

# Data mining with Ensembl Biomart (answers to questions)

Stéphanie Le Gras  
([slegras@igbmc.fr](mailto: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            |
|---------------------------------|---------------------------------|----------------------|
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000345146</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000446179</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000415913</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000484575</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000415282</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000462386</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000417583</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000451391</a> | <a href="#">IDH1</a> |
| <a href="#">ENSG00000138413</a> | <a href="#">ENST00000481557</a> | <a href="#">IDH1</a> |

- 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.

The screenshot shows the Ensembl BioMart interface. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. The main content area is titled "Please restrict your query using criteria below (If filter values are truncated in any lists, hover over the list item to see the full text)". The "GENE" section is expanded, showing options for "Limit to genes (external references)..." and "Input external references ID list [Max 500 advised]". The "Input external references ID list" option is selected, and the text field contains a list of Gene Stable IDs: ENSG00000178568, ENSG00000275183, ENSG00000100024, and ENSG00000134851. The "Count" button is highlighted in the top left. Arrows from the text instructions point to the "Filters" panel, the "GENE" section, the "Input external references ID list" checkbox, the text field, and the "Count" button.

In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.

# 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.

The screenshot shows the Ensembl BioMart interface. The left panel is titled 'Dataset 3762 / 68005 Genes' and 'Human genes (GRCh38.p13)'. It has sections for 'Filters', 'Attributes', and 'Dataset'. The 'Attributes' section is expanded, showing 'Gene stable ID', 'Chromosome/scaffold name', 'Gene start (bp)', 'Gene end (bp)', 'Strand', 'Gene name', and 'Gene type'. The 'Dataset' section shows '[None Selected]'. The right panel is titled 'Please select columns to be included in the output and hit 'Results' when ready'. It has a sub-section 'Missing non coding genes in your mart query output, please check the following FAQ'. Below this, there are radio buttons for 'Features' (selected), 'Structures', and 'Homologues (Max select 6 orthologues)'. There are also radio buttons for 'Variant (Germline)' and 'Sequences'. The 'GENE' section is expanded, showing 'Ensembl' attributes: 'Gene stable ID' (checked), 'Gene stable ID version', 'Transcript stable ID', 'Transcript stable ID version', 'Protein stable ID', 'Protein stable ID version', 'Exon stable ID', 'Gene description', 'Chromosome/scaffold name' (checked), 'Gene start (bp)' (checked), 'Gene end (bp)' (checked), 'Strand' (checked), 'Karyotype band', and 'Transcript start (bp)'. There are also 'Other' attributes: 'APPRIS annotation', 'Ensembl Canonical', 'RefSeq match transcript (MANE Select)', 'RefSeq match transcript (MANE Plus Clinical)', 'Gene name' (checked), 'Source of gene name', 'Transcript name', 'Source of transcript name', 'Transcript count', 'Gene % GC content', 'Gene type' (checked), 'Transcript type', 'Source (gene)', and 'Source (transcript)'. At the bottom, there is a note: 'In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.'

# 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)

The screenshot shows the Ensembl BioMart interface. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, and a search bar. The main content area is divided into a left sidebar and a main panel. The sidebar shows the dataset 'Human genes (GRCh38.p13)' and a list of attributes. The main panel has a 'Results' button (1) and a table of gene annotations. Below the table, there are options to export results as a 'Compressed file (.gz)' (2) and a 'Go' button (3).

| Gene stable ID   | Chromosome/scaffold name | Gene start (bp) | Gene end (bp) | Strand | Gene name | Gene type      |
|------------------|--------------------------|-----------------|---------------|--------|-----------|----------------|
| ENSG000000000971 | 1                        | 196652043       | 196747504     | 1      | CFH       | protein_coding |
| ENSG000000001461 | 1                        | 24415802        | 24472976      | 1      | NIPAL3    | protein_coding |
| ENSG000000002330 | 11                       | 64269830        | 64284704      | -1     | BAD       | protein_coding |
| ENSG000000002549 | 4                        | 17577198        | 17607972      | 1      | LAP3      | protein_coding |
| ENSG000000002586 | X                        | 2691187         | 2741309       | 1      | CD99      | protein_coding |
| ENSG000000002834 | 17                       | 38869859        | 38921770      | 1      | LASP1     | protein_coding |
| ENSG000000002919 | 17                       | 48103357        | 48123601      | 1      | SNX11     | protein_coding |
| ENSG000000003137 | 2                        | 72129238        | 72147862      | -1     | CYP26B1   | protein_coding |
| ENSG000000003436 | 2                        | 187464230       | 187565760     | -1     | TFPI      | protein_coding |
| ENSG000000003756 | 3                        | 50088919        | 50119021      | 1      | RBM5      | protein_coding |

In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.

# 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)

1. 

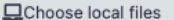
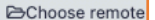




Download from web or upload from disk

Regular Composite Collection Rule-based

You added 2 file(s) to the queue. Add more files or click 'Start' to proceed.

| Name   | Size     | Type    | Genome           | Settings  | Status |
|--|----------|---------|------------------|---|--------|
|  mart_export.txt.gz   | 70 KB    | tabular | Human Dec. 20... |  | 0%     |
|  siMitfvssiLuc.up.txt | 587.1 KB | tabular | Human Dec. 20... |  | 0%     |

Type (set all): tabular Genome (set all): Human Dec. 20...

