Data mining with Ensembl Biomart

Stéphanie Le Gras
(slegras@igbmc.fr)

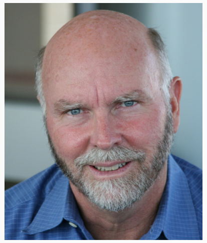
Guidelines

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

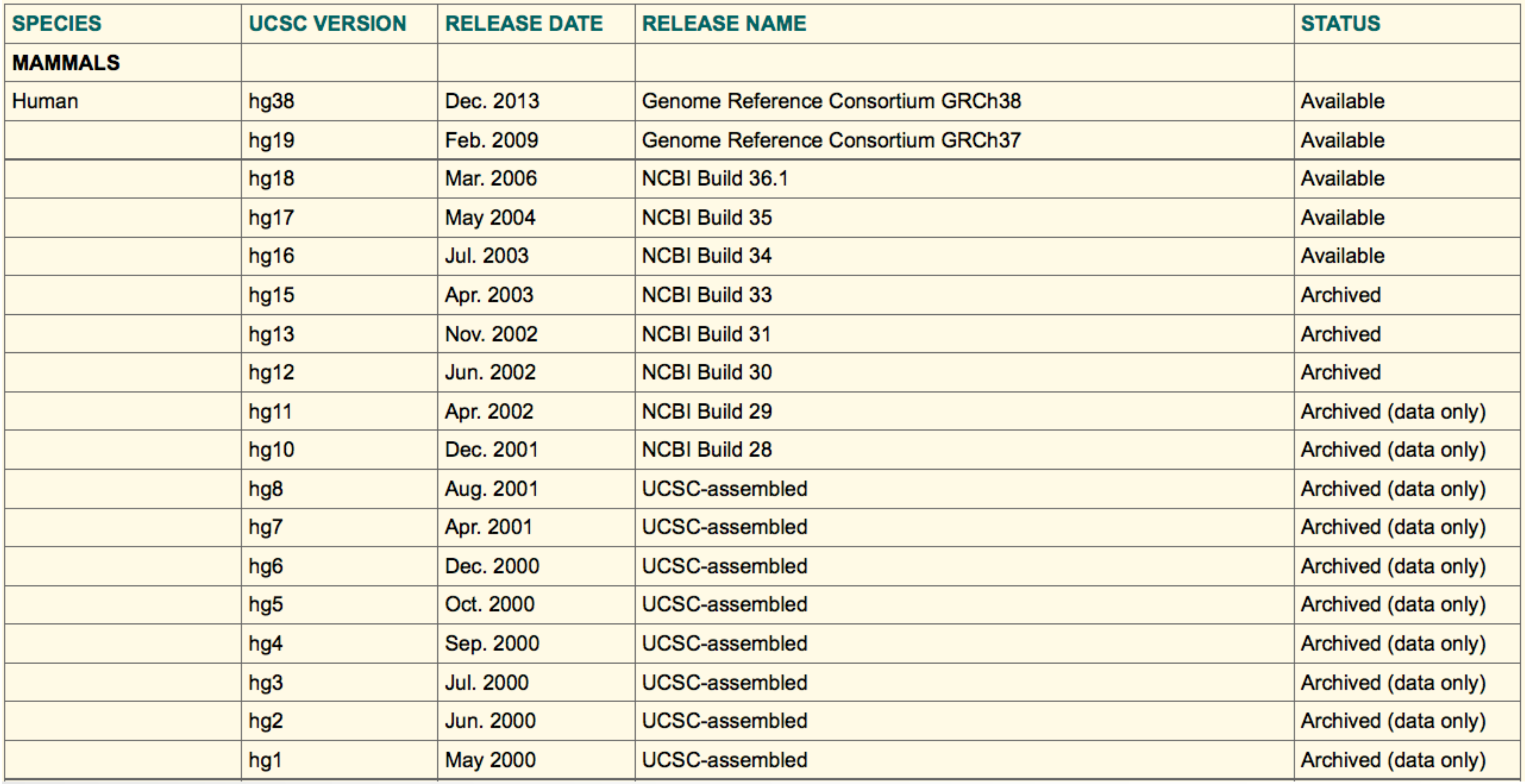
Genome Sequencing

Example: Human genome

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



Genome builds

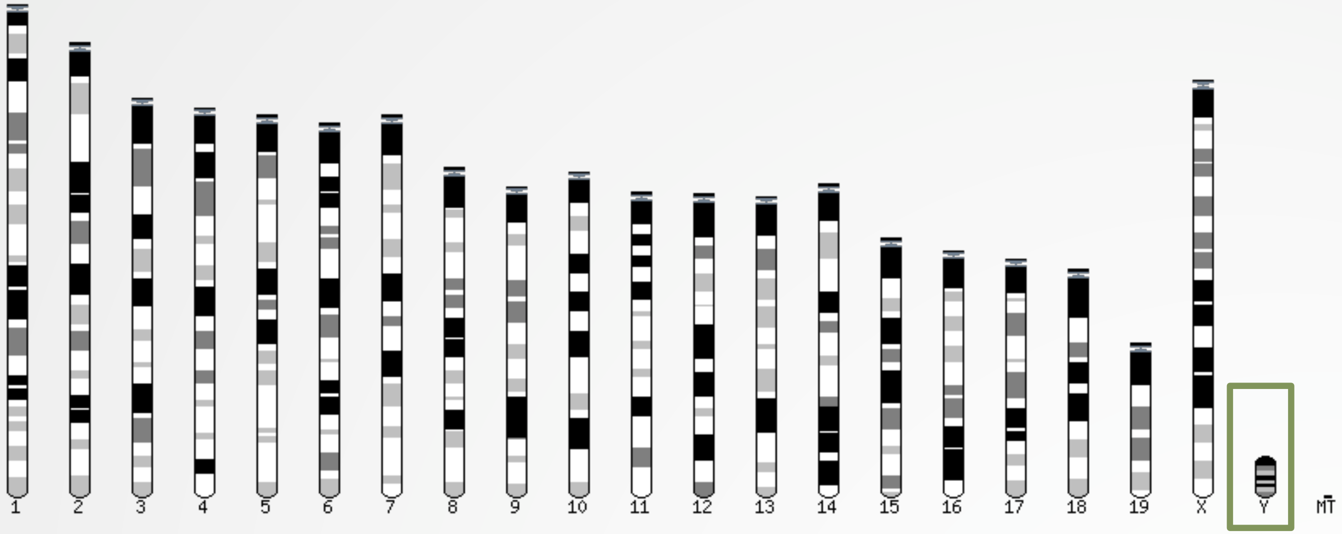


SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
MAMMALS				
Human	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 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)

