Genômica Computacional

Linux e processamento de dados de sequenciamento

Professor: Ricardo A. Vialle

CS31 - Genômica Computacional [11,18,25/10 e 1,8/11/23 - 12h00-14h00 - 4ªs. feiras]

18 de Outubro de 2023

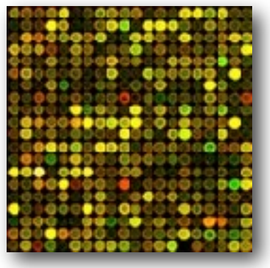
Cronograma

Data	Tema
~~11-Oct~~	~~Introdução a Genômica, Sequenciamento (teórica)~~
18-Oct	Bioinformática - Linux - Processamento de dados de sequenciamento (teórico-prática)
1-Nov	Montagem de genomas (teórico-prática)
8-Nov	Anotação de genomas (teórico-prática)
22-Nov	Analise de variabilidade genética (teórico-prática)

Genomics technology

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

Since ~2007

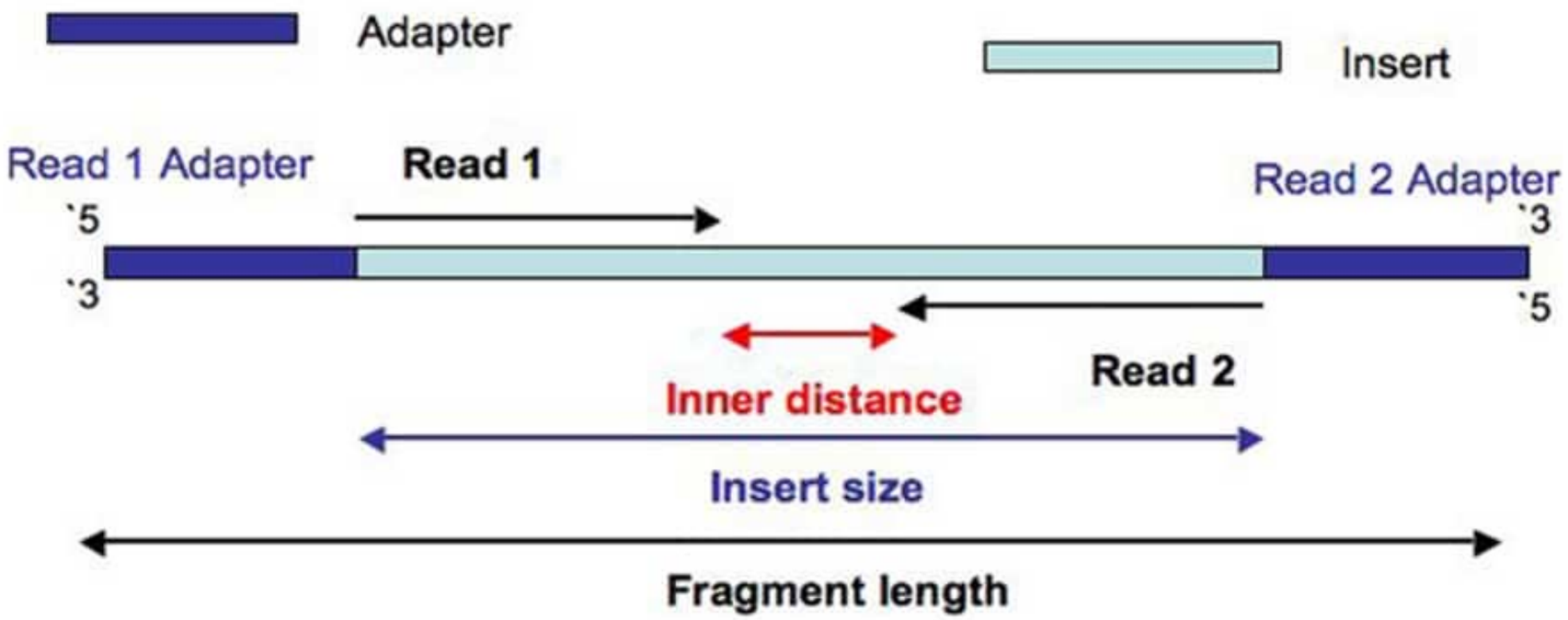
3rd-generation & single-molecule DNA sequencing

