

Data mining with Ensembl Biomart

Stéphanie Le Gras
(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



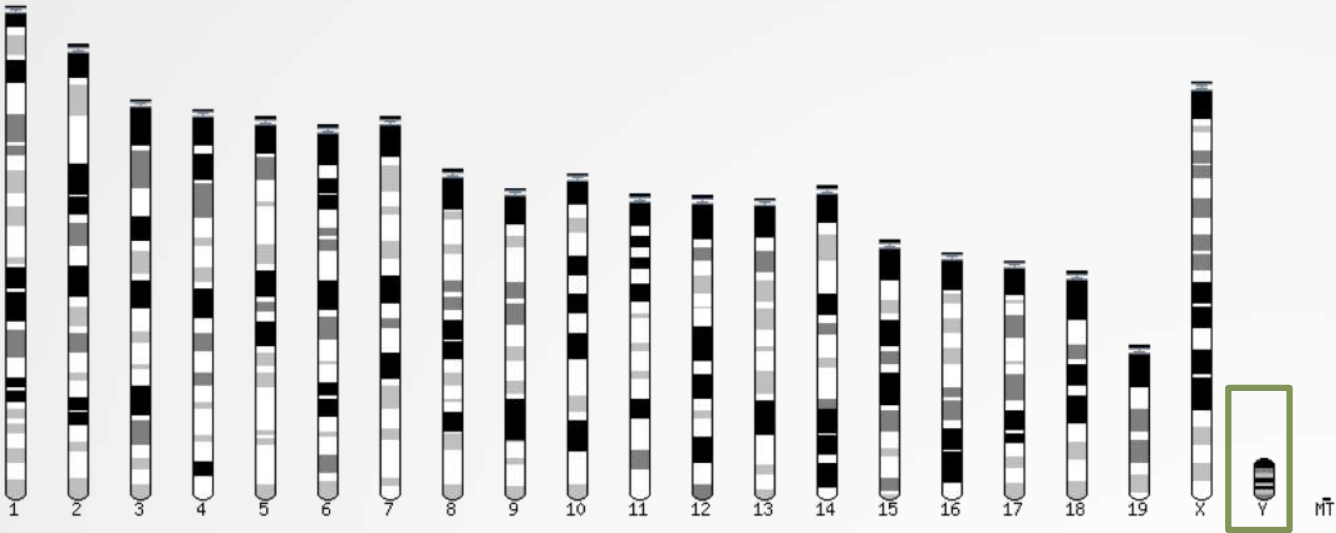
Genome builds

SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
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)

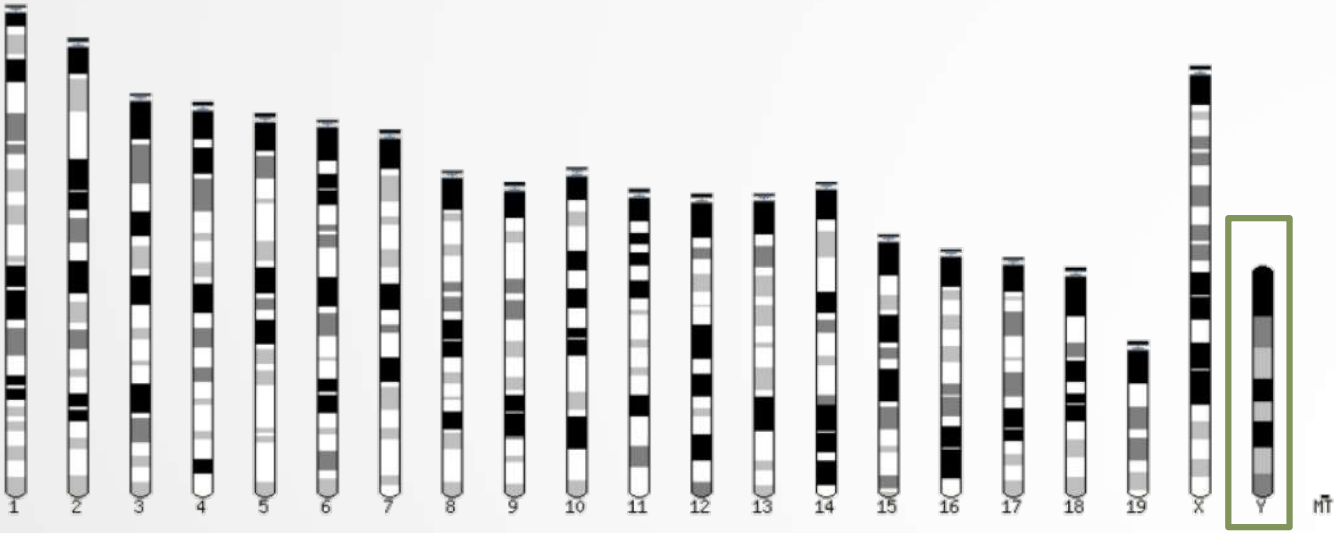
Source: <https://genome.ucsc.edu/FAQ/FAQreleases.html>

