FastQC Report

- FastQC tool –Introduction

  Introduction:- Modern high throughput sequencers can generate tens of millions of sequences in a single run. Before analysing this sequence to draw biological conclusions we should always perform some simple quality control checks to ensure that the raw data looks good and there are no problems or biases in our data.

  Most sequencers will generate a QC report as part of their analysis pipeline, but this is usually only focused on identifying problems which were generated by the sequencer itself.

  Input :- FastQC supports files in the following formats FastQ (all quality encoding variants)

  - Casava FastQ files*
  - Colorspace FastQ
  - GZip compressed FastQ
  - SAM
  - BAM
  - SAM/BAM Mapped only (normally used for colorspace data)

  Evaluating results:- After several analysis modules analysis of FastQC file is generated. An interactive HTML output file is generated. In a sequence several parameters are applied on the given input sequence file.

  Output file content:-

  The below mentioned modules are given as output, each plays an important role for further processing of data in future.

- 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

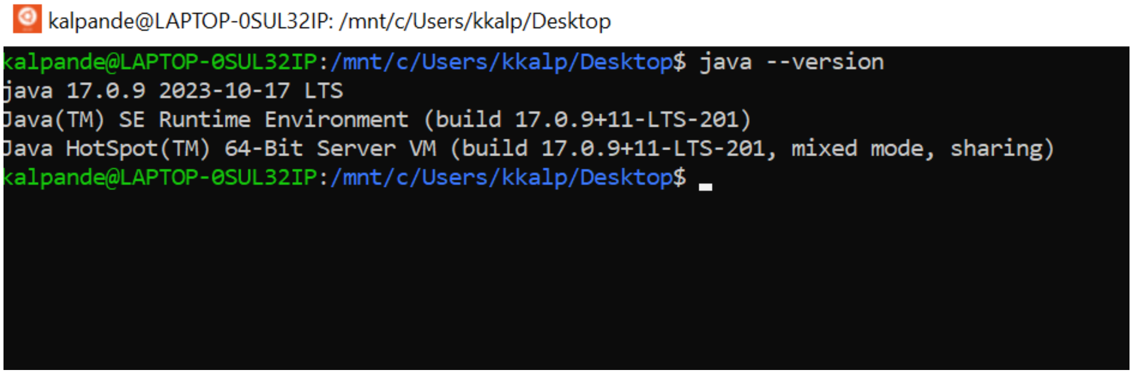
Purpose:- FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the source of organism.

Bash Script Assignment: FastQC Analysis for SRA Data

❖ Methodology

1) Installation of FastQC

- First check wether system having Java Jdk install as FastQC is a java based tool.



kalpande@LAPTOP-0SUL32IP: /mnt/c/Users/kkalp/Desktop

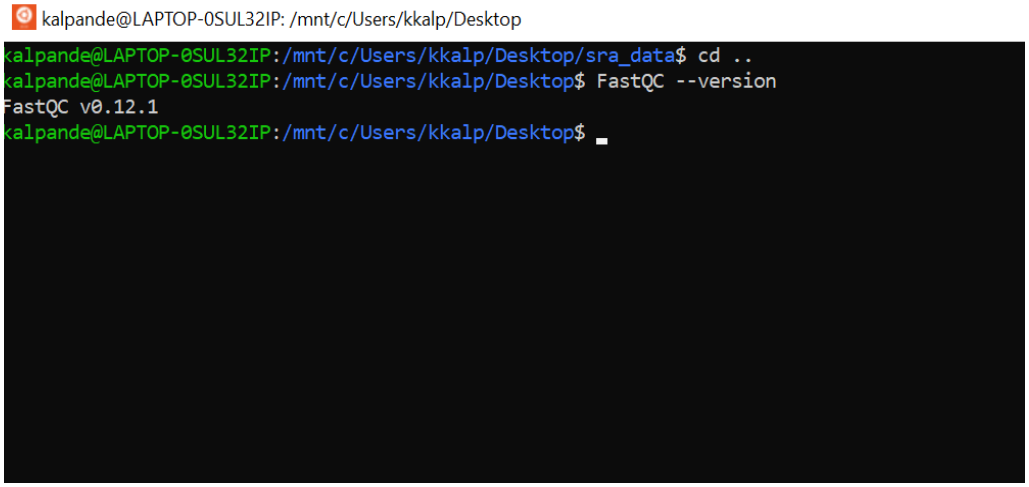
```
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$ java --version
java 17.0.9 2023-10-17 LTS
Java(TM) SE Runtime Environment (build 17.0.9+11-LTS-201)
Java HotSpot(TM) 64-Bit Server VM (build 17.0.9+11-LTS-201, mixed mode, sharing)
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$
```

- Download the FastQC tool from Babraham Bioinformatics site in zip format.
- Using ubuntu unzip it by following below command:

```
unzip FastQC
```

- Using below command check the version of FastQC to ensure proper installation of tool:

```
FastQC  --version
```



kalpande@LAPTOP-0SUL32IP: /mnt/c/Users/kkalp/Desktop

```
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ cd ..
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$ FastQC --version
FastQC v0.12.1
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$
```

2) Download FastQ Data from SRA [sratoolkit/ wget ftp link]

- We need to do first SRA toolkit installation
- Created a new directory in which sratoolkit is installed.

```
mkdir test
```

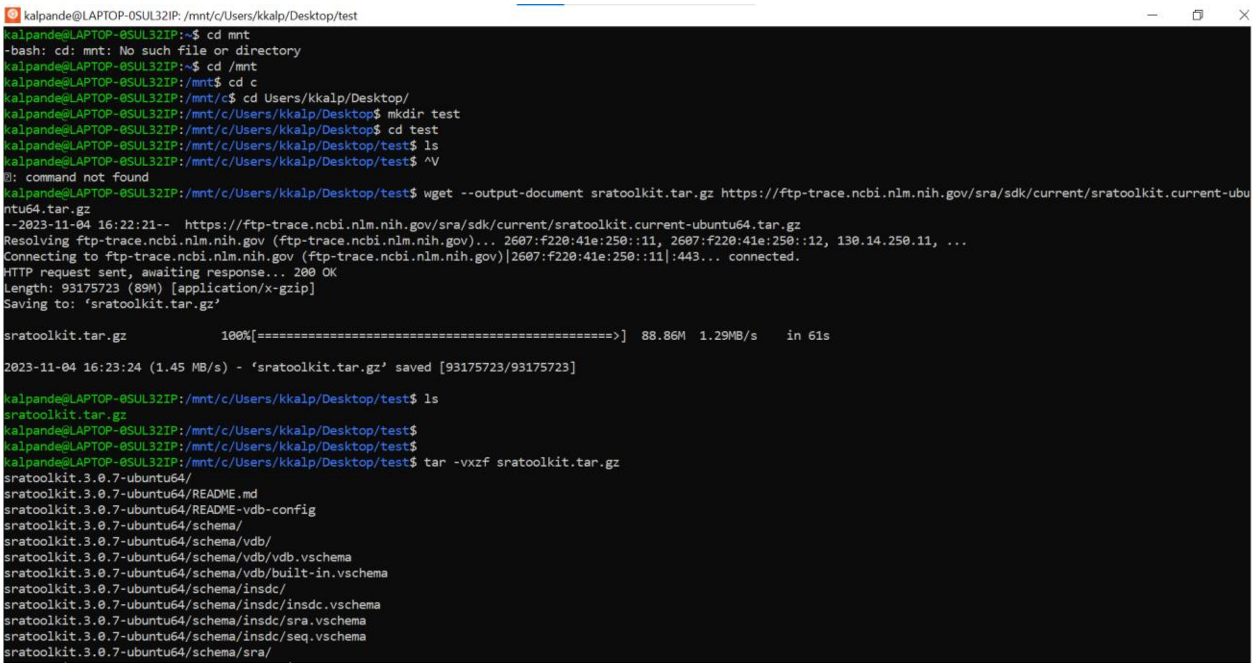
- From the below image we can see the steps:

- Using following command we install sratoolkit.

-
  ```
  wget --output-document sratoolkit.tar.gz https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz
  ```

- Then check for the file by ls
- Now extract the content of tar file by following command:

-
  ```
  tar -vxzf sratoolkit.tar.gz
  ```



