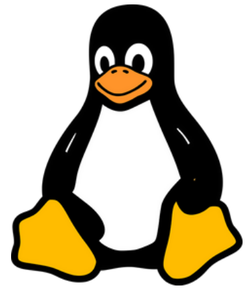
BIGA BIOINFORMATICS FOR GENOME ANALYSIS LABORATORY

BASIC LINUX FOR GENOME ANALYSIS



VEGA: Game Changer For Clinical Genomics

15 May 2025

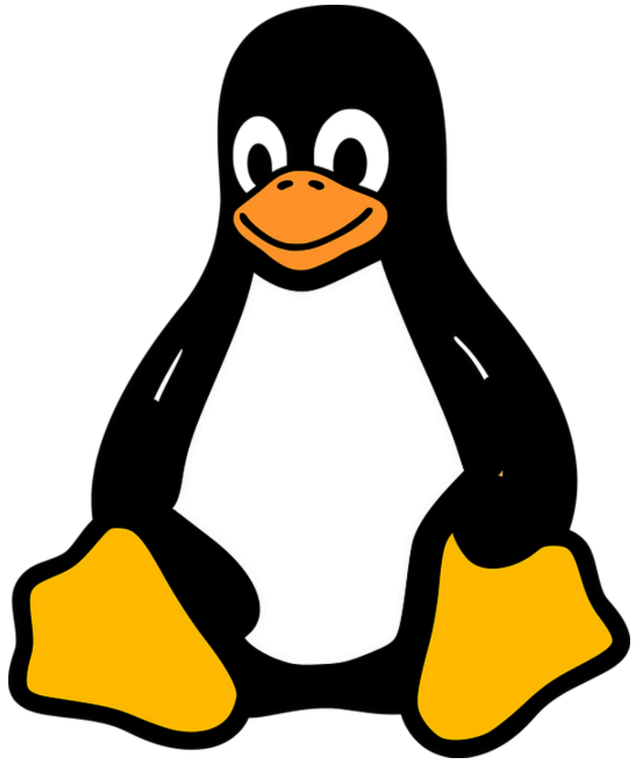
Disclaimer

- The sequencing data were provided by the Bioinformatics for Genome Analysis Laboratory.
- Research used only

What is a LINUX?

Linux is an operating system (OS) like Windows or macOS.

An OS is the software that lets you interact with your computer hardware (CPU, memory, storage, etc.) and run applications



How Linux Relates to Bioinformaticians?

- Most bioinformatics tools (e.g., **BWA, SAMtools, STAR, GATK**) are Linux-based.
- Genomics, transcriptomics, and proteomics pipelines typically run on Linux servers or HPC clusters.
- Linux efficiently handles large genomic files (**FASTQ, BAM, VCF**) using command-line tools for sorting, filtering, and mapping.

Genomic files size

sequence.fastq.gz

file size

```
-rw-r--r--  1  thoranin  thoranin  1.5G  Aug  25  2020  P6B_1.fastq.gz
-rw-r--r--  1  thoranin  thoranin  1.6G  Aug  25  2020  P6B_2.fastq.gz
-rw-r--r--  1  thoranin  thoranin  2.5G  Aug  25  2020  P7B_1.fastq.gz
-rw-r--r--  1  thoranin  thoranin  2.6G  Aug  25  2020  P7B_2.fastq.gz
-rw-r--r--  1  thoranin  thoranin  1.5G  Aug  25  2020  P8B_1.fastq.gz
-rw-r--r--  1  thoranin  thoranin  1.6G  Aug  25  2020  P8B_2.fastq.gz
-rw-r--r--  1  thoranin  thoranin  1.5G  Aug  25  2020  P9B_1.fastq.gz
-rw-r--r--  1  thoranin  thoranin  1.5G  Aug  25  2020  P9B_2.fastq.gz
```

sequence.fastq

```
drwxrwxr-x  1  thoranin  thoranin  100M  Sep  17  2024  ta_47.fastq
drwxrwxr-x  1  thoranin  thoranin  86M   Sep  17  2024  ta_48.fastq
drwxrwxr-x  1  thoranin  thoranin  96M   Sep  17  2024  ta_49.fastq
drwxrwxr-x  1  thoranin  thoranin  92M   Sep  17  2024  ta_50.fastq
drwxrwxr-x  1  thoranin  thoranin  102M  Sep  17  2024  ta_51.fastq
drwxrwxr-x  1  thoranin  thoranin  92M   Sep  17  2024  ta_52.fastq
drwxrwxr-x  1  thoranin  thoranin  117M  Sep  17  2024  ta_53.fastq
drwxrwxr-x  1  thoranin  thoranin  164M  Sep  17  2024  ta_54.fastq
```

Linux shell

- **Shell**: A program that interprets and executes user commands, serving as an interface between the user and the operating system.
  Eg. bash, sh, ksh, tcsh, zsh, etc

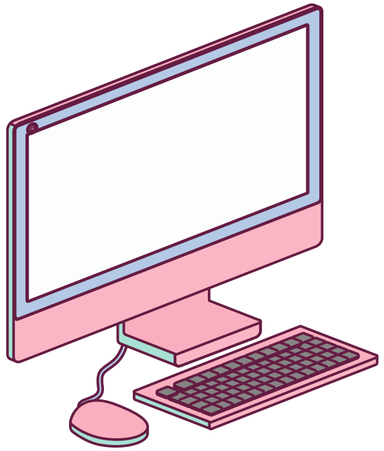
- **Terminal**: A program that run a shell (‘command line’ in window)

- **Directory**: Folder or location of file

Remote Access Using SSH

ssh: Opens SSH client (remote login program)

Example: ssh user@example.com
ssh biga@10.208.103.12



Via ssh to access

personal laptop/ local server

remote server/ cloud

Practice I

HUAWEI CLOUD

- Remote Access to cloud
  - username: **root**
  - password **Biga_workshop**
- check if conda is available

```
ssh root@101.44.63.67
conda --version
```

Note:

If Conda is available, it will show something like: **conda 24.1.2**

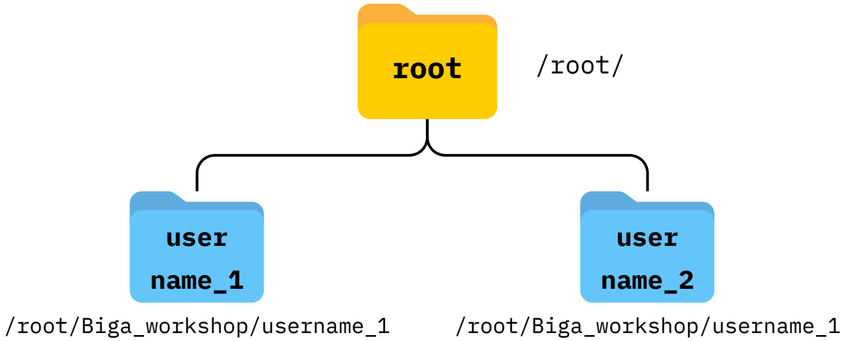
If not, you'll see a message like: **conda: command not found**

Basic Linux Commands

Command	Description	Example
ls	**Lists** files and directories	**ls**
cd	**Changes** the current directory	**cd** /home/user/Documents
pwd	Displays the **current directory path**	**pwd**
mkdir	**Creates** a new directory	**mkdir** new_directory
rmdir	**Removes** directories	**rmdir** directory

Practice II

- Displays the current directory path
- Create a new directory called: **username**
- Lists files and directories
- Change directory to **username**
- Displays the current directory path
- Change directory to Parent directory




```
pwd
mkdir username_folder
ls
cd username_folder
pwd
cd ..
```

**** Please change the username to your name. ****

Note:

~	Home directory
.	Current directory
..	Parent directory
-	Last directory

📌 Use **rm -r folder** if the folder is not empty.

