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

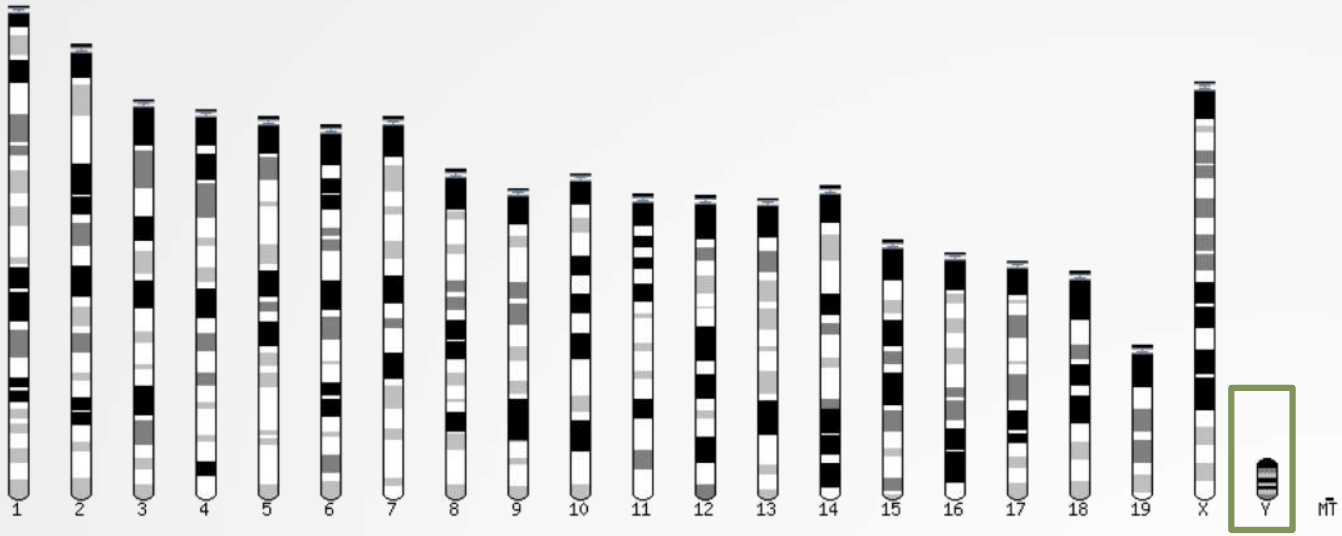
Genome builds

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

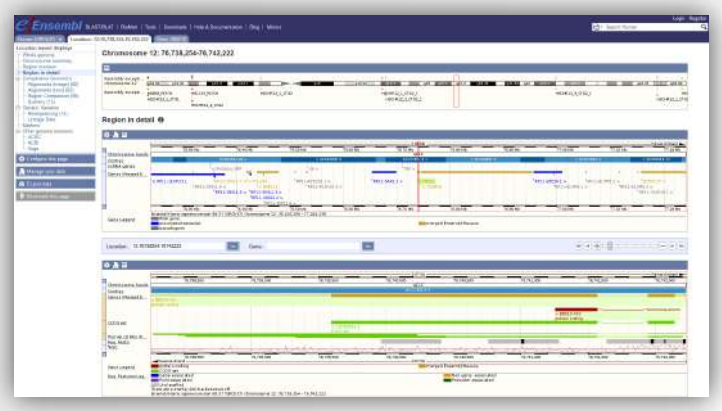
Genome browsers

Genome Browsers

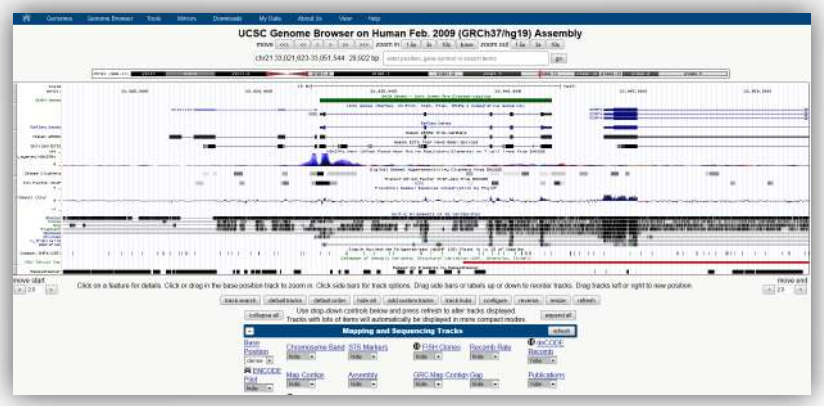
- Graphical interface to display genomic data
- Visualize and browse entire genomes with annotated data
  - Gene prediction and structure
  - Proteins,
  - Expression,
  - Regulation,
  - Variation,
  - Comparative analysis…

