

Next generation sequencing analysis of microRNAs

Jessy Slota

PhD Student

Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Canada

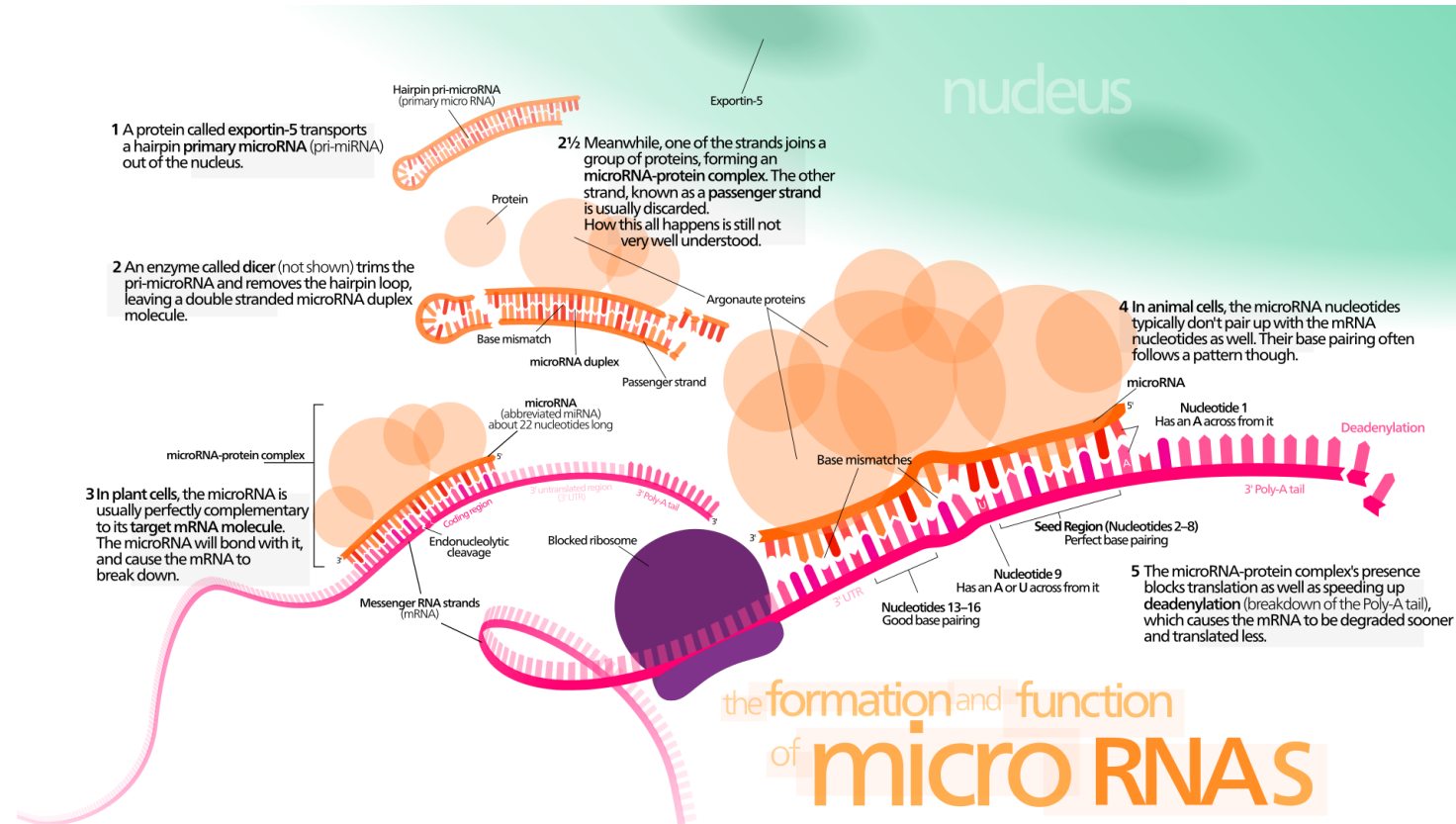
Prion Diseases Section, National Microbiology Laboratory, Public Health Agency of Canada

jessy.slota@phac-aspc.gc.ca; slotaj@myumanitoba.ca

Micro-RNAs (miRNAs)

- Small, non-coding RNAs
- ~22 nt in length
- Regulate gene expression through RISC
- Found in circulating fluids

Bartel, David P. *Cell* vol. 173,1 (2018): 20-51. doi:10.1016/j.cell.2018.03.006



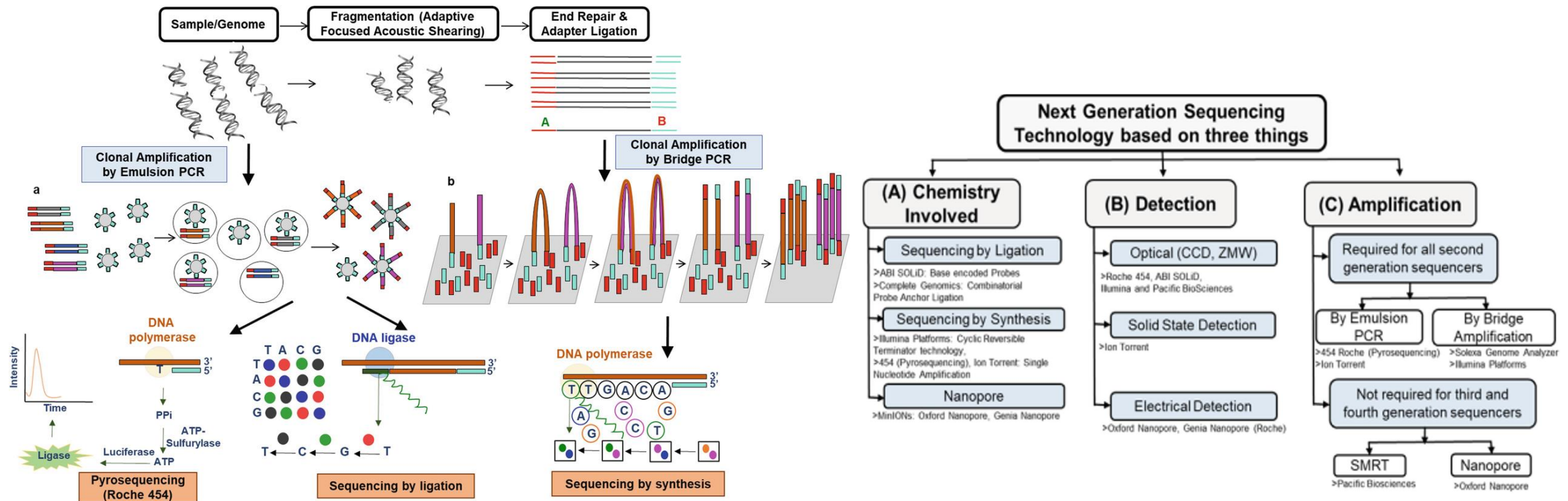
By Kelvinsong - Own work, CC BY 3.0,

<https://commons.wikimedia.org/w/index.php?curid=23311105>

Next generation sequencing (NGS)

- Allows for high-throughput sequencing of DNA or RNA
- Sequencing of RNA can be used to measure gene expression
- Can also measure abundance of miRNAs through small-RNA sequencing
- **Why sequence miRNAs?**
 - Global identification of miRNAs with altered abundance
 - Biomarker discovery
 - Combine with mRNA seq for miRNA-target identification

Next Generation Sequencing (NGS)



Gupta N., Verma V.K. (2019) Microorganisms for Sustainability, vol 17. Springer, Singapore.

https://doi.org/10.1007/978-981-13-8844-6_15

NGS of miRNAs

1. RNA extraction... **depends on tissue** being used
 - E.g., TRIzol reagent, various kits for different types of tissues or biological fluids
 - Assess quality with Agilent Bioanalyzer
2. Library preparation
 - E.g., Illumina small RNA library preparation
 - Check quality of library on Agilent Bioanalyzer or TapeStation etc.
 - Library Fractionation (size selection **122–166 bp**) with Pippin HT (SageScience)
3. Sequencing
 - E.g., Illumina NextSeq 500/550
 - usually performed by sequencing core
4. Primary output is raw `fastq` files

Kolanowska M., Kubiak A., Jażdżewski K., Wójcicka A.
(2018) In: Ørom U. (eds) miRNA Biogenesis. Methods in
Molecular Biology, vol 1823. Springer, New York, NY.
https://doi-org.uml.idm.oclc.org/10.1007/978-1-4939-8624-8_8

NGS data analysis

- “I have some data, **now what?**”
- Different tools available... commercially available vs publicly available
- The following tutorial will only use publicly available tools:
- Pre-processing with **Galaxy** – mapping reads to miRNAs
- Downstream analysis with **R** – statistics and data visualization (e.g., differential expression analysis)

Tutorial: basic NGS analysis of miRNAs

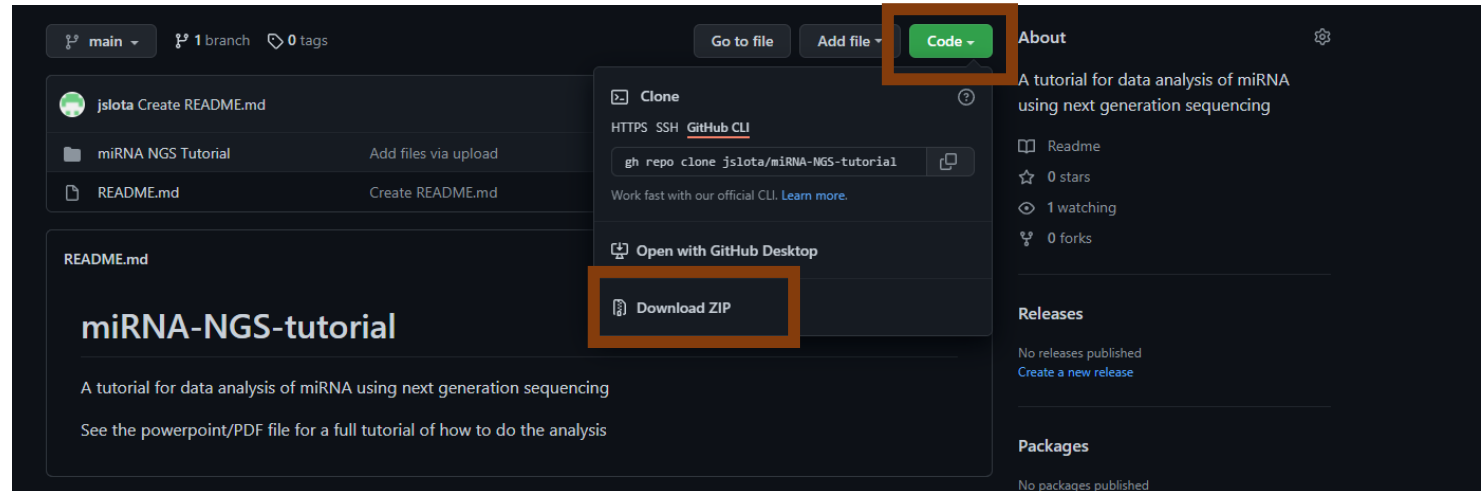
- **Part I: Preprocessing in Galaxy**

- Assessing sequencing read quality with FastQC
- Removing sequencing adapters with Cutadapt
- Cleaning sequencing reads with Trimmomatic
- Aligning to reference genome with Bowtie2
- Mapping reads to miRNAs and counting with FeatureCounts

- **Part II: Downstream Analysis with R**

- Raw reads count processing
- Normalization and differential expression analysis with DESeq2
- Examples of common data visualizations

Getting started: download tutorial data directory



- Find the tutorial at: <https://github.com/jslota/miRNA-NGS-tutorial>
- Can manually download Zip file with all tutorial materials
- Unzip to your preferred location to get folder with all materials

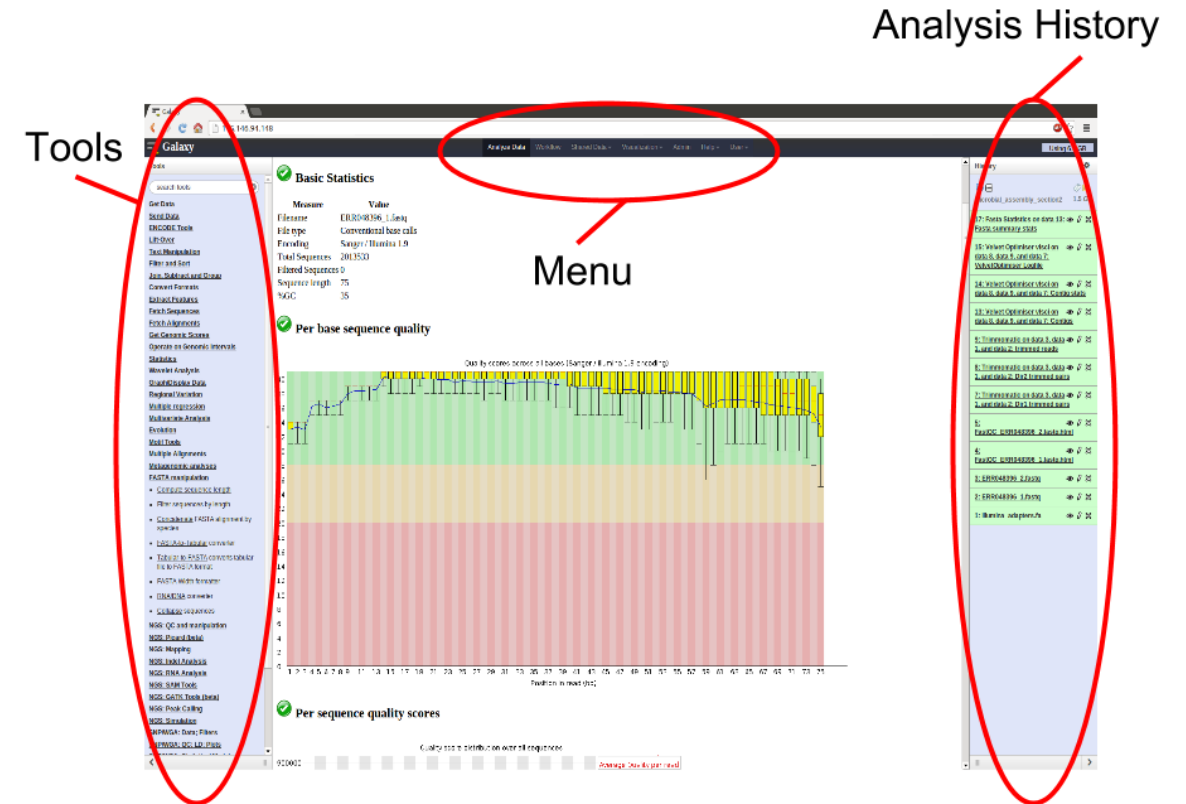
Part I: Pre-processing sequencing reads in Galaxy

- **Part I: Preprocessing in Galaxy**

1. Downloading data from NCBI SRA
2. Assessing sequencing read quality with FastQC
3. Removing sequencing adapters with Cutadapt
4. Cleaning sequencing reads with Trimmomatic
5. Aligning to reference genome with Bowtie2
6. Mapping reads to miRNAs and counting with FeatureCounts
7. Download read count files (and re-upload as individual files)
8. Differential expression analysis with DESeq2 within Galaxy

Galaxy platform

- Publicly available
- Different versions... can try free version online
- Need to set up free account for this tutorial
- Access at: <https://usegalaxy.org/>



Step 1: Collect some data

**SCIENTIFIC
REPORTS**
nature research

We will use some of the
data from this publication...

