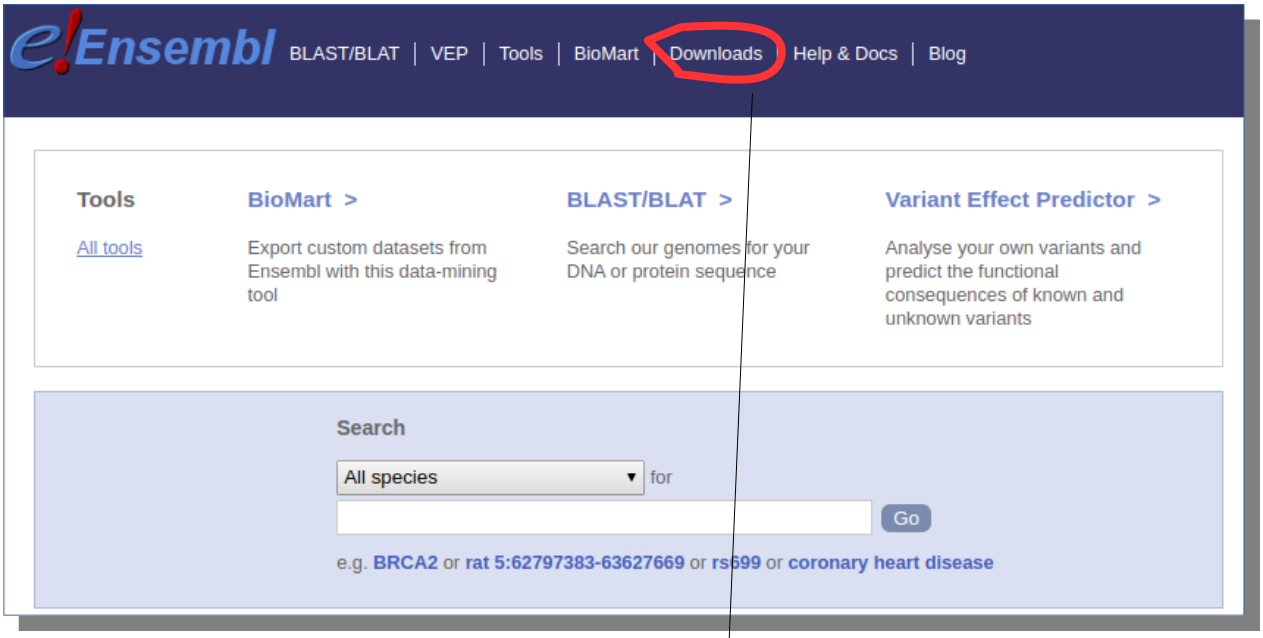
RNA-Seq Analysis: Geting Data

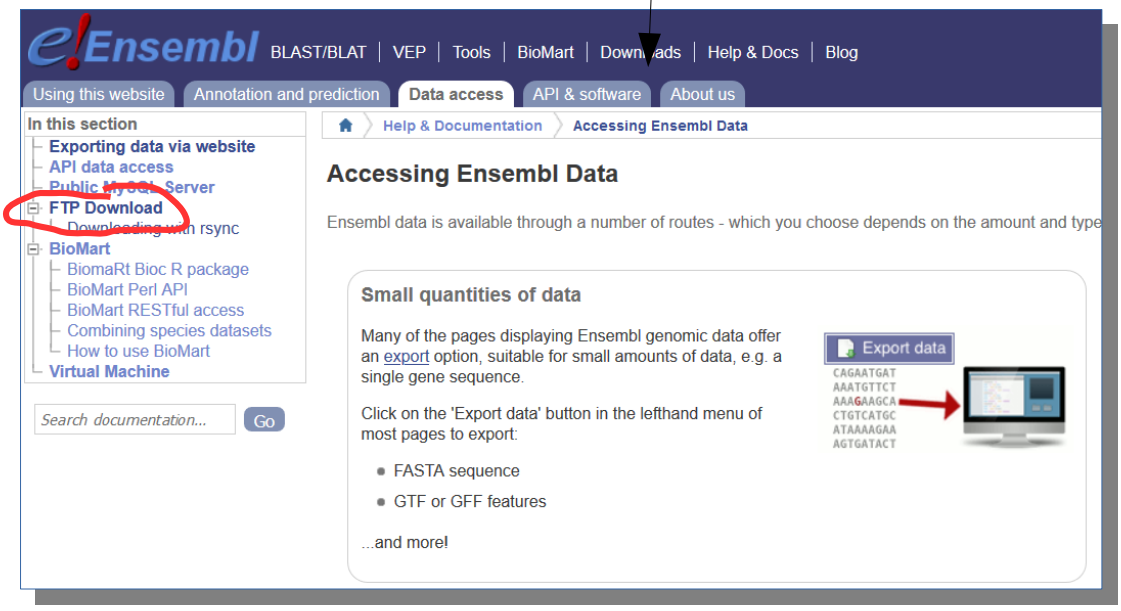
Collect the FTP link to get the Gene and Genome references:

- Collect the files needed to prepare the sequence reference and the gene definitions.
- Ensembl Genome Browser is a good source for references and gene definitions.

http://www.ensembl.org/



Go to the Downloads section



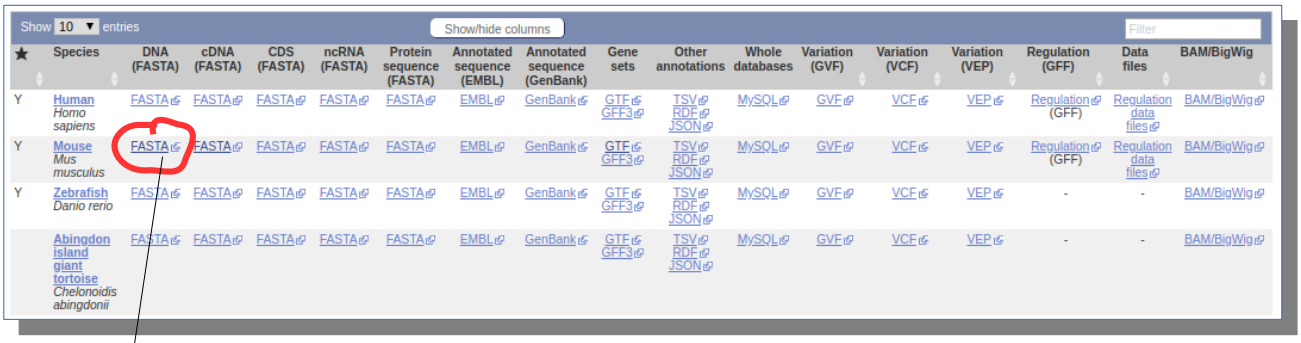
Go to Download Data Via FTP

RNA-Seq Analysis: Geting Reference Data

This table provides links to the different ftp sites with data related to each species in different formats.

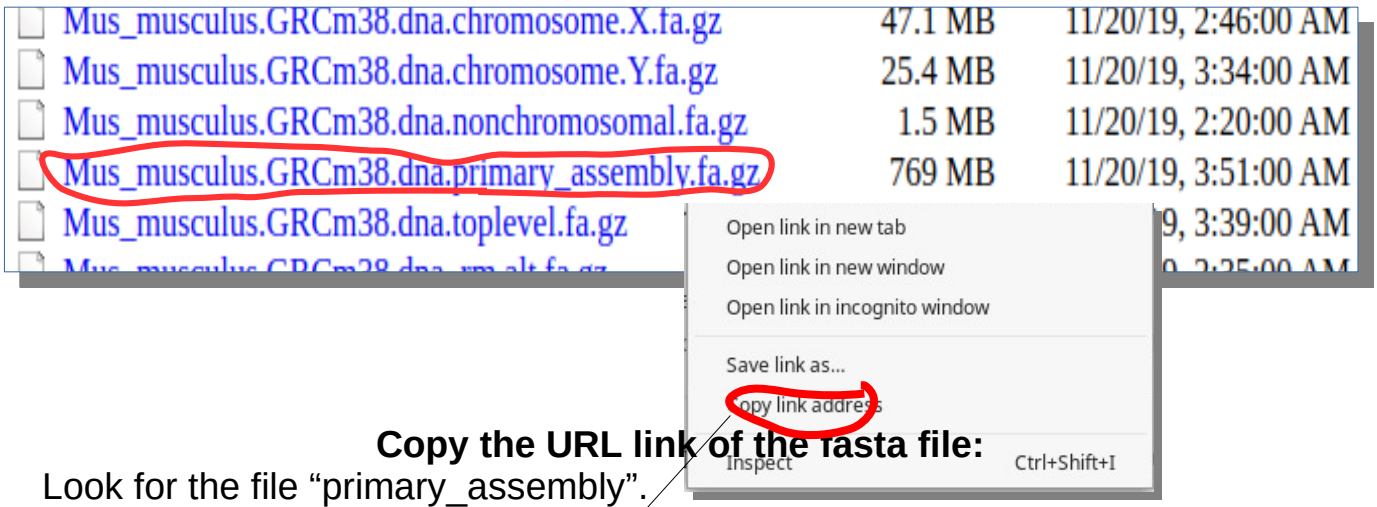
- **DNA**: ftp site to get genome reference sequence in FASTA format
- **CDNA**: ftp site to get transcripts sequences in FASTA format
- **Protein**: ftp site to get Protein sequences in FASTA format
- **Gene**: ftp site to get gene definition data in GTF format

...



