Script-One: Sample Data Collection

INSDC:

The International Nucleotide Sequence Database Collaboration (INSDC) is a database system to collect and disseminate DNA and RNA sequence databases. It involves the following computerized databases:

- DNA Data Bank of Japan (Japan)
- GenBank (USA)
- European Nucleotide Archive (UK)

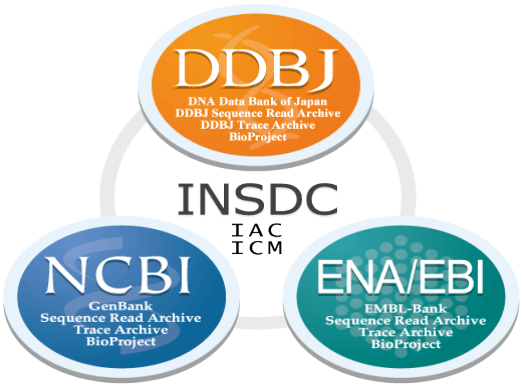
DNA Data Bank of Japan (DDBJ): The DDBJ is a biological database that collects DNA sequences. It is located at the National Institute of Genetics (NIG) in Japan. It exchanges its data with European Molecular Biology Laboratory at the European Bioinformatics Institute and with GenBank at the National Center for Biotechnology Information (NCBI).

GenBank (USA): The GenBank database is an open access, annotated collection of all publicly available nucleotide sequence and their protein translations. It is produced and maintained by the National Center for Biotechnology Information (NCBI). NCBI is a part of the National Institutes of Health in the United States as a part of the INSDC.

European Nucleotide Archive (ENA): The ENA is a repository, providing free and unrestricted access to annotated DNA and RNA sequences. The archive is composed of three main databases:

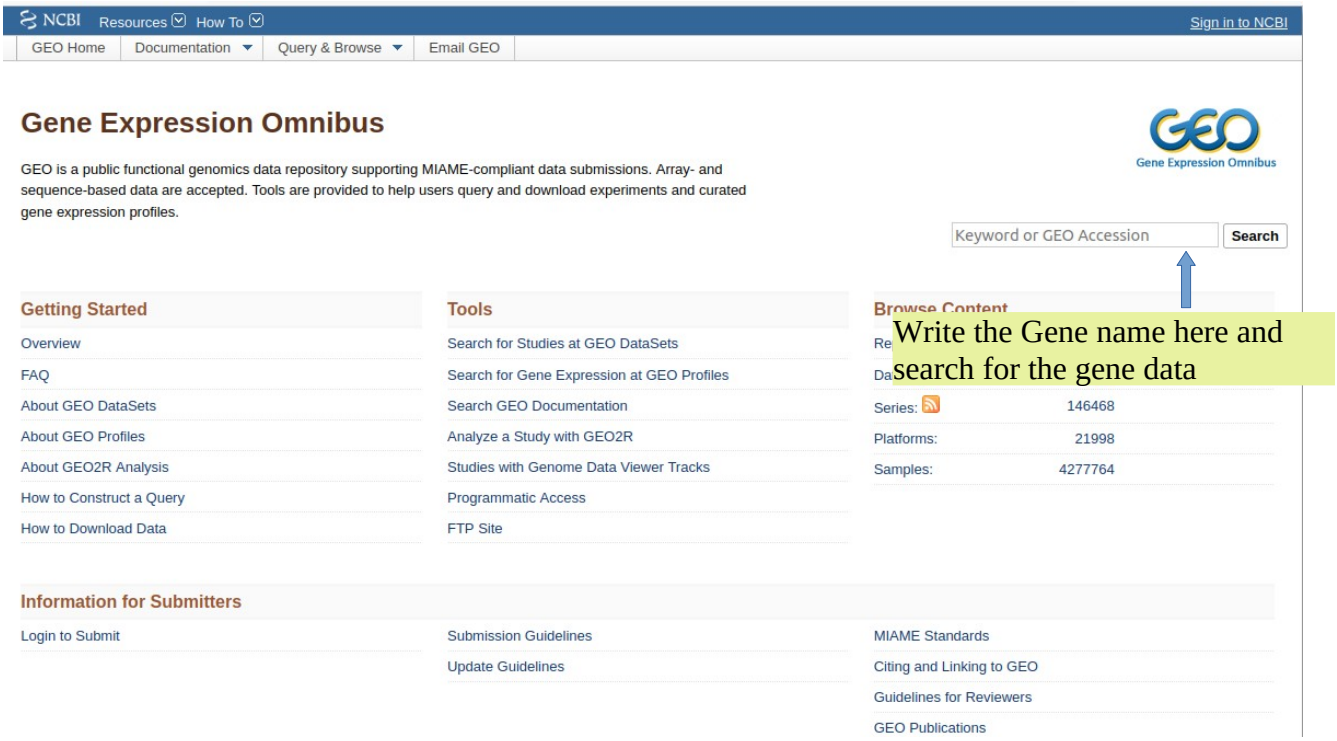
1. The Sequence Read Archive
2. The Trace Archive
3. The EMBL (European Molecular Biology Laboratory) Nucleotide Sequence Database (also known as EMBL-bank)

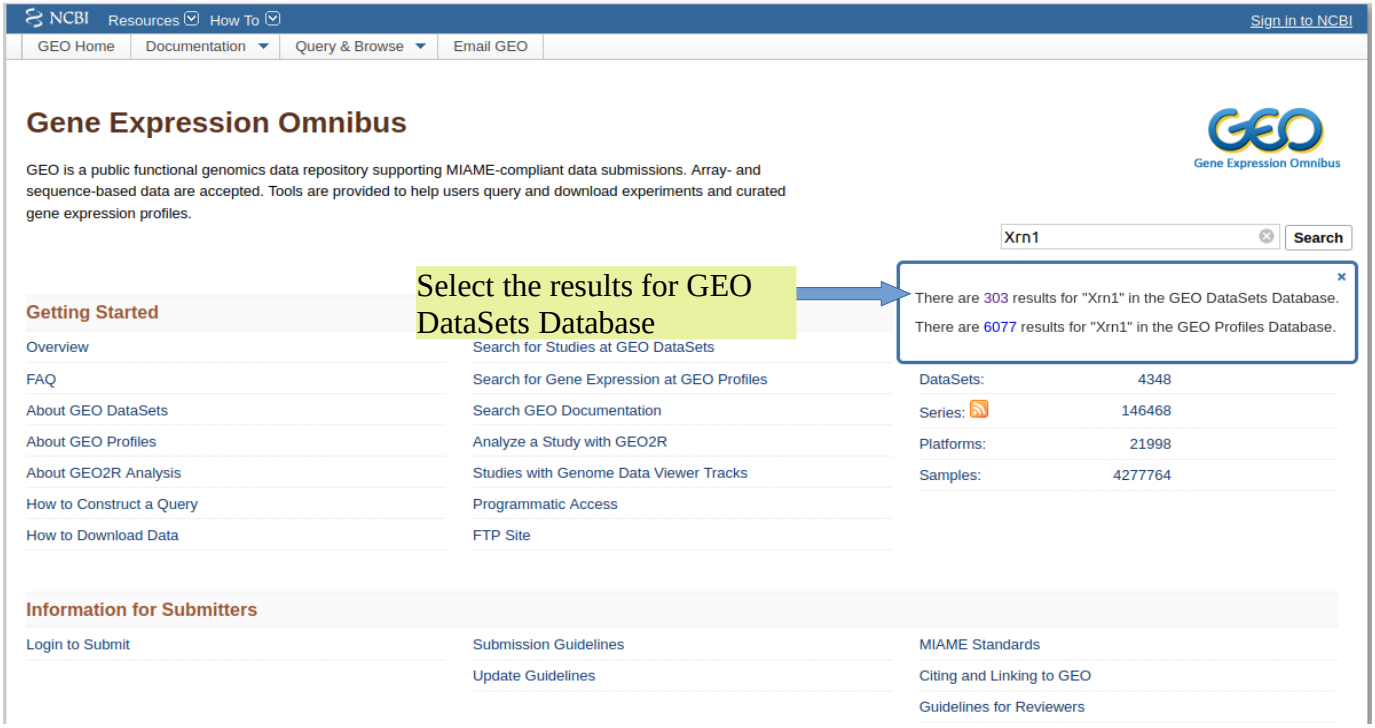
The ENA is produced and maintained by the European Bioinformatics Institute (EBI).



Collection of Data from NCBI:

Gene Expression Omnibus (GEO) is a database supported by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) that accepts raw and processed data with written descriptions of experimental design, sample attributes, and methodology for studies of high-throughput gene expression and genomics.

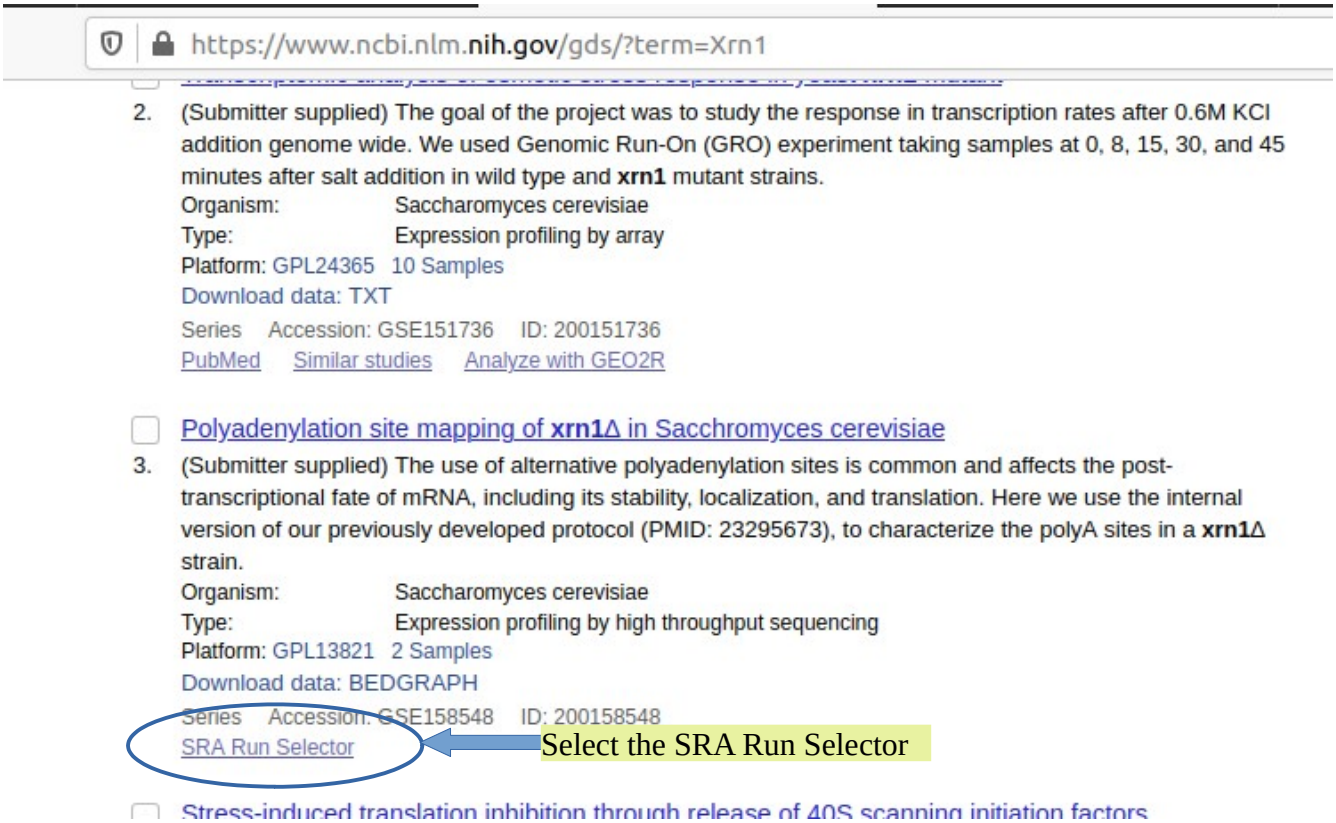




Metadata Preparation :

For the preparation of Metadata sheet by the NCBI data, first of all the data have to be collected from the GEO database. The data sheet has to contained the Gene names, GEO accessions, SRR numbers, organism names, library types (single or pair-end), data types (for example., RNA Seq), reference of the papers, strains, genotypes and types (wild or mutated).

1. The gene which has searched and used for collecting data has to be written in the first column of the sheet.
2.For collecting other information such as GEO Accession, SRR Number, organism Name, library type (single or pair-end), data type, reference of the paper, strain, genotype and type search the gene name in the GEO database. Then select the only experiments with SRA Run Selector.



Sequence Read Archive (SRA): The Sequence Read Archive (SRA, previously known as the Short Read Archive) is a bioinformatics database that provides a public repository for DNA sequencing data, especially the "short reads" generated by high-throughput sequencing, which are typically less than 1,000 base pairs in length. The archive accepts the data from all branches of life.

SRA Run: SRA run accession is an object that contains actual sequencing data for a particular sequencing experiment. The SRA Run Selector is used to download or analyze SRA data with SRA Toolkit. The SRA Toolkit from NCBI is a collection of tools and libraries for using data in the INSDC Sequence Read Archives. The SRA Toolkit contains a series of independent data-“dump” utilities that allow to convert SRA data into different file formats. For example, fastq-dump, sam-dump, sff-dump, abi-dump, illumina-dump and vdb-dump.

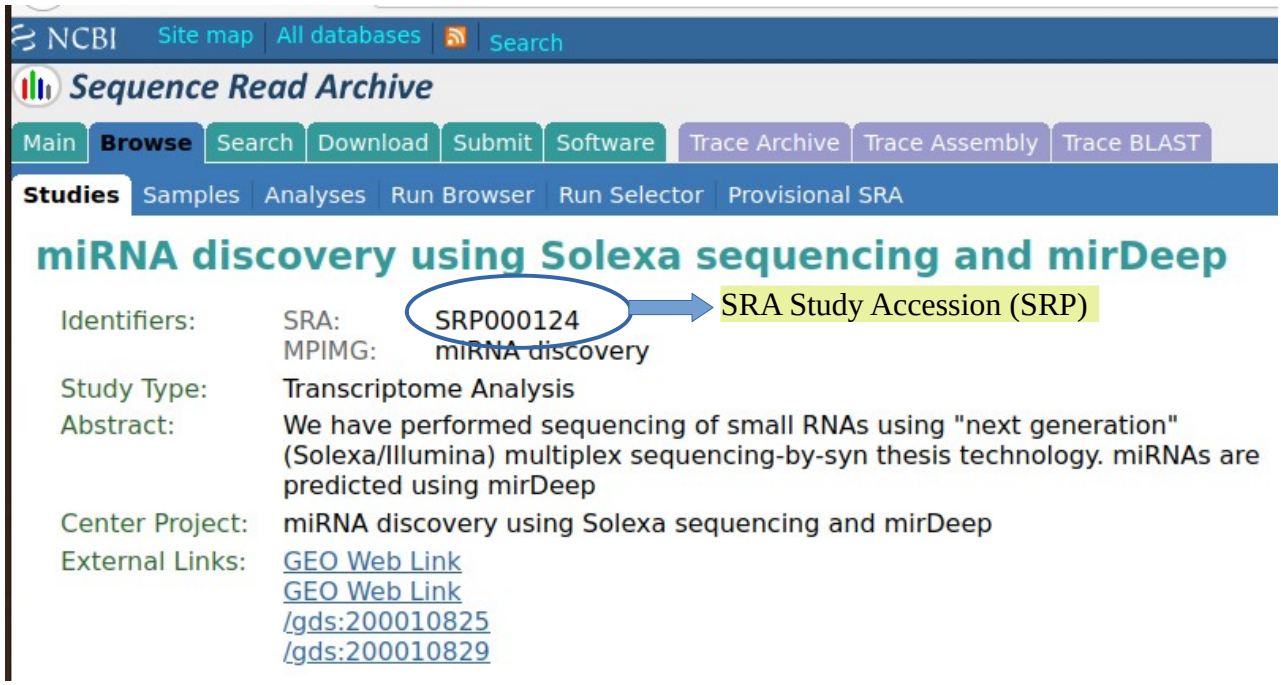
There are six different SRA Accession Types:

1. SRA: The SRA Submission Accession represents a virtual container that holds the objects represented by the other five accessions and is used to track the submission in the archive.

Example: Since the SRA accession number is an artificial packaging construct, there is no example available since the SRA accession number has no specific response page

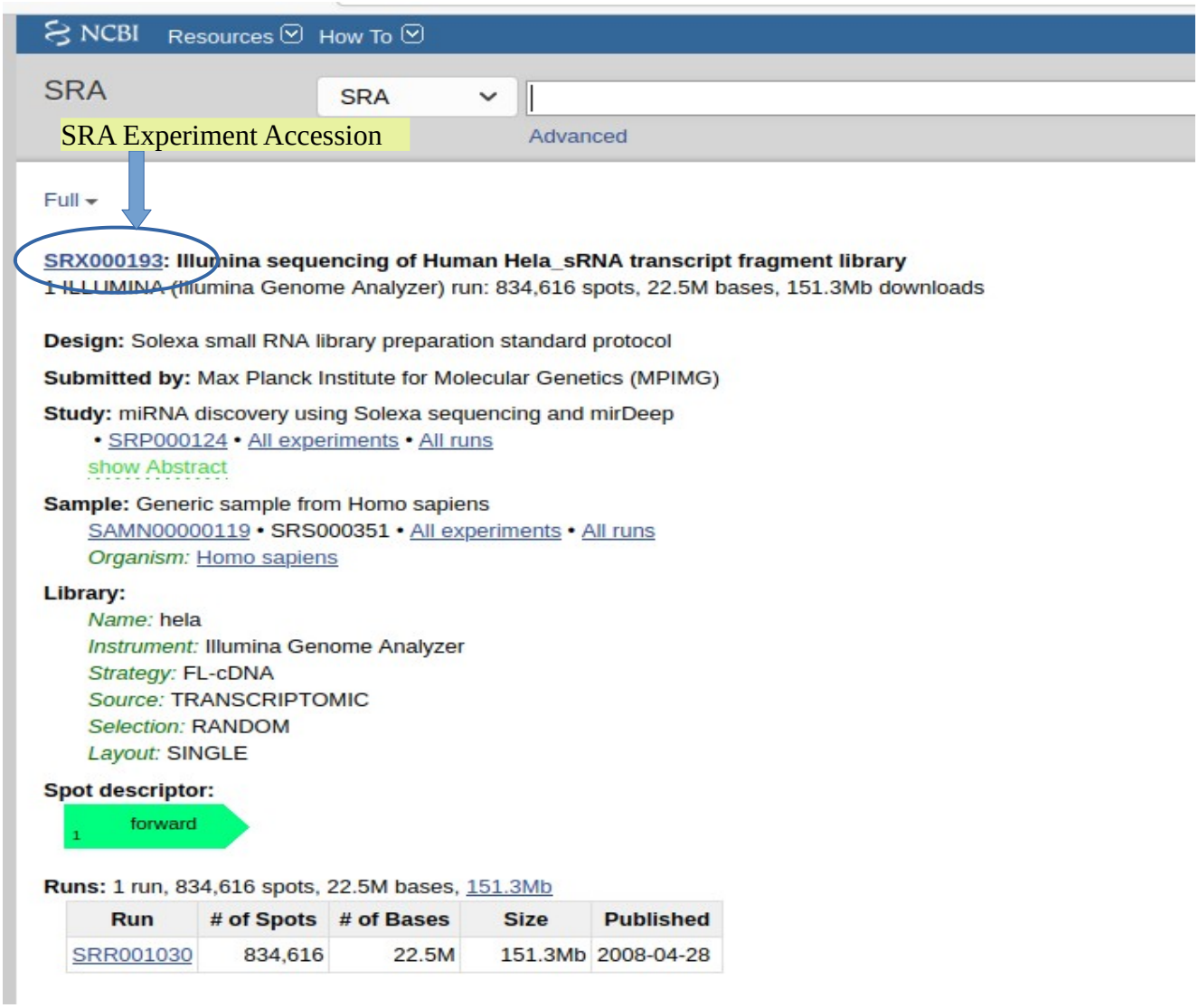
2. SRP: The SRA Study Accession is an object that contains the project metadata describing a sequencing study or project imported from BioProject.

Example:



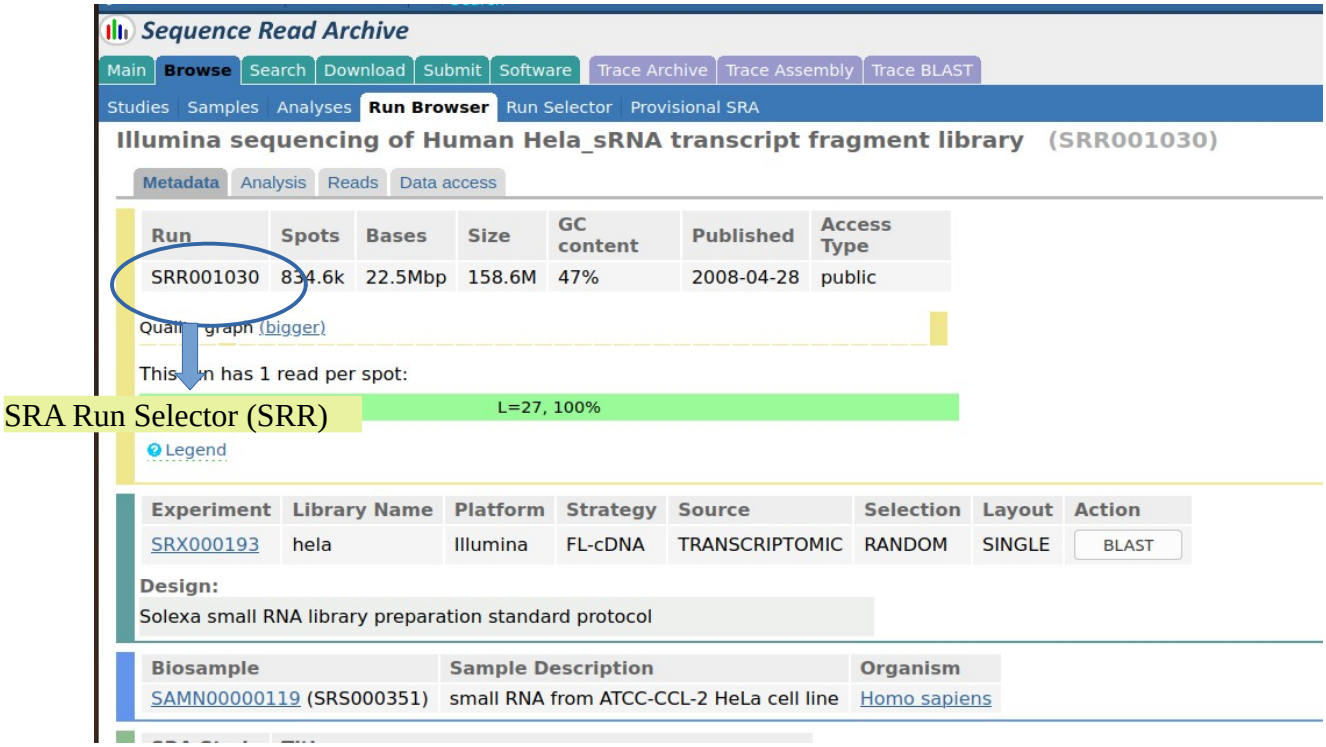
3. SRX: SRA Experiment Accession is an object that contains the metadata describing the library, platform selection, and processing parameters involved in a particular sequencing experiment.

Example:



4. SRR: SRA Run Accession is an object that contains actual sequencing data for a particular sequencing experiment. Experiments may contain many Runs depending on the number of sequencing instrument runs that were needed.

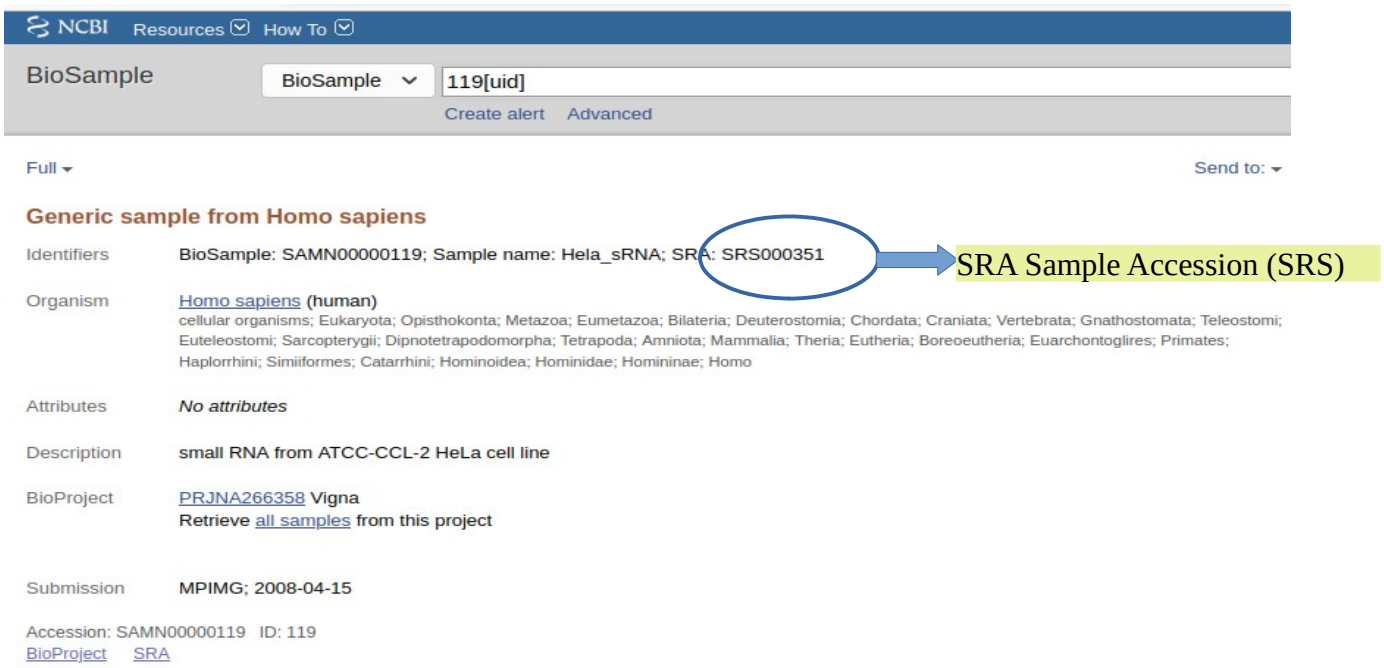
Example:



SRA Run Selector (SRR)

5. SRS: SRA Sample Accession is an object that contains the metadata describing the physical sample upon which a sequencing experiment was performed. Imported from BioSample.

