Last Updated on 10/1/2025 by RAA

Taxonomic and Functional Profiling of Shotgun Metagenomes

We use KneadData 0.12.1 [1], MetaPhlAn4.2.2 [2], and HUMAnN3.9 [3] for taxonomic and functional profiling of metagenomes via shotgun metagenomic data.

The BioBakery suite of the Huttenhower Lab at Harvard is the current gold standard set of pipelines for open-source, not-for-profit processing and alignment of shotgun metagenomics data. Other alternatives exist that use modified versions of their profiling tool, but may not have the open-source or shared database that underlies Metaphlan4.2.2 and are thus not recommended as they are for-profit and locked behind hiring certain companies.

Initially, we use KneadData 0.12.1 to perform QC: trimmomatic to trim reads for adapter content, via a sliding window for read quality, and read length; then host reads (typically human) are removed from the .fastq.gz files via alignment and Samtools, and QC reports are generated via a combination of FastQC and proprietary scripts. We will share the amount of host reads removed compared to the overall reads to show a proportion of bacterial or other target reads retained.

Taxonomic profiling and assignment will be conducted as deeply as the data and underlying database allow for, with the possibility of species level hits but a noted focus on genus level in the literature, using Metaphlan4.2.2. Some taxa have species-level resolution, but far and away, the most common level of reporting is still at the genus stage. As of this writing, Metaphlan4.2.2 and its underlying database also curated by the authors, ChocoPhlan, boasts ~5.1M unique clade-specific marker genes taken from ~1M microbial genomes (consisting of approximately 236,600 references and 771,500 MAGs or metagenomic assembled genomes inferred from data output, likely unculturables) that span 26,970 species-level genome bins or SGBs [2]. For more information on the database and how the SGB approach is used within metaphlan4.2.2, please see the reference paper. Upon each update to the Metaphlan package, the Chocophlan database is also updated, so the two are always in sync. A taxonomic output table is generated based on your data, typically in relative abundance due to the nature of metagenomic profiling making it uncertain to get an exact number of reads per 'hit' [2], and as of Metaphlan4.2.2, a new viral profiling option has been introduced. Within the metaphlan4.2.2 package, a sub-pipeline is run for the assignment of viral sequence clusters (VSCs) using GeNomad via the --profile_vsc flag. Note that the focus on the BioBakery pipeline and the underlying database(s) is heavily on bacterial metagenomics, but that eukaryotic, archaeal, and viral hits will be detected but are far less abundant in the database(s) used due to the ever-evolving nature of the shotgun metagenomics experiment itself and the amount of data generated, curated, and added to the databases. The analysis will always be only as good as the databases are, just as in 16S microbiome experiments.

HUMAnN3.9 (HMP Unified Metabolic Analysis Network) is a sub-pipeline that utilizes the MetaCyc database, as well as the UniRef gene family catalog for nucleotide alignment, and DIAMOND for translated protein alignments to strengthen the quality of the results to characterize the microbial pathways present in samples. HUMAnN3.9 generates three outputs: 1) gene families based on UniRef proteins and their abundances, 2) MetaCyc pathways and their coverage, and 3) MetaCyc pathways and their relative abundances.

Of note is that the HUMAnN3.9 sub-pipeline takes a considerable amount of time per sample, so if you do not need information about pathways or gene families, it is recommended to skip this stage. Metaphlan4.2.2 is relatively quick, while KneadData takes a moderate amount of time but both of these are mandatory. The amount of computational power and time needed scales heavily with the number of samples, so experiments with *n* in the 100s or 1000s will take far longer and require much more computational power than those with less *n*. This is on a scale of increasing from days to weeks/months even with modern HPC cluster usage due to the extreme amount of I/O and database comparisons that need to be made.

Required:
1. Raw data files (fastq.gz)
2. Metadata spreadsheet connecting your samples to groups of interest

Deliverables of shotgun metagenomics data analysis processing:

1. Per-sample taxonomic composition as derived from Metaphlan4.2.2
    a. Relative abundance table for all samples
    b. A small descriptive table showing the amount of host/target (human/bacteria) reads per sample
    c. QC information
2. Functional composition by HUMAnN3.9 – pathways and gene families tables
3. Analysis methods
4. Optional downstream statistical analysis with LDM (Linear Decomposition Model)

References:

1. https://huttenhower.sph.harvard.edu/kneaddata/ [authors are still prepping manuscript]
2. Aitor Blanco-Miguez, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, Leonard Dubois, Kun D. Huang, Andrew Maltez Thomas, Gianmarco Piccinno, Elisa Piperni, Michal Punčochář, Mireia Valles-Colomer, Adrian Tett, Francesca Giordano, Richard Davies, Jonathan Wolf, Sarah E. Berry, Tim D. Spector, Eric A. Franzosa, Edoardo Pasolli, Francesco Asnicar, Curtis Huttenhower, Nicola Segata. **Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4.** Nature Biotechnology (2023).
3. Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltez Thomas, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A Franzosa, Nicola Segata (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3 eLife 10:e65088. https://doi.org/10.7554/eLife.65088

For sequencing data acquisition please contact Lyra Griffiths, Core Director of the Emory Integrated Genomics Core (EIGC@emory.edu).

Questions? Comments?
Please email us at EICC@emory.edu