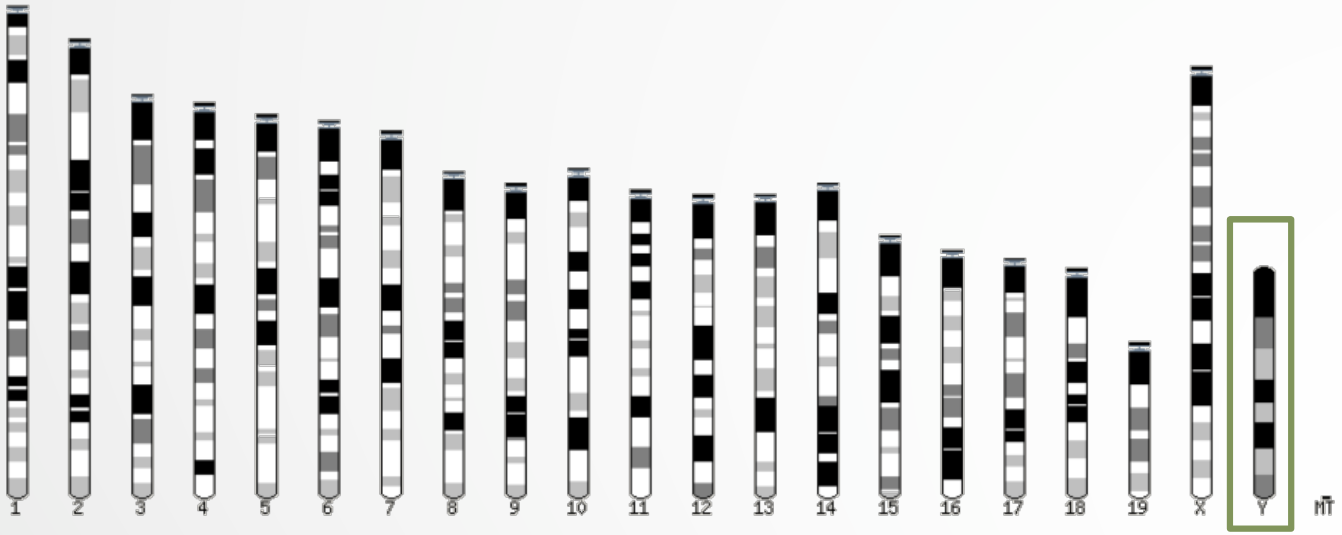
Source: https://genome.ucsc.edu/FAQ/FAQreleases.html

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

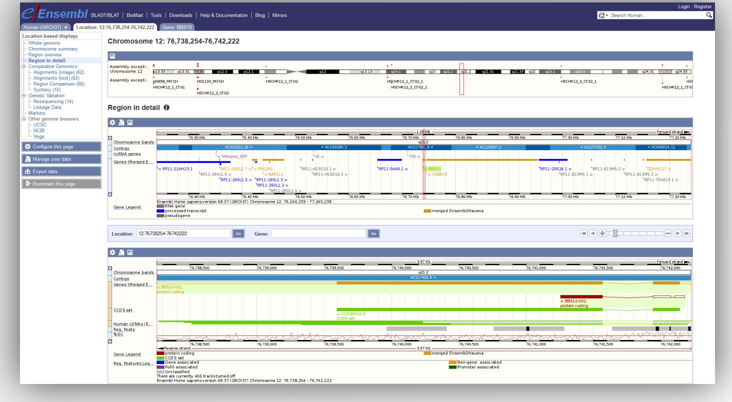
Genome browsers

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



Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh37) · Location: 12:76,738,254-76,742,222 · Gene: BBS10

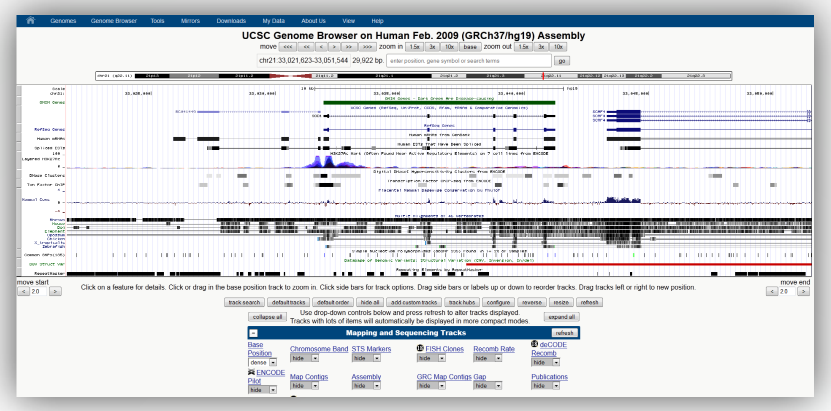
Location-based displays: Whole genome, Chromosome summary, Region overview, Region in detail, Comparative Genomics, Alignments (image) (62), Alignments (text) (62), Region Comparison (66), Synteny (15), Genetic Variation, Resequencing (14), Linkage Data, Markers, Other genome browsers, UCSC, NCBI, Vega

Configure this page · Manage your data · Export data · Bookmark this page

Chromosome 12: 76,738,254-76,742,222

Region in detail

UCSC – Genome Browser



Genomes · Genome Browser · Tools · Mirrors · Downloads · My Data · About Us · View · Help

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

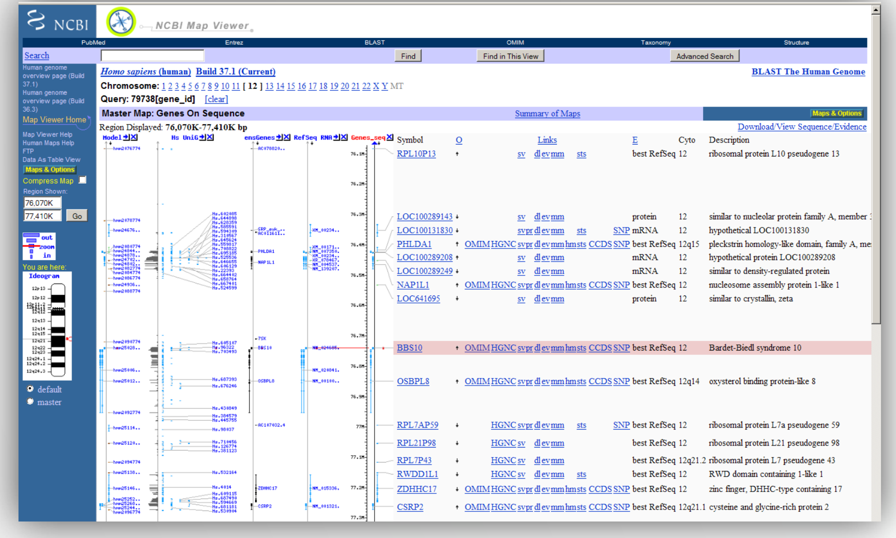
chr21:33,021,623-33,051,544 29,922 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

move · <<< · << · < · > · >> · >>> · zoom in · 1.5x · 3x · 10x · base · zoom out · 1.5x · 3x · 10x · go
move start · move end
track search · default tracks · default order · hide all · add custom tracks · track hubs · configure · reverse · resize · refresh
collapse all · expand all
Base Position · Chromosome Band · STS Markers · FISH Clones · Recomb Rate · deCODE Recomb · ENCODE Pilot · Map Contigs · Assembly · GRC Map Contigs · Gap · Publications



