

Read-based Metagenomics

Freed Ahmad

Who are they?

Taxonomic profiling

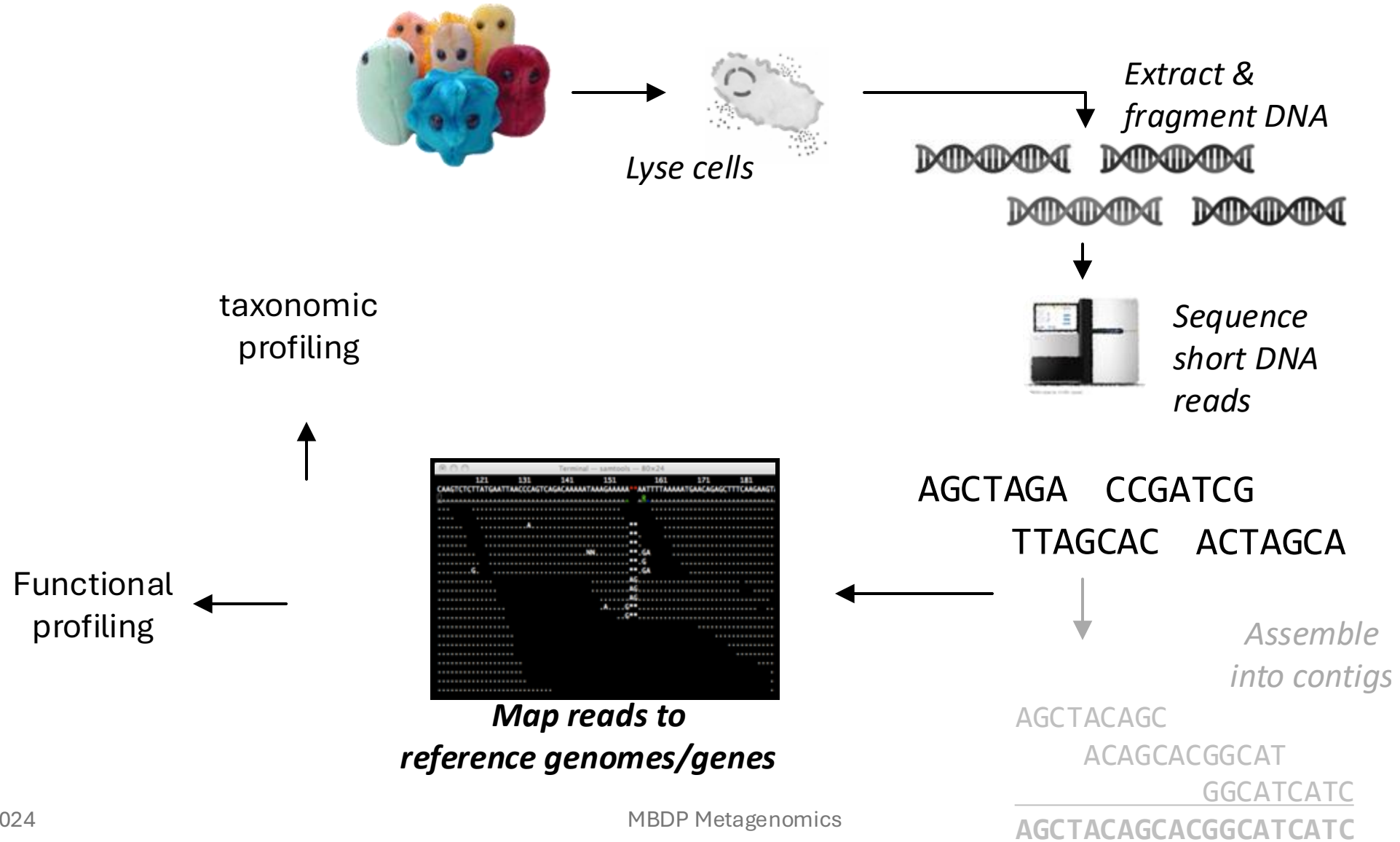


microbial community

What are they doing?

