Data mining with Ensembl Biomart

Stéphanie Le Gras
(slegras@igbmc.fr)

Guidelines

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

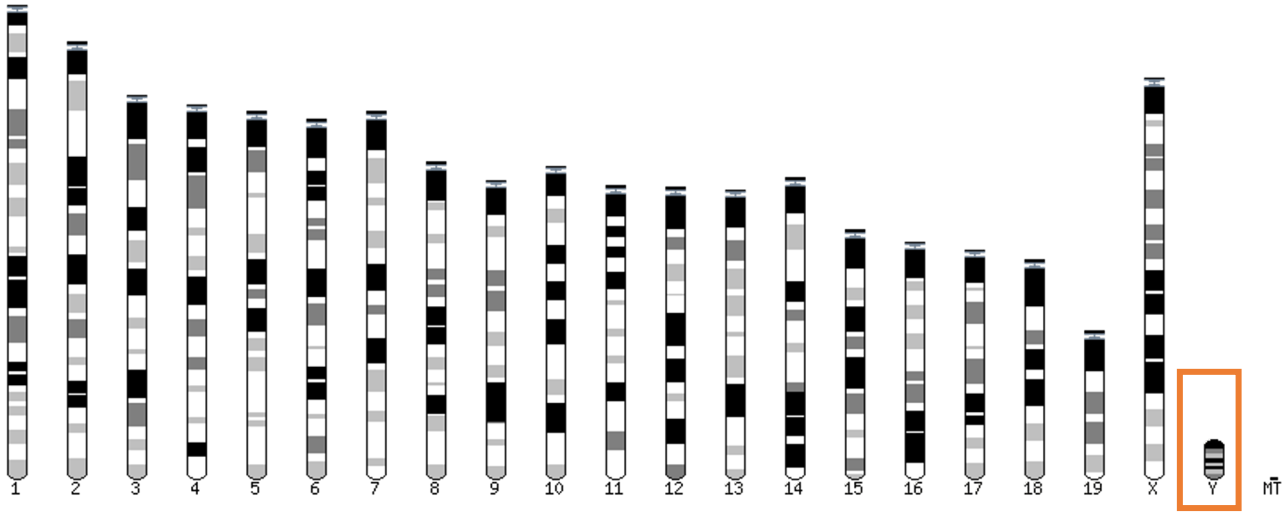
Genome builds

SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
MAMMALS				
Human	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)

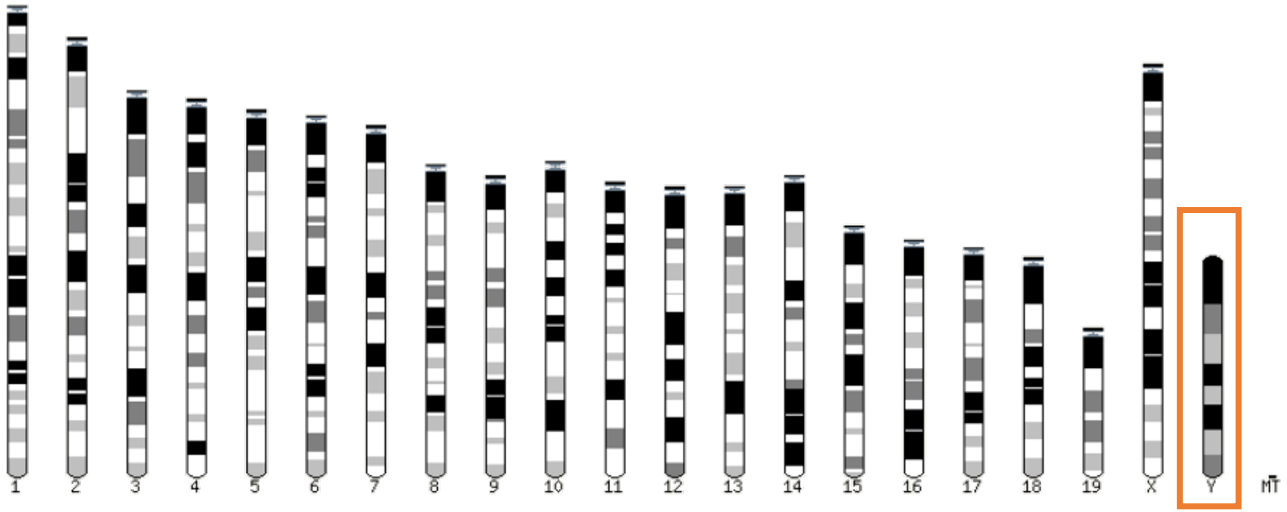
Source: https://genome.ucsc.edu/FAQ/FAQreleases.html

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

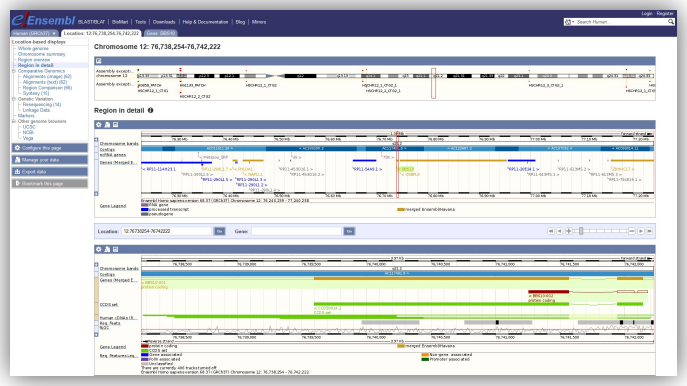
Genome browsers

Genome Browsers

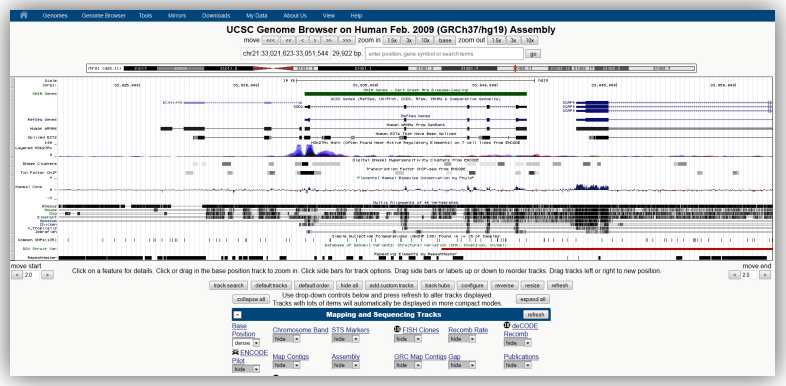
- Graphical interface to display genomic data
- Visualize and browse entire genomes with annotated data
  - Gene prediction and structure
  - Proteins,
  - Expression,
  - Regulation,
  - Variation,
  - Comparative analysis…

