

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.

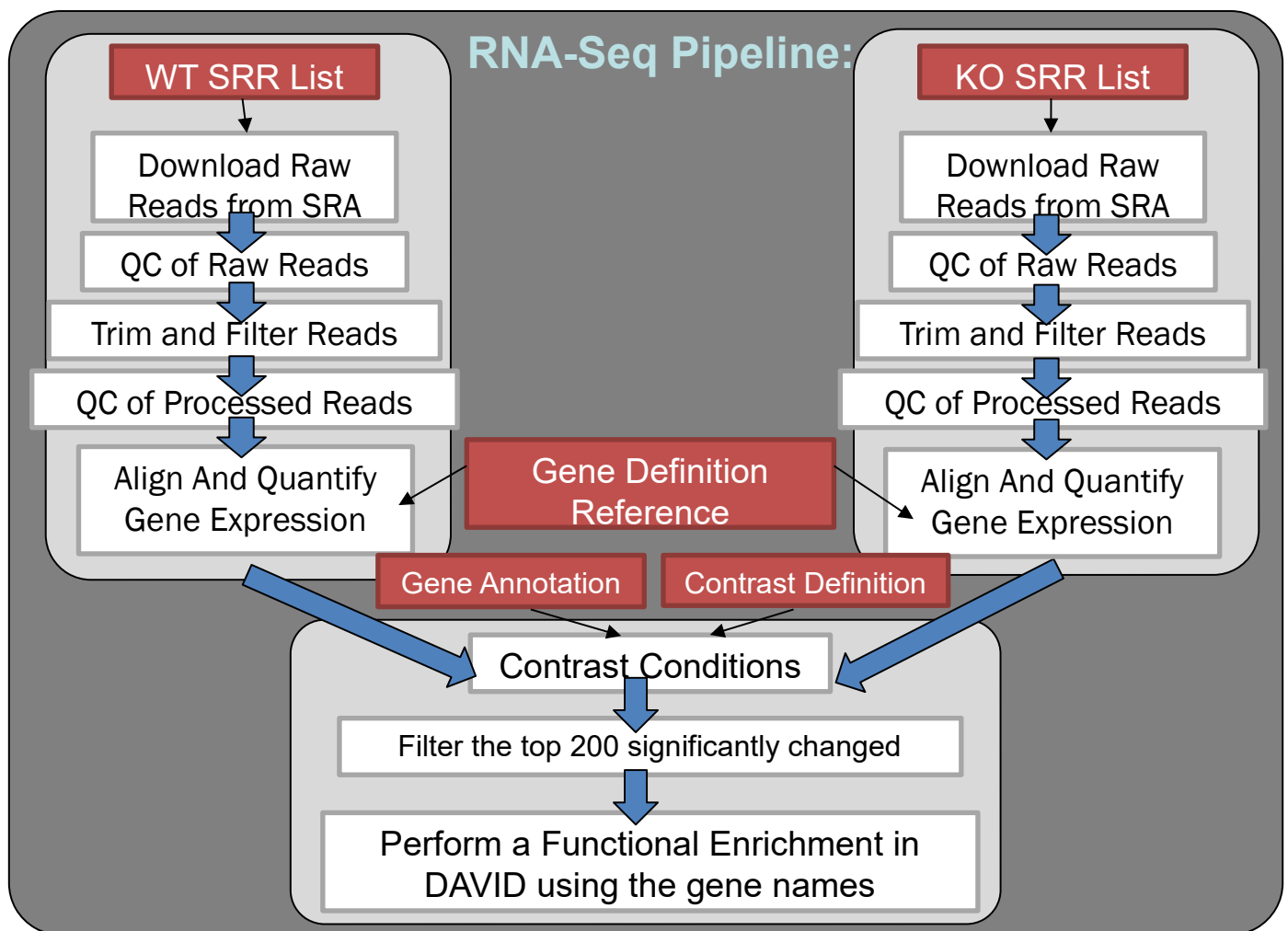
This is the paper from which we will get the data

