

# Data mining with Ensembl Biomart

Stéphanie Le Gras  
([slegras@igbmc.fr](mailto:slegras@igbmc.fr))

# Guidelines

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

# Genome Sequencing

Example: Human genome

- 2000: First draft of the human genome
- 2003: Human genome sequencing complete



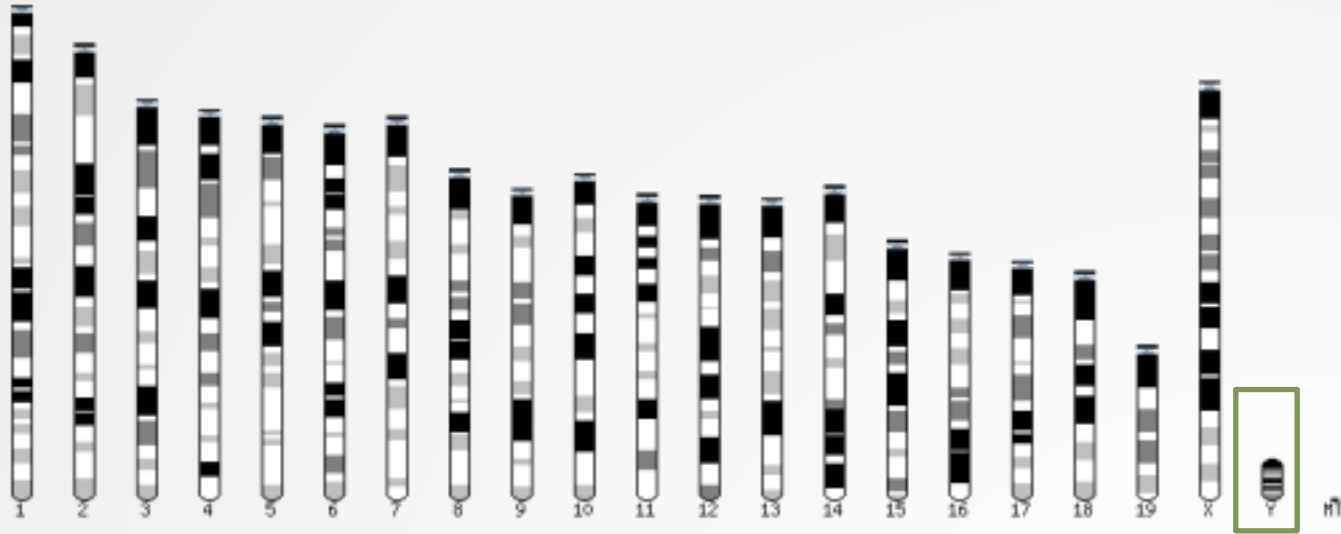
# Genome builds

| SPECIES        | UCSC VERSION | RELEASE DATE | RELEASE NAME                       | STATUS               |
|----------------|--------------|--------------|------------------------------------|----------------------|
| <b>MAMMALS</b> |              |              |                                    |                      |
| 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) |