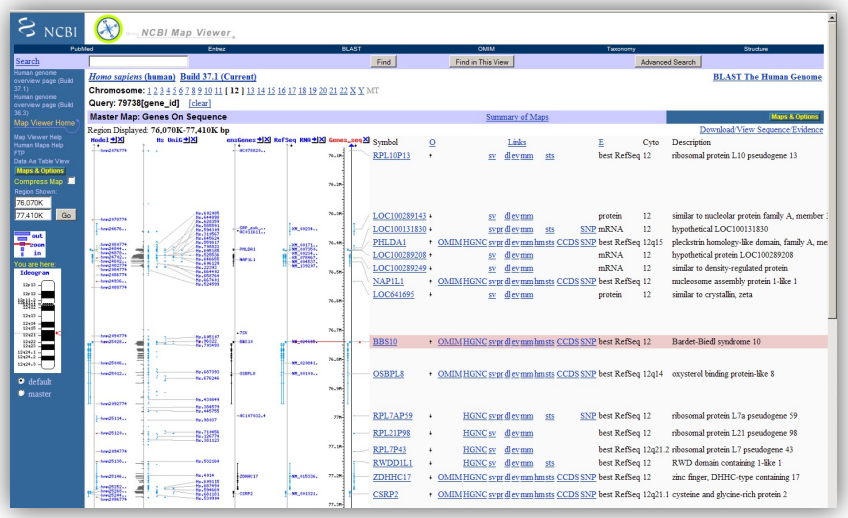
There are Genome Browsers…

EBI - Ensembl



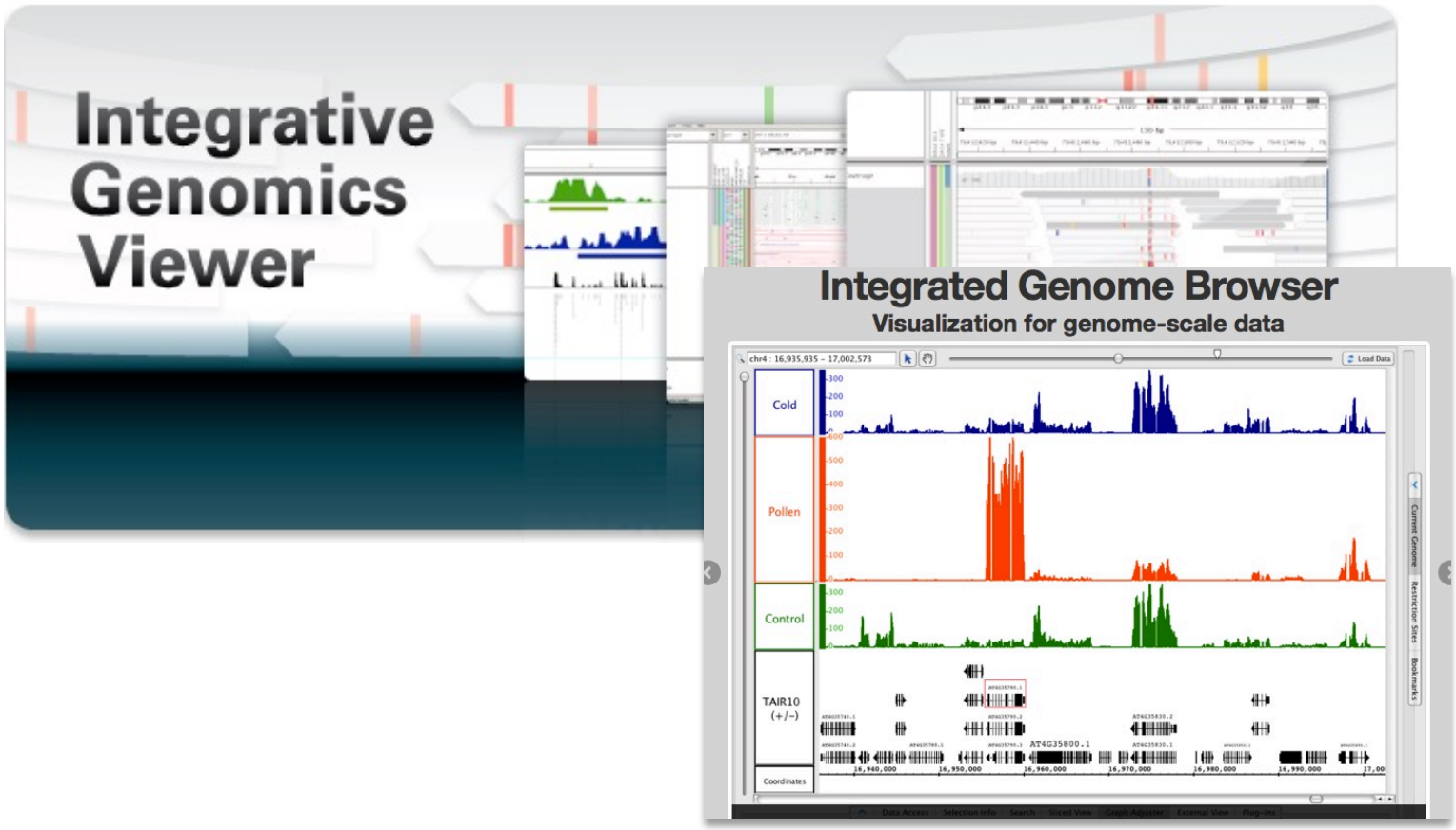
UCSC – Genome Browser





NCBI – Genome Data Viewer

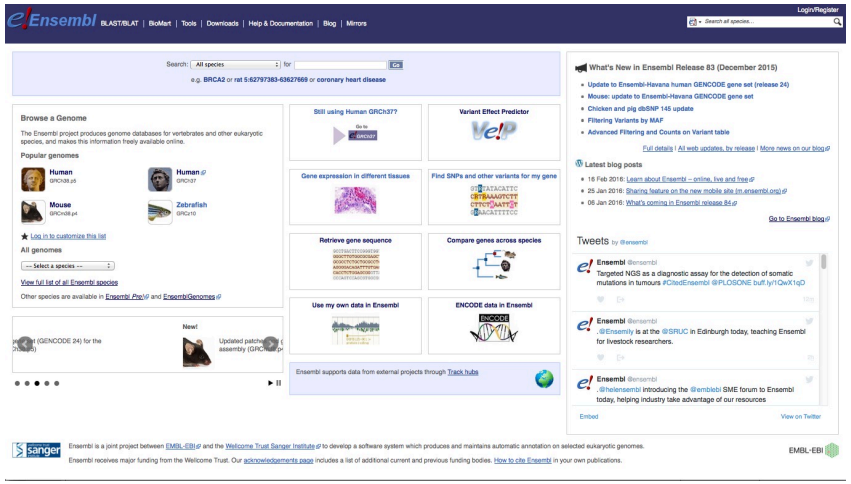
And Genome browsers…



Getting access to genomic data: ENSEMBL/BIOmart

