

Data mining with Ensembl Biomart

Stéphanie Le Gras
(slegras@igbmc.fr)

Guidelines

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

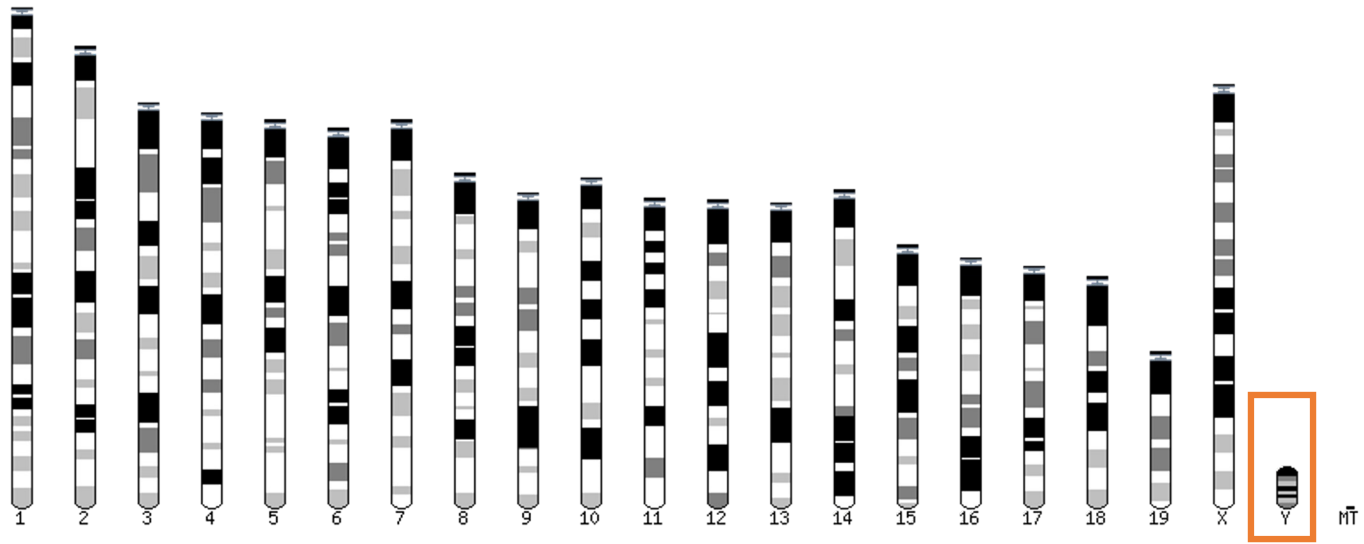
Genome builds

SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
MAMMALS				
Human	hs1	Jan. 2022	T2T Consortium CHM13v2.0	Available
	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)

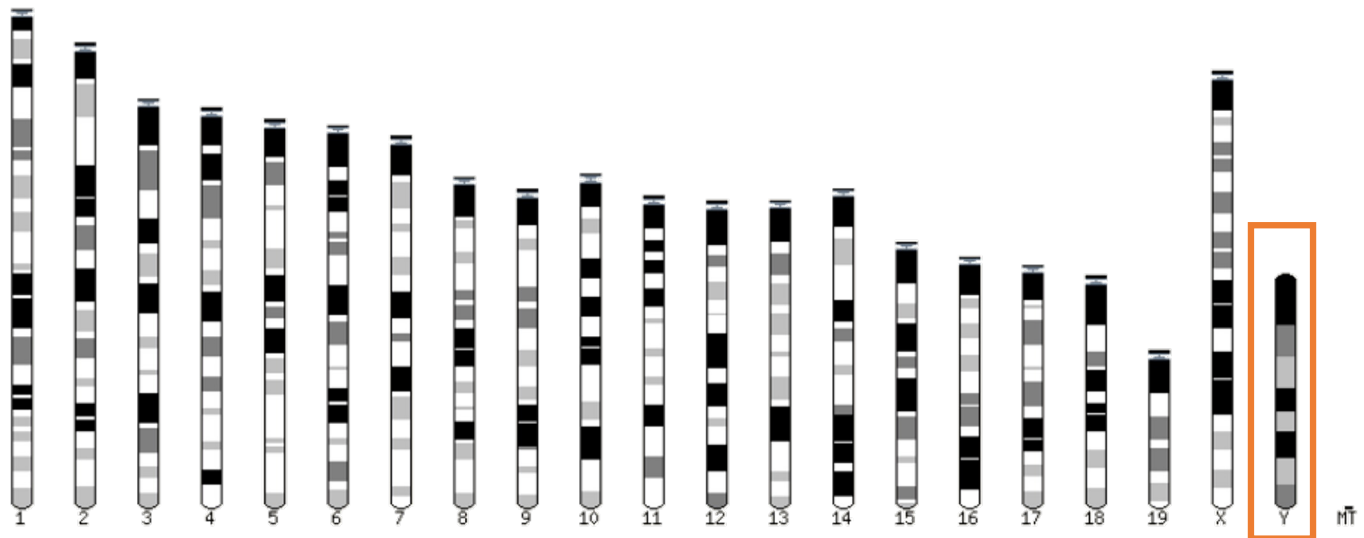
Source: <https://genome.ucsc.edu/FAQ/FAQreleases.html>