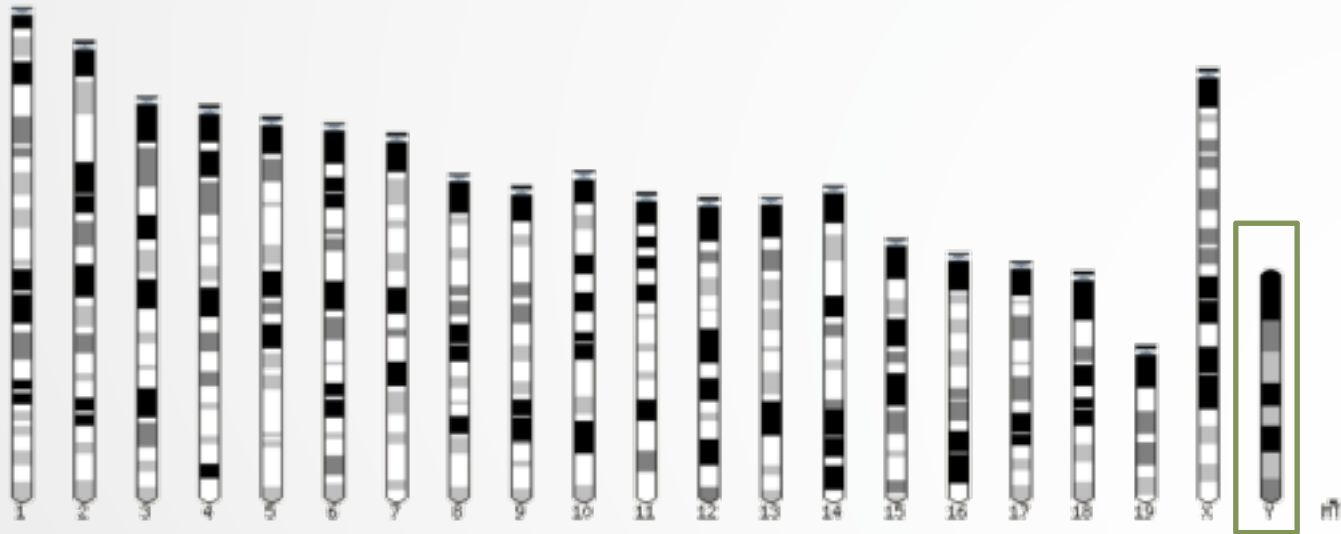
Source: <https://genome.ucsc.edu/FAQ/FAQreleases.html>

# 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

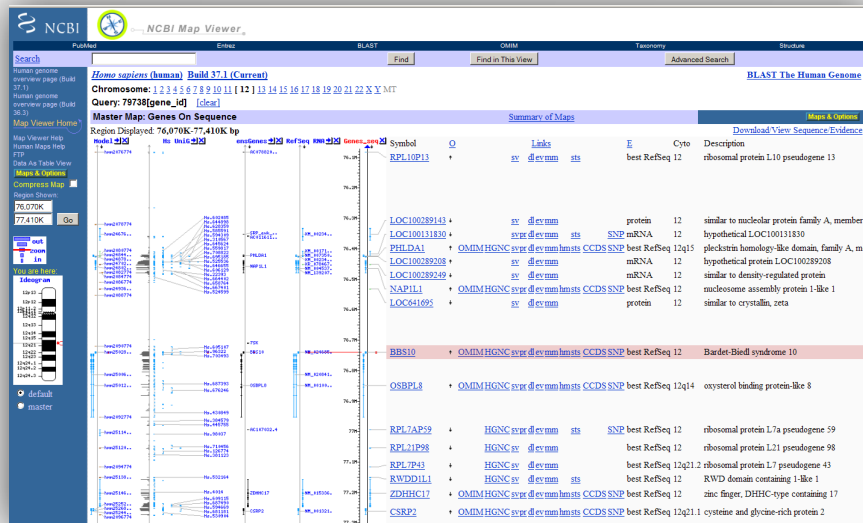
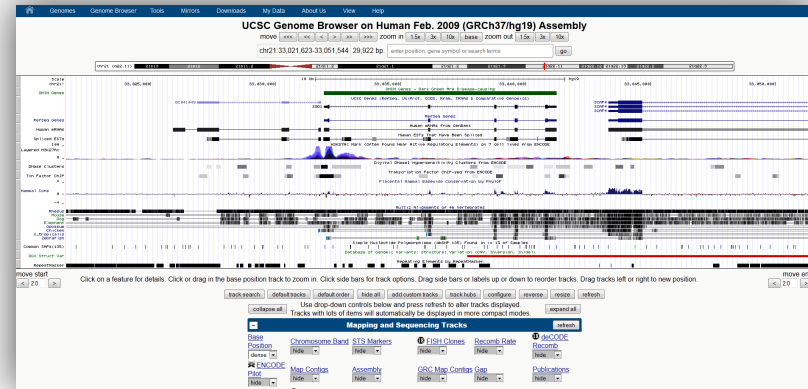
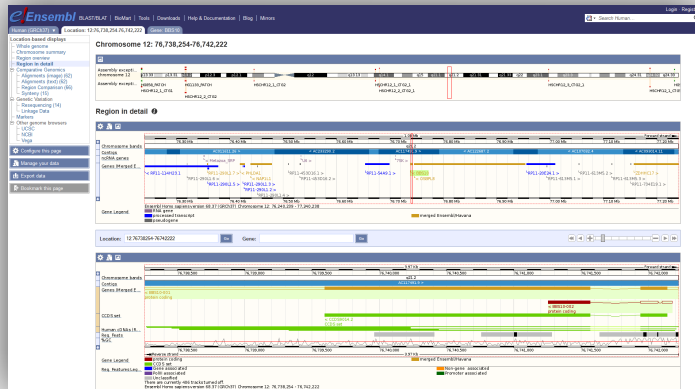
- Graphical interface to display genomic data
- Visualize and browse entire genomes with annotated data
  - Gene prediction and structure
  - Proteins,
  - Expression,
  - Regulation,
  - Variation,
  - Comparative analysis...