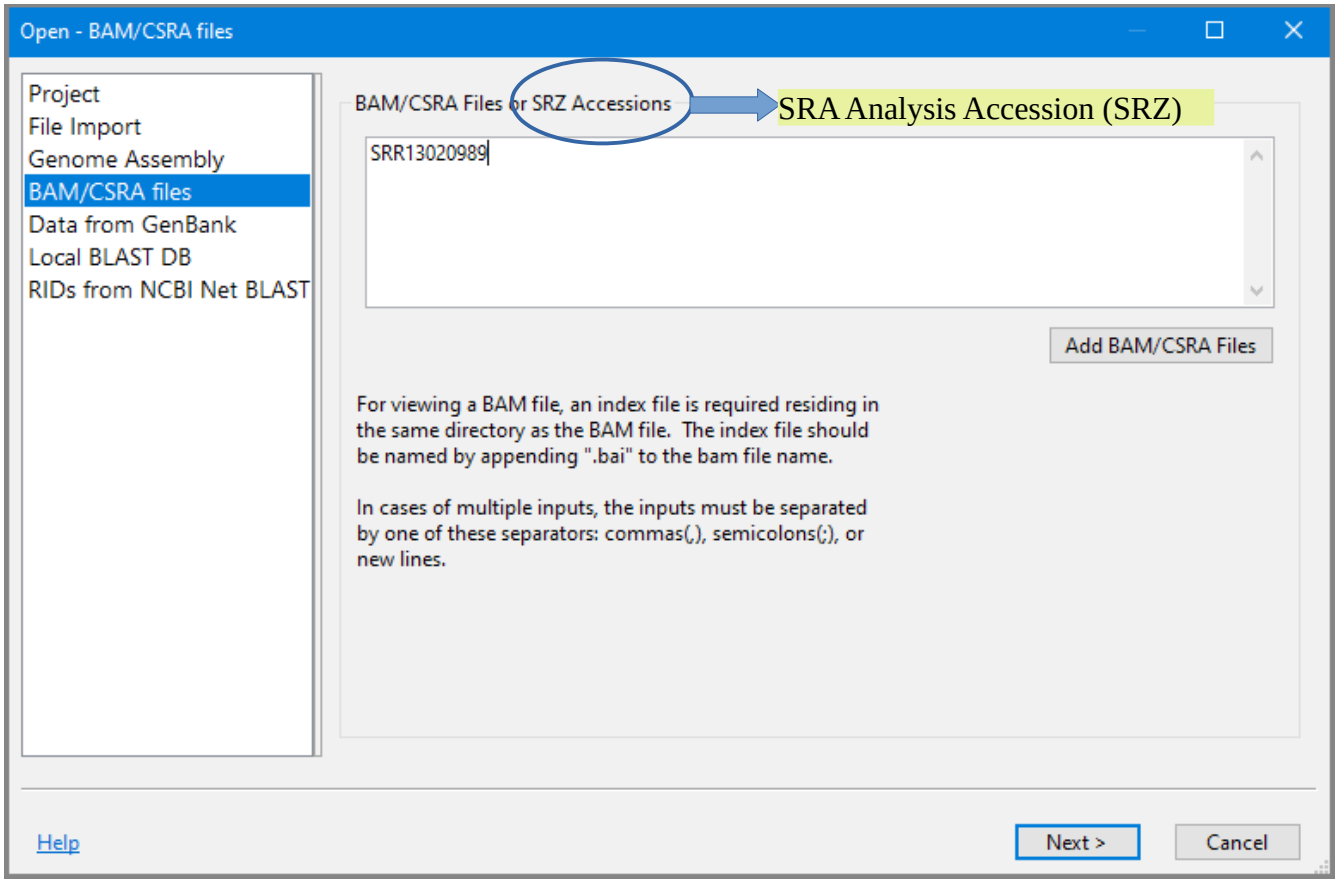
Example:

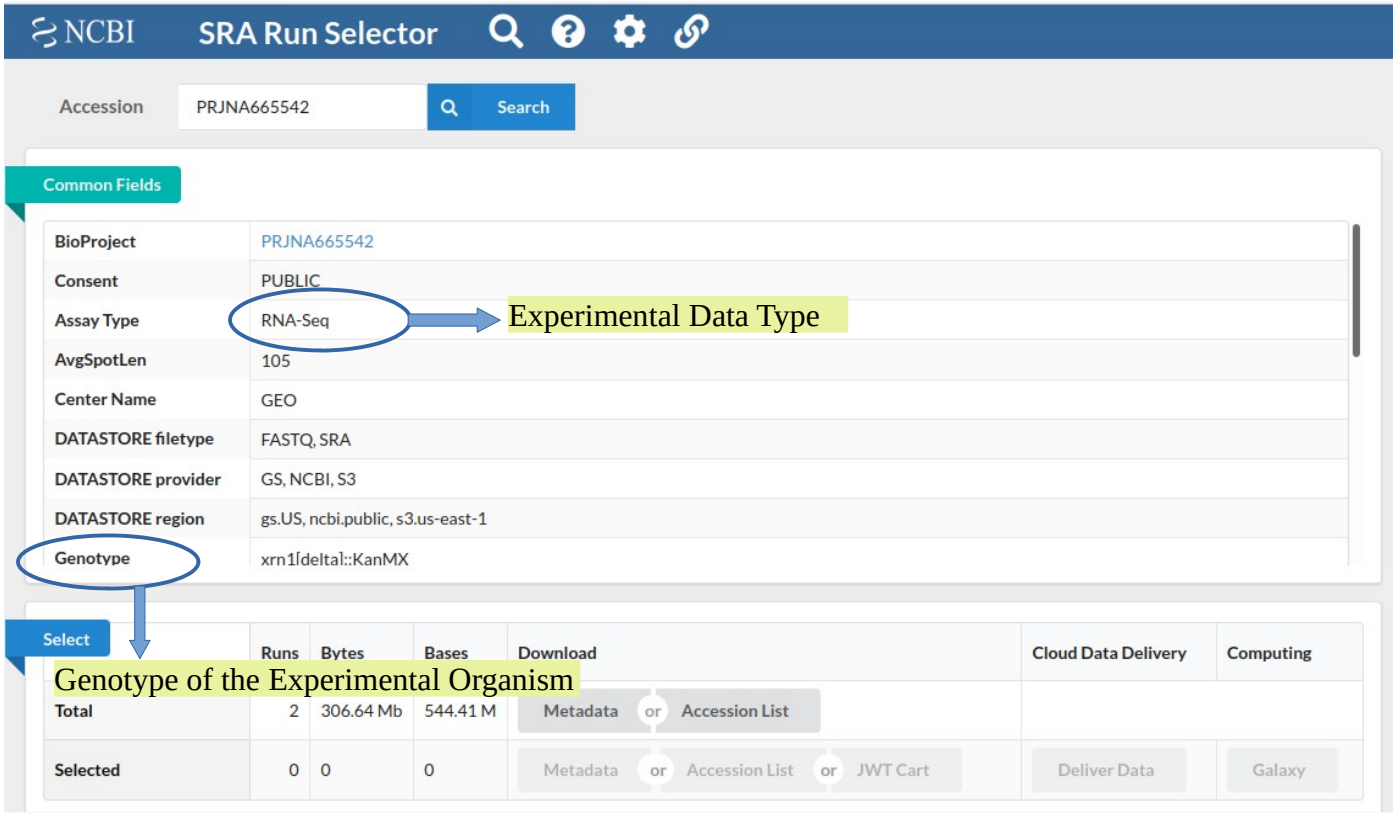


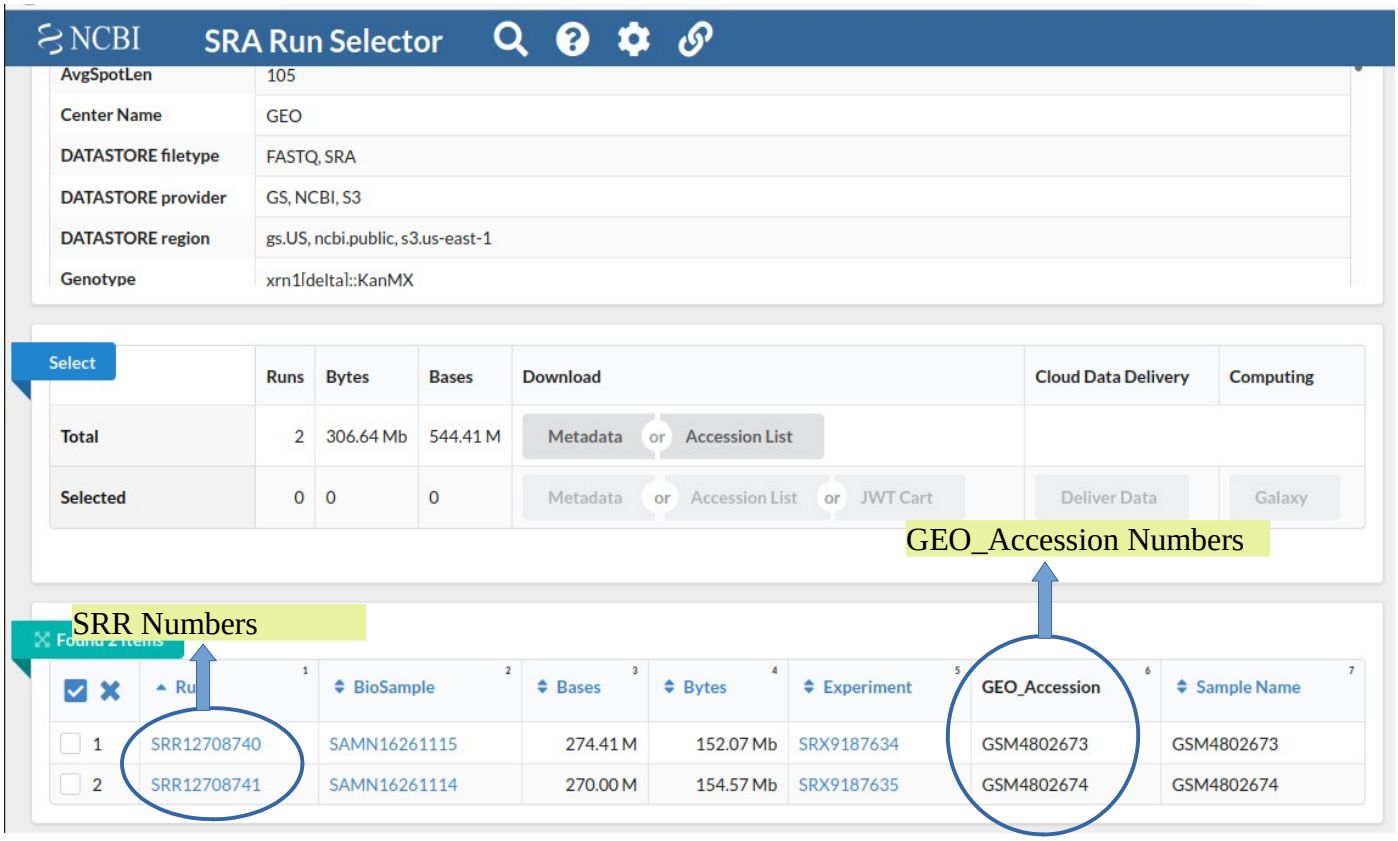
SRA Sample Accession (SRS)

6. SRZ: SRA Analysis Accession is an object that contains a sequence data analysis of BAM file and the metadata describing the sequence analysis.

Example:




NCBI
SRA Run Selector
Accession
PRJNA665542
Search
Common Fields
BioProject
PRJNA665542
Consent
PUBLIC
Assay Type
RNA-Seq
Experimental Data Type
AvgSpotLen
105
Center Name
GEO
DATASTORE filetype
FASTQ, SRA
DATASTORE provider
GS, NCBI, S3
DATASTORE region
gs.US, ncbi.public, s3.us-east-1
Genotype
xrn1[delta]::KanMX
Select
Genotype of the Experimental Organism
Runs
Bytes
Bases
Download
Cloud Data Delivery
Computing
Total
2
306.64 Mb
544.41 M
Metadata
or
Accession List
Selected
0
0
0
Metadata
or
Accession List
or
JWT Cart
Deliver Data
Galaxy


NCBI
SRA Run Selector
AvgSpotLen
105
Center Name
GEO
DATASTORE filetype
FASTQ, SRA
DATASTORE provider
GS, NCBI, S3
DATASTORE region
gs.US, ncbi.public, s3.us-east-1
Genotype
xrn1[delta]::KanMX
Select
Runs
Bytes
Bases
Download
Cloud Data Delivery
Computing
Total
2
306.64 Mb
544.41 M
Metadata
or
Accession List
Selected
0
0
0
Metadata
or
Accession List
or
JWT Cart
Deliver Data
Galaxy
GEO_Accession Numbers
SRR Numbers
BioSample
Bases
Bytes
Experiment
GEO_Accession
Sample Name
1
SRR12708740
SAMN16261115
274.41 M
152.07 Mb
SRX9187634
GSM4802673
GSM4802673
2
SRR12708741
SAMN16261114
270.00 M
154.57 Mb
SRX9187635
GSM4802674
GSM4802674

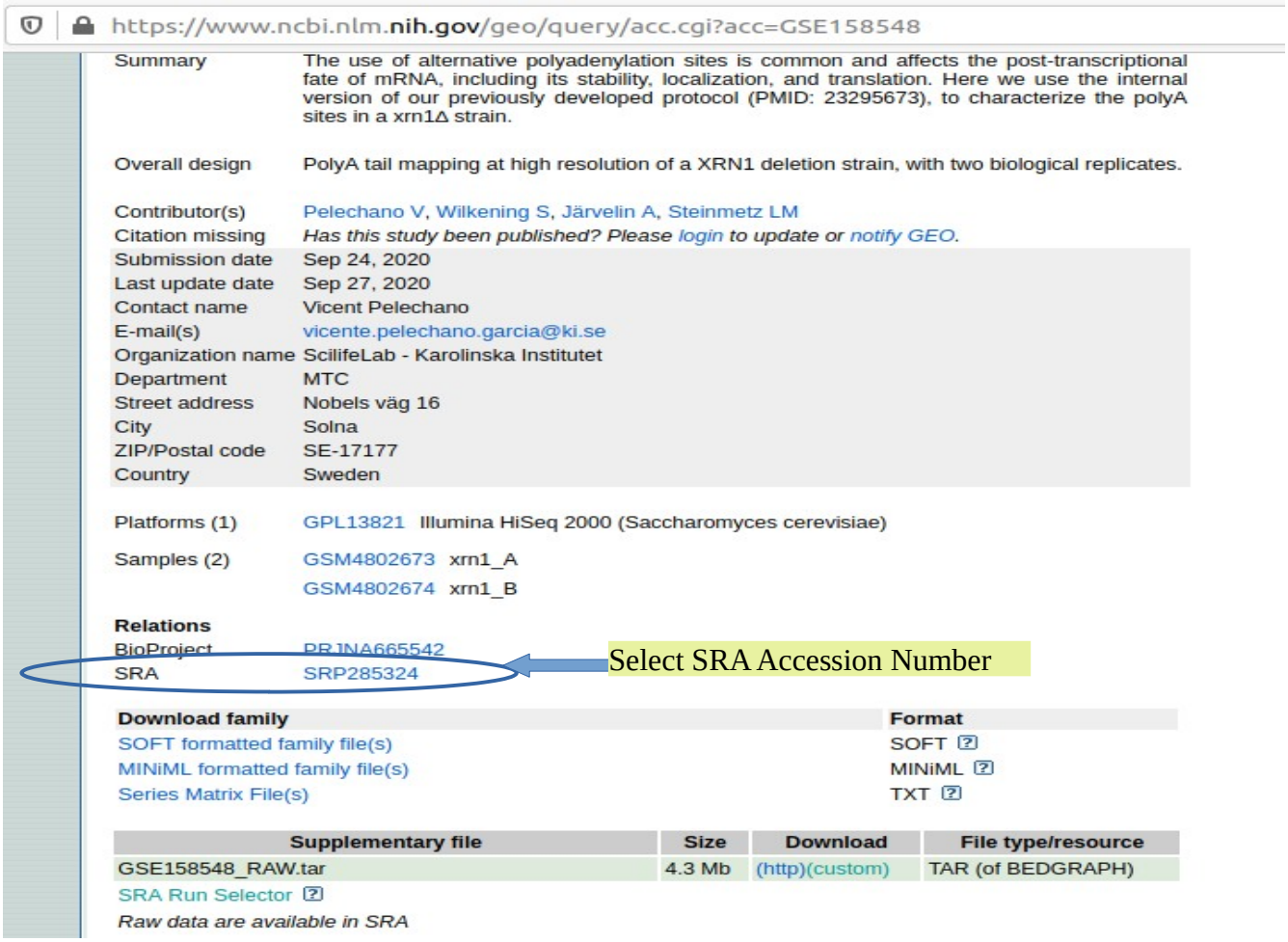
- For getting enough and details information select the paper of the Experiment that has SRA Run Selector.

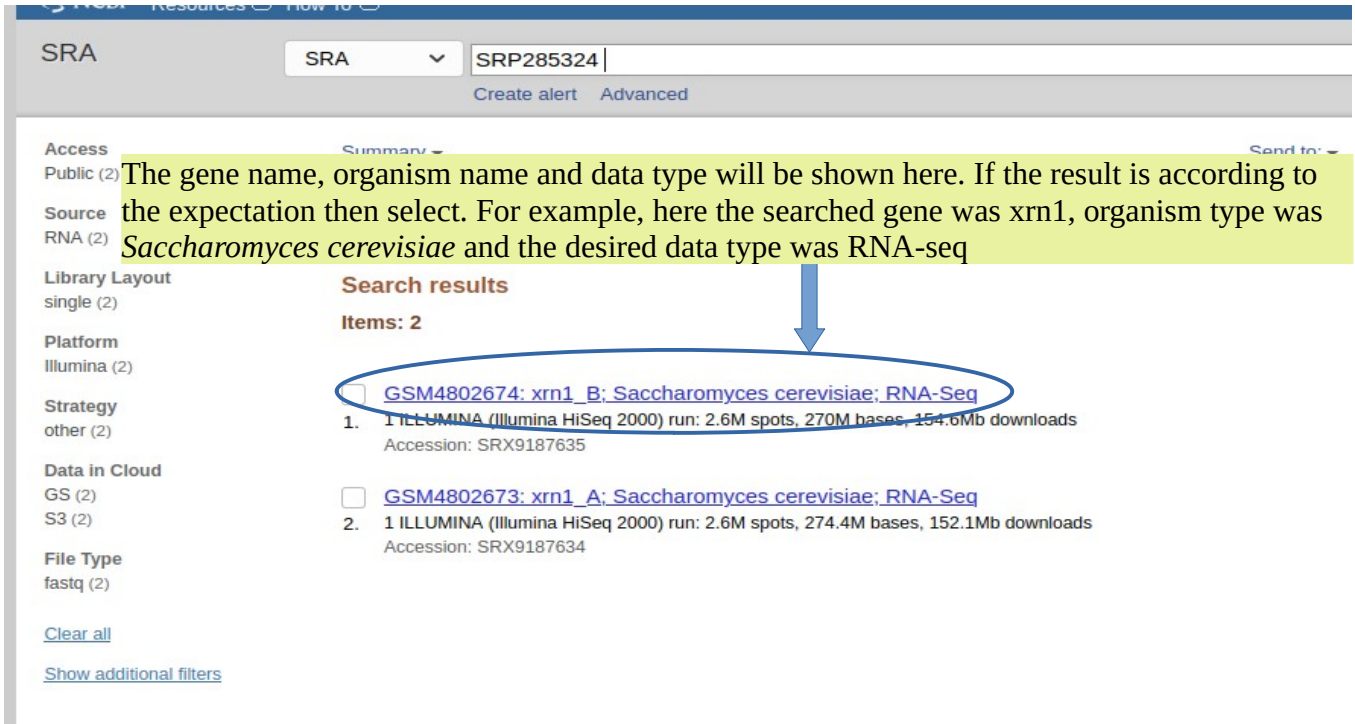
2. (Submitter supplied) The goal of the project was to study the response in transcription rates after 0.6M KCl addition genome wide. We used Genomic Run-On (GRO) experiment taking samples at 0, 8, 15, 30, and 45 minutes after salt addition in wild type and **xrn1** mutant strains.
Organism: Saccharomyces cerevisiae
Type: Expression profiling by array
Platform: GPL24365 10 Samples
Download data: TXT
Series Accession: GSE151736 ID: 200151736
PubMed Similar studies Analyze with GEO2R

☐ Polyadenylation site mapping of **xrn1Δ** in Sacchromyces cerevisiae → Experiment Paper

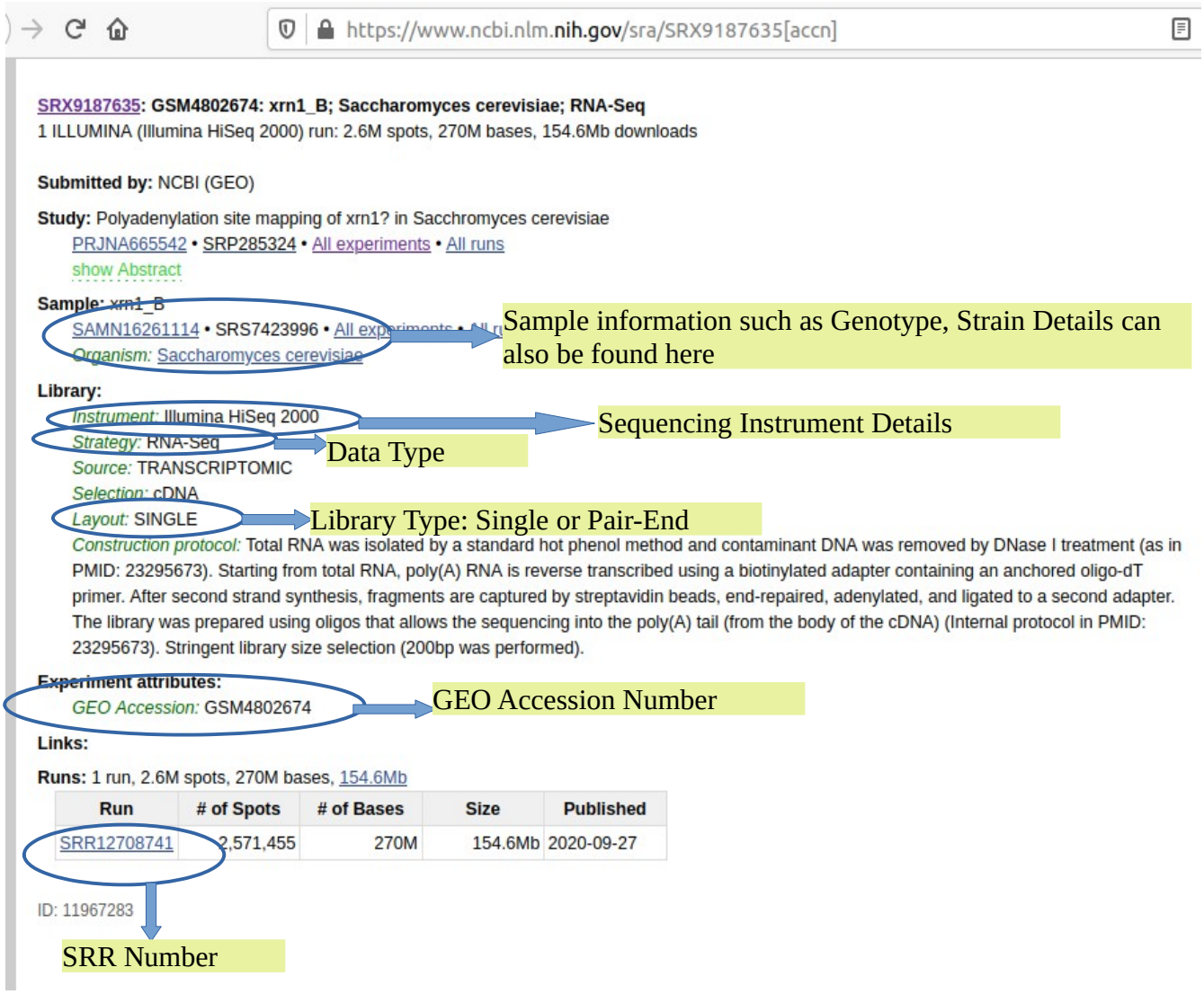
3. (Submitter supplied) The use of alternative polyadenylation sites is common and affects the post-transcriptional fate of mRNA, including its stability, localization, and translation. Here we use the internal version of our previously developed protocol (PMID: 23295673), to characterize the polyA sites in a **xrn1Δ** strain.
Organism: Saccharomyces cerevisiae
Type: Expression profiling by high throughput sequencing
Platform: GPL13821 2 Samples
Download data: BEDGRAPH
Series Accession: GSE158548 ID: 200158548
SRA Run Selector

→ Experiment details after clicking on the given paper


https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158548
Summary
The use of alternative polyadenylation sites is common and affects the post-transcriptional fate of mRNA, including its stability, localization, and translation. Here we use the internal version of our previously developed protocol (PMID: 23295673), to characterize the polyA sites in a xrn1Δ strain.
Overall design
PolyA tail mapping at high resolution of a XRN1 deletion strain, with two biological replicates.
Contributor(s)
Pelechano V, Wilkening S, Järvelin A, Steinmetz LM
Citation missing
Has this study been published? Please login to update or notify GEO.
Submission date
Sep 24, 2020
Last update date
Sep 27, 2020
Contact name
Vicent Pelechano
E-mail(s)
vicente.pelechano.garcia@ki.se
Organization name
ScilifeLab - Karolinska Institutet
Department
MTC
Street address
Nobels väg 16
City
Solna
ZIP/Postal code
SE-17177
Country
Sweden
Platforms (1)
GPL13821 Illumina HiSeq 2000 (Saccharomyces cerevisiae)
Samples (2)
GSM4802673 xrn1_A
GSM4802674 xrn1_B
Relations
BioProject
PRJNA665542
SRA
SRP285324
Select SRA Accession Number
Download family
Format
SOFT formatted family file(s)
SOFT
MINiML formatted family file(s)
MINiML
Series Matrix File(s)
TXT
Supplementary file
Size
Download
File type/resource
GSE158548_RAW.tar
4.3 Mb
(http)(custom)
TAR (of BEDGRAPH)
SRA Run Selector
Raw data are available in SRA

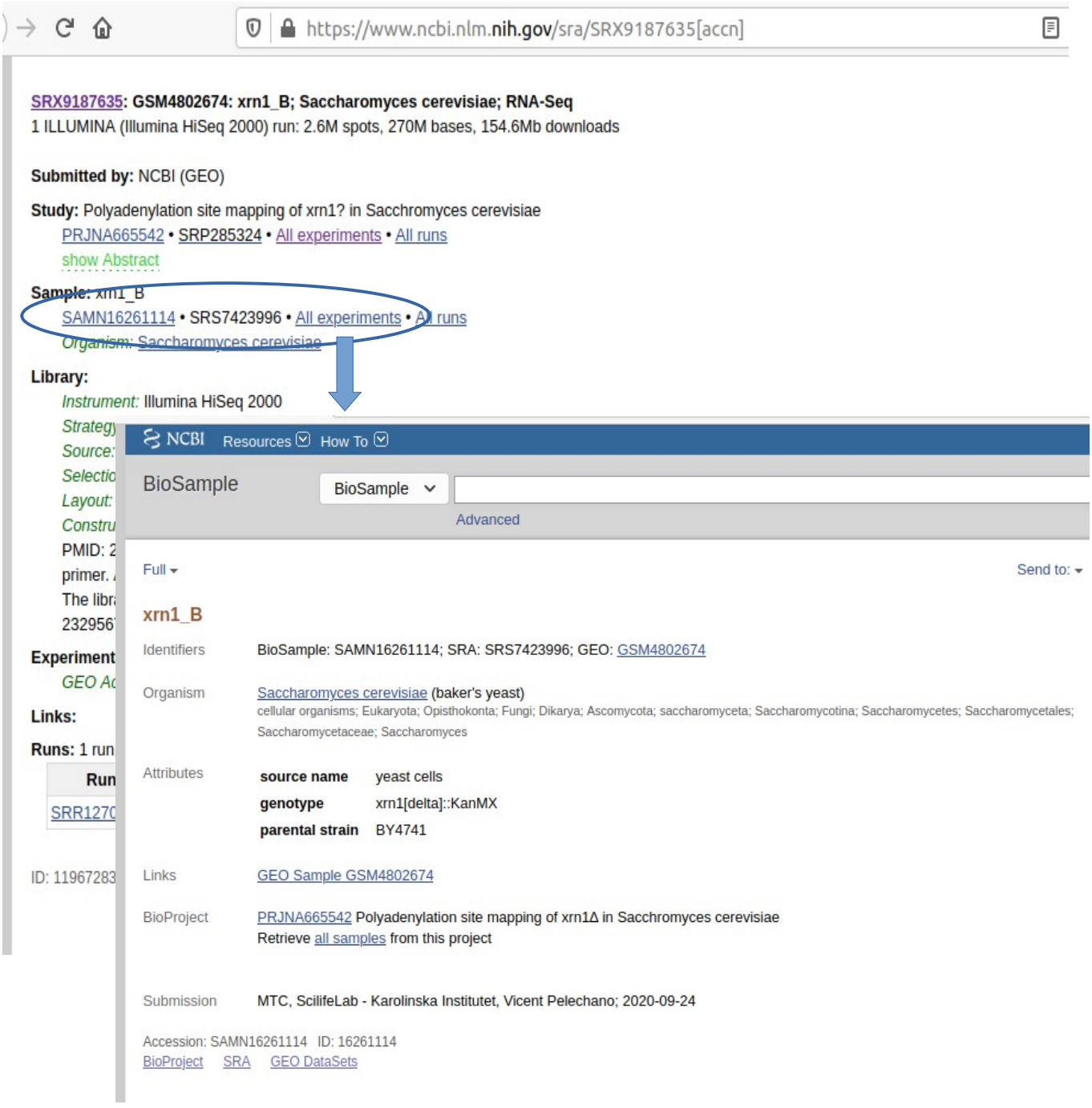

SRA
SRA
SRP285324
Create alert
Advanced
Access
Public (2)
Source
RNA (2)
Library Layout
single (2)
Platform
Illumina (2)
Strategy
other (2)
Data in Cloud
GS (2)
S3 (2)
File Type
fastq (2)
Clear all
Show additional filters
The gene name, organism name and data type will be shown here. If the result is according to the expectation then select. For example, here the searched gene was xrn1, organism type was Saccharomyces cerevisiae and the desired data type was RNA-seq
Search results
Items: 2
GSM4802674: xrn1_B; Saccharomyces cerevisiae; RNA-Seq
1. 1 ILLUMINA (Illumina HiSeq 2000) run: 2.6M spots, 270M bases, 154.6Mb downloads
Accession: SRX9187635
GSM4802673: xrn1_A; Saccharomyces cerevisiae; RNA-Seq
2. 1 ILLUMINA (Illumina HiSeq 2000) run: 2.6M spots, 274.4M bases, 152.1Mb downloads
Accession: SRX9187634

After selecting the link of the circle the additional information will be appeared here:

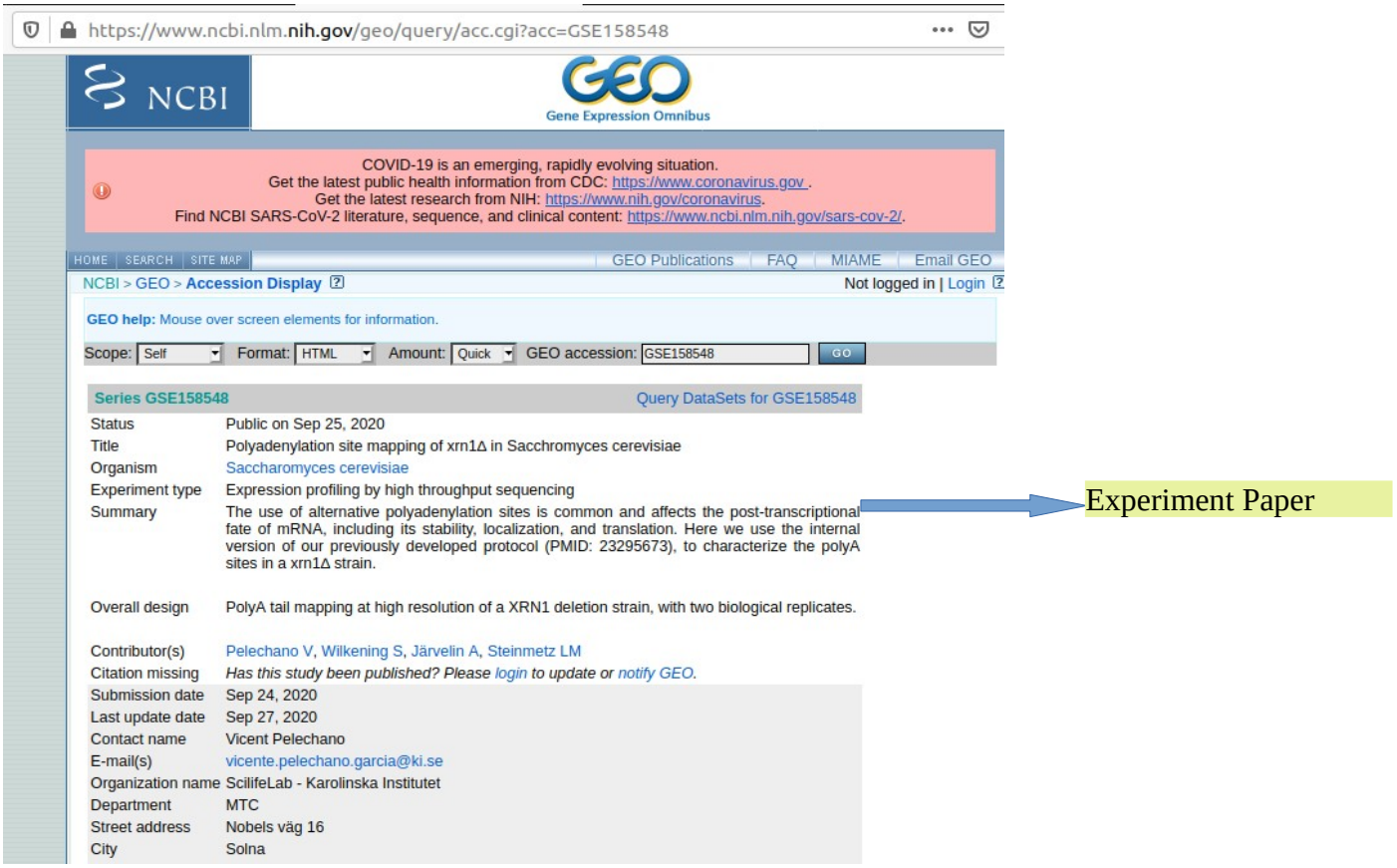


*****GEO Dataset is a study based database which contains Platform, Samples, Series and the corrected data from the curators. These curated dataset form the root of GEO's advanced data display and analysis features, with gene expression level identifying tools and cluster heat maps. These dataset are assigned with a unique and stable GEO accession number GDSxxx.

