UiT The Arctic University of Norway

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

What is Linux?

Linux is a family of free and open sorce 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 top500 biggest supercomputers, stock exchange servers and of course desktop computers.

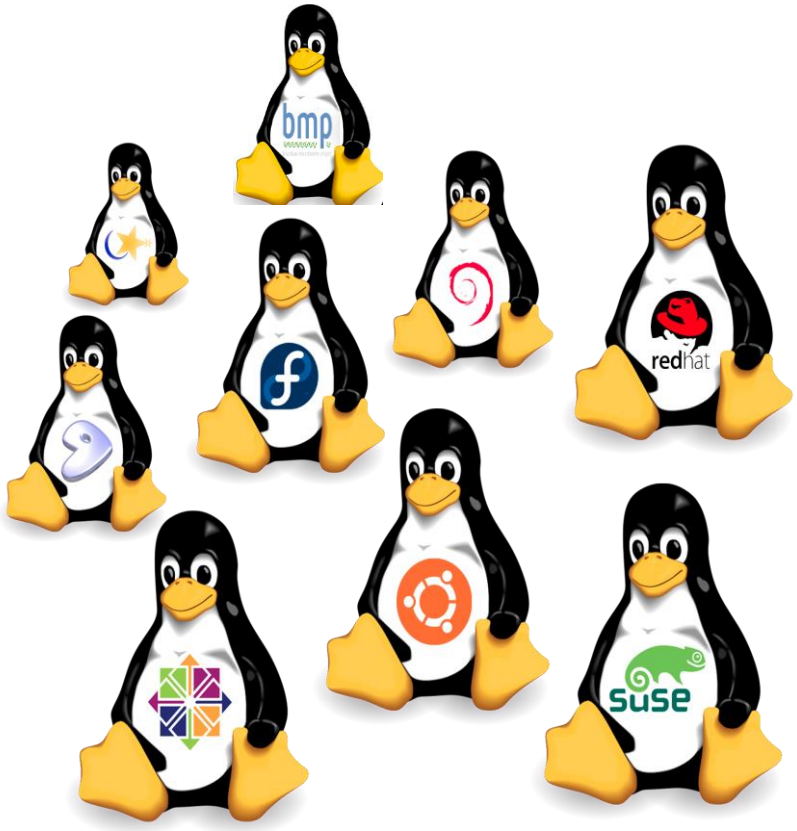
It got that big, thanks to voluntary colaborations around the world.

