



# FSK-2053 Data science & bioinformatics for fisheries and aquaculture

## Lecture 6 – Introduction to bioinformatics:

### Linux systems and biological data

Daniel Kumazawa Morais

[daniel.morais@uit.no](mailto:daniel.morais@uit.no)



# What is Linux?

Linux is a family of free and open source operating systems. It was released in 1991 and developed by Linus Torvalds.

This operating system is in your phones, smart TVs, home appliances, most of the internet servers, the top 500 biggest supercomputers, stock exchange servers and of course desktop computers.

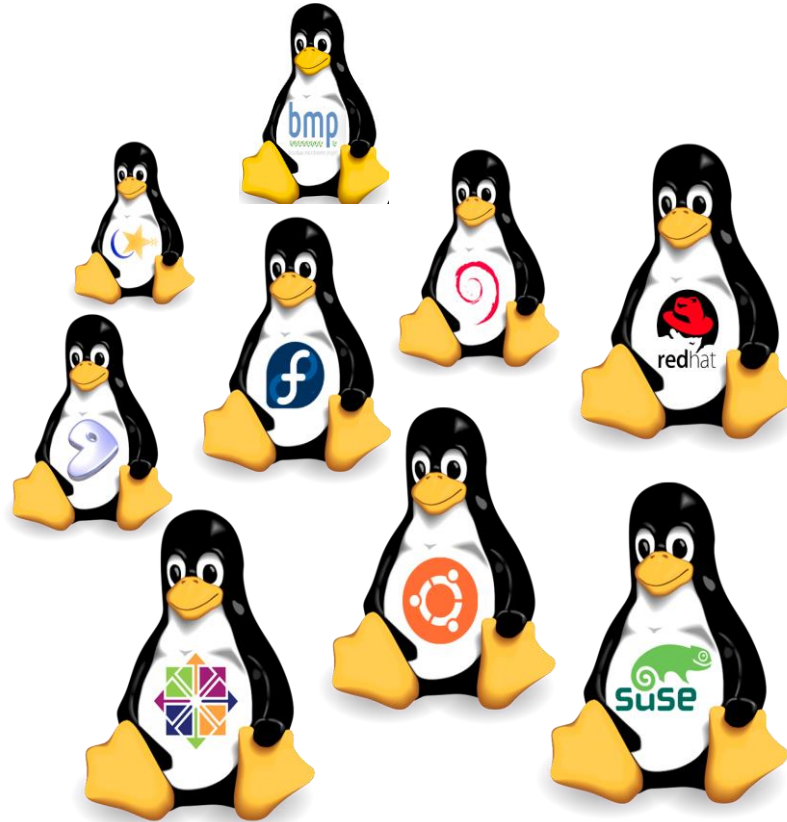
It got that big, thanks to voluntary collaborations around the world.



# Linux

- Linux is free
- Most bioinformatics platforms are developed for Unix systems
- High performance and easy control of processes
- Easy share of resources
- Highly adopted by the scientific community
- Possibility of deep modifications aiming specific interests: Eg. Genome Assembly Courses
- High number of Open Source/Free tools

# Linux Distros

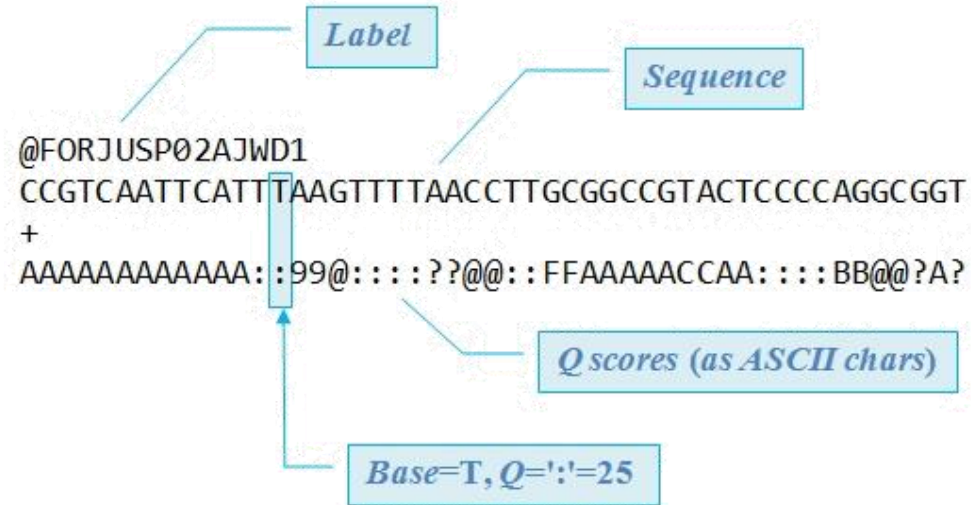




# Types biological data we have now

## Fastq files and phred score

Quality Score  
 $P = 10^{-Q/10}$   
 $Q = -10 \log_{10}(P)$



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			

# Fastq file header

```
@M02149:53:000000000-AANLH:1:1101:14924:1701 1:N:0:0
TACGGAGGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACGTA
GGCTGTCTGGTAAGTCAGGGGTGAAATCCCGCGGCTCACCCGCGGAATT
GCCCTTGATACTGCTGGACTTGAGTTCGGGAGAGGGTGGCGGAATTCCAG
GTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATCAGGGGGCGAAGG
CGGCCACCTGGACCGATACTGACGCTGAGGTGCGAAAGCGTGGGGGAGGA
AACAGG
```

+

```
AAA??1>DDAAA11AFEGF00BGCEA0F1A1F10AAAFAB//BAAA/AAB00ABGFF
@F10BB@DGG2B00/B//1@BF1F/>>>EEA<1B</<>///?F?DD<FGF>??<F1<F<
??<FGHF?G<?CHHHHHFF<::/0GHFB;:BFF0F;<1GG>BF2HHEB//?F@HGB@
B110FFHFHGB1B0FB>/EE>HGFEEAA0/1A011EEBA/2D2D/AEEABB1FHE00
AAGFFEA1A1GGFFFB3@F>1AAA
```