OPEN

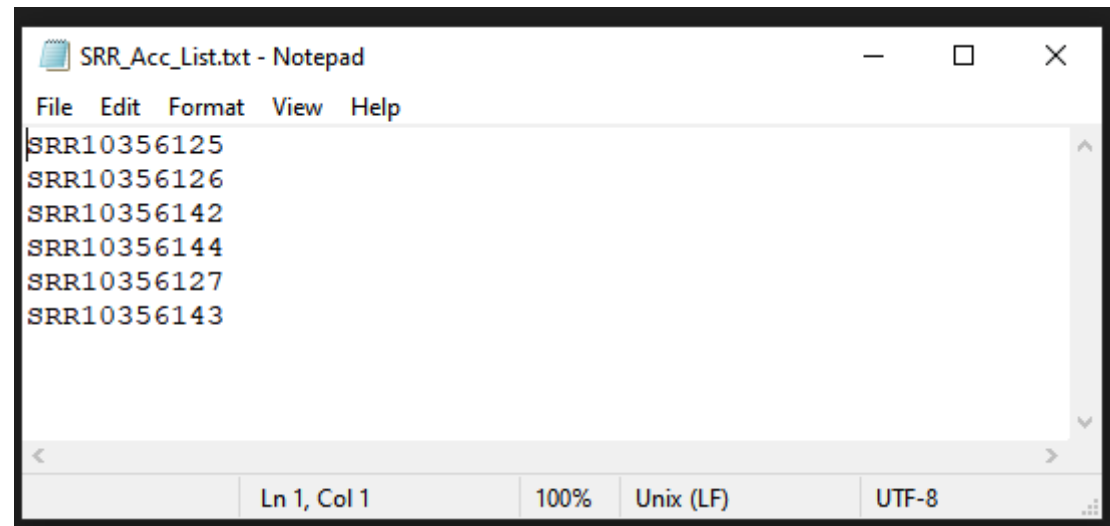
Identification of circulating microRNA signatures as potential biomarkers in the serum of elk infected with chronic wasting disease

Jessy A. Slota^{1,2}, Sarah J. Medina¹, Megan Klassen¹, Damian Gorski⁴, Christine M. Mesa¹, Catherine Robertson¹, Gordon Mitchell⁵, Michael B. Coulthart³, Sandra Pritzkow⁴, Claudio Soto⁴ & Stephanie A. Booth^{1,2*}

Chronic wasting disease (CWD) is an emerging infectious prion disorder that is spreading rapidly in wild populations of cervids in North America. The risk of zoonotic transmission of CWD is as yet unclear but a high priority must be to minimize further spread of the disease. No simple diagnostic tests are available to detect CWD quickly or in live animals; therefore, easily accessible biomarkers may be useful in identifying infected animals. MicroRNAs (miRNAs) are a class of small, non-coding RNA molecules that circulate in blood and are promising biomarkers for several infectious diseases. In this study we used next-generation sequencing to characterize the serum miRNA profiles of 35 naturally infected elk that tested positive for CWD in addition to 35 elk that tested negative for CWD. A total of 21 miRNAs that are highly conserved amongst mammals were altered in abundance in sera, irrespective of hemolysis in the samples. A number of these miRNAs have previously been associated with prion diseases. Receiver operating characteristic (ROC) curve analysis was performed to evaluate the discriminative potential of these miRNAs as biomarkers for the diagnosis of CWD. We also determined that a subgroup of 6 of these miRNAs were consistently altered in abundance in serum from hamsters experimentally infected with scrapie. This suggests that common miRNA candidate biomarkers could be selected for prion diseases in multiple species. Furthermore, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses pointed to a strong correlation for 3 of these miRNAs, miR-148a-3p, miR-186-5p, miR-30e-3p, with prion disease.

Slota, J.A., Medina, S.J., Klassen, M. *et al.* *Sci Rep* **9**, 19705 (2019).
<https://doi.org/10.1038/s41598-019-56249-6>

Step 1: Collect some data

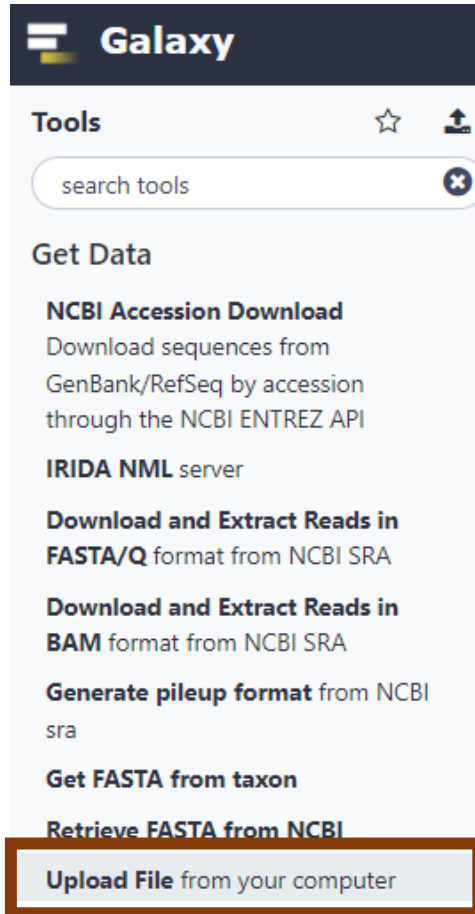


A screenshot of a Notepad window titled "SRR_Acc_List.txt - Notepad". The window contains a list of six SRA accession numbers: SRR10356125, SRR10356126, SRR10356142, SRR10356144, SRR10356127, and SRR10356143. The status bar at the bottom indicates "Ln 1, Col 1", "100%", "Unix (LF)", and "UTF-8".

```
SRR10356125
SRR10356126
SRR10356142
SRR10356144
SRR10356127
SRR10356143
```

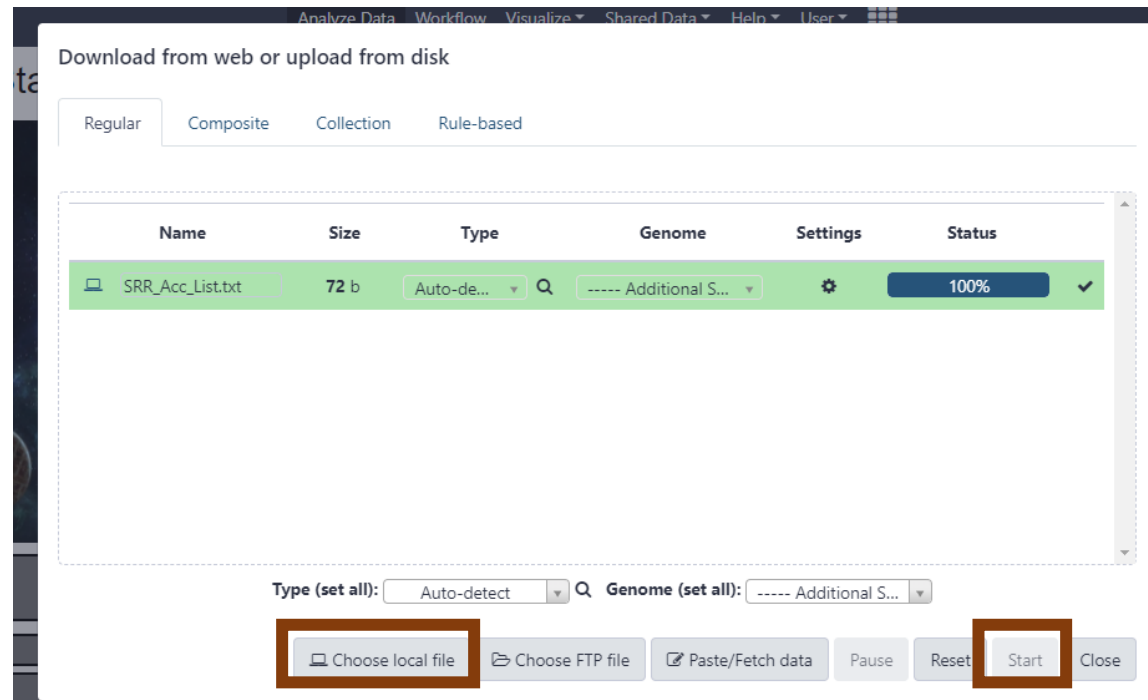
- Publicly available data from NCBI SRA (<https://www.ncbi.nlm.nih.gov/sra>)
- Data will be imported into Galaxy using this list of SRA accession numbers that can be found in the downloaded tutorial directory

Step 1: Collect some data



The Galaxy Tools sidebar is shown. At the top is the 'Galaxy' logo. Below it is a 'Tools' section with a search bar. Under 'Get Data', several tools are listed: 'NCBI Accession Download', 'IRIDA NML server', 'Download and Extract Reads in FASTA/Q format from NCBI SRA', 'Download and Extract Reads in BAM format from NCBI SRA', 'Generate pileup format from NCBI sra', 'Get FASTA from taxon', and 'Retrieve FASTA from NCBI'. At the bottom, the 'Upload File from your computer' option is highlighted with a brown border.

Type “**Upload file**” into
Tools search bar

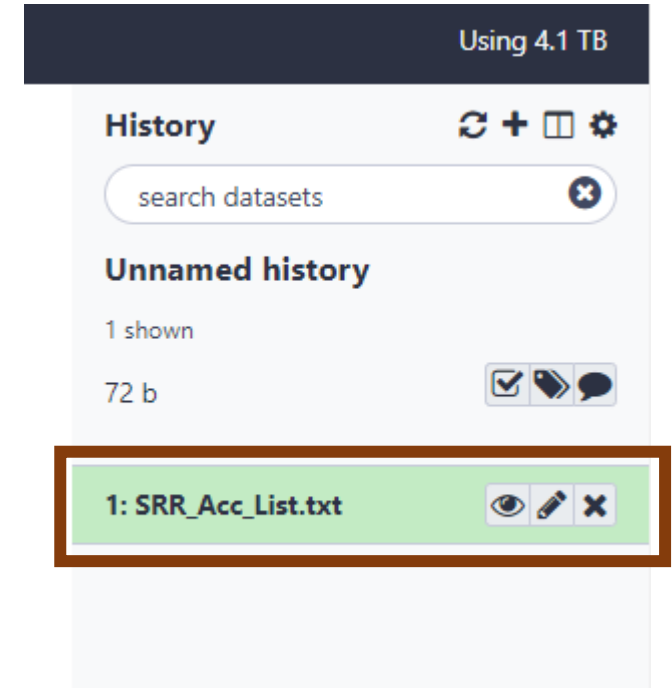


The 'Download from web or upload from disk' dialog is open. It has tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based'. A table lists the upload details:

Name	Size	Type	Genome	Settings	Status
SRR_Acc_List.txt	72 b	Auto-de...	----- Additional S...		100%

At the bottom, there are buttons for 'Choose local file', 'Choose FTP file', 'Paste/Fetch data', 'Pause', 'Reset', 'Start', and 'Close'. The 'Choose local file' and 'Start' buttons are highlighted with brown borders.

Browse and select “**SRR_Acc_List.txt**”
from wherever it is saved



The Galaxy History sidebar is shown. At the top, it says 'Using 4.1 TB'. Below is the 'History' section with a search bar. Under 'Unnamed history', it shows '1 shown' and '72 b'. A list item '1: SRR_Acc_List.txt' is highlighted with a brown border, showing icons for view, edit, and delete.

Step 1: Collect some data

The screenshot displays the Galaxy web interface. On the left, the 'Tools' sidebar shows the search bar with 'download and extract reads' entered. Below it, the 'Get Data' section lists several tools, with 'Download and Extract Reads in FASTA/Q format from NCBI SRA' highlighted by a red box. The main panel shows the configuration for this tool. The 'select input type' dropdown is set to 'List of SRA accession, one per line', and the 'sra accession list' field contains '1: SRR_Acc_List.txt'. The 'Select output format' section has 'gzip compressed fastq' selected. Below this, there is an 'Email notification' section with 'Yes' and 'No' buttons, and an 'Execute' button at the bottom. A red box highlights the 'Execute' button. On the right, the 'History' panel shows a list of datasets, with '3: Single-end data (fastq-dump)' and '2: Pair-end data (fastq-dump)' highlighted by a red box. The top of the interface shows the Galaxy logo and the tool title 'Download and Extract Reads in FASTA/Q format from NCBI SRA (Galaxy Version 2.8.1.3)'.

Galaxy

Tools

download and extract reads

Get Data

NCBI Accession Download

Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API

Download and Extract Reads in FASTA/Q format from NCBI SRA

Download and Extract Reads in BAM format from NCBI SRA

Generate pileup format from NCBI SRA

Download and Extract Reads in FASTA/Q format from NCBI SRA (Galaxy Version 2.8.1.3)

select input type

List of SRA accession, one per line

Select “List of SRA accession...” from drop-down menu

sra accession list

1: SRR_Acc_List.txt

Select output format

☒ gzip compressed fastq

☐ Uncompressed fastq

☐ bzip2 compressed fastq

Compression will greatly reduce the amount of space occupied by downloaded data. Downstream applications such as a short-read mappers will accept compressed data as input. Consider this example: an uncompressed 400 Mb fastq datasets compresses to 100 Mb or 80 Mb by gzip or bzip2, respectively. (--gzip --bzip2)

Advanced Options

Email notification

Yes No

Send an email notification when the job completes.

Execute

Using 4.1 TB

History

search datasets

Unnamed history

3 shown

72 b

3: Single-end data (fastq-dump)

a list

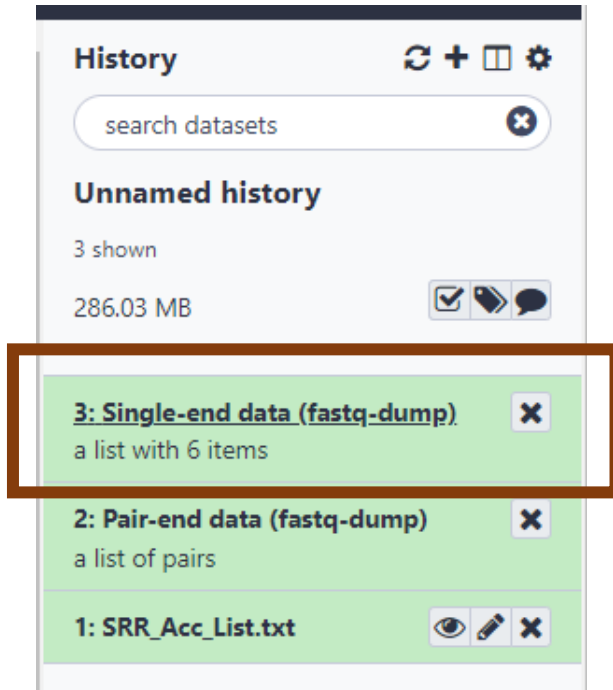
2: Pair-end data (fastq-dump)

a list of pairs

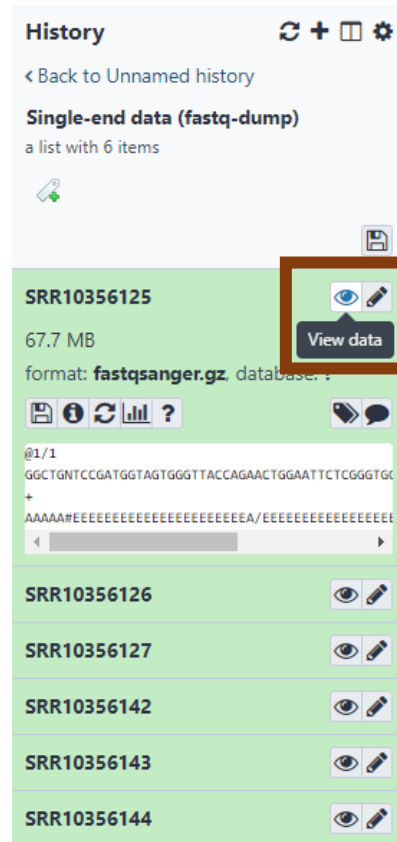
1: SRR_Acc_List.txt

Type “Download and Extract Reads in FASTA/Q” into tools search bar

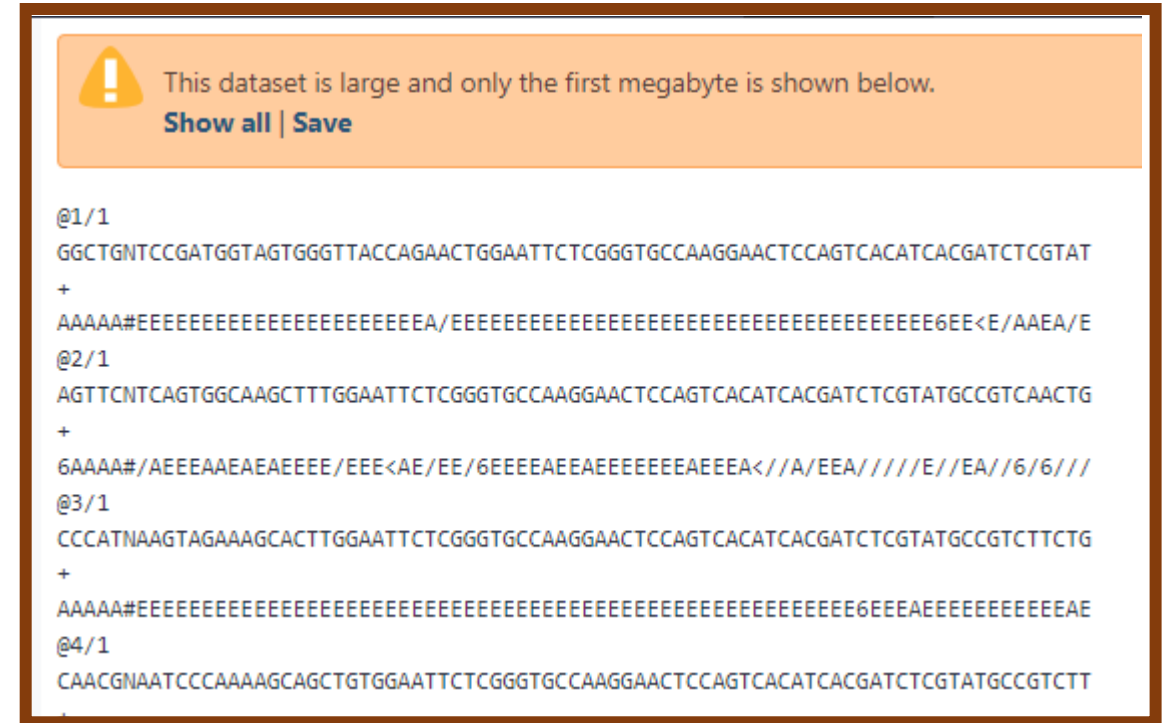
Step 1: Collect some data



Data is in the format of a **collection**
Click on the collection to examine individual datasets