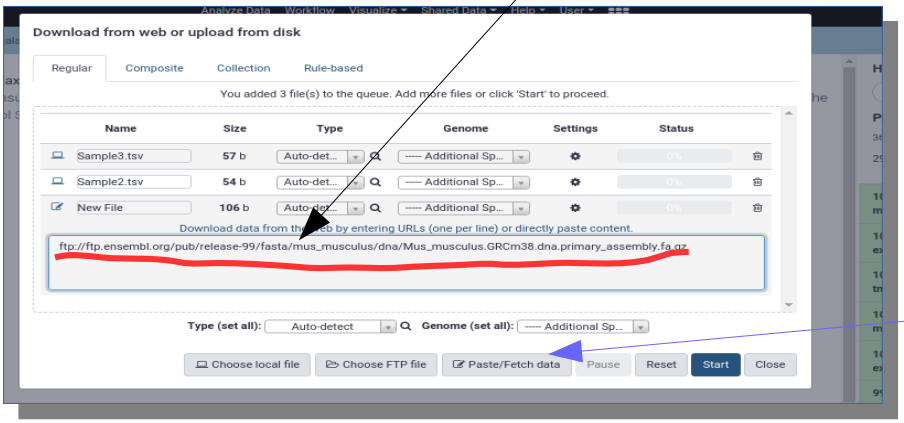
Select DNA link for the Mouse:

To find the genome reference sequence for the mouse.



Copy the URL link of the fasta file:

Look for the file “primary_assembly”.
In the contextual menu (right click in the mouse) select the option to copy the URL link of the file.
Paste the URL into the Upload Tool → Paste/Fetch Box in Galaxy.

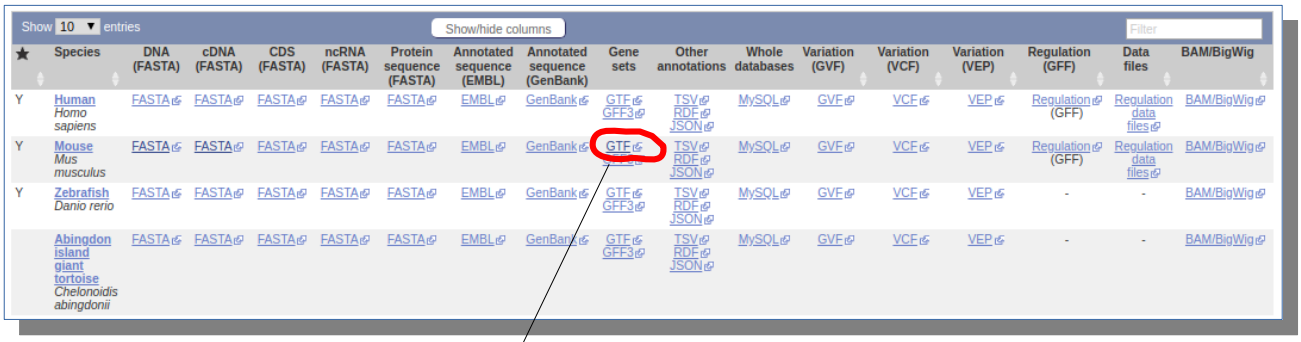


Use this option (Paste/Fetch) to transfer data from another server using an URL link.