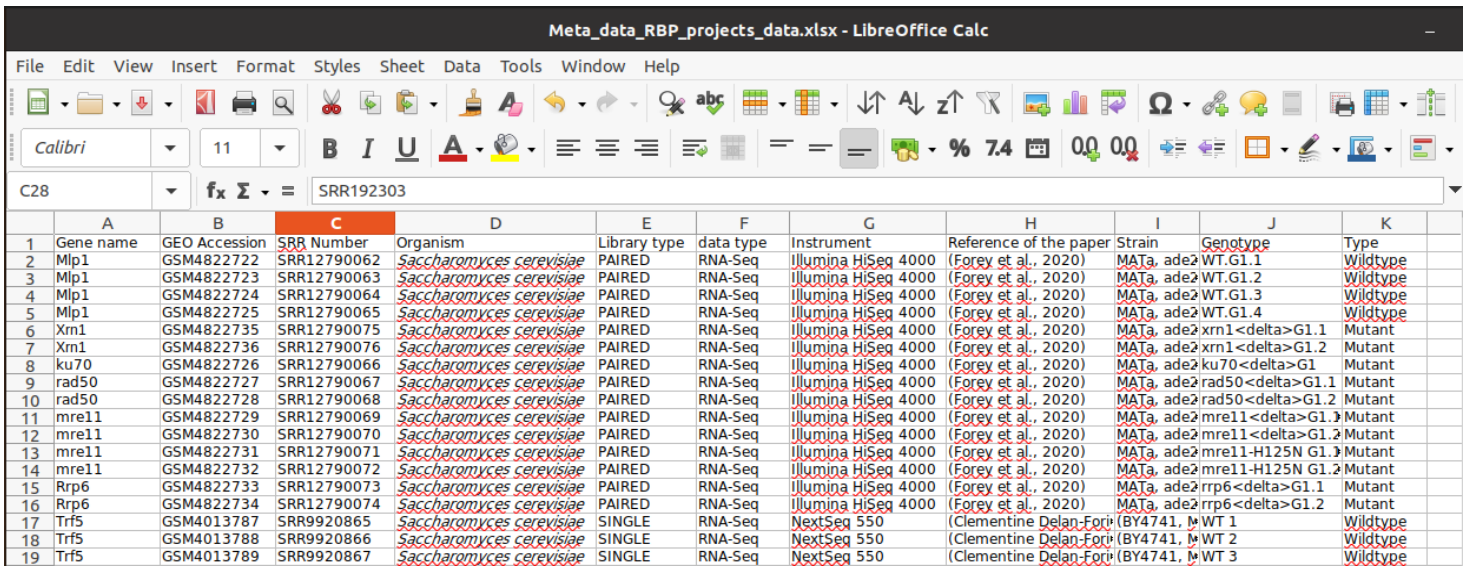
- Platform contains information about the array/sequencer.
- Samples contains the sample, the changes and manipulation it has undergone and the abundance measurement of each element derived from it.
- Series contains a group of related samples with its description, tables, extracted data, summary, conclusion and analysis.
- Cluster heatmaps reveal hierarchical clusters in data matrices.


https://www.ncbi.nlm.nih.gov/sra/SRX9187635[accn]
SRX9187635: GSM4802674: xrn1_B; Saccharomyces cerevisiae; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2000) run: 2.6M spots, 270M bases, 154.6Mb downloads
Submitted by: NCBI (GEO)
Study: Polyadenylation site mapping of xrn1? in Sacchromyces cerevisiae
PRJNA665542 • SRP285324 • All experiments • All runs
show Abstract
Sample: xrn1_B
SAMN16261114 • SRS7423996 • All experiments • All runs
Organism: Saccharomyces cerevisiae
Library:
Instrument: Illumina HiSeq 2000
Experiment
Links:
Runs: 1 run
Run
ID: 11967283
NCBI Resources How To
BioSample
BioSample
Advanced
Full
Send to:
xrn1_B
Identifiers
BioSample: SAMN16261114; SRA: SRS7423996; GEO: GSM4802674
Organism
Saccharomyces cerevisiae (baker's yeast)
cellular organisms; Eukaryota; Opisthokonta; Fungi; Dikarya; Ascomycota; saccharomyceta; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces
Attributes
source name yeast cells
genotype xrn1[delta]::KanMX
parental strain BY4741
Links
GEO Sample GSM4802674
BioProject
PRJNA665542 Polyadenylation site mapping of xrn1Δ in Sacchromyces cerevisiae
Retrieve all samples from this project
Submission
MTC, ScilifeLab - Karolinska Institutet, Vicent Pelechano; 2020-09-24
Accession: SAMN16261114 ID: 16261114
BioProject SRA GEO DataSets

For Further information such as the sample type (wild type or mutant), experiment type and other information check the experiment paper with citation:



The overall data sheet will look like the following figure in the excel file:



	A	B	C	D	E	F	G	H	I	J	K
1	Gene name	GEO Accession	SRR Number	Organism	Library type	data type	Instrument	Reference of the paper	Strain	Genotype	Type
2	Mlp1	GSM4822722	SRR12790062	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	WT.G1.1	Wildtype
3	Mlp1	GSM4822723	SRR12790063	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	WT.G1.2	Wildtype
4	Mlp1	GSM4822724	SRR12790064	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	WT.G1.3	Wildtype
5	Mlp1	GSM4822725	SRR12790065	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	WT.G1.4	Wildtype
6	Xrn1	GSM4822735	SRR12790075	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	xrn1<delta>G1.1	Mutant
7	Xrn1	GSM4822736	SRR12790076	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	xrn1<delta>G1.2	Mutant
8	ku70	GSM4822726	SRR12790066	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	ku70<delta>G1	Mutant
9	rad50	GSM4822727	SRR12790067	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	rad50<delta>G1.1	Mutant
10	rad50	GSM4822728	SRR12790068	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	rad50<delta>G1.2	Mutant
11	mre11	GSM4822729	SRR12790069	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	mre11<delta>G1.1	Mutant
12	mre11	GSM4822730	SRR12790070	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	mre11<delta>G1.2	Mutant
13	mre11	GSM4822731	SRR12790071	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	mre11-H125N G1.1	Mutant
14	mre11	GSM4822732	SRR12790072	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	mre11-H125N G1.2	Mutant
15	Rrp6	GSM4822733	SRR12790073	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	rrp6<delta>G1.1	Mutant
16	Rrp6	GSM4822734	SRR12790074	Saccharomyces cerevisiae	PAIRED	RNA-Seq	Illumina HiSeq 4000	(Forey et al., 2020)	MATa, ade2	rrp6<delta>G1.2	Mutant
17	Trf5	GSM4013787	SRR9920865	Saccharomyces cerevisiae	SINGLE	RNA-Seq	NextSeq 550	(Clementine Delan-Fori	(BY4741, M	WT 1	Wildtype
18	Trf5	GSM4013788	SRR9920866	Saccharomyces cerevisiae	SINGLE	RNA-Seq	NextSeq 550	(Clementine Delan-Fori	(BY4741, M	WT 2	Wildtype
19	Trf5	GSM4013789	SRR9920867	Saccharomyces cerevisiae	SINGLE	RNA-Seq	NextSeq 550	(Clementine Delan-Fori	(BY4741, M	WT 3	Wildtype

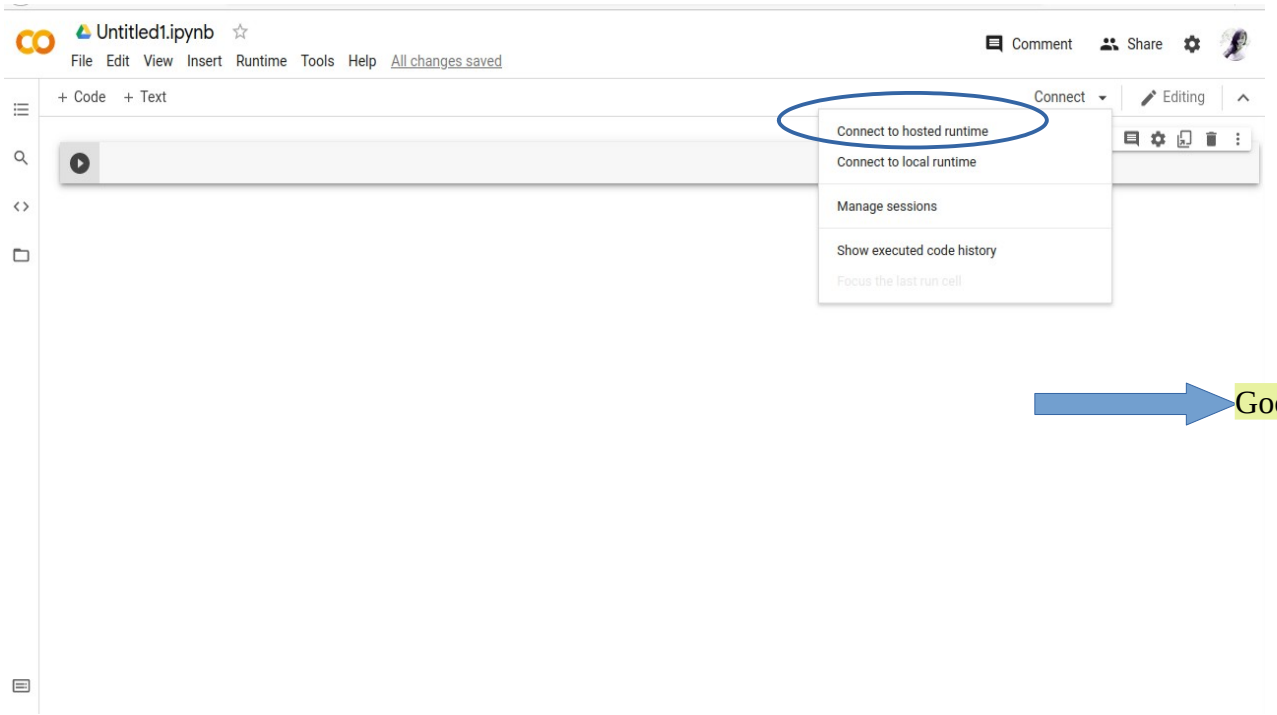
- All the information can also be saved in the text file format.

Script-Two: Command-lines for Downloading SRR Files

SRR files can be downloaded by two systems:

1. Google Colab
2. Linux Operating System

1. Google Colab: The first option is Colaboratory, or “Colab”. It is a free, easy accessible product from Google Research, that allows anyone to write and execute the programming languages. Google Colab can be used if Linux operating system is not available by creating a virtual Linux environment in the Google system to execute the codes for Bioinformatics. It can provide around 100 GB memory at a time for working. The colab notebooks are stored in Google Drive, or it can be loaded from GitHub. For getting access to the Google Colab first search Google Colab in Google search engine. Then after opening it connect the colab to the hosted runtime. Now it is ready to use. The upper code bar is to write the codes and the text bar for writing the texts in the colab system.



Downloading SRR files in Google Colab:

The following command languages and steps are followed for this purpose:

The !-sign is used for running the Linux command-lines in the Colab system.

Step-1: Downloading Miniconda

Miniconda is an installer for Conda, that is free and easy to access. The Conda is an open source package management system and environment management system that can runs on Windows, Linux and macOS. Conda in Implemented here mainly because Conda can installs,runs and updates packages and their dependencies very quickly. Conda also can install and manage the numerous packages that are built,reviewed and maintained by Anaconda® at repo.anaconda.com.

For downloading Miniconda and creating python environment the following command-lines are uses:

```
!wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
! chmod +x Miniconda3-latest-Linux-x86_64.sh
! bash ./Miniconda3-latest-Linux-x86_64.sh -b -f -p /usr/local
import sys
sys.path.append('/usr/local/lib/python3.6.9/site-packages/')
```

Here,

- Wget = The Wget stands for World Wide Web and get. Wget is a Linux command-line utility for downloading files from the web by using HTTPS, HTTP, and FTP protocols.
- repo.anaconda.com/miniconda = To access the Miniconda from Anaconda® Installers and Packages.
- Miniconda3-latest-Linux-x86_64.sh = In this case we downloaded Miniconda3 version for Linux in 64-bit processor. x86_64 means only the 64-bit computers alone can run it.
- chmod + x = The chmod stands for Change the Mode. chmod modifies the file permissions. For each set of permissions there are three characters. They are either a dash (-) or a letter. If the character is a dash (-) that permission is not granted or if r ,w, or x, character are used, the permission has to be granted. Each letters represent the following meaning:

  - r : Read permissions. The file can be opened, and its content can be viewed.
  - w: Write permissions. The file can be edited,modified,and deleted.
  - x: Execute permissions. If the file is a script or a program,it can be run (executed).

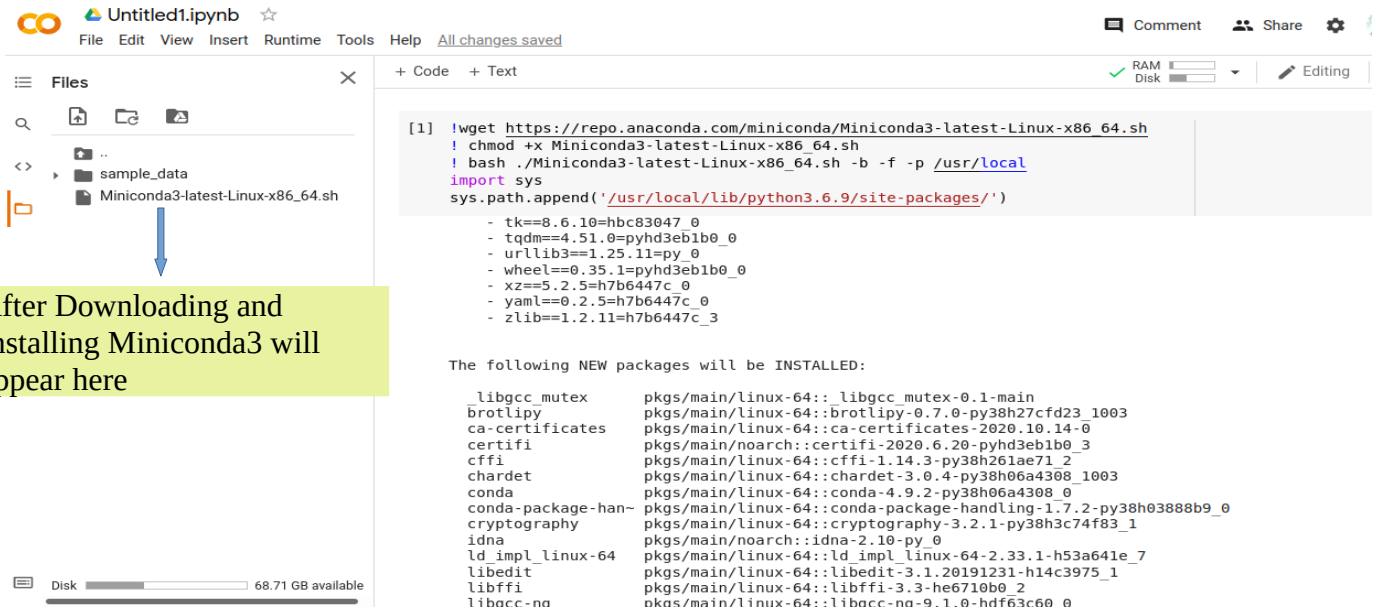
So, the chmod + x command-line is to execute permissions to run the Miniconda3.

- Bash = The Linux command-line is used execute (to run) the Miniconda3 in the system. The bash is an sh-compatible command language interpreter.
- import =import statement is to access the python modules.
- sys=The sys module is a set of functions which provides crucial information about how the python codes are interacting with the host systems in which its running.
- sys.path.append= Command-line to import the python file to the system path. Here, python 3.6.9 version was appended to the system path.

Step-2: Installation of Miniconda3

After downloading the following command-line is used to install the Miniconda3:

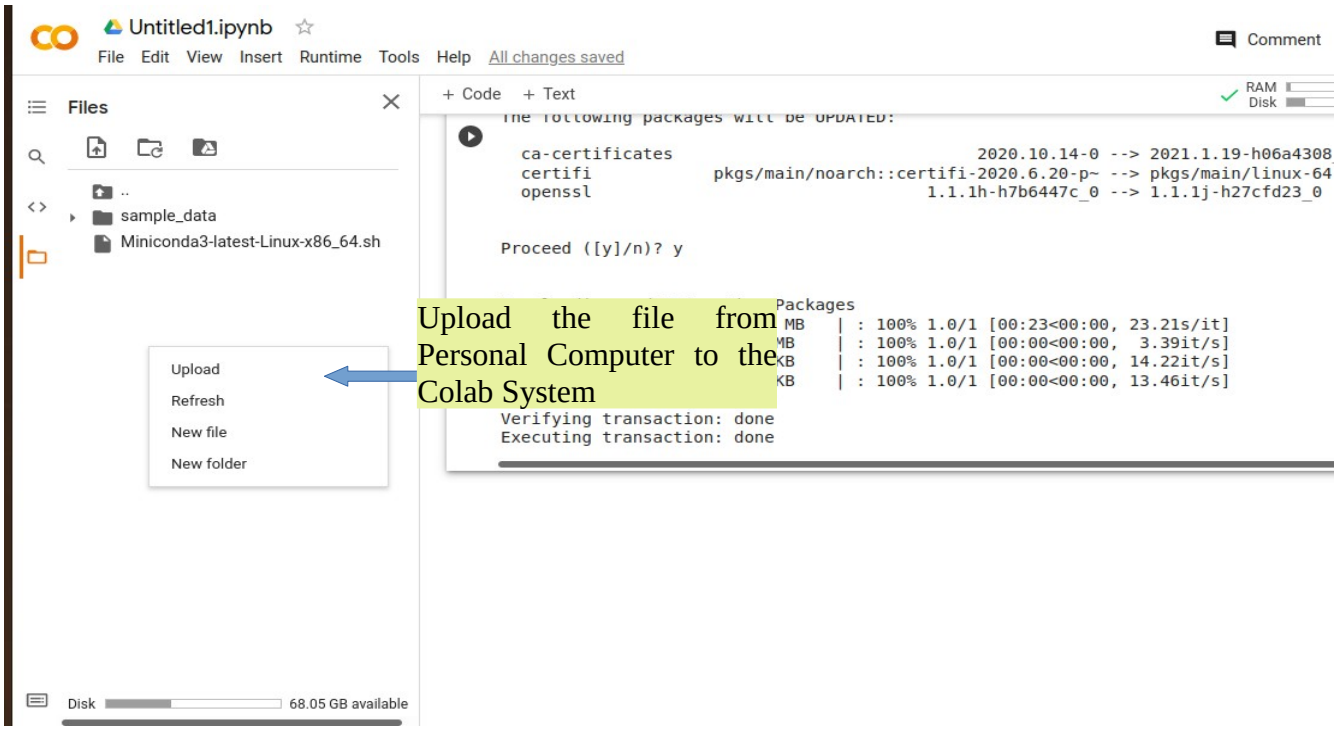
```
!sh ./Miniconda3-latest-Linux-x86_64.sh
```



Step-3: Installing SRA-Tools

The conda use -c flag for specifying the appropriate channel such as bioconda. Bioconda is a channel for conda package specializing in management of Bioinformatics software (in this case SRA-Tools). Here, SRA-Tools is used to restore the original data from NCBI for example as fastq,by fast access to the reference sequences that the original data was aligned to. The SRA-Tools packages can be installed by running the following code:

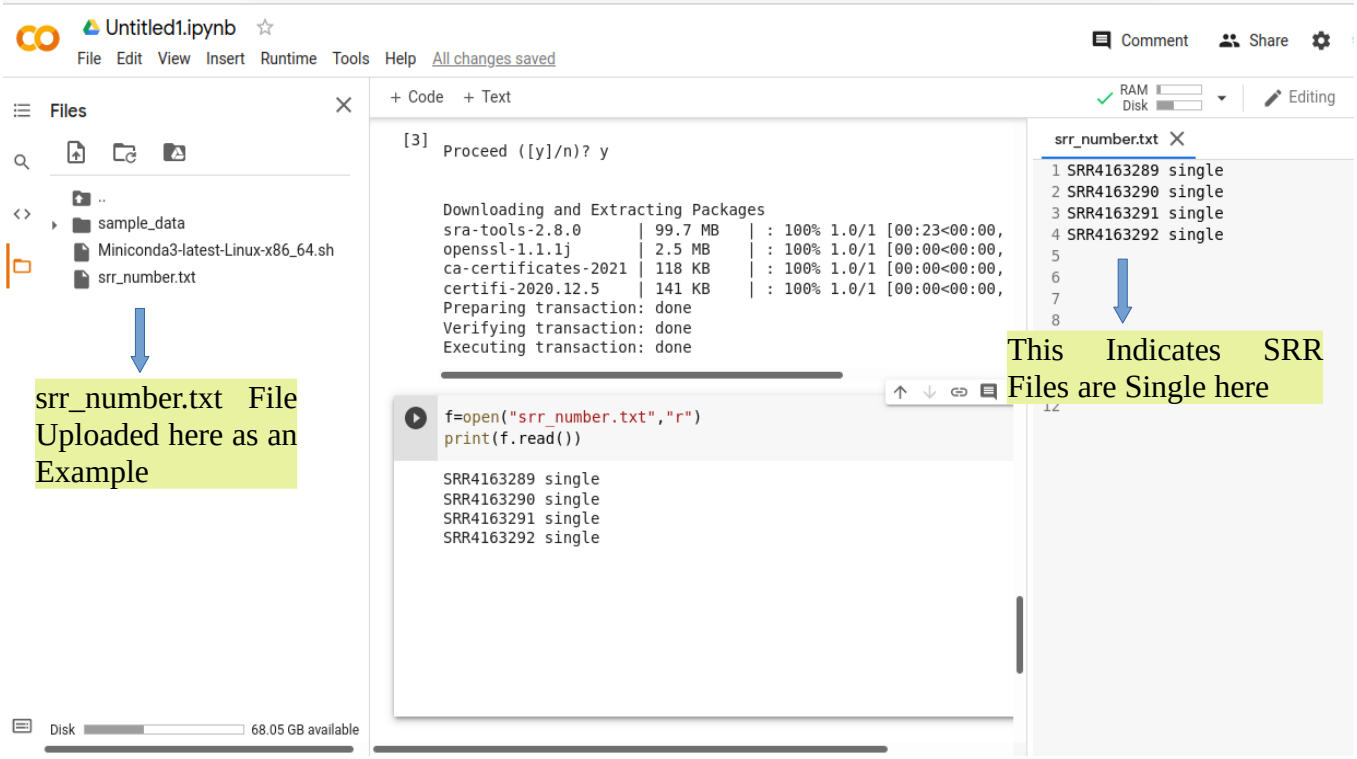
!conda install -c bioconda sra- tools



Step-4: File Read
For file read in Google Colab the following command-lines are used:

```
f = open("file_ name.txt", "r")
print(f.read())
```

Here,
f = File pointer
r = “r” is a mode that opens an existing text file only for reading purpose.
read()= Python file method that reads at most size bytes from the file.



Step-5.1: Downloading SRR Files in Single Number
For Paired End the following command line is used:

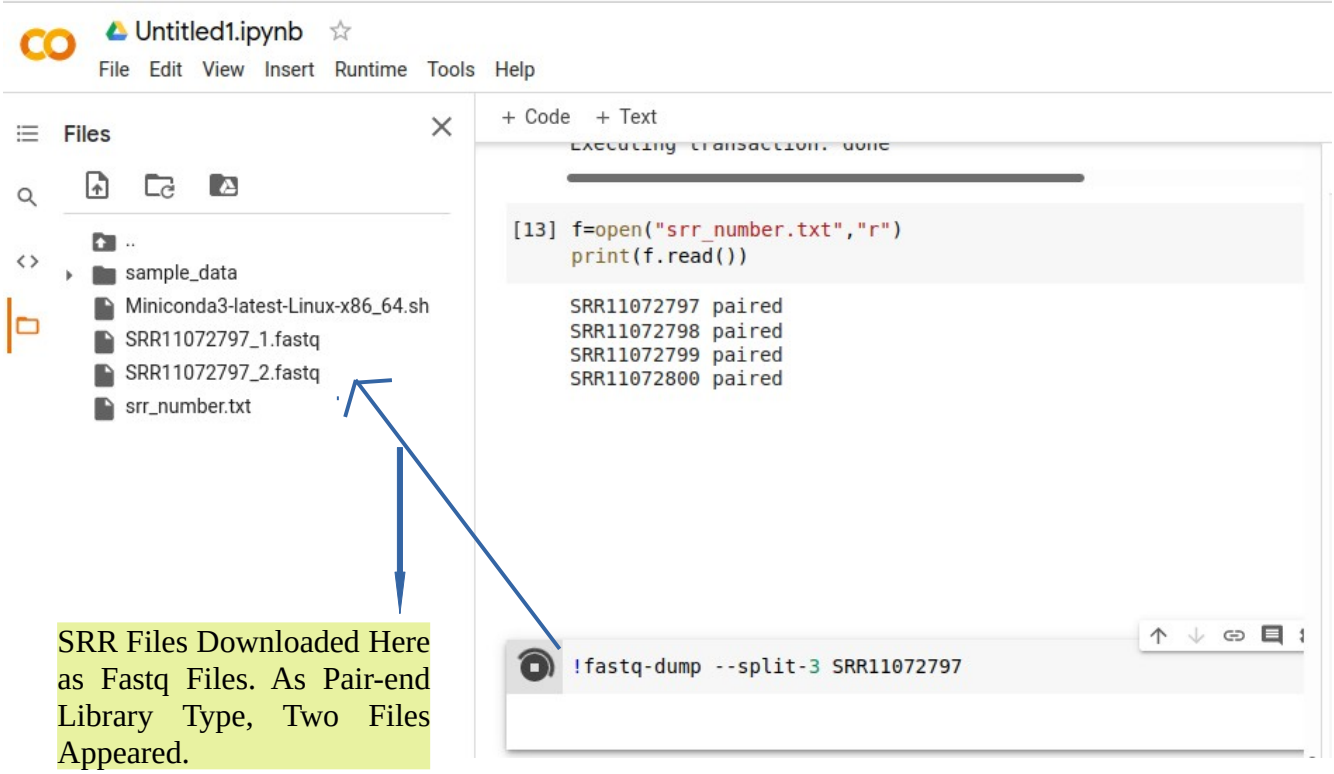
```
!fastq-dump --split-3 SRR11072797 (Sample SRR Number)
```

Here, --split-3 separates the read into left and right ends.

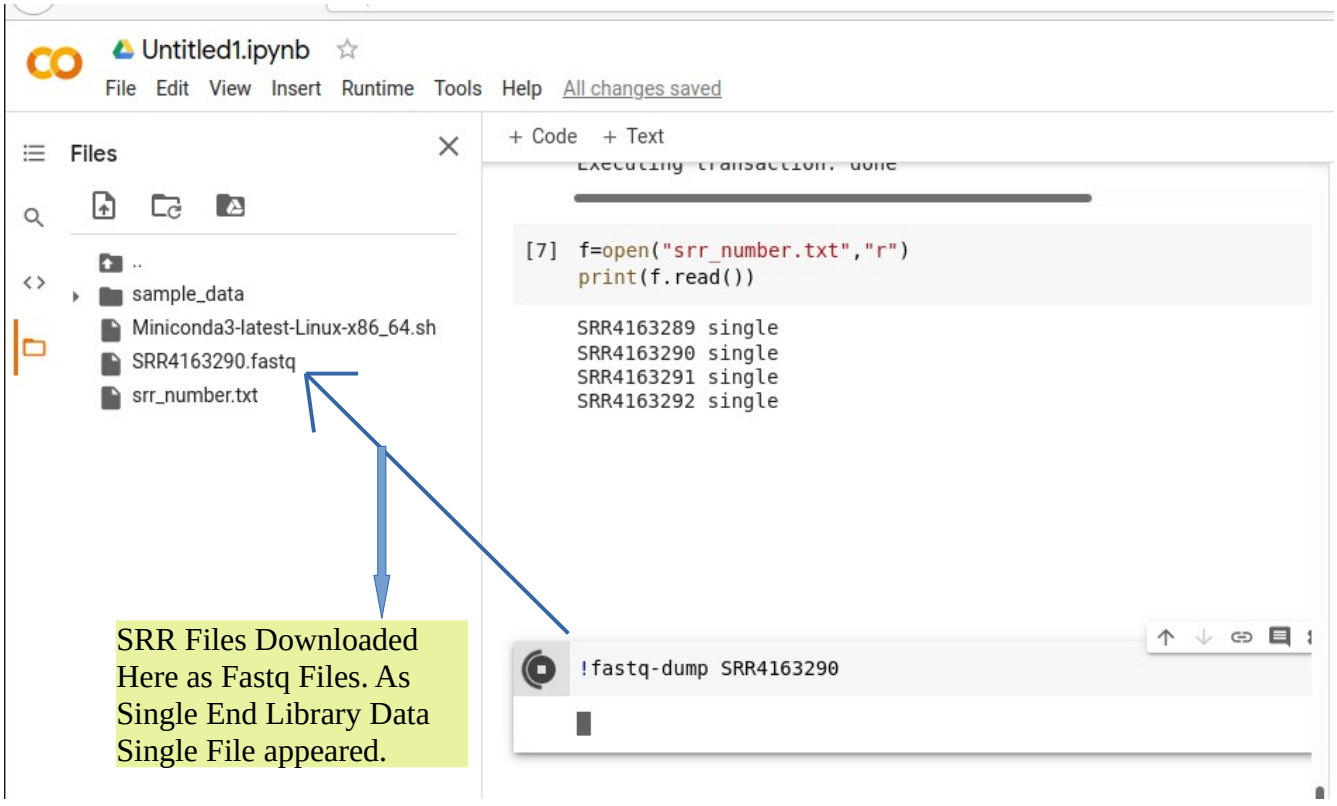
For Single End the following command line is used:

```
!fastq-dump SRR4163290 (Sample SRR Number)
```

- Fastq-Dump is SRA Tool-kit that converts data to fastq and Fasta format.



SRR Files Downloaded Here as Fastq Files. As Pair-end Library Type, Two Files Appeared.

