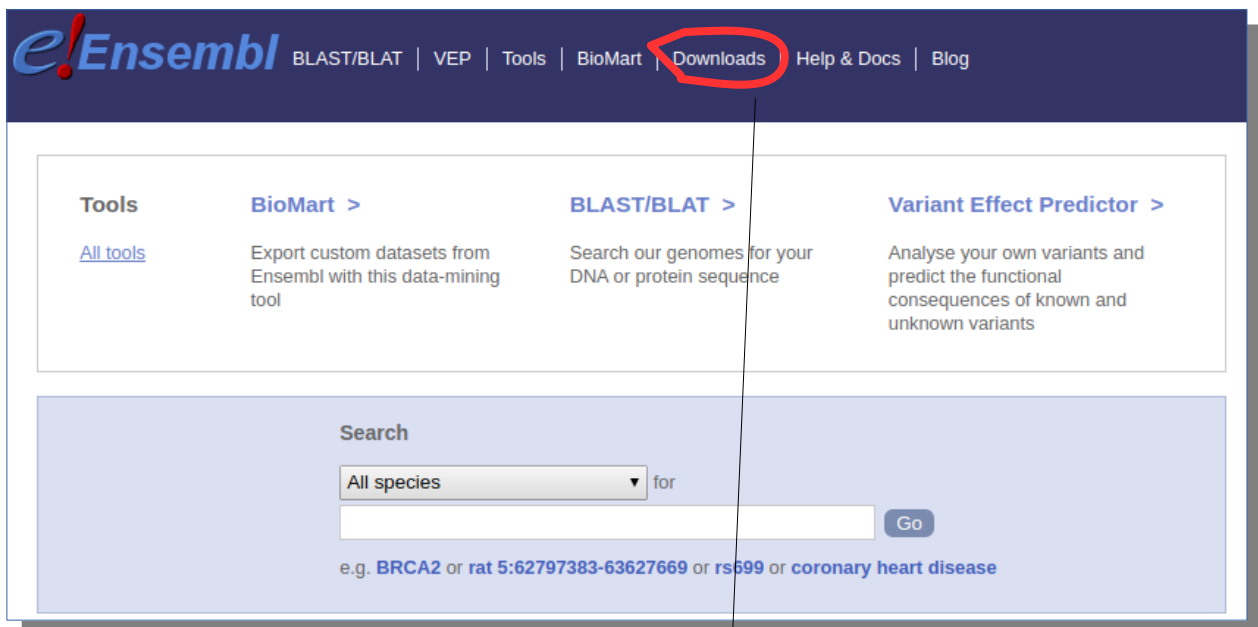


RNA-Seq Analysis: Getting Data

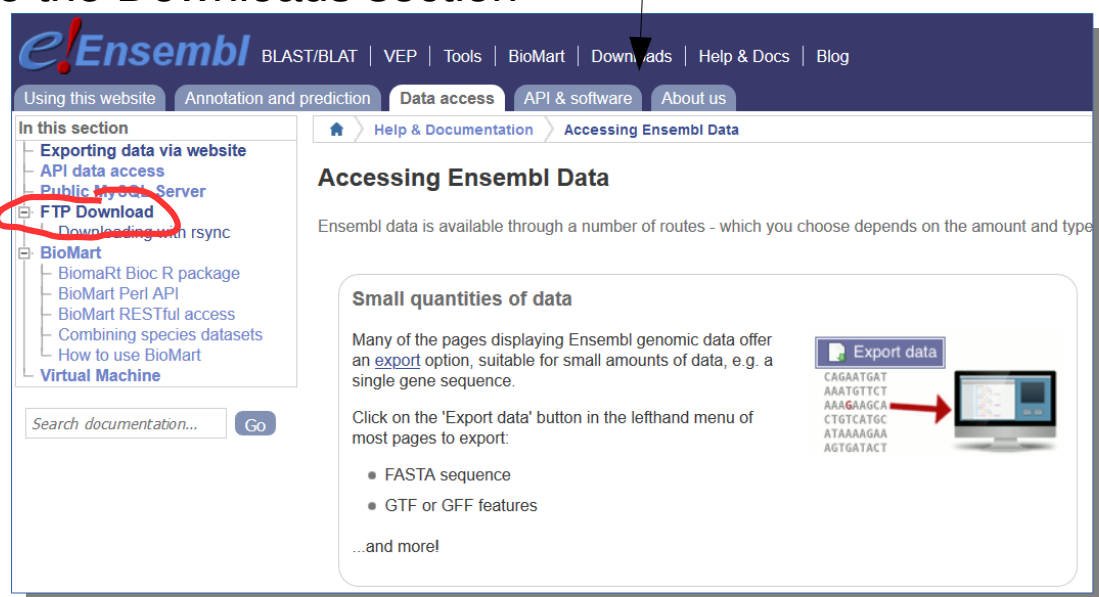
Collect the FTP link to get the Gene and Genome references:

- 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: Getting 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

...