Illumina header

```
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>
<read>:<is filtered>:<control number>:<sample number>
```

Lets use these tools to  
understand and manipulate our sequences



# **Brazilian Microbiome Project: Revealing the Unexplored Microbial Diversity—Challenges and Prospects**

**Victor Satler Pylro • Luiz Fernando Wurdig Roesch • José Miguel Ortega •  
Alexandre Morais do Amaral • Marcos Rogério Tótola • Penny Ruth Hirsch •  
Alexandre Soares Rosado • Aristóteles Góes-Neto • Artur Luiz da Costa da Silva •  
Carlos Augusto Rosa • Daniel Kumazawa Morais • Fernando Dini Andreote •  
Gabriela Frois Duarte • Itamar Soares de Melo • Lucy Seldin • Márcio Rodrigues Lambais •  
Mariangela Hungria • Raquel Silva Peixoto • Ricardo Henrique Kruger  
Siu Mui Tsai • Vasco Azevedo •  
The Brazilian Microbiome Project Organization Committee**



- **What linux programs our users needed?**

- QIIME 1.9.0 - <http://qiime.org/index.html>
- USEARCH 7/8 (UPARSE) - <http://drive5.com/uparse/>
- UPARSE Python scripts – <http://drive5.com/python/>
- BMP Scripts – <https://github.com/BMP>
- BIOM format scripts – <http://biom-format.org/>
- FASTX Toolkit – [http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)
- ITSx – <http://microbiology.se/software/itsx/>
- R packages - <https://www.bioconductor.org/>

**- We really needed to find a way to make it easier for beginner users!!!**

METHODS

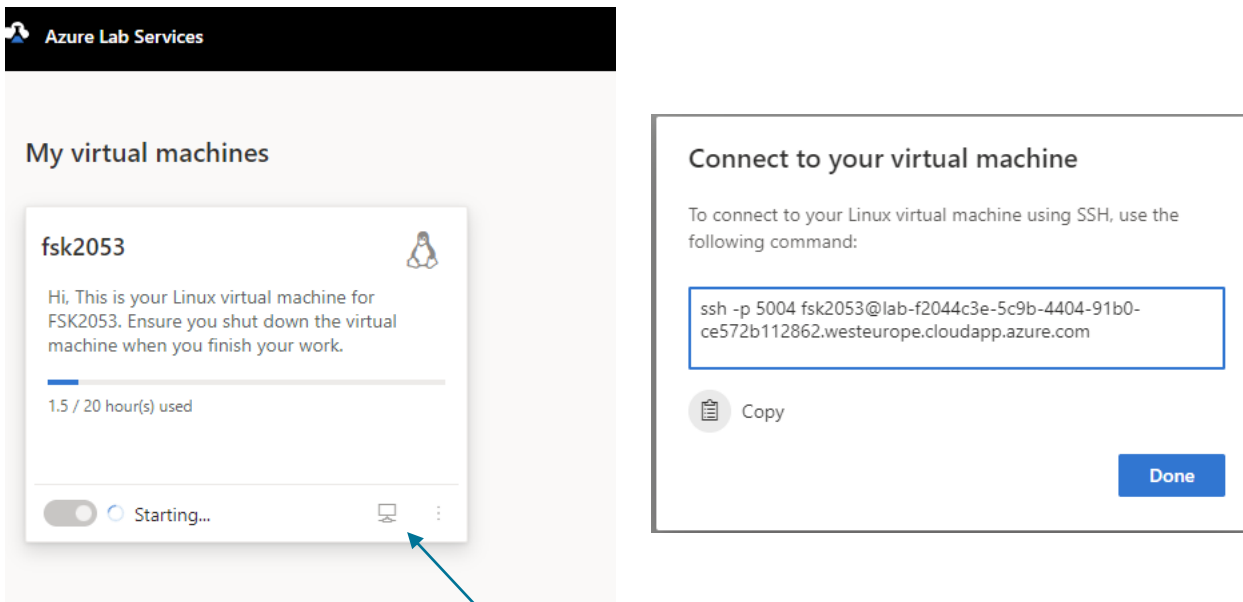
# BMPOS: a Flexible and User-Friendly Tool Sets for Microbiome Studies

Victor S. Pylro<sup>1</sup> • Daniel K. Morais<sup>1</sup> • Francislton S. de Oliveira<sup>1</sup> • Fausto G. dos Santos<sup>1</sup> •  
Leandro N. Lemos<sup>2</sup> • Guilherme Oliveira<sup>3</sup> • Luiz F. W. Roesch<sup>4</sup>



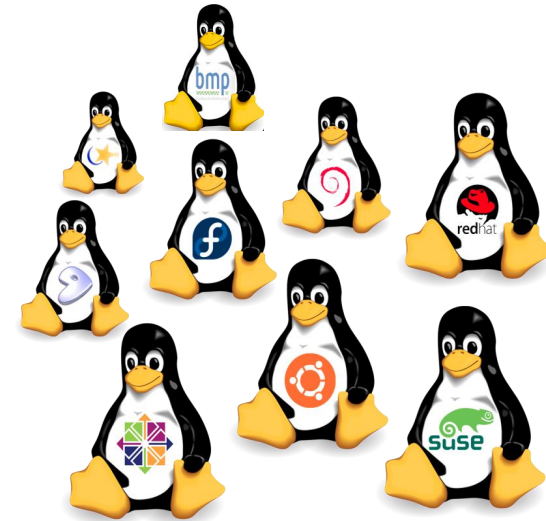
# Linux Connection

- Connection to our cloud Linux machines:
  - <https://labs.azure.com/register/detno9f2n>



