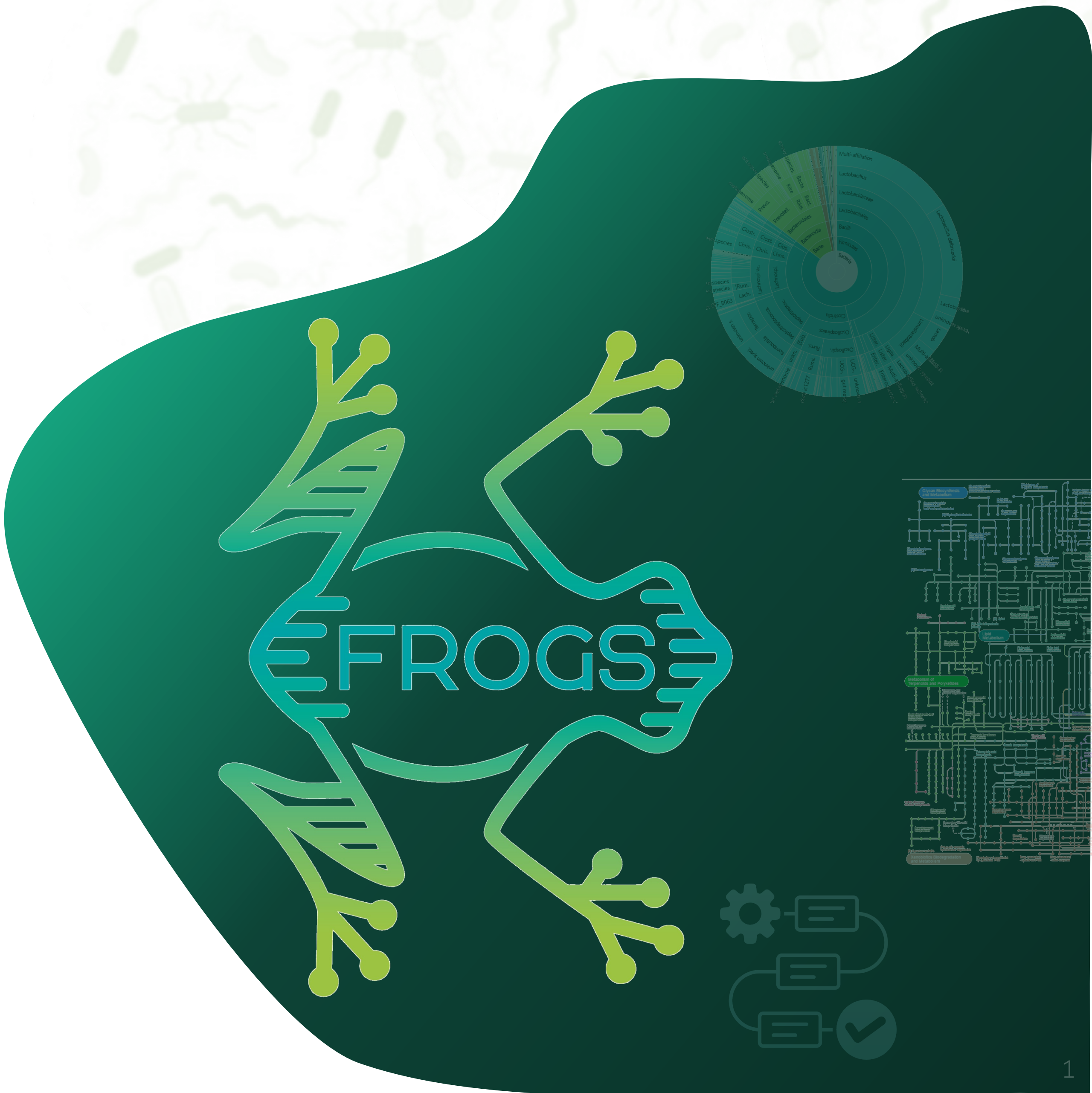
Metabarcoding overview

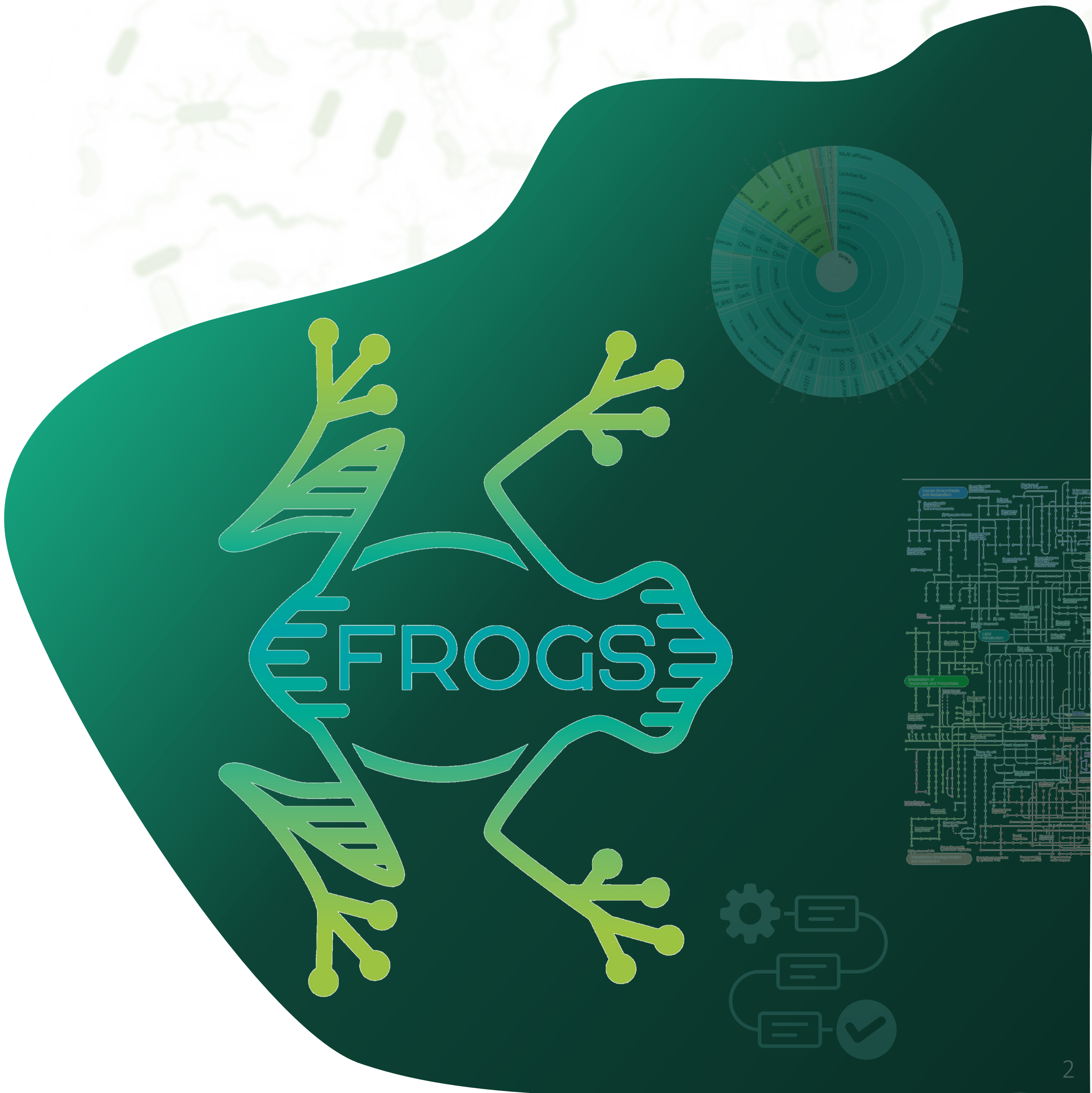
Key concepts

Lucas Auer, Gabryelle Agoutin, Maria Bernard, Géraldine Pascal, Maëlle Pomiès & Olivier Rué

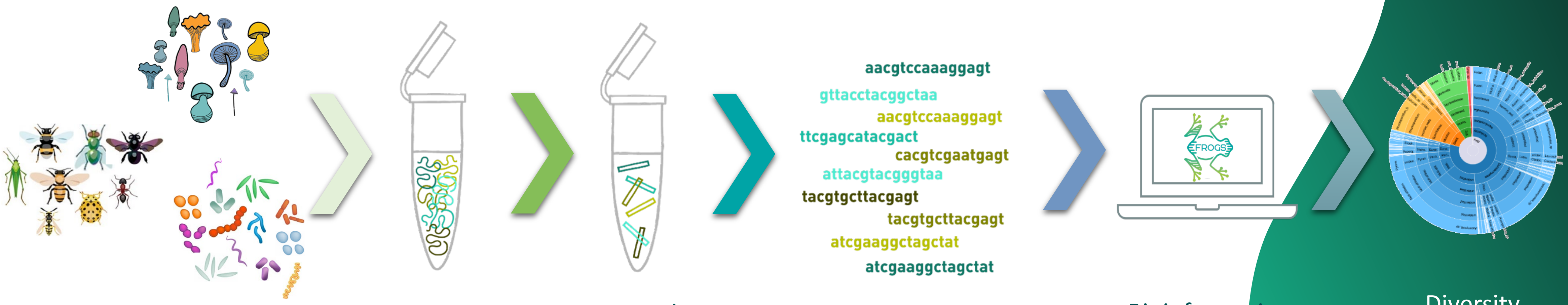


Metabarcoding overview

Barcode



Objectives



aacgtccaaaggagt
gttacctacggctaa
aacgtccaaaggagt
ttcgagcatacgact
cacgtcgaatgagt
attacgtacgggtaa
tacgtgcttacgagt
tacgtgcttacgagt
atcgaaggctagctat
atcgaaggctagctat

Community analysis

DNA extraction

Barcode amplification

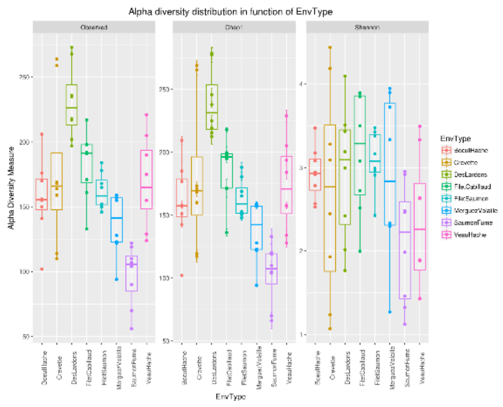
Amplicon sequencing

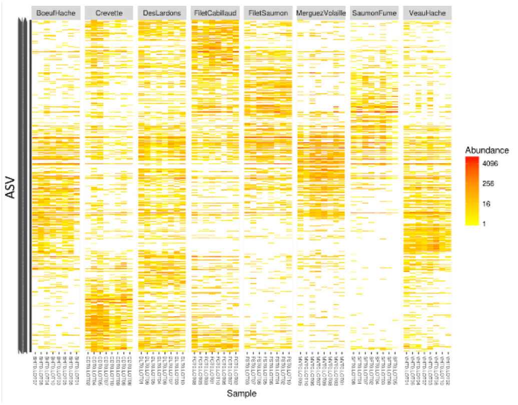
Bioinformatics process

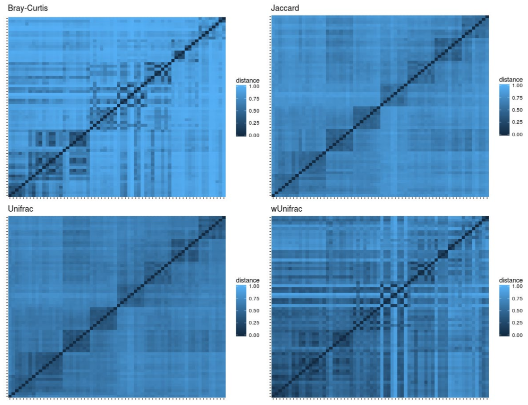
Diversity characterisation

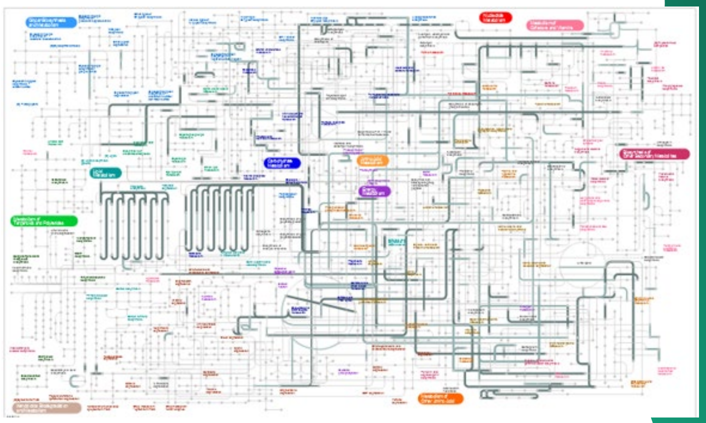
Objectives: a count table for statistics analysis

	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

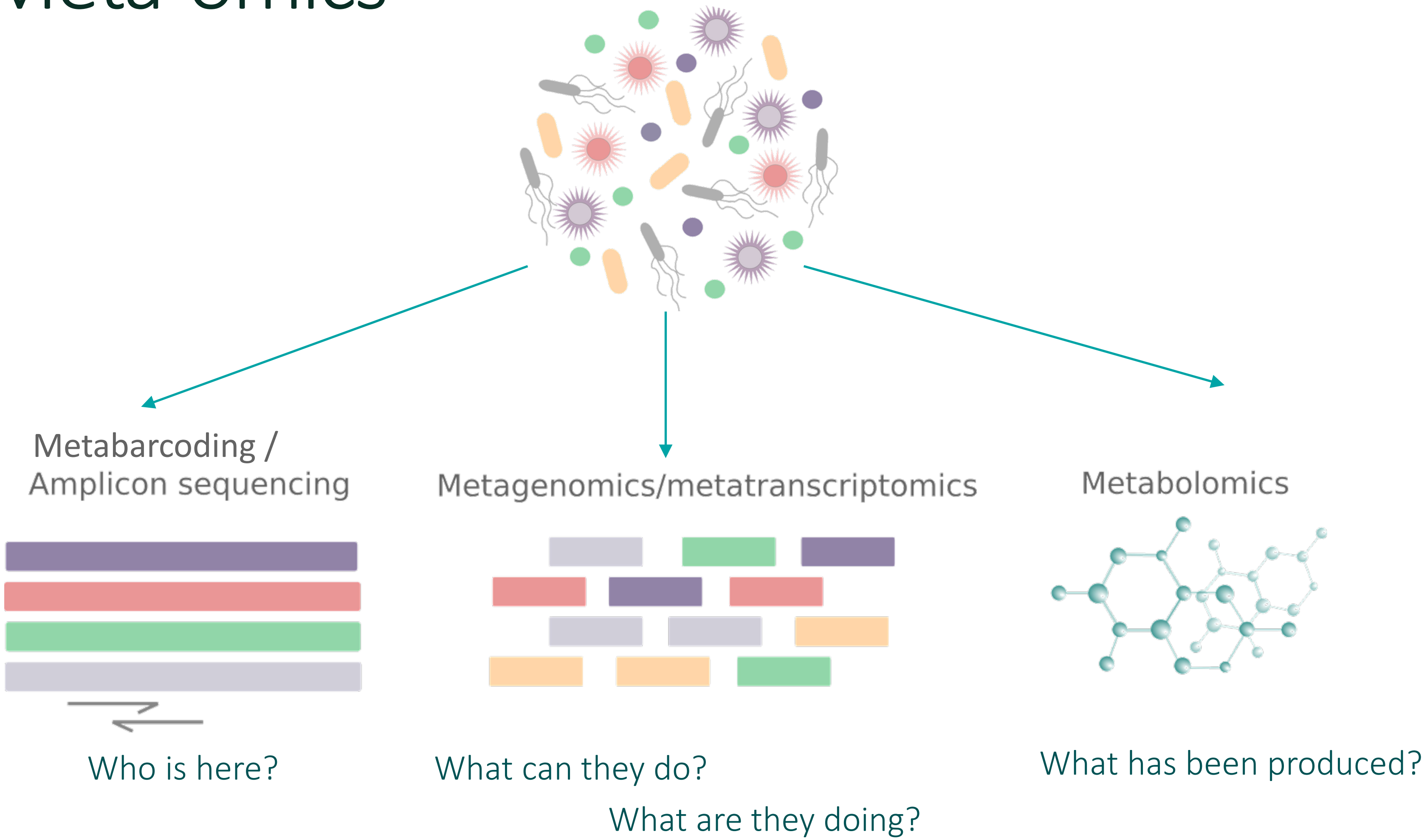








Meta-omics



Metabarcoding / Amplicon sequencing

Who is here?

