b
UNIVERSITÄT
BERN

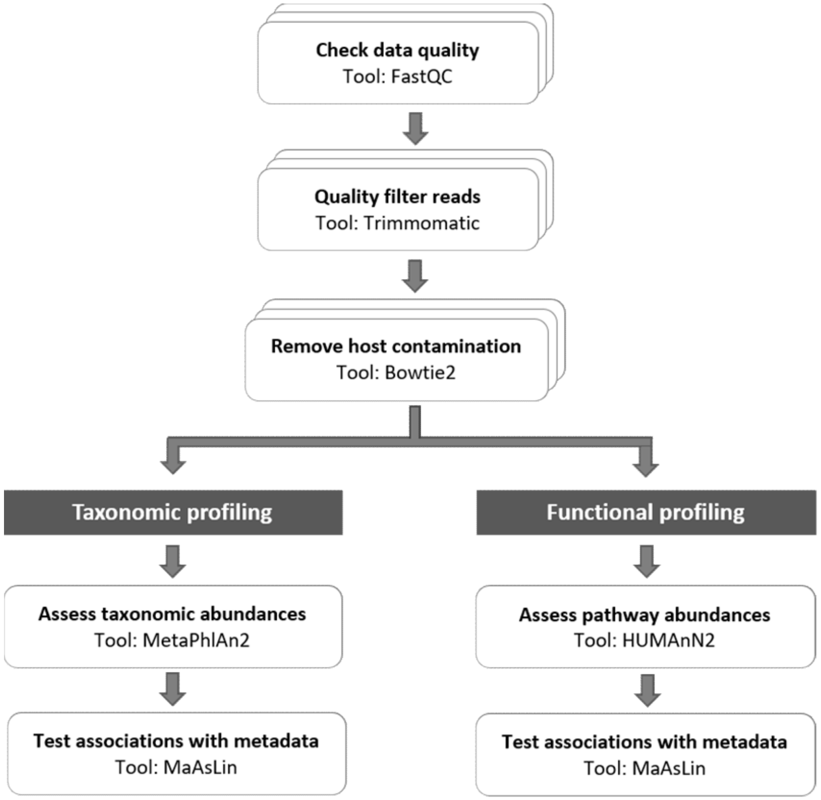
Metagenomics: Taxonomic and functional profiling

Interfaculty Bioinformatics Unit, University of Bern

07.06.2021

SIB
Swiss Institute of
Bioinformatics

Workflow



The pipeline for taxonomic and functional profiling from metagenomics data has been implemented using the Snakemake workflow engine. First, reads are quality filtered and host contamination is removed. Then, for taxonomical analysis, the reads are mapped against a set of clade-specific marker sequences using Metaphlan2. HUMAnN2 is used to assess the abundance of gene families and pathways in each sample and provide a functional interpretation of the metagenomic sequences. We then use multivariate association with linear models (MaAsLin2) to test for associations between user-defined metadata variables and the abundance of bacterial taxa and/or pathways.

Links

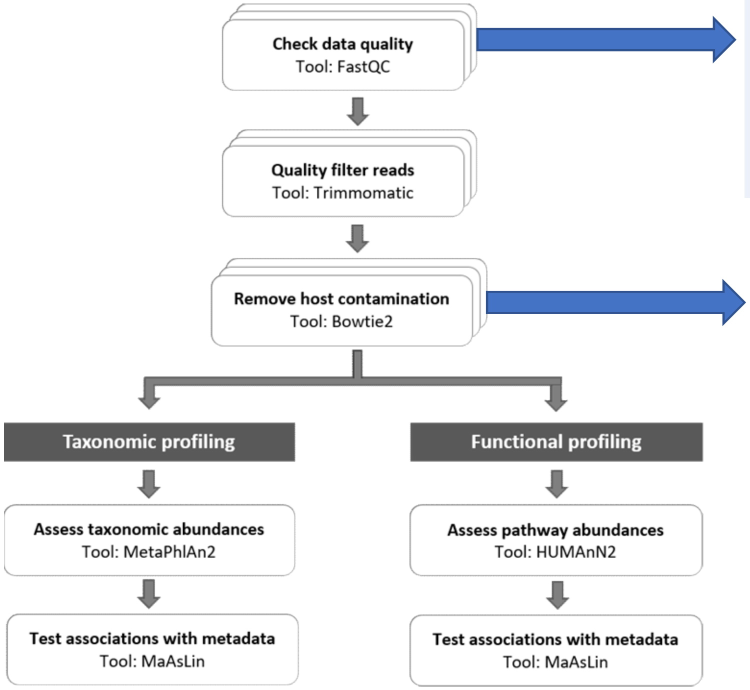
Metaphlan2: https://huttenhower.sph.harvard.edu/metaphlan2/
HUMAnN2: https://huttenhower.sph.harvard.edu/humann2/
MaAsLin2: https://huttenhower.sph.harvard.edu/maaslin/