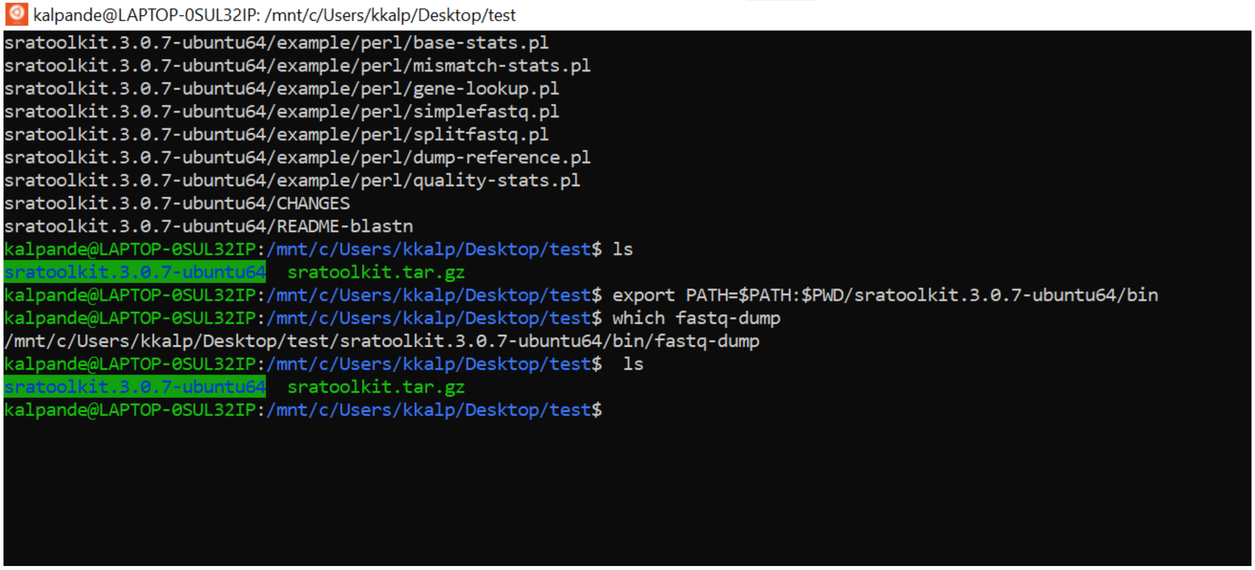
```
kalpande@LAPTOP-0SUL32IP:~$ cd mnt
-bash: cd: mnt: No such file or directory
kalpande@LAPTOP-0SUL32IP:~$ cd /mnt
kalpande@LAPTOP-0SUL32IP:/mnt$ cd c
kalpande@LAPTOP-0SUL32IP:/mnt/c$ cd Users/kkalp/Desktop/
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$ mkdir test
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$ cd test
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ ls
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ ^V
: command not found
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ wget --output-document sratoolkit.tar.gz https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubu
ntu64.tar.gz
--2023-11-04 16:22:21--  https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz
Resolving ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)... 2607:f220:41e:250::11, 2607:f220:41e:250::12, 130.14.250.11, ...
Connecting to ftp-trace.ncbi.nlm.nih.gov (ftp-trace.ncbi.nlm.nih.gov)|2607:f220:41e:250::11|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 93175723 (89M) [application/x-gzip]
Saving to: ‘sratoolkit.tar.gz’

sratoolkit.tar.gz            100%[==================================================>]  88.86M  1.29MB/s    in 61s

2023-11-04 16:23:24 (1.45 MB/s) - ‘sratoolkit.tar.gz’ saved [93175723/93175723]

kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ ls
sratoolkit.tar.gz
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ tar -vxzf sratoolkit.tar.gz
sratoolkit.3.0.7-ubuntu64/
sratoolkit.3.0.7-ubuntu64/README.md
sratoolkit.3.0.7-ubuntu64/README-vdb-config
sratoolkit.3.0.7-ubuntu64/schema/
sratoolkit.3.0.7-ubuntu64/schema/vdb/
sratoolkit.3.0.7-ubuntu64/schema/vdb/vdb.vschema
sratoolkit.3.0.7-ubuntu64/schema/vdb/built-in.vschema
sratoolkit.3.0.7-ubuntu64/schema/insdc/
sratoolkit.3.0.7-ubuntu64/schema/insdc/insdc.vschema
sratoolkit.3.0.7-ubuntu64/schema/insdc/sra.vschema
sratoolkit.3.0.7-ubuntu64/schema/insdc/seq.vschema
sratoolkit.3.0.7-ubuntu64/schema/sra/
```

- Check for the extracted files .
- Now set path of folder by following command:

```
export PATH=$PATH:$PWD/sratoolkit.3.0.7-ubuntu64/bin
```

- Then to again verify do

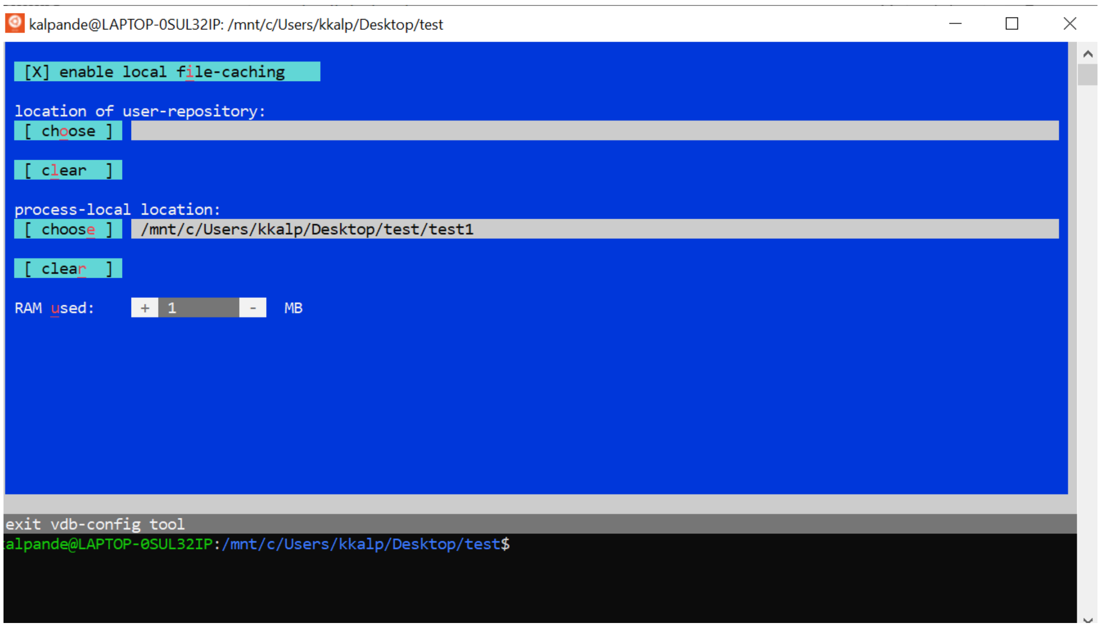
```
which fastq -dump
```



```
sratoolkit.3.0.7-ubuntu64/example/perl/base-stats.pl
sratoolkit.3.0.7-ubuntu64/example/perl/mismatch-stats.pl
sratoolkit.3.0.7-ubuntu64/example/perl/gene-lookup.pl
sratoolkit.3.0.7-ubuntu64/example/perl/simplefastq.pl
sratoolkit.3.0.7-ubuntu64/example/perl/splitfastq.pl
sratoolkit.3.0.7-ubuntu64/example/perl/dump-reference.pl
sratoolkit.3.0.7-ubuntu64/example/perl/quality-stats.pl
sratoolkit.3.0.7-ubuntu64/CHANGES
sratoolkit.3.0.7-ubuntu64/README-blastn
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ ls
sratoolkit.3.0.7-ubuntu64  sratoolkit.tar.gz
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ export PATH=$PATH:$PWD/sratoolkit.3.0.7-ubuntu64/bin
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ which fastq-dump
/mnt/c/Users/kkalp/Desktop/test/sratoolkit.3.0.7-ubuntu64/bin/fastq-dump
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$  ls
sratoolkit.3.0.7-ubuntu64  sratoolkit.tar.gz
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$
```