Basic Linux Commands

Command	Description	Example
cp	Copies files or directories	cp file.txt /home/user/
mv	Moves or renames files and directories	mv old_name new_name
rm	Removes files	rm file.txt

Basic Linux Commands

Copying Files and Directories with cp:

```
cp [options] source destination
```

Note: **copy file**

- cp file.txt /path/to/destination

copy and rename newfile

- cp file.txt newname_file.txt

copy directories

- cp -r my_folder /path/to/destination

Basic Linux Commands

Moving files and Directories:

```
mv [options] source destination
```

Note: **Move a File to Another Directory**

- mv file.txt /path/to/destination

Move a Directory

- mv my_folder new_folder_name

📌 **No -r** needed with mv — it moves both files and directories by default.
(cp needs -r for directories.)

Practice III

- Change directory to **data**
- Lists files
- copy **data** to your **username** directory
- change directory to your home directory
- Lists files

Note:

After running each command, `ls` ls to check if the expected files or directories were created successfully.

```
cd data
ls
cp n_3000_vega.fastq /root/user_folder
cp t_3000_vega.fastq /root/user_folder
cd /root/user_folder
ls
```

Text Processing Commands in Linux

Command	Description	Example
cat	Displays the contents of a file	cat file.txt
grep	Searches for a pattern in a file	grep "error" log.txt

Practice IV

- Displays the contents of n_3000_vega.fastq
- Searches “>” in a n_3000_vega.fastq and t_3000_vega.fastq

```
cat n_3000_vega.fastq
grep "@" n_3000_vega.fastq
grep "CATACCT" t_3000_vega.fastq
```

Note:

You can use **cat n_3000_vega.fastq | head** or **head n_3000_vega.fastq** to preview the first few lines (including the headers) of a FASTQ file.

FASTA Format

FASTA format is a text-based format for representing nucleotide or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.

Reference: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQ Format

A text-based format used to store biological sequence data (usually from high-throughput sequencing) along with quality scores for each base.

1. @FORJUSP02AJWD1 → Identifier line (starts with "@")
2. CCGTCAATTCATTTAA... → Raw sequence (A, T, C, G, N)
3. + → Separator line (starts with "+")
4. :!'@((***+))%%%... → Quality scores (ASCII characters)

Reference: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Phred Score

- A quality score used in DNA sequencing to represent the accuracy of each base call (A, T, C, G).
- In **fastq files**, these scores are stored as ASCII characters, not numbers directly

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

where: P = probability that the base is incorrect

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 10000	99.99%
50	1 in 100000	99.999%

Higher Phred score = more confident that the base call is correct.

Xie, S. (2025). Understanding Phred Scores for FASTQ format. Medium.

ASCII table

ASCII stands for American Standard Code for Information Interchange. It is a standardized system that assigns numerical values to characters, allowing computers to store and communicate text.

ASCII uses 7 bits (values from 0 to 127) to represent:

- Letters (A–Z, a–z)
- Digits (0–9)
- Punctuation (e.g., !, @, #)
- Control characters (e.g., newline, tab, backspace)

ASCII table

Sanger, Illumina v1.3 to 1.7 (ASCII_BASE=64)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	A	0.79433	12	L	0.06310	23	W	0.00501	34	b	0.00040
2	B	0.63096	13	M	0.05012	24	X	0.00398	35	c	0.00032
3	C	0.50119	14	N	0.03981	25	Y	0.00316	36	d	0.00025
4	D	0.39811	15	O	0.03162	26	Z	0.00251	37	e	0.00020
5	E	0.31623	16	P	0.02512	27	[	0.00200	38	f	0.00016
6	F	0.25119	17	Q	0.01995	28	\	0.00158	39	g	0.00013
7	G	0.19953	18	R	0.01585	29	]	0.00126	40	h	0.00010
8	H	0.15849	19	S	0.01259	30	^	0.00100			
9	I	0.12589	20	T	0.01000	31	_	0.00079			
10	J	0.10000	21	U	0.00794	32	`	0.00063			
11	K	0.07943	22	V	0.00631	33	a	0.00050			

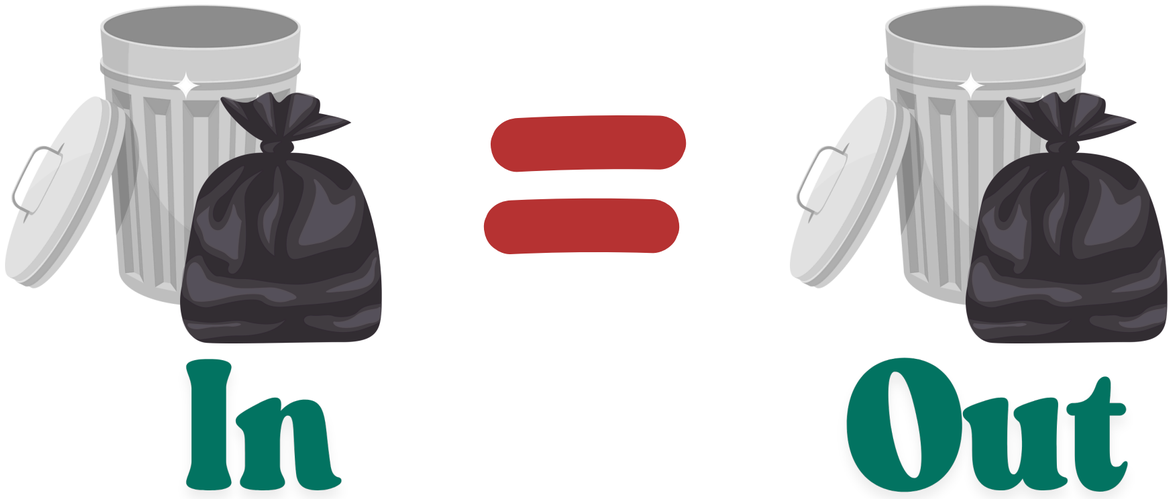
Illumina v1.8 and later (ASCII_BASE=33)

Q	ASCII	P	Q	ASCII	P	Q	ASCII	P	Q	ASCII	P
1	"	0.79433	12	-	0.06310	23	8	0.00501	34	C	0.00040
2	#	0.63096	13	.	0.05012	24	9	0.00398	35	D	0.00032
3	$	0.50119	14	/	0.03981	25	:	0.00316	36	E	0.00025
4	%	0.39811	15	0	0.03162	26	;	0.00251	37	F	0.00020
5	&	0.31623	16	1	0.02512	27	<	0.00200	38	G	0.00016
6	'	0.25119	17	2	0.01995	28	=	0.00158	39	H	0.00013
7	(	0.19953	18	3	0.01585	29	>	0.00126	40	I	0.00010
8	)	0.15849	19	4	0.01259	30	?	0.00100	41	J	0.00008
9	*	0.12589	20	5	0.01000	31	@	0.00079			
10	+	0.10000	21	6	0.00794	32	A	0.00063			
11	,	0.07943	22	7	0.00631	33	B	0.00050			

Xie, S. (2025). Understanding Phred Scores for FASTQ format. Medium.