NCBI · NCBI Map Viewer

PubMed · Entrez · BLAST · OMIM · Taxonomy · Structure

Homo sapiens (human) Build 37.1 (Current) · BLAST The Human Genome

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 [12] 13 14 15 16 17 18 19 20 21 22 X Y MT

Query: 79738[gene_id] [clear]

Master Map: Genes On Sequence · Summary of Maps · Maps & Options

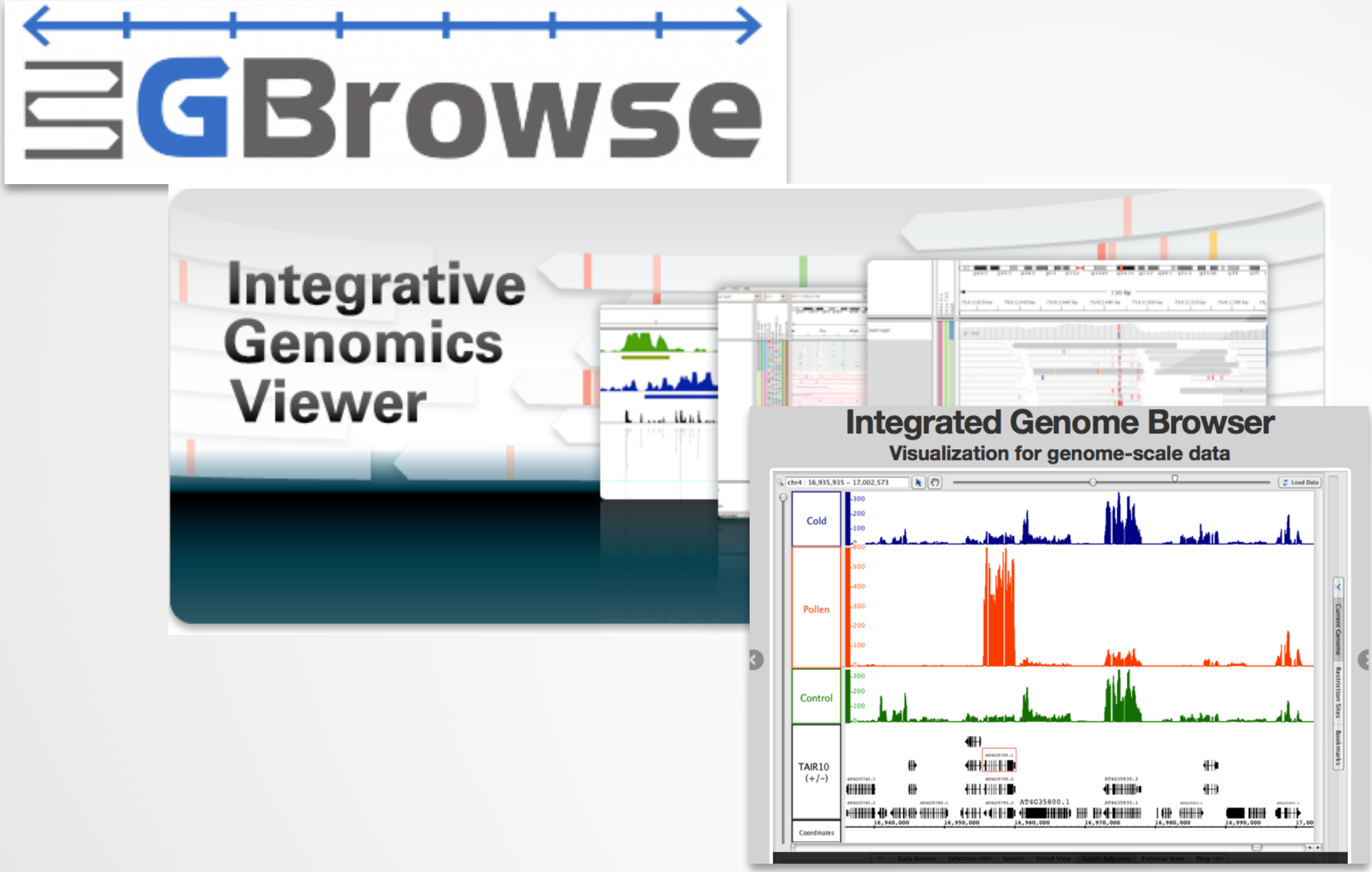
Region Displayed: 76,070K-77,410K bp · Download/View Sequence/Evidence

Symbol	O	Links	E	Cyto	Description
RPL10P13	↑	sv dl ev mm sts	best RefSeq	12	ribosomal protein L10 pseudogene 13
LOC100289143	↓	sv dl ev mm	protein	12	similar to nucleolar protein family A, member
LOC100131830	↓	sv pr dl ev mm sts SNP	mRNA	12	hypothetical LOC100131830
PHLDA1	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	12q15	pleckstrin homology-like domain, family A, me
LOC100289208	↑	sv dl ev mm	mRNA	12	hypothetical protein LOC100289208
LOC100289249	↓	sv dl ev mm	mRNA	12	similar to density-regulated protein
NAP1L1	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	12	nucleosome assembly protein 1-like 1
LOC641695	↓	sv dl ev mm	protein	12	similar to crystallin, zeta
BBS10	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	12	Bardet-Biedl syndrome 10
OSBPL8	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	12q14	oxysterol binding protein-like 8
RPL7AP59	↓	HGNC sv pr dl ev mm sts SNP	best RefSeq	12	ribosomal protein L7a pseudogene 59
RPL21P98	↓	HGNC sv dl ev mm	best RefSeq	12	ribosomal protein L21 pseudogene 98
RPL7P43	↓	HGNC sv dl ev mm	best RefSeq	12q21.2	ribosomal protein L7 pseudogene 43
RWDD1L1	↓	HGNC sv dl ev mm sts	best RefSeq	12	RWD domain containing 1-like 1
ZDHHC17	↓	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	12	zinc finger, DHHC-type containing 17
CSRP2	↑	OMIM HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	12q21.1	cysteine and glycine-rich protein 2

Search · Find · Find in This View · Advanced Search
Human genome overview page (Build 37.1)
Human genome overview page (Build 36.3)
Map Viewer Home
Map Viewer Help
Human Maps Help
FTP
Data As Table View
Maps & Options
Compress Map
Region Shown:
76,070K
77,410K
Go
out zoom in
You are here:
Ideogram
default
master
Model · Hs UniG · ensGenes · RefSeq RNA · Genes_seq

