MICROBA

Microba Research Data Deliverable

Introduction

This document describes the bioinformatics methods and data files for your project data. We thank you for choosing Microba, and trust that you gain valuable insights from these results. We welcome feedback on these outputs, so that we can improve the quality of this service to meet your needs.

The Microba Analysis Platform

MAP consists of several key elements, these are the Microba Genome Database (MGDB), the Microba Genes Database (MGENES), the Microba Community Profiler (MCP), and the Microba Genes and Pathway Profiler (MGPP). Each of these key proprietary bioinformatic systems will be described briefly.

The Microba Genome Database (MGDB)

MGDB is Microba's high-quality and expansive genome database, collated and curated from public repositories such as GenBank, and supplemented with faecal derived species that Microba has mined from public and private metagenomic data. A core concept of MGDB is the species cluster. Many hundreds of thousands of candidate genomes are dereplicated based on average nucleotide alignment (ANI) distance between other candidate genomes and selected for inclusion in the database based on genome quality and annotation information, such as type-strain information. Microba uses the Genome Taxonomy Database (GTDB) as the primary source of taxonomic annotation of the resulting representative genomes (http://gtdb.ecogenomic.org/). When a species cluster does not have a taxonomic description, Microba assigns a species name directly. Such names are prefixed with *MIC* and followed with an incremented number, for example *s_Bifidobacterium MIC4613*. The current version of MGDB (v2.0.0), contains 28,246 species clusters consisting of 73,646 representative genomes. Of these, 3,540 species clusters have no GTDB species taxonomy annotation, and are assigned a MIC ID. It is important to note that taxonomy is an ever-changing system, and as the GTDB taxonomy updates, so too will the taxonomy that Microba provides. This may result in changes to these species names for future MGDB releases.

The Microba Community Profiler (MCP)

MCP quantifies the relative abundance of species within a metagenomic sample using the MGDB as a genome reference. Profiling is performed after all data are quality controlled. The results of the MCP are both the relative abundances of species, as well as read counts assigned to each species. Reads that cannot be identified are listed as "Unmapped" (the read had no high-quality alignment in the database), or "Unassigned" (the read had a high-quality alignment, however there was insufficient evidence to determine that the species was present). Reads may remain unassigned because the species they originated from was present at very low abundance, or because the region the read mapped to is highly conserved and is not specific to a single species. The MCP is parameterised conservatively and has been validated to report less than 0.1% false positive species.

The Microba Genes Database (MGENES)

Microba maintains a database of Open Reading Frames (ORFs) from all genomes in MGDB. From the 73,646 genomes in MGDB, over 200 million ORFs have been identified. In order to functionally annotate these ORFs, they are clustered into clusters of >90% sequence similarity, over at least 80% of the length of the ORF. This is initially done by searching the genes against all representative proteins in the database UniRef90 (https://www.uniprot.org/help/uniref). In the case that a gene cluster contains a sequence annotated in UniRef90, that cluster is annotated with available UniRef90 database information. The remaining ORFS were *denovo* clustered into 90% sequencing similarity clusters, and annotated separately, where possible. These two methods resulted in 73.9M unique gene clusters. Each cluster is linked to either a UniRef90 identifier, or an internal Microba identifier, and are identified as the M* IDs (for example: MIG2010879907). Using UniRef90 annotations allows Microba to annotate these gene clusters with the gene functions assigned to Uniprot proteins that share sequence similarity. In this way, the gene abundances provided in your data package have assigned functional annotations. Annotations currently available are listed in **Table 1**. Any annotation available in the UniProt ID Mapping service can be annotated onto Microba Gene IDs via UniRef90 annotations if needed (please see here: https://www.uniprot.org/uploadlists/).

Table 1: External databases for which annotation information is provided when available in the MGENES (Microba Genes Database).