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: Getting Experiment Data

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

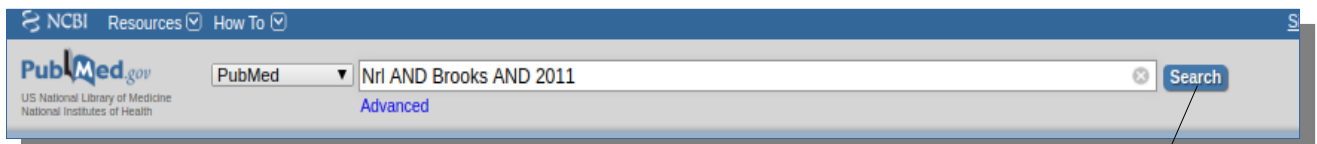
Most of the papers published now with NGS analysis will provide an accession number like "GSE33141" or an ENA accession number like "ERP123456".

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

Got to PubMed:

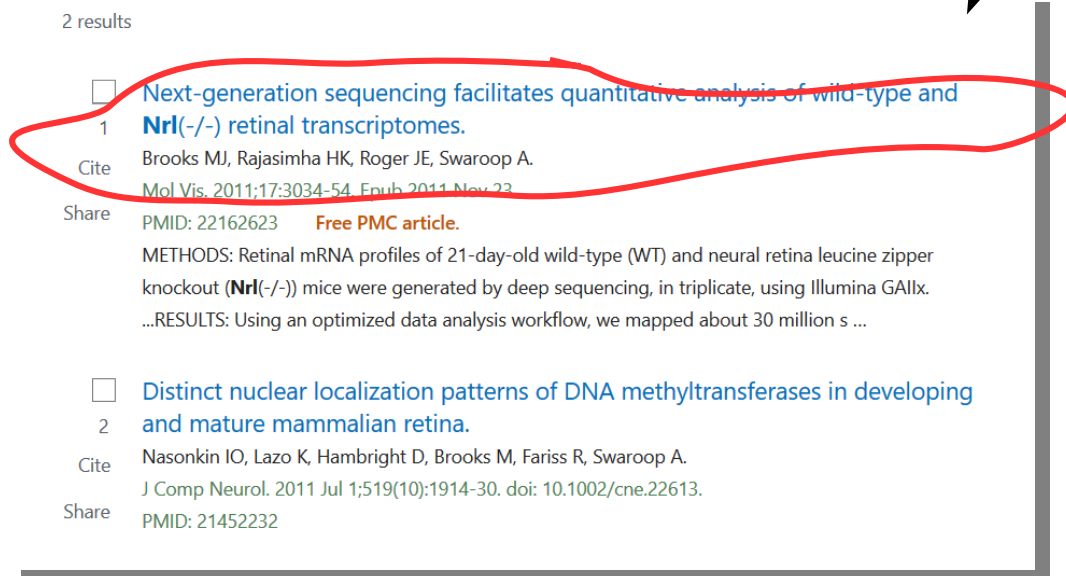
<https://www.ncbi.nlm.nih.gov/pubmed/>

Search for this terms: Nrl AND Brooks AND 2011



Select the entry by clicking the title:

There will be two results. Our paper will be the first one



RNA-Seq Analysis: Getting Experiment Data

Select the GEO Datasets link:

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

Next-generation sequencing facilitates quantitative analysis of wild-type and Nrl(-/-) retinal transcriptomes

Matthew J Brooks¹, Harsha K Rajasimha, Jerome E Roger, Anand Swaroop

Affiliations + expand
PMID: 22162623 PMCID: PMC3233386
Free PMC article

Abstract

Purpose: Next-generation sequencing (NGS) has revolutionized systems-based analysis of cellular pathways. The goals of this study are to compare NGS-derived retinal transcriptome profiling (RNA-seq) to microarray and quantitative reverse transcription polymerase chain reaction (qRT-PCR) methods and to evaluate protocols for optimal high-throughput data analysis.

Methods: Retinal mRNA profiles of 21-day-old wild-type (WT) and neural retina leucine zipper knockout (Nrl(-/-)) mice were generated by deep sequencing, in triplicate, using Illumina GAIIx. The sequence reads that passed quality filters were analyzed at the transcript isoform level with two methods: Burrows-Wheeler Aligner (BWA) followed by ANOVA (ANOVA) and TopHat followed by Cufflinks. qRT-PCR validation was performed using TaqMan and SYBR Green assays.