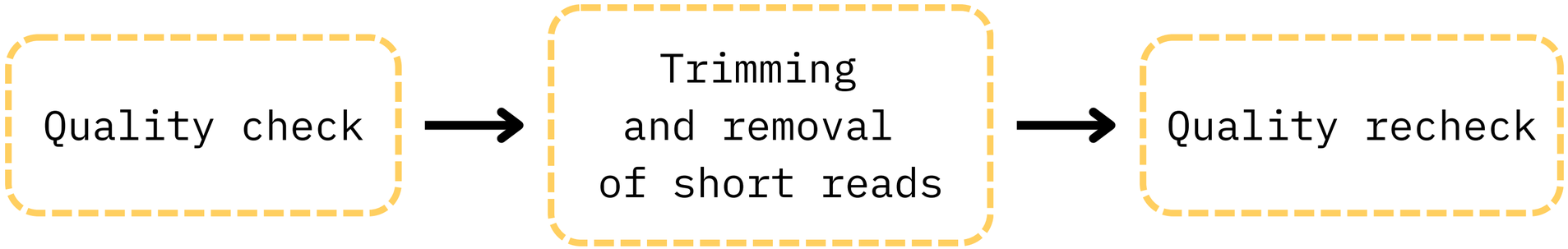
- Older sequencing machines (early Illumina) used Phred+64
- Newer machines (modern Illumina, Sanger, etc.) all use Phred+33
- Today, **Phred+33** is the standard in most pipelines

Quality Control: Fastqc



“garbage in, garbage out”: If poor-quality data enters your analysis pipeline, it can lead to misleading or incorrect conclusions.

Quality Control: Fastqc



After the initial quality check, Low-quality bases and short reads are trimmed, followed by a second quality check to ensure the remaining reads are suitable for analysis.

Fastqc: installation

1. Create a new Conda environment

```
conda create -n fastqc
```

2. Activate the environment

```
conda activate fastqc
```

3. Install FastQC from Bioconda

```
conda install bioconda::fastqc
```

4. Verify the installation

```
fastqc --help
```

Fastqc: how to run

1.single file

```
fastqc file1.fastq
```

2. Multiple Files

```
fastqc file1.fastq file2.fastq
```

Practice V

- activate fastqc environment
- change directory to **username_folder**
- Lists files
- run fastqc

```
conda activate fastqc
cd /root/username_folder
ls
fastqc n_3000_vega.fastq
fastqc n_3000_vega.fastq t_3000_vega.fastq -o output_qc -t 4
```

Note:

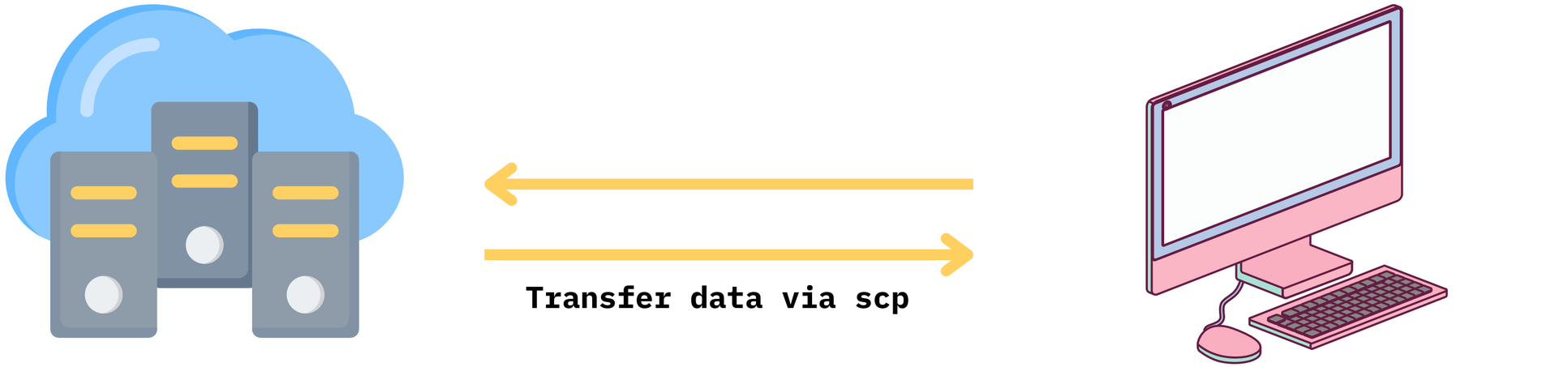
-o output_fastqc: specifies output directory for results. Make sure output_fastqc directory exists before running command.

-t 4: uses 4 threads to speed up analysis

SCP

scp: Downloading a file from a remote server to local

```
scp [options] source destination
```



remote server/ cloud

personal laptop/ local server

mount directories

- The WSL has access to your PC’s file system through /mnt/directories (or mount points).
- For example, your C: and D: root directories in Windows would be available through /mnt/c/ and /mnt/d/ respectively in the WSL

```
cd /mnt/directories
```

Practice VI

- Change directory to C
- Creates a directory named output_fastqc in /mnt/c/
- Change directory to output_fastqc
- Displays the current directory path (should be /mnt/c/output_fastqc)
- transfer data to /mnt/c/output_fastqc
- Lists files and directories

```
cd /mnt/c/
mkdir output_fastqc
cd output_fastqc
pwd
scp -r root@101.44.63.67:/root/username/ /mnt/c/output_fastqc
ls
```

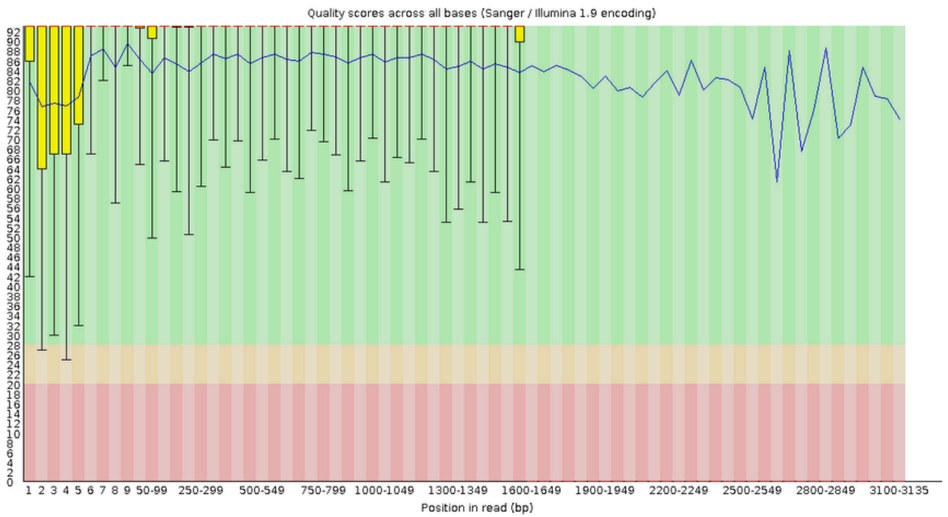
Fastqc: Quality Control Modules

- **Basic Statistics**: Provides a summary of the input data.
- **Per Base Sequence Quality**: Visualizes the quality scores across all bases.
- **Per Sequence Quality Scores**: Displays the distribution of quality scores across all sequences.
- **Per Base N Content**: Shows the percentage of unknown bases across all bases.
- **Sequence Length Distribution**: Illustrates the distribution of sequence lengths.
- **Duplicated Sequences**: Identifies the percentage of sequences with potential duplicates.
- **Adapter Content**: Detects adapter sequences that may have been included during library preparation.