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



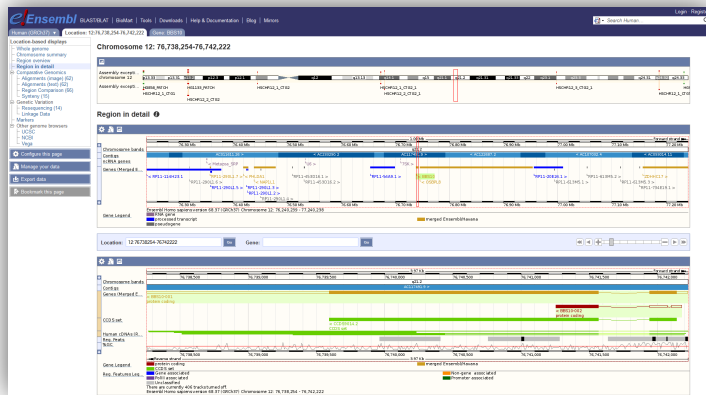
Genome browsers

Genome Browsers

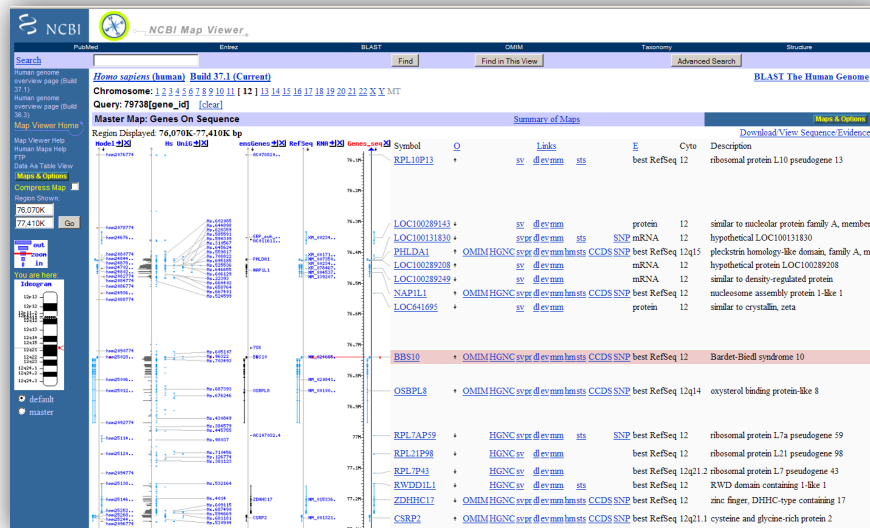
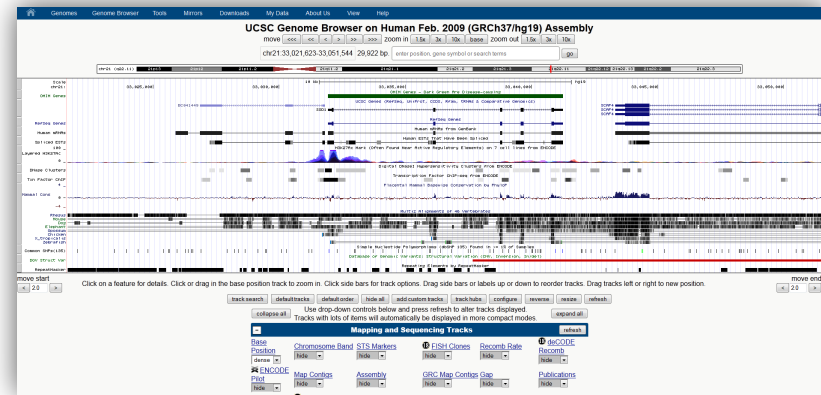
- Graphical interface to display genomic data
- Visualize and browse entire genomes with annotated data
 - Gene prediction and structure
 - Proteins,
 - Expression,
 - Regulation,
 - Variation,
 - Comparative analysis...