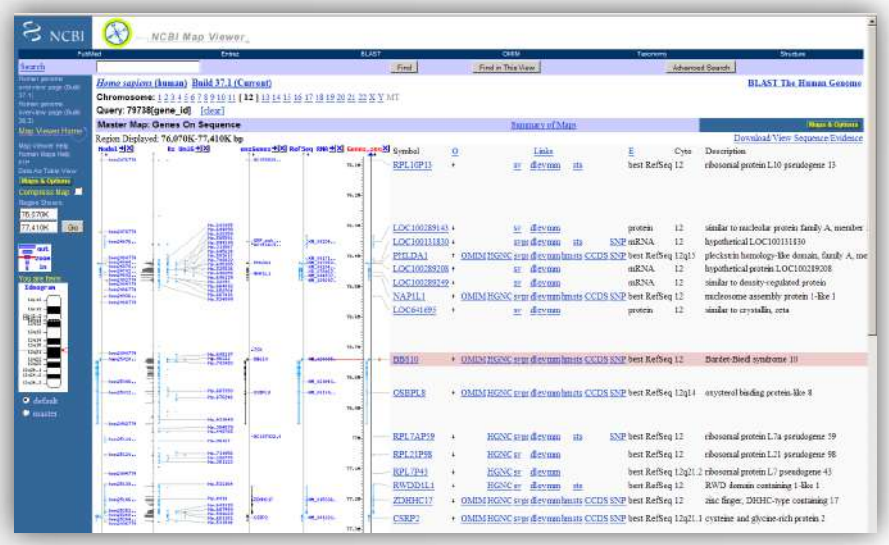
There are Genome Browsers…

EBI - Ensembl



UCSC – Genome Browser

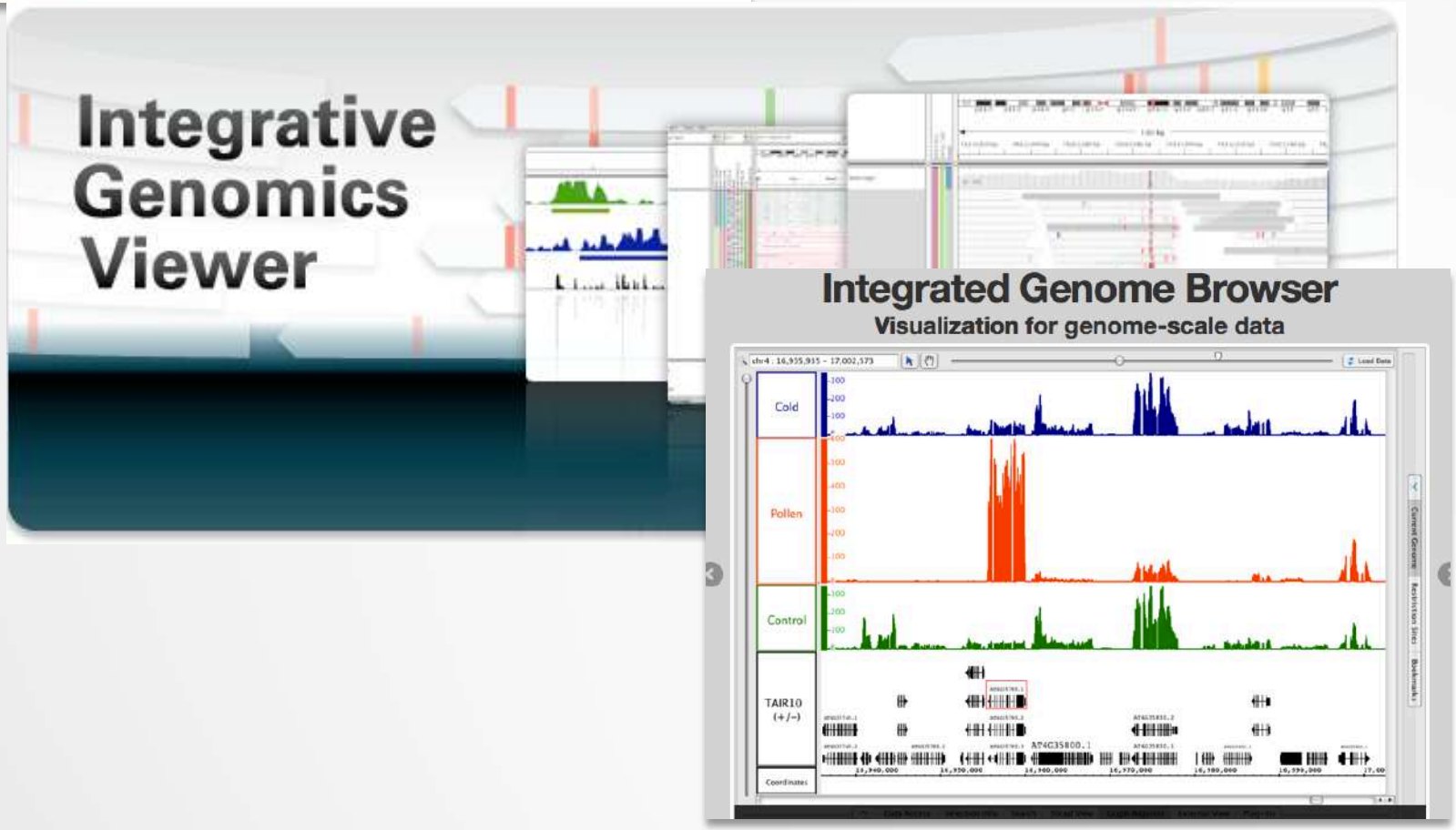




NCBI – Map Viewer

And Genome browsers…

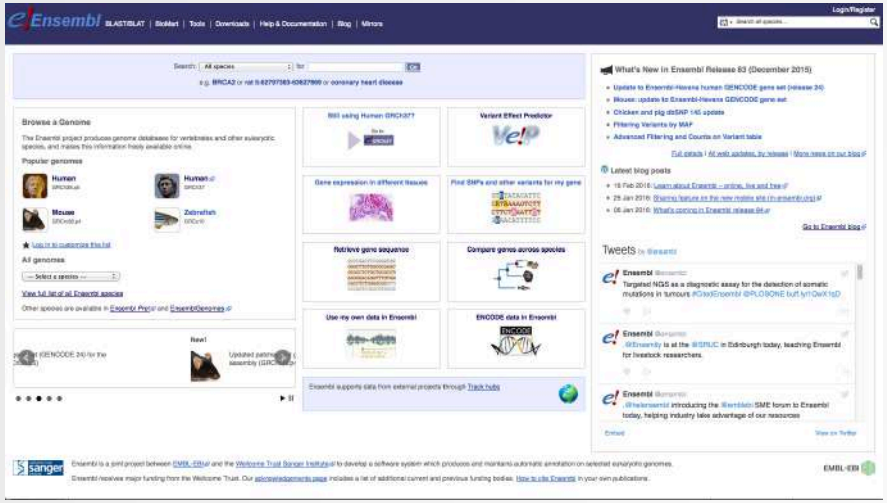
