RNA Seq Analysis Pipeline:

RNA-Seq Exercise:

We are going to analyse an RNA-Seq experiment of 6 mouse samples from two conditions. A WT condition and a Nrl KO condition in which some segments of the Nrl gene have been removed from the mouse genome.

We will use GRCm38 assembly and the latest Gene Build from Ensemble Genome Browser.

We will create a pipeline using the following tools:

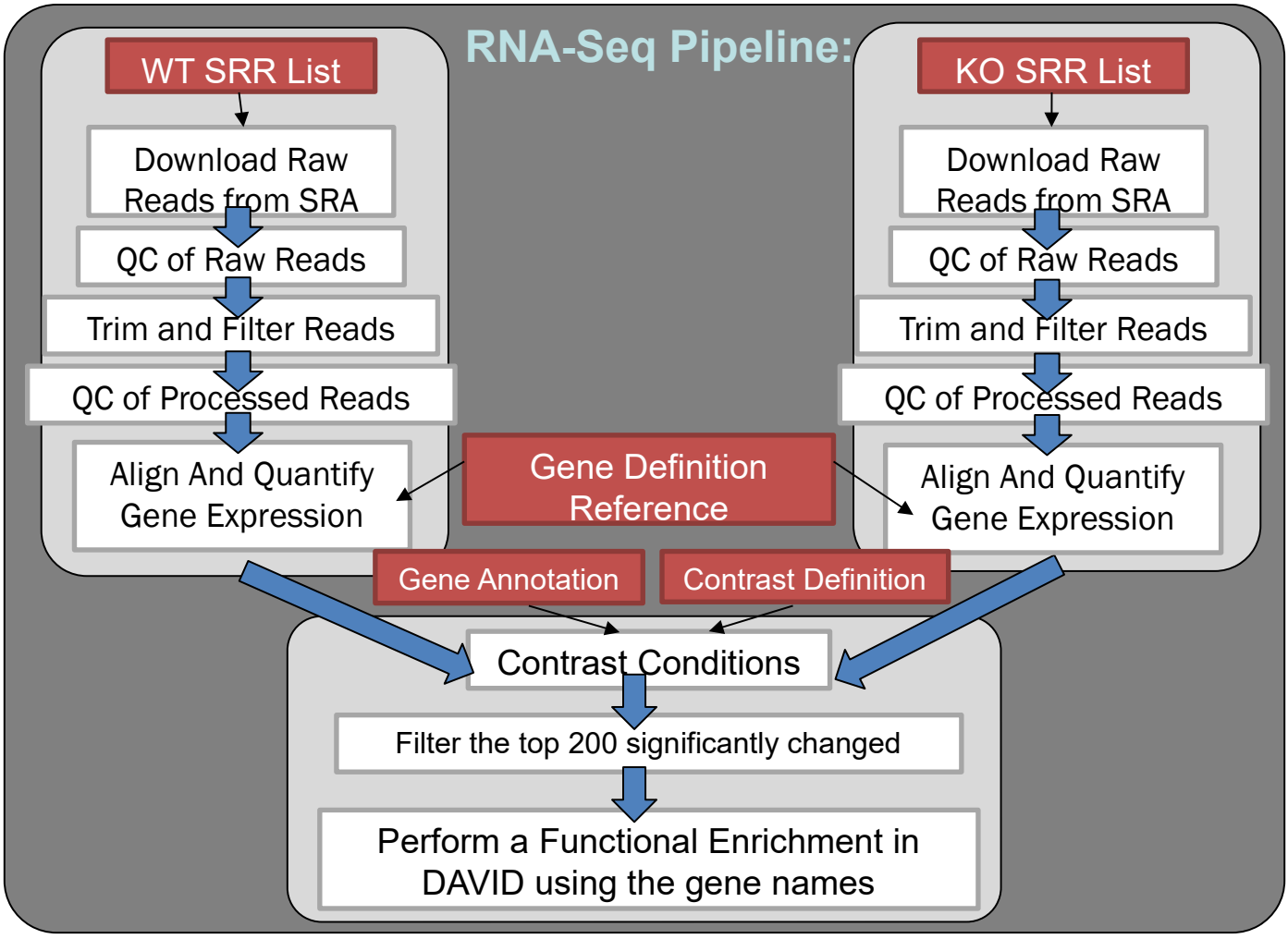
- **Text manipulation , Collection of Sorting tools:** to prepare the input files, reformat some intermediate files and output files.
- **Faster Download and Extract Reads in FASTQ:** tool to get reads from SRA
- **FastQC** for qc read quality
- **Trim-Galore** for trimming adaptors and filter low quality reads
- **Salmon** for alignment and Quantification of gene expresion
- **limma** to evaluate the differential expression between KO and WT conditions
- **DAVID** Tool for functional analysis.

This is the paper from which we will get the data

Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl- retinal transcriptomes. Brooks MJ, Rajasimha HK, Roger JE and Swarrop A. Molecular Vision 2011.

RNA Seq Analysis Pipeline:

This is a graphical representation of all the steps for this pipeline. We will split the the first steps of the analysis in two lines of processing, one for WT samples and one for KO samples, although it could be done in just one line.
The brown boxes are datasets that we need to provide.
The rest will be generated by the tools.



1) The first step will download the data to galaxy
2) The raw reads should go through a Quality Check to evaluate its quality.
3) After we will preprocess the the reads to remove adapter contaminations and reads of low sequencing quality.
4) Then a new QC will tell us how well the preprocessing improved the quality of the sequencing
5) The we can proceed with alignment and quantification and subsequent contrast and functional analysis.

RNA-Seq Analysis: Geting Experiment Data

Most of the publicly available NGS datasets can be found at GEO/SRA (USA NCBI database) or ENA (European EBI database).

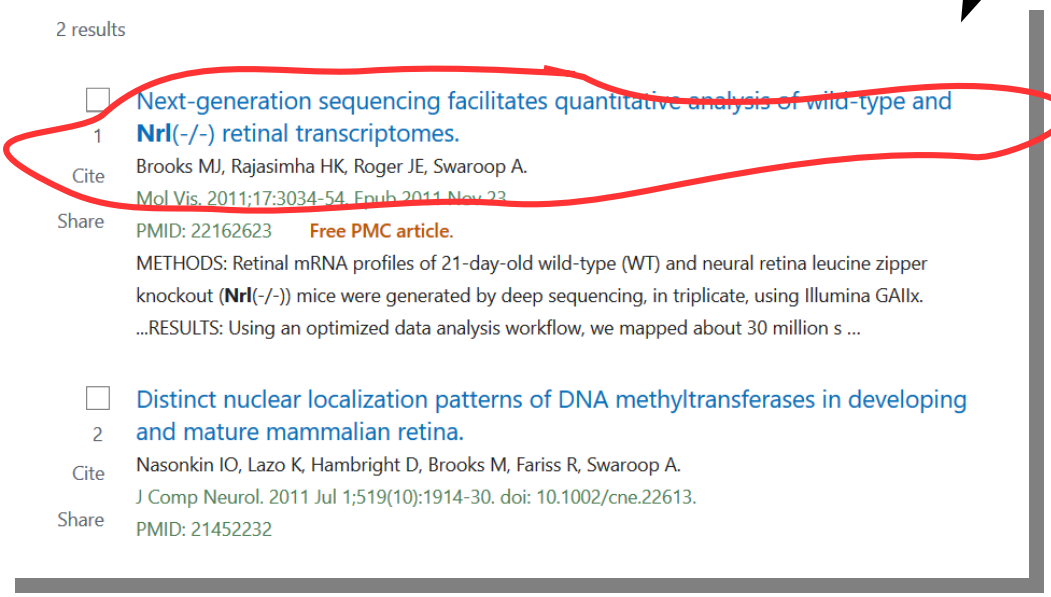
Most of the papers published now with NGS analysis will provide an accesion number like “GSE33141” or an ENA accesion number like “ERP123456”.

