

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



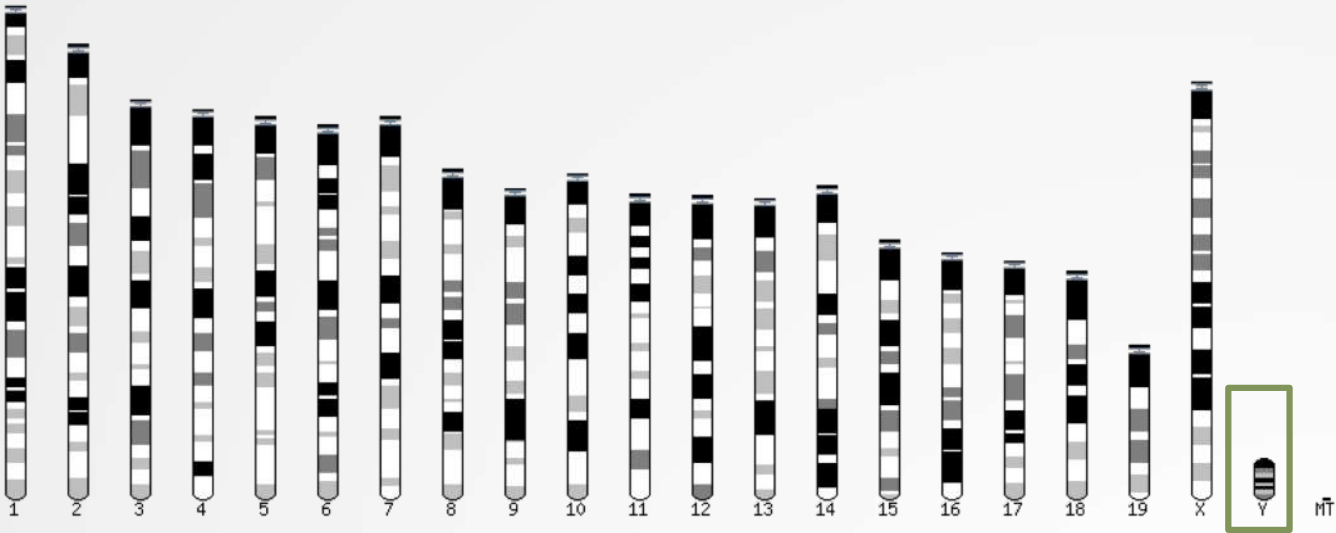
Genome builds

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)

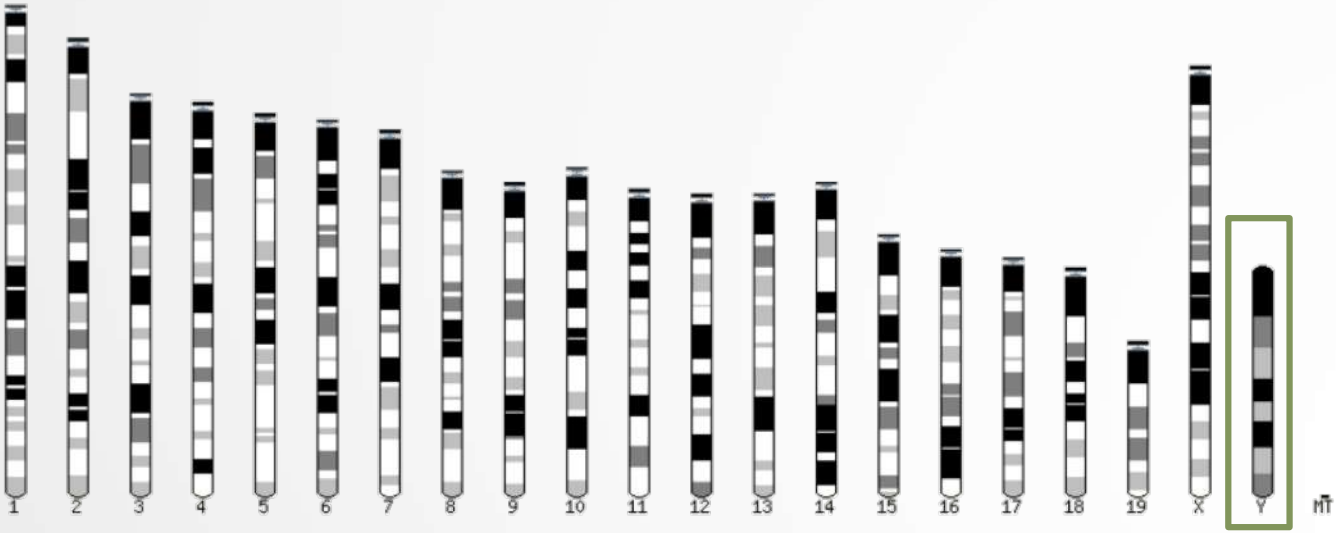
Source: <https://genome.ucsc.edu/FAQ/FAQreleases.html>

