

Data mining with Ensembl Biomart

Stéphanie Le Gras
(slegras@igbmc.fr)

Guidelines

- Genome data
- Genome browsers
- Getting access to genomic data: Ensembl/BioMart

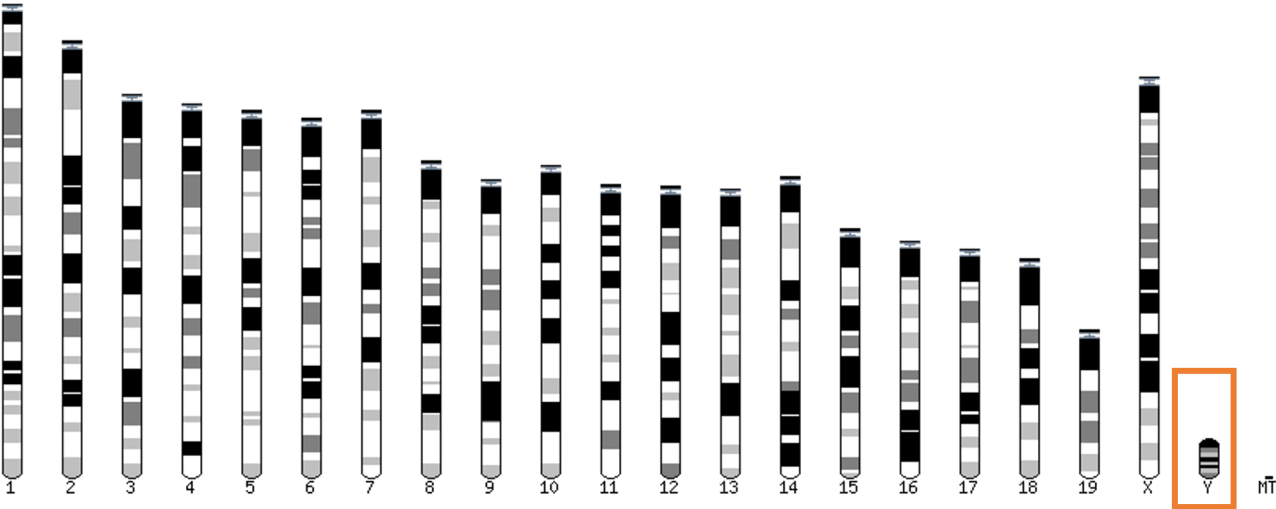
Genome builds

| SPECIES | UCSC VERSION | RELEASE DATE | RELEASE NAME | STATUS |
|----------------|--------------|--------------|------------------------------------|----------------------|
| MAMMALS | | | | |
| Human | hg38 | Dec. 2013 | Genome Reference Consortium GRCh38 | Available |
| | hg19 | Feb. 2009 | Genome Reference Consortium GRCh37 | Available |
| | hg18 | Mar. 2006 | NCBI Build 36.1 | Available |
| | hg17 | May 2004 | NCBI Build 35 | Available |
| | hg16 | Jul. 2003 | NCBI Build 34 | Available |
| | hg15 | Apr. 2003 | NCBI Build 33 | Archived |
| | hg13 | Nov. 2002 | NCBI Build 31 | Archived |
| | hg12 | Jun. 2002 | NCBI Build 30 | Archived |
| | hg11 | Apr. 2002 | NCBI Build 29 | Archived (data only) |
| | hg10 | Dec. 2001 | NCBI Build 28 | Archived (data only) |
| | hg8 | Aug. 2001 | UCSC-assembled | Archived (data only) |
| | hg7 | Apr. 2001 | UCSC-assembled | Archived (data only) |
| | hg6 | Dec. 2000 | UCSC-assembled | Archived (data only) |
| | hg5 | Oct. 2000 | UCSC-assembled | Archived (data only) |
| | hg4 | Sep. 2000 | UCSC-assembled | Archived (data only) |
| | hg3 | Jul. 2000 | UCSC-assembled | Archived (data only) |
| | hg2 | Jun. 2000 | UCSC-assembled | Archived (data only) |
| | hg1 | May 2000 | UCSC-assembled | Archived (data only) |

