

Core Bioinformatics **Workshops 2024**

Metagenomic profiling
with BioBakery
MetaPhlAn and HuMANN



Session overview

- Introduction
- Biobakery Tools
- Taxonomic profiling using MetaPhlAn 4
- Functional Profiling using HuMANN3
- Questions!

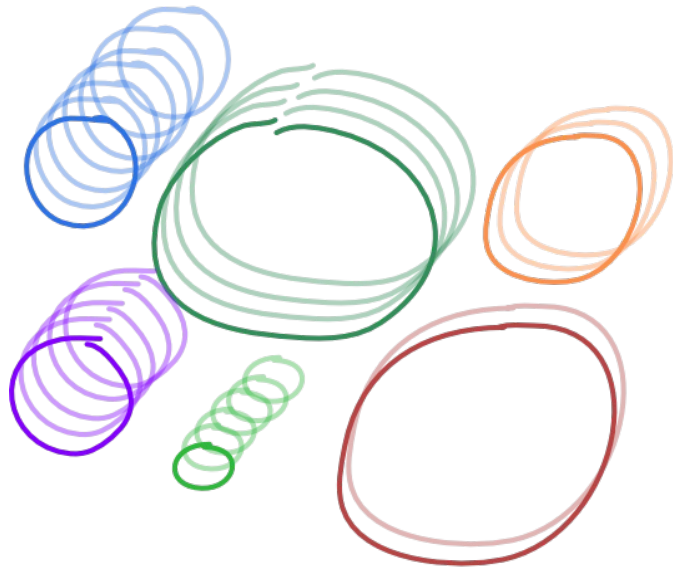


Alise Ponsero

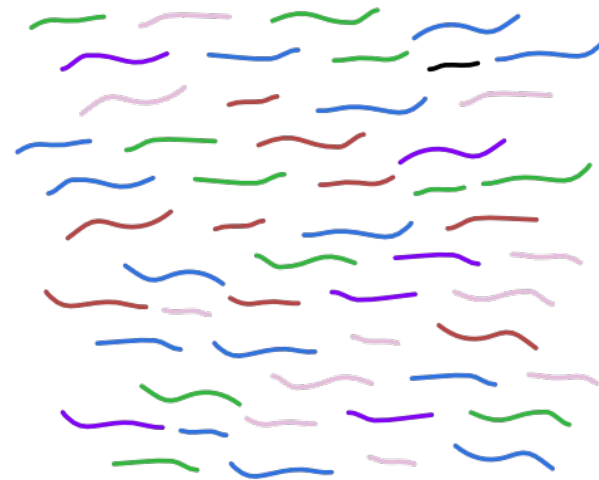


Introduction to short reads profiling

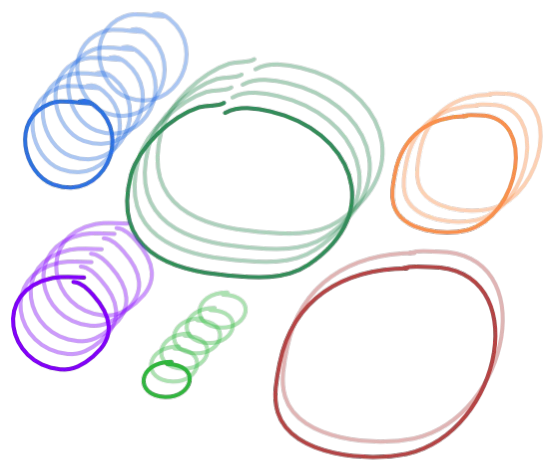
github.com/quadram-institute-bioscience/biobakery-2024



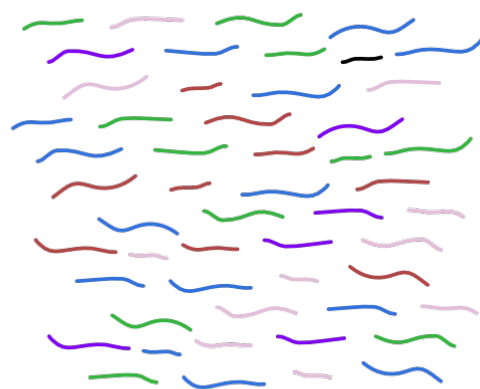
A "Metagenome"



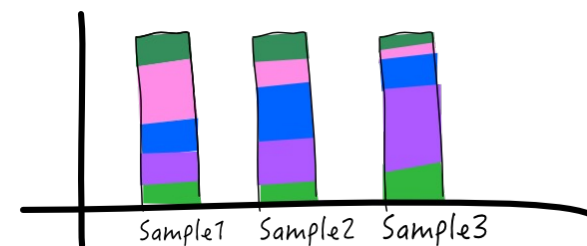
Whole Metagenome Shotgun
(WMS)



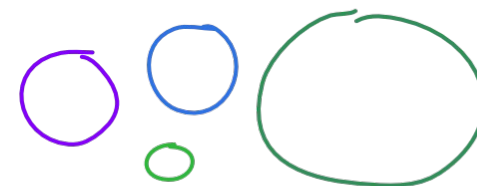
A "Metagenome"