RNA-Seq Analysis: Geting Reference Data

Select the Gene GTF link from the previous table

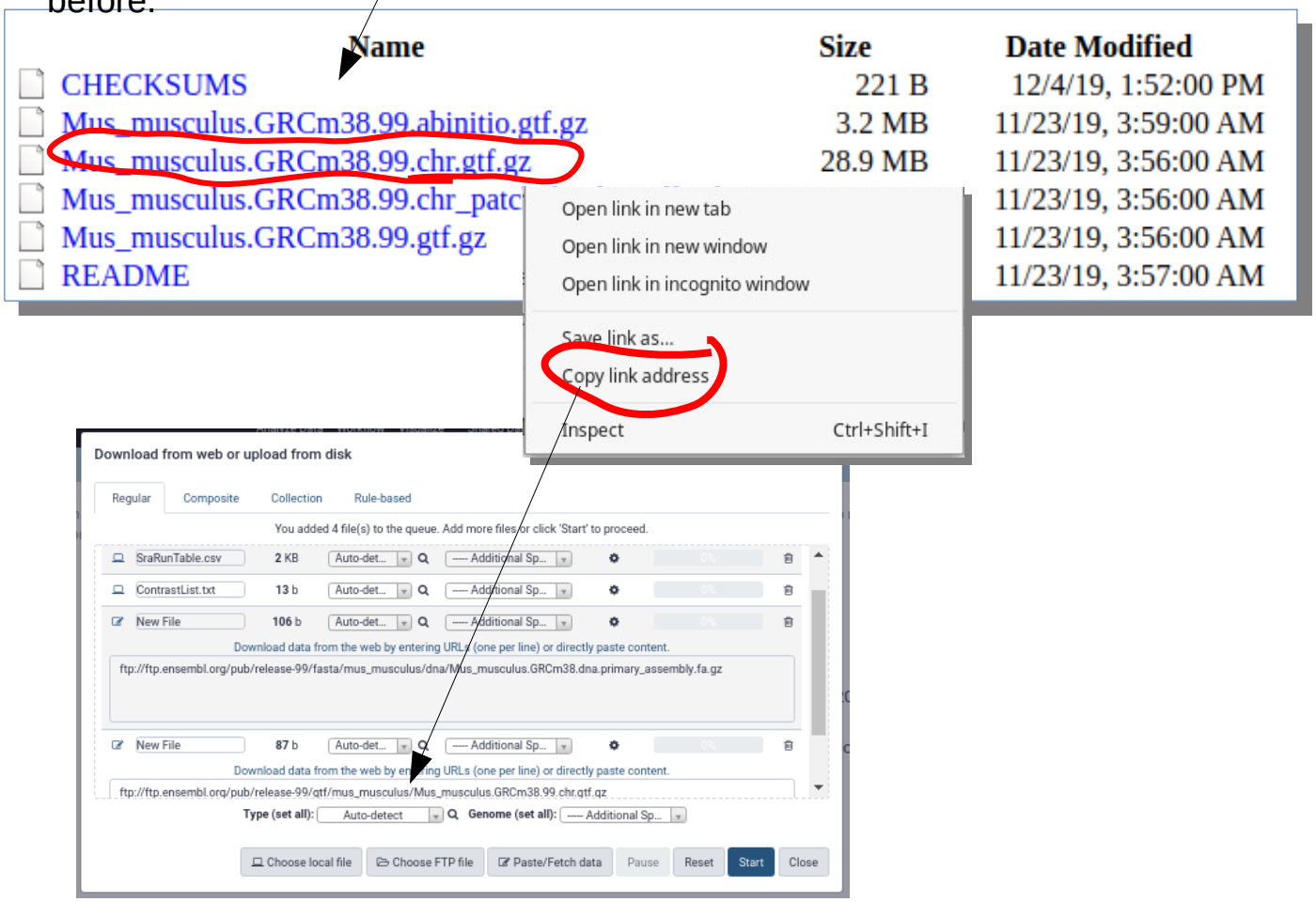
To find the gtf file with the gene definitions for mouse genome



★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Other annotations	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM/BigWig
Y	Human *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BigWig
Y	Mouse *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BigWig
Y	Zebrafish *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP	-	-	BAM/BigWig
	Abingdon island giant tortoise *Chelonoidis abingdonii*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP	-	-	BAM/BigWig

Copy the URL link of the gtf file:

Look for the file “Mus_musculus.GRCm38.99.chr.gtf”.
This is the Gene Build definition version 99 on mouse genome assembly version GRCm38.
In the contextual menu (right click in the mouse) select the option to copy the URL link of the file.
Paste the URL into the Upload Tool → Paste/Fetch Box in Galaxy as we did before.