GBrowse



Getting access to genomic data: ENSEMBL/BIOmart

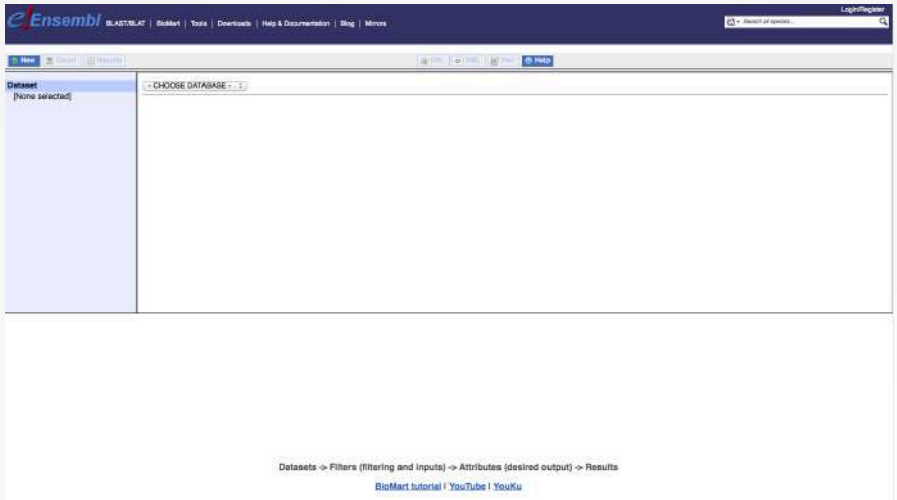
Access Ensembl's data

Web site



- User friendly
- Straightforward
- Only one request at once

Mining tool: BioMart