In our case, this paper is old and didn’t provide the number in the manuscript. However we can find it trough navigation across PubMed-GEO-SRA databases.

Got to PubMed:
https://www.ncbi.nlm.nih.gov/pubmed/

Search for this terms: Nrl AND Brooks AND 2011

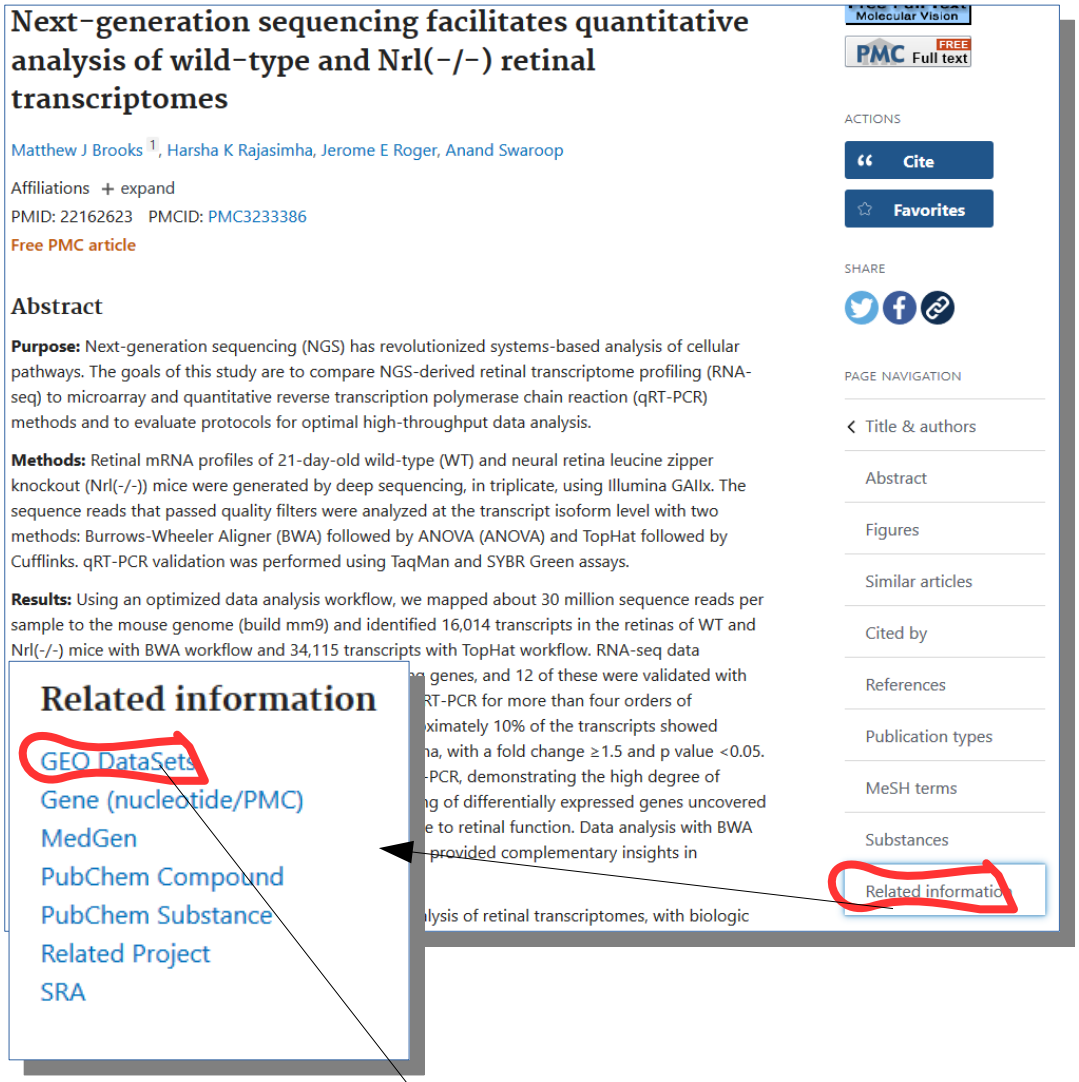
Select the entry by clicking the title:
There will be two resuls. Our paper will be the first one



RNA-Seq Analysis: Geting Experiment Data

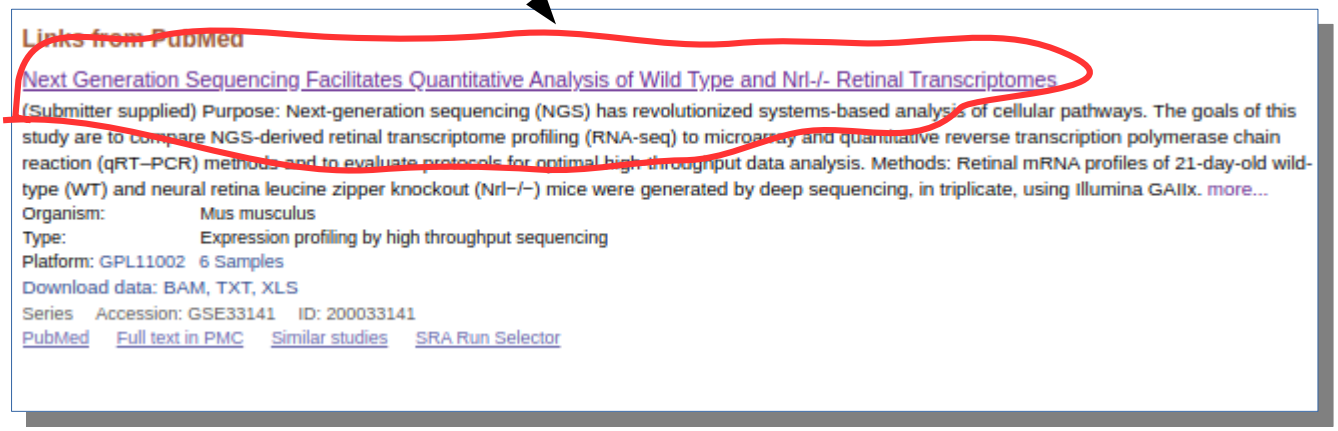
Select the GEO Datasets link:

On the Summary page for the paper look for a link to GEO DataSets tha will be present unther the related information section whenever a paper has some data in GEO Database.



Get into the GEO record details:

We have moved now to the Gene Expression Ommnibus Database (GEO), where the details of the experiment are stored. Select the title entry to look a it in detail.