- Go for configuration using

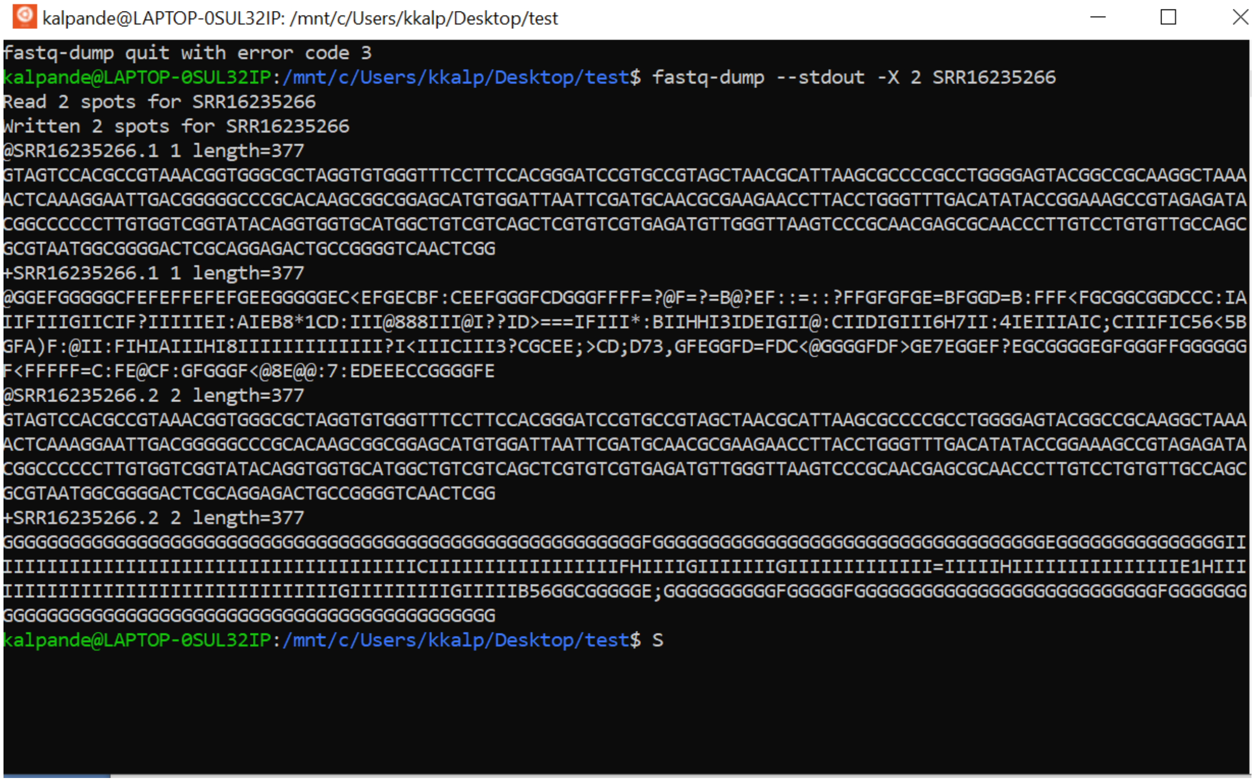
```
vdb-config -i
```



```
[X] enable local file-caching

location of user-repository:
[ choose ]

[ clear  ]

process-local location:
[ choose ]  /mnt/c/Users/kkalp/Desktop/test/test1

[ clear  ]

RAM used:    + 1        -    MB

exit vdb-config tool
alpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$
```

- To test wether tool is functional or not we use following command for our given id:

```
fastq-dump --stdout -X 2 SRR162352266
```

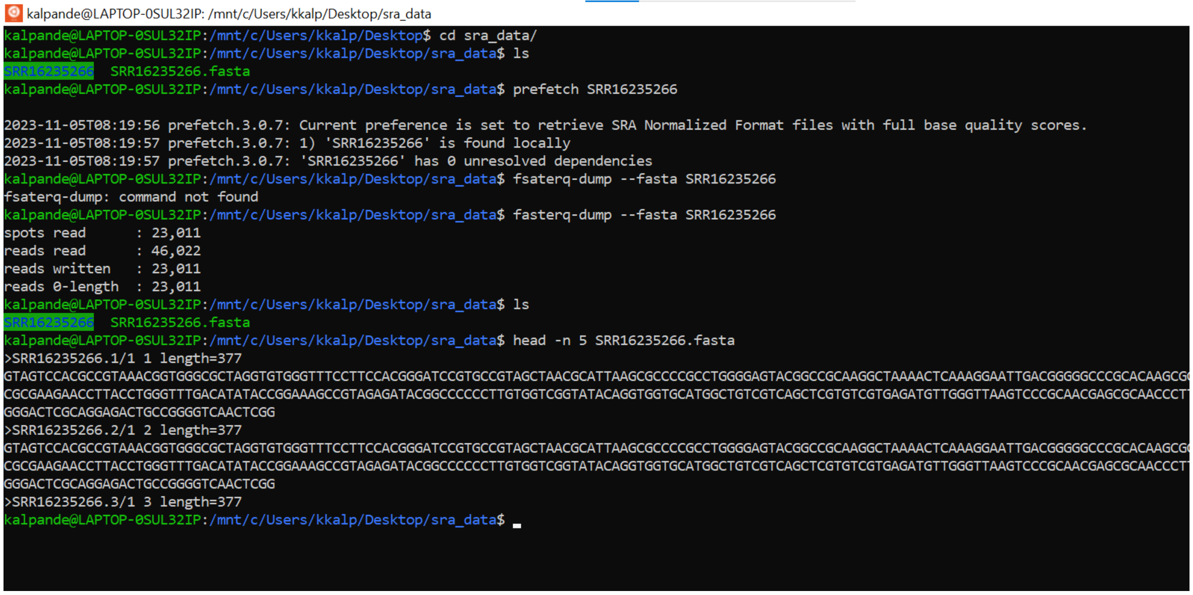


```
fastq-dump quit with error code 3
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ fastq-dump --stdout -X 2 SRR16235266
Read 2 spots for SRR16235266
Written 2 spots for SRR16235266
@SRR16235266.1 1 length=377
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGGGATCCGTGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAA
ACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGCGGAGCATGTGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGTTTGACATATACCGGAAAGCCGTAGAGATA
CGGCCCCCCTTGTGGTCGGTATACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTCCTGTGTTGCCAGC
GCGTAATGGCGGGGACTCGCAGGAGACTGCCGGGGTCAACTCGG
+SRR16235266.1 1 length=377
@GGEFGGGGGCFEFEFFEFEFGEEGGGGGEC<EFGECBF:CEEFGGGFCDGGGFFFF=?@F=?=B@?EF::=::?FFGFGFGE=BFGGD=B:FFF<FGCGGCGGDCCC:IA
IIFIIIGIICIF?IIIIIEI:AIEB8*1CD:III@888III@I??ID>===IFIII*:BIIHHI3IDEIGII@:CIIDIGIII6H7II:4IEIIIAIC;CIIIFIC56<5B
GFA)F:@II:FIHIAIIIHI8IIIIIIIIIIII?I<IIICIII3?CGCEE;>CD;D73,GFEGGFD=FDC<@GGGGFDF>GE7EGGEF?EGCGGGGEGFGGGFFGGGGGG
F<FFFFF=C:FE@CF:GFGGGF<@8E@@:7:EDEEECCGGGGFE
@SRR16235266.2 2 length=377
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGGGATCCGTGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAA
ACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGCGGAGCATGTGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGTTTGACATATACCGGAAAGCCGTAGAGATA
CGGCCCCCCTTGTGGTCGGTATACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTCCTGTGTTGCCAGC
GCGTAATGGCGGGGACTCGCAGGAGACTGCCGGGGTCAACTCGG
+SRR16235266.2 2 length=377
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGGGGGGGGGGGGGII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIICIIIIIIIIIIIIIIIIIIFHIIIIGIIIIIIIGIIIIIIIIIIIII=IIIIIHIIIIIIIIIIIIIIIIIIE1HIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIGIIIIIB56GGCGGGGGE;GGGGGGGGGGFGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ S
```

- After this we will do fasterq-dump to extract sequence data from sra and converting it into FastQ format.