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 Browser

Genome browsers

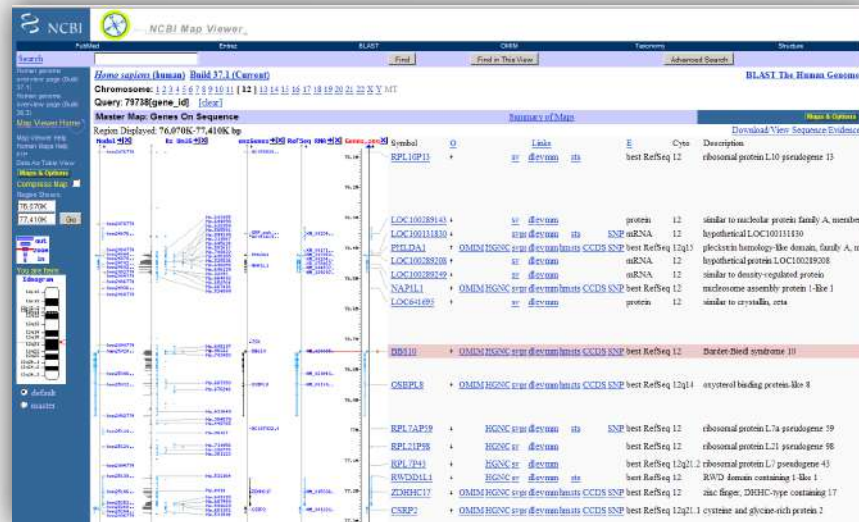
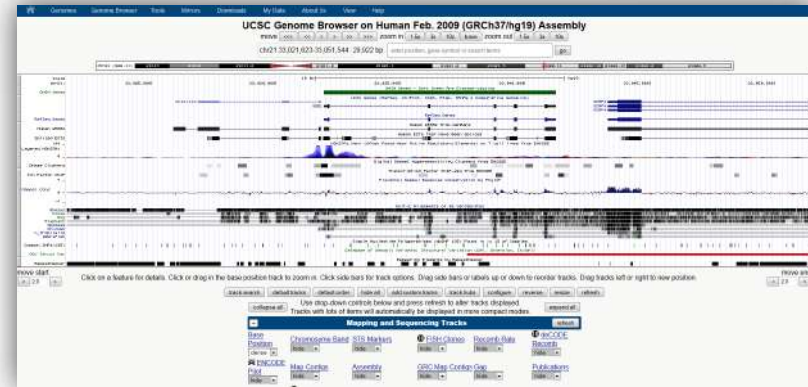
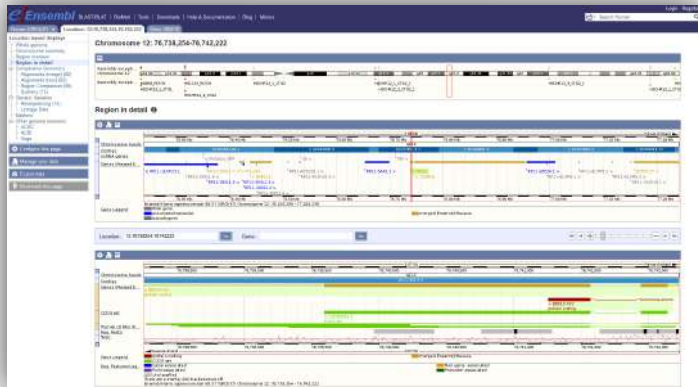
Genome Browsers

- 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...