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

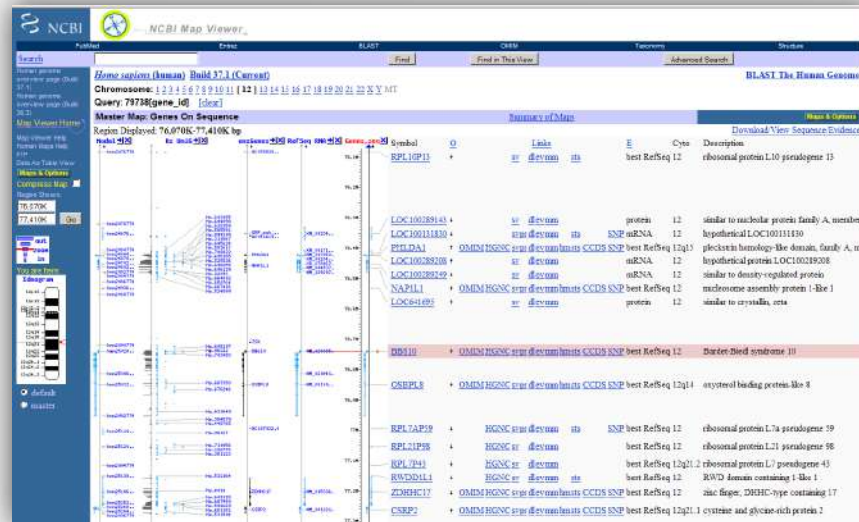
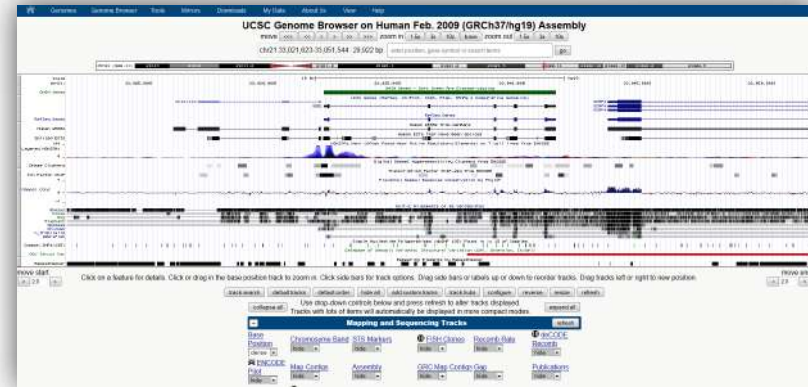
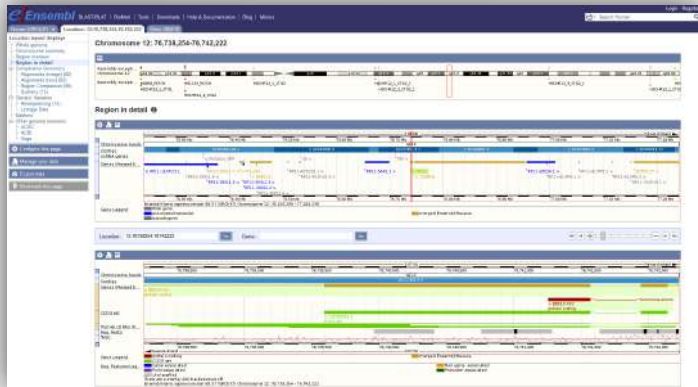
Genome Browsers

- 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...