Whole Metagenome Shotgun
(WMS)



Profiling



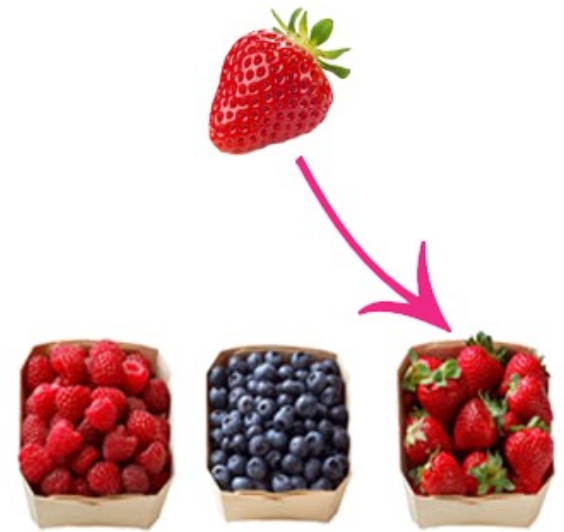
Assembly and MAGs

Classifying reads

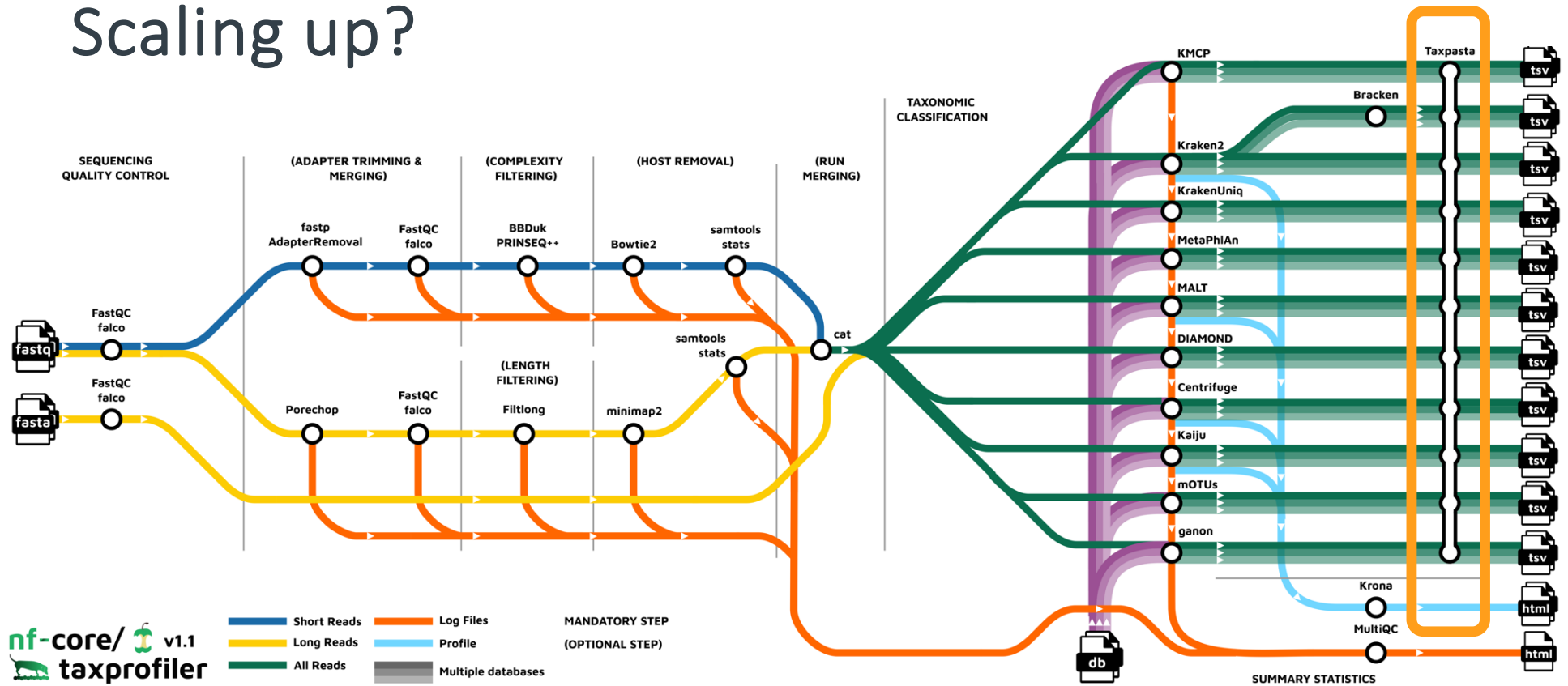
- Mapping against known genomes?