# There are Genome Browsers...

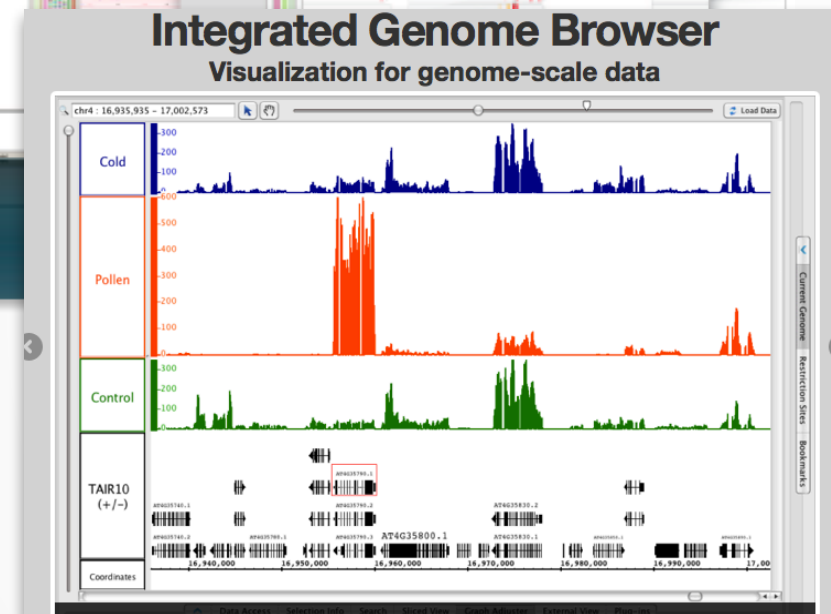
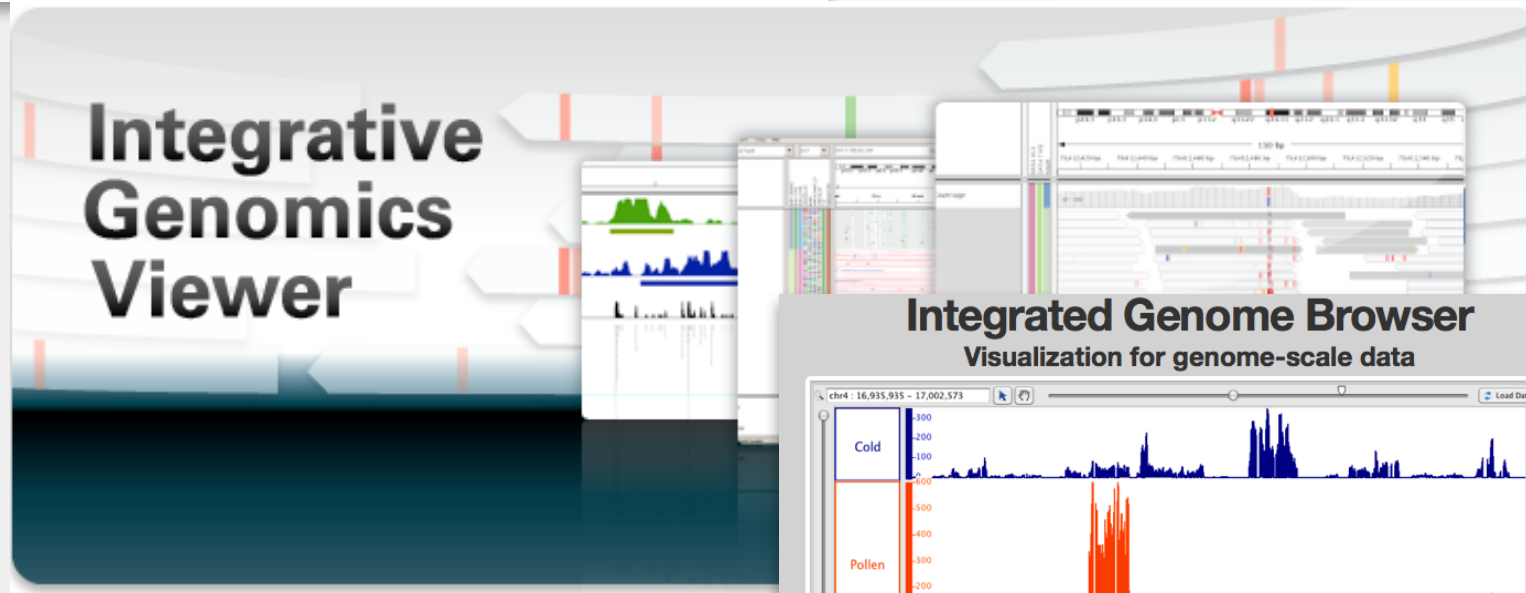
EBI - Ensembl