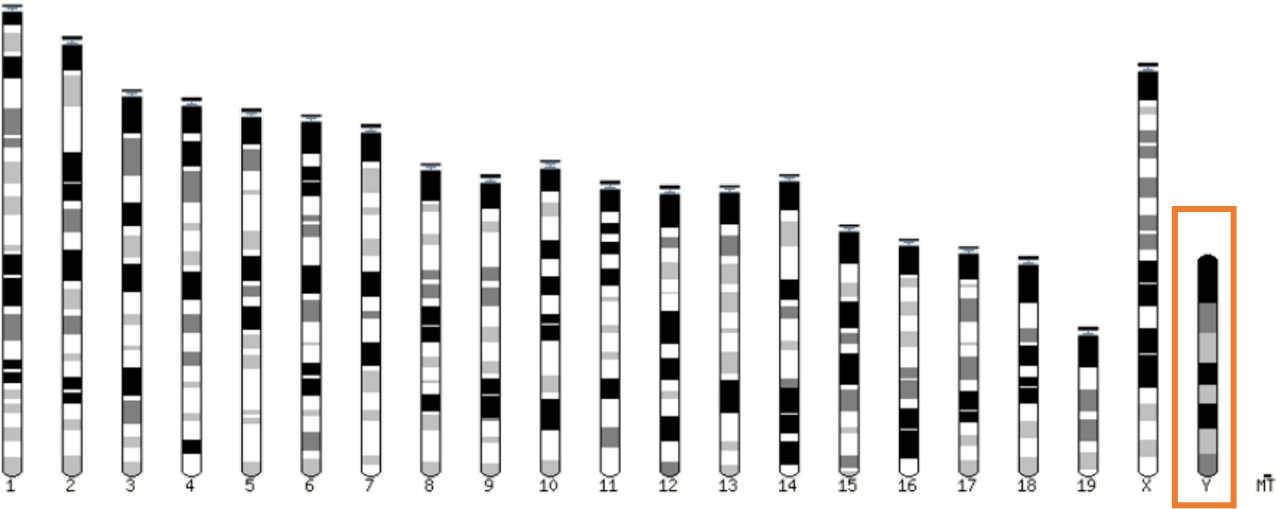
Source: <https://genome.ucsc.edu/FAQ/FAQreleases.html>

Genome builds

mm9



mm10



Get access to genomic data

- Need a way to gather all genomic information in one place
- Availability of the data
- Accessibility to the data