Linus Torvalds - https://www.ted.com/talks/linus_torvalds_the_mind_behind_linux

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



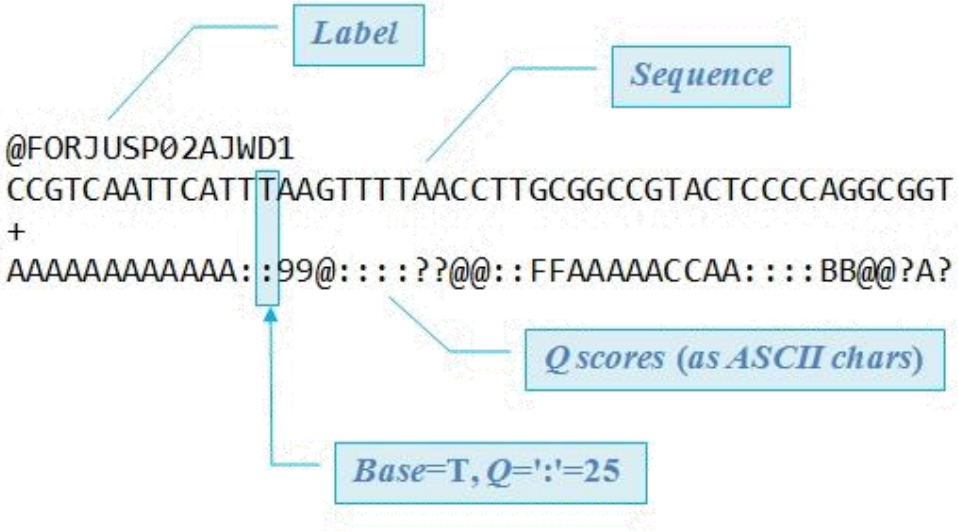
Types biological data we have now

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
TACGGAGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACGTA
GGCTGTCTGGTAAGTCAGGGGTGAAATCCCGCGGCTCACCCGCGGAATT
GCCCTTGATACTGCTGGACTTGAGTTCGGGAGAGGGTGGCGGAATTCCAG
GTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATCAGGGGCGAAGG
CGGCCACCTGGACCGATACTGACGCTGAGGTGCGAAAGCGTGGGGAGGA
AACAGG

+

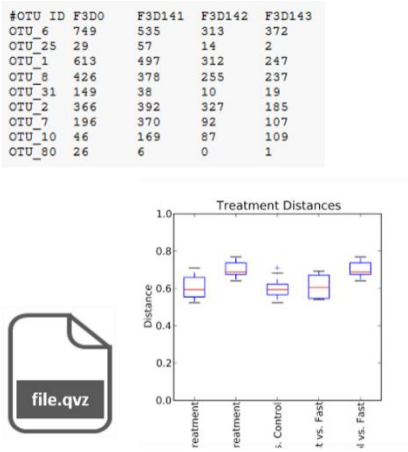
AAA??1>DDAAA11AFEGF00BGCEA0F1A1F10AAAFA//BAAA/AAB00ABGFF
@F10BB@DGG2B00/B//1@BF1F/>>>EEA<1B</<>///?F?DD<FGF>??<F1<F<
??<FGHF?G<?CHHHHHFF<::/0GHFB;:BFF0F;<1GG>BF2HHEB//?F@HGB@
B110FFHFHGB1B0FB>/EE>HGFEEAA0/1A011EEBA/2D2D/AEEABB1FHE00
AAGFFEA1A1GGFFFB3@F>1AAA

Illumina header

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

http://support.illumina.com/content.htm

Lets use these tools to understand and manipulate our sequences



Microb Ecol
DOI 10.1007/s00248-013-0302-4

SHORT COMMENTARY

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

Microb Ecol. 2014 Feb;67(2):237-41

- **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/
- 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!!!**

Microb Ecol
DOI 10.1007/s00248-016-0785-x

METHODS

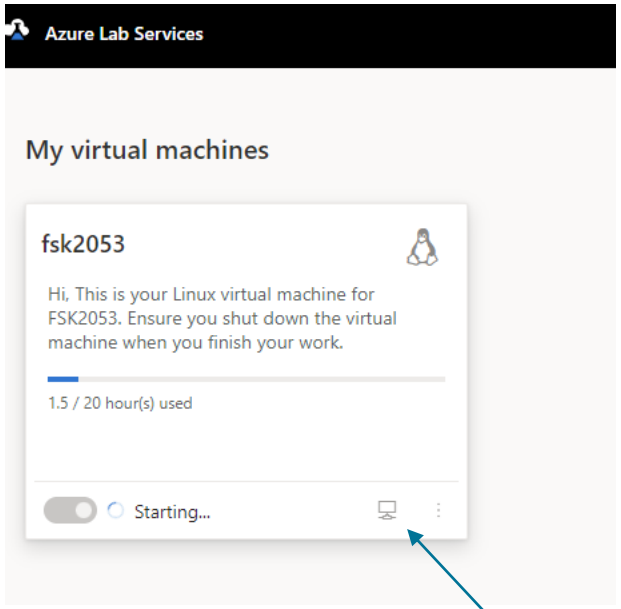
BMPOS: a Flexible and User-Friendly Tool Sets for Microbiome Studies