RNA-Seq Analysis: Geting Experiment Data

Explore the Information about the Experiment:

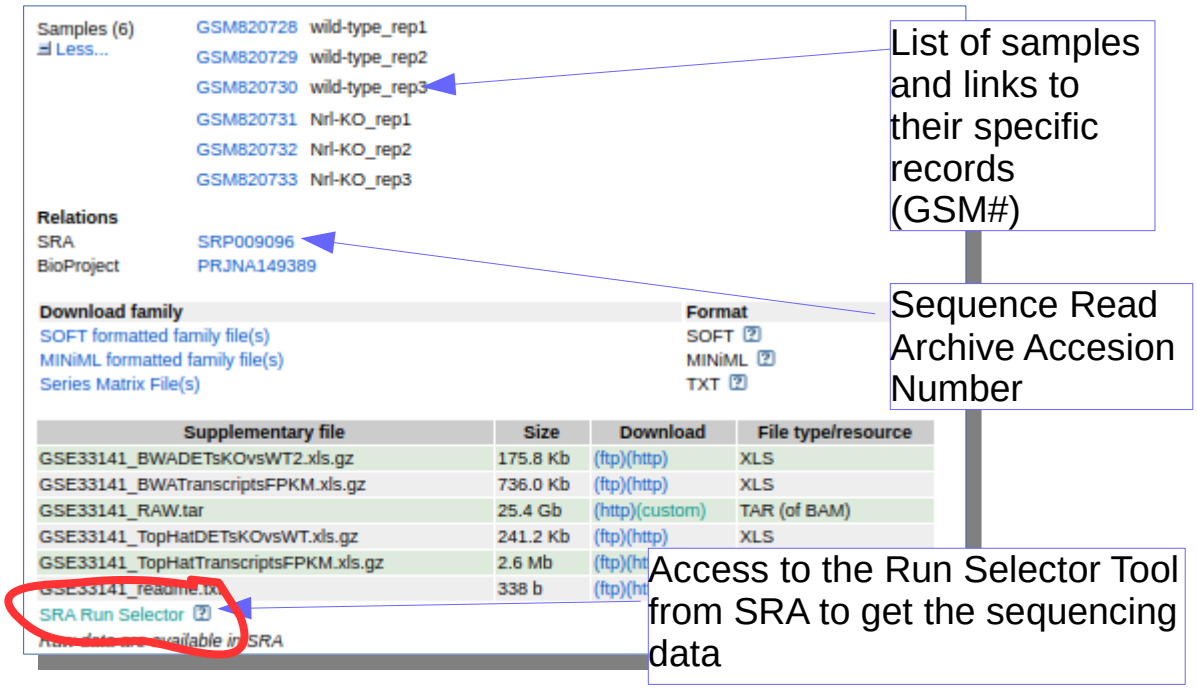
The record has detailed information of how the samples were processed, how they qere sequenced and analysed. There are also links to the records of the biological samples (GSM accesion number) and some datasets are provided with the results of their analysis.

Experiment accesion number (GSE#)

GSE33141

Select the SRA Run Selector Tool:

This link will get you to the Run Selector Tool from the Sequence Read Archive Database, were all the NGS sequences are stored. The tool will help you select and retrieve the sequencing reads



List of samples and links to their specific records (GSM#)

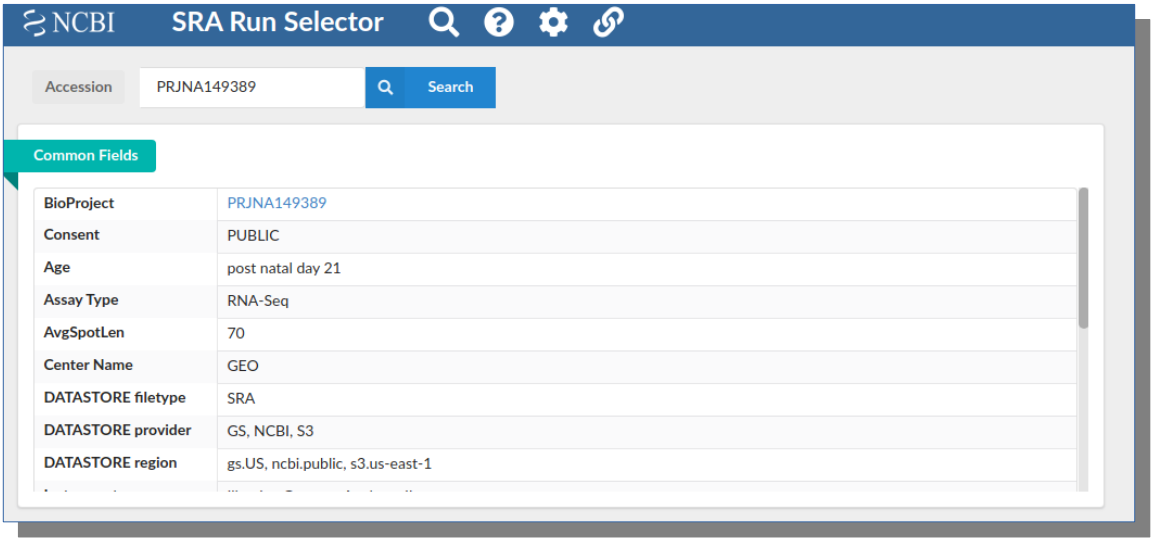
Sequence Read Archive Accesion Number

Access to the Run Selector Tool from SRA to get the sequencing data

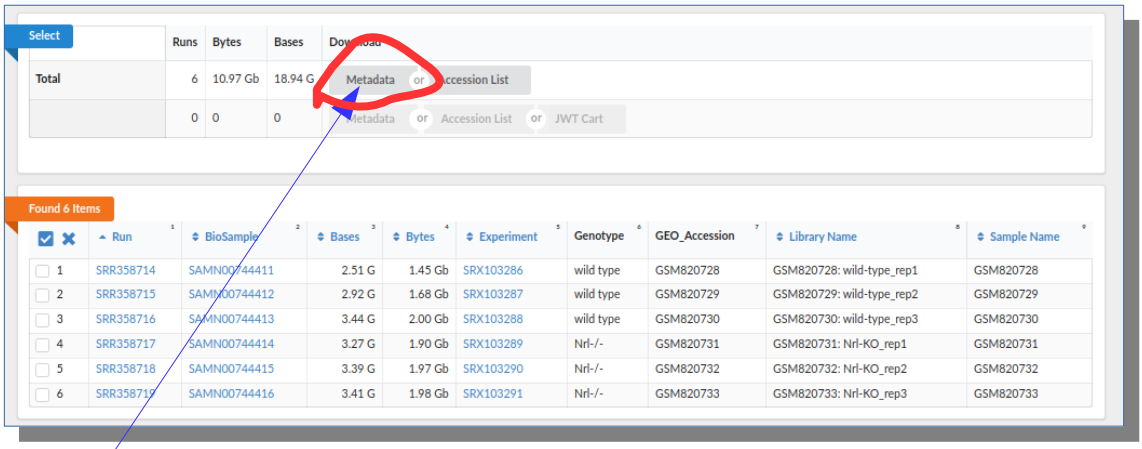
RNA-Seq Analysis: Geting Experiment Data

SRA Databe Run Selector Tool:

- Displays the metadata related to the sequences of a sequencing experiment (PRJNA149389 in our case).
- The first section diplays all fields that have the same values for all samples in the experiment.