Refreence: https://diphyx.com/stories/how-to-use-fastqc-cli/

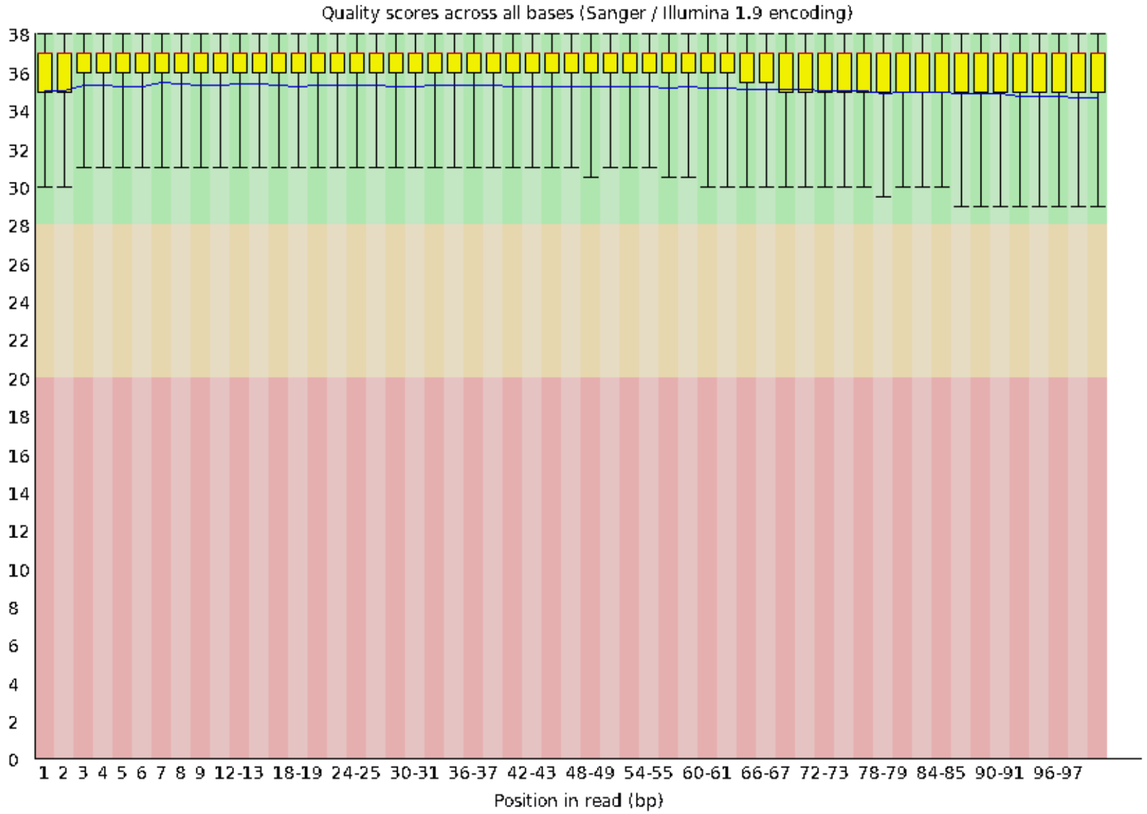
Fastqc: Output Reports

- **HTML Report:** An interactive HTML file summarizing all quality checks.
- **Text Report:** A detailed text file providing the raw data for each analysis.
- **Zipped File:** A compressed file containing all the outputs for easy sharing and storage.