Results: Using an optimized data analysis workflow, we mapped about 30 million sequence reads per sample to the mouse genome (build mm9) and identified 16,014 transcripts in the retinas of WT and Nrl(-/-) mice with BWA workflow and 34,115 transcripts with TopHat workflow. RNA-seq data identified 1,200 genes, and 12 of these were validated with qRT-PCR for more than four orders of magnitude. Approximately 10% of the transcripts showed differential expression, with a fold change ≥ 1.5 and p value < 0.05 . qRT-PCR, demonstrating the high degree of agreement of differentially expressed genes uncovered by RNA-seq, provided complementary insights into the role of Nrl in the development and function of the retina. Data analysis with BWA and TopHat provided complementary insights into the role of Nrl in the development and function of the retina. Data analysis with BWA and TopHat provided complementary insights into the role of Nrl in the development and function of the retina.

Related information

- [GEO DataSets](#)
- [Gene \(nucleotide/PMC\)](#)
- [MedGen](#)
- [PubChem Compound](#)
- [PubChem Substance](#)
- [Related Project](#)
- [SRA](#)

ACTIONS

- Cite
- Favorites

SHARE

- Twitter
- Facebook
- Link

PAGE NAVIGATION

- Title & authors
- Abstract
- Figures
- Similar articles
- Cited by
- References
- Publication types
- MeSH terms
- Substances
- Related information

Get into the GEO record details:

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

Links from PubMed

[Next Generation Sequencing Facilitates Quantitative Analysis of Wild Type and Nrl-/- Retinal Transcriptomes](#)

(Submitter supplied) Purpose: Next-generation sequencing (NGS) has revolutionized systems-based analysis of cellular pathways. The goals of this study are to compare NGS-derived retinal transcriptome profiling (RNA-seq) to microarray and quantitative reverse transcription polymerase chain reaction (qRT-PCR) methods and to evaluate protocols for optimal high-throughput data analysis. Methods: Retinal mRNA profiles of 21-day-old wild-type (WT) and neural retina leucine zipper knockout (Nrl(-/-)) mice were generated by deep sequencing, in triplicate, using Illumina GAIIx. more...

Organism: Mus musculus
Type: Expression profiling by high throughput sequencing
Platform: GPL11002 6 Samples
Download data: BAM, TXT, XLS
Series Accession: GSE33141 ID: 200033141
[PubMed](#) [Full text in PMC](#) [Similar studies](#) [SRA Run Selector](#)

RNA-Seq Analysis: Getting Experiment Data

Explore the Information about the Experiment:

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

The screenshot shows the NCBI GEO Accession Display page for GSE33141. The page includes a header with the NCBI logo and a COVID-19 notice. The main content area displays the experiment details for GSE33141, including the title, organism, experiment type, and summary. The summary describes the purpose and methods of the study, which involved next-generation sequencing (NGS) of retinal transcriptomes in wild-type and Nrl-/- mice. The results section mentions that approximately 30 million sequence reads per sample were mapped to the mouse genome, identifying 16,014 transcripts.

NCBI > GEO > Accession Display

Not logged in | Login

GEO help: Mouse over screen elements for information.

Scope: Format: Amount: GEO accession:

Series GSE33141 [Query DataSets for GSE33141](#)

Status: Public on Oct 25, 2011

Title: Next Generation Sequencing Facilitates Quantitative Analysis of Wild Type and Nrl-/- Retinal Transcriptomes

Organism: [Mus musculus](#)

Experiment type: Expression profiling by high throughput sequencing

Summary: Purpose: Next-generation sequencing (NGS) has revolutionized systems-based analysis of cellular pathways. The goals of this study are to compare NGS-derived retinal transcriptome profiling (RNA-seq) to microarray and quantitative reverse transcription polymerase chain reaction (qRT-PCR) methods and to evaluate protocols for optimal high-throughput data analysis.