- The second section displays all fields that contain some differences between the samples.



Download the table of sample metadata:

- Select the option Metadata to download the table in comma separated format.
- Explore the table using Excel and modify or remove all those columns with problematic symbols on it like -, +, =, ?, !, /, \, “,”, “.” or spaces. I.e: The field with Genotype information, change -/- by “KO” or “mutant” word.
- The same with the spaces like “wild type”. You could replace the space with “_”: “wild_type” or “WT” directly
- From the sample library column remove the GSM id, the colon, the spaces and the “-”. (In example: “GSM820728: wild-type_rep1” to “wildtype_rep1”)
- Keep only the columns Run, Library Name and Genotyp.e
- Export the table in comma or tab separated text file.

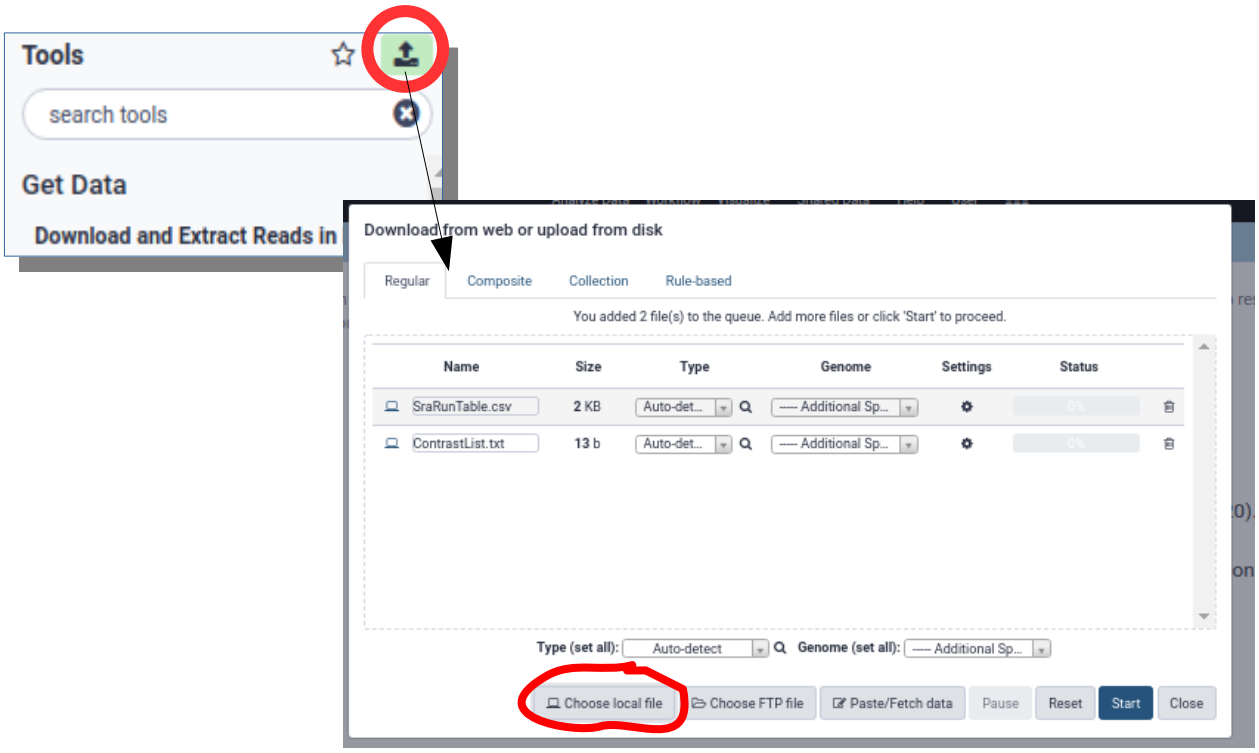
RNA-Seq Analysis: Geting Experiment Data

Prepare a File with Contrast Definitions:

This will be a simple txt file with the contrasts we want to perform.
In our case:
“NrlKO-WT”

If we had more conditions, we could add more contrasts, one per line:
“NrlKO-WT”
“NrlKO_Treated-WT_Treated”
...

Put those files into the Upload Tool From Galaxy:



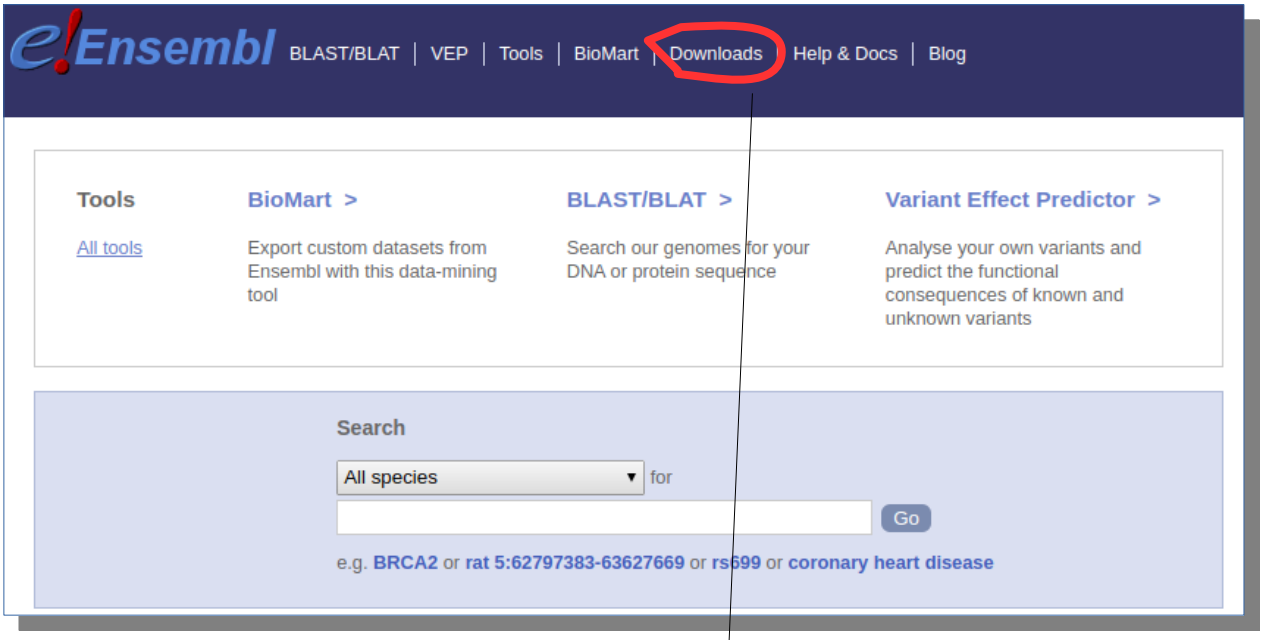
There is no need to select “Start” right now. We can prepare all the files to upload and external links to get data from and the click “Start” for all together.