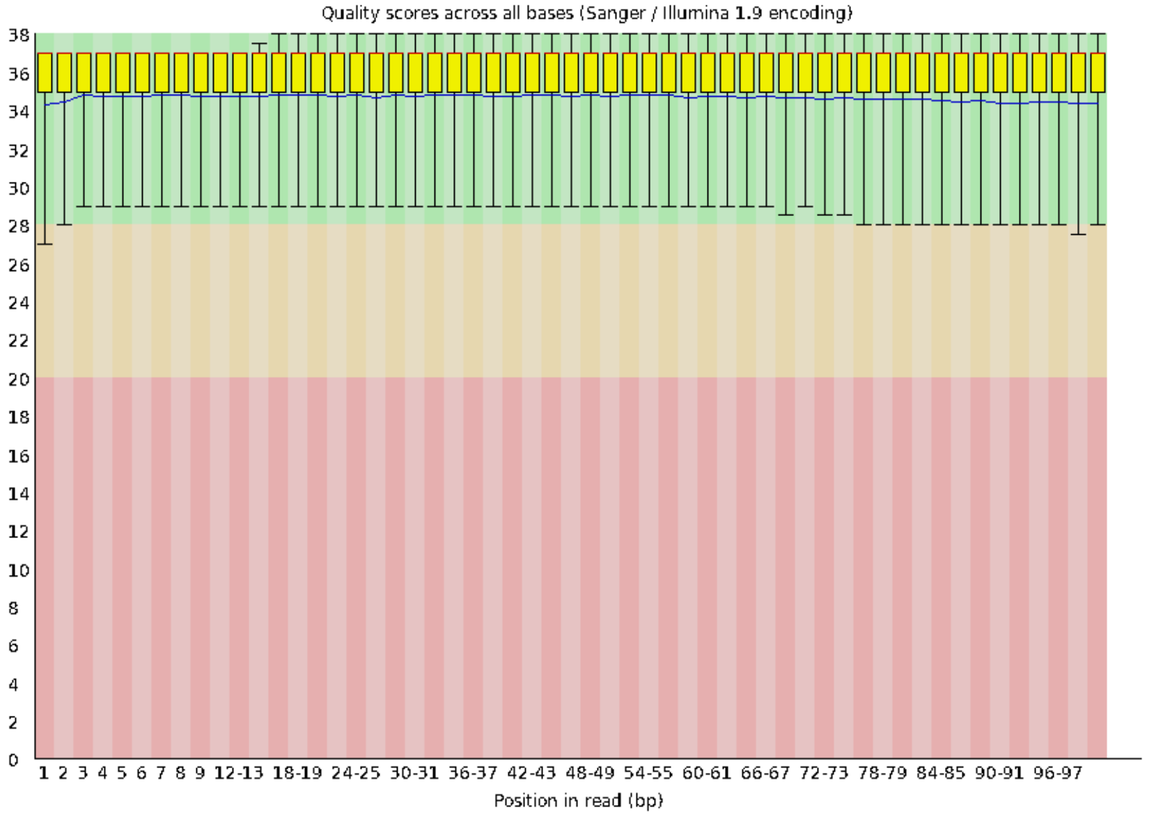


Quality Control: Fastqc

company b



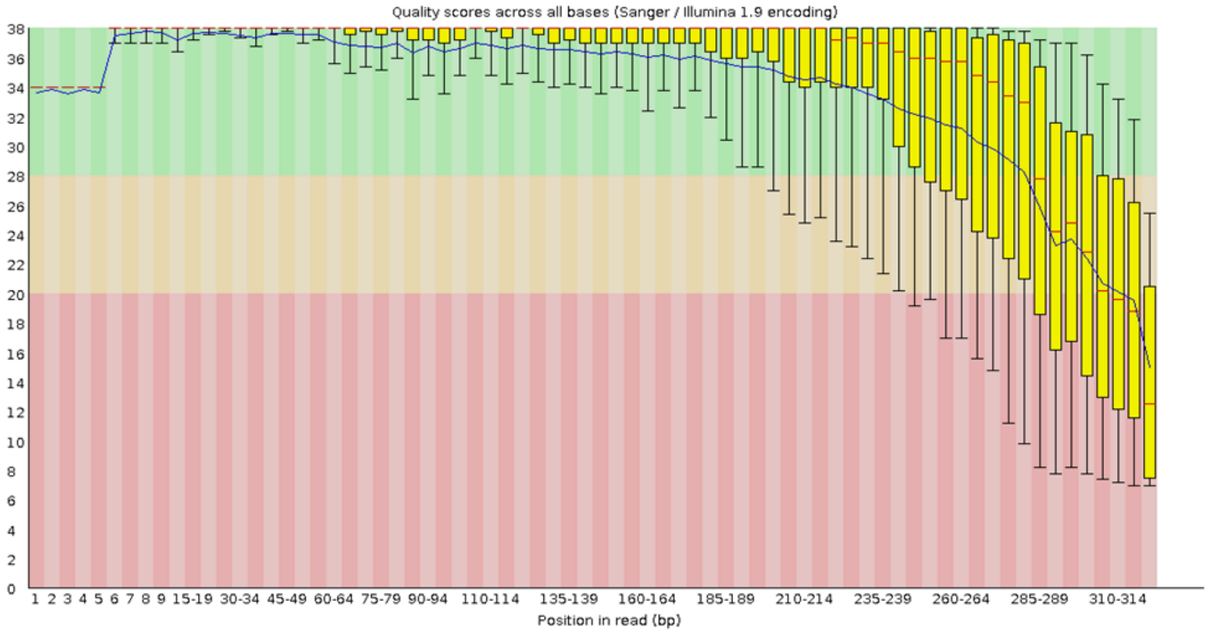
forward



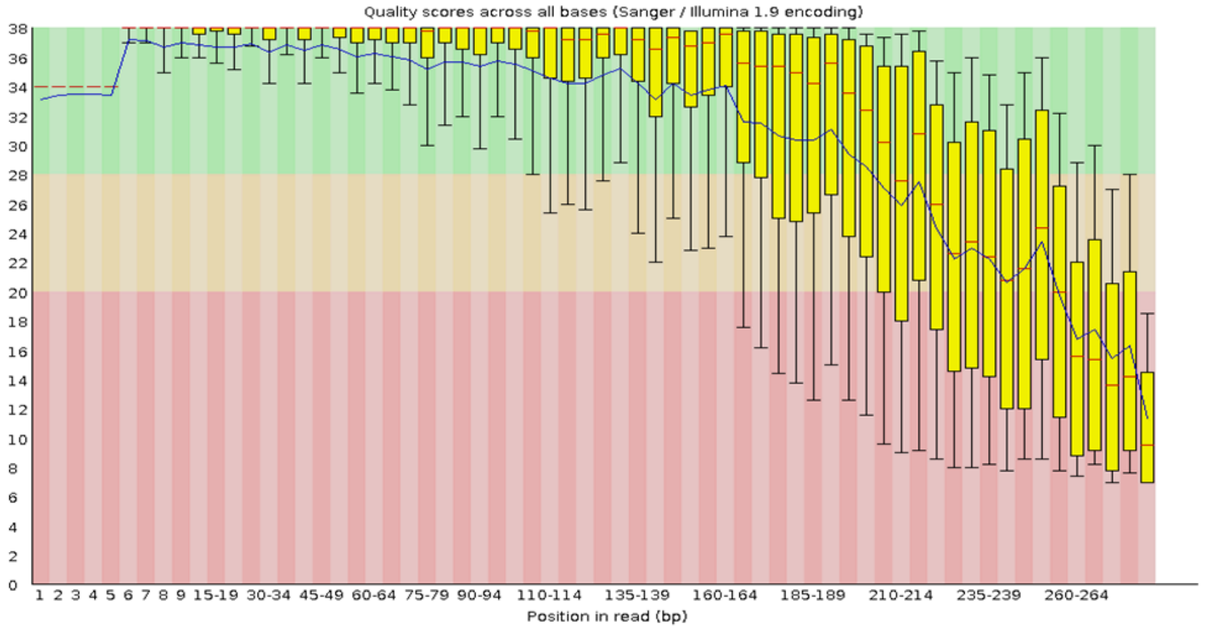
reverse

Quality Control: Fastqc

company i



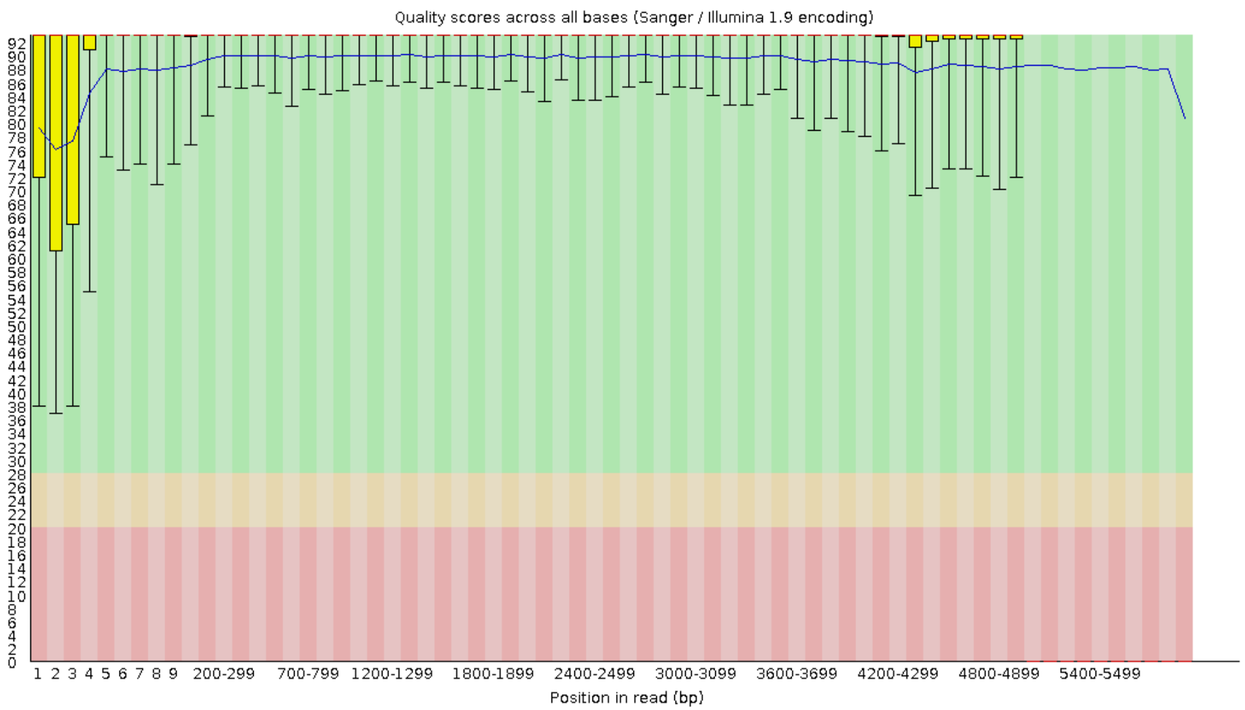
forward

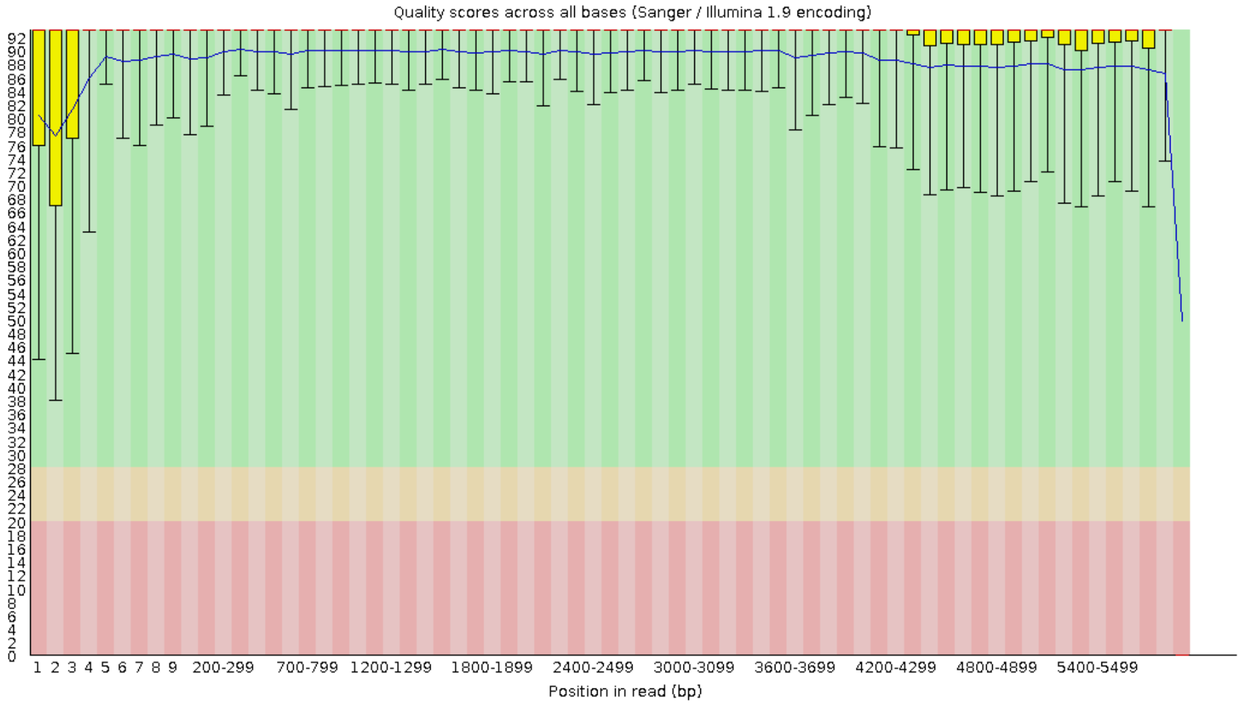


reverse

Quality Control: Fastqc

company p