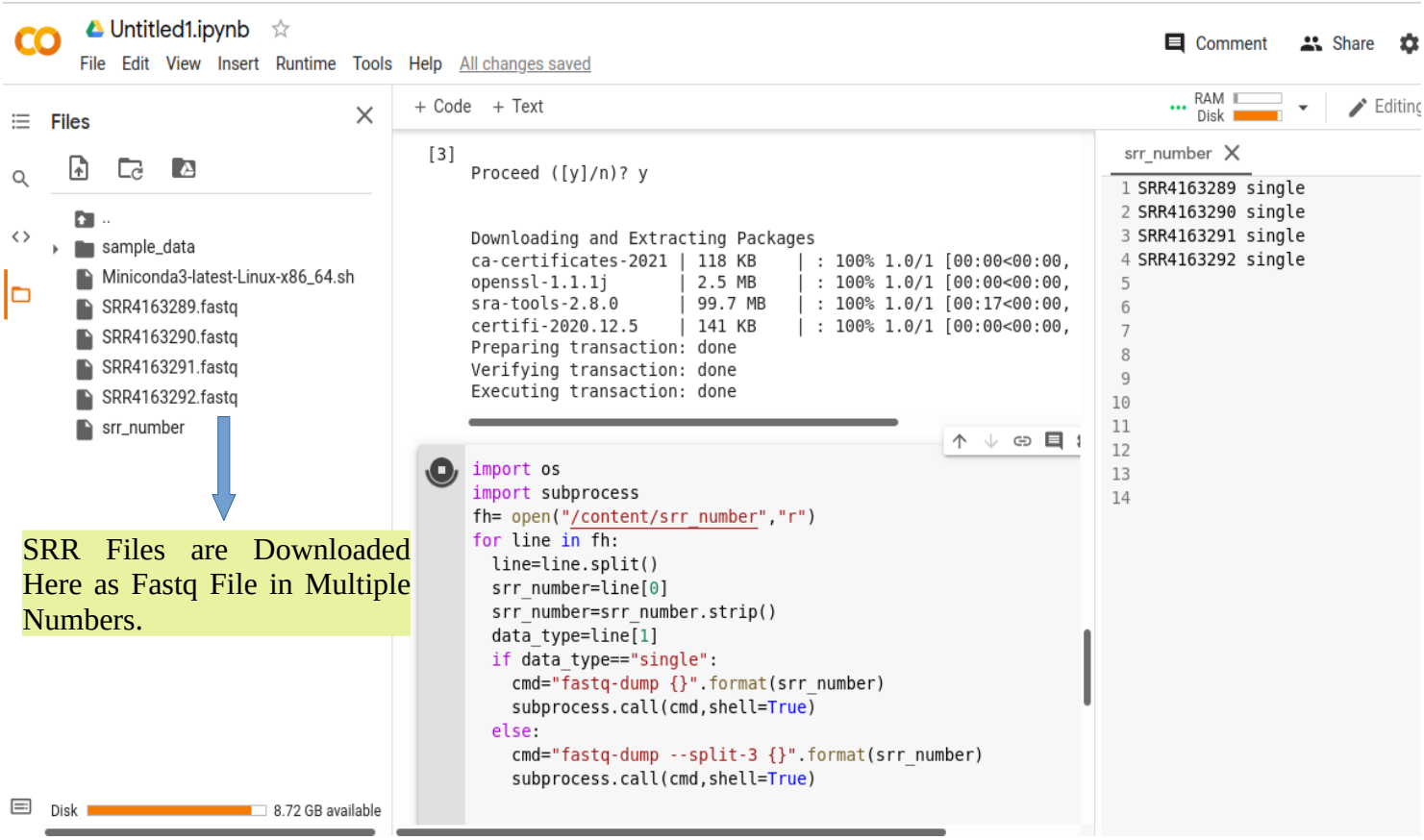


SRR Files Downloaded Here as Fastq Files. As Single End Library Data Single File appeared.

Step-5.2: Downloading SRR Files in Multiple Numbers
The following command-lines are effective only when the multiple SRR files are uniformally Single or Paired.

```
import os
import subprocess
fh = open("/content/sample_file.txt","r")
for line in fh:
  line = line.split()
  sample_file = line[0]
  sample_file = sample_file.strip
  data_type = line[1]
  if data_type == "pair":
  cmd="fastq-dump --split-3 {}".format(sample_file)
  subprocess.call(cmd, shell=True)
  else:
  cmd="fastq-dump {}".format(sample_file)
  subprocess.call(cmd, shell=True)
```

Here,
os = A Python module that provides functions for interacting with the operating system.
subprocess = The module is for connecting input/output/error pipes and to obtain the return codes.
fh,data_type and cmd = File pointers.
for loop = Implied to execute the statement defined by the file pointer.
split() = To break the text file/string into list.
strip = This method returns a copy of the string by removing both the leading (spaces at the beginning) and the trailing (spaces at the end) characters.
if/else=Python conditional statement.
format() = This method has been introduced for handling complex string to give great flexibility over reading, writing and maintaining the output of the string.
subprocess.cal= subprocess has a method called call that is used to start a program by executing the code through the python shell (if shell=True).



SRR Files are Downloaded Here as Fastq File in Multiple Numbers.

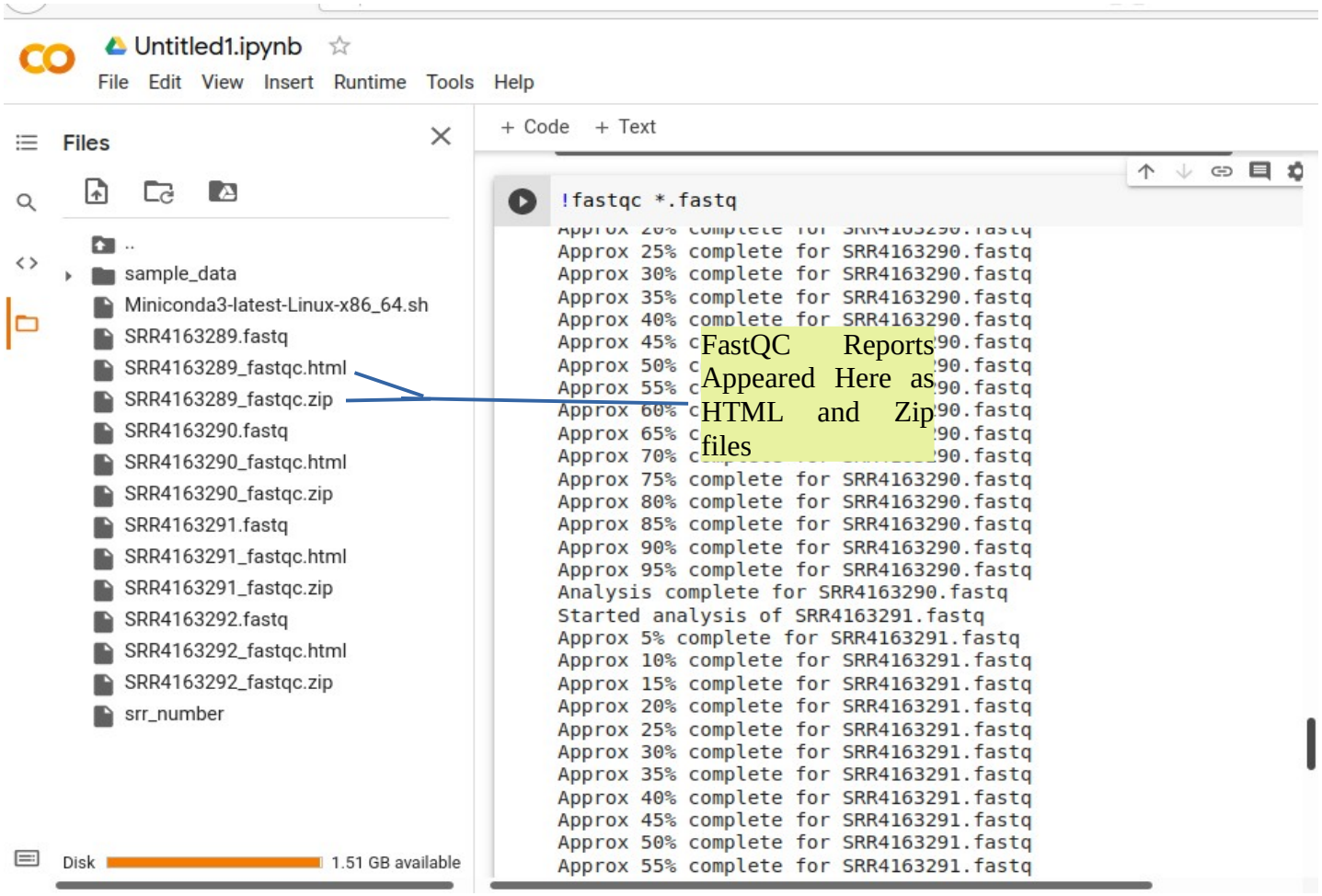
Step-6: Installing fastQC:
fastQC is an quality control tool for high throughput sequence data.fastQC can be installed by running the following code through accessing the bioconda channel:

```
!conda install -c bioconda fastqc
```

Step-7: Accessing Quality with fastQC:

```
!fastqc *.fastq
```

The * character is a special type of character called a wildcard, which can be used to represent any number of any type of character. Thus, *.fastq matches every file that ends with .fastq.


Untitled1.ipynb
File Edit View Insert Runtime Tools Help
Files
+ Code + Text
!fastqc *.fastq
sample_data
Miniconda3-latest-Linux-x86_64.sh
SRR4163289.fastq
SRR4163289_fastqc.html
SRR4163289_fastqc.zip
SRR4163290.fastq
SRR4163290_fastqc.html
SRR4163290_fastqc.zip
SRR4163291.fastq
SRR4163291_fastqc.html
SRR4163291_fastqc.zip
SRR4163292.fastq
SRR4163292_fastqc.html
SRR4163292_fastqc.zip
srr_number
Disk 1.51 GB available
Approx 25% complete for SRR4163290.fastq
Approx 30% complete for SRR4163290.fastq
Approx 35% complete for SRR4163290.fastq
Approx 40% complete for SRR4163290.fastq
Approx 75% complete for SRR4163290.fastq
Approx 80% complete for SRR4163290.fastq
Approx 85% complete for SRR4163290.fastq
Approx 90% complete for SRR4163290.fastq
Approx 95% complete for SRR4163290.fastq
Analysis complete for SRR4163290.fastq
Started analysis of SRR4163291.fastq
Approx 5% complete for SRR4163291.fastq
Approx 10% complete for SRR4163291.fastq
Approx 15% complete for SRR4163291.fastq
Approx 20% complete for SRR4163291.fastq
Approx 25% complete for SRR4163291.fastq
Approx 30% complete for SRR4163291.fastq
Approx 35% complete for SRR4163291.fastq
Approx 40% complete for SRR4163291.fastq
Approx 45% complete for SRR4163291.fastq
Approx 50% complete for SRR4163291.fastq
Approx 55% complete for SRR4163291.fastq
FastQC Reports Appeared Here as HTML and Zip files

2.Linux Operating System:
The second option is to use the Linux. An operating system just like Windows, MacOs, and ios, which is highly flexible with Bioinformatics software tools. By using Linux there is no necessity of using any virtual environment and files are directly stored in the users system.

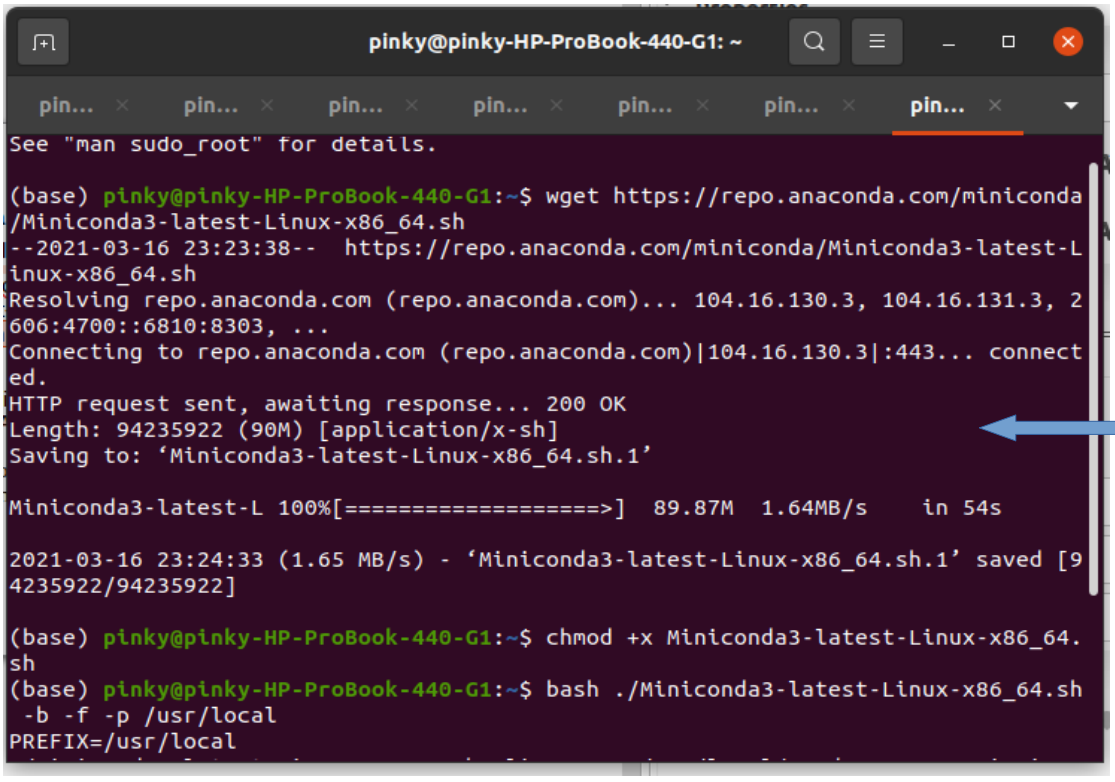
Downloading SRR Files in Linux Operating System:
For downloading SRR Files only few command languages are used & the command-lines are mostly similar to that of the Google Colab.

Step-1: Downloading and Installing Miniconda
For downloading and installing Miniconda the following command lines are used:

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
chmod +x Miniconda3-latest-Linux-x86_64.sh
bash ./Miniconda3-latest-Linux-x86_64.sh -b -f -p /usr/local
```

Here,

- Wget = The Wget stands for World Wide Web and get. Wget is a Linux command-line utility for downloading files from the web by using HTTPS, HTTP, and FTP protocols.
- repo.anaconda.com/miniconda = To access the Miniconda from Anaconda® Installers and Packages.
- Miniconda3-latest-Linux-x86_64.sh = In this case we downloaded Miniconda3 version for Linux in 64-bit processor. x86_64 means only the 64-bit computers alone can run it.
- chmod + x = The chmod stands for Change the Mode. chmod modifies the file permissions. The chmod + x command-line is to execute permissions to run the Miniconda3.
- Bash = The Linux command-line is used execute (to run) the Miniconda3 in the system. The bash is an sh-compatible command language interpreter.



Downloading and Installing Miniconda3 in Linux Operating System.

As an optional step Miniconda3 installation can be checked by the following command-line:

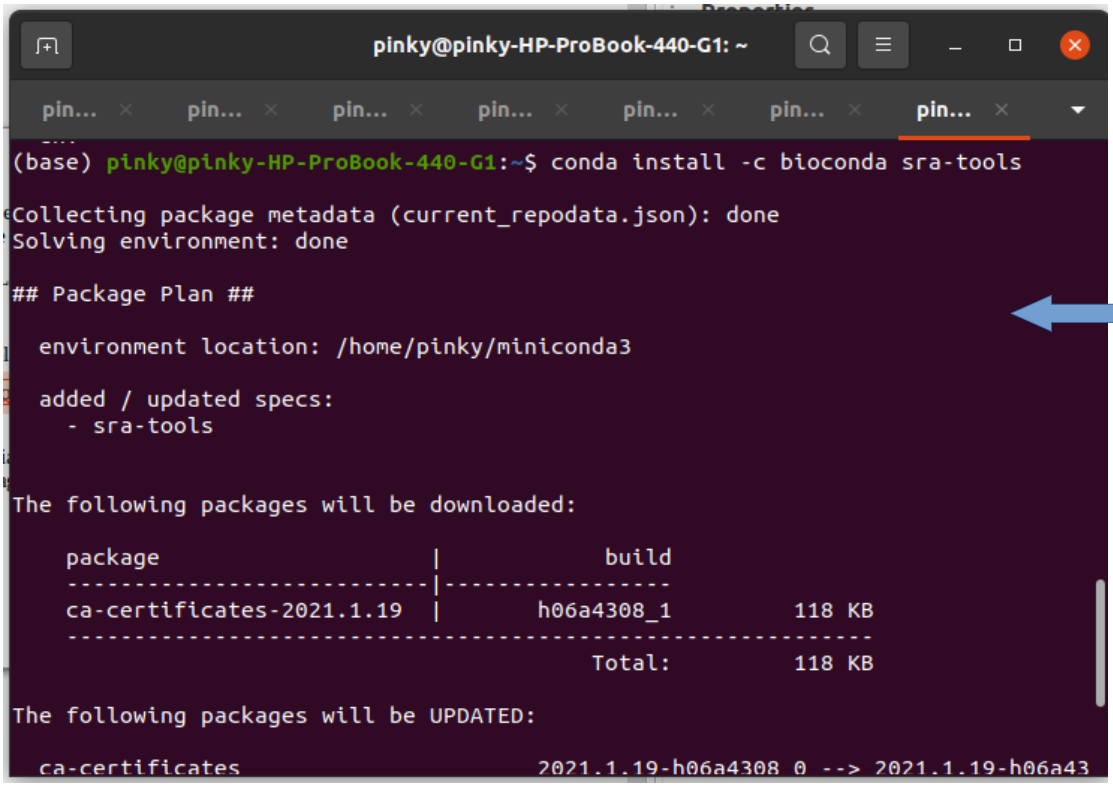
sh./Miniconda3-latest-Linux-x86_64.sh

Step-2: Installing SRA-Tools
SRA-Tools are installed by running the following code:

conda install -c bioconda sra-tools

Here,
-c = -c flag for specifying the appropriate channel such as bioconda.
bioconda = A channel for conda package specializing in management of Bioinformatics software.
sra-tools = Restore the original data from NCBI for example as fastq.

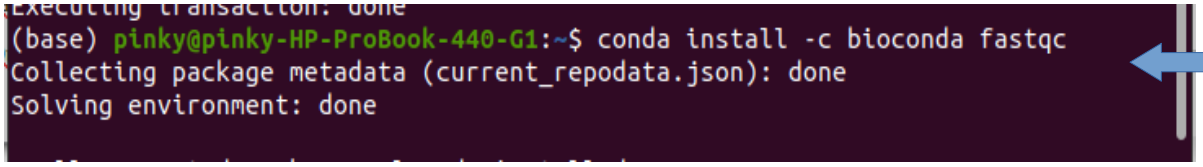
In this case the conda is channeling the bioconda by -c flag to get the access to the bioinformatics tools (sra- tools, fastqc), which are carried by bioconda.



Installing sra-tools in Linux by Accessing Bioconda though Conda System.

Step-3: Installing fastQC
fastQC can be installed by running the following code through the Bioconda Channel:

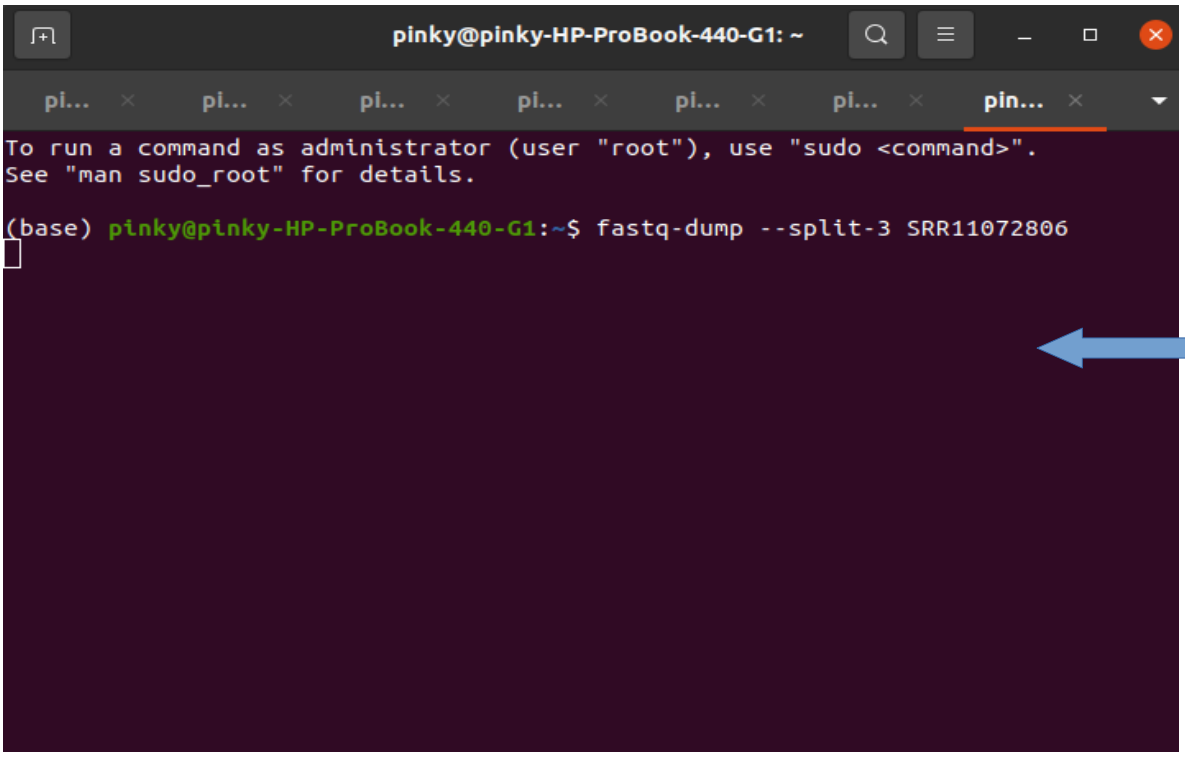
conda install -c bioconda fastqc



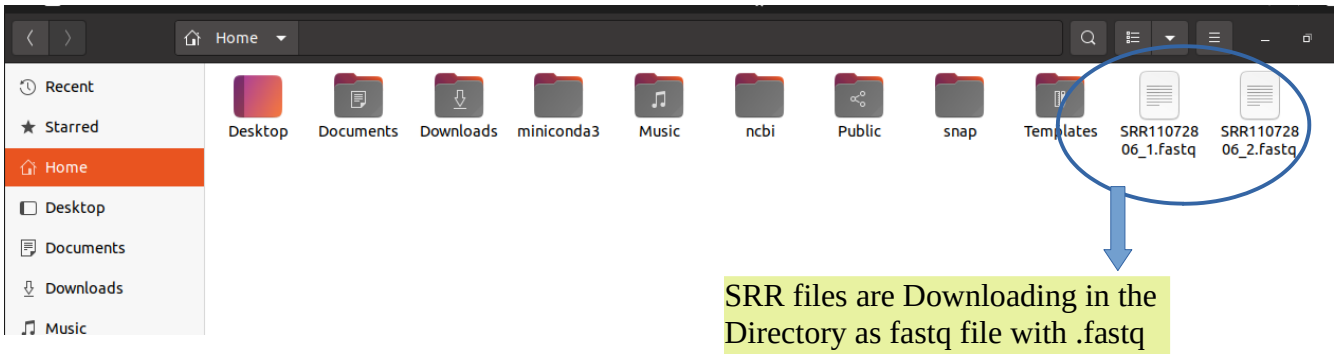
Installing Bioinformatics Tools Such as FastQC in Linux by Accessing Bioconda through Conda System.

Step-4.1: Downloading SRR Files as decompressed files
For Paired End the following command line is used:

fastq-dump --split-3 SRR11072806 (Sample SRR Number)



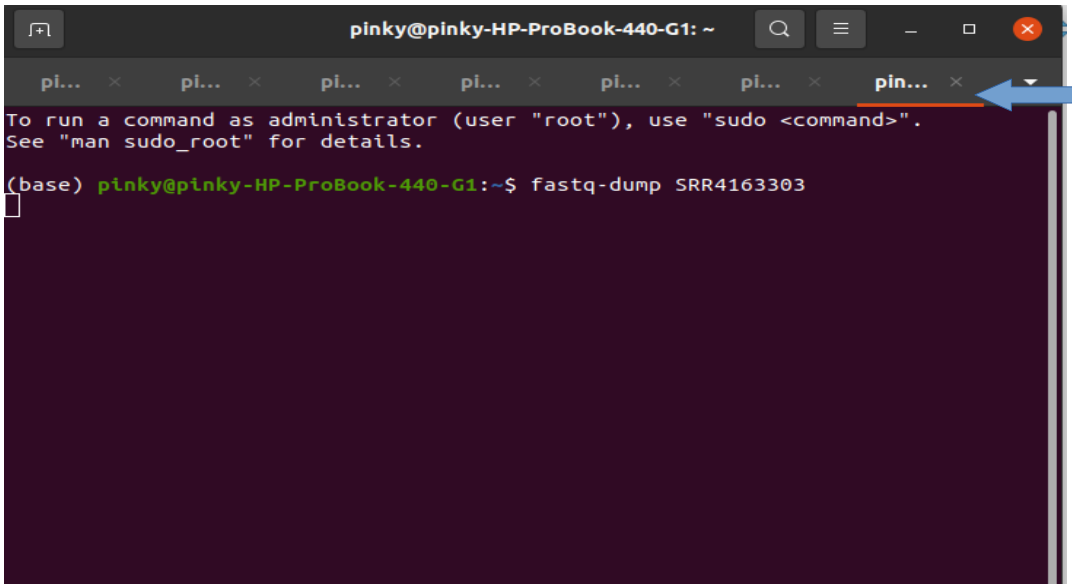
Downloading SRR file (paired data type) as decompressed file in the system. The file is downloaded here as fastq file in the text file format.--split-3 separates the read into left and right ends.



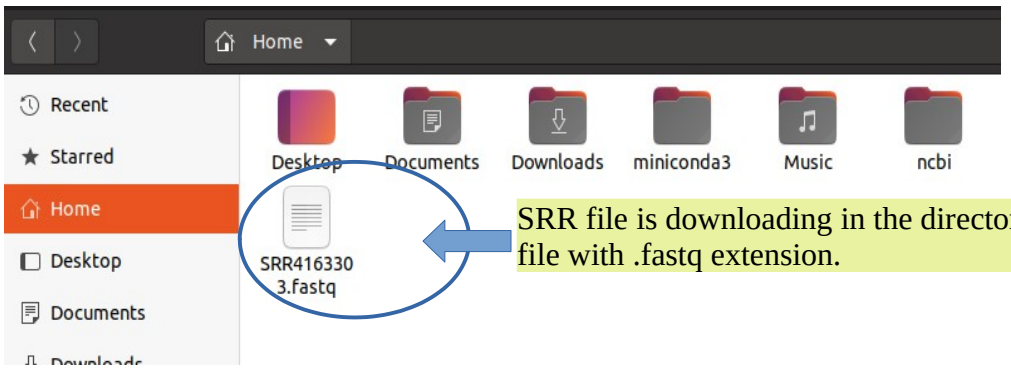
SRR files are Downloading in the Directory as fastq file with .fastq extension.

For single end the following command-line is used:

fastq-dump SRR4163303 (Sample SRR Number)



Downloading SRR file (Single Library type) as Decompressed file in the system. The file is downloaded here as fastq file in the text file format.



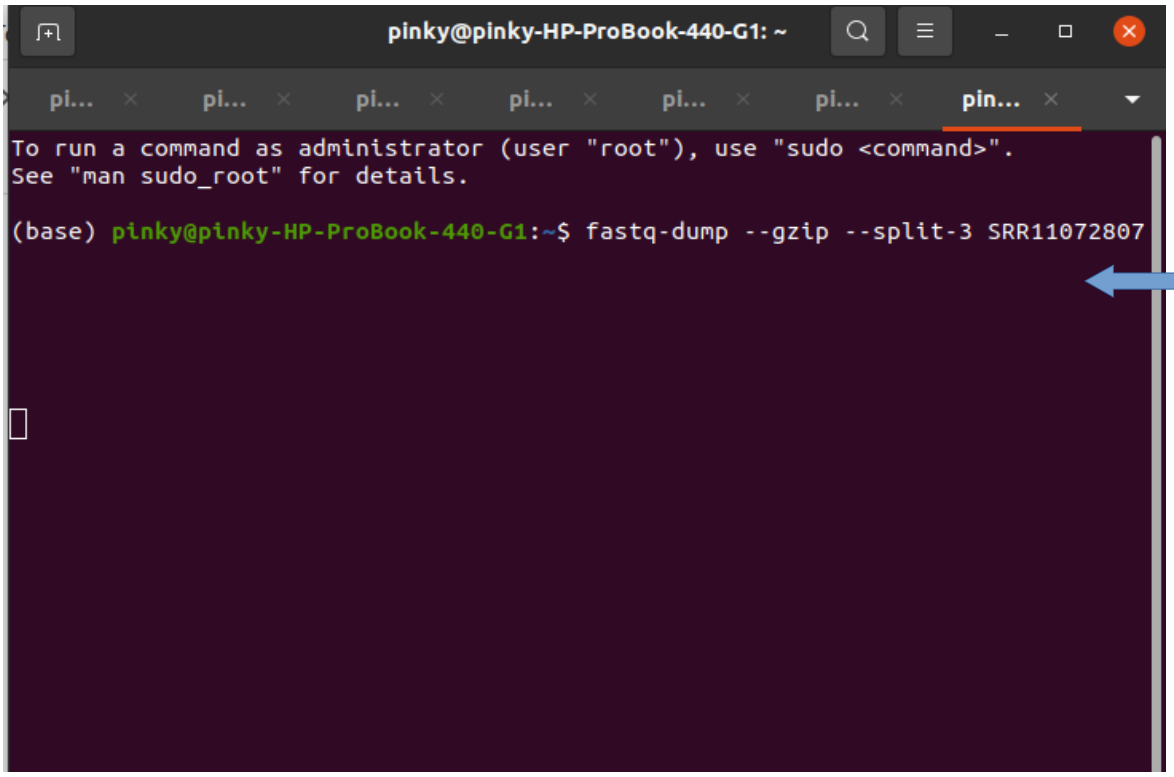
SRR file is downloading in the directory as fastq file with .fastq extension.

Step-4.2: Downloading SRR Files as Zip File

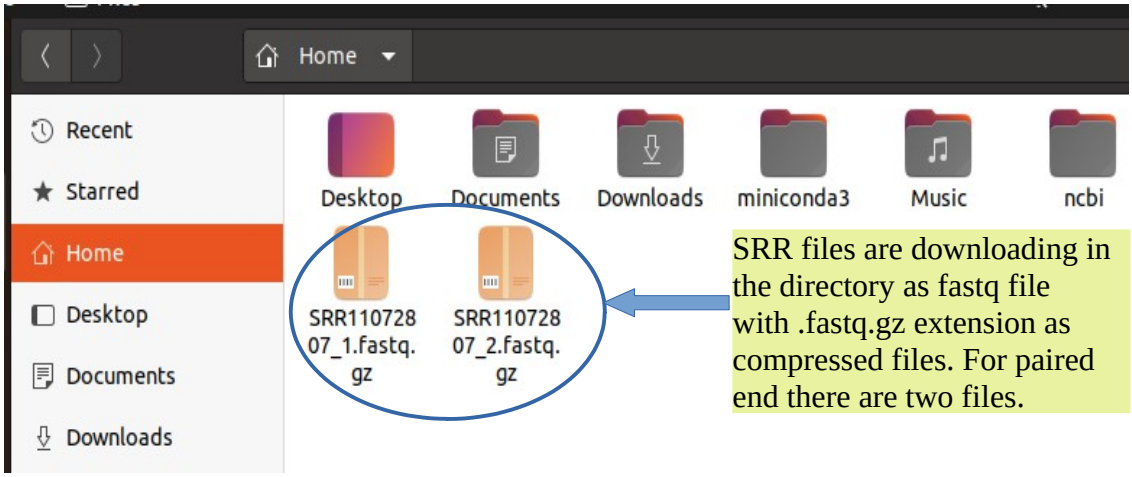
For downloading SRR files as Zip file the following command-lines are used:

For Paired-End:

fastq-dump --gzip --split-3 SRR11072807 (Sample SRR Number)



Downloading SRR file (Pair Library Type) as a zip file to Reduce the size of the original file. --gzip compressed the file as zip.--split-3 separates the read into left and right ends.

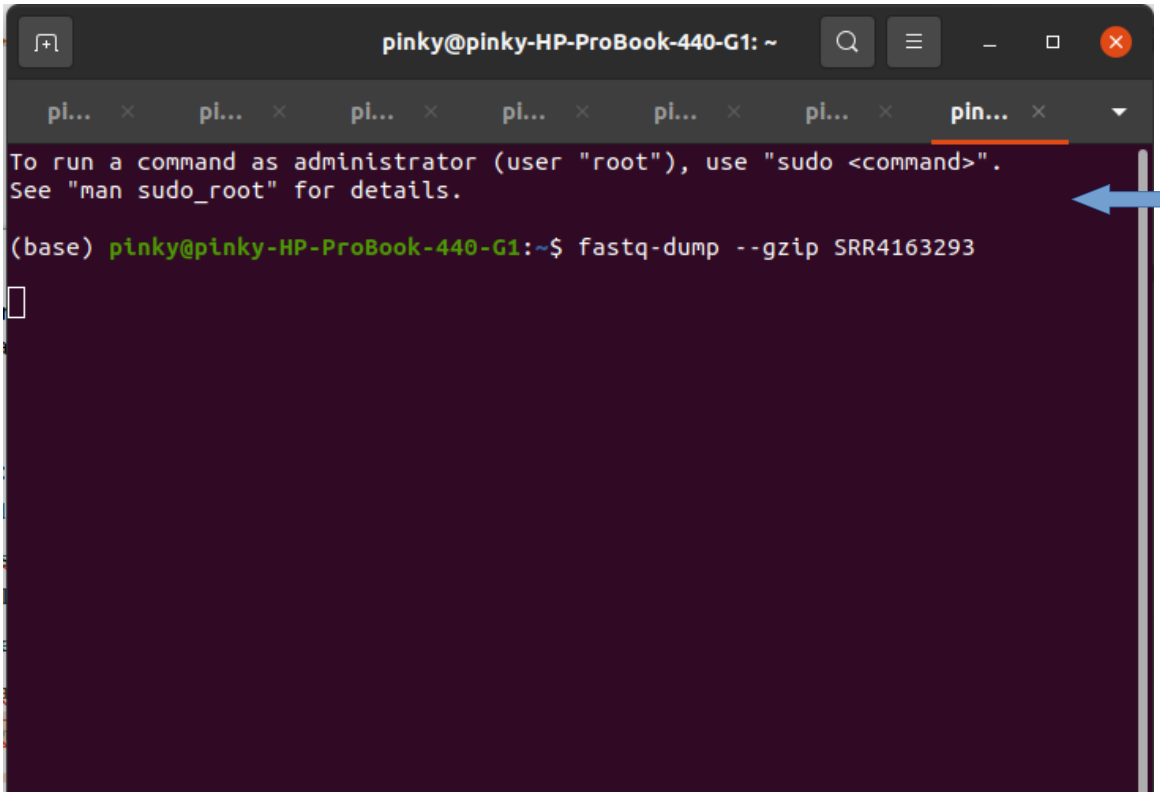


SRR files are downloading in the directory as fastq file with .fastq.gz extension as compressed files. For paired end there are two files.

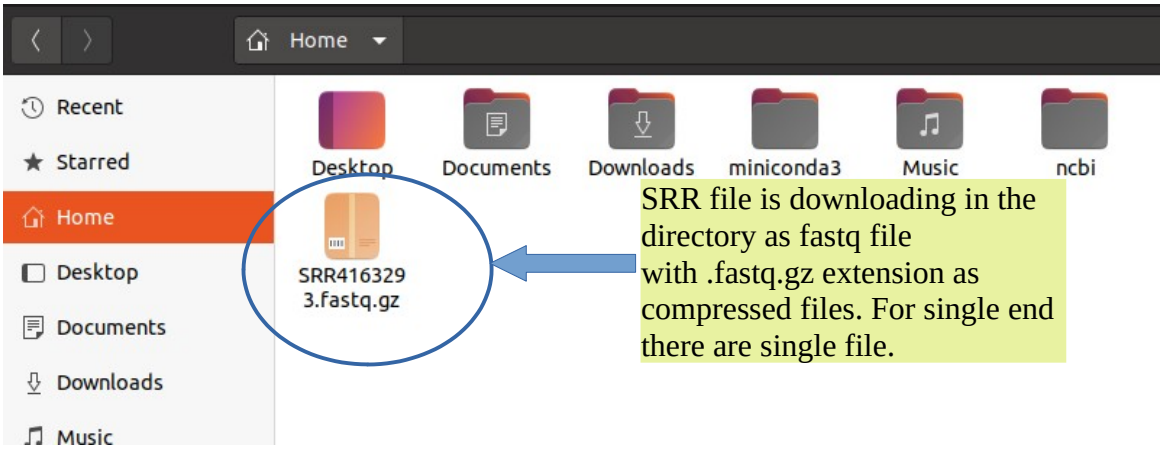
For Single End:

fastq-dump –gzip SRR4163293 (Sample SRR Number)

- Fastq-Dump is SRA Tool-kit that converts data to Fastq and Fasta format.



Downloading SRR file (Single Library Type) as Compressed file in the system. The file is downloaded here as fastq file in the text file format. --gzip is converting the large file into zip file.



SRR file is downloading in the directory as fastq file with .fastq.gz extension as compressed files. For single end there are single file.

Step-5 : Accessing the Quality of fastq Files

The FastQC is a bioinformatics software is used to analyse the quality of fastq files. The quality is expressed as HTML file report, which has to be downloaded. The following command line is used for analyzing the quality of decompressed fastq files:

fastqc*.fastq

