Microbiome Analysis

Introduction to sequencing technologies data pre-processing and quality control

La Serena 3-7.03.2025

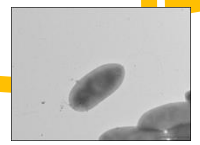
https://github.com/AdamOss88

Adam Ossowicki

Few things about myself

IFB Gdańsk, PL

Traits involved in biological control of plant pathogens by *Pseudomoas*

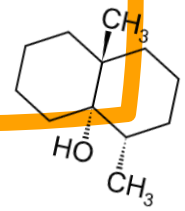


NIOO-KNAW
Wageningen, NL

PhD thesis June 2021: Mechanisms of soil disease suppressivene

BC CAS
Ceske Budejovice , CZ

Interactions between Actinobacteria and soil fauna via volatiles

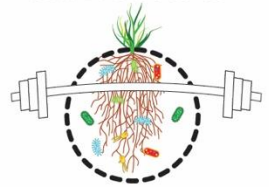


Leiden University
Leiden , NL

The functions of core microbiome in Jacobaea vulgaris herbivores core microbiome

IHSM
Malaga , ES

MicroTRIAS



Microbiomes !!!

How many microbes ?

$4x10^3$ to $5x10^4$ species per gram

$1x10^8$ to $1x10^9$ bacteria cells

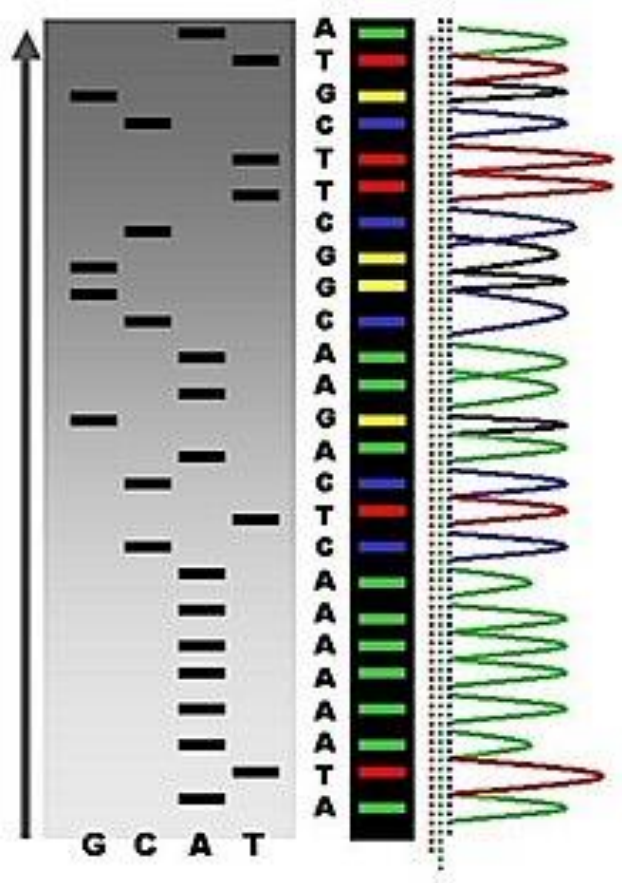
How to study something we cannot see ?

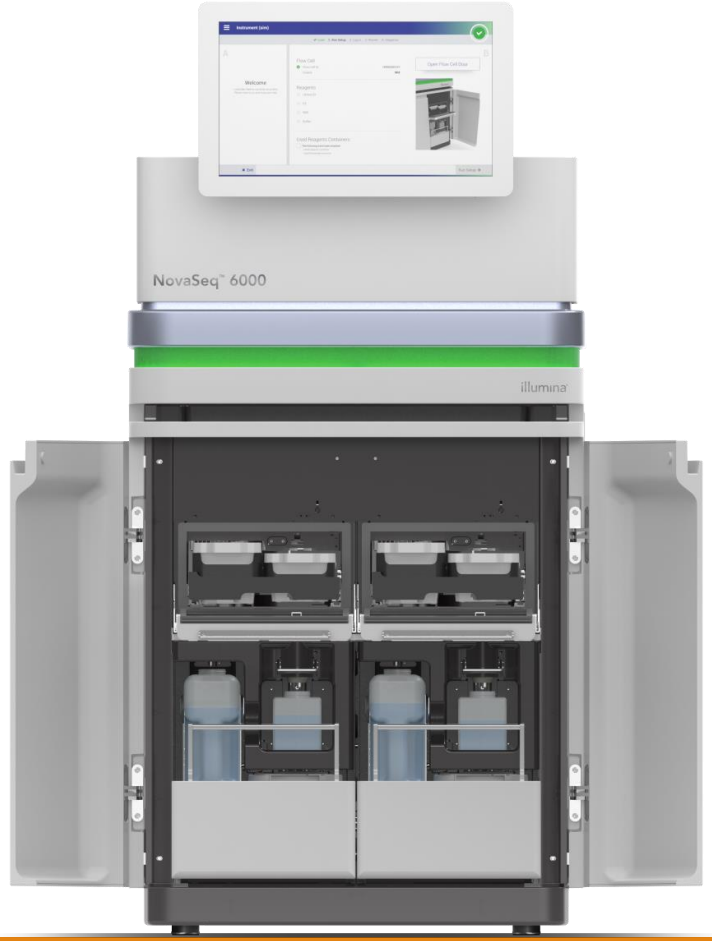
30,000 formally named bacteria species (NCBI, 2022)

Bates et al. ISME 2012 Nunan N., PLoS One 2014

Sanger, Illumina , PacBio, 10X, Oxford nanopore….

- Homogenous samples
- Obligatory PCR step (specific)
- **What do we use it for ?**