NCBI – Map Viewer

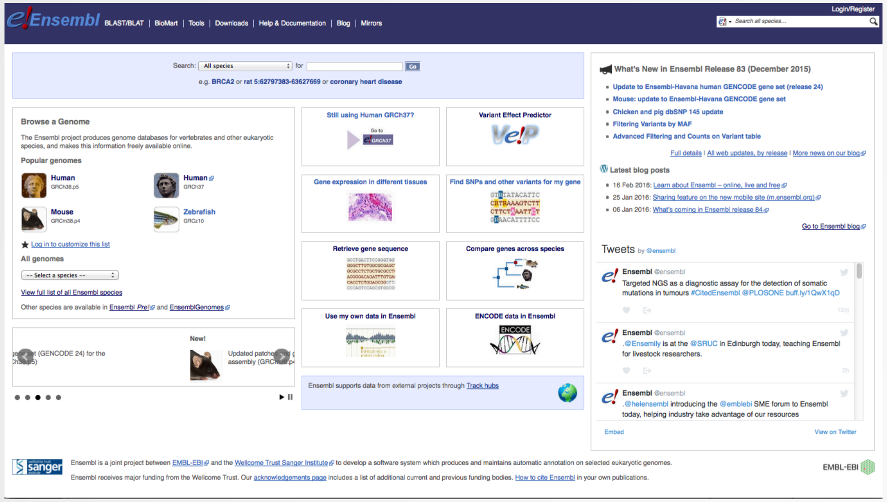
And Genome browsers…



Getting access to genomic data: ENSEMBL/BIOmart

Access Ensembl’s data

Web site



e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
Go
Still using Human GRCh37?
Go to e! GRCh37
Variant Effect Predictor
VE!P
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
Compare genes across species
Use my own data in Ensembl
ENCODE data in Ensembl
ENCODE
New!
Embed
View on Twitter
sanger
EMBL-EBI

Search: All species for

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

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh38.p5

Human GRCh37

Mouse GRCm38.p4

Zebrafish GRCz10

★ Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl Pre! and EnsemblGenomes

t (GENCODE 24) for the h38.p5)

Updated patches and g assembly (GRCm38.p4

Ensembl supports data from external projects through Track hubs

What's New in Ensembl Release 83 (December 2015)

- Update to Ensembl-Havana human GENCODE gene set (release 24)
- Mouse: update to Ensembl-Havana GENCODE gene set
- Chicken and pig dbSNP 145 update
- Filtering Variants by MAF
- Advanced Filtering and Counts on Variant table

Full details | All web updates, by release | More news on our blog

Latest blog posts

- 16 Feb 2016: Learn about Ensembl – online, live and free
- 25 Jan 2016: Sharing feature on the new mobile site (m.ensembl.org)
- 06 Jan 2016: What's coming in Ensembl release 84

Go to Ensembl blog

Tweets by @ensembl

Ensembl @ensembl
Targeted NGS as a diagnostic assay for the detection of somatic mutations in tumours #CitedEnsembl @PLOSONE buff.ly/1QwX1qD
12m

Ensembl @ensembl
.@Ensemily is at the @SRUC in Edinburgh today, teaching Ensembl for livestock researchers.
2h

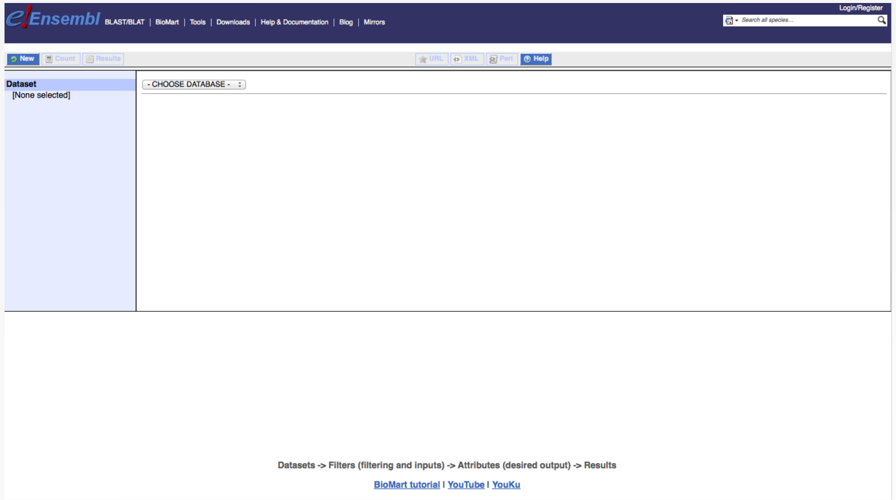
Ensembl @ensembl
.@helensembl introducing the @emblebi SME forum to Ensembl today, helping industry take advantage of our resources

Ensembl is a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes.

Ensembl receives major funding from the Wellcome Trust. Our acknowledgements page includes a list of additional current and previous funding bodies. How to cite Ensembl in your own publications.

🙂 User friendly
🙂 Straightforward
🙁 Only one request at once

Mining tool: BioMart



e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
New
Count
Results
URL
XML
Perl
Help

Dataset
[None selected]

- CHOOSE DATABASE -

Datasets -> Filters (filtering and inputs) -> Attributes (desired output) -> Results