```
(base) pinky@pinky-HP-ProBook-440-G1:~$ fastqc *.fastq
Started analysis of SRR7587575.fastq
Approx 5% complete for SRR7587575.fastq
Approx 10% complete for SRR7587575.fastq
Approx 15% complete for SRR7587575.fastq
Approx 20% complete for SRR7587575.fastq
Approx 25% complete for SRR7587575.fastq
Approx 30% complete for SRR7587575.fastq
Approx 35% complete for SRR7587575.fastq
Approx 40% complete for SRR7587575.fastq
Approx 45% complete for SRR7587575.fastq
Approx 50% complete for SRR7587575.fastq
Approx 55% complete for SRR7587575.fastq
Approx 60% complete for SRR7587575.fastq
Approx 65% complete for SRR7587575.fastq
Approx 70% complete for SRR7587575.fastq
Approx 75% complete for SRR7587575.fastq
Approx 80% complete for SRR7587575.fastq
Approx 85% complete for SRR7587575.fastq
Approx 90% complete for SRR7587575.fastq
Approx 95% complete for SRR7587575.fastq
Analysis complete for SRR7587575.fastq
(base) pinky@pinky-HP-ProBook-440-G1:~$
```

Downloading FastQC HTML report of decompressed fastq file. Here, SRR7587575 is a sample fastq file.

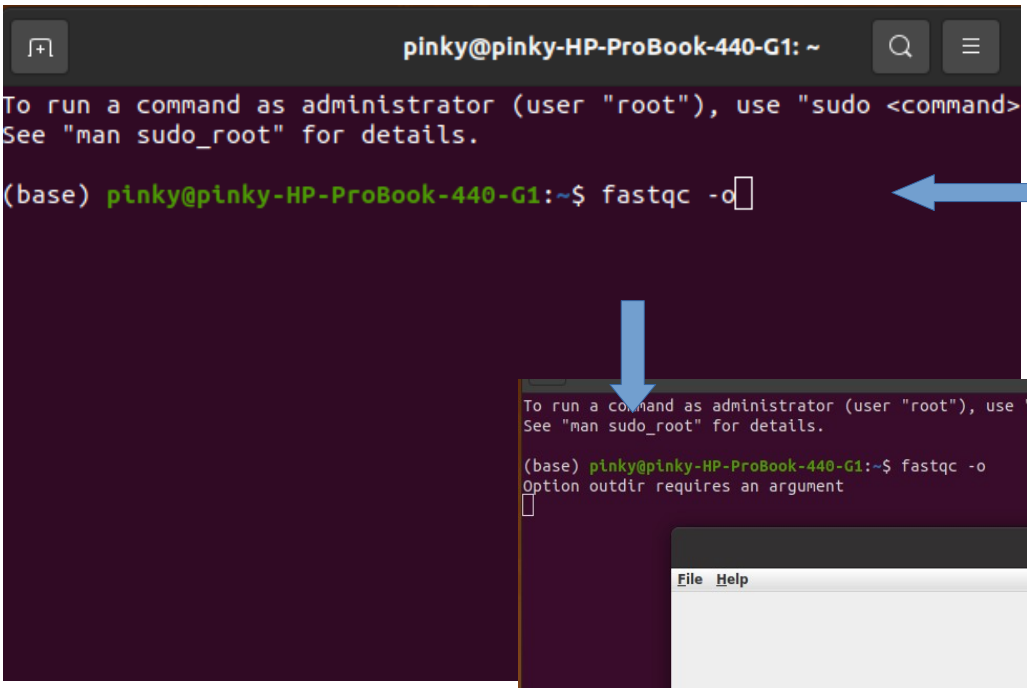
The following command line is used for analyzing the quality of zip fastq files:

fastqc*.fastq,gz

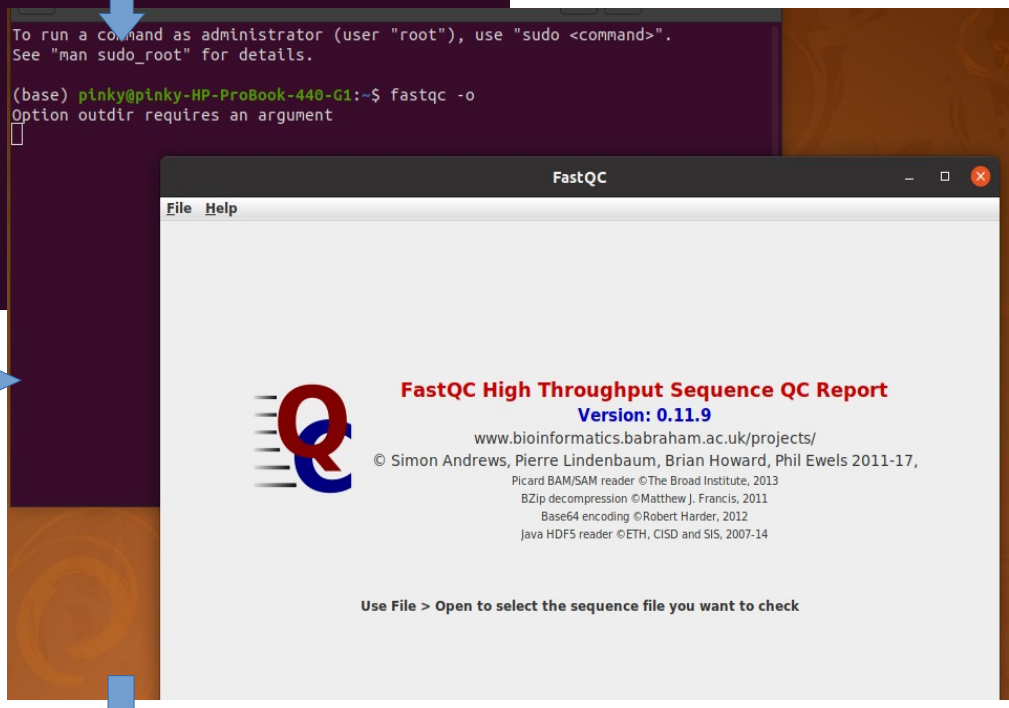
```
fastqc *.fastq.gzRead 10056607 spots for SRR9033892
Written 10056607 spots for SRR9033892
(base) pinky@pinky-HP-ProBook-440-G1:~$ fastqc *.fastq.gz
Started analysis of SRR9033890.fastq.gz
Approx 5% complete for SRR9033890.fastq.gz
Approx 10% complete for SRR9033890.fastq.gz
Approx 15% complete for SRR9033890.fastq.gz
Approx 20% complete for SRR9033890.fastq.gz
Approx 25% complete for SRR9033890.fastq.gz
Approx 30% complete for SRR9033890.fastq.gz
Approx 35% complete for SRR9033890.fastq.gz
Approx 40% complete for SRR9033890.fastq.gz
Approx 45% complete for SRR9033890.fastq.gz
Approx 50% complete for SRR9033890.fastq.gz
Approx 55% complete for SRR9033890.fastq.gz
Approx 60% complete for SRR9033890.fastq.gz
Approx 65% complete for SRR9033890.fastq.gz
Approx 70% complete for SRR9033890.fastq.gz
Approx 75% complete for SRR9033890.fastq.gz
Approx 80% complete for SRR9033890.fastq.gz
Approx 85% complete for SRR9033890.fastq.gz
Approx 90% complete for SRR9033890.fastq.gz
Approx 95% complete for SRR9033890.fastq.gz
Analysis complete for SRR9033890.fastq.gz
```

Downloading fastQC HTML report of compressed/zip fastq file. Here, SRR9033890 is a sample fastq zip file.

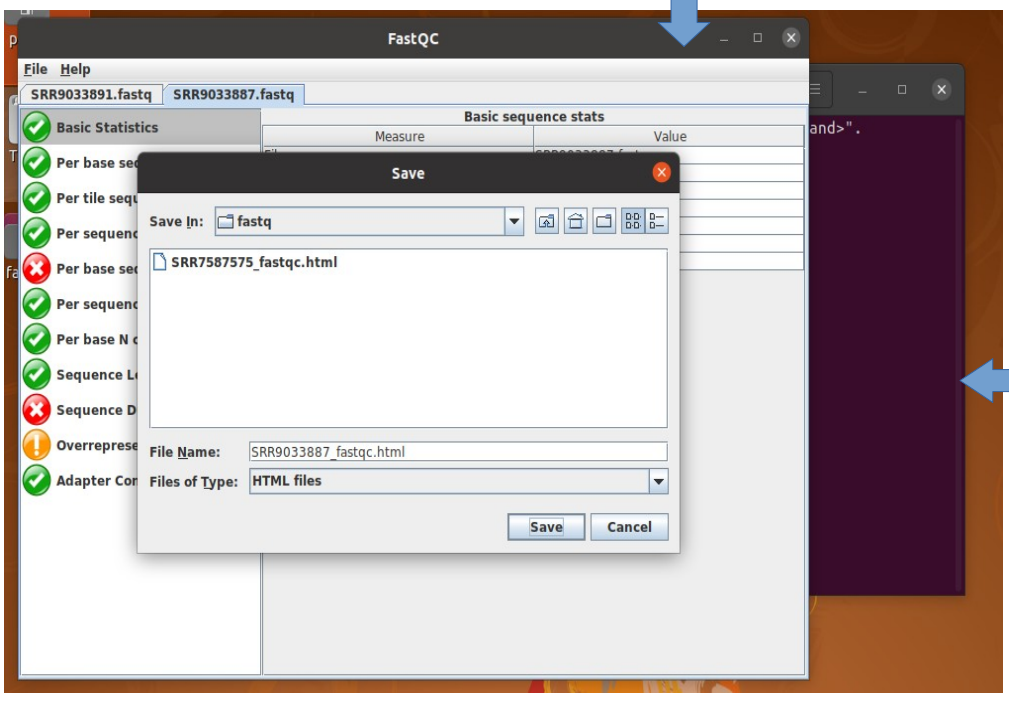
Storing the HTML fastQC report in the particular desired folder:



Write the command-line fastqc -o in the terminal to get the fastqc files, i.e., zip and HTML file.



This FastQC software then appeared. Selected the file, choose and open the fastq files.



After analysing the HTML files, save the only HTML reports in the desired folder/directory all together specifically.

Script-Three: Analysing FastQC Reports

Quality control of fastq files:
The SRR files are downloaded as fastq file. The fastq is a text based format for storing both a nucleiotide sequence and its corresponding quality scores. The sequence letter and quality scores are encoded with ASCII (American Standard Code for Information Interchange) character. ASII is a character encoding standard for electronic communication. In computer, character is a unit of information or symbol such as alphabet or syllabary in the written form of a natural language.

Fastq format of NCBI sequence read archive:
A Fastq file normally uses fours lines per sequence.
Line-1: Begins with a '@' character and is followed by an optional description and sequence identifier.
Line-2: Consists of raw sequence letters.
Line-3: Begins with a '+' character and is followed by an optional description and same sequence identifier again.
Line-4: Encodes the quality values for the sequence in Line-2, and must contain the same number of symbols as the raw sequence letters in Line-2.

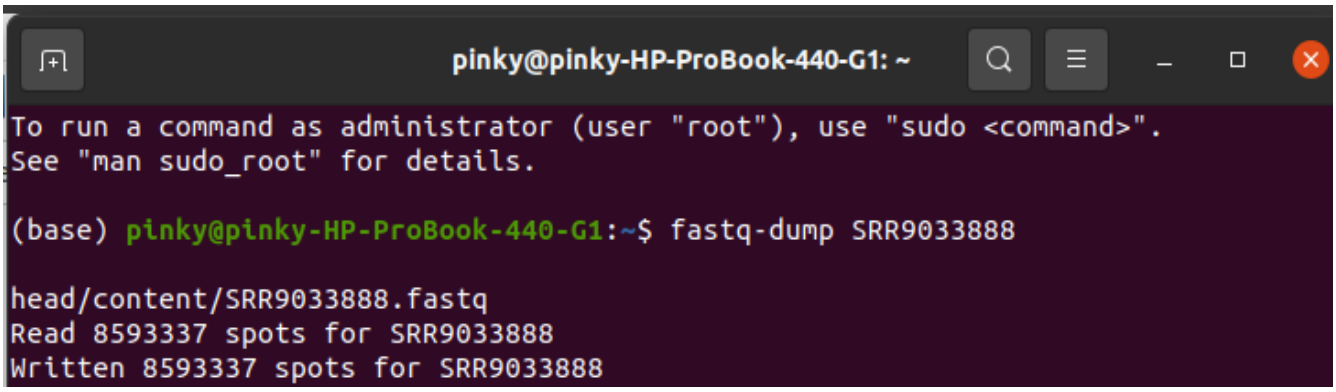
```
@SEQ_ID
AATTGTGGTTCAAAGCAGTGTCGATCAAATAGTAAATCCAGTTGTTCAAGTCACAGTCT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CaC%CCC65
```

Fastq Format In General