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/BioWig
Human	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	TSV	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BioWig
Mouse	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	TSV	MySQL	GVF	VCF	VEP	Regulation (GFF)	Regulation data files	BAM/BioWig
Zebrafish	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	TSV	MySQL	GVF	VCF	VEP	-	-	BAM/BioWig
Abingdon island giant tortoise	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	TSV	MySQL	GVF	VCF	VEP	-	-	BAM/BioWig

Select DNA link for the Mouse:

To find the genome reference sequence for the mouse.

Mus_musculus.GRCm38.dna.chromosome.X.fa.gz 47.1 MB 11/20/19, 2:46:00 AM

Mus_musculus.GRCm38.dna.chromosome.Y.fa.gz 25.4 MB 11/20/19, 3:34:00 AM

Mus_musculus.GRCm38.dna.nonchromosomal.fa.gz 1.5 MB 11/20/19, 2:20:00 AM

Mus_musculus.GRCm38.dna.primary_assembly.fa.gz 769 MB 11/20/19, 3:51:00 AM

Mus_musculus.GRCm38.dna.toplevel.fa.gz 9, 3:39:00 AM

Mus_musculus.GRCm38.dna.alt.fa.gz 0, 3:25:00 AM

Open link in new tab

Open link in new window

Open link in incognito window

Save link as...

Copy link address

Inspect Ctrl+Shift+I

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.

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 3 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
Sample3.tsv	57 b	Auto-det...	Additional Sp...		
Sample2.tsv	54 b	Auto-det...	Additional Sp...		
New File	106 b	Auto-det...	Additional Sp...		

Download data from the web by entering URLs (one per line) or directly paste content.

ftp://ftp.ensembl.org/pub/release-99/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.primary_assembly.fa.gz

Type (set all): Auto-detect Genome (set all): Additional Sp...

Choose local file Choose FTP file **Paste/Fetch data** Pause Reset Start Close

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

RNA-Seq Analysis: Getting Reference Data

Select the Gene GTF link from the previous table
To find the gtf file with the gene definitions for mouse genome

Show	10 entries	Show/hide columns															Filter
★	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 <i>Homo sapiens</i>	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 <i>Mus musculus</i>	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 <i>Danio rerio</i>	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF	VCF	VEP	-	-	BAM/BigWig
	Abingdon island giant tortoise <i>Chelonoidis abingdonii</i>	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.

	Name	Size	Date Modified
	CHECKSUMS	221 B	12/4/19, 1:52:00 PM
	Mus_musculus.GRCm38.99.abinitio.gtf.gz	3.2 MB	11/23/19, 3:59:00 AM
	Mus_musculus.GRCm38.99.chr.gtf.gz	28.9 MB	11/23/19, 3:56:00 AM
	Mus_musculus.GRCm38.99.chr_patch.gtf.gz		11/23/19, 3:56:00 AM
	Mus_musculus.GRCm38.99.gtf.gz		11/23/19, 3:56:00 AM
	README		11/23/19, 3:57:00 AM

Open link in new tab
Open link in new window
Open link in incognito window

Save link as...
Copy link address

Inspect Ctrl+Shift+I

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 4 file(s) to the queue. Add more files or click 'Start' to proceed.

	SraRunTable.csv	2 KB	Auto-det...	Q	Additional Sp...	
	ContrastList.txt	13 b	Auto-det...	Q	Additional Sp...	
	New File	106 b	Auto-det...	Q	Additional Sp...	

Download data from the web by entering URLs (one per line) or directly paste content.

ftp://ftp.ensembl.org/pub/release-99/fasta/mus_musculus/dna/Mus_musculus.GRCm38.dna.primary_assembly.fa.gz

New File | 87 b | Auto-det... | Q | Additional Sp... | |

Download data from the web by entering URLs (one per line) or directly paste content.

ftp://ftp.ensembl.org/pub/release-99/gtf/mus_musculus/Mus_musculus.GRCm38.99.chr.gtf.gz

Type (set all): Auto-detect Q Genome (set all): Additional Sp...

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 the 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.

Select lines that match an expression (Galaxy Version 1.0.1) ☆ Favorite ▼ Options

Select lines from

5: Mus_musculus.GRCm38.99.chr.gtf.gz

that

NOT Matching

the pattern

(pseudogene|miRNA)

here you can enter text or regular expression (for syntax check id

Execute

Select the gtf file with the gene definition downloaded from ensembl

Change the option to not matching!