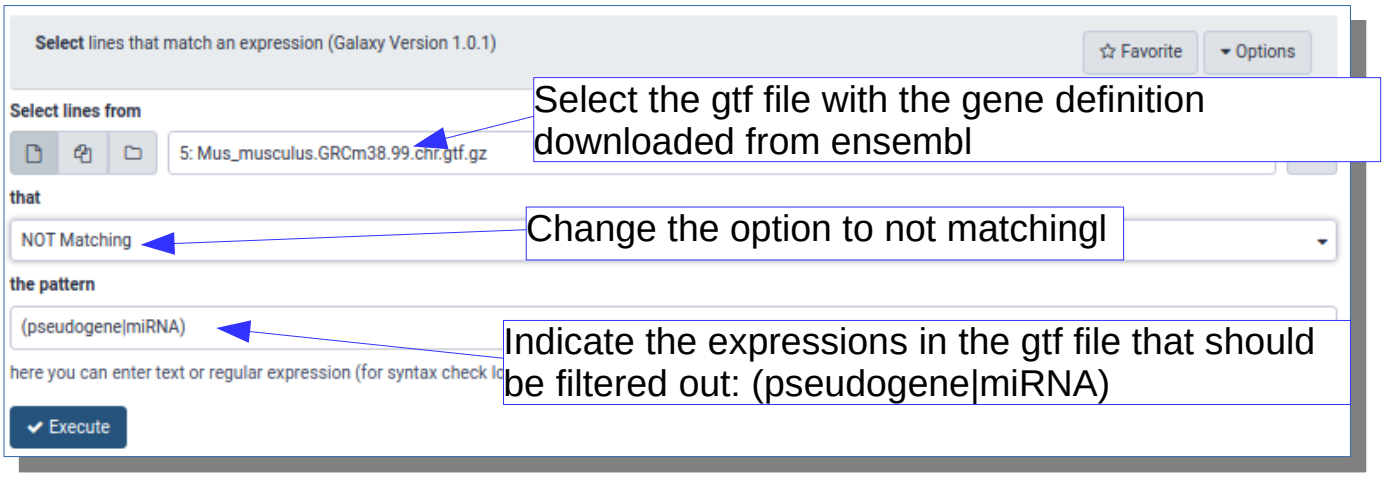
Now, you can click the Start button and Galaxy will collect the four datasets (the two files from your computer and the references from Ensembl ftp links).

RNA-Seq Analysis: Prepare Reference

- We are going to use “Salmon” as a tool to align reads to the reference and quantify the gene expression.
- Salmon needs as reference the sequences of the transcripts and a file with the relation of the transcripts to the genes.
- We will first prepare de transcript reference and the two column table with the transcript IDs in the first column and the gene IDs they belong to in the second column.

Select: lines that match an expression.

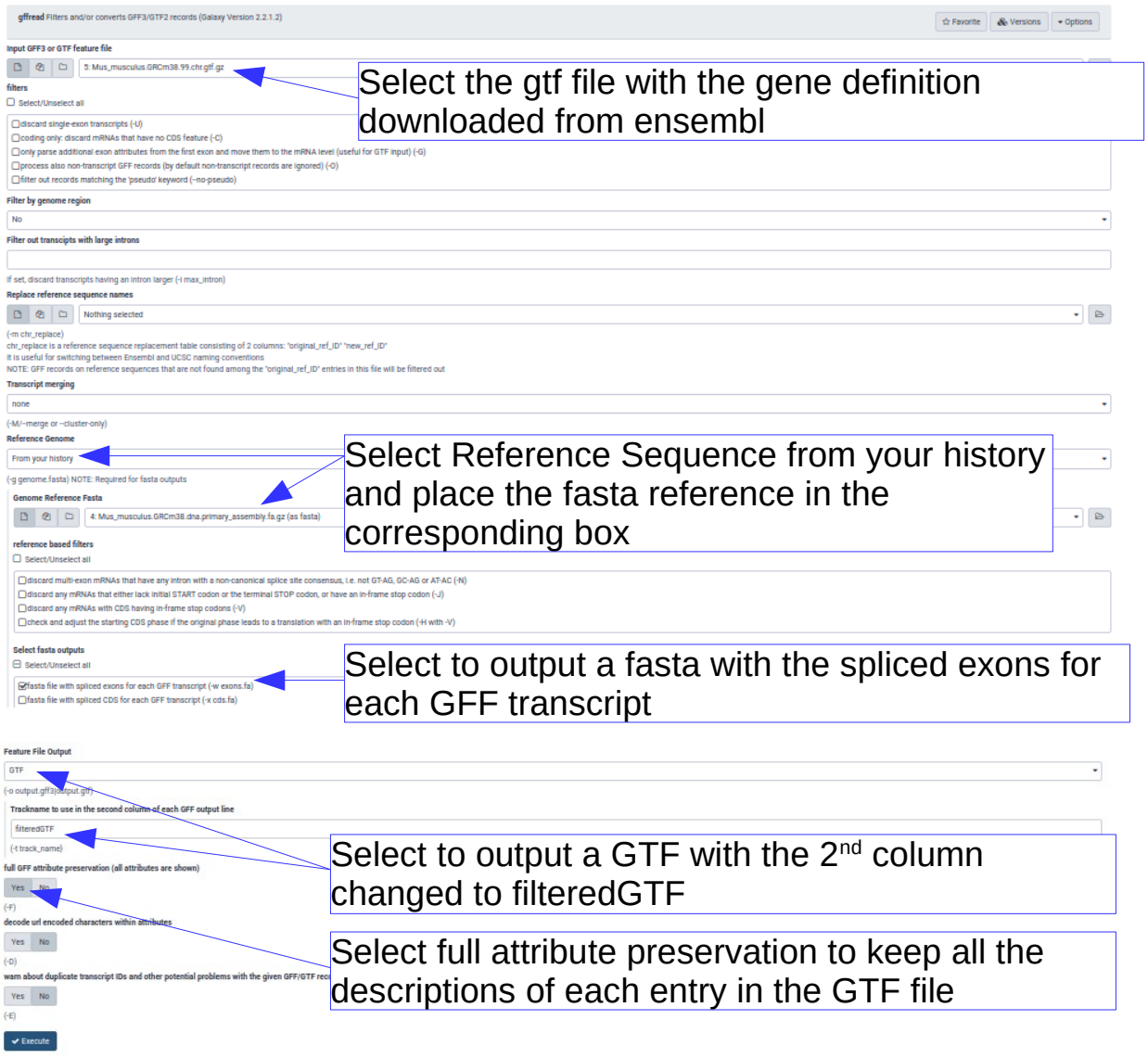
- To improve the quality of the quantifications.
- Use the annotation of the gtf file, in particular the transcript_biotype annotation to remove those transcripts annotated as pseudogenes or miRNAs.



