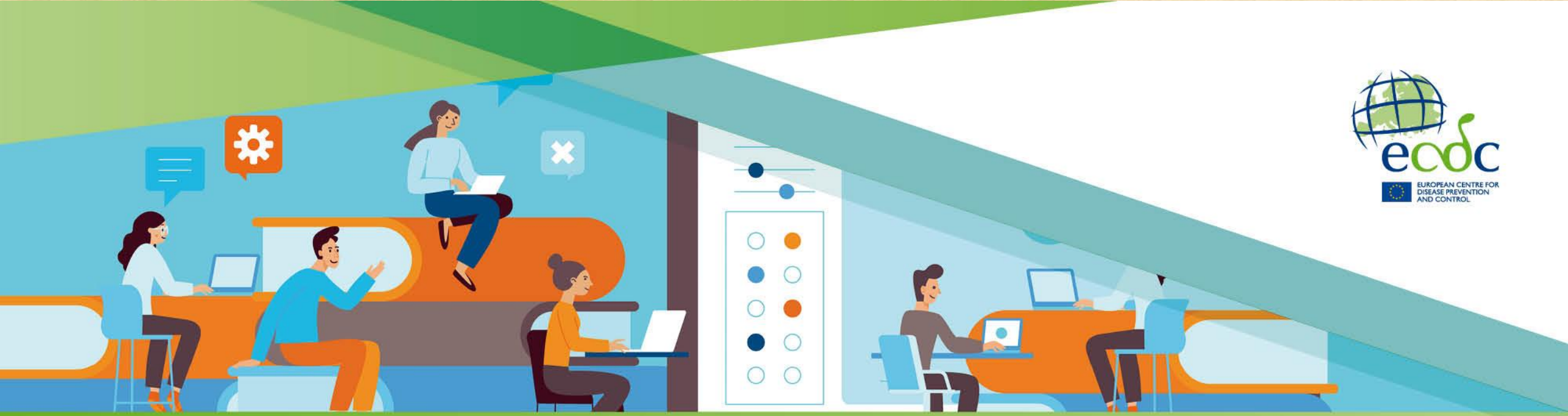


GenEpi-BioTrain Virtual Training 27

Clinical application of Nanopore for diagnostic 16S rDNA sequencing analysis

2026-06-01

Anna Norén, Elin Loo, Lili Andersson-Li, Samuel Lampa, Sofia Stamouli

Session 1 Part 2

Open source tools for 16S analysis

Aim and learning objectives

Aim

This session covers Nanopore full-length 16S rDNA sequencing for bacterial identification and community profiling, focusing on the bioinformatics part – from quality control and taxonomic classification to reporting - using the EMU algorithm and TRANA pipeline.

Specific learning objectives

At the end of this session, you will be able to:

- Process Nanopore 16S rDNA data from raw reads to taxonomic profiles
- Describe EMU's probabilistic read assignment and expectation-maximization approach
- Explain the TRANA pipeline's main stages: QC, filtering, classification, and reporting
- Understand abundance tables and KRONA-based visualizations

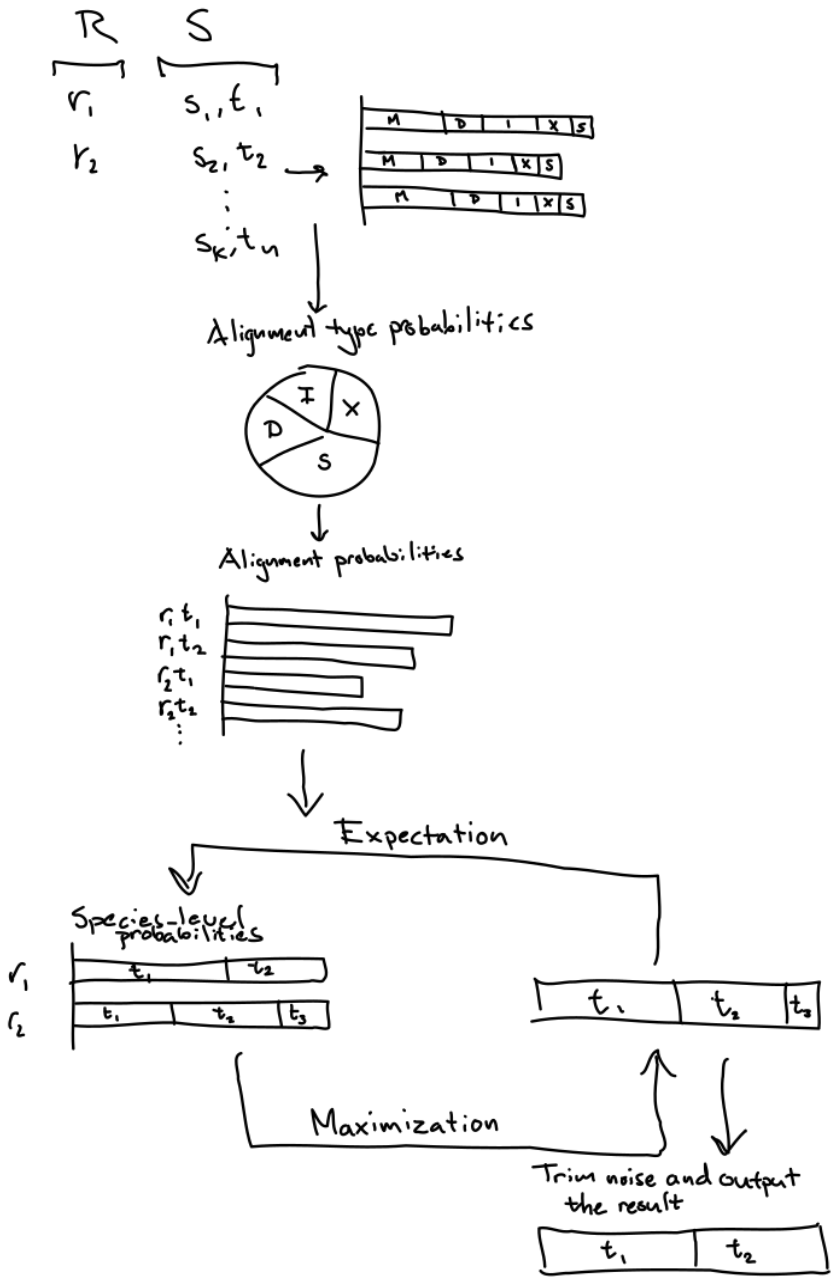
Part 2: The Emu tool and the Trana pipeline

Tools for analysis of Nanopore 16S rDNA data:

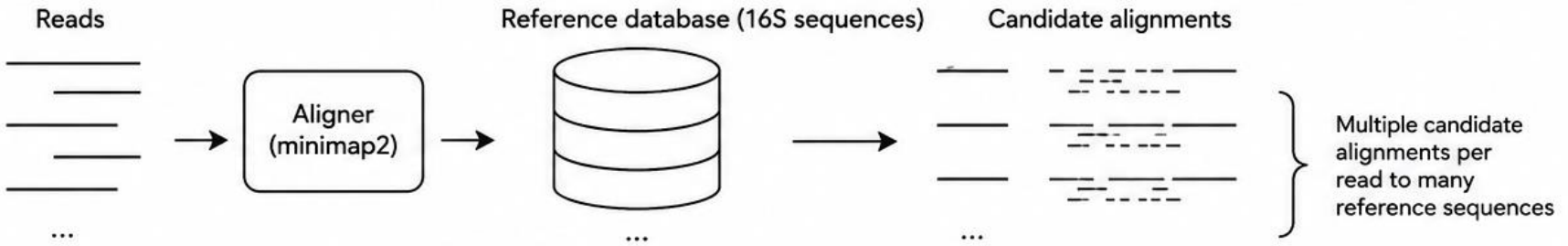
- Emu – Abundance estimation tool
- Trana - Emu-based pipeline
- TranaVy - Reporting tool for Trana