UCSC - Genome Browser



NCBI - Map Viewer

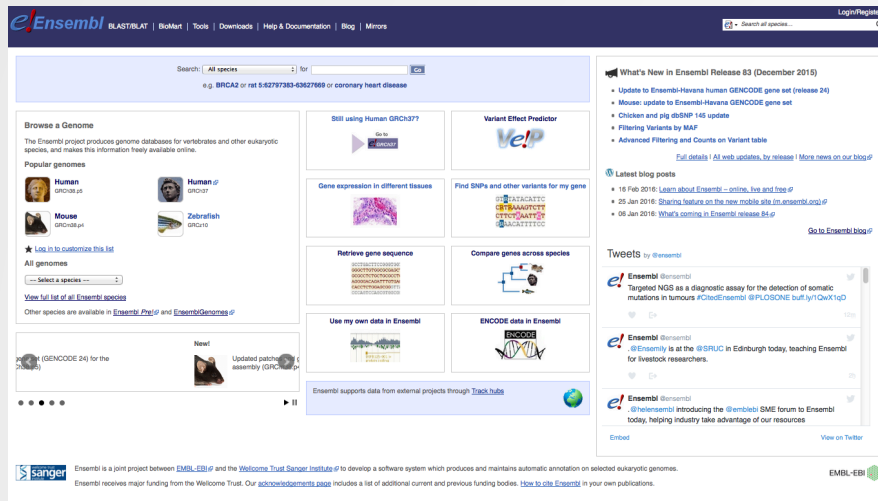
# And Genome browsers...



