Pipeline output details

Global files

- `WGSA2_output_details.pdf` : contains these pipeline output details.
- `logfile.txt` : The main log file log from the pipeline run on this dataset. If the pipeline failed or halted due to error, search for "Error" terms in this log.
- `for_analyze_with_microbiomedb.biom` : TAX community abundance matrix of the dataset, which can be uploaded to MicrobiomeDB.
- `_corrected.txt.no_gz` : a copy of the mapping file submitted with the pipeline. The "corrected" & ".no_gz" labels are not referring to your metadata, but internal processing labels. Feel free to manually remove these labels, for easier access to the file.

Aggregated (Main) results

`TAXprofiles` folder

Contains dataset-wide taxonomic results

Although listed first here, this folder is one of the last produced from the pipeline as it contains the dataset-wide taxonomic results (visualizations, summaries and community matrixes) produced from each processed sample.

Complete contents of this folder will depend on the options elected for the pipeline.

- `readsTAX_<DBname>` folder (default): Contains logs and reports from the taxonomic classification of TED reads. Sub-folders:
    - `bin` folder: files used by Krona for creating the `TAXplots` html chart. These are the text versions of the per-sample taxonomic reports from the TED reads, and they are visualized in the `TAXplots_ReadsTAX_<DBname>.html` report.
    - `bioms` folder: The files from the `bin` folder presented in a json biom format. These can be manually uploaded to MicrobiomeDB or any other analytical platform, if

needed.

- `reports` folder: contains Kraken2 report files with detailed taxonomic classifications produced for each sample
- `TAXplots_readsTAX_<DBname>.html` file: interactive Krona chart of the TED read-based taxonomic profiles, collected in the `bin` folder.
- `merged_tables` folder (**conditional of ≥ 2 samples**): a folder containing the collated taxonomic abundance matrix from the entire dataset. The TAX abundance matrix is reported in 6 tab-delimited text files and 1 biom file. The files are named based on content and format of certain profile information: **TAX-labels** designate files that contain the taxonomy in separate columns per rank, the **Lineage-labels** designate files representing the lineage of each taxonomy (all tax ranks) in 1 column (ranks separated by ';' within the column). **Counts-label** indicates files representing the abundance counts of each taxonomy for each sample (collated across dataset). The **biom file** represents counts for each sample and lineage column. This is the source file for the `for_analyze_with_microbiomedb.biom` in main folder.
    - `DivPlots` folder (**conditional of ≥ 2 samples**): Diversity plots and exploratory statistics for the dataset (**note new location** is within `ReadsTAX` folder as it summarizes the data from the ReadTAX profiles). The plots include alpha diversity indices table and boxplot (`TAX_AlphaDiv.txt` and `TAX_AlphaDiv.pdf`), beta diversity PCoA and nMDS ordination plots (`TAX_BetaDiv_nMDS.pdf` & `TAX_BetaDiv_PCoA.pdf` , require >=3 samples), TAX content heatmap (`TAX_Profile_Heatmap.pdf`), Rank Abundance Curve plot (`TAX_RankAbundanceCurve.pdf`) and Rarefaction Curves (`TAX_RarefactionCurve.pdf`). In some cases (e.g. ordination plots), some exploratory statistics will fail due to the content of the dataset.

- `geneTAX_<DBname>` folder **(based on user-elections)**: Conditional on the choice of classification DB, visual summary of the gene-based taxonomic annotations for each sample, are presented here. Note: **MGBCdb does not produce visual summaries**. This folder contains a `bin` folder with text versions of the profiles from each sample and a `TAXplots_genetax.html` interactive krona plot summary file of the gene-based taxonomic profiles from each sample.

- `MAGs_TAX` folder **(based on user-elective)**: This folder is generated only if MAGs creation and taxonomic profiling is enabled. It contains the TAX classification and abundances of the MAGs. It also contains a `bin` folder with text versions of the MAG-based taxonomic profiles and visualizations by KronaTools (`TAXplots_MAGx.html`). It will also contain a `MAG-based_Counts+TAX.biom` file summarizing the MAG-based taxonomic profiles of all samples.