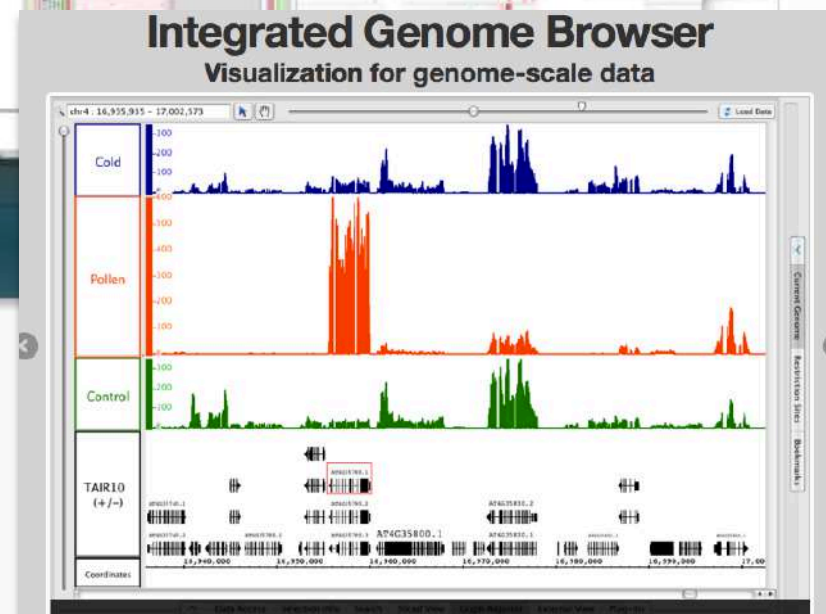
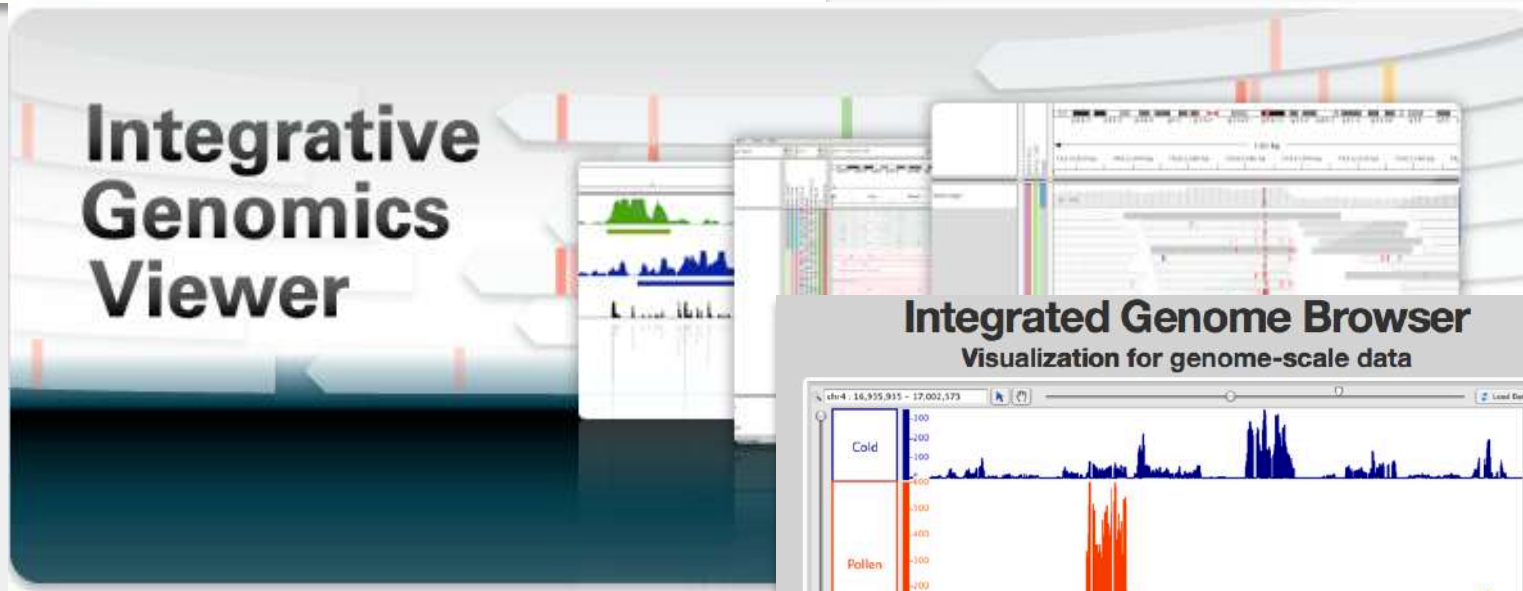
EBI - Ensembl