Genome browsers

Genome Browsers

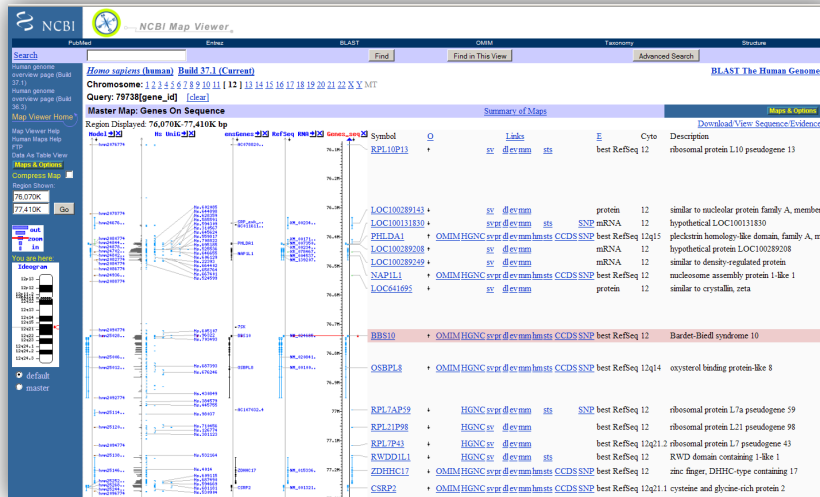
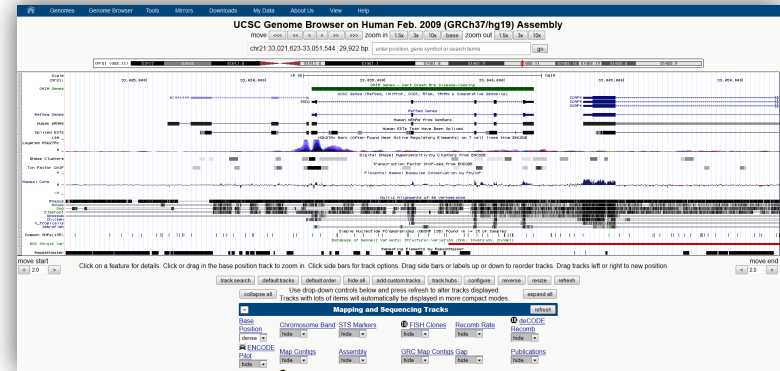
- Graphical interface to display genomic data
- Visualize and browse entire genomes with annotated data
 - Gene prediction and structure
 - Proteins,
 - Expression,
 - Regulation,
 - Variation,
 - Comparative analysis...

There are Genome Browsers...

EBI - Ensembl

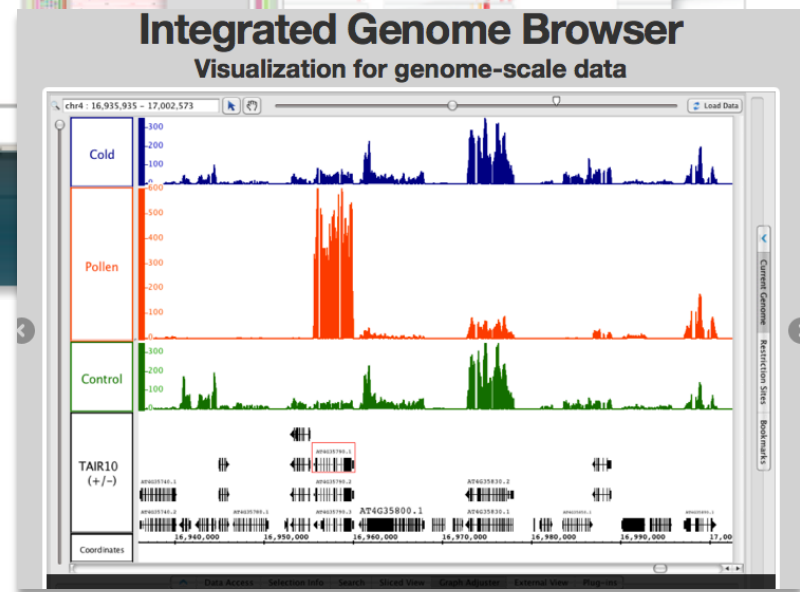
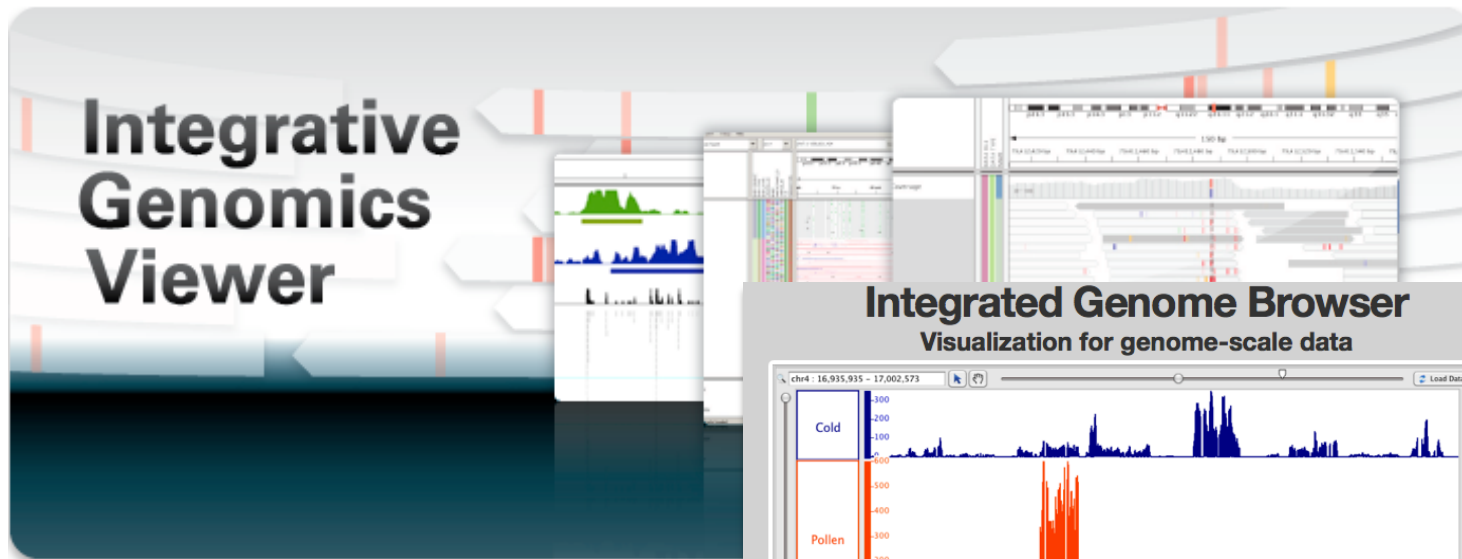