Metagenomics/metatranscriptomics

What can they do?

What are they doing?

Metabolomics

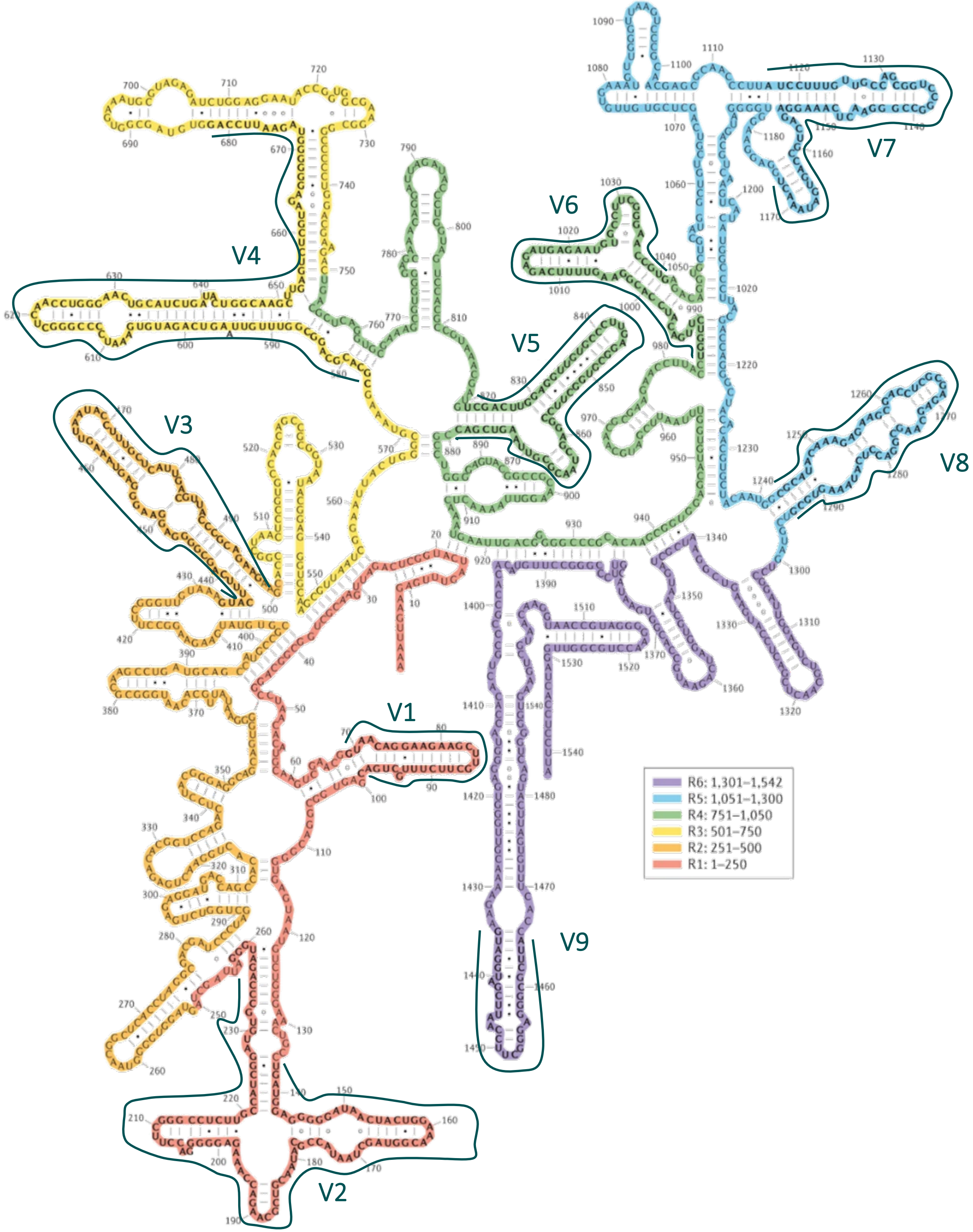
What has been produced?

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

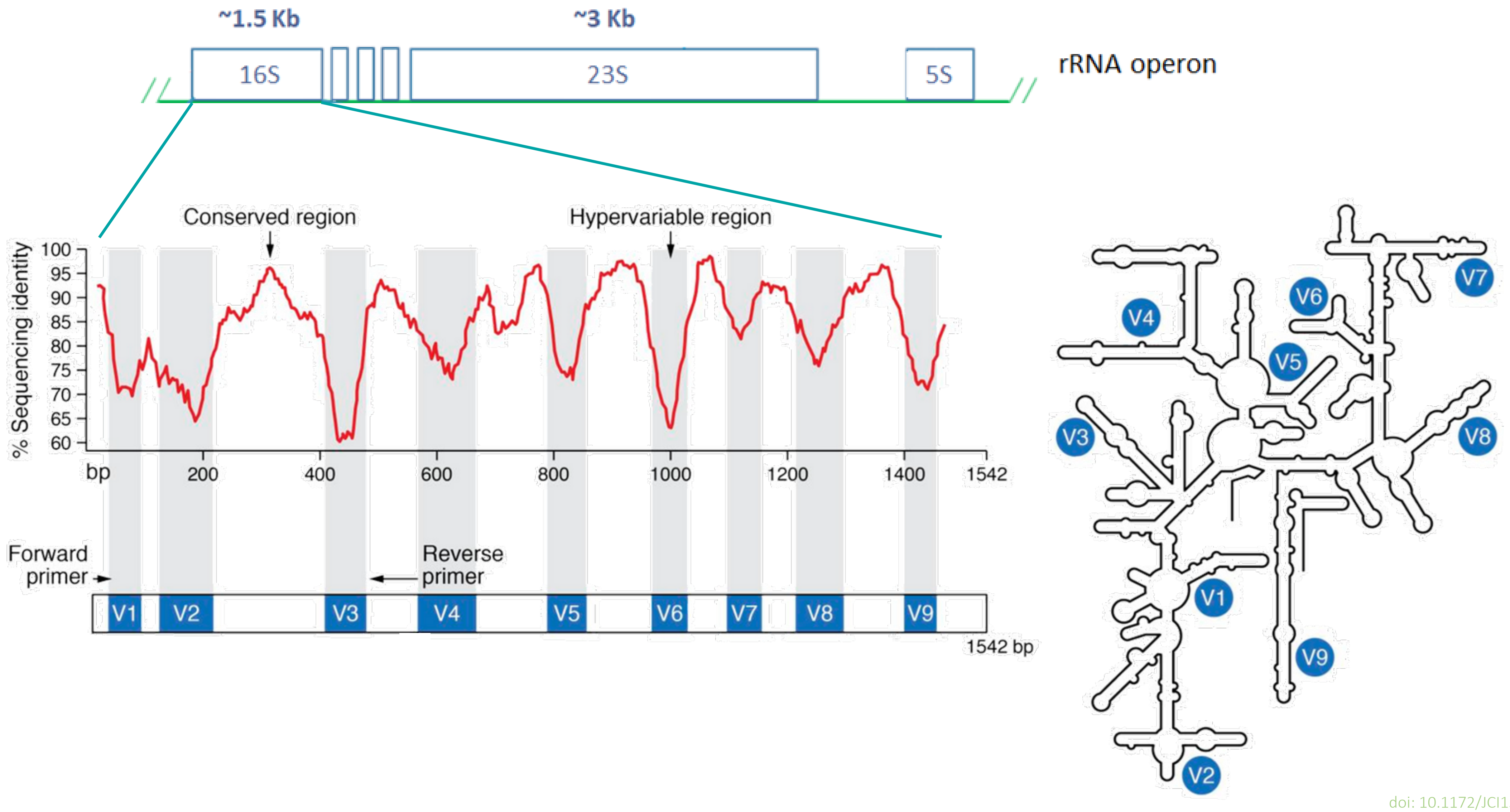


doi:10.1038/nrmicro3330

16S

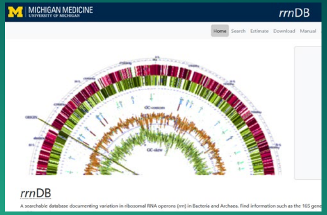
- Ubiquist gene
- No lateral gene transfer
- Molecular phylogenetic marker
- Availability of databases
  GTDB_220 (2024) 863832
  SILVA_138.2 (2024) 451555

16S rRNA structure



doi: 10.1172/JCI154944

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

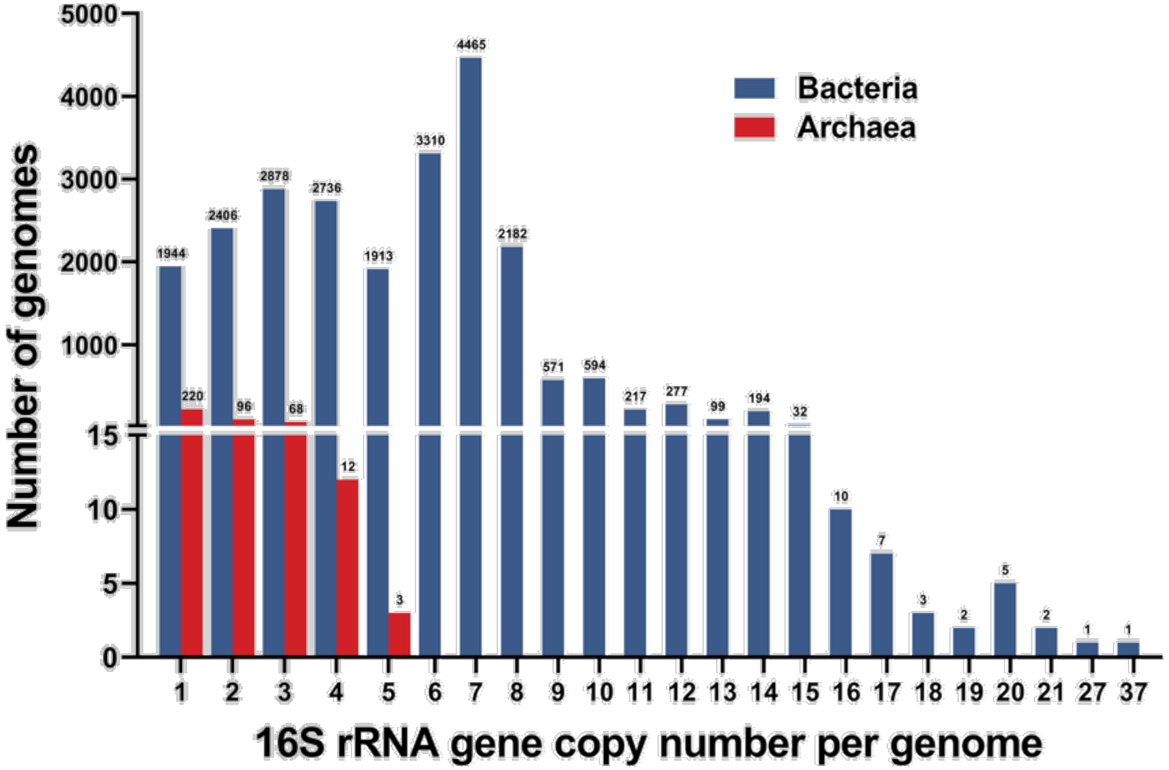


*rrn*DB

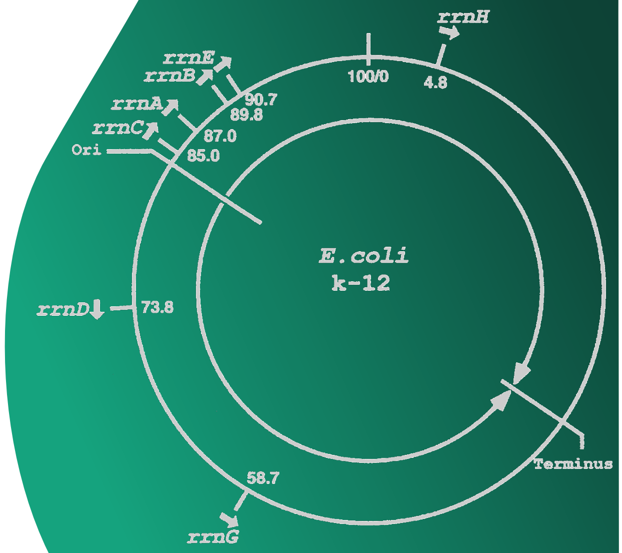
A searchable database documenting variation in ribosomal RNA operons (*rrn*) in Bacteria and Archaea. Find information such as the 16S gene