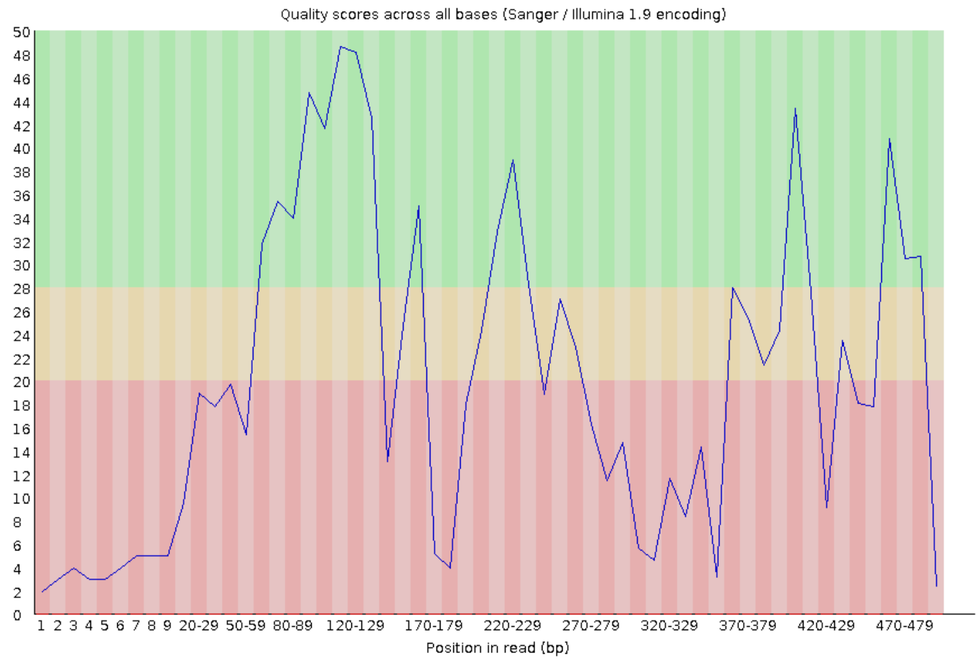
repeat gene

Quality Control: Fastqc

A quality control tool for high throughput sequence data

Basic Statistics

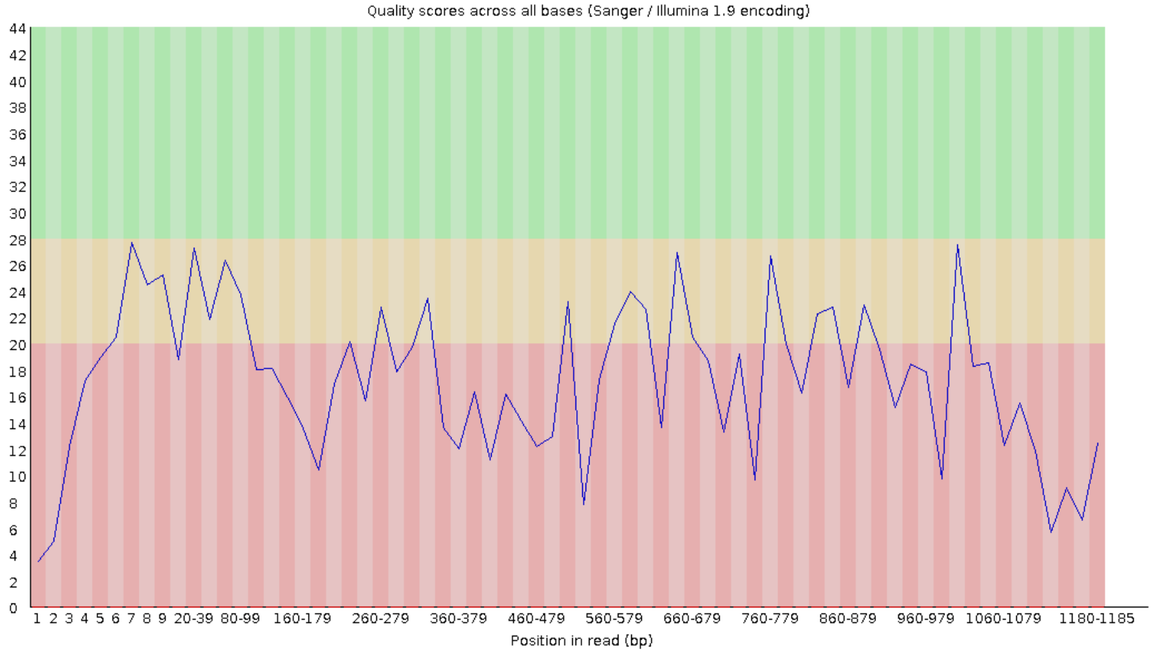
Measure	Value
Filename	FAY61511_pass_501d0f7c_c65828f8_96.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1
Sequences flagged as poor quality	0
Sequence length	498
%GC	98