UCSC – Genome Browser



NCBI – Genome Data Viewer

And Genome browsers...









Getting access to genomic data: ENSEMBL/BIOmart

Access Ensembl's data

Web site

Mining tool: BioMart


-  User friendly
-  Straightforward
-  Only one request at once

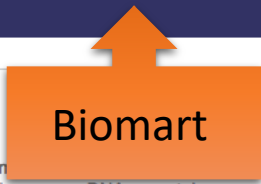
-  Get answers to complex queries
-  Very fast
-  Need training

BioMart

- <http://www.biomart.org/>
- Joint development between EBI and Cold Spring Harbor Laboratory (CSHL)
- Open source project
- BioMart can access diverse databases from a single interface
- It is a search engine that can find multiple terms and put them into a table format
- No programming required!

BioMart/Ensembl

 [BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#) [Login/Register](#)



Tools [All tools](#)

BioMart > Export custom datasets from Ensembl with this data-mining tool

Biomart for your DNA or protein sequence

Variant Effect Predictor > Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 95 (January 2019)

- New regulatory build for human, incorporating new data from ENCODE
- Update to GENCODE M20 for mouse
- New genomes: donkey, polar bear, black bear, red fox, koala, dingo, tuatara, painted turtle and desert tortoise
- Updated genomes for chicken, cow and horse
- New protein structure variation view

[More release news](#) on our blog

Search

All species for

e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

All genomes

Favourite genomes

 **Human**
GRCh38.p12

• [View full list of all Ensembl species](#)

• [Still using GRCh37?](#)

Other news from our blog

- 01 Mar 2019: [Getting to know us: Guy from Ensembl Plants](#)
- 27 Feb 2019: [Job: Ensembl Infrastructure Project Leader](#)
- 27 Feb 2019: [Custom data upload: creating URLs for large](#)

- Get access to :
 - Genomic annotation (genes, SNPs)
 - Functional annotation
 - Expression data

Example: Step 1 (Select datasets)