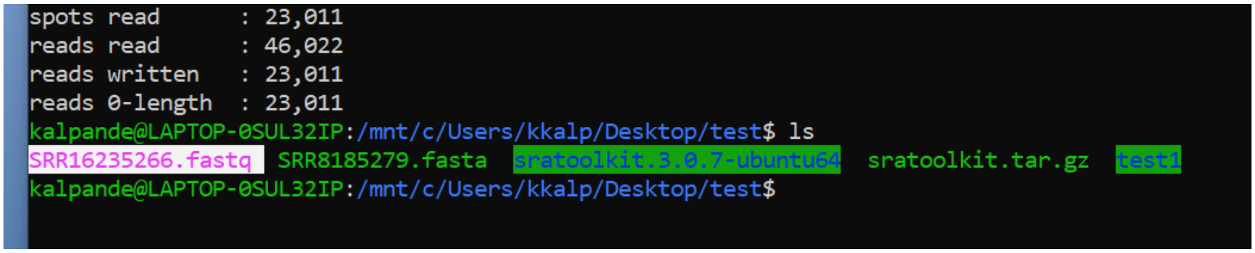
```
fasterq-dump SRR16235266
```

- Below image shows the result, in addition I have done head for first lines.

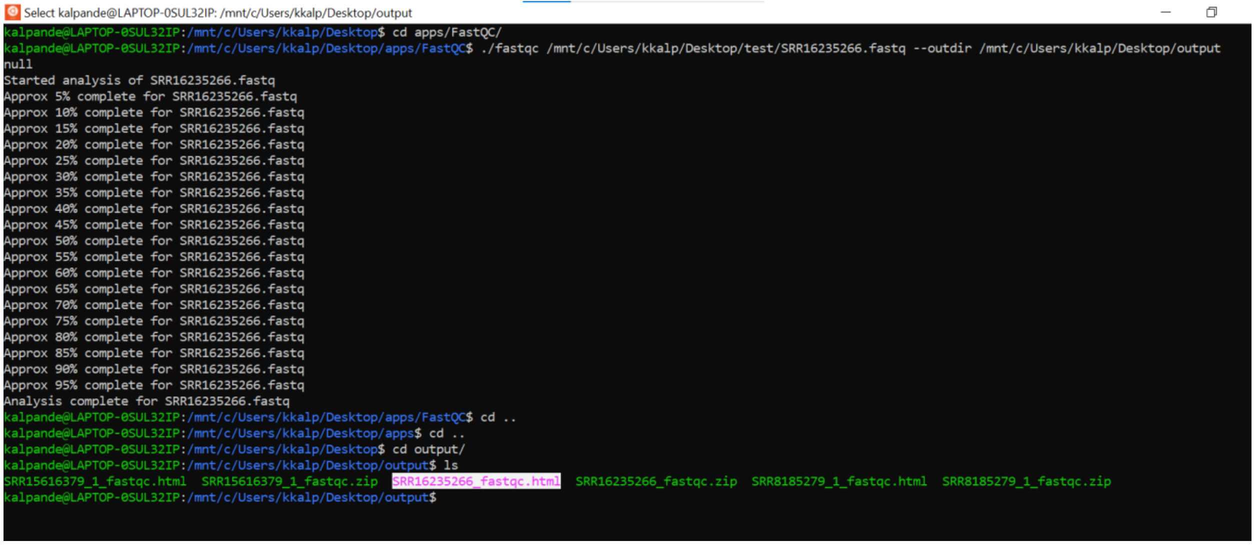


kalpande@LAPTOP-0SUL32IP: /mnt/c/Users/kkalp/Desktop/sra_data

```
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$ cd sra_data/
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ ls
SRR16235266  SRR16235266.fasta
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ prefetch SRR16235266

2023-11-05T08:19:56 prefetch.3.0.7: Current preference is set to retrieve SRA Normalized Format files with full base quality scores.
2023-11-05T08:19:57 prefetch.3.0.7: 1) 'SRR16235266' is found locally
2023-11-05T08:19:57 prefetch.3.0.7: 'SRR16235266' has 0 unresolved dependencies
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ fsaterq-dump --fasta SRR16235266
fsaterq-dump: command not found
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ fasterq-dump --fasta SRR16235266
spots read      : 23,011
reads read      : 46,022
reads written   : 23,011
reads 0-length  : 23,011
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ ls
SRR16235266  SRR16235266.fasta
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ head -n 5 SRR16235266.fasta
>SRR16235266.1/1 1 length=377
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGGGATCCGTGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCG
CGCGAAGAACCTTACCTGGGTTTGACATATACCGGAAAGCCGTAGAGATACGGCCCCCCTTGTGGTCGGTATACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCT
GGGACTCGCAGGAGACTGCCGGGGTCAACTCGG
>SRR16235266.2/1 2 length=377
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGGGATCCGTGCCGTAGCTAACGCATTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCG
CGCGAAGAACCTTACCTGGGTTTGACATATACCGGAAAGCCGTAGAGATACGGCCCCCCTTGTGGTCGGTATACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCT
GGGACTCGCAGGAGACTGCCGGGGTCAACTCGG
>SRR16235266.3/1 3 length=377
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/sra_data$ _
```



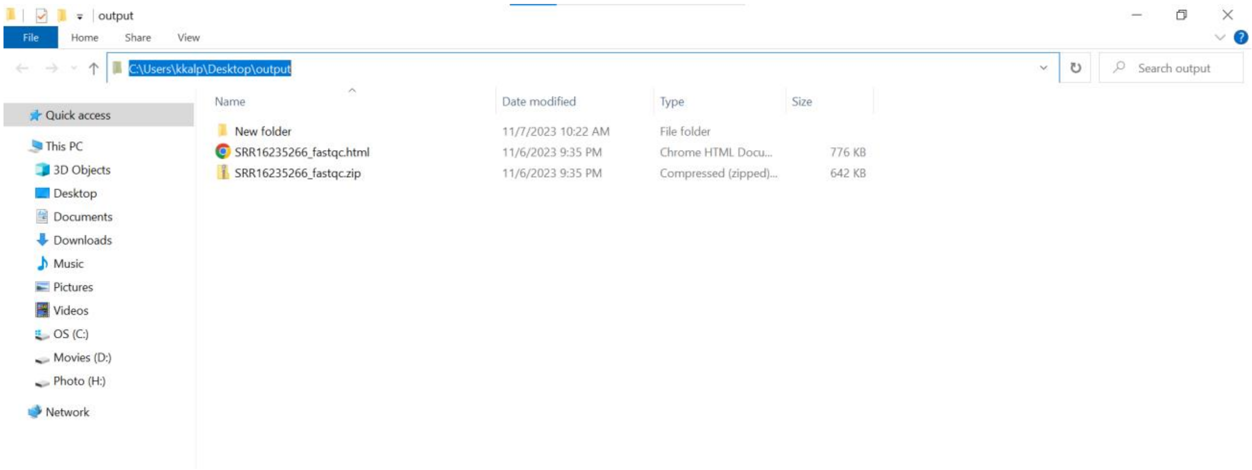
```
spots read      : 23,011
reads read      : 46,022
reads written   : 23,011
reads 0-length  : 23,011
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$ ls
SRR16235266.fastq  SRR8185279.fasta  sratoolkit.3.0.7-ubuntu64  sratoolkit.tar.gz  test1
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/test$
```