Database	Abbreviation	Further information
UniRef90	UniRef90	https://www.uniprot.org/help/uniref
Enzyme Commission	EC	https://enzyme.expasy.org/
Transporter Classification Database	TCDB	http://www.tcdb.org/
Gene Name	Gene_Name	https://www.uniprot.org/help/gene_name
Search Tool for the Retrieval of Interacting Genes	STRING	https://string-db.org
Evolutionary genealogy of genes: Non-supervised Orthologous Groups	eggNOG	http://eggnogdb.embl.de

Microba Gene and Pathway Profiler (MGPP)

Gene clusters are quantified by counting the reads that map to each gene within the cluster. Because the Microba Analysis Platform is genome-centric, it is possible for Microba to provide not only the counts of genes present across an entire sample, but also the counts for each gene *per species* that is identified in each sample. This powerful feature allows very high-resolution analysis of the gene data. Furthermore, gene annotations are performed on the full length ORF, and gene abundances calculated based on high-specificity genomic (nucleotide-space) alignments. This ensures very high specificity in read functional annotation, much higher than can be attained by a translated blast search such as blastx.

Gene count matrices are provided per sample and per species for all Microba gene clusters (MGENES), Enzyme Commission entries (EC) and also Transporter Classification Database (TCDB) entries.

Additionally, an annotation file is provided which links, when possible, the Microba gene clusters to all databases listed in **Table 1** above.

The Microba gene functional data is also used to quantify higher level metabolic pathways. These pathway annotations are provided based on the MetaCyc database (https://metacyc.org/). The MetaCyc pathway abundance pipeline summarises an input Enzyme Commission (EC) matrix into pathways (modular metabolic pathways such as glycolysis), and metabolic groups (groups of pathways that share a role in metabolism e.g. Carbohydrate metabolism).

Quantification of pathways in the metagenomic samples is a two-step process. First, each genome was annotated with the fraction of all genes present in all pathways in MetaCyc (based on EC annotations). Pathways that were complete or near complete (completeness > 80%), were stored as a reference database for further analysis. Second, for each pathway present within a species, the abundance of all EC IDs associated with that pathway are averaged. If a pathway is present within the species identified in that sample, but no reads are detected mapping to that pathway, a value of zero is included in the calculation. If a pathway is not present within a species, the value in the output matrix is an NA. In this way, it is possible to distinguish between pathways that are likely present, but at very low abundance, to those that are not present in the sample. MetaCyc provides annotations grouping pathways into higher level classifications, called Groups. MetaCyc Group abundances are quantified as the summed abundance of the pathways within each group (available in the file MetaCyc_group.samples.tsv). Pathway and metabolic group data files are available sample wide, and per species in each sample.

Listing of output files

The data are provided as tab separated value files (.tsv). These are plain text files. Large files are compressed into .gz files (.tsv.gz). Uncompressed files of modest size can be opened in excel by importing them through File->Import. Larger files files must be uncompressed to be read (for example, with WinZip on Windows, or gunzip on Linux or Mac), or loaded directly into an analysis script such as with R or python.

Directory	Description	Files
./annotations	Annotation files for databases used in the systems.	TCDB.families.tsv enzyme.dat.gz metacyc_heirarchy.tsv metacyc_pathways.tsv microba_annotations.tsv.gz versions
./taxonomic_profiles	Species profiles for all samples (relative abundance)	profile_summary_stats.tsv profiles.tsv
./taxonomic_profiles_counts	Species profiles for all samples (counts)	profile_summary_stats.tsv profiles.tsv
./functional_profiles/by_sample	Gene and pathway profiles aggregated at the sample level.	EC.samples.tsv.gz MICROBA.samples.tsv.gz MetaCyc_group.samples.tsv.gz

		MetaCyc_pathways.samples.tsv.gz TCDB.samples.tsv.gz
./functional_profiles/by_species	Gene and pathway profiles, aggregated for each sample, at the species level	SAMPLEID.EC.species.tsv.gz SAMPLEID.MICROBA.species.tsv.gz SAMPLEID.MetaCyc_group.tsv.gz SAMPLEID.MetaCyc_pathway.tsv.gz SAMPLEID.TCDB.species.tsv.gz