The Emu tool

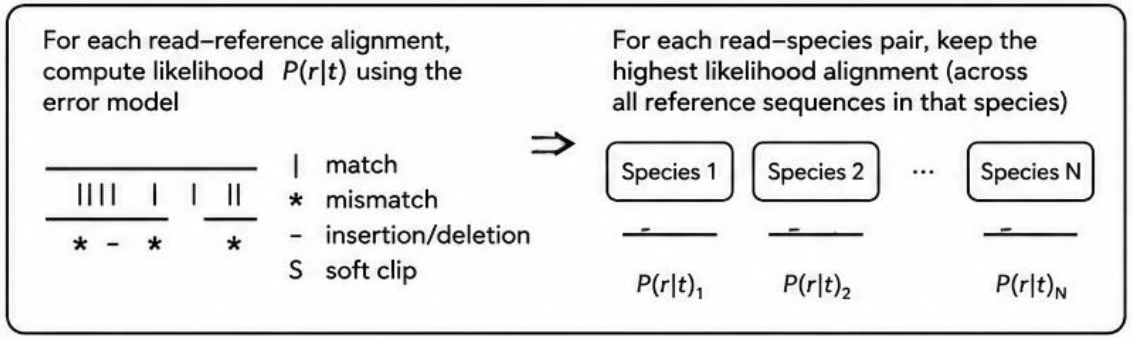
- Aligns every read separately to a full-length 16S reference database
- Computes likelihood for each read-species pair, keeping the best scoring reference sequence within each species
- Calculates and uses a probabilistic model of ONT sequencing errors to score alignments
- Iteratively applies an expectation maximization algorithm to refine both the overall species abundances and the probability estimate of different species for each read



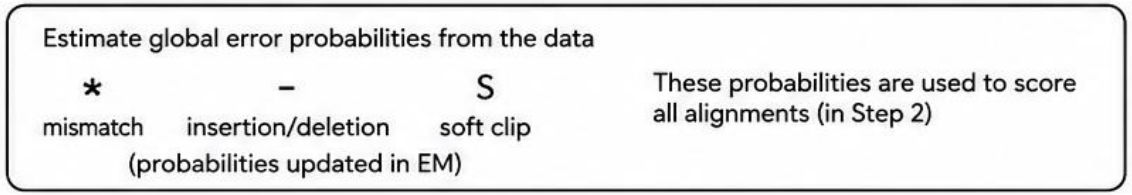
1 Align reads to a full-length 16S reference database



2 Compute alignment likelihoods for each read–species pair



3 Use a probabilistic model of ONT sequencing errors



4 Iterative Expectation–Maximization (EM) algorithm

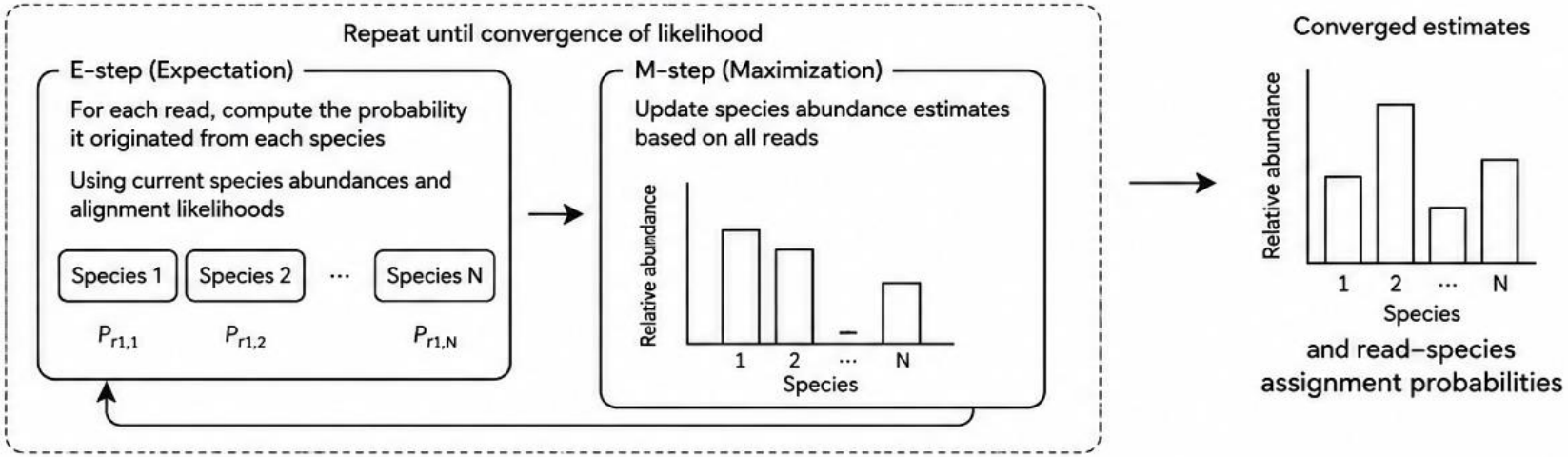
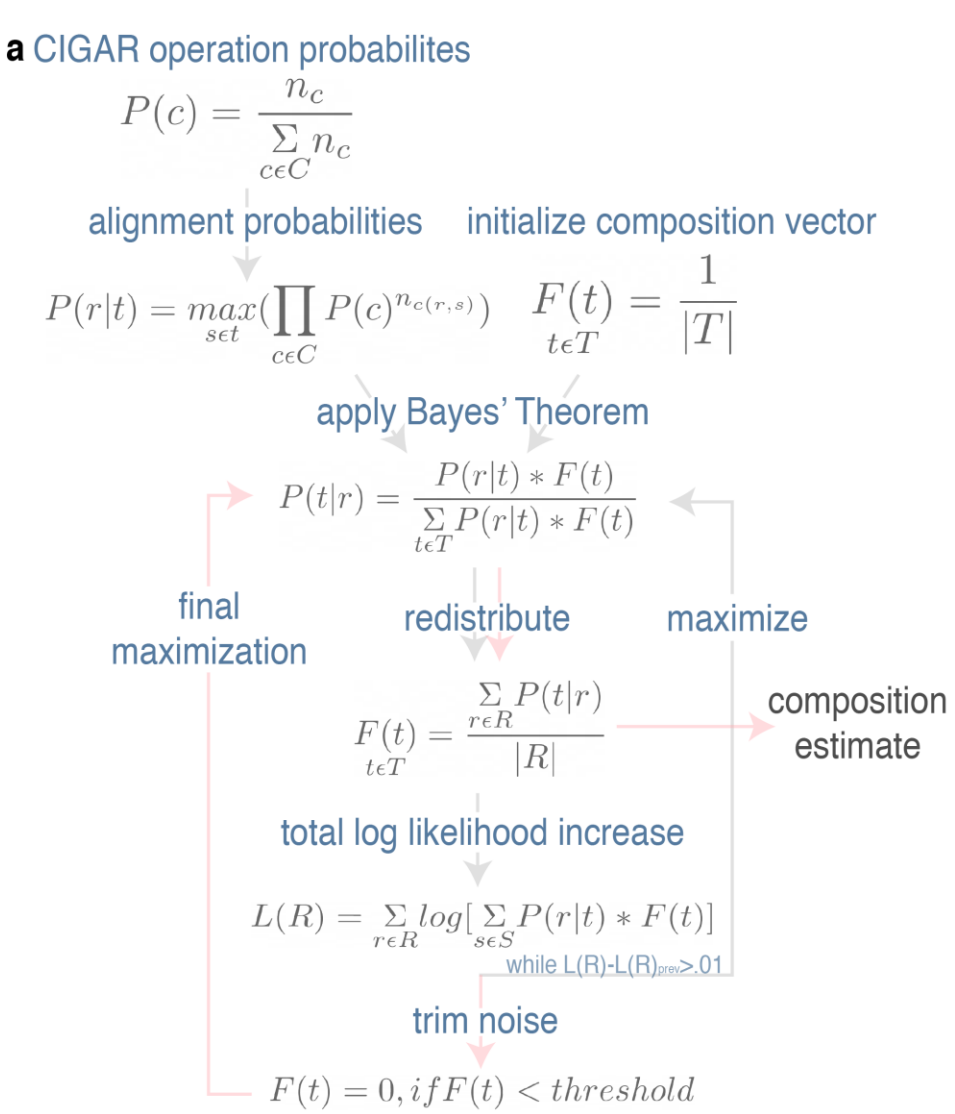


Image: Generated with ChatGPT / Dall-E based on a detailed description of the process. The image has been manually verified for accuracy.