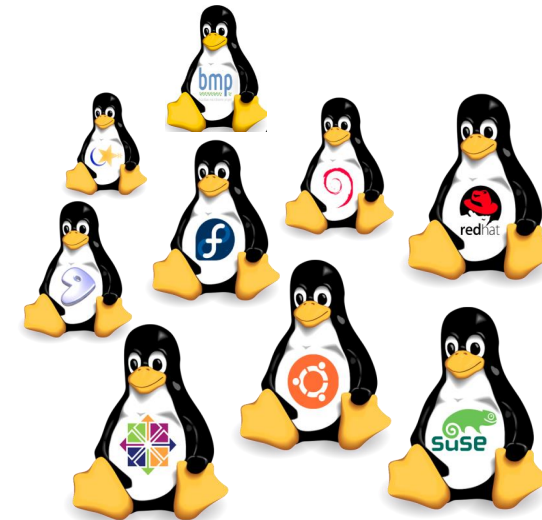
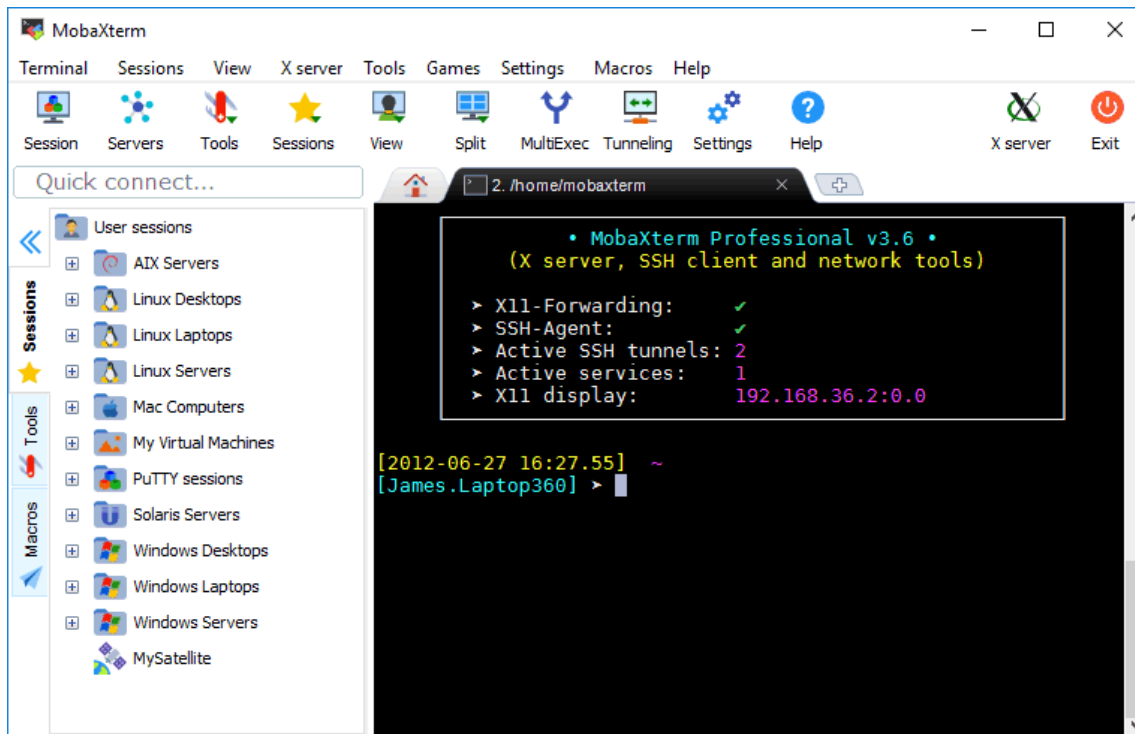
The screenshot shows the Azure Lab Services interface. On the left, under 'My virtual machines', there is a card for a machine named 'fsk2053'. The card includes a message: 'Hi, This is your Linux virtual machine for FSK2053. Ensure you shut down the virtual machine when you finish your work.' and a progress bar showing '1.5 / 20 hour(s) used'. At the bottom of the card, there is a status indicator 'Starting...' and a small icon of a terminal window. A blue arrow points from this icon to the right-hand panel. The right-hand panel, titled 'Connect to your virtual machine', provides instructions on how to connect using SSH and displays the following command in a text box: `ssh -p 5004 fsk2053@lab-f2044c3e-5c9b-4404-91b0-ce572b112862.westeurope.cloudapp.azure.com`. Below the command box is a 'Copy' button and a 'Done' button.