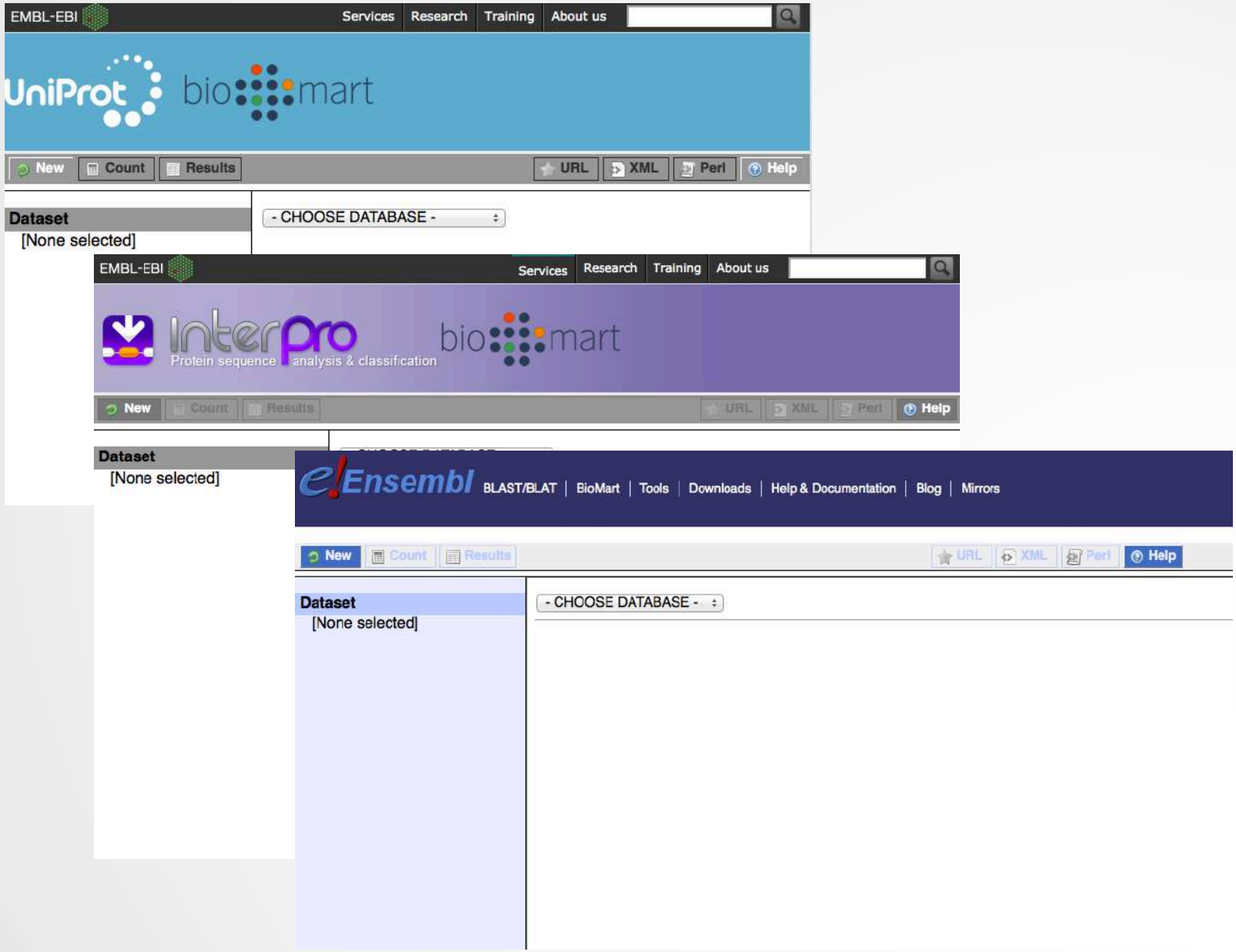


- Get answer to complex query
- Very fast
- Need training

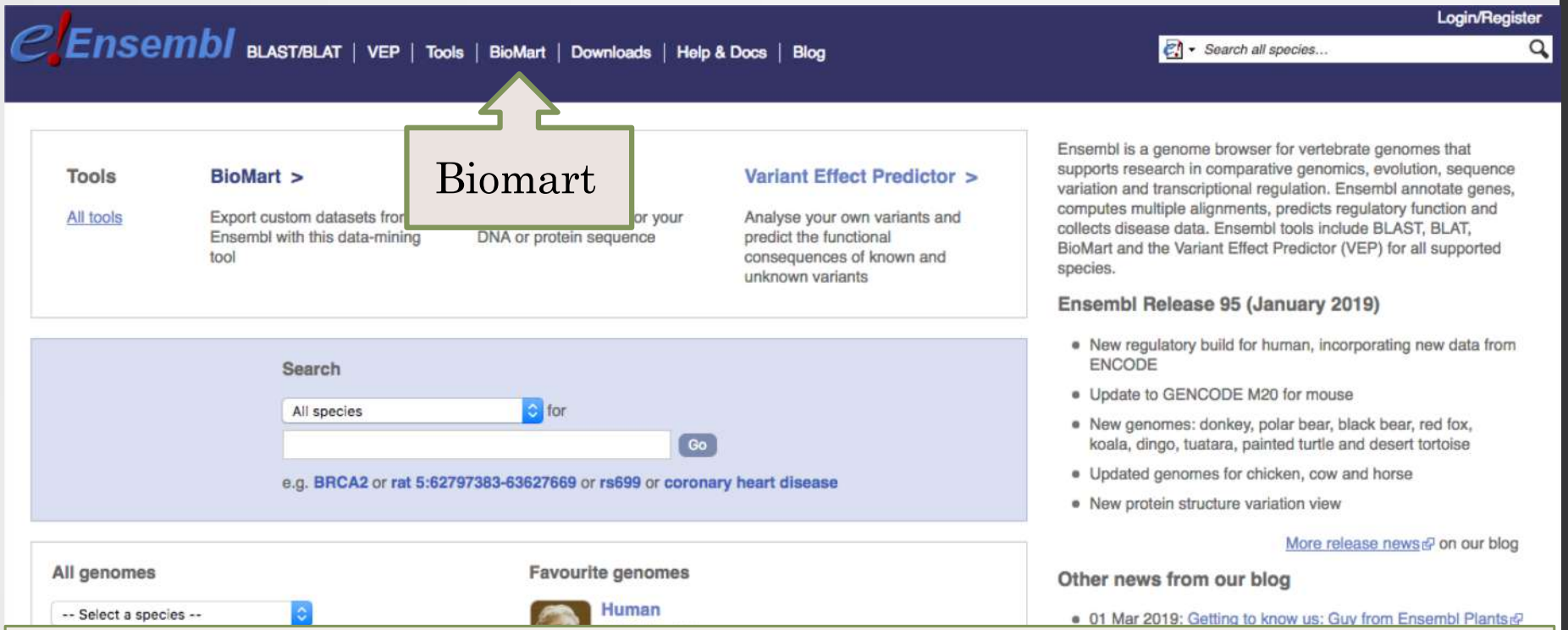
BioMart

- 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