Click "View data"



Data is in **Fastq** format

Step 2: Assess data quality with FastQC

Tools

fastqc

NGS: QC and manipulation

fastp - fast all-in-one preprocessing for FASTQ files

FastQC Read Quality reports

FastQC Summary Provide a one line summary of a FastQC report(s)

FastQC Read Quality reports (Galaxy Version 0.72+galaxy1)

Short read data from your current history

3: Single-end data (fastq-dump)

Dataset collection this is a batch mode input field. Separate jobs will b

Contaminant list

No tabular dataset available.

tab delimited file with 2 columns: name and sequence. For example: Illumin

Adapter list

No tabular dataset available.

list of adapters adapter sequences which will be explicitly searched against t

Submodule and Limit specifying file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and :

Disable grouping of bases for reads > 50bp

Using this option will cause fastqc to crash and burn if you use it on really l

Lower limit on the length of the sequence to be shown in the report

As long as you set this to a value greater or equal to your longest read leng
comparable statistics from datasets with somewhat variable read lengths. t

length of Kmer to look for

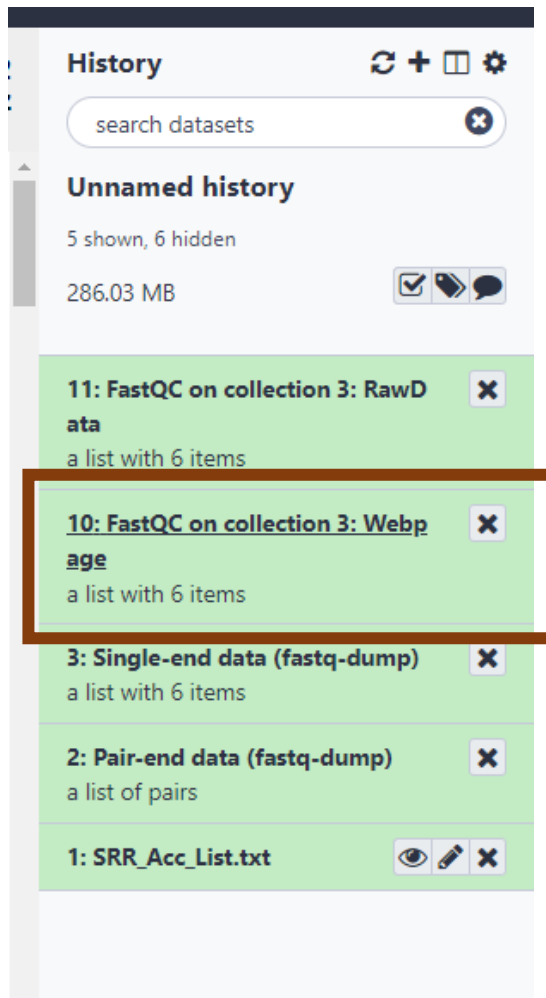
note: the Kmer test is disabled and needs to be enabled using a custom Sut

Email notification

Send an email notification when the job completes.

Select “**dataset collection**” Icon
Then select “**Single-end data...**” from drop-down

Step 2: Assess data quality



History

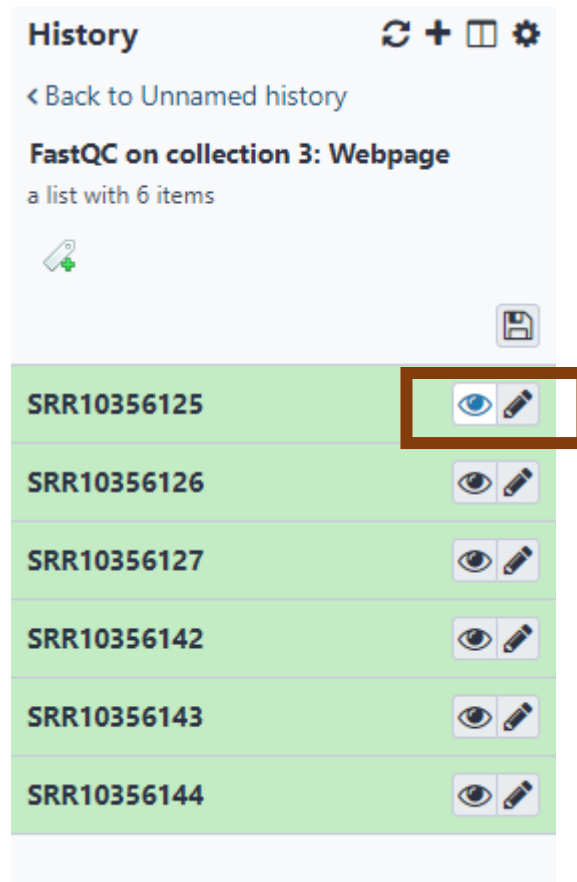
search datasets

Unnamed history

5 shown, 6 hidden

286.03 MB

- 11: FastQC on collection 3: RawData
a list with 6 items
- 10: FastQC on collection 3: Webpage**
a list with 6 items
- 3: Single-end data (fastq-dump)
a list with 6 items
- 2: Pair-end data (fastq-dump)
a list of pairs
- 1: SRR_Acc_List.txt



History

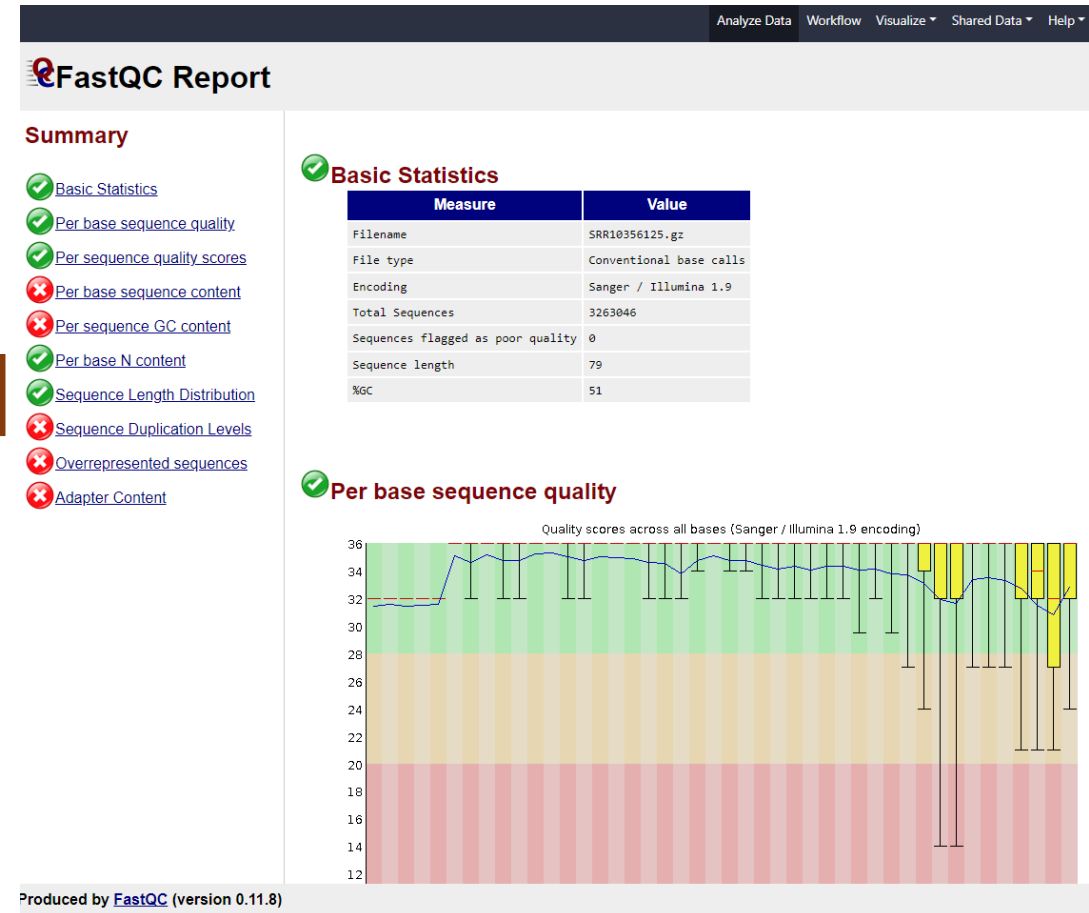
< Back to Unnamed history

FastQC on collection 3: Webpage

a list with 6 items

- SRR10356125
- SRR10356126
- SRR10356127
- SRR10356142
- SRR10356143
- SRR10356144

Click “View data”



FastQC Report

Analyze Data Workflow Visualize Shared Data Help

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

Measure	Value
Filename	SRR10356125.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3263046
Sequences flagged as poor quality	0
Sequence length	79
%GC	51

Per base sequence quality

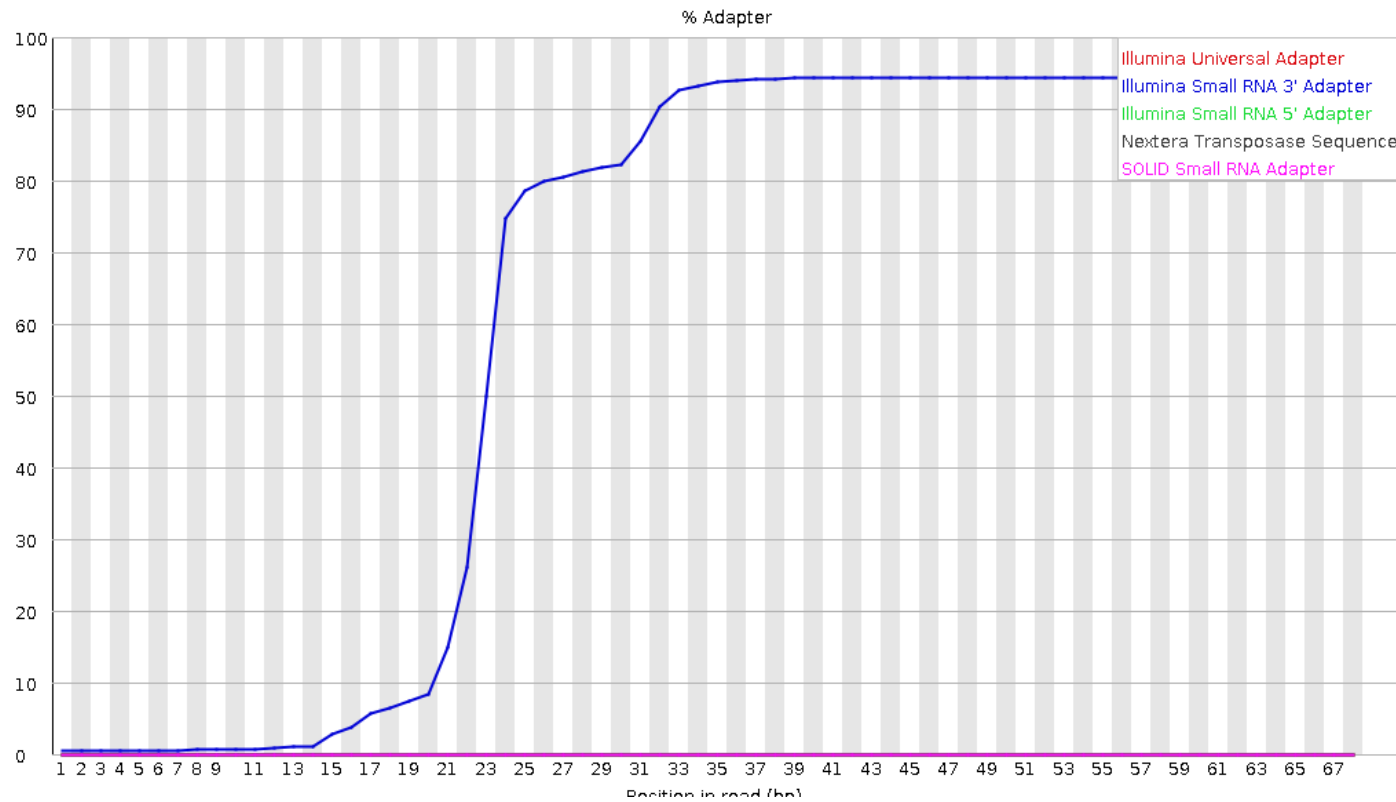
Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Produced by FastQC (version 0.11.8)

Explore various quality reports...

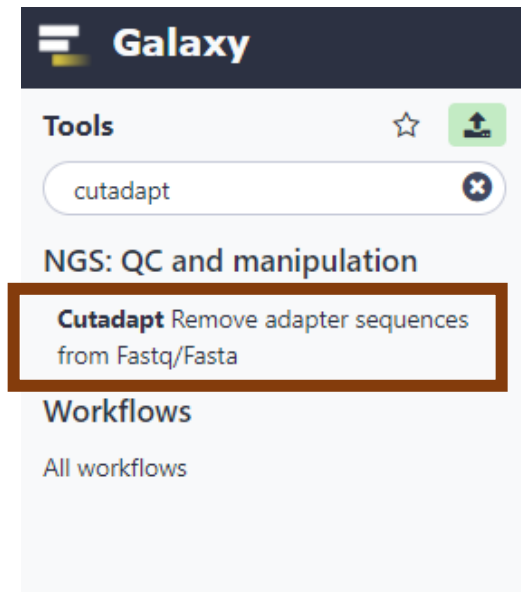
Step 2: Assess data quality

❌ Adapter Content

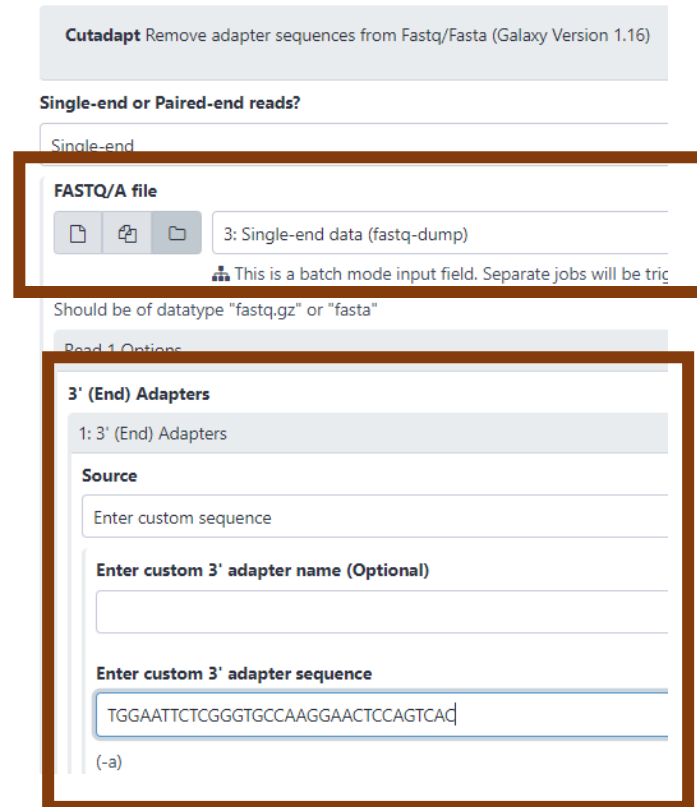


From FastQC results, we can see Illumina adapter contamination...

Step 3: Remove sequencing adapters with Cutadapt



The image shows the Galaxy web interface's left sidebar. At the top is the 'Galaxy' logo. Below it is a 'Tools' section with a search bar containing 'cutadapt'. Underneath the search bar, the category 'NGS: QC and manipulation' is listed. The 'Cutadapt Remove adapter sequences from Fastq/Fasta' tool is highlighted with a brown rectangular box. Below this, there is a 'Workflows' section with a link to 'All workflows'.



The image shows the configuration page for the 'Cutadapt Remove adapter sequences from Fastq/Fasta' tool. The title bar indicates 'Cutadapt Remove adapter sequences from Fastq/Fasta (Galaxy Version 1.16)'. The first section, 'Single-end or Paired-end reads?', has 'Single-end' selected. The 'FASTQ/A file' section is highlighted with a brown box and contains a file selection icon, a folder icon, and a text input field with '3: Single-end data (fastq-dump)'. Below this is a note: 'This is a batch mode input field. Separate jobs will be triggered for each line.' The next section, 'Read 1 Options', is also highlighted with a brown box and contains a sub-section '3' (End) Adapters' with '1: 3' (End) Adapters' listed. Under 'Source', there is a text input field for 'Enter custom sequence'. Below that is a text input field for 'Enter custom 3' adapter name (Optional)'. At the bottom, there is a text input field for 'Enter custom 3' adapter sequence' containing the sequence 'TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC'.

Select “**dataset collection**” Icon
Then select “**Single-end data..**” from drop-down

Copy/paste the 3' adapter sequence:
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC

Step 3: Remove sequencing adapters with Cutadapt

The image shows the Cutadapt web interface. On the left, the configuration panel includes a text input for 'Cut bases from reads before adapter' set to 0, with a description 'Remove bases from each read (first read)'. Below this are links for 'Adapter Options', 'Filter Options', 'Read Modification Options', and 'Output Options'. An 'Email notification' section has 'Yes' and 'No' buttons. A blue 'Execute' button with a checkmark is highlighted with a brown box. Below the button is an information icon and the text 'What it does'. At the bottom, a partial description of Cutadapt's function is visible.