Select kalpande@LAPTOP-0SUL32IP: /mnt/c/Users/kkalp/Desktop/output

```
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$ cd apps/FastQC/
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/apps/FastQC$ ./fastqc /mnt/c/Users/kkalp/Desktop/test/SRR16235266.fastq --outdir /mnt/c/Users/kkalp/Desktop/output
null
Started analysis of SRR16235266.fastq
Approx 5% complete for SRR16235266.fastq
Approx 10% complete for SRR16235266.fastq
Approx 15% complete for SRR16235266.fastq
Approx 20% complete for SRR16235266.fastq
Approx 25% complete for SRR16235266.fastq
Approx 30% complete for SRR16235266.fastq
Approx 35% complete for SRR16235266.fastq
Approx 40% complete for SRR16235266.fastq
Approx 45% complete for SRR16235266.fastq
Approx 50% complete for SRR16235266.fastq
Approx 55% complete for SRR16235266.fastq
Approx 60% complete for SRR16235266.fastq
Approx 65% complete for SRR16235266.fastq
Approx 70% complete for SRR16235266.fastq
Approx 75% complete for SRR16235266.fastq
Approx 80% complete for SRR16235266.fastq
Approx 85% complete for SRR16235266.fastq
Approx 90% complete for SRR16235266.fastq
Approx 95% complete for SRR16235266.fastq
Analysis complete for SRR16235266.fastq
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/apps/FastQC$ cd ..
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/apps$ cd ..
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop$ cd output/
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/output$ ls
SRR15616379_1_fastqc.html  SRR15616379_1_fastqc.zip  SRR16235266_fastqc.html  SRR16235266_fastqc.zip  SRR8185279_1_fastqc.html  SRR8185279_1_fastqc.zip
kalpande@LAPTOP-0SUL32IP:/mnt/c/Users/kkalp/Desktop/output$
```

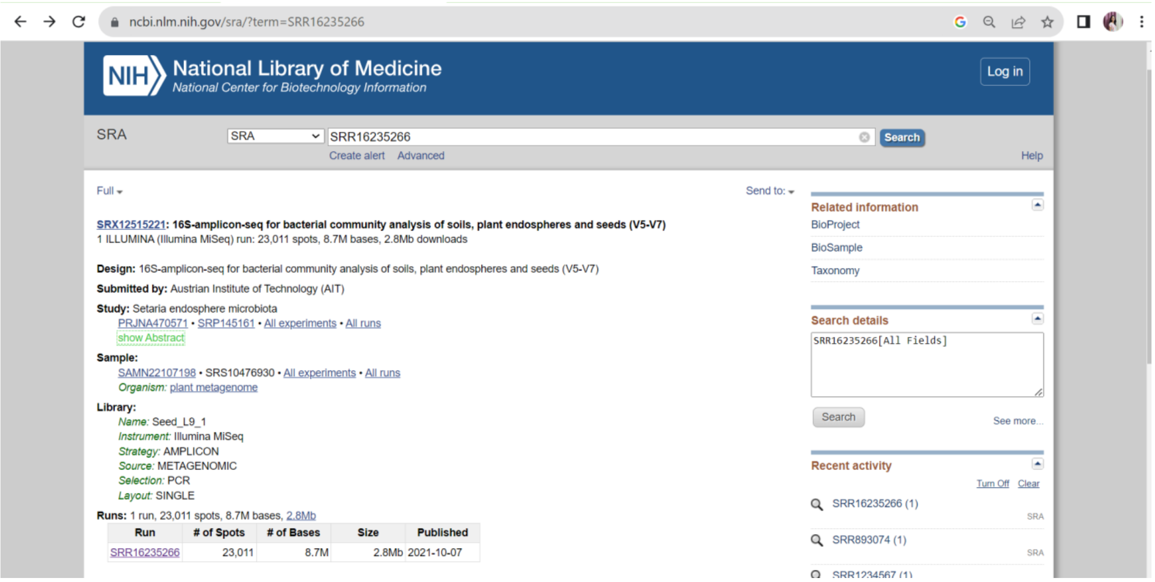
.html file is created



C:\Users\kkalp\Desktop\output

Name	Date modified	Type	Size
New folder	11/7/2023 10:22 AM	File folder	
SRR16235266_fastqc.html	11/6/2023 9:35 PM	Chrome HTML Docu...	776 KB
SRR16235266_fastqc.zip	11/6/2023 9:35 PM	Compressed (zipped)...	642 KB

Below is the detail regarding used Accession Id-



SRR16235266

SRX12515221: 16S-amplicon-seq for bacterial community analysis of soils, plant endospheres and seeds (V5-V7)
1 ILLUMINA (Illumina MiSeq) run: 23,011 spots, 8.7M bases, 2.8Mb downloads

Design: 16S-amplicon-seq for bacterial community analysis of soils, plant endospheres and seeds (V5-V7)

Submitted by: Austrian Institute of Technology (AIT)

Study: Setaria endosphere microbiota
PRJNA470571 • SRP145161 • All experiments • All runs
show Abstract

Sample:
SAMN22107198 • SRS10476930 • All experiments • All runs
Organism: plant metagenome

Library:
Name: Seed_L9_1
Instrument: Illumina MiSeq
Strategy: AMPLICON
Source: METAGENOMIC
Selection: PCR
Layout: SINGLE

Runs: 1 run, 23,011 spots, 8.7M bases, 2.8Mb

Run	# of Spots	# of Bases	Size	Published
SRR16235266	23,011	8.7M	2.8Mb	2021-10-07

SRR16235266[All Fields]

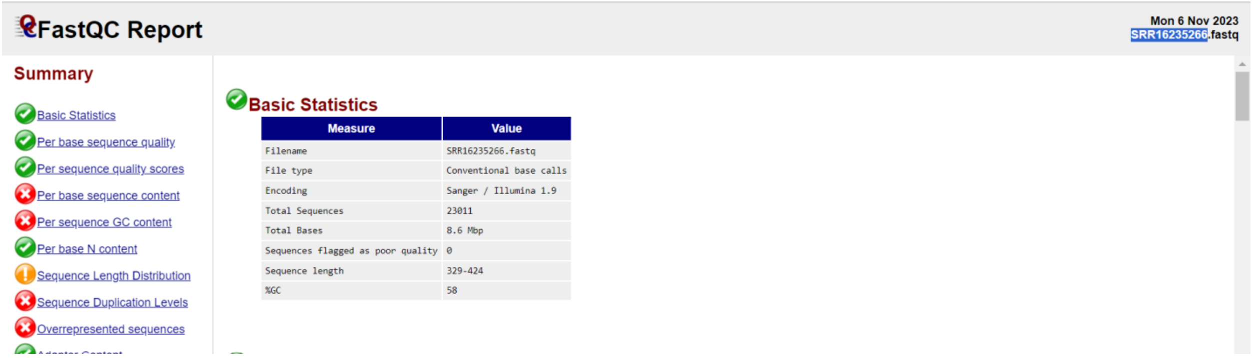
- ❖ Results

3) Run FastQC Analysis

A .html file is generated for the given fastq file.
Several modules are checked and an interactive graphical file id generated.
Link of output file is: click

- The green tick represents that given sequence is within the expected or acceptable range, suggesting good data quality.
- The orange tick is used to highlight areas of moderate concern. It suggests that there may be some issues or deviations from ideal quality, but they are not severe enough to be an immediate cause for alarm.
- The red tick is used to flag critical issues or areas of significant concern. If a metric or result is marked in red, it indicates a potential problem that may require attention or further investigation.

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



FastQC Report

Mon 6 Nov 2023
SRR16235266.fastq

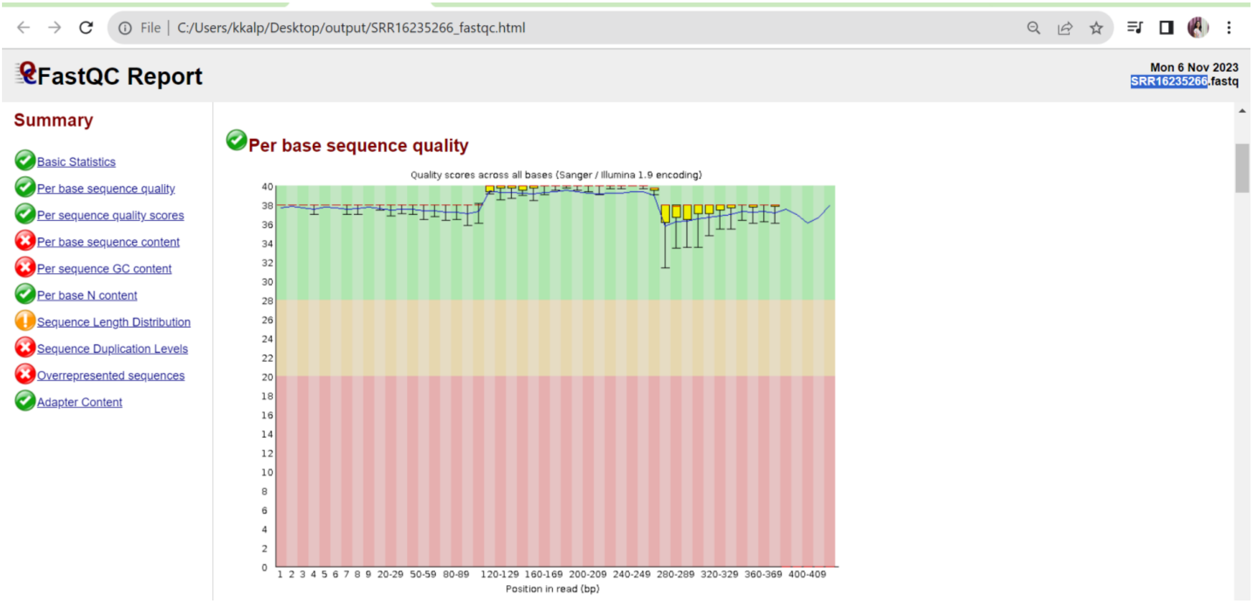
Summary

- ✓ 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

✓ Basic Statistics

Measure	Value
Filename	SRR16235266.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	23011
Total Bases	8.6 Mbp
Sequences flagged as poor quality	0
Sequence length	329-424
%GC	58