UCSC - Genome Browser



NCBI - Map Viewer

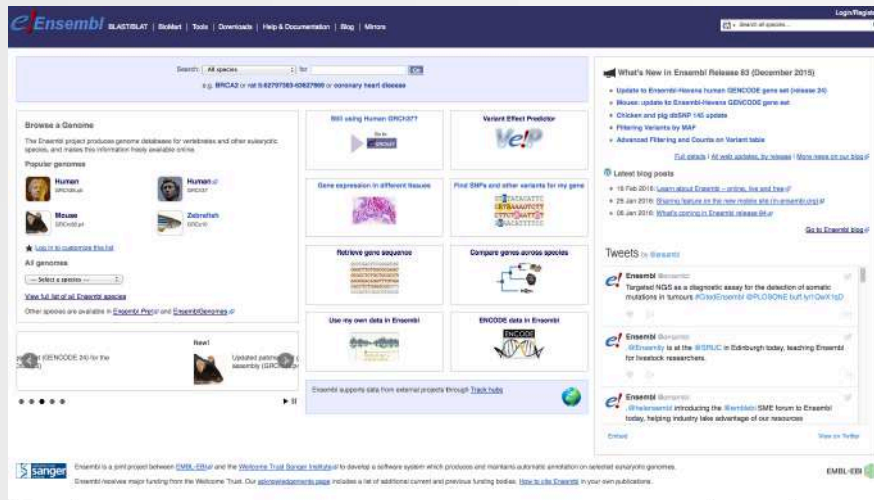
And Genome browsers...



Getting access to genomic
data:
ENSEMBL/BIOmart

Access Ensembl's data

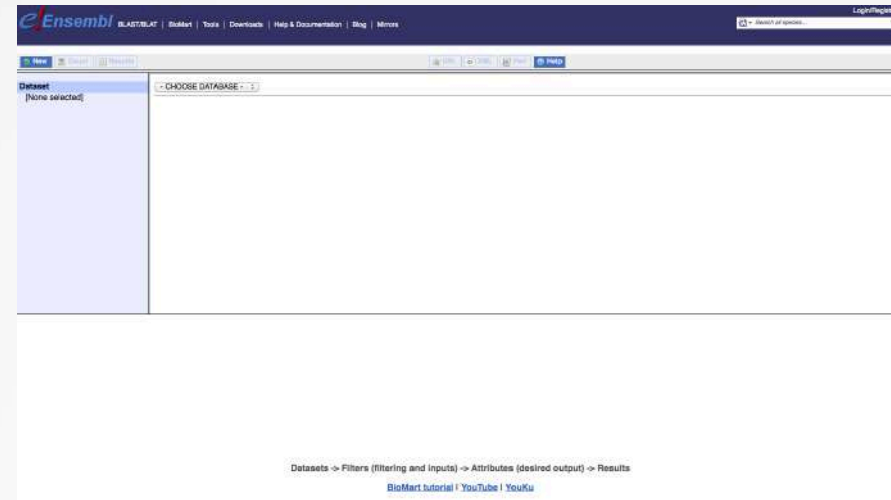
Web site






The screenshot shows the Ensembl web site homepage. It features a search bar at the top with the text "Search: All species" and a search button. Below the search bar, there are several sections: "Browse a Genome" with a list of popular genomes (Human, Mouse, Zebrafish), "What's New in Ensembl Release 83 (December 2015)", "Gene expression in different tissues", "Use my own data in Ensembl!", and "ENCODE data in Ensembl!". The page is designed to be user-friendly and straightforward.

-  User friendly
-  Straightforward
-  Only one request at once

Mining tool: BioMart



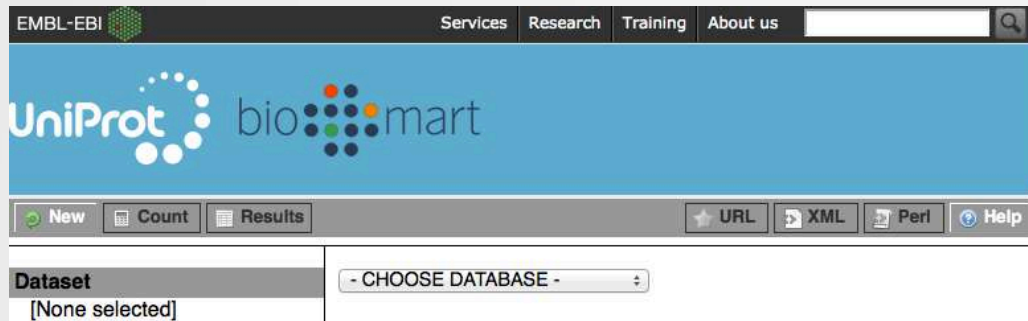
The screenshot shows the BioMart web site interface. It features a search bar at the top with the text "Search: All species" and a search button. Below the search bar, there is a "Dataset" section with a dropdown menu labeled "CHOOSE DATABASE". The page is designed to be user-friendly and straightforward.

