National Institute of Allergy and Infectious Diseases

A New Assembly-based Shotgun Data Analysis Pipeline for Microbiome Exploration in Online Analysis Platform, Nephele

Angelina G Angelova, Duc Doan, Poorani Subramanian, Mariam Quiñones, Lewis Kim, Michael Dolan 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

ABSTRACT

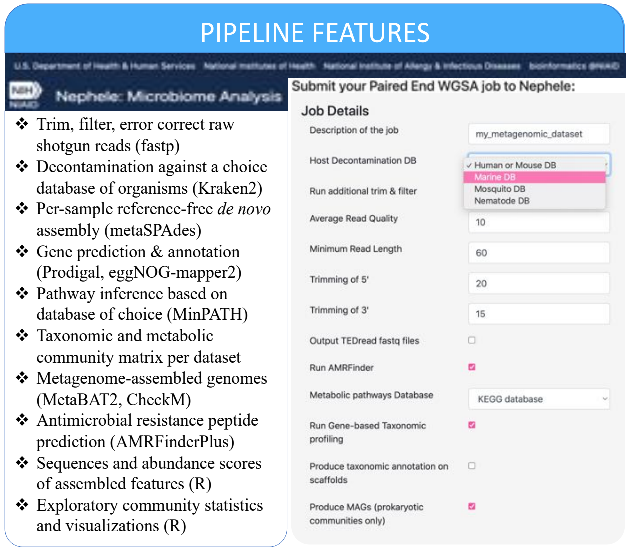
Many researchers lack the computational training or resources to process large metagenomic datasets to the level optimal for extracting biological data. To address this problem, our Nephele2 team at NIAID, developed a new user-friendly command line-free pipeline, Whole metaGenome Sequence Assembly pipeline, version 2 (WGSA2) designed to bridge the computational gap between short shotgun reads and metagenomic assemblies. Nephele's WGSA2 pipeline can process metagenomic datasets from complex communities of various environments including human and animal-associated or environmental microbiomes, inclusive of prokaryotic, micro-eukaryotic and viral organisms. Overall, the WGSA2 pipeline allows the user to easily gain understanding of their metagenomic samples, without investment in time, effort or command line computation.

INTRODUCTION

A plethora of amazing tools are available online (e.g. HUMAnN3.0, MG-RAST, IDseq) that can perform taxonomic and/or functional classification on short shotgun metagenomic reads. However, due to their short nature, shotgun reads often fail to provide relevant biological information. Long-read based metagenomic tools are also abundant online (e.g. GhostKoala, iPath, MetaCyc, NCBI BLAST) and can capture copious amounts of biological information, but require computational power and effort to obtain. To bridge the gap between these two analytical strategies, the National Institute of Allergy and Infectious Diseases (NIAID) has engaged their cloud-based microbiome analysis platform, Nephele2, to provide new user-friendly command line-free Whole Metagenome Sequence Assembly-based pipeline, WGSA2. The pipeline is designed as a researcher's shortcut through the computational requirements of shotgun data processing, read assembly and assembly evaluation, allowing for focusing researcher's time and efforts towards project-specific downstream investigations (gene mining, community & statistical explorations, etc.).

PIPELINE FEATURES

- ❖ Trim, filter, error correct raw shotgun reads (fastp)
- ❖ Decontamination against a choice database of organisms (Kraken2)
- ❖ Per-sample reference-free *de novo* assembly (metaSPAdes)
- ❖ Gene prediction & annotation (Prodigal, eggNOG-mapper2)
- ❖ Pathway inference based on database of choice (MinPATH)
- ❖ Taxonomic and metabolic community matrix per dataset
- ❖ Metagenome-assembled genomes (MetaBAT2, CheckM)
- ❖ Antimicrobial resistance peptide prediction (AMRFinderPlus)
- ❖ Sequences and abundance scores of assembled features (R)
- ❖ Exploratory community statistics and visualizations (R)


U.S. Department of Health & Human Services National Institutes of Health National Institute of Allergy & Infectious Diseases bioinformatics @NIAID
Nephele: Microbiome Analysis

Submit your Paired End WGSA job to Nephele:

Job Details

Description of the job: my_metagenomic_dataset

Host Decontamination DB: ✓ Human or Mouse DB / Marine DB / Mosquito DB / Nematode DB

Run additional trim & filter

Average Read Quality: 10

Minimum Read Length: 60

Trimming of 5': 20

Trimming of 3': 15

Output TEDread fastq files ☐

Run AMRFinder ☑

Metabolic pathways Database: KEGG database

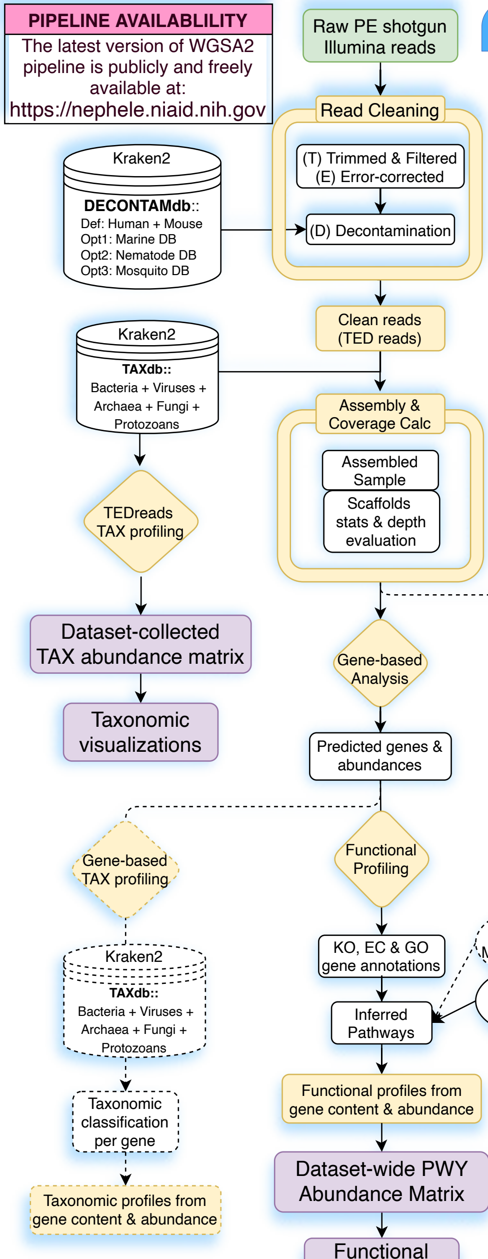
Run Gene-based Taxonomic profiling ☑

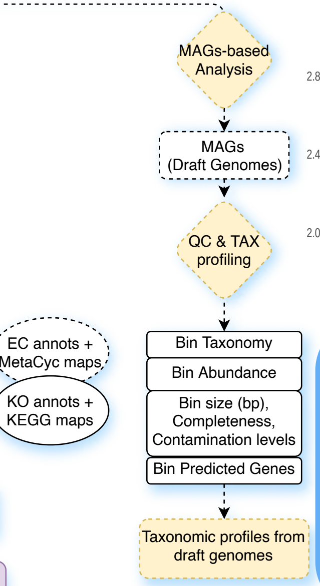
Produce taxonomic annotation on scaffolds ☐

Produce MAGs (prokaryotic communities only) ☑

PIPELINE AVAILABLILITY

The latest version of WGSA2 pipeline is publicly and freely available at:
https://nephele.niaid.nih.gov