Output files - Summary stats

Stats.txt: Table with summary statistics for each sample. This provides an overview of the data quantity and quality and, in particular, allows you to identify samples that are very different from the average. The columns contain the following information:

Sample	Sample ID
Total reads	Total number of reads (for single-end data) or read pairs (for paired-end data)
Host contamination (%)	% of reads that map to the host genome
Nbr reads mapped to Metaphlan db	Total number of reads mapped to the MetaPhlAn database. Important: MetaPhlAn2 uses markers not entire genomes to identify taxa. This is why these numbers are relatively low.
Nbr orders	Total number of orders detected by MetaPhlAn2
Nbr genera	Total number of genera detected by MetaPhlAn2
Nbr pathways	Total number of pathways detected by HUMAnN2
% unmapped	HUMAnN2: Proportion of reads that cannot be assigned to a pathway (e.g. 0.9 = 90%)
% unintegrated	HUMAnN2: Proportion of reads that are assigned to a gene family but not to a pathway

Output files - Quality checks



Folder “**fastqc**”:

One html file per sample and mate (i.e. 2 files for paired-end data) with various quality checks as outlined here: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

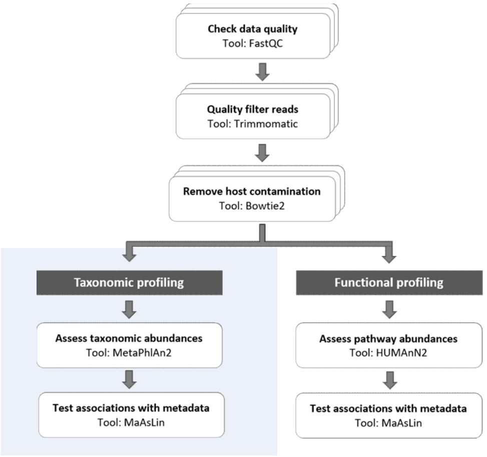
Folder “**host_contamination**”:

One file per sample with the mapping statistics from mapping the reads to the host genome. We expect to see low overall alignment rates.

Output files - Metaphlan

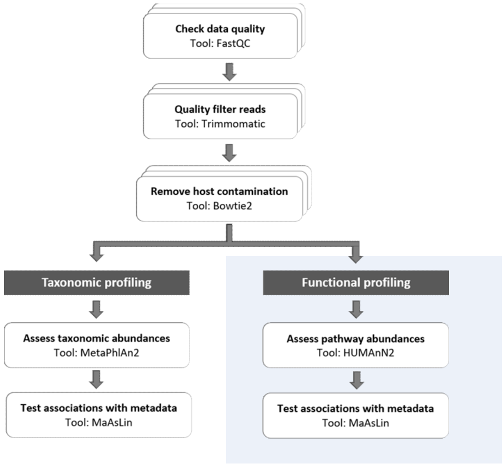
The following results are produced by the Metaphlan analysis:

- File ending in **mergedAbundances.txt**: This contains the relative abundances of the different taxa across samples as output by Metaphlan
- File ending in **metaphlan.phyla.pdf**: Barplot of the relative abundances of phyla in each sample.
- File ending in **metaphlan.PCA.pdf**: Plot of the first two axes from a principal component analysis based on the 100 most variable taxa. For samples belonging to distinct categories, 95% confidence ellipses are shown (based on multivariate t-distribution).
- Folder **maaslin**: Contains all output files from Maaslin as described in the Maaslin tutorial (https://github.com/biobakery/biobakery/wiki/maaslin2). Pdf files with boxplots will be output only for tests significant at P-adjusted < 0.25. Additionally, the folder contains one table per variable where the full Maaslin test results are merged with the relative abundances (files ending in **withRelAbundances.txt**). These tables will be the most useful for further analysis.



Output files - Humann

- **pathcoverage.tsv**: Pathway coverage across samples
- **genefamilies.tsv**: Gene family abundance across samples
- **pathabundance.tsv**: Relative abundances of pathways from. Maaslin analysis will be run only on the community totals from this file.
- Folder **maaslin**: Contains all output files from Maaslin as described in the maaslin tutorial (https://github.com/biobakery/biobakery/wiki/maaslin2). Pdf files with boxplots will be output only for tests significant at P-adjusted < 0.25. Note that pathways are recoded as documented in pathwayTranslationTable.txt.
  Additionally, the folder contains one table per variable where the full Maaslin test results are merged with the relative pathway abundances (files ending in **withRelAbundances.txt**). These tables will be the most useful for further analysis.



Check data quality
Tool: FastQC

Quality filter reads
Tool: Trimmomatic

Remove host contamination
Tool: Bowtie2

Taxonomic profiling

Assess taxonomic abundances
Tool: MetaPhlAn2

Test associations with metadata
Tool: MaAsLin

Functional profiling

Assess pathway abundances
Tool: HUMAnN2

Test associations with metadata
Tool: MaAsLin