  **3.**    

2.

2.

3.




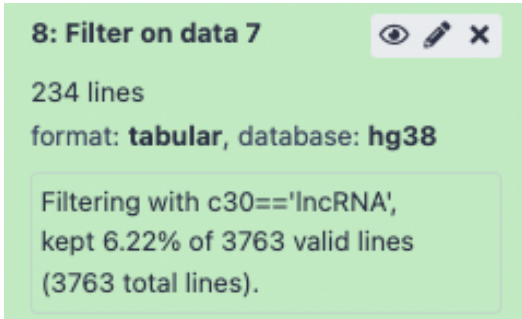
# Exercise 2: get annotations for a set of genes




- 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

The screenshot shows the Galaxy web interface for the 'Join two Datasets' tool. The tool title is 'Join two Datasets side by side on a specified field (Galaxy Version 2.1.3)'. The interface includes several configuration sections: 'Join' with a dropdown menu set to '6: siMitfvssiLuc.up.txt'; 'using column' with a dropdown menu set to 'Column: 1'; 'with' with a dropdown menu set to '5: mart\_export.txt.gz'; 'and column' with a dropdown menu set to 'Column: 1'; 'Keep lines of first input that do not join with second input' with a dropdown menu set to 'No'; 'Keep lines of first input that are incomplete' with a dropdown menu set to 'No'; 'Fill empty columns' with a dropdown menu set to 'No'; 'Keep the header lines' with a dropdown menu set to 'Yes'; and 'Email notification' with a toggle switch turned off. At the bottom, there is a blue 'Execute' button with a checkmark icon.

# 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



8: Filter on data 7   

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

Filtering with c30=="lncRNA",  
kept 6.22% of 3763 valid lines  
(3763 total lines).

## 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!

The screenshot shows the Ensembl BioMart interface. At the top, the Ensembl logo is on the left, and navigation links (BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, Blog) and a search bar (Search all species...) are on the right. Below the navigation bar, there are buttons for 'New', 'Count', and 'Results'. The main content area is titled 'Please select columns to be included in the output and hit 'Results' when ready'. A message states: 'Missing non coding genes in your mart query output, please check the following FAQ'. The interface is divided into several sections:

- Dataset:** 61487 / 68005 Genes, Human genes (GRCh38.p13)
- Filters:** Chromosome/scaffold: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, MT, X, Y
- Attributes:** Chromosome/scaffold name, Gene start (bp), Gene end (bp), Gene stable ID, Gene name, Strand. An orange arrow points to this section.
- Dataset:** [None Selected]
- Feature Selection:**
  - Features
  - Structures
  - Homologues (Max select 6 orthologues)
  - Variant (Germline)
  - Sequences
- GENE:**
  - Ensembl:**
    - Gene stable ID
    - Gene stable ID version
    - Transcript stable ID
    - Transcript stable ID version
    - Protein stable ID
    - Protein stable ID version
    - Exon stable ID
    - Gene description
    - Chromosome/scaffold name
    - Gene start (bp)
    - Gene end (bp)
    - Strand
    - Karyotype band
    - Transcript start (bp)
  - APPRIS annotation
  - Ensembl Canonical
  - RefSeq match transcript (MANE Select)
  - RefSeq match transcript (MANE Plus Clinical)
  - 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)

At the bottom, a message states: 'In order to maintain service for all users, BioMart browser sessions running for more than 5 minutes are terminated. If you have queries that you think will run longer than this, please choose have the results emailed to you.'

# Exercise 3: get annotations in the genome

1.

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Search all species...

Results URL XML Perl Help

Export all results to  TSV  Unique results only

Email notification to

View  rows as   Unique results only

| Chromosome/scaffold name | Gene start (bp) | Gene end (bp) | Gene stable ID  | Gene name | Strand |
|--------------------------|-----------------|---------------|-----------------|-----------|--------|
| 1                        | 1211340         | 1214153       | ENSG00000186827 | TNFRSF4   | -1     |
| 1                        | 1203508         | 1206592       | ENSG00000186891 | TNFRSF18  | -1     |
| 1                        | 1471765         | 1497848       | ENSG00000160072 | ATAD3B    | 1      |
| 1                        | 1249777         | 1251334       | ENSG00000260179 |           | -1     |
| 1                        | 2212523         | 2220738       | ENSG00000234396 |           | 1      |
| 1                        | 629062          | 629433        | ENSG00000225972 | MTND1P23  | 1      |
| 1                        | 8786211         | 8786913       | ENSG00000224315 | RPL7P7    | -1     |
| 1                        | 634376          | 634922        | ENSG00000198744 | MTCO3P12  | 1      |
| 1                        | 182696          | 184174        | ENSG00000279928 | DDX11L17  | 1      |
| 1                        | 2581560         | 2584533       | ENSG00000228037 |           | 1      |

Dataset 61487 / 68005 Genes  
Human genes (GRCh38.p13)

Filters  
Chromosome/scaffold: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, MT, X, Y

Attributes  
Chromosome/scaffold name  
Gene start (bp)  
Gene end (bp)  
Gene stable ID  
Gene name  
Strand

Dataset  
[None Selected]

mart\_export (1).txt.gz Tout afficher

- 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.