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 Darrell 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

National Institute of Allergy and Infectious Diseases

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).

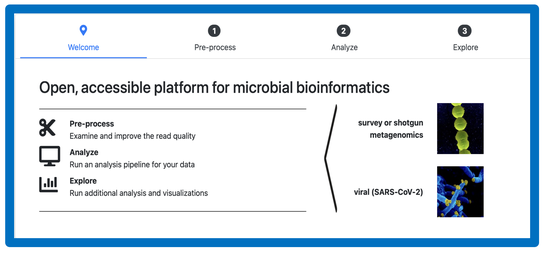
Results:
In tandem with our recently released long read Nanopore QC pipeline, the LoRA pipeline provides an automated, flexible and reproducible workflow through the fundamental, computationally and time demanding sequence processing steps of metagenomic analysis. LoRA outputs leads researchers directly 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

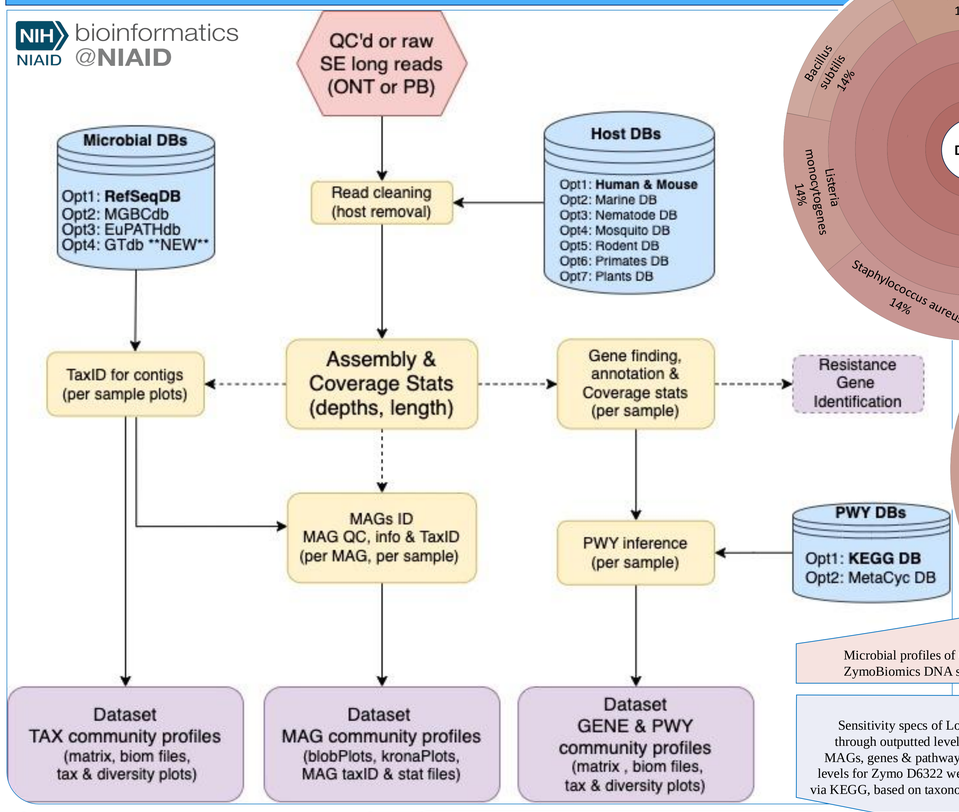


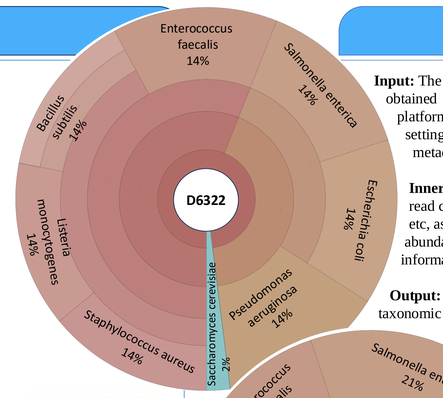
US. Department of Health & Human Services
National Institutes of Health
National Institutes of Allergy & Infectious Diseases

