A Long Read Assembly pipeline (LoRA) for microbiome data, on NIAID's free cloud service, Nephele

Angelina G Angelova, Poorani Subramanian, Kathryn E. McCauley, Duc Doan, Mariam Quiñones, and Darrel E. Hurt

All authors are affiliated with Bioinformatics and Computational Biosciences Branch, Office of Cyber Infrastructure and Computational Biology,
National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA

ABSTRACT

Background:
In recent years, the increased access to long read sequencing technologies (e.g. minION), has produced an interest within the scientific field, in long read metagenomics specific tools and processing workflows. Our team at NIAID, has met this opportunity by developing a long read, assembly-based metagenomics data processing pipeline (LoRA), which can process both minION and PacBio data, to biologically meaningful community profiles, matrices, and visualizations.

Methods:
The LoRA pipeline is a complete automated Snakemake workflow, which runs through steps for host read removal, long read assembly, microbial taxonomic content classifications, feature predictions, annotations, and abundance scoring, as well as functional inference, community stats and visualizations. Additionally, the pipeline is equipped with user electable databases specific for host decontamination, taxonomic classifications (RefSeq, GTdb, MGBC) or functional inference (KEGG, MetaCyc), as well as features such as resistance gene finding, production and quality assessment of metagenome assembled genome (MAGs).

Results:
In tandem with our recently released long read Nanopore QC pipeline, the LoRA pipeline provides an automated, flexible and reproducible workflow through the fundamental, computationally and time demanding sequence processing steps of metagenomic analysis. LoRA outputs leads researchers directly into the information needed for their customized analyses, targeting their research questions. LoRA is currently available as a stand-alone CONDA version at https://github.com/niaid/LoRA_pipeline. Its public release through NIAID's microbiome analysis cloud platform, Nephele, is anticipated by the end of 2024.
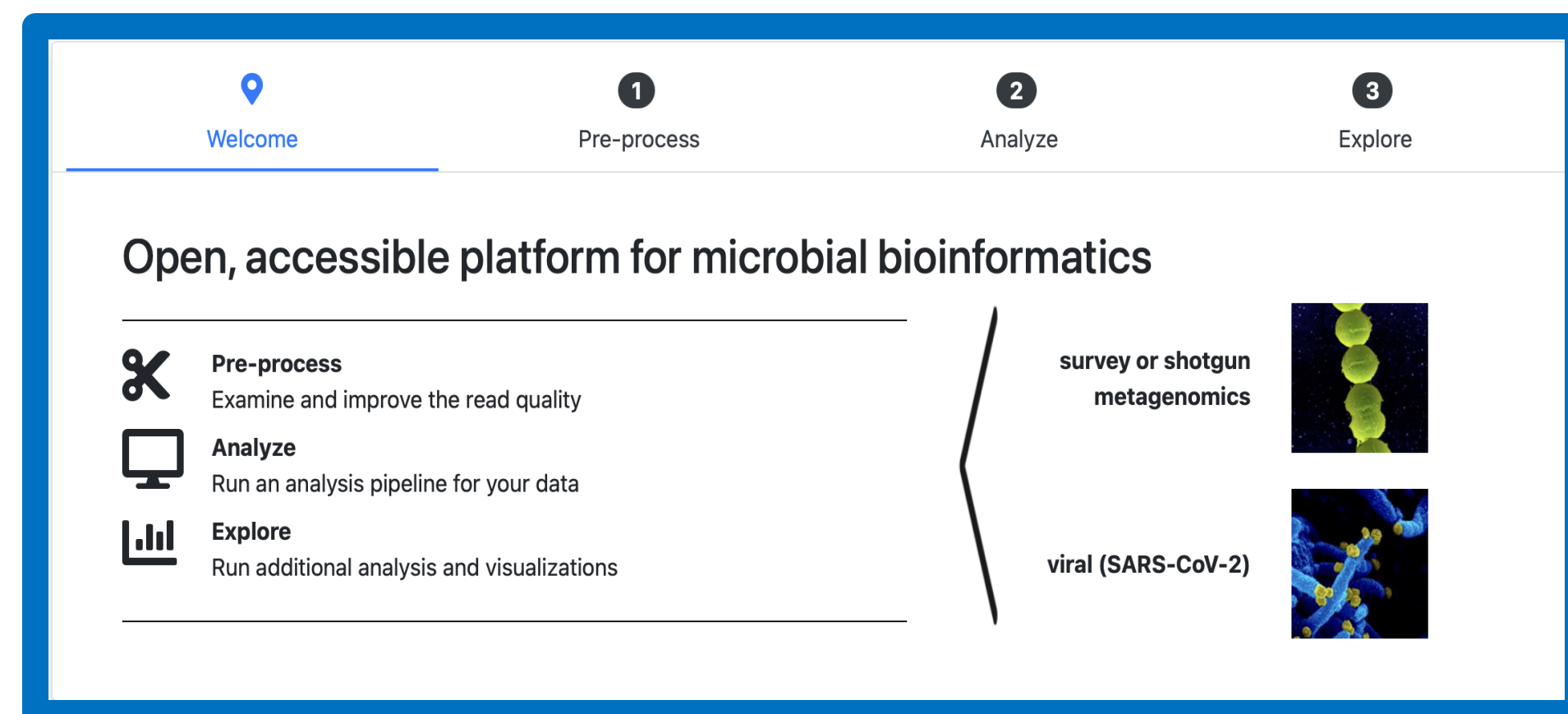
ABOUT NEPHELE