There are Genome Browsers...

EBI - Ensembl

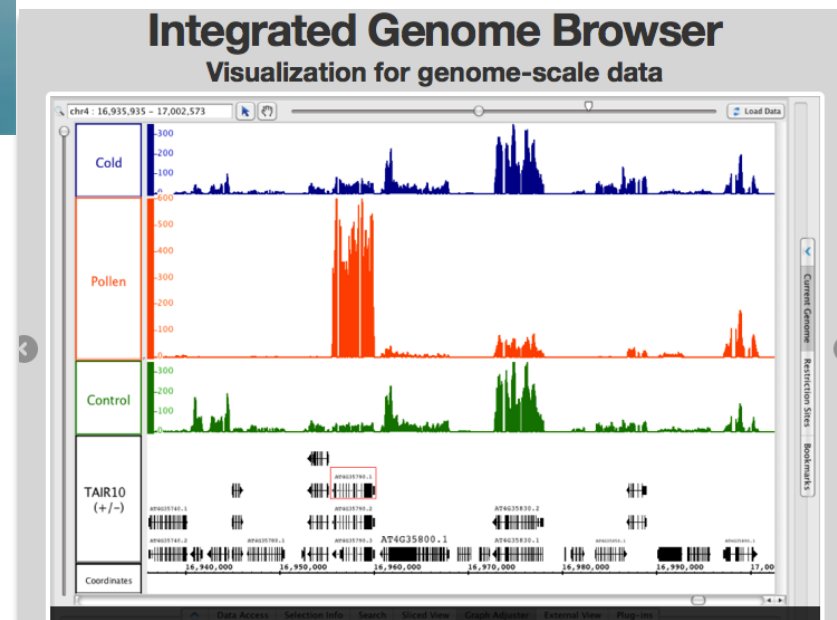


UCSC – Genome Browser



NCBI – Genome Data Viewer

And Genome browsers...






Getting access to genomic data: ENSEMBL/BIOmart

Access Ensembl's data




Web site

The screenshot shows the Ensembl website homepage. At the top, there is a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar is located in the top right corner. Below the navigation bar, there are several sections: 'Tools' with links to BLAST/BLAT and VEP; 'BioMart' with a description of expert custom datasets; 'Variant Effect Predictor' with a description of analyzing variants; and 'Ensembl Release 109 (Feb 2023)' with a list of updates. There is also a search bar for species and a section for 'All genomes' with a dropdown menu and a list of favourite genomes (Human, Mouse, Zebrafish). At the bottom, there are several icons representing different tools and services, and a footer with the EMBL-EBI logo and a URL.

-  User friendly
-  Straightforward
-  Only one request at once

Mining tool: BioMart

The screenshot shows the Ensembl BioMart interface. At the top, there is a navigation bar with links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar is located in the top right corner. Below the navigation bar, there is a 'Dataset' section with a dropdown menu for 'CHOOSE DATABASE'. The main area is currently empty, showing 'None selected'. At the bottom, there is a small text box with a warning message: '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. Note that queries that run for longer than 5 hours will be terminated even when submitted this way. If this happens please reformat your query or contact us for details on how to reformat this.'

-  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

The screenshot shows the Ensembl website 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 search bar with the text "Search all species..." and a "Login/Register" link. Below the navigation, there are three main sections: "Tools" with a link to "All tools", "BioMart" (highlighted with an orange arrow and a box labeled "Biomart"), and "Variant Effect Predictor" with a link to "Variant Effect Predictor". The BioMart section description reads: "Export custom datasets from Ensembl with this data-mining tool...". The VEP section description reads: "Analyse your own variants and predict the functional consequences of known and unknown variants". Below these sections is a search box with a dropdown menu set to "All species" and a "Go" button. Below the search box, there are examples of search terms: "e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease". To the right of the search box, there is a paragraph of text about Ensembl's capabilities and a list of "Ensembl Release 109 (Feb 2023)" updates, including new gene sets for donkey and horse, updated SIFT and PolyPhen-2 missense variant pathogenicity, new VEP plugins for UTR annotation, and new ATAC-seq tracks for fish species. Below the release news is a link to "More release news". At the bottom of the screenshot, there is a section titled "Ensembl Rapid Release" with a "Go" button and text about new assemblies with gene and protein annotation every two weeks.