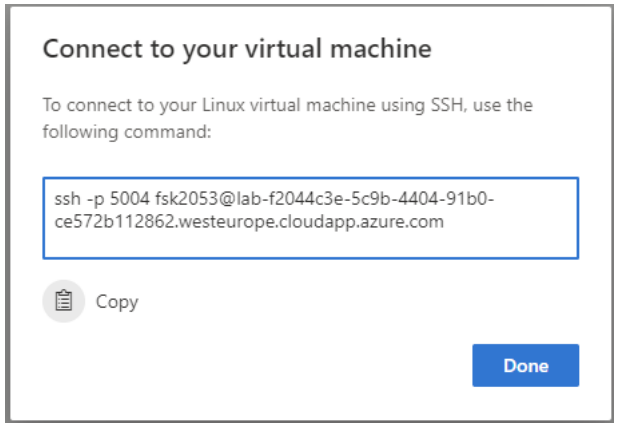
Victor S. Pylro[1] • Daniel K. Morais[1] • Francislon S. de Oliveira[1] • Fausto G. dos Santos[1] • Leandro N. Lemos[2] • Guilherme Oliveira[3] • Luiz F. W. Roesch[4]

The Linux-based BMP Operating System

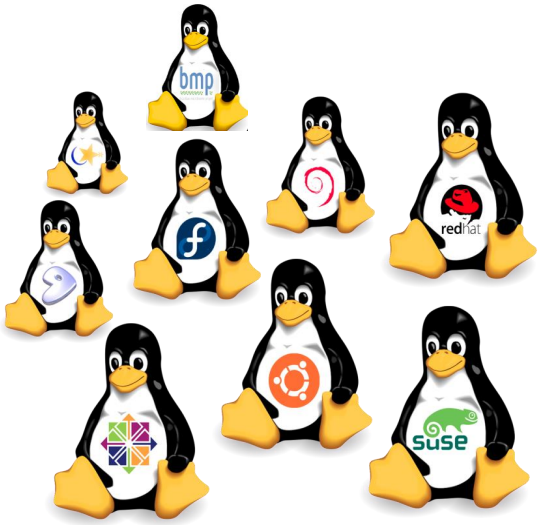
Linux Connection