The screenshot shows the Ensembl genome browser interface. At the top, the Ensembl logo is on the left, and navigation links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog are in the center. On the right, there is a 'Login/Register' link and a search bar with the text 'Search all species...'. Below the navigation bar, there are buttons for 'New', 'Count', and 'Results'. A secondary bar contains 'URL', 'XML', 'Perl', and 'Help' buttons. The main content area is divided into a left sidebar and a main panel. The sidebar has a 'Dataset' section with '[None selected]'. The main panel has a dropdown menu currently showing 'Ensembl Genes 105'. An orange arrow points from the text 'First choose the database and dataset' to the dropdown menu. The dropdown menu is open, showing a list of species and their genome versions. The first item is '- CHOOSE DATASET -'. The second item, 'Human genes (GRCh38.p13)', is highlighted in blue. Other items include 'Chicken genes (GRCg6a)', 'Mouse genes (GRCm39)', 'Rat genes (mRatBN7.2)', 'Zebrafish genes (GRCz11)', and many others. A text box at the bottom right contains the text: '...ning for more than 5 minutes are terminated. If you have choose have the results emailed to you.'

Ensembl Genes 105

- ✓ - CHOOSE DATASET -
- Chicken genes (GRCg6a)
- Human genes (GRCh38.p13)**
- Mouse genes (GRCm39)
- Rat genes (mRatBN7.2)
- Zebrafish genes (GRCz11)
-
- Abingdon island giant tortoise genes (ASM359739v1)
- African ostrich genes (ASM69896v1)
- Algerian mouse genes (SPRET_EIJ_v1)
- Alpaca genes (vicPac1)
- Alpine marmot genes (marMar2.1)
- Amazon molly genes (Poecilia_formosa-5.1.2)
- American bison genes (Bison_UMD1.0)
- American black bear genes (ASM334442v1)
- American mink genes (NNQGG.v01)
- Arabian camel genes (CamDro2)
- Arctic ground squirrel genes (ASM342692v1)
- Argentine black and white tegu genes (HLtupMer3)
- Armadillo genes (Dasnov3.0)
- Asian bonytongue genes (fSclFor1.1)
- Atlantic cod genes (gadMor3.0)
- Atlantic herring genes (Ch_v2.0.2)
- Atlantic salmon genes (ICSASG_v2)
- Australian saltwater crocodile genes (CroPor_comp1)

First choose the database and dataset

...ning for more than 5 minutes are terminated. If you have choose have the results emailed to you.

Example: Step 2 (Filter)

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: 1
Start: 78895
End: 224561

Attributes
Gene stable ID
Gene stable ID version
Transcript stable ID
Transcript stable ID version

Dataset
[None Selected]

8
9
10
11
12
13
14
15
16
17
18
19
20

Coordinates
Start: 78895
End: 224561

Karyotype band
Band Start
Band End

Marker

Limit to chromosome 1

Limit to given coordinates

Example: Step 3 (Count results)

Compute match count

The screenshot displays the Ensembl BioMart interface during the 'Count' step. The left sidebar contains the following information:

- Dataset 12 / 68005 Genes**
- Human genes (GRCh38.p13)
- Filters**
- Chromosome/scaffold: 1
- Start: 78895
- End: 224561
- Attributes**
- Gene stable ID
- Gene stable ID version
- Transcript stable ID
- Transcript stable ID version
- Dataset**
- [None Selected]

The main content area shows the following options and input fields:

- Coordinates**
- Start:
- End:
- Multiple regions (Chr:Start:End:Strand) [Max 500 advised]**
e.g. 1:100:10000:-1, 1:100000:200000:1
-
- Aucun fichier choisi

At the bottom of the page, a notice 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."

Example: Step 4 (Select attributes)

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

Search all species...

New Count Results URL XML Perl Help

Dataset 12 / 68005 Genes
Human genes (GRCh38.p13)

Filters
Chromosome/scaffold: 1
Start: 78895
End: 224561

Attributes
Gene stable ID
Transcript stable ID

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready
Missing non coding genes in your mart query output, please check the following [FAQ](#)

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

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)
- 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)

Select attributes to be output

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.