Alternatives

- DNA *k*-mer based approaches (notably: **Kraken2**)
- Marker based approaches (notably: **MetaPhlAn**)
- Protein sequence alignment (e.g.: Kaiju)
- Pseudomapping (e.g. KMCP)

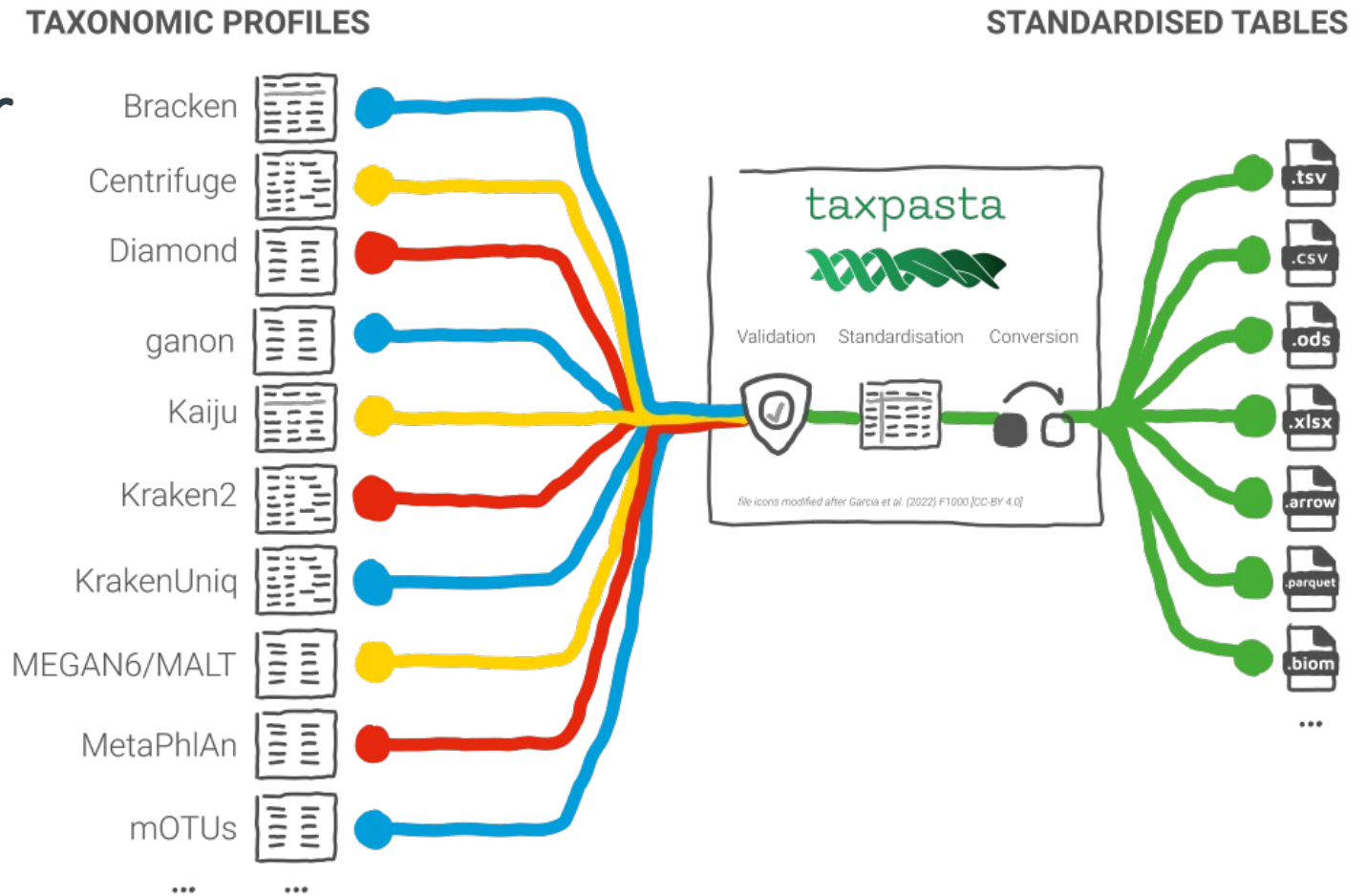


Scaling up?



nf-core/taxprofiler v1.1

Taxprofiler USP



What is the Biobakery?

The Biobakery

The bioBakery tools methods for microbial community profiling developed by the [Huttenhower lab](#).

- Most tools are supported both as individual software packages or as workflows
- The tools cover methods for microbial community profiling but also downstream analysis and statistical methods
- The community is supported by a forum where users can ask questions or requests

The Biobakery

Microbial community profiling tools



Note: not all tools of the Biobakery can work together...

They are mostly meant to be run independent of each other!

The Biobakery

Visualization and statistical analysis tools

HALLA
Perform well-powered comparisons of paired, high-dimensional datasets

ARepA
Extract and normalize 'omics data from online repositories

CCREPE
Assess the significance of similarity measures in compositional data

LefSe
Associate up to two metadata with microbiome features

MaAsLin
Associate arbitrarily complex metadata with microbiome features

MMUPHin
Correct batch effects, meta-analyze microbes, genes, and pathways across multiple studies