RNA-Seq Analysis: Geting Data

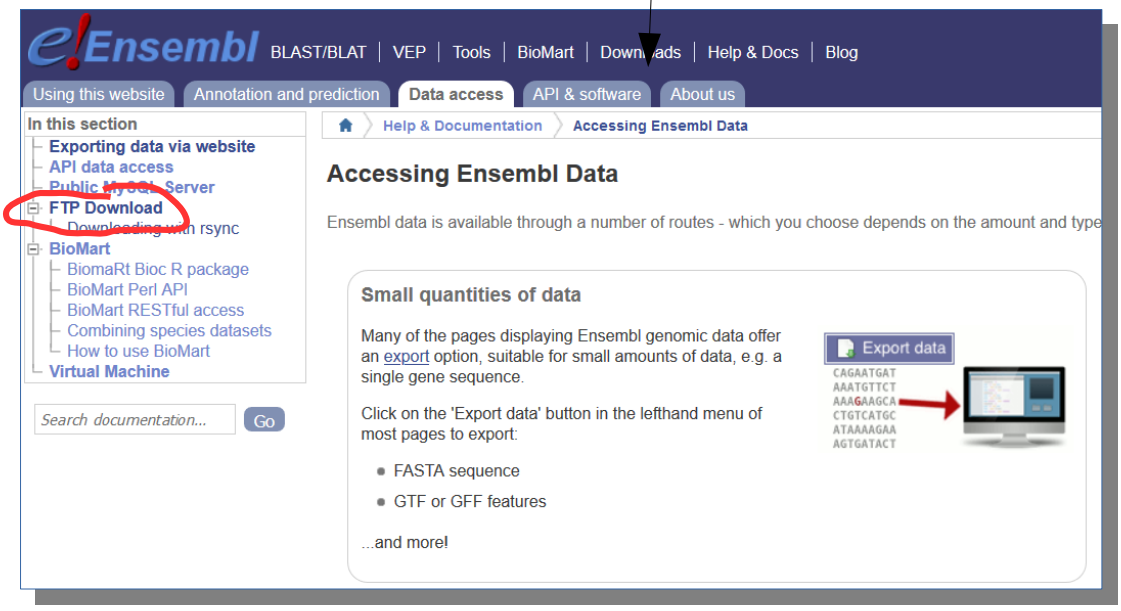
Collect the FTP link to get the Gene and Genome references:

- 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



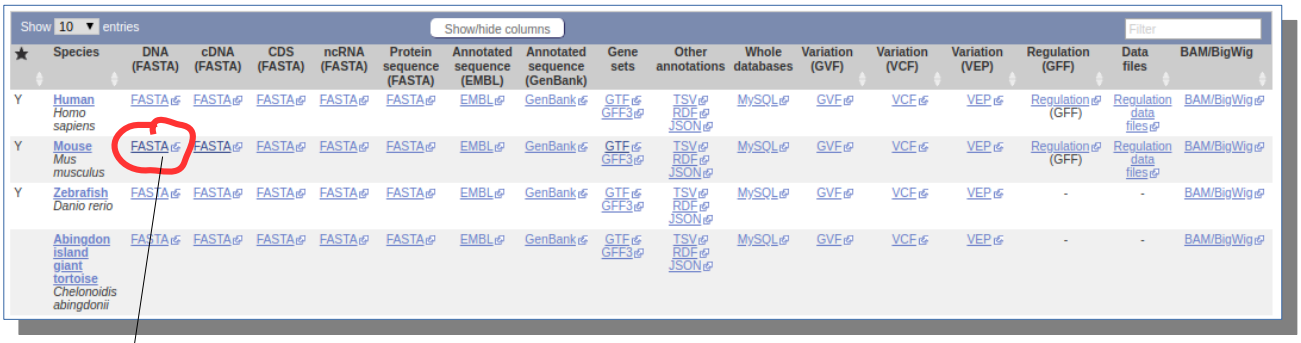
Go to Download Data Via FTP

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

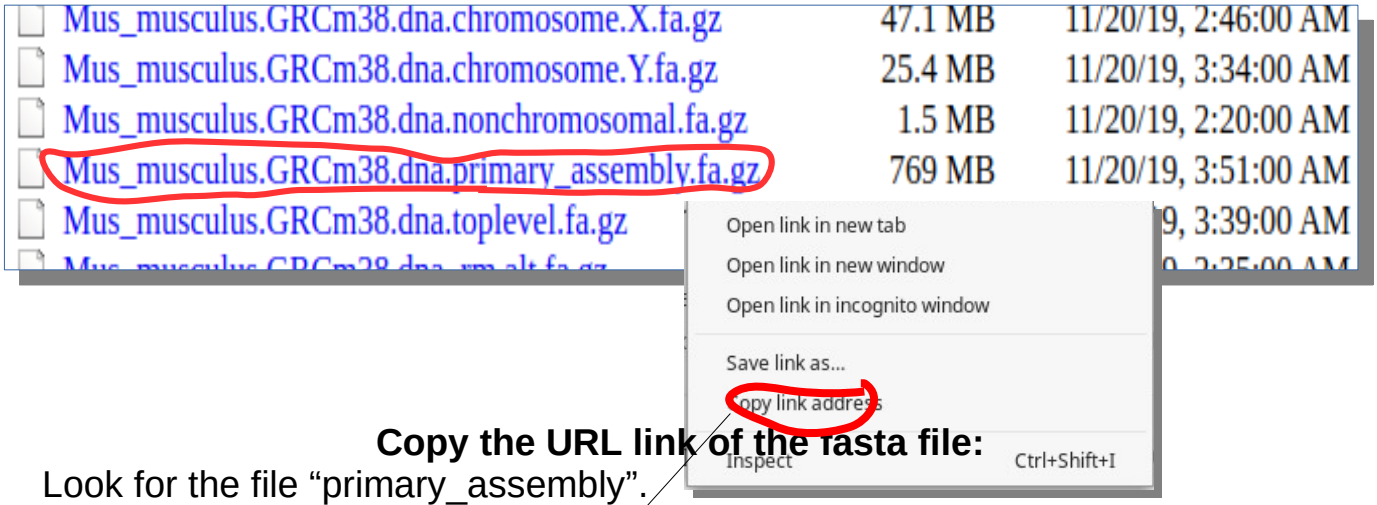
...



★	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

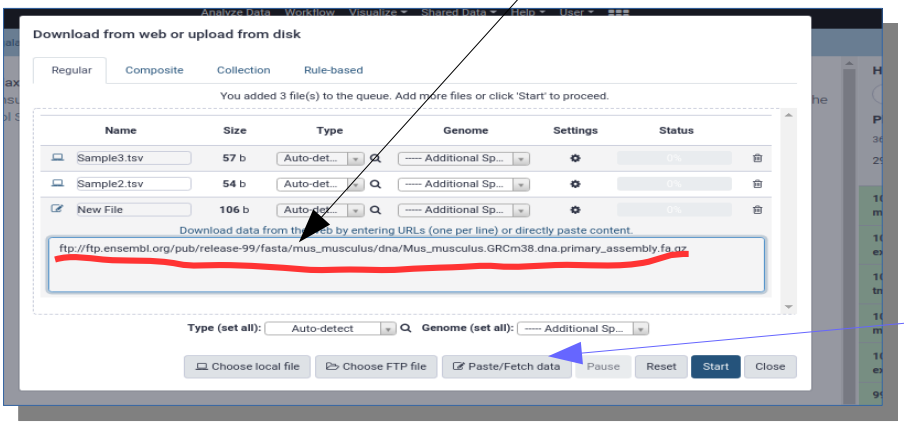
Select DNA link for the Mouse:

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.

