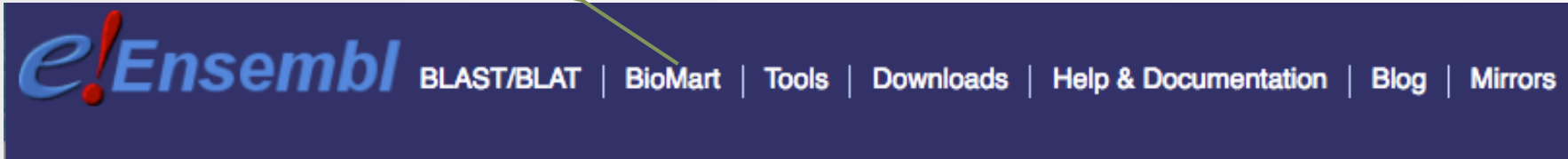


Data mining with Ensembl Biomart (answers to questions)

Stéphanie Le Gras
(slegras@igbmc.fr)



Start using Ensembl/BioMart

- Go to Ensembl website
<http://www.ensembl.org/index.html>
- Click on BioMart in the top menu



- **CHOOSE DATABASE**: select “Ensembl Genes 87”
- **CHOOSE DATASET**: select “Homo sapiens genes (GRCh38.p7)”



Exercise 1: get annotations of a gene

- 1.
 - Click on Filters (left panel),
 - Expand the “GENE” section
 - Select “Input external references ID list”, select “HGNC symbol(s) in the drop down list and enter IDH1.
 - Click on Count in the top left panel  . You should get **Dataset 1 / 63305 Genes**
 - Click on Attributes (left menu)
 - Select “Features” (selected by default)
 - Select Gene ID, Transcript ID and associated Gene Name
 - Click on Results (top left menu) 



Gene ID	Transcript ID	Associated 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 ID” (Gene Information)
 - Select Associated gene name (Gene Information), transcript IDs (Transcript Information) and Exon 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 Associated gene name (Gene Information) and select transcript ID (Transcript Information) and Exon IDs (Exon Information).
 - Click on Results 


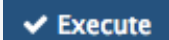
Exercise 1: get annotations of a gene

- 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 ID, Transcript ID and Associated 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 ID, Transcript ID and Associated 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


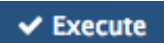
Exercise 2: get annotations for a set of genes

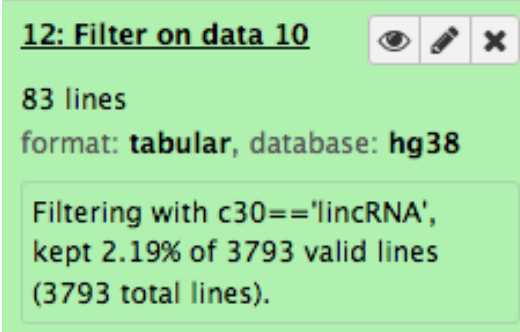
- 2.
 - In Ensembl/BioMart, create a new request (see slide 2.)
 - Click on Filters (left panel)
 - Expand the GENE section
 - Select “Input external references ID list” and select “Gene ID(s)” in the drop down list
 - Click on “Browse” and select the file siMitfvssiLuc.up.txt
 - Click on “Count” (top left button)  **Count**. You should have the number of genes you have in your file generated by SARTools: 3792
 - Click on Attributes (left panel)
 - Select “Features” (selected by default), select Ensembl Gene ID, Chromosome/scaffold name, Gene Start (bp), Gene End (bp), Strand, Associated Gene Name and Gene type.
 - Click on Results  **Results**
 - Select Compressed file (.gz) in the drop down menu. Click on Go  **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”
 - Number of header lines to skip: 1
 - Click on Execute 



12: Filter on data 10   

83 lines
format: **tabular**, database: **hg38**

Filtering with c30=='lincRNA',
kept 2.19% of 3793 valid lines
(3793 total lines).

Exercise 2: get annotations for a set of genes

- 7.
 - Don't change Dataset and Filters – simply click on Attributes.
 - Click on Attributes (left panel)
 - Select “Sequences”
 - Expand the SEQUENCES section
 - Select Flank (Transcript) and enter 2000 in the Upstream flank text box
 - Expand the Header information section
 - Select, in addition to the default selected attributes, Description and Associated 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 (see slide 2.)
 - Click on Filters (left panel)
 - Expand the REGION section
 - Select “Multiple Chromosomal 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 (see slide 2.)
 - Click on Filters (left panel)
 - Expand the REGION section
 - Select “Chromosome” and multiple select 1 -> MT (click and drag). This corresponds to 57992 / 63305 Genes
 - Click on Attributes (left panel)
 - Select “Features” (selected by default)
 - In GENE, select Gene ID, Chromosome/scaffold name, Gene Start (bp), Gene End (bp), strand and associated Gene Name
 - Click on Results 