Emu algorithm in depth



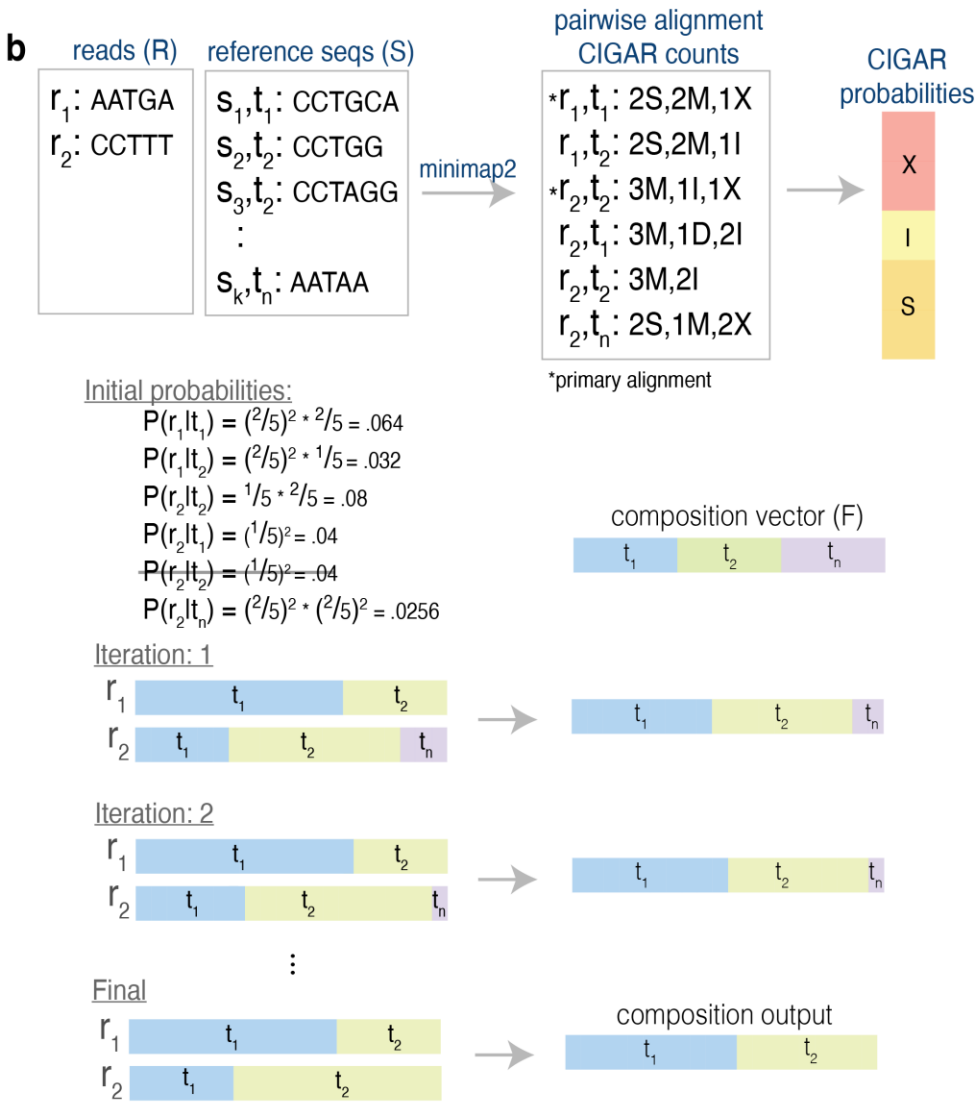


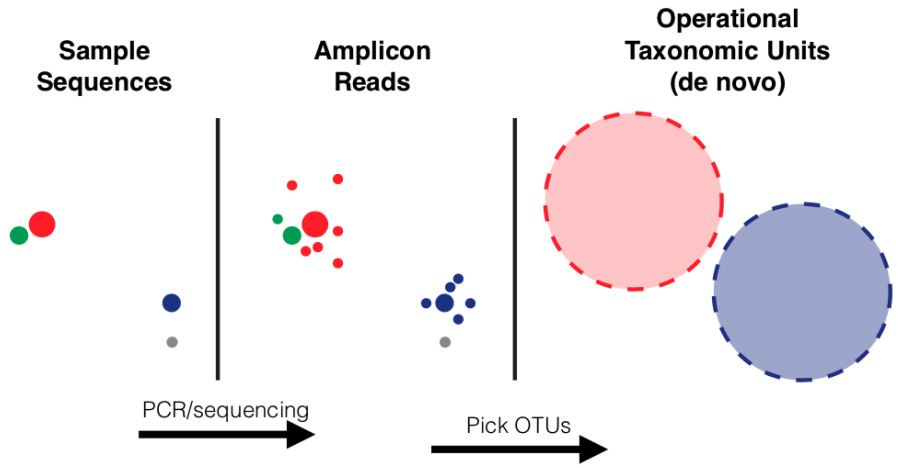
Image source: Curry et al., *Emu: Species-Level Microbial Community Profiling for Full-Length Nanopore 16S Reads*, bioRxiv (2021), https://doi.org/10.1101/2021.05.02.442339, CC BY-ND 4.0

How does Emu handle Nanopore sequencing errors?

🤔

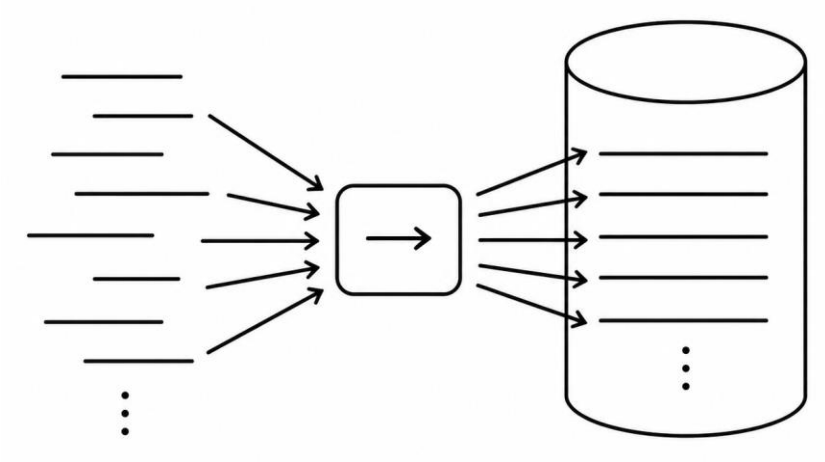
How does Emu handle Nanopore sequencing errors?

OTUs vs Emu approach



Classical OTU approach
(OTU = Operational Taxonomical Unit)

- **Approach:** Clustering and creating consensus sequences for alignment with database
- **Pros:** Easier interpretation of accuracy
- **Cons:** Information lost in the consensus sequence generation



Emu approach

- **Approach:** Align individual reads against database. Using stats on errors and an EM algorithm to optimize species probability per read
- **Pros:** Better accuracy for Nanopore data
- **Cons:** More computationally demanding + somewhat more complicated accuracy metrics

What is the main benefit of Emu compared to classical OTU-based approaches?

🤔

What is the main benefit of Emu compared to classical OTU-based approaches?

