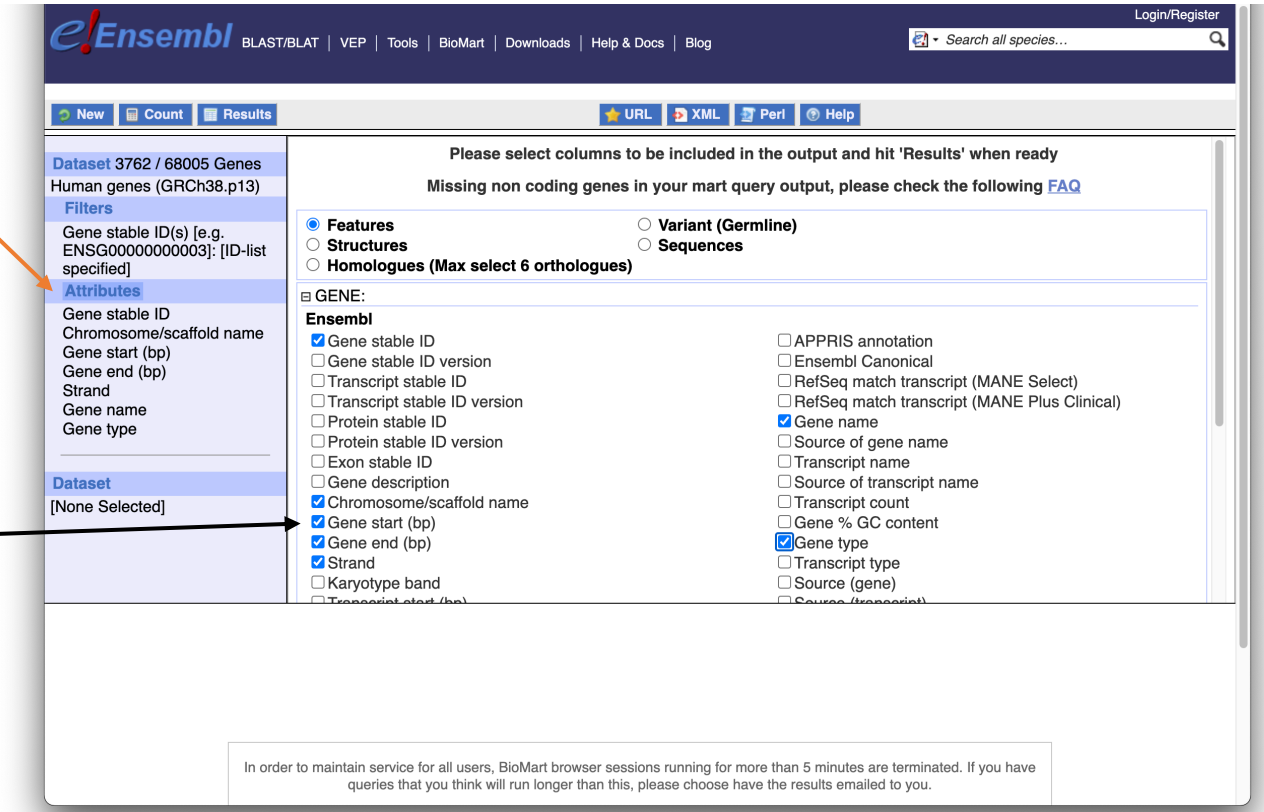
- 2.
  1. In Ensembl/BioMart, create a new request
  2. Click on Filters (left panel)
  3. Expand the GENE section
  4. Select “Input external references ID list” and select “Gene stable ID(s)” in the drop down list
  5. Open the file siMitfvssiLuc.up.txt in Excel and copy the content of the first column (ENSG**) without the title and paste it all into the text field (Input external references ID list) of the Ensembl Biomart filter page
  6. Click on “Count” (top left button). You should have the number of genes you have in your file generated by SARTools: 3762 (6)

6.



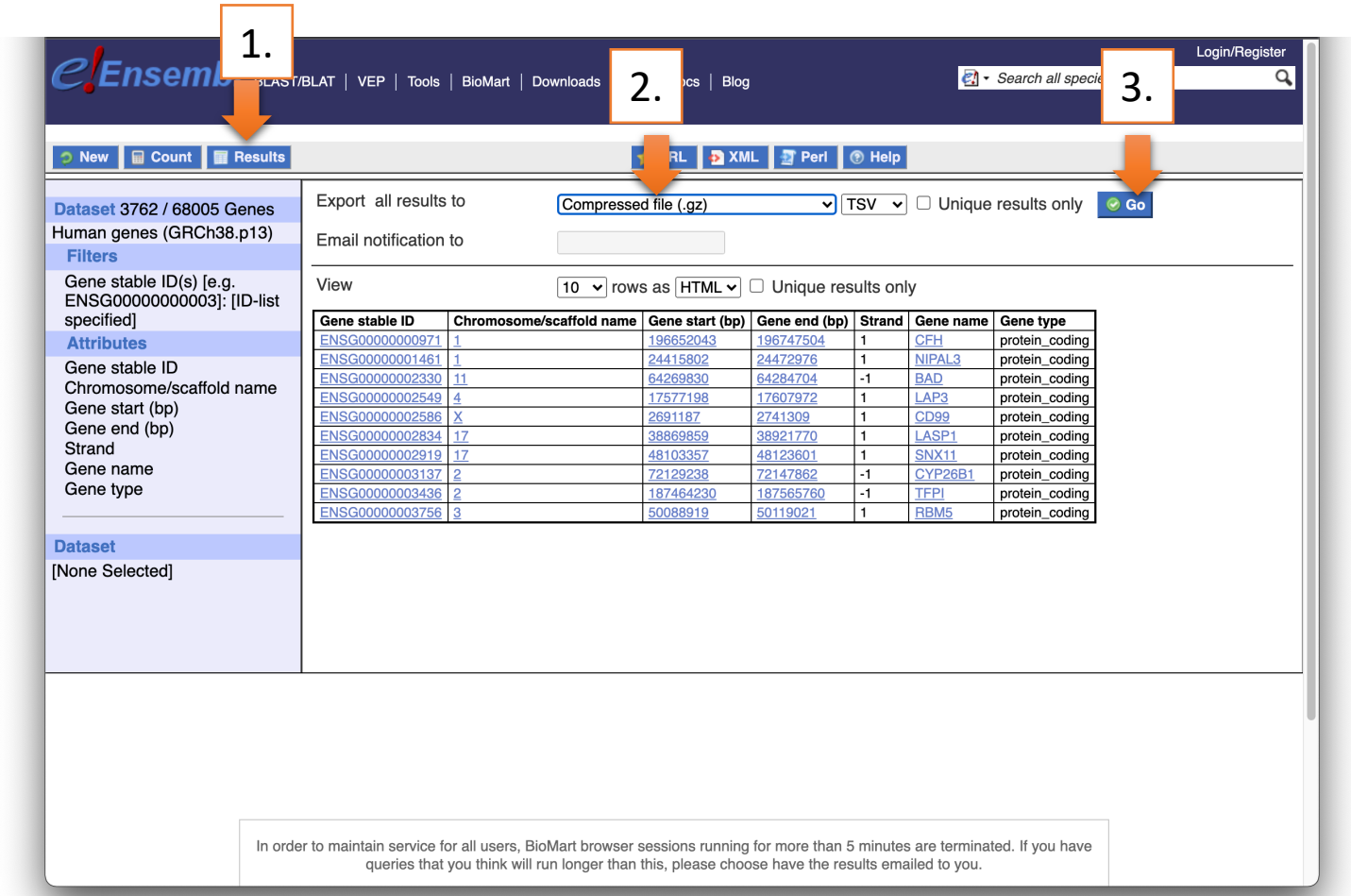
Exercise 2: get annotations for a set of genes

- 2
  - Click on Attributes (left panel)
  - Select “Features” (selected by default), Expand the GENE section, select
    - Gene stable ID,
    - Chromosome/scaffold name,
    - Gene Start (bp),
    - Gene End (bp),
    - Strand,
    - Gene Name
    - Gene type.



Exercise 2: get annotations for a set of genes

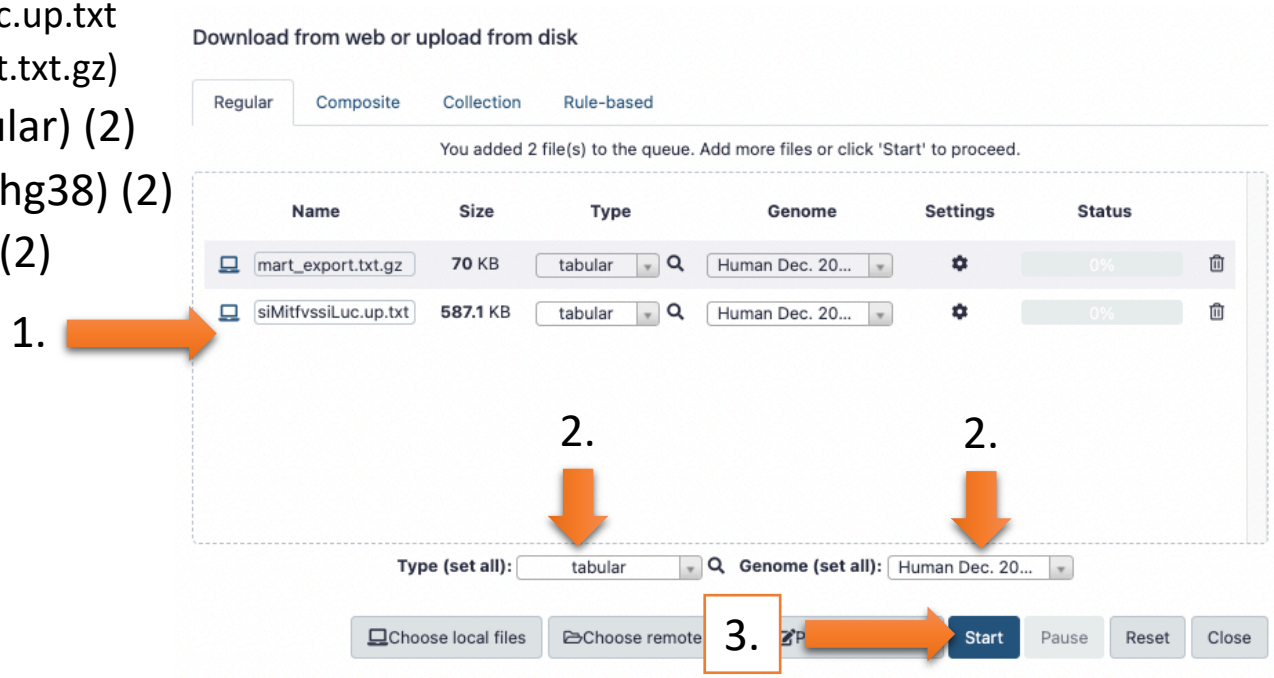
- 2
  - Click on Results (1)
  - Select Compressed file (.gz) in the drop down menu. (2)
  - Click on Go to download the resulting file. (3)



Gene stable ID	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Strand	Gene name	Gene type
ENSG00000000971	1	196652043	196747504	1	CFH	protein_coding
ENSG00000001461	1	24415802	24472976	1	NIPAL3	protein_coding
ENSG00000002330	11	64269830	64284704	-1	BAD	protein_coding
ENSG00000002549	4	17577198	17607972	1	LAP3	protein_coding
ENSG00000002586	X	2691187	2741309	1	CD99	protein_coding
ENSG00000002834	17	38869859	38921770	1	LASP1	protein_coding
ENSG00000002919	17	48103357	48123601	1	SNX11	protein_coding
ENSG00000003137	2	72129238	72147862	-1	CYP26B1	protein_coding
ENSG00000003436	2	187464230	187565760	-1	TFPI	protein_coding
ENSG00000003756	3	50088919	50119021	1	RBM5	protein_coding

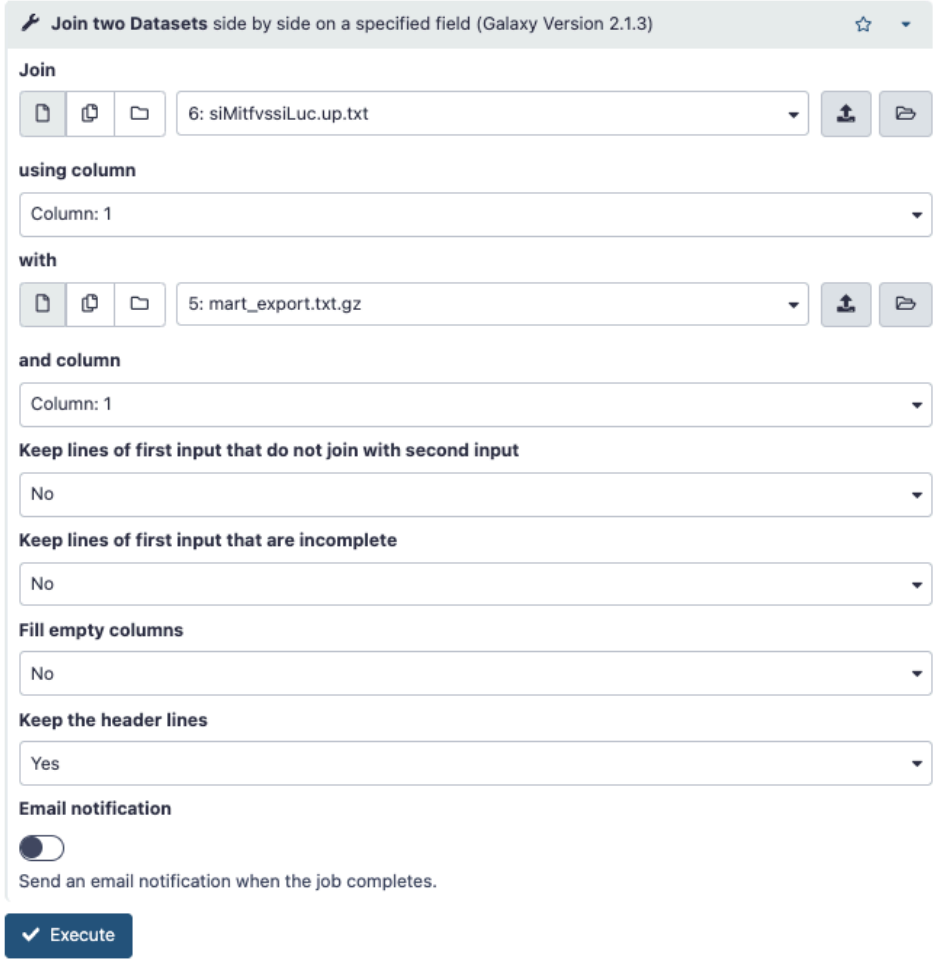
Exercise 2: get annotations for a set of genes

- 3.
  - Go to Galaxy France (https://usegalaxy.fr/)
  - Open the upload utility: Upload Data
  - Drag and drop your files (1)
    - siMitfvssiLuc.up.txt
    - mart_export.txt.gz)
  - Set type (tabular) (2)
  - Set Genome (hg38) (2)
  - Click on Start (2)



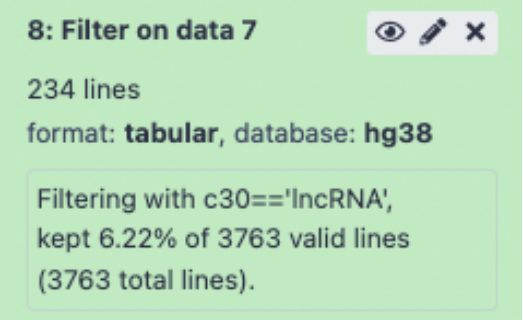
Exercise 2: get annotations for a set of genes

- 4.
  - Enter “Join” in the search field of the tool panel
  - Click on **Join two Datasets** side by side on a specified field
  - Run the tool “Join Two Datasets”
    - **Join**: siMitfvssiLuc.up.txt
    - **Using column**: Column: 1
    - **With**: mart_export.txt.gz
    - **And column**: Column: 1
    - **Keep the header lines**: Yes
    - Click on Execute



Exercise 2: get annotations for a set of genes

- …4.
  - Click on ✎ of the dataset you've just generated [join two datasets on * and data *]
  - In the "Attributes" tab, enter siMitfvssiLuc.up.annot.txt in the text box "Name".
  - Click on Save
- 5.
  - Run the tool "Filter data on any column using simple expressions" with the following parameters
    - **Filter**: siMitfvssiLuc.up.annot.txt
    - **With following condition:** c30=="lncRNA" (check which column contains Gene type)
    - **Number of header lines to skip**: 1
    - Click on Execute



Exercise 2: get annotations for a set of genes

- Bonus question.
  - Don’t change Dataset and Filters – simply click on Attributes.
  - Click on Attributes (left panel)
  - Select “Sequences”
  - Expand the SEQUENCES section
  - Select Flank (Gene) and enter 200 in the text box Upstream flank
  - Expand the Header information section
  - Select, in addition to the default selected attributes, Gene description and Gene Name
  - Note: Flank (Transcript) will give the flanks for all transcripts of a gene with multiple transcripts. Flank (Gene) will give the flanks for one possible transcript in a gene (the most 5’ coordinates for upstream flanking)

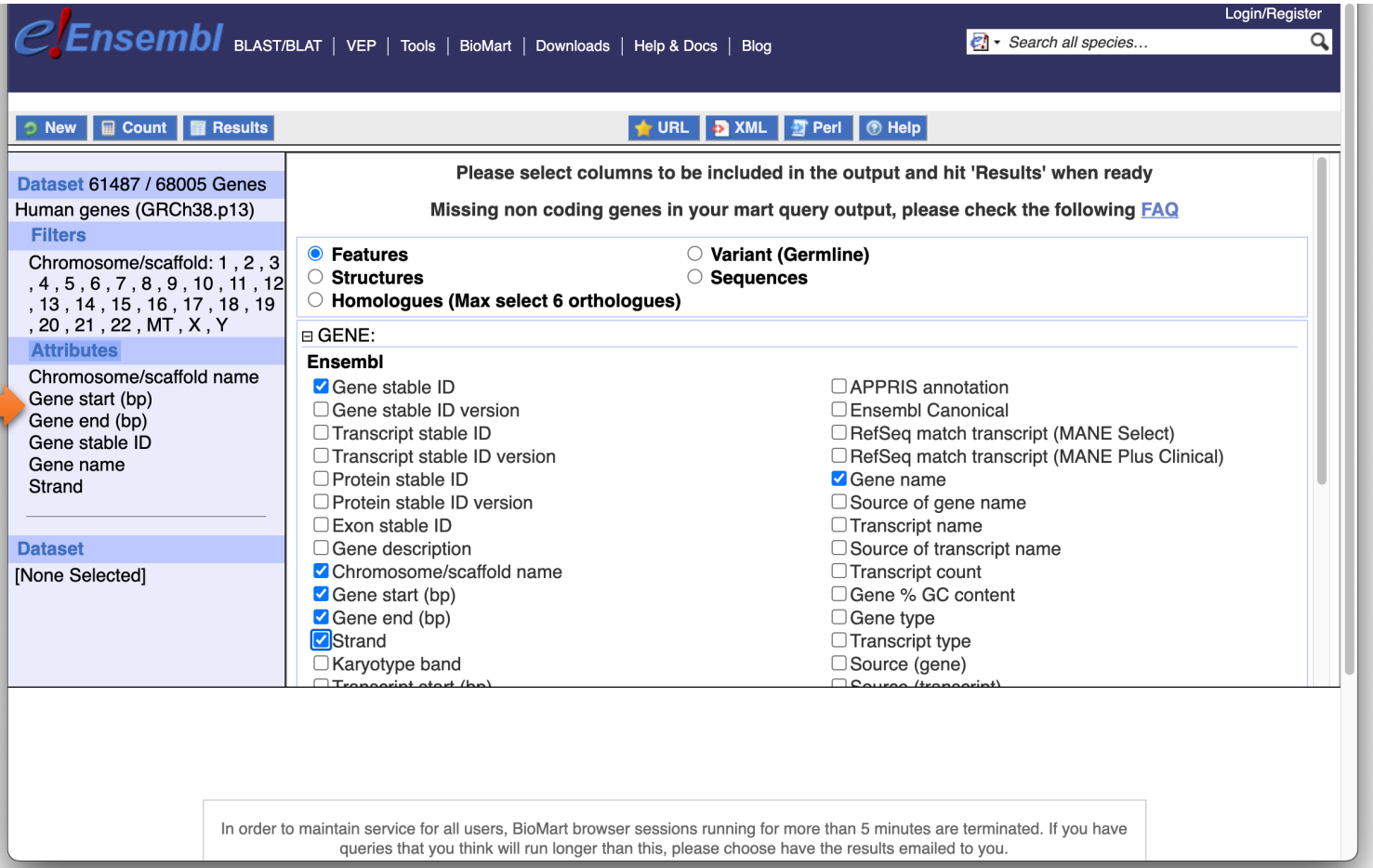
Exercise 3: get annotations in the genome

- 1.
  - In Ensembl/BioMart, create a new request
  - Click on Filters (left panel)
  - Expand the REGION section
  - Select “Multiple regions” and enter 2:208226227:208276270 in the text box
  - Click on count. **4 genes are found.**
- 2.
  - In Ensembl/BioMart, create a new request
  - Click on Filters (left panel)
  - Expand the REGION section
  - Select “Chromosome/scaffold” and multiple select 1 -> MT (click and drag). This corresponds to 61487 / 68005 Genes
  - Click on Attributes (left panel)
  - Select “Features” (selected by default)
  - In GENE, select Chromosome/scaffold name, Gene Start (bp), Gene End (bp), Gene stable ID, Gene Name and strand (**in that specific order!)**

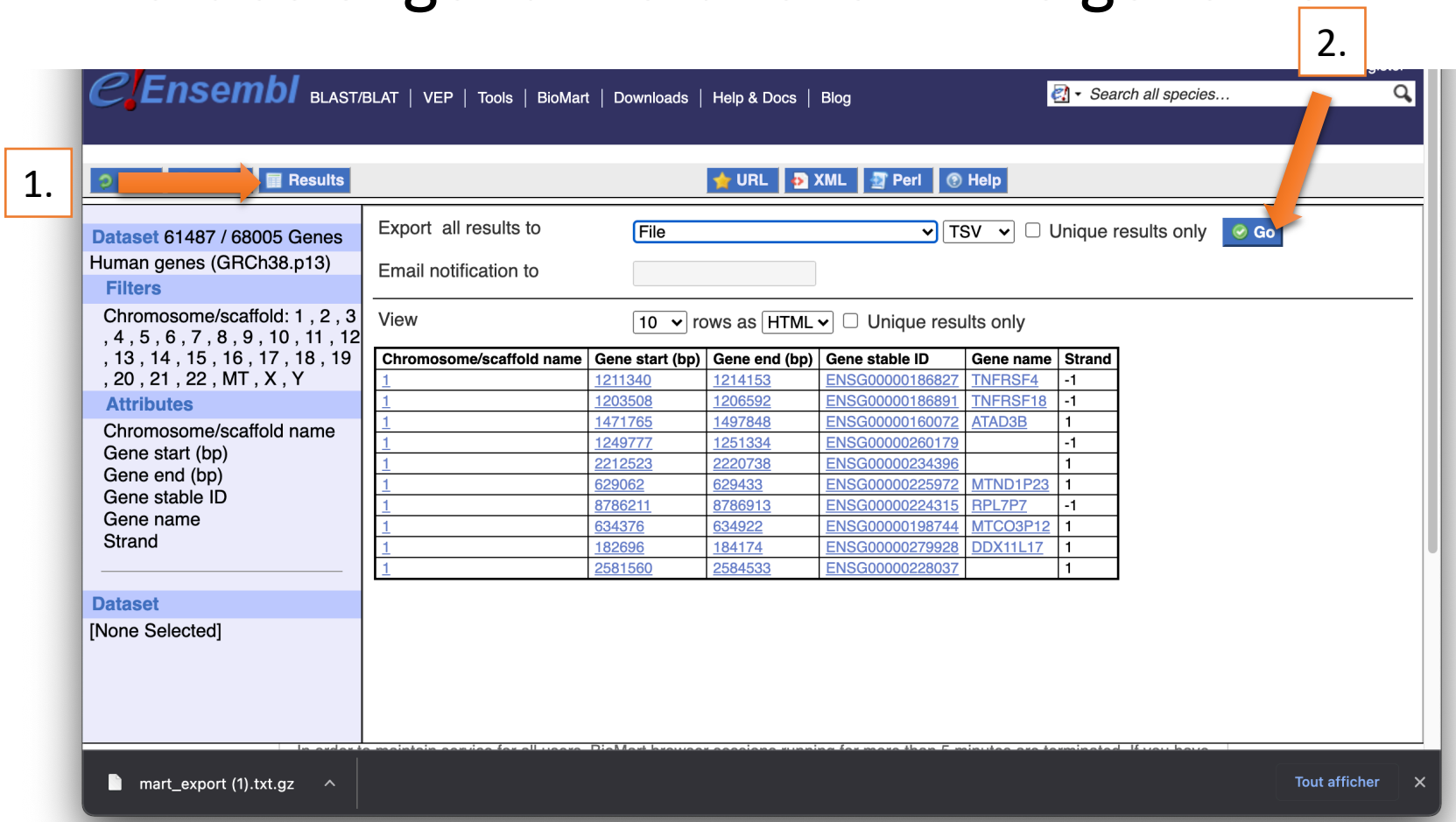
  Download the file and rename it hg38_ens105.bed

Exercise 3: get annotations in the genome

Check the order. Extracting fields in that specific order makes you create a BED file!



Exercise 3: get annotations in the genome



- Click on Results (1) and download the file as a TSV file (2). **Rename the file hg38_ens105.bed**
- Open the file and remove the first line. Save change.