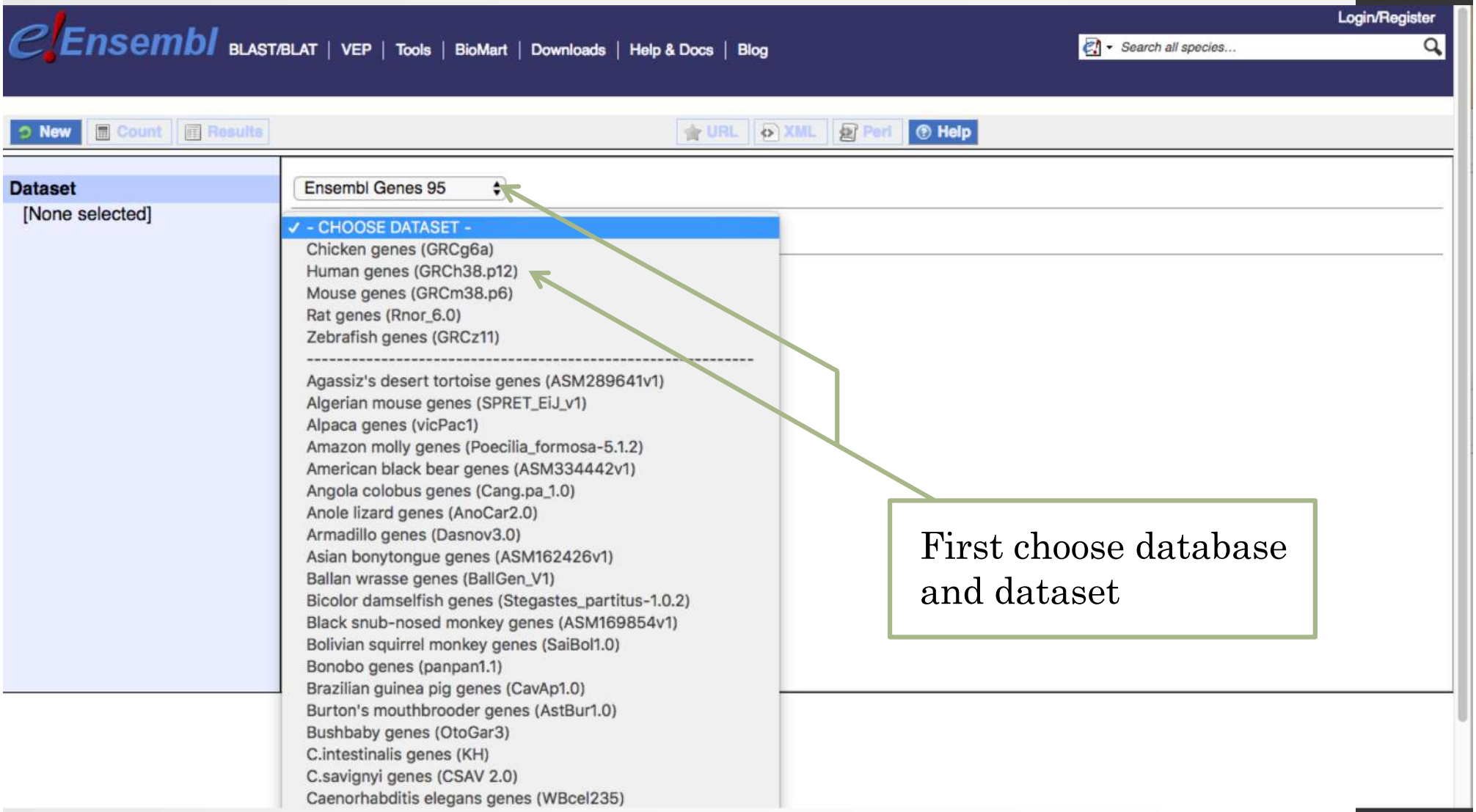


BioMart/Ensembl

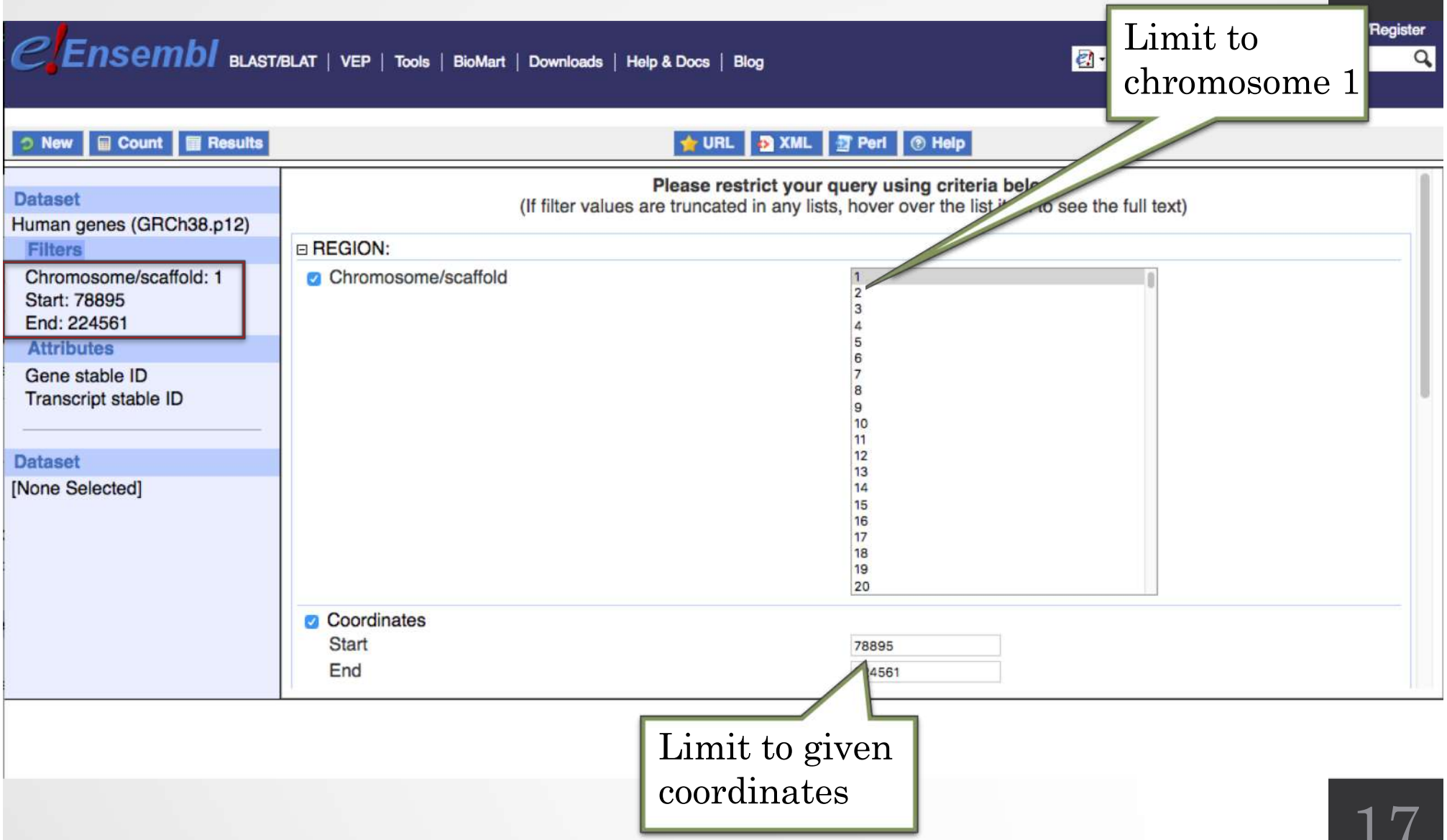


- Get access to :
  - Genomic annotation (genes, SNPs)
  - Functional annotation
  - Expression data

Example: Step 1 (Select datasets)