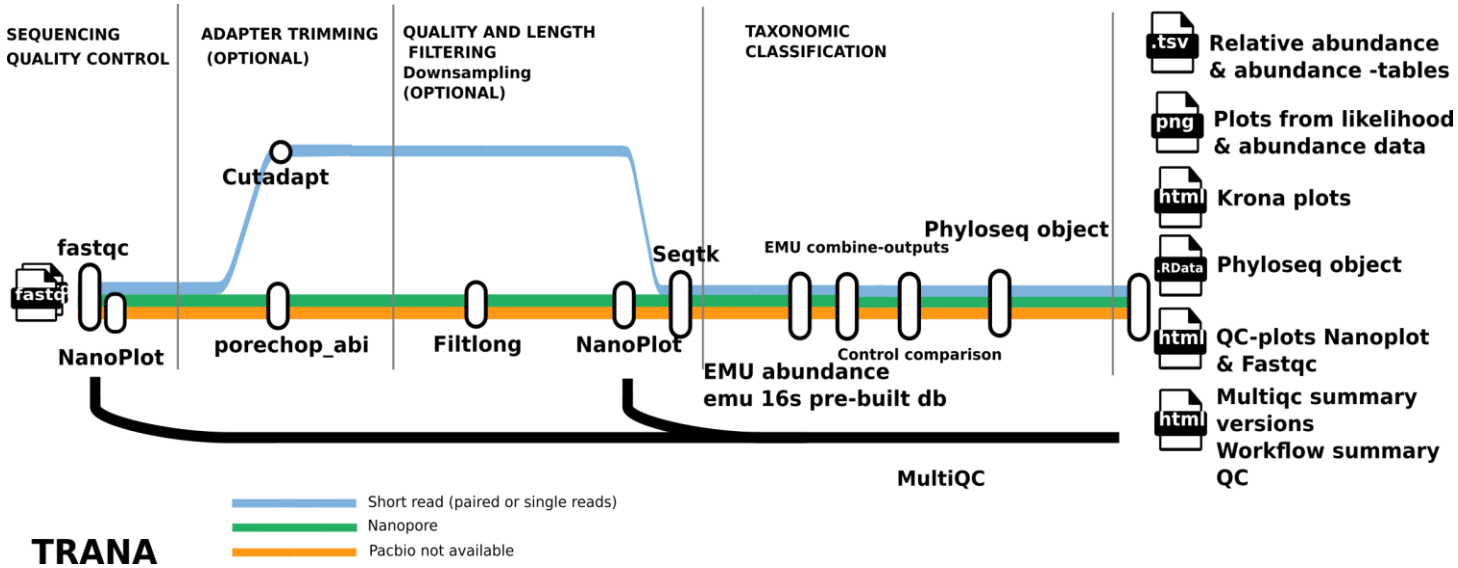
Emu computational requirements

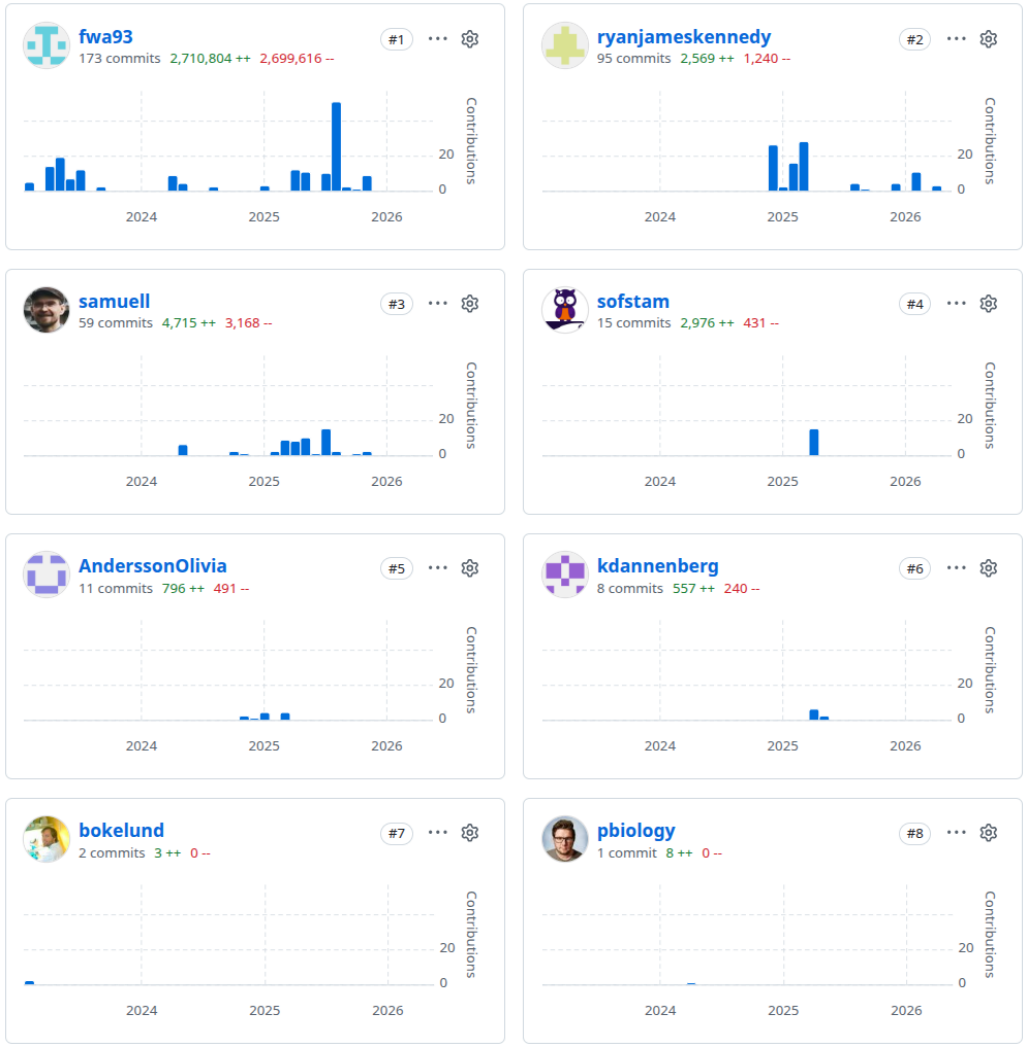
- Ca **0.4 - 0.5 (CPU) core seconds / read**
- For a sample with 30k reads, this is a little more than 4 core hours
- For 10 samples, that is ca 42 core hours
- Spread on 8 cores gives ca 5 hours in total

- Majority of time spent on the alignment step (minimap2), so parameters to minimap can affect runtimes.
- Note: These are very rough estimates, and differs wildly based on CPU model, sample type and minimap2 settings.

The TRANA pipeline

- A Nextflow pipeline based on the Emu software (github.com/treangenlab/emu)
- Initially created by Frans Wallin and subsequently developed collaboratively by Frans, Ryan Kennedy & others within the Genomic Medicine Sweden (GMS) collaboration
- Available open source on GitHub: https://github.com/genomic-medicine-sweden/TRANA