Note: scaffold-based taxonomic annotations option will not produce visualizations, but annotations are reported within each sample's assembly folder.

`PWYprofiles` folder

Contains dataset-wide functional results

Contains dataset-wide the functional analysis results from the pipeline, including pathway inferred profiles (PWY) and gene-based profiles. The folder will contain one folder, named based on the **metabolic pathway database** chosen at submission: `keggPWYs.MP` (default choice) or `metacycPWYs.MP` (alternative choice). Either folder will have the same structure:

- `PWYplots_[ko2gg or ec2cc].html` : an interactive krona html report with the functional profiles for each sample (KronaTools), based on inferred PWYs
- `pwybin` folder: per-sample text reports of the PWYs identity & abundances files. the files are used to create the html report & biom files
- `bioms` folder: the json biom versions of the pwybin files.
- `genebin` folder **NEW** : per-sample text reports of identity and abundances of each unique gene found in the sample.
- `merged_tables` folder (**conditional of ≥ 2 samples**): Dataset-wide PWY abundance summary abundance matrix tables representing the collated results each sample functional annotation and abundance scoring. The PWY abundance matrix is reported in 6 tab-delimited text files and 1 biom file. The files are named based on content and format of certain profile information: **PWY-labels** designate files that contain the pathway information as separate columns per tiers, the **allTiers-labels** designate files representing the Tiers of each pathway in 1 ';'-delimited column. **Counts-label** indicates files representing the abundance counts of each pathway for each sample (collated across dataset). The **biom file** represents counts for each sample and tiers column. NOTE: As of the latest WGSA2 update, the folder also contains a **NEW** file - a gene-based abundance matrix (`merged_geneTPMtable.txt`) produced from the collated gene abundance information in the `genebin` folder.
- `DivPlotsMP` (**conditional of ≥ 2 samples**): Simple exploratory plots of PWY functional dataset-wide profiles. Visualizations include a community PCoA and nMDS functional composition ordination plots (`PWY_BetaDiv_nMDS.pdf` & `PWY_BetaDiv_PCoA.pdf` ; (>=3 samples)) produced from the inferred PWY matrix in `merged_tables.MP` as well as a pathway abundance heatmap plot (`PWY_Profile_Heatmap.pdf` (>=2 samples)).

Detailed results

Contains per-sample results & analyses

`TEDreads` folder

Output of Steps 1 & 2. These log files track the stats of samples through the trimming, filtering (T), error correction (E) and decontamination (D) Steps (TED Steps). If you chose the **Output TEDread FASTQ files** user option, we recommend downloading & storing the TEDreads to this folder. For each sample you will see the following files:

- `<SampName>_fastpog.[html,txt]:` log files with filtering, trimming, and error-correction stats, in HTML and TXT format
- `<SampName>_kr2_decontam[REPORT,LOG].txt:` files reporting abundance and classification details of the host reads, removed from the dataset.

`asmbMetaSpades` folder

Contain per-sample assembly-related results

Although discussed last in this text, the `asmbMetaSpades` folder is the dataset's most detailed per-sample results. Each sample with sufficient reads for assembly will have its own folder here (`<SampName>_asmb`), the content of which will be consistent across samples, but will depend on the **user options** elected (e.g. Run AMRFinder, Produce MAGs, etc). The aggregated results of the pipeline (see above) will be produced from select files and information from here (for each sample).

Organized into sub-folders based on the chronological order of the workflow steps, the `asmbMetaSpades` folder contain the following:

assembly-related files

within `<SampName>_asmb` folder

- `contigs.FASTA` & `final.assembly.FASTA` : FASTA files of the produced

assembly from sample. The assembly represented in independent contigs (contiguous consensus sequences are stored in `contigs.FASTA`, while the scaffold assembly (the same contigs, oriented, arranged and connected with gaps (Ns in the sequence), based on PE information from the TED reads) are in `final.assembly.FASTA`. It is the scaffold assembly that is used for all subsequent assembly-based analyses.