Variations of 16S rRNA gene copy number per genome

Several 16S rRNA genes in genomes



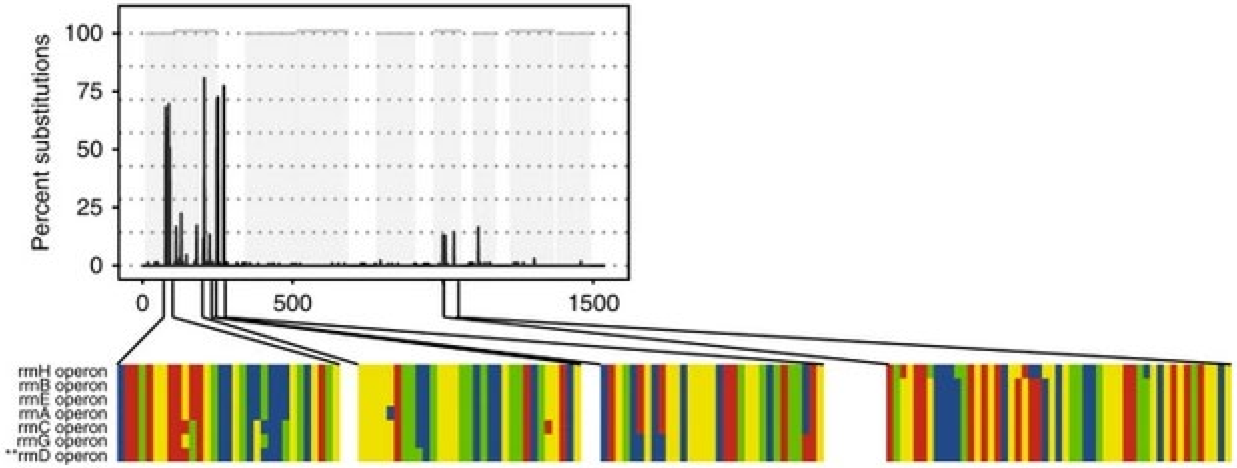
2023 - doi: 10.1128/aem.02108-22



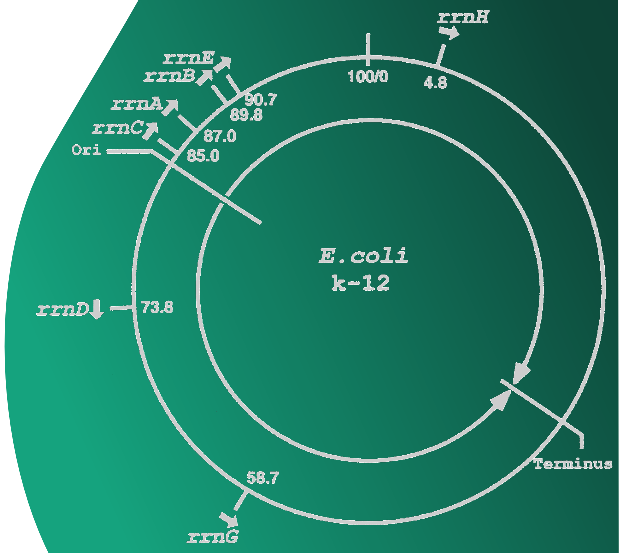
doi/10.1073/pnas.96.5.1820

Intragenomic variations of 16S rRNA genes

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



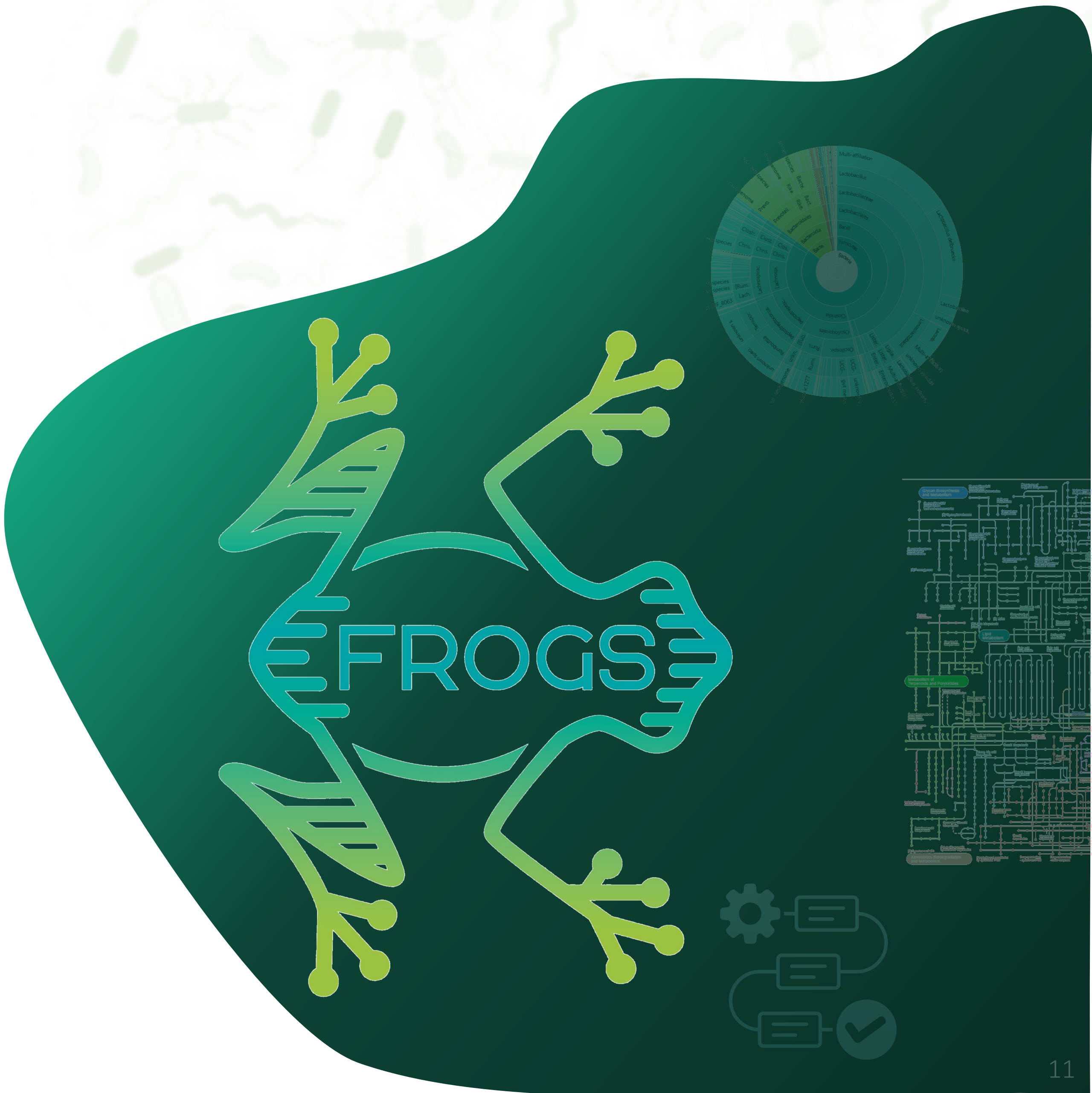
https://doi.org/10.1038/s41467-019-13036-1



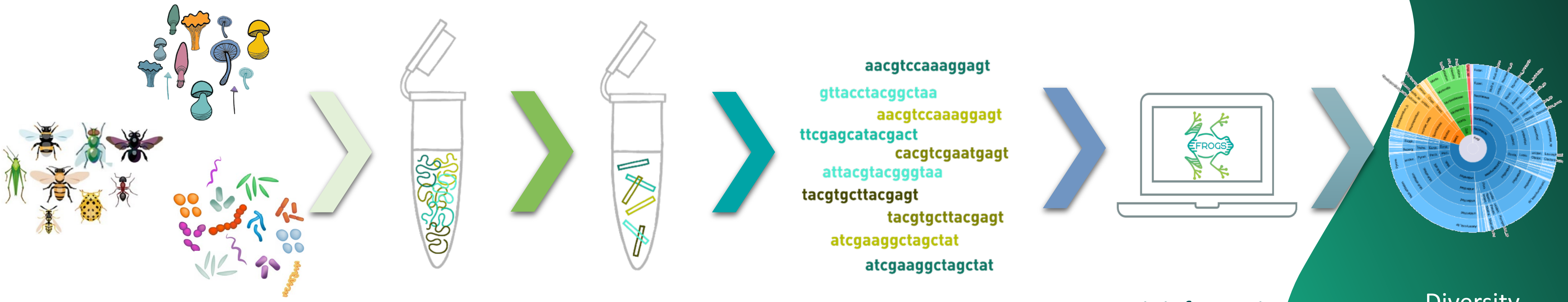
doi/10.1073/pnas.96.5.1820

Metabarcoding overview

Technical sources of noise and errors



Metabarcoding analysis steps



Community analysis

DNA extraction

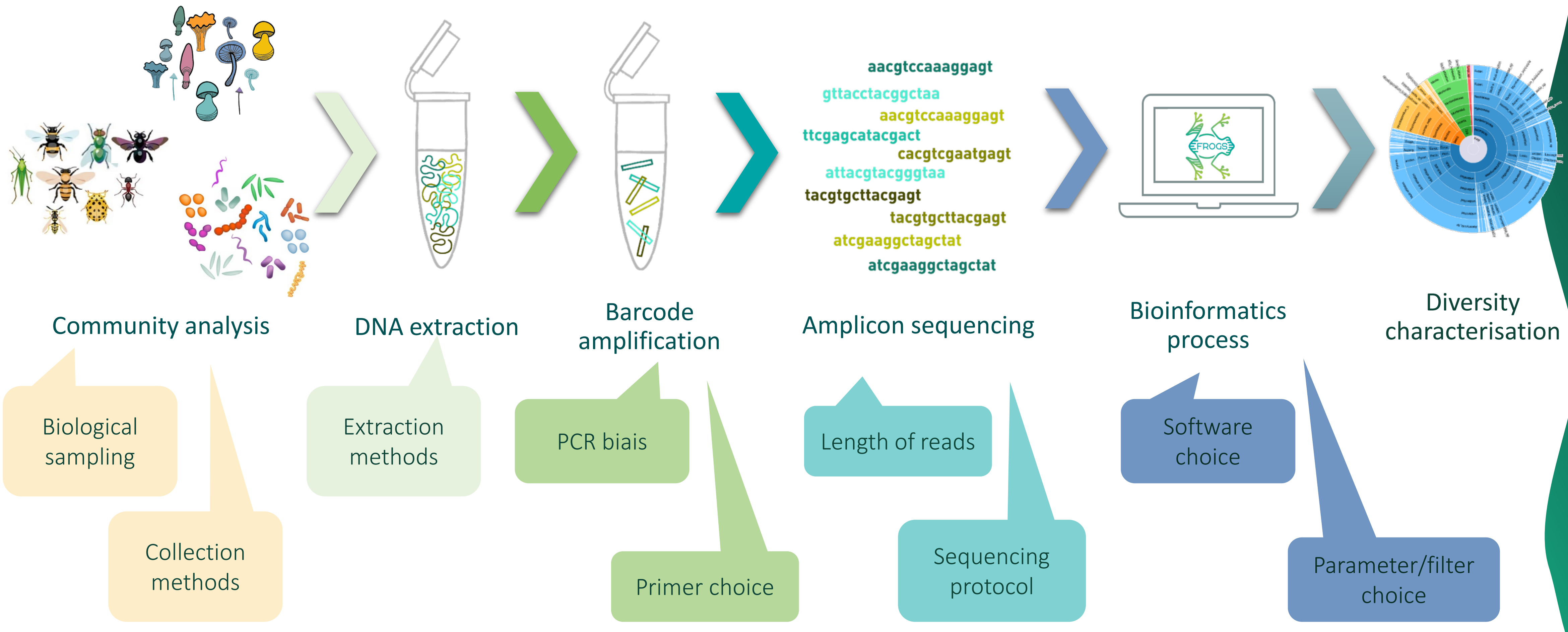
Barcode amplification

Amplicon sequencing

Bioinformatics process

Diversity characterisation

Noise/errors are inevitable at every stage.



Community analysis

- Biological sampling
- Collection methods

DNA extraction

- Extraction methods

Barcode amplification

- PCR biais
- Primer choice

Amplicon sequencing

- Length of reads
- Sequencing protocol

Bioinformatics process

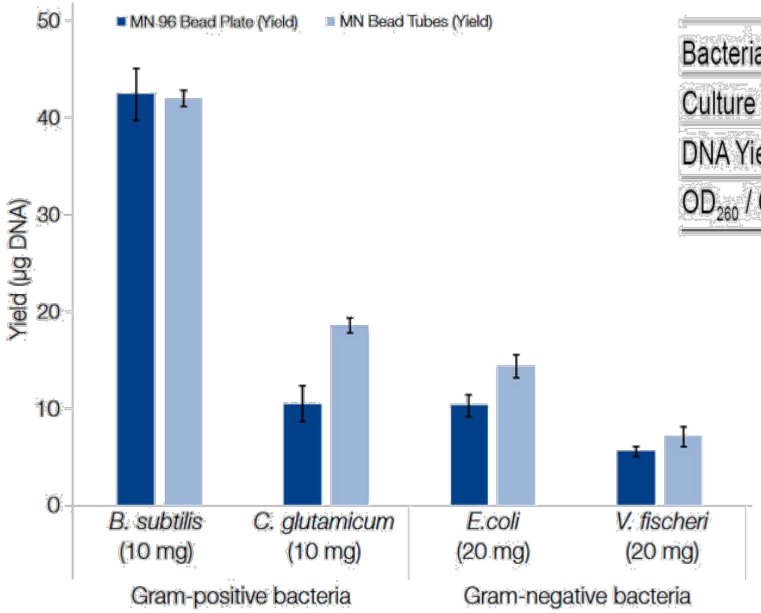
- Software choice
- Parameter/filter choice

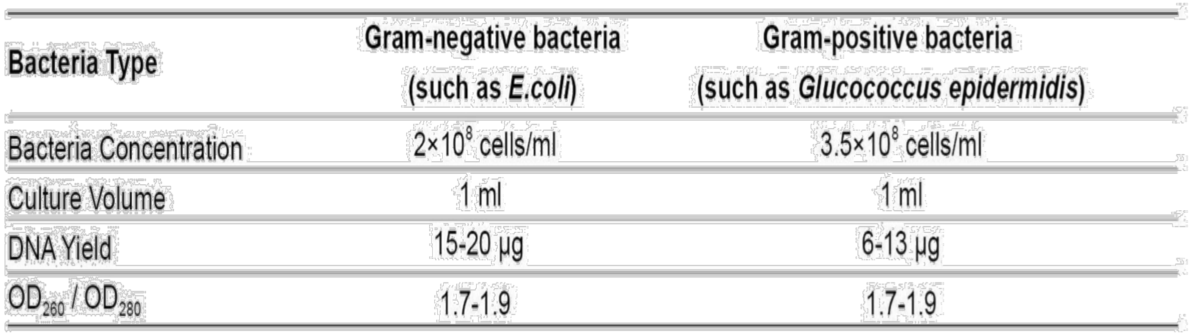
Diversity characterisation

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 *E.coli*)	Gram-positive bacteria (such as *Glucococcus epidermidis*)
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

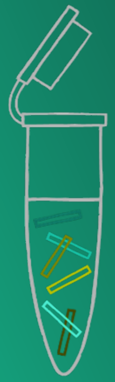
https://en.tiangen.com/content/details_43_4220.html

Amplification

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

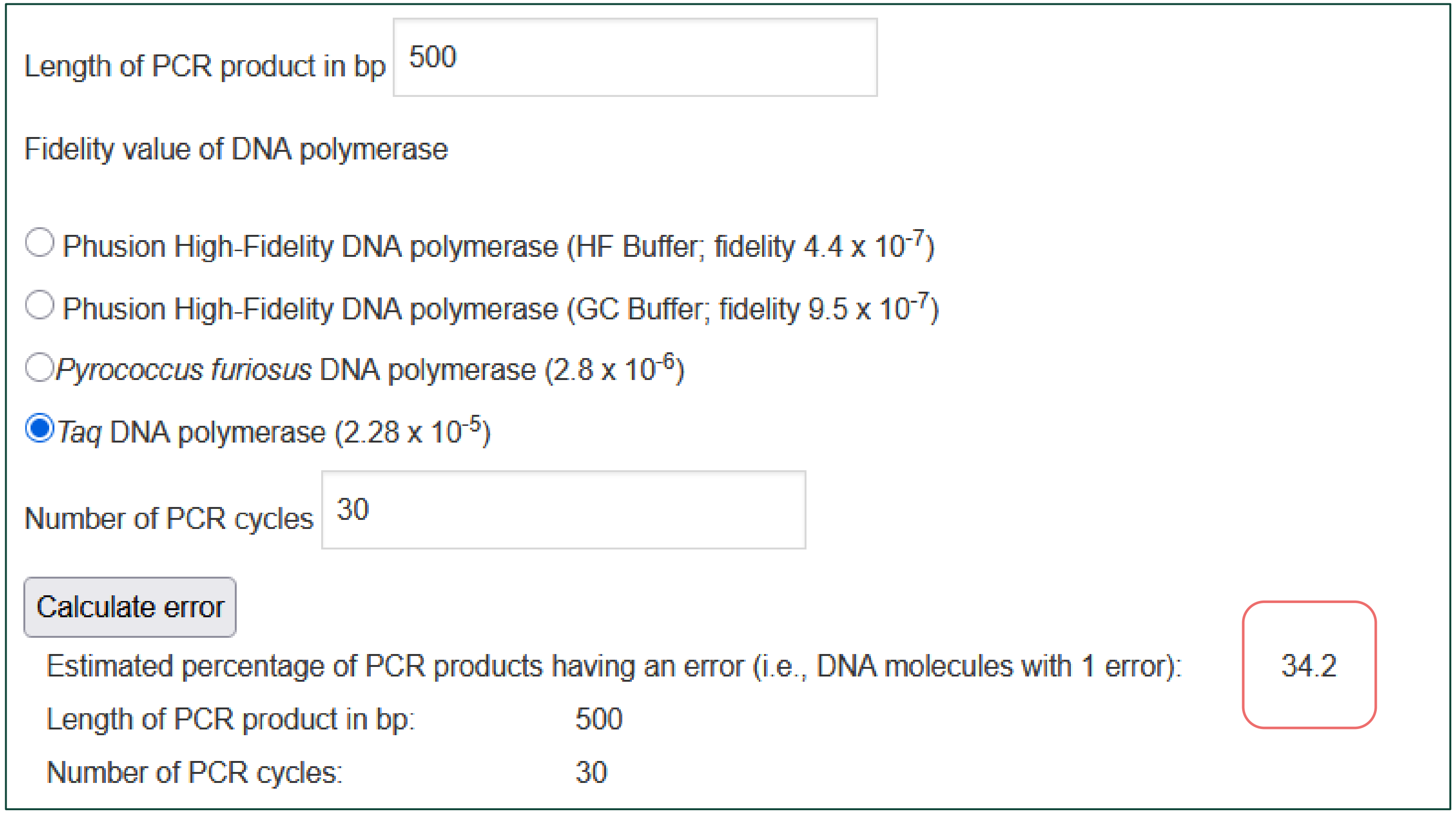
Polymerase	1 kb template	3 kb template
Phusion High-Fidelity DNA Polymerases (HF Buffer)	1.32%	3.96%
Phusion High-Fidelity DNA Polymerases (GC Buffer)	2.85%	8.55%
Pyrococcus furiosus DNA polymerase	8.4%	25.2%
Taq DNA polymerase	68.4%	205.2%

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.



https://www.thermofisher.com/fr/fr/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/pcr-fidelity-calculator.html

Amplification polymerase choice



Length of PCR product in bp 500

Fidelity value of DNA polymerase

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

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

○ *Pyrococcus furiosus* DNA polymerase (2.8 x 10^{-6})

● *Taq* DNA polymerase (2.28 x 10^{-5})

Number of PCR cycles 30

Estimated percentage of PCR products having an error (i.e., DNA molecules with 1 error): 34.2