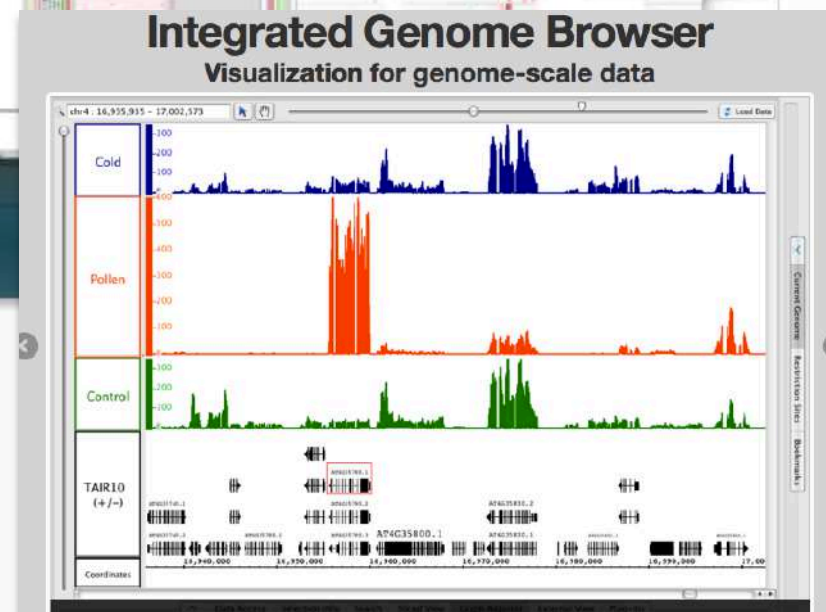
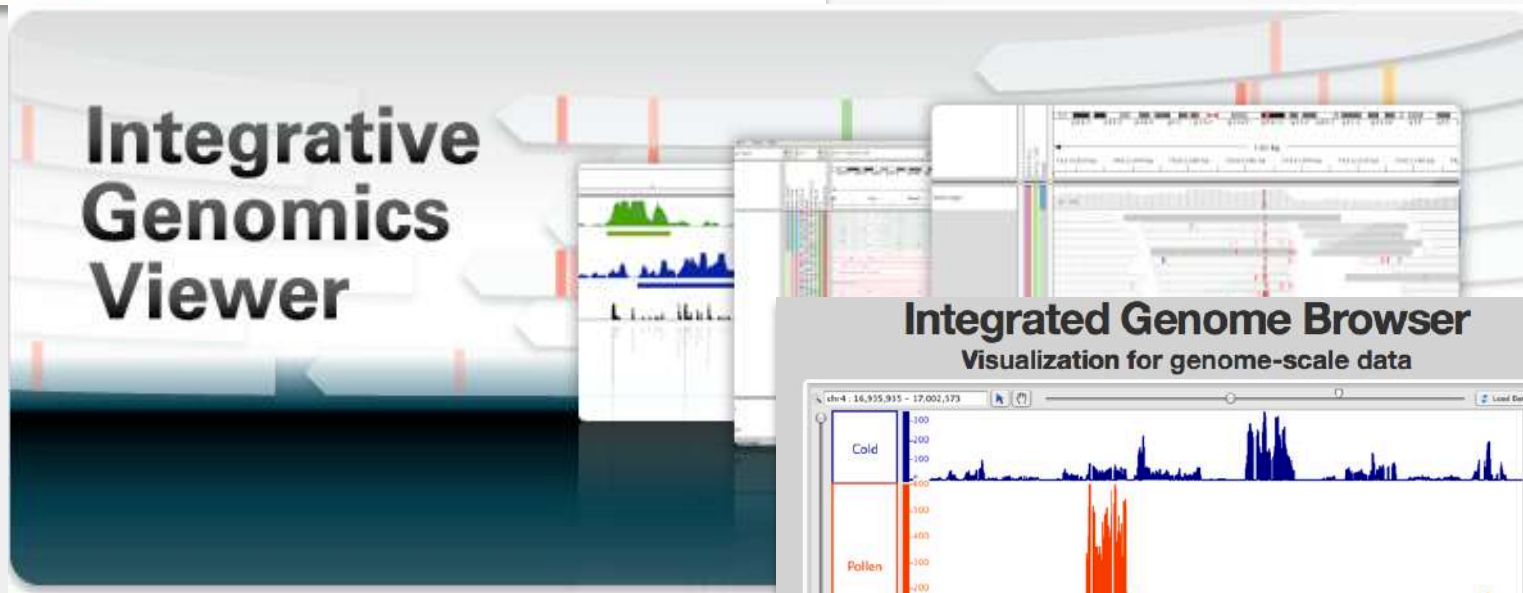
EBI - Ensembl