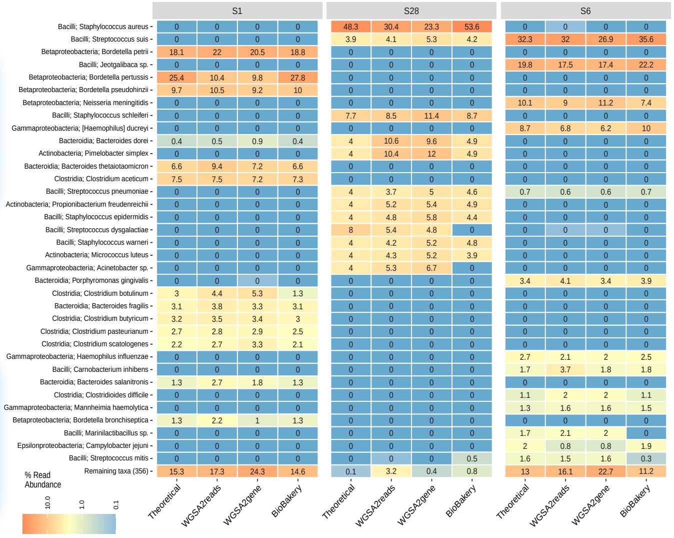

Raw PE shotgun Illumina reads
Read Cleaning
(T) Trimmed & Filtered
(E) Error-corrected
(D) Decontamination
Kraken2
DECONTAMdb::
Def: Human + Mouse
Opt1: Marine DB
Opt2: Nematode DB
Opt3: Mosquito DB
Clean reads (TED reads)
Kraken2
TAXdb::
Bacteria + Viruses + Archaea + Fungi + Protozoans
TEDreads TAX profiling
Dataset-collected TAX abundance matrix
Taxonomic visualizations
Assembly & Coverage Calc
Assembled Sample
Scaffolds stats & depth evaluation
Gene-based Analysis
Predicted genes & abundances
Gene-based TAX profiling
Kraken2
TAXdb::
Bacteria + Viruses + Archaea + Fungi + Protozoans
Taxonomic classification per gene
Taxonomic profiles from gene content & abundance
Functional Profiling
KO, EC & GO gene annotations
EC annots + MetaCyc maps
KO annots + KEGG maps
Inferred Pathways
Functional profiles from gene content & abundance
Dataset-wide PWY Abundance Matrix
Functional visualizations



CONTACTS

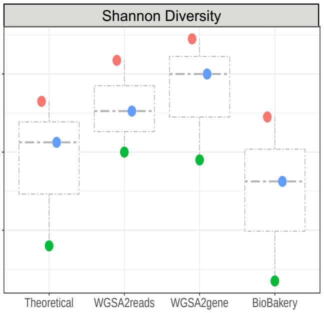
Nephele: nephelesupport@nih.gov
Bioinformatics: bioinformatics@niaid.nih.gov