Since ~2010

Solexa

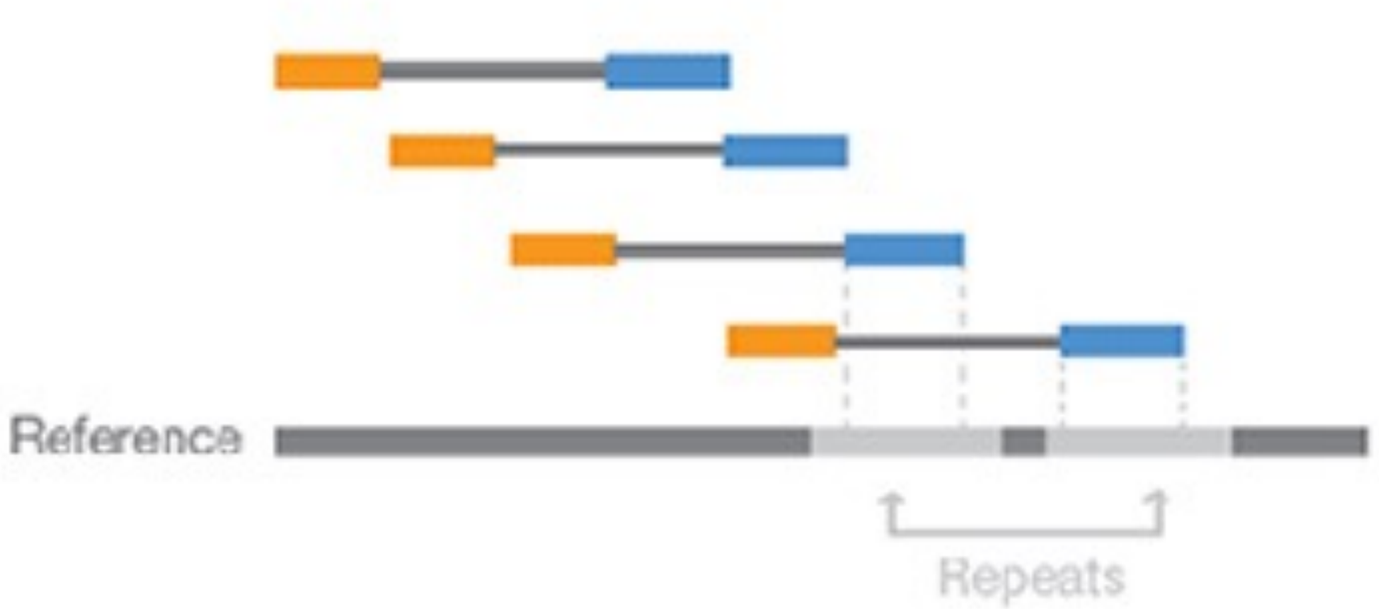
illumina®

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf



https://thesequencingcenter.com/knowledge-base/what-are-paired-end-reads/

Alignment to the Reference Sequence



https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

A read in FASTQ format

Name	@ERR194146.1 HSQ1008:141:D0CC8ACXX:3:1308:20201:3607 2:Y:18:ATCACG
Sequence	ACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAAT
(ignore)	+
Base qualities	?@@FFBFFDDHHBCEAFGEGIDHGH@GDHHHGEHID@C?GGDG@FHIGGH@FHBEGG

Always starts with “@”
ERR194146.1HSQ1008:141:D0CC8ACXX:3 – Machine, Run, Flowcell, Lane
1308:20201:3607 – Tile, X-pos, Y-pos
2:Y:18:ATCACG – Direction, Filtered?, Control bits, Index/Sample

Base qualities

Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
||||||||||||||||||||||||||||||||
IHHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

Base quality is ASCII-encoded version of $Q = -10 \log_{10} p$

Base quality

Probability that base call is incorrect

$Q = 10 \rightarrow$ 1 in 10 chance call is incorrect

$Q = 20 \rightarrow$ 1 in 100

$Q = 30 \rightarrow$ 1 in 1,000

Slide from: Bem Langmead Lab
https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf

Value	Character	Value	Character	Value	Character	Value	Character	Value	Character	Value	Character
0	Null	22	Synchronous Idle	44	,	66	B	88	X	110	n
1	Start of Heading	23	End of Transmission Block	45	-	67	C	89	Y	111	o
2	Start of Text	24	Cancel	46	.	68	D	90	Z	112	p
3	End of Text	25	End of Medium	47	/	69	E	91	[	113	q
4	End of Transmission	26	Substitute	48	0	70	F	92	\	114	r
5	Enquiry	27	Escape	49	1	71	G	93	]	115	s
6	Acknowledgement	28	File Separator	50	2	72	H	94	^	116	t
7	Bell (Causes an alert sound)	29	Group Separator	51	3	73	I	95	_	117	u
8	Backspace	30	Record Separator	52	4	74	J	96	@	118	v
9	Horizontal Tab	31	Unit Separator	53	5	75	K	97	a	119	w
10	Line Feed	32	space	54	6	76	L	98	b	120	x
11	Vertical Tab	33	!	55	7	77	M	99	c	121	y
12	Form Feed	34	"	56	8	78	N	100	d	122	z
13	Carriage Return	35	#	57	9	79	O	101	e	123	{

Base 33 (typical)

Bases and qualities line up:

```
AGCTCTGGTGACCCATGGGCAGCTGCTAGGGA
|||||||||||||||||||||||||||||||||
IHHHHHHHHHHHHHGCGC5FEFFFGHHHHHH
```