UCSC - Genome Browser



NCBI - Map Viewer

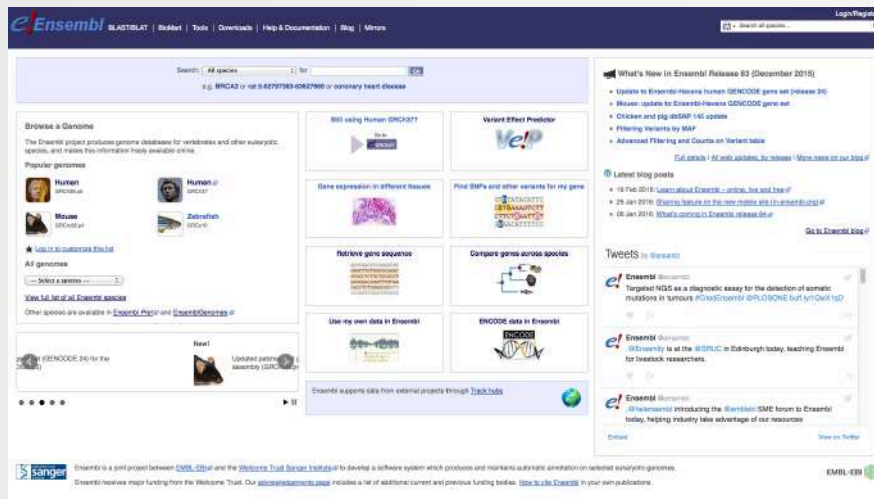
And Genome browsers...