MicroPITA
Select samples for follow-up analysis in two-stage tiered studies

SparseDOSSA
A hierarchical model of microbial ecological population structure

BAnOCC
Bayesian assessment of association in compositional data

GraPhlAn
Generate cladograms and decorate with metadata

Utilities

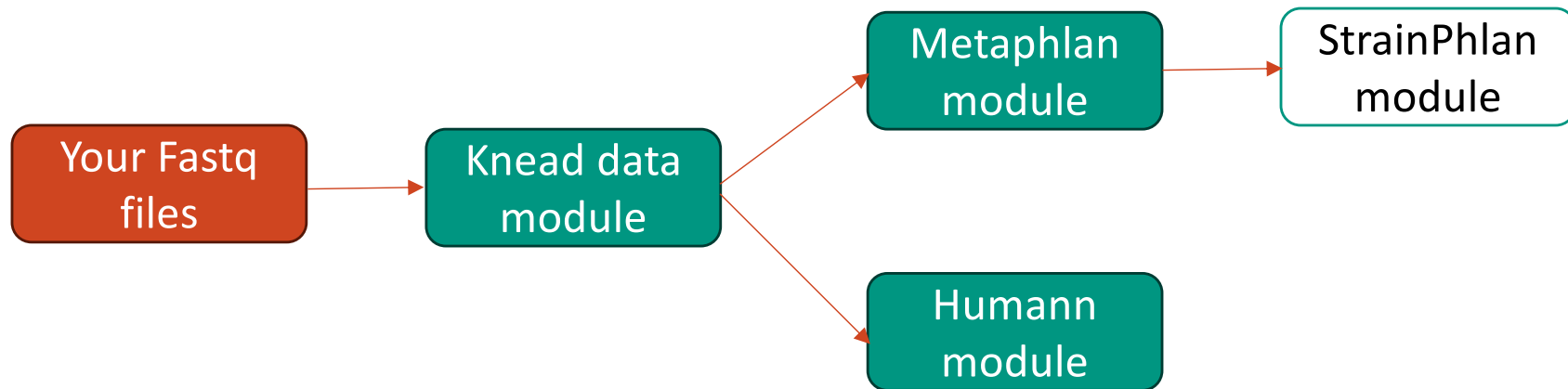
KneadData
Perform quality control of meta-omic reads, including host read removal

AnADAMA
Perform and document automated scientific workflows

Workflows
A collection of meta-omic data processing and visualization workflows

The Biobakery workflows

Several workflows are proposed to link the main Biobakery tools together:



The Biobakery help forum

The bioBakery help forum

Sign Up Log In

categories Categories Top Latest

Category	Topics
Downstream analysis and statistics Tools for microbial community modeling, significance calculations, differential abundance testing, and more. LEfSe MaAsLin microPITA SparseDOSSA BAnOCC MMUPHin HAIIA ARepA CCREPE anpan PARATHAA	12 / month
Microbial community profiling Tools for bioinformatics on raw microbial community data, typically quantifying features such as taxa, genes, functions, metabolites, or other molecular or cellular activities. HUMANN MetaPhlAn PhyloPhlAn PICRUST ShortBRED PPANINI StrainPhlAn MelonnPan WAAFLA PanPhlAn MetaWIBELE MACARRoN FUGAsseM	22 / month
Infrastructure and utilities General bioBakery or infrastructure utilities for microbial community or other computational biology research. KneadData bioBakery workflows GraPhlAn AnADAMA2	4 / month
Data resource IBDMDB	1 / month

Bug, requests, and suggestions are handled in the Biobakery help forum (not through Github)

Taxonomic profiling with Metaphlan4

concepts and quirks

Taxonomic profiler or classifier?

Taxonomic Classifiers (Kraken2):

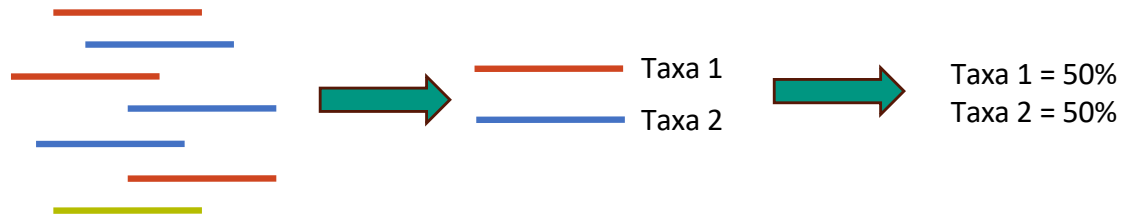
- Aim to classify each sequencing reads independently from each other
- Do not account for genome size
- Do not aim to give you the relative abundance of taxa in the community



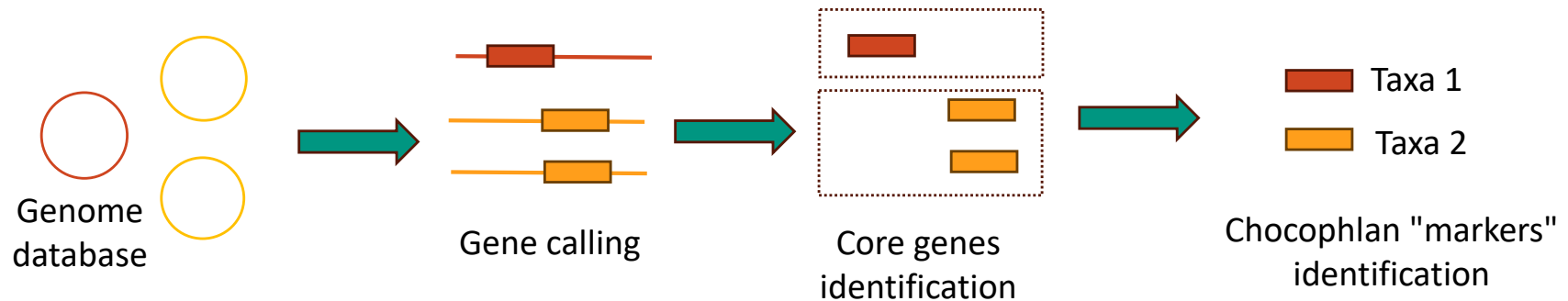
Taxonomic profiler or classifier?