The connection command to your machines

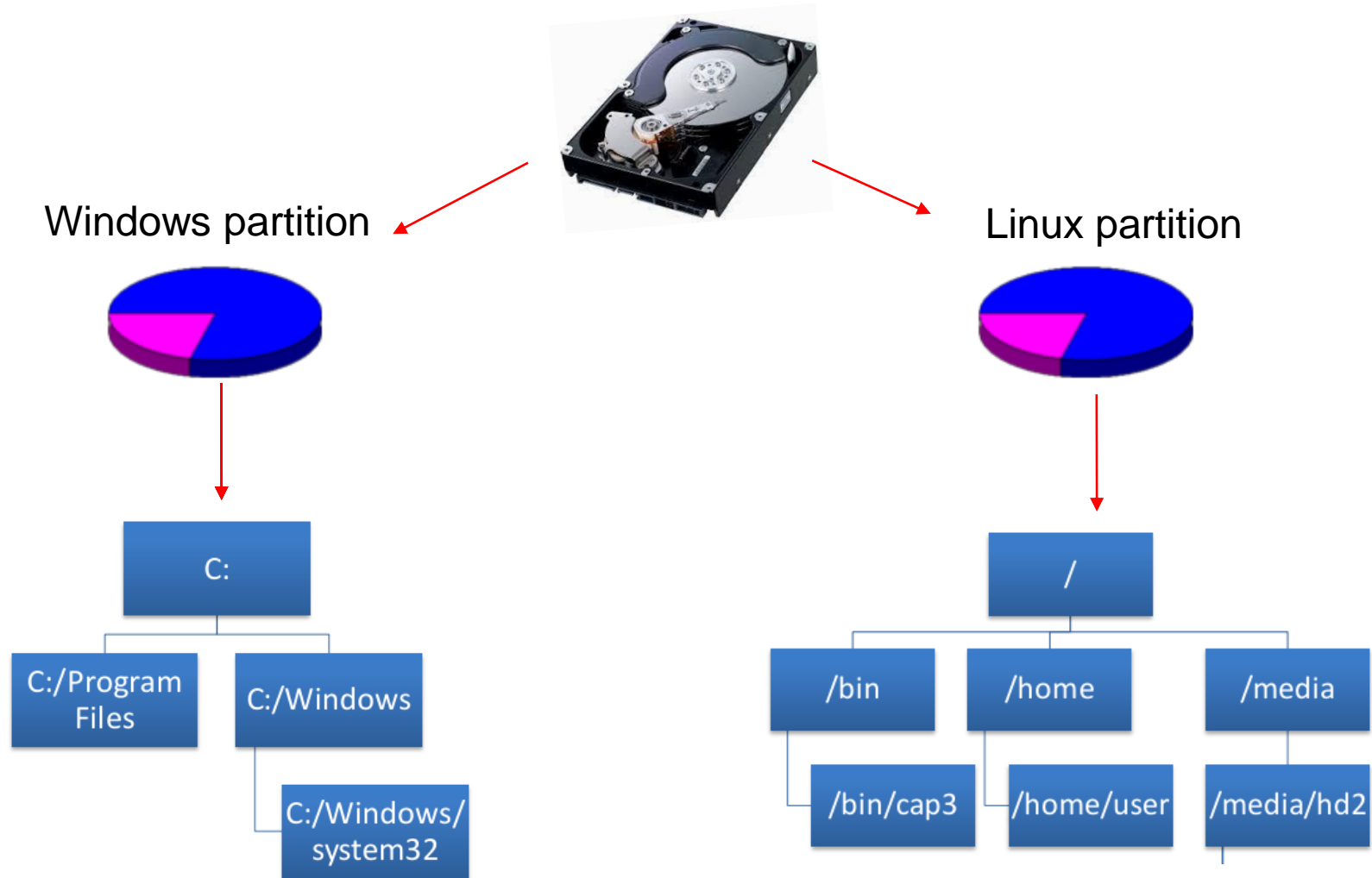


# Linux Connection

- Connection to our cloud Linux machines:
  - <https://labs.azure.com/register/detno9f2n>