```
1 @SRR9033888.1 HWI-ST865:628:H72TMBCXY:1:1101:1232:1920 length=50
2 NACCGTTCTGTAAGCCTTCATCCCTTCGAATAGCTCAAAAGCATAATGGA
3 +SRR9033888.1 HWI-ST865:628:H72TMBCXY:1:1101:1232:1920 length=50
4 #<DDDHHIIIIHHHHIIIHIIHE@HHIC<<DCGGHGHIIIIGHIHGHIG@
```

The above example represent single sequence of fastq file SRR9033888. The sequencing was performed in single-end mode and this is an NCBI-assigned identifier as the sequencing gave the original read name that starts with SRR. The Fastq files from the INSDC Sequence Read Archive often include a description such as here, the description holds the original identifier from Illumina(HWI) plus the read length. The second line is about the raw sequence letters, the third line contains the information like as first line with '+' character except the '@' character, and the fourth line encodes the quality values of the line two sequence according to the ASCII value.

The modern usage of fastq almost always involves splitting the spot into its biological read, as described in submitter-provided metadata:



Encoding the Quality Values:

The quality values are encoding by the Phred score, also known as Q- Score is the quality score of a base. This is an integer value that represent the estimate probability of an error of a base. If P is the error probability, then :

$$P = 10^{-Q/10}$$

$$Q = -10 \log_{10}(P) \text{ ----------(1)}$$

The Q Scores are often represented as ASCII character. The following chart is representing how the Q-scores are determining from the ASCII character & what is the error rate:

ASCII_BASE=33, is now almost universally used, and ASCII_BASE 64 is used in some older Illumina data.

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

From the ASCII value table the higher the Q- value the lower the probability of error rate. Here, we find out that Q=3 (p=0.5), meaning that there is a 50% chance the base is wrong, and lower values represent the higher probabilities of error. When Q=2 (p=0.63), means the base call is more likely to be wrong than correct.

According to equation (1) the Phred quality scores are logarithmically linked to error probabilities are represented in the following chart:

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Accessing Quality With FastQC:

FastQC is quality control control checks to ensure the good quality of raw data without any biases. FastQC aims to provide a Quality Control report which can spot problems originated either in the sequencer or in the starting library material.

The main functions of FastQC are:

- Import data from FastQ, SAM or BAM files (any varient).
- Finding out the data problem by providing a quick overview.
- Summarize the graphs and tables for assessing the data quickly.
- Export the results to an HTML based permanent report.
- Offline operation can run to allow automated generation of reports without running the interactive application.

FastQC supports files in the following formats:

- ✓ fastq (all quality encoding variants)
- ✓ Casava fastq files
- ✓ Colorspace fastq
- ✓ Gzip compressed fastq
- ✓ SAM
- ✓ BAM

fastq: A text-based format for storing both a biological sequence and its corresponding quality scores.

Casava fastq files: The Casava fastq file format is same as regular fastq file except the data is usually split across multiple files for a single sample. In Casava mode the program exclude the poor quality sequences which have been flagged to be removed from the report.

Colorspace fastq: Colorspace fastq is the sequencing for SOLiD (Sequencing by Oligonucleotide Ligation and Detection) data, except the first position. The quality values are those of Sanger format. Some include a quality score (set to 0, i,e. '!') for the leading nucleotide, others do not. The sequence read archive includes this quality score. The single end Colorspace fastq file ends with .csfastq and paired end Colorspace fastq end with pair.1.csfastq and pair.2.csfastq.

Gzip compressed fastq: Gzip (GNU zip) is a free and open source algorithm for the file compression to reduce the large size of fastq files. The compression is done to deliberate reduction of the file size to save storage space or increase the data transfer rate.

SAM: Sequence Alignment Map (SAM) is a text-based format for storing the biological sequences aligned to a reference sequence (RefSeq) database build by NCBI,. It is applied to the data generated by next generation sequencing technologies and the standard has been boardened to include unmapped sequences. The SAM format consists of a header and an alignment section. SAM files can be analysed and edited with the SAMtools software. The heading begins with the '@' symbol.

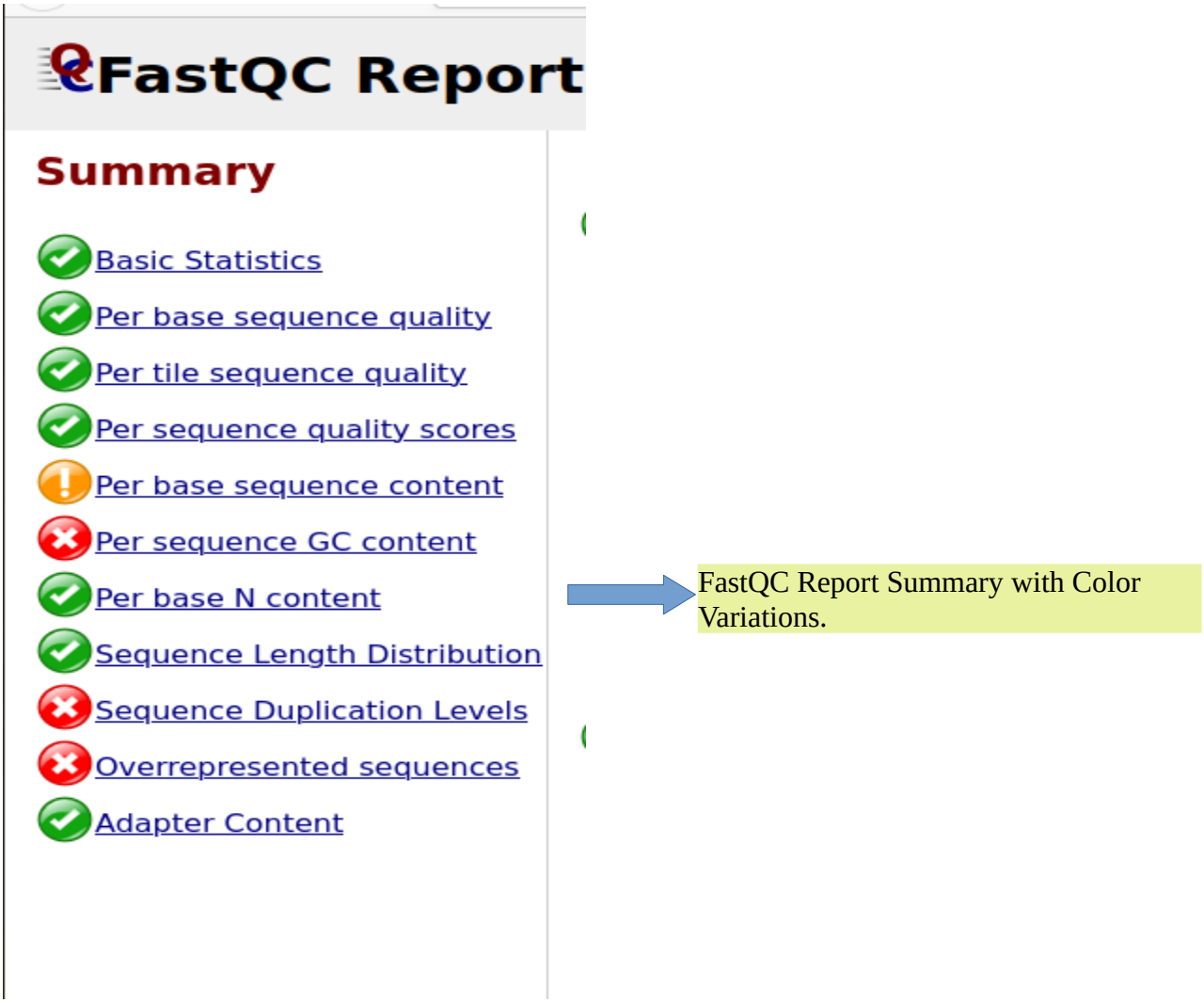
BAM: Binary Alignment Map (BAM) is the comprehensive binary representation of the Sequence Alignment Map (SAM) files. BAM consists of compressed raw data of genome-sequencing.

fastQC analyze the SAM and BAM files with .SAM and .BAM extensions.

Evaluating the FastQC Results:

The FastQC analysis is performed by a series of analysis modules. The left hand side of the main interactive display or the top of the HTML report show a summary of the modules which were run. The evaluation reports are showing different color of meaning about the results.

- Entirely Normal (Green Tick)
- Slightly Abnormal (Orange Triangle)
- Very unusual (Red Cross)



FastQC Report Summary with Color Variations.

Analysis Modules:

1. Basic Statistics:

The Basic Statistics module generates some simple composition statistics for the file analysed.

- Filename: The original filename of the analysed file.
- File type: Says whether the file appeared to contain conventional base calls or colorspace data which had to be converted to base calls
- Encoding: Says which ASCII encoding of quality values was found in this file.
- Total Sequences: A count of the total number of sequences processed. There are two values reported, actual and estimated.
- Filtered Sequences: The sequences flagged as poor quality will be filtered. If running in Casava mode sequences flagged to be filtered will be removed from all analyses. The number of such sequences removed will be reported here. The total sequences count above will not include these filtered sequences and will the number of sequences actually used for the rest of the analysis.

- Sequence Length: Provides the length of the shortest and longest sequence in the set. If all sequences are the same length only one value is reported.
- %GC: The overall %GC of all bases in all sequences
- Warning and Error: Basic statistics never raises a warning and an error.

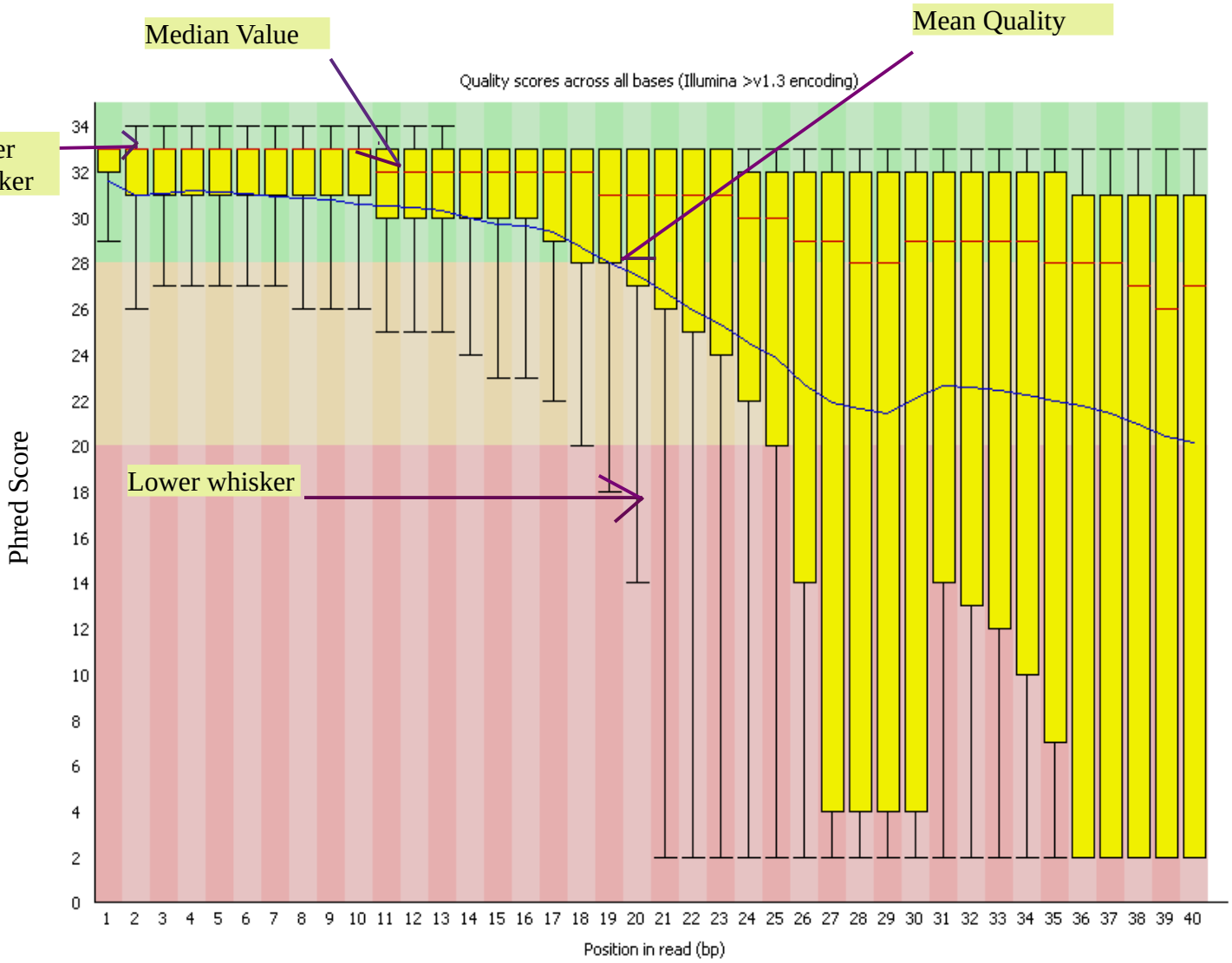
Basic Statistics

Measure	Value
Filename	SRR5259850.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	48889529
Sequences flagged as poor quality	0