History ↺ + □ ⚙

search datasets ✕

Unnamed history

6 shown, 18 hidden

305.1 MB ✓ 🔍 💬

- 24: Cutadapt on collection 3: Read 1 Output** ✕
a list with 6 items
- 11: FastQC on collection 3: Raw Data** ✕
a list with 6 items
- 10: FastQC on collection 3: Webpage** ✕
a list with 6 items
- 3: Single-end data (fastq-dump)** ✕
a list with 6 items
- 2: Pair-end data (fastq-dump)** ✕
a list of pairs
- 1: SRR_Acc_List.txt** 👁 ✎ ✕

Step 4: Clean reads with Trimmomatic

Galaxy

Tools

trimmo

NGS: QC and manipulation

fastp - fast all-in-one preprocessing for FASTQ files

Trimmomatic flexible read trimming tool for Illumina NGS data

NGS: Assembler

Shovill Faster SPAdes assembly of Illumina reads

Workflows

All workflows

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.5)

Single-end or paired-end reads?

Single-end

Input FASTQ file

24: Cutadapt on collection 3: Read 1 Output

This is a batch mode input field. Separate jobs will be triggered for each input file.

Perform initial Trimmomatic step?

Yes No

Cut adapter and other illumina-specific sequences from the read

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

20

+ Insert Trimmomatic Operation

+ Add new Trimmomatic Operation block

Send an email notification when the job completes.

Execute

Select “**dataset collection**” Icon
Then select “**Cutadapt on collection...**”
from drop-down

Click “**Insert Trimmomatic Operation**”
twice for a total of 3 “operations”

Step 4: Clean reads with Trimmomatic

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

20

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads with average quality lower than a specified

Minimum average quality required to keep a read

20

3: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept

18

+ Insert Trimmomatic Operation

Email notification

Yes No

Send an email notification when the job completes.

✓ Execute

1st operation – leave as **defaults**

2nd operation – “**Drop reads with average quality...**” Min average quality = **20**

3rd operation – “**Drop reads below length...**” min length = **18**
(How long should miRNA reads be?)

History

search datasets

Unnamed history

7 shown, 24 hidden rename history

418.95 MB

31: Trimmomatic across collection 24
a list with 6 items

24: Cutadapt on collection 3: Read 1 Output
a list with 6 items

11: FastQC on collection 3: RawData
a list with 6 items

10: FastQC on collection 3: Webpage
a list with 6 items

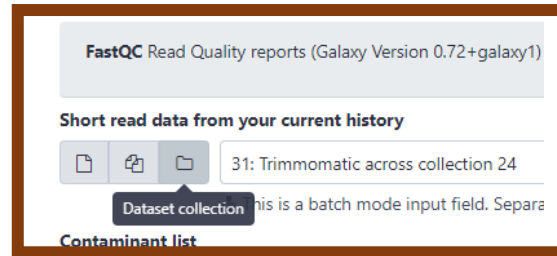
3: Single-end data (fastq-dump)
a list with 6 items

2: Pair-end data (fastq-dump)
a list of pairs

1: SRR_Acc_List.txt




Step 4: Clean reads with Trimmomatic

Use **FastQC** to
check quality of
cleaned reads
(Same as Step 2)






FastQC Read Quality reports (Galaxy Version 0.72+galaxy1)

Short read data from your current history

   31: Trimmomatic across collection 24




Dataset collection is a batch mode input field. Separate multiple datasets by comma.

Contaminant list

   No tabular dataset available.




tab delimited file with 2 columns: name and sequence. For example:

Adapter list

   No tabular dataset available.

list of adapters adapter sequences which will be explicitly searched

Submodule and Limit specifying file

   Nothing selected

a file that specifies which submodules are to be executed (default is all)

Disable grouping of bases for reads >50bp

Yes No

Using this option will cause fastqc to crash and burn if you use

Lower limit on the length of the sequence to be shown in the

As long as you set this to a value greater or equal to your long

length of Kmer to look for

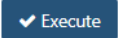
7

note: the Kmer test is disabled and needs to be enabled using

Email notification

Yes No

Send an email notification when the job completes.



Select “**dataset collection**” Icon
Then select “**Trimmomatic across
collection...**” from drop-down

Step 4: Clean reads with Trimmomatic

Examine **FastQC** results as in step 2:

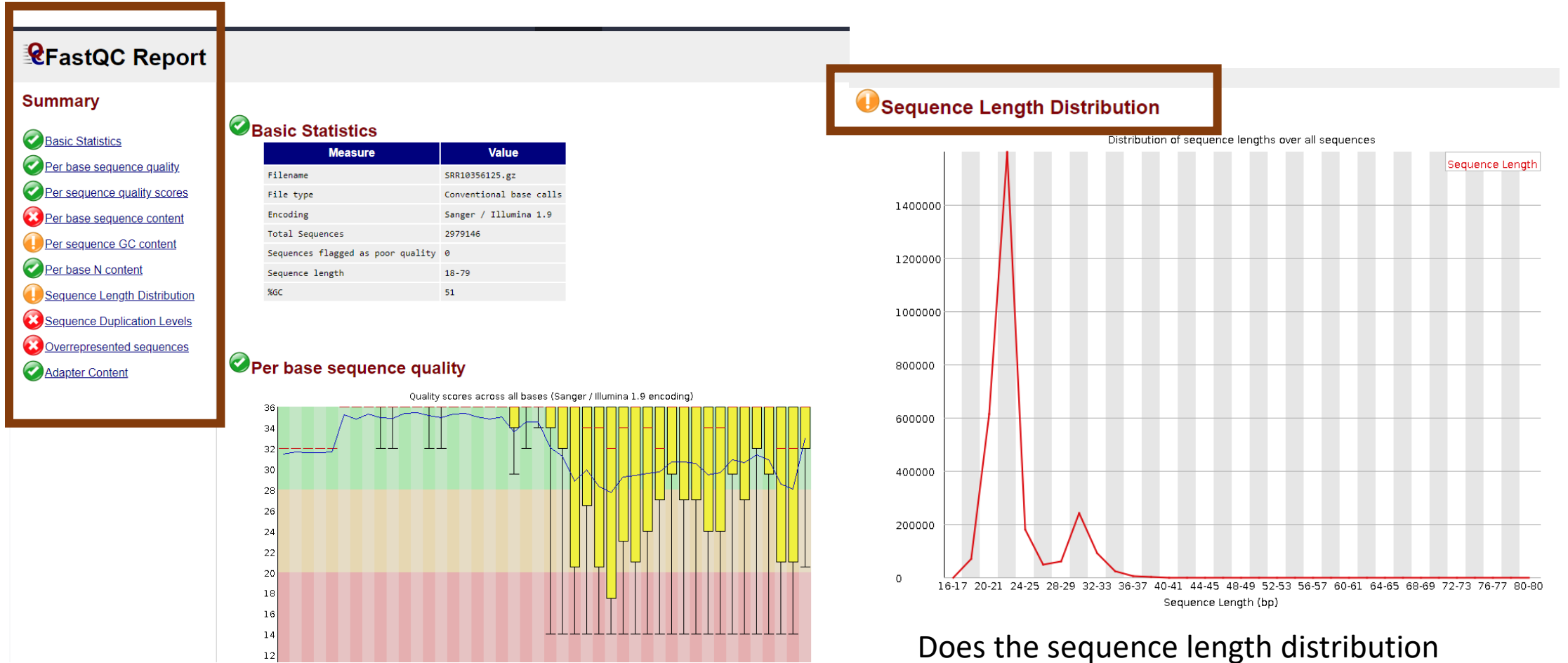
The History panel shows a list of datasets. The entry '38: FastQC on collection 31: Web page' is highlighted with a brown box. Below it, the entry '31: Trimmomatic across collection 24' is also visible. The list includes various FastQC and Trimmomatic outputs, along with raw data and pair-end data.

Dataset Name	Description	Actions
39: FastQC on collection 31: Raw Data	a list with 6 items	Close (X)
38: FastQC on collection 31: Web page	a list with 6 items	Close (X)
31: Trimmomatic across collection 24	a list with 6 items	Close (X)
24: Cutadapt on collection 3: Read 1 Output	a list with 6 items	Close (X)
11: FastQC on collection 3: Raw Data	a list with 6 items	Close (X)
10: FastQC on collection 3: Web page	a list with 6 items	Close (X)
3: Single-end data (fastq-dump)	a list with 6 items	Close (X)
2: Pair-end data (fastq-dump)	a list of pairs	Close (X)
1: SRR_Acc_List.txt		View, Edit, Close (X)

The FastQC results page shows a list of items. The entry 'SRR10356125' is highlighted with a brown box. The page includes a 'Back to Unnamed history' link and a 'FastQC on collection 31: Webpage' title. The list contains six items, each with a 'View' (eye) and 'Edit' (pencil) icon.

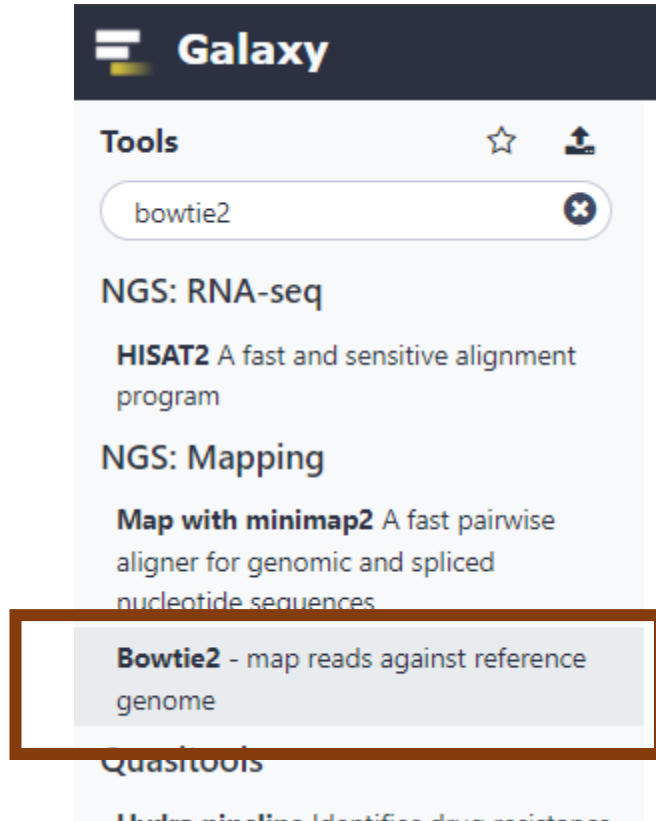
Item ID	Actions
SRR10356125	View, Edit
SRR10356126	View, Edit
SRR10356127	View, Edit
SRR10356142	View, Edit
SRR10356143	View, Edit
SRR10356144	View, Edit

Step 4: Clean reads with Trimmomatic



Does the sequence length distribution make sense for **miRNAs**?

Step 5: Align to reference genome with Bowtie2



The screenshot shows the Bowtie2 tool configuration page in Galaxy. The 'FASTA/Q file' section is highlighted with a brown box, showing the input '31: Trimmomatic across collection 24'. The 'Will you select a reference genome from your history or use a built-in index?' section is also highlighted with a brown box, showing the selection of 'human_GRCh38' as the built-in genome index.

Analyze Data

Bowtie2 - map reads against reference genome (Galaxy Version 2.3.4.3+galaxy0)

Is this single or paired library

Single-end

FASTA/Q file

31: Trimmomatic across collection 24

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Must be of datatype "fastqsanger" or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

Yes No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

Yes No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See 'Indexes' section of help below

Select reference genome

human_GRCh38

If your genome of interest is not listed, contact the Galaxy team

Set read groups information?

Do not set

Specifying read group information can greatly simplify your downstream analyses by allowing combining multiple datasets.

Select analysis mode

1: Default setting only

Do you want to use samtools?

Select **"dataset collection"** Icon
Then select **"Trimmomatic across collection..."** from drop-down

Select **"Built in genome index"** and choose **Human genome** from dropdown
***Note, we are using human genome, because elk annotation is not available*

Step 5: Align to reference genome with Bowtie2

Do you want to use presets?

☒ No, just use defaults

☐ Very fast end-to-end (--very-fast)

☐ Fast end-to-end (--fast)

☐ Sensitive end-to-end (--sensitive)

☐ Very sensitive end-to-end (--very-sensitive)

☐ Very fast local (--very-fast-local)

☐ Fast local (--fast-local)

☐ Sensitive local (--sensitive-local)

☐ Very sensitive local (--very-sensitive-local)

Allow selecting among several preset parameter settings

Do you want to tweak SAM/BAM Options?

No

See "Output Options" section of Help below for information

Save the bowtie2 mapping statistics to the history

Email notification

Send an email notification when the job completes.

Bowtie2 Overview

Using 4.1 TB

History

search datasets

Unnamed history

11 shown, 24 hidden

418.95 MB

53: Bowtie2 on collection 31: mapping stats
a list with 6 items

52: Bowtie2 on collection 31: alignments
a list with 6 items

39: FastQC on collection 31: RawData
a list with 6 items

38: FastQC on collection 31: Webpage
a list with 6 items

31: Trimmomatic across collection 24
a list with 6 items

History

< Back to Unnamed history

Bowtie2 on collection 31: mapping stats

a list with 6 items

View data

SRR10356125

SRR10356126

SRR10356127

SRR10356142

SRR10356143

SRR10356144

2979146 reads; of these:

- 2979146 (100.00%) were unpaired; of these:
- 346086 (11.62%) aligned 0 times
- 1270093 (42.63%) aligned exactly 1 time
- 1362967 (45.75%) aligned >1 times

88.38% overall alignment rate

View mapping statistics... notice that there is a high level of multi-mapped reads (Does this make sense for miRNAs?)

Step 5: Align to reference genome with Bowtie2

Using 4.1 TB

History

search datasets

11 shown, 24 hidden

418.95 MB

53: Bowtie2 on collection 31: mapping stats

a list with 6 items

52: Bowtie2 on collection 31: a alignments

a list with 6 items

39: FastQC on collection 31: RawData

a list with 6 items

38: FastQC on collection 31: Webpage

a list with 6 items

31: Trimmomatic across collection 31

a list with 6 items

History

[◀ Back to Unnamed history](#)

Bowtie2 on collection 31: alignments
 a list with 6 items

SRR10356125	
SRR10356126	<div>View data</div>

View alignment file in “BAM” format

@SQ SN:chrUn_GL000216v2 LN:176608 @SQ SN:chrUn_GL000218v1 LN:161147 @SQ SN:chrEBV LN:171823 @PG ID:bowtie2 PN:bowtie2 VN:2.3.4.1 CL:/Drives/P/Galaxies/main_nm/galaxy-common/deps/_conda/envs/mulled-v1-5bee08a20f60a5597c4ecd54735d608dc6a44cf6f433cd2f23c80aa5a38d02/bin/bowtie2-align-s --wrapper basic-0 -p 4 -x /Drives/P/Galaxies/main_nm/galaxy-common/tool-data/human_GRCh											
3192/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAA/EEEEAAEEEEEEEEEE/EAAE	ASit-5 XSi-5 XNi-0 XMit-1 X
3467/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAAEEEEEEEEEEEEEEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
4171/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAAEEEEEEEEEEEEEEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
6286/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	AAAAAAEEEEEEEEEE/EEEEEEEEEEA	ASit-5 XSi-5 XNi-0 XMit-1 X
8602/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	AAAAAAEEEEEEEEEEA/EEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
9612/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAAEEEEEEEEEE/EEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
12020/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAAEEEEEEEEEEEEEEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
12816/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	AAAAAAEEEEEEEEEEEEAAEEEEEEEE<A	ASit-5 XSi-5 XNi-0 XMit-1 X
13931/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	A/AAAAAEEEEEEEEEEAAEEEEAAEEA<	ASit-5 XSi-5 XNi-0 XMit-1 X
20161/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAAEEEEEEEEEEEE/EE<E/EE<<	ASit-5 XSi-5 XNi-0 XMit-1 X
30122/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	AAAAAAEEEEEEEEEEEEEEEEEEEEEE/E	ASit-5 XSi-5 XNi-0 XMit-1 X
32636/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAAEEEEEEEEEEEEEE/EEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
33359/1	0	chr1	629572	1	29M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGC	AAAAAAEEEEEEEEEEEEEEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
41153/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	AAAAAAEEEEEEEEEEEEAAEEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
42249/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	AAAAAAEEEEEEEEEEEEEEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
43004/1	0	chr1	629572	1	30M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCC	AAAAAAEEEEEEEEEEEEEEEEEEEEEE	ASit-5 XSi-5 XNi-0 XMit-1 X
53652/1	0	chr1	629572	1	31M	*	0	0	AGTAAGGTCAGCTAATTAAGCTATCGGGCCC	AAAAAAEEEEEEEEEEEEAAEEEEEEAE	ASit-5 XSi-5 XNi-0 XMit-1 X

Step 6: Map and count miRNA reads with FeatureCounts

The screenshot shows the Galaxy web interface. On the left sidebar, under the 'Tools' section, the 'Upload File from your computer' option is highlighted with a brown box. The main area is titled 'Download from web or upload from disk'. It has tabs for 'Regular', 'Composite', 'Collection', and 'Rule-based'. Below the tabs, it says 'You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.' A table with columns 'Name', 'Size', 'Type', 'Genome', 'Settings', and 'Status' is shown. The first row has 'New File' in the Name column, '52 b' in the Size column, 'Auto-de...' in the Type column, and '0%' in the Status column. A text input field below the table contains the URL 'https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3', which is highlighted with a brown box. Below the input field, it says 'Download data from the web by entering URLs (one per line) or directly paste content.' At the bottom of the interface, there are buttons for 'Choose local file', 'Choose FTP file', 'Paste/Fetch data' (highlighted with a brown box), 'Pause', 'Resume', 'Start' (highlighted with a brown box), and 'Close'.