Nephele is NIAID's **free web-based platform** for **automated microbiome data processing**, making microbiome sequence processing more **streamlined and accessible** to researchers worldwide. Our pipelines include workflows for quality check (QC) of short and long sequencing data (e.g. our NanoporeQC pipeline), well **established pipelines** for amplicon and short sequence read processing (e.g. DADA2, BioBakery), as well as **internally developed assembly-based** pipelines inclusive of **computationally demanding steps** (e.g. Whole Genome Sequence Assembly pipeline, WGSA2).

Scan here to check out Nephele!

https://nephele.niaid.nih.gov

Open, accessible platform for microbial bioinformatics

- Pre-process: Examine and improve the read quality
- Analyze: Run an analysis pipeline for your data
- Explore: Run additional analysis and visualizations
- survey or shotgun metagenomics
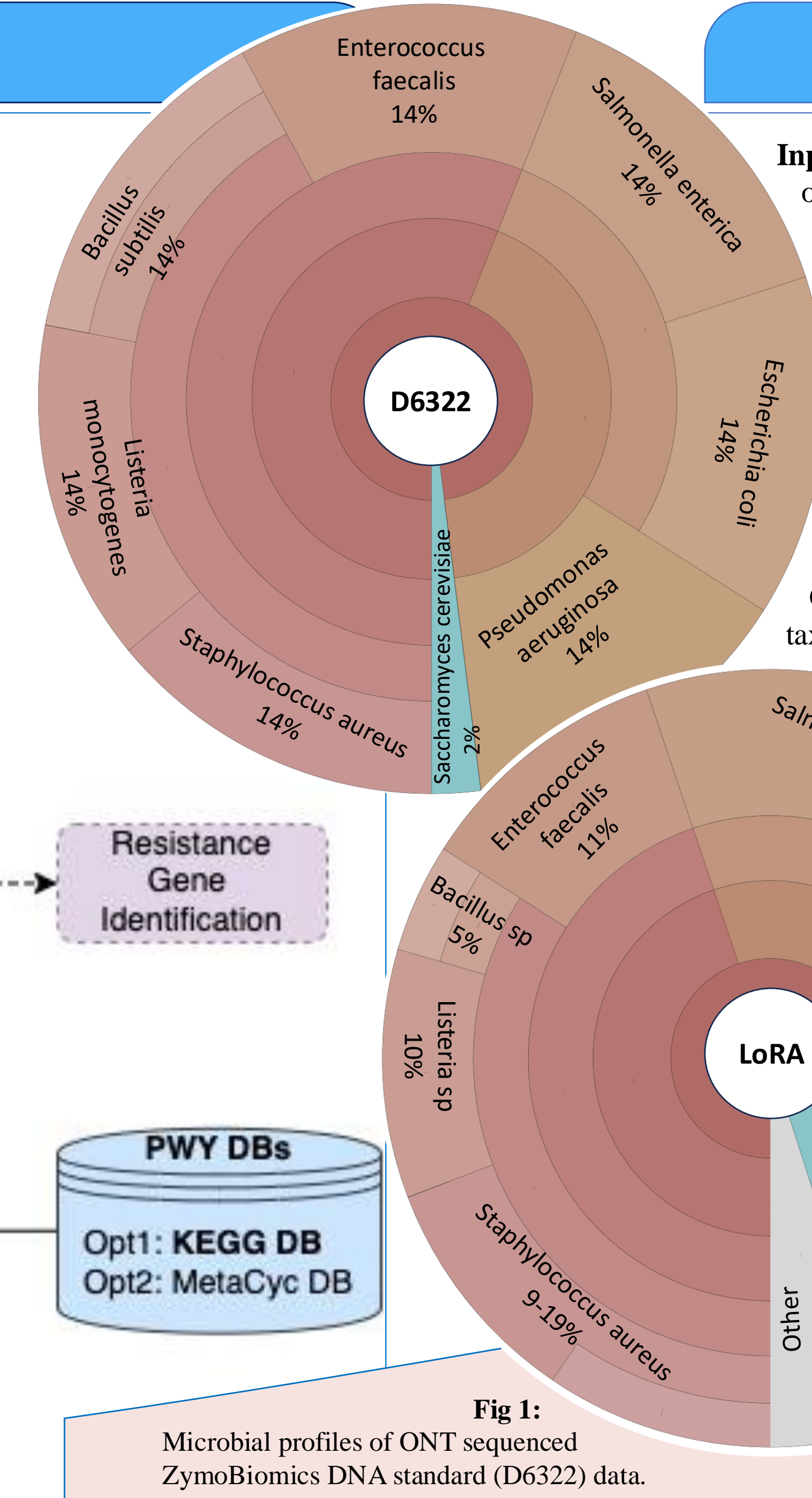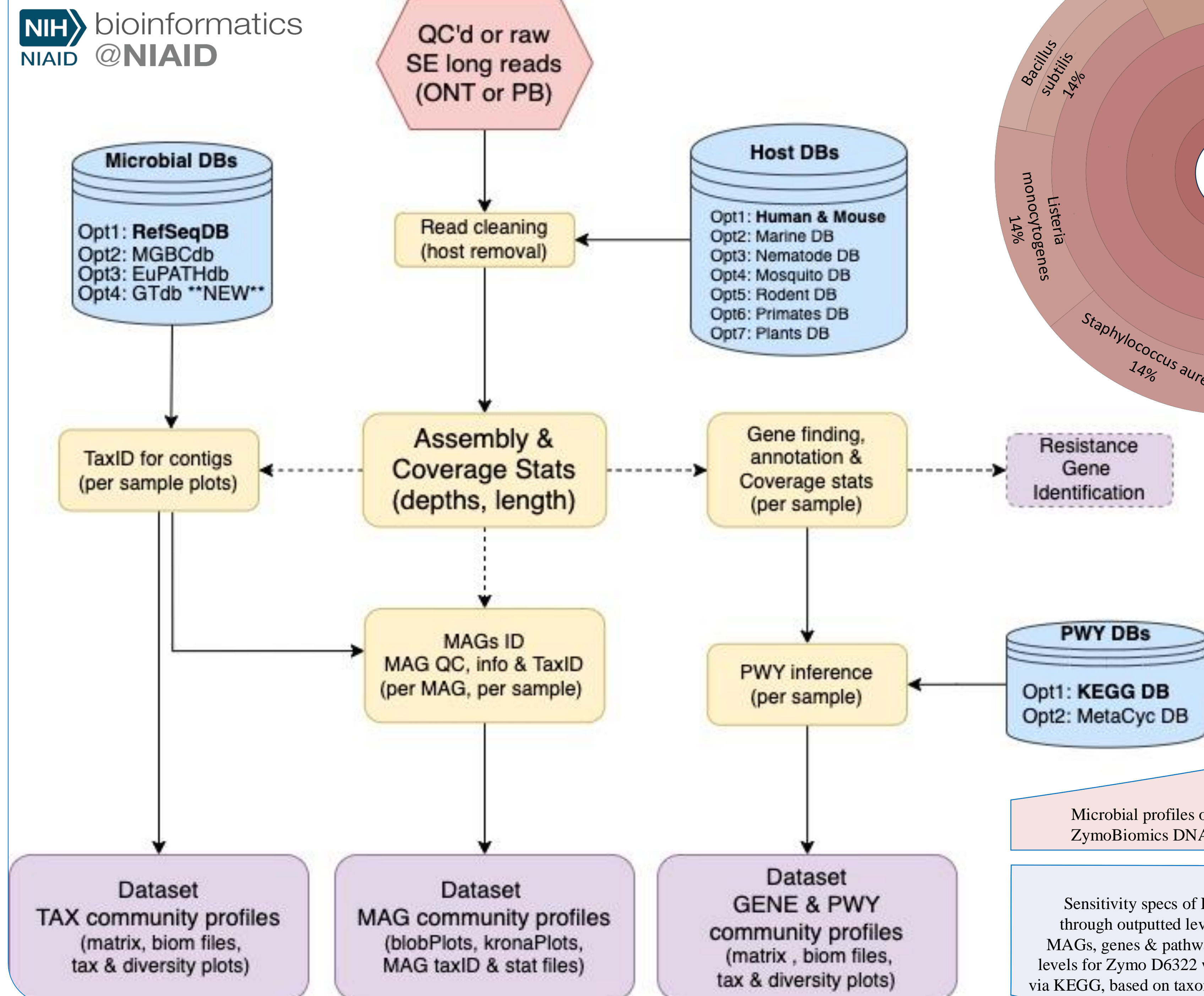- viral (SARS-CoV-2)