Basic Statistics

Measure	Value
Filename	FAY61511_pass_501d0f7c_c65828f8_98.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4
Sequences flagged as poor quality	0
Sequence length	169-1185
%GC	65

company N

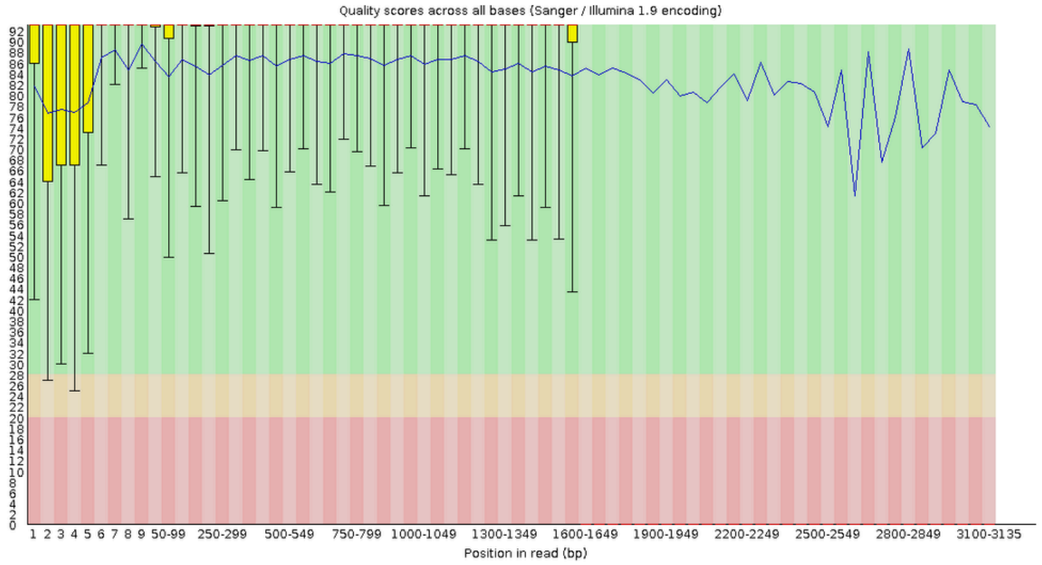


https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Quality Control: Fastqc

Basic Statistics

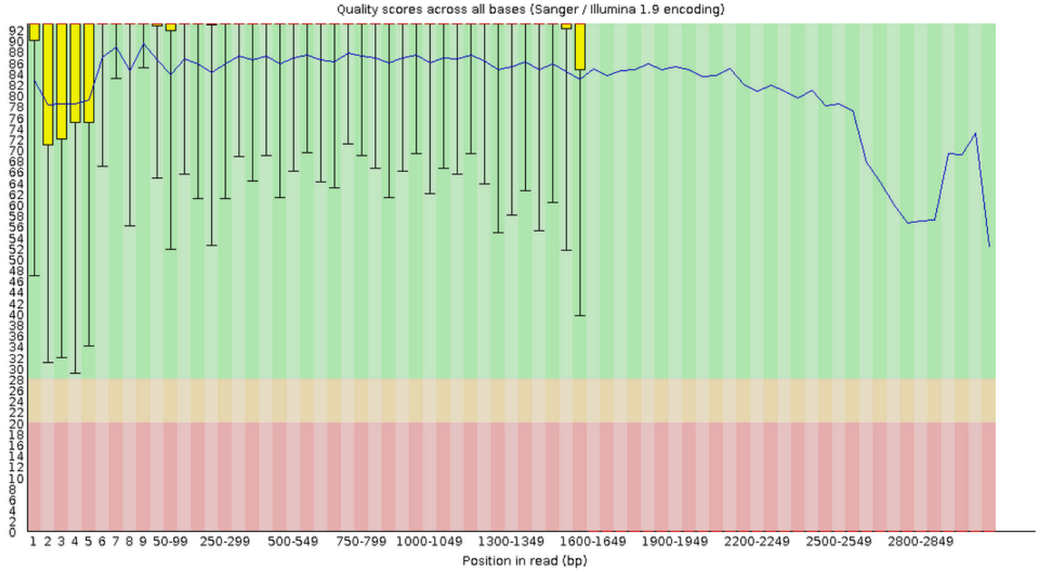
Measure	Value
Filename	sample_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8803
Sequences flagged as poor quality	0
Sequence length	590-3135
%GC	52



Basic Statistics

Measure	Value
Filename	sample_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	11311
Sequences flagged as poor quality	0
Sequence length	1219-3059
%GC	53

company P



Carney Complex disease

A

7 yrs	left atrial myxoma resection, atrial repair
10 yrs	testicular mass, right orchiectomy
13 yrs	pituitary tumor
16 yrs	recurrence of myxomata, resection, adrenal microadenoma
18 yrs	recurrence of ventricular myxomata, resection, VT
19 yrs	ACTH-independent Cushing's disease, thyroid nodules
21 yrs	transphenoidal resection of pituitary
present	recurrence of myxomata, under consideration for heart transplant