Bioinformatics @NIAID bioinformatics@niaid.nih.gov
Nephele Support: nephelesupport@nih.gov

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





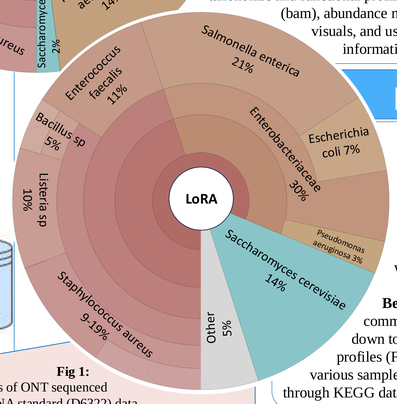


Fig 1: Microbial profiles of ONT sequenced ZymoBiomics DNA standard (D6322) data.

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

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.

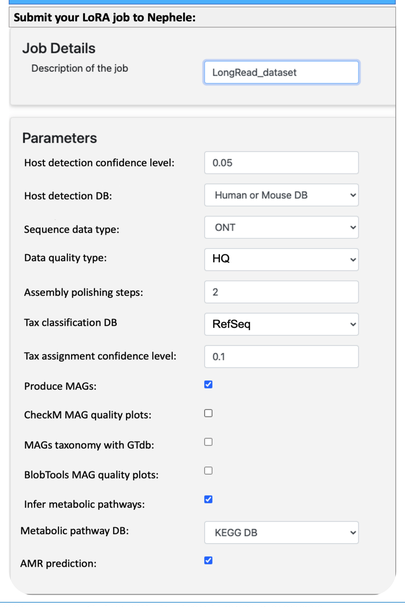
Output: LoRA outputs sequences of assembled contigs, and genes, taxonomic and functional profiles based on elected databases, mapping files (bam), abundance matrices for entire dataset, diversity statistics visuals, and user elected features such as MAGs and AMR information.

BENCHMARKING

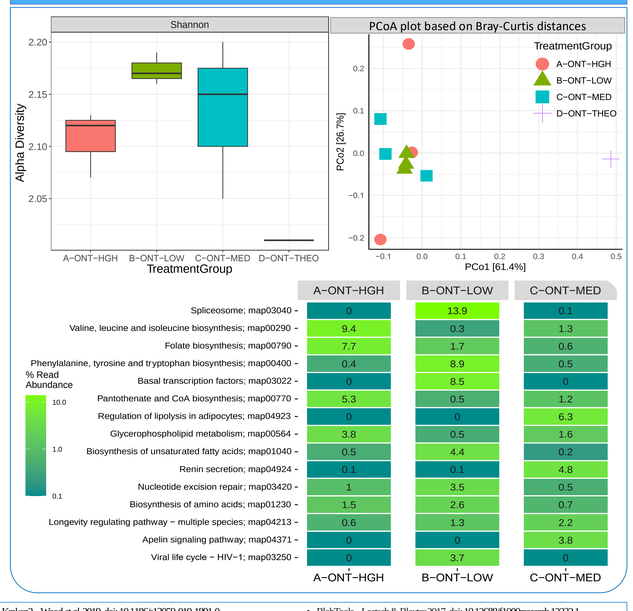
LoRA was **benchmarked** against theoretical taxonomic and functional profiles of ZymoBiomics HMW DNA standard sequenced by Simon *et. al* in 2023 with long read ONT platform. Three technical replicates from 3 libraries of varying DNA amounts (3x 1000ng, 3x 500ng, 3x 350ng) were processed through LoRA.

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).

LoRA FEATURES



LoRA VISUALIZATIONS



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.

Funding: This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services under BCBB Support Services Contract HHSN316201300006W/75N93022F00001 to MSC, Inc - a Guidehouse Company

- 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.186072.114
- 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
- MetaFlye – Kolmogorov *et al* 2019. doi: 10.1038/s41587-019-0072-8