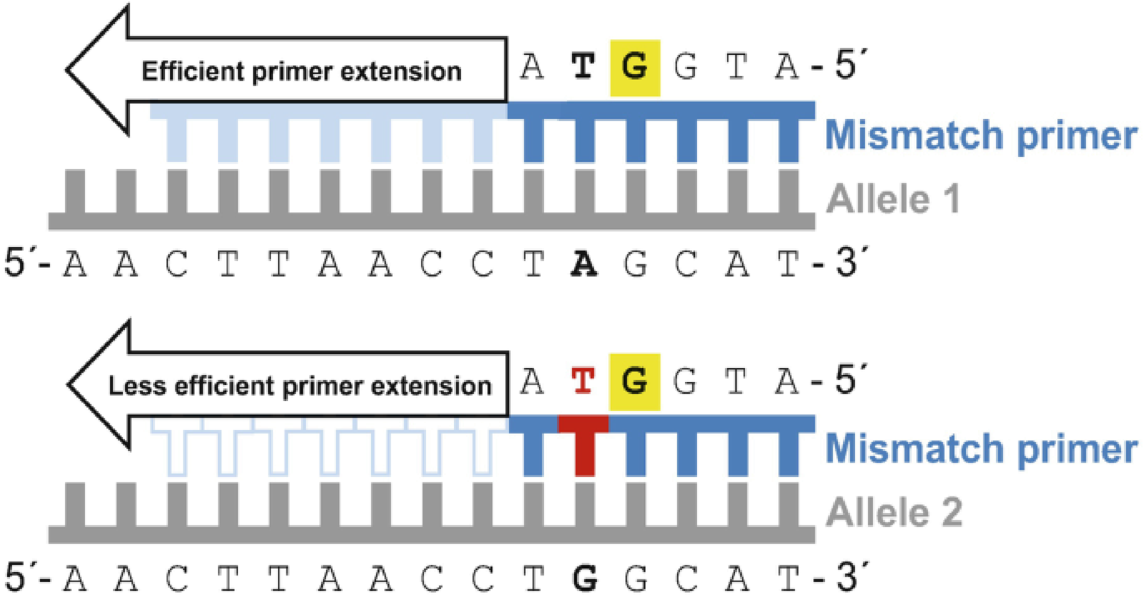
Length of PCR product in bp: 500

Number of PCR cycles: 30

https://www.thermofisher.com/fr/fr/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/pcr-fidelity-calculator.html

Amplification: efficiency

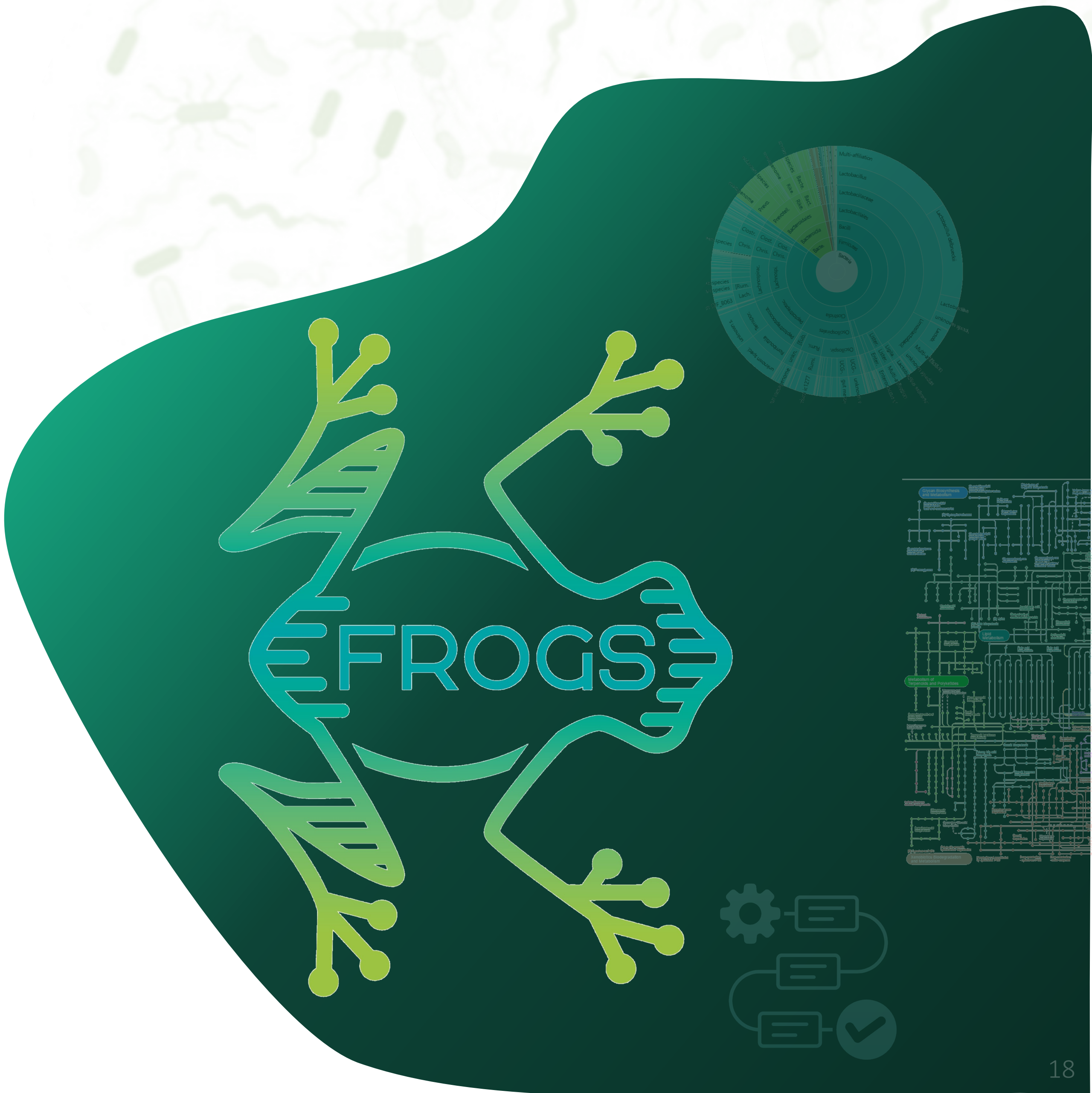
Not all microorganisms have the same amplification efficiency in PCR (depending on the primers). They may not even be amplified at all.


Efficient primer extension
A T G G T A - 5´
Mismatch primer
Allele 1
5´- A A C T T A A C C T A G C A T - 3´
Less efficient primer extension
A T G G T A - 5´
Mismatch primer
Allele 2
5´- A A C T T A A C C T G G C A T - 3´

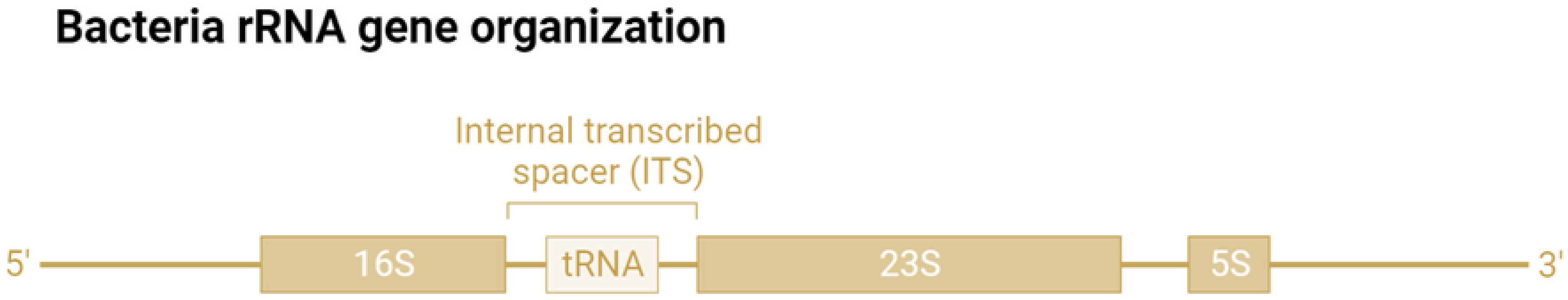
https://doi.org/10.1007/978-1-0716-1799-1_5

Metabarcoding overview

Short reads



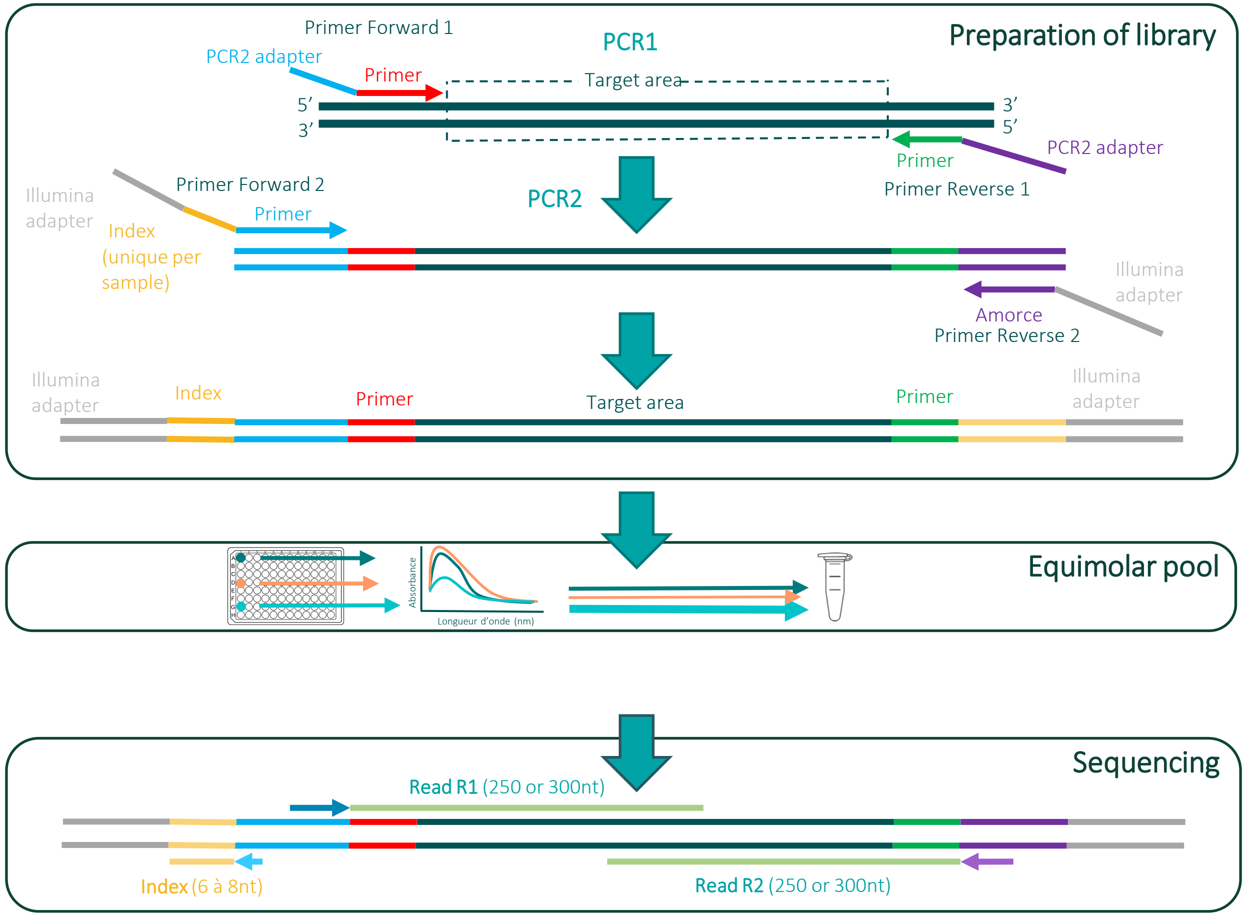
Short reads

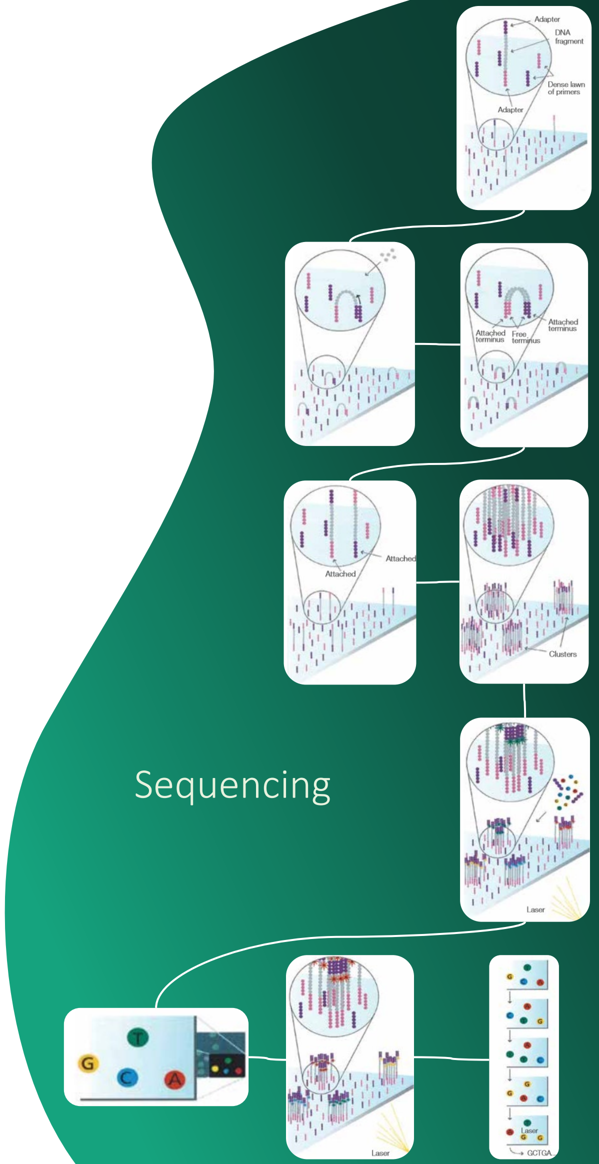

Bacteria rRNA gene organization
Internal transcribed spacer (ITS)
5'
16S
tRNA
23S
5S
3'

~500 nt

Steps for Illumina sequencing

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



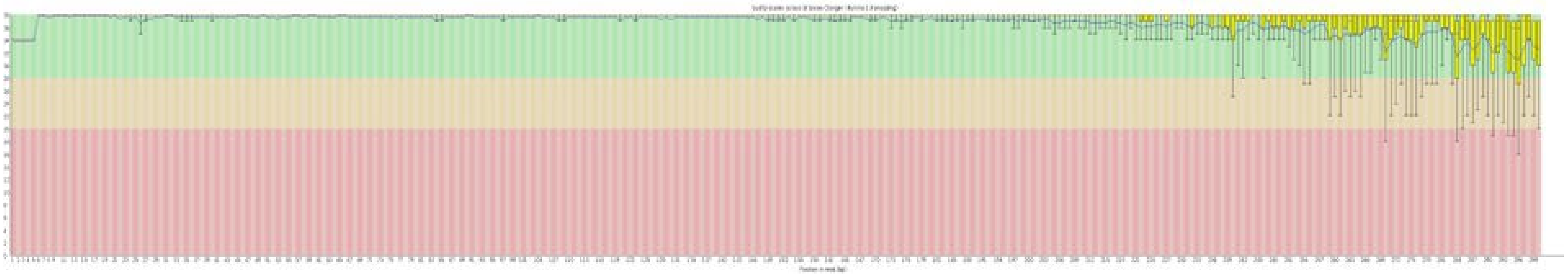


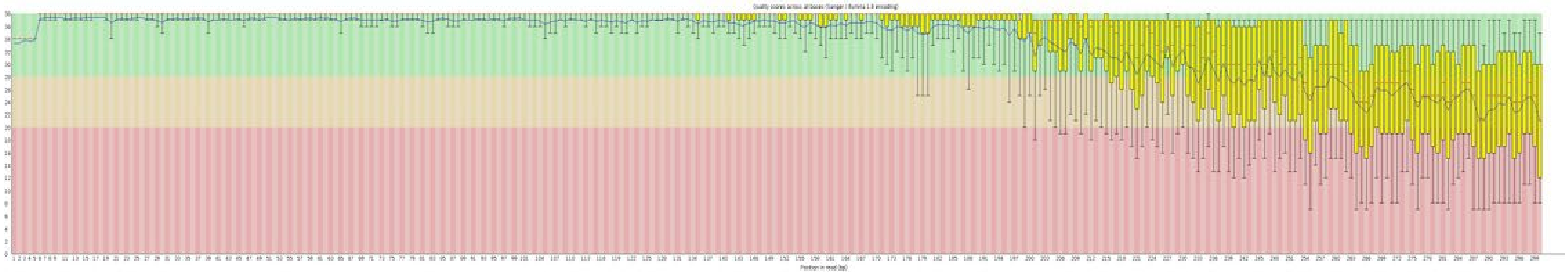
DOI:10.13140/RG.2.1.1003.1129

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



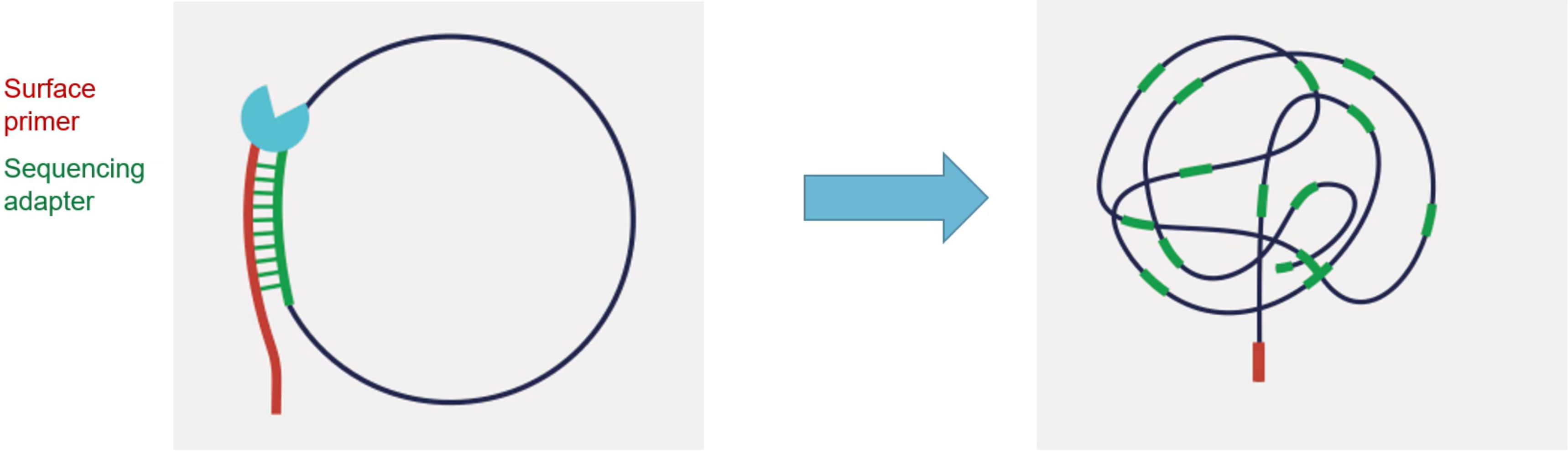


- ➕ Proven technologies that are easy to implement
- ➖ Data quality tends to decline at the end of the read
- ➖ Requires diversity input *via* phix → data loss

