

# WGSA2+ Pipeline output details (version 0.2.7, March 2026)

---

## Global Files

- `WGSA2plus_Output_Details_${Date}.pdf` : This file with useful summary of outputted files
- `logfile.txt` : The pipeline log. If the pipeline fails or receives an error, this is where the information will be logged for tracing it.
- `for_analyze_with_microbiomedb.biom` : BIOM file with the TAX community abundance matrix of the dataset, which can be [uploaded to MicrobiomeDB](#).
- `mapping_file.txt` : a copy of the mapping file submitted with the pipeline.

## Dataset-wide (Main) Results

### `TAXprofiles` folder

#### Contains dataset-wide taxonomic results

Although listed first here, this folder is one of the last produced by the pipeline as it contains the dataset-wide taxonomic results (visualizations, summaries and community matrices) produced from each processed sample. Exact contents of this folder will depend on the tax-classification options selected for the pipeline (e.g. read, gene, scaffold, MAG taxonomy), and the database options (\$DBname). For convenience files and folders that are the main outputs of interest to researchers, will be highlighted in this text.

- `readsTAX_${DBname}` folder (default): Contains logs and reports from the default taxonomic classification of the pipeline, based on the TED reads ([see TEDreads section below](#)). Sub-folders:
  - `bin/` folder: these are the text versions of the per-sample taxonomic reports from the TED reads, used for taxonomic profile visualization presented in the `TAXplots_ReadsTAX_${DBname}.html` report.
  - `bioms/` folder: The files from the `bin` folder presented in a *BIOM V1 (JSON)* format. These can be manually uploaded to MicrobiomeDB or any other analytical platform, if

needed.

- reports/ folder: contains taxonomic classifications reports produced by Kraken2 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 for dataset with  $\geq 2$  samples**): a folder containing the collated taxonomic abundance matrix from the entire dataset.
- The full matrix is represented in 2 files - tab-delimited text file ( merged\_Counts+TAX.txt ) and 2 lineage based files( merged\_Counts+Lineage{.txt,.biom} ). The .biom file is JSON format and is the source file for the for\_analyze\_with\_microbiomedb.biom file in the primary folder). RData and log files are outputted from R for troubleshooting.
- DivPlots/ folder (**conditional for dataset with  $\geq 2$  samples**): Diversity plots and basic exploratory statistics for the dataset, summarizing and providing basic visualization for the data from the merged\_tables/ folder. Provided are alpha diversity tab-delimited text file alongside a boxplot ( TAX\_AlphaDiv.txt and TAX\_AlphaDiv.pdf ), beta diversity PCoA and nMDS ordination plots ( TAX\_BetaDiv\_nMDS.pdf & TAX\_BetaDiv\_PCoA.pdf ), taxonomic profile heatmaps of most abundant taxa ( TAX\_Profile\_Heatmap.pdf ) as well as other basic community measures. Please note that some exploratory statistics and visualizations require  $\geq 3$  samples in dataset to produce.
- geneTAX\_\$DBname/ folder: (**based on user selections**): a visual summary of the gene-based taxonomic annotations for each sample. Each folder would contain a bin/ with text versions of the per-sample profiles and a TAXplots\_<gene>TAX.html interactive krona plot summary file of the assembled-feature-based taxonomic profiles. (Note **Some database do not produce visual summaries for genes**)
- magsTAX\_checkM/ folder (**conditional and optional**): Generated only if "MAGs & MAG TAX annotations" is selected, and is **Metagenomics Mode ONLY**. It contains the TAX classification and abundances of the MAGs produced by CheckM. It also contains a bin/ folder with text versions of the MAG-based taxonomic profiles and visualizations by KronaTools ( TAXplots\_MAGx.html ) as well as a MAG-based\_Counts+TAX.biom biom file summarizing the MAG-based taxonomic profiles of all samples.
- scaffTAX\_\$DBname/ folder: the scaffold-based taxonomic annotations option will not produce dataset-wide visualizations. Annotations are reported only within each sample's assembly folder.