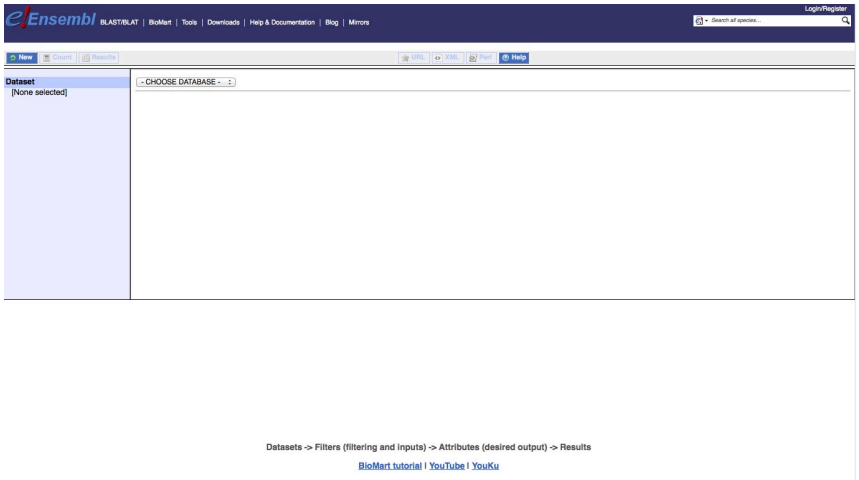
Access Ensembl's data

Web site



- User friendly
- Straightforward
- Only one request at once

Mining tool: BioMart

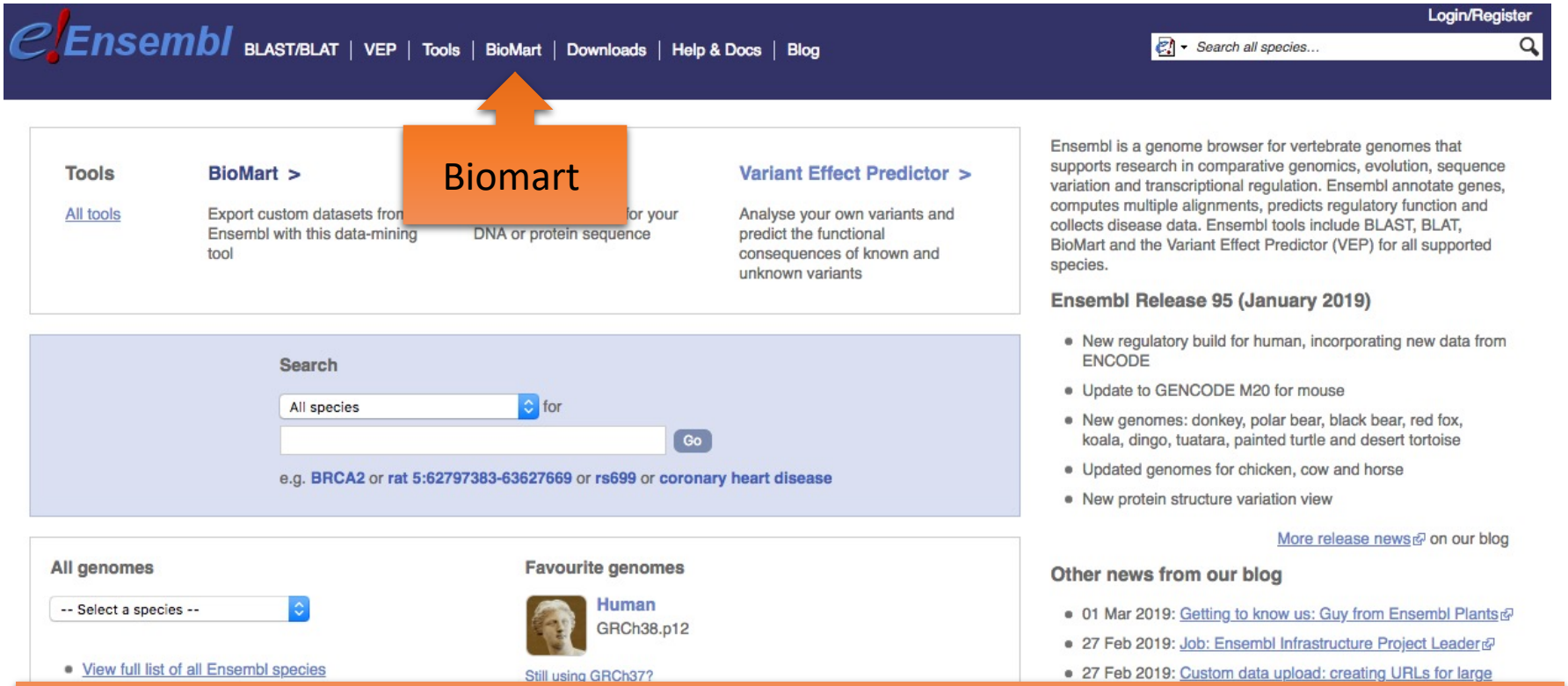


- Get answer to complex query
- Very fast
- Need training

BioMart

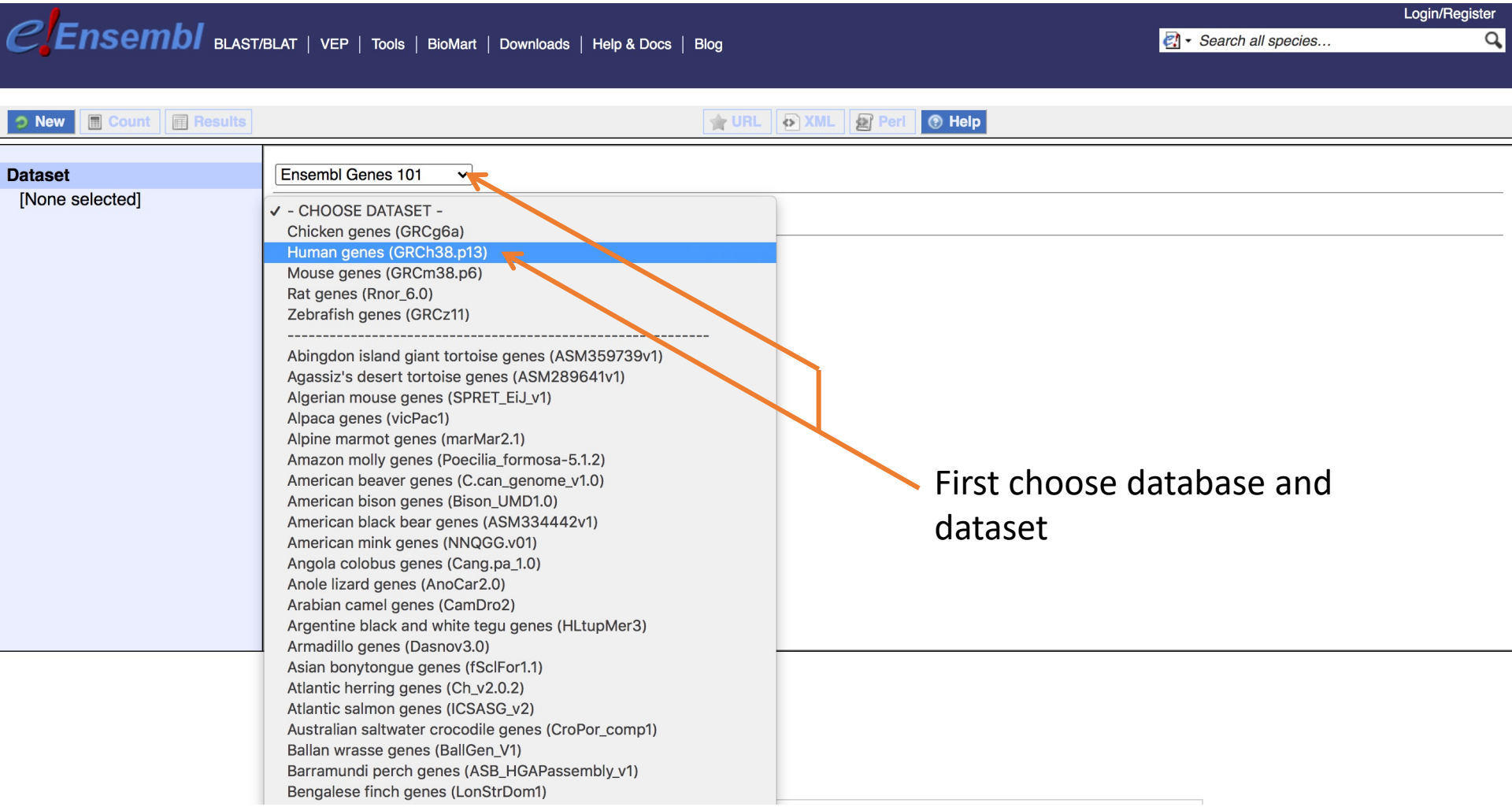
- http://www.biomart.org/
- Joint development between EBI and Cold Spring Harbor Laboratory (CSHL)
- Open source project
- BioMart can access diverse databases from a single interface
- It is search engine that can find multiple terms and put them into a table format
- No programming required!

BioMart/Ensembl

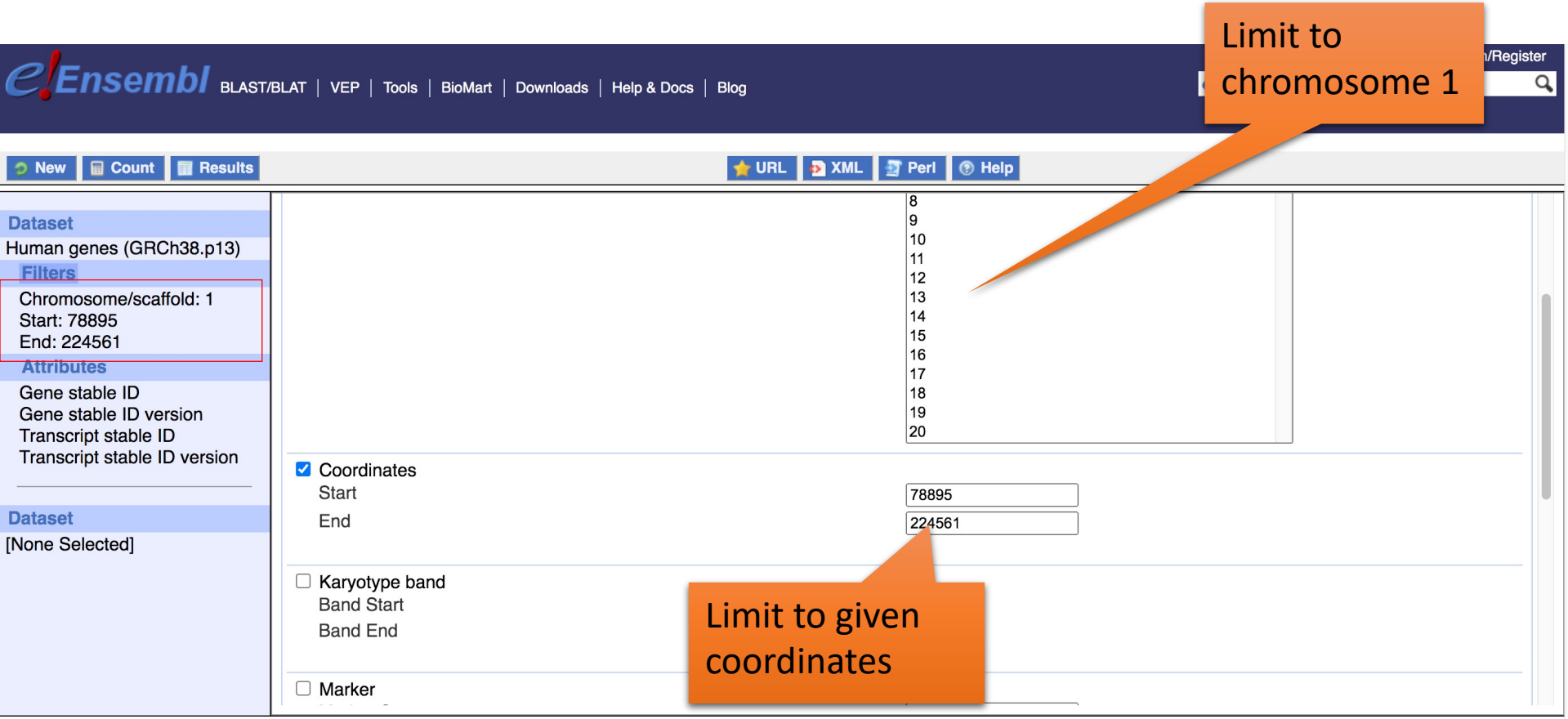


- Get access to :
  - Genomic annotation (genes, SNPs)
  - Functional annotation
  - Expression data

Example: Step 1 (Select datasets)

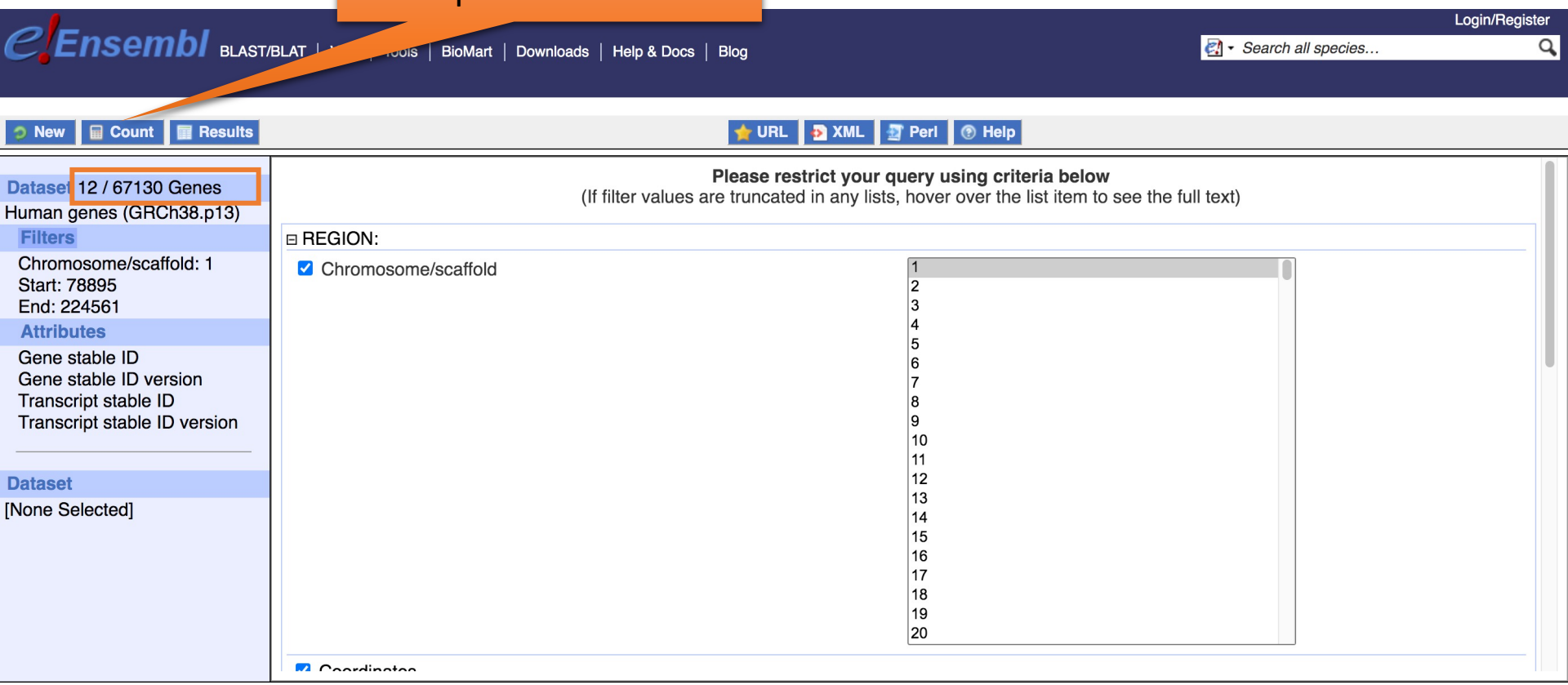


Example: Step 2 (Filter)

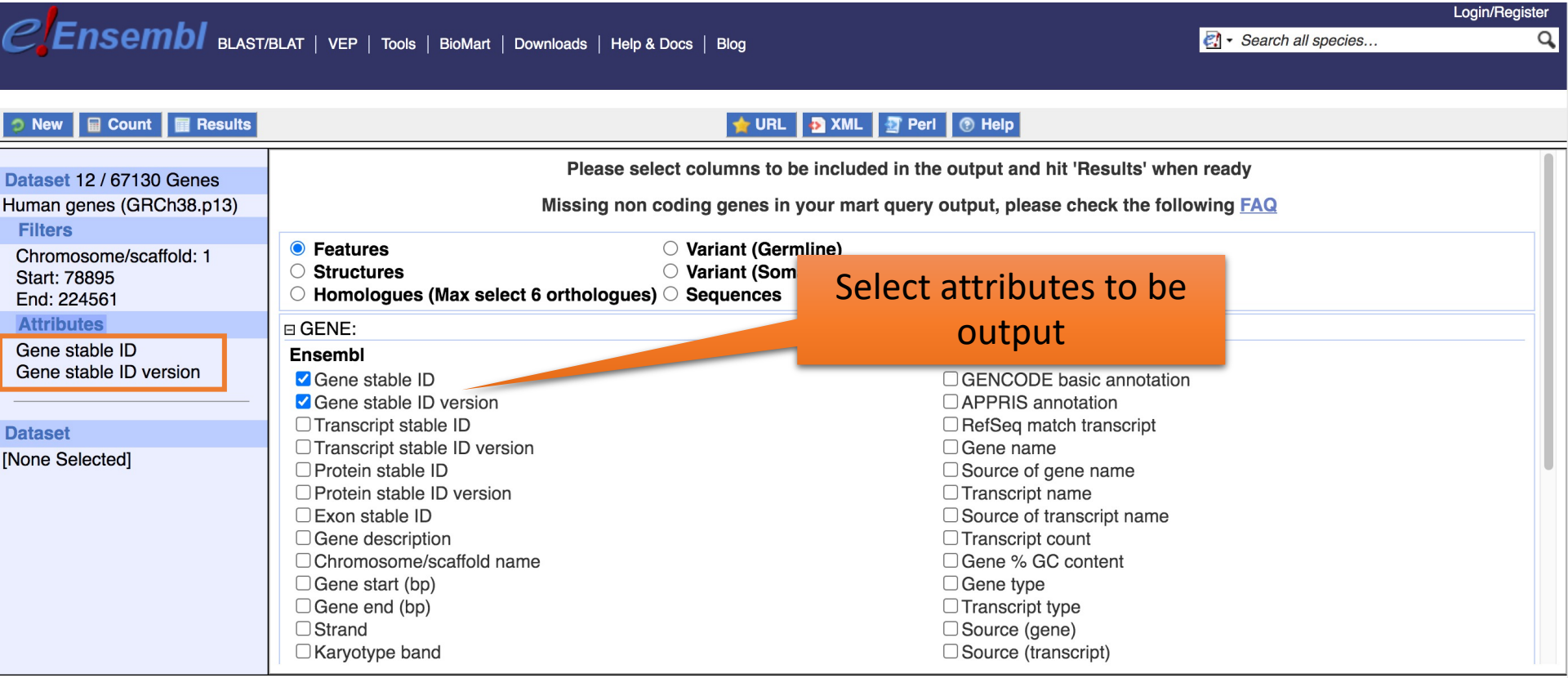


Example: Step 3 (Count results)

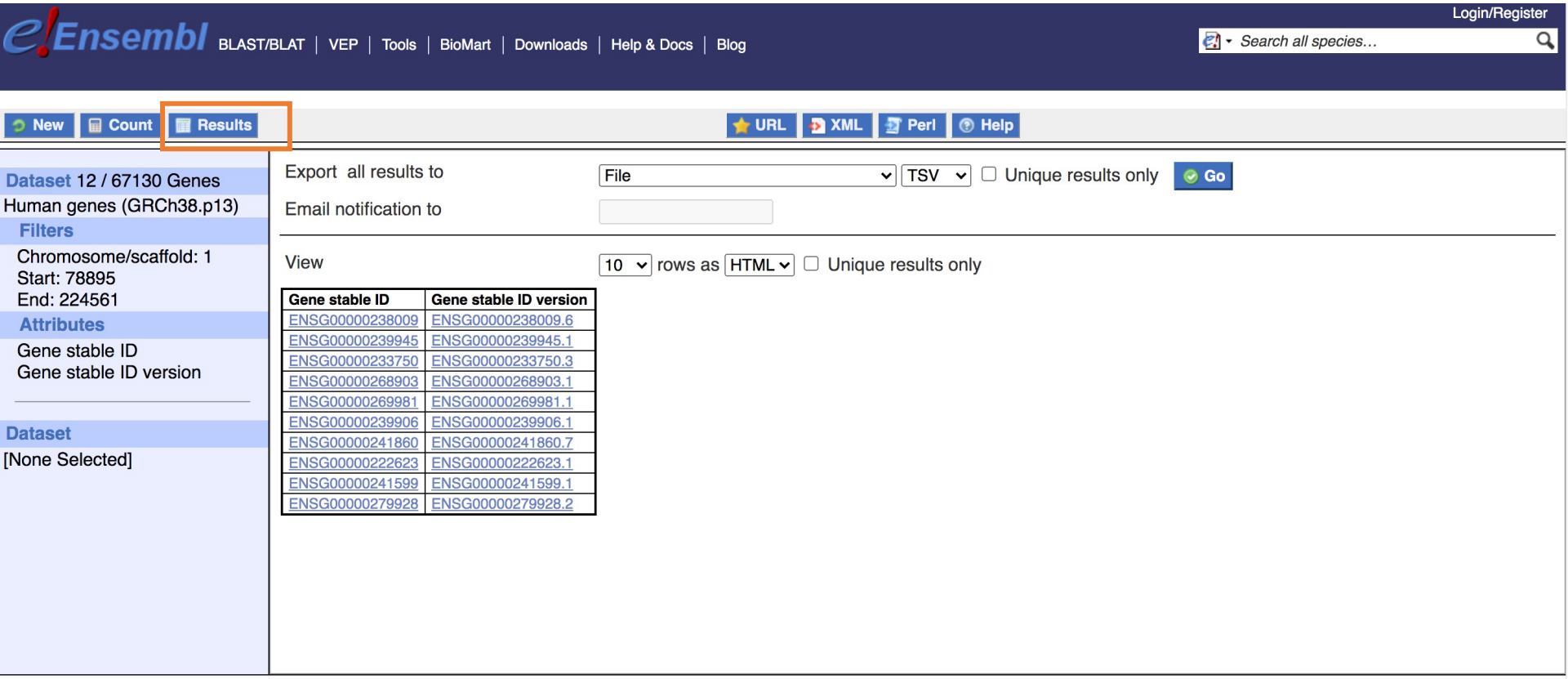
Compute result count



Example: Step 4 (Select attributes)



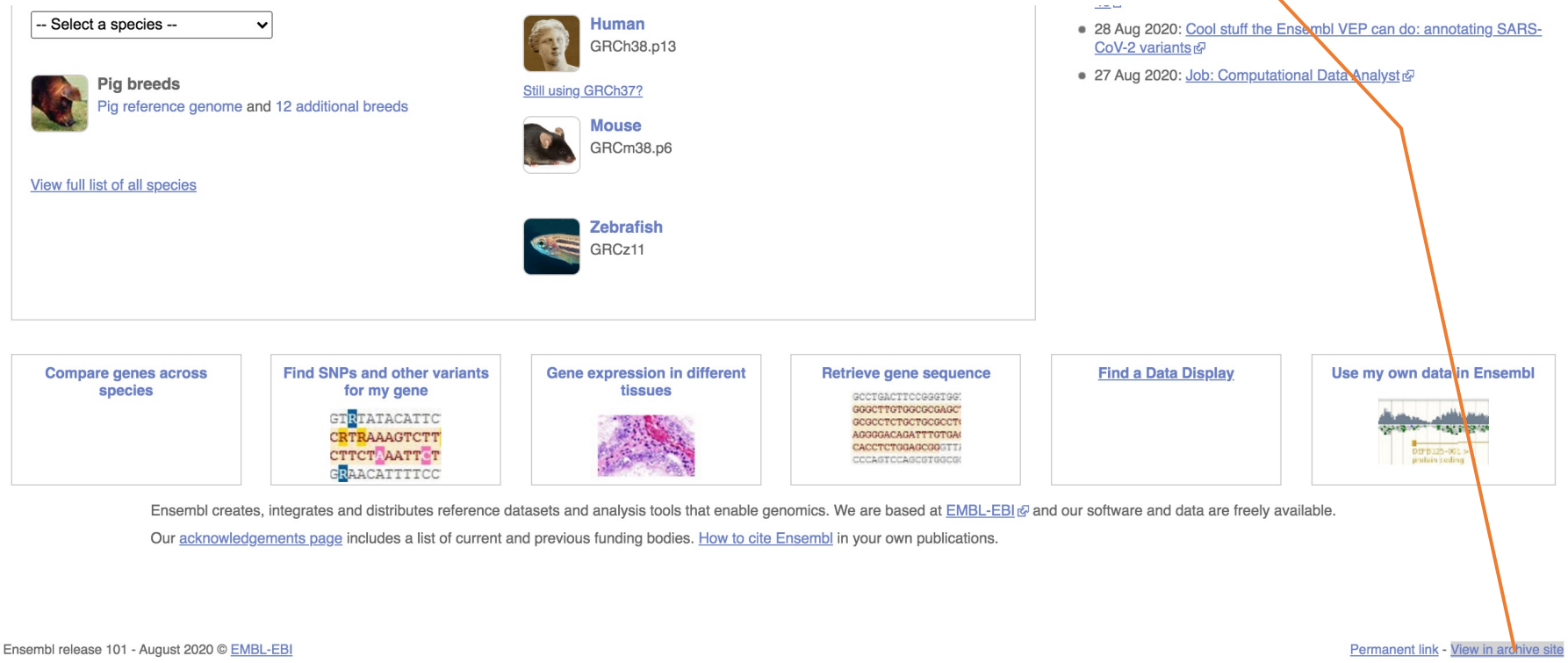
Example: Step 5 (get results)



Gene stable ID	Gene stable ID version
ENSG00000238009	ENSG00000238009.6
ENSG00000239945	ENSG00000239945.1
ENSG00000233750	ENSG00000233750.3
ENSG00000268903	ENSG00000268903.1
ENSG00000269981	ENSG00000269981.1
ENSG00000239906	ENSG00000239906.1
ENSG00000241860	ENSG00000241860.7
ENSG00000222623	ENSG00000222623.1
ENSG00000241599	ENSG00000241599.1
ENSG00000279928	ENSG00000279928.2

Start using Ensembl/BioMart

- We are going to use Ensembl/Biomart for **Ensembl v95**.
- On the main page of Ensembl website, click on View in archive site



Start using Ensembl/BioMart

- Choose Ensembl 95: Jan 2019 (GRCh38.p12)

View in archive site

Search

Help topics
- Frequently Asked Questions
- Video Tutorials
- Glossary

Contact HelpDesk

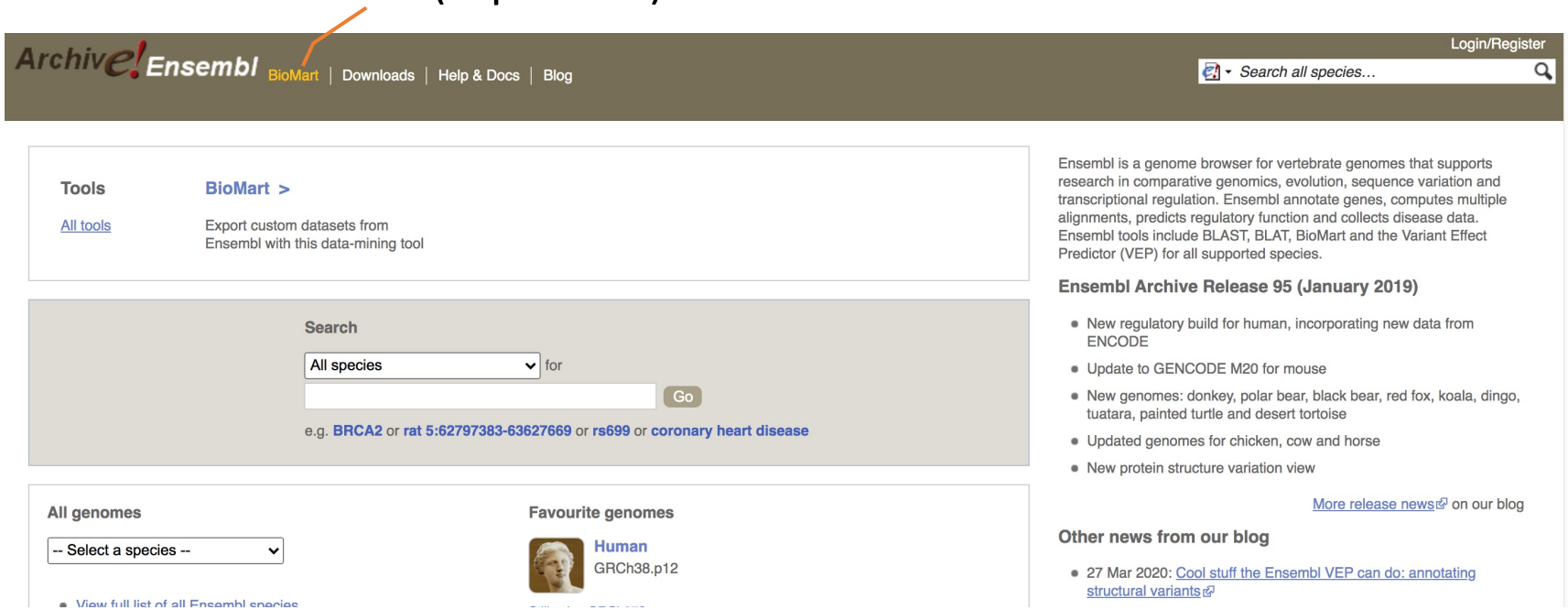
The following archives are available for this page:

- **Ensembl GRCh37**: Full Feb 2014 archive with BLAST, VEP and BioMart
- Ensembl 103: Feb 2021 (GRCh38.p13) - patched/updated gene set Aug 2020
- Ensembl 102: Nov 2020 (GRCh38.p13) - patched/updated gene set Sep 2020
- Ensembl 101: Aug 2020 (GRCh38.p13) - patched/updated gene set Mar 2020
- Ensembl 100: Apr 2020 (GRCh38.p13) - patched/updated gene set Jun 2019
- Ensembl 99: Jan 2020 (GRCh38.p13) - patched/updated gene set Aug 2019
- Ensembl 98: Sep 2019 (GRCh38.p13) - patched/updated gene set Jun 2019
- 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 80: May 2015 (GRCh38.p2) - patched/updated gene set Jan 2015
- 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

Start using Ensembl/BioMart

- Click on Biomart (top menu)



- CHOOSE DATABASE : select “Ensembl Genes 95”
- CHOOSE DATASET : select “Human genes (GRCh38.p12)”

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 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
  - 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

- **Annotate the file siMitfvssiLuc.up.txt you have generated using SARTools with gene annotations you extract from Ensembl/BioMart**

Exercise 2: get annotations for a set of genes

siMitfvssiLuc.up.txt

Gene stable ID	siLuc2	siLuc...
ENSG00000018408	4685	...
ENSG00000081189	1716	...
ENSG00000106772	3063	...
ENSG00000124942	309	...
ENSG00000142871	243	...
ENSG00000143341	3760	...
ENSG00000154556	352	...
ENSG00000185565	679	...
ENSG00000163328	136	...
ENSG00000064042	1160	...
ENSG00000114423	2293	...

mart_export.txt (from Ensembl/Biomart)

Gene stable ID	Gene name	Chro...	
ENSG00000000971	CFH	1	19665187...
ENSG00000001461	NIPAL3	1	2441...
ENSG00000124942	AHNAK	11	624...
ENSG00000002330	BAD	11	642698...
ENSG00000002549	LAP3	4	175771...
ENSG00000002586	CD99	X	269113...
ENSG00000002834	LASP1	17	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...	
ENSG00000124942	309	...		AHNAK	11	624...
ENSG00000018408	4685	...		WWTR1	3	1495...

Exercise 2: get annotations for a set of genes

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.*

- 1. Click on 👁 to display the content of the dataset SARTools DESeq2 table and download the file siMitfvssiLuc.up.txt (click right, save …)
- 2. Open the file siMitfvssiLuc.up.txt and change the name of the column which contains "Id" to "Gene stable ID" (first word, first line). Save the change.
- 3. 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: colums are in the same order as columns are selected
- 4. 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 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.
  - Tip 1: Search "lincRNA" in the column containing Gene types
  - Tip 2: c3 refers to column 3 of a dataset.
  - Tip3 : 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: Gene stable ID, Chromosome/scaffold name, Gene Start (bp), Gene End (bp), strand and Gene Name