Getting access to genomic  
data:  
ENSEMBL/BIOmart

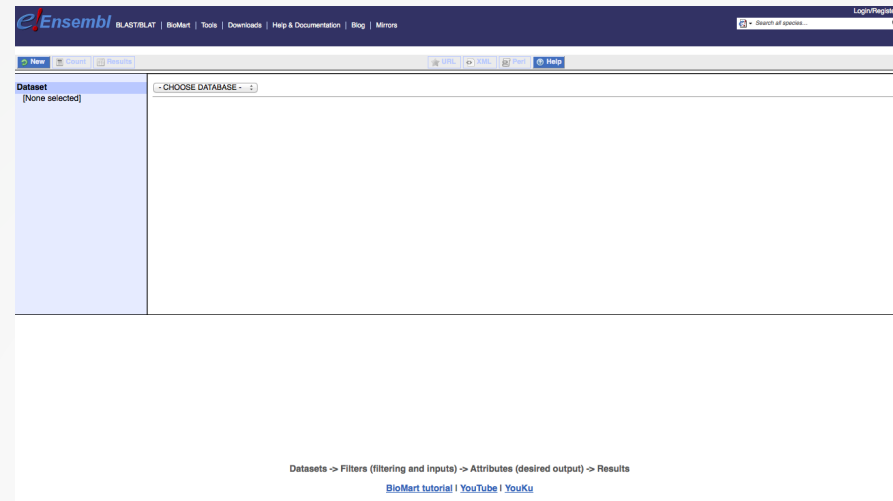
# Access Ensembl's data

Web site






The screenshot shows the Ensembl web site homepage. It features a search bar at the top with the text "Search all species" and a "Login/Register" link. Below the search bar, there are several navigation links: "BLAST/BLAT", "BioMart", "Tools", "Downloads", "Help & Documentation", "Blog", and "Mirrors". The main content area is divided into several sections: "Browse a Genome" with a search bar and a list of popular genomes (Human, Mouse, Zebrafish); "What's New in Ensembl Release 83 (December 2015)" with a list of updates; "Latest blog posts" with a list of recent posts; "Tweets by @ensembl" with a list of tweets; and "Ensembl supports data from external projects through Track hubs".

Mining tool: BioMart



The screenshot shows the Ensembl BioMart interface. It features a search bar at the top with the text "Search all species" and a "Login/Register" link. Below the search bar, there are several navigation links: "BLAST/BLAT", "BioMart", "Tools", "Downloads", "Help & Documentation", "Blog", and "Mirrors". The main content area is divided into several sections: "Dataset" with a "CHOOSE DATABASE" dropdown menu; "Filters" with a list of filters; "Attributes" with a list of attributes; and "Results" with a list of results. The interface is designed for complex queries and data mining.

-  User friendly
-  Straightforward
-  Only one request at once

-  Get answer to complex query
-  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 search engine that can find multiple terms and put them into a table format
- No programming required!

# Many uses of BioMart

The image displays three overlapping screenshots of the BioMart web interface, demonstrating its integration with different data sources:

- Top Screenshot (UniProt):** Shows the UniProt logo and the BioMart interface. The navigation bar includes "Services", "Research", "Training", and "About us". The main header features the UniProt and BioMart logos. Below the header, there are buttons for "New", "Count", and "Results". A dropdown menu labeled "Dataset" is currently set to "[None selected]".
- Middle Screenshot (InterPro):** Shows the InterPro logo and the BioMart interface. The navigation bar includes "Services", "Research", "Training", and "About us". The main header features the InterPro logo and the BioMart logo. Below the header, there are buttons for "New", "Count", and "Results". A dropdown menu labeled "Dataset" is currently set to "[None selected]".
- Bottom Screenshot (Ensembl):** Shows the Ensembl logo and the BioMart interface. The navigation bar includes "BLAST/BLAT", "BioMart", "Tools", "Downloads", "Help & Documentation", "Blog", and "Mirrors". The main header features the Ensembl logo and the BioMart logo. Below the header, there are buttons for "New", "Count", and "Results". A dropdown menu labeled "Dataset" is currently set to "[None selected]".

# BioMart/Ensembl

Search: **Biomart** for

63627669 or rs699 or coronary heart disease

### Browse a Genome

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.

### Favourite genomes

- Human** GRCh38.p7
- Human** GRCh37
- Mouse** GRCm38.p5
- Zebrafish** GRCz10

[Edit favourites](#)

### Still using Human GRCh37?

Go to

### Variant Effect Predictor

### Gene expression in different tissues

### Find SNPs and other variants for my gene

```
GTGATACATTC  
CRTAAAAGTCTT  
CTTCTAAATTCT  
GRRACATTTCC
```

### Retrieve gene sequence

```
GCCTGACTCCGGGTGG  
GGGCTTGTGGGGAGC  
GGGCTCTGCTGGGCT  
AGGGGACAGATTTGTGA  
CACCTCTGGAGCGGTTI
```

### Compare genes across species

### What's New in Ensembl Release 87 (December 2016)

- **Chicken:** Gene set update
- **Mouse:** update to Ensembl-Havana GENCODE gene set
- **New dbSNP data for Sheep**
- **Zebrafish:** update to Ensembl-Havana merged gene set
- **New table on Regulation Summary page**