Example: Step 5 (get results)

Ensembl

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

Login/Register

Search all species...

New Count Results

URL XML Perl Help

Export all results to: File TSV Unique results only Go

Email notification to:

View: 10 rows as HTML Unique results only

| Gene stable ID | Transcript stable ID |
|---------------------------------|---------------------------------|
| ENSG00000238009 | ENST00000466430 |
| ENSG00000238009 | ENST00000477740 |
| ENSG00000238009 | ENST00000471248 |
| ENSG00000238009 | ENST00000610542 |
| ENSG00000238009 | ENST00000453576 |
| ENSG00000239945 | ENST00000495576 |
| ENSG00000233750 | ENST00000442987 |
| ENSG00000268903 | ENST00000494149 |
| ENSG00000269981 | ENST00000595919 |
| ENSG00000239906 | ENST00000493797 |

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 1: get annotations of a gene (1/2)

- 1. Using Ensembl/BioMart, retrieve all transcripts IDs and the gene ID of IDH1 gene (human). How many transcripts does the gene IDH1 have?
 - Use Ensembl Gene **v105**, for Human genes (GRCh38.p13)
 - Click on Filters :
 - Expand the GENE section
 - Select « Input external references ID list »
 - Select Gene Name(s) in the drop down menu
 - Enter IDH1 in the text box
 - Click on Attributes :
 - Select “Features” (top panel, selected by default)
 - Expand GENE:
 - Select Gene stable ID, Transcript stable ID, Gene Name
 - Deselect Gene stable ID version, Transcript stable ID version
 - Click on Results

Exercise 1: get annotations of a gene (2/2)

- 2. Extract all exon sequences of the IDH1 gene in fasta format. Headers will contain the Gene names, transcript stable IDs and Exon stable IDs.
- 3. Extract all coding sequences of the IDH1 gene in fasta format. Headers will contain the transcript stable IDs and Exon stable IDs.
- 4. Retrieve GO-terms associated to the IDH1 gene (select GO Term Name, GO domain and GO Term Accession along with Gene stable ID, Transcript stable ID and Gene Name)
- 5. Retrieve the germline variations found in this gene. Annotations to be found (Variant Name, Variant Alleles, Minor allele frequency, Chromosome/scaffold name, Chromosome/scaffold position start (bp), Chromosome/scaffold position end (bp), Variant Consequence along with Gene stable ID, Transcript stable ID and Gene Name)

Exercise 2: get annotations for a set of genes

- The file siMitfvssiLuc.up.txt you generated using SARtools lacks meaningful annotation. Annotate the file siMitfvssiLuc.up.txt with gene annotations you'll extract from Ensembl/BioMart. To do so:
 1. We are going to extract annotation [Ensembl/BioMart]
 2. Then, we are going to join the two datasets (tabular text file) based on a common field. [Galaxy]

Exercise 2: get annotations for a set of genes

siMitfvssiLuc.up.txt

| Id | siLuc2 | siLuc... |
|-----------------|--------|----------|
| ENSG00000018408 | 4685 | ... |
| ENSG00000081189 | 1716 | ... |
| ENSG00000106772 | 3063 | ... |
| ENSG00000124942 | 309 | ... |
| ENSG00000142871 | 243 | ... |
| ENSG00000143341 | 3760 | ... |
| ENSG00000154556 | 352 | ... |
| ENSG00000185565 | 679 | ... |
| ENSG00000163328 | 136 | ... |
| ENSG00000064042 | 1160 | ... |
| ENSG00000114423 | 2293 | ... |

mart_export.txt (from Ensembl/Biomart)


| Gene stable ID | Gene name | Chro... |
|-----------------|-----------|-------------|
| ENSG00000000971 | CFH 1 | 19665187... |
| ENSG00000001461 | NIPAL3 | 1 2441... |
| ENSG00000124942 | AHNAK | 11 624... |
| ENSG00000002330 | BAD 11 | 642698... |
| ENSG00000002549 | LAP3 4 | 175771... |
| ENSG00000002586 | CD99 X | 269113... |
| ENSG00000002834 | LASP1 17 | 3886... |
| ENSG00000002919 | SNX11 | 17 4810... |
| ENSG00000003137 | CYP26B1 | 2 7212... |
| ENSG00000003436 | TFPI 2 | 187464... |
| ENSG00000018408 | WWTR1 | 3 1495... |