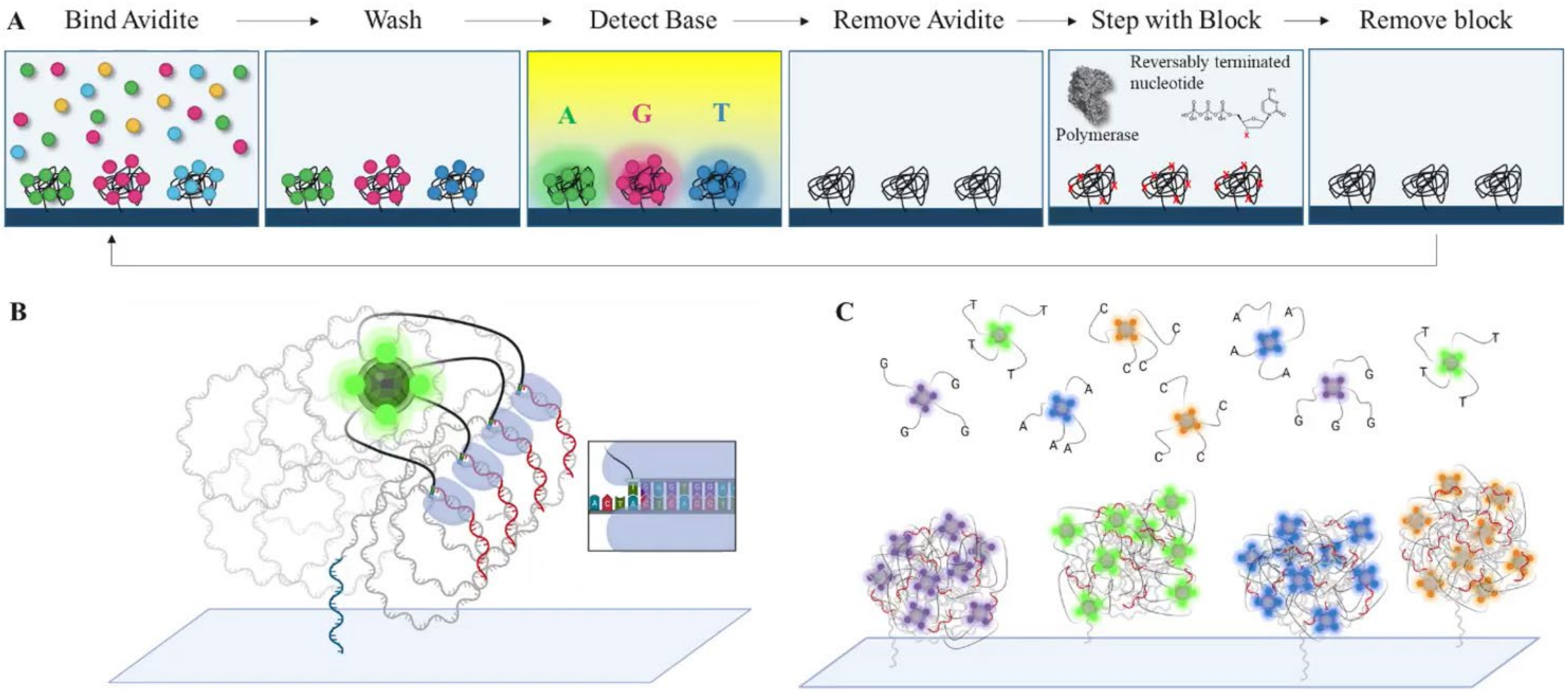
AVITI : polonies polymerization

- The AVITI generates **polonies** with rolling circle amplification (RCA) from circular templates

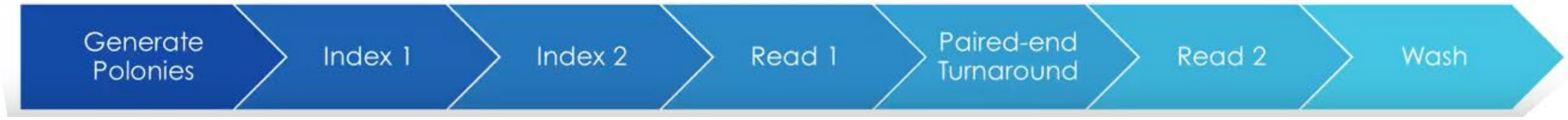
- Each **polony** is a continuous DNA strand with many sequencing start sites
- No on-instrument PCR reduces errors by only copying from the original



AVITI : sequencing

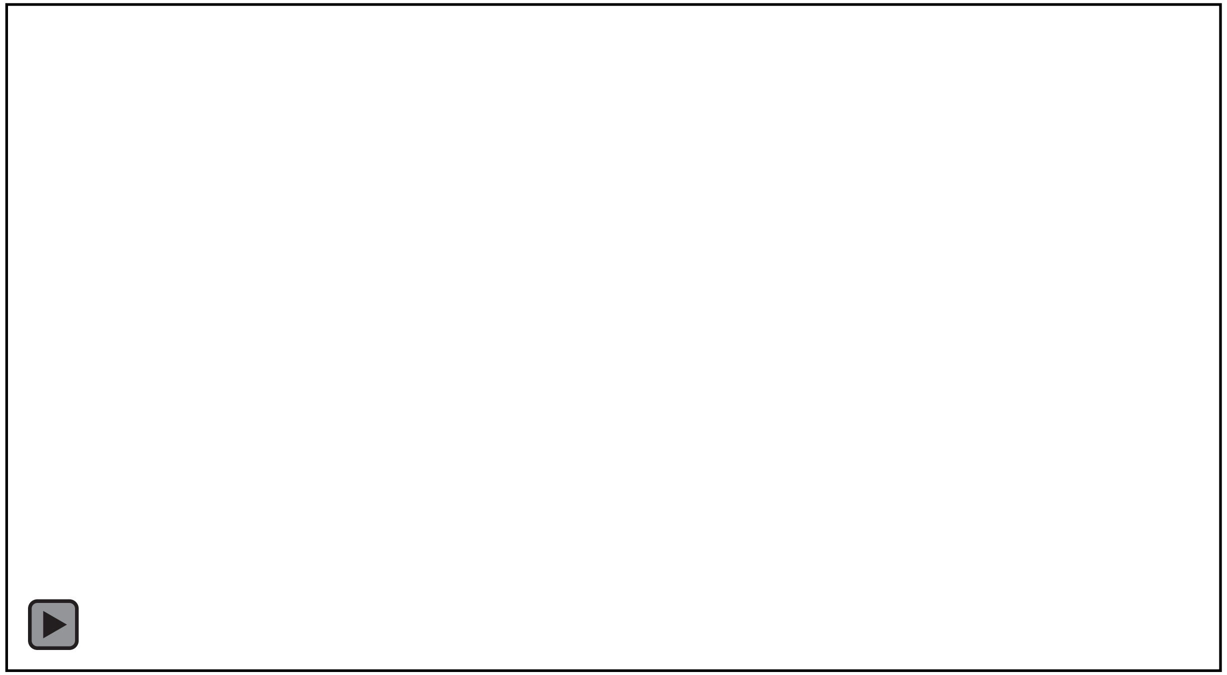


Instrument Workflow:


Generate Polonies → Index 1 → Index 2 → Read 1 → Paired-end Turnaround → Read 2 → Wash

https://bit.ly/3vAh6gE

AVITI sequencing



AVITI sequencing (Element Biosciences)

2x300 bp

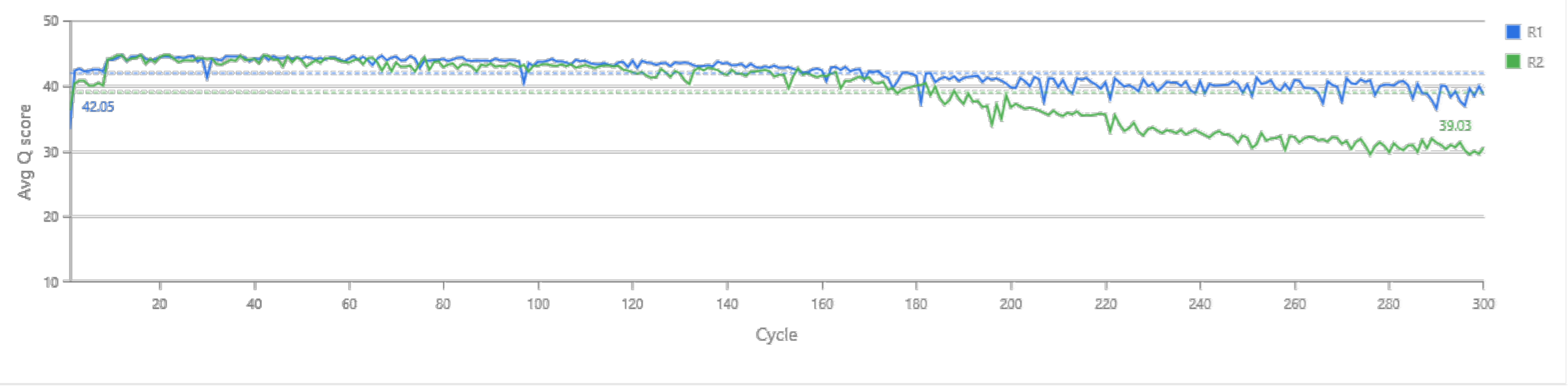
7.5 to 150 Gb per lane

2 lanes per Flowcell

38-hour run

Compatible with Illumina libraries

90% bases > Q30



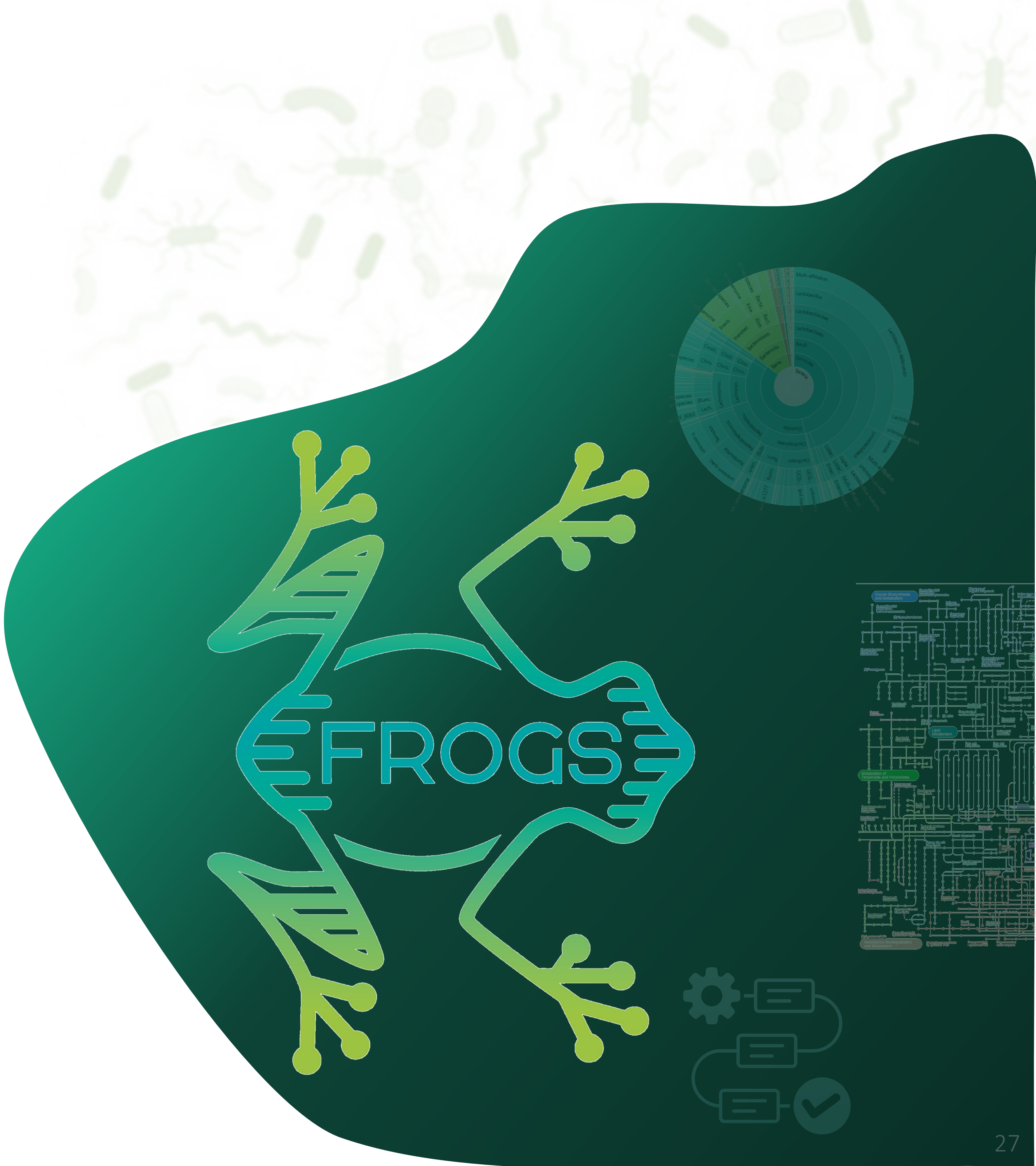
⊕ Excellent data quality throughout the reads.

⊖ Very large volume of data.

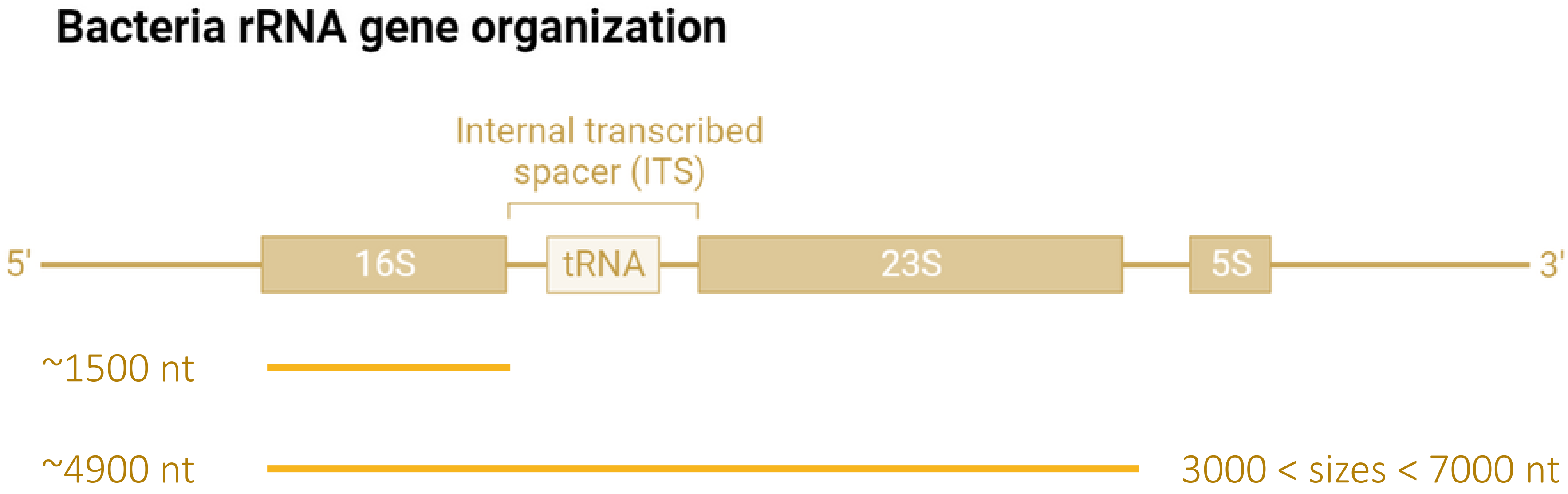
⚠ Different technology → different analysis results.