ii. **Per base sequence quality**- This view shows an overview of the range of quality values across all bases at each position in the FastQ file.The good quality reads are checked by Phred parameter whose **good value should be above 28.** Whereas **poor quality ranges below 20**.



File | C:/Users/kkalp/Desktop/output/SRR16235266_fastqc.html

FastQC Report

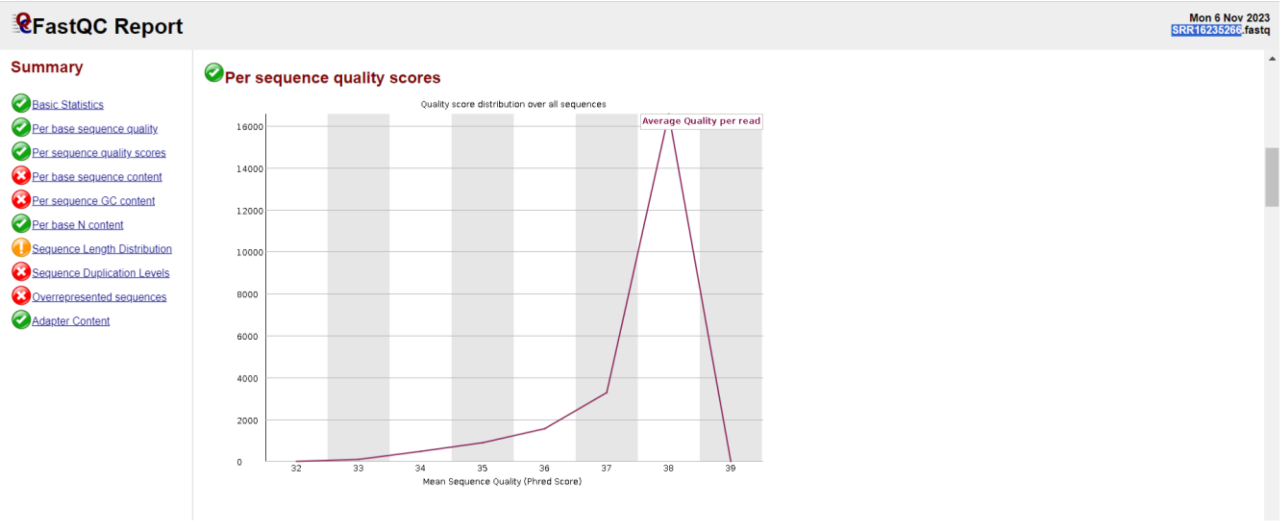
Mon 6 Nov 2023
SRR16235266.fastq

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

✓ Per base sequence quality

iii. **Per sequence quality scores**-The per sequence quality score report allows you to see if a **subset of your sequences have universally low quality values**. It is often the case that a subset of sequences will have universally poor quality, often because they are poorly imaged , however these should represent only a small percentage of the total sequences.



FastQC Report

Mon 6 Nov 2023
SRR16235266.fastq

Summary

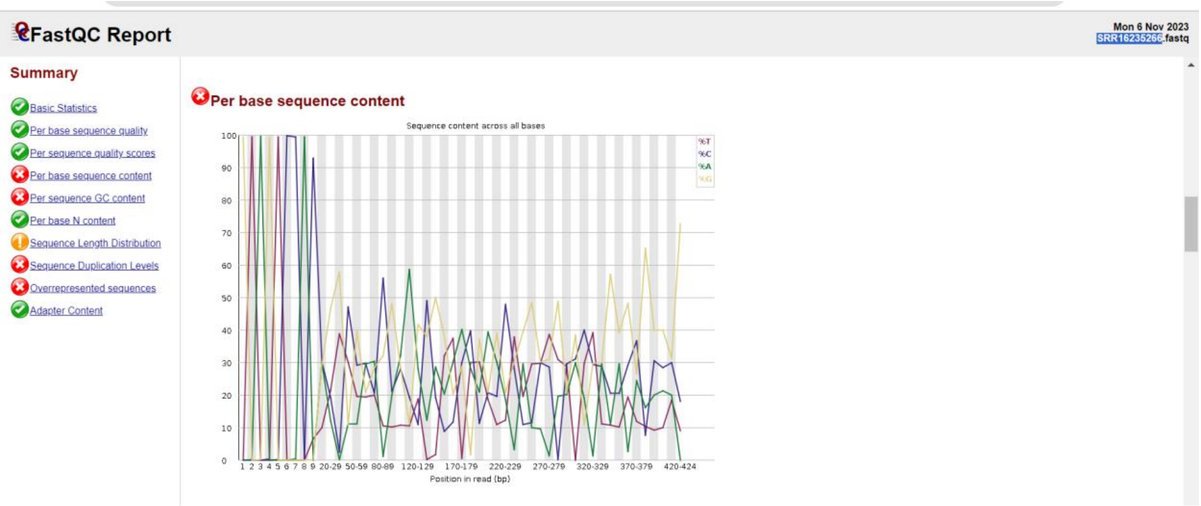
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter Content

✓ Per sequence quality scores

Quality score distribution over all sequences
Average Quality per read
16000
14000
12000
10000
8000
6000
4000
2000
0
32
33
34
35
36
37
38
39
Mean Sequence Quality (Phred Score)

iv. **Per base sequence content** -Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called. In a random library you would **expect that there would be little to 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 should reflect the overall amount of these bases in your genome, but in any case they should not be hugely imbalanced from each other.

In our case, the bases are not in parallel with each other.



FastQC Report

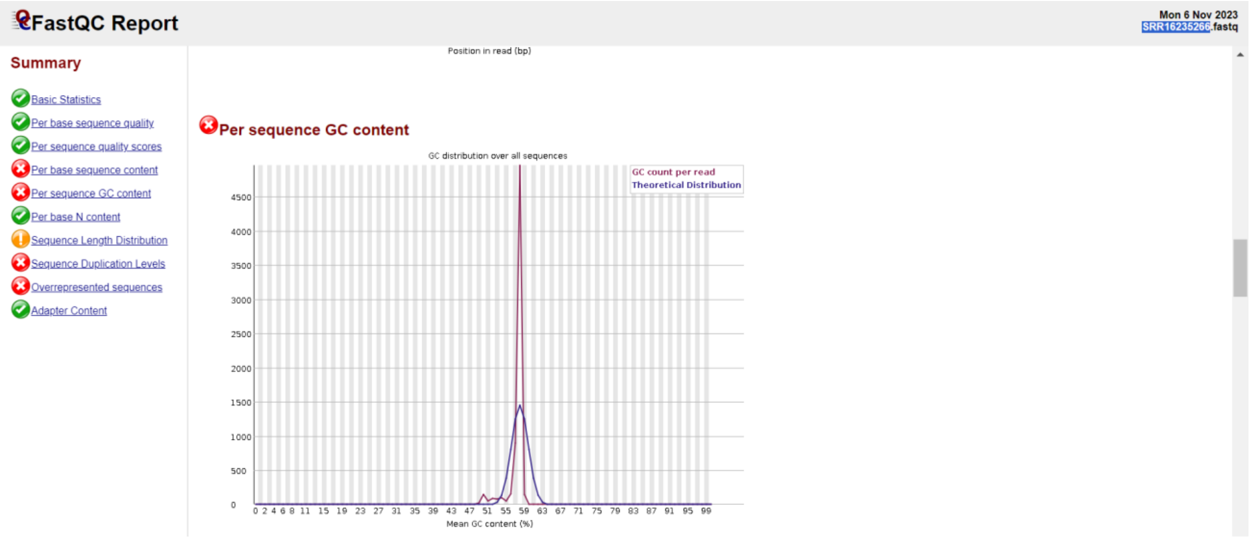
Mon 6 Nov 2023
SRR16235266.fastq

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

✗ Per base sequence content

v. **Per sequence GC content-**This module measures the GC content across the whole length of each sequence in a file and compares it to a modelled normal distribution of GC content.In a normal random library you would expect to see a roughly normal distribution of GC content **where the central peak corresponds to the overall GC content of the underlying genome**. Since we don't know the the GC content of the genome the modal GC content is calculated from the observed data and used to build a reference distribution.