Methods: Retinal mRNA profiles of 21-day-old wild-type (WT) and neural retina leucine zipper knockout (Nrl-/-) mice were generated by deep sequencing, in triplicate, using Illumina GAIIX. The sequence reads that passed quality filters were analyzed at the transcript isoform level with two methods: Burrows-Wheeler Aligner (BWA) followed by ANOVA (ANOVA) and TopHat followed by Cufflinks. qRT-PCR validation was performed using TaqMan and SYBR Green assays.

Results: Using an optimized data analysis workflow, we mapped about 30 million sequence reads per sample to the mouse genome (build mm9) and identified 16,014 transcripts in the

Experiment accession 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, where all the NGS sequences are stored. The tool will help you select and retrieve the sequencing reads

The screenshot shows the SRA Run Selector tool interface. It displays a list of samples (GSM820728 to GSM820733) and their corresponding SRA and BioProject accession numbers. Below the list, there are links to download the family file in SOFT, MINIML, or Series Matrix File format. A table of supplementary files is also shown, including GSE33141_BWADETsKOvsWT2.xls.gz, GSE33141_BWATranscriptsFPKM.xls.gz, GSE33141_RAW.tar, GSE33141_TopHatDETsKOvsWT.xls.gz, and GSE33141_TopHatTranscriptsFPKM.xls.gz. The SRA Run Selector link is highlighted with a red circle.

Samples (6) [Less...](#)

- GSM820728 wild-type_rep1
- GSM820729 wild-type_rep2
- GSM820730 wild-type_rep3
- GSM820731 Nrl-KO_rep1
- GSM820732 Nrl-KO_rep2
- GSM820733 Nrl-KO_rep3

Relations

SRA: [SRP009096](#)

BioProject: [PRJNA149389](#)

Download family

SOFT formatted family file(s) [Format](#) [?](#)

MINIML formatted family file(s) [Format](#) [?](#)

Series Matrix File(s) [Format](#) [?](#)

Supplementary file	Size	Download	File type/resource
GSE33141_BWADETsKOvsWT2.xls.gz	175.8 Kb	(ftp) (http)	XLS
GSE33141_BWATranscriptsFPKM.xls.gz	736.0 Kb	(ftp) (http)	XLS
GSE33141_RAW.tar	25.4 Gb	(http) (custom)	TAR (of BAM)
GSE33141_TopHatDETsKOvsWT.xls.gz	241.2 Kb	(ftp) (http)	XLS
GSE33141_TopHatTranscriptsFPKM.xls.gz	2.6 Mb	(ftp) (http)	XLS
GSE33141_readme.txt	338 b	(ftp) (http)	XLS

[SRA Run Selector](#) [?](#)

[Raw data available in SRA](#)

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

Sequence Read Archive Accession Number

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

RNA-Seq Analysis: Getting Experiment Data

SRA Database Run Selector Tool:

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

The screenshot shows the SRA Run Selector interface. At the top, the NCBI logo and 'SRA Run Selector' title are visible. Below the title, there is a search bar with 'PRJNA149389' entered and a 'Search' button. A 'Common Fields' section is expanded, showing a table of metadata fields that are consistent across all samples in the experiment.

Field	Value
BioProject	PRJNA149389
Consent	PUBLIC
Age	post natal day 21
Assay Type	RNA-Seq
AvgSpotLen	70
Center Name	GEO
DATASTORE filetype	SRA
DATASTORE provider	GS, NCBI, S3
DATASTORE region	gs.US, ncbi.public, s3.us-east-1

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

The screenshot shows the 'Found 6 Items' section of the SRA Run Selector. A table lists 6 items with columns for Run, BioSample, Bases, Bytes, Experiment, Genotype, GEO_Accession, Library Name, and Sample Name. A red circle highlights the 'Download' button in the top right corner of the table, with a blue arrow pointing to it from the text 'Download the table of sample metadata:' below.