Functional profiling

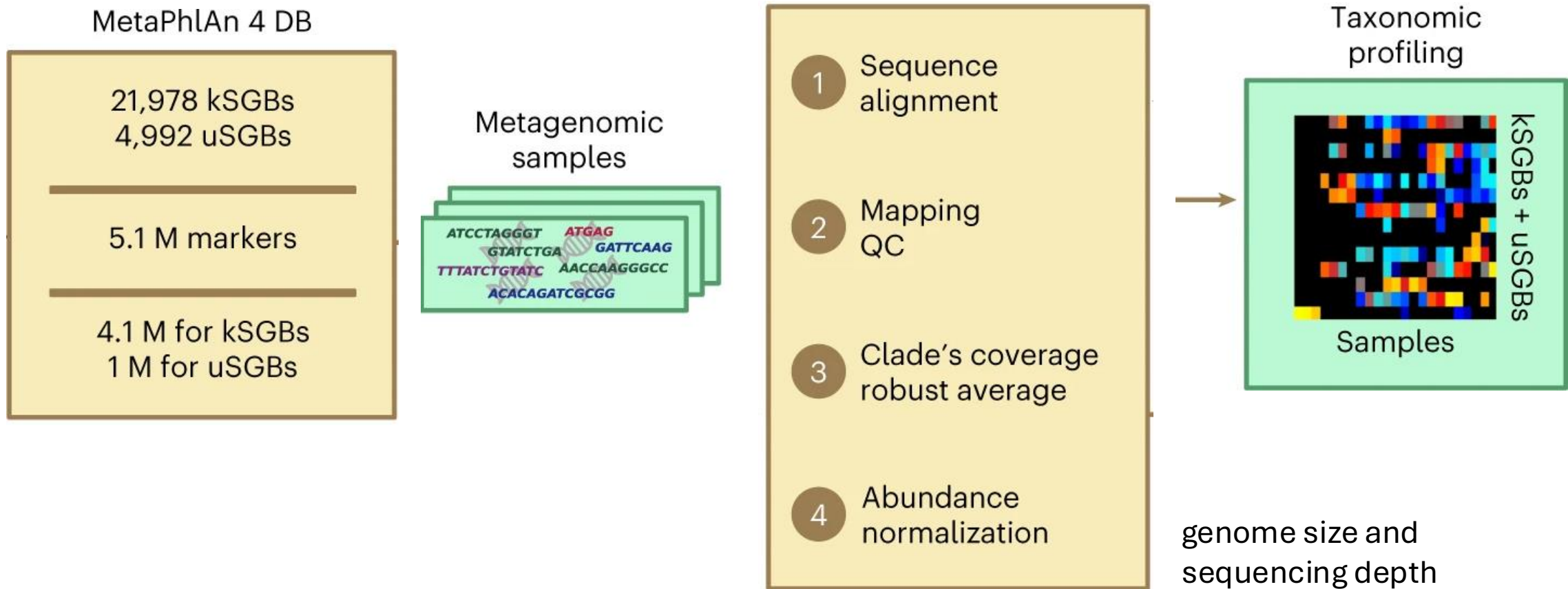
Read-based Metagenomics



MetaPhlAn Metagenomic Phylogenetic Analysis

- Estimates the relative abundance of microbial taxa using the coverage of clade-specific marker genes.
- Criteria for marker genes:
 - Present in all strains in a clade (species or otherwise), AND
 - No other clade contains homologs close enough to incorrectly map metagenomic reads.