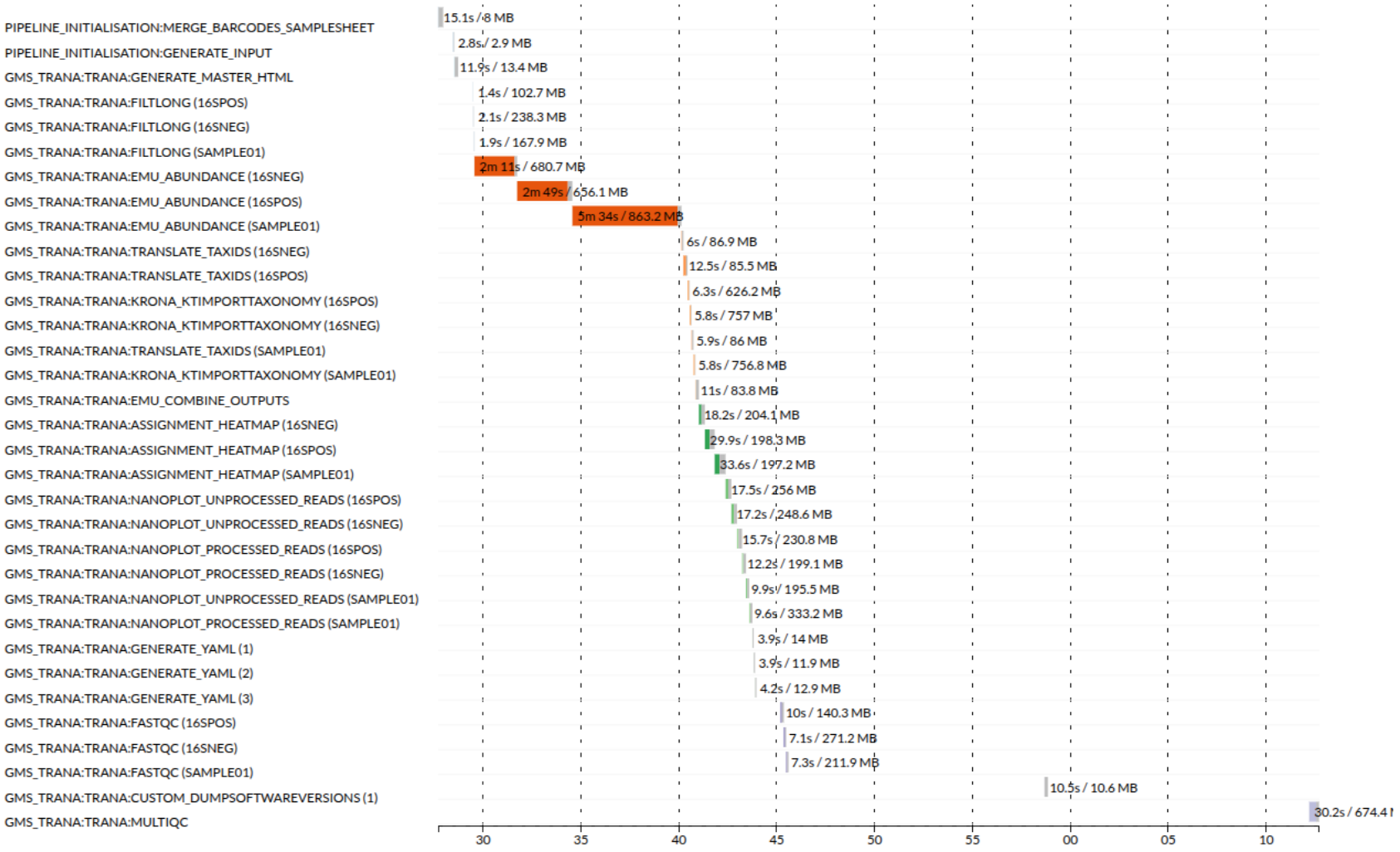




Trana resource requirements

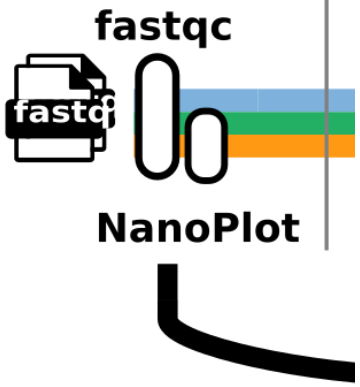
Processes execution timeline

Launch time: 30 May 2026 17:27
Elapsed time: 45m 8s
Legend: job wall time / memory usage (RAM)



Quality control

SEQUENCING QUALITY CONTROL



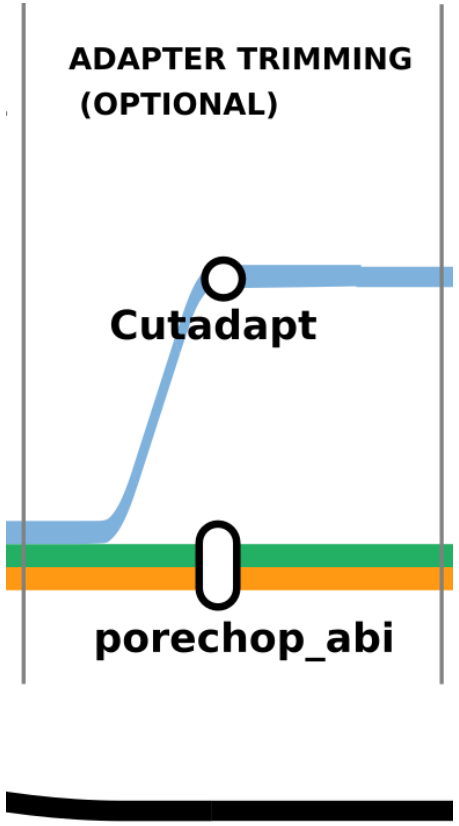
fastqc - Quality metrics of e.g. distribution of length of reads distribution, GC content, quality scores etc.

Nanoplot (unprocessed)

Summary statistics

Plots of e.g. length vs quality

Adapter trimming

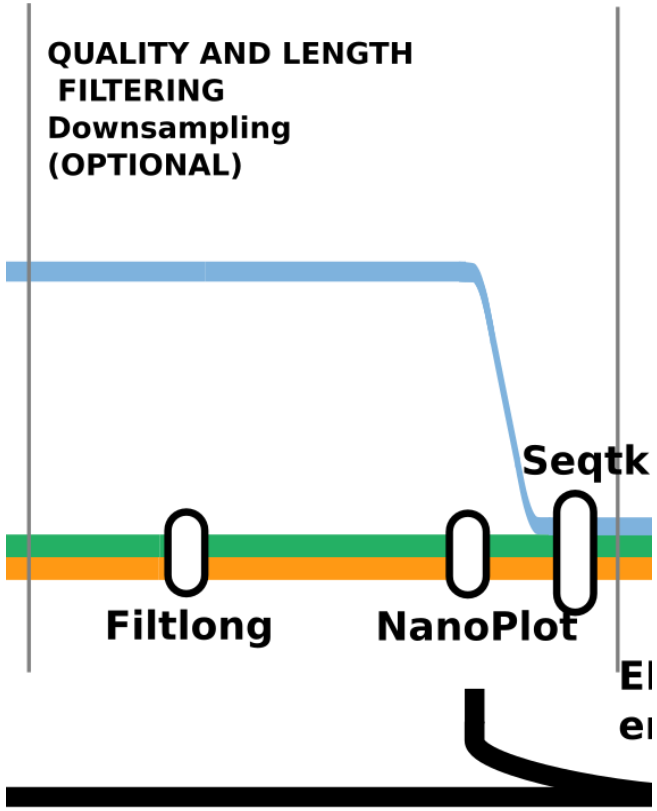


Cutadapt – trimming adapters on short read sequencing data

Porechop_abi – trimming adapters on long read sequencing data

`--adapter_trimming` (default: false)

Quality and length filtering, downsampling



Filtlong - Filtering based on read length and quality

```
--quality_filtering \
--longread_qc_qualityfilter_minlength 1200 \
--longread_qc_qualityfilter_maxlength 1800 \
```

Nanoplot (processed)

Seqtk - Downsampling

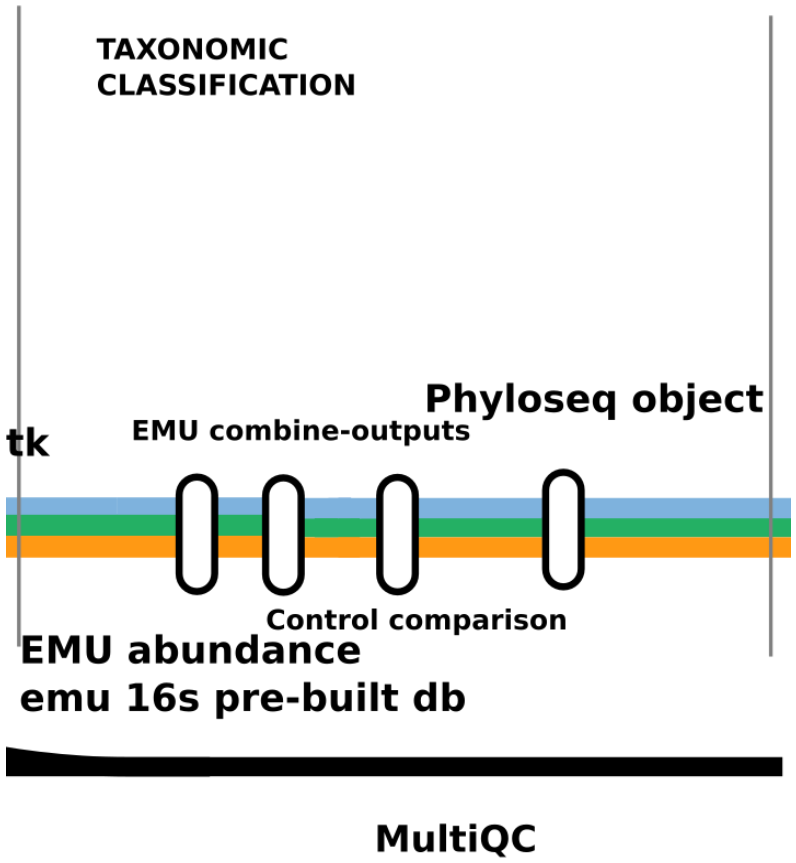
```
--downsample_n_reads 5000
```

slido

What do you think is a good number of reads to downsample to? 🤔

What do you think is a good number of reads to downsample to?