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



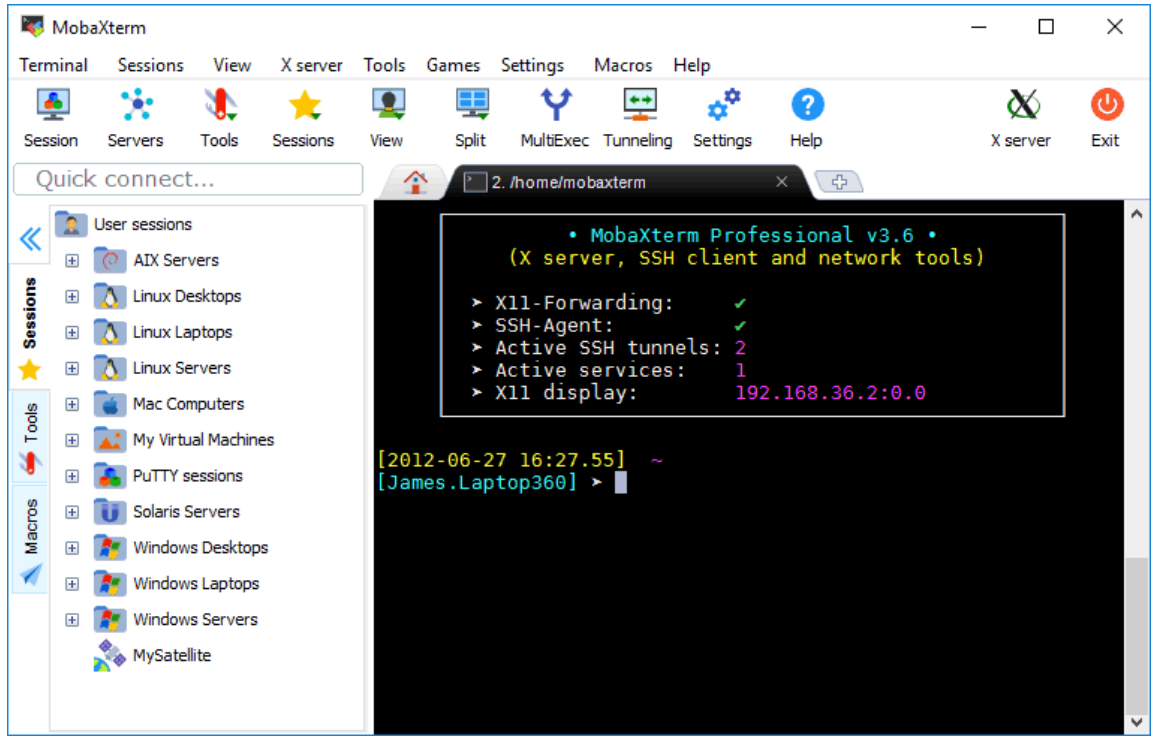


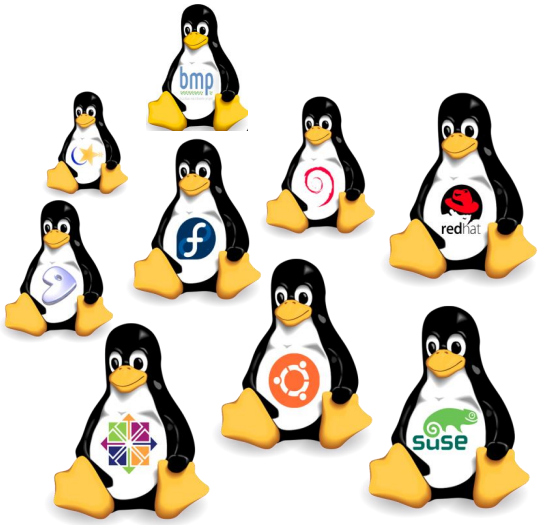
The connection command to your machines



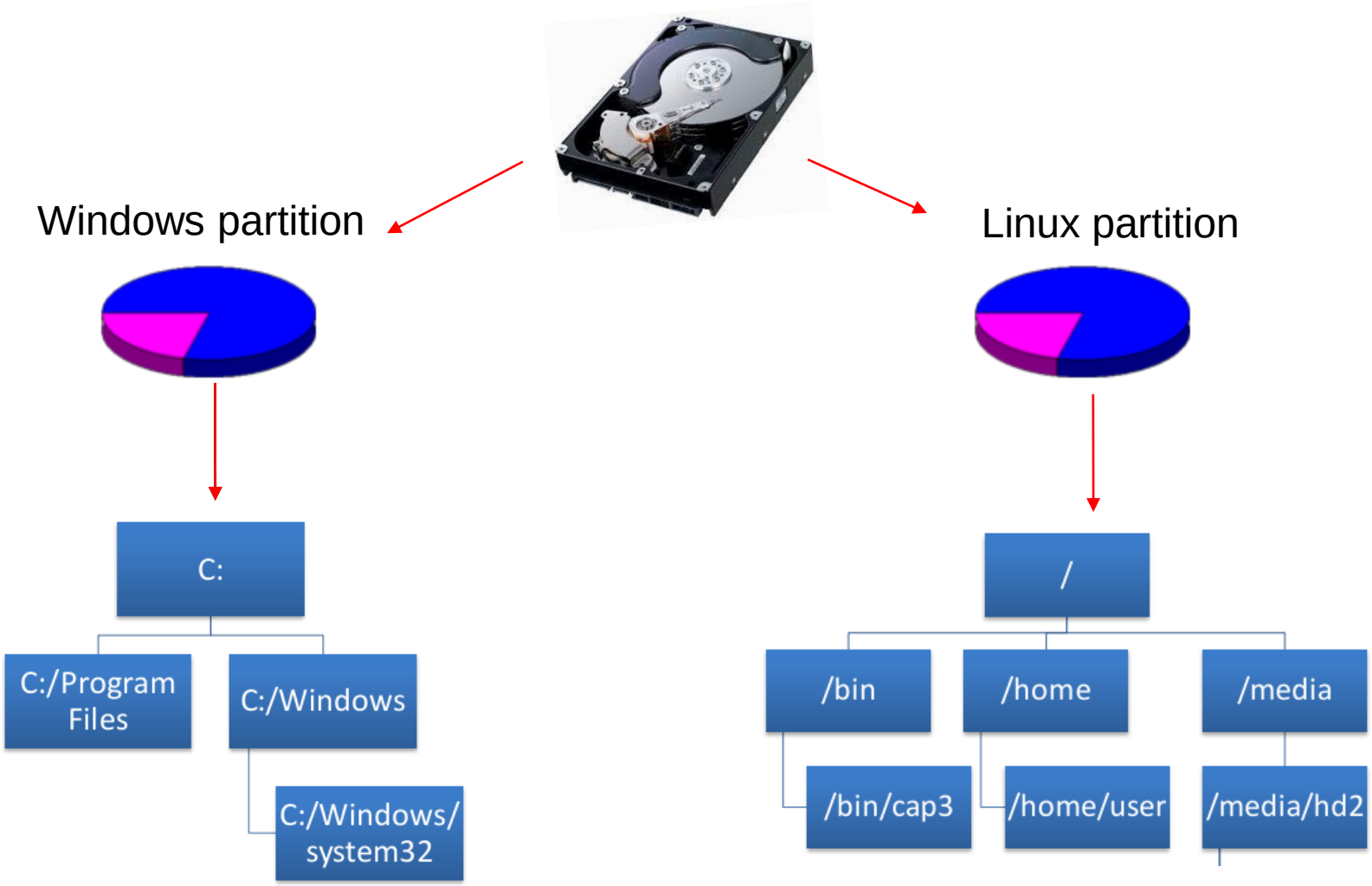
Linux Connection

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