Sequence length	50
%GC	42

2. Per Base Sequence Quality

It shows an overview of the range of quality, measured as Phred Scores across all bases at each position in the raw reads. The score can be divided into different categories according to the color and Phred scores:

- Green (Phred Score more than 30)= Very good quality calls
- Orange (Phred Score between 20-30) = Reasonable quality calls
- Red (Phred Score less than 20) = Poor quality calls



According to the figure:

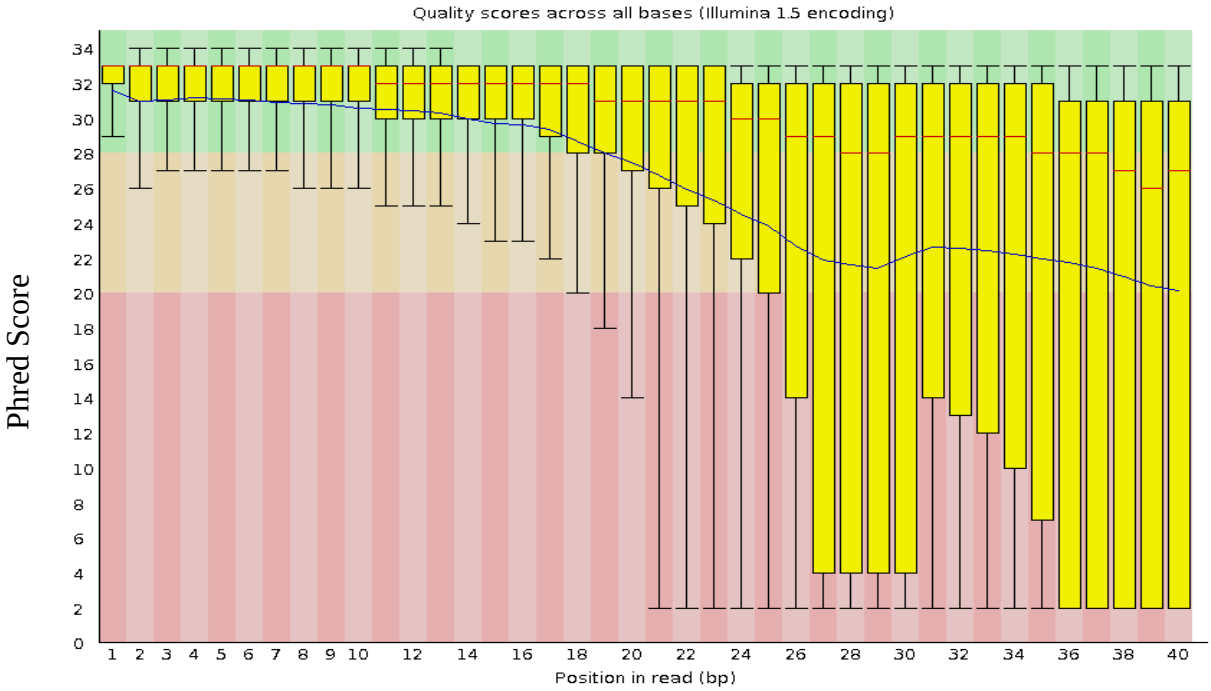
Central red line = Median Value

Blue line= Mean Quality

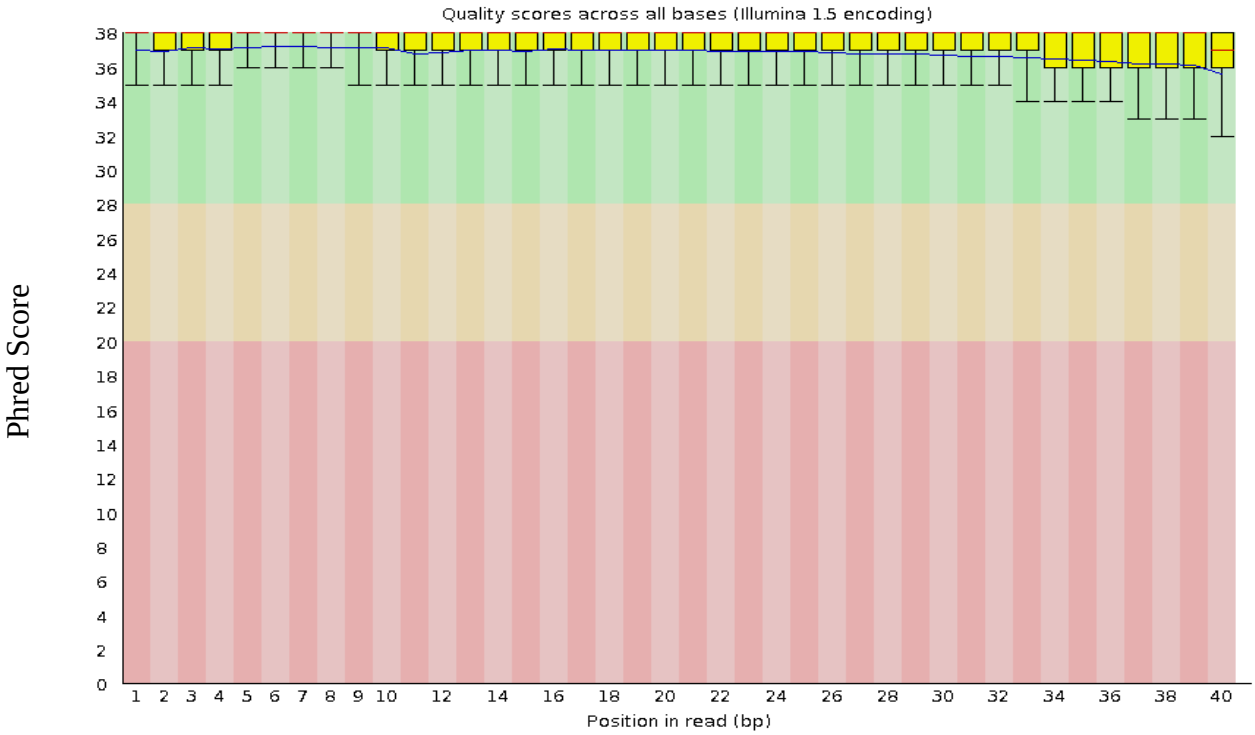
Yellow box = Inter-quartile range (25-70 %)

Upper and lower whiskers = 10% and 90% points. The whiskers drawn up to the 10th percentile and and drawn down to the 90^{h} percentile.

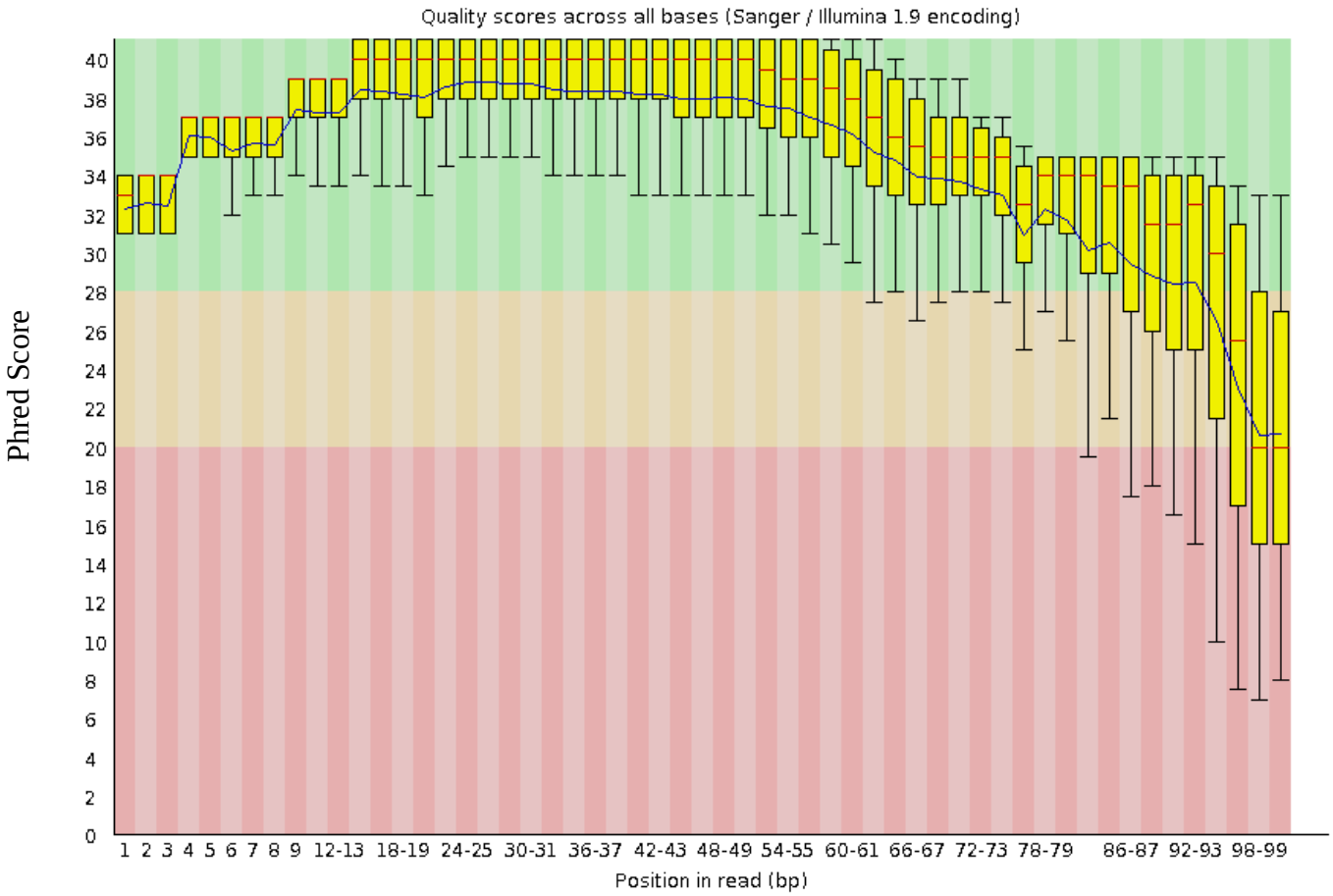
Per base sequence quality



Per base sequence quality



Per base sequence quality



Warning (Orange Triangle):

- Lower quartile for any bases < 10 or,
- Median for any bases < 25

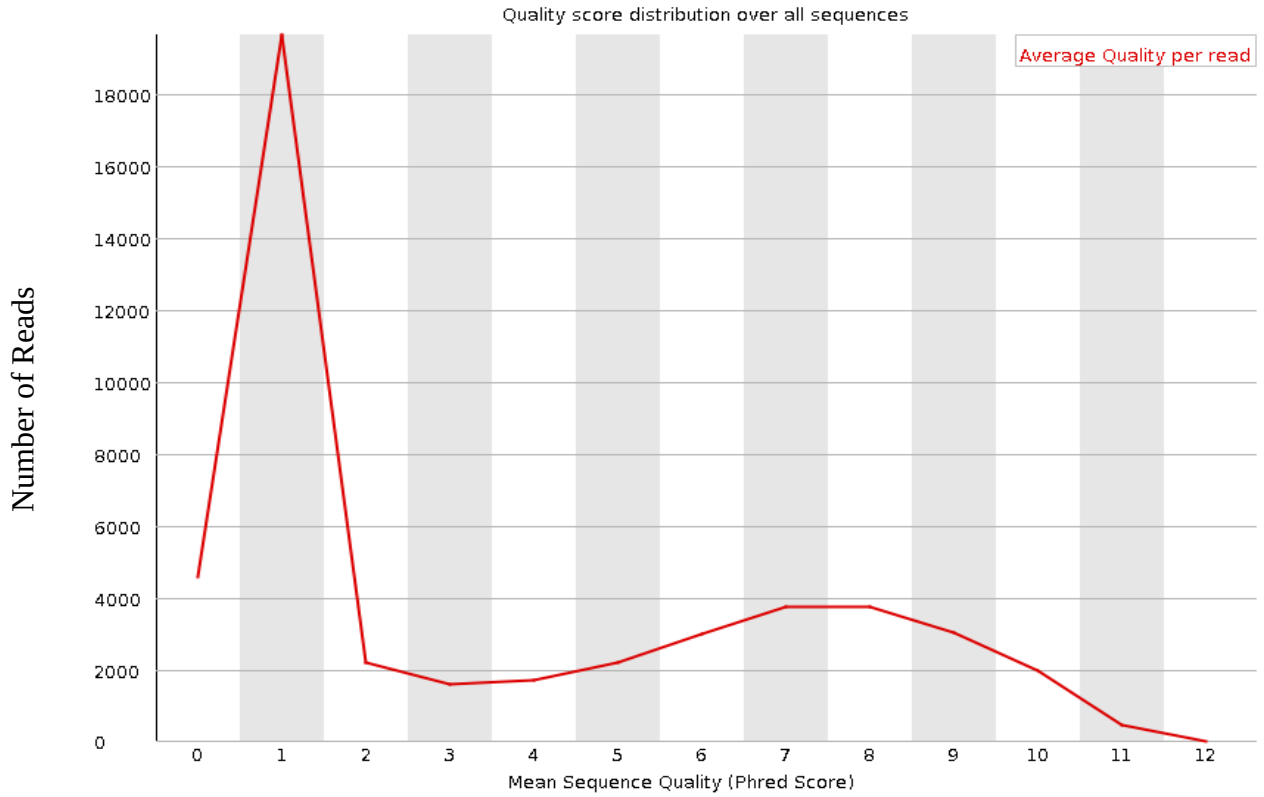
Failure (Red Cross):

- Lower quartile for any bases < 5 or,
- Median for any bases < 20

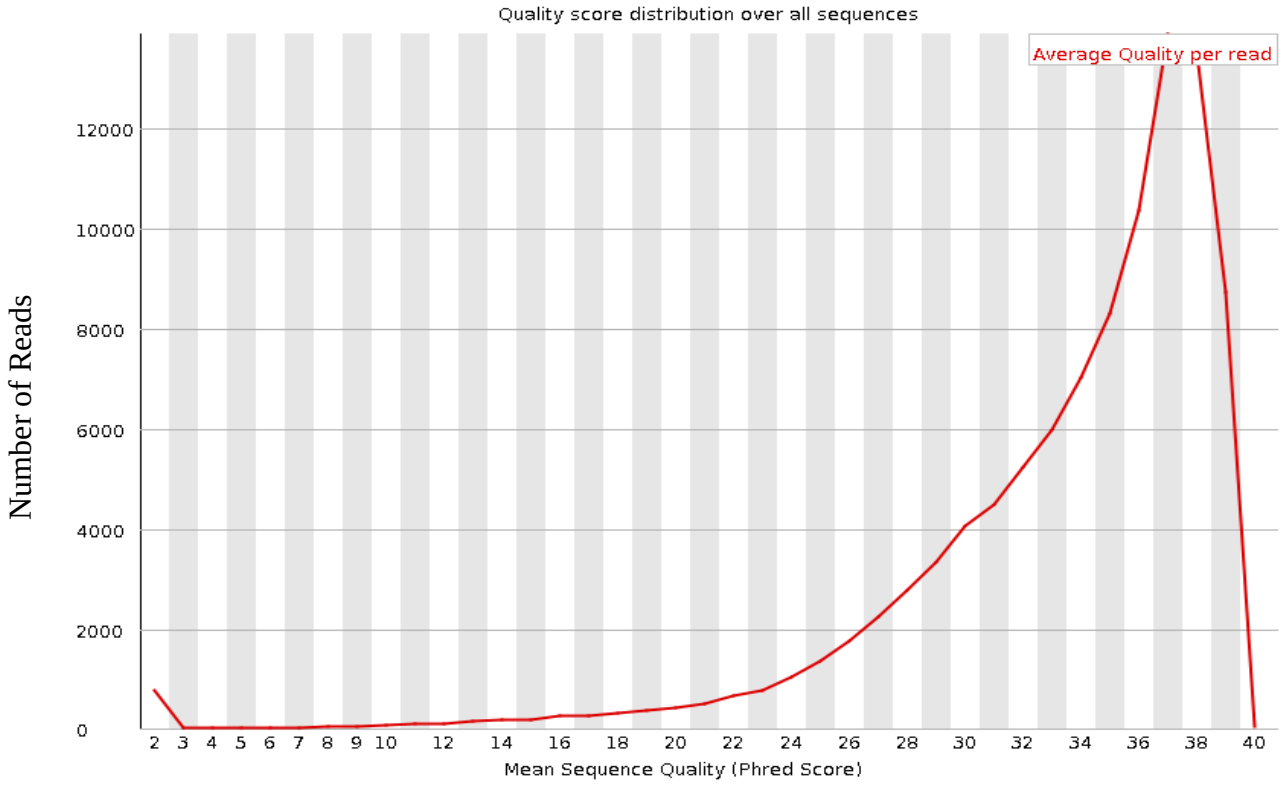
3. Per Sequence Quality Scores:

It is a plot of the total number of reads vs the average quality score (Phred Score) over full length of that read. The distribution of average read quality should be fairly tight in the upper range of the plot and these should represent only a small percentage of the total sequences.

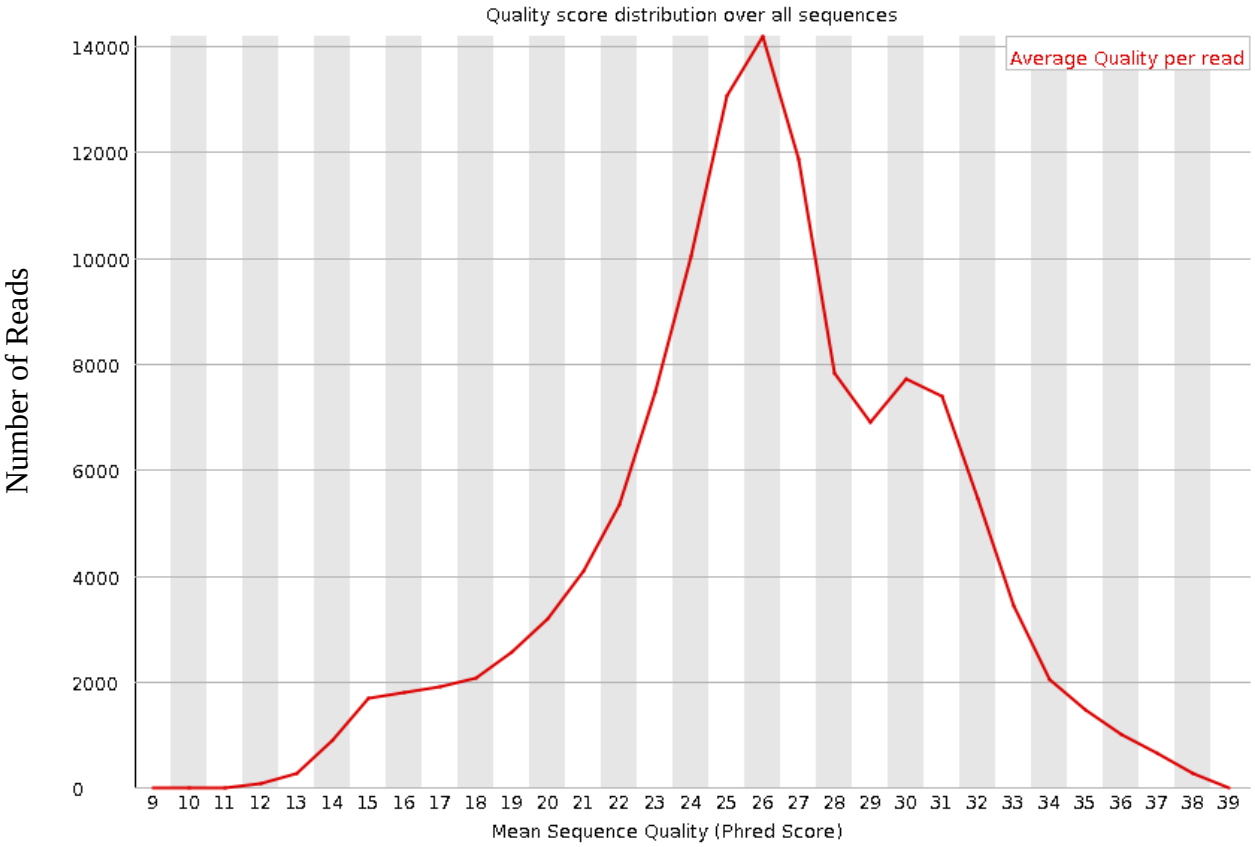
Per sequence quality scores



Per sequence quality scores



Per sequence quality scores



Warning (Orange Triangle) :

Most frequently observed mean quality < 27 (Equates to a 0.2% error rate).

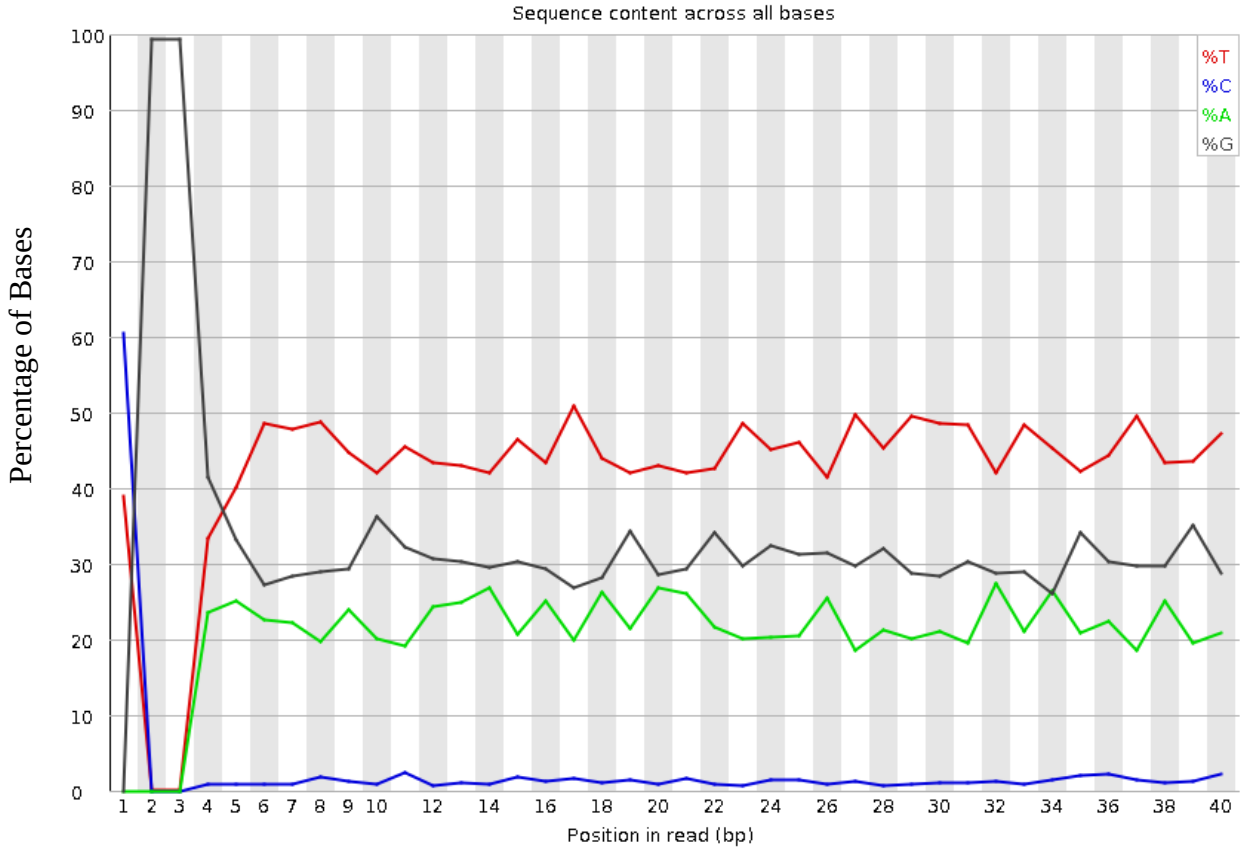
Failure (Red Cross):

Most frequently observed mean quality < 20 (Equates to a 1% error rate).

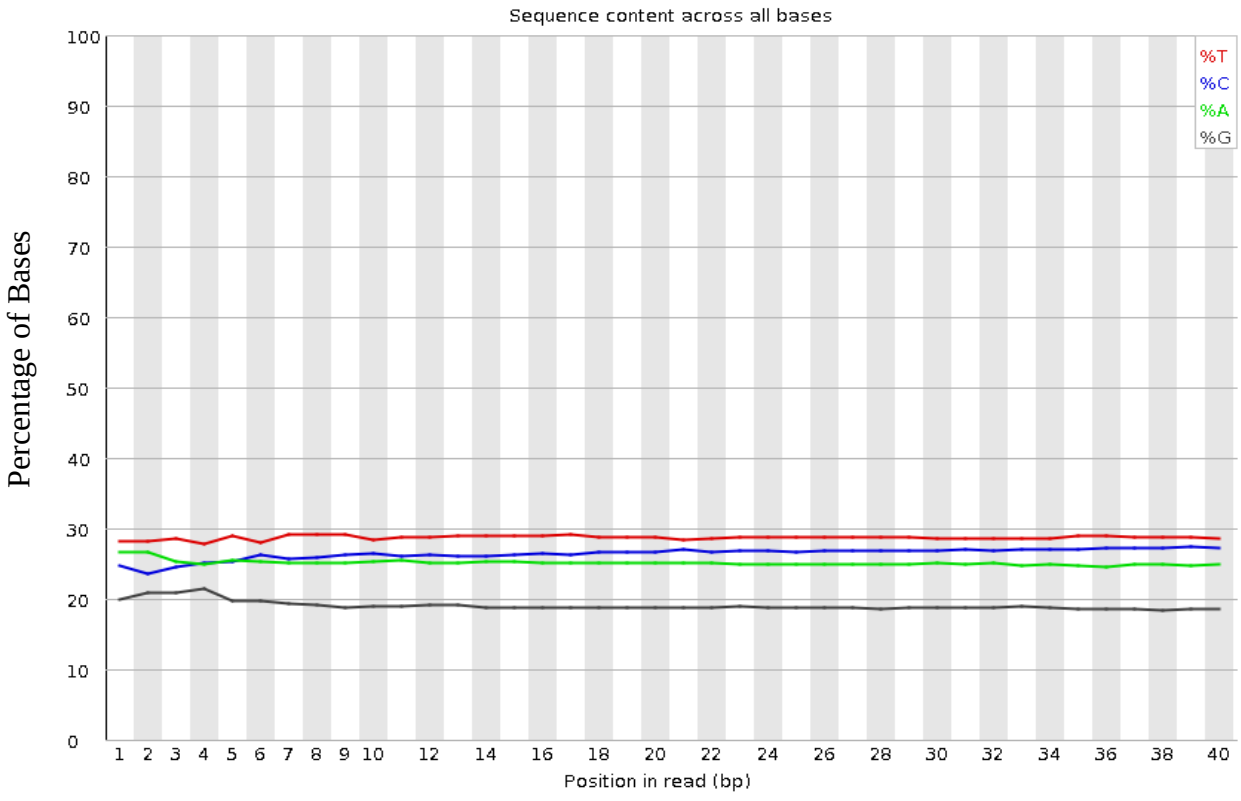
4. Per Base Sequence Content

This plots represent the percentage of bases called for each base position for each of the four nucleotides (A,T,G,C) across all read files in the file. In random library it is expect that there would be little or no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other. The relative amount of each base need to reflect the overall amount of these bases in the genome, but the lines should not be largely imbalanced from each other. If there is strong imbalances among the bases it usually indicates that overrepresented sequence contaminated the library. A bias which is consistent across all bases indicates that the original library was sequenced biased, or that there was a systematic problem during the sequencing of the library. Therefore, for whole genome sequencing the proportion of each of the four bases should remain relatively constant over the length of the read with %A=%T and %G= %C.

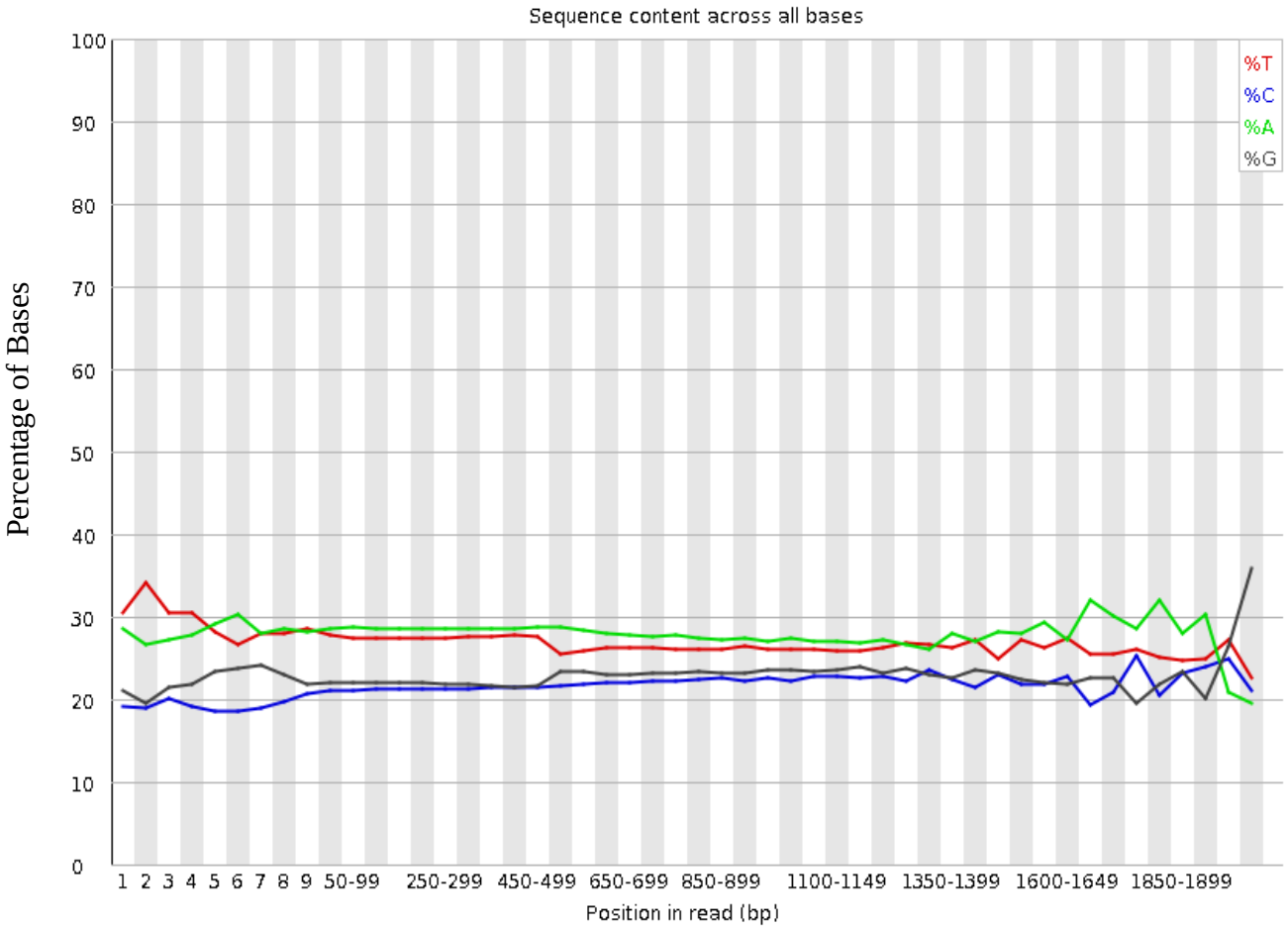
Per base sequence content



Per base sequence content



Per base sequence content



Warning (Orange Triangle) :

The differences between A and T, or G and C is greater than 10% in any position.

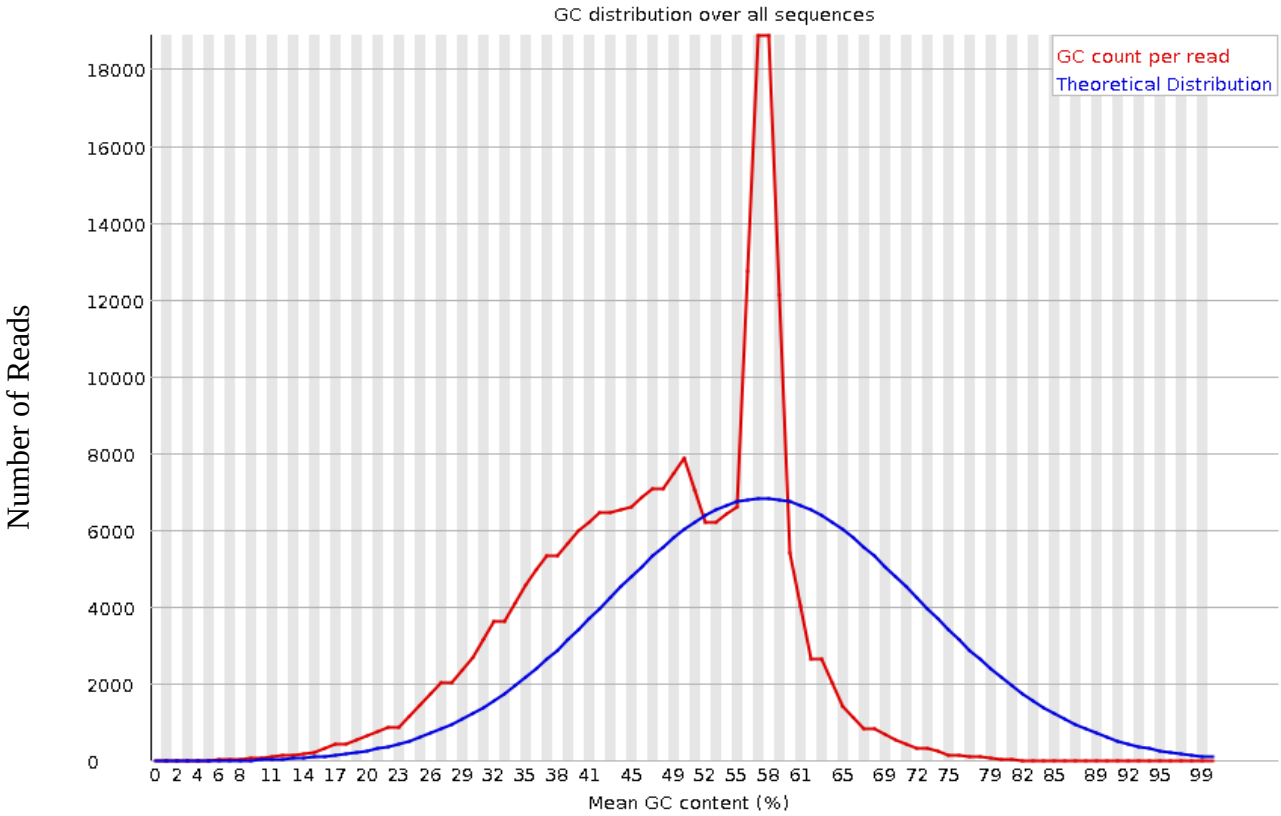
Failure (Red Mark):

The differences between A and T, or G and C is greater than 20% in any position.

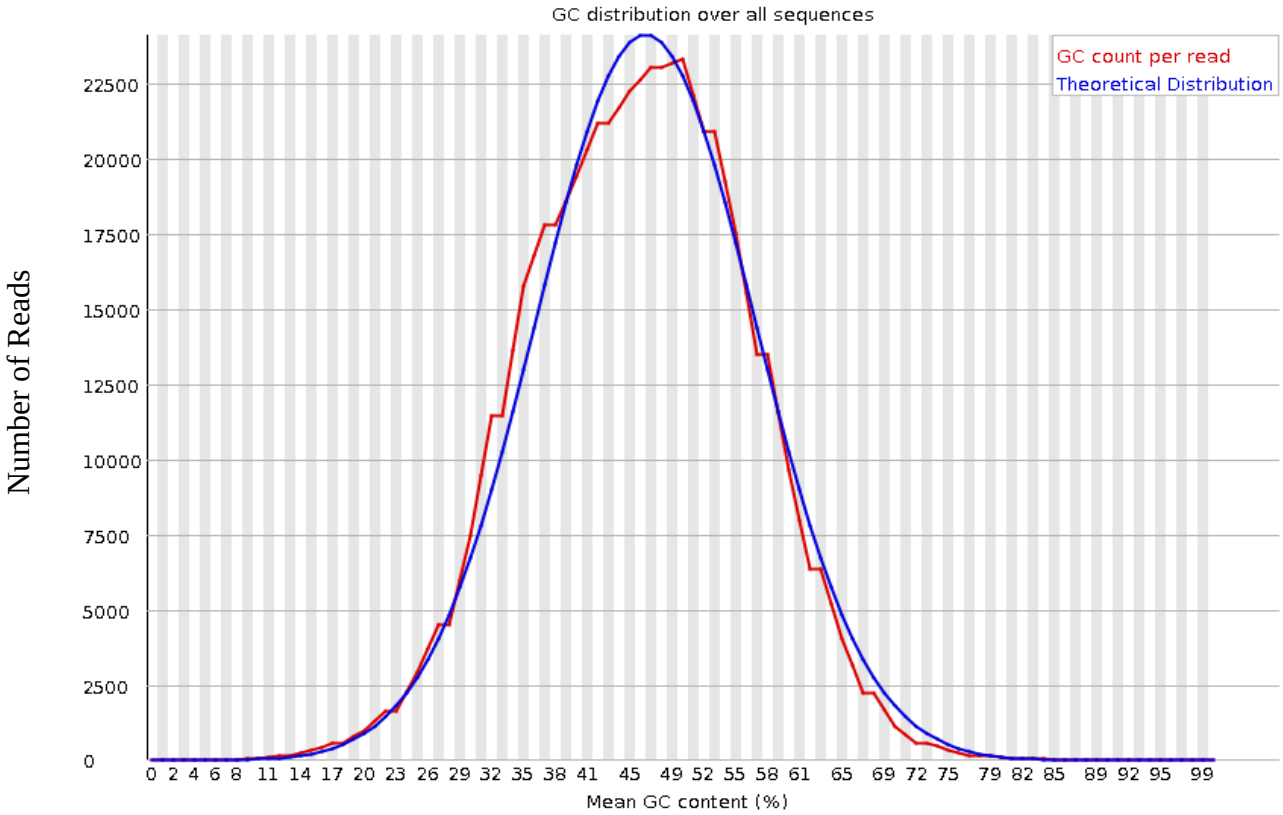
5. Per Sequence GC Content

It measures the GC content across the whole length of each sequence in a raw read file and compares it to a modelled normal distribution of GC content. It is expected that the GC content of all reads should form a normal distribution with the peak of the theoretical distributed curve at the mean GC content for the organism sequenced. If the observed distribution deviates too far from the theoretical, fastQC will call a fail.

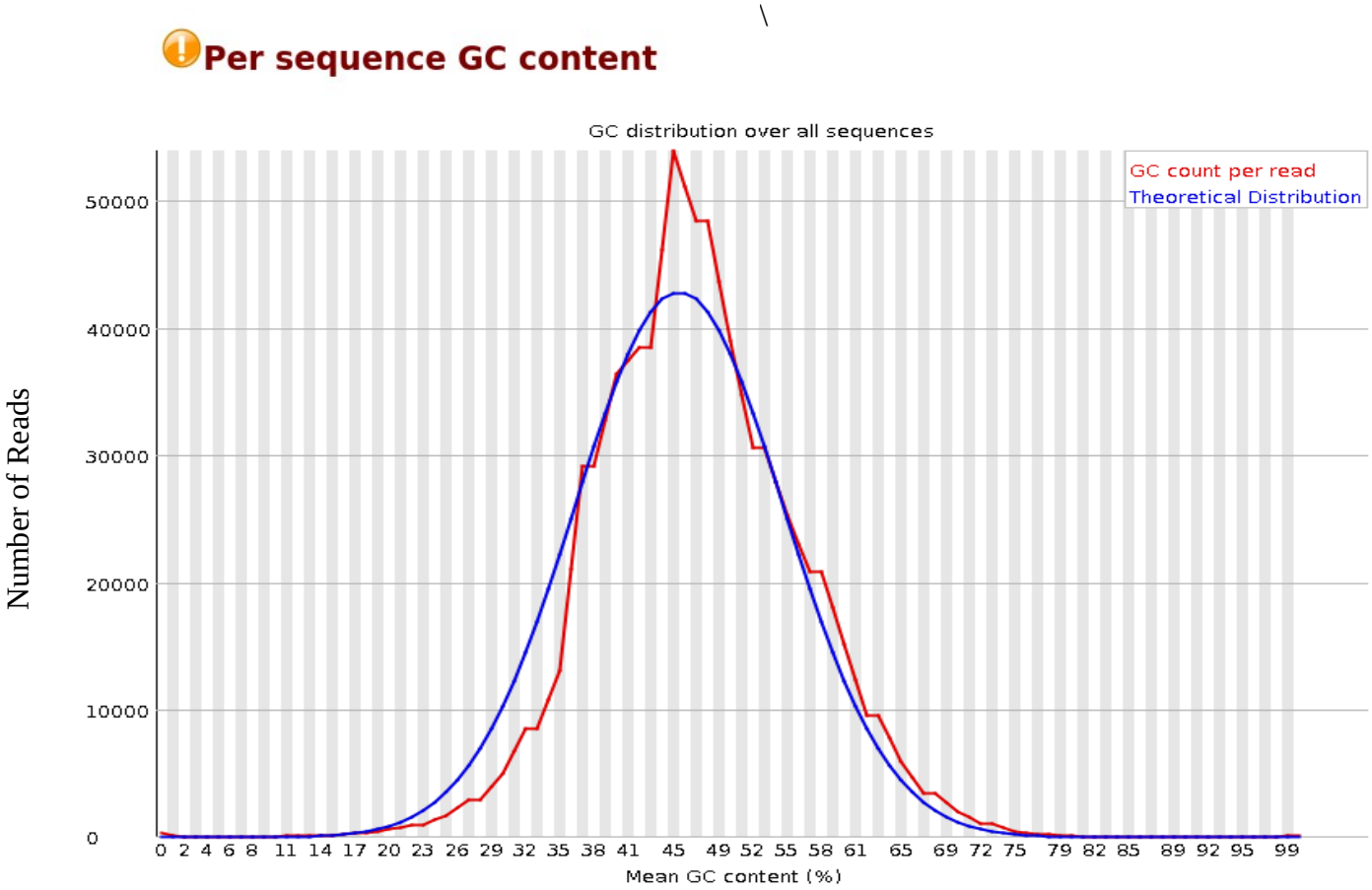
Per sequence GC content



Per sequence GC content



\



Warning (Orange Triangle):

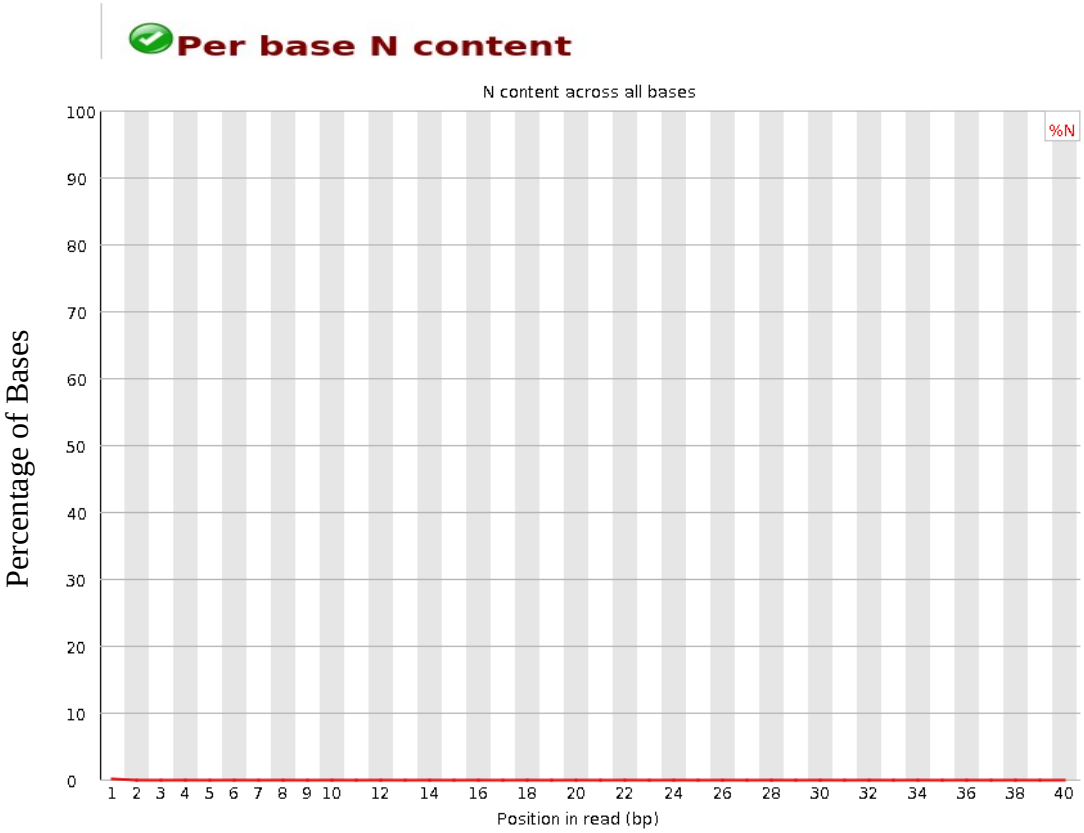
If the sum of the deviations from the normal distribution represents more than 15% of the reads.

Failure (Red Cross):

If the sum of the deviations from the normal distribution represents more than 30% of the reads.

6. Per Base N Content

Percent of bases at each position with no base call, i.e. ‘N’.



Warning (Orange Triangle):

If any position shows an N content of >5%.

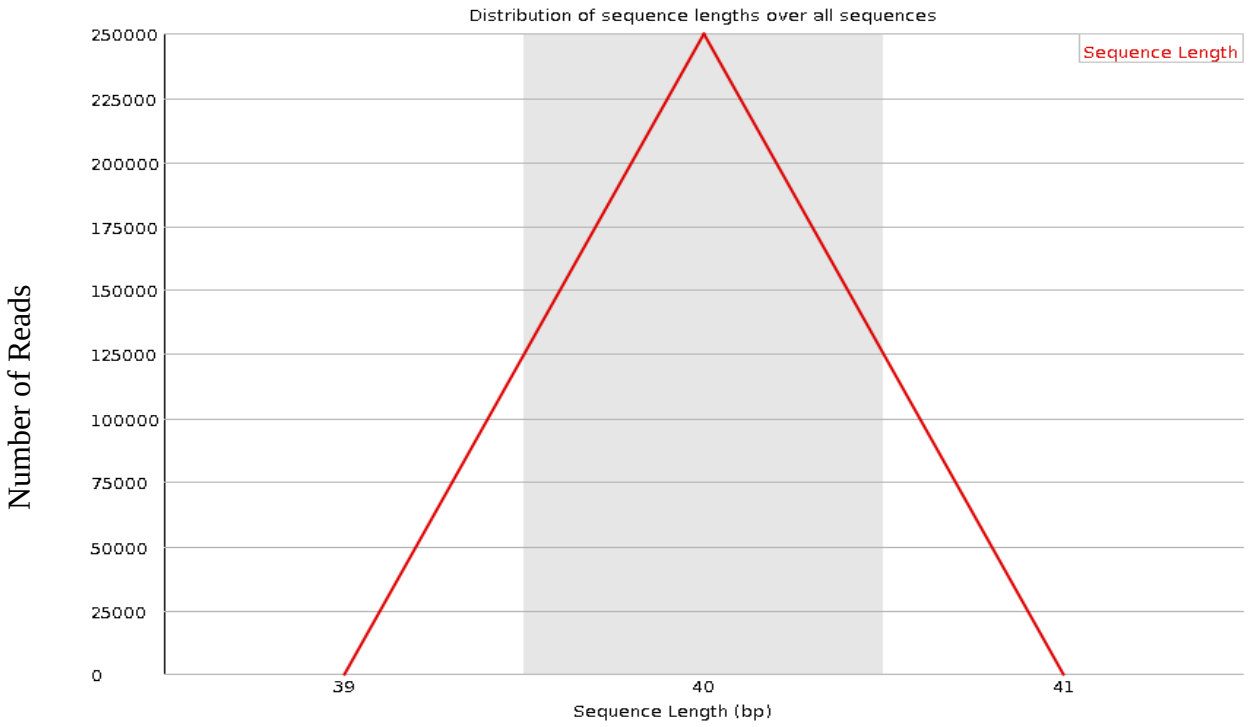
Failure (Red Cross):

If any position shows an N content of >20%.

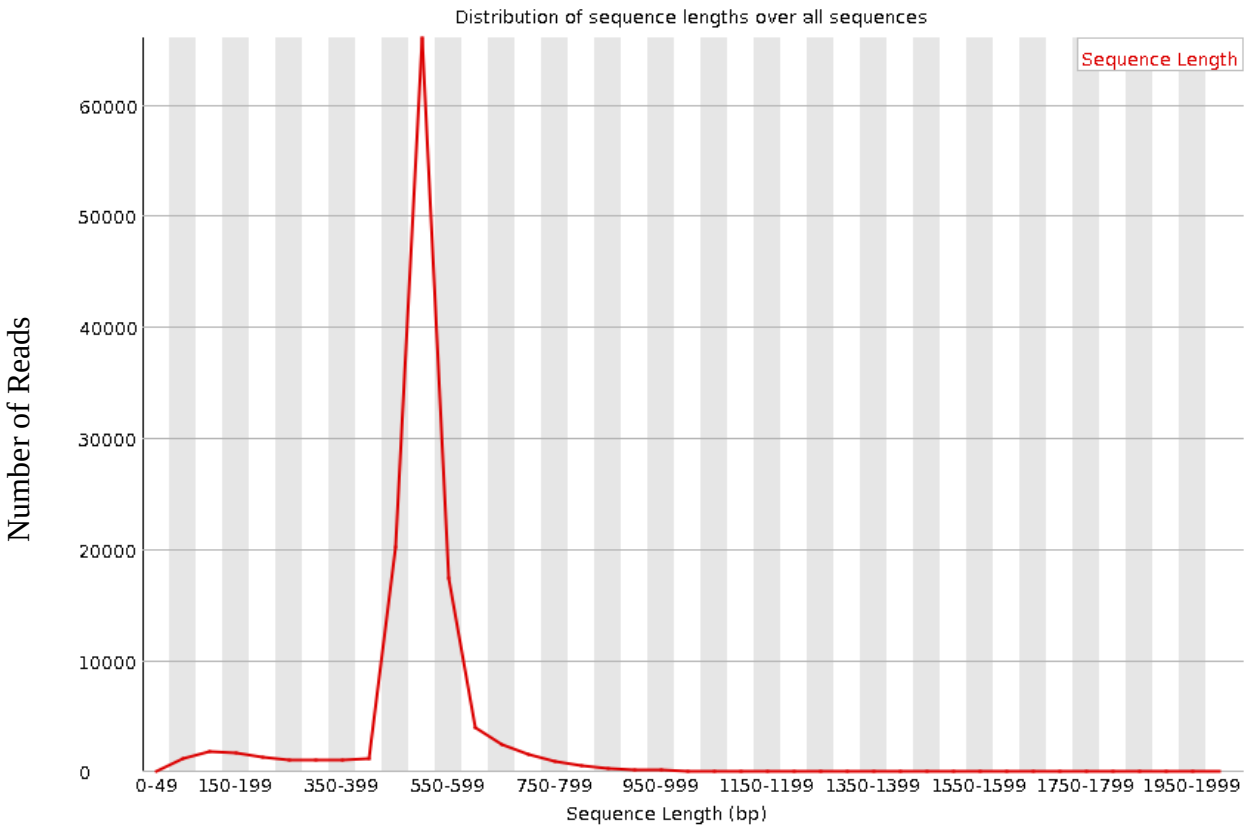
7. Sequence Length Distribution

Some high throughput sequencers generate sequence fragments of uniform length, but others can contain reads of wildly varying lengths. Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end.

Sequence Length Distribution



Sequence Length Distribution



Warning (Orange Triangle):

If all sequences are not the same length.

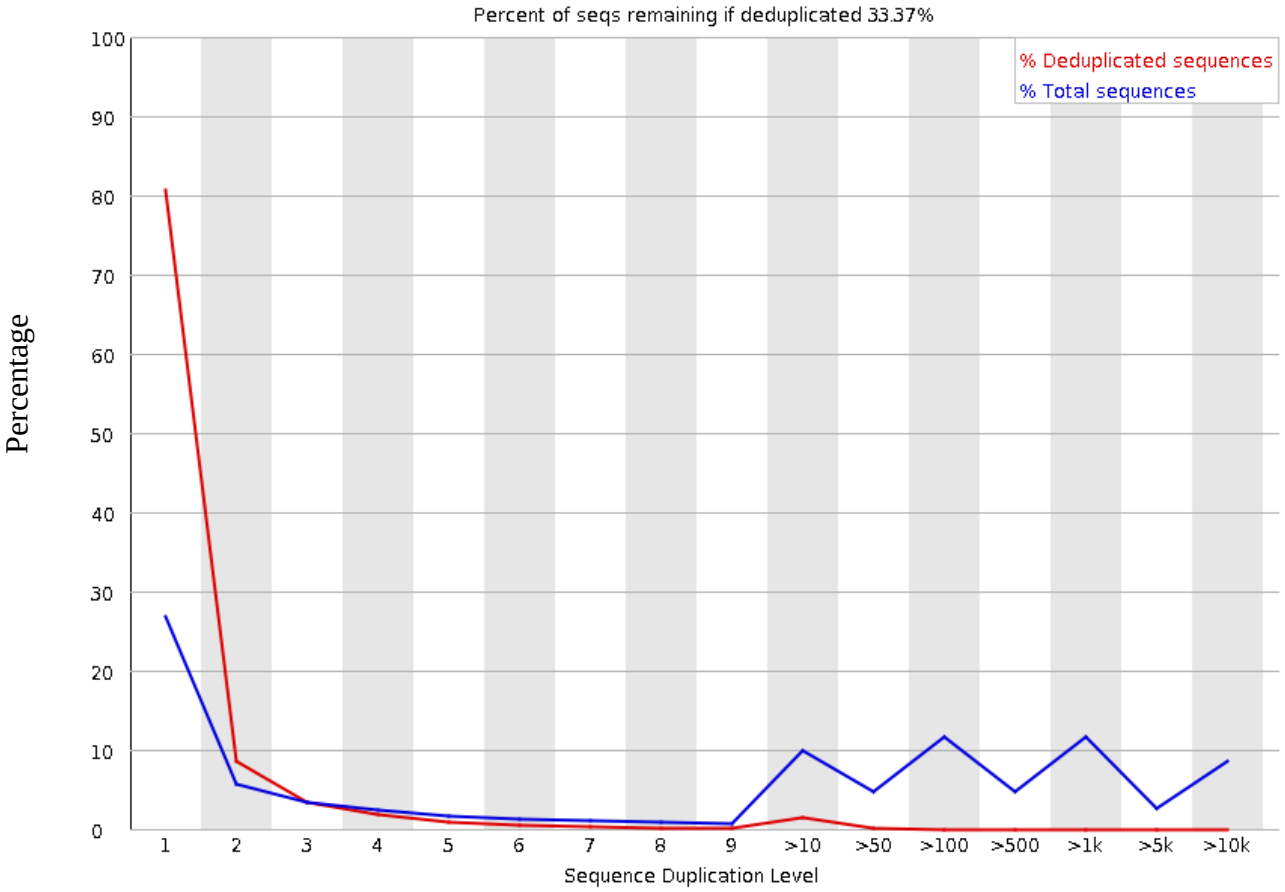
Failure (Red Cross) :

If any of the sequences have zero length.

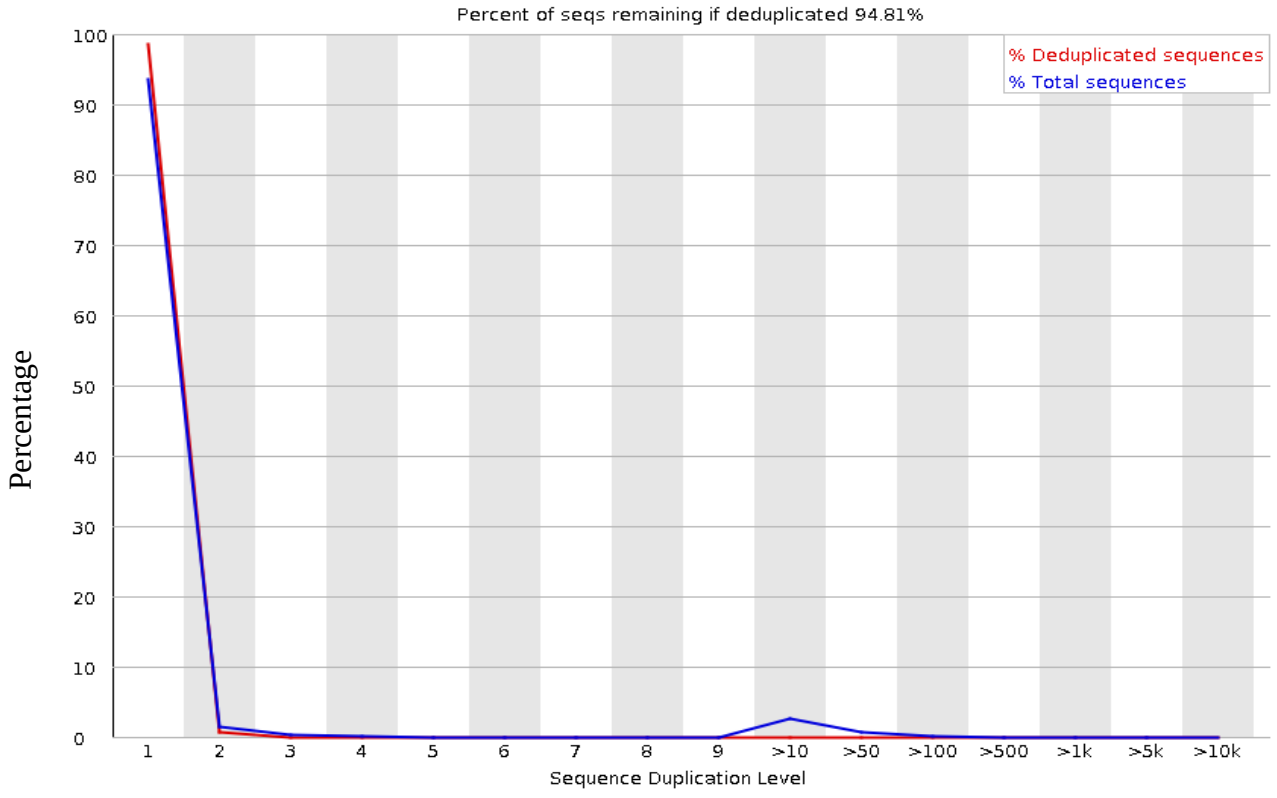
8. Duplicated Sequences

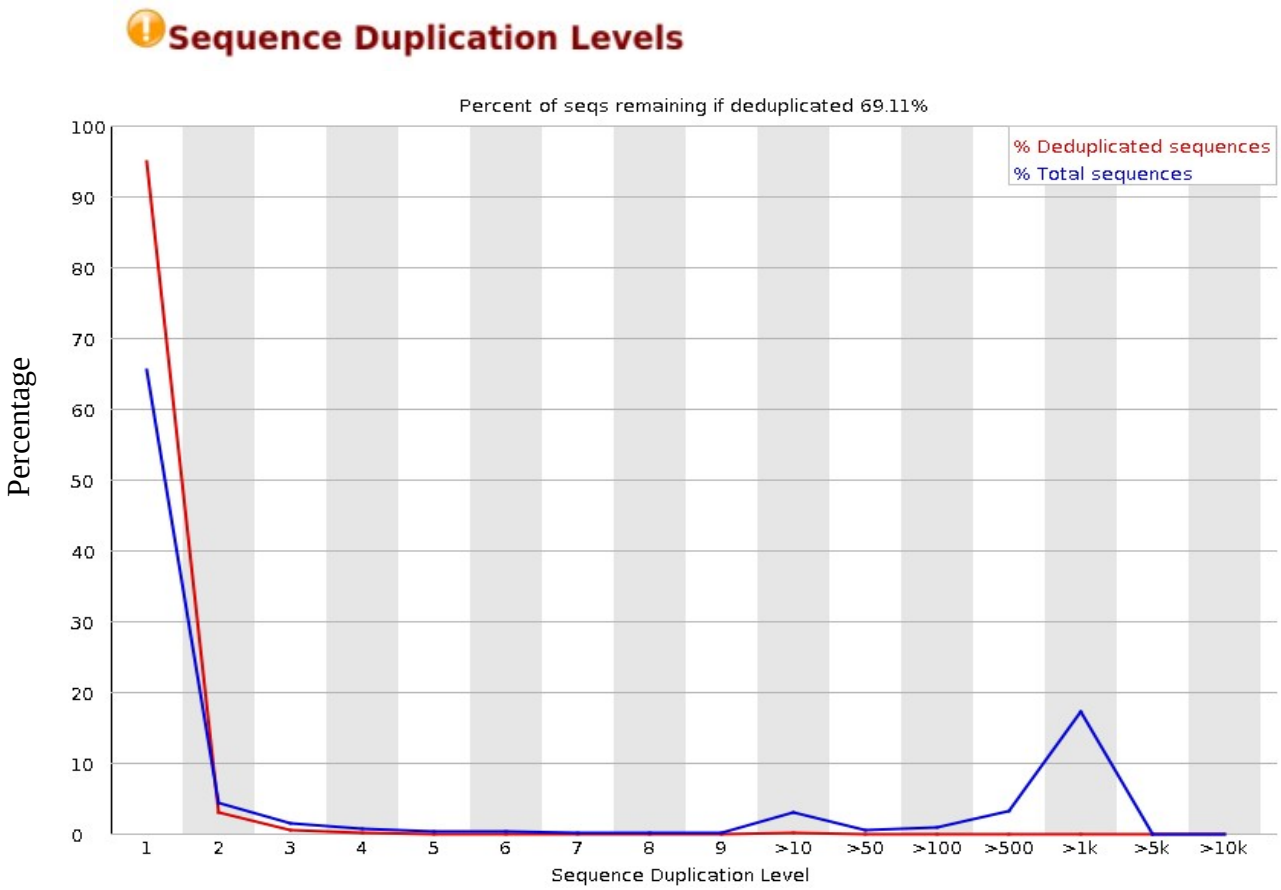
For whole genome sequencing data it is expected that nearly 100% of all reads will be unique (appearing only 1 time in the sequence data). This indicates a highly diverse library that was not over sequenced. A low level of duplication may indicate a very high level of coverage of the target sequence, but a high level of duplication is more likely to indicate enrichment bias (e.g. PCR over amplification). This module counts the degree of duplication for every sequence in a library and creates a plot showing the relative number of sequences with different degrees of duplication.

Sequence Duplication Levels



Sequence Duplication Levels





Warning (Orange Triangle):

If non-unique (duplicated) sequences make up more than 20% of the total.

Failure (Red Mark):

If non-unique (duplicated) sequences make up more than 50% of the total.

9. Overrepresented Sequences

List of sequences which appear more than expected in the file. Only the first 50bp are considered for the purpose of analysis. A sequence is considered overrepresented if it accounts for ≥ 0.1% of the total reads. Each overrepresented sequence is compared to a list of common contaminants to try to identify it. For each overrepresented sequence the program will look for matches in a database of common contaminants and will report the best hit it finds. Hits must be at least 20bp in length and have no more than 1 mismatch. Finding a hit doesn't necessarily mean that this is the source of the contamination, but may point the right direction. It's also worth pointing out that many adapter sequences are very similar to each other so it is possible to get a hit reported which isn't technically correct, but which has very similar sequence to the actual match. For good Illumina data there is no overrepresented sequences.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TGAGGTAGTAGATTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	10865	4.346	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TAGCTTATCAGACTGATGTTGACAGATCGGAAGAGCACACGTCTGAACTC	10845	4.338	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TCTTTGGTTATCTAGCTGTATGAGATCGGAAGAGCACACGTCTGAACTCC	7062	2.8247999999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATGAAGATCGGAAGAGCACACGTCTGAACTC	4056	1.6223999999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TGAGGTAGTAGTTTGTGCTGTTAGATCGGAAGAGCACACGTCTGAACTCC	3737	1.4948	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGTTTGTACAGTTAGATCGGAAGAGCACACGTCTGAACTCC	3549	1.4196	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGGTTGTATGGTTAGATCGGAAGAGCACACGTCTGAACTCC	2931	1.1724	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