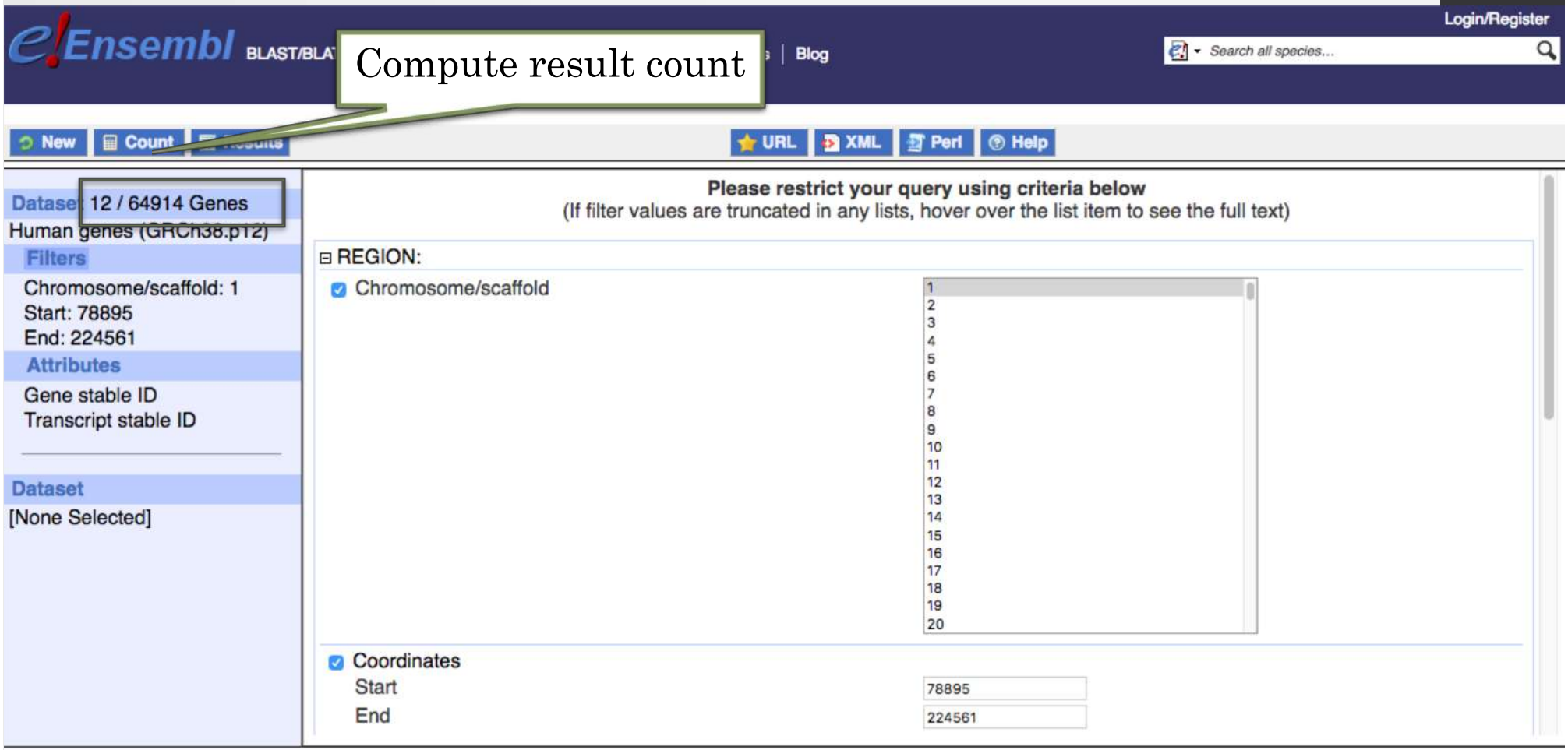
Example: Step 2 (Filter)