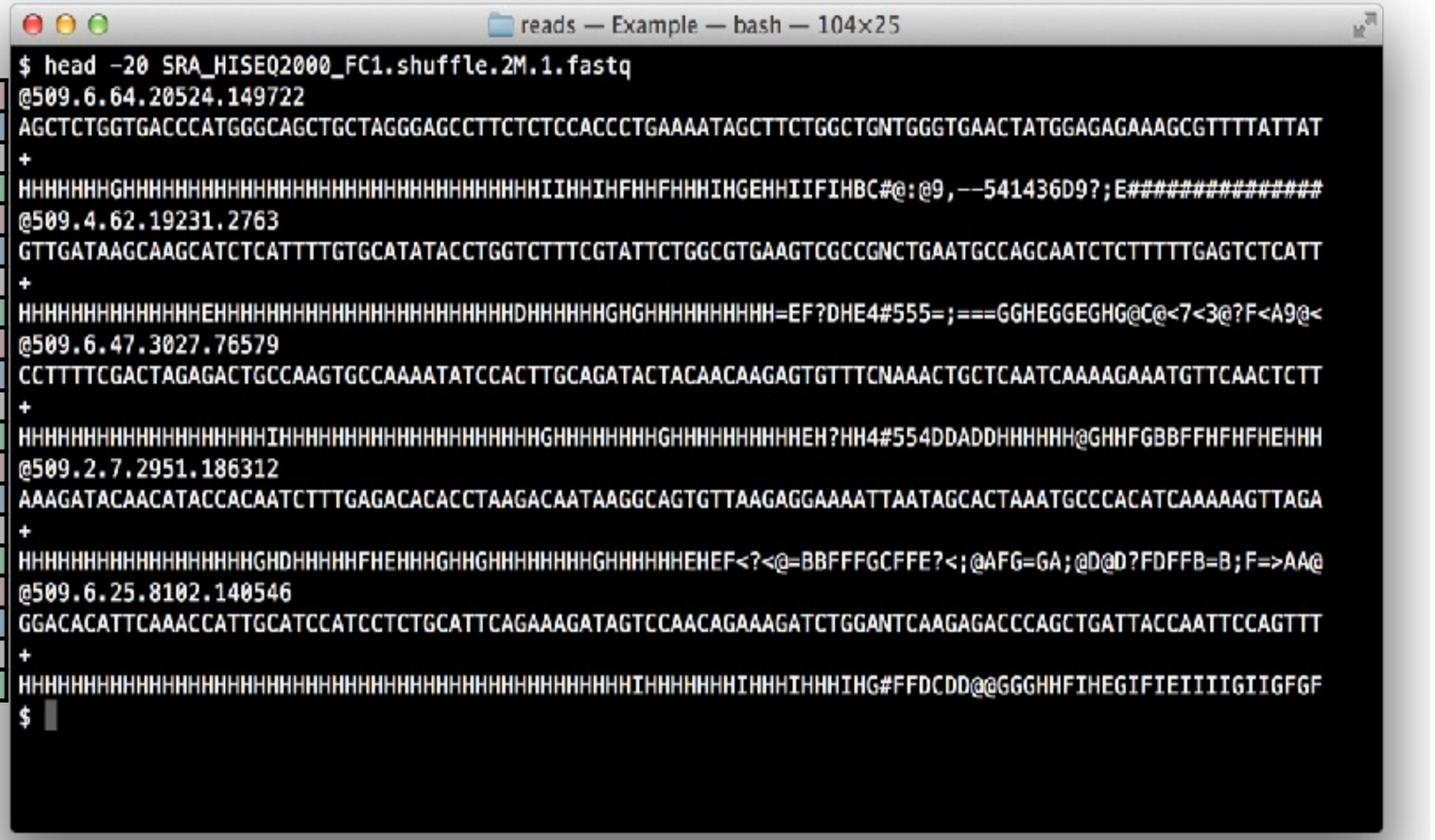
```
Value - base = Phred
  73  -  33  = 40
```

$Q = -10 \log_{10} p$

Phred	Error Probability	Confidence
0	1 / 1	0%
10	1 / 10	90%
20	1 / 100	99%
30	1 / 1000	99.9%
40	1 / 10000	99.99%
50	1 / 100000	99.999%
60	1 / 1000000	99.9999%

FASTQ

Read 1	Name
	Sequence
	(placeholder)
	Base qualities
Read 2	Name
	Sequence
	(placeholder)
	Base qualities
Read 3	Name
	Sequence
	(placeholder)
	Base qualities
Read 4	Name
	Sequence
	(placeholder)
	Base qualities
Read 5	Name
	Sequence
	(placeholder)
	Base qualities