## PWYprofiles folder

## Contains dataset-wide functional results

Contains the dataset-wide functional analysis results from the pipeline, including pathway inferred profiles (PWY) and gene-based profiles. This folder contains `keggPWYs.MP/` folder or `mtccPWYs.MP/` folder (default vs alternative user choice of **metabolic pathway database**). Either folder will have the same structure:

- `PWYplots_$mapping.html` : an interactive krona html report with the functional profiles for each sample (KronaTools), based on inferred PWYs. Gene mapping is either designated `ko2gg` (KO annotations to KEGG pathways) or `ec2mc` (EC annotations to metaCyc pathways) and is part of the file name.
- `pwybin/` folder: per-sample text reports of the PWYs identity & abundances files, used to create the html report & biom files.
- `bioms/` folder: the json biom versions of the `pwybin` files.
- `genebin/` folder: per-sample text reports of identity and abundances of each unique gene found in the sample.
- `merged_tables/` folder (**conditional for datasets with  $\geq 2$  samples**): a folder containing the collated pathway annotations and abundance matrix from the entire dataset. The file content is analogous to the taxonomic one.
  - the full matrix is represented in 2 files - tab-delimited text file ( `merged_Counts+PWY.txt` ) and 2 lineage based files ( `merged_Counts+allTiers{.txt,.biom}` ). RData and log files are outputted from R for troubleshooting.
  - If AMR gene prediction was elected, this folder will also contain a `merged_argSTPMtable.txt` with collated information from all samples about the presence and abundance of AMR genes identified in each sample.
- `DivPlots/` (**conditional on  $\geq 2$  samples**): 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` ; ( $\geq 3$  samples)) produced from the inferred PWY matrix in `merged_tables/` as well as a pathway abundance heatmap plot ( `PWY_Profile_Heatmap.pdf` ( $\geq 2$  samples)).

## External results

---

### `TEDreads_fqs` folder (optional and external)

Contains the trimmed, filtered, error-corrected, host-decontaminated (TED), rna-cleaned fastq files

Due to the size of these TED files, they are omitted from the main output of the pipeline. User needs to elect **Output TEDread FASTQ files** to produce the `TEDreads_fqs/` folder containing those files, available as a separate download. If chosen to download, we recommend storing this folder within their original home, the `TEDreads_$mode/` folder. Note: the normalized reads are *not* contained in the `TEDreads_fqs/` folder, as they were used only for assembly.

### `asmb_files` folder

**Contains the per-sample final assembly files ( `.fasta` and `.bam` bam index and depth files) needed for external processing**

For easy access to main assembly files from each sample, the pipeline moves the `fasta` and `bam` files from each assembly into the `asmb_files/` folder. These can then easily accessed when submitting to external pipelines such as Nephele's DiscoVir or for custom downstream processing. The

`${SampleName}_asmbFinal.{fasta,bam,bam.bai,depths.txt}` files are linked to their original location in the per-sample assembly folders (at

`{asmbMGX/asmbMTX}/${SampleName}_asmb/` [see below for details](#))

## Per-sample (detailed) results

---

### Folders containing per-sample details

#### `TEDreads_$mode` folder

This folder contain log files tracking the stats of each samples through the TED steps (trimming, filtering (T), error correction (E), decontamination from host (D)), RNA removal (optional) and normalization (optional). The folders are tagged `mgx` or `mtx` based on pipeline data processing mode chosen by user (MetaTranscriptomiX vs MetaGenomiX) and thus only 1 of the folders will exist in any pipeline output. Their log content will also vary based on user options chosen. For each sample the following files may appear:

- `${SampleName}_fastplog.[html,txt]` : log file with filtering, trimming and error-correction stats, in HTML and TXT format.
- `${SampleName}_contam[REPORT,LOG].txt` : files reporting abundance and classification details of the host reads removed from the dataset.
- `${SampleName}_rna_filter_LOG.txt` (if **mtx** mode): raw log files about rRNA content and removal steps provided by RiboDetector.