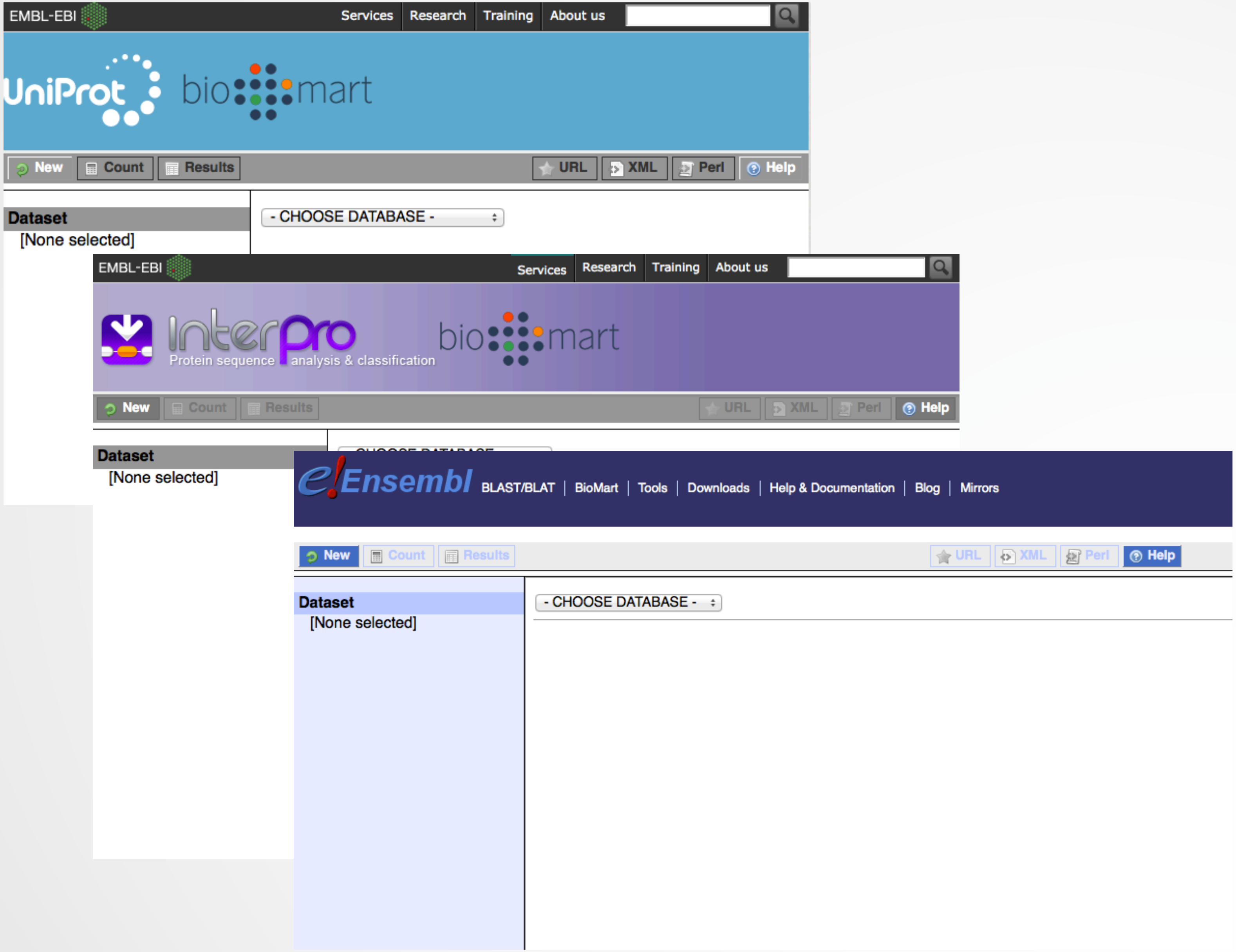
BioMart tutorial | YouTube | YouKu

🙂 Get answer to complex query
🙂 Very fast
🙁 Need training

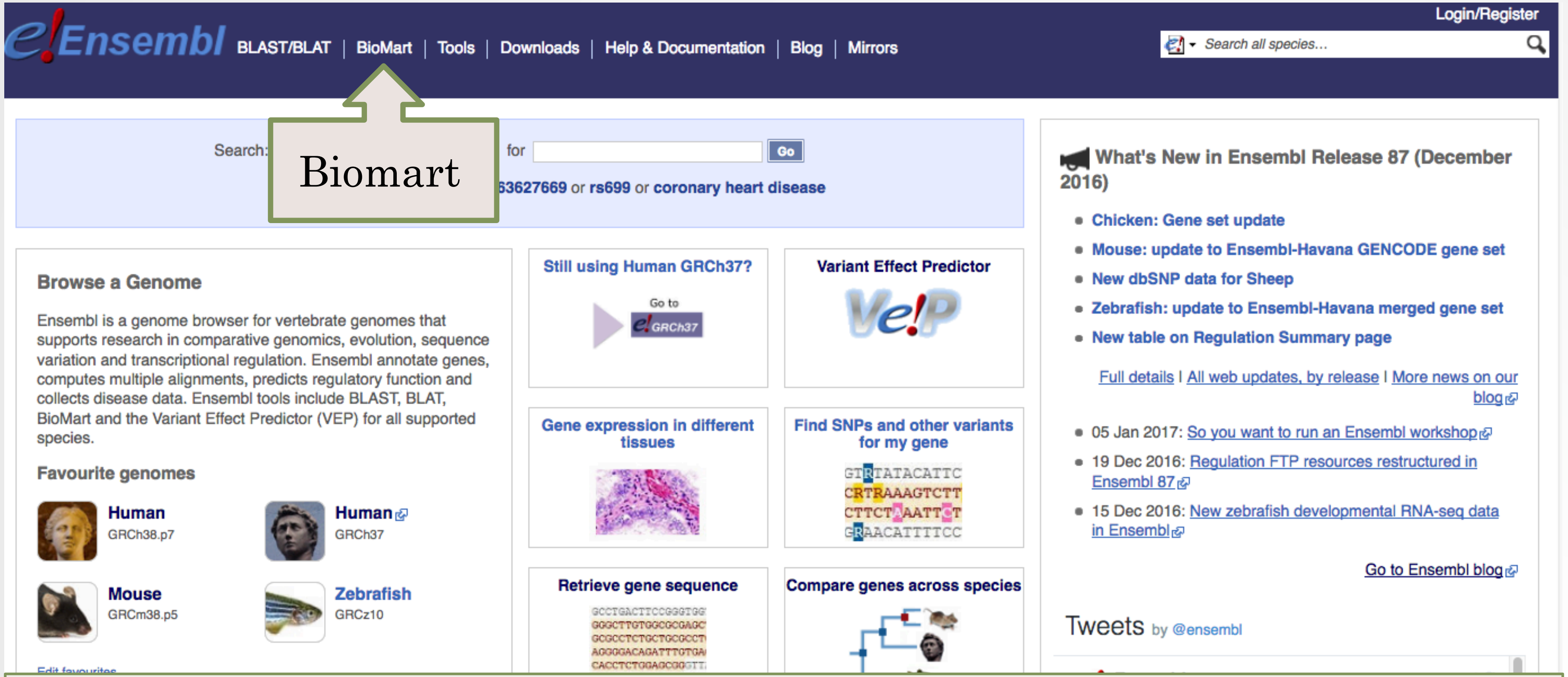
BioMart

- 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



Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Login/Register

Search all species...