- `final.assembly_scaffCoverage.txt` file: per-scaffold stats and abundance scores such as average read count, % scaffold covered by reads, number covered bases, GC content, etc. **NEW** the file now contains RPM abundance values (reads per million) for each scaffold (it is this value that is used to assess taxonomic abundance in the elective scaffold-based taxonomic profiles).
- `final.assembly_stats.txt` file: assembly-wide stats such as size of assembly, N50 an L50 stats, number of scaffolds created, etc. File also contains TED read mapping stats and alignment de-replication stats.
- `final.assembly.bam` & `.bam.bai` : assembly alignment and index of alignment files (.bam file is big)
- `spades.log` : log file of the assembly software SPAdes (also captured by the main WGSA logfile.txt)
- `scaffTAX_<DBname>` folder **(user-elective)**: This folder contains the taxonomic annotations of each scaffold (if elected) within the sample, the raw Kraken2 report, the produced human & Krona-readable `<SampName>_4krona` file, and a taxonomy-supplemented scaffold coverage file `final.assembly_scaffCovr+tax.txt`. Note: The database used for taxonomic classification of the scaffolds, is the same as the one elected for main taxonomic classification of WGSA2 (TED read TAX classification).

gene-related files

within `<SampName>_asmb/genes` folder

This folder contains files with information about predicted features (likely genes) within each assembly. These files have the prefix `PREDgenes`:

- `PREDgenes.{faa, fna, gff, gtf}` text files: contain information about the predicted features/genes from the sample, described as amino acid sequences (.faa), nucleotide sequences (.fna), or scaffold names and coordinates {.gff & .gtf} files.
- `PREDgenes_ABUNtab.txt` text file: **provides feature abundance values** with information about feature length, coverage and estimated abundances in RPK (reads per kilobases) and TPM (transcripts per million) score for each feature (instance of gene; iTPM).

- `PREDgenes_stats.txt` : overall stats of predicted features including information about annotation yields, etc.
- `genesTAX_<DBname>` folder **(user-elective)**: Taxonomic association of each predicted feature (gene) within the sample. If elected, the Kraken2 report is provided (`taxREPORT.txt`), along with produced human & Krona-readable `<SampName>_4krona.txt` file (used to produce Krona charts in `TAXprofiles/geneTAX` folder). Additionally, the PREDgenes_ABUNtab.txt file is re-provided amended with gene-based taxonomic affiliation (`PREDgenes_ABUNtab+tax.txt` file).
- `AMRs` foler **(user-elective)**: A folder containing results from AMRFinding algorithm, if elected during submission. Files `AMRs{.faa & _info.txt}` contain the amino-acid sequences and characterizing information (element name, symbol, type, coverage and other stats) for each predicted AMR.

Note: The `<SampName>_asmb/genes/PREDgenes.faa` file can be manually provided to the GhostKoala online annotation & pathway mapping tool for sample-specific metabolic and taxonomic annotation outside of the WGSA2 pipeline.

annotation-related files

within `<SampName>_asmb/genes/annotations`

Folder containing information about successfully annotated features - features recognized as homologues of known genes. Note: from the latest release of WGSA2, there is a lot of **NEW** files here:

- `annots.emapper.{annotations, hits, seed_orthologs}.txt` files: raw outputs of EggNOG-mapper2, from which various annotation types are extracted for analysis. These files also provide detailed (too much) and interweaved information about each annotated gene.
- `annots.{ko, ec, KEGGmap, COG}.txt` files: extracted annotations for each annotation type (KO, EC, KEGGmap and COG annotation respectively) from the e-mapper raw output files.
- `annots.{ko,ec}.{fna,faa}` files: conveniently extracted AA and NT sequences from the annotated genes
- `ANNOTgenes_ABUNDtab.{ko,ec}.txt` files: a subset of the `PREDgenes_ABUNtab.txt` text file from `<SampName>_asmb/genes` , containing only the predicted features with successfully assigned KO or EC annotations (Note: the

provided gene information (length, coverage, RPK and iTPM) is again, per instance of the gene in the sample => multiple features with the same annotation exist).