[Full details](#) | [All web updates, by release](#) | [More news on our blog](#)

- 05 Jan 2017: [So you want to run an Ensembl workshop](#)
- 19 Dec 2016: [Regulation FTP resources restructured in Ensembl 87](#)
- 15 Dec 2016: [New zebrafish developmental RNA-seq data in Ensembl](#)

[Go to Ensembl blog](#)

### Tweets by @ensembl

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

# Example: Step 1 (Select datasets)

**e!Ensembl** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors Login/Register

[New](#) [Count](#) [Results](#) [URL](#) [XML](#) [Perl](#) [Help](#)

**Dataset**  
[None selected]

Ensembl Genes 87

- CHOOSE DATASET -
- CHOOSE DATASET -
- Chicken genes (Gallus\_gallus-5.0)
- Human genes (GRCh38.p7)
- Mouse genes (GRCm38.p5)
- Rat genes (Rnor\_6.0)
- Zebrafish genes (GRCz10)
- 
- Alpaca genes (vicPac1)
- Amazon molly genes (Poecilia\_formosa-5.1.2)
- Anole lizard genes (AnoCar2.0)
- Armadillo genes (Dasnov3.0)
- Bushbaby genes (OtoGar3)
- C.intestinalis genes (KH)
- C.savignyi genes (CSAV 2.0)
- Caenorhabditis elegans genes (WBcel235)
- Cat genes (Felis\_catus\_6.2)
- Cave fish genes (AstMex102)
- Chimpanzee genes (CHIMP2.1.4)
- Chinese softshell turtle genes (PelSin\_1.0)
- Cod genes (gadMor1)

First choose database and dataset



# Example: Step 2 (Filter)



New Count Results

★ URL XML Perl Help

Dataset  
Human genes (GRCh38.p7)

Filters  
Chromosome/scaffold: 1  
Gene Start (bp): 78895  
Gene End (bp): 10000000

Attributes  
Gene ID  
Transcript ID

Dataset  
[None Selected]

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)

REGION:

Chromosome/scaffold

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

Base pair  
Gene Start (bp)  
Gene End (bp)

78895

104561

Band  
Band Start  
Band End

Limit to chromosome 1

Limit to given coordinates

# Example: Step 3 (Count results)

**Dataset** 2 / 63305 Genes  
Human genes (GRCh36.p7)

**Filters**  
Chromosome/scaffold: 1  
Gene Start (bp): 78895  
Gene End (bp): 104561

**Attributes**  
Gene ID  
Transcript ID

**Dataset**  
[None Selected]

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)

REGION:

- Chromosome/scaffold   
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20
- Base pair  
Gene Start (bp)   
Gene End (bp)
- Band  
Band Start   
Band End
- Marker

# Example: Step 4 (Select attributes)

**Dataset**  
Human genes (GRCh38.p7)

**Filters**  
Chromosome/scaffold: 1  
Gene Start (bp): 78895  
Gene End (bp): 104561

**Attributes**  
 Gene ID  
 Transcript 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**     **Variant (Somatic)**  
 **Homologues**     **Sequences**

GENE:

**Ensembl**

- Gene ID
- Transcript ID
- Protein ID
- Exon ID
- Description
- Chromosome/scaffold name
- Gene Start (bp)
- Gene End (bp)
- Strand
- Band
- Transcript Start (bp)
- Transcript End (bp)
- Transcription Start Site (TSS)
- Transcript length (including UTRs and CDS)
- Transcript Support Level (TSL)
- GENCODE basic annotation

**Phenotype**

- Phenotype description
- Source name
- Study External Reference
- APPRIS annotation
- Associated Gene Name
- Associated Gene Source
- Associated Transcript Name
- Associated Transcript Source
- Transcript count
- % GC content
- Gene type
- Transcript type
- Source (gene)
- Source (transcript)
- Status (gene)
- Status (transcript)
- Version (gene)
- Version (transcript)
- Strain name
- Strain gender
- P value

Select attributes to be output

# Example: Step 4 (get results)

**e!Ensembl** BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors Login/Register

Search all species...