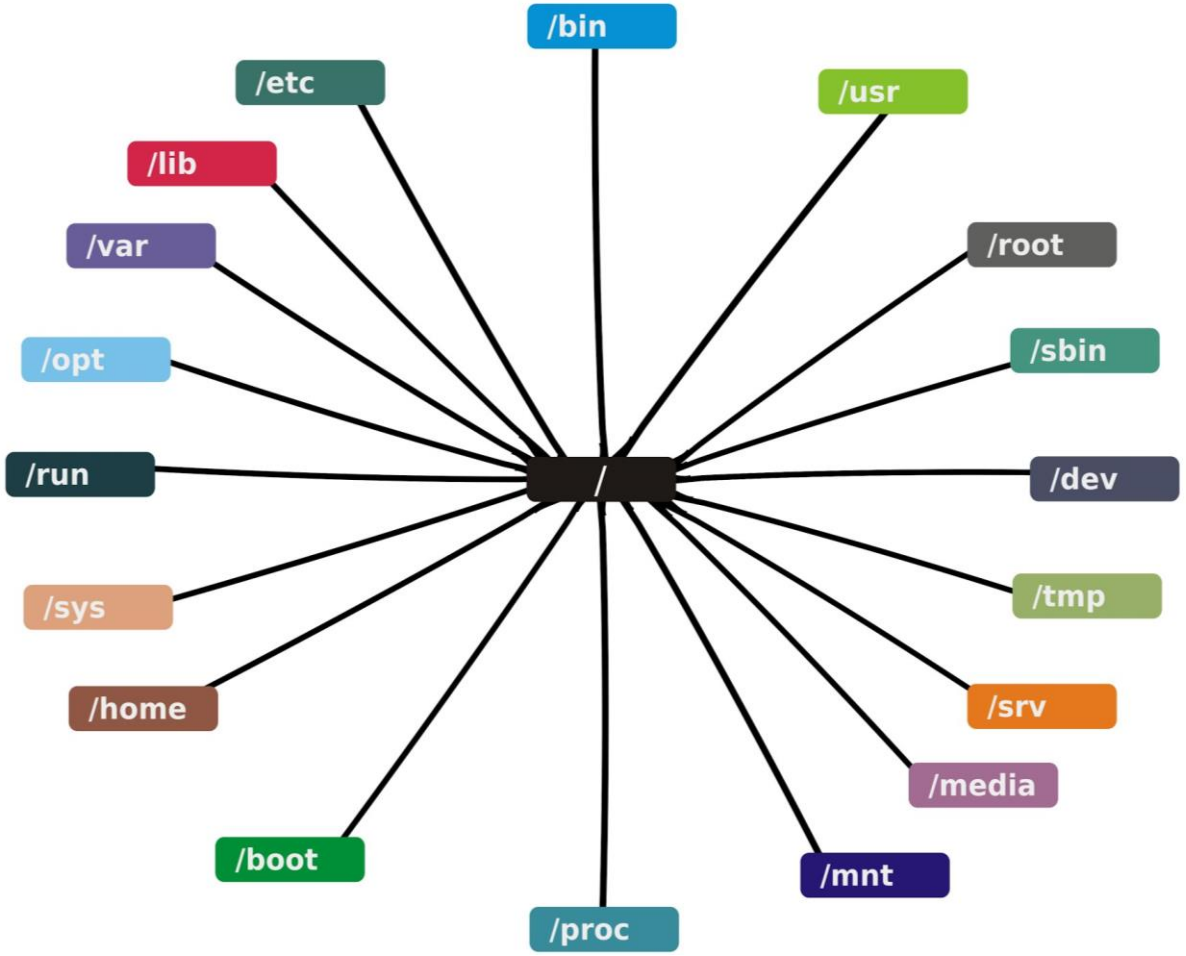




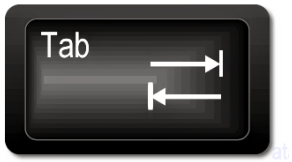
Directory organization system



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

Terminal multiplexers

Screen
https://help.ubuntu.com/community/Screen
Tmux
https://manpages.ubuntu.com/manpages/bionic/man1/tmux.1.html