Taxonomic classification



EMU abundance – Runs the abundance estimation

Phyloseq object – Combines abundance output files into R structure

MultiQC – Generates HTML report from multiple analysis output

OUTPUT

Wide range of output files

Metagenomics output:

- Bacterial community profile
- Reported in abundances
- KRONA plots

.tsv **Relative abundance & abundance -tables**

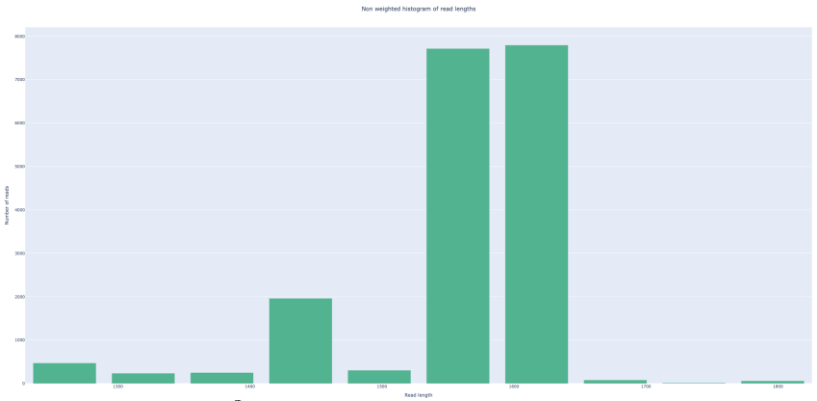
png **Plots from likelihood & abundance data**

html **Krona plots**

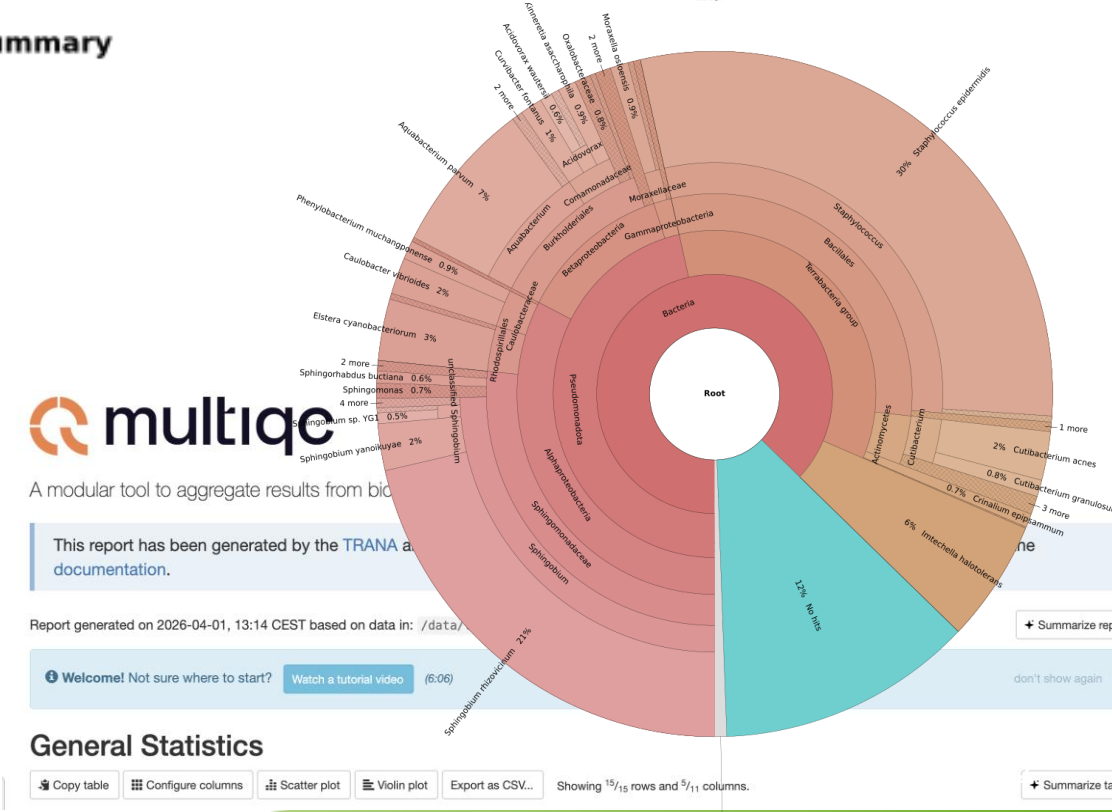
.RData **Phyloseq object**

html **QC-plots Nanoplot & Fastqc**

html **Multiqc summary versions Workflow summary QC**



tax_id	abundance	species	genus	estimated counts
1282	0.3367365785035574	Staphylococcus epidermidis	Staphylococcus	5584.7761544815
432308	0.24390134693179355	Sphingobium rhizovicinum	Sphingobium	4045.103838863796
1886786	0.0016240798215598068	Sphingobium naphthae	Sphingobium	26.935363840569394
86193	0.0020781089663317893	Sphingobium abikonense	Sphingobium	34.46543720661273
70584	0.08268364569480453	Aquabacterium parvum	Aquabacterium	1371.3082638483331
450365	0.0020801841894162137	Aquabacterium fontiphilum	Aquabacterium	34.4998547814679
155892	0.019596020497615364	Caulobacter vibrioides	Caulobacter	324.9999999529508
1090573	0.0007046317305563299	Sphingobium czechense	Sphingobium	11.686317251276732
1834038	0.004220681338558939	Undibacterium amnicola	Undibacterium	70.0
1165090	0.06579848261329183	Imtechella halotolerans	Imtechella	1091.2678341414448

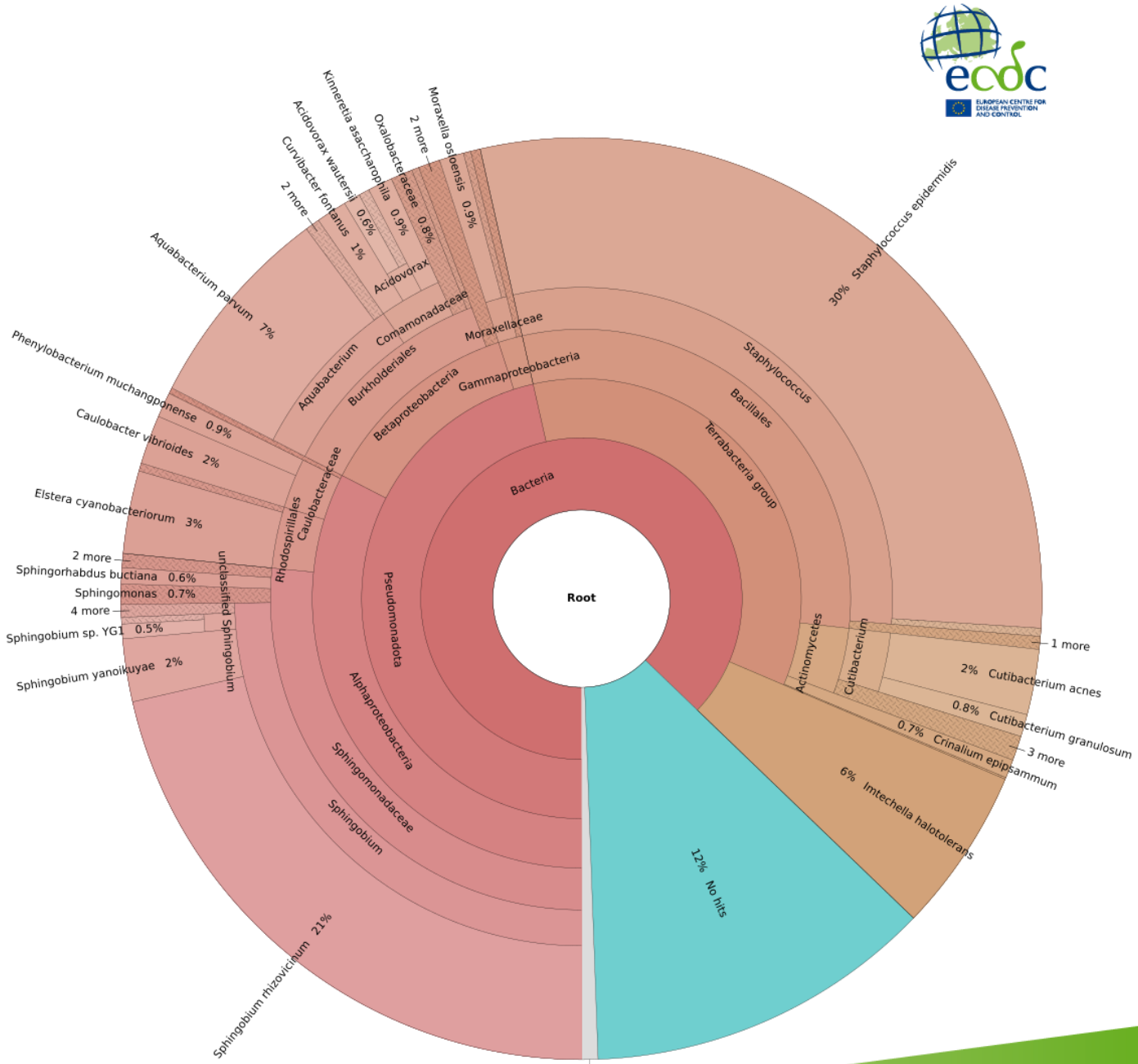


KRONA

Abundances displayed in percentages for species with hierarchical order, allowing to zoom in on taxon groups at any level (order, genus, species etc)