	Run	BioSample	Bases	Bytes	Experiment	Genotype	GEO_Accession	Library Name	Sample Name
1	SRR358714	SAMN00744411	2.51 G	1.45 Gb	SRX103286	wild type	GSM820728	GSM820728: wild-type_rep1	GSM820728
2	SRR358715	SAMN00744412	2.92 G	1.68 Gb	SRX103287	wild type	GSM820729	GSM820729: wild-type_rep2	GSM820729
3	SRR358716	SAMN00744413	3.44 G	2.00 Gb	SRX103288	wild type	GSM820730	GSM820730: wild-type_rep3	GSM820730
4	SRR358717	SAMN00744414	3.27 G	1.90 Gb	SRX103289	Nrl-/-	GSM820731	GSM820731: Nrl-KO_rep1	GSM820731
5	SRR358718	SAMN00744415	3.39 G	1.97 Gb	SRX103290	Nrl-/-	GSM820732	GSM820732: Nrl-KO_rep2	GSM820732
6	SRR358719	SAMN00744416	3.41 G	1.98 Gb	SRX103291	Nrl-/-	GSM820733	GSM820733: Nrl-KO_rep3	GSM820733

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 Genotype.e
- Export the table in comma or tab separated text file.

RNA-Seq Analysis: Getting 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:

“Nr1KO-WT”

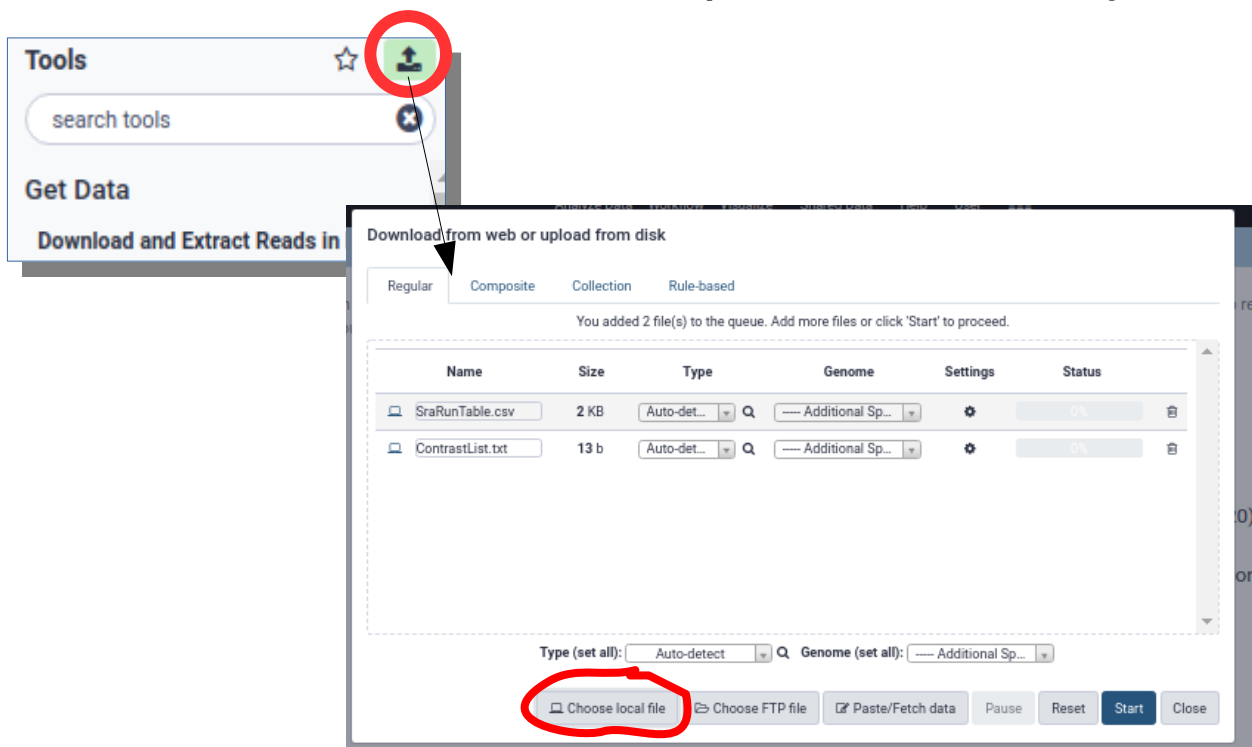
If we had more conditions, we could add more contrasts, one per line:

“Nr1KO-WT”

“Nr1KO_Treated-WT_Treated”

...