Basic Linux terminal usage concepts

$

User mode and *variable calls*.

```
echo $HOME
```

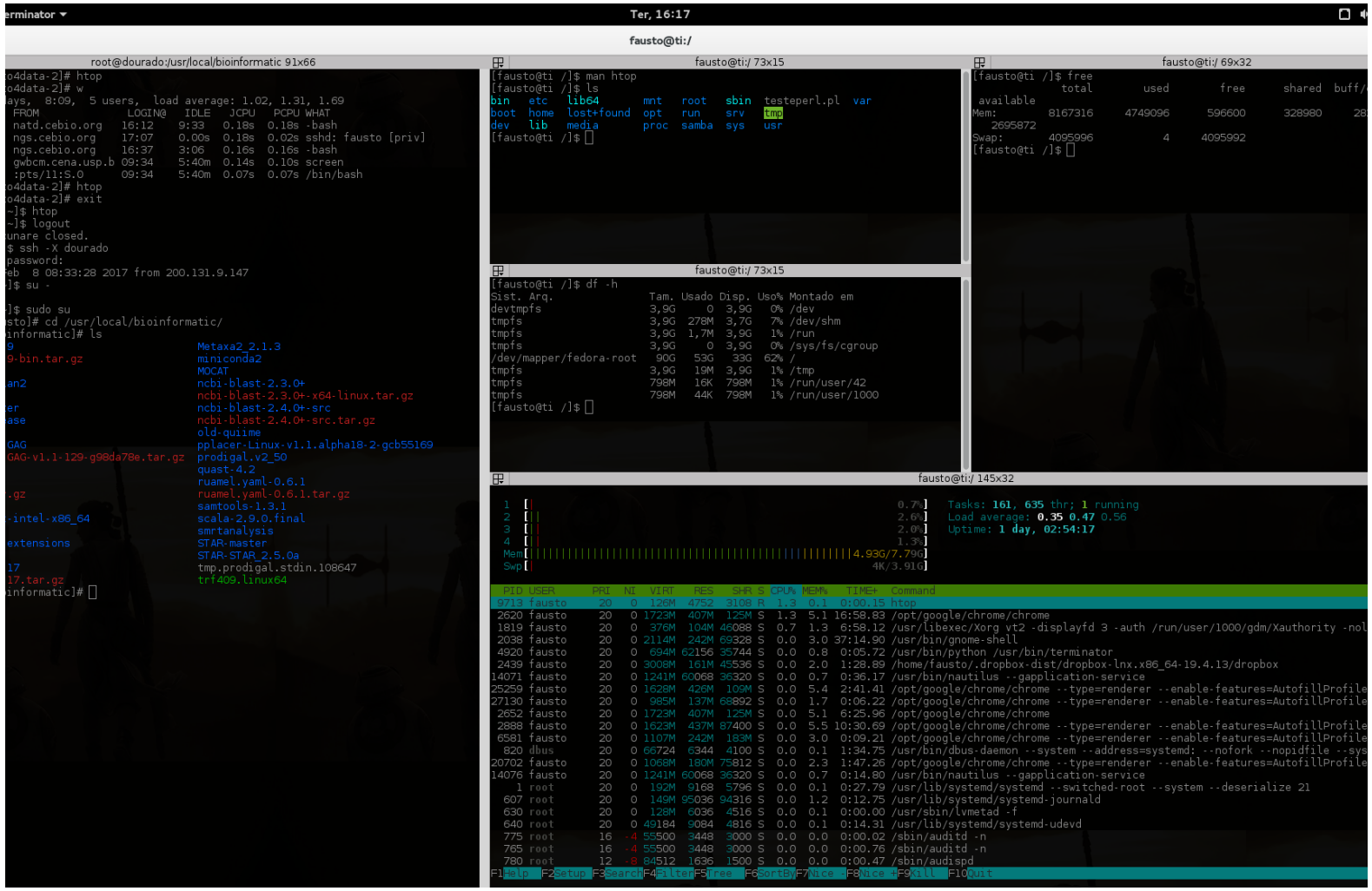
```
fsk2053@lab000000:/home/adminfsk2053$ pwd
/home/adminfsk2053
fsk2053@lab000000:/home/adminfsk2053$ ls ./*
./data:
fasta_test1.fasta  results2.txt  unknown1.txt

./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.nto
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


Ter, 16:17
fausto@ti:/
root@dourado:/usr/local/bioinformatic 91x66
fausto@ti:/ 73x15
fausto@ti:/ 69x32
fausto@ti:/ 145x32

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
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 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 exibition

cat	Exibit and contatenate	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]
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 exibition

cat	Exibit and contatenate	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]