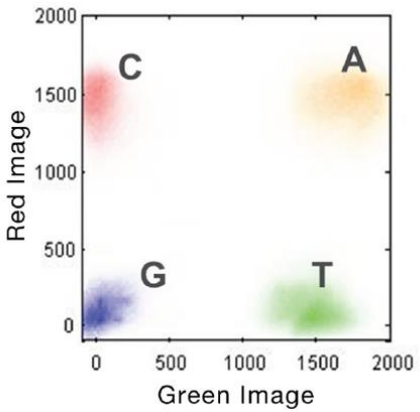


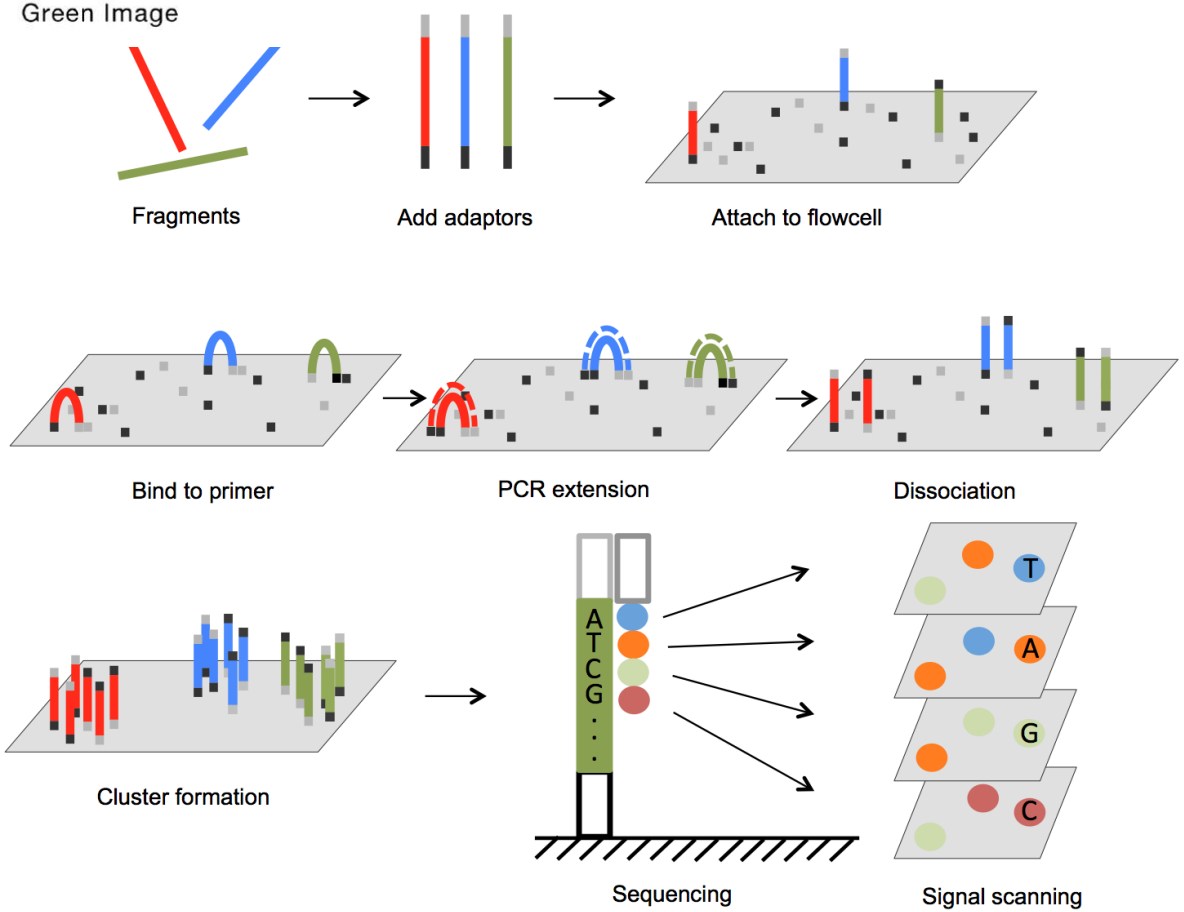


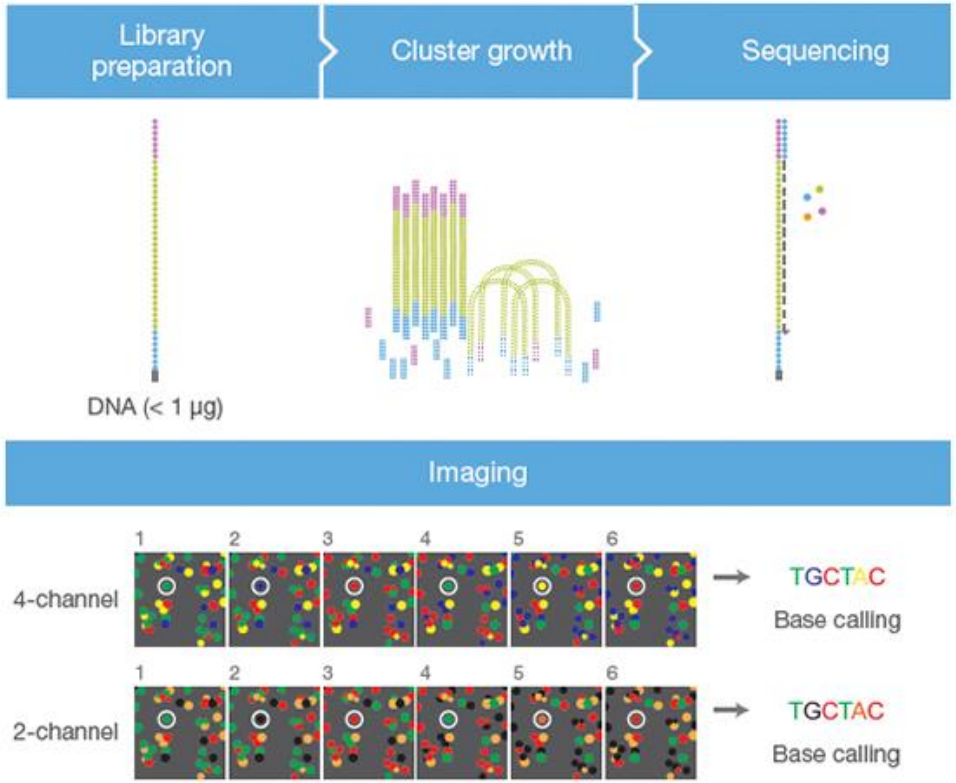
https://emea.illumina.com/

NovaSeq 6000 Sequencing System :up to 6 Tb and 20B single reads in < 2 days

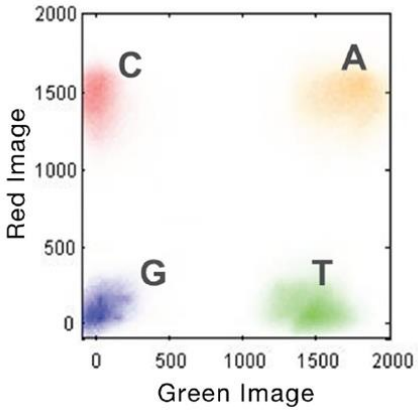
What is Illumina ? And how does it work ?





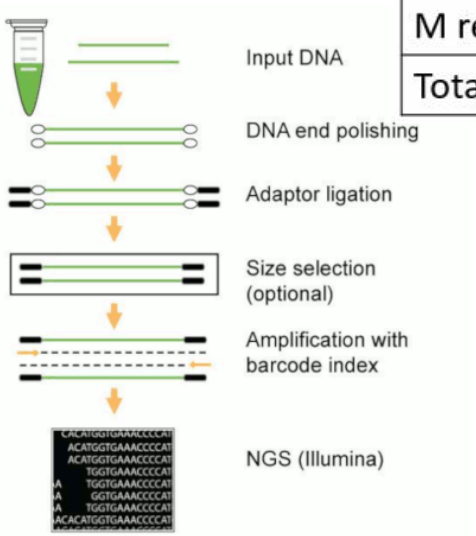


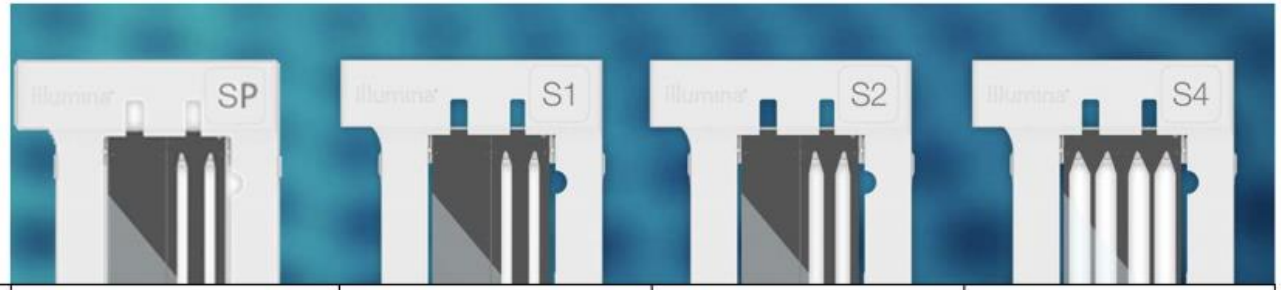
Source: Illumina; Next Generation Sequencing Edited by Jerzy K Kulski, DOI: 10.5772/61657



What is Illumina ? And how does it work ?

A flowcell





Flow cell	SP	S1	S2	S4
Lane number	2	2	2	4
M reads per lane	400	800	2'000	2'500
Total M reads	800	1'600	4'000	10'000

When do we barcode and why ?

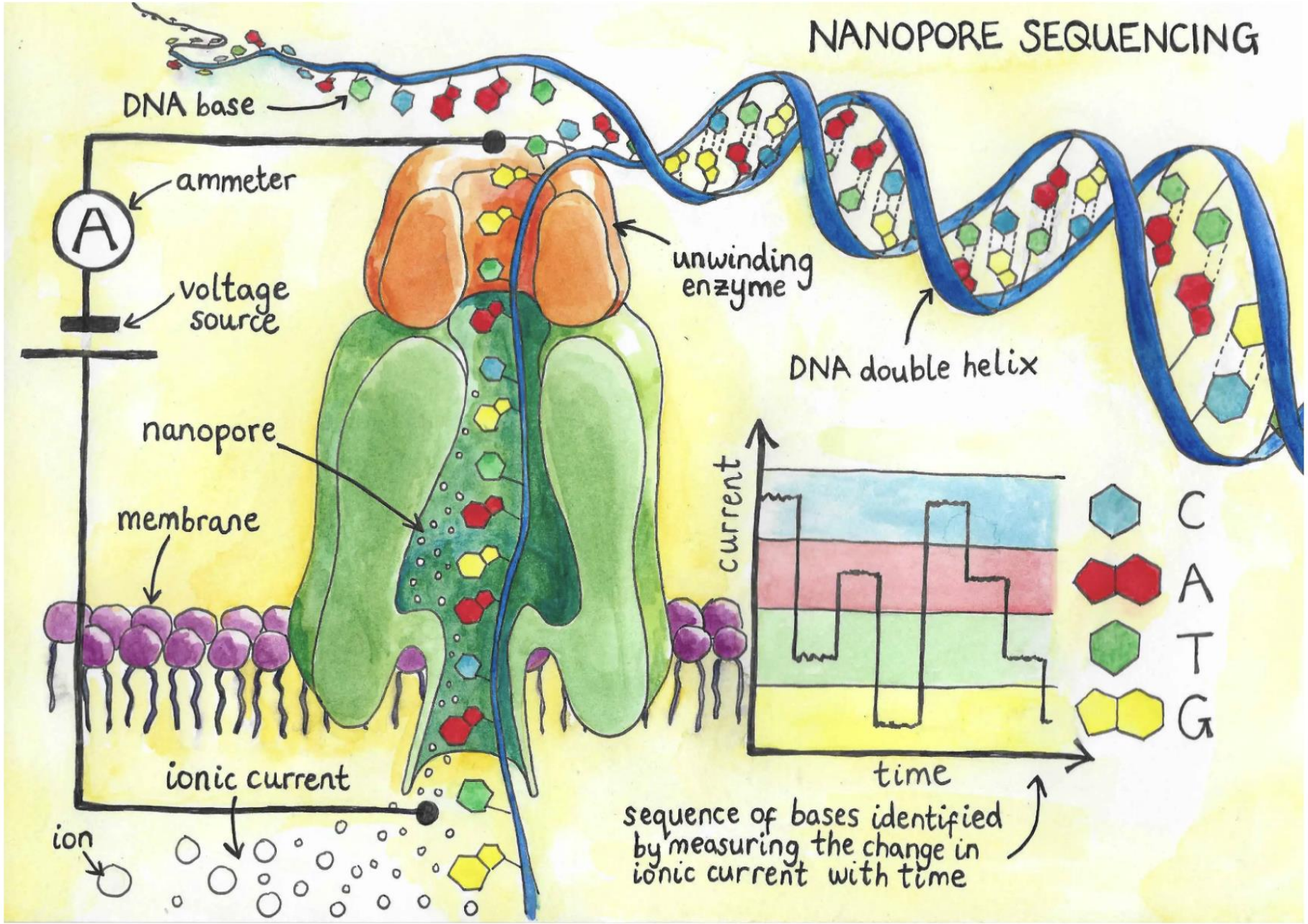
Source: Illumina; Next Generation Sequencing Edited by Jerzy K Kulski, DOI: 10.5772/61657

Why most sequencing we do is outsourced ?

What if I want to sequence myself ?

https://emea.illumina.com/

Where is the sequencer ???



Nanopore vs. Ilumina

- Size of sequencers
- Longer reads
  - how much longer ?
- More sequencing errors ?

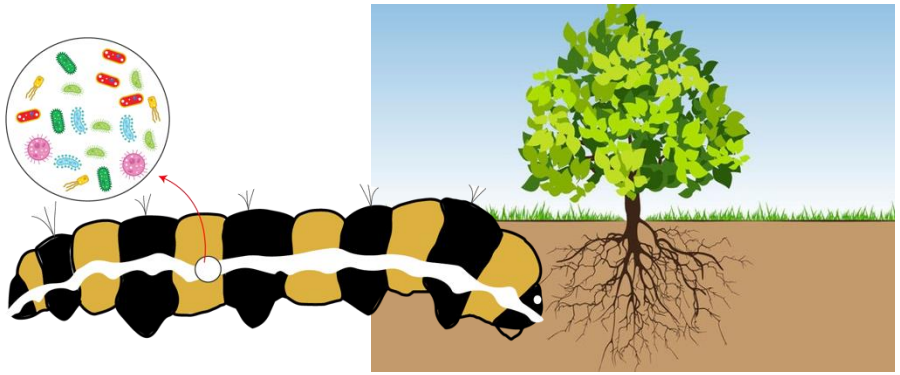
https://oxsci.org/pore-over-this-advances-in-dna-sequencing/

What is metagenomics?

Answer: Depends who you ask !

- Unspecific

- Requires computing infrastructure *
- Analysis requires (more) knowledge and experience
- More versatile



shotgun metagenomics

taxonomy profiles, relative abundance, functional profiles (cds), BGCs, MAGs, metabolic pathway reconstruction

genetic potential

- Specific (is it ?)

- Not necessarily requires computing infrastructure *
- Analysis requires (less) knowledge and experience
- Gets deeper into microbiome *
- Less versatile

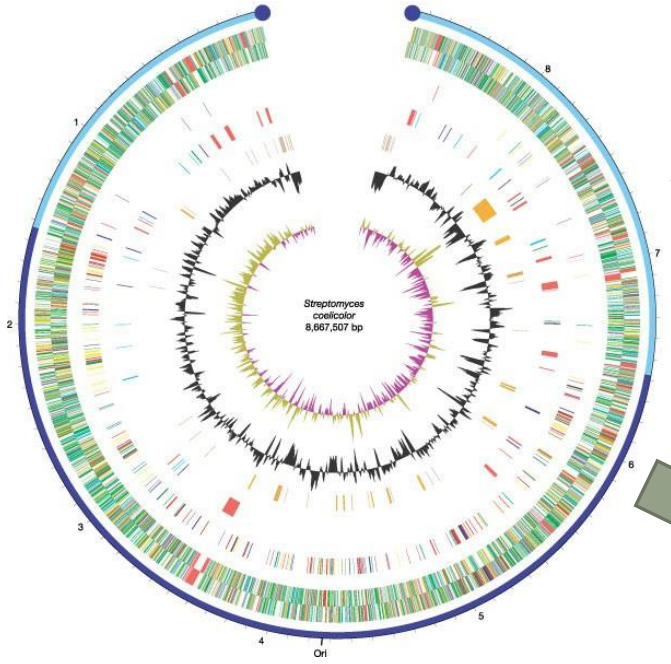
amplicon sequencing

taxonomy profiles, relative abundance

* Most of the times

Gets deeper into microbiome… example

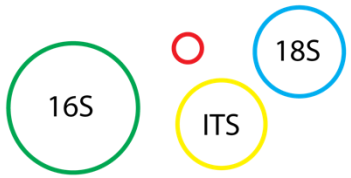
Streptomyces coelicolor A3(2) ~ 8 Mb



shotgun metagenomics

We can sequence the whole genome <60x

amplicon sequencing

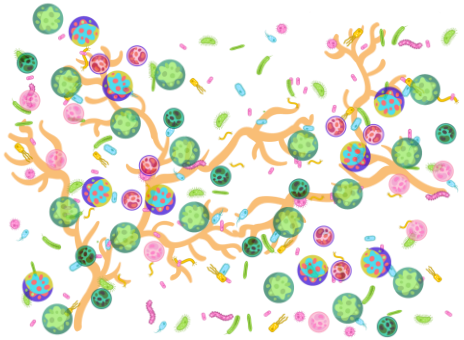


We can sequence the 16S fragment ~1666x

In 1 Gb sequencing output

.. So in a simplified microbiome

You will find 60 taxa

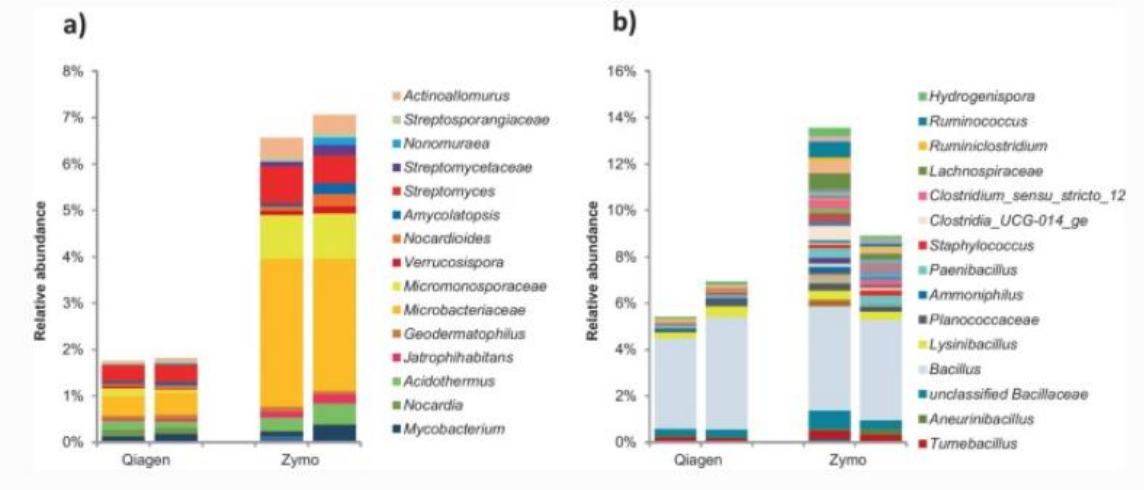


You will find 1666 taxa

Let’s sequence something…

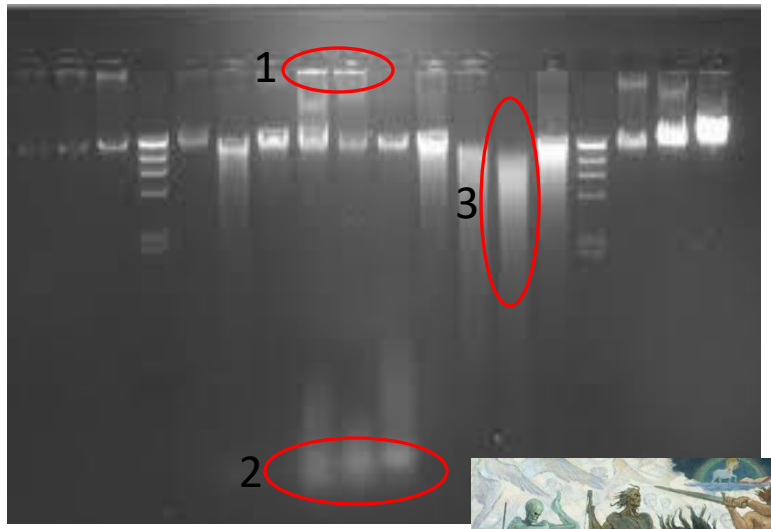
What you give and what you get in return ?

isolation → quality check → shipping



https://doi.org/10.1007/s12223-021-00866-0

Effects of DNA preservation solution and DNA extraction methods on microbial community profiling of soil

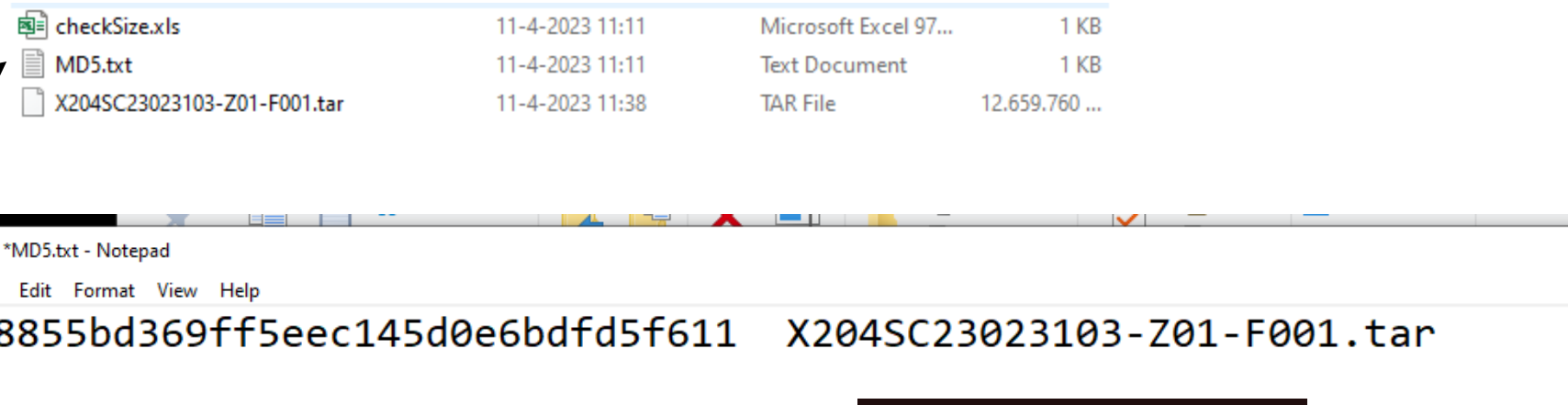


Source of the gel: JGI

Who is the 4th horseman ?

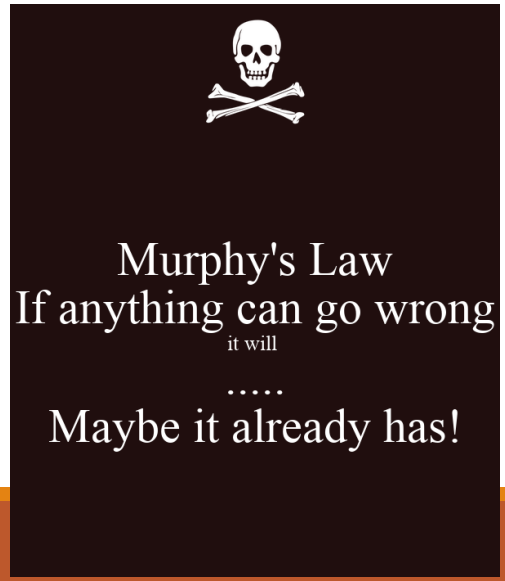
What you get in return…

..a link



What is this ?

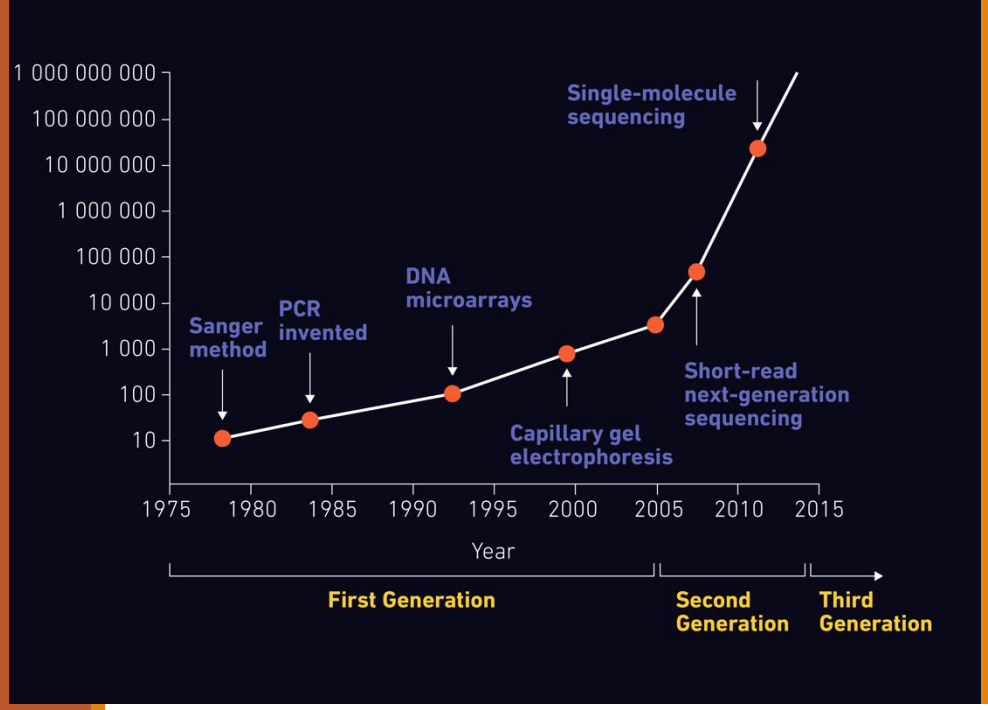
Windows: WinMD5Free or certutil (command line) or other
Linux/Mac: has in-built software called … md5sum



Where (not) to store my data ?

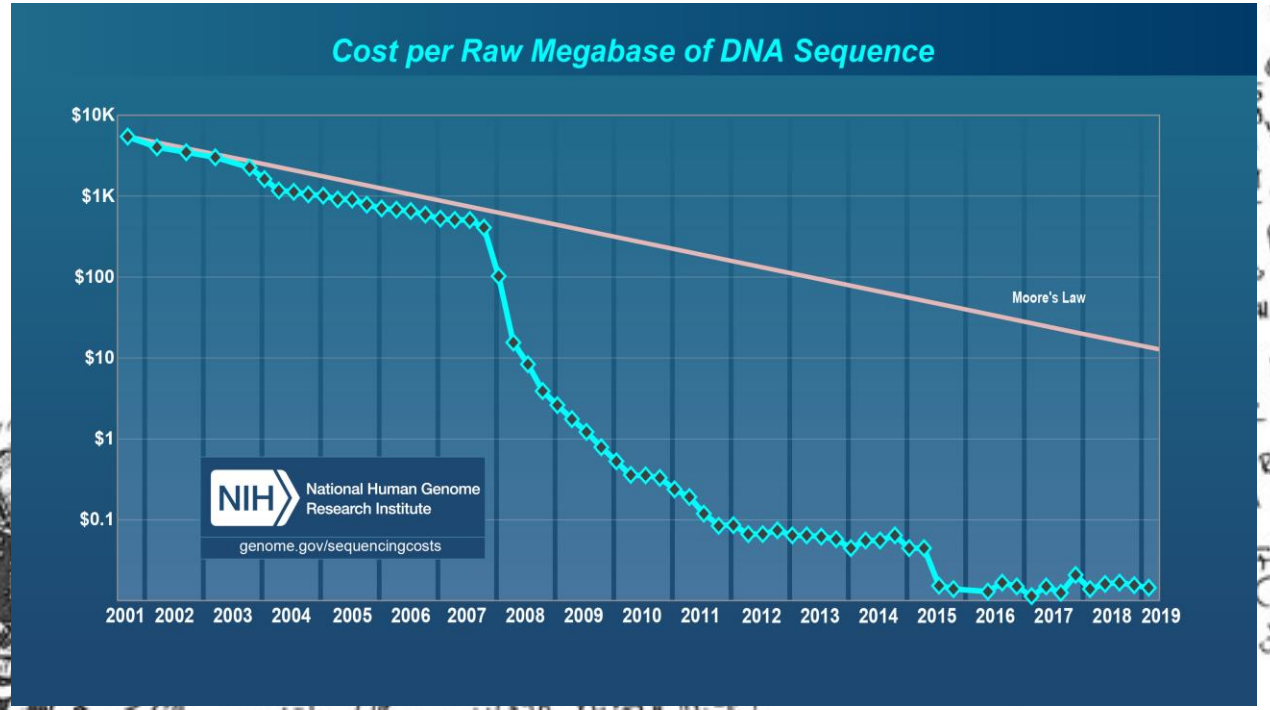
- Backed-up storage servers
- Repositories
- Clouds ?

Computing power



To work with big data we need

Computing infrastructure like servers and clusters.



What are they ?

Why do we need them ?

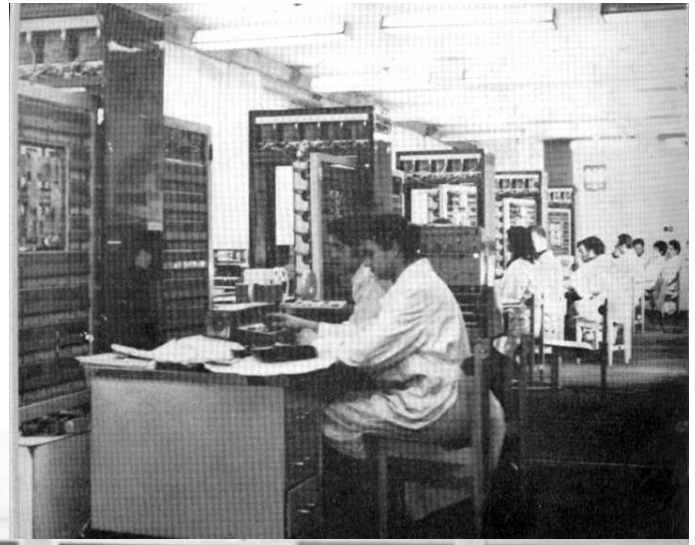
How do we use them ?

What are they ?

computer (n.)
1640s, "one who calculates, a reckoner, one whose occupation is to make arithmetical calculations," agent noun from **compute** (v.).

Lp	Nr ścieżki 1	2	3	4	5	Litery	Cyfry i znaki
1	●	●	•			A	–
2	●		•		●	B	?
3		●	•	●		C	:
4	●		•		●	D	kto tam
5	●		•			E	3
6	●		•	●	●	F	wolny
7		●	•		●	G	wolny
8			•	●		H	wolny
9		●	•	●		I	8
10	●	●	•		●	J	dzwonek
11	●	●	•	●	●	K	(
12		●	•			L	)
13			•	●	●	M	.
14			•	●	●	N	,
15			•		●	O	9
16		●	•	●		P	0

Lp	Nr ścieżki 1	2	3	4	5	Litery	Cyfry i znaki
17	●	●	•	●		Q	1
18		●	•		●	R	4
19	●		•	●		S	'
20			•		●	T	5
21	●	●	•	●		U	7
22		●	•	●	●	V	=
23	●	●	•		●	W	2
24	●		•	●	●	X	/
25	●		•	●		Y	6
26	●		•		●	Z	+
27			•	●		powrót karetki	
28		●	•			obrot wałka	
29	●	●	•	●	●	litery	
30	●	●	•		●	cyfry i znaki	
31			•	●		odstęp	
32			•				



They are **just** computers….

What are they ?

Laptop

performance

portability

easy to use

Server

performance

portability

easy to use

HPC – high performance computing

performance

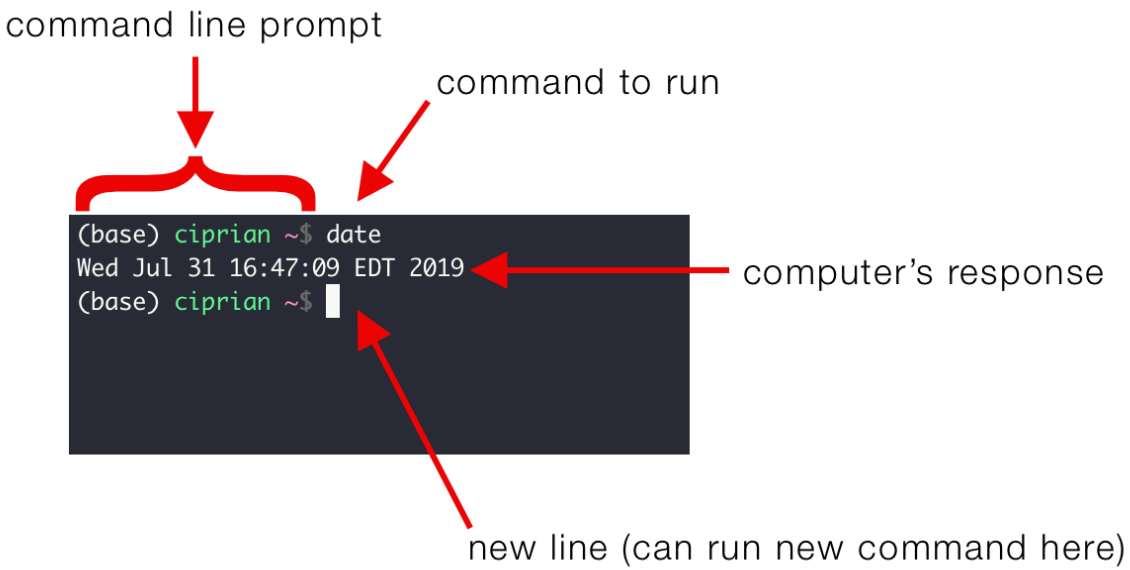
portability

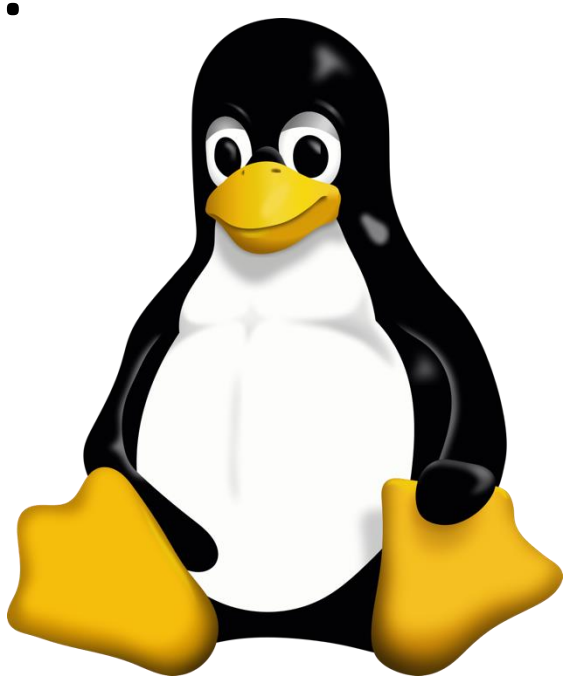
easy to use

Why do we need them ?

Performance portability or easy to use ?

Why do they use Linux based operating systems* ?





*most do but there are exceptions

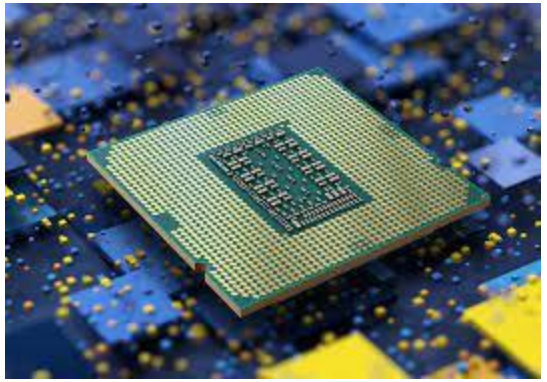
Parameters !

RAM

memory

Disc space

cores



What I can do using…

Laptop

Process “small” amplicon dataset

If you have a fast computer with =>16Gb ram, process a small metagenome

Work on genomes

Do the most of downstream analysis

What I can do using…

Server

Process big amplicon dataset

Process and annotate moderate size metagenomes

Comparative genomics

What I can do using…

HPC – high performance computing

Do a metanalysis

Do comparative metagenomics

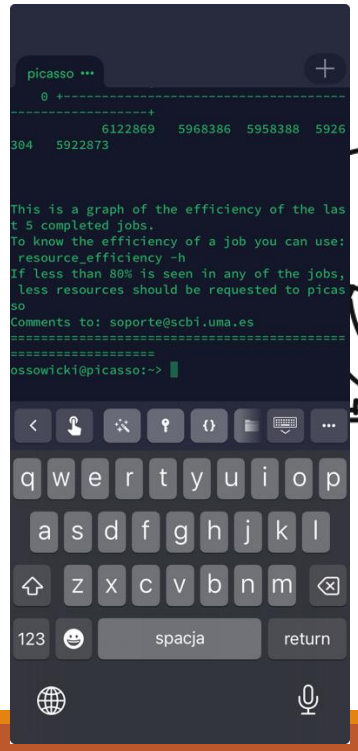
Well…

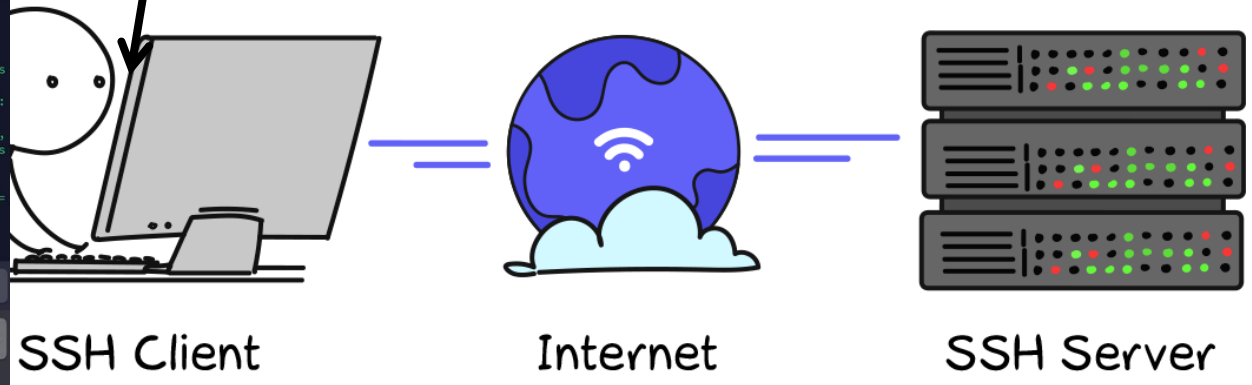
“Sky is the limit”

How do we use them ?

Where is the screen ?

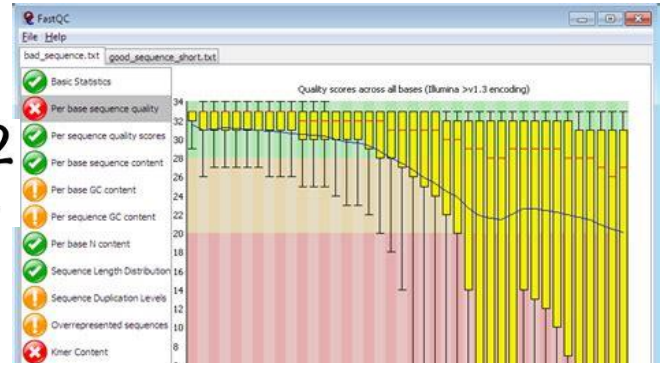
SSH - Secure Shell Protocol

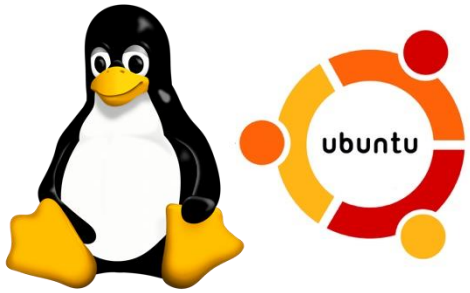




Is it gonna work on my… ?

DADA2
Amplicon Sequencing. Exactly.





- Most probably yes

marcelm/**cutadapt**

MultiQC

- Probably no …but you can emulate Linux in WSL (Windows Subsystem for Linux) and then probably yes

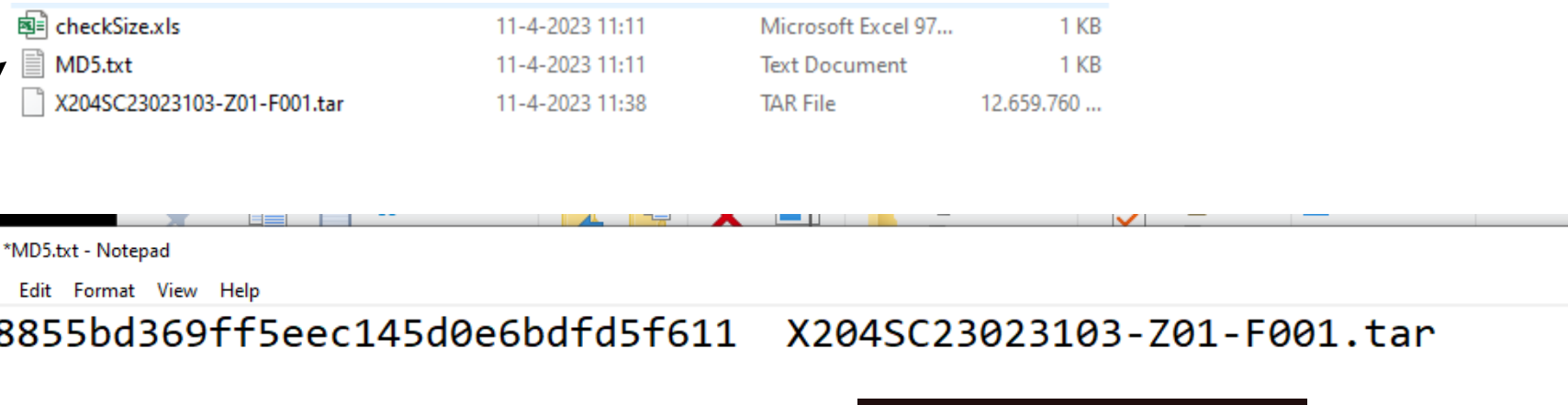
qiime2

BOW TIE

- Probably yes … with exceptions

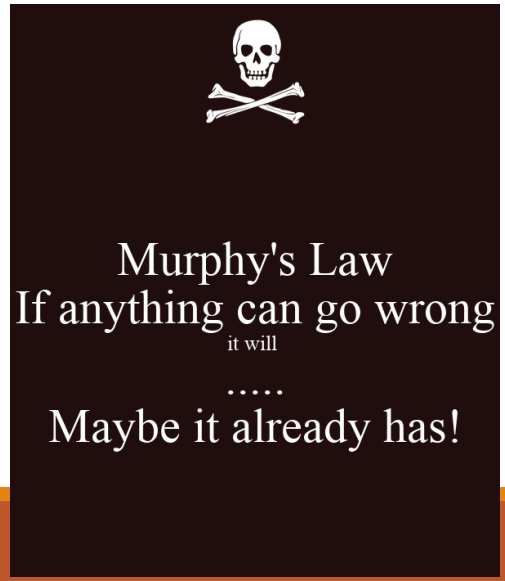
What you get in return…

..a link



What is this ?

Windows: WinMD5Free or certutil (command line) or other
Linux/Mac: has in-built software called … md5sum



What you get in return…

Sequence identifier, description

Sequence

Optional field

Quality

Next sequence

```
@A00881:1079:HTJF2DRX2:1:210
ATTGAGGAGTGTCAGCAGCCGCGGTAAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00881:1079:HTJF2DRX2:1:216
ATTGAGGAGTGCCAGCCGCCGCGGTAAT
+
F FFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Symbol	ASCII Code	Q-Score
6	54	21
7	55	22
8	56	23
9	57	24
:	58	25
;	59	26
<	60	27
=	61	28
>	62	29
?	63	30
@	64	31
A	65	32
B	66	33
C	67	34
D	68	35
E	69	36
F	70	37
G	71	38
H	72	39
I	73	40

@HWI-Mxxxx or @Mxxxx - MiSeq
@HWUSI - GAIIx
@HWI-Dxxxx - HiSeq 2000/2500
@Kxxxx - HiSeq 3000(?)/4000
@Nxxxx - NextSeq 500/550
@Axxxxx - NovaSeq
@Vxxxxx = NextSeq 2000
@AAxxxxx - NextSeq 2000 P1/P2/P3
@Hxxxxxx - NovaSeq S1/S2/S4

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

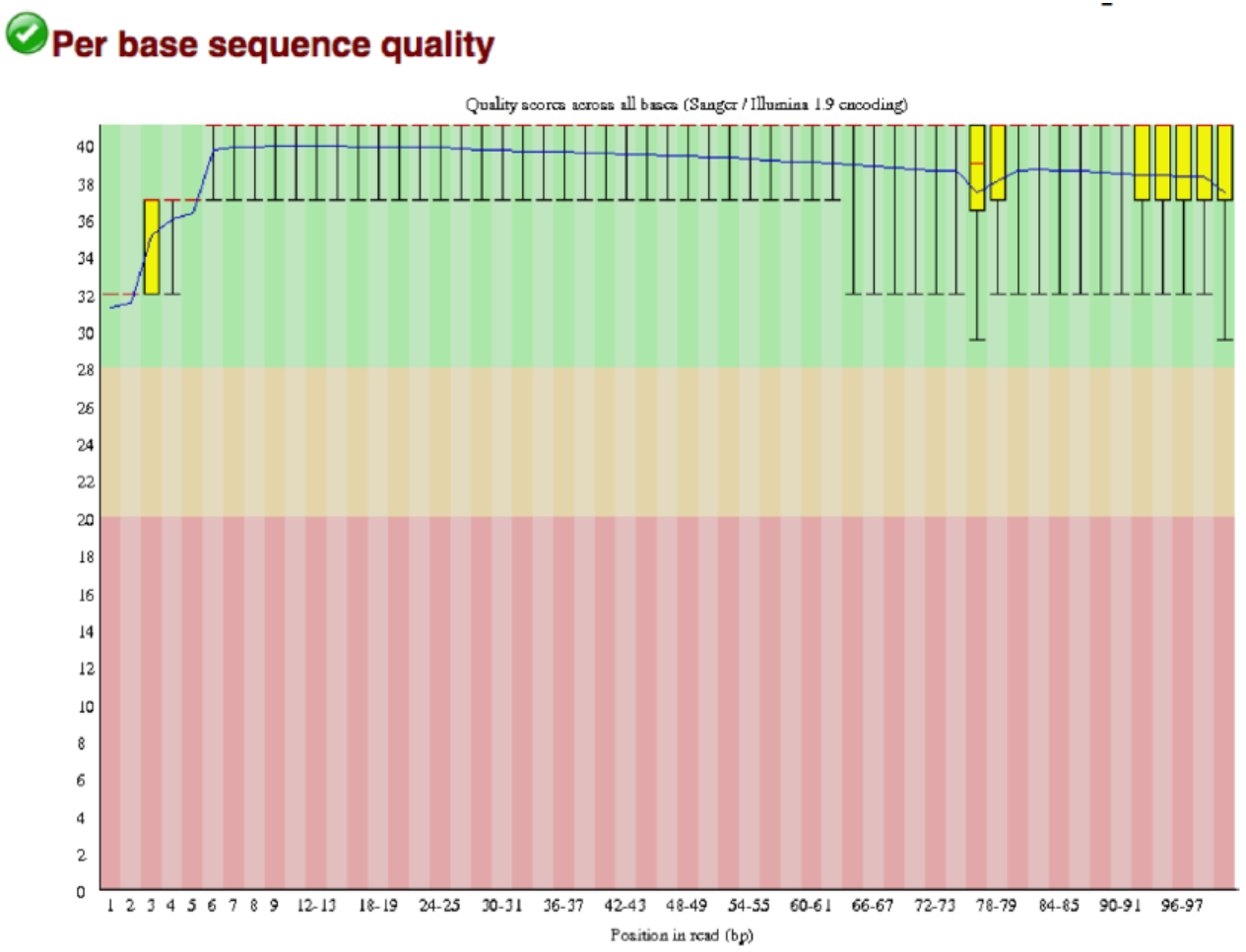
Table 1- The probability values of a Phred quality score and the interpretation of their base-calling accuracy.

Quality is encoded ☺
Mostly using 2 different systems

Instrument identifier
run id flowcell id lane tiles … Index sequence