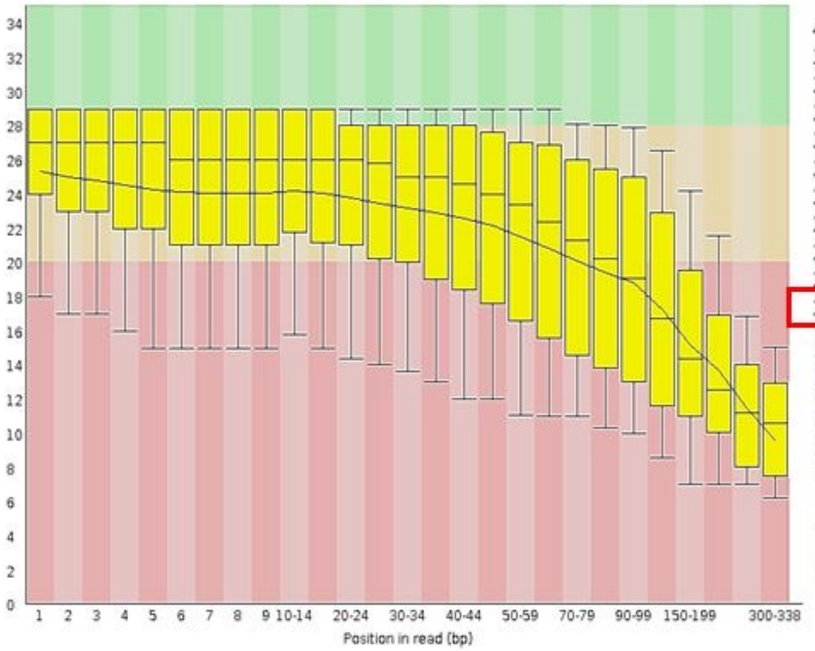


Sample_S1_L001_R1_001.fastq.gz

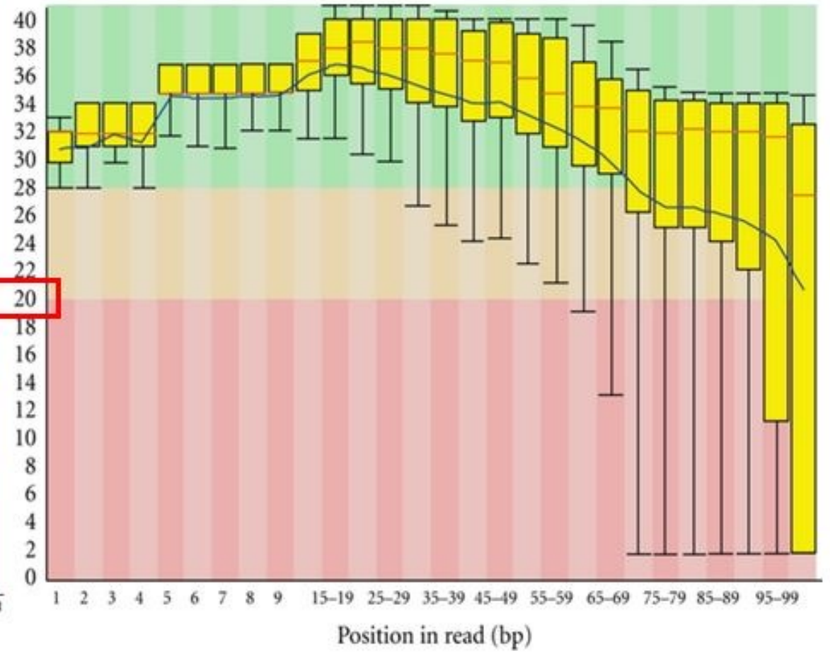
Sample: Sample name
S1: Sample number
L001: Lane number
R1: Read direction (If starts with “I”, Index read direction)
001: File number (always 001 on modern systems)