MetaPhlAn4 Metagenomic Phylogenetic Analysis

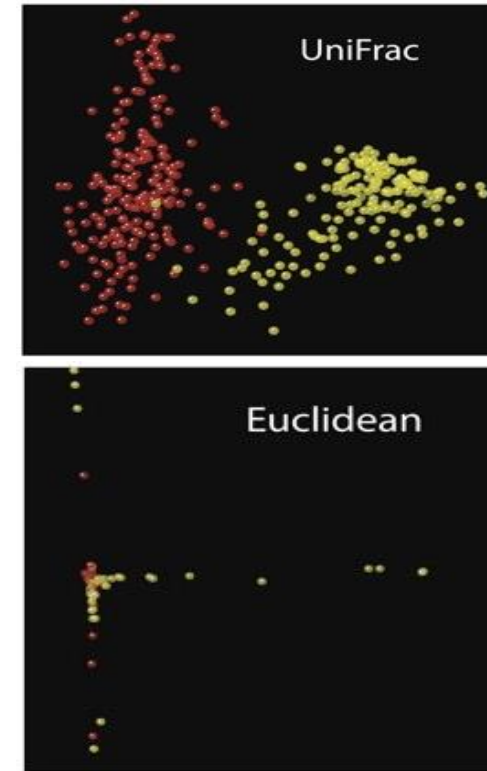
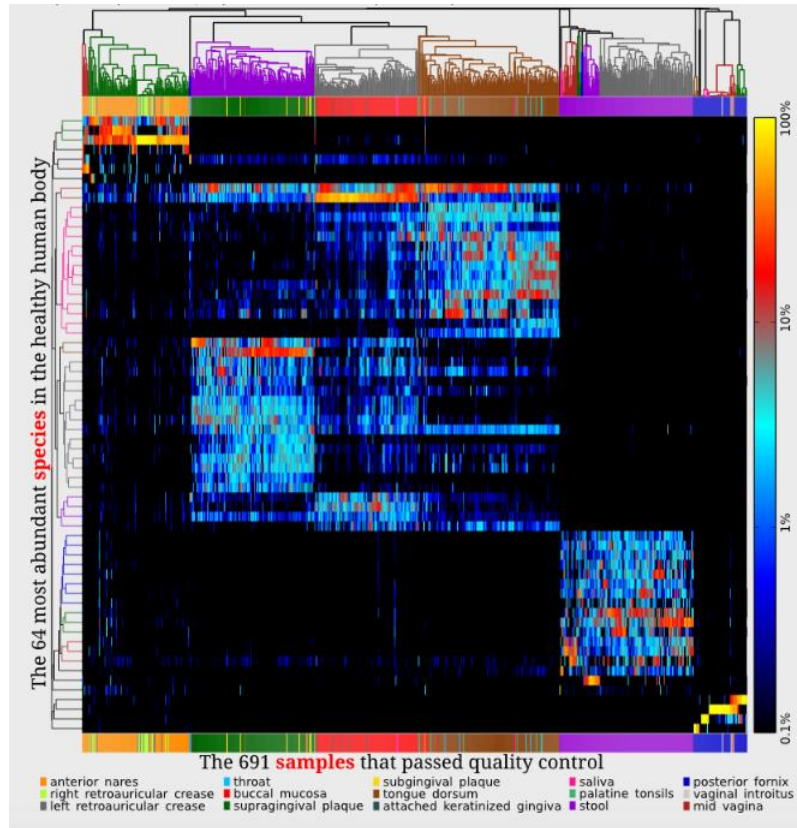