FastQC Report

Mon 6 Nov 2023
SRR16235266.fastq

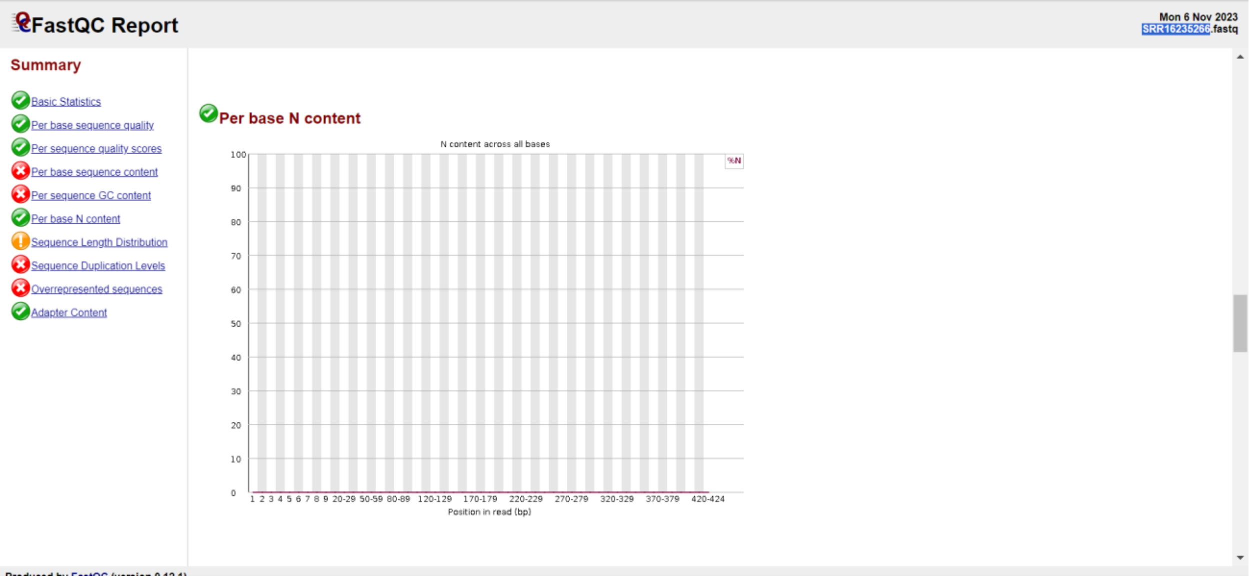
Summary

- ✓ 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

Position in read (bp)

✗ Per sequence GC content

vi. **Per base N content**- If a sequencer is unable to make a base call with sufficient confidence then **it will normally substitute an N rather than a conventional base call.** This module plots out the percentage of base calls at each position for which an N was called.



FastQC Report

Mon 6 Nov 2023
SRR16235266.fastq

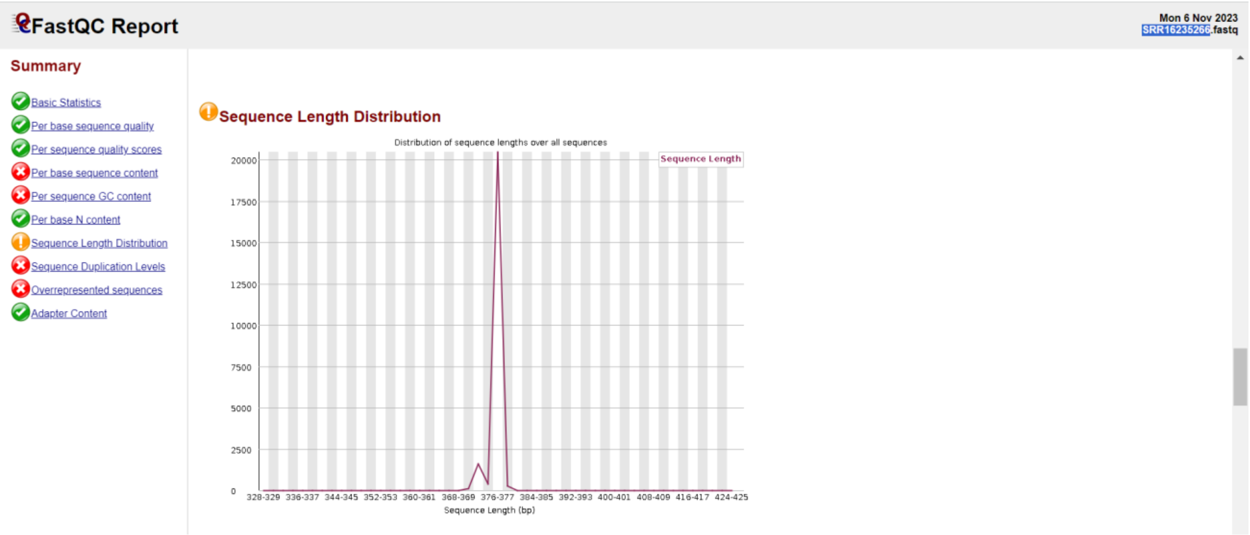
Summary

- ✓ 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

✓ Per base N content

Produced by FastQC (version 0.12.1)

vii. **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.



FastQC Report

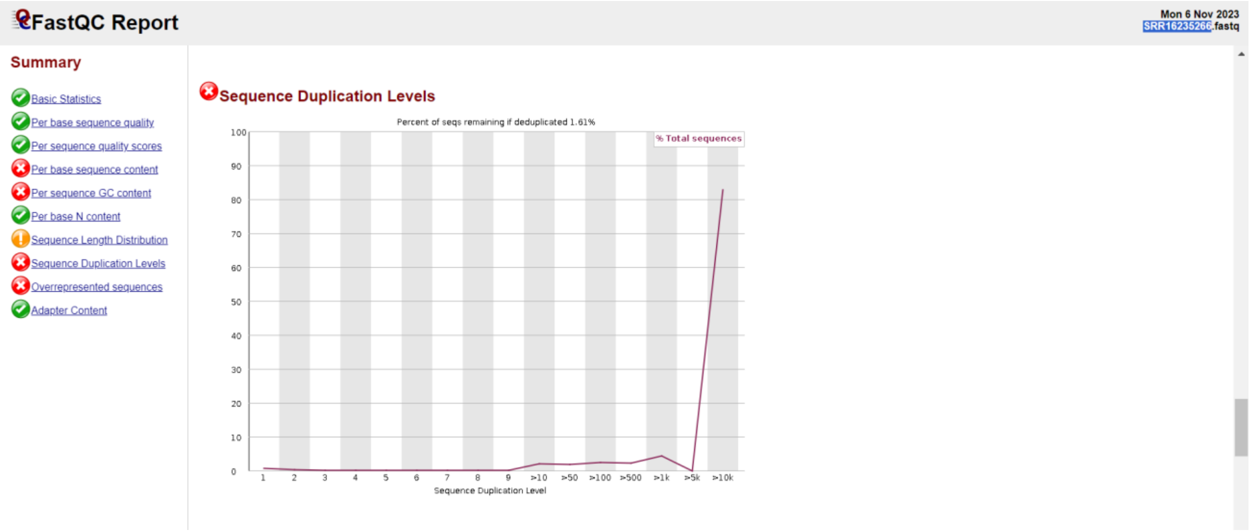
Mon 6 Nov 2023
SRR16235266.fastq

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

! Sequence Length Distribution

viii. **Sequence Duplication Levels**-In a diverse library most sequences will occur only once in the final set. 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 some kind of enrichment bias. This module **counts the degree of duplication for every sequence in the set** and creates a plot showing the relative number of sequences with different degrees of duplication.