Comparison in sequencing quality

Ion Torrent PGM



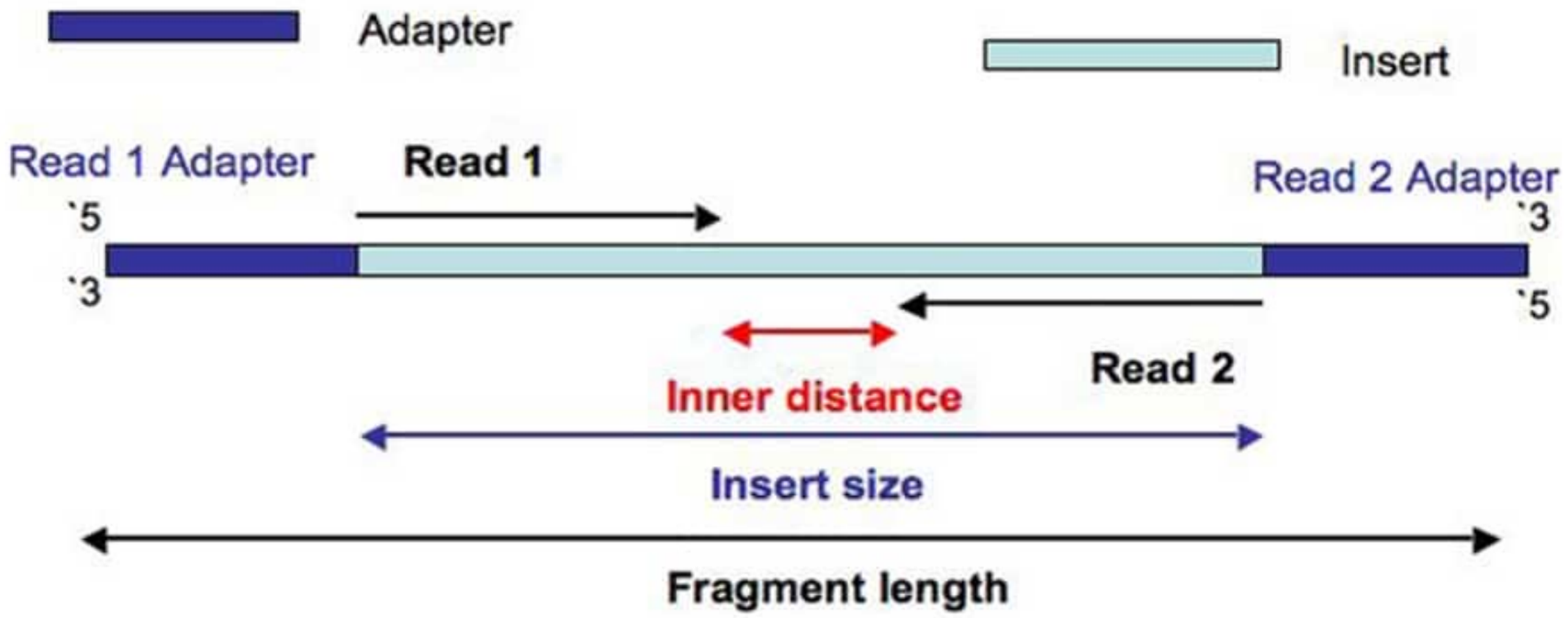
Illumina Hiseq 2000



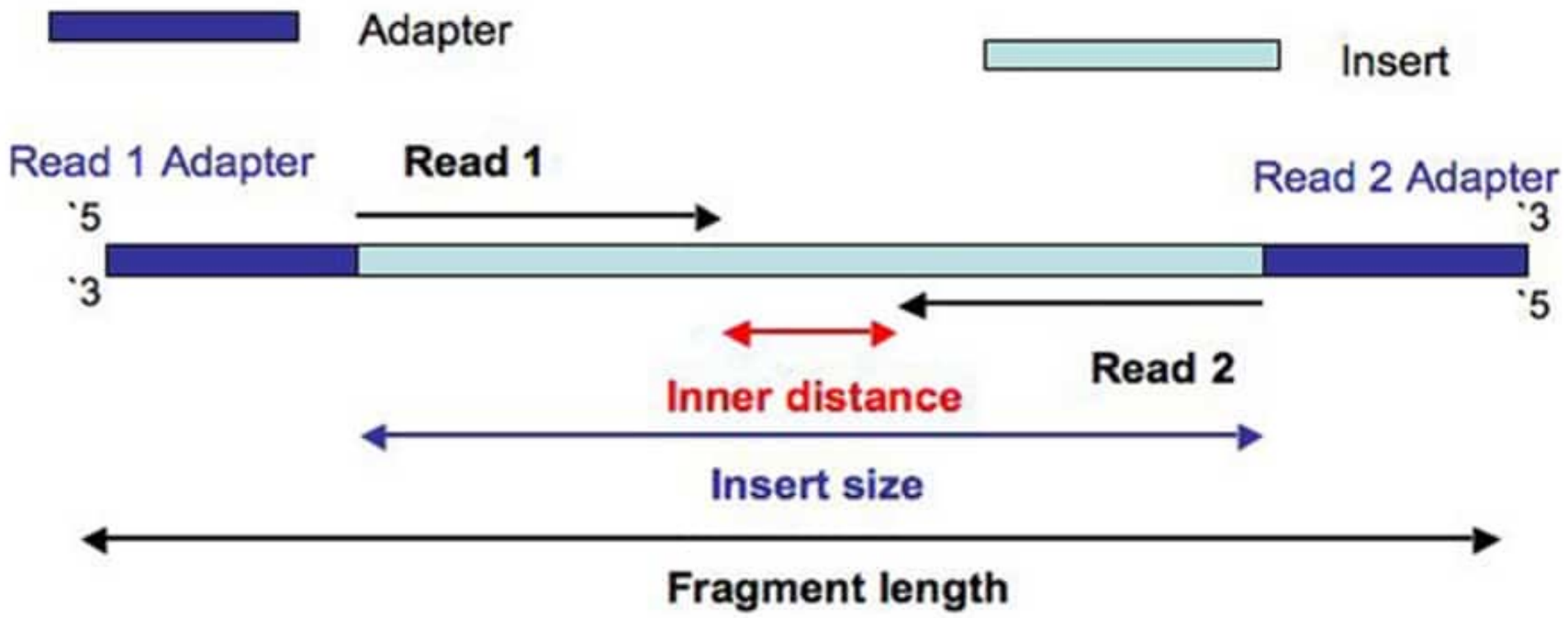
Per base sequence quality of samples generated by FASTQC. The yellow box show the base-calling quality scores across all sequencing reads. The blue line indicates the mean quality score. Q20=99% accuracy. Q30=99.9% accuracy…

https://www.hadriengourle.com/tutorials/qc/

- Trimming
  - Barcode & adapter sequences



- Trimming
  - Barcode & adapter sequences



  - Poor quality sequence at the starts/ends of reads

Which trimming threshold? Examples

RNAseq

- Gentle trimming
- Q > 5 should be enough
- Too aggressive trimming => losing part of the dataset

SNP calling

- You need to be sure of the bases
- Q > 20

MacManes, M. D. (2014). On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in genetics*, *5*, 13.

https://www.hadriengourle.com/tutorials/qc/

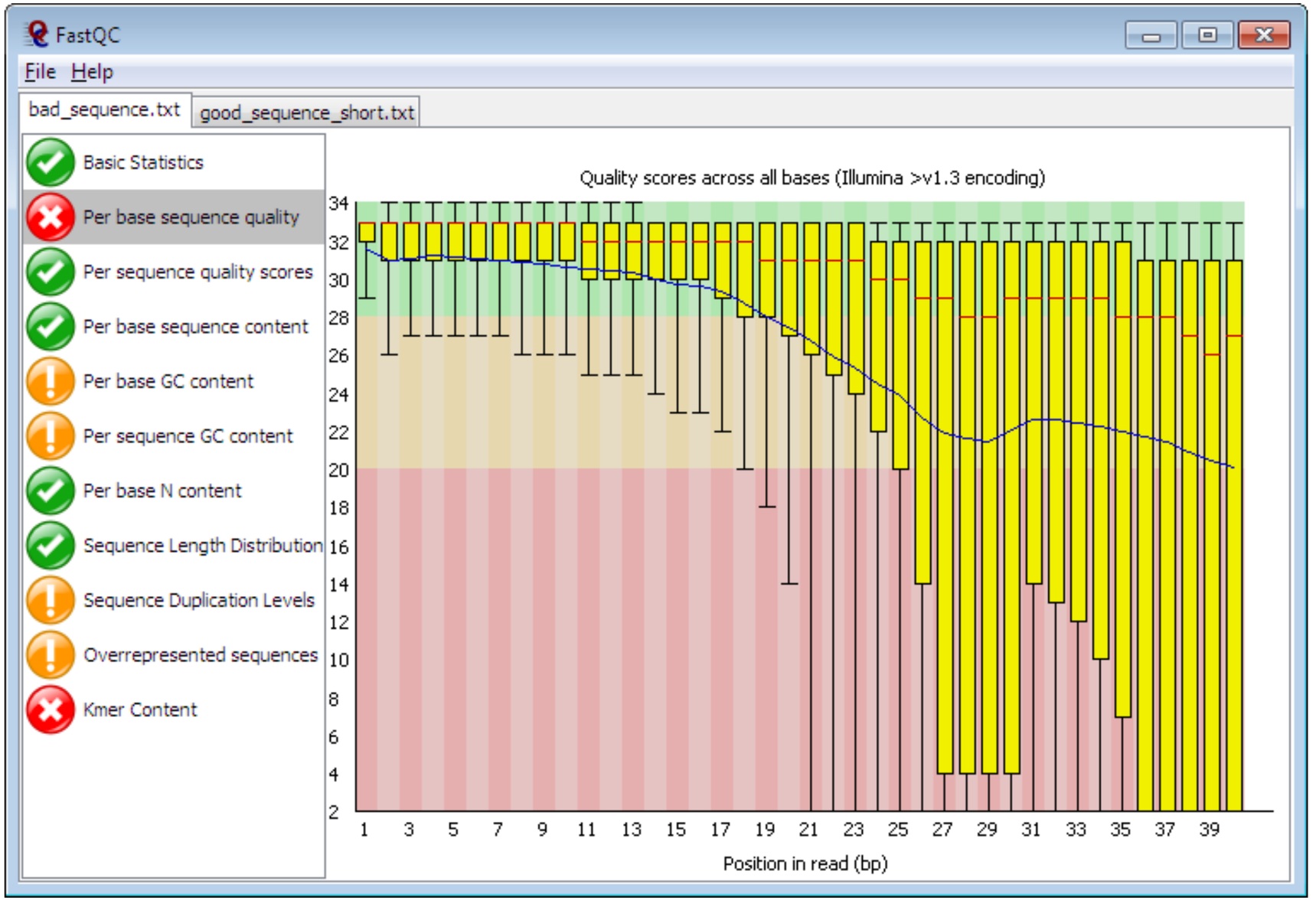
Adapter trimmer [Scythe]