SGB: species-level genome bin

A typical output for a single sample

#clade_name	NCBI_tax_id	relative_abundance	additional_species
k_Bacteria		2	100
k_Bacteria p_Firmicutes	2 1239	47.66271	
k_Bacteria p_Bacteroidota	2 976	40.118	
k_Bacteria p_Actinobacteria	2 201174	10.03857	
k_Bacteria p_Proteobacteria	2 1224	2.04322	
k_Bacteria p_Verrucomicrobia	2 74201	0.1375	
k_Bacteria p_Bacteroidota c_Bacteroidia	2 976 200643	40.118	
k_Bacteria p_Actinobacteria c_Actinomycetia	2 201174 1760	8.5446	
k_Bacteria p_Proteobacteria c_Gammaproteobacteria	2 1224 1236	2.04322	
k_Bacteria p_Actinobacteria c_Coriobacteriia	2 201174 84998	1.49397	
k_Bacteria p_Firmicutes c_Negativicutes	2 1239 909932	0.99389	
k_Bacteria p_Firmicutes c_CFGB14936	2 1239	0.22013	
k_Bacteria p_Verrucomicrobia c_Verrucomicrobiae	2 74201 203494	0.1375	
k_Bacteria p_Firmicutes c_Bacilli	2 1239 91061	0.03224	
k_Bacteria p_Firmicutes c_CFGB1872	2 1239	0.00725	
k_Bacteria p_Firmicutes c_CFGB1354	2 1239	0.00179	
k_Bacteria p_Bacteroidota c_Bacteroidia o_Bacteroidales	2 976 200643 171549	40.118	
k_Bacteria p_Actinobacteria c_Actinomycetia o_Bifidobacteriales	2 201174 1760 85004	8.5446	
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Enterobacterales	2 1224 1236 91347	1.45491	
k_Bacteria p_Actinobacteria c_Coriobacteriia o_Coriobacteriales	2 201174 84998 84999	1.44348	
k_Bacteria p_Firmicutes c_Negativicutes o_Veillonellales	2 1239 909932 1843489	0.99389	
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Pasteurellales	2 1224 1236 135625	0.58831	
k_Bacteria p_Firmicutes c_CFGB14936 o_OFGB14936	2 1239	0.22013	
k_Bacteria p_Verrucomicrobia c_Verrucomicrobiae o_Verrucomicrobiales	2 74201 203494 48461	0.1375	
k_Bacteria p_Actinobacteria c_Coriobacteriia o_Eggerthellales	2 201174 84998 1643822	0.05049	
k_Bacteria p_Firmicutes c_Bacilli o_Lactobacillales	2 1239 91061 186826	0.03224	
k_Bacteria p_Firmicutes c_CFGB1872 o_OFGB1872	2 1239	0.00725	
k_Bacteria p_Firmicutes c_CFGB1354 o_OFGB1354	2 1239	0.00179	
k_Bacteria p_Bacteroidota c_Bacteroidia o_Bacteroidales f_Bacteroidaceae	2 976 200643 171549 815	39.74706	
k_Bacteria p_Firmicutes c_Clostridia o_Eubacteriales f_Lachnospiraceae	2 1239 186801 186802 186803	35.46703	
k_Bacteria p_Actinobacteria c_Actinomycetia o_Bifidobacteriales f_Bifidobacteriaceae	2 201174 1760 85004 31953	8.5446	
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Enterobacterales f_Enterobacteriaceae	2 1224 1236 91347 543	1.45491	
k_Bacteria p_Actinobacteria c_Coriobacteriia o_Coriobacteriales f_Coriobacteriaceae	2 201174 84998 84999 84107	1.44348	
k_Bacteria p_Firmicutes c_Clostridia o_Eubacteriales f_Eubacteriales_unclassified	2 1239 186801 186802	1.38614	
k_Bacteria p_Firmicutes c_Clostridia o_Eubacteriales f_Clostridiaceae	2 1239 186801 186802 31979	1.1541	
k_Bacteria p_Firmicutes c_Negativicutes o_Veillonellales f_Veillonellaceae	2 1239 909932 1843489 31977	0.99389	
k_Bacteria p_Proteobacteria c_Gammaproteobacteria o_Pasteurellales f_Pasteurellaceae	2 1224 1236 135625 712	0.58831	
k_Bacteria p_Firmicutes c_Clostridia o_Eubacteriales f_Peptostreptococcaceae	2 1239 186801 186802 186804	0.4335	