Put those files into the Upload Tool From Galaxy:



The screenshot shows the Galaxy web interface. In the top left, there is a 'Tools' sidebar with a search bar and a 'Get Data' section. A red circle highlights the upload icon (a green square with a white arrow) in the 'Tools' sidebar. An arrow points from this icon to the 'Choose local file' button at the bottom of the upload tool window. The upload tool window is titled 'Download from web or upload from disk' and has tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based'. It shows a table of files added to the queue:

Name	Size	Type	Genome	Settings	Status
SraRunTable.csv	2 KB	Auto-det...	— Additional Sp...	⚙	0%
ContrastList.txt	13 b	Auto-det...	— Additional Sp...	⚙	0%

At the bottom of the window, there are buttons for 'Choose local file', 'Choose FTP file', 'Paste/Fetch data', 'Pause', 'Reset', 'Start', and 'Close'. The 'Choose local file' button is circled in red.

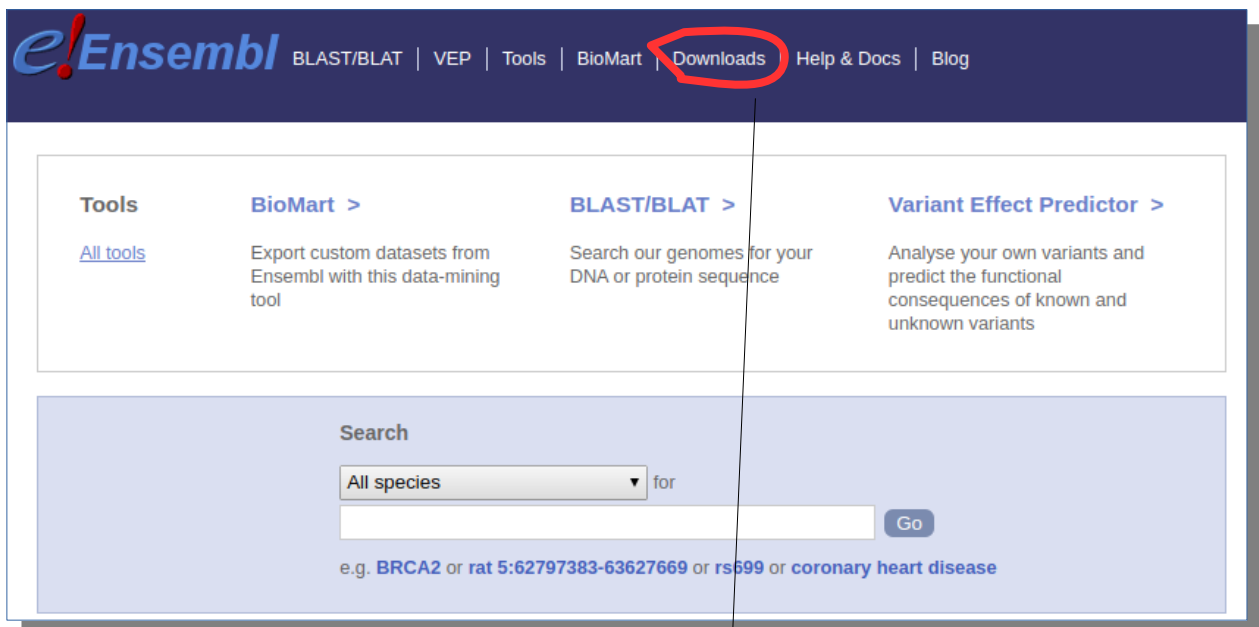
There is no need to select “Start” right now. We can prepare all the files to upload and external links to get data from and then click “Start” for all together.