Taxonomic profiler (MetaPhlan):

- Takes into account genome size
- Aims to give you the relative abundance of the taxa in the community
- Can overlook the unclassified/unknowns in the community



Chocophlan3 database



Chocophlan is a **pangenome** database used in Metaphlan and Humann

--> Most of the Chocophlan databases do not include viral/eukaryotic sequences

--> Does not include a human reference

Chocophlan3 database



Chocophlan3 database

Chocophlan1 = 400,141 species markers from 2,834 NCBI/IMG genomes

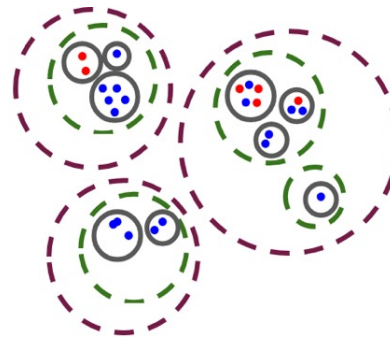


Chocophlan3 = ~7.3M unique clade-specific marker genes
from 36,822 species-level genome bins (SGB)
(11,062 of them taxonomically unidentified at the species level)

Cell

Volume 176, Issue 3, 24 January 2019, Pages 649-662.e20

Extensive Unexplored Human Microbiome
Diversity Revealed by Over 150,000 Genomes from
Metagenomes Spanning Age, Geography, and
Lifestyle



Estimation of the unknown in Metaphlan4

%uncl.reads =

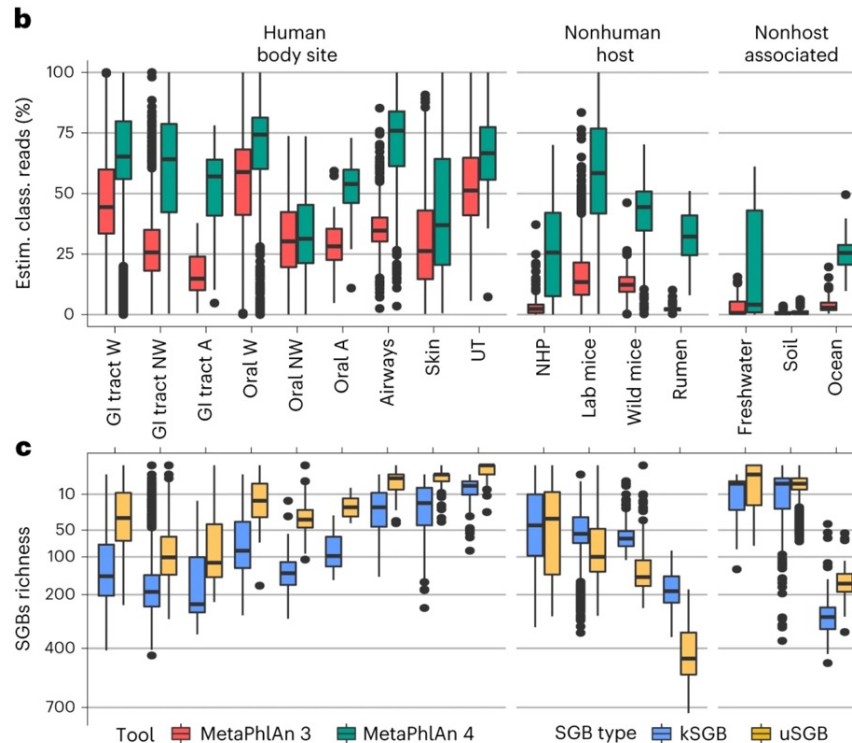
$$\frac{\text{Total reads} - \left(\sum_{sp=0}^n (\text{avg nonzero markers coverage}_{sp} \times \text{avg genome length}_{sp}) \right) / \text{avg read length}}{\text{Total reads}}$$

sp = indices of all the SGBs reported in the MetaPhlAn profile

MetaPhlAn 4 includes a feature for estimating the fraction of input reads that **cannot be assigned to taxa in the database**.