Visualizing microbiome data



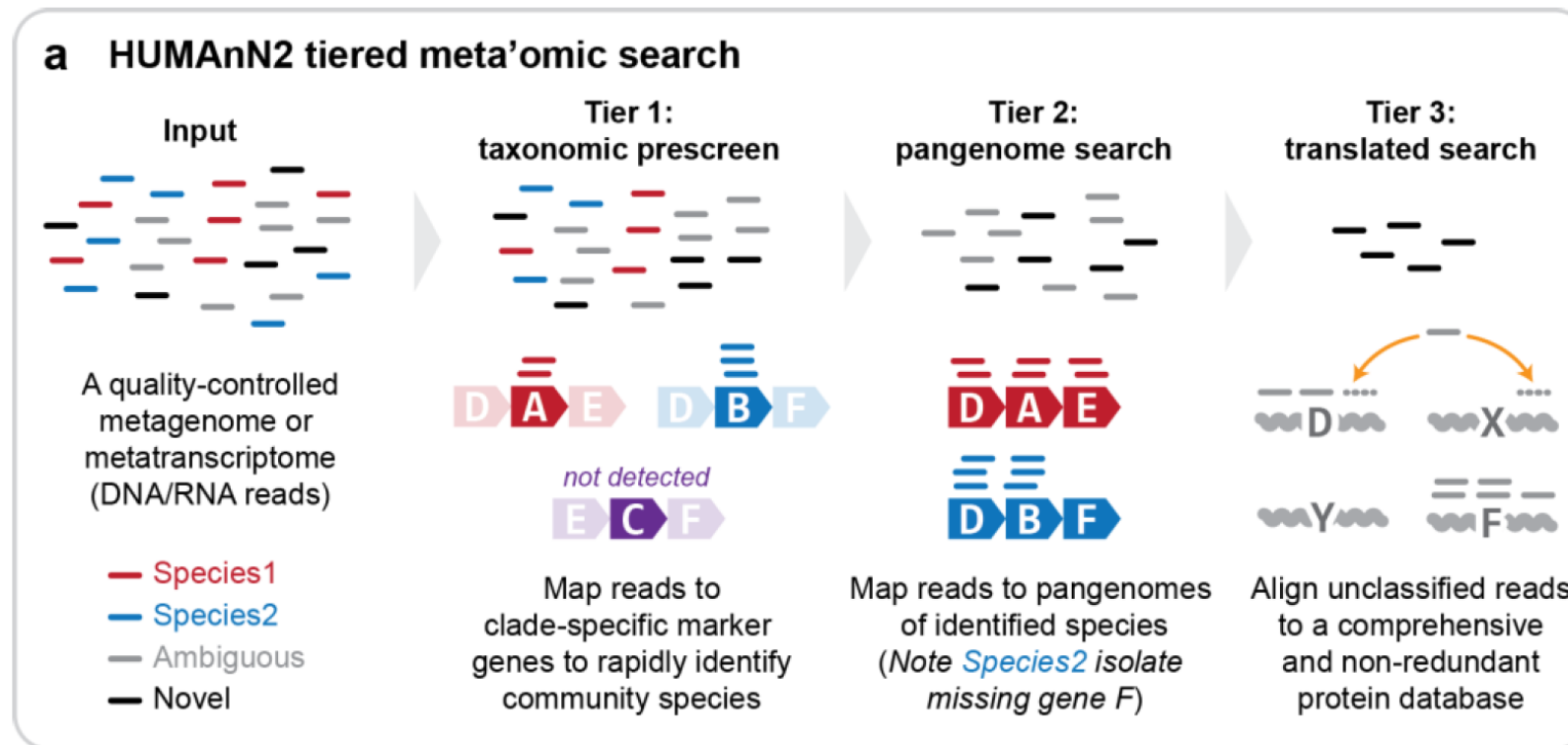
Hamady, 2009

Limitations

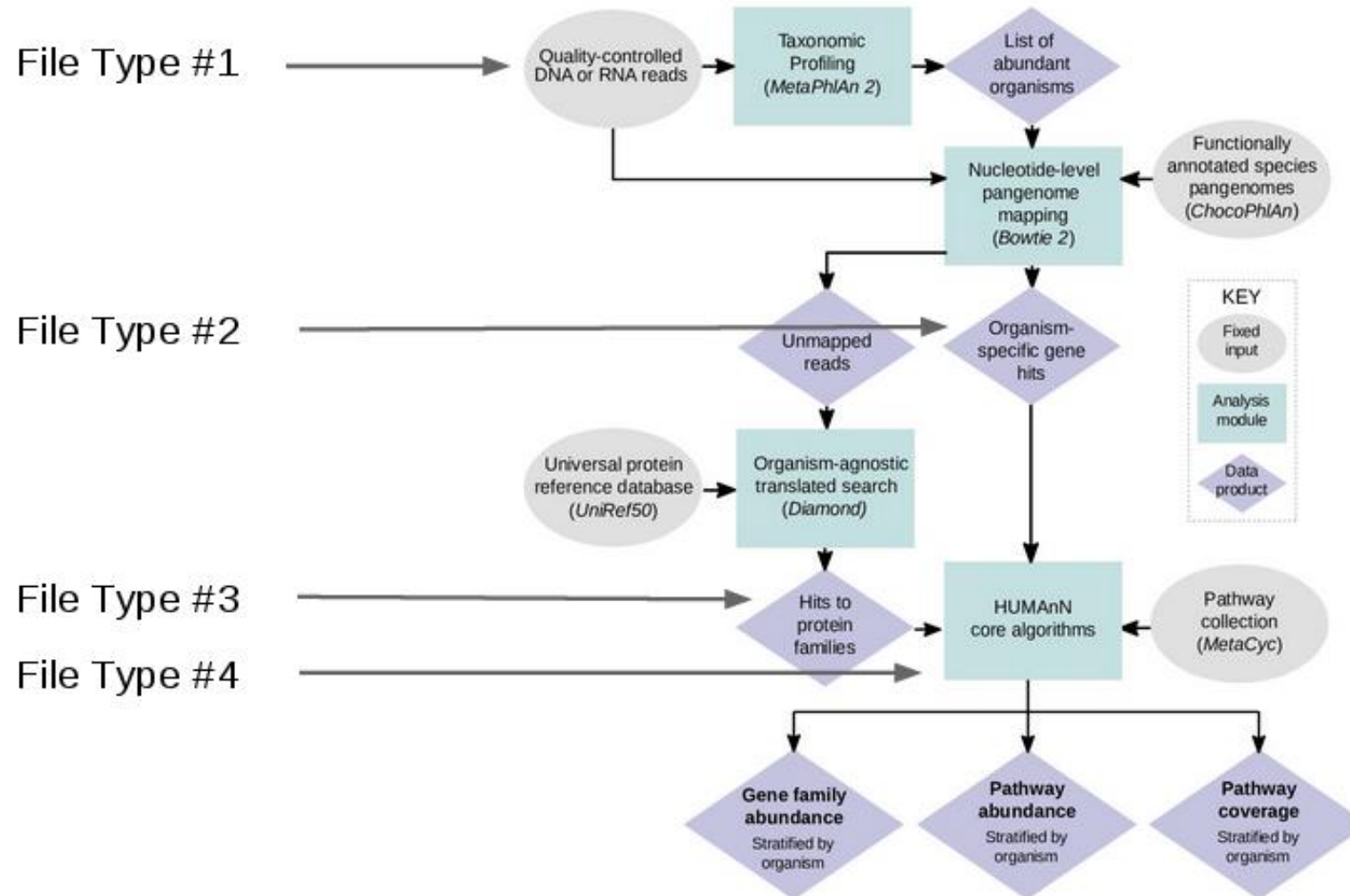
- Database/ Marker Gene Dependence
 - If the database lacks certain microbial genomes, those organisms won't be detected.
- Viral or eukaryotic microbial sequences
 - Does not extensively incorporate viral or eukaryotic microbial sequences
- Environmental Bias
 - under-characterized habitats possess more unknown sequences.

Functional profiling- HUMAnN

- HUMAnN(HMP Unified Metabolic Analysis Network) is a pipeline commonly used for profiling the presence/absence and abundance of microbial pathways in a microbial community.



HUMAnN2



less src/metaphlan.sh

```
#!/bin/bash -l
#SBATCH --job-name metaphlan
#SBATCH --output 00_LOGS/metaphlan_%A_%a.out
#SBATCH --error 00_LOGS/metaphlan_%A_%a.err
#SBATCH --time 00:30:00
#SBATCH --nodes 1
#SBATCH --ntasks-per-node 1
#SBATCH --cpus-per-task 12
#SBATCH --mem 20G
#SBATCH --account project_2001499
#SBATCH --array=1-9
#SBATCH --gres=nvme:100
```

```
module load metaphlan/4.1.1
cd /scratch/project_2001499/$USER/MBDP_Metagenomics_2024
SAMPLE_ACC=$(sed -n ${SLURM_ARRAY_TASK_ID}p 01_DATA/DF16_accessions.txt)

metaphlan \
  01_DATA/${SAMPLE_ACC}_1.fastq.gz,01_DATA/${SAMPLE_ACC}_2.fastq.gz \
  --nproc $SLURM_CPUS_PER_TASK \
  --unclassified_estimation \
  --sample_id ${SAMPLE_ACC} \
  --input_type fastq \
  --output 05_TAXONOMY/${SAMPLE_ACC}.txt \
  --bowtie2db /scratch/project_2001499/Databases/metaphlan \
  --bowtie2out 05_TAXONOMY/${SAMPLE_ACC}.bowtie2.bz2
```

Thank you!