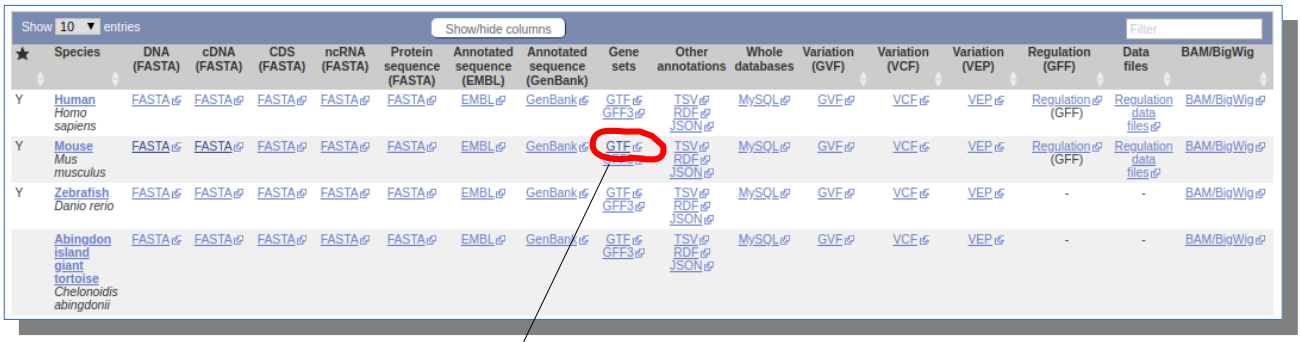


Use this option (Paste/Fetch) to transfer data from another server using an URL link.

RNA-Seq Analysis: Geting Reference Data

Select the Gene GTF link from the previous table

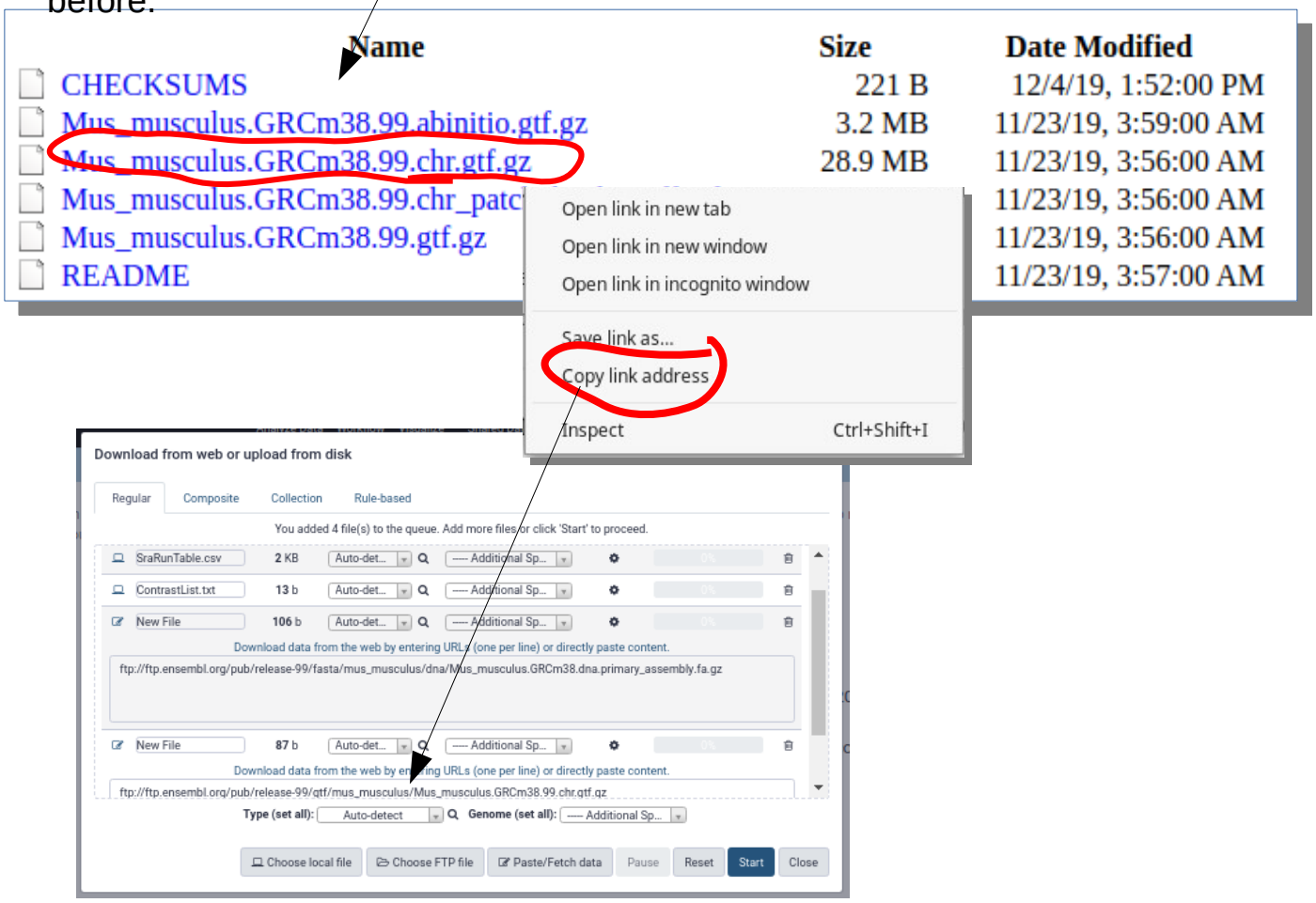
To find the gtf file with the gene definitions for mouse genome



★	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	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.