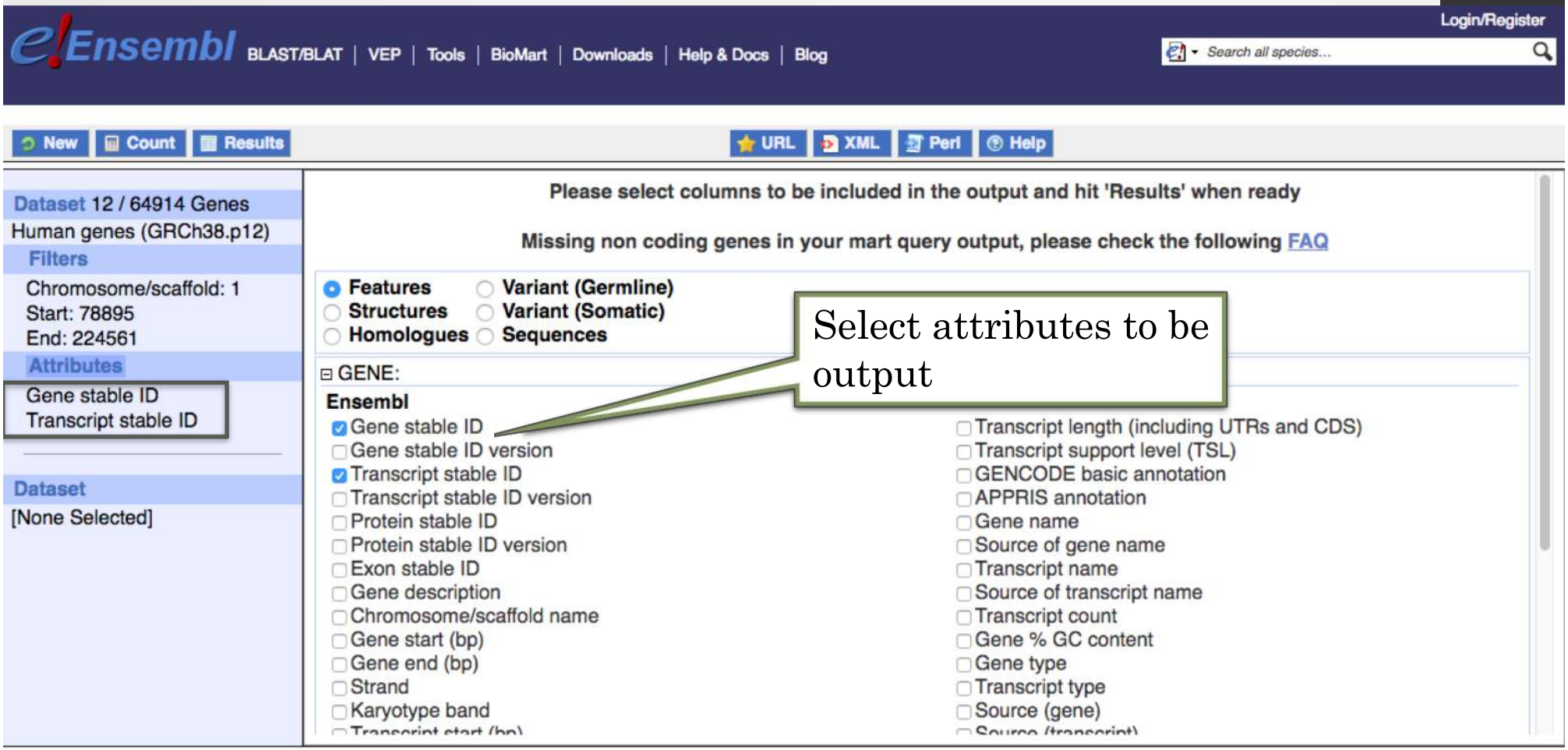
Limit to chromosome 1

Limit to given coordinates

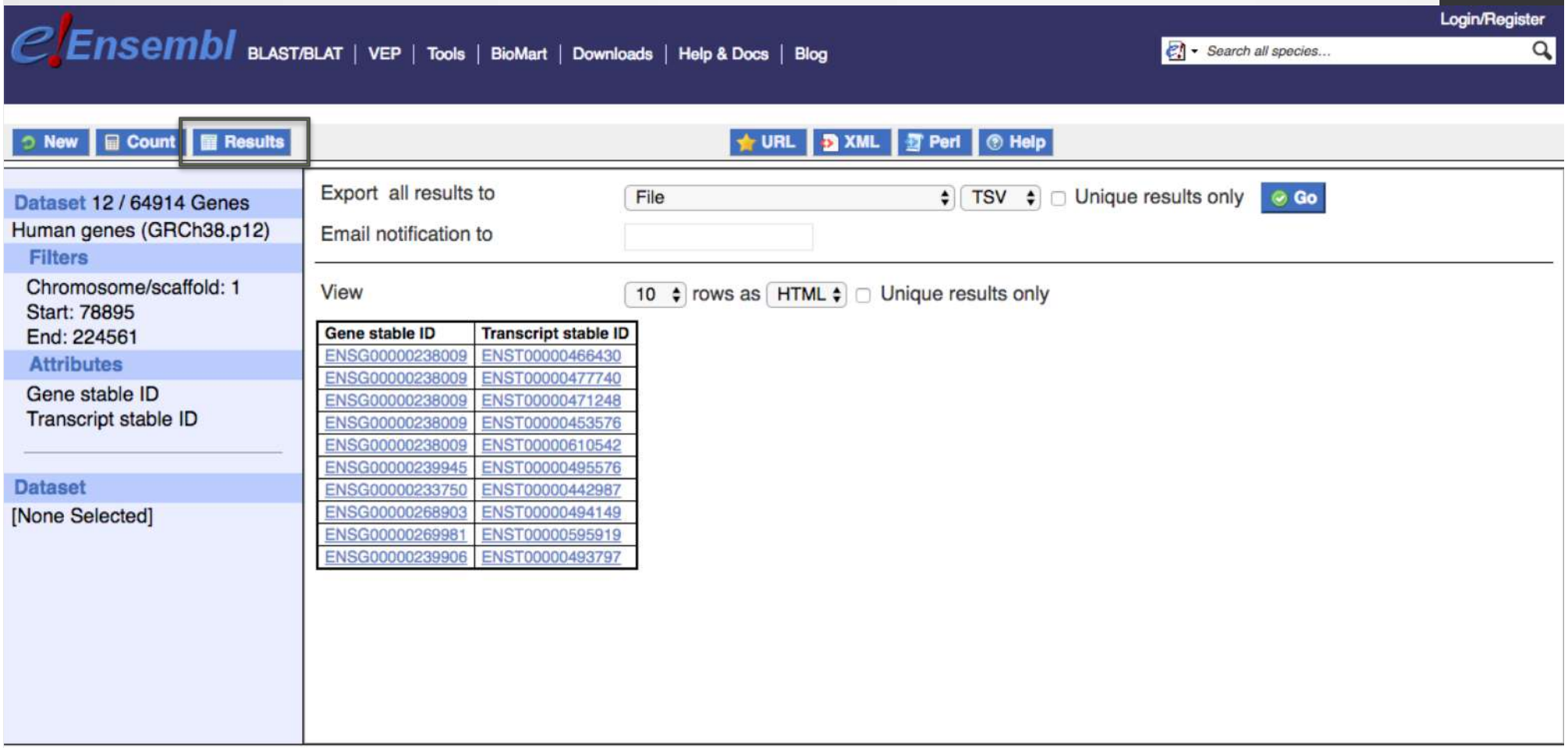
Example: Step 3 (Count results)



Example: Step 4 (Select attributes)



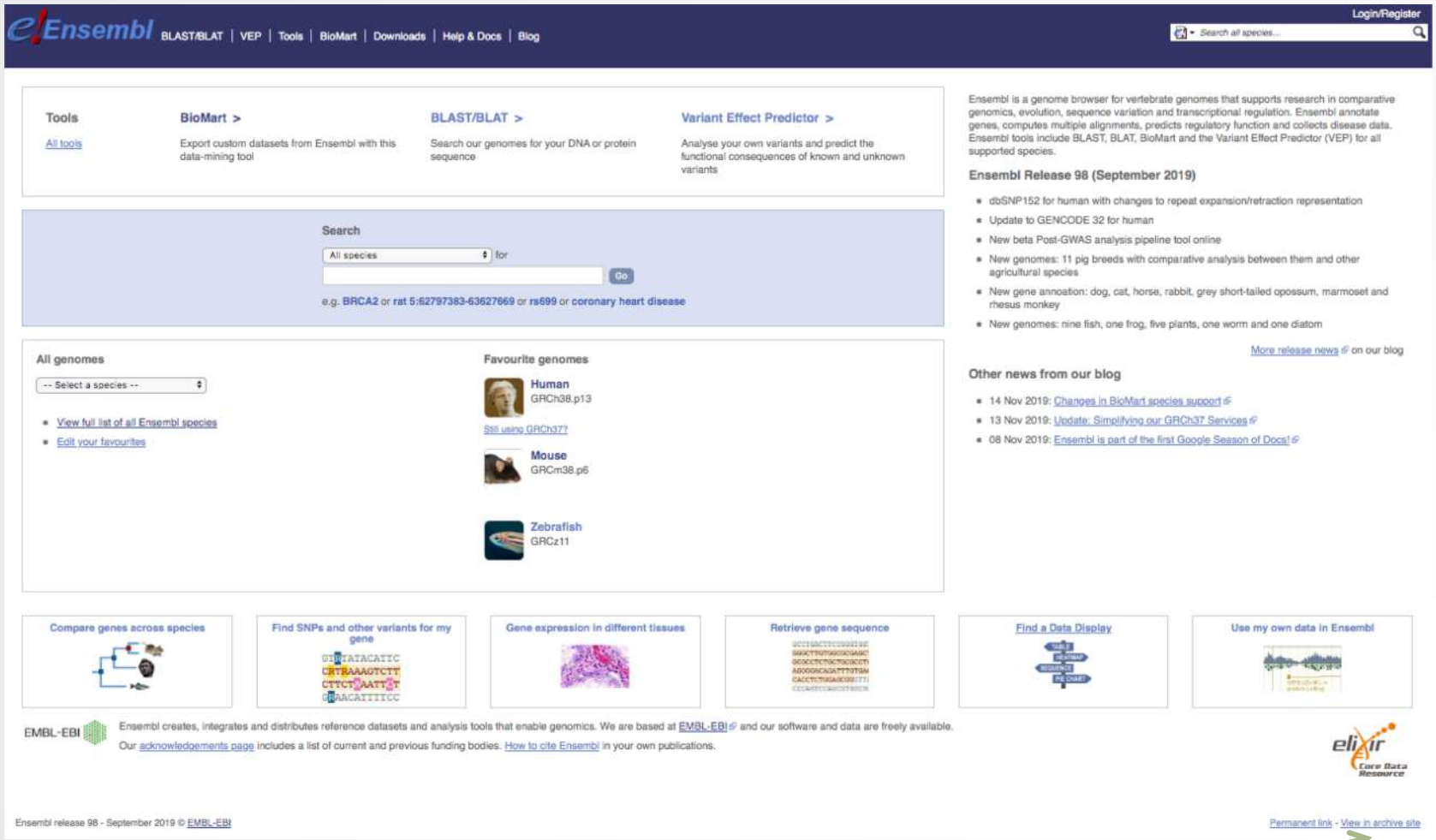
Example: Step 5 (get results)