Uses the quality information in a FASTQ entry and a prior to decide whether a 3' substring is adapter.

Very basically, it takes a naïve Bayesian approach to classifying 3'-end contaminants only. Because these are the most poor quality bases and most likely to be contaminated (especially as reads get longer and longer), Scythe is designed to specifically remove these contaminants.

Low quality trimmer [Sickle]

Sickle is a sliding window quality trimmer, designed to be used after Scythe. Unlike *cutadapt* and other tools, this pipeline remove adapter contaminants before quality trimming, as removing poor quality bases throws away any useful information that could be used in identifying a 3'-end adapter contaminant.



FastQC

[See each report detail here]

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

MultiQC
v1.3

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2017-11-19, 21:42 based on data in: `/Users/hadrien/Documents/workspace/ar`

General Statistics

Copy table | Configure Columns | Plot | Showing 4/4 rows and 4/5 columns.

Sample Name	% Dups	% GC	Length	M Seqs
SRR957824_500K_R1	16.2%	49%	150 bp	0.5
SRR957824_500K_R2	7.2%	50%	150 bp	0.5
SRR957824_trimmed_R1	2.9%	51%	142 bp	0.4
SRR957824_trimmed_R2	2.7%	51%	136 bp	0.4

FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Quality Histograms 3 1

Toolbox

https://www.hadriengourle.com/tutorials/qc/

Improving quality: toolbox

- Trimmomatic
- Cutadapt
- Scythe
- Sickle
- Atropos

Fastp:

https://github.com/OpenGene/fastp

https://www.hadriengourle.com/tutorials/qc/

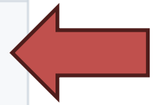
https://github.com/RushAlz/IAMSPE-CS31-Genomica_Computacional

CS31 - Genômica Computacional

Esse repositório contem materiais de aula para a disciplina de Genômica Computacional da IAMSPE.

Para facilitar execução dos tutoriais, utilizaremos o Google Cloud Shell.

Aula	Data	Tema	Slides	Tutoriais
1	2023-10-11	Introdução a Genômica e Sequenciamento	Slides	NA
2	2023-10-18	Bioinformática, Linux e Processamento de dados de sequenciamento	[Slides]	Fastq Quality-Control (QC) tutorial



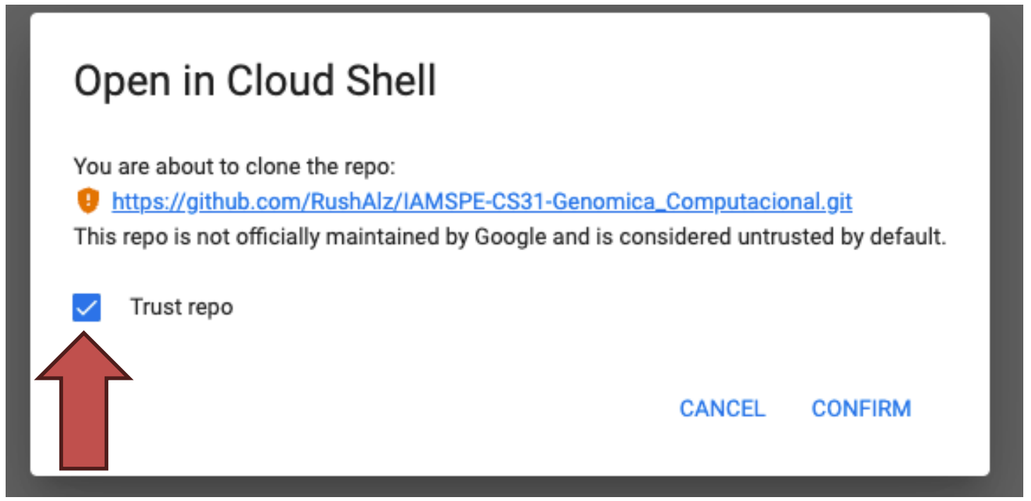
https://github.com/RushAlz/IAMSPE-CS31-Genomica_Computacional

CS31 - Genômica Computacional

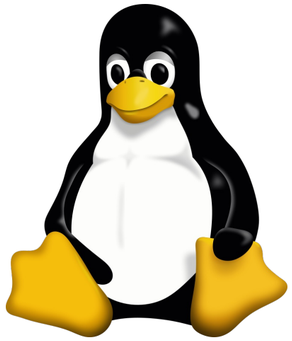
Esse repositório contem materiais de aula para a disciplina de Genômica Computacional da IAMSPE.