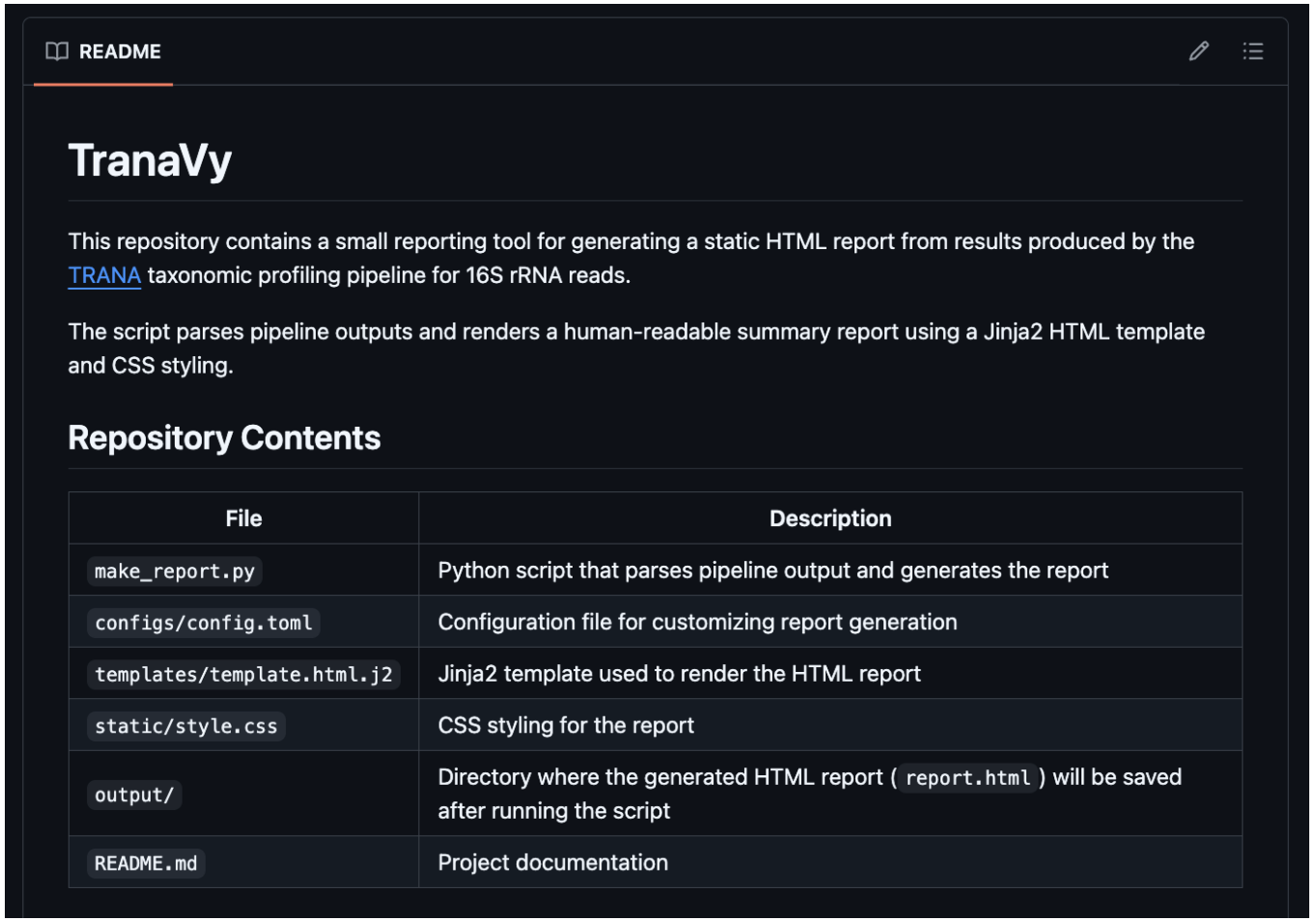
The primary tool used by clinicians to explore and visualize the abundance estimates

github.com/marbl/Krona/wiki



TranaVy

- Small reporting tool for generating static HTML reports
- Developed as an extension to the TRANA pipeline at Clinical microbiology, Karolinska University Hospital
- Available open source on GitHub: github.com/kclinmicro/TranaVy



README

TranaVy

This repository contains a small reporting tool for generating a static HTML report from results produced by the TRANA taxonomic profiling pipeline for 16S rRNA reads.

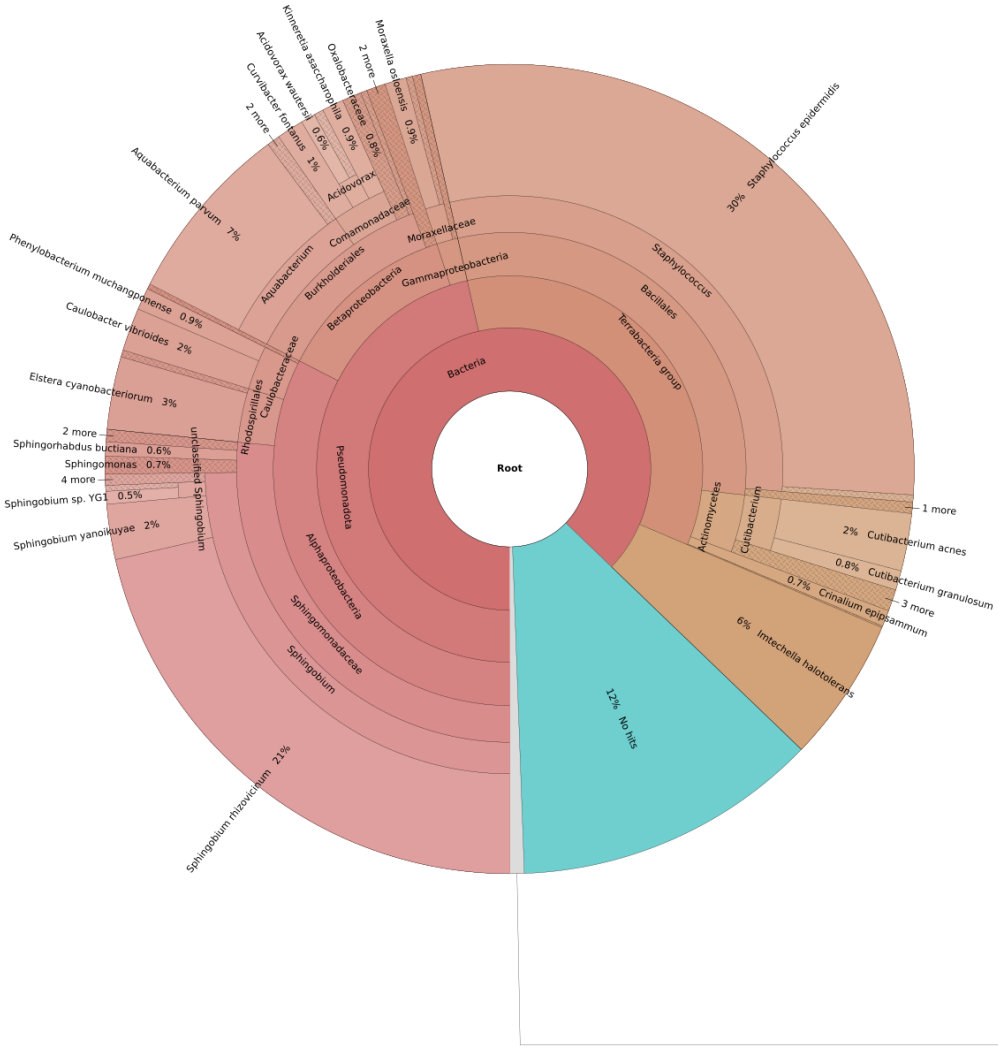
The script parses pipeline outputs and renders a human-readable summary report using a Jinja2 HTML template and CSS styling.