Result file

| Gene stable ID | siLuc2 | siLuc3 | ... | Gene name | Chro... |
|-----------------|--------|--------|-------|-----------|---------|
| ENSG00000124942 | 309 | ... | AHNAK | 11 | 624... |
| ENSG00000018408 | 4685 | ... | WWTR1 | 3 | 1495... |


Exercise 2: get annotations for a set of genes

- 1. Click on  to display the content of the dataset [SARTools DESeq2 tables] (1) (*from your history « RNA-seq data analysis »*) and download the file siMitfvssiLuc.up.txt (click right, save ...) (2)

1.



2.

| Output File Name (click to view) | Size |
|--|----------|
| siMitfvssiLuc.complete.txt | 6.1 MB |
| siMitfvssiLuc.down.txt | 521.9 KB |
|  siMitfvssiLuc.up.txt | 587.0 KB |

Exercise 2: get annotations for a set of genes

- 2. Use the file `siMitfvssiLuc.up.txt` to extract gene annotations for those genes. Annotation to extract are : gene stable IDs, Chromosome/scaffold name, Gene start, Gene end, strand, Gene name, Gene type. Save the results to a compressed TSV file. (don't close the Ensembl/Biomart window once done)
 - Tip: columns are in the same order as columns are selected
- 3. Upload the file `siMitfvssiLuc.up.txt` and the annotation file (`mart_export.txt.gz`) you obtained from Ensembl/BioMart to Galaxy into your current history "RNA-seq data analysis".
 - **Type:** tabular
 - **Genome:** hg38

Exercise 2: get annotations for a set of genes

- 4. Use the tool “**Join two Datasets**” to merge the two datasets (**siMitfvssiLuc.up.txt** and **mart_export.txt.gz**) based on the column that contains Ensembl Gene IDs in each dataset.
 - Ensembl Gene IDs are used as unique identifiers common to the two datasets. For a given gene, data spread in the two files are going to be merged in the same line in the newly generated file.
 - Tip 1: Keep the header lines

Rename the dataset siMitfvssiLuc.up.annot.txt

- 5. Is there lncRNAs in the upregulated genes? Use the tool “**Filter** data on any column using simple expressions” to search for “lncRNA” (<- this exact case) in the dataset siMitfvssiLuc.up.annot.txt.
 - Tip 1: Search “lncRNA” in the column containing Gene types
 - Tip 2: c3 refers to column 3 of a dataset.
 - Tip 3 : look at examples below the form to help you find the correct syntax

Exercise 2: get annotations for a set of genes

- Bonus question: go back to Ensembl/BioMart. You want to extract sequences of all promoters of the up-regulated genes (the ones from the file siMitfvssiLuc.up.txt) to run a *de novo* motif discovery and search for over represented nucleotide sequence. Retrieve the 200nt upstream of these genes. Header should contain Gene stable ID, Transcript stable ID, Gene name and Gene description.

Exercise 3: get annotations in the genome

- 1. How many genes are located in the genomic region: **2:208226227-208276270**
- 2. Extract the coordinates of all human genes located on chromosomes (exclude scaffolds). Information to extract for each gene (**beware of the order you tick the features to extract**): Chromosome/scaffold name, Gene Start (bp), Gene End (bp), Gene stable ID, Gene Name and strand.
 1. Download the resulting file on your computer as a TSV file.
 2. Once downloaded rename the file **hg38_ens105.bed**
 3. Open the file with a text editor and remove the first line (the one with headers)
 - **Congrats, you've just created a BED file!**