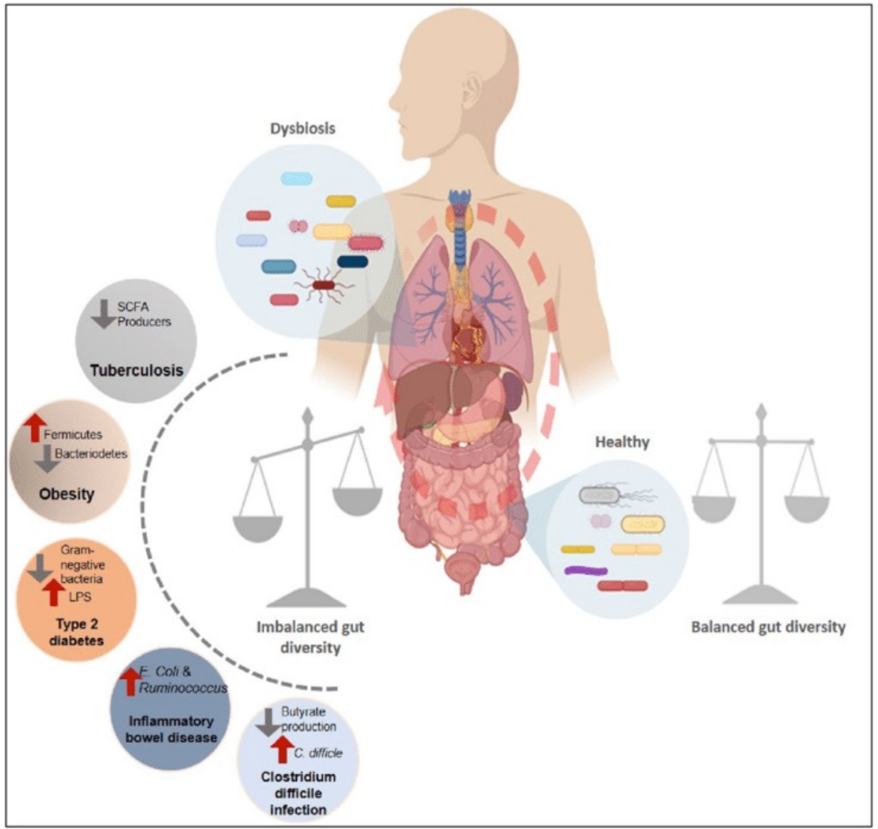
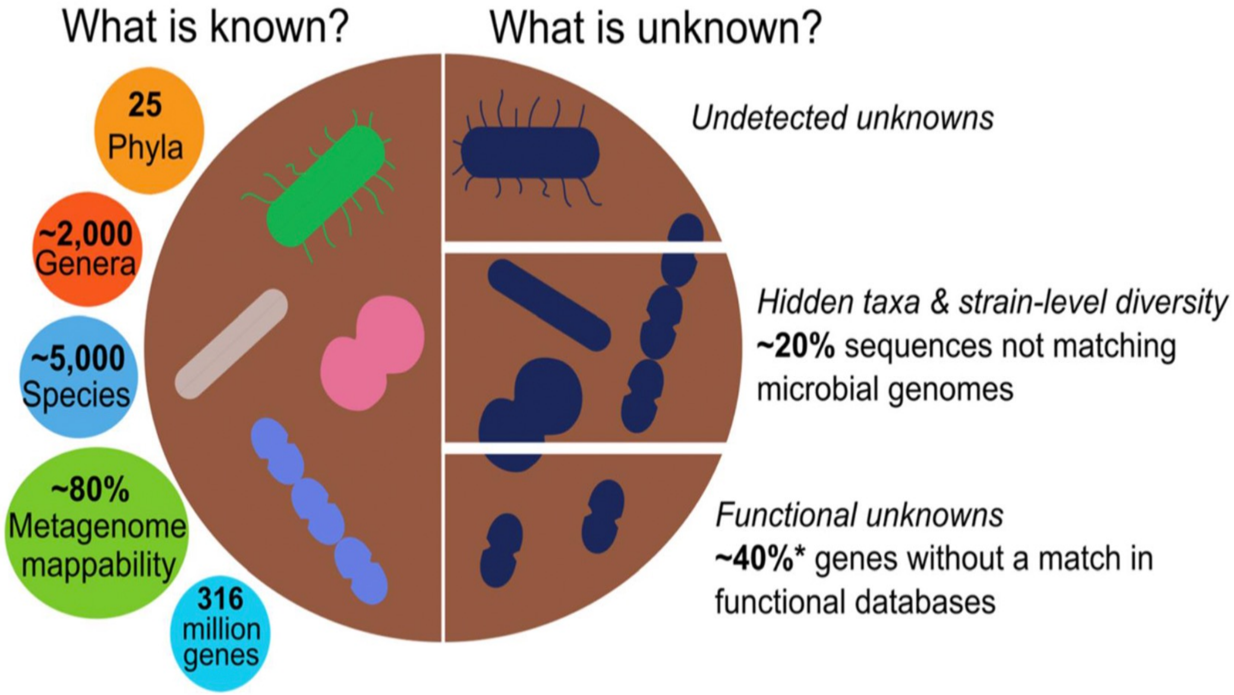
METAPHLAN2 FOR ENHANCED METAGENOMIC TAXONOMIC PROFILING.