Galaxy

Tools

search tools

Get Data

NCBI Accession Download
Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API

IRIDA NML server

Download and Extract Reads in FASTA/Q format from NCBI SRA

Download and Extract Reads in BAM format from NCBI SRA

Generate pileup format from NCBI sra

Get FASTA from taxon

Retrieve FASTA from NCBI

Upload File from your computer

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	52 b	Auto-de...	----- Additional S...		0%

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

https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3

Copy/paste the following URL:
<https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3>





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


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

Upload **reference annotation** file in GFF format

Step 6: Map and count miRNA reads with FeatureCounts




Using 4.1 TB




History    

search datasets 

Unnamed history







12 shown, 24 hidden

418.95 MB   

66: <https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3>   

4,801 lines, 13 comments
format: **gff3**, database: ?


uploaded gff3 file

display with IGV local

1. Seqid

```
##gff-version 3
##date 2018-3-5
#
# Chromosomal coordinates of Homo sapiens microRNAs
# microRNAs: mirBase v22
```

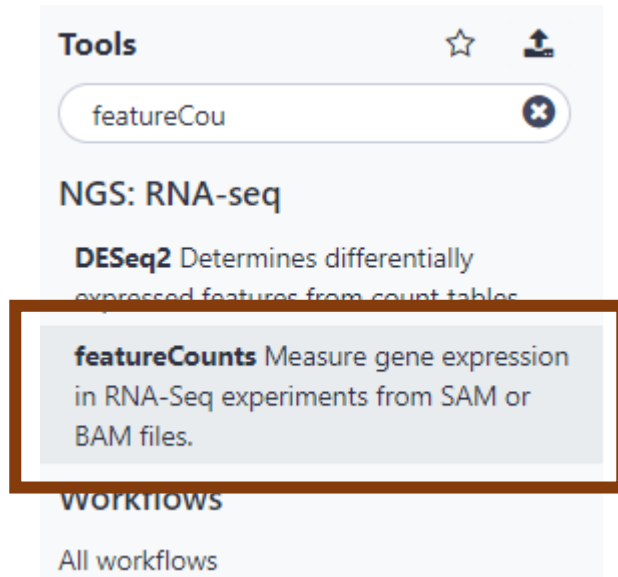
52: Bowtie2 on collection 31: a 

alignments
a list with 6 items

Seqid	Source	Type	Start	End	Score	Strand	Phase	Attributes
##gff-version 3								
##date 2018-3-5								
#								
# Chromosomal coordinates of Homo sapiens microRNAs								
# microRNAs: mirBase v22								
# genome-build-id: GRCh38								
# genome-build-accession: NCBI_Assembly:GCA_000001405.15								
#								
# Hairpin precursor sequences have type "miRNA_primary_transcript".								
# Note, these sequences do not represent the full primary transcript.								
# rather a predicted stem-loop portion that includes the precursor								
# miRNA. Mature sequences have type "miRNA".								
#								
chr1	.	miRNA_primary_transcript	17369	17436	.	-	.	ID=MI0022705;Alias=MI0022705;Name=hsa-mir-6859-1
chr1	.	miRNA	17409	17431	.	-	.	ID=MIMAT0027618;Alias=MIMAT0027618;Name=hsa-miR-6859-5p;Derives_from=MI0022705
chr1	.	miRNA	17369	17391	.	-	.	ID=MIMAT0027619;Alias=MIMAT0027619;Name=hsa-miR-6859-3p;Derives_from=MI0022705
chr1	.	miRNA_primary_transcript	30366	30503	.	+	.	ID=MI0006363;Alias=MI0006363;Name=hsa-mir-1302-2
chr1	.	miRNA	30438	30458	.	+	.	ID=MIMAT0005890;Alias=MIMAT0005890;Name=hsa-miR-1302;Derives_from=MI0006363
chr1	.	miRNA_primary_transcript	187891	187958	.	-	.	ID=MI0026420;Alias=MI0026420;Name=hsa-mir-6859-2
chr1	.	miRNA	187931	187953	.	-	.	ID=MIMAT0027618_1;Alias=MIMAT0027618;Name=hsa-miR-6859-5p;Derives_from=MI0026420
chr1	.	miRNA	187891	187913	.	-	.	ID=MIMAT0027619_1;Alias=MIMAT0027619;Name=hsa-miR-6859-3p;Derives_from=MI0026420
chr1	.	miRNA_primary_transcript	632615	632685	.	-	.	ID=MI0039740;Alias=MI0039740;Name=hsa-mir-12136
chr1	.	miRNA	632668	632685	.	-	.	ID=MIMAT0049032;Alias=MIMAT0049032;Name=hsa-miR-12136;Derives_from=MI0039740
chr1	.	miRNA_primary_transcript	1167104	1167198	.	+	.	ID=MI0000342;Alias=MI0000342;Name=hsa-mir-200b
chr1	.	miRNA	1167124	1167145	.	+	.	ID=MIMAT0004571;Alias=MIMAT0004571;Name=hsa-miR-200b-5p;Derives_from=MI0000342
chr1	.	miRNA	1167160	1167181	.	+	.	ID=MIMAT0000318;Alias=MIMAT0000318;Name=hsa-miR-200b-3p;Derives_from=MI0000342
chr1	.	miRNA_primary_transcript	1167863	1167952	.	+	.	ID=MI0000737;Alias=MI0000737;Name=hsa-mir-200a
chr1	.	miRNA	1167878	1167899	.	+	.	ID=MIMAT0001620;Alias=MIMAT0001620;Name=hsa-miR-200a-5p;Derives_from=MI0000737
chr1	.	miRNA	1167916	1167937	.	+	.	ID=MIMAT0000682;Alias=MIMAT0000682;Name=hsa-miR-200a-3p;Derives_from=MI0000737
chr1	.	miRNA_primary_transcript	1169005	1169087	.	+	.	ID=MI0001641;Alias=MI0001641;Name=hsa-mir-429
chr1	.	miRNA	1169055	1169076	.	+	.	ID=MIMAT0001536;Alias=MIMAT0001536;Name=hsa-miR-429;Derives_from=MI0001641
chr1	.	miRNA_primary_transcript	1296110	1296170	.	-	.	ID=MI0022571;Alias=MI0022571;Name=hsa-mir-6726
chr1	.	miRNA	1296145	1296165	.	-	.	ID=MIMAT0027353;Alias=MIMAT0027353;Name=hsa-miR-6726-5p;Derives_from=MI0022571
chr1	.	miRNA	1296110	1296129	.	-	.	ID=MIMAT0027354;Alias=MIMAT0027354;Name=hsa-miR-6726-3p;Derives_from=MI0022571

View annotation file in GFF format

Step 6: Map and count miRNA reads with FeatureCounts



The screenshot shows the configuration page for the 'featureCounts' tool. The page has a light blue header with the tool name and description. The main content area is divided into several sections, each with a title and input fields. The 'Alignment file' section is highlighted with a brown box and contains a text input field with the value '52: Bowtie2 on collection 31: alignments'. Below this field is a small icon of a person and the text 'This is a batch mode input field. Separate jobs will be triggered for each dataset'. The 'Gene annotation file' section is also highlighted with a brown box and contains a text input field with the value 'in your history'. Below this field is a small icon of a person and the text 'The program assumes that the provided annotation file is in GTF format. Make sure that the gene'. The 'Output format' section contains a text input field with the value 'Gene-ID "\t" read-count (DESeq2 IUC wrapper compatible)'. Below this field is the text 'The output format will be tabular, select the preferred columns here'. The 'Create gene-length file' section contains two radio buttons, 'Yes' and 'No', with 'No' selected. Below this section is the text 'Creates a tabular file that contains the effective (nucleotides used for counting reads) length of the fe:'. At the bottom of the page, there are two expandable sections: 'Options for paired-end reads' and 'Advanced options'.

Select “**dataset collection**” Icon
Then select “**Bowtie2 on collection...**” from drop-down

Select “in your history”
Then select **miRbase GFF file** from drop-down

Step 6: Map and count miRNA reads with FeatureCounts

Creates a tabular file that contains the effective (nucleotides used for mapping)

[Options for paired-end reads](#)

[Advanced options](#)

GFF feature type filter

miRNA

Specify the feature type. Only rows which have the matched match will be included.

GFF gene identifier

Name

Specify the attribute type used to group features (eg. exons) into groups.

On feature level

☒ Yes ☐ No

If specified, read summarization will be performed at the feature level.

Allow read to contribute to multiple features

☒ Yes ☐ No

If specified, reads (or fragments if -p is specified) will be allowed to contribute to multiple features.

Strand specificity of the protocol

Stranded (forwards)

Indicate if strand-specific read counting should be performed. (-) indicates reverse strand.

Count multi-mapping reads/fragments

Enabled; multi-mapping reads are included

If specified, multi-mapping reads/fragments will be counted (ie. 1/n).

Assign fractions to multimapping reads

☒ Yes ☐ No

If specified, a fractional count 1/n will be generated for each mapping.

Minimum mapping quality per read

0

The minimum mapping quality score a read must satisfy in order to be counted.

Change GFF feature type to “**miRNA**”

Change GFF gene identifier to “**Name**”

Change strand specificity to “**Stranded (forwards)**”

Enable **multi-mapped** read counting

Set minimum read quality to 0
(This is because of how bowtie2 deals with multi-mapped read)

Step 6: Map and count miRNA reads with FeatureCounts

The image shows two parts of a bioinformatics workflow interface. The left part is a configuration panel for a job, and the right part is a history panel.

Configuration Panel (Left):





- Text: "If specified, reads that were marked as duplicates will be ignored."
- Section: **Ignore unspliced alignments**
- Buttons:
- Text: "If specified, only split alignments (CIGAR strings containing T)"
- Section: **Email notification**
- Buttons:
- Text: "Send an email notification when the job completes."
- Button: (highlighted with a brown box)


History Panel (Right):

- Header: "Using 4.1 TB"
- Section: **History**
- Search bar: "search datasets" with a clear button (X)
- Section: **Unnamed history**
- Text: "14 shown, 54 hidden"
- Text: "619.38 MB"
- Buttons: ☒ ☐ ☐
- Job list (highlighted with a brown box):
 - 68: featureCounts on collection n 52: summary
a list with 6 items
 - 67: featureCounts on collection n 52
a list with 6 items
 - 66: <https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3>
 - 53: Bowtie2 on collection 31: mapping stats
a list with 6 items

Step 6: Map and count miRNA reads with FeatureCounts




Using 4.1 TB


History    

search datasets 


Unnamed history

14 shown, 54 hidden




619.38 MB   


68: featureCounts on collection 52: summary 

a list with 6 items





67: featureCounts on collection 52 

a list with 6 items

66: <https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3>   

53: Bowtie2 on collection 31: mapping stats 

a list with 6 items


History    


[Back to Unnamed history](#)



featureCounts on collection 52:



summary



a list with 6 items









SRR10356125  

SRR10356126  

SRR10356127  

SRR10356142  

SRR10356143  

SRR10356144

View data

Status	SRR10356125
Assigned	2149693
Unassigned_Unmapped	346086
Unassigned_MappingQuality	0
Unassigned_Chimera	0
Unassigned_FragmentLength	0
Unassigned_Duplicate	0
Unassigned_MultiMapping	0
Unassigned_Secondary	0
Unassigned_Nonjunction	0
Unassigned_NoFeatures	483367
Unassigned_Overlapping_Length	0
Unassigned_Ambiguity	0

Examine mapping statistics

Step 6: Map and count miRNA reads with FeatureCounts

Using 4.1 TB

History

search datasets

Unnamed history

14 shown, 54 hidden
619.38 MB

- 68: featureCounts on collection 52: summary
a list with 6 items
- 67: featureCounts on collection 52**
a list with 6 items
- 66: <https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3>
- 53: Bowtie2 on collection 31: mapping stats
a list with 6 items

History

< Back to Unnamed history

featureCounts on collection 52

a list with 6 items

- SRR10356125
- SRR10356126
- SRR10356127
- SRR10356142
- SRR10356143
- SRR10356144





View data


Geneid	SRR10356125
hsa-miR-6859-5p	0
hsa-miR-6859-3p	0
hsa-miR-1302	0
hsa-miR-12136	0
hsa-miR-200b-5p	0
hsa-miR-200b-3p	1
hsa-miR-200a-5p	0
hsa-miR-200a-3p	5
hsa-miR-429	0
hsa-miR-6726-5p	0
hsa-miR-6726-3p	0
hsa-miR-6727-5p	0
hsa-miR-6727-3p	0
hsa-miR-6808-5p	0
hsa-miR-6808-3p	0
hsa-miR-4251	0
hsa-miR-551a	0
hsa-miR-4689	0
hsa-miR-4252	0
hsa-miR-6728-5p	0
hsa-miR-6728-3p	0
hsa-miR-34a-5p	2

Examine read count files

Step 7: Download read count files for further analysis




Using 4.1 TB







History    





search datasets 

Unnamed history

14 shown, 54 hidden

619.38 MB   



- 68: featureCounts on collection n 52: summary
a list with 6 items 
- 67: featureCounts on collection n 52**
a list with 6 items 
- 66: <https://www.mirbase.org/ftp/CURRENT/genomes/hsa.gff3>   
- 53: Bowtie2 on collection 31: mapping stats
a list with 6 items 

History    













< Back to Unnamed history

featureCounts on collection 52

a list with 6 items

Download Collection

- SRR10356125  
- SRR10356126  
- SRR10356127  
- SRR10356142  
- SRR10356143  
- SRR10356144  

hsa-miR-4003-3p 0

hsa-miR-6507-5p 0

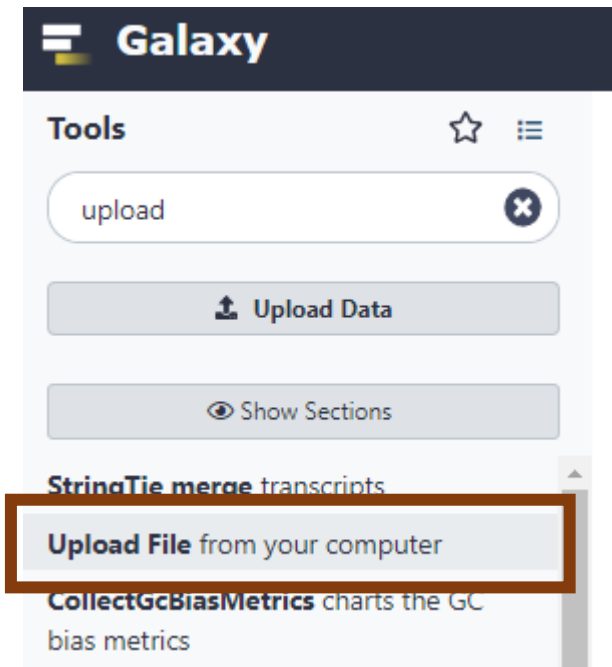
hsa-miR-6507-3p 0

hsa-miR-608 0

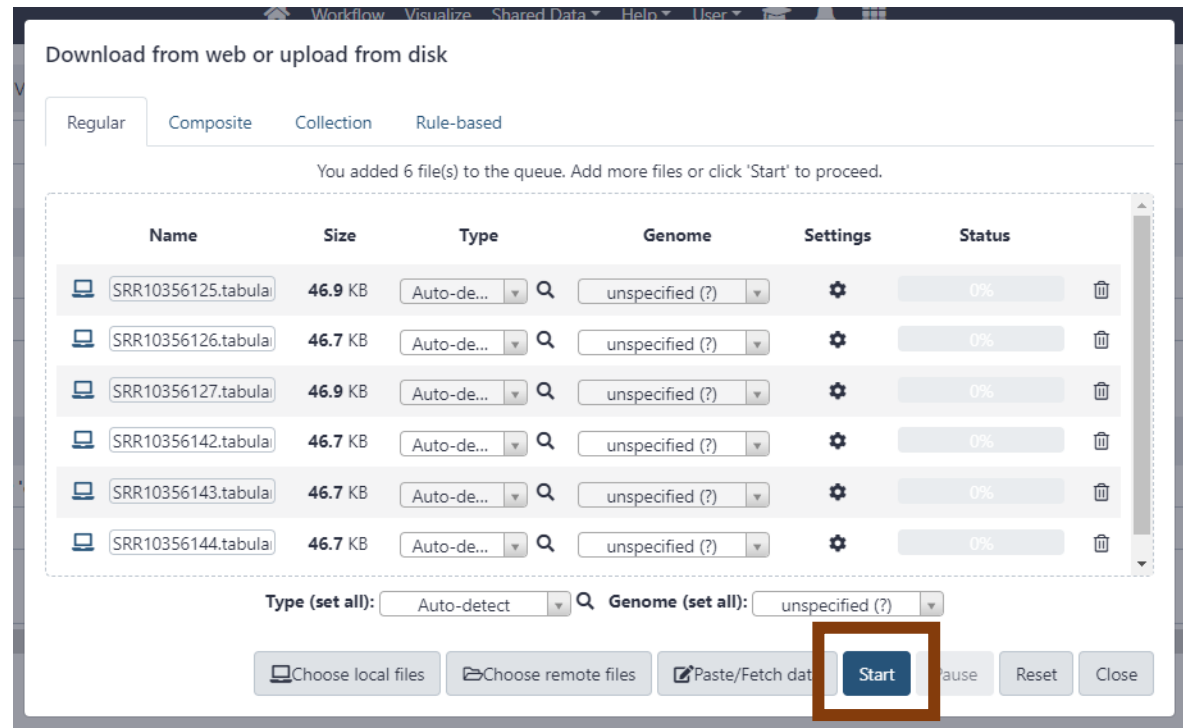
67_featureCounts....tgz

Download and unzip/extract raw read count files

Step 8: Differential expression analysis with DESeq2 in Galaxy



The Galaxy Tools sidebar shows various tools. The 'upload' button is highlighted with a red box. Below it, the 'Upload Data' button is visible. Further down, the 'StrinaTie merae transcripts' tool is listed, and the 'Upload File from your computer' button is highlighted with a red box. Below that, the 'CollectGcBiasMetrics' tool is listed.