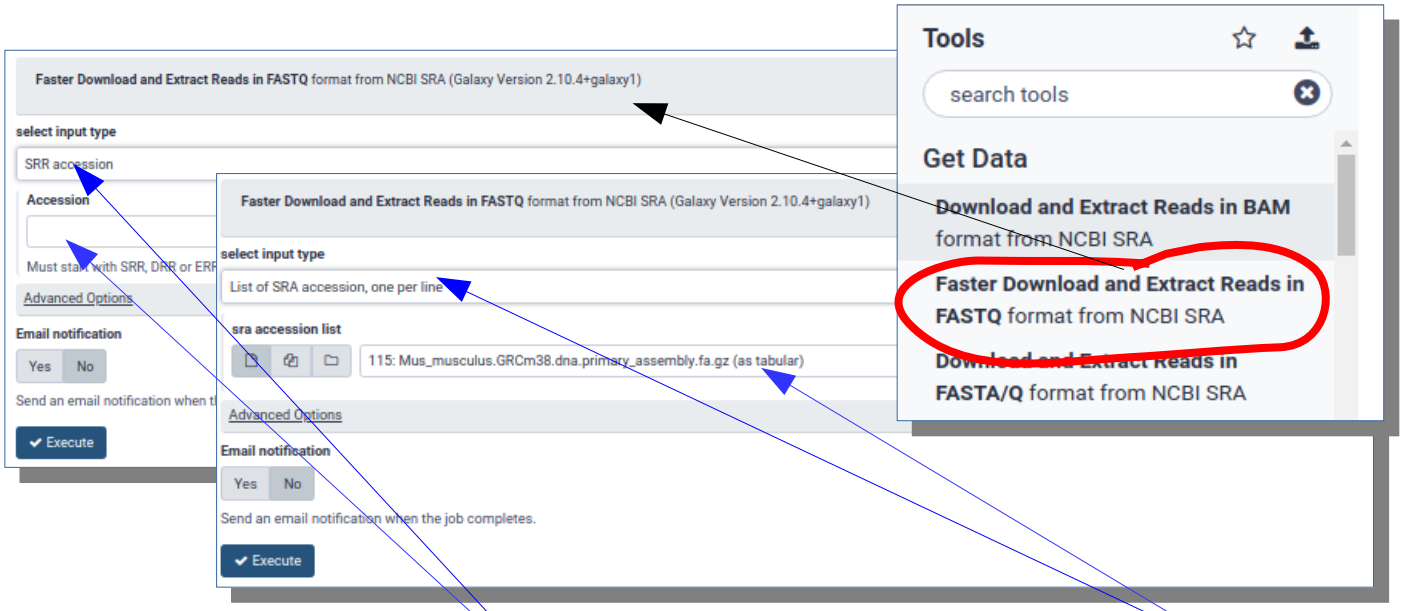


Now, you can click the Start button and Galaxy will collect the four datasets (the two files from your computer and the references from Ensembl ftp links).

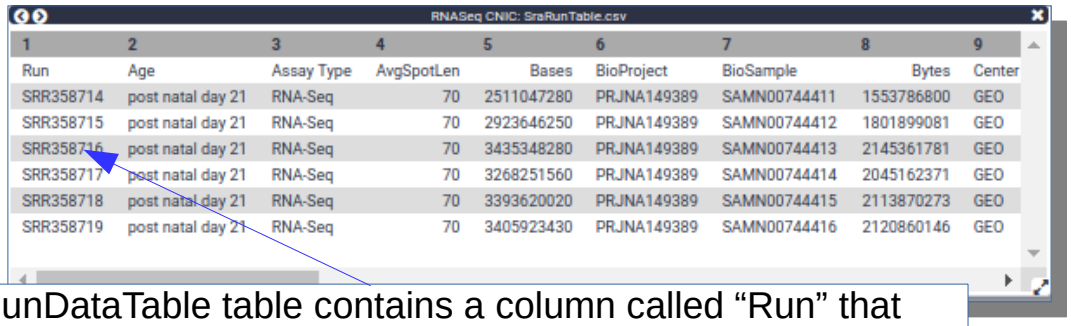
RNA-Seq Analysis: Getting Sequence Data

We will use the Tool “**Faster Download and Extract Reads in FASTQ format**”

- Select the tool and explore the form of the tool to see what it needs.



- The tool asks for a SRR accession number or a list of SRR numbers in a text file if we change the “input type” menu.



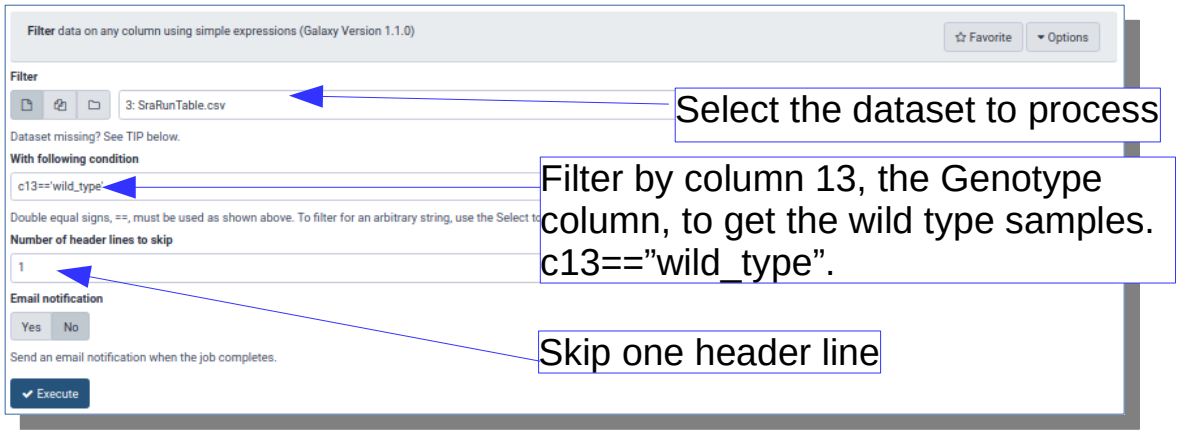
RNASeq CNIC: SraRunTable.csv

1	2	3	4	5	6	7	8	9
Run	Age	Assay Type	AvgSpotLen	Bases	BioProject	BioSample	Bytes	Center
SRR358714	post natal day 21	RNA-Seq	70	2511047280	PRJNA149389	SAMN00744411	1553786800	GEO
SRR358715	post natal day 21	RNA-Seq	70	2923646250	PRJNA149389	SAMN00744412	1801899081	GEO
SRR358716	post natal day 21	RNA-Seq	70	3435348280	PRJNA149389	SAMN00744413	2145361781	GEO
SRR358717	post natal day 21	RNA-Seq	70	3268251560	PRJNA149389	SAMN00744414	2045162371	GEO
SRR358718	post natal day 21	RNA-Seq	70	3393620020	PRJNA149389	SAMN00744415	2113870273	GEO
SRR358719	post natal day 21	RNA-Seq	70	3405923430	PRJNA149389	SAMN00744416	2120860146	GEO