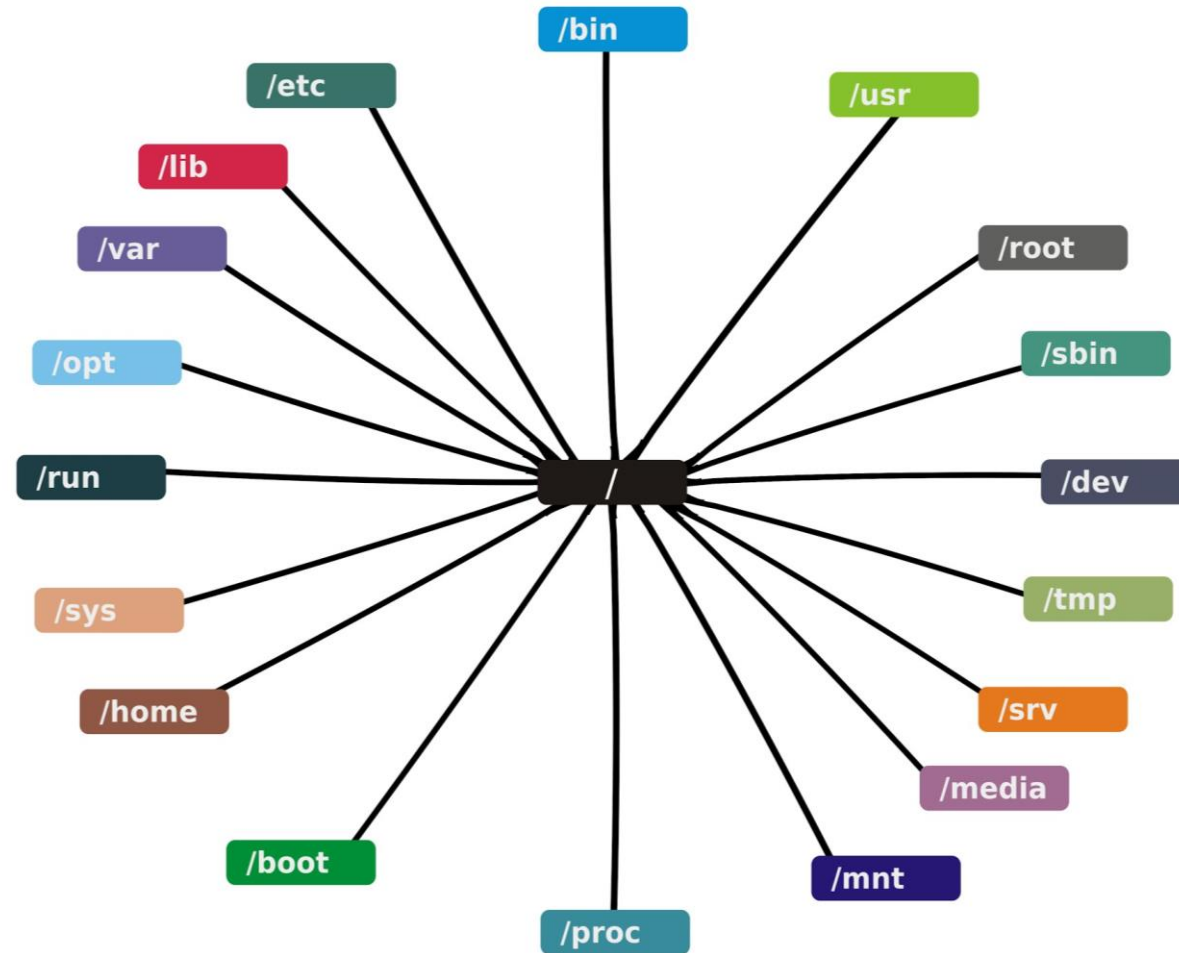


# Directory organization system

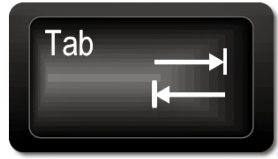




# Directory organization system



# Basic Linux terminal usage concepts



## Key TAB

The autocomplete key



## Path: /usr/local/bin/blastn

Path to the blastn tool



## *User mode and variable calls.*

```
echo $HOME
```



## Root/Power mode

Screen

<https://help.ubuntu.com/community/Screen>

Tmux

<https://manpages.ubuntu.com/manpages/bionic/man1/tmux.1.html>

Terminal  
multiplexers

# Basic Linux terminal usage concepts



*User mode and variable calls.*

```
echo $HOME
```

```
fsk2053@lab000000:/home/adminfsk2053$ pwd
/home/adminfsk2053
fsk2053@lab000000:/home/adminfsk2053$ ls ./*
./data:
fasta_test1.fasta  results2.txt  unknown1.txt

./database:
nt_teleost_16112020.00.nhr  nt_teleost_16112020.04.nhr  nt_teleost_16112020.08.nhr  nt_teleost_16112020.12.nhr
nt_teleost_16112020.00.nin  nt_teleost_16112020.04.nin  nt_teleost_16112020.08.nin  nt_teleost_16112020.12.nin
nt_teleost_16112020.00.nog  nt_teleost_16112020.04.nog  nt_teleost_16112020.08.nog  nt_teleost_16112020.12.nog
nt_teleost_16112020.00.nsq  nt_teleost_16112020.04.nsq  nt_teleost_16112020.08.nsq  nt_teleost_16112020.12.nsq
nt_teleost_16112020.01.nhr  nt_teleost_16112020.05.nhr  nt_teleost_16112020.09.nhr  nt_teleost_16112020.13.nhr
nt_teleost_16112020.01.nin  nt_teleost_16112020.05.nin  nt_teleost_16112020.09.nin  nt_teleost_16112020.13.nin
nt_teleost_16112020.01.nog  nt_teleost_16112020.05.nog  nt_teleost_16112020.09.nog  nt_teleost_16112020.13.nog
nt_teleost_16112020.01.nsq  nt_teleost_16112020.05.nsq  nt_teleost_16112020.09.nsq  nt_teleost_16112020.13.nsq
nt_teleost_16112020.02.nhr  nt_teleost_16112020.06.nhr  nt_teleost_16112020.10.nhr  nt_teleost_16112020.nal
nt_teleost_16112020.02.nin  nt_teleost_16112020.06.nin  nt_teleost_16112020.10.nin  nt_teleost_16112020.ndb
nt_teleost_16112020.02.nog  nt_teleost_16112020.06.nog  nt_teleost_16112020.10.nog  nt_teleost_16112020.nos
nt_teleost_16112020.02.nsq  nt_teleost_16112020.06.nsq  nt_teleost_16112020.10.nsq  nt_teleost_16112020.not
nt_teleost_16112020.03.nhr  nt_teleost_16112020.07.nhr  nt_teleost_16112020.11.nhr  nt_teleost_16112020.ntf
nt_teleost_16112020.03.nin  nt_teleost_16112020.07.nin  nt_teleost_16112020.11.nin  nt_teleost_16112020.nton
nt_teleost_16112020.03.nog  nt_teleost_16112020.07.nog  nt_teleost_16112020.11.nog
nt_teleost_16112020.03.nsq  nt_teleost_16112020.07.nsq  nt_teleost_16112020.11.nsq

./software:
R-4.2.3  R-4.2.3.tar.gz  bioawk  bioinfo_functions.sh  ncbi-blast-2.13.0+-src  ncbi-blast-2.13.0+-src.tar.gz
fsk2053@lab000000:/home/adminfsk2053$
```