The Galaxy Data Upload dialog shows the 'Regular' tab selected. It displays a table of 6 files added to the queue. The 'Start' button is highlighted with a red box.

Name	Size	Type	Genome	Settings	Status
SRR10356125.tabular	46.9 KB	Auto-de...	unspecified (?)		0%
SRR10356126.tabular	46.7 KB	Auto-de...	unspecified (?)		0%
SRR10356127.tabular	46.9 KB	Auto-de...	unspecified (?)		0%
SRR10356142.tabular	46.7 KB	Auto-de...	unspecified (?)		0%
SRR10356143.tabular	46.7 KB	Auto-de...	unspecified (?)		0%
SRR10356144.tabular	46.7 KB	Auto-de...	unspecified (?)		0%



The Galaxy File List shows 6 files added to the queue. The files are listed in a table with columns for file name, size, type, genome, settings, and status. The 'Start' button is highlighted with a red box.

Name	Size	Type	Genome	Settings	Status
13: SRR10356144.tabular	46.7 KB	Auto-de...	unspecified (?)		0%
12: SRR10356143.tabular	46.7 KB	Auto-de...	unspecified (?)		0%
11: SRR10356142.tabular	46.9 KB	Auto-de...	unspecified (?)		0%
10: SRR10356127.tabular	46.9 KB	Auto-de...	unspecified (?)		0%
9: SRR10356126.tabular	46.7 KB	Auto-de...	unspecified (?)		0%
8: SRR10356125.tabular	46.9 KB	Auto-de...	unspecified (?)		0%

Re-upload raw read count files as individual files, instead of a collection

Step 8: Differential expression analysis with DESeq2 in Galaxy

Factor Name = **CWD_Status**

Tools

deseq2

Upload Data

Show Sections

sequencing reads

Descriptors calculated with RDKit

DESeq2 Determines differentially expressed features from count tables

Parsimony Describes whether two or more communities have the same structure

DESeq2 Determines differentially expressed features from count tables (Galaxy Version 2.11.40.7+galaxy1)

Select datasets per level

Factor

Specify a factor name, e.g. effects_drug_x or cancer_markers

CWD_status

Factor level

1: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

Positive

Only letters, numbers and underscores will be retained in this field

Counts file(s)

13: SRR10356144.tabular
12: SRR10356143.tabular
11: SRR10356142.tabular
10: SRR10356127.tabular
9: SRR10356126.tabular
8: SRR10356125.tabular

2: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

Negative

Only letters, numbers and underscores will be retained in this field

Counts file(s)

13: SRR10356144.tabular
12: SRR10356143.tabular
11: SRR10356142.tabular
10: SRR10356127.tabular
9: SRR10356126.tabular
8: SRR10356125.tabular

These 3 samples are “positive”

These 3 samples are “negative”

Step 8: Differential expression analysis with DESeq2 in Galaxy

You can produce this file using RUVSeq or svaseq.

Files have header?
☒ No

If this option is set to yes, the tool will assume that the count files have column headers in the first row. Default: Yes

Choice of Input data

Count data (e.g. from HTSeq-count, featureCounts or StringTie)

Advanced options

Output options

Job Resource Parameters

























Use default job resource parameters

Email notification
☒ No

Send an email notification when the job completes.

What it does

Input files do not have a header

17: DESeq2 plots on data 10, data 9, and others	  
16: DESeq2 result file on data 10, data 9, and others	  
15: SRR10356144.tabular	  
12: SRR10356143.tabular	  
11: SRR10356142.tabular	  
10: SRR10356127.tabular	  
9: SRR10356126.tabular	  
8: SRR10356125.tabular	  

Step 8: Differential expression analysis with DESeq2 in Galaxy

17: DESeq2 plots on data 10, data 9, and others

16: DESeq2 result file on data 10, data 9, and others

13: SRR10356144.tabular

12: SRR10356143.tabular

11: SRR10356142.tabular

10: SRR10356127.tabular

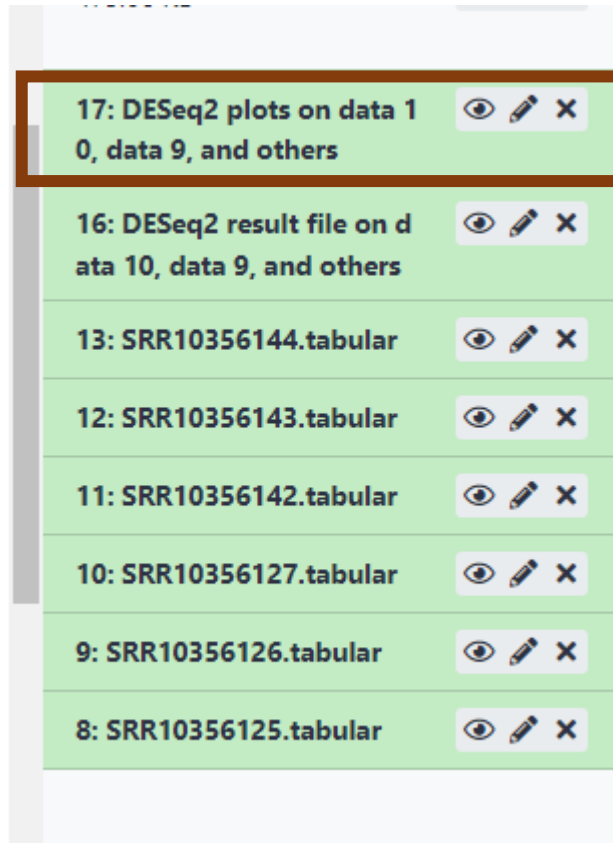
9: SRR10356126.tabular

8: SRR10356125.tabular

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
hsa-miR-144-5p	256.334411435762	4.03572974461951	0.509595099187083	7.91948303870543	2.38500096580741e-15	5.60475226964741e-13
hsa-miR-375-3p	1530.82399828163	4.02965836923229	0.606270106563184	6.64663872687961	2.99861824441712e-11	2.94775520856156e-09
hsa-miR-185-5p	200.697555088063	4.50851675336146	0.681753599736636	6.61311763532032	3.76309175561051e-11	2.94775520856156e-09
hsa-miR-486-5p	767656.041359639	3.29959068102232	0.540860163445729	6.10063543966926	1.05647618400484e-09	6.20679758102844e-08
hsa-miR-107	4529.11854670314	2.7340159694676	0.478737653288701	5.71088559816885	1.12389780540217e-08	5.28231968539021e-07
hsa-miR-103a-3p	9291.40871253028	2.64615614586663	0.470716535901716	5.62154915759989	1.89252629252658e-08	7.41239464572909e-07
hsa-miR-125b-5p	42.6781725115144	-3.81025135724564	0.705115328070426	-5.4037278804767	6.52699409200655e-08	2.19120515945934e-06
hsa-miR-125a-5p	225.362195319651	-2.78984505055407	0.547866040035147	-5.09220292313627	3.5392700475753e-07	1.03966057647525e-05
hsa-miR-185-3p	22.1888417148245	3.58467677028873	0.7252146880113	4.94291804833506	7.69618412421234e-07	2.00955918798878e-05
hsa-miR-223-3p	1246.71302037906	-2.06816457103587	0.453311111717051	-4.56235136880294	5.05839009801464e-06	0.000118872167303344
hsa-miR-199b-5p	112.822032004768	2.08988529051403	0.462872154480673	4.5150378355743	6.33054624514435e-06	0.000135243487964448
hsa-miR-451a	10083.3521501164	1.81253011494869	0.422331834059908	4.29172032220423	1.77294168302527e-05	0.000320595551386441
hsa-miR-25-3p	55069.7606639701	2.26925756216252	0.528761157782266	4.29164950708605	1.77350730554201e-05	0.000320595551386441

Examine results file

Step 8: Differential expression analysis with DESeq2 in Galaxy



17: DESeq2 plots on data 10, data 9, and others

16: DESeq2 result file on data 10, data 9, and others

13: SRR10356144.tabular

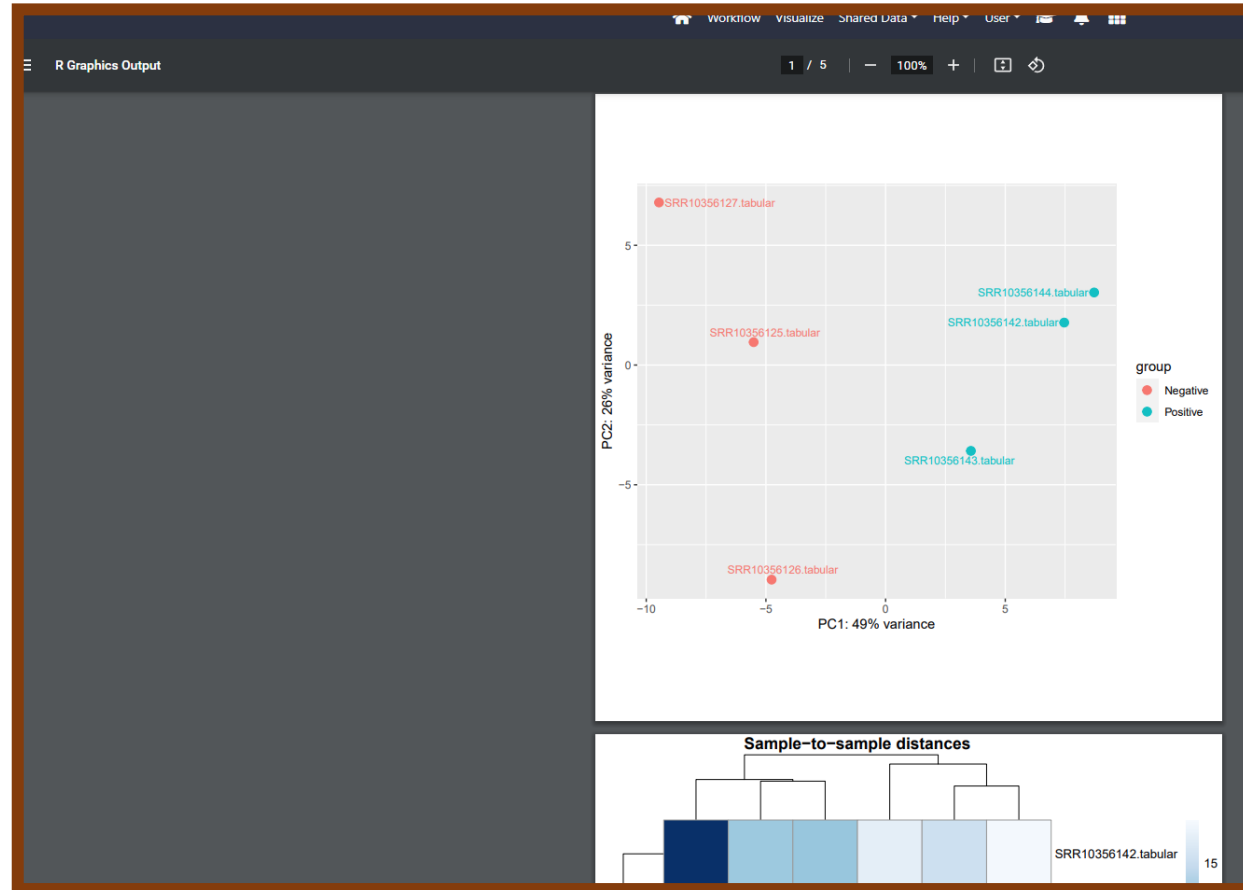
12: SRR10356143.tabular

11: SRR10356142.tabular

10: SRR10356127.tabular

9: SRR10356126.tabular

8: SRR10356125.tabular



Examine plots

Part II: Downstream analysis in R (more advanced)

1. Raw read count processing
2. Normalization and differential expression analysis with DESeq2
3. Common data visualizations

R coding environment & R Studio

What is R?

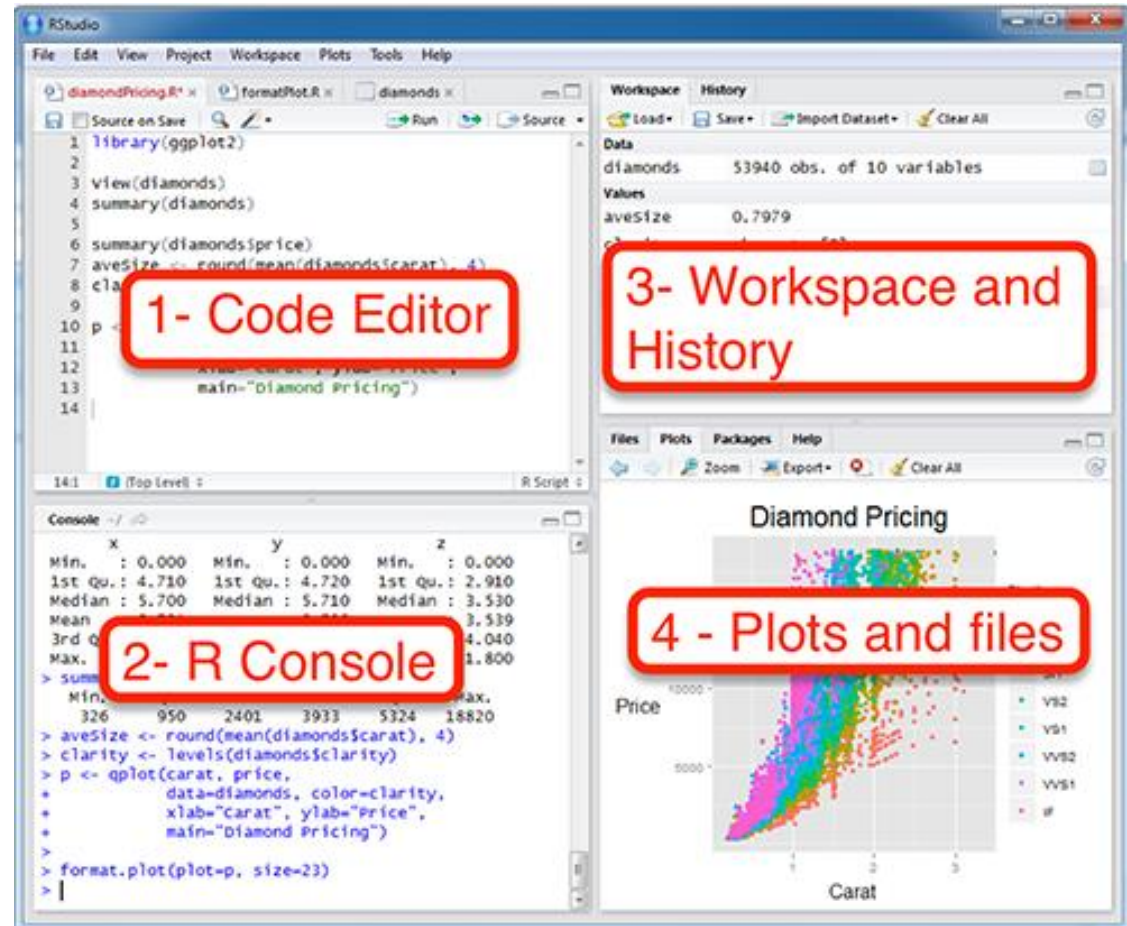
"R is a programming language for statistical computing and graphics"

Download **R** from <https://www.r-project.org/>

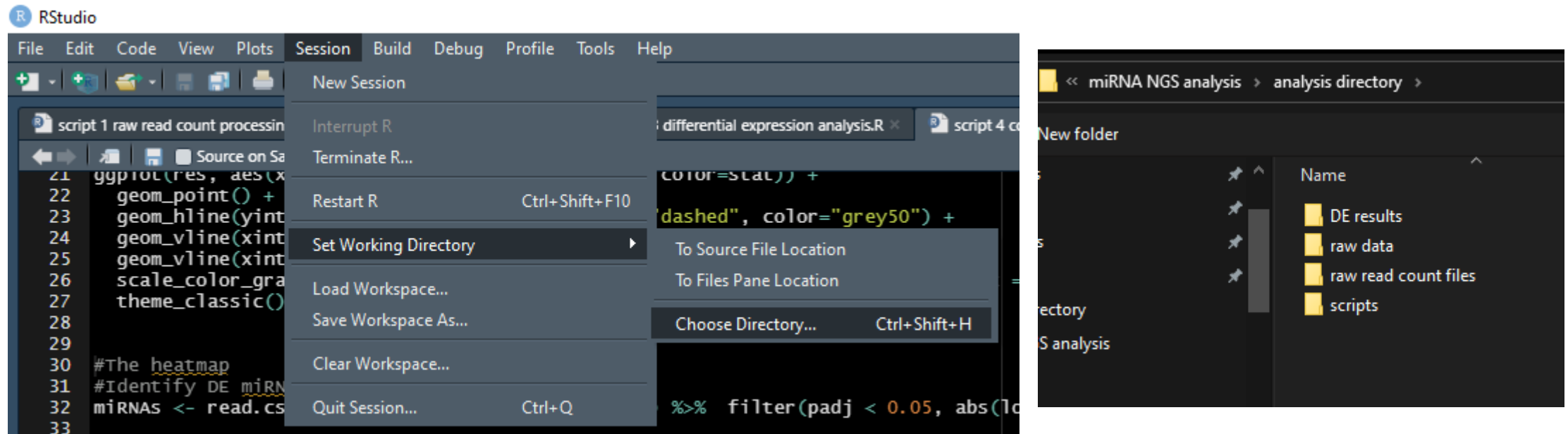
Download **R studio** from
<https://www.rstudio.com/products/rstudio/download/>

New to R?