- `$SampleName_normLog.txt` (if **read\_normalization** elected): raw log files provided by bbnorm with read normalization stats.

## **asmbMGX / asmbMTX folder**

### **Contain per-sample assembly results**

The folders are tagged `mgx` or `mtx` based on pipeline data processing mode chosen by user (MetaTranscriptomiX vs MetaGenomiX) and thus only 1 of the folders will exist in any pipeline output. This folder is the dataset's most detailed per-sample output. **Each sample that has successfully assembled will have its own sub-folder here (labeled**

**`$SampleName_asmb` ),** the content of which will be consistent across samples, but may vary per run, depend on the **user options** selected for the run. These folders and results will be the source of the main / dataset-wide pipeline results. The per-sample assembly folder has the following content:

### **Main assembly files**

- `$SampleName_asmbFinal.{bam,bam.bai,fasta}` : sample assembly FASTA file along with the TED read alignment file (.bam; large) and its index file (.bam.bai). These are used for all subsequent assembly-based analyses. **Note: within the `$SampleName_asmb/` folder these files are replaced by soft links pointing to the real files collected in `asmb_files/` (done for easy access and forwarding to external pipelines).**
- `$SampleName_{contigs,scaffolds}.fasta` are final assembly files represented as independent contigs (contiguous consensus sequences) and/or scaffolds (oriented, arranged and N-connected contigs). In **Metagenomics Mode**, the scaffold assembly produces the `$SampleName_asmbFinal.fasta`. In **Metatranscriptomics Mode** the final assembly stems from the contigs file as those present assembled transcripts.
- `$SampleName_{spades/trinity}LOG.txt` log file of the assembly software used (metaSPAdes in **Metagenomics Mode**, or Trinity in **Metatranscriptomics Mode**).

### **assembly stat and metrics files**

- `$SampleName_asmbCvrg.txt` file: per-node stats and abundance scores such as average read count, % scaffold covered by reads, number covered bases, GC content, etc. The file also contains RPM (reads per million) abundance values for each node (it is this value that is used to produce taxonomic abundance **estimations** in the elective scaffold-based taxonomic profiles).
- `$SampleName_asmbStats.txt` file: assembly-wide stats such as size of assembly,

N50 and L50 stats, number of scaffolds/contigs, etc. File also amended with TEDread mapping stats and alignment de-replication stats. As noted above, in **Metatranscriptomics Mode**, these statistics will relate to assembled transcripts (contigs) and not scaffolds.

## assembly taxonomy files

Those are presented in a `$SampleName_asmb/taxonomic/` folder, if any of the assembly-based taxonomic options are **selected by user (Gene-based TAX annotations, Scaffold-based TAX annotations or MAGs & MAG TAX annotations)**

- `scafTAX_$DBname/` folder (**optional**): This folder contains the taxonomic assignments of each scaffold/transcript within the sample, the Kraken2 report of taxonomic annotations, the human- & Krona-readable `$SampleName_4krona.txt` file, and a taxonomy-extended asmbCovrg file `$SampleName_asmbCovrg_wTAX.txt`. **Note:** The **scaffold-based taxonomic profiles are not an accurate biological representation of the community composition**, but an estimate of taxonomic affiliation and abundance of assembled contigs/scaffolds. For biologically accurate taxonomic profiles, see `geneTAX$DBname` or `readTAX$DBname` folders.
- `geneTAX_$DBname/` folder (**optional**): This folder contains taxonomic annotations of each predicted gene (PREDgene) within the sample. The Kraken2 report is provided ( `taxREPORT.txt` ), along Krona-readable `$SampleName_4krona.txt` file. Additionally, the `PREDgenes_ABUNtab.txt` file is provided with per-gene taxonomic affiliation ( `PREDgenes_ABUNtab_wTAX.txt` file).
- `MAGs/` folder (**optional**): This folder contains output from metagenome assembled genomes (MAGs) if the **MAGs and MAG TAX Annotations** option was selected. It contains draft genomes (binned scaffolds) and their quality assessments. The number of files will vary based on the community composition, organismal abundance and sequence representation. MAG analysis is only available in **Metagenomics Mode**. Files include:
  - `mag.{number}.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 criteria of the tool (MetaBat2) (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. Files of note are:
  - `mags_SUMMARY.txt` : a file containing a summary of the taxonomic, abundance, coverage and quality assessment and stats about each MAG extracted from the

