Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

SUS-MIRRI.IT

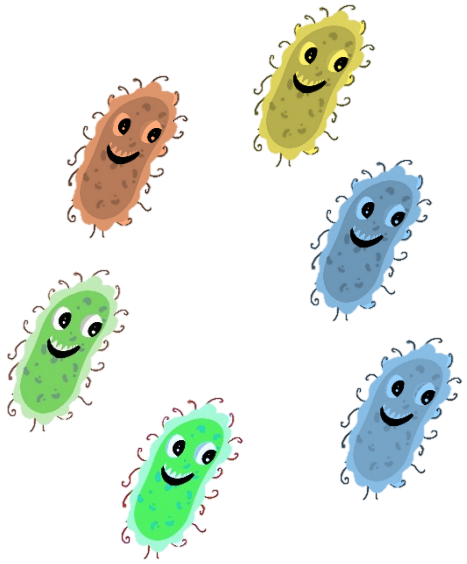
PIPELINES FOR NGS DATA

27/11/2023

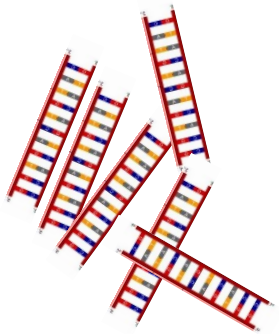
Dr. Andrea Visca

Why am I sequencing DNA?

Microbial community



Genetic information (DNA)



Sequencing

Escherichia coli

Pseudomonas sp.

Xanthomonas sp.

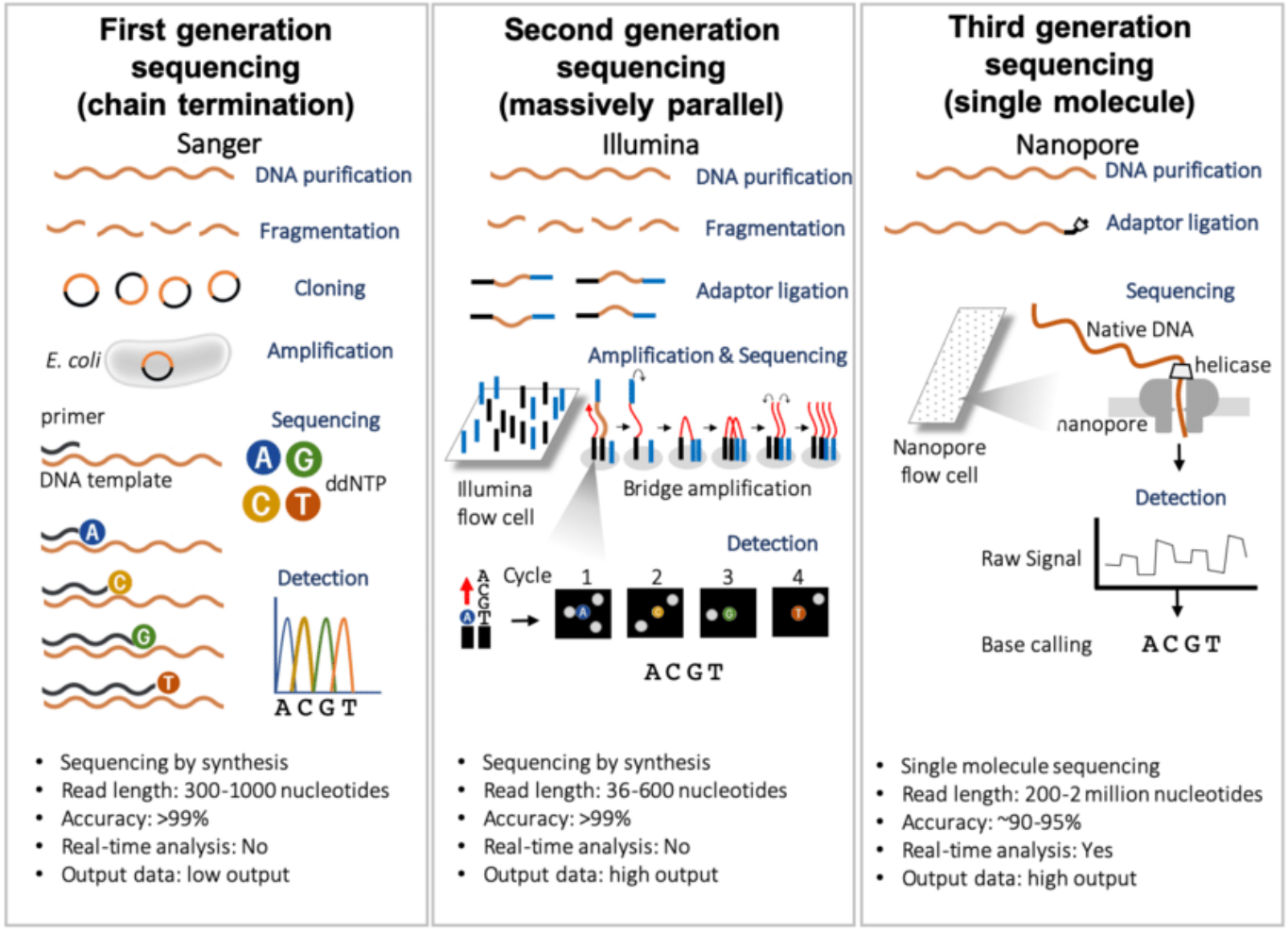
Streptomyces sp.

Clostridium sp.

Clostridium sp.

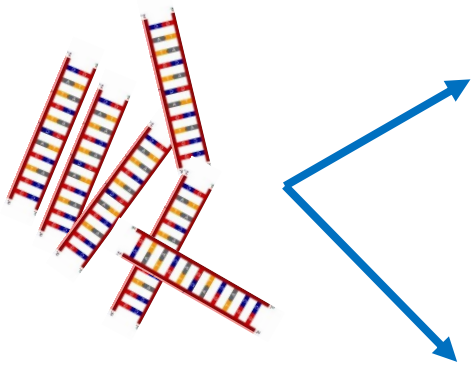
NGS technologies

Next-generation sequencing (NGS) is a massively parallel sequencing technology that offers ultra-high throughput, scalability, and speed. The technology is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA.

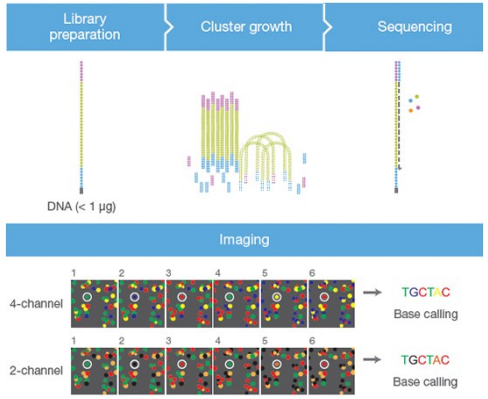


Workflow of a sequencing: Illumina and Nanopore

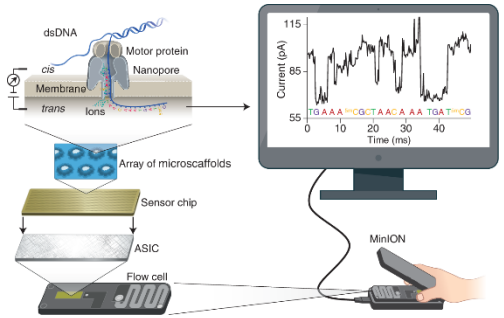
DNA library



Illumina

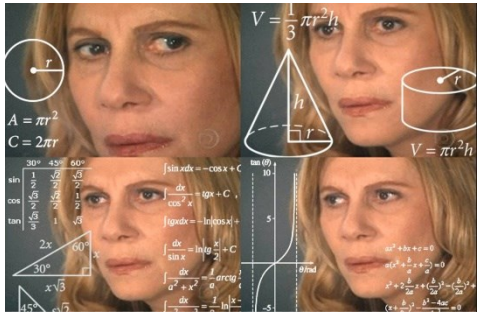


Nanopore



Sequencing output

```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCGGATTCACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCCFFFCADBGGGGGGGGGGHHGHGGFHGHHHGHGGGAFFHGGGGGHHHHHHHHGGGGGHHGGGGGGGGGHGGEGGGGGHHHHHHHHGGHGGGHHHHHGGG
@M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBBBBGGGGGGGGGFHHHHHGGHGGGGGGGGGGGHHGGEGFHHHHHHHHGGGGHFHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHGGGG
@M04743:199:000000000-CGG4F:1:1101:13893:1760 1:N:0:233
GGTGCCAGCAGCCGCGGTACTACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
BBBBBFFFB4CCGGGGGGGCFFHGHHHGGHGGGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEHGGGGGGDHHEHGHGGGGGG
F9FFFFFFFFFFFEFFBBBBFEB;-@DFB-BBBFFFFEFF/EBBEFFF/BADFFDFFF.;
@M04743:199:000000000-CGG4F:1:1101:14830:1795 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTCTTTAAGTCAGTGGTGAAATACAGCCG
+
ABBABFBFB?AAEE?EGEFCGGHHFFHGEHFFHHGHGGGCFHHGEEGGDFGDHHHGGGFGDGHGGFEGFGGDFGGGGGHHFFFBGFH34FGBFFHGHHHGHFFC
9BD?99-9/9@-BD.;ADFFFBF///BBF:FFFFFFED?DFDFF?A.
@M04743:199:000000000-CGG4F:1:1101:14968:1984 1:N:0:233
AGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTCTTTAAGTCAGTGGTGAAATACAGCCG
+
BBBBBFFFBABBGGGGGGGGGGHHGFHGHHGHHHGHGGGCFHHGGEGGHHHHHHGGGGHHHHHGHGGGGGGGHGGGGGHHHHHHHHHHHHHGFFHHHHHHHHG
FCHHHGGHHHHHHHHHHHHHHHHHHHHHHFHHHGFHHGEGGFHHGHGGGFEGG9FGGAEGGGGAFDGEFFGGFFFBFEFFFFFFFFFFFFFFFFF>DFDBFFI
@M04743:199:000000000-CGG4F:1:1101:12706:2099 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGGAGGGAGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGTTTTTTAAGTCAGAGGTGAAAGCCCGGGG
+
BCCCCFFFCCCCGGGGGGGGGGHHEGGGGDFGGHHHGGGGGHHGGGGFHHGHHHHHGGGGHHHGGGGGHHHGHGGGGGGGGACGHHHHHHGHHGHHFHHHGGGG
BFFFFFFFF9FFFFFFFFFFFFFF/
@M04743:199:000000000-CGG4F:1:1101:13747:2260 1:N:0:233
CGTGCCAGCCGCCGCGGTAATACGAAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
CCCCCCFFCABCGGGGGGGGGGHHFCEGDGGGGHHHGGGEFHHGGGFFHHFHHHHGGGGGHH@GHHHGGHGGHGGGGGGFH</>CFCGGGGHHHHHFHGGGGGG
A@@FFFFFFFFFFFFFBF9C;=CF.@;CDFFFFFBDFFFFFF?BEFFFFFFFFFFFFFFF?
@M04743:199:000000000-CGG4F:1:1101:20151:2263 1:N:0:233
TGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBAADGGGGGGGGGGHHHHHGGHGGGGGGGGGGGHHGGDFFHHHHHHHGGGGGHGHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHGGFG
@M04743:199:000000000-CGG4F:1:1101:17232:2363 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGGAGGGGGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGATCGGAAAGTCAGAGGTGAAATCCCAGGG
+
BBBBBFFFBBBBGGGGGGGGGGHHGDGGGGGGGHHHGGG0FGHGGEGFHHHHHHHHHGGGGHHHGGGGGHHHGHGGGGGGGGGGHHHHHHHGHHHGHHHGGG
```



FastQ file

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

Header Sequence Quality

```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?######
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?
```

cBio and Sanger

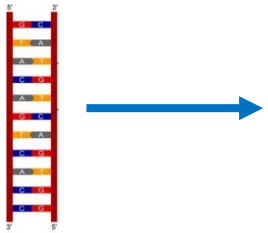
II	Q	P_error	ASCII	Q	P_error	ASCII
,	22	0.00631	55 7	33	0.00050	66 B
-	23	0.00501	56 8	34	0.00040	67 C
.	24	0.00398	57 9	35	0.00032	68 D
/	25	0.00316	58 :	36	0.00025	69 E
0	26	0.00251	59 ;	37	0.00020	70 F
1	27	0.00200	60 <	38	0.00016	71 G
2	28	0.00158	61 =	39	0.00013	72 H
3	29	0.00126	62 >	40	0.00010	73 I
4	30	0.00100	63 ?	41	0.00008	74 J
5	31	0.00079	64 @	42	0.00006	75 K
6	32	0.00063	65 A			

II	Q	P_error	ASCII	Q	P_error	ASCII
K	22	0.00631	86 V	33	0.00050	97 a
L	23	0.00501	87 W	34	0.00040	98 b
M	24	0.00398	88 X	35	0.00032	99 c
N	25	0.00316	89 Y	36	0.00025	100 d
O	26	0.00251	90 Z	37	0.00020	101 e
P	27	0.00200	91 [	38	0.00016	102 f
Q	28	0.00158	92 \	39	0.00013	103 g
R	29	0.00126	93]	40	0.00010	104 h
S	30	0.00100	94 ^	41	0.00008	105 i
T	31	0.00079	95 _	42	0.00006	106 j
U	32	0.00063	96 `			

How to classify?

The taxonomy classification of the sequenced reads is made by aligning the DNA reads with a database of know sequences

DNA sequenced read



Database of known sequences

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len
☑	Pseudomonas sp. strain E8 16S ribosomal RNA gene, partial sequence	Pseudomonas sp.	2459	2459	100%	0.0	99.85%	1413
☑	Bacterium strain G12 16S ribosomal RNA gene, partial sequence	bacterium	2459	2459	100%	0.0	99.85%	1411
☑	Pseudomonas sp. Eqa60 gene for 16S ribosomal RNA, partial sequence	Pseudomonas sp.	2459	2459	100%	0.0	99.85%	1420
☑	Pseudomonas protegens strain 58B7 16S ribosomal RNA gene, partial sequence	Pseudomonas protegens	2459	2459	100%	0.0	99.85%	1413
☑	Pseudomonas sp. strain LPH60 16S ribosomal RNA gene, partial sequence	Pseudomonas sp.	2459	2459	100%	0.0	99.85%	1386
☑	Pseudomonas sp. strain LJ-A13 16S ribosomal RNA gene, partial sequence	Pseudomonas sp.	2459	2459	100%	0.0	99.85%	1444
☑	Pseudomonas protegens strain 18P_8.2_Bac1 16S ribosomal RNA gene, partial sequence	Pseudomonas protegens	2459	2459	100%	0.0	99.85%	1386

Which one is the correct classification?

In this way I aligned one sequence.. But with NGS I usually have **10-100 x 10^6 sequences**!

The sequences in the fastQ are not all the same, I have to **consider the quality** of the sequencing!

We need to use **dedicated (bio)informatic tools**

Filtering and trimming

The quality check might have shown the number of reads that have low quality scores. These reads will probably not align very well because of the potential mistakes in base calling, or they may align to wrong places in the genome.

```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCGGATTCACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCCFFFCADBGGGGGGGGGGHHGHGGFHGHHHGHGGGAFFHGGGGGHHHHHHHHHGGGGGHHGGGGGGGGGHGGEGGGGGHHHHHHHHGGHGGGHHHHHHGGG
@M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBBBBGGGGGGGGGFHHHHHGGHGGGGGGGGGGHHGGEGFHHHHHHHHHGGGGHFHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHGGGG
@M04743:199:000000000-CGG4F:1:1101:13893:1760 1:N:0:233
GGTGCCAGCAGCCGCGGTACTACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
BBBBBFFFB4CCGGGGGGGGCFFHGHHHGGHGGGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEHGGGGGGGDHHEHGHGGGGGG
F9FFFFFFFFFFFEFFBBBBFEB;-@DFB-BBBFFFFEFF/EBBEFFF/BADFFDFFF.;
@M04743:199:000000000-CGG4F:1:1101:14830:1795 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGTTCTTTAAGTCAGTGGTGAAATACAGCCG
+
ABBABFBFB?AAEE?EGEFCGGHHFFHGEHFFHHGHGGGCFHHGEEGGDFGDHHHGGGFGDGHGGFEGFGGDFGGGGGHHFFFBGFH34FGBFFHGHHHGHFFC
9BD?99-9/9@-BD.;ADFFFBF///BBF:FFFFFFED?DFDFF?A.
```

Q30

Q33

Q31

Q20

Filtering

```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCGGATTCACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCCFFFCADBGGGGGGGGGGHHGHGGFHGHHHGHGGGAFFHGGGGGHHHHHHHHHGGGGGHHGGGGGGGGGHGGEGGGGGHHHHHHHHGGHGGGHHHHHHGGG
@M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBBBBGGGGGGGGGFHHHHHGGHGGGGGGGGGGHHGGEGFHHHHHHHHHGGGGHFHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHGGGG
@M04743:199:000000000-CGG4F:1:1101:13893:1760 1:N:0:233
GGTGCCAGCAGCCGCGGTACTACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
BBBBBFFFB4CCGGGGGGGGCFFHGHHHGGHGGGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEHGGGGGGGDHHEHGHGGGGGG
F9FFFFFFFFFFFEFFBBBBFEB;-@DFB-BBBFFFFEFF/EBBEFFF/BADFFDFFF.;
```

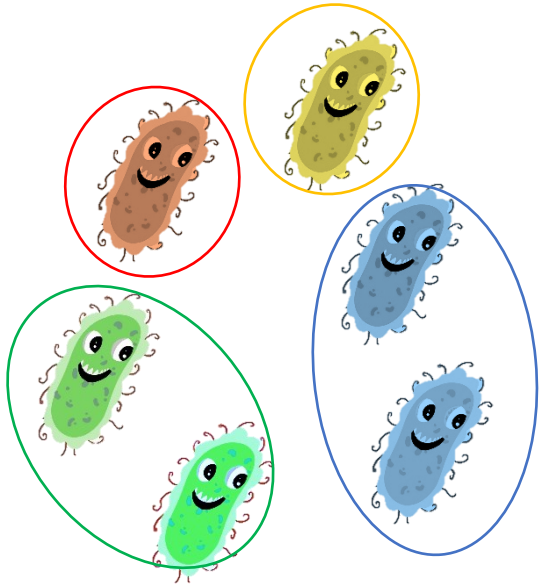
Trimming out adapters/primers

```
@M04743:199:000000000-CGG4F:1:1101:16145:1655 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGAAGGTGGCAAGCGTTGTTCGGATTCACTGGGCGTACAGGGAGCGTAGGCGGTTGGGTAAGCCCTCCGTGAAATCTCCGGG
+
ABCCCFFFCADBGGGGGGGGGGHHGHGGFHGHHHGHGGGAFFHGGGGGHHHHHHHHHGGGGGHHGGGGGGGGGHGGEGGGGGHHHHHHHHGGHGGGHHHHHHGGG
@M04743:199:000000000-CGG4F:1:1101:18938:1729 1:N:0:233
GGTGCCAGCCGCCGCGGTAATACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGCGCAGGCTGTTTTGTAAGTCAGATGTGAAATCCCCGAG
+
BBBBBFFFBBBBGGGGGGGGGFHHHHHGGHGGGGGGGGGGHHGGEGFHHHHHHHHHGGGGHFHGGGGGGGGGGHHHHHHHHHHHHHHHHHHHHHHHHGGGG
@M04743:199:000000000-CGG4F:1:1101:13893:1760 1:N:0:233
GGTGCCAGCAGCCGCGGTACTACGTAGGGTGCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGTTCGTAGGCGGTTTGTCGCGTCGTTTGTGAAAACCCGGGG
+
BBBBBFFFB4CCGGGGGGGGCFFHGHHHGGHGGGGGGGGGAFGHGG?EFHFEHHHHHGGGGFHFHFGHGGHGG3EEEGGGHHEHGGGGGGGDHHEHGHGGGGGG
F9FFFFFFFFFFFEFFBBBBFEB;-@DFB-BBBFFFFEFF/EBBEFFF/BADFFDFFF.;
```

OTU vs ASV clustering

OTU Clusters are generated using a similarity threshold of 97% sequence identity. This approach carries with it the risk that multiple similar species can be grouped into a single OTU, with their individual identifications being lost to the abstract of a cluster.

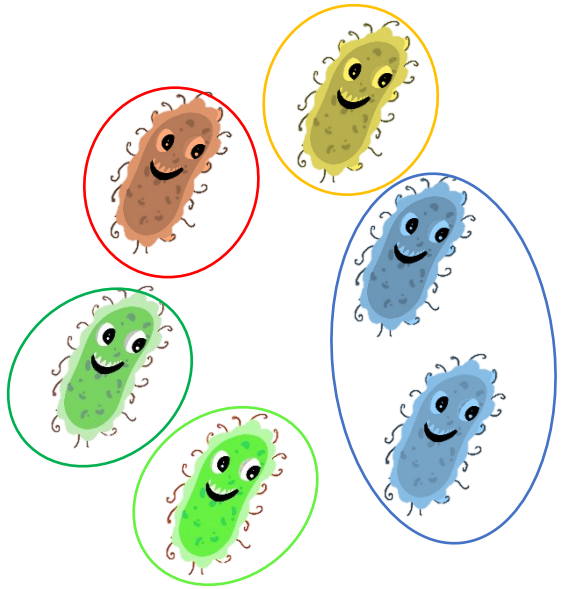
Operational Taxonomic Unit



Output: 4 OTUs

OTU calling is based on similarity and can overlook small biological variations by grouping sequences together.

Amplicon Sequence Variant



Output: 5 ASVs

ASVs can preserve biological sequence variation in output reads.

The ASV approach determine which exact sequences were read and how many times each exact sequence was read. These data will be combined with an error model for the sequencing run, enabling the comparison of similar reads to **determine the probability that a given read at a given frequency is not due to sequencer error**.

How does it works?

These steps are not suitable for Nanopore reads!

Usearch

Read preparation
Assemble paired reads, quality filter, trim lengths, find unique sequences

OTU clustering / denoising
Select OTU sequences

Construct OTU table
Map reads to OTUs to get counts per sample

Quality control
Check OTU sequences and analyze control samples

Diversity and taxonomy analysis
Calculate alpha and beta diversity from OTU table
Predict taxonomy for OTU sequences

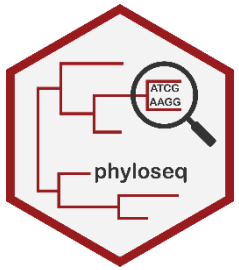
```
#!/bin/bash

# Quality filter
$usearch -fastq_filter ex_min_reads.fq -fastq_maxee 1.0 \
  -relabel Filt -fastaout filtered.fa

# Find unique read sequences and abundances
$usearch -fastx_uniques filtered.fa -sizeout -relabel Uniq -fastaout uniques.fa

# Make 97% OTUs and filter chimeras
$usearch -cluster_otus uniques.fa -otus otus.fa -relabel Otu
```

DADA2
Amplicon Sequencing. **Exactly.**

1. Filter and trim: `filterAndTrim()`
2. Dereplicate: `derepFastq()`
3. Learn error rates: `learnErrors()`
4. Infer sample composition: `dada()`

`)`
`nceTable()`
`Denovo()`

The use of these tools require a good knownledge of programming languages



Diversity and taxonomy analysis

```
errF <- learnErrors(filtFs, nbases = 1e8, multithread = TRUE, randomize = TRUE)
errR <- learnErrors(filtRs, nbases = 1e8, multithread = TRUE, randomize = TRUE)

errF_plot <- plotErrors(errF, nominalQ = TRUE)
errR_plot <- plotErrors(errR, nominalQ = TRUE)

saveRDS(errF_plot, paste0(filtpathF, "/errF_plot.rds"))
saveRDS(errR_plot, paste0(filtpathR, "/errR_plot.rds"))

ggsave(plot = errF_plot, filename = paste0(filtpathF, "/errF_plot.png"),
       width = 10, height = 10, dpi = "retina")
ggsave(plot = errR_plot, filename = paste0(filtpathR, "/errR_plot.png"),
       width = 10, height = 10, dpi = "retina")

mergers <- vector("list", length(sample.names))
names(mergers) <- sample.names
ddF <- vector("list", length(sample.names))
names(ddF) <- sample.names
ddR <- vector("list", length(sample.names))
names(ddR) <- sample.names

# For each sample, get a list of merged and denoised sequences
for(sam in sample.names) {
  cat("Processing:", sam, "\n")
  # Dereplicate forward reads
```

User «friendly» bioinformatic tools

QIIME 2 is a completely re-engineered microbiome bioinformatics platform based on the popular QIIME platform, which it has replaced. QIIME 2 facilitates comprehensive and fully reproducible microbiome data science, improving accessibility to diverse users by adding multiple user interfaces

mothur is an open-source software package for bioinformatics data processing and it is capable of processing data generated from several DNA sequencing methods including 454 pyrosequencing, Illumina HiSeq and MiSeq, Sanger, PacBio, and IonTorrent. The first release of mothur occurred in 2009

EPI2ME Labs is a bioinformatics notebook environment and will work with sequence data from Flongle, MinION, GridION and PromethION.

Custom pipeline: how to?

To validate a tool, we need:

A known da
DNA reads

1) *In silico* generated dataset;
2) *In vitro* generated dataset.

Mock communities

Comparison wit
already validate

Let's try validate a custom tool called «Wooey»

A common database to reduce the variabilty in the analysis

Statistic and graphical visualization

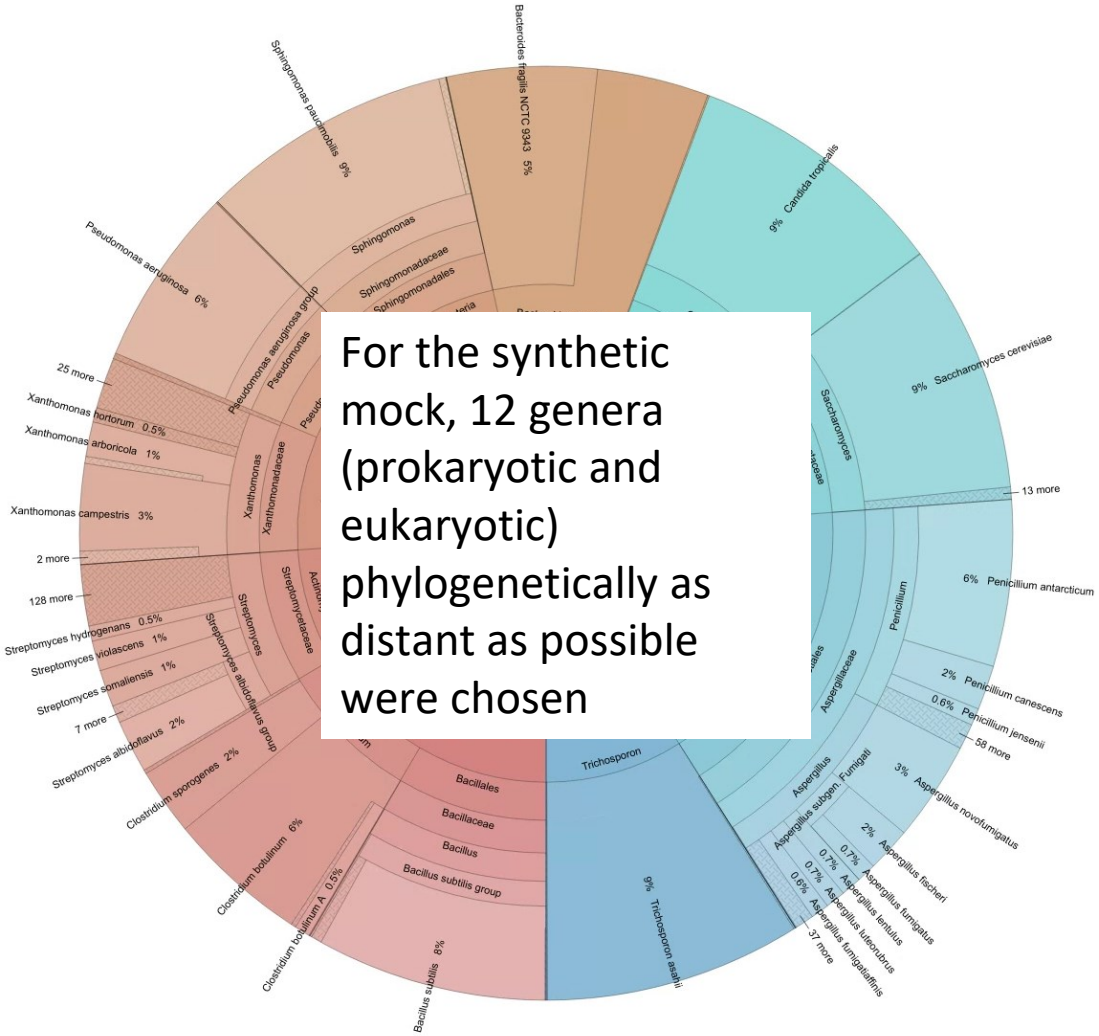
Bioinformatic analysis for reads classification

Sequence data manipulation and statistical analysis

What is a mock community?

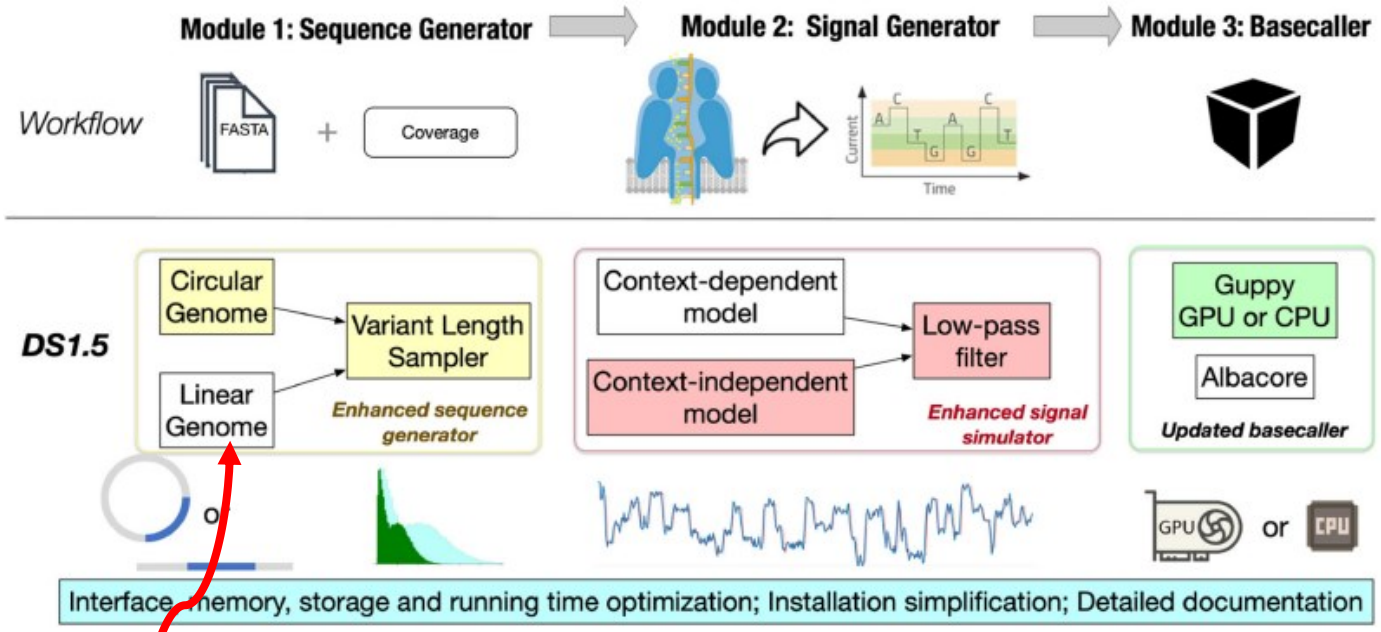
Mock community: A defined mixture of microbial cells and/or viruses or nucleic acid molecules created *in vitro* to simulate the composition of a microbiome sample or the nucleic acid isolated therefrom.

Genus simulated dataset (12 genera): *Aspergillus, Bacillus, Bacteroides, Candida, Clostridium, Penicillium, Pseudomonas, Saccharomyces, Sphingomonas, Streptomyces, Trichosporon, Xanthomonas.*



For the synthetic mock, 12 genera (prokaryotic and eukaryotic) phylogenetically as distant as possible were chosen

In silico mock: simulating Nanopore sequencing with DeepSimulator



FastQ from simulated sequencing

Manually added header for downstream analysis

```
@Aspergillusa32211ab-3185-4f05-a046-9d58231492a2 runid=c2d19c21188
TTCAGGTGAACCTGCGGAAGGATCATTACCGAGTGAGGGCCCTCGGGTCAACCTCCCACCCGTGTC
+
($((.**-/5440022817+2464')+*05*%))+*13-293,,*2+)+()))),72+077,-*.0
&($)&'&&%%+++**&(&$)(-63)-.0.677552(%&)-/95/151),+*../&22+(0./.383
@Aspergillusb9f43e95-18cf-43f9-a41a-83bdb306eb5e runid=c2d19c21188
CTACAATAATTGCTGGGGATAGGCATTTACAGTTATTGCTCTTCAACGAGGAATGCTAGTAGGCTT
+
$"$$)*+((*++.3216+.1'(%($)0&&*&%&'&-('(-.2:8402-)02-32..-+&'')-'%&
116543253;82.1523)$&'(24:422,5-6.3038/22.3:2-,,.,53/36+,-,1++,-.-+
131.*-0341-)'*0(.(.0('/2+0,/**-1/,.,..-./-2-3,..(+(+4,*,./*0/0/65'
(%%(,32*03++-+)1))**)4014-6.25--,%+0555711:.,055495364602276/6'0)/
214742/30)+..*/+00-,3/,.'-)64604441,.+.*+,%('/-1.()812003)+/,(()3
@Bacillus15b27301-16b0-4cd4-923b-304683548912 runid=c2d19c211888bc
ATGTAAGACTGGGATAACTCCGGGAAACCGGGGCTAATACCGGATGGTTGTTTGAACCGCATGGTT
+
#(&&*/-,5649632493/4&-..*///4-31...242'/*001632,714;846742&.&,).04
()'&)*.&*+55511,,,*1,(*)))+,0331097+,4-12(*(2.&,--))$%#&&*-5.52075
```

Shell script (part of) for simulating reads:

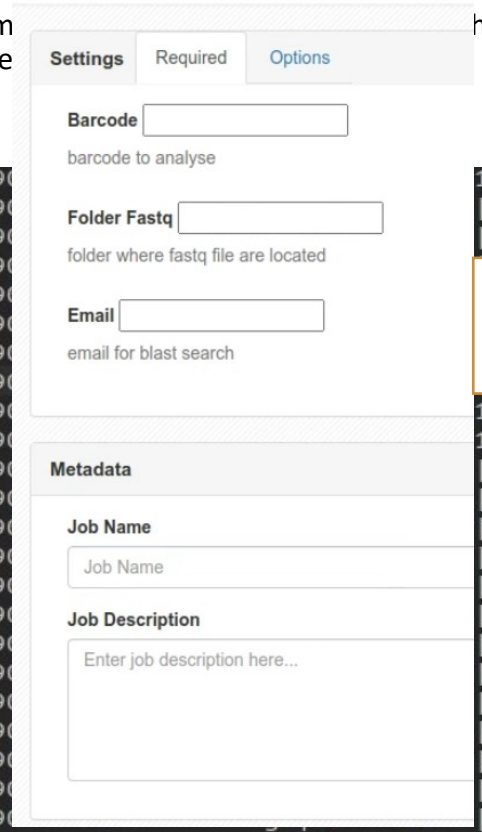
bash ./deep_simulator.sh -i genome.fasta -n numberofreads -B 1 -o outputdir -c numberofCPUs -l readlenght

Fasta generated genomes

Wooey: the tool to be validated [1]

Wooey aligns (using blast algorithm) the reads against the database that is provided.

Correction step (optional): all reads are aligned vs all reac ITS_orignalConsesus nce is genera None conser alignm he datase



The best alignment generates the taxonomy assignment to the examined read.

Called species

Alignments are ranked by **scores** and **e-values**.

Score

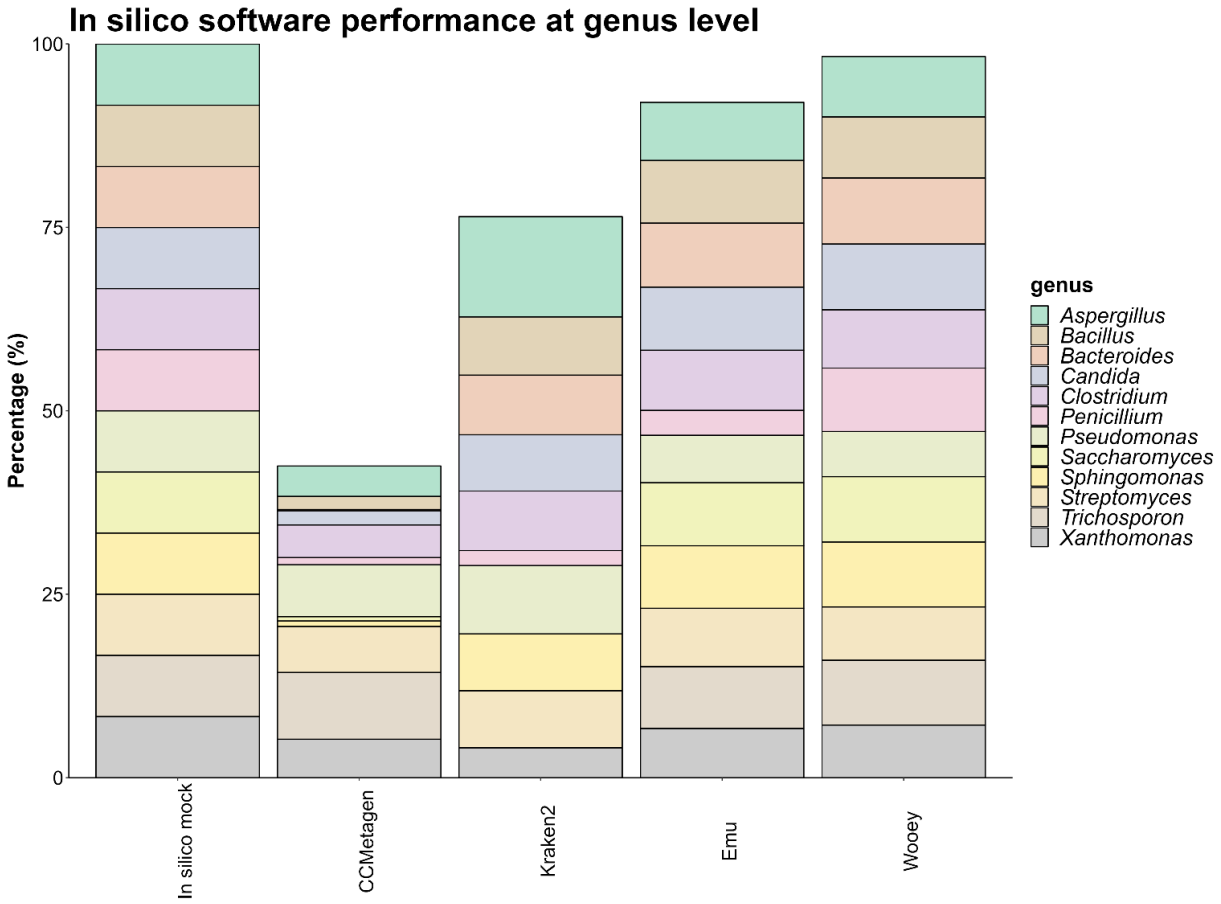
Reads

Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	1| 843	Aspergillus laciniosus	97.222	0.0	388753	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 843	Aspergillus spinosus	97.222	0.0	36631	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 843	Aspergillus novofumigatus	97.222	0.0	340412	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59		akii	97.222	0.0	2079208	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59		iniosus	97.222	0.0	388753	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59		ofumigatus	97.222	0.0	340412	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59		cheri	97.222	0.0	36630	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59		tulus	97.030	0.0	293939	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	1| 839	Aspergillus lentulus	97.030	0.0	293939	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	1| 839	Aspergillus lentulus	97.030	0.0	293939	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	fungal sp. AM0077	97.030	0.0	1578594	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus spinosus	97.030	0.0	36631	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus lentulus	97.030	0.0	293939	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus lentulus	97.030	0.0	293939	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus fischeri	97.030	0.0	36630	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus spinosus	97.030	0.0	36631	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus lentulus	97.030	0.0	293939	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus sp. 97.030	0.0	5065	503	
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus fumisynnematus	97.030	0.0	286432	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus lentulus	97.030	0.0	293939	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus botucatensis	97.030	0.0	1907480	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 839	Aspergillus fischeri	97.030	0.0	36630	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59	| 837	Aspergillus spinosus	97.024	0.0	36631	503
Aspergillus196257e6-3fe4-4e5a-b15e-b6e59033449drr	emb|HE578059.1| 833	Aspergillus fumigatiaffinis	96.832	0.0	340414	503

It is very user friendly!

[1] the Wooey tools has been created by prof. Faino Luigi, Università di Roma «La Sapienza»

Validation of the tool *in silico:* comparison with already validate tools



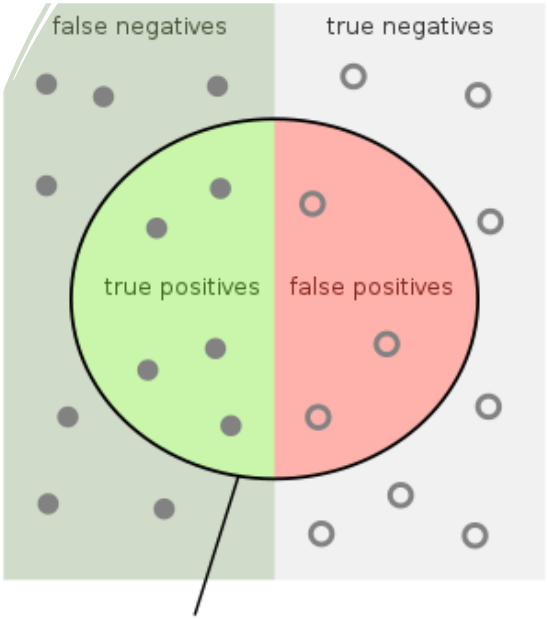
Kraken2 (https://github.com/DerrickWood/kraken2) *(Least Common Ancestor: LCA)*: while Kraken 1 used a sorted list of *k*-mer/LCA pairs indexed by minimizers, Kraken 2 introduces a probabilistic, compact hash table to map minimizers to LCAs.

CCMetagen (https://github.com/vrmarcelino/CCMetagen): processes sequence alignments produced with KMA (k-mer alignment), which implements the ConClave sorting scheme to achieve highly accurate read mappings.

Emu (https://gitlab.com/treangenlab/emu): is a homology-aware alignment likelihood approach in which read classification probabilities are adaptively updated based on read alignments to multiple reference sequences and the current community profile estimate.

Validation of the tool *in silico*

- Precision: True Positive / (True Positive + False Positive)
- Recall: True Positive / (True Positive + False Negative)
- F-score: (2 x precision x recall) / (precision + recall)



My dataset

Reads identity
Pseudomonas
Bacillus
Bacteroides
Aspergillus

Reads identity	Called taxonomy	Results
Pseudomonas	*Pseudomonas*	*True Positive*
Pseudomonas	*Bacillus*	*False Positive*
Pseudomonas	*Clostridium*	*False Negative*

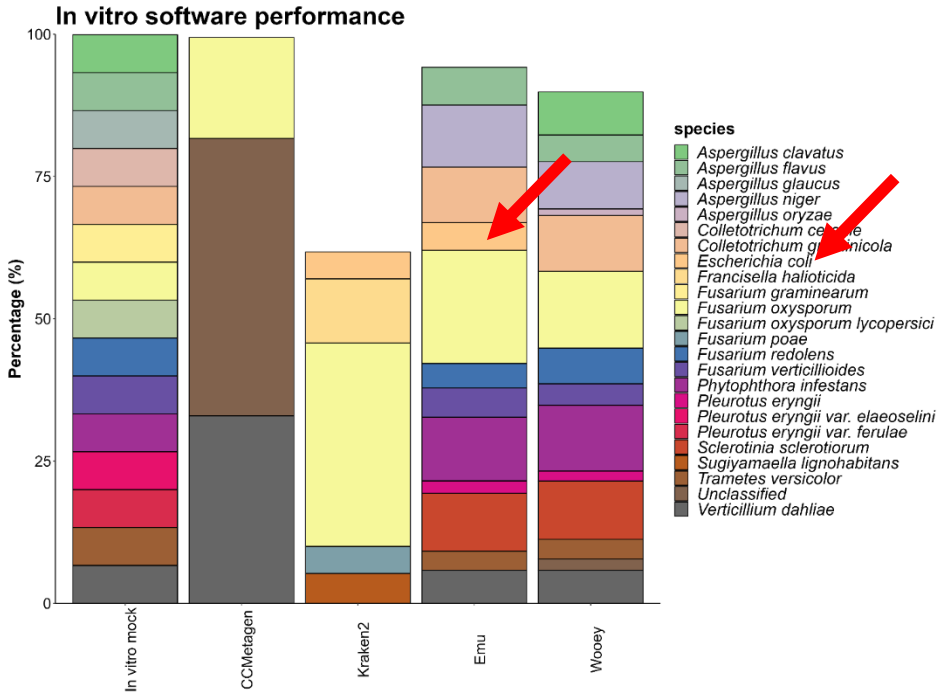
In silico dataset with 120,000 reads

Software	N. of reads aligned	Precision	Recall	F-score	True Positive	False Positive	False Negative
Wooey_genus	116,994	1.00	0.93	0.97	109,281	99	7,614
Emu_genus	22,609	0.99	0.92	0.96	20,662	158	1,788
Wooey_species	119,718	1.00	0.56	0.72	66,882	0	52,836
Emu_species	10,948	1.00	0.21	0.35	2,337	0	8,610

Wooey software has better F-score than Emu, especially at species level and Wooey classify more reads than Emu, which is more conservative.

In vitro mock construction and validation

Eukaryotic mock: DNA has been first amplified with ITS1 (Fw) ITS4 (Rv) primers for each species, then mixed together to have the same concentration in the mock.



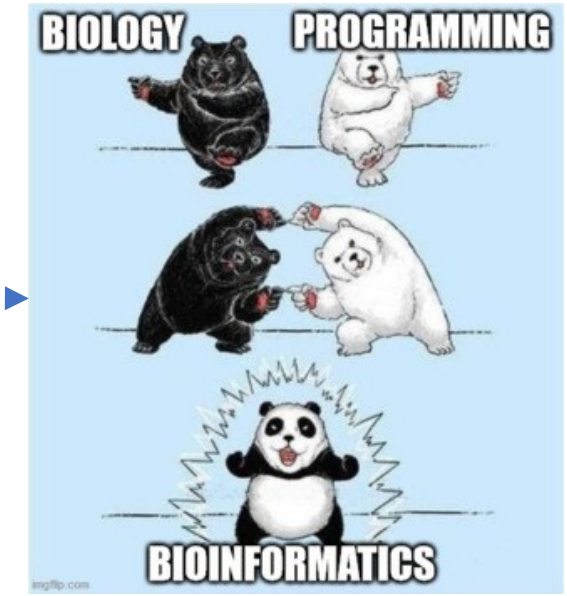
Mock species	Emu	Wooey
Aspergillus clavatus	*Aspergillus clavatus*	*Aspergillus clavatus*
Aspergillus flavus	*Aspergillus flavus*	*Aspergillus flavus*
Aspergillus glaucus	*Aspergillus*	*Aspergillus*
Trametes versicolor	*Trametes versicolor*	*Trametes versicolor*
Fusarium graminearum	*Fusarium*	*Fusarium*
Pleurotus eryngii var. ferulae	*Pleurotus eryngii*	*Pleurotus eryngii var. ferulae*
Pleurotus eryngii var. elaeoselini	*Pleurotus eryngii*	*Pleurotus eryngii*
Fusarium oxysporum	*Fusarium oxysporum*	*Fusarium oxysporum*
Phytophtora infestans	*Phytophtora infestans*	*Phytophtora infestans*
Fusarium verticillioides	*Fusarium verticillioides*	*Fusarium verticillioides*
Fusarium oxysporum lycopersici	*Fusarium oxysporum*	*Fusarium oxysporum lycopersici*
Verticillium dahliae	*Verticillium dahliae*	*Verticillium dahliae*
Colletotrichum cereale	*Colletotrichum cereale*	*Colletotrichum cereale*
Colletotrichum graminicola	*Colletotrichum graminicola*	*Colletotrichum graminicola*
Fusarium redolens	*Fusarium redolens*	*Fusarium redolens*
Sclerotinia sclerotium	*Sclerotinia sclerotium*	*Sclerotinia sclerotium*

Take home message

There are plenty of bioinformatic tools for taxonomy classification of microbial communities

Some tools require knowledge of programming languages

Others tools are more user friendly



Validation of a custom tool requires both *in silico* and *in vitro* tests

Statistic expressed as ***F-score*** *is an importan value to consider*

The tool presented here showed better F-score than other validated tools

Even though the tool can reach sub-species levels, the high number of False Positives should be fixed

Should bioinformatic be more user friendly?