Indicate the expressions in the gtf file that should be filtered out: (pseudogene|miRNA)

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 screenshot shows the Gffread tool interface with several annotations pointing to specific fields:

- Select the gtf file with the gene definition downloaded from ensembl:** Points to the "Input GFF3 or GTF feature file" field, which contains "3: Mus_musculus.GRCm38.99 chr.gtf.gz".
- Select Reference Sequence from your history and place the fasta reference in the corresponding box:** Points to the "Reference Genome" field, which contains "4: Mus_musculus.GRCm38.dna.primary_assembly.fa.gz (as fasta)".
- Select to output a fasta with the spliced exons for each GFF transcript:** Points to the "Select fasta outputs" section, where the "fasta file with spliced exons for each GFF transcript (-w exons.fa)" option is selected.
- Select to output a GTF with the 2nd column changed to filteredGTF:** Points to the "Feature File Output" field, which contains "filteredGTF".
- Select full attribute preservation to keep all the descriptions of each entry in the GTF file:** Points to the "full GFF attribute preservation (all attributes are shown)" section, where the "Yes" option is selected.

The interface also includes sections for "filters", "Filter by genome region", "Filter out transcripts with large introns", "Replace reference sequence names", "Transcript merging", "reference based filters", and "Select fasta outputs".

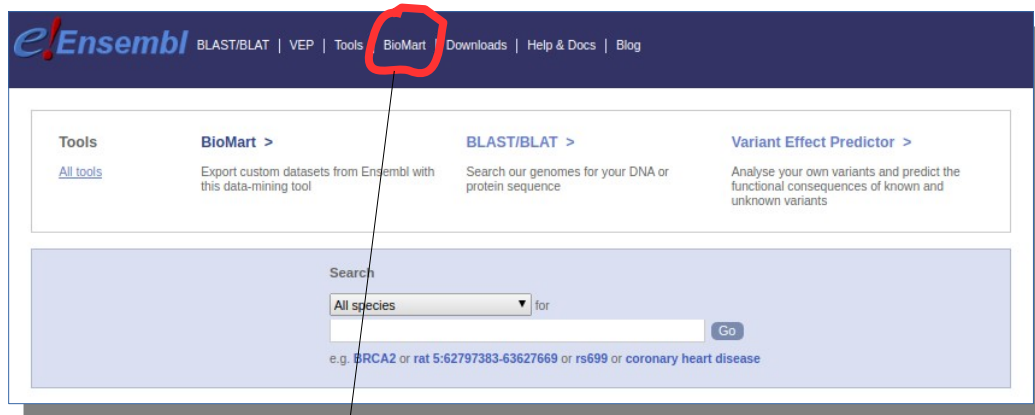
The tool will output two datasets:

- 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.