Biomart

Search: for Go

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 GRCh37

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

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

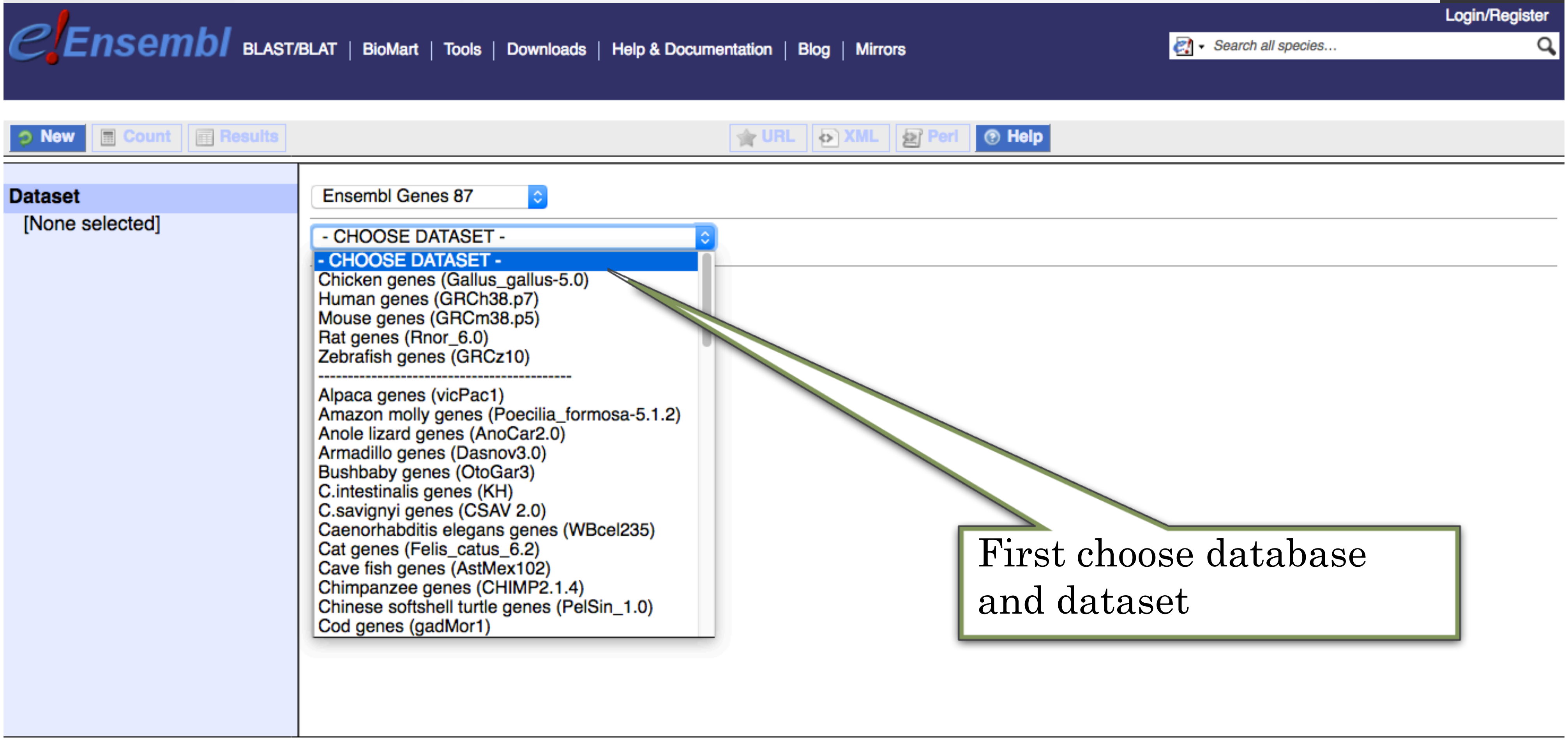
- 05 Jan 2017: So you want to run an Ensembl workshop
- 19 Dec 2016: Regulation FTP resources restructured in Ensembl 87
- 15 Dec 2016: New zebrafish developmental RNA-seq data in Ensembl

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
Search all species...
New
Count
Results
URL
XML
Perl
Help
- 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)

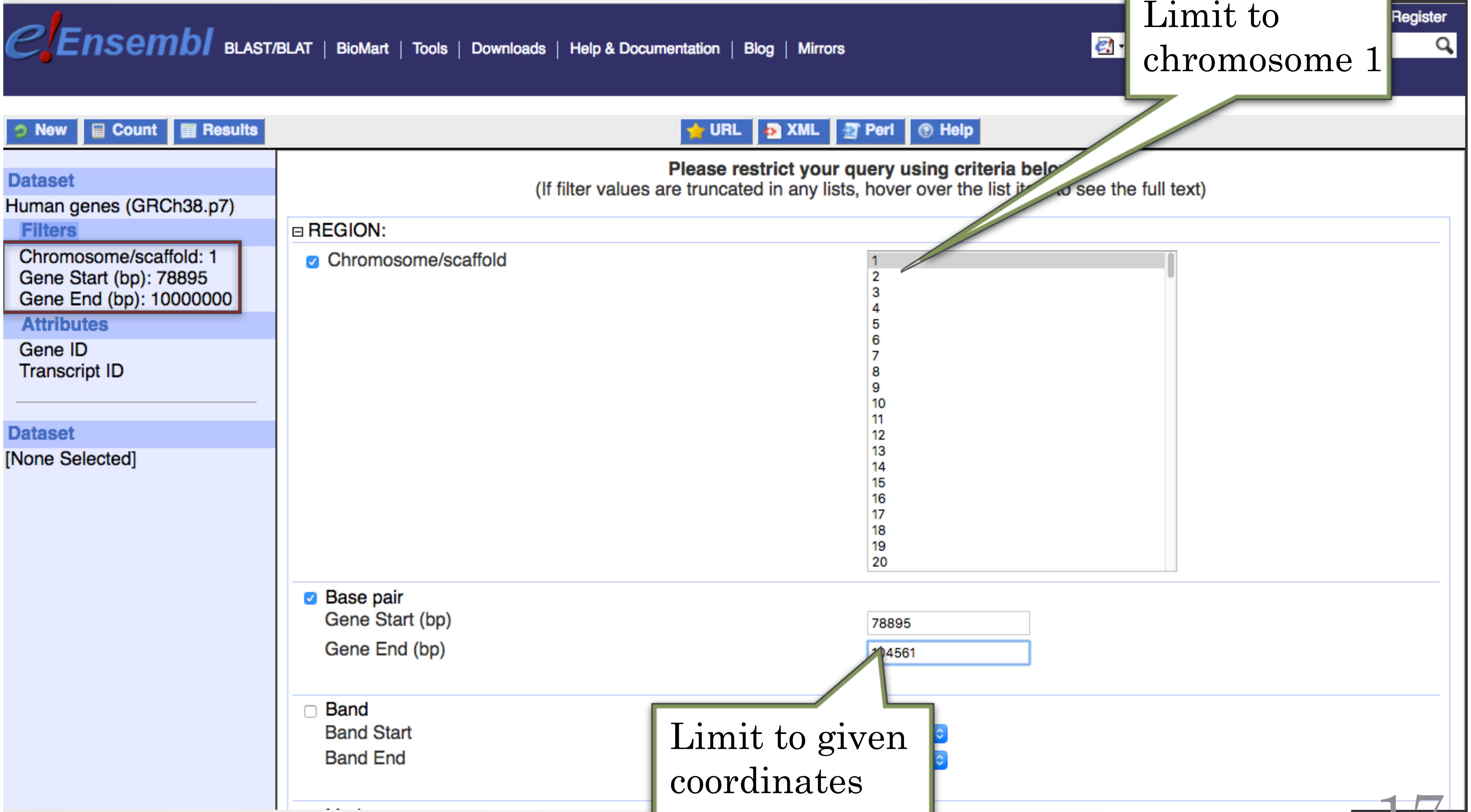
Dataset
[None selected]