User name: *fsk2053*

Machine: *lab000000*

Working directory: */home/adminfsk2053*

# Terminal

```
terminator
root@dourado:usr/local/bioinformatic 91x66
o4data-2]# htop
o4data-2]# w
ays, 8:09, 5 users, load average: 1.02, 1.31, 1.69
FROM LOGIN# IDLE JCPU PCPU WHAT
nstd.cebio.org 16:12 9:33 0.18s 0.18s -bash
ngs.cabio.org 17:07 0:00s 0.18s 0.02s sshd: fausto [priv]
ngs.cabio.org 16:37 3:06 0.16s 0.16s -bash
gwbcm.cena.usp.br 09:34 5:40m 0.14s 0.10s screen
:pts/11:5.0 09:34 5:40m 0.07s 0.07s /bin/bash
o4data-2]# htop
o4data-2]# exit
~]$ htop
~]$ logout
unare closed.
$ ssh -X dourado
password:
ab 8 08:33:28 2017 from 200.131.9.147
i$ su -

]$ sudo su
sto)# cd /usr/local/bioinformatic/
informatic]# ls
g
  Metaxa2_2.1.3
  mixtcond2
  MOCAT
an2
  ncbi-blast-2.3.0+
  ncbi-blast-2.3.0+-x64-linux.tar.gz
er
  ncbi-blast-2.4.0+src
ase
  ncbi-blast-2.4.0+src.tar.gz
  old-quime
GAG
  pplacer-Linux-v1.1.alpha18-2-gcb55169
GAG-v1.1-129-g98da78e.tar.gz
  quast-v4.2
  ruamel.yaml-0.6.1
  ruamel.yaml-0.6.1-linux.tar.gz
  samtools-1.3.1
  scala-2.9.0.final
extensions
  smrtanalysis
  STAR-master
  STAR-STAR_2.5.0a
  tmp.prodigal.stdin.108647
  trf409.Linuxx64
17.tar.gz
informatic]#

fausto@ti:/
fausto@ti:/ 73x15
[fausto@ti /]$ man htop
[fausto@ti /]$ ls
bin etc lib64 mnt root sbin testperl.pl var
boot home lost+found opt run srv sys tmp
dev lib media proc samba sys usr
[fausto@ti /]$

fausto@ti:/ 73x15
[fausto@ti /]$ df -h
Sist. Arq. Tam. Usado Disp. Uso% Montado em
devtmpfs 3,9G 0 3,9G 0% /dev
tmpfs 3,9G 278M 3,7G 7% /dev/shm
tmpfs 3,9G 1,7M 3,9G 1% /run
tmpfs 3,9G 0 3,9G 0% /sys/fs/cgroup
/dev/mapper/fedora-root 92G 53G 39G 62% /
tmpfs 3,9G 19M 3,9G 1% /tmp
tmpfs 798M 16K 798M 1% /run/user/42
tmpfs 798M 44K 798M 1% /run/user/1000
[fausto@ti /]$

fausto@ti:/ 145x32
1 1 Tasks: 161, 635 thr: 1 running
2 2 Load average: 0.35 0.47 0.56
3 3 Uptime: 1 day, 02:54:17
4 4
Mem 4.93G/7.79G
Swp 4K/3.91G

PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
6713 fausto 20 0 126M 4752 3108 R 1.3 0.1 0:00.15 htop
2620 fausto 20 0 1723M 407M 125M S 1.3 5.1 16:58.83 /opt/google/chrome/chrome
1819 fausto 20 0 376M 104M 46088 S 0.7 1.3 6:58.12 /usr/libexec/Xorg vt2 -displayfd 3 -auth /run/user/1000/gdm/Xauthority -no
2038 fausto 20 0 2114M 242M 69328 S 0.0 3.0 37:14.90 /usr/bin/gnome-shell
4920 fausto 20 0 694M 62156 35744 S 0.0 0.8 0:05.72 /usr/bin/python /usr/bin/terminator
2439 fausto 20 0 3008M 161M 45536 S 0.0 2.0 1:28.89 /home/fausto/.dropbox-dist/dropbox.lnx.x86_64-19.4.13/dropbox
14071 fausto 20 0 1241M 60068 36320 S 0.0 0.7 0:36.17 /usr/bin/nautilus --application-service
26259 fausto 20 0 1628M 426M 106M S 0.0 5.4 2:41.41 /opt/google/chrome/chrome --type=renderer --enable-features=AutofillProfile
27130 fausto 20 0 985M 137M 68892 S 0.0 1.7 0:06.22 /opt/google/chrome/chrome --type=renderer --enable-features=AutofillProfile
2652 fausto 20 0 1723M 407M 125M S 0.0 5.1 6:25.96 /opt/google/chrome/chrome
2888 fausto 20 0 1623M 437M 87400 S 0.0 5.5 10:30.69 /opt/google/chrome/chrome --type=renderer --enable-features=AutofillProfile
6581 fausto 20 0 1107M 242M 183M S 0.0 3.0 0:09.21 /opt/google/chrome/chrome --type=renderer --enable-features=AutofillProfile
820 dbus 20 0 66724 6344 4100 S 0.0 0.1 1:34.75 /usr/bin/dbus-daemon --system --address=systemd: --nofork --nopidfile --sys
20702 fausto 20 0 1068M 180M 75812 S 0.0 2.3 1:47.26 /opt/google/chrome/chrome --type=renderer --enable-features=AutofillProfile
14076 fausto 20 0 1241M 60068 36320 S 0.0 0.7 0:14.80 /usr/bin/nautilus --application-service
1 root 20 0 192M 9168 5796 S 0.0 0.1 0:27.79 /usr/lib/systemd/systemd --switched-root --system --deserialize 21
607 root 20 0 145M 95036 94316 S 0.0 1.2 0:12.75 /usr/lib/systemd/systemd-journald
630 root 20 0 128M 6036 4516 S 0.0 0.1 0:00.00 /usr/sbin/lvmstat -f
640 root 20 0 49184 9084 4816 S 0.0 0.1 0:14.31 /usr/lib/systemd/systemd-udev
775 root 16 -4 55500 3448 3000 S 0.0 0.0 0:00.02 /sbin/auditd -n
765 root 16 -4 55500 3448 3000 S 0.0 0.0 0:00.76 /sbin/auditd -n
780 root 12 -8 84512 1636 1500 S 0.0 0.0 0:00.47 /sbin/audispd
F1 Help F2 Setup F3 Search F4 Filter F5 Free F6 Sort By F7 Nice F8 Force F9 Kill F10 Quit
```

# Basic commands

CMD	Action	Usage
pwd	Show current address	pwd
man	Show command manual	man chosen_command
cd	Change directory	cd directory_path
ls	List files and directories	ls directory_path
mkdir	Create a directory	mkdir folder_Name



# Basic commands

CMD	Action	Usage
pwd	Show current address	pwd
man	Show command manual	man chosen_command
cd	Change directory	cd directory_path
ls	List files and directories	ls directory_path
mkdir	Create a directory	mkdir folder_Name

Try all the commands **carefully**.

\*pwd

"/home/your\_user"

\* Attention to how the **paths** are **written**. Eg. /home/your\_username

Create a directory called "carnival". Check what is inside your new directory.