Calculated by subtracting from the total number of input reads the average read depth of each reported SGB normalized by its SGB-specific average genome length

Metaphlan3 vs Metaphlan4



Metaphlan4 new database significantly improves the % of classified reads in particular for human associated ecosystems

Blanco-Míguez, A., Beghini, F., Cumbo, F. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* **41**, 1633–1644 (2023).
<https://doi.org/10.1038/s41587-023-01688-w>

Conclusions & take-home messages

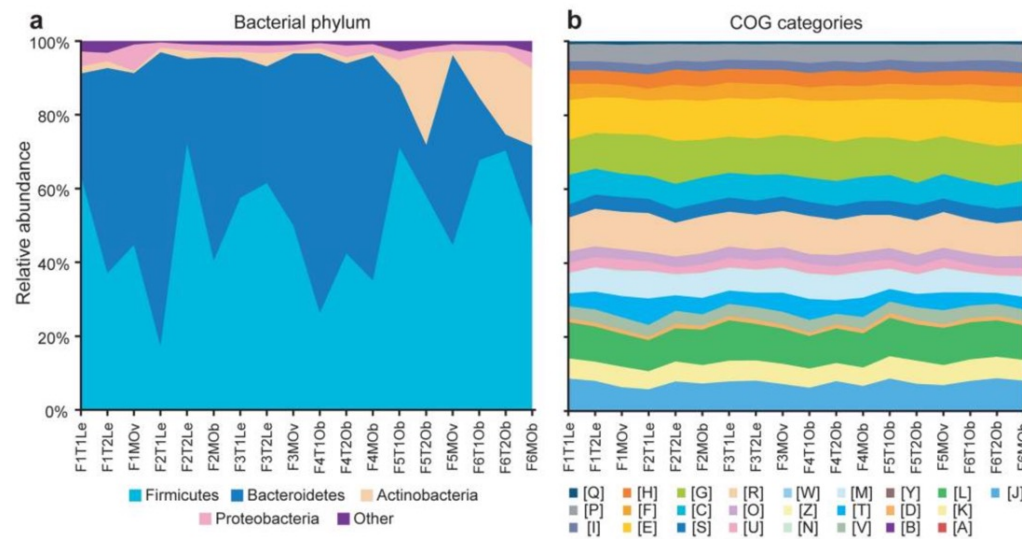
- Metaphlan4 is a metagenomic profiler
- Relies on the Chocophlan pangenome database for its predictions
- Chocophlan3 includes a large proportion of MAGs clustered *de-novo*
- Metaphlan4 is able to give an estimate of the proportion of unknowns in the metagenome (but not by default)

Functional profiling with HuMaNN3

concepts and quirks

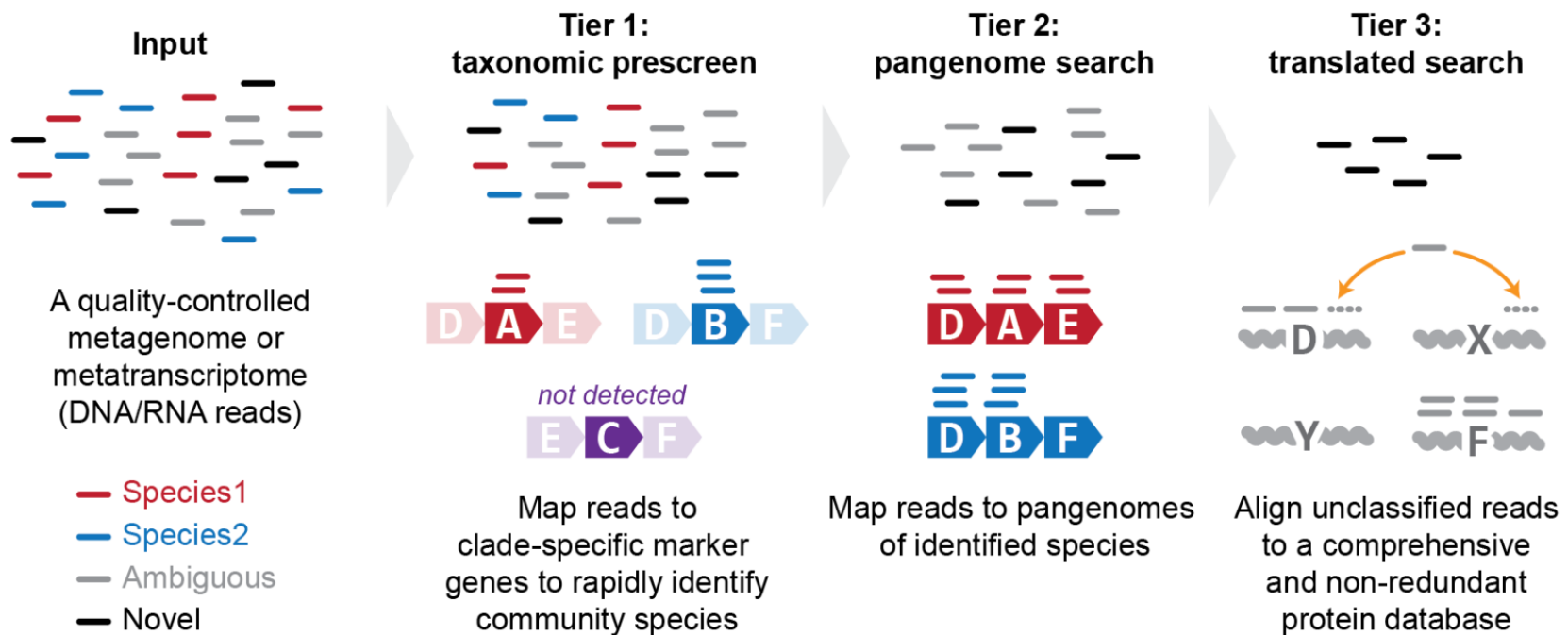
Functional profiling principles

While taxonomic composition and functional potential of the microbial communities are linked...