B

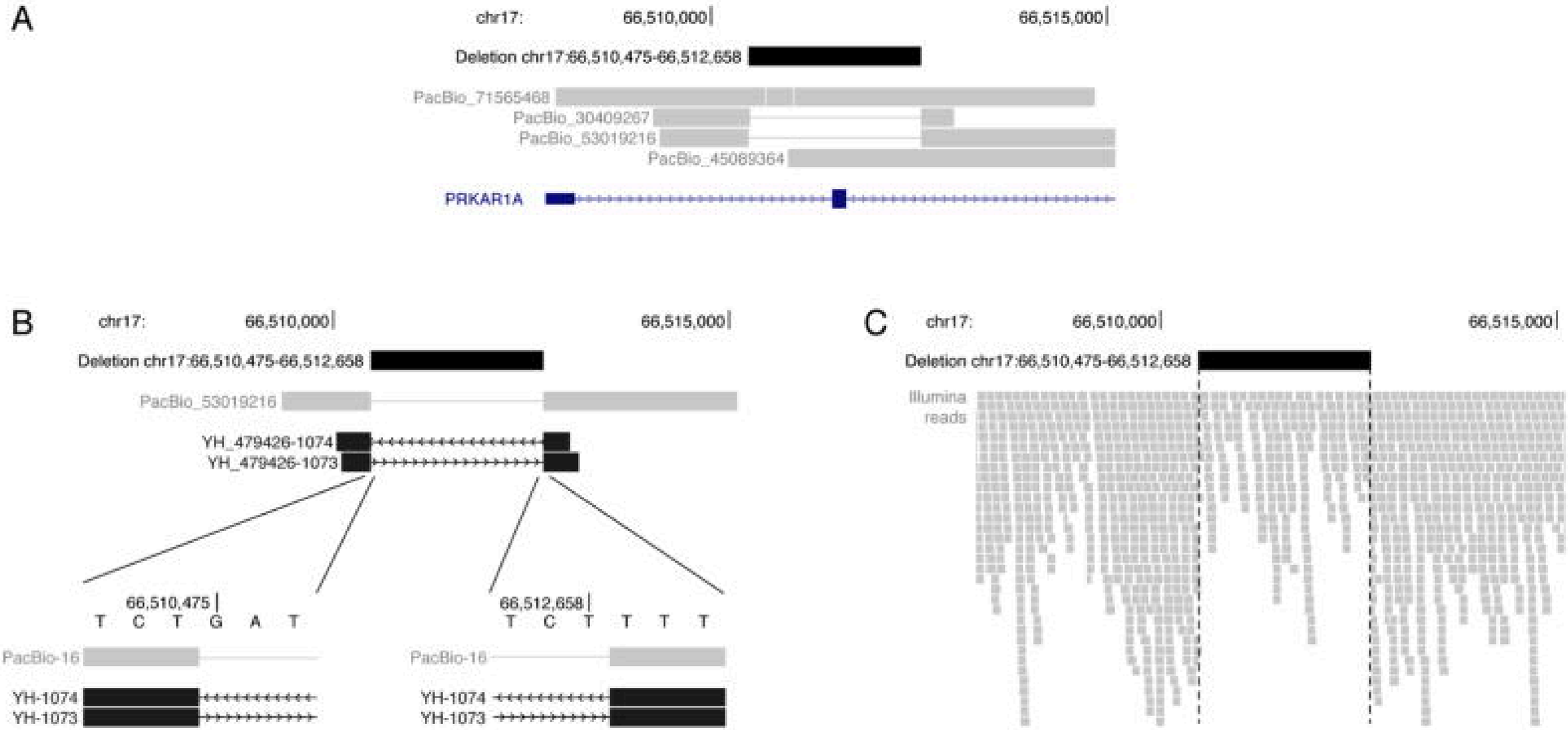
C

The patient is an Asian/Hispanic male was hospitalized for the first 10 days of life for cardiac and respiratory issues

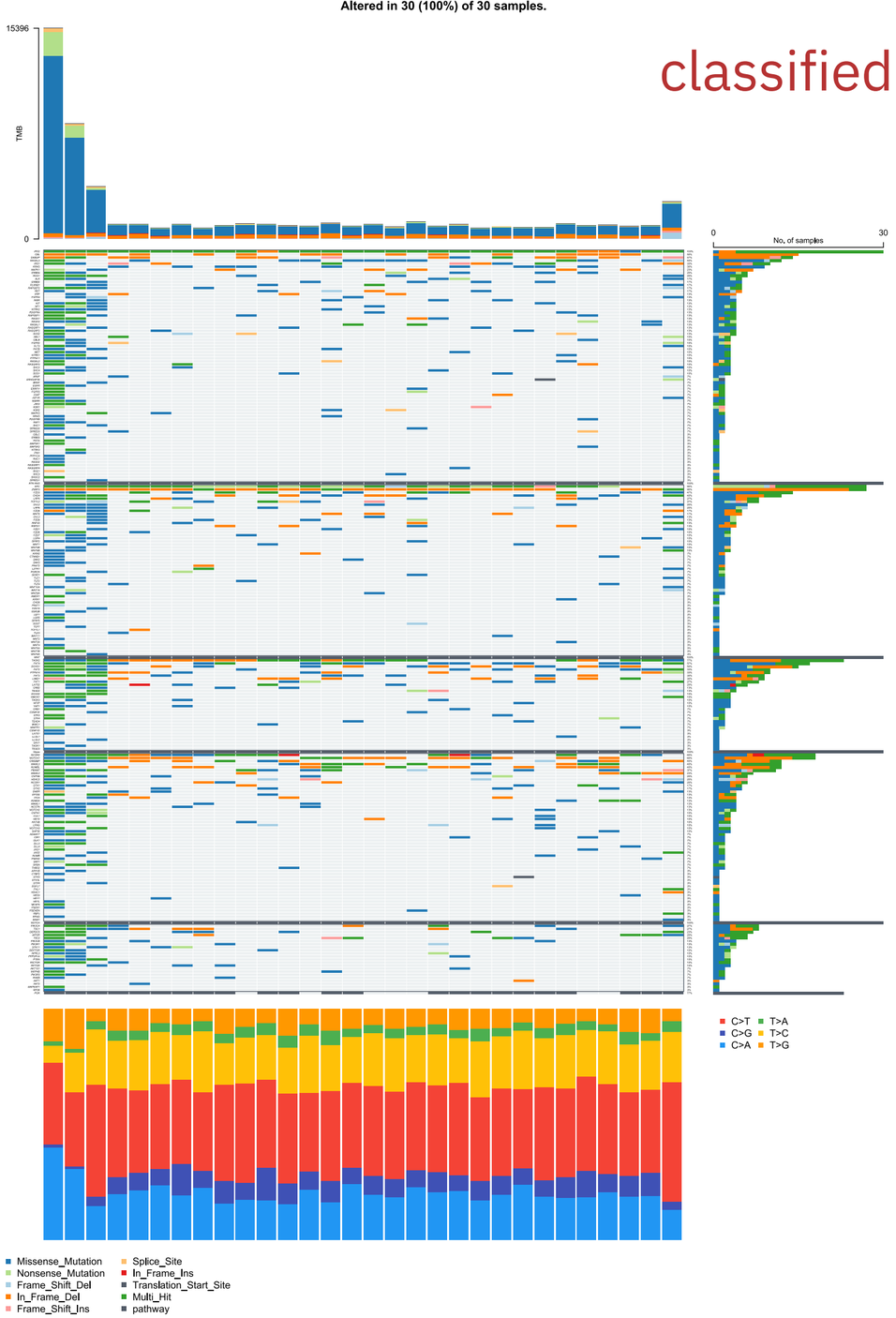
a genetics evaluation suggested the possibility of Carney complex but clinical sequencing of PRKAR1A was negative for disease causing variation

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., ... & Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine, 20(1), 159-163.

Carney Complex disease



Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., ... & Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genetics in Medicine, 20(1), 159-163.


Altered in 30 (100%) of 30 samples.
15396
TMB
0
0
30
No. of samples
C>T
C>G
C>A
T>A
T>C
T>G
Missense_Mutation
Nonsense_Mutation
Frame_Shift_Del
In_Frame_Del
Frame_Shift_Ins
Splice_Site
In_Frame_Ins
Translation_Start_Site
Multi_Hit
pathway

classified

"If you're not too busy..."

"Janet, you don't need to suit up if you're just assembling a virus genome."

Himelblau, E. (2021). A cartoon guide to bioinformatics by a novice coder. Nature.

References

- Linux Foundation. (n.d.). What is Linux? Retrieved from https://www.linux.com/what-is-linux/
- CompGenOMR. (n.d.). FASTA and FASTQ formats. Retrieved from https://compgenomr.github.io/book/fasta-and-fastq-formats.html
- RTSF MSU. (2017). FastQC Tutorial and FAQ. Retrieved from https://rtsf.natsci.msu.edu/sites/_rtsf/assets/File/FastQC_TutorialAndFAQ_080717.pdf
- Usadel, B. (2014). Trimmomatic Manual v0.32. Retrieved from http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf