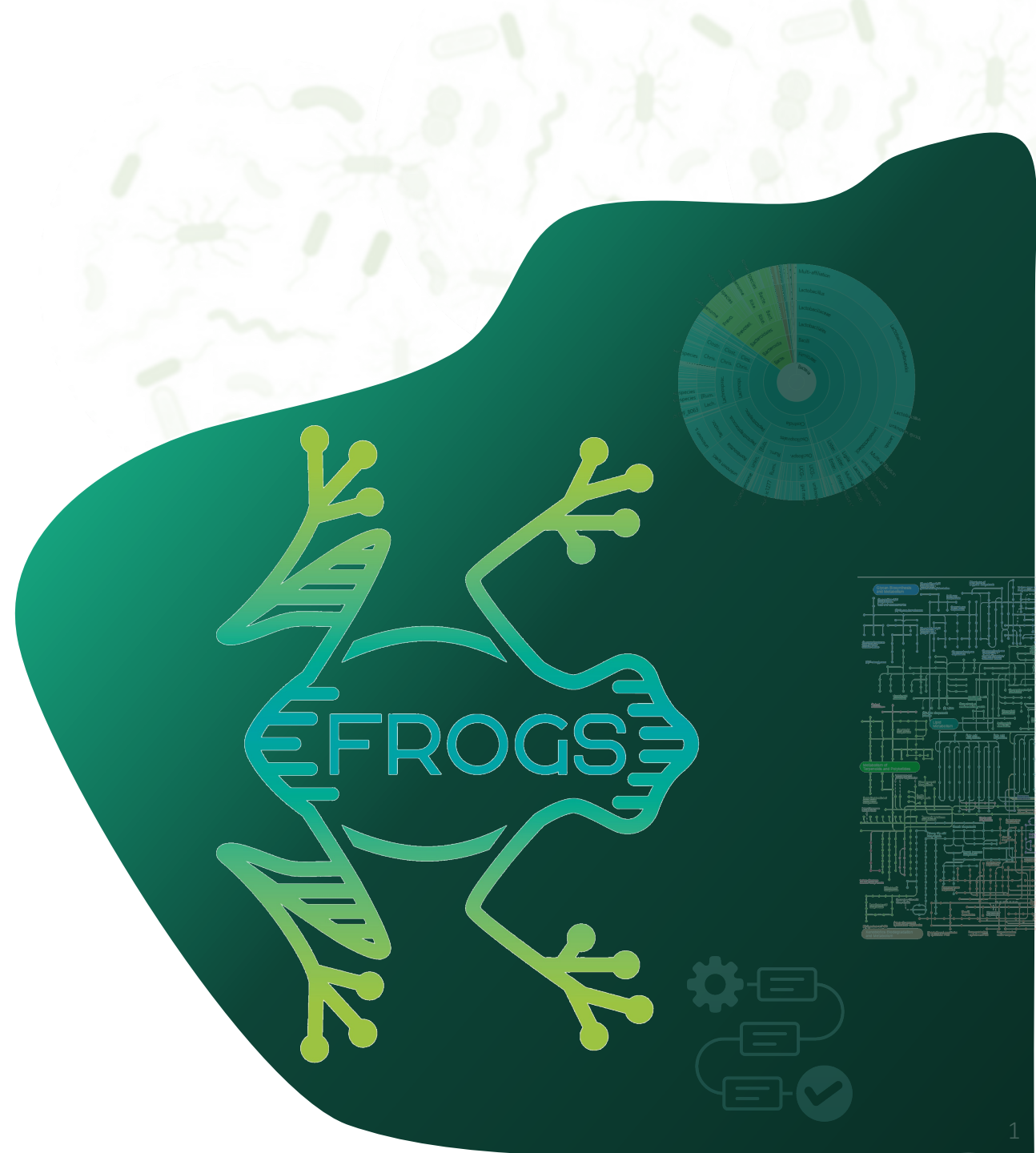


Metabarcoding overview

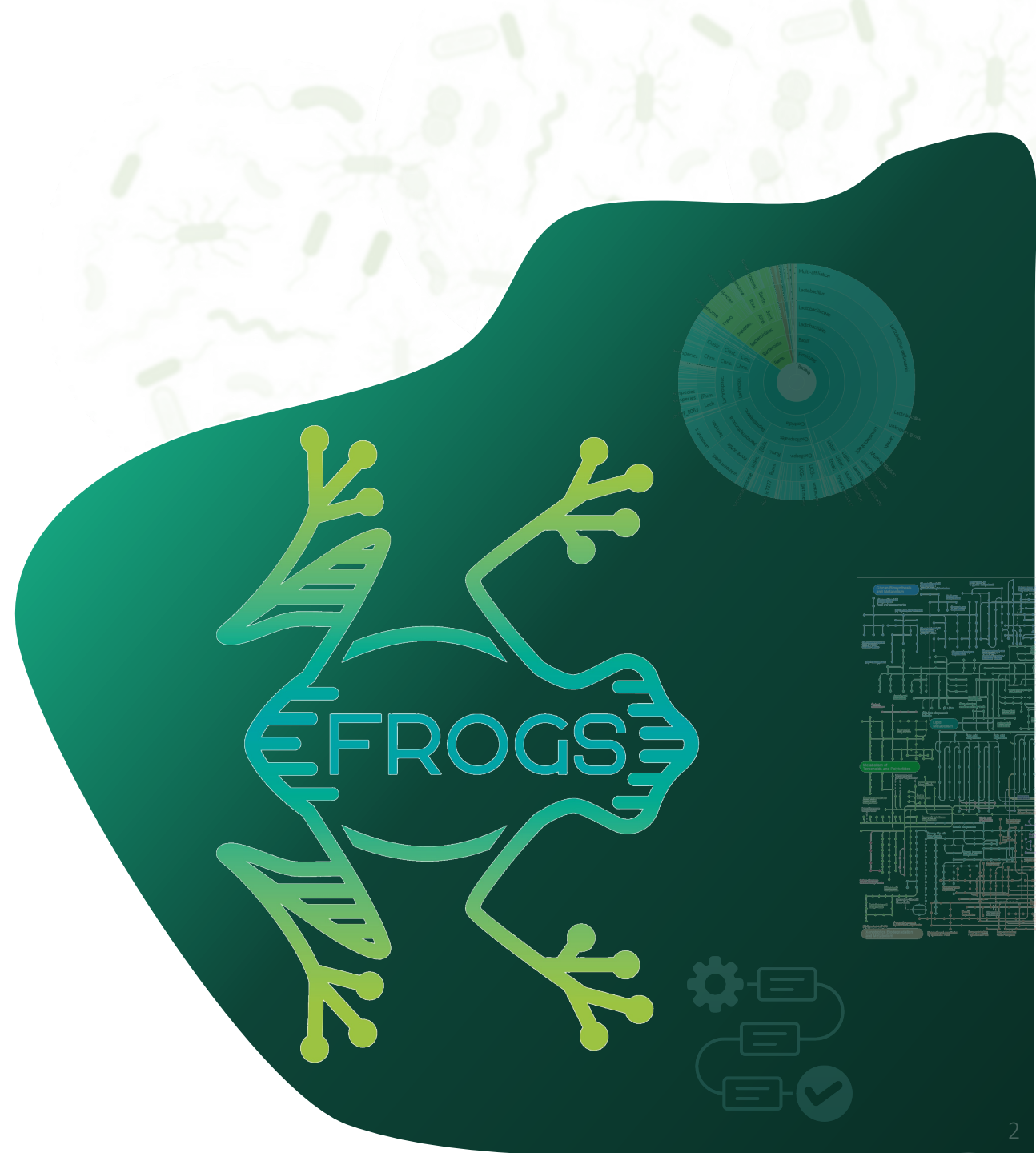
Key concepts

Lucas Auer, Gabryelle Agoutin,
Maria Bernard, [Géraldine Pascal](#),
Maëlle Pomiès & Olivier Rué

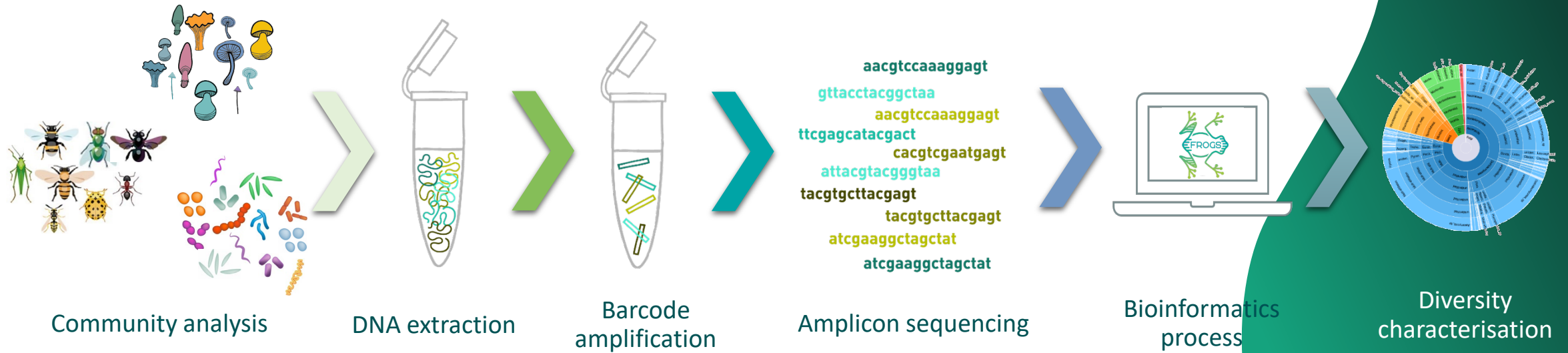


Metabarcoding overview

Barcode

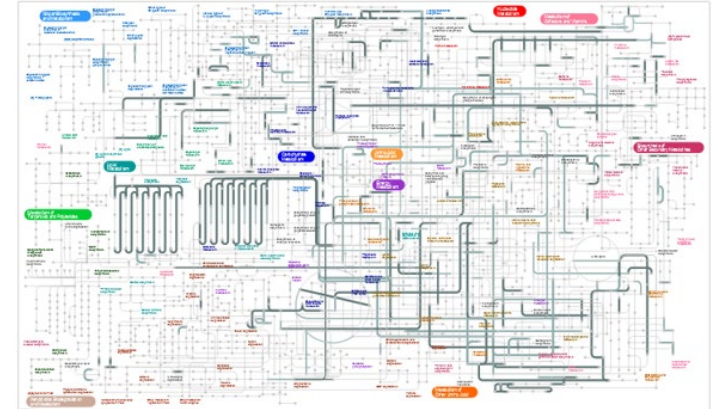
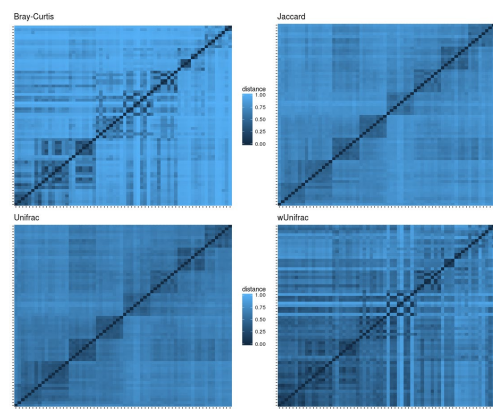
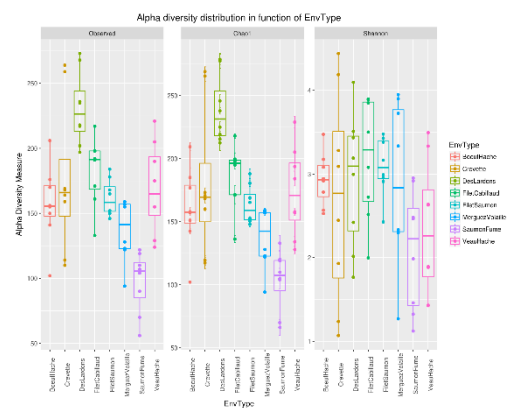


Objectives

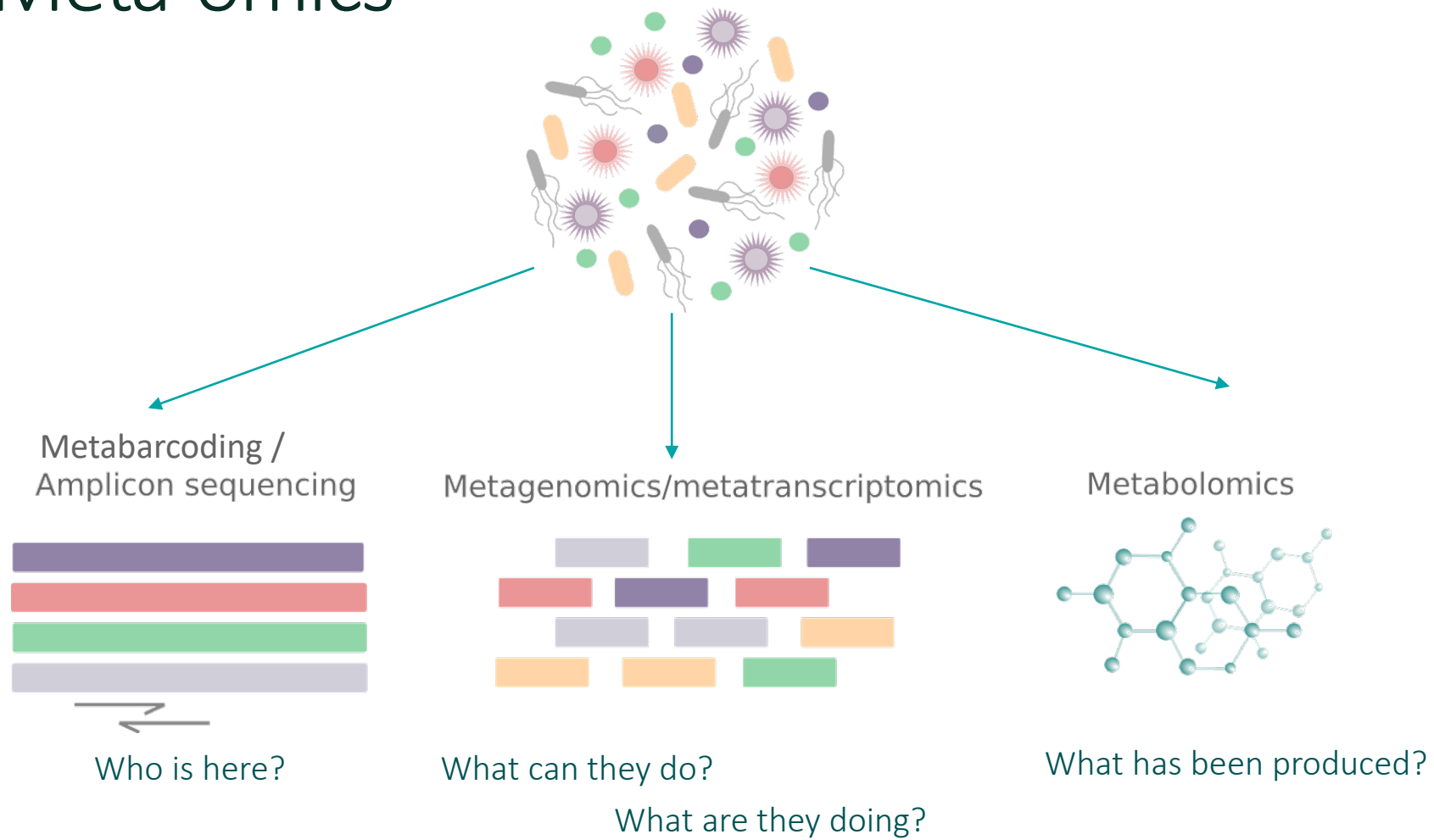


Objectives: a count table for statistics analysis

	Affiliation	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
ASV1	Species A	0	100	0	45	75
ASV2	Species B	741	0	456	4421	1255
ASV3	Species C	12786	45	3	0	0
ASV4	Species D	127	4534	80	456	756
ASV5	Species E	8766	7578	56	0	0



Meta-omics

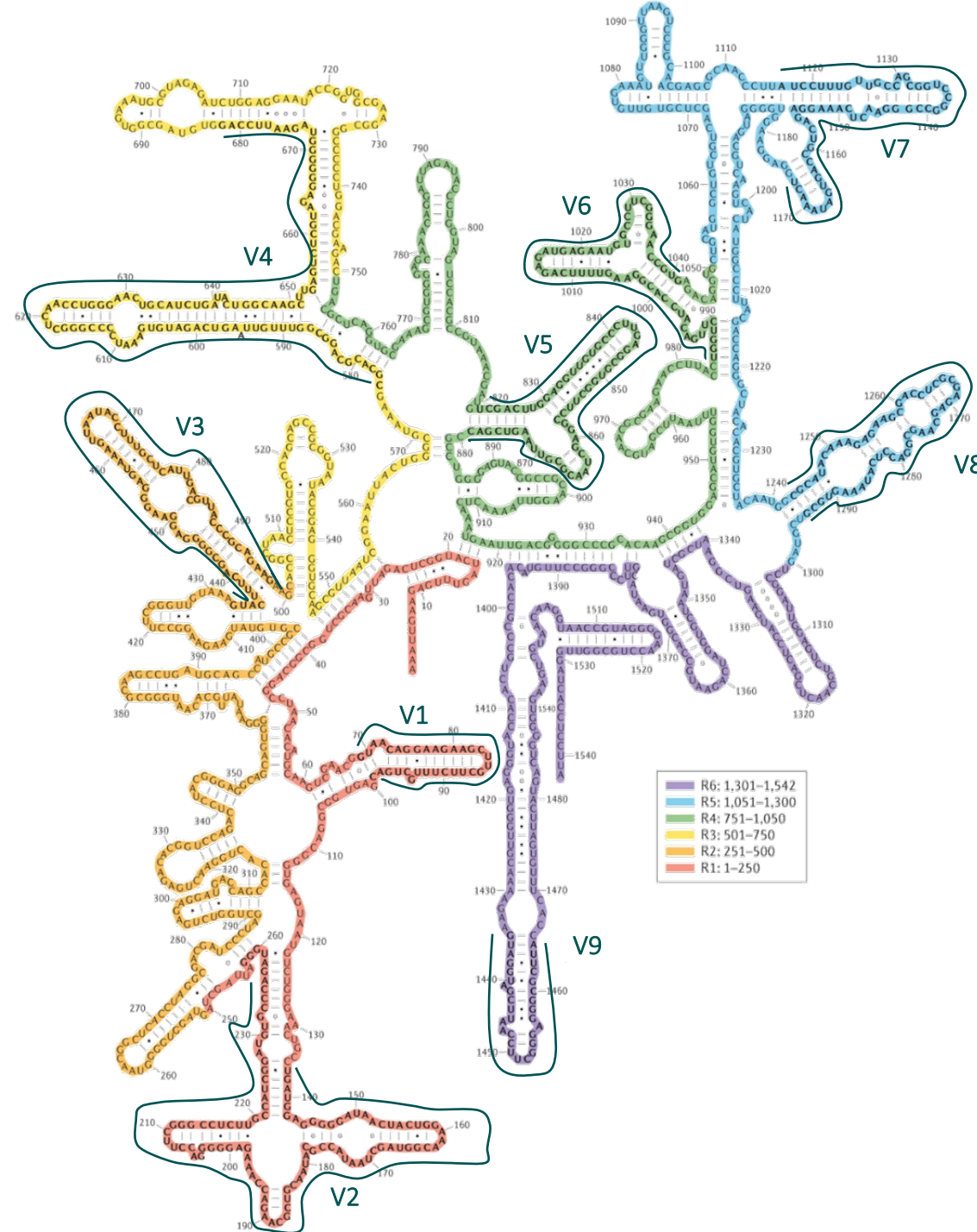


Choose the marker according to the ecological question

Goal	Recommended marker
Microbiome	16S / rpoB / gyrB
Mycobiome	ITS1 / ITS2
Whole eukaryotic community	18S
Animals (invertebrates)	COI
Plants	rbcL / nifD
degraded DNA / sediment / plants	Chloroplast trnL P6 loop

Secondary structure of
the 16S rRNA of
Escherichia coli

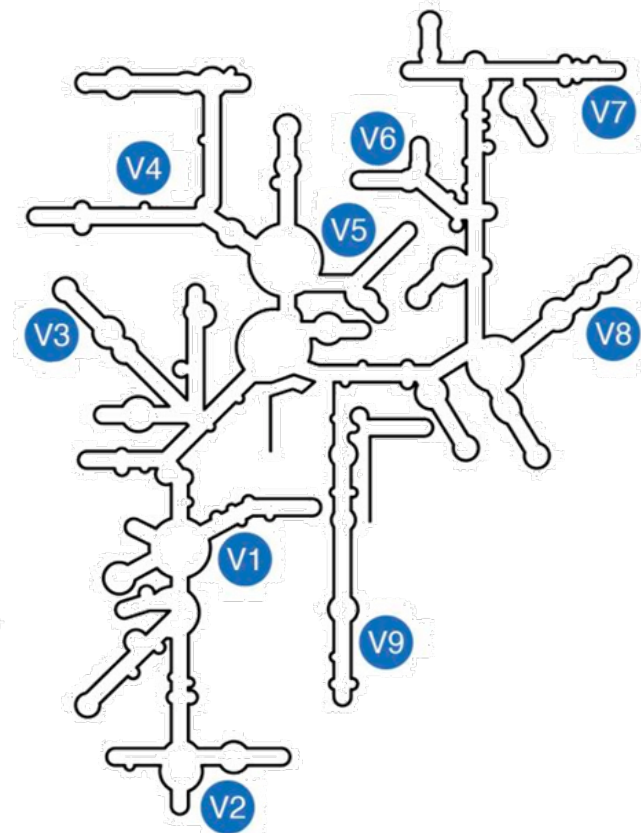
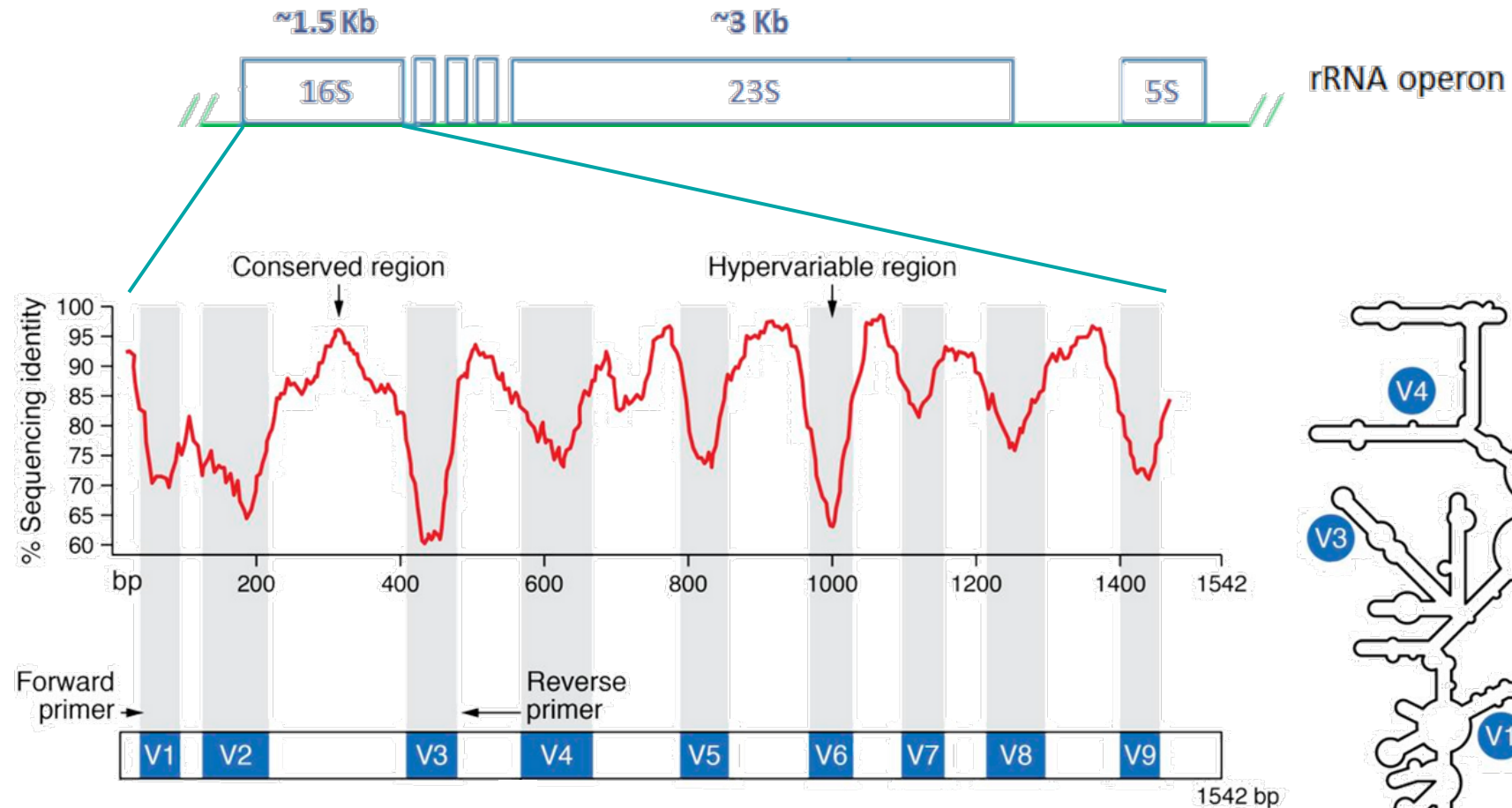
9 variable regions



16S

- Ubiquist gene
- No lateral gene transfer
- Molecular phylogenetic marker
- Availability of databases
GTDB_220 (2024) 863832
SILVA_138.2 (2024) 451555

16S rRNA structure

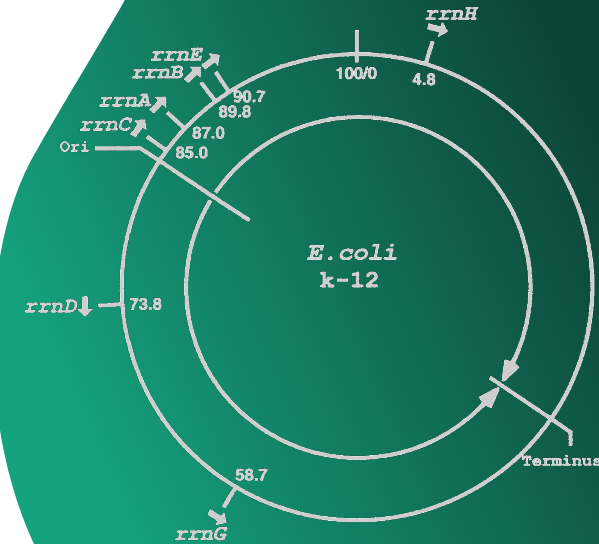
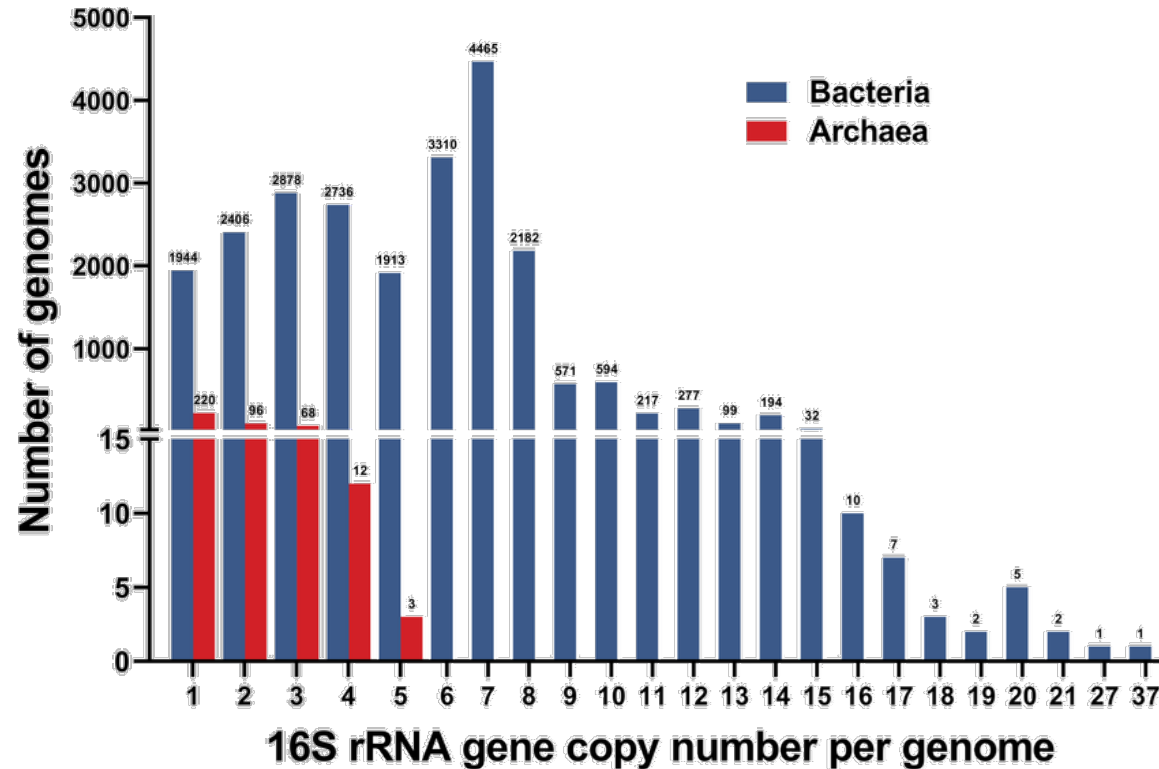


Explore this database:
<https://rrndb.umms.med.umich.edu/genomes/>



Variations of 16S rRNA gene copy number per genome

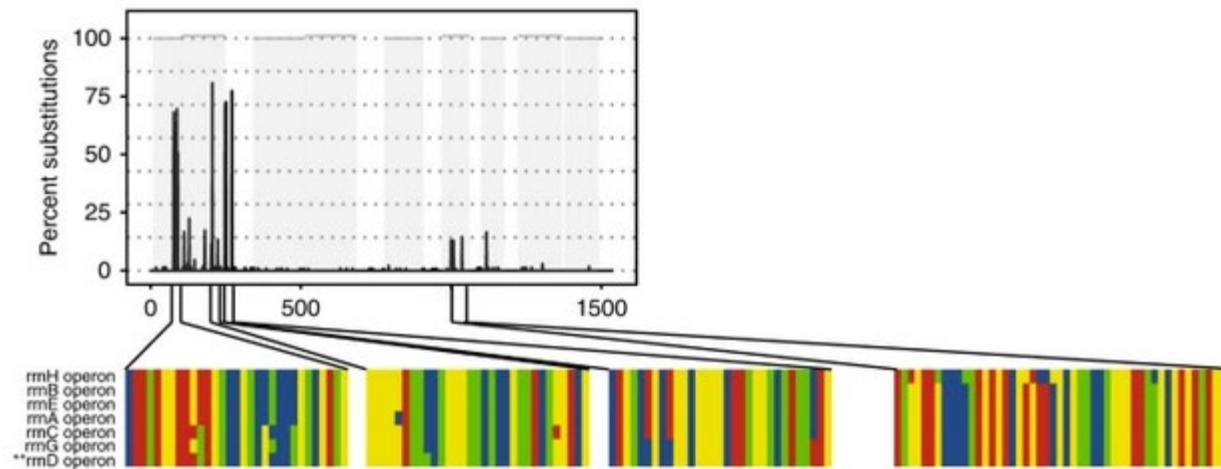
Several 16S rRNA genes in genomes



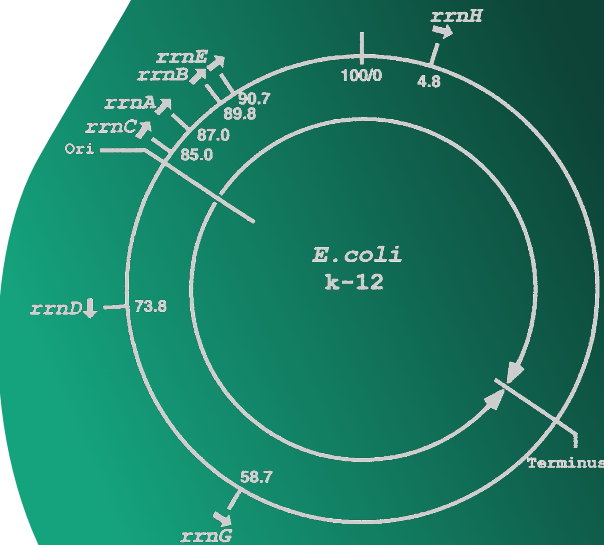
doi/10.1073/pnas.96.5.1820

Intragenomic variations of 16S rRNA genes

Example: *E. coli* strain K-12 MG1655



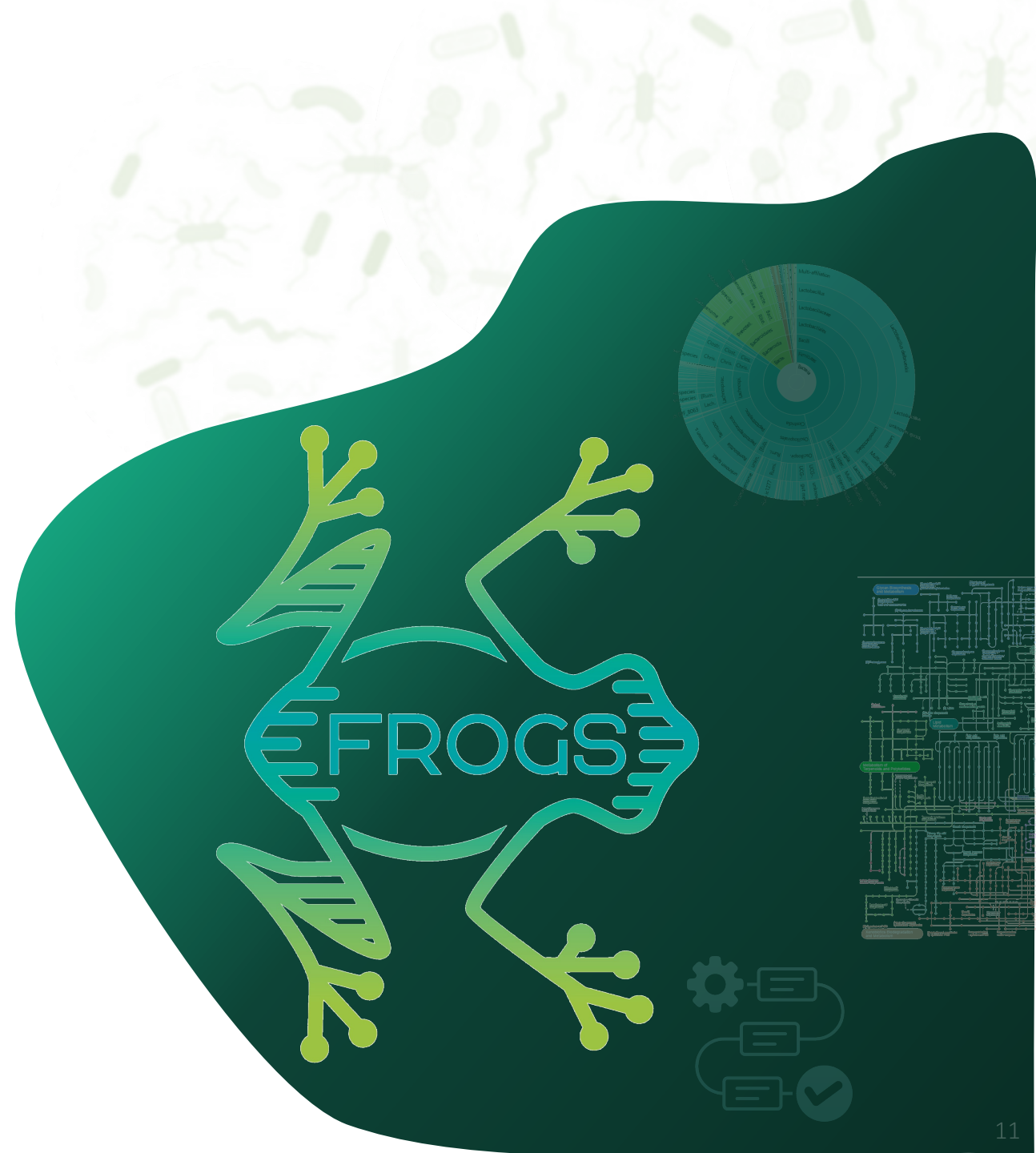
<https://doi.org/10.1038/s41467-019-13036-1>



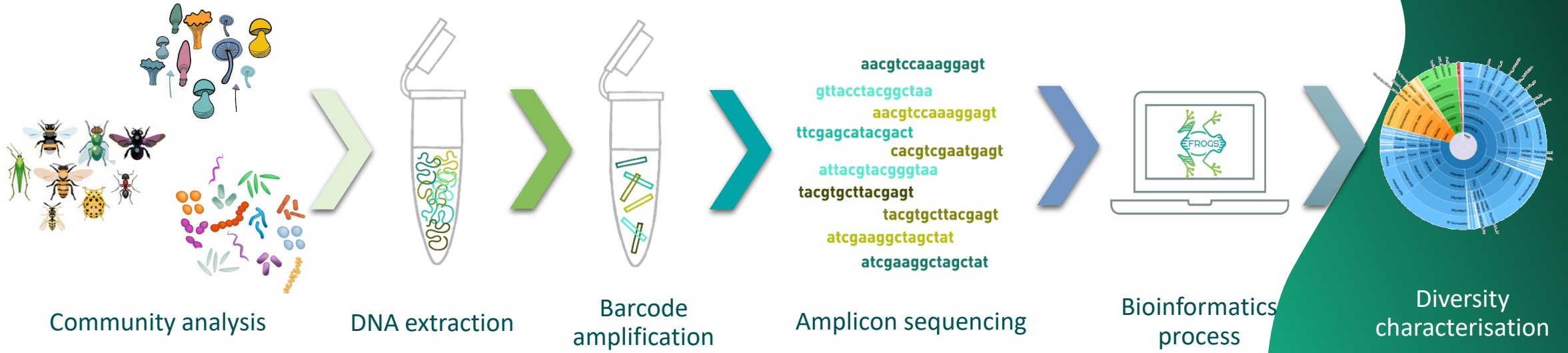
[doi/10.1073/pnas.96.5.1820](https://doi.org/10.1073/pnas.96.5.1820)

Metabarcoding overview

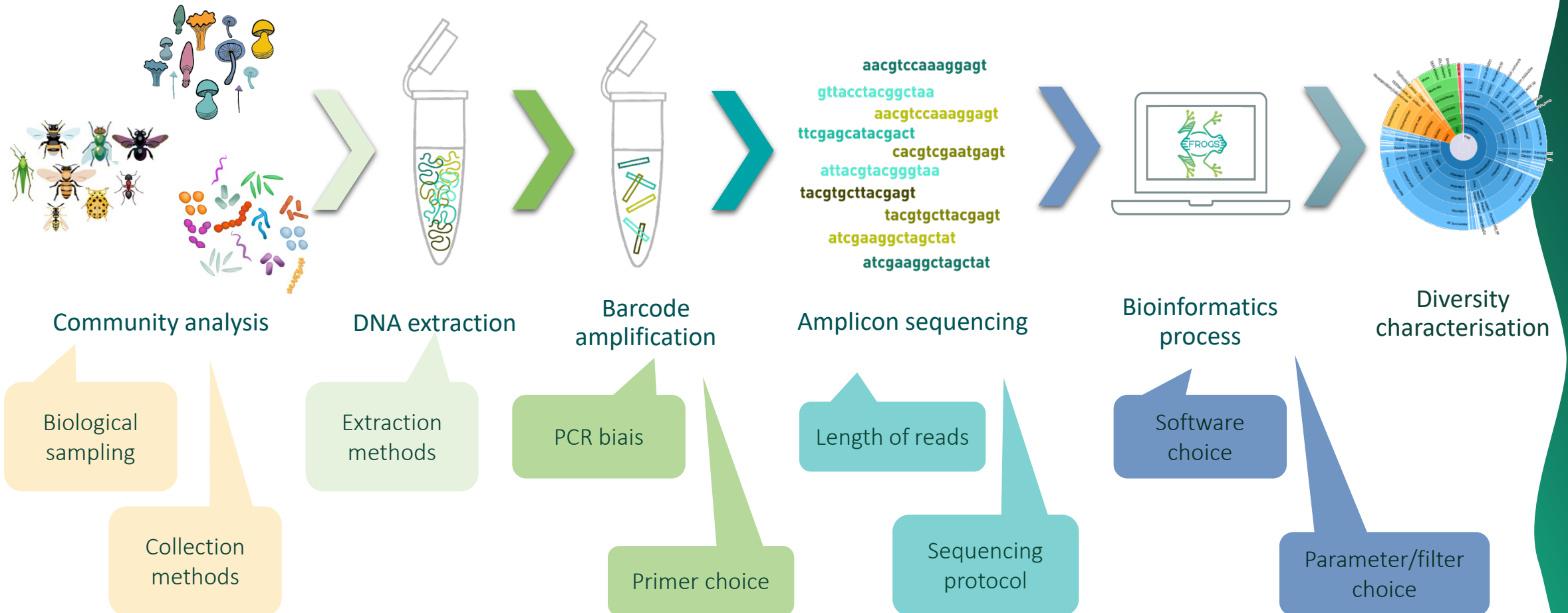
Technical
sources of noise
and errors



Metabarcoding analysis steps

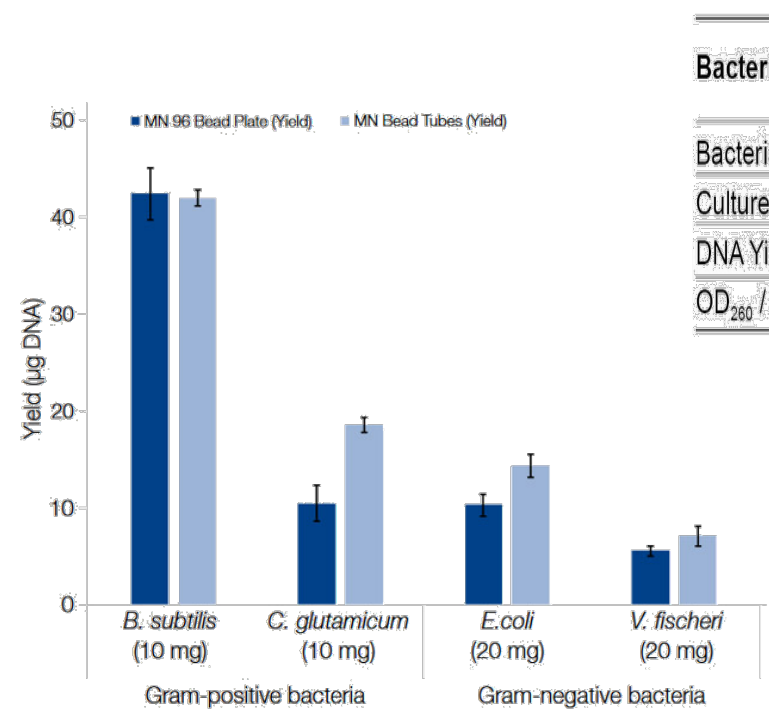


Noise/errors are inevitable at every stage.



Extraction

Not all microorganisms have the same yield in terms of DNA extraction.
This depends on the protocols used!



Bacteria Type	Gram-negative bacteria (such as <i>E.coli</i>)	Gram-positive bacteria (such as <i>Glucococcus epidermidis</i>)
Bacteria Concentration	2×10^8 cells/ml	3.5×10^8 cells/ml
Culture Volume	1 ml	1 ml
DNA Yield	15-20 µg	6-13 µg
OD ₂₆₀ / OD ₂₈₀	1.7-1.9	1.7-1.9

https://en.tiangen.com/content/details_43_4220.html



Amplification

PCR polymerase fidelity: What percent of the product molecules contain an error after PCR (30 cycles) with different polymerases?

Polymerase	1 kb template	3 kb template
Phusion High-Fidelity DNA Polymerases (HF Buffer)	1.32%	3.96%
Phusion High-Fidelity DNA Polymerases (GC Buffer)	2.85%	8.55%
<i>Pyrococcus furiosus</i> DNA polymerase	8.4%	25.2%
<i>Taq</i> DNA polymerase	68.4%	205.2%

After 30 cycles of PCR amplifying a 3 kb template, only 3.96 % of the product DNA molecules contain 1 (nucleotide) error each. This means that 96.04 % of the product molecules are entirely error-free. In contrast, after the same PCR protocol performed with *Taq* DNA polymerase, every product molecule contains an average of 2 errors.

Amplification polymerase choice

Length of PCR product in bp

Fidelity value of DNA polymerase

☐ Phusion High-Fidelity DNA polymerase (HF Buffer; fidelity 4.4×10^{-7})

☐ Phusion High-Fidelity DNA polymerase (GC Buffer; fidelity 9.5×10^{-7})

☐ *Pyrococcus furiosus* DNA polymerase (2.8×10^{-6})

☒ Taq DNA polymerase (2.28×10^{-5})

Number of PCR cycles

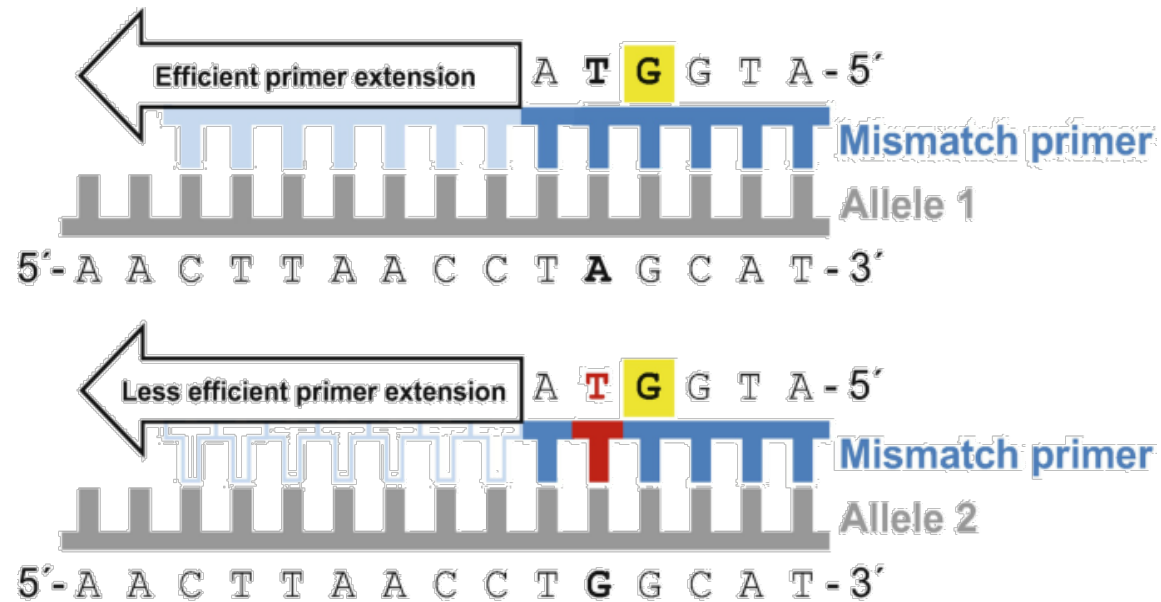
Estimated percentage of PCR products having an error (i.e., DNA molecules with 1 error):

Length of PCR product in bp:	500
Number of PCR cycles:	30

34.2

Amplification: efficiency

Not all microorganisms have the same amplification efficiency in PCR (depending on the primers). They may not even be amplified at all.

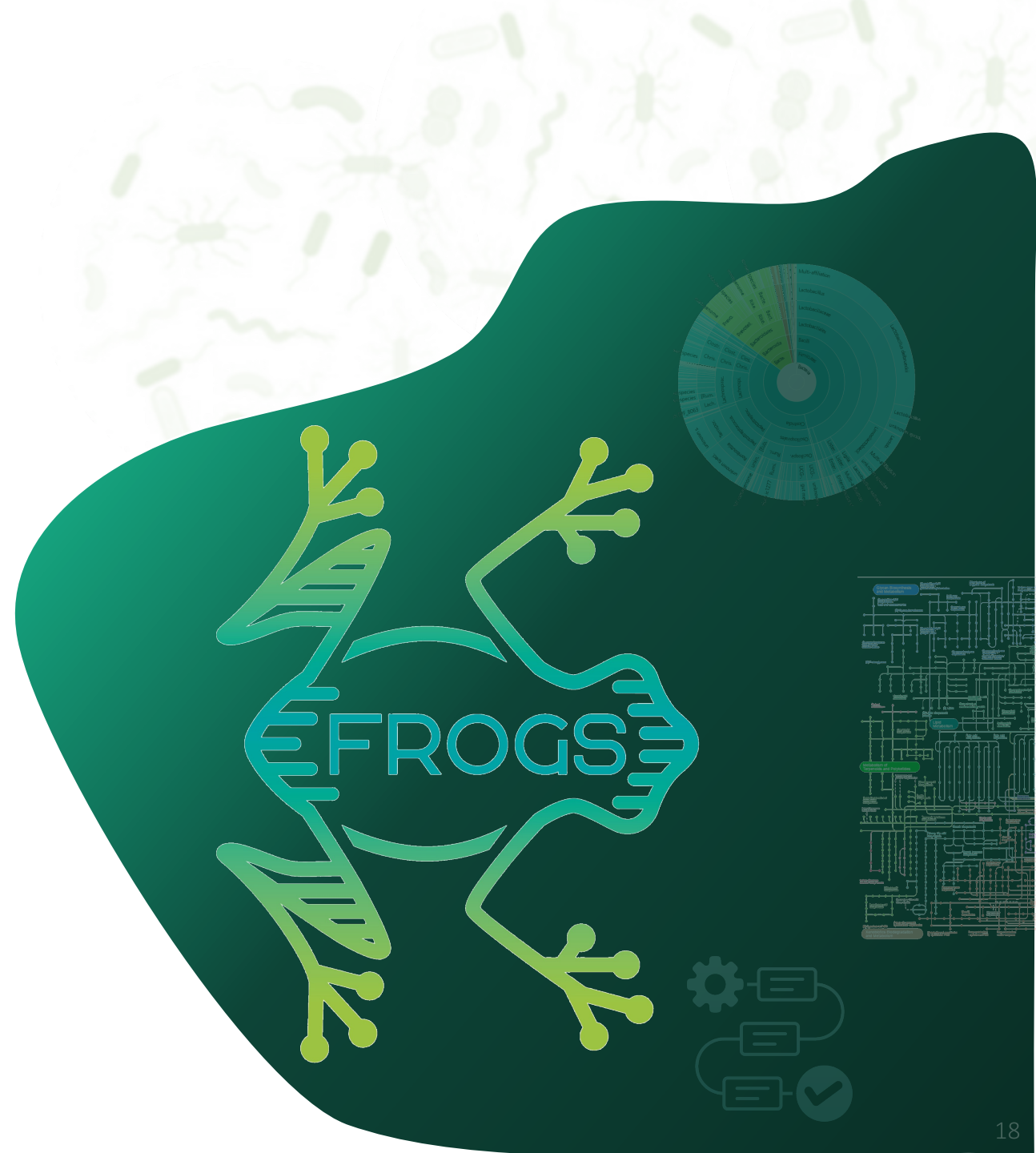


https://doi.org/10.1007/978-1-0716-1799-1_5



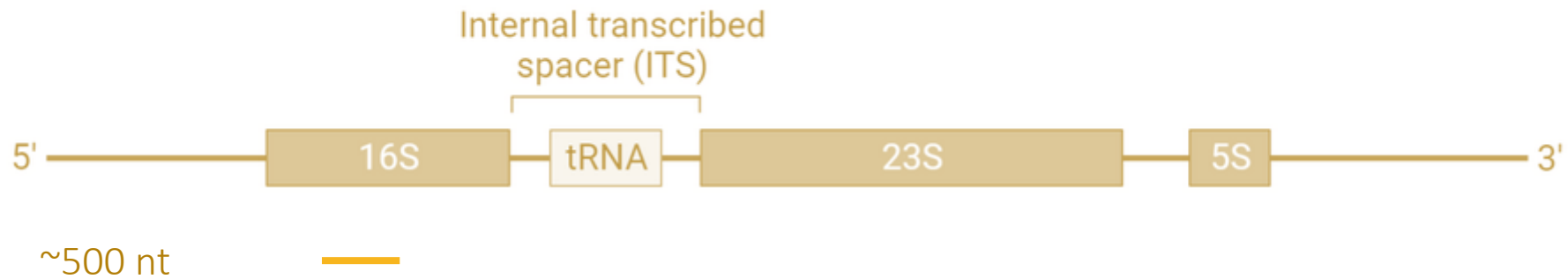
Metabarcoding overview

Short reads



Short reads

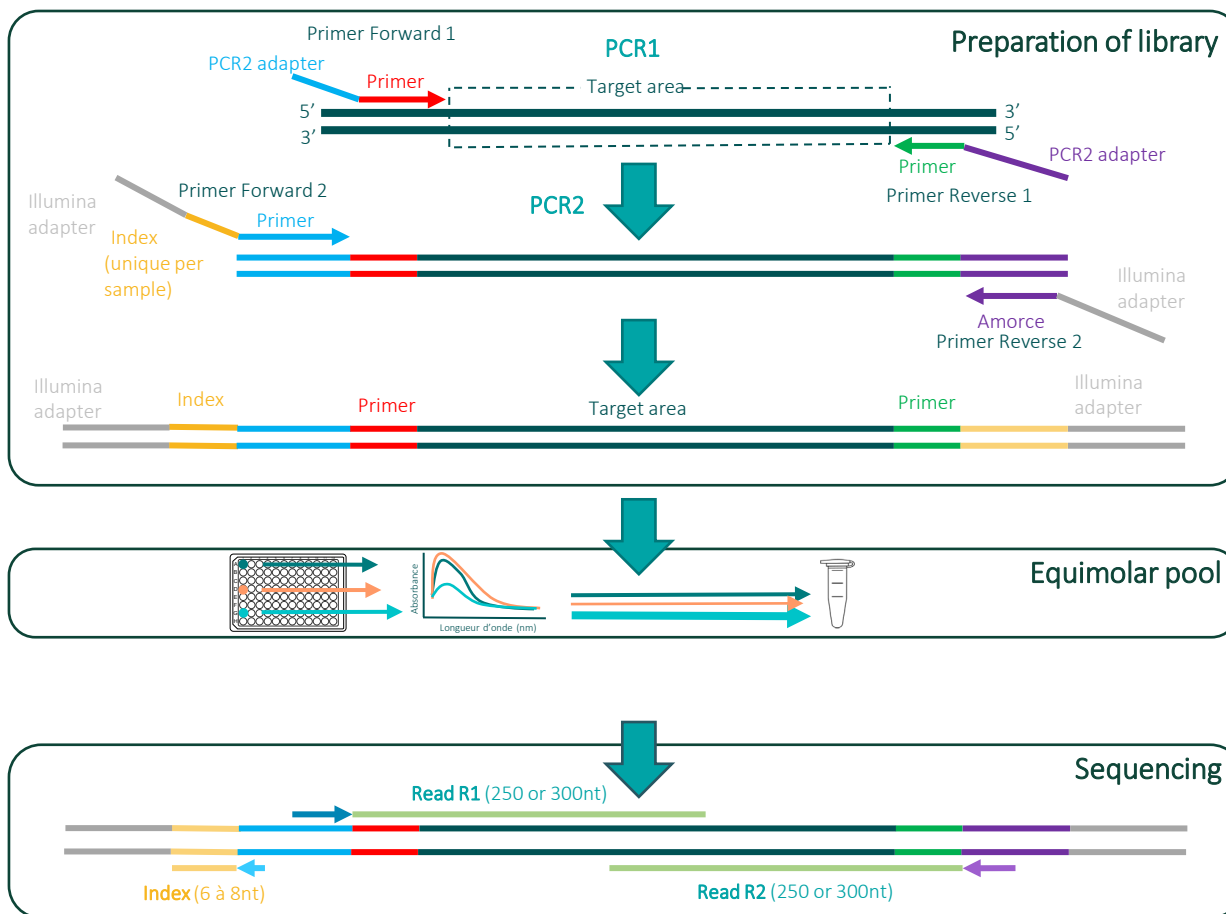
Bacteria rRNA gene organization



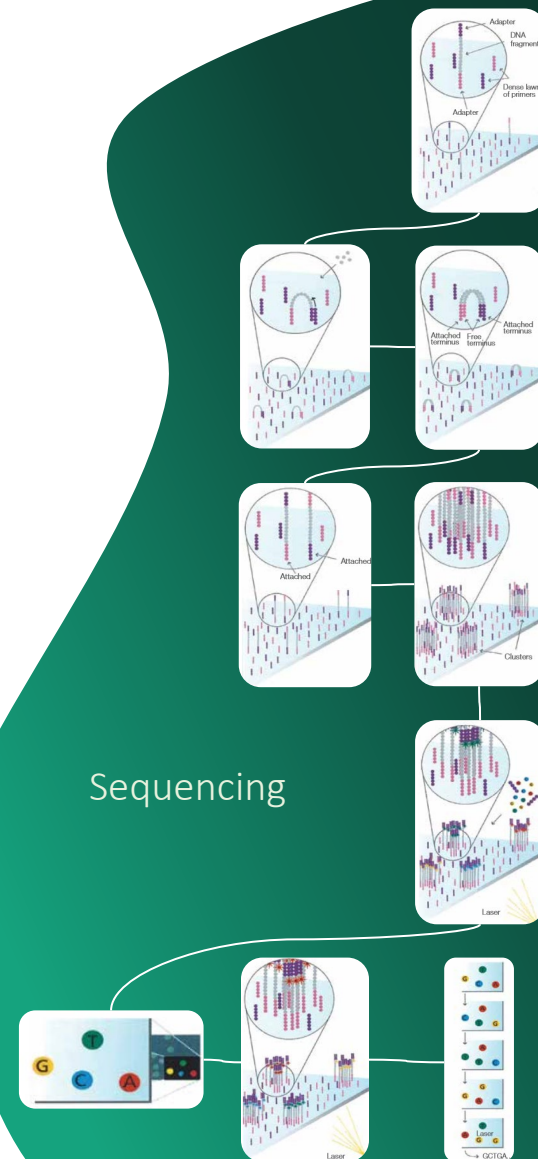


<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

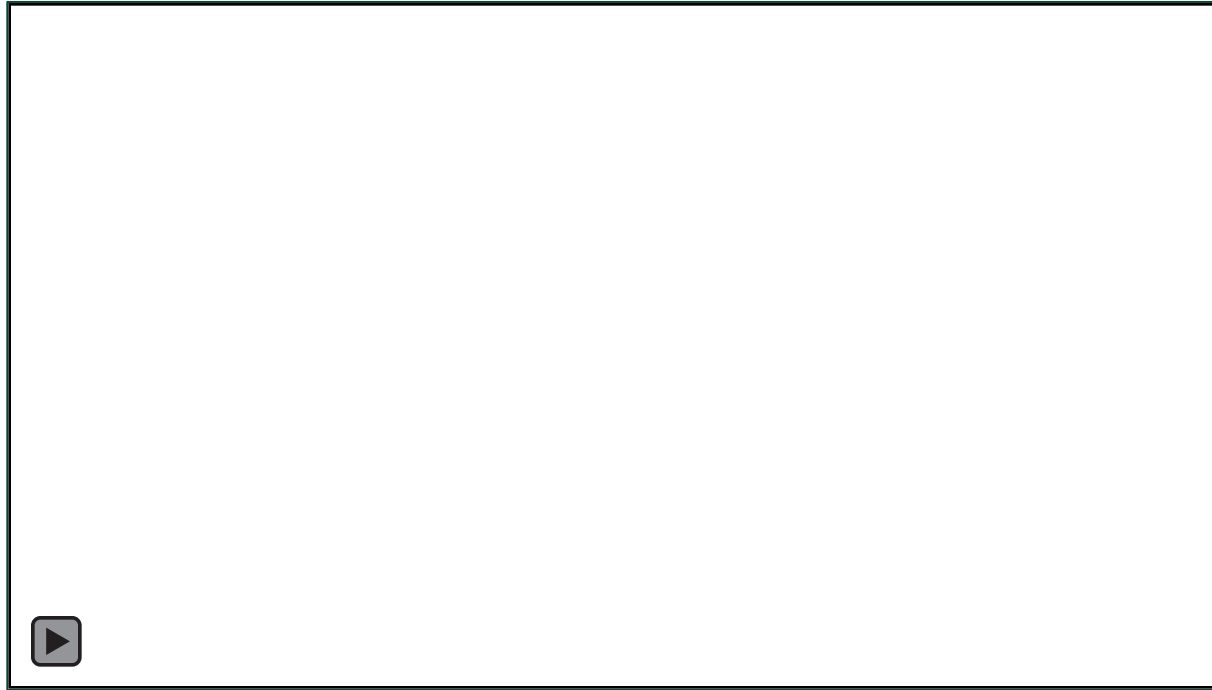
Steps for Illumina sequencing



Sequencing



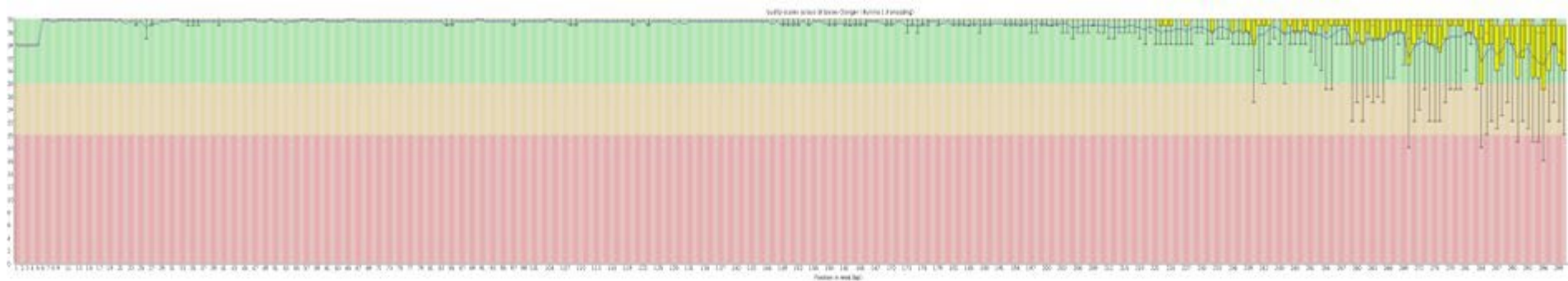
Illumina sequencing



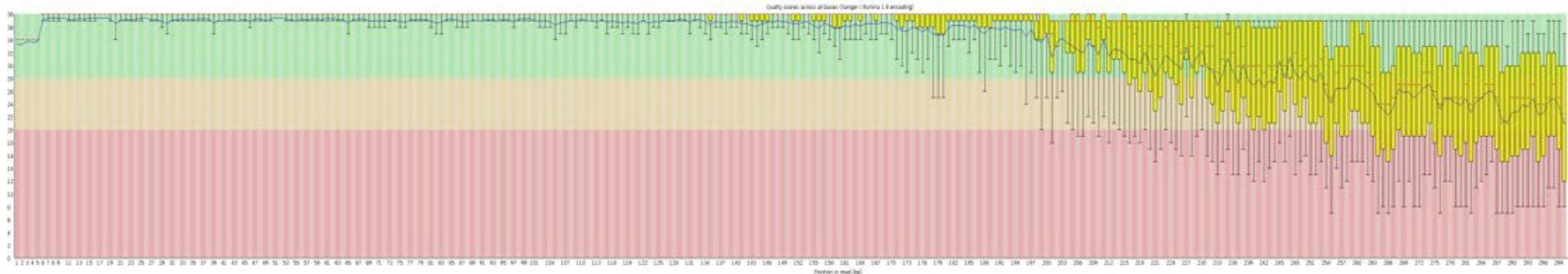
Illumina sequencing



From 2x250 to 2x300 bp
From 1 to 25 million read pairs per Flowcell
1 line per Flowcell
56-hour run



Reads 1



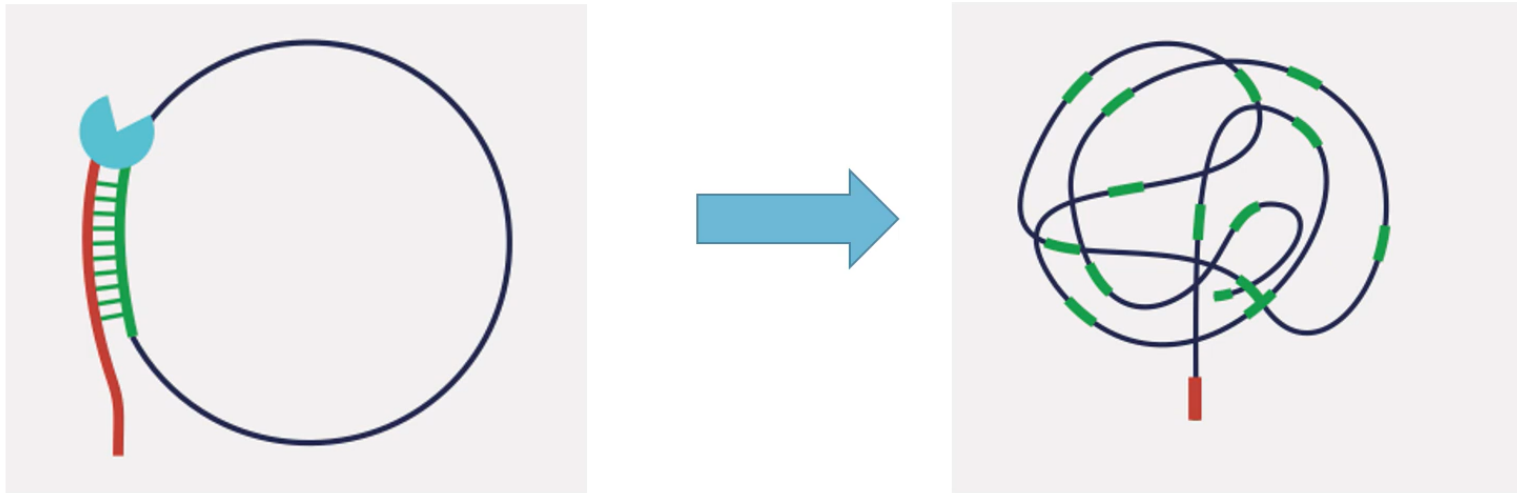
Reads 2

- + Proven technologies that are easy to implement
- Data quality tends to decline at the end of the read
- Requires diversity input *via* phix → data loss

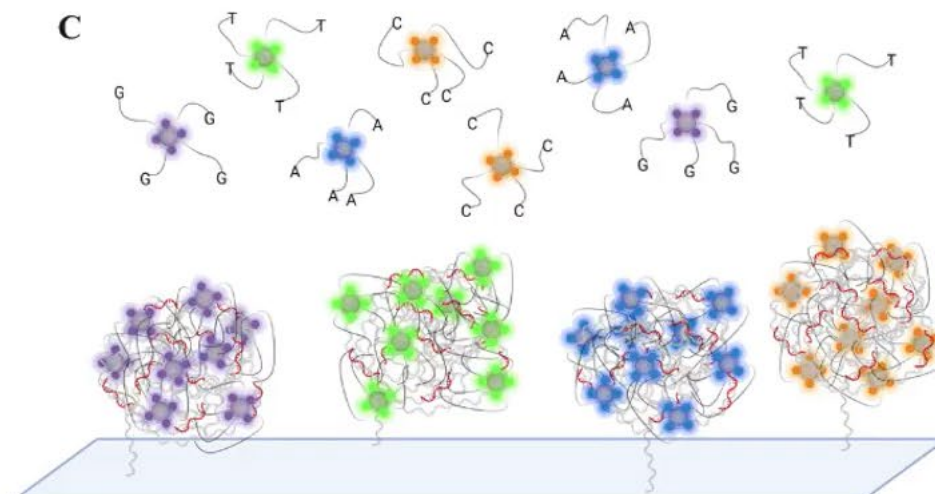
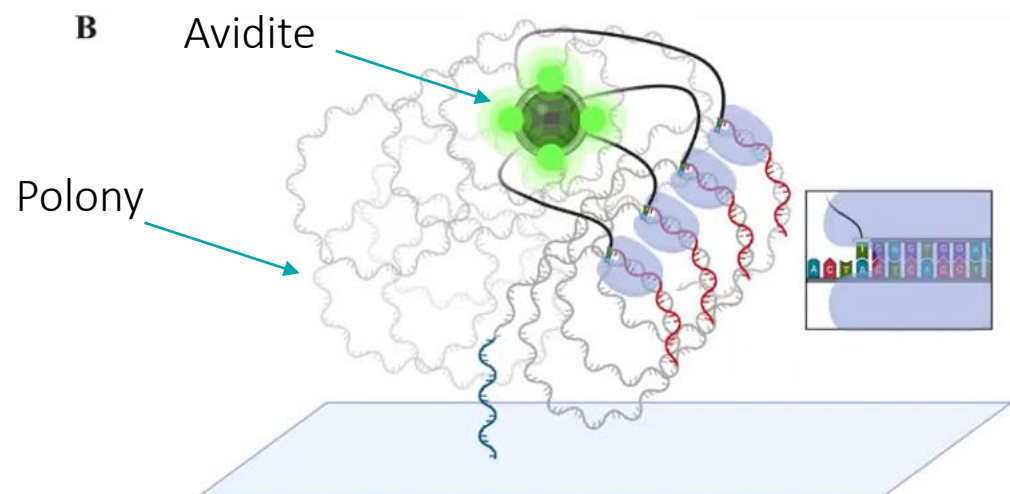
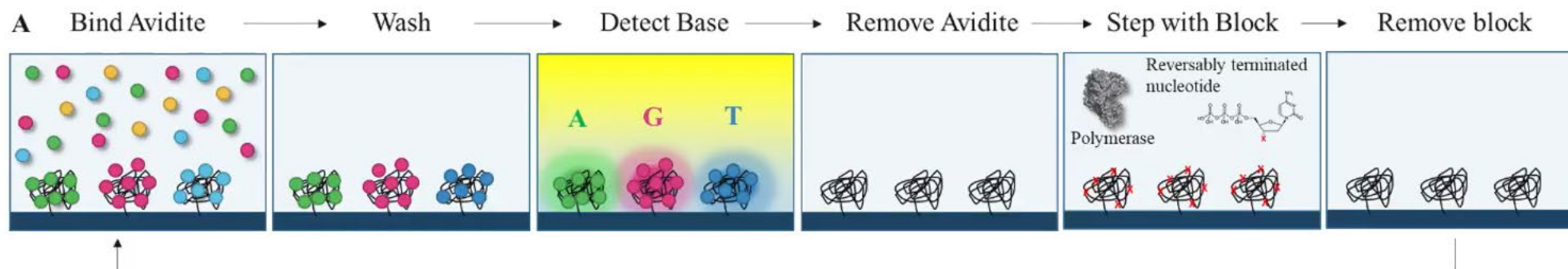
AVITI : colonies polymerization

- The AVITI generates **colonies** with rolling circle amplification (RCA) from circular templates
- Each **colony** is a continuous DNA strand with many sequencing start sites
- No on-instrument PCR reduces errors by only copying from the original

Surface
primer
Sequencing
adapter



AVITI : sequencing



Instrument Workflow:





<https://bit.ly/3vAh6gE>

AVITI sequencing



AVITI sequencing (Element Biosciences)

2x300 bp

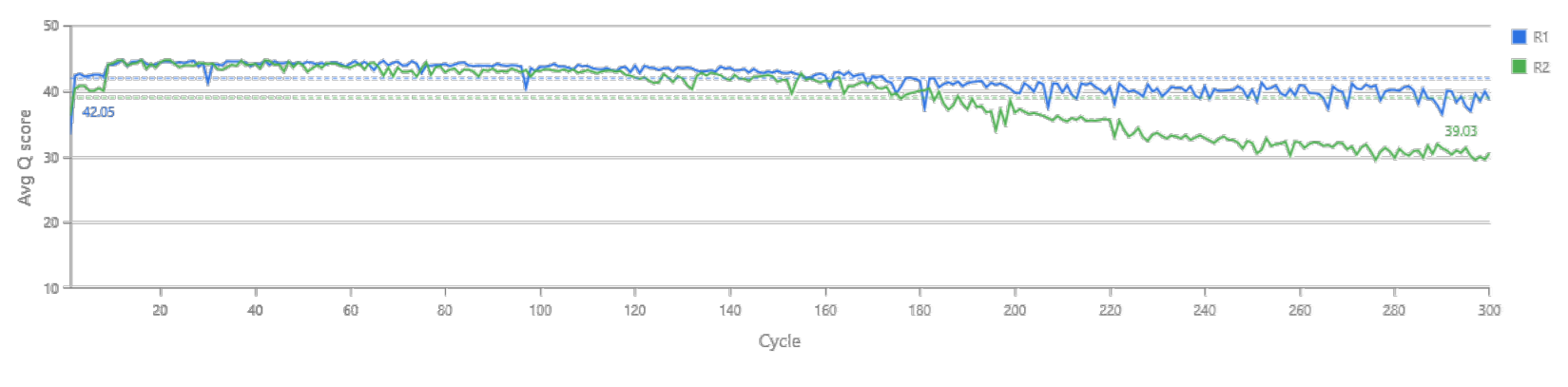
7.5 to 150 Gb per lane

2 lanes per Flowcell

38-hour run

Compatible with Illumina libraries

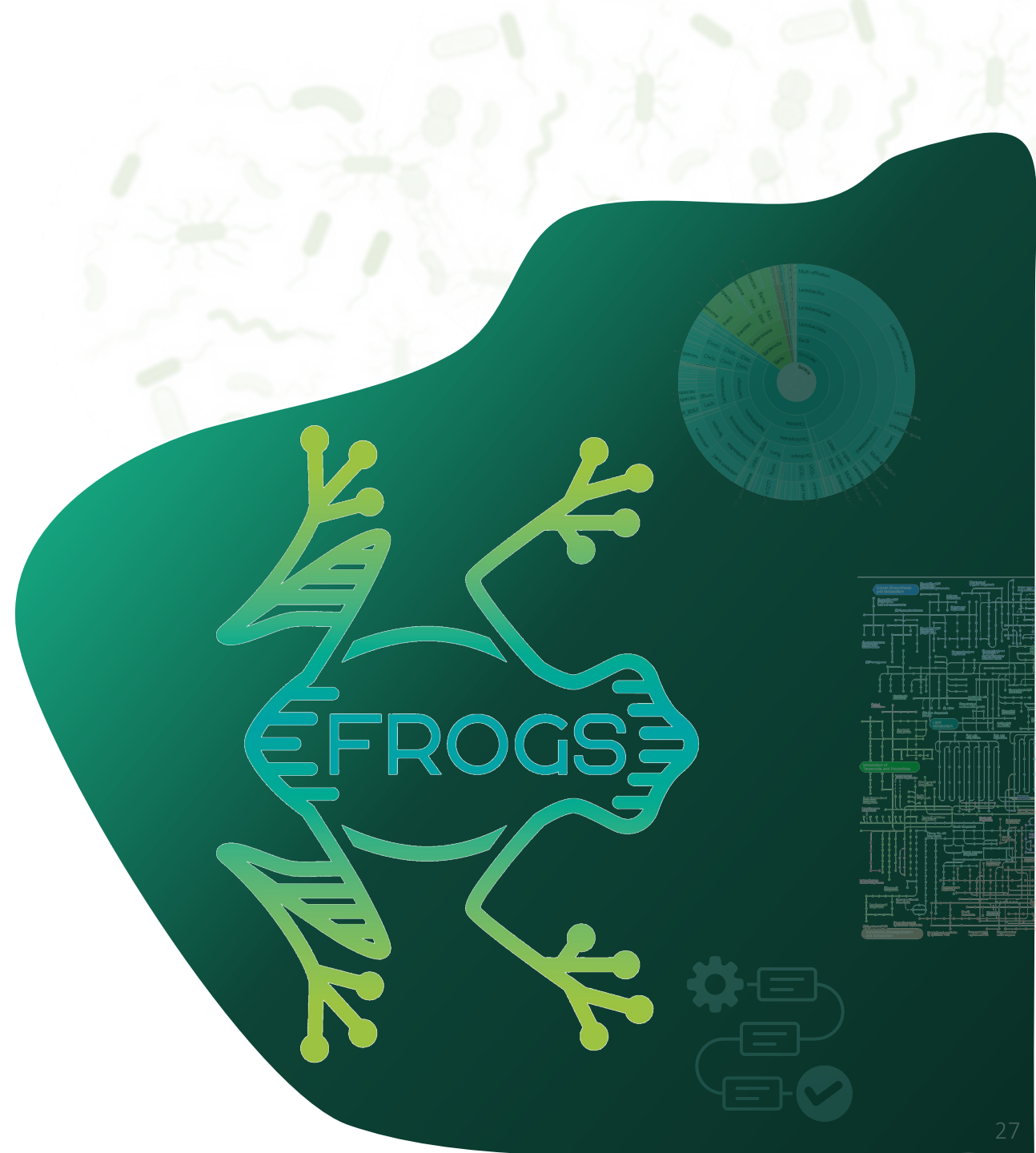
90% bases > Q30



- + Excellent data quality throughout the reads.
- Very large volume of data.
- Requires diversity input *via* phix → data loss
- ! Different technology → different analysis results.

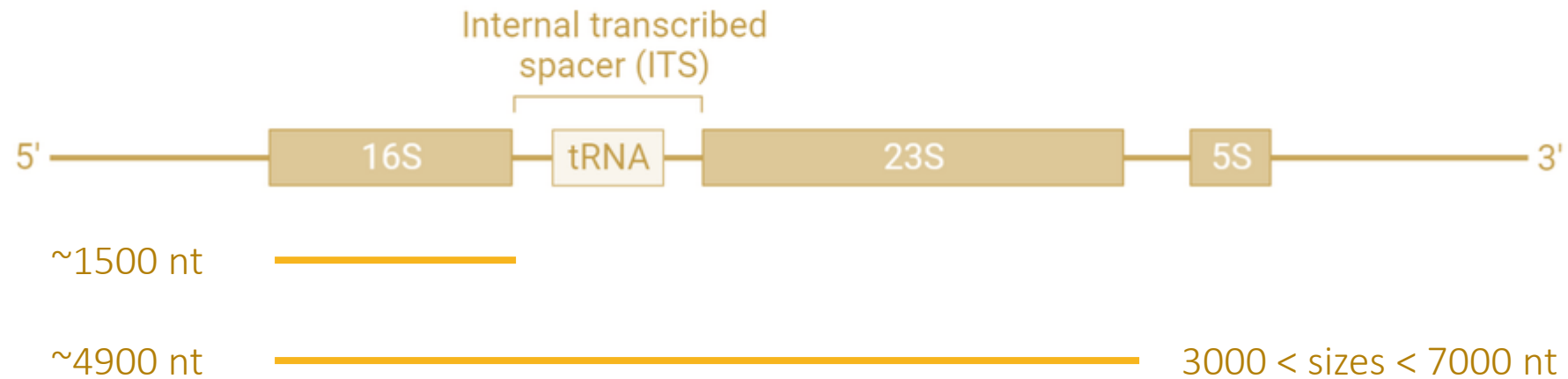
Metabarcoding overview

Long reads



Long reads

Bacteria rRNA gene organization

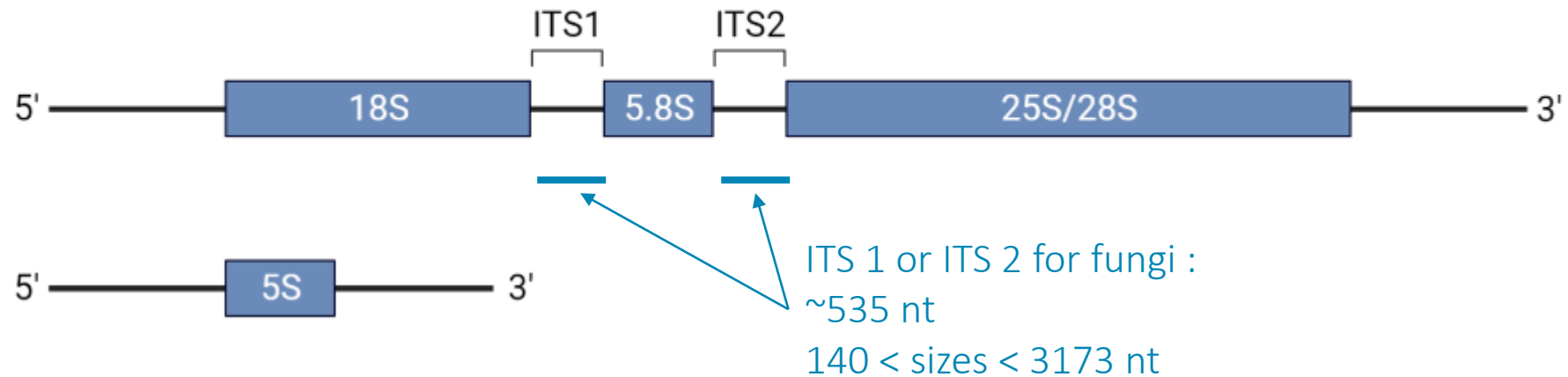


Long reads

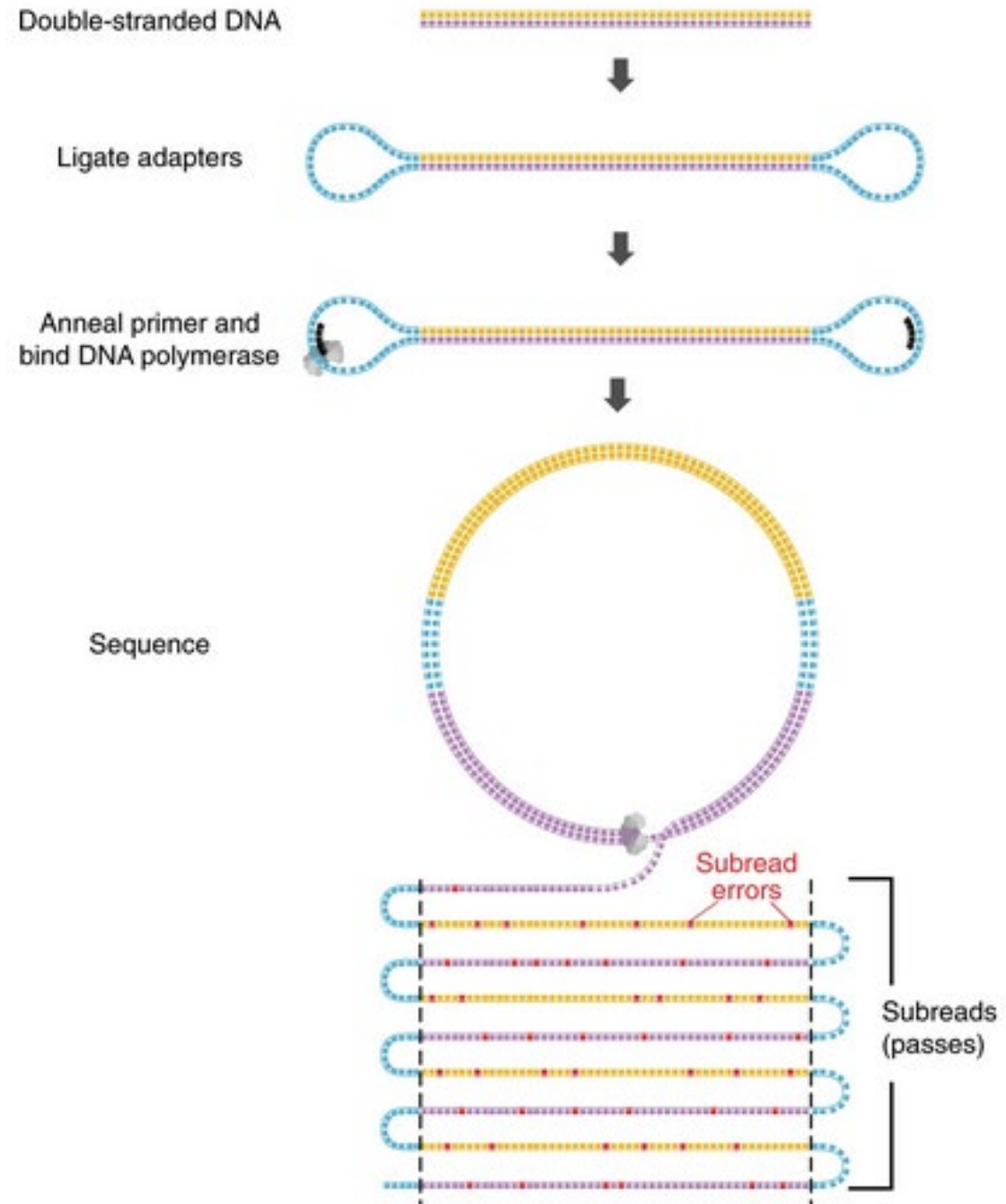
Bacteria rRNA gene organization



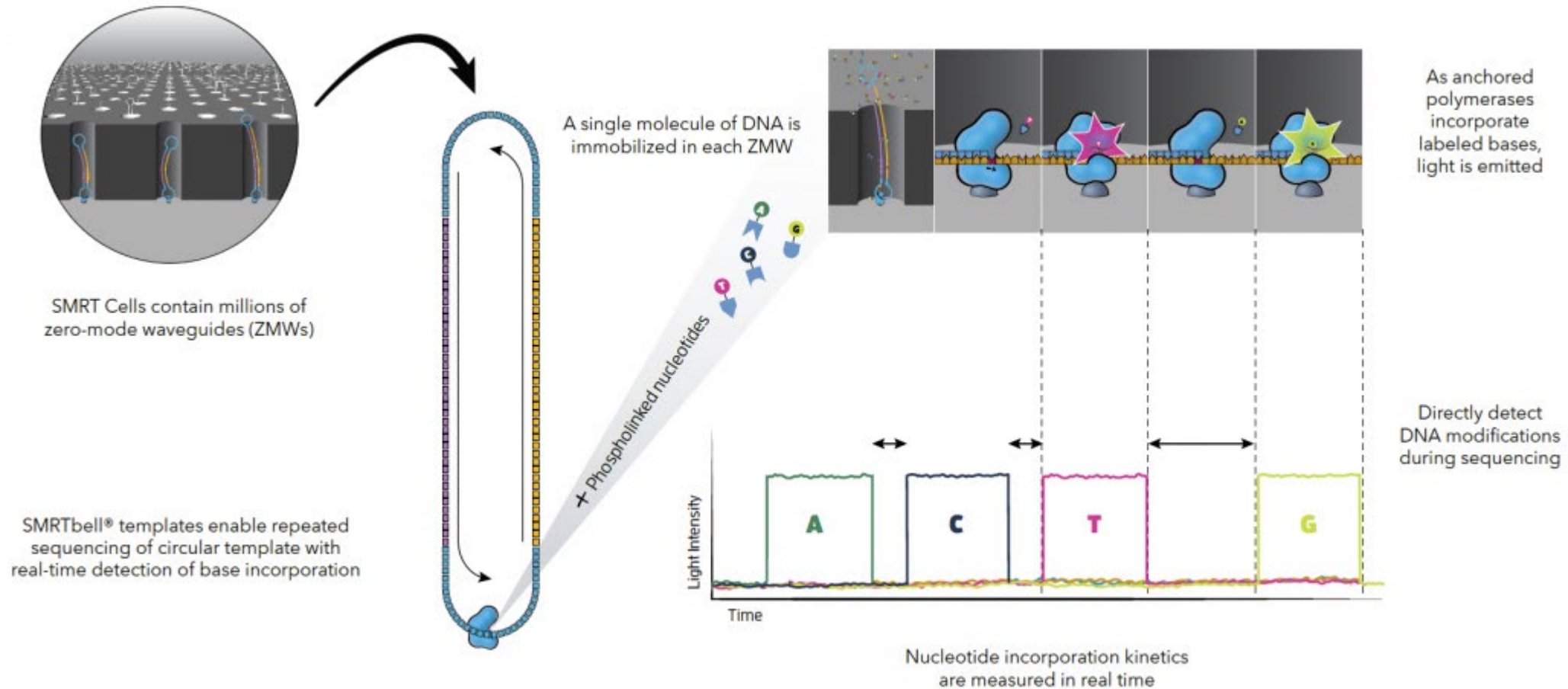
Eukaryotes rRNA gene organization



PacBio Sequencing



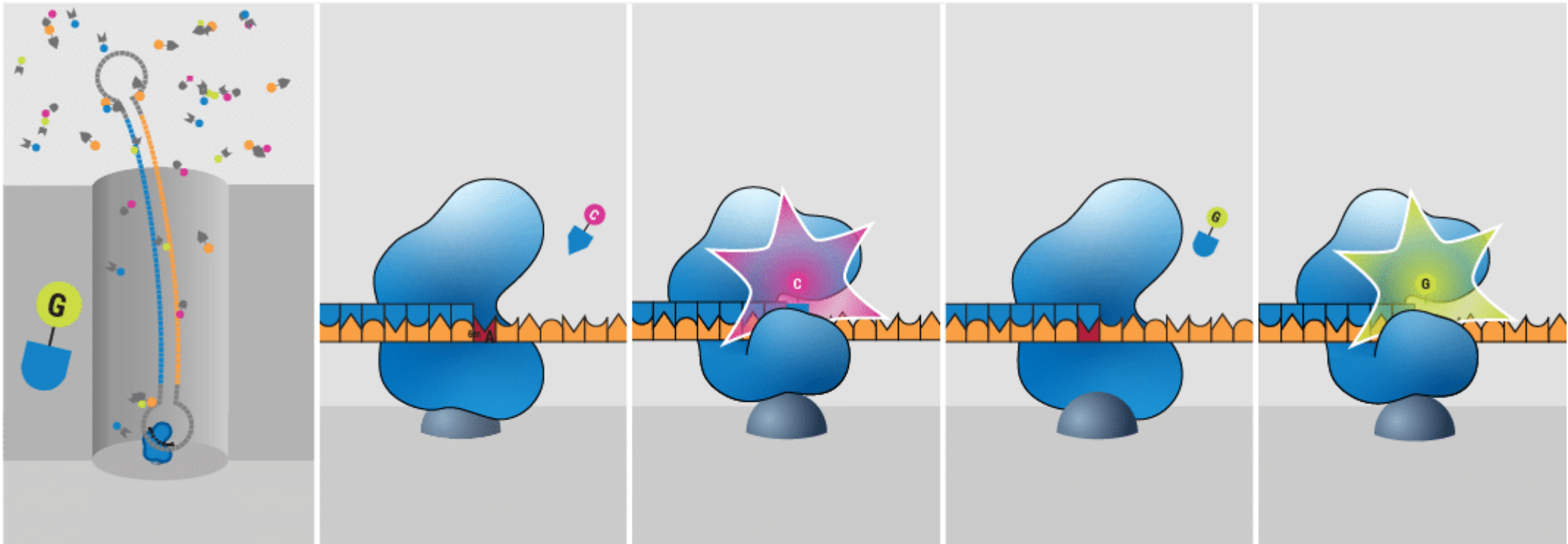
PacBio Sequencing



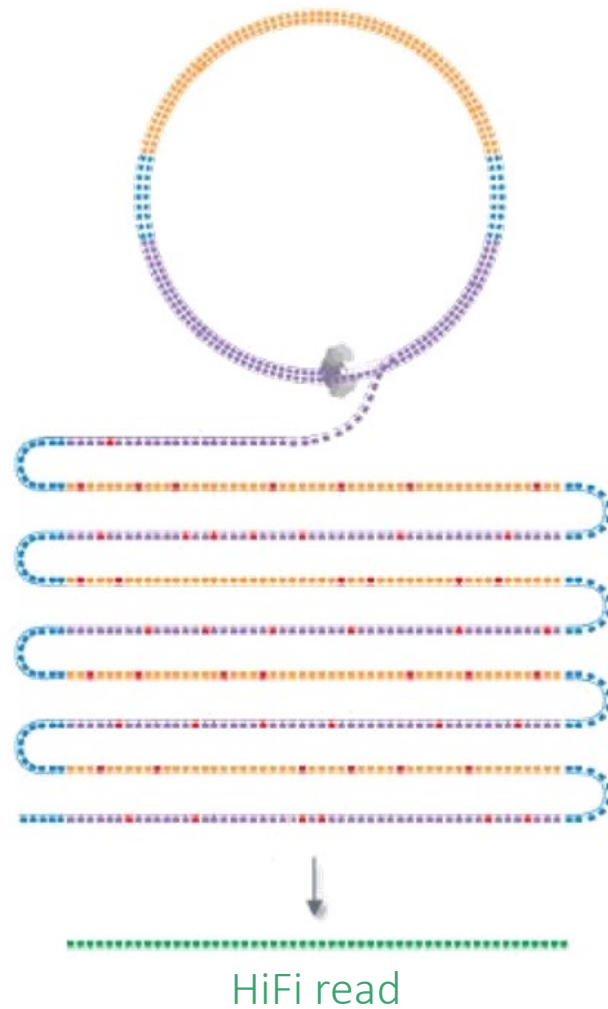
Here, the internal control is not phiX but an internal sequencing control consisting of a 2 kb DNA insert (unknown sequence).

Single-Molecule Real-Time sequencing technology (SMRT)

PacBio Sequencing



PacBio Sequencing



Single-pass sequencing read errors are rapidly washed out with increasing number of passes (intra-molecular coverage)

1 pass → 90 % accuracy

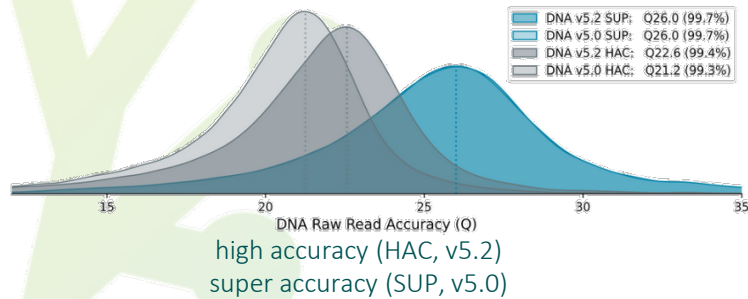
2 passes → 96 % accuracy

3 passes → 99 % accuracy



ONT sequencing

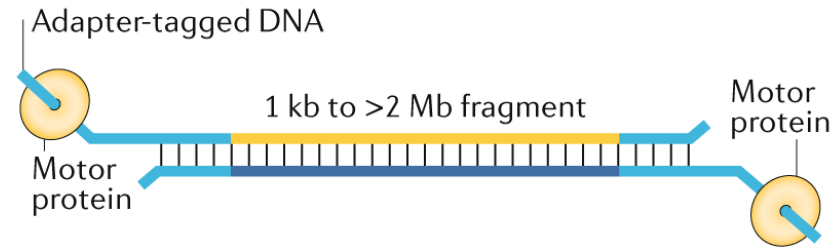
From 0.5 to 25Kb
From 25 to 90 Gb per Flowcell
72 hours of runs



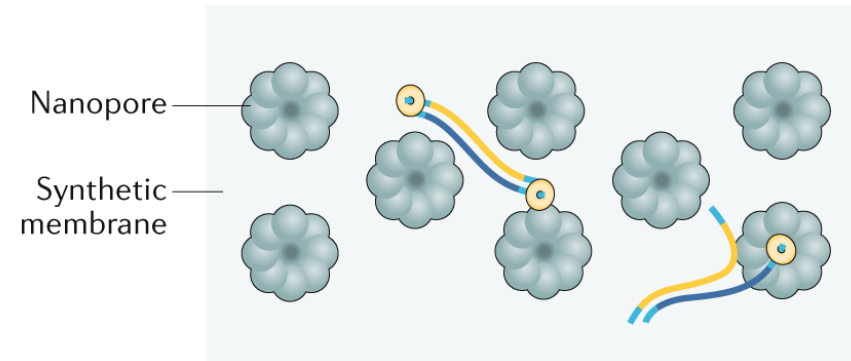
Here, the internal control is not phiX but the YHR174W gene.

<https://doi.org/10.1038/s41576-020-0236-x>

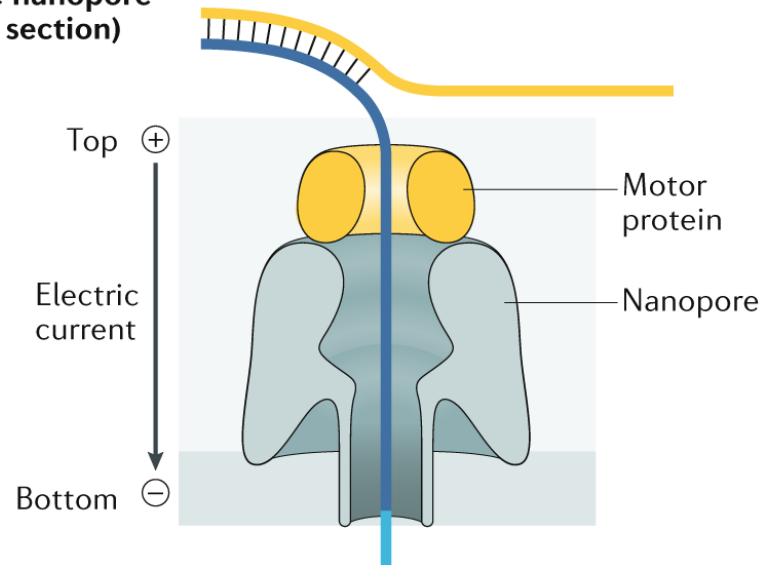
Template topology



Flow cell (top view)

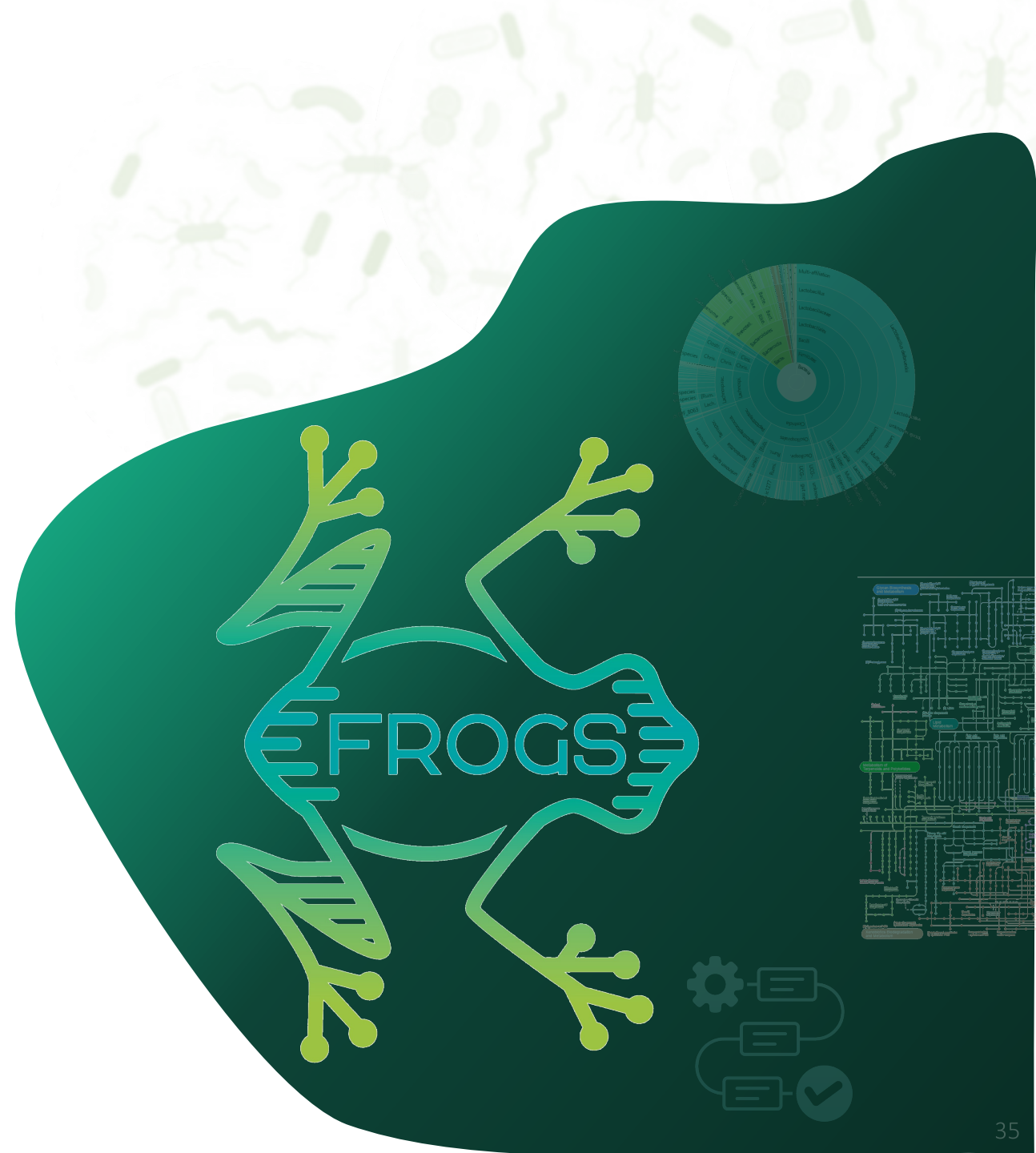


Single nanopore (cross section)

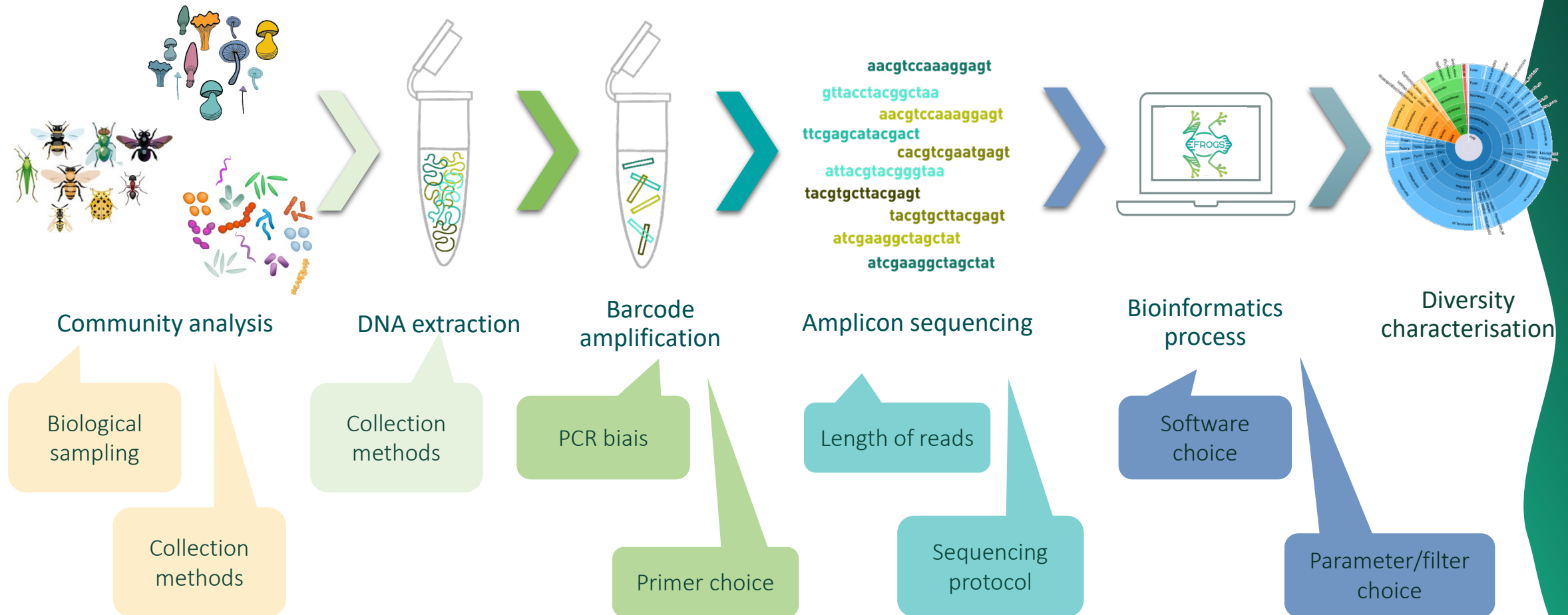


Metabarcoding overview

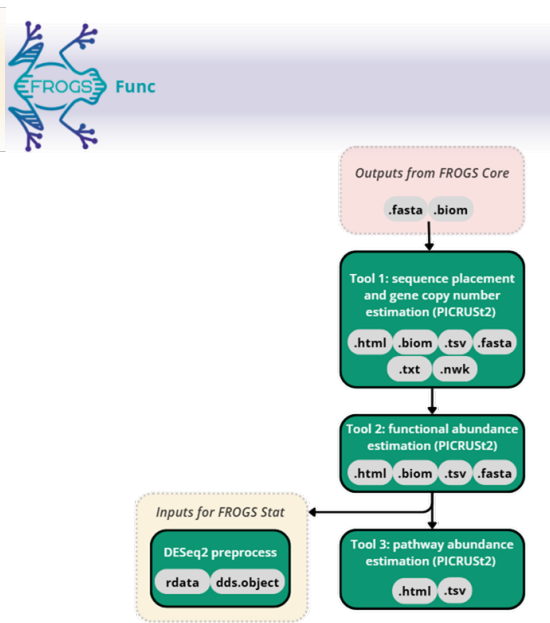
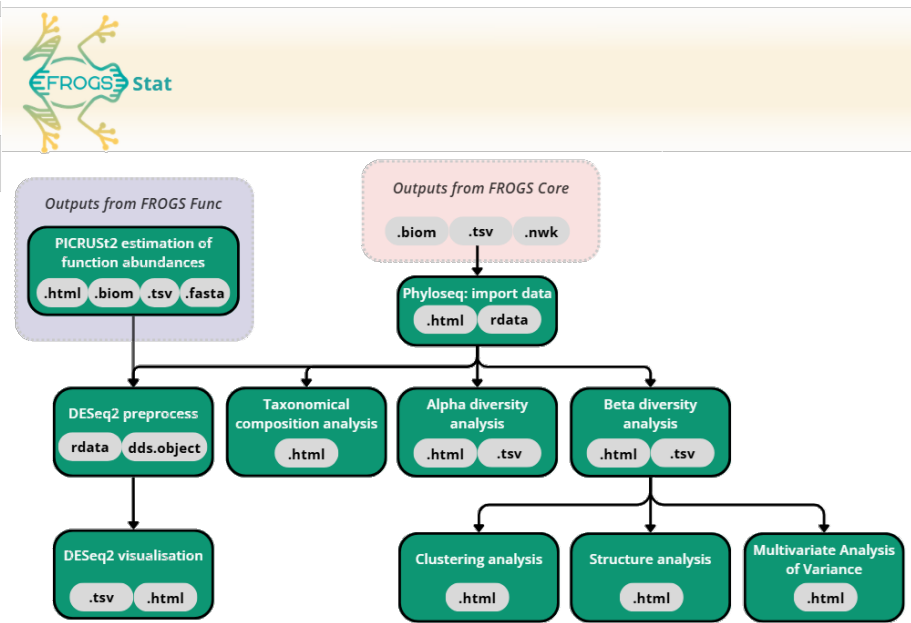
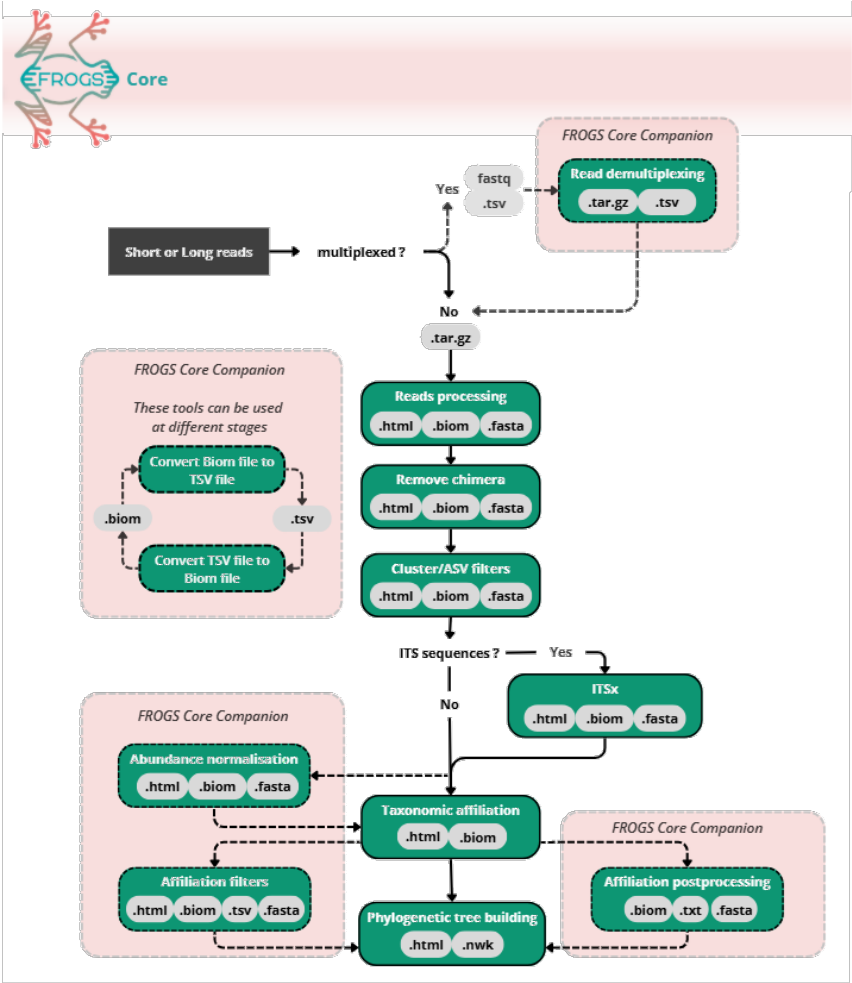
Bioinformatics analysis



Methods and parameter of tools impact final diversity view

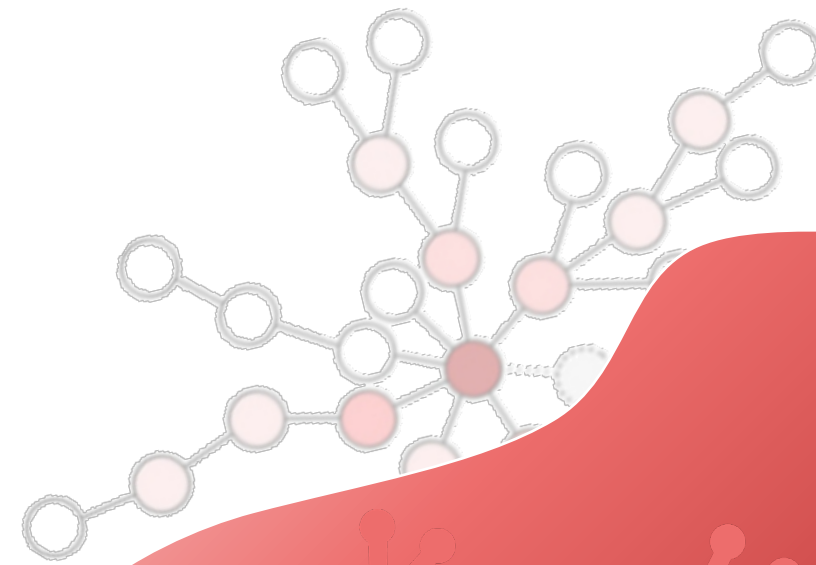
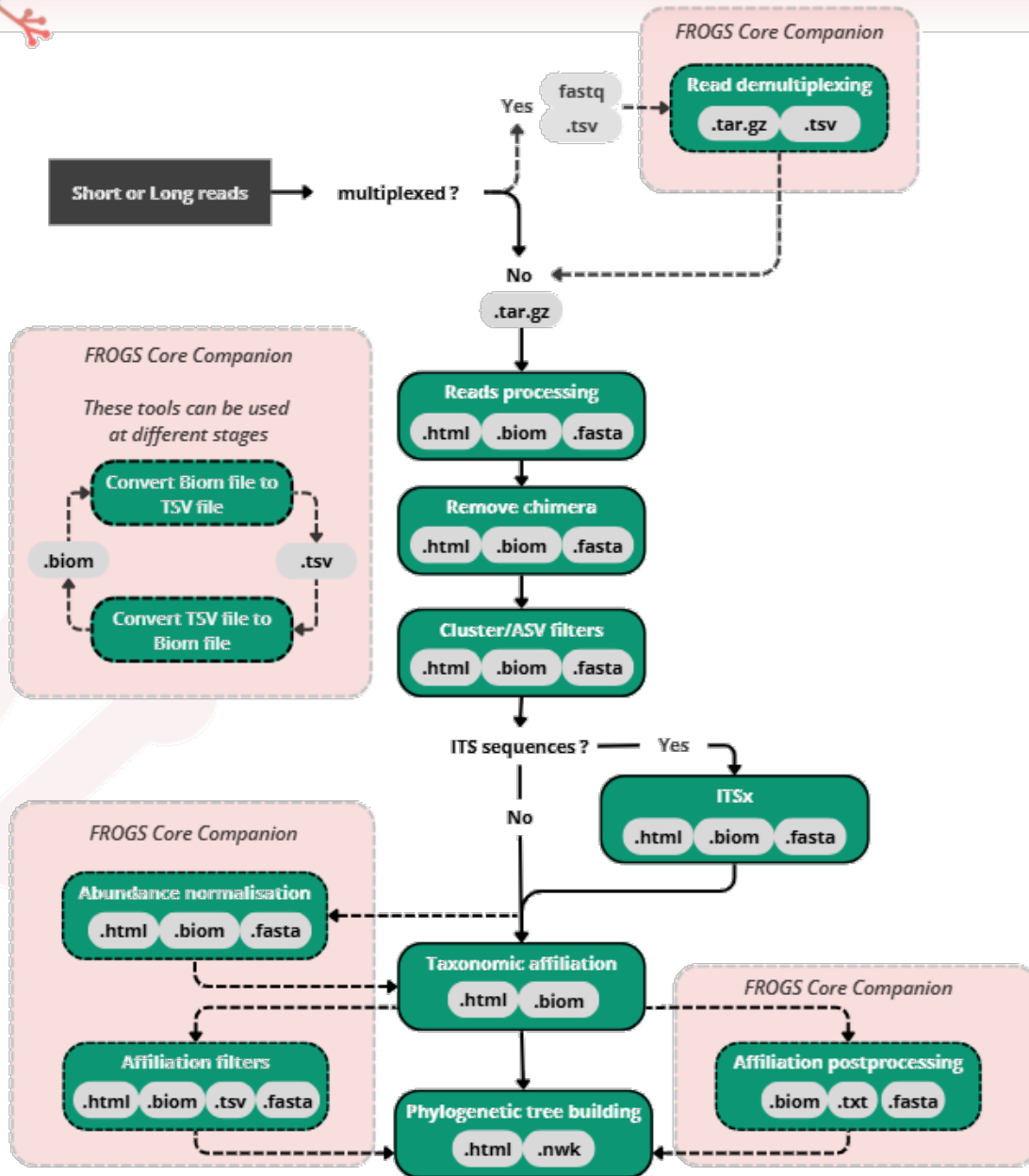
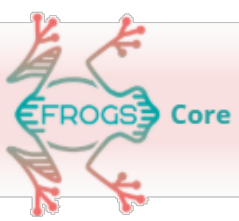


FROGS = 3 tool groups



For 454 data, Illumina, AVITI, PACBIO, ONT – 17 markers – 135 databases.

- 12S
- 16S
- 16S-ITS-23S
- 18S
- 23S
- 28S
- COI
- EF1,18S
- ITS
- ITS2
- SSU-ITS-LSU
- gyrb
- matK
- rbcl
- rpoB
- trnH
- trnL



aacgtccaaaggagt
gttacctacggctaa
aacgtccaaaggagt
ttcgagcatagact
cacgtcgaatgagt
attacgtacgggtaa
tacgtgcttacgagt
tacgtgcttacgagt
atcgaaggctagctat
atcgaaggctagctat