Repository Contents

File	Description
`make_report.py`	Python script that parses pipeline output and generates the report
`configs/config.toml`	Configuration file for customizing report generation
`templates/template.html.j2`	Jinja2 template used to render the HTML report
`static/style.css`	CSS styling for the report
`output/`	Directory where the generated HTML report (`report.html`) will be saved after running the script
`README.md`	Project documentation

TranaVy

- Complementary to KRONA output



16S TRANA Report

Sample Name: SAMPLE_01
Run: RUN_01
Date: 2026-05-05
Pipeline version: TRANA v0.6.0

Summary Statistics

Number of reads before downsampling	18898
Mean Read Length	1525.3 bp
Mean Read Quality	16.5
Median Read Length	1545.0 bp
Median Read Quality	17.2
Read Length N50	1548.0 bp
STDEV Read Length	78.7 bp
Total Bases	28.8 Mbp
Number of mapped reads	18898

Phred Quality Score (Q)

Q10: 10% error rate
Q20: 1% error rate
Q30: 0.1% error rate
Q40: 0.01% error rate

< Q15 Not acceptable	> Q15 Good	> Q20 Excellent

Abundance

Species with low abundance (<0.5%) are shown in light grey.

SAMPLE_01

row	abundance	species	genus	family	tax id	estimated read counts	median aligned identity	median aligned coverage
1	33.67%	Staphylococcus epidermidis	Staphylococcus	Staphylococcaceae	1282	5585	99.30%	96.33%
2	24.39%	Sphingobium rhizovicinum	Sphingobium	Sphingomonadaceae	432308	4045	98.61%	98.07%
3	8.27%	Aquabacterium parvum	Aquabacterium	nan	70584	1371	98.45%	100.00%
4	6.58%	Imtechella halotolerans	Imtechella	Flavobacteriaceae	1165090	1091	98.77%	100.00%
5	3.30%	Elstera cyanobacteriorum	Elstera	Rhodospirillaceae	2022747	548	98.63%	99.57%
6	2.61%	Cutibacterium acnes	Cutibacterium	Propionibacteriaceae	1747	433	98.78%	95.77%
7	2.51%	Sphingobium yanoikuyae	Sphingobium	Sphingomonadaceae	13690	417	98.54%	95.91%
8	1.96%	Caulobacter vibrioides	Caulobacter	Caulobacteraceae	155892	325	98.80%	96.63%
9	1.23%	Curvibacter fontanus	Curvibacter	Comamonadaceae	247486	203	96.85%	98.72%

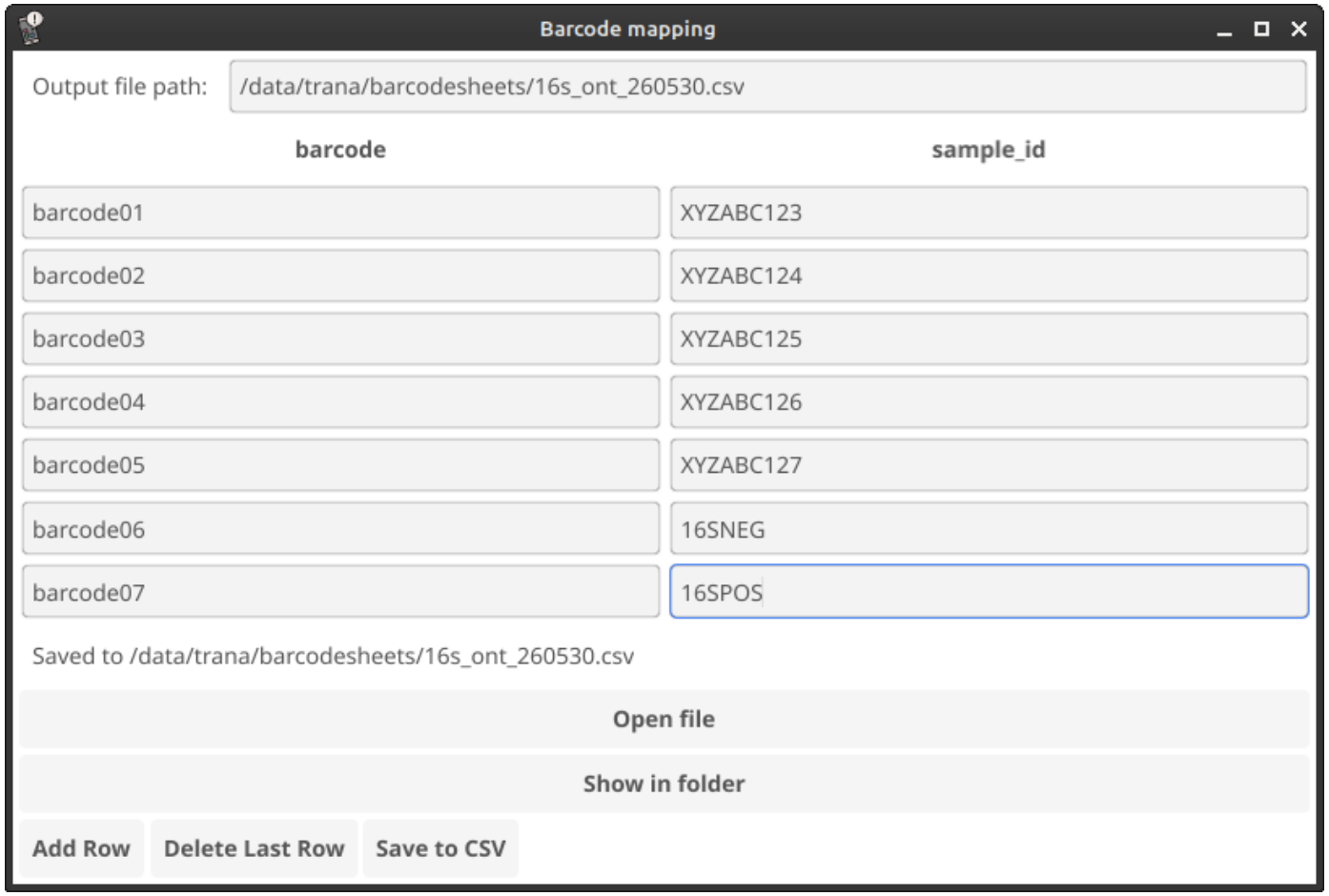
Purple rows indicate spike species
Green rows indicate species not found in negative control or with a normalised abundance ratio ≥ 25.

Bartender

A simple graphical user interface for generating a TRANA compatible barcode-sheet in a valid format, with features to simplify use with hand scanners (auto-jumps to next field etc)

Runs as a native app on the GridIon instrument computer

Open source and available at: github.com/kclinmicro/bartender



In summary

List of learning points in this session:

- EMU enables species-level profiling of Nanopore full-length 16S data
- EMU improves taxonomic resolution compared with classical OTU approaches
- TRANA provides an automated, open-source workflow for Nanopore 16S analysis
- Beyond Emu, it adds steps for Quality control, filtering and reporting and visualisation
- Krona and TranaVy support visualization, interpretation, and reporting of abundance estimates

Recommended material to explore further

- Kristen D. Curry et al.,
  "Emu: Species-Level Microbial Community Profiling of Full-Length 16S RRNA Oxford Nanopore Sequencing Data,"
  Nature Methods, June 30, 2022, 1–9,
  https://doi.org/10.1038/s41592-022-01520-4
- Aja-Macaya, et al. 2025.
  "Nanopore Full Length 16S rRNA Gene Sequencing Increases Species Resolution in Bacterial Biomarker Discovery."
  Scientific Reports 15 (1): 26486. https://doi.org/10.1038/s41598-025-10999-8.

Acknowledgements

The creation of this training material was commissioned by ECDC to Karolinska University Hospital with the direct involvement of Anna Norén, Elin Loo, Lili Andersson-Li, Samuel Lampa, Sofia Stamouli

The revision and update of this training material was commissioned by ECDC to Karolinska University Hospital with the direct involvement of Anna Norén, Elin Loo, Lili Andersson-Li, Samuel Lampa, Sofia Stamouli