-  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



EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

UniProt bio:mart

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

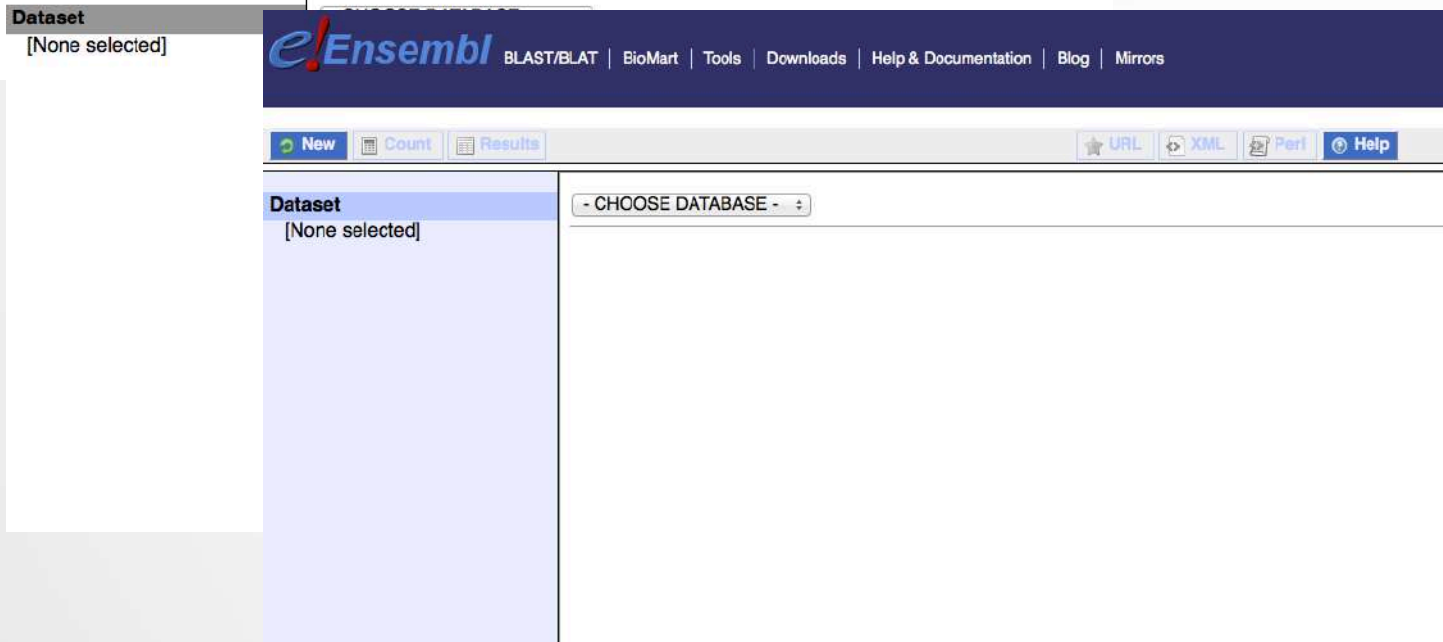
Dataset
[None selected] - CHOOSE DATABASE -



EMBL-EBI [Services](#) [Research](#) [Training](#) [About us](#)

InterPro bio:mart
Protein sequence analysis & classification

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



Dataset
[None selected]

e!Ensembl [BLAST/BLAT](#) [BioMart](#) [Tools](#) [Downloads](#) [Help & Documentation](#) [Blog](#) [Mirrors](#)

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

Dataset
[None selected] - CHOOSE DATABASE -

BioMart/Ensembl

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

Search all species...