# Copying, moving and deleting

CMD	Action	Usage
cp	Copy file or folder	cp original_file copy_file
mv	Move/rename file or folder	mv file directory
rm	Remove file or directory	rm file1 rm -r directory
rmdir	Remove directory	rmdir dir1
touch	Create empty file	touch file.txt

Create three new empty files called “rain.txt”, “people.txt” and “samba.txt”

Now move rain.txt, people.txt and samba.txt to the carnival!

```
mv rain.txt people.txt samba.txt carnival
```

# Copying, moving and deleting

cp	Copy file or folder	cp original_file copy_file
mv	Move/rename file or folder	mv file directory
rm	Remove file or directory	rm file1 rm -r directory
rmdir	Remove directory	rmdir dir1
touch	Create empty file	touch file.txt

Check what is inside the carnival with the command “ls”.

The rain is ruining it. Remove the rain from the carnaval.  
See how this is fun?

Now, lets practice with some biological data:  
Change the name of the “**carnival**” directory to “**data**” and remove the samba and the people.

Use the command *git clone* to download full github projects, and download the FSK2053 project from github:

```
git clone https://github.com/bioinfo-arctic/FSK2053.git
```

# Copying, moving and deleting

cp	Copy file or folder	cp original_file copy_file
<b>mv</b>	Move/rename file or folder	<b>mv file directory/</b>
rm	Remove file or directory	rm file1 rm -r directory
rmdir	Remove directory	rmdir dir1
touch	Create empty file	touch file.txt

Check what is inside the directory FSK2053/ with the command “ls”.

You will find a data/ directory inside the Spring\_2023/ directory. Get the sequencing data from this directory:

*A4\_A006\_R1\_FSK2053.fastq*

*A4\_A006\_R2\_FSK2053.fastq*

Move them to your data/ (/home/fsk2053/data) directory, outside of the FSK2053/ (/home/FSK2053/Spring\_2053/data) directory.

# File exhibition

cat	Exhibit and contatenate	cat file.txt
<b>less</b>	Read a file	<b>less file.txt</b>
more	Read a file	more file.txt
wc	Count the number of lines, characters and bytes of a file	wc -l file.txt [lines of the file] ls   wc -l [how many files are there]
head	First 21 lines of the file	head -n 21 file.txt
tail	Last 15 lines of the file	tail -n 15 file.txt
sort	Order the lines of the file by the user's definition.	sort names > names.sorted

Check the sequences inside the data directory “/home/your\_user/data” using the command *less*.

Use the arrow keys in your keyboard to navigate through this file and the key “q” to exit the *less* command.



# File exhibition

cat	Exhibit and concatenate	cat file.txt
less	Read a file	less file.txt
more	Read a file	more file.txt
wc	Count the number of lines, characters and bytes of a file	wc -l file.txt [lines of the file] ls   wc -l [how many files are there]
<b>head</b>	First 21 lines of the file	<b>head -n 21 file.txt</b>
tail	Last 15 lines of the file	tail -n 15 file.txt
sort	Order the lines of the file by the user's definition.	sort names > names.sorted

Count the lines of your fastq files.

Take a look in the first 16 lines of each fastq file.

Take a look in the last 4 lines of each fastq file.

# File manipulation

>	Forward standard output	<b>Your command file &gt; report.txt</b>
<	Modify the standard input to a file	Your command < gene.fasta
	Allow the combination of commands	<b>ls   wc -l</b>
>>	Adds the output to the end of the file.	programa.pl >> relatorios

Now we are going to generate a sample of our data set. Take the 16 first lines of the file "A4\_A006\_R1\_FSK2053.fastq" and forward to a new file called R1.fastq in your work directory. If the file doesn't exist, the forward sign will create it automatically.

Do the same with the other file and generate R2.fastq inside the work directory.

Use the pipe sign "|" to combine the command ls and wc and check how many files are inside the work directory

# Text manipulation

cat	Concatenate and exhibit	<code>cat text1 text2 &gt; text1text2</code>
<b>grep</b>	Search the file line by line for defined expressions	<b><code>grep "&gt;" genes.fasta</code></b>
uniq	Remove duplicated lines	<code>sort alfa   uniq -c</code>
cut	Cut input files. Ideal for tables.	<code>cut -d " " f1 alfa</code>
awk	Programming language for text manipulation	<code>awk -F '{print \$2 \$1}' table.csv</code>
sed	Used to manipulate and transform text.	<code>sed 's/t/u/g' dna.seq &gt; rna.seq</code>

Lets inspect our sequences.

Use grep to search for the a pattern of your interest in your new fastq files. For example something from their header. “@M”

With this command you can compare if the paired reads are still paired.

Choose a pattern, for example the start of the read name “@M” and use grep to count how many times it show up in your

# Text manipulation

cat	Concatenate and exhibit	<code>cat text1 text2 &gt; text1text2</code>
grep	Search the file line by line for defined expressions	<code>grep "&gt;" genes.fasta</code>
uniq	Remove duplicated lines	<code>sort alfa   uniq -c</code>
cut	Cut input files. Ideal for tables.	<code>cut -d " " f1 alfa</code>
awk	Programming language for text manipulation	<code>awk -F '{print \$2 \$1}' table.csv</code>
sed	Used to manipulate and transform text.	<code>sed 's/t/u/g' dna.seq &gt; rna.seq</code>

Now to finish, lets see how many commands we learned, using the command history and forwarding the output to the file commands.txt

```
history > commands.txt
```

To inspect this file, use the commands:

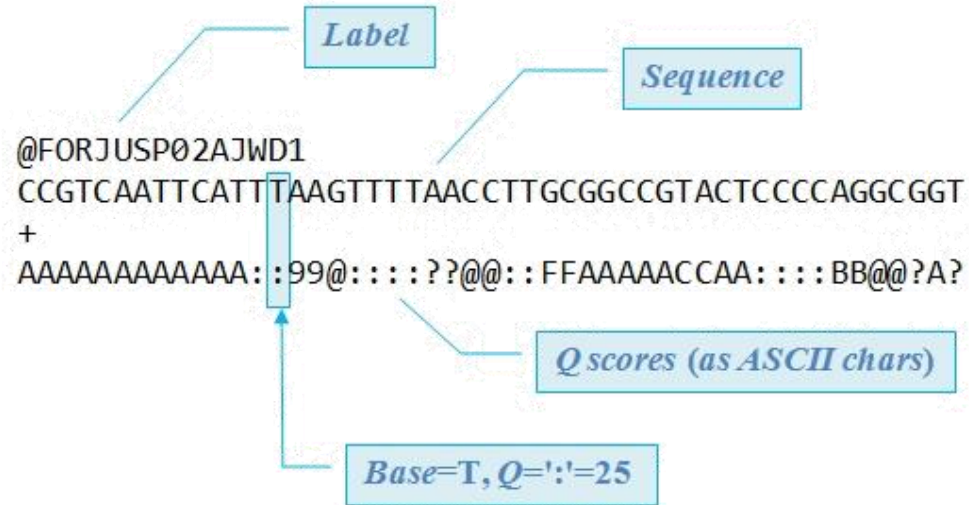
```
head commands.txt
```

```
cut -f 7 -d " " commands.txt | sort | uniq | wc -l #what does it show you?
```

What if you want to count how many times you used each one?

# Fastq files and phred score

Quality Score  
 $P = 10^{-Q/10}$   
 $Q = -10 \log_{10}(P)$



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			



# Fastq file header

```
@M02149:53:000000000-AANLH:1:1101:14924:1701 1:N:0:0
TACGGAGGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACGTA
GGCTGTCTGGTAAGTCAGGGGTGAAATCCCGCGGCTCACCCGCGGAATT
GCCCTTGATACTGCTGGACTTGAGTTCGGGAGAGGGTGGCGGAATTCCAG
GTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATCAGGGGGCGAAGG
CGGCCACCTGGACCGATACTGACGCTGAGGTGCGAAAGCGTGGGGGAGGA
AACAGG
```

+

```
AAA??1>DDAAA11AFEGF00BGCEA0F1A1F10AAAFAB//BAAA/AAB00ABGFF
@F10BB@DGG2B00/B//1@BF1F/>>>EEA<1B</<>///?F?DD<FGF>??<F1<F<
??<FGHF?G<?CHHHHHFF<::/0GHFB;:BFF0F;<1GG>BF2HHEB//?F@HGB@
B110FFHFHGB1B0FB>/EE>HGFEEAA0/1A011EEBA/2D2D/AEEABB1FHE00
AAGFFEA1A1GGFFFB3@F>1AAA
```

Illumina header

```
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>
<read>:<is filtered>:<control number>:<sample number>
```

# Fasta file

```
>M02149:53:000000000-AANLH:1:1101:14924:1701 1:N:0:0  
TACGGAGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACG  
TAGGCTGTCTGGTAAGTCAGGGGGTGAAATCCCGCGGGCTCACCCGCGGA  
ATTGCCCTTGATACTGCTGGACTTGAGTTCGGGAGAGGGGTGGCGGAAT  
TCCAGGTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATCAGGGG  
CGAAGGCGGGCCACCTGGACCGATACTGACGCTGAGGTGCGAAAGCGT  
GGGGAGGAAACAGG
```

# Quality check

## FastQC Report

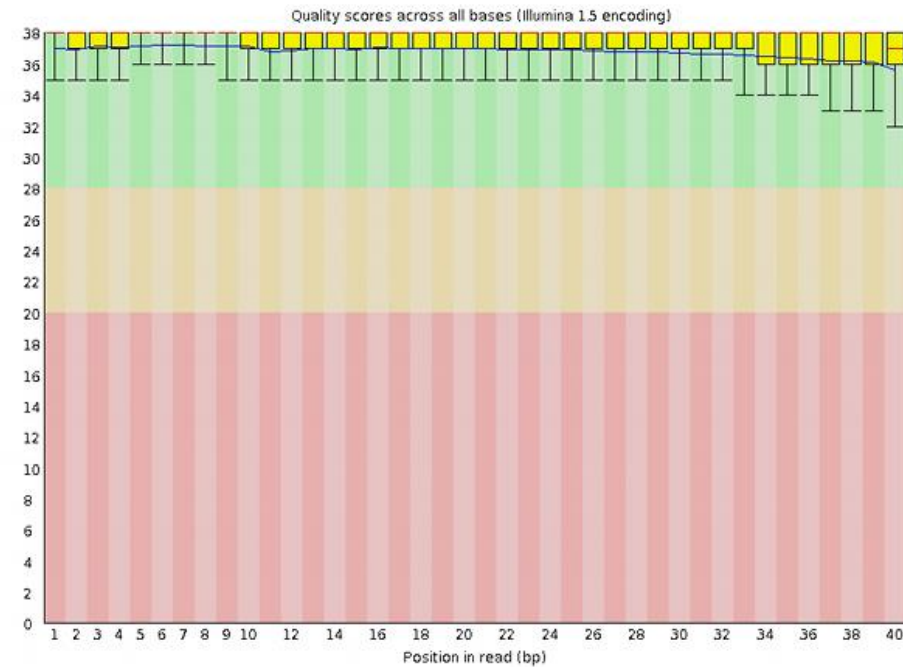
### Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ! [Per base sequence content](#)
- ✓ [Per base GC content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)
- ! [Kmer Content](#)

### ✓ Basic Statistics

Measure	Value
Filename	good_sequence_short.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Filtered Sequences	0
Sequence length	40
%GC	45

### ✓ Per base sequence quality



# “Reality” check

