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 1/64914 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

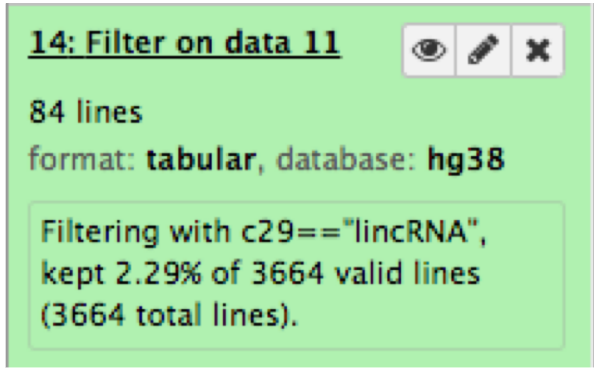
- 2.
  - In Ensembl/BioMart, create a new request
  - Click on Filters (left panel)
  - Expand the GENE section
  - Select “Input external references ID list” and select “Gene stable ID(s)” in the drop down list
  - Open the file siMitfvssiLuc.up.txt in Excel and copy the content of the first column (ENSG**) without the title and paste it all into the text field (Input external references ID list) of the Ensembl Biomart filter page
  - Click on “Count” (top left button). You should have the number of genes you have in your file generated by SARTools: 3663
  - 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 and Gene type.
  - Click on Results
  - Select Compressed file (.gz) in the drop down menu. Click on Go to download the resulting file.

Exercise 2: get annotations for a set of genes

- 3.
  - Go to GalaxEast (http://use.galaxeast.fr)
  - Open the upload utility: click on  in top of the tool panel and drag and drop your files (siMitfvssiLuc.up.txt and mart_export.txt.gz) into the opened window
  - Click on Start
- 4.
  - Run the tool “Join Two Datasets”
    - Join: siMitfvssiLuc.up.txt
    - Using column: Column: 1
    - With: mart_export.txt
    - And column: Column: 1
    - Keep lines of first input that do not join with second input: No
    - Keep lines of first input that are incomplete: No
    - Fill empty columns: No
    - Click on Execute

Exercise 2: get annotations for a set of genes

- 5.
  - Click on the button  of the dataset you've just generated "join two datasets on (…)"
  - In the "Attributes" tab, enter siMitfvssiLuc.up.annot.txt in the text box "Name".
  - Click on Save
- 6.
  - Run the tool "Filter data on any column using simple expressions" with the following parameters
    - Filter: siMitfvssiLuc.up.annot.txt
    - With following condition: c30=="lincRNA" (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 Upstream flank text box
  - 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 58676 / 64914 Genes
  - Click on Attributes (left panel)
  - Select “Features” (selected by default)
  - In GENE, select Gene stable ID, Chromosome/scaffold name, Gene Start (bp), Gene End (bp), strand and Gene Name
  - Click on Results