RNA-Seq Analysis: Prepare Reference

Gffread: To get the fasta file of the transcripts

This tool will use a genome reference fasta file and a gtf gene definition file to build a transcript reference fasta file and another gtf with reduced annotation.



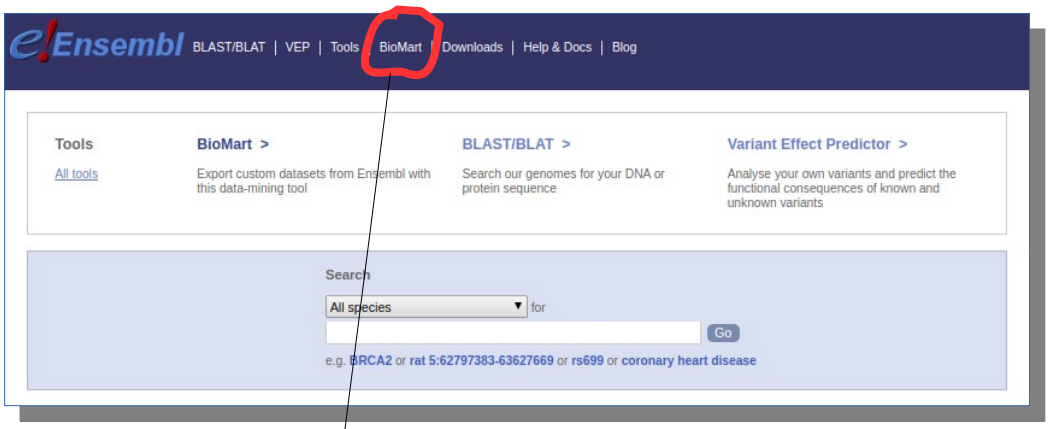
The tool will output two datasets:

- A GTF gene definition file with the lines of the gtf file used to build the transcripts sequences
- A FASTA file with the sequences of the transcripts.

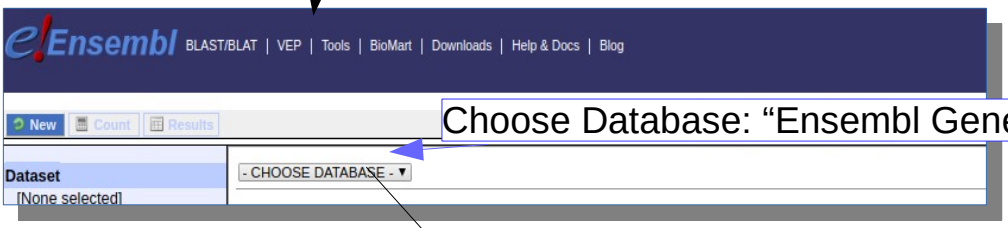
Transcript to Gene mapping file

- Salmon needs a file of two columns with the relation of the transcripts to the genes.
- The first column will have the IDs of the transcripts used as reference
- The second column will have the IDs of the genes they belong to.
- To build this file we will use the BioMart tool from Ensembl web site.