- `geneTPMtab.{ko,ec}.txt` files: files summarizing **gene abundances for each unique gene** with KO or EC annotation types. Note: It is this file that is collected from each sample in the `PWYprofiles/<DBname>PWY.MP/genebin` folder, and used to produce the gene-based community matrix in the `PWYprofiles/<DBname>PWY.MP/merged_tables`

pathway-related information

within `<SampName>_asmb/genes/pathways`

This folder contains inferred pathways information (from MinPATH) based on the predicted genes (`annots.{ko,ec}.txt`) and calculated abundance scores (iTPM values from `ANNOTgenes_ABUNDtab.{ko,ec}.txt` file) for each sample. Depending on the user-elected `<DBname>` , the file names will have `<prefix>` of `ko2gg` (for ko annotations mapped to KEGGdb) or `ec2mc` (for ec annotations mapped to MetaCycDB).

- `<prefix>.{report,details,log}.txt` files: raw output from MinPath pathway prediction tool. See: How to read MinPath output and the MinPATH publication for more information about how to read files.
- `<prefix>_4krona.txt` file: Information from the `report` file is used to create this file, describing each complete pathway and corresponding calculated abundance score (from iTPM values). The `_4krona.txt` file from each assembled sample is copied in `<PWYprofiles>/<DBname>PWYs.MP/pwybin` and is used to produce functional profile plots and matrixes in the `<PWYprofiles>/<DBname>PWYs.MP/` folder.
- `missing_pathways.txt` : a raw output from MinPATH tool, which would provide information about expected but missing pathways (in most cases, this file will be empty).

MAGs-related information

within `<SampName>_asmb/MAGs` optional folder

This folder contains output from MAGs if the **MAGs** option was elected at submission. It contains binned scaffolds forming draft genomes and their quality assessments. The number of files will vary based on the community composition, organismal abundance and sequence representation.

- mags FASTA files (draft genomes):
    - `mag.{count}.fa` : FASTA file containing scaffolds assumed to be representing the same organism or lineage
    - `mag.{lowDepth, tooShort, unbinned}.fa` : FASTA files containing scaffolds that do not meet the binning tool's (MetaBAT2) criteria for binning (see MetaBAT2 publication). These files are also likely to contain scaffolds affiliated with unicellular eukaryotic organisms (if such are present in the sample).
- `magsQA` folder: Output of CheckM's bin quality assessment tool. In this folder are: `bins/` , `QCplots/` , `storage/` folders as well as `checkm.log` , `lineage.ms` files. More information can be found on CheckM's website. Of note are:
    - `magsQA/bins/` : AA and NT sequences for the predicted genes from that bin, along with gene coordinates, annotations (.gff) and other files
    - `magsQA/QCplots` : plots describing QC stats for each bin more info).

    The WGSA2 pipeline extracts the more descriptive information from CheckM's output to create the following reports:

    - `mags_SUMMARY.txt` : taxonomic, abundance, coverage and other information and stats about each MAG. This file is also used to create the MAGs-based TAX profiles reported in *TAXprofiles/MAGs_TAX* folder.
    - `mags_coverage.txt` : per scaffold coverage and read mapping stats for each MAG
    - `mags_profiles.txt` : coverage & abundance assessment stats for each MAG more info
    - `mags_qa.txt` : quality assessment stats for each MAG
    - `mags_tax.txt` : more detailed taxonomic information about each MAG

Contact us

For feedback, questions, problems or concerns with WGSA2, please contact us at **nephelesupport@nih.gov**, with specific mention of the WGSA2 pipeline in subject. Thank you.