Metabarcoding overview

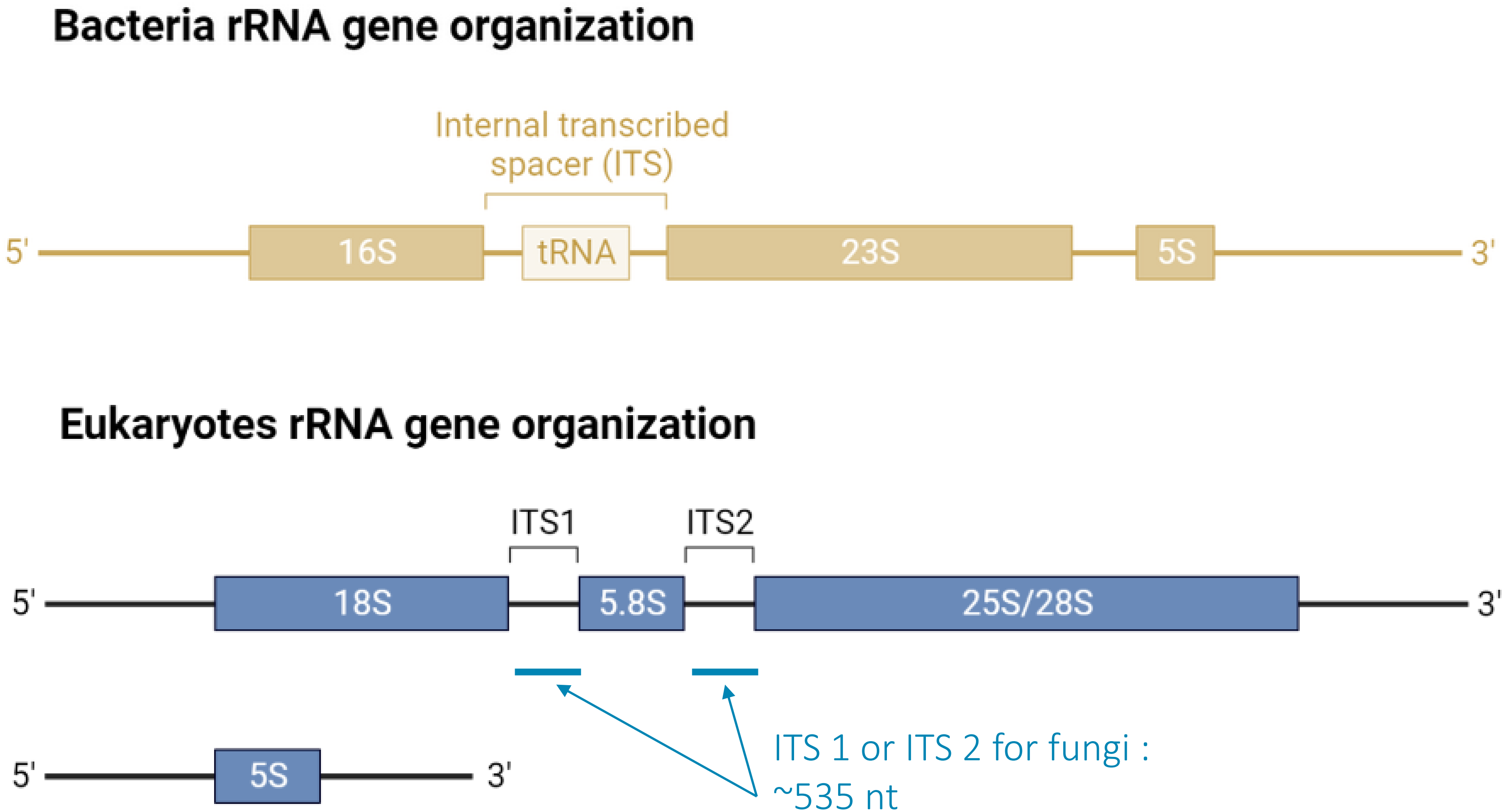
Long reads



Long reads

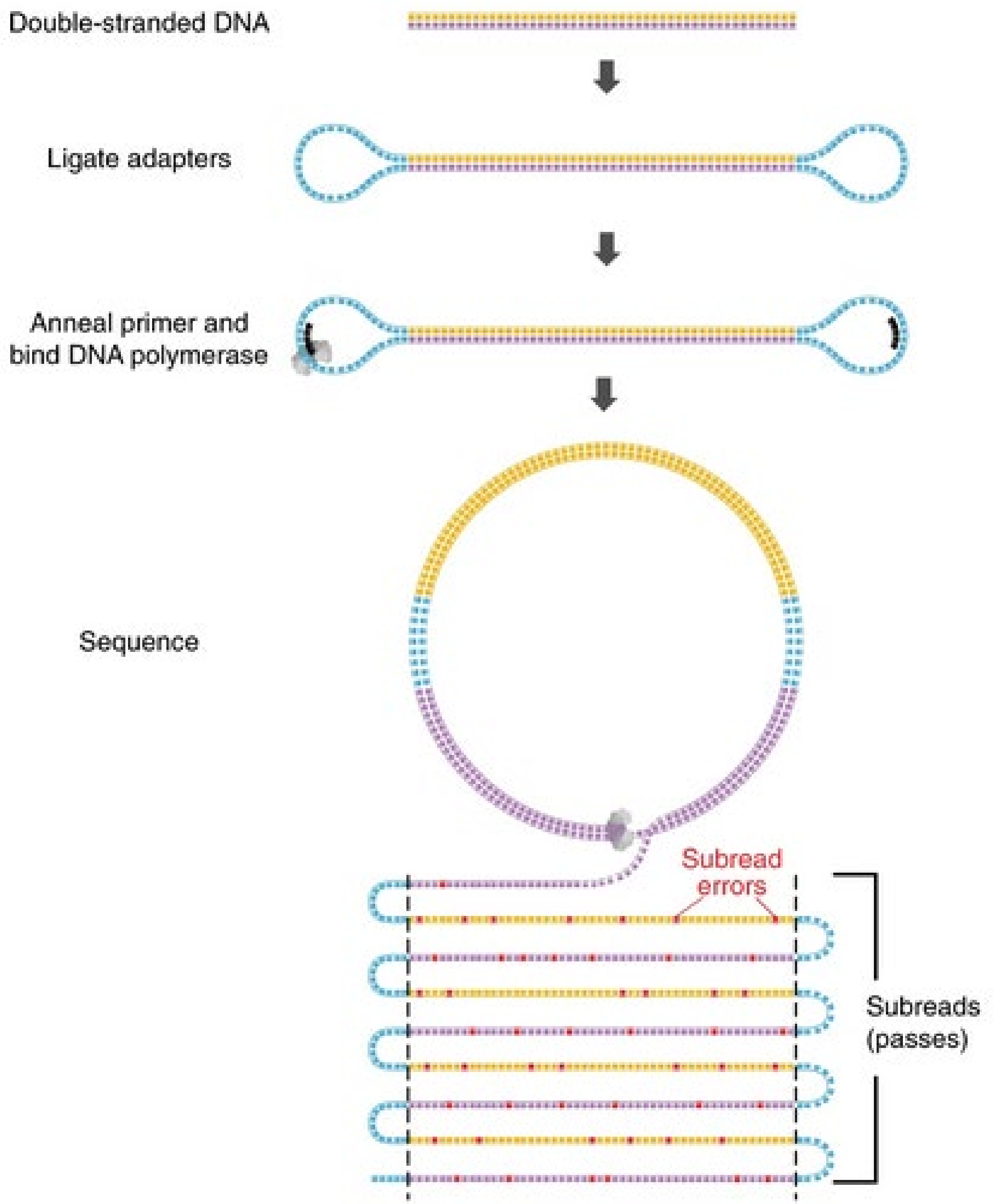

Bacteria rRNA gene organization
Internal transcribed spacer (ITS)
5'
16S
tRNA
23S
5S
3'
~1500 nt
~4900 nt
3000 < sizes < 7000 nt

Long reads

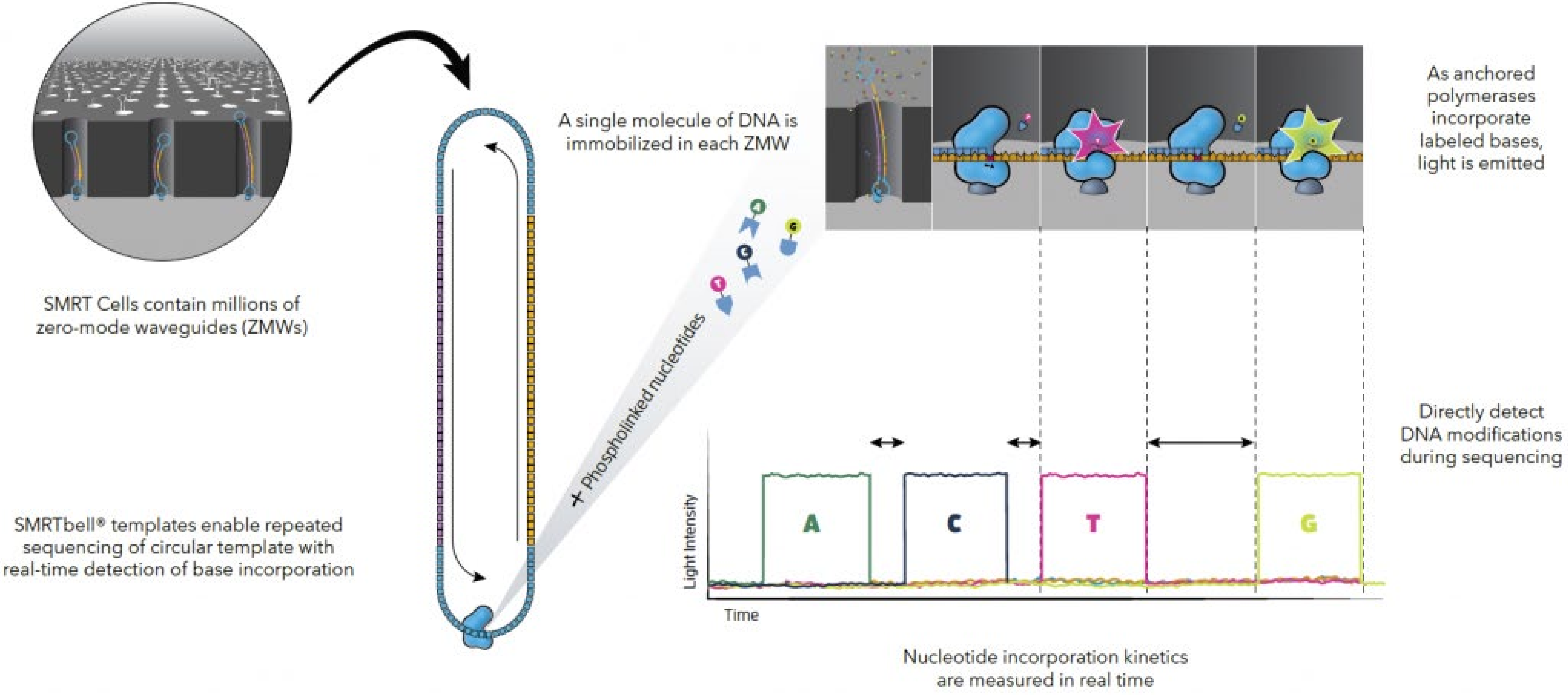

Bacteria rRNA gene organization
Internal transcribed spacer (ITS)
5'
16S
tRNA
23S
5S
3'
Eukaryotes rRNA gene organization
ITS1
ITS2
5'
18S
5.8S
25S/28S
3'
5'
5S
3'

ITS 1 or ITS 2 for fungi :
~535 nt
140 < sizes < 3173 nt

PacBio Sequencing



PacBio Sequencing



SMRT Cells contain millions of zero-mode waveguides (ZMWs)

SMRTbell® templates enable repeated sequencing of circular template with real-time detection of base incorporation

A single molecule of DNA is immobilized in each ZMW

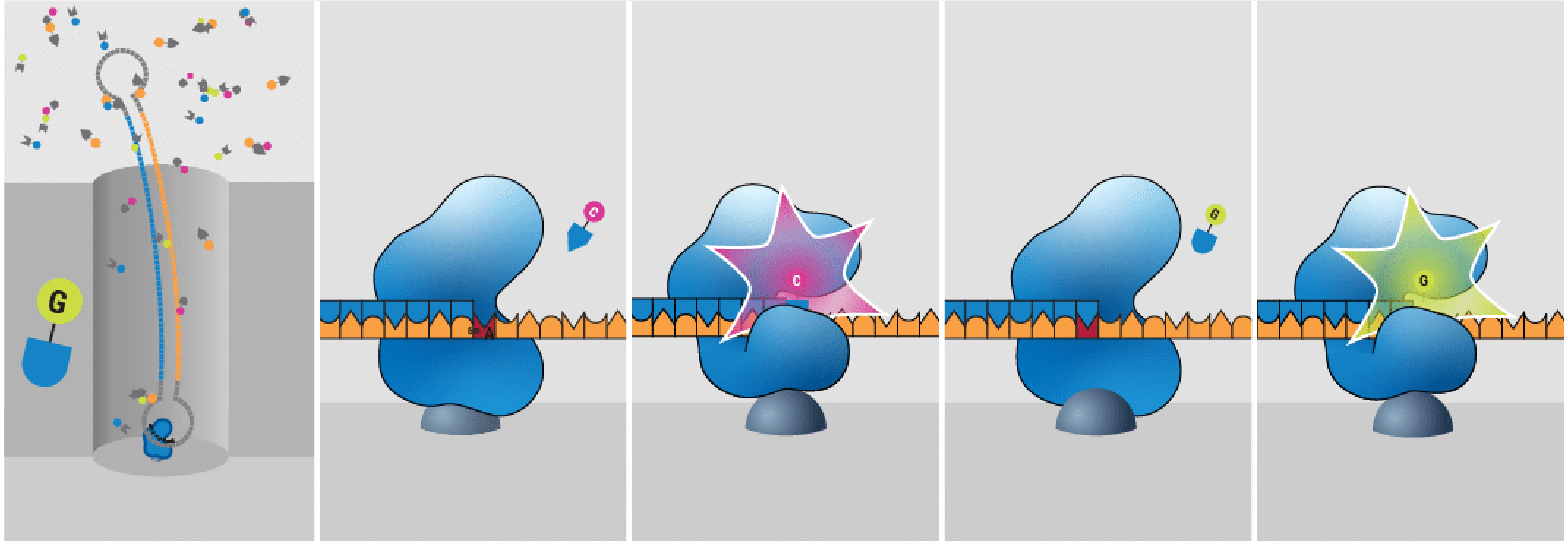
As anchored polymerases incorporate labeled bases, light is emitted

Directly detect DNA modifications during sequencing

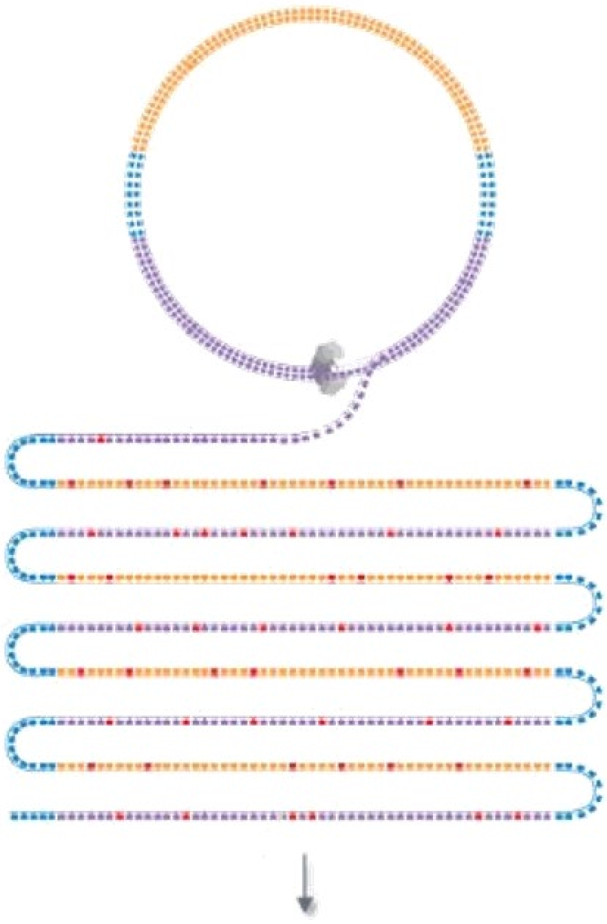
Nucleotide incorporation kinetics are measured in real time