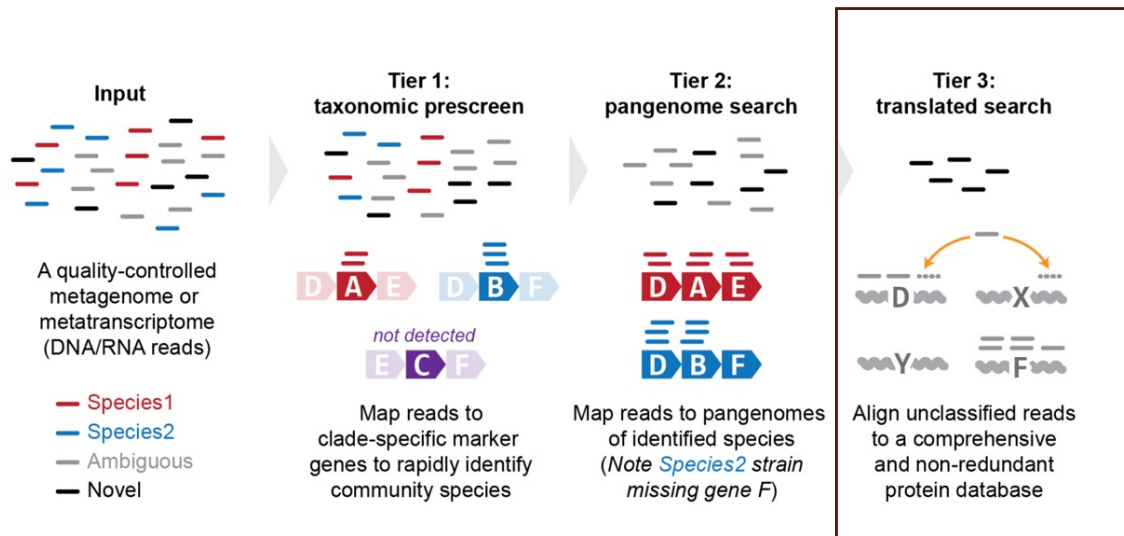
A change in the microbial taxonomic composition does not mean a change in the functional potential!

Humann3 tiered search approach



<https://github.com/biobakery/biobakery/wiki/humann3>

Using UniRef90 vs UniRef50



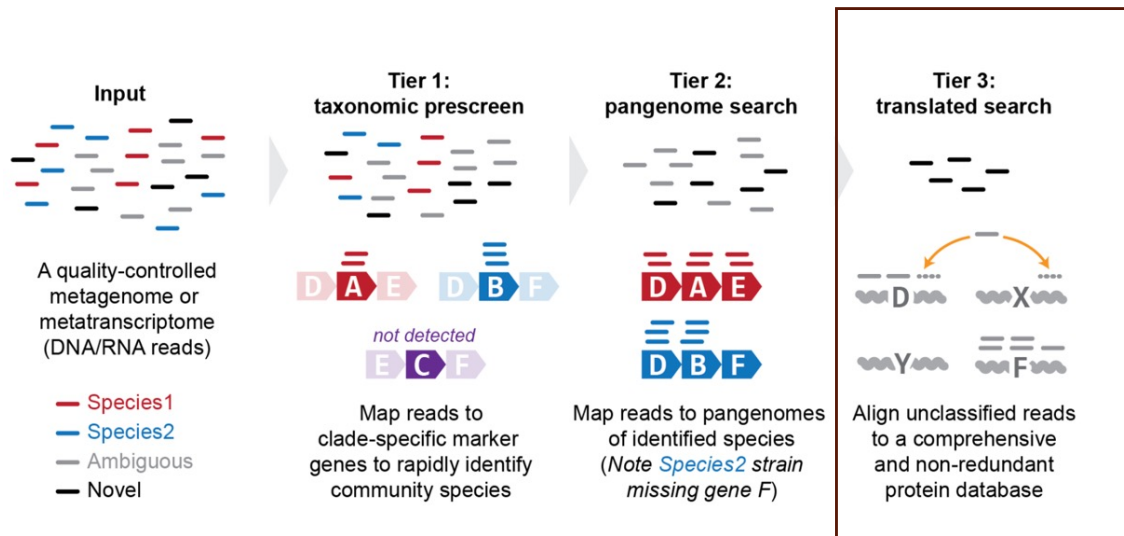
Humann3 uses UniRef protein clusters as a gene family system.

Clustering proteins from UniProt :

- **Uniref90:** clustering non-redundant proteins at 90% identity and selecting representative sequences
- **UniRef50:** clustering UniRef90 representative sequences at 50% identity to produce broader clusters.

The representative of a UniRef cluster is the best-annotated member of the cluster.