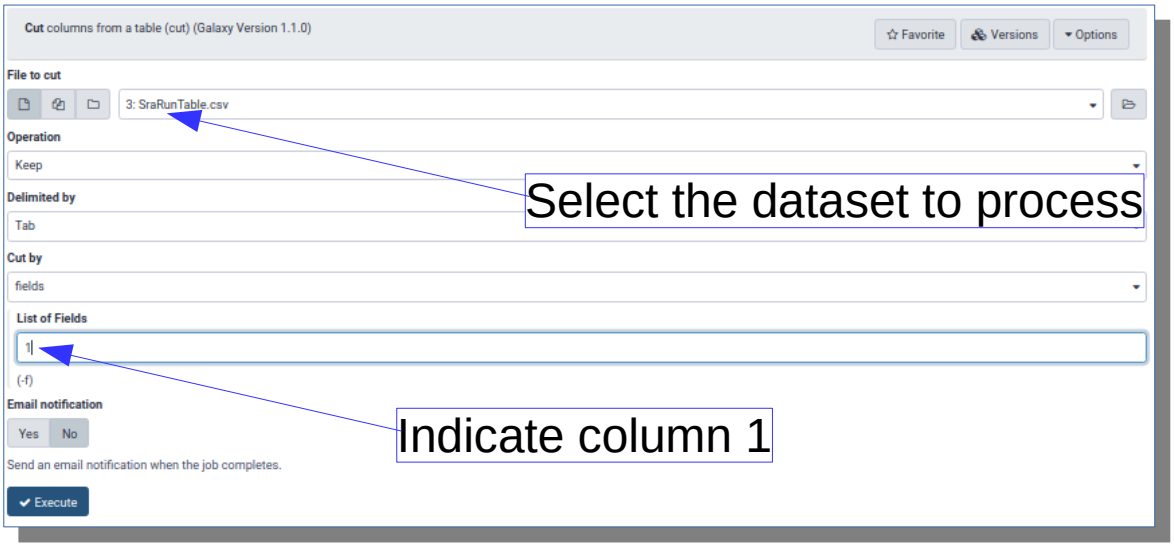
- Our RunDataTable table contains a column called “Run” that contains SRR accesion numbers.
- But first we are going to split our table in two datasets, one for the WTs and one for the KOs
- We can build a file with just that column by using some Text Manipulation tools.

RNA-Seq Analysis: Geting Sequence Data

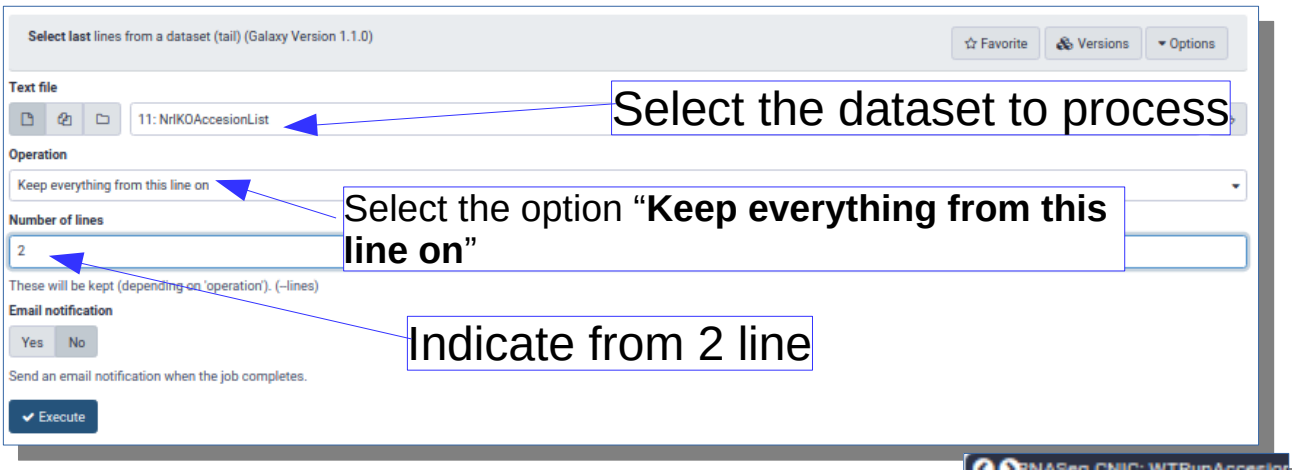
Filter: The SRADataTable file to extract a dataset for WT samples and another for KO samples



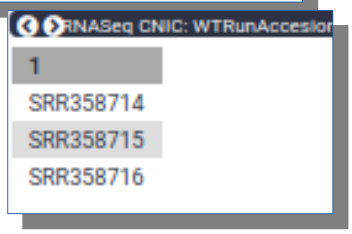
Cut: cut the 1st column



Select Last (tail): Select the lines from the 2nd onwards to remove the header



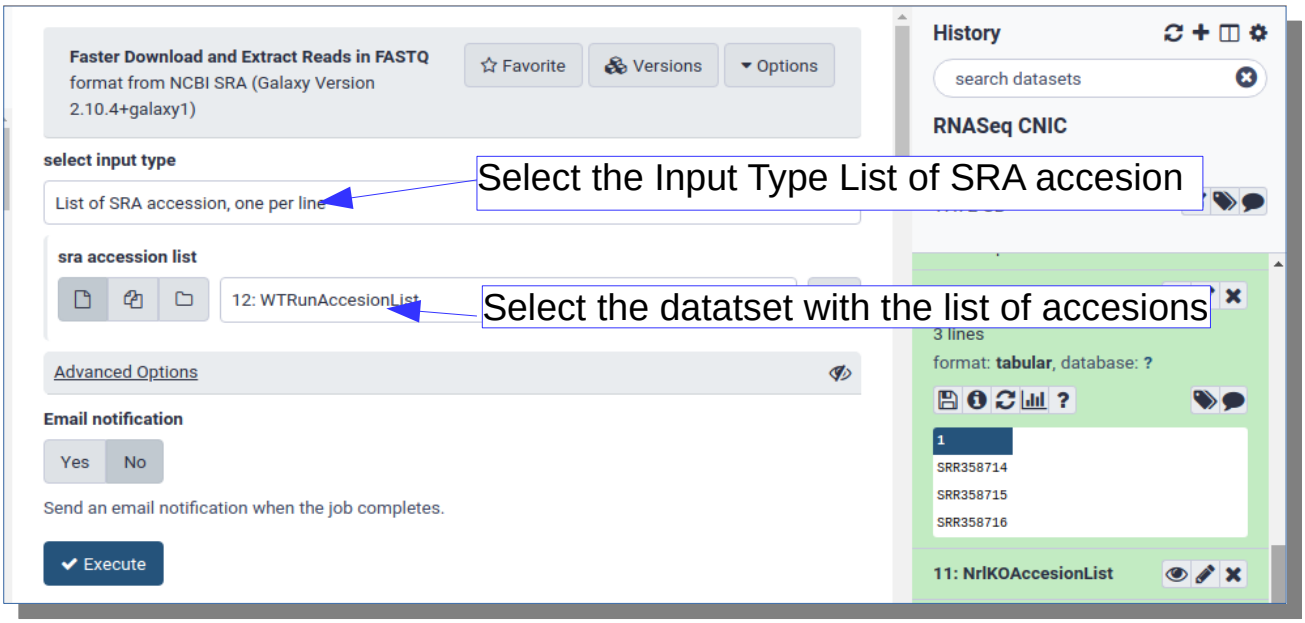
Now we have a file with a list of SRR accesion numbers



Repeat the same **filter, cut and Select Last** steps for the KO samples

RNA-Seq Analysis: Geting Sequence Data

Now we can feed the list of SRR accesion numbers to the **Faster Download and Extract Reads in FASTQ** and **Execute**



Repeat the Faster Download Step for the WT and KO SRA Accesion list datasets

This tool will generate several dataset collections:

- **Single-End data**: Will contain the reads in Fastq format if the samples downloaded are single-end sequences.
- **Pair-End data:** Will contain the reads if in Fastq format if the sequencing was done in pair-end mode.
- **Other data:** Will contain other relevant data for the experiment. Usually empty
- **Log:** will contain log information about the process of extracting the Fastq data.