Tools **BioMart >** **Biomart** **Variant Effect Predictor >**

[All tools](#) Export custom datasets from Ensembl with this data-mining tool or your DNA or protein sequence

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes **Favourite genomes**

-- Select a species -- Human

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

Other news from our blog

- 01 Mar 2019: [Getting to know us: Guy from Ensembl Plants](#)

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

Example: Step 1 (Select datasets)

The screenshot shows the Ensembl genome browser 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 for species and a Login/Register link. The main content area is titled 'Dataset' and shows '[None selected]'. A dropdown menu is open, displaying a list of datasets. The first two items are highlighted in blue: 'Chicken genes (GRCg6a)' and 'Human genes (GRCh38.p12)'. A green box highlights these two items, and a callout box with a green border points to them, containing the text 'First choose database and dataset'.

Ensembl Genes 95

✓ - CHOOSE DATASET -

- Chicken genes (GRCg6a)
- Human genes (GRCh38.p12)
- Mouse genes (GRCm38.p6)
- Rat genes (Rnor_6.0)
- Zebrafish genes (GRCz11)

- Agassiz's desert tortoise genes (ASM289641v1)
- Algerian mouse genes (SPRET_EiJ_v1)
- Alpaca genes (vicPac1)
- Amazon molly genes (Poecilia_formosa-5.1.2)
- American black bear genes (ASM334442v1)
- Angola colobus genes (Cang.pa_1.0)
- Anole lizard genes (AnoCar2.0)
- Armadillo genes (Dasnov3.0)
- Asian bonytongue genes (ASM162426v1)
- Ballan wrasse genes (BallGen_V1)
- Bicolor damselfish genes (Stegastes_partitus-1.0.2)
- Black snub-nosed monkey genes (ASM169854v1)
- Bolivian squirrel monkey genes (SaiBol1.0)
- Bonobo genes (panpan1.1)
- Brazilian guinea pig genes (CavAp1.0)
- Burton's mouthbrooder genes (AstBur1.0)
- Bushbaby genes (OtoGar3)
- C.intestinalis genes (KH)
- C.savignyi genes (CSAV 2.0)
- Caenorhabditis elegans genes (WBcel235)

First choose database and dataset

Example: Step 2 (Filter)

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

Register

New Count Results URL XML Perl Help

Dataset
Human genes (GRCh38.p12)

Filters

Chromosome/scaffold: 1
Start: 78895
End: 224561

Attributes
Gene stable ID
Transcript stable 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

Coordinates

Start: 78895
End: 224561

Limit to chromosome 1

Limit to given coordinates

Example: Step 3 (Count results)

e!Ensembl BLAST/BLAT | [Blog](#) | [Login/Register](#)

Search all species...

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

Dataset: 12 / 64914 Genes
Human genes (GRCh38.p12)

Filters

Chromosome/scaffold: 1
Start: 78895
End: 224561

Attributes

Gene stable ID
Transcript stable 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

Coordinates

Start:
End:

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 / 64914 Genes
Human genes (GRCh38.p12)

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 **Variant (Somatic)**
 Homologues **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)
- Transcript length (including UTRs and CDS)
- Transcript support level (TSL)
- GENCODE basic annotation
- APPRIS annotation
- 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

Example: Step 5 (get results)

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

Search all species...

New **Count** **Results** **URL** **XML** **Perl** **Help**

Dataset 12 / 64914 Genes
Human genes (GRCh38.p12)

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	ENST00000453576
ENSG00000238009	ENST00000610542
ENSG00000239945	ENST00000495576
ENSG00000233750	ENST00000442987
ENSG00000268903	ENST00000494149
ENSG00000269981	ENST00000595919
ENSG00000239906	ENST00000493797

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 **v95**, for Human GRCh38.p12
 - 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)
 - Select Gene stable ID, Transcript stable ID, Gene Name
- 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

- 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 / NGS data analysis 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 “Id” to “**Gene stable ID**”. Save the change.
- 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)
- 3. Upload the file siMitfvssiLuc.up.txt and the annotation file (mart_export.txt.gz) you obtained from Ensembl/BioMart to GalaxEast 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 then mart_export.txt) based on the “Gene stable IDs” field.
 - Gene stable 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 lincRNAs in the upregulated genes? Use the tool “Filter data on any column using simple expressions” to search for “lincRNA” (<- this exact case) in the dataset siMitfvssiLuc.up.annot.txt.
 - Hint 1: Search “lincRNA” in the column containing Gene types
 - Hint 2: c3 refers to column 3 of a dataset.
- 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. 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: Gene stable ID, Chromosome/scaffold name, Gene Start (bp), Gene End (bp), strand and Gene Name