**New** **Count** **Results** [URL](#) [XML](#) [Perl](#) [Help](#)

**Dataset**  
Human genes (GRCh38.p7)

**Filters**  
Chromosome/scaffold: 1  
Gene Start (bp): 78895  
Gene End (bp): 104561

**Attributes**  
Gene ID  
Transcript ID

**Dataset**  
[None Selected]

Export all results to    Unique results only

Email notification to


View  rows as   Unique results only

| Gene ID                         | Transcript ID                   |
|---------------------------------|---------------------------------|
| <a href="#">ENSG00000238009</a> | <a href="#">ENST00000466430</a> |
| <a href="#">ENSG00000238009</a> | <a href="#">ENST00000477740</a> |
| <a href="#">ENSG00000238009</a> | <a href="#">ENST00000471248</a> |
| <a href="#">ENSG00000238009</a> | <a href="#">ENST00000453576</a> |
| <a href="#">ENSG00000238009</a> | <a href="#">ENST00000610542</a> |
| <a href="#">ENSG00000239945</a> | <a href="#">ENST00000495576</a> |

# Exercise 1: get annotations of a gene

- 1. Using Ensembl/BioMart, retrieve all transcripts IDs and the gene ID of IDH1 gene (human). How many transcripts the gene IDH1 has?
  - Use Ensembl Gene v87, for Human GRCh38.p7
  - Click on Filters :
    - Expand the GENE section
    - Select « Input external references ID list »
    - Select HGNC symbol(s) in the drop down menu
    - Enter IDH1 in the text box
  - Click on Attributes :
    - Select “Features” (top panel, selected by default)
    - Select Gene ID, Transcript ID, Associated Gene Name
- 2. Extract all exon sequences of the IDH1 gene in fasta format. Headers will contain the Associated gene names, transcript IDs and Exon IDs.
- 3. Extract all coding sequences of the IDH1 gene in fasta format. Headers will contain the transcript IDs and Exon IDs.
- 4. Retrieve GO-terms associated to the IDH1 gene (select GO Term Name, GO domain and GO Term Accession along with Gene ID, Transcript ID and Associated 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 ID, Transcript ID and Associated Gene Name)

## Exercise 2: get annotations for a set of genes

- Annotate the file siMitfvssiLuc.up.txt you have generated using SARTools with gene annotations extracted from Ensembl/BioMart
  - If you encountered any trouble with the generation of the dataset
    - go to GalaxEast (<http://use.galaxeast.fr>)
    - go to Shared Data/ Data Libraries/ CNRS training / RNAseq / statistical\_analysis.
    - Import the dataset SARTools\_DESeq2\_tables to your history.
    - Click on  to display the content of the dataset and download the file siMitfvssiLuc.up.txt (click right, save ...)
- 1. Open the file siMitfvssiLuc.up.txt and change the name of the column which contains Ensembl Gene IDs to “Gene ID”. Save the change.
- 2. Use the file siMitfvssiLuc.up.txt to extract gene annotations for those genes. Annotation to extract are : gene IDs, chromosome, start of gene, end of gene, strand, associated gene name, gene type. Save the results to a compressed TSV file. (don't close the Ensembl/Biomart window once done)
- 3. Upload the file siMitfvssiLuc.up.txt and the annotation file you obtained from Ensembl/BioMart to GalaxEast into your current history “CNRS training”.
  - 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 based on the Gene IDs.
  - 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.
- 5. rename the generated dataset in 4. to `siMitfvssiLuc.up.annot.txt`
- 6. Is there lncRNAs in the upregulated genes? Use the tool “Filter data on any column using simple expressions” to search for “lincRNA” in the dataset `siMitfvssiLuc.up.annot.txt`
- 7. Go back to Ensembl/BioMart. You want to run a *de novo* motif discovery on all promoters of the up-regulated genes (the ones from the file `siMitfvssiLuc.up.txt`). Extract the promoter sequences of all up-regulated genes: retrieve the 2kb upstream of the transcripts of these genes.

## 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: Gene ID, Chromosome/scaffold name, Gene Start (bp), Gene End (bp), strand and associated Gene Name