FastQC Report

Mon 6 Nov 2023
SRR16235266.fastq

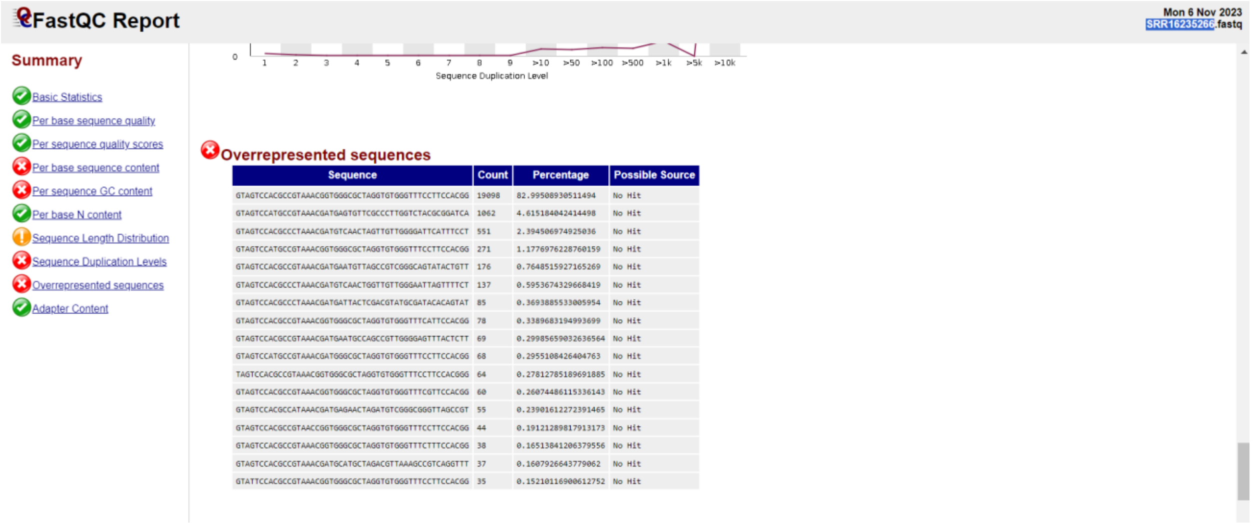
Summary

- ✓ 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

☒ Sequence Duplication Levels

Percent of seqs remaining if deduplicated 1.61%
% Total sequences
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k
Sequence Duplication Level

ix. **Overrepresented sequences**-A normal high-throughput library will contain a diverse set of sequences, with no individual sequence making up a tiny fraction of the whole. Finding that a **single sequence is very overrepresented in the set either means that it is highly biologically significant, or indicates that the library is contaminated, or not as diverse as you expected.**



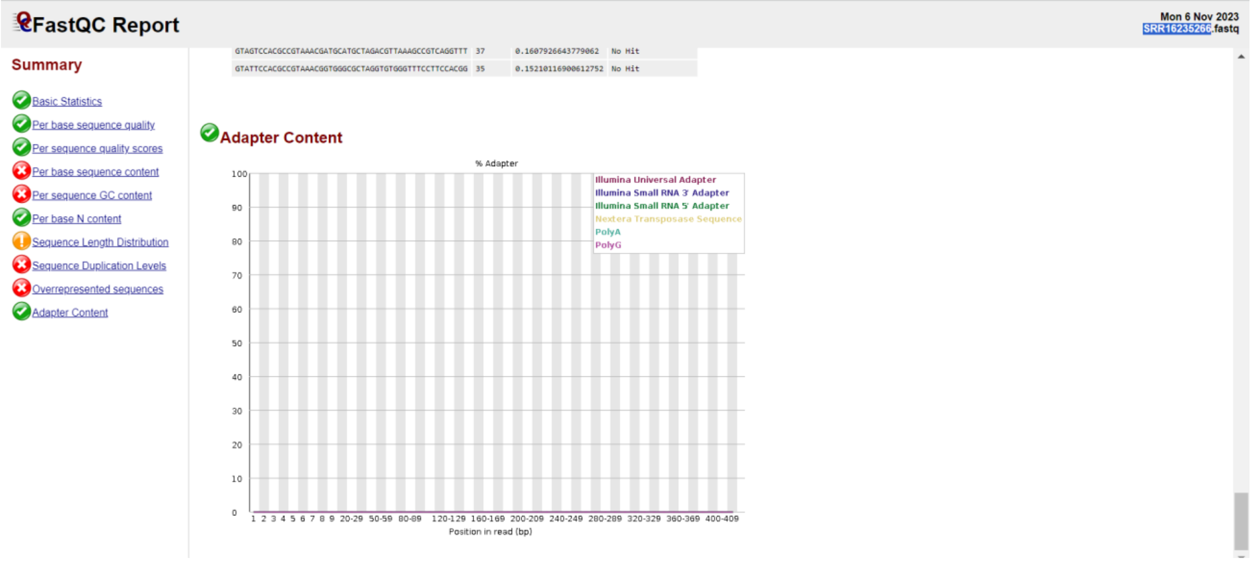
Summary

- ✓ 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

✗ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGG	19098	82.99508930511494	No Hit
GTAGTCCATGCCGTAAACGATGAGTGTTCGCCCTTGGTCTACGCGGATCA	1062	4.615184042414498	No Hit
GTAGTCCACGCCCTAAACGATGTCAACTAGTTGTTGGGGATTCATTTCCT	551	2.394506974925036	No Hit
GTAGTCCATGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGG	271	1.1776976228760159	No Hit
GTAGTCCACGCCGTAAACGATGAATGTTAGCCGTCGGGCAGTATACTGTT	176	0.7648515927165269	No Hit
GTAGTCCACGCCCTAAACGATGTCAACTGGTTGTTGGGAATTAGTTTTCT	137	0.5953674329668419	No Hit
GTAGTCCACGCCCTAAACGATGATTACTCGACGTATGCGATACACAGTAT	85	0.3693885533005954	No Hit
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCATTCCACGG	78	0.3389683194993699	No Hit
GTAGTCCACGCCGTAAACGATGAATGCCAGCCGTTGGGGAGTTTACTCTT	69	0.29985659032636564	No Hit
GTAGTCCATGCCGTAAACGATGGGCGCTAGGTGTGGGTTTCCTTCCACGG	68	0.2955108426404763	No Hit
TAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGGG	64	0.27812785189691885	No Hit
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCGTTCCACGG	60	0.26074486115336143	No Hit
GTAGTCCACGCCATAAACGATGAGAACTAGATGTCGGGCGGGTTAGCCGT	55	0.23901612272391465	No Hit
GTAGTCCACGCCGTAACCGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGG	44	0.19121289817913173	No Hit
GTAGTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCTTTCCACGG	38	0.16513841206379556	No Hit
GTAGTCCACGCCGTAAACGATGCATGCTAGACGTTAAAGCCGTCAGGTTT	37	0.1607926643779062	No Hit
GTATTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGG	35	0.15210116900612752	No Hit

x. **Adapter Content**- Adapters are specific module or section that assesses the presence of adapter sequences in sequencing data.The adapter content analysis in FASTQC checks for the **presence of these adapter sequences in your sequencing data and evaluates how much of your data might be affected by adapter contamination**. This information is important for quality control and preprocessing steps in NGS data analysis, as adapter contamination can lead to issues such as reduced mapping efficiency, erroneous variant calls.



GTAGTCCACGCCGTAAACGATGCATGCTAGACGTTAAAGCCGTCAGGTTT	37	0.1607926643779062	No Hit
GTATTCCACGCCGTAAACGGTGGGCGCTAGGTGTGGGTTTCCTTCCACGG	35	0.15210116900612752	No Hit

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

✓ Adapter Content