Getting access to genomic
data:
ENSEMBL/BIOmart

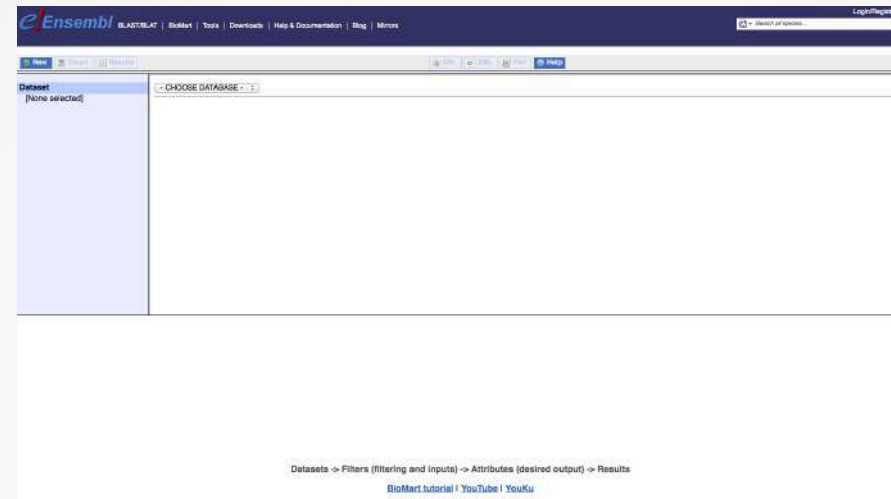
Access Ensembl's data




Web site



-  User friendly
-  Straightforward
-  Only one request at once

Mining tool: BioMart

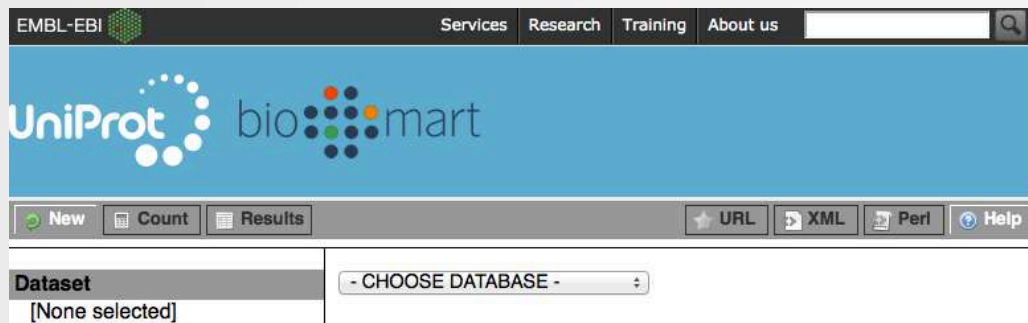


-  Get answer to complex query
-  Very fast
-  Need training

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

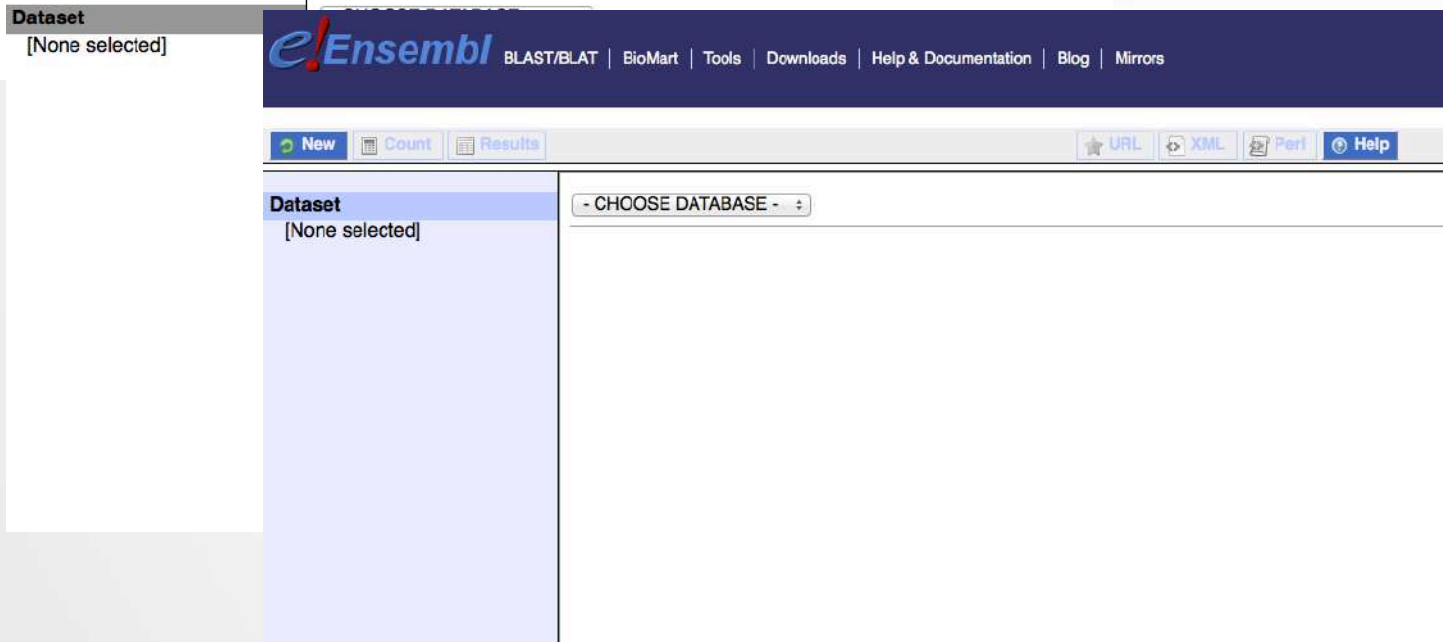
[None selected]