Para facilitar execução dos tutoriais, utilizaremos o Google Cloud Shell.

Aula	Data	Tema	Slides	Tutoriais
1	2023-10-11	Introdução a Genômica e Sequenciamento	Slides	NA
2	2023-10-18	Bioinformática, Linux e Processamento de dados de sequenciamento	[Slides]	Fastq Quality-Control (QC) tutorial



https://www.hostinger.com/tutorials/linux-commands



1. ls command

The **ls** command lists files and directories in your system. Here's the syntax:

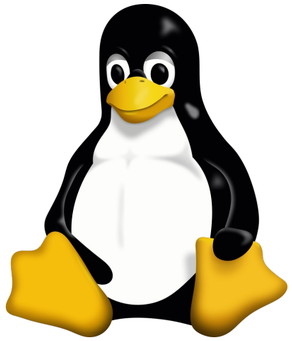
```
ls [/directory/folder/path]
```

```
root@srv:/# ls /directory/folder/path
file1.txt
```

If you remove the path, the **ls** command will show the current working directory's content. You can modify the command using these options:

- **-R** – lists all the files in the subdirectories.
- **-a** – shows all files, including hidden ones.
- **-lh** – converts sizes to readable formats, such as **MB**, **GB**, and **TB**.

https://www.hostinger.com/tutorials/linux-commands



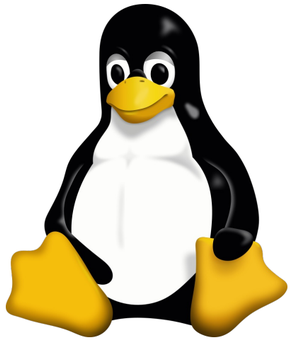
2. pwd command

The **pwd** command prints your current working directory's path, like **/home/directory/path**. Here's the command syntax:

```
pwd [option]
```

It supports two options. The **-L** or **--logical** option prints environment variable content, including symbolic links. Meanwhile, **-P** or **--physical** outputs the current directory's actual path.

```
root@srv:/directory/folder/path# pwd
/directory/folder/path
```



3. cd command

Use the **cd** command to navigate the Linux files and directories. To use it, run this syntax with sudo privileges:

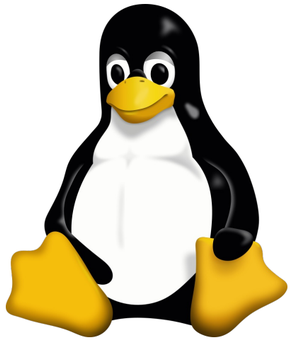
```
cd /directory/folder/path
```

```
root@srv:/# cd /directory/folder/path
root@srv:/directory/folder/path#
```

Depending on your current location, it requires either the full path or the directory name. For example, omit **/username** from **/username/directory/folder** if you are already within it.

Omitting the arguments will take you to the home folder. Here are some navigation shortcuts:

- **cd ~[username]** – goes to another user's home directory.
- **cd ..** – moves one directory up.
- **cd-** – switches to the previous directory.



https://www.hostinger.com/tutorials/linux-commands

4. mkdir command

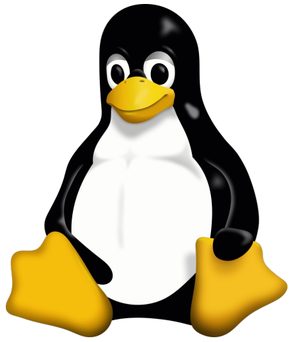
Use the **mkdir** command to create one or multiple directories and set their permissions. Ensure you are authorized to make a new folder in the parent directory. Here's the basic syntax:

```
mkdir [option] [directory_name]
```

To create a folder within a directory, use the path as the command parameter. For example, **mkdir music/songs** will create a **songs** folder inside **music**. Here are several common **mkdir** command options:

- **-p** – creates a directory between two existing folders. For example, **mkdir -p Music/2023/Songs** creates a new **2023** directory.
- **-m** – sets the folder permissions. For instance, enter **mkdir -m777 directory** to create a directory with read, write, and execute permissions for all users.
- **-v** – prints a message for each created directory.

```
root@srv:/# mkdir -v new-folder
mkdir: created directory 'new-folder'
```



6. rm command

Use the rm command to permanently delete files within a directory. Here's the general syntax:

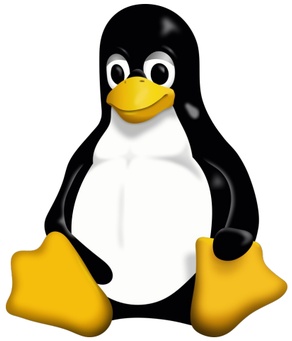
rm [filename1] [filename2] [filename3]

Adjust the number of files in the command according to your needs. If you encounter an error, ensure you have the **write** permission in the directory.

To modify the command, add the following options:

- **-i** – prompts a confirmation before deletion.
- **-f** – allows file removal without a confirmation.
- **-r** – deletes files and directories recursively.