<https://support.rstudio.com/hc/en-us/articles/201141096-Getting-Started-with-R>

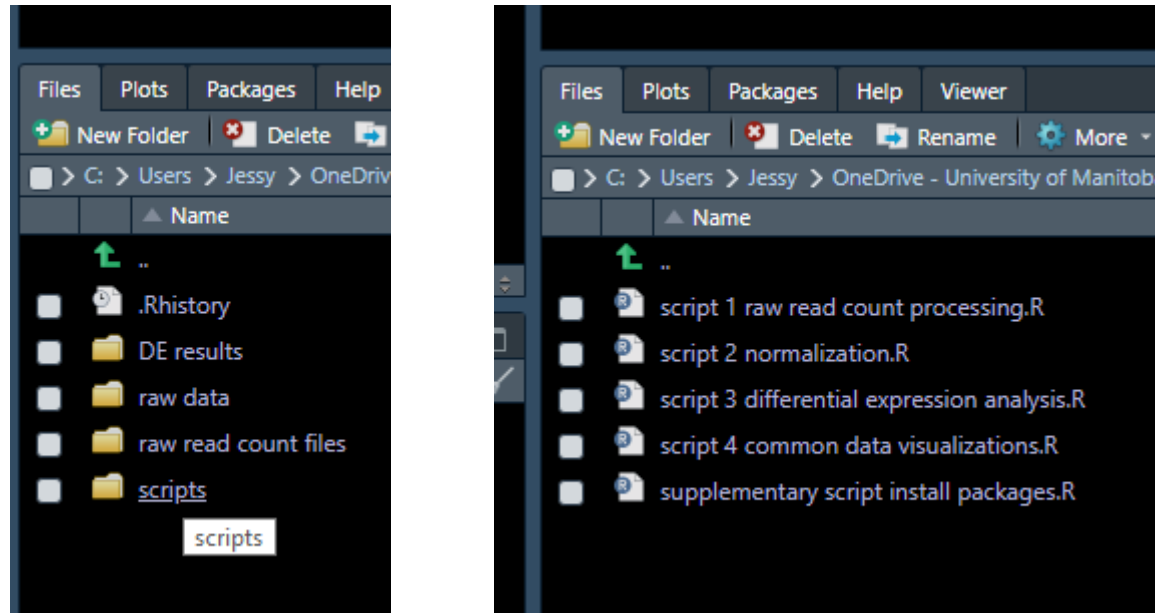


Set working directory in R Studio



The **working directory** lets R know where your files are saved, you simply specify the folder that will be used for the analysis. In this case it is the **`analysis directory`** folder provided in the tutorial.

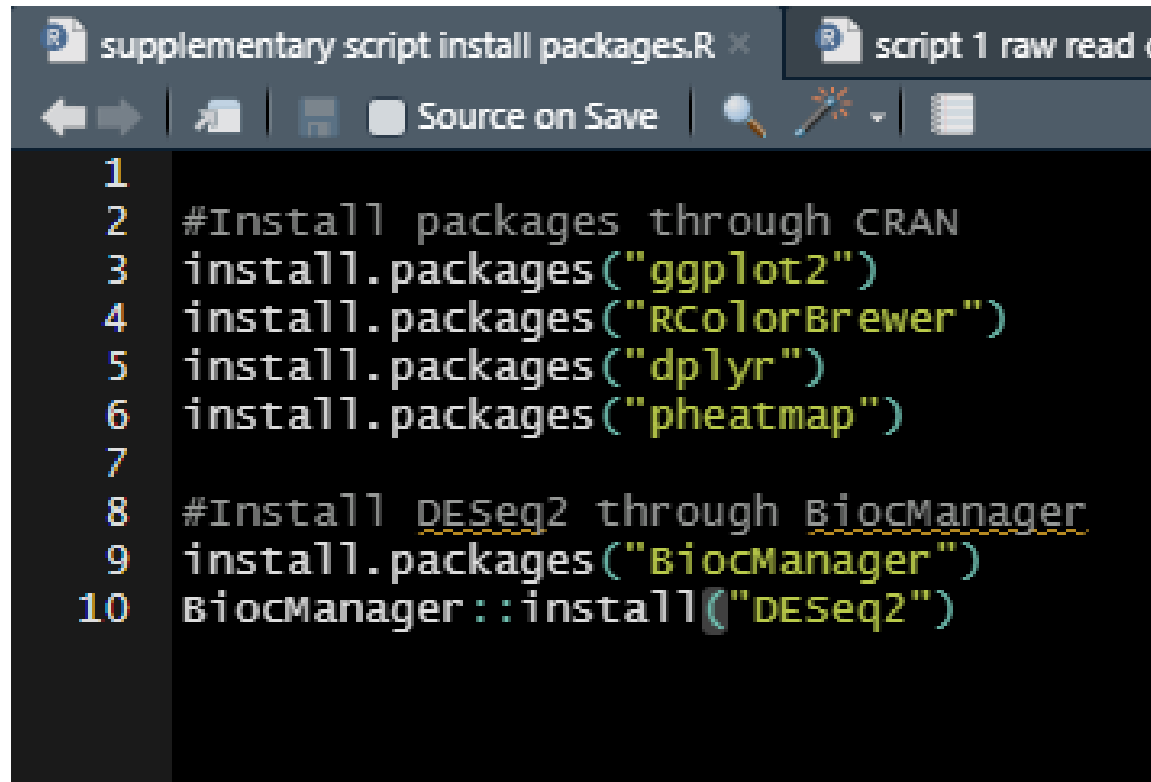
Open R script files into R Studio



Files can be examined and opened in the “Plots and files” pane within R Studio. Open the provided **R scripts** by simply clicking on the file.

Download required R packages

An **R package** is a reproducible unit of R code that usually contains a set of pre-defined functions used to complete a particular type of analysis.

A screenshot of the R Studio interface. The top pane shows two open scripts: 'supplementary script install packages.R' and 'script 1 raw read c'. The bottom pane displays R code for installing packages. The code is as follows:

```
1  
2 #Install packages through CRAN  
3 install.packages("ggplot2")  
4 install.packages("RColorBrewer")  
5 install.packages("dplyr")  
6 install.packages("pheatmap")  
7  
8 #Install DESeq2 through BiocManager  
9 install.packages("BiocManager")  
10 BiocManager::install("DESeq2")
```

Scripts are opened in R studio. Individual lines of code can be run by pressing ``ctrl + enter``

Script 1: Pre-processing raw read count files

```
script 1 raw read count processing.R × script 2 normalization.R × script 3 differential expression analysis.R × script 4 common data visualizations.R ×
1 #miRNA NGS data analysis
2 #
3 #Script 1
4 #
5 #Raw read count processing
6
7 ###Collect all raw read count files and merge into one matrix
8 data_files <- Sys.glob("raw read count files/*.tabular") #store paths for all raw read count files
9 tmp <- list() #create an empty list to store each file
10 for (i in data_files) {
11   x <- gsub(".tabular.*", "", gsub(".*raw read count files/", "", i)) #extract sample name from file name and store in "x"
12   tmp[[X]] <- read.delim(i, row.names = 1, header = FALSE) #load read count file "i"
13   colnames(tmp[[X]]) <- x #rename column with sample name
14   print(x) #print sample name to track progress in console
15 }
16 read_counts <- do.call(cbind, tmp) #do.call function collapses all objects within list into one data frame
17
18 #Clean up read count file
19 read_counts <- read_counts[rowMeans(read_counts)>0,] # remove all transcripts that were not detected
20 read_counts <- read_counts[order(rowMeans(read_counts), decreasing = TRUE),] # order transcripts based on average raw read count
21
22 #Manually make a data frame containing sample information
23 sample_info <- data.frame(sample = colnames(read_counts),
24                           CWD_status = c(rep("Neg", 3), rep("Pos", 3)))
25
26 #Save files for further analysis
27 if (dir.exists("raw data") == FALSE) { dir.create("raw data") }
28 write.csv(read_counts, "raw data/raw_read_counts.csv")
29 write.csv(sample_info, "raw data/sample_info.csv")
30
```

Run each line of code by pressing `ctrl + enter`. Pay attention to what is happening in the console and environment. Files that are loaded in the `environment` of R studio can be examined by clicking on them.

Script 1: Pre-processing raw read count files

Input: (x6)

	V2
hsa-miR-6859-5p	0
hsa-miR-6859-3p	0
hsa-miR-1302	0
hsa-miR-12136	0
hsa-miR-200b-5p	0
hsa-miR-200b-3p	0
hsa-miR-200a-5p	0
hsa-miR-200a-3p	1
hsa-miR-429	0
hsa-miR-6726-5p	0
hsa-miR-6726-3p	0
hsa-miR-6727-5p	0
hsa-miR-6727-3p	0
hsa-miR-6808-5p	0

Output:

	SRR10356125	SRR10356126	SRR10356127	SRR10356142	SRR10356143	SRR10356144
hsa-miR-486-5p	398669	70960	316241	1175615	405887	748382
hsa-miR-191-5p	449776	124957	248597	65468	72379	45908
hsa-miR-423-5p	316543	33840	226684	169837	66163	127436
hsa-miR-142-5p	139973	45982	109468	28867	44975	24944
hsa-miR-22-3p	122713	38722	105241	13749	44072	14978
hsa-miR-25-3p	56498	9574	44564	88626	33123	36732
hsa-miR-16-5p	76560	8317	60443	23148	18099	27028
hsa-miR-92a-3p	61304	14322	41601	32927	18763	12571
hsa-miR-148a-3p	41482	12246	51872	4224	5529	1414
hsa-miR-143-3p	14738	81	96933	92	228	120
hsa-miR-27b-3p	31915	8268	43163	3723	9890	3260
hsa-miR-6529-5p	31697	3540	25879	15659	6682	7850
hsa-miR-192-5p	19760	5274	17513	12242	10896	5767

This script takes the 6 separate read-count files from feature counts and merges them together into a single matrix, which is then saved as a `.csv` file and can be examined in excel.

Script 2: Normalization with DESeq2

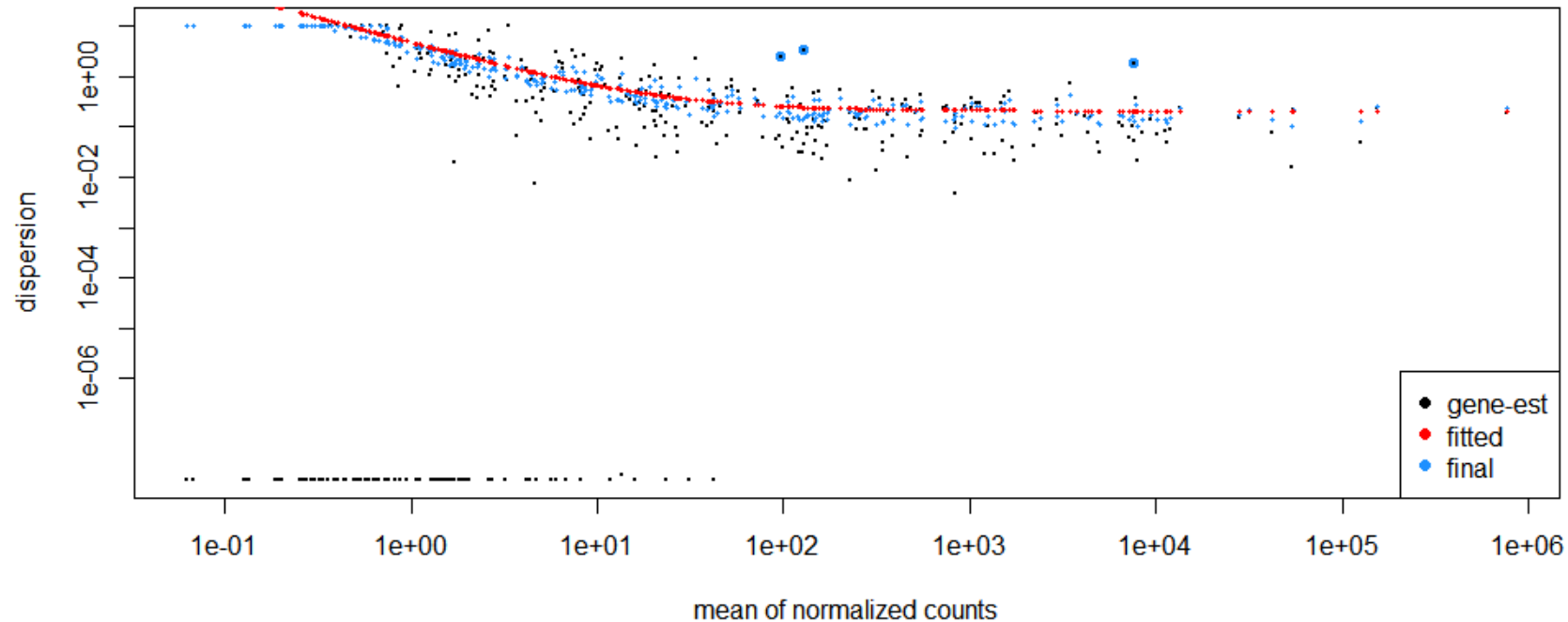
```
1 #miRNA NGS data analysis
2 #
3 #Script 2
4 #
5 #Normalizing the data and assessing variation
6
7
8 library(DESeq2)
9 library(ggplot2)
10
11 #load raw data
12 read_counts <- read.csv("raw data/raw_read_counts.csv", row.names = 1)
13 sample_info <- read.csv("raw data/sample_info.csv", row.names = 1)
14 rownames(sample_info) <- sample_info$sample
15
16 #make sure samples are in order
17 summary(colData(DESeqDataSetFromMatrix(countData, colData, design, tidy = FALSE, ignoreRank =
18 FALSE, ...))
19 #make DESeq data object
20 dds <- DESeqDataSetFromMatrix(countData = read_counts, colData = sample_info, design = ~CWD_status)
21 dds <- DESeq(dds)
22
23 #plot dispersion estimates to examine normalization
24 plotDispEsts(dds)
25
```

Run each line of code by pressing ``ctrl + enter``. Pay attention to what is happening in the console and environment. Files that are loaded in the ``environment`` of R studio can be examined by clicking on them.

The **DESeq2** R package is popular for differential expression analysis of RNAseq data. A full description of how to use DESeq2 can be found at

<http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>

Script 2: Normalization with DESeq2



This plot illustrates the “negative binomial model of gene-fitted dispersion estimates” used by DESeq2 to normalize the data. This is a global normalization method that is used by other RNAseq analysis R packages such as edgeR.