Single-Molecule Real-Time sequencing technology (SMRT)

PacBio Sequencing



PacBio Sequencing



HiFi read

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

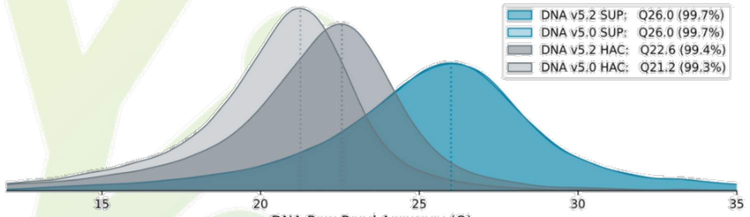
1 pass → 90 % accuracy

2 passes → 96 % accuracy

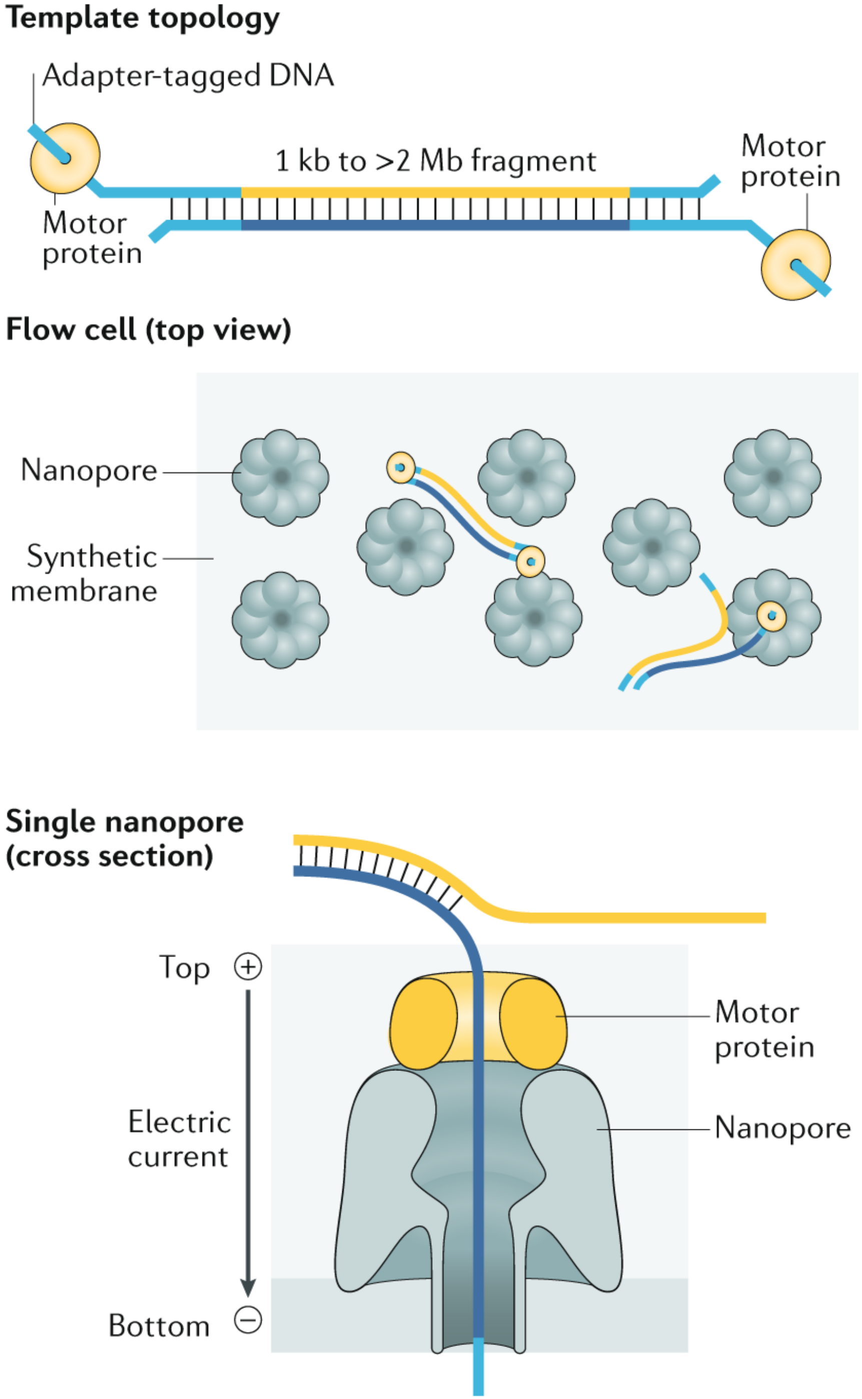
3 passes → 99 % accuracy

ONT sequencing

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



high accuracy (HAC, v5.2)
super accuracy (SUP, v5.0)



Template topology

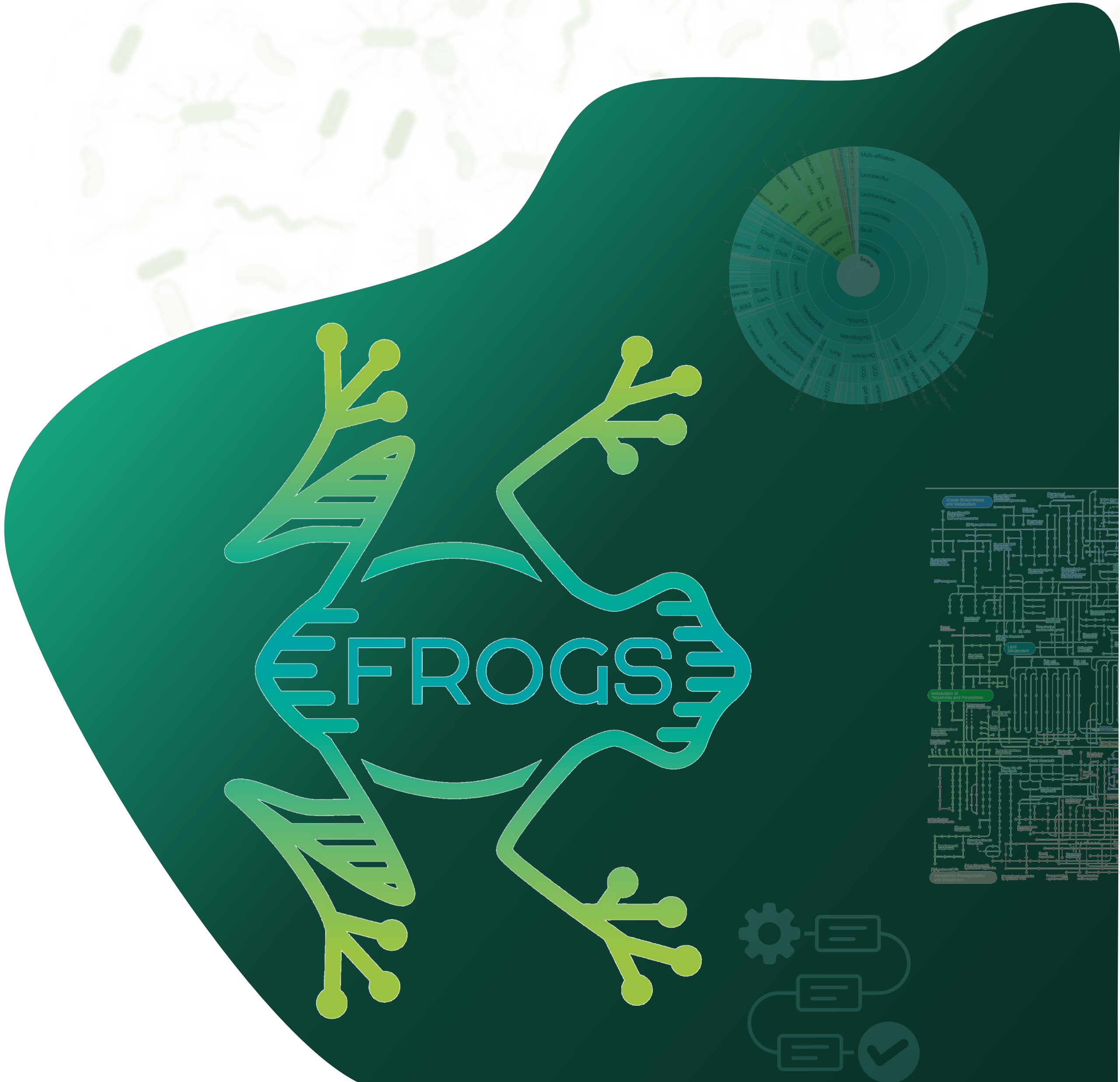
Flow cell (top view)

Single nanopore (cross section)

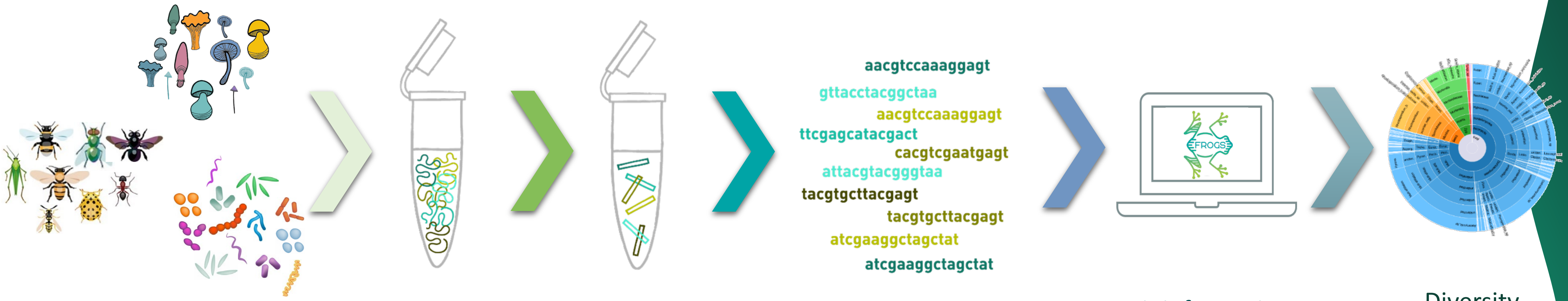
https://doi.org/10.1038/s41576-020-0236-x

Metabarcoding overview

Bioinformatics analysis



Methods and parameter of tools impact final diversity view



Community analysis

- Biological sampling
- Collection methods

DNA extraction

- Collection methods

Barcode amplification

- PCR biais
- Primer choice

Amplicon sequencing

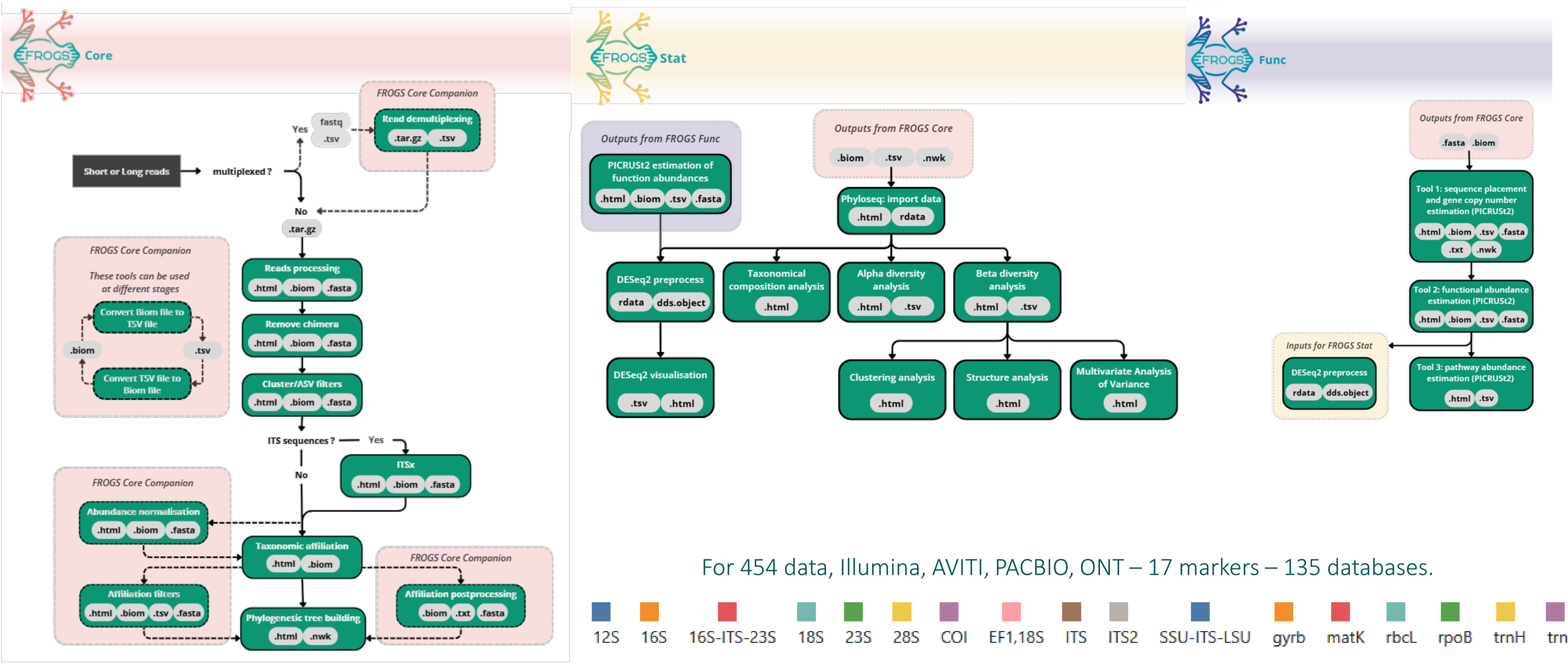
- Length of reads
- Sequencing protocol

Bioinformatics process

- Software choice
- Parameter/filter choice

Diversity characterisation

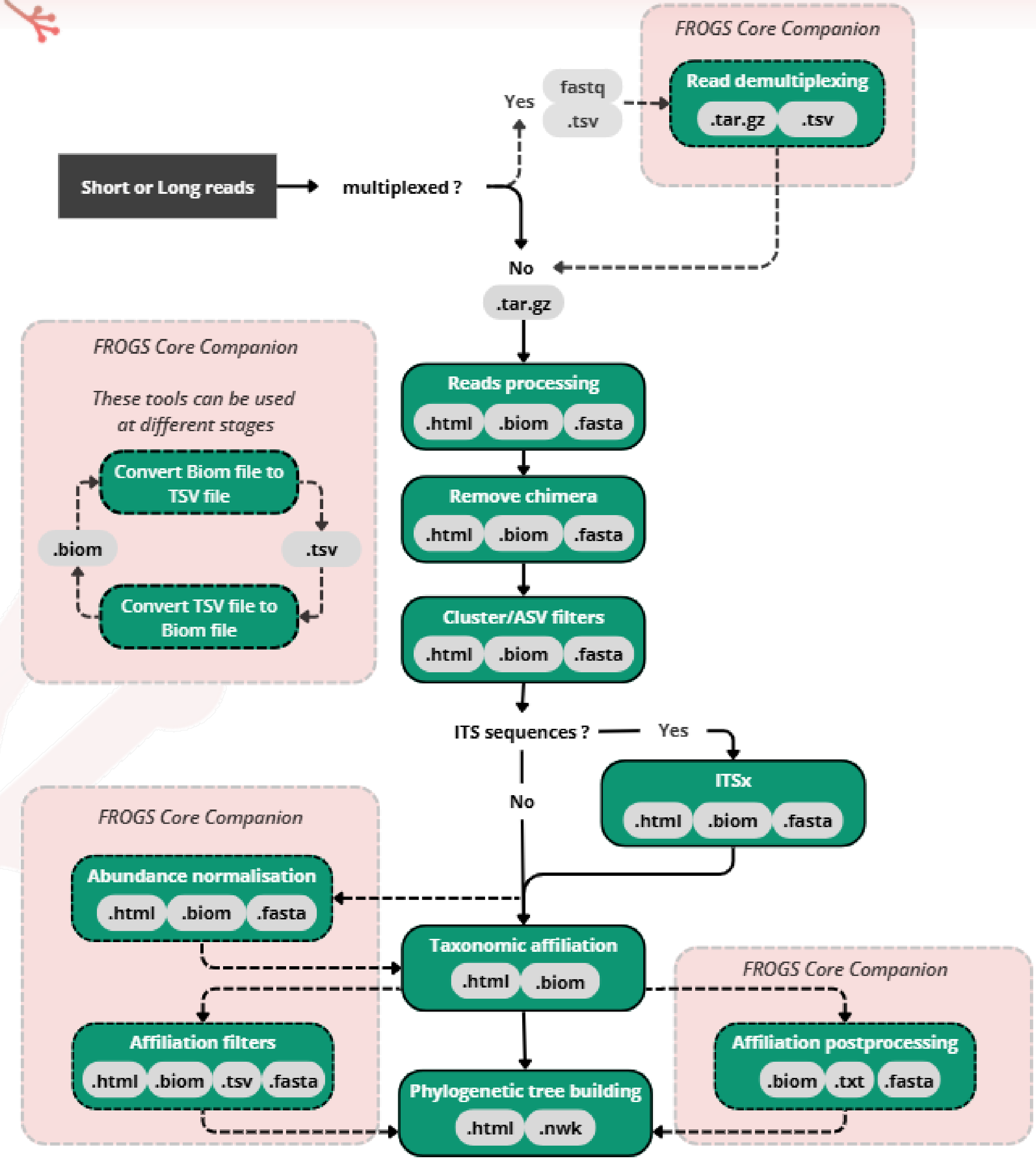
FROGS = 3 tool groups



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

FROGS Core



Short or Long reads → multiplexed ?