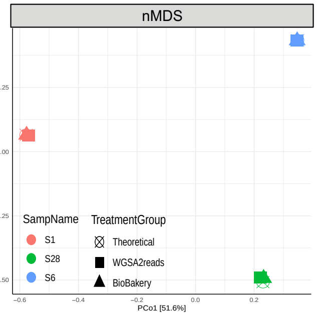
BENCHMARKING



	S1				S28				S6			
	Theoretical	WGSA2reads	WGSA2gene	BioBakery	Theoretical	WGSA2reads	WGSA2gene	BioBakery	Theoretical	WGSA2reads	WGSA2gene	BioBakery
Bacilli; Staphylococcus aureus	0	0	0	0	48.3	30.4	23.3	53.6	0	0	0	0
Bacilli; Streptococcus suis	0	0	0	0	3.9	4.1	5.3	4.2	32.3	32	26.9	35.6
Betaproteobacteria; Bordetella petrii	18.1	22	20.5	18.8	0	0	0	0	0	0	0	0
Bacilli; Jeotgalibaca sp.	0	0	0	0	0	0	0	0	19.8	17.5	17.4	22.2
Betaproteobacteria; Bordetella pertussis	25.4	10.4	9.8	27.8	0	0	0	0	0	0	0	0
Betaproteobacteria; Bordetella pseudohinzii	9.7	10.5	9.2	10	0	0	0	0	0	0	0	0
Betaproteobacteria; Neisseria meningitidis	0	0	0	0	0	0	0	0	10.1	9	11.2	7.4
Bacilli; Staphylococcus schleiferi	0	0	0	0	7.7	8.5	11.4	8.7	0	0	0	0
Gammaproteobacteria; [Haemophilus] ducreyi	0	0	0	0	0	0	0	0	8.7	6.8	6.2	10
Bacteroidia; Bacteroides dorei	0.4	0.5	0.9	0.4	4	10.6	9.6	4.9	0	0	0	0
Actinobacteria; Pimelobacter simplex	0	0	0	0	4	10.4	12	4.9	0	0	0	0
Bacteroidia; Bacteroides thetaiotaomicron	6.6	9.4	7.2	6.6	0	0	0	0	0	0	0	0
Clostridia; Clostridium aceticum	7.5	7.5	7.2	7.3	0	0	0	0	0	0	0	0
Bacilli; Streptococcus pneumoniae	0	0	0	0	4	3.7	5	4.6	0.7	0.6	0.6	0.7
Actinobacteria; Propionibacterium freudenreichii	0	0	0	0	4	5.2	5.4	4.9	0	0	0	0
Bacilli; Staphylococcus epidermidis	0	0	0	0	4	4.8	5.8	4.4	0	0	0	0
Bacilli; Streptococcus dysgalactiae	0	0	0	0	8	5.4	4.8	0	0	0	0	0
Bacilli; Staphylococcus warneri	0	0	0	0	4	4.2	5.2	4.8	0	0	0	0
Actinobacteria; Micrococcus luteus	0	0	0	0	4	4.3	5.2	3.9	0	0	0	0
Gammaproteobacteria; Acinetobacter sp.	0	0	0	0	4	5.3	6.7	0	0	0	0	0
Bacteroidia; Porphyromonas gingivalis	0	0	0	0	0	0	0	0	3.4	4.1	3.4	3.9
Clostridia; Clostridium botulinum	3	4.4	5.3	1.3	0	0	0	0	0	0	0	0
Bacteroidia; Bacteroides fragilis	3.1	3.8	3.3	3.1	0	0	0	0	0	0	0	0
Clostridia; Clostridium butyricum	3.2	3.5	3.4	3	0	0	0	0	0	0	0	0
Clostridia; Clostridium pasteurianum	2.7	2.8	2.9	2.5	0	0	0	0	0	0	0	0
Clostridia; Clostridium scatologenes	2.2	2.7	3.3	2.1	0	0	0	0	0	0	0	0
Gammaproteobacteria; Haemophilus influenzae	0	0	0	0	0	0	0	0	2.7	2.1	2	2.5
Bacilli; Carnobacterium inhibens	0	0	0	0	0	0	0	0	1.7	3.7	1.8	1.8
Bacteroidia; Bacteroides salanitronis	1.3	2.7	1.8	1.3	0	0	0	0	0	0	0	0
Clostridia; Clostridioides difficile	0	0	0	0	0	0	0	0	1.1	2	2	1.1
Gammaproteobacteria; Mannheimia haemolytica	0	0	0	0	0	0	0	0	1.3	1.6	1.6	1.5
Betaproteobacteria; Bordetella bronchiseptica	1.3	2.2	1	1.3	0	0	0	0	0	0	0	0
Bacilli; Marinilactibacillus sp.	0	0	0	0	0	0	0	0	1.7	2.1	2	0
Epsilonproteobacteria; Campylobacter jejuni	0	0	0	0	0	0	0	0	2	0.8	0.8	1.9
Bacilli; Streptococcus mitis	0	0	0	0	0	0	0	0.5	1.6	1.5	1.6	0.3
Remaining taxa (356)	15.3	17.3	24.3	14.6	0.1	3.2	0.4	0.8	13	16.1	22.7	11.2

To assess the accuracy of the WGSA2 taxonomic profiling, the pipeline was run with 3 samples containing mock microbial communities from the 2nd CAMI Toy Human Microbiome Project Dataset (Sczyrba et al. 2017). The profiles were obtained using the Kraken2 classification tool against its standard database, and 2 of the available WGSA2 TAX profiling strategies - from cleaned short reads and from predicted genes (Prodigal), in assembled samples (metaSPAdes). These profiles were compared against each other, their theoretical counterparts and against another pipeline's shotgun-based profiles - BioBakery (McIver et al, 2018). Results indicate of closely related community profiles, composition and structure.





Poster #
EEB1107

NIH bioinformatics

bioinformatics.niaid.nih.gov/metagenomics