Warning! Use the **rm** command with caution since deletion is irreversible. Avoid using the **-r** and **-f** options since they may wipe all your files. Always add the **-i** option to avoid accidental deletion.



7. cp command

Use the **cp** command to copy files or directories, including their content, from your current location to another. It has various use cases, such as:

- Copying one file from the current directory to another folder. Specify the file name and target path:

```
cp filename.txt /home/username/Documents
```

- Duplicating multiple files to a directory. Enter the file names and the destination path:

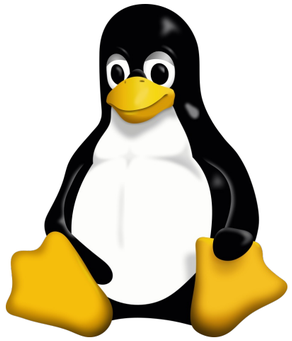
```
cp filename1.txt filename2.txt filename3.txt /home/username/Documents
```

- Copying a file's content to another within the same directory. Enter the source and the destination file:

```
cp filename1.txt filename2.txt
```

- Duplicating an entire directory. Pass the **-R** flag followed by the source and destination directory:

```
cp -R /home/username/Documents /home/username/Documents_backup
```



8. mv command

Use the **mv** command to move or rename files and directories. To move items, enter the file name followed by the destination directory:

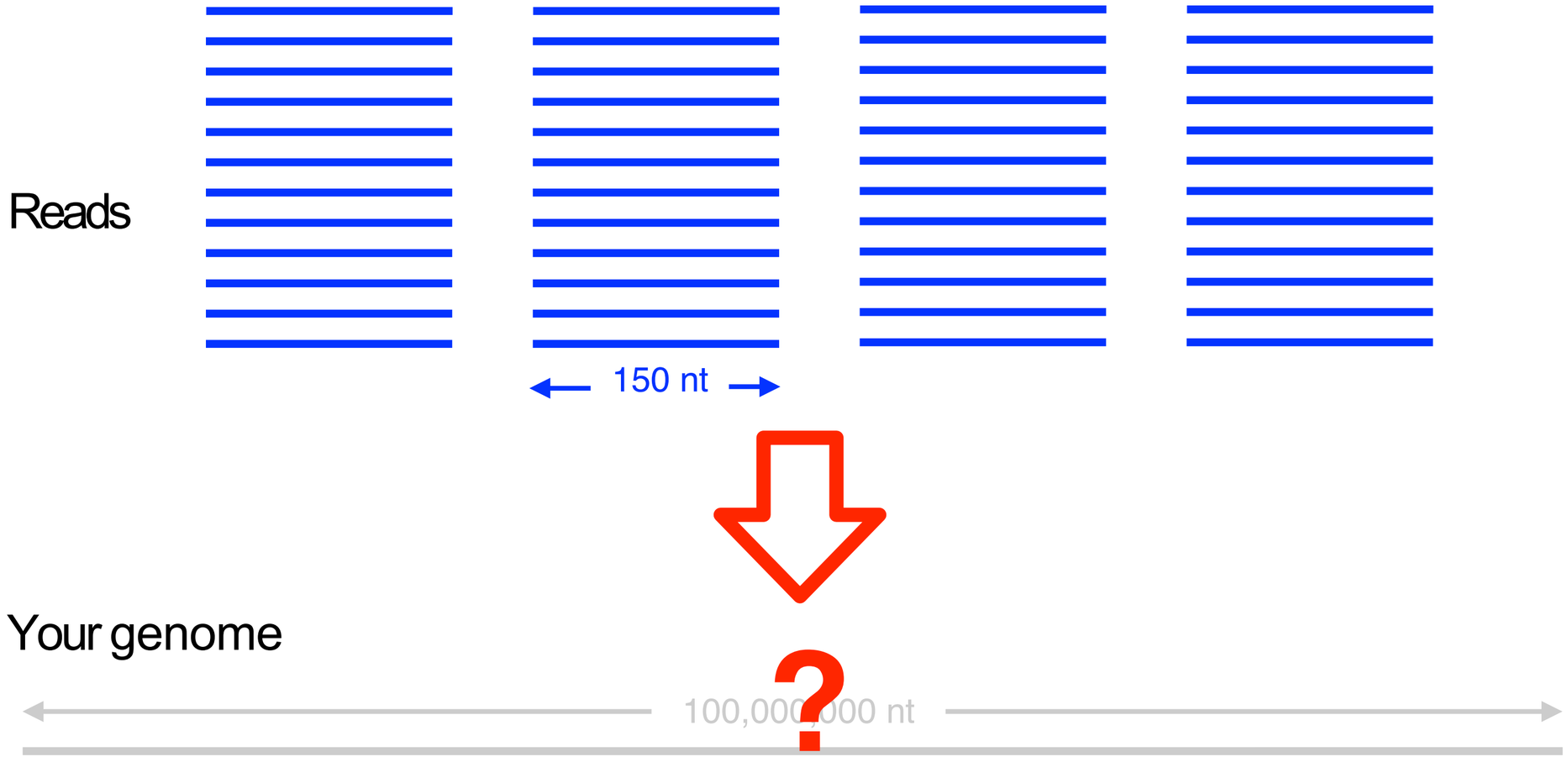
```
mv filename.txt /home/username/Documents
```

Meanwhile, use the following syntax to **rename a file in Linux** with the **mv** command:

```
mv old_filename.txt new_filename.txt
```

Proxima aula…

25-Oct Montagem de genomas (teórico-prática)



https://www.cs.jhu.edu/~langmea/resources/lecture_notes/01_genomics_comp_genomics_v2.pdf