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

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

>	Foward standar output	**Your command file > report.txt**
<	Modify the standar input to a file	Your command < gene.fasta
|	Allow the combination of commands	**ls | wc -l**
>>	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 exibit	cat text1 text2 > text1text2
grep	Search the file line by line for defined expressions	**grep ">" genes.fasta**
uniq	Remove duplicated lines	sort alfa | uniq -c
cut	Cut input files. Ideal for tables.	cut -d " " f1 alfa
awk	Programming language for text manipulation	awk -F '{print $2 $1}' table.csv
sed	Used to manipulate and transform text.	sed 's/t/u/g' dna.seq > rna.seq

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

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?
```

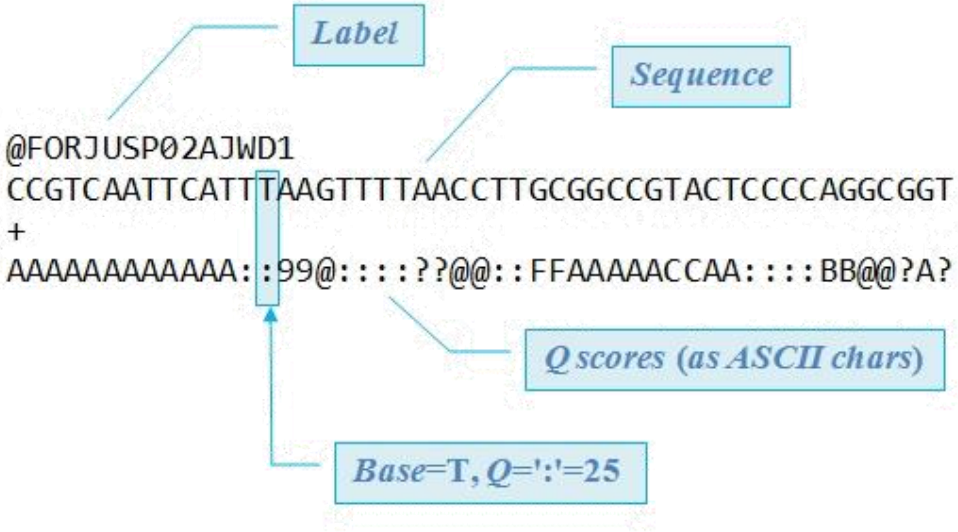
What if you want to count how many times you used each one?