Ensembl Genes 87

- CHOOSE DATASET -

First choose database and dataset

Example: Step 2 (Filter)



e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Register
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 bel
(If filter values are truncated in any lists, hover over the list it o 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

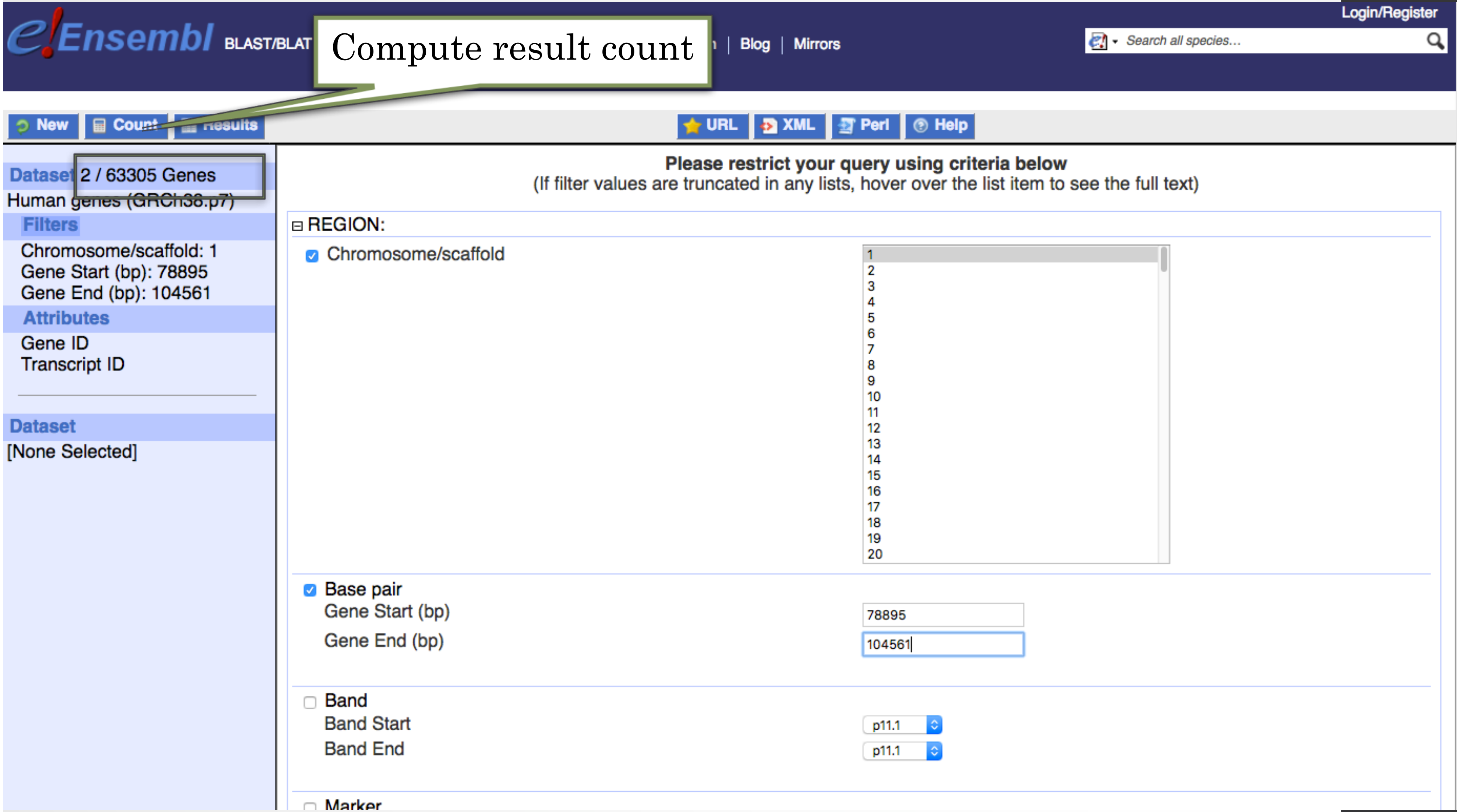
☑ Base pair
Gene Start (bp): 78895
Gene End (bp): 104561

□ Band
Band Start
Band End

Limit to chromosome 1

Limit to given coordinates

Example: Step 3 (Count results)



Compute result count

Dataset 2 / 63305 Genes

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 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:

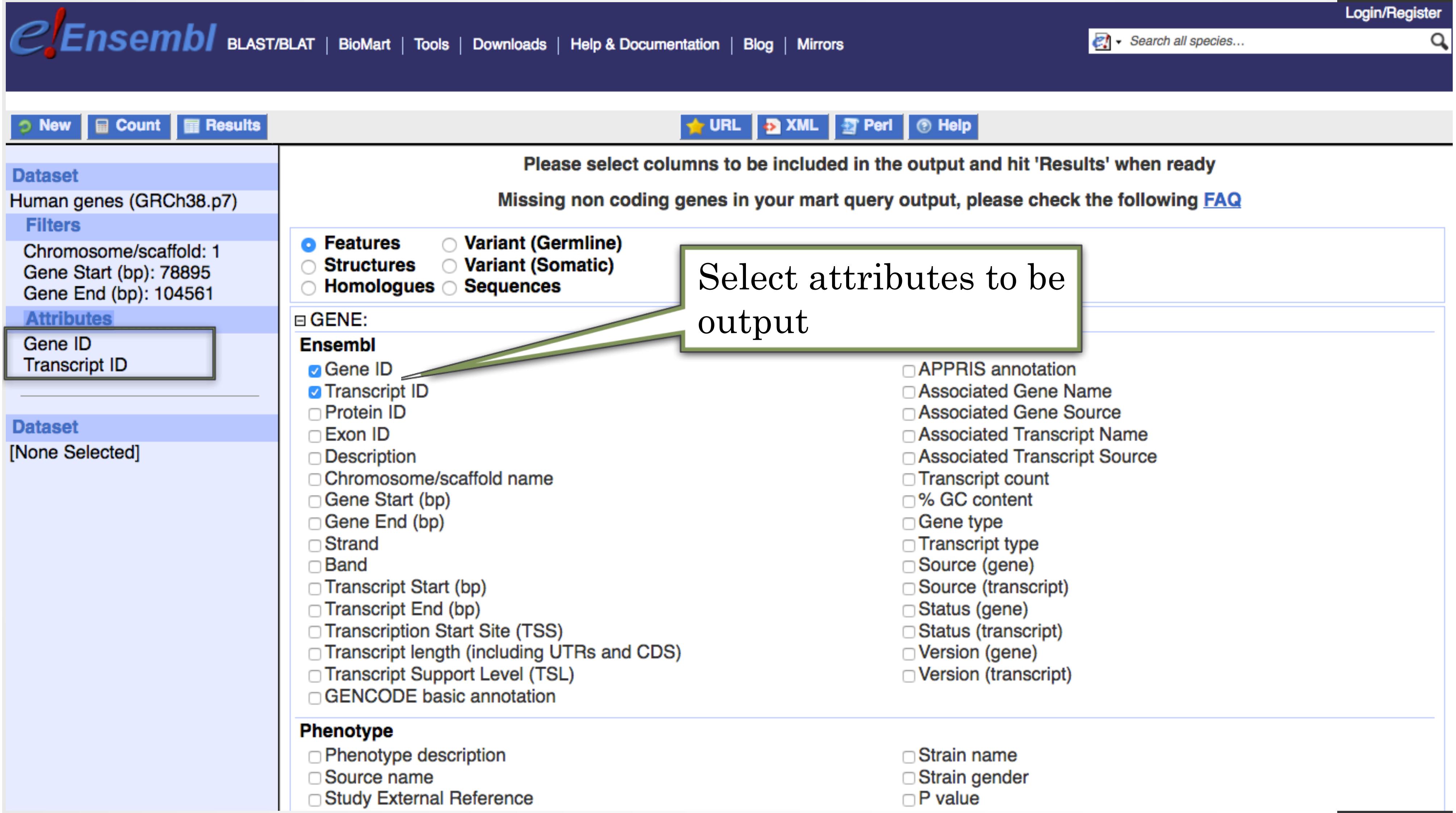
☑ Chromosome/scaffold: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

☑ Base pair
Gene Start (bp) 78895
Gene End (bp) 104561

☐ Band
Band Start p11.1
Band End p11.1

☐ Marker

Example: Step 4 (Select attributes)



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]

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

Select attributes to be output

⊟ GENE:

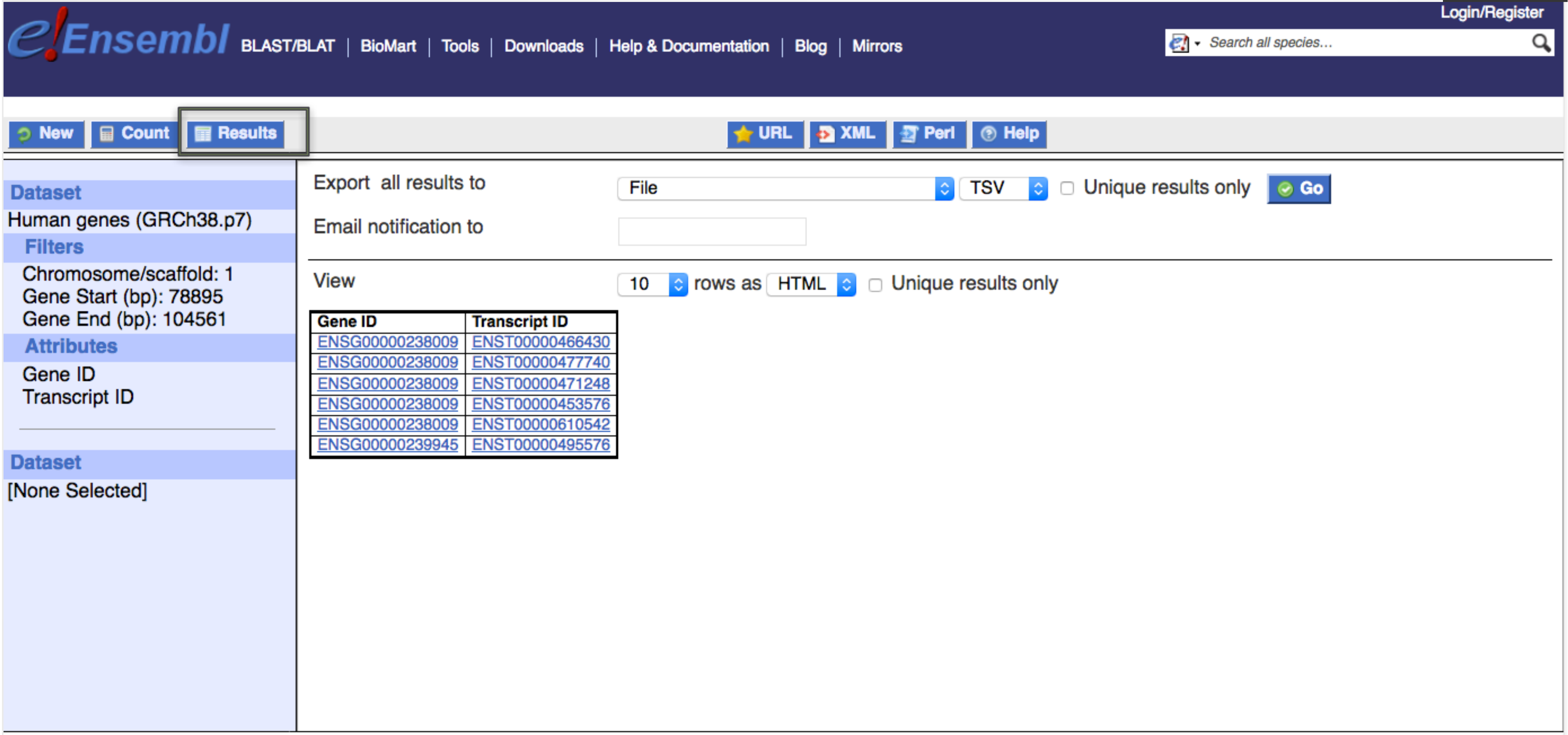
Ensembl

☑ 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
□ 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)

Phenotype

□ Phenotype description
□ Source name
□ Study External Reference
□ Strain name
□ Strain gender
□ P value

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
Go

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 File TSV □ Unique results only

Email notification to

View 10 rows as HTML □ Unique results only

Gene ID	Transcript ID
ENSG00000238009	ENST00000466430
ENSG00000238009	ENST00000477740
ENSG00000238009	ENST00000471248
ENSG00000238009	ENST00000453576
ENSG00000238009	ENST00000610542
ENSG00000239945	ENST00000495576

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 **v85**, 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 siMitfvssıLuc.up.txt (click right, save …)

- 1. Open the file siMitfvssiLuc.up.txt and change the name of the column which contains “Id” 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 “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 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