Nephele pipelines are **actively supported, free to use and publicly available through our website.** For updates and release news, make a Nephele account and **sign up to our news letter**!

LoRA WORKFLOW

QC'd or raw SE long reads (ONT or PB)

Microbial DBs
- Opt1: RefSeqDB
- Opt2: MGBCdb
- Opt3: EuPATHdb
- Opt4: GTdb **NEW**

Host DBs
- Opt1: Human & Mouse
- Opt2: Marine DB
- Opt3: Nematode DB
- Opt4: Mosquito DB
- Opt5: Rodent DB
- Opt6: Primates DB
- Opt7: Plants DB

Read cleaning (host removal)

TaxID for contigs (per sample plots)

Assembly & Coverage Stats (depths, length)

Gene finding, annotation & Coverage stats (per sample)

Resistance Gene Identification

MAGs ID MAG QC, info & TaxID (per MAG, per sample)

PWY DBs
- Opt1: KEGG DB
- Opt2: MetaCyc DB

PWY inference (per sample)

Dataset TAX community profiles (matrix, biom files, tax & diversity plots)

Dataset MAG community profiles (blobPlots, kronaPlots, MAG taxID & stat files)

Dataset GENE & PWY community profiles (matrix, biom files, tax & diversity plots)

Fig 1: Microbial profiles of ONT sequenced ZymoBiomics DNA standard (D6322) data.

METHODS

Input: The input of LoRA are raw or quality treated long sequence reads obtained from Oxford Nanopore Technologies (ONT) or PacBio (PB) platforms. With submission, the pipeline takes in the user-electable settings custom data processing, along with a dataset-specific metadata file, designating samples into specific comparison groups.

Inner workings: LoRA was written in Snakemake and includes long read compatible tools such as metaFlye, MiniMap2, GTdb-tk, subread, etc, as well as custom python, UNIX and R scripts for estimating abundances, collating sample-based information, into basic statistical information and generating visualizations.

Output: LoRA outputs sequences of assembled contigs, and genes, taxonomic and functional profiles based on elected databases, mapping files (bam), abundance matrices for entire dataset, diversity statistics visuals, and user elected features such as MAGs and AMR information.

BENCHMARKING

LoRA was **benchmarked** against theoretical taxonomic and functional profiles of ZymoBiomics HMW DNA standard sequenced by Simon et al in 2023 with long read ONT platform. Three technical replicates from 3 libraries of varying DNA amounts (3x 1000ng, 3x 500ng, 3x 350ng) were processed through LoRA.

Benchmarking results showed accurate community composition assessment of contigs down to genus level comparable to the theoretical profiles (Fig.1) and gene and metabolic profiles of various sample sizes, comparable to theoretical estimations through KEGG database (Table 1).

Table 1:
Sensitivity specs of LoRA assessed through outputted levels of Species, MAGs, genes & pathway. Theoretical levels for Zymo D6322 were estimated via KEGG, based on taxonomic content.

Samp Size	Species	MAGs	Genes (KEGG)	PWYs (KEGG)
D6233	8	8	~35.1K	~130
>4.0G	5	8	26.3K	~190
~2.3G	6	8	23.8K	~170
~1.7G	6	9	20.9K	~100
~1.3G	4	6	21.9K	~130
<1.0G	4	7	21.5K	~118

LoRA FEATURES

Submit your LoRA job to Nephele:

Job Details
Description of the job: LongRead_dataset

Parameters
Parameter	Value
Host detection confidence level	0.05
Host detection DB	Human or Mouse DB
Sequence data type	ONT
Data quality type	HQ
Assembly polishing steps	2
Tax classification DB	RefSeq
Tax assignment confidence level	0.1
Produce MAGs	☑
CheckM MAG quality plots	☐
MAGs taxonomy with GTdb	☐
BlobTools MAG quality plots	☐
Infer metabolic pathways	☑
Metabolic pathway DB	KEGG DB
AMR prediction	☑

LoRA VISUALIZATIONS

Shannon — Alpha Diversity — TreatmentGroup: A-ONT-HGH, B-ONT-LOW, C-ONT-MED, D-ONT-THEO

PCoA plot based on Bray-Curtis distances — PCo2 [26.7%] vs PCo1 [61.4%]

Pathway	A-ONT-HGH	B-ONT-LOW	C-ONT-MED
Spliceosome; map03040	0	13.9	0.1
Valine, leucine and isoleucine biosynthesis; map00290	9.4	0.3	1.3
Folate biosynthesis; map00790	7.7	1.7	0.6
Phenylalanine, tyrosine and tryptophan biosynthesis; map00400	0.4	8.9	0.5
Basal transcription factors; map03022	0	8.5	0
Pantothenate and CoA biosynthesis; map00770	5.3	0.5	1.2
Regulation of lipolysis in adipocytes; map04923	0	0	6.3
Glycerophospholipid metabolism; map00564	3.8	0.5	1.6
Biosynthesis of unsaturated fatty acids; map01040	0.5	4.4	0.2
Renin secretion; map04924	0.1	0.1	4.8
Nucleotide excision repair; map03420	1	3.5	0.5
Biosynthesis of amino acids; map01230	1.5	2.6	0.7
Longevity regulating pathway – multiple species; map04213	0.6	1.3	2.2
Apelin signaling pathway; map04371	0	0	3.8
Viral life cycle - HIV-1; map03250	0	3.7	0

% Read Abundance (10.0, 1.0, 0.1)

RESULTS

Tools/methodology:
- Utilization of long-read compatible tools (e.g. metaFlye, MiniMap2) and strategies (e.g. coverage-based abundance estimations)
- Decontamination, assembly and taxonomic and functional profiling
- Processing flexibility (databases and features)
- Individual processing of samples
- Dataset-wide summary of results
- Provision of quality statistics and summarizing visualizations

Sensitivity and Benchmarking:
- High level of sensitivity for taxonomic and functional features for datasets
- Stable detection through varying read depths

User experience:
- Fully automated, CLI-free
- Computationally outsourced
- Standardized, reproducible, customizable
- Long read metagenomic data processing

Outputs:
- Abundance matrices for taxonomic, genomic and pathway profiles
- Read mapping files to assembly, assembly quality statistics and information
- Sequence of assembled contigs, predicted and annotated genes per sample, and for dataset
- Optional outputs of MAGs, related QA information and visualizations
- Optional outputs of AMRs found per sample within dataset
- Visualizations of dataset-wide community profiles and diversity metrics

Future work and considerations:
- Improvement of processing efficiency
- Improving taxonomic resolution and abundance estimations
- Maintenance of databases
- Improved benchmarking

CONCLUSION

Overall, LoRA is a shortcut through time and effort consuming, computationally demanding standard bioinformatic steps, commonly experienced in processing of any raw sequence data. User is provided with various fundamental biological information that can be used directly or for subsequent customized processing or analytical steps. Thus, LoRA allows users to dive directly into their specific biological questions.

- eggNOG-mapper2 – Carlos et al 2021. doi: 10.1101/2021.06.03.446934
- Prodigal – Hyatt et al. 2010. doi: 10.1186/1471-2105-11-119
- CheckM – Parks et al. 2015. doi: 10.1101/gr.186072114
- CheckM2 – Chklovski et al. 2023. doi: 10.1038/s41592-023-01940-w
- Zymo ONT data – Simon et al. 2023. doi: 10.1186/s12864-023-09853-w
- GTDB-tk – Chaumeil et al. 2020. doi: 10.1093/bioinformatics/btz848
- Kraken2 – Wood et al. 2019. doi: 10.1186/s13059-019-1891-0
- R language – R Core Team 2021. https://www.r-project.org/
- minimap2 – Li et al 2018. doi: 10.1093/bioinformatics/bty191
- BlobTools – Laetsch & Blaxter 2017. doi: 10.12688/f1000research.12232.1
- fastp – Chen et al. 2018. doi: 10.1093/bioinformatics/bty560
- MtaFlye – Kolmogorov et al 2019. doi: 10.1038/s41587-019-0072-8