Using UniRef90 vs UniRef50



You can choose to use either UniRef90 or UniRef50 in Humann3:

- **Default** : UniRef90
- **Poorly characterized environments**: UniRef50

expected to explain a larger portion of sample read

but reduced functional resolution

Humann3 gene and pathways abundance

Feature	RPK
GeneA	2
GeneA Species1	2
GeneB	3
GeneB Species2	3
Σ GeneD	8
GeneD Species1	2
GeneD Species2	3
GeneD unclassified	3
GeneE	2
GeneE Species1	2
GeneF	5
GeneF unclassified	5

Process mapping results to estimate per-species and community total gene family abundance, weighting by 1) alignment quality, 2) gene length, and 3) gene coverage

RPK : Reads per kilobases

Normalize count per gene length

--> Utilities to convert the counts to relative abundances (RPKM)

RPKM : RPK per million reads

Normalize your RPK by the total number of reads in your sample

Note : in Humann3, RPKM is referred as "CPM"

Humann3 gene and pathways abundance

Feature	RPK
GeneA	2
GeneA Species1	2
GeneB	3
GeneB Species2	3
Σ GeneD	8
GeneD Species1	2
GeneD Species2	3
GeneD unclassified	3
GeneE	2
GeneE Species1	2
GeneF	5
GeneF unclassified	5

Process mapping results to estimate per-species and community total gene family abundance, weighting by 1) alignment quality, 2) gene length, and 3) gene coverage

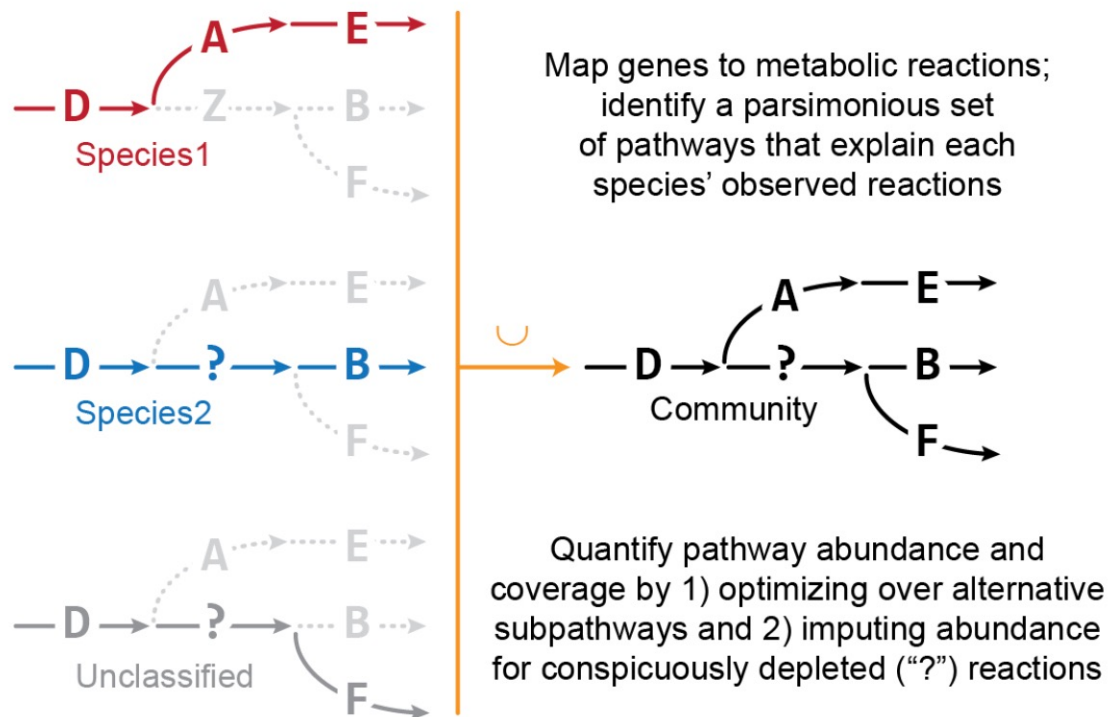
Mappings are available for both UniRef90 and UniRef50 gene families to :

- MetaCyc Reactions
- **KEGG Orthogroups (KOs)**
- Pfam domains
- Level-4 enzyme commission (EC) categories
- EggNOG (including COGs)
- Gene Ontology (GO)
- Informative GO

Importantly: some genes will not be mapped to the new gene families (and will be labeled as "unmapped")

<https://github.com/biobakery/biobakery/wiki/humann3>

Humann3 gene and pathways abundance



HUMAnN3 uses MetaCyc pathway definitions and MinPath by default.

--> The user can provide a custom pathways database

Humann3 quirks

- Humann3 (just like Metaphlan4) does not account for the information of paired-end reads:
 - The authors suggest to concatenate your two files
 - In my opinion, this inflates the computational runtime and I suggest using only the forward pair
- Human genome decontamination is important to do before running Humann3
- Can be used on meta-transcriptomes

<https://github.com/biobakery/biobakery/wiki/humann3>

Thank you



Quadram Institute
Norwich Research Park
Norfolk NR4 7UQ

quadram.ac.uk

Follow us     

 **Quadram**
Institute
Science • Health • Food • Innovation