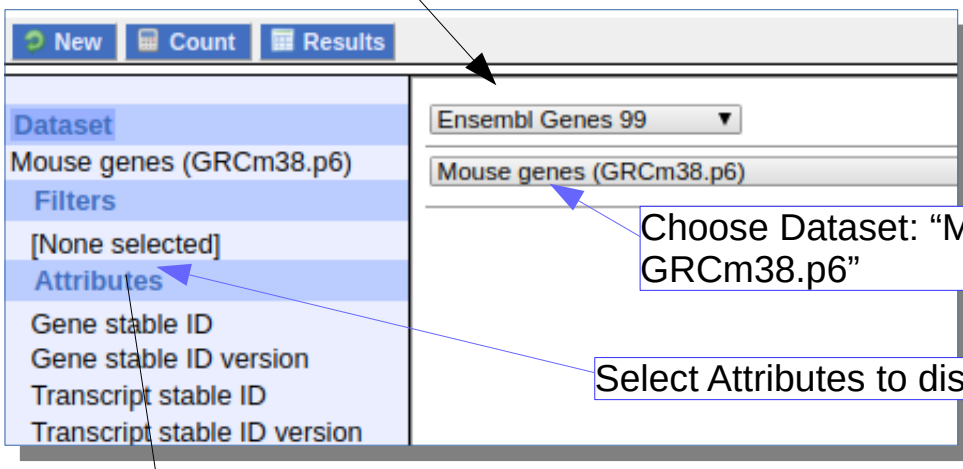
Go to ensembl.org



Go to BioMart

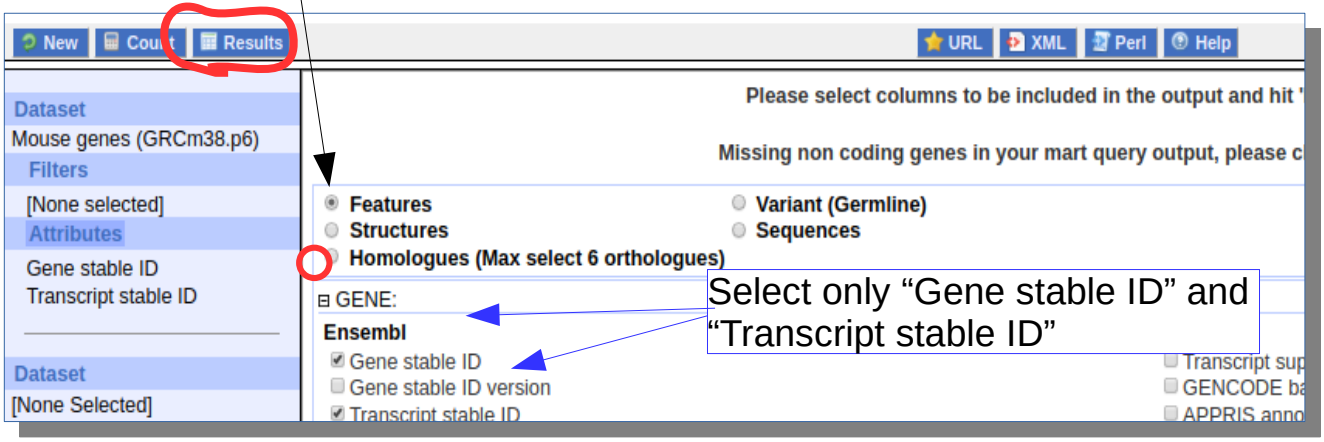


Choose Database: “Ensembl Genes 99”



Choose Dataset: “Mouse Genes GRCm38.p6”