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

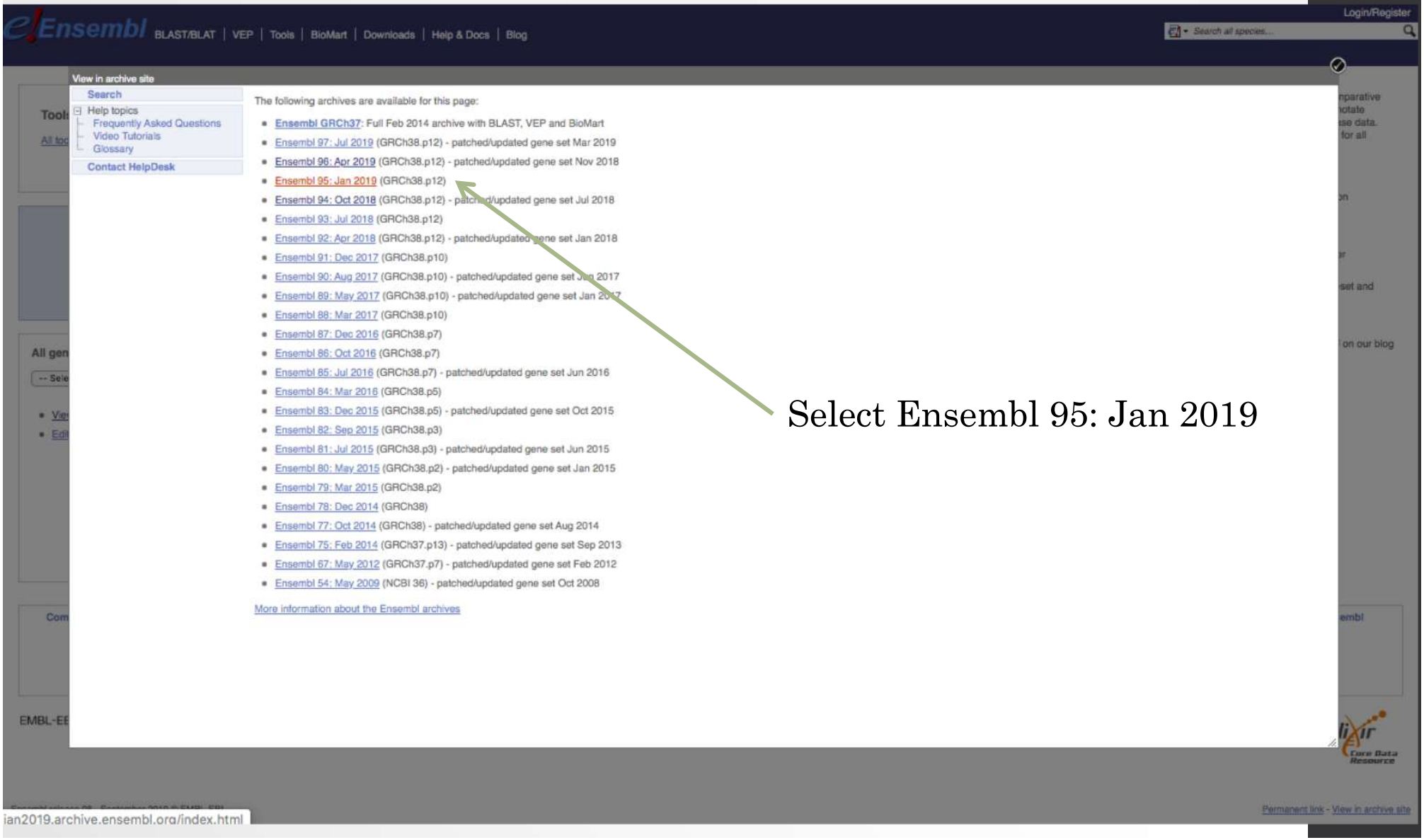
Using a previous version of Ensembl

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



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

Using a previous version of Ensembl



Using a previous version of Ensembl

Go to BioMart

Archive Ensembl BioMart | Downloads | Help & Docs | Blog

Login/Register

Search all species...

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

Search

All species for

Go

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 lncRNAs 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