CheckM-produced `mags_coverage.txt`, `mags_profiles.txt`, `mags_ga.txt` and `mags_tax.txt` files ([more info here](#)). This file is also used to create the MAGs-based TAX profiles reported in `TAXprofiles/MAGs_TAX` folder.

- `bins/` : folder with AA and NT sequences for the predicted genes from that bin, along with gene coordinates, annotations (.gff) and other files
- `QCplots/` : folder with plots describing QC stats for each bin; [more info here](#).
- others: `checkm.log` and `lineage.ms` files and `/storage/` folder. See [CheckM's website](#)

## assembly functional files

This folder contains files associated with the functional assessment of the assembly, including gene predictions, gene annotations, pathway predictions and AMR identification (if selected by user)

- `genes/` folder: predicted genes are outputted in files with prefix `PREDgenes` and include:
  - `PREDgenes.{faa,fna,gff,gtf}` files: contain information about the predicted features/genes from the sample, described as amino acid sequences (.faa), nucleotide (NT) sequences (.fna), or 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/transcript; iTPM).
  - `PREDgenes_stats.txt` : stats from feature/gene predictions annotation yields and number of genes predicted
- `RGI/` folder (if **find ARGs** option is selected): A folder containing results for Antimicrobial Resistance Genes (AMRs) produced by Resistance Gene Identifier tool. File `RGI_main.txt`, contains the main information about the ARGs based on the [CARD database](#) (name, symbol, drug type, antibiotic resistance etc). `RGI_seq.fna` holds the amino acid sequences for each predicted AMR.
- `annotations/` folder: contains information about successfully annotated genes - features recognized as homologues of known genes.
  - `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 highly-detailed 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 NT and AA sequences from the annotated genes.
- `ANNOTgenes_ABUNDtab.{ko, ec}.txt` : a subset of the `PREDgenes_ABUNTtab.txt` text file from `$SampleName_asmb/genes` , containing only the predicted features with successfully assigned KO or EC annotations (*Note*: the provided gene information (length, cov, RPK and iTPM) is again, per instance of the gene in the sample => multiple features with the same annotation exist).
- `geneTPMtab.{ko, ec}.txt` : files summarizing **gene abundances for each unique gene** with KO or EC annotation types. *Note*: It is this file that is collected from each sample in the `PWYprofiles/$DBnamePWY.MP/genebin` folder, and used to produce the gene-based community matrix in the `PWYprofiles/$DBnamePWY.MP/merged_tables` folder.
- `pathways/` folder: contains inferred pathways information (from MinPATH) based on the predicted genes ( `annots.{ko, ec}.txt` ) and calculated abundance scores (iTPM values from `ANNOTgenes_ABUNDtab.{ko, ec}.txt` file) for each sample. Depending on the `$DBname` chosen, the file names will have a `<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/$DBnamePWYs.MP/pwybin` and is used to produce functional profile plots and matrices in the `PWYprofiles/$DBnamePWYs.MP/` folder.

## Contact Us

---

For feedback, questions, problems or concerns with WGS2+, please contact us at [nephelesupport@nih.gov](mailto:nephelesupport@nih.gov), with specific mention of the WGS2+ pipeline in the subject line. Thank you!