Yes → fastq, .tsv → FROGS Core Companion: Read demultiplexing (.tar.gz, .tsv)

No → .tar.gz

Reads processing (.html, .biom, .fasta)

Remove chimera (.html, .biom, .fasta)

Cluster/ASV filters (.html, .biom, .fasta)

ITS sequences ? Yes → ITSx (.html, .biom, .fasta); No

Taxonomic affiliation (.html, .biom)

Phylogenetic tree building (.html, .nwk)

FROGS Core Companion

These tools can be used at different stages

Convert Biom file to TSV file (.biom → .tsv)

Convert TSV file to Biom file (.tsv → .biom)

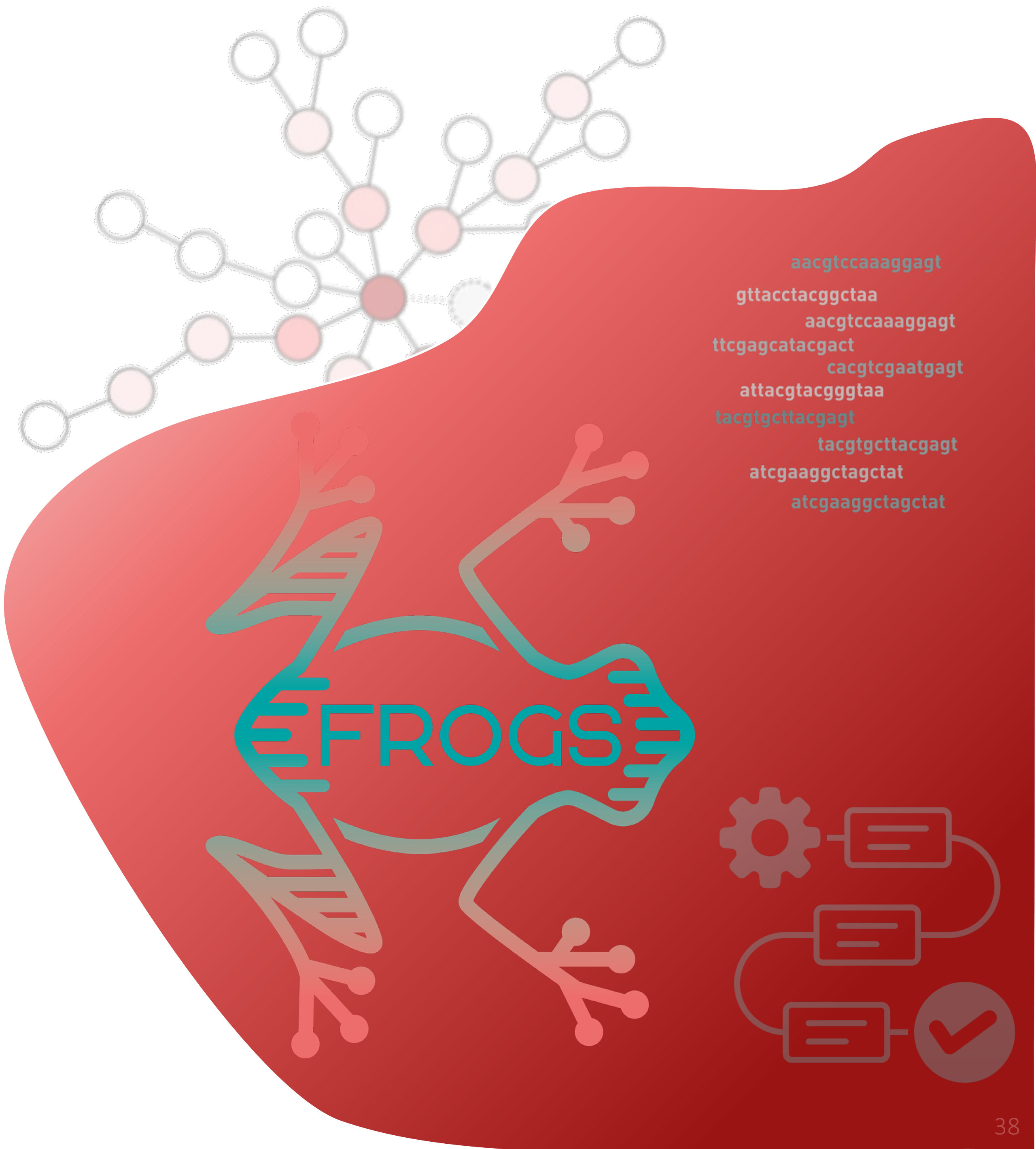
FROGS Core Companion

Abundance normalisation (.html, .biom, .fasta)

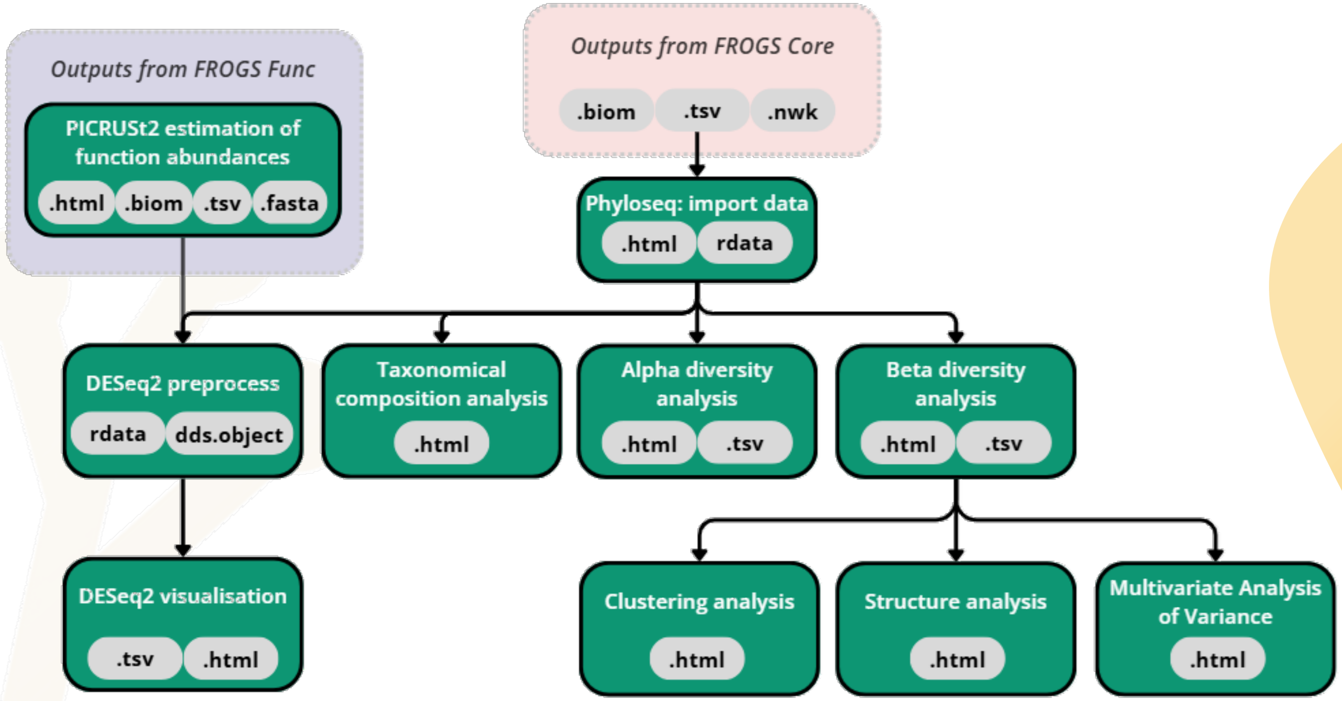
Affiliation filters (.html, .biom, .tsv, .fasta)

FROGS Core Companion

Affiliation postprocessing (.biom, .txt, .fasta)



FROGS Stat



Outputs from FROGS Func

PICRUSt2 estimation of function abundances
.html .biom .tsv .fasta

Outputs from FROGS Core

.biom .tsv .nwk

Phyloseq: import data
.html rdata

DESeq2 preprocess
rdata dds.object

Taxonomical composition analysis
.html

Alpha diversity analysis
.html .tsv

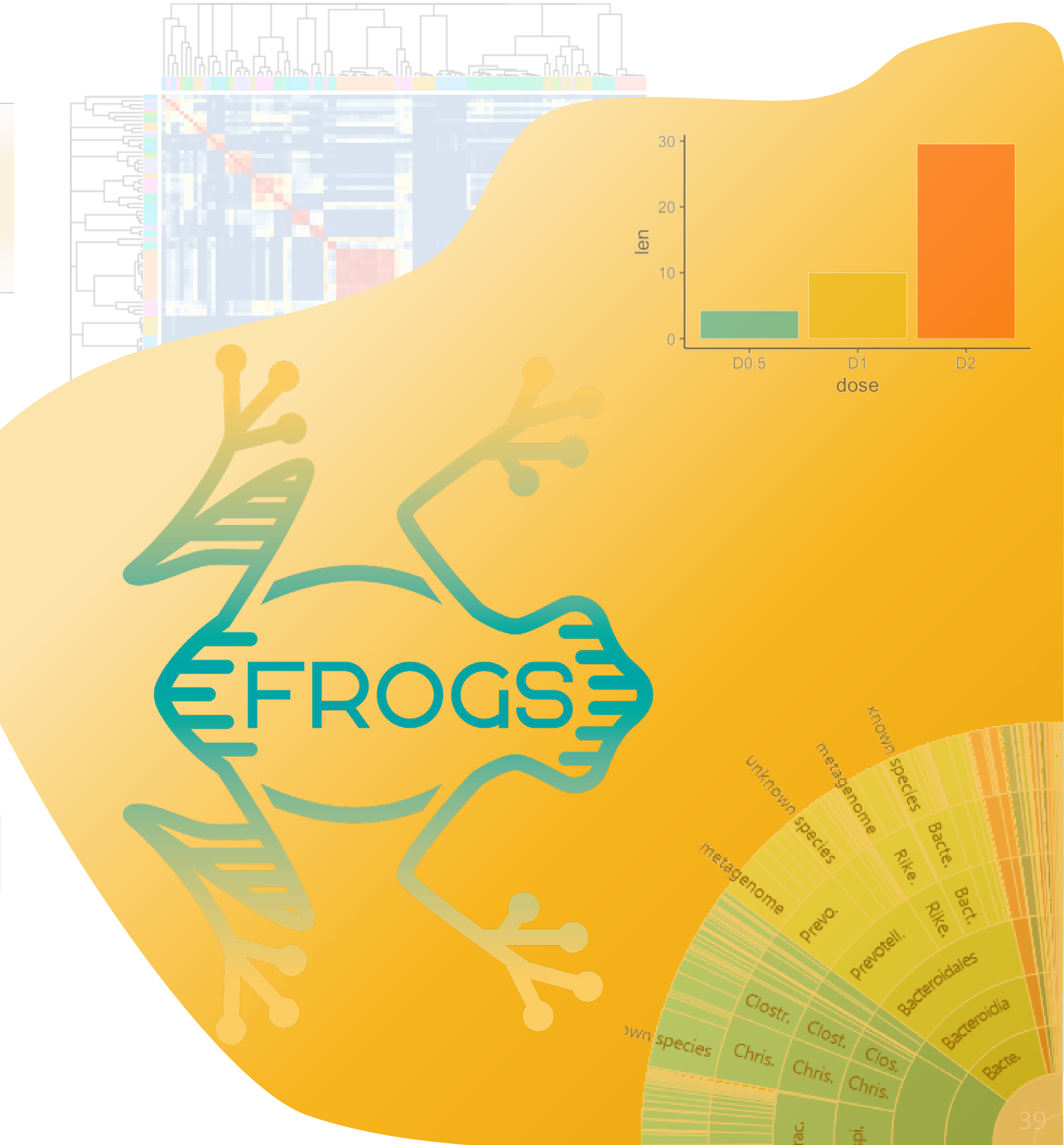
Beta diversity analysis
.html .tsv

DESeq2 visualisation
.tsv .html

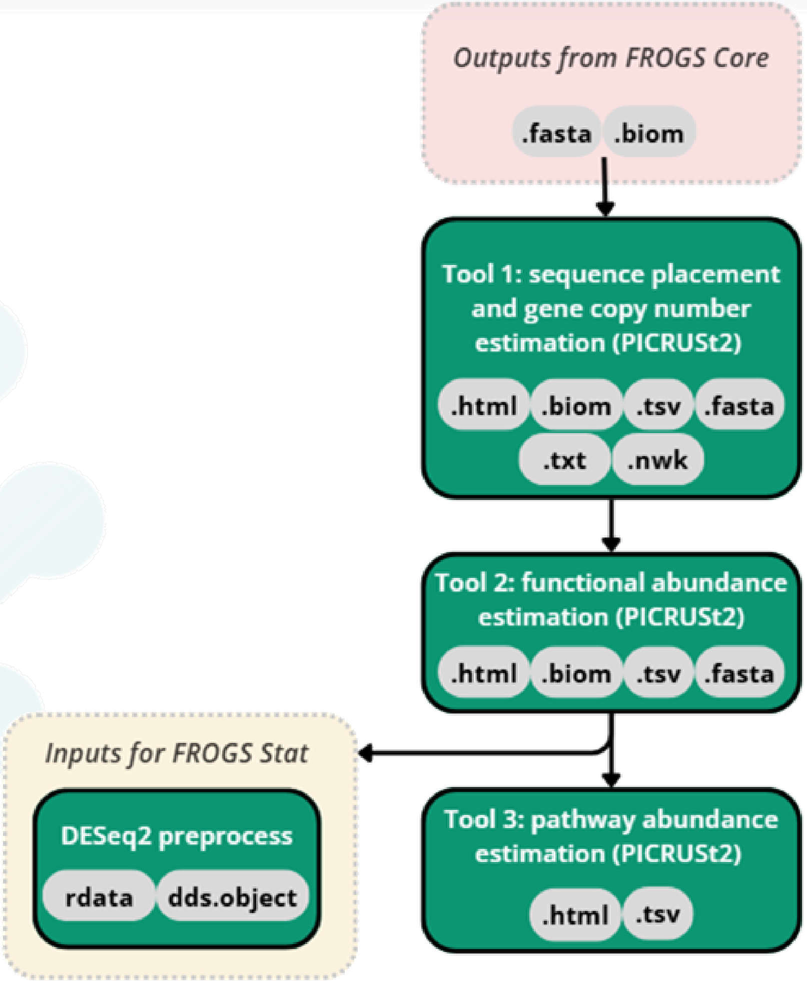
Clustering analysis
.html

Structure analysis
.html

Multivariate Analysis of Variance
.html



FROGS Func



Outputs from FROGS Core

.fasta .biom

Tool 1: sequence placement and gene copy number estimation (PICRUSt2)

.html .biom .tsv .fasta .txt .nwk

Tool 2: functional abundance estimation (PICRUSt2)

.html .biom .tsv .fasta

Tool 3: pathway abundance estimation (PICRUSt2)

.html .tsv

Inputs for FROGS Stat

DESeq2 preprocess

rdata dds.object