AACCCGTAGATCCGATCTTGTAGATCGGAAGAGCACACGTCTGAACTCCA	1910	0.764	Illumina Multiplexing PCR Primer 2.01 (100% over 29bp)
CGCGACCTCAGATCAGACGTAGATCGGAAGAGCACACGTCTGAACTCCAG	1749	0.6996	Illumina Multiplexing PCR Primer 2.01 (100% over 30bp)
TGAGGTAGTAGGTTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	1647	0.6588	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTC	2065	0.5224039181558763	No Hit
GATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATG	2047	0.5178502762542754	No Hit
ATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCATGA	2014	0.5095019327680071	No Hit
CGATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTAT	1913	0.4839509420979134	No Hit
GTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGA	1879	0.47534961850600066	No Hit
AAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCT	1846	0.4670012750197325	No Hit
TGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCAT	1841	0.46573637449150995	No Hit
AACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAGTTAA	1836	0.46447147396328753	No Hit
GATAAAAATGATTGGCGTATCCAACCTGCAGAGTTTTATC	1831	0.4632065734350651	No Hit
AAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTC	1779	0.45005160794155147	No Hit
ATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCCA	1779	0.45005160794155147	No Hit
AATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTTCC	1760	0.4452449859343061	No Hit
AAAATGATTGGCGTATCCAACCTGCAGAGTTTTATCGCTT	1729	0.4374026026593269	No Hit
CGTATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAG	1713	0.43335492096901496	No Hit
ATCCAACCTGCAGAGTTTTATCGCTTCCATGACGCAGAAG	1708	0.43209002044079253	No Hit

Warning (Orange Triangle) :

If any sequence is found to represent more than 0.1% of the total.

Failure (Red Cross) :

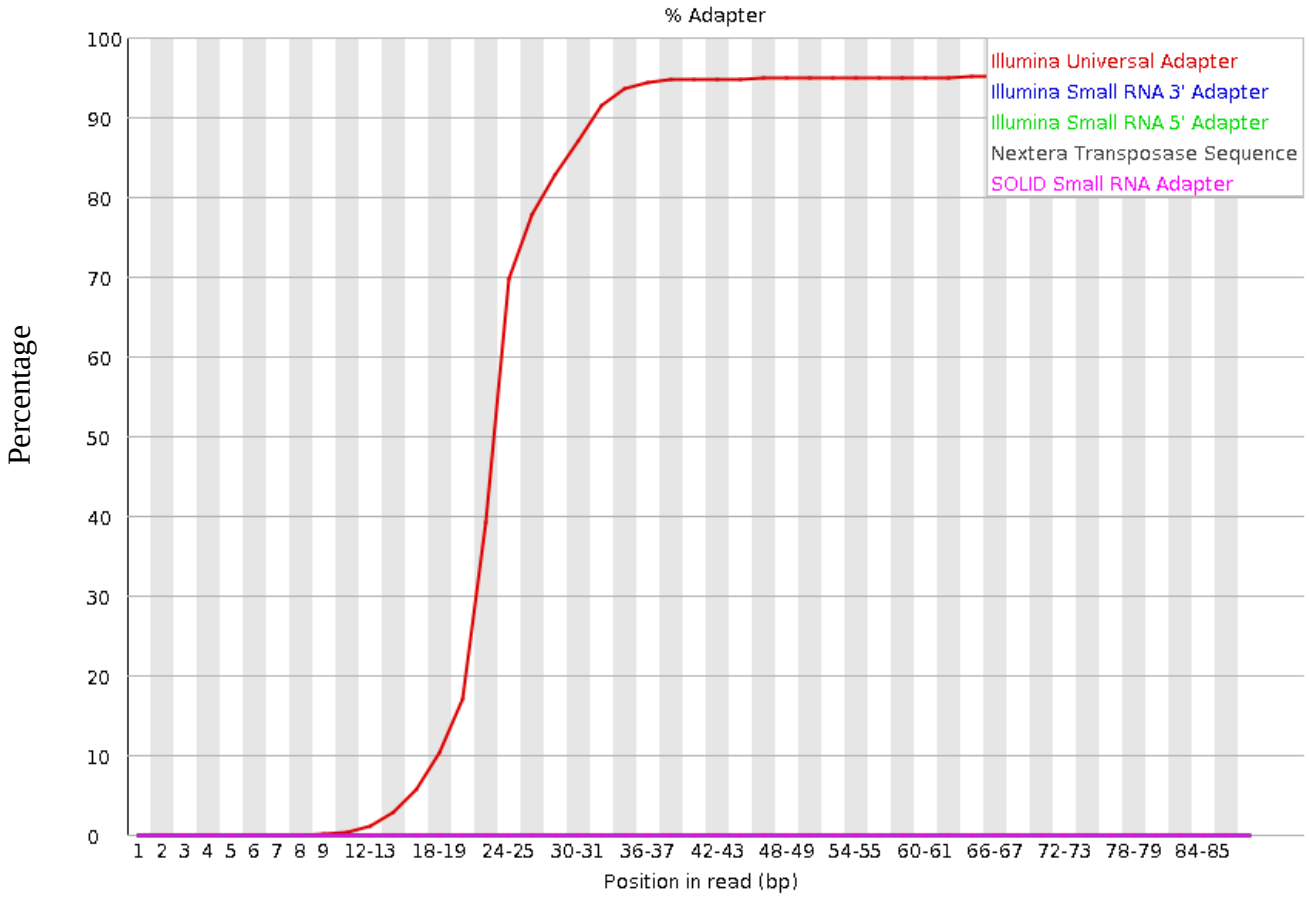
If any sequence is found to represent more than 1% of the total.

10. Adapter Content:

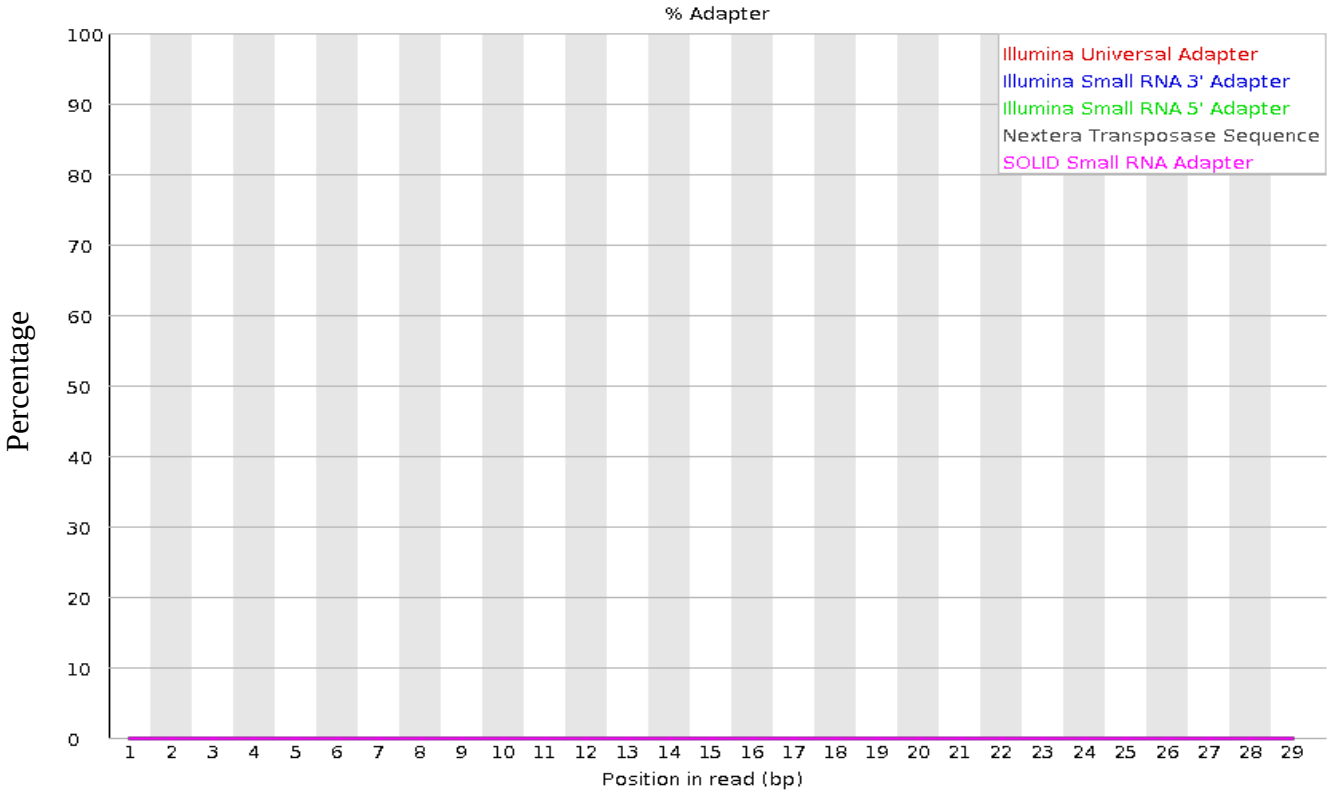
Cumulative plot of the fraction of reads where the sequence library adapter sequence is identified at the base position. Sequence adaptors are any kind of short DNA sequence. Only adapters specific to the library type are searched. Ideally Illumina sequence data should not have any adapter sequence present. However when using long read lengths it is possible that some of the library inserts are shorter than the read length. This ultimately resulting in the read-through adapters at the 3' end of the read. This is more likely to occur with RNA-Seq libraries where the distribution of library insert sizes is more varied and likely to include some short inserts.

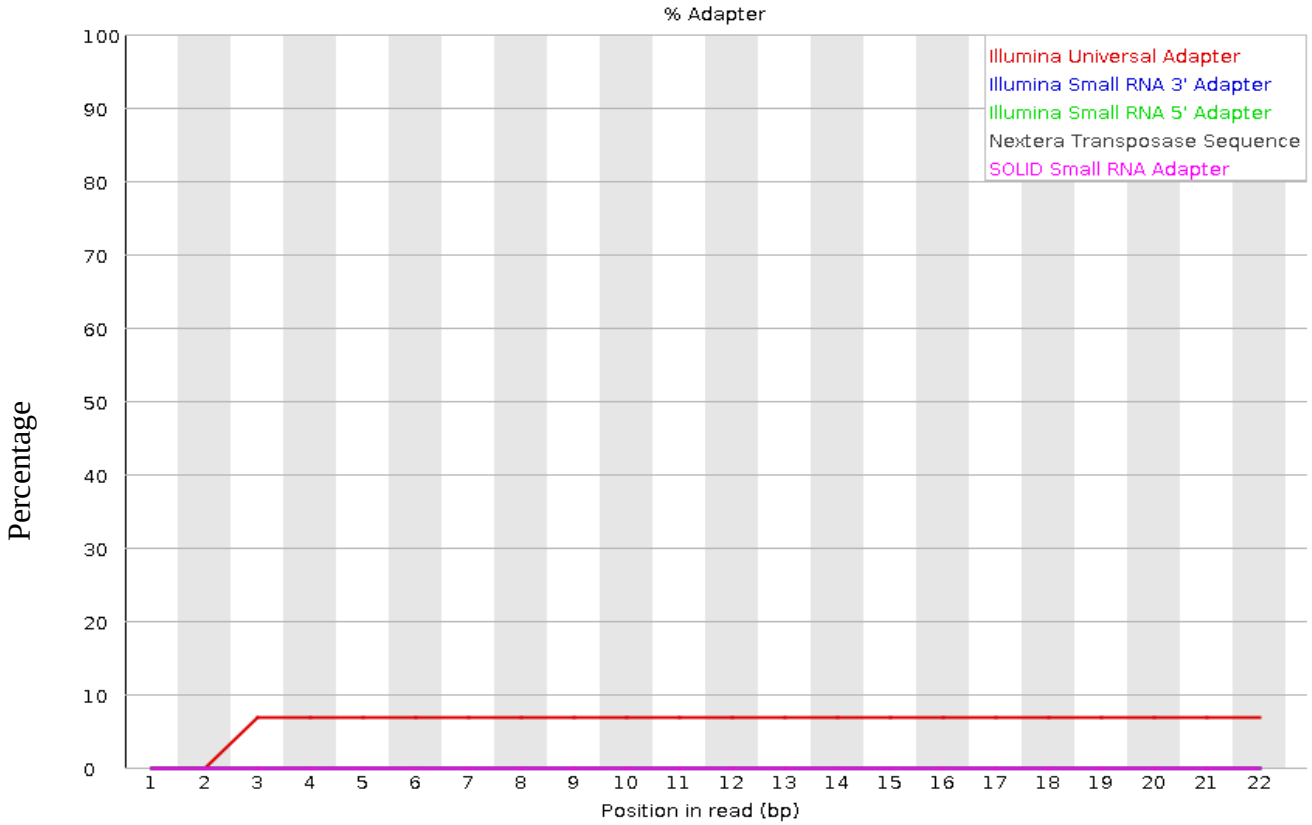
***Read-through adapters: In some cases the length of the read inserts in the library might be quite short – short enough that in some cases they are shorter than the read length of the sequencing run. Where the read length exceeds the length of the insert then the read will run off the end of the desired sequence and into the adapter on the other end. This will lead to the addition of non-native sequence on the end of some reads in the library. These will appear at various positions within the length of the reads, but will always consist of the same sequence.

Adapter Content



Adapter Content





Warning (Orange Triangle):

If any sequence is present in more than 5% of all reads.

Failure (Red Cross) :

If any sequence is present in more than 10% of all reads.

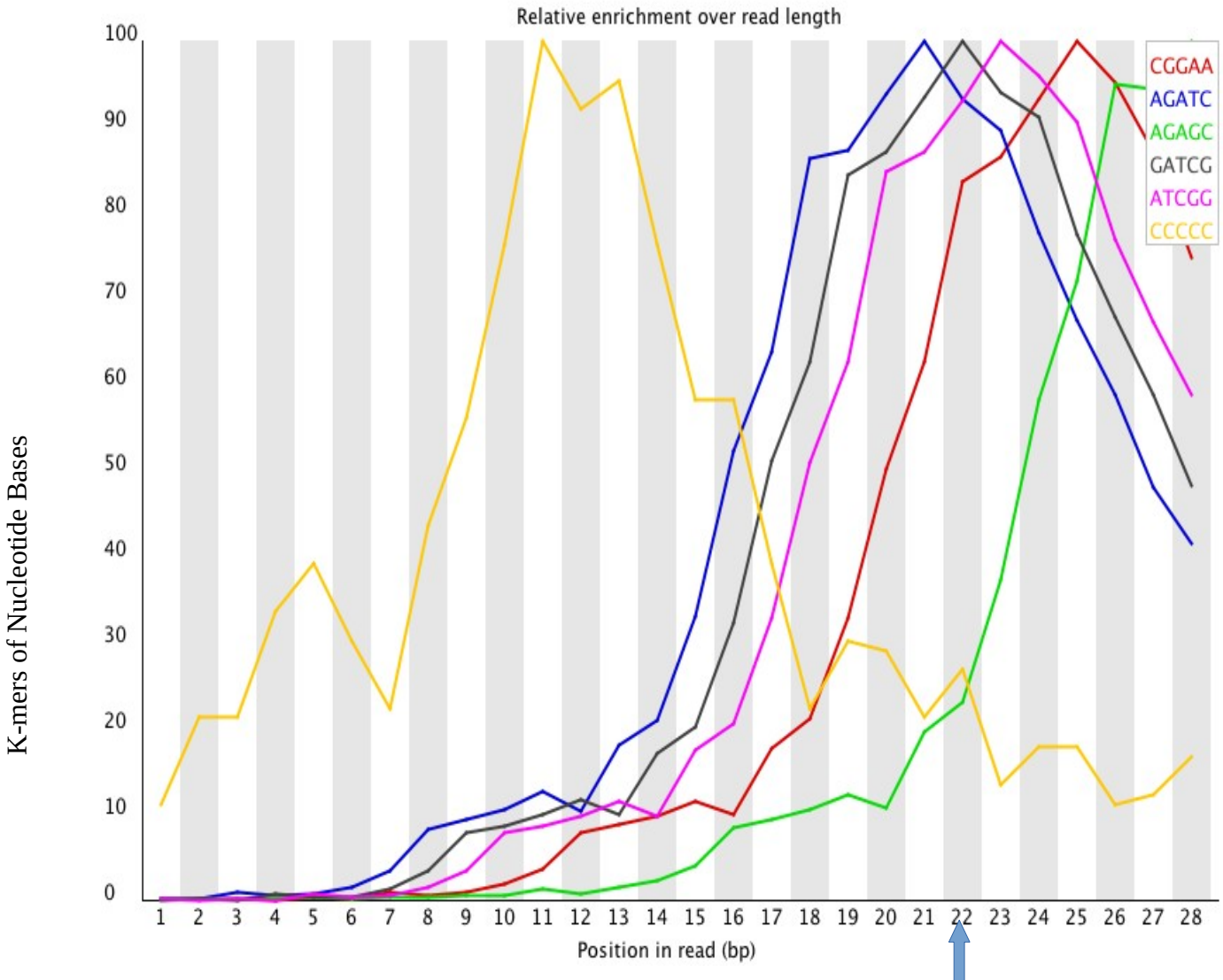
11. K-mer Content:

K-mers are sequence length subsequences. For example K-mers for GTAGAGCTGT

k	***k*mers**
1	G, T, A, G, A, G, C, T, G, T
2	GT, TA, AG, GA, AG, GC, CT, TG, GT
3	GTA, TAG, AGA, GAG, AGC, GCT, CTG, TGT
4	GTAG, TAGA, AGAG, GAGC, AGCT, GCTG, CTGT
5	GTAGA, TAGAG, AGAGC, GAGCT, AGCTG, GCTGT
6	GTAGAG, TAGAGC, AGAGCT, GAGCTG, AGCTGT
7	GTAGAGC, TAGAGCT, AGAGCTG, GAGCTGT
8	GTAGAGCT, TAGAGCTG, AGAGCTGT
9	GTAGAGCTG, TAGAGCTGT

(wikipedia)

K-mer content count each short nucleotide of length k starting at each position along the read. Any given K-mer should be evenly represented across the length of the read. Therefore, this module measures the number of each k-mer at each position in selected library and then uses a binomial test to look for significant deviations from an even coverage at all positions. Sequences longer than 500bp are truncated to 500bp for this analysis. A list of k-mers which appear at specific positions with greater than expected frequency are reported. The positions for the six most biased k-mers are plotted.



A plot of the relative enrichment of k-mers of five nucleotides in a single library from an example dataset.

Warning:

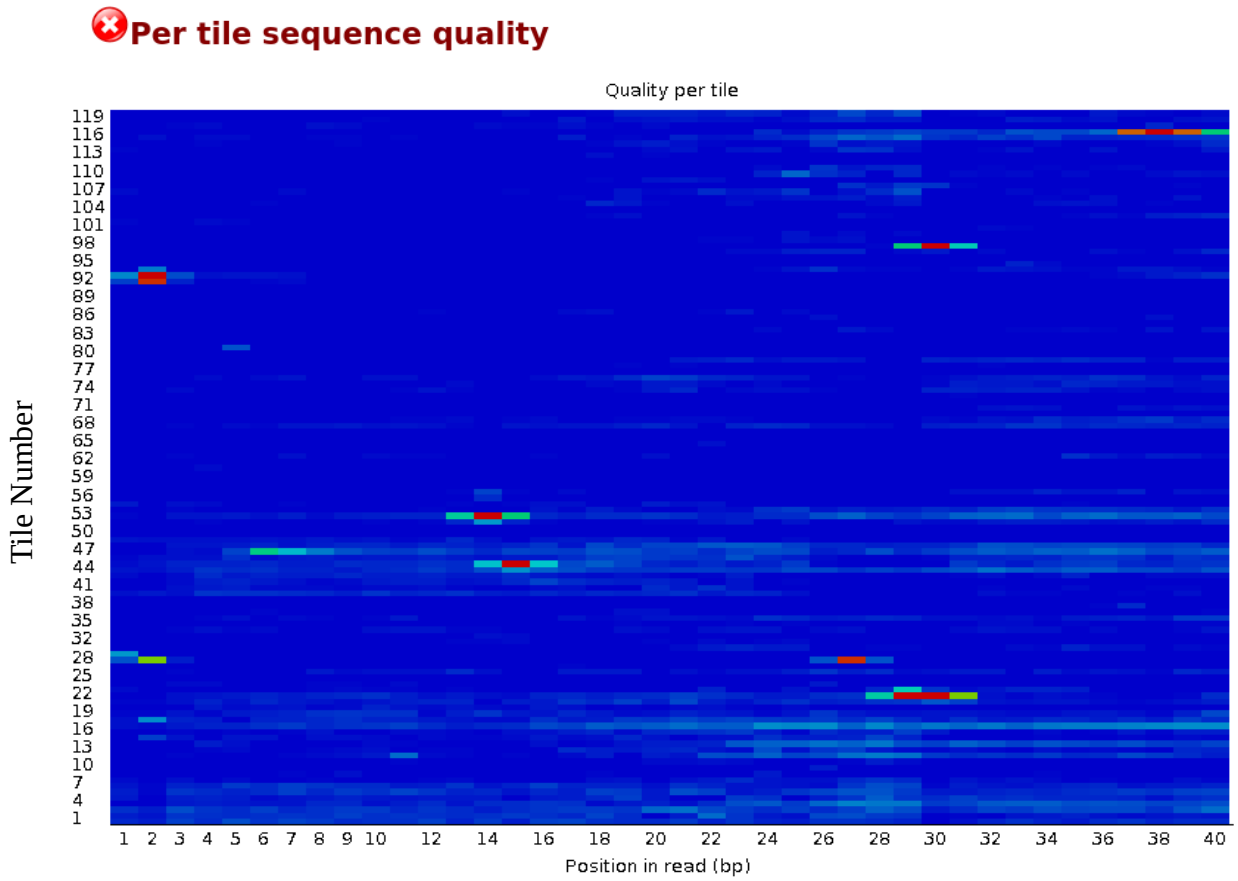
If any k-mer is imbalanced with a binomial p-value <0.01.

Failure:

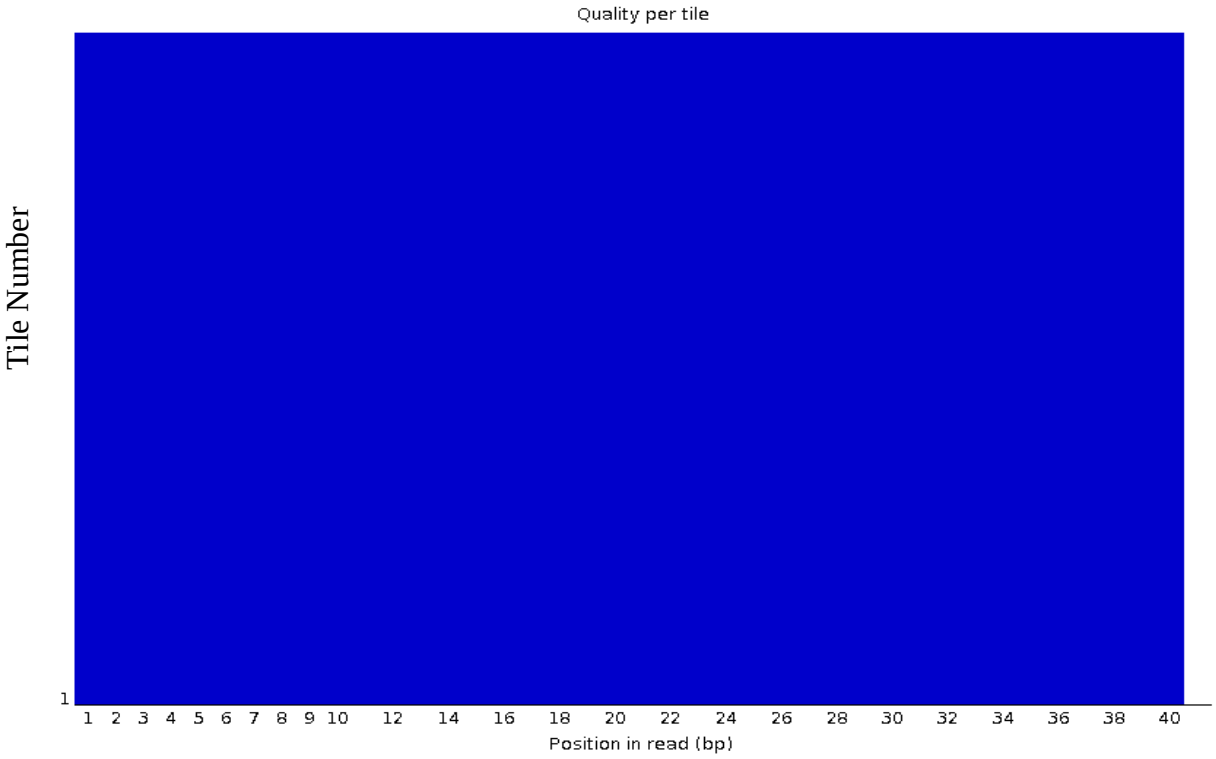
If any k-mer is imbalanced with a binomial p-value < 10^-5.

12. Per Tile Sequence Quality:

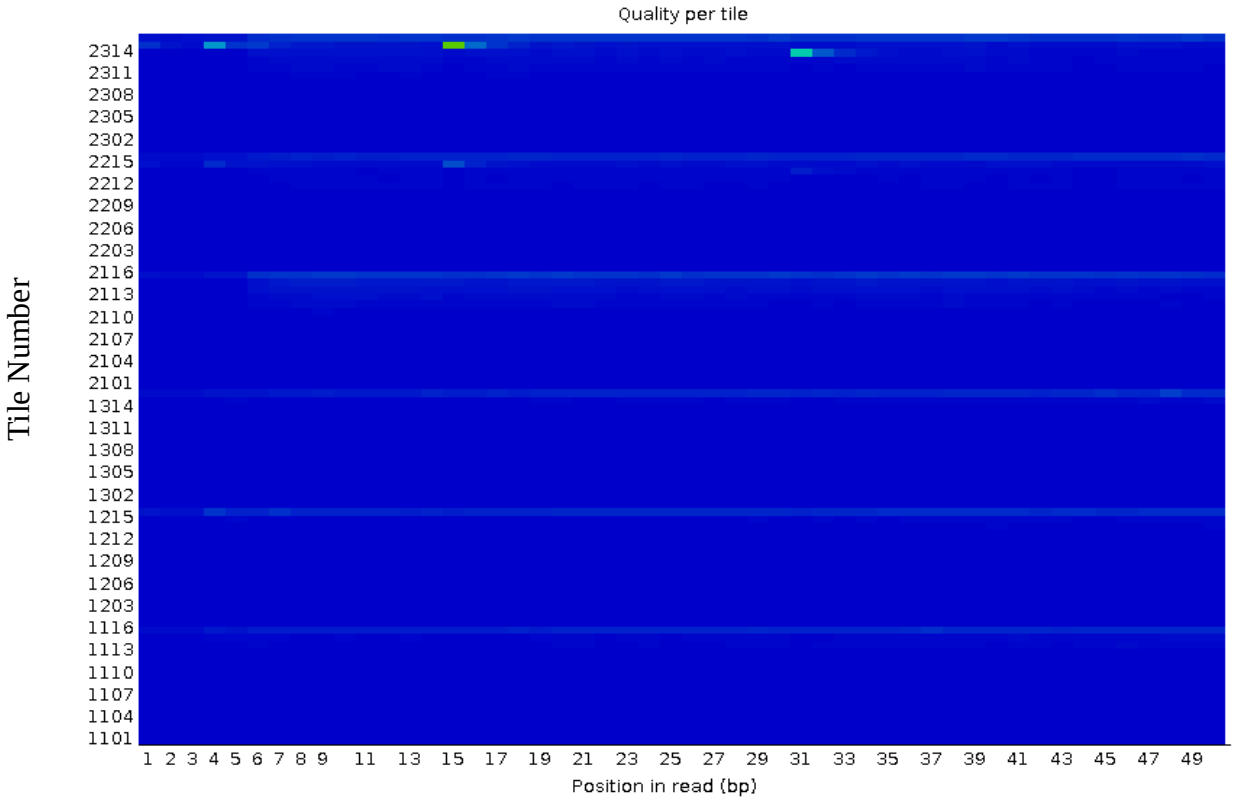
The graph allows to look at the average quality scores (mean Phred Scores) from each tile across all of the bases to see if there was a loss in quality associated with only one part of the flow cell. The flow cells on which the sequencing is performed are subdivided into "tiles". This graph will only appear in the analysis results if Illumina library is used. The plot shows the deviation from the average quality for each tile. The colors are on a cold to hot scale, with cold colors being positions where the quality was at or above the average for that base in the run, and hotter colors indicate that a tile had worse qualities than other tiles for that base. A good plot should be blue all over.



Per tile sequence quality



Per tile sequence quality



Warning (Orange Traingle):

If any tile shows a mean Phred score more than 2 less than the mean for that base across all tiles.

Failure (Red Cross):

If any tile shows a mean Phred score more than 5 less than the mean for that base across all tiles.