Common data visualizations: PCA (Script 2)

```
#extract normalized read counts
norm_counts <- varianceStabilizingTransformation(dds)

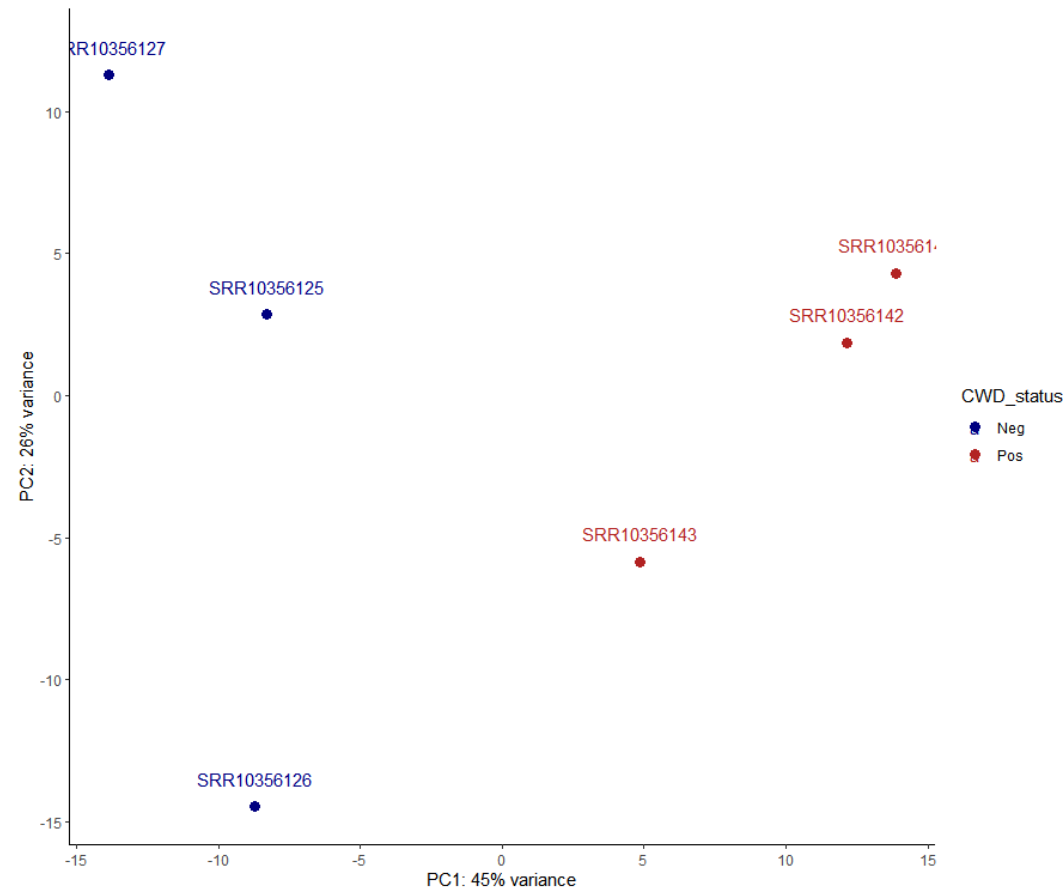
#make a basic PCA plot
plotPCA(norm_counts, intgroup="CWD_status")

#make a custom PCA plot with ggplot
pcaData <- plotPCA(norm_counts, intgroup=c("Sample", "CWD_status"), returnData=TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"))
ggplot(pcaData, aes(PC1, PC2, color=CWD_status, label=Sample)) +
  geom_point(size=3) +
  geom_text(nudge_y = 1) +
  xlab(paste0("PC1: ", percentVar[1], "% variance")) +
  ylab(paste0("PC2: ", percentVar[2], "% variance")) +
  scale_color_manual(values = c("navy", "firebrick")) +
  coord_fixed() +
  theme_classic()

#Save normalized read counts for visualization later
write.csv(assay(norm_counts), "raw data/normalized_read_counts.csv")
```

Run each line of code by pressing `ctrl + enter`. Pay attention to what is happening in the console and environment. Files that are loaded in the `environment` of R studio can be examined by clicking on them.

Common data visualizations: PCA (Script 2)



Principle component analysis, or PCA, is a dimensionality reduction technique used to visualize variation in the data in 2-dimensions. It is often used to examine sources of variation in gene expression data.

Script 3: Differential expression analysis with DESeq2

```
6
7 library(DESeq2)
8
9 #load raw data
10 read_counts <- read.csv("raw data/raw_read_counts.csv", row.names = 1)
11 sample_info <- read.csv("raw data/sample_info.csv", row.names = 1)
12 rownames(sample_info) <- sample_info$Sample
13
14 #make DEseq data object
15 dds <- DESeqDataSetFromMatrix(countData = read_counts, colData = sample_info, design = ~CWD_status)
16 dds <- DESeq(dds)
17
18 #get differential expression results
19 resultsNames(dds)
20 res <- results(object = dds, contrast = c("CWD_status", "Pos", "Neg"))
21
22 #clean up results file
23 res <- res[order(res$padj),]
24 res <- na.omit(res)
25 res <- as.data.frame(res)
26
27 summary(res$padj < 0.05)
28 |
29 #Save differential expression results
30 if (dir.exists("DE results")==FALSE) { dir.create("DE results") }
31 write.csv(res, "DE results/CWD_DE_miRNAs.csv")
32
```

Run each line of code by pressing **`ctrl + enter`**. Pay attention to what is happening in the console and environment. Files that are loaded in the **`environment`** of R studio can be examined by clicking on them.

Script 3: Differential expression analysis with DESeq2

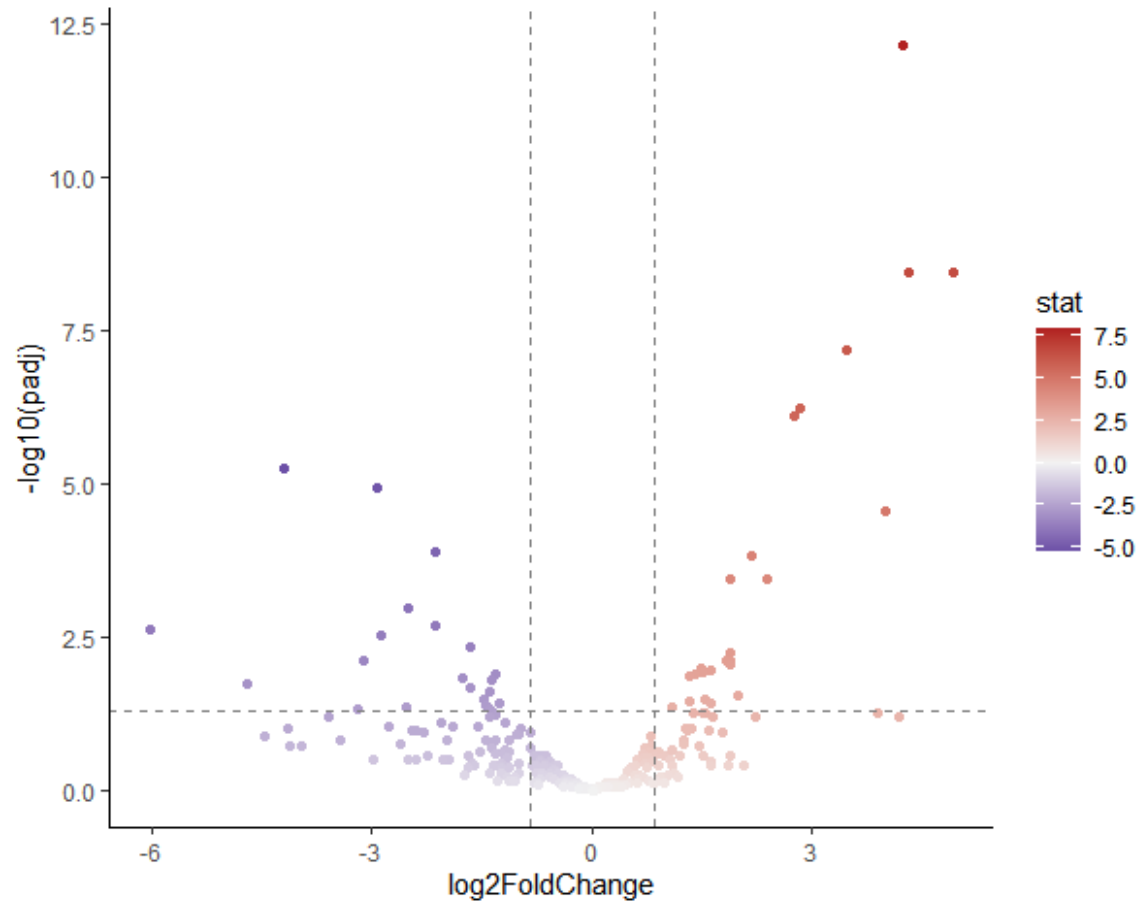
Results file:

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
hsa-miR-144-5p	2.563344e+02	4.2183261	0.5337497	7.9031908	2.718531e-15	6.905068e-13
hsa-miR-375-3p	1.530824e+03	4.3010739	0.6471957	6.6457082	3.017625e-11	3.608444e-09
hsa-miR-185-5p	2.006976e+02	4.9145387	0.7452286	6.5946730	4.261942e-11	3.608444e-09
hsa-miR-486-5p	7.676560e+05	3.4711205	0.5689435	6.1009932	1.054114e-09	6.693623e-08
hsa-miR-107	4.529119e+03	2.8427122	0.4977763	5.7108228	1.124312e-08	5.711507e-07
hsa-miR-103a-3p	9.291409e+03	2.7475877	0.4887584	5.6215661	1.892341e-08	8.010911e-07
hsa-miR-125b-5p	4.267817e+01	-4.1982603	0.8006333	-5.2436741	1.574101e-07	5.711738e-06
hsa-miR-125a-5p	2.253622e+02	-2.9383044	0.5781449	-5.0822977	3.728963e-07	1.183946e-05
hsa-miR-185-3p	2.218884e+01	3.9829770	0.8127359	4.9007024	9.549461e-07	2.695070e-05
hsa-miR-223-3p	1.246713e+03	-2.1411961	0.4693917	-4.5616407	5.075544e-06	1.289188e-04
hsa-miR-199b-5p	1.128220e+02	2.1673227	0.4801270	4.5140616	6.359774e-06	1.468530e-04
hsa-miR-25-3p	5.506976e+04	2.3813877	0.5548757	4.2917498	1.772706e-05	3.463799e-04
hsa-miR-451a	1.008335e+04	1.8675975	0.4351612	4.2917368	1.772810e-05	3.463799e-04
hsa-miR-222-3p	4.586697e+02	-2.5124478	0.6242641	-4.0246550	5.705890e-05	1.035212e-03
hsa-miR-99b-5p	2.297306e+02	-2.1312760	0.5546841	-3.8423238	1.218749e-04	2.063748e-03
hsa-miR-143-3p	7.673291e+03	-6.0147614	1.5887402	-3.7858684	1.531728e-04	2.288581e-03

Common data visualizations: Volcano plot (Script 4)

```
6
7 library(ggplot2)
8 library(RColorBrewer)
9 library(pheatmap)
10 library(dplyr)
11
12 #The volcano plot
13 #Load differential expression results from 150 dpi
14 res <- read.csv("DE results/CWD_DE_miRNAs.csv")
15
16 #a basic volcano plot
17 ggplot(res, aes(x=log2FoldChange, y=-log10(padj))) +
18   geom_point()
19
20 #a nicer volcano plot
21 ggplot(res, aes(x=log2FoldChange, y=-log10(padj), color=stat)) +
22   geom_point() +
23   geom_hline(yintercept = -log10(0.05), linetype="dashed", color="grey50") +
24   geom_vline(xintercept = 0.85, linetype="dashed", color="grey50") +
25   geom_vline(xintercept = -0.85, linetype="dashed", color="grey50") +
26   scale_color_gradient2(low = "navy", high = "firebrick", mid="grey95", midpoint = 0) +
27   theme_classic()
28
```

Common data visualizations: Volcano plot (Script 4)

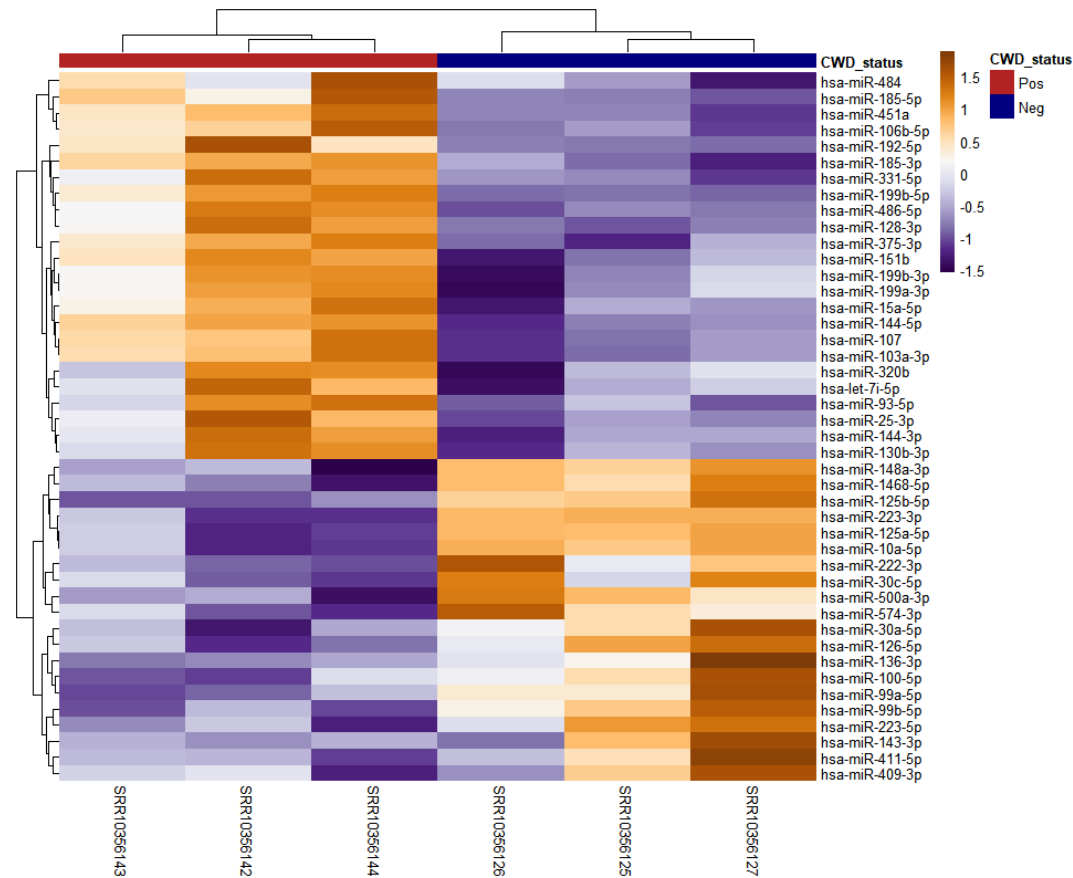


A volcano plot show statistical significance (p-value) versus magnitude (fold change) and can quickly show changes in large datasets

Common data visualizations: Heatmap (Script 4)

```
29
30 #The heatmap
31 #Identify DE miRNAs
32 miRNAs <- read.csv("DE results/CWD_DE_miRNAs.csv") %>% filter(padj < 0.05, abs(log2FoldChange) > 0.85, baseMean > 15) %>% pull(x)
33
34 #we will use normalized read-counts to calculate z-scores
35 zscores <- read.csv("raw data/normalized_read_counts.csv", row.names = 1)
36 zscores <- as.matrix(zscores[miRNAs,])
37 zscores <- (zscores-rowMeans(zscores))/matrixStats::rowSds(zscores)
38
39 #basic heatmap with hierachichal clustering
40 pheatmap(zscores)
41
42 #nicer heatmap
43
44 #specify additional variables required by pheatmap
45 plot_colors <- rev(colorRampPalette(brewer.pal(11,"PuOr"))(100))
46 column_annotation <- read.csv("raw data/sample_info.csv", row.names = 1)
47 column_annotation <- data.frame(row.names = column_annotation$sample,
48                                CWD_status = column_annotation$CWD_status)
49 annotation_colors <- list(`CWD_status`=c(`Pos`="firebrick", `Neg`="navy"))
50
51 pheatmap(zscores, color = plot_colors, annotation_col = column_annotation, annotation_colors = annotation_colors,
52          treeheight_row = 25, treeheight_col = 25, border_color = FALSE)
53
```

Common data visualizations: Heatmap (Script 4)



Heatmaps are commonly used to examine gene expression across samples. Hierarchical clustering of genes and samples can make it easier to identify patterns in the gene expression data.

Further analysis

- Which genes are targeted by these miRNAs?
- miRNA target prediction online tools:
 - http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=tarbasev8%2Findex
 - <http://snf-515788.vm.oceanos.grnet.gr/>
 - <http://mirdb.org/>
 - http://www.targetscan.org/vert_80/

Summary

- **Part I: Preprocessing in Galaxy**
 - Assessing sequencing read quality with FastQC
 - Removing sequencing adapters with Cutadapt
 - Cleaning sequencing reads with Trimmomatic
 - Aligning to reference genome with Bowtie2
 - Mapping reads to miRNAs and counting with FeatureCounts
- **Part II: Downstream Analysis with R**
 - Raw reads count processing
 - Normalization and differential expression analysis with DESeq2
 - Examples of common data visualizations