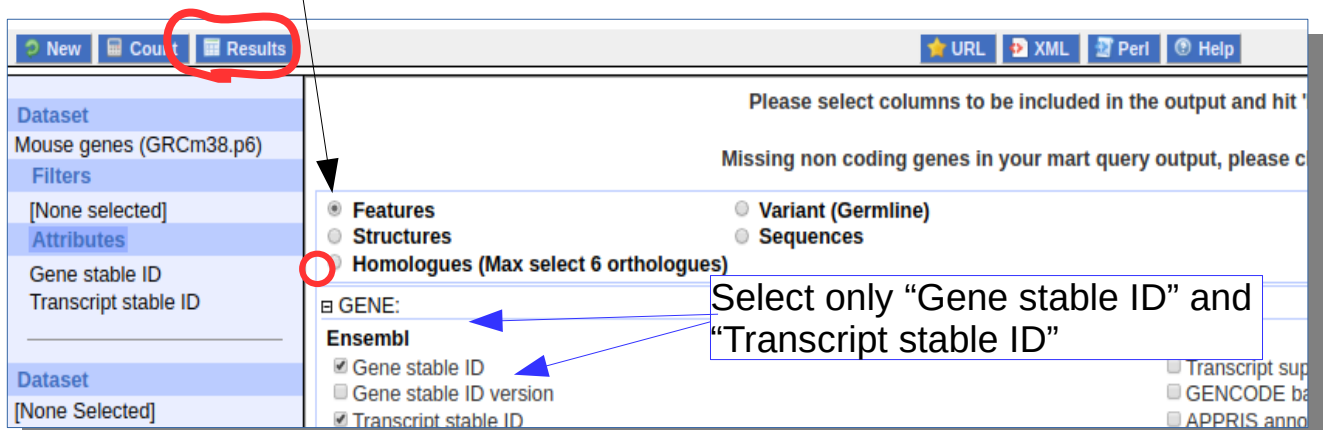
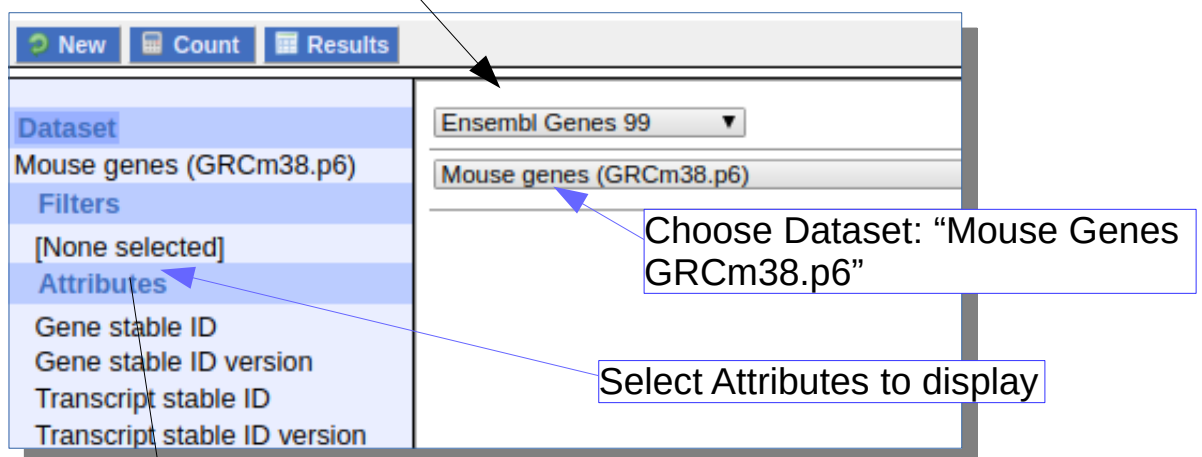
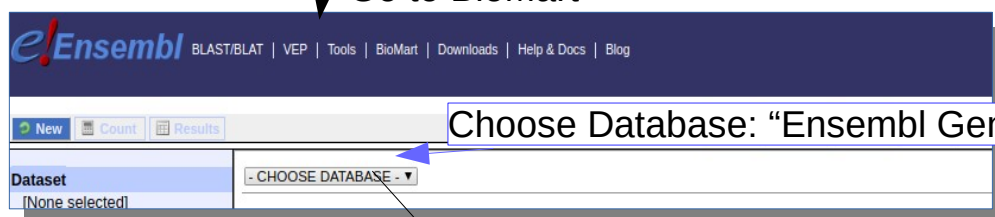
Transcript to Gene mapping file

- 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



Then click RESULTS

RNA-Seq Analysis: Prepare Reference

Export all results to

File: TSV ☒ Unique results only **Go**

Email notification to

View: 10 rows as HTML ☐ Unique results only

Gene stable ID	Transcript stable ID
ENSMUSG00000064372	ENSMUST000000082423
ENSMUSG00000064371	ENSMUST000000082422
ENSMUSG00000064370	ENSMUST000000082421
ENSMUSG00000064369	ENSMUST000000082420
ENSMUSG00000064368	ENSMUST000000082419
ENSMUSG00000064367	ENSMUST000000082418
ENSMUSG00000064366	ENSMUST000000082417
ENSMUSG00000064365	ENSMUST000000082416
ENSMUSG00000064364	ENSMUST000000082415
ENSMUSG00000064363	ENSMUST000000082414

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

Sort Column Order by heading (Galaxy Version 0.0.1)

Favorite Options

Tabular file: 39: TranscriptToGeneIDs

Identifier column: Column: 2

This column will be made left-most.

Execute

Now, we should have a file like this one

1	2
Transcript stable ID	Gene stable ID
ENSMUST000000082423	ENSMUSG000
ENSMUST000000082422	ENSMUSG000
ENSMUST000000082421	ENSMUSG000
ENSMUST000000082420	ENSMUSG000
ENSMUST000000082419	ENSMUSG000
ENSMUST000000082418	ENSMUSG000
ENSMUST000000082417	ENSMUSG000
ENSMUST000000082416	ENSMUSG000
ENSMUST000000082415	ENSMUSG000
ENSMUST000000082414	ENSMUSG000
ENSMUST000000084013	ENSMUSG000
ENSMUST000000082412	ENSMUSG000
ENSMUST000000082411	ENSMUSG000