- 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, there is a navigation bar with the Ensembl logo and links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar on the right contains the text "Search all species...". Below the navigation bar, there are buttons for "New", "Count", and "Results". A secondary bar contains buttons for "URL", "XML", "Perl", and "Help".

The main content area is titled "Dataset" and shows "[None selected]". A dropdown menu is open, displaying a list of datasets. The first item is "Ensembl Genes 109", which is highlighted in blue. Below it, there is a section titled "CHOOSE DATASET -" with a list of species and their corresponding genome assemblies. The list includes:

- Chicken genes (bGalGal1.mat.broiler.GRCg7b)
- Human genes (GRCh38.p13)
- Mouse genes (GRCm39)
- Rat genes (mRatBN7.2)
- Zebrafish genes (GRCz11)

Below this list, there is a dashed line and a long list of other species and their genome assemblies, including Abingdon island giant tortoise, African ostrich, Algerian mouse, Alpaca, Alpine marmot, Amazon molly, American bison, American black bear, American mink, Arabian camel, Arctic ground squirrel, Argentine black and white tegu, Armadillo, Asian bonytongue, Atlantic cod, Atlantic herring, Atlantic salmon, and Australian saltwater crocodile.

Two orange arrows point from the text "First choose the database and dataset" to the "Ensembl Genes 109" dropdown and the "CHOOSE DATASET -" section header.

At the bottom of the page, there is a small text box that reads: "Warning for more than 5 minutes are terminated. If you have choose have the results emailed to you."

Example: Step 2 (Filter)

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 are on the right. Below the navigation bar, there are buttons for 'New', 'Count', and 'Results'. The main content area is divided into several sections:

- Dataset:** Human genes (GRCh38.p12)
- Filters:** A box containing the following settings:
 - Chromosome/scaffold: 1
 - Start: 78895
 - End: 224561
- REGION:** A section with a checked checkbox for 'Chromosome/scaffold' and a list of chromosomes from 1 to 20. Chromosome 1 is highlighted.
- Coordinates:** A section with a checked checkbox for 'Coordinates' and two input fields: 'Start' (78895) and 'End' (224561).

Limit to chromosome 1

Limit to given coordinates

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 3 (Count results)

e!Ensembl BLAST/BLAT Login/Register

Search all species...

Compute match count

New Count Results URL XML Perl Help

Dataset 12 / 69299 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]

REGION:
 Chromosome/scaffold

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Coordinates
Start
End

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)

The screenshot shows the Ensembl BioMart interface. The top navigation bar includes the Ensembl logo, links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog, along with a search bar and a Login/Register link. Below the navigation bar, there are buttons for 'New', 'Count', and 'Results', and a secondary bar with 'URL', 'XML', 'Perl', and 'Help' options.

The main content area is titled 'Please select columns to be included in the output and hit 'Results' when ready'. Below this, 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:** 12 / 69299 Genes, Human genes (GRCh38.p13)
- Filters:** Chromosome/scaffold: 1, Start: 78895, End: 224561
- Attributes:** A list of attributes is shown, with 'Gene stable ID' and 'Transcript stable ID' highlighted in an orange box.
- Ensembl:** A list of attributes is shown, with 'Gene stable ID' and 'Transcript stable ID' checked. An orange callout bubble points to these two attributes with the text 'Select attributes to be output'.

The 'Ensembl' section includes the following attributes:

- 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)

The 'Ensembl' section also includes the following attributes:

- 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)

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

Dataset 12 / 69299 Genes
Human genes (GRCh38.p13)

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

Attributes
Gene stable ID
Transcript stable ID

Dataset
[None Selected]

Export all results to Unique results only

Email notification to

View rows as 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	...
ENSG000000081189	1716	...
ENSG000000106772	3063	...
ENSG000000124942	309	...
ENSG000000142871	243	...
ENSG000000143341	3760	...
ENSG000000154556	352	...
ENSG000000185565	679	...
ENSG000000163328	136	...
ENSG000000064042	1160	...
ENSG000000114423	2293	...

mart_export.txt (from Ensembl/Biomart)


Gene stable ID	Gene name	Chro...
ENSG00000000971	CFH 1	19665187...
ENSG00000001461	NIPAL3	1 2441...
ENSG000000124942	AHNAK	11 624...
ENSG00000002330	BAD 11	642698...
ENSG00000002549	LAP3 4	175771...
ENSG00000002586	CD99 X	269113...
ENSG00000002834	LASP117	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...
ENSG000000124942	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.

SARTools DESeq2 tables    

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!**