```
@A00881:1079:HTJF2DRX2:1:2103:10348:4038 1:N:0:GCCAATAC+ACCGACAA
ATTGAGGAGTGTCAGCAGCCGCGGTAATACGGAGGGAGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Source: Illumina



DATA PREPROCESSING - QUALITY CONTROL

Software:

FastQC

FastP

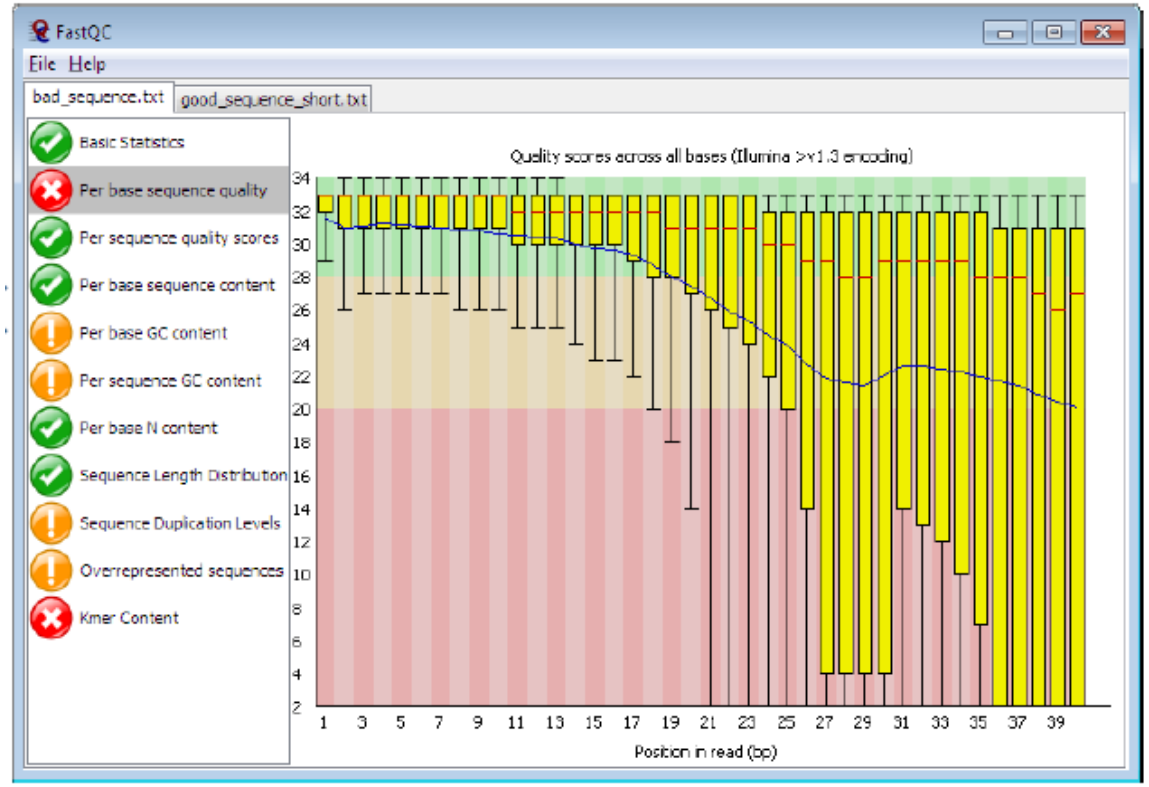
FastX

PRINSEQ

TagCleaner

…

And very useful MultiQC for consolidation



DATA PREPROCESSING - ADAPTER TRIMMING, QUALITY FILTERING

- From QC data you may notice that adapters or primers (amplicon sequencing) are still present in your sequence. You should remove them either by providing the adapter/primer sequence or using a de-novo search.

- Recommended tools: Trimmomatic, Cutadapt, Fastp Dada2, bbduk, PRINSEQ++, AfterQC, Sickle, Flexbar, Seqtk…

Quality filtering – typical commands

- **Trimmomatic**

```
trimmomatic PE -phred33  raw_R1.fastq.gz raw_R2.fastq.gz  trimmed_R1_paired.fastq.gz\
trimmed_R1_unpaired.fastq.gz trimmed_R2_paired.fastq.gz trimmed_R2_unpaired.fastq.gz \
ILLUMINACLIP:adapters.fa:2:30:10 LEADING:25 TRAILING:25 SLIDINGWINDOW:4:20 MINLEN:50
```

SLIDINGWINDOW:4:20 - Scans using a 4-base window, cutting when average quality < 20.

- **Cutadapt**

```
cutadapt -a <primer F> -A <primer R> -o trimmed_R1.fastq.gz  -p trimmed_R2.fastq.gz \
-m 50 -q 25,25 -n 2 raw_R1.fastq.gz raw_R2.fastq.gz
```

- **Fastp**

```
fastp -i raw_R1.fastq.gz -I raw_R2.fastq.gz -o trimmed_R1.fastq.gz -O \
trimmed_R2.fastq.gz --detect_adapter_for_pe --cut_right  --cut_mean_quality 25 \  --
length_required 50 --n_base_limit 0 --html fastp_report.html --json fastp_report.json
```

examples

Thank you for your attention