RNA-Seq Analysis: Getting Data

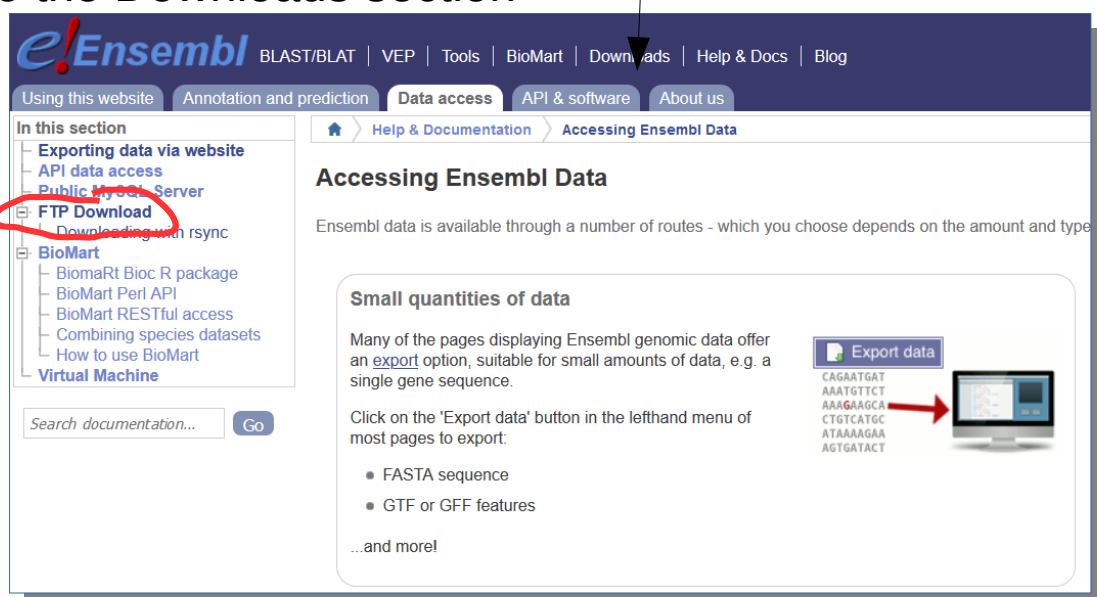
Collect the FTP link to get the Gene and Genome references:

- 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/fastg/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_patc		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.

File Name	Size	Auto-detect	Additional Sp...	Progress	Actions
SraRunTable.csv	2 KB	Auto-det...	Additional Sp...	0%	
ContrastList.txt	13 b	Auto-det...	Additional Sp...	0%	
New File	106 b	Auto-det...	Additional Sp...	0%	

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... Additional Sp... Progress 0%

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 Genome (set all): Additional Sp...

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

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 screenshot shows the Galaxy tool interface for "Faster Download and Extract Reads in FASTQ format from NCBI SRA (Galaxy Version 2.10.4+galaxy1)". The tool is selected in the "Tools" sidebar, which is circled in red. The main interface shows the "select input type" dropdown set to "SRR accession", an "Accession" text field, and an "Advanced Options" section with an "Email notification" checkbox. A "Tools" sidebar on the right lists the tool, which is circled in red. Blue arrows point from the sidebar to the tool's input fields and the "Advanced Options" section.

- 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 accession numbers.
- But first we are going to split our table in two datasets, one for the WT's and one for the KO's
- We can build a file with just that column by using some Text Manipulation tools.

RNA-Seq Analysis: Getting Sequence Data

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

Filter data on any column using simple expressions (Galaxy Version 1.1.0)

File to filter: 3: SraRunTable.csv

With following condition: c13=="wild_type"

Number of header lines to skip: 1

Email notification: Yes

Execute

Cut: cut the 1st column

Cut columns from a table (cut) (Galaxy Version 1.1.0)

File to cut: 3: SraRunTable.csv

Operation: Keep

Delimited by: Tab

Cut by: fields

List of Fields: 1

Email notification: Yes

Execute

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

Select last lines from a dataset (tail) (Galaxy Version 1.1.0)

Text file: 11: NrIKOAccessionList

Operation: Keep everything from this line on

Number of lines: 2

Email notification: Yes

Execute

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

1
SRR358714
SRR358715

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

RNA-Seq Analysis: Getting Sequence Data

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

The screenshot shows the Galaxy web interface for the tool "Faster Download and Extract Reads in FASTQ" (format from NCBI SRA (Galaxy Version 2.10.4+galaxy1)).

select input type

List of SRA accession, one per line

sra accession list

12: WTRunAccesionList

Advanced Options

Email notification

Yes No

Send an email notification when the job completes.

Execute

History

search datasets

RNASeq CNIC

3 lines

format: **tabular**, database: ?

1

SRR358714

SRR358715

SRR358716

11: NrlKOAccesionList

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.