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
TACGGAGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACGTA
GGCTGTCTGGTAAGTCAGGGGTGAAATCCCGCGGCTCACCCGCGGAATT
GCCCTTGATACTGCTGGACTTGAGTTCGGGAGAGGGTGGCGGAATTCCAG
GTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATCAGGGGCGAAGG
CGGCCACCTGGACCGATACTGACGCTGAGGTGCGAAAGCGTGGGGAGGA
AACAGG

+

AAA??1>DDAAA11AFEGF00BGCEA0F1A1F10AAAFA//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>

http://support.illumina.com/content.htm

Fasta file

>M02149:53:000000000-AANLH:1:1101:14924:1701 1:N:0:0
TACGGAGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACG
TAGGCTGTCTGGTAAGTCAGGGGTGAAATCCCGCGGCTCACCCGCGGA
ATTGCCCTTGATACTGCTGGACTTGAGTTCGGGAGAGGGTGGCGGAAT
TCCAGGTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATCAGGGG
CGAAGGCGGCCACCTGGACCGATACTGACGCTGAGGTGCGAAAGCGT
GGGGAGGAAACAGG

Quality check

FastQC Report

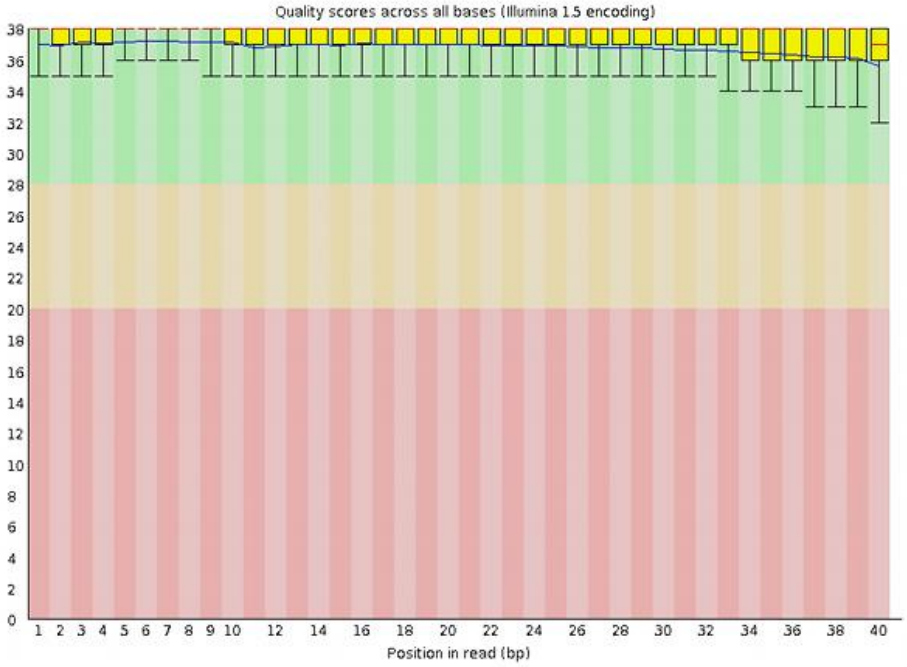
Summary

- 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