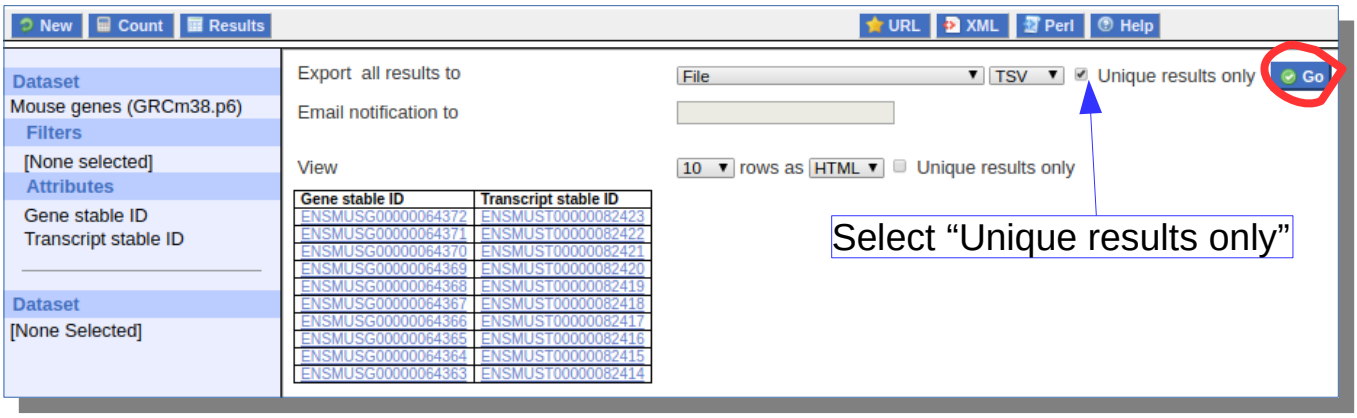
Select Attributes to display



Select only “Gene stable ID” and “Transcript stable ID”

Then click RESULTS

RNA-Seq Analysis: Prepare Reference

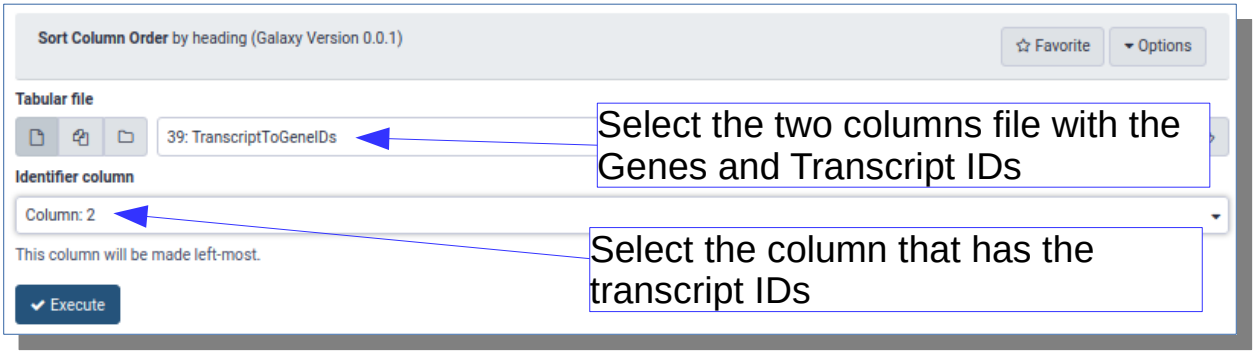


Click **GO**:

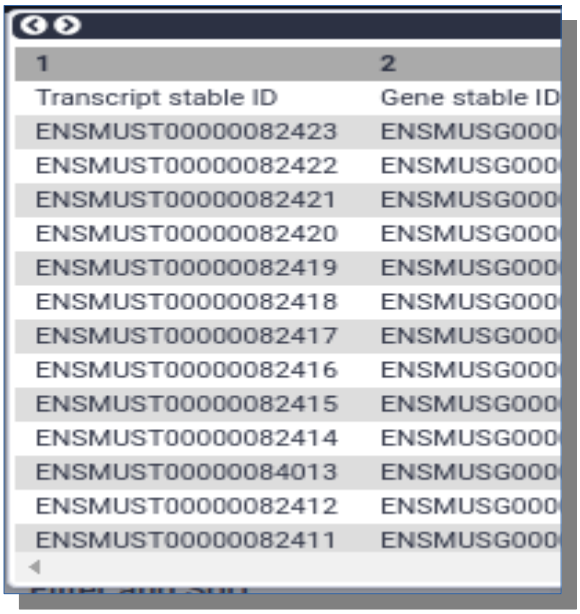
- This will download a tab delimited file with those two columns
- Upload the file to Galaxy
- We will need to invert the order of the columns, first the transcript column and second the gene column

Sort Column Order: by heading

This tool will place the column selected as “Identifier” in the first place



Now, we should have a file like this one



1	2
Transcript stable ID	Gene stable ID
ENSMUST00000082423	ENSMUSG000
ENSMUST00000082422	ENSMUSG000
ENSMUST00000082421	ENSMUSG000
ENSMUST00000082420	ENSMUSG000
ENSMUST00000082419	ENSMUSG000
ENSMUST00000082418	ENSMUSG000
ENSMUST00000082417	ENSMUSG000
ENSMUST00000082416	ENSMUSG000
ENSMUST00000082415	ENSMUSG000
ENSMUST00000082414	ENSMUSG000
ENSMUST00000084013	ENSMUSG000
ENSMUST00000082412	ENSMUSG000
ENSMUST00000082411	ENSMUSG000