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](#)

BioMart/Ensembl

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

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

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. The main content area has a 'Dataset' section on the left with '[None selected]'. A dropdown menu is open, showing a list of datasets. The first two items are highlighted in blue: 'Ensembl Genes 95' and 'Chicken genes (GRCg6a)'. A green box highlights these two items, and a callout box with a green border contains the text 'First choose database and dataset'.

Dataset
[None selected]

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)

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]

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](#)

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

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]

Using a previous version of Ensembl

- During this course, we are going to use a previous version of Ensembl: Ensembl v95.

The screenshot shows the Ensembl genome browser homepage. 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 is located on the right side of the navigation bar. Below the navigation bar, there are four main tool categories: Tools, BioMart, BLAST/BLAT, and Variant Effect Predictor. The BioMart section includes a search box and a 'Go' button. Below the search box, there are sections for 'All genomes' and 'Favourite genomes'. The 'All genomes' section has a dropdown menu to select a species. The 'Favourite genomes' section lists Human (GRCh38.p13), Mouse (GRCm38.p5), and Zebrafish (GRCz11). At the bottom of the page, there are several service tiles: 'Compare genes across species', 'Find SNPs and other variants for my gene', 'Gene expression in different tissues', 'Retrieve gene sequence', 'Find a Data Display', and 'Use my own data in Ensembl'. The footer contains the EMBL-EBI logo, a paragraph of text about Ensembl's mission and funding, and the Elixir Core Data Resource logo. A permanent link to the archive site is also provided.

On ensembl.org/index.html, click on View in archive site



Using a previous version of Ensembl

The screenshot shows the Ensembl website 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 on the right side of the top bar. Below the navigation bar, there is a section titled 'View in archive site' with a search box and a list of links: Help topics, Frequently Asked Questions, Video Tutorials, Glossary, and Contact HelpDesk. The main content area displays a list of archives available for the page, starting with 'Ensembl GRCh37: Full Feb 2014 archive with BLAST, VEP and BioMart' and ending with 'Ensembl 54: May 2009 (NCBI 36) - patched/updated gene set Oct 2008'. A green arrow points to the entry 'Ensembl 95: Jan 2019 (GRCh38.p12)'. Below the list, there is a link for 'More information about the Ensembl archives'. The footer of the page includes the URL 'ian2019.archive.ensembl.org/index.html' and a 'Permanent link - View in archive site' link.

The following archives are available for this page:

- Ensembl GRCh37: Full Feb 2014 archive with BLAST, VEP and BioMart
- Ensembl 97: Jul 2019 (GRCh38.p12) - patched/updated gene set Mar 2019
- Ensembl 96: Apr 2019 (GRCh38.p12) - patched/updated gene set Nov 2018
- Ensembl 95: Jan 2019 (GRCh38.p12)**
- Ensembl 94: Oct 2018 (GRCh38.p12) - patched/updated gene set Jul 2018
- Ensembl 93: Jul 2018 (GRCh38.p12)
- Ensembl 92: Apr 2018 (GRCh38.p12) - patched/updated gene set Jan 2018
- Ensembl 91: Dec 2017 (GRCh38.p10)
- Ensembl 90: Aug 2017 (GRCh38.p10) - patched/updated gene set Jun 2017
- Ensembl 89: May 2017 (GRCh38.p10) - patched/updated gene set Jan 2017
- Ensembl 88: Mar 2017 (GRCh38.p10)
- Ensembl 87: Dec 2016 (GRCh38.p7)
- Ensembl 86: Oct 2016 (GRCh38.p7)
- Ensembl 85: Jul 2016 (GRCh38.p7) - patched/updated gene set Jun 2016
- Ensembl 84: Mar 2016 (GRCh38.p5)
- Ensembl 83: Dec 2015 (GRCh38.p5) - patched/updated gene set Oct 2015
- Ensembl 82: Sep 2015 (GRCh38.p3)
- Ensembl 81: Jul 2015 (GRCh38.p3) - patched/updated gene set Jun 2015
- Ensembl 80: May 2015 (GRCh38.p2) - patched/updated gene set Jan 2015
- Ensembl 79: Mar 2015 (GRCh38.p2)
- Ensembl 78: Dec 2014 (GRCh38)
- Ensembl 77: Oct 2014 (GRCh38) - patched/updated gene set Aug 2014
- Ensembl 75: Feb 2014 (GRCh37.p13) - patched/updated gene set Sep 2013
- Ensembl 67: May 2012 (GRCh37.p7) - patched/updated gene set Feb 2012
- Ensembl 54: May 2009 (NCBI 36) - patched/updated gene set Oct 2008

More information about the Ensembl archives

Select Ensembl 95: Jan 2019

ian2019.archive.ensembl.org/index.html

Permanent link - View in archive site

Using a previous version of Ensembl

Go to BioMart

Archive! **Ensembl** [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#) [Login/Register](#)

Tools [BioMart >](#)

[All tools](#) Export custom datasets from Ensembl with this data-mining tool

Search

All species for




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

All genomes

-- Select a species --

- [View full list of all Ensembl species](#)
- [Edit your favourites](#)

Favourite genomes

-  **Human**
GRCh38.p12
[Still using GRCh37?](#)
-  **Mouse**
GRCm38.p6
-  **Zebrafish**
GRCz11

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 Archive 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

- 06 Sep 2019: [GSoC 2019: Our students and their projects](#)
- 05 Sep 2019: [Job: Ensembl Web Back-end Developer](#)
- 02 Sep 2019: [Job: Ensembl Applications Developer](#)

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:
 - Gene names
 - transcript stable IDs
 - Exon stable IDs

Exercise 1: get annotations of a gene

- 3. Extract all coding sequences of the IDH1 gene in fasta format. Headers will contain:
 - transcript stable IDs
 - Exon stable IDs.
- 4. Retrieve GO-terms associated to the IDH1 gene. Select
 - GO Term Name
 - GO domain
 - GO Term Accession
 - Gene stable ID
 - Transcript stable ID
 - Gene Name

Exercise 1: get annotations of a gene

- 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
- Gene stable ID
- Transcript stable ID
- Gene Name

Exercise 2: get annotations for a set of genes

We want to annotate the file `siMitfvssiLuc.up.txt` you have generated using SARTools with gene annotations extracted from Ensembl/BioMart.

The file can be found in the directory `ensemble` on your computer. **Take this file.**

- 1. Use the file `siMitfvssiLuc.up.txt` to extract gene annotations for those genes. **Save the results to a compressed TSV file.** Annotation to extract are:
 - Gene stable IDs,
 - Chromosome/scaffold name,
 - Gene start,
 - Gene end,
 - strand,
 - Gene name,
 - Gene type.

To limit extraction to upregulated genes found in the `siMitfvssiLuc.up.txt` file, go to Filters (left panel)/GENE/ Input external references ID list, select Gene stable IDs in the drop down list and select the file `siMitfvssiLuc.up.txt`.

Once done you can click on Count. You should get 3663 / 64914 genes.

(!) it will only work because the first column of the table contains Ensembl gene IDs!

Don't close the Ensembl/Biomart window once done

Exercise 2: get annotations for a set of genes

- 2. 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
- 3. 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 i.e the first column in both datasets.
 - 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.
- 4. rename the generated dataset in 4. to `siMitfvssiLuc.up.annot.txt`

Exercise 2: get annotations for a set of genes

- 5. 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.
 - Hint 3: there is 1 header line
- 6. 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