TRUONG, Duy Tin, et al. *Nature methods*,
2015, 12.10: 902-903.

1. Introduction

- Profiling the taxonomic and phylogenetic compositions of such communities is critical for understanding their biology and characterizing complex disorders that do not appear to be associated with any individual microbes.




What is known?
What is unknown?
25 Phyla
~2,000 Genera
~5,000 Species
~80% Metagenome mappability
316 million genes
Undetected unknowns
Hidden taxa & strain-level diversity
~20% sequences not matching microbial genomes
Functional unknowns
~40%* genes without a match in functional databases

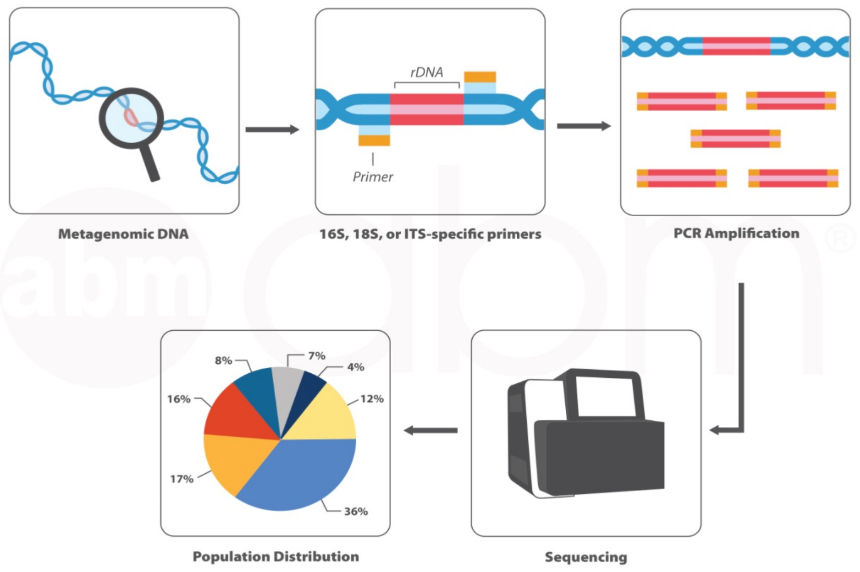
SINGHVI, Nirjara, et al. Interplay of human gut microbiome in health and wellness. Indian journal of microbiology, 2020, 60.1: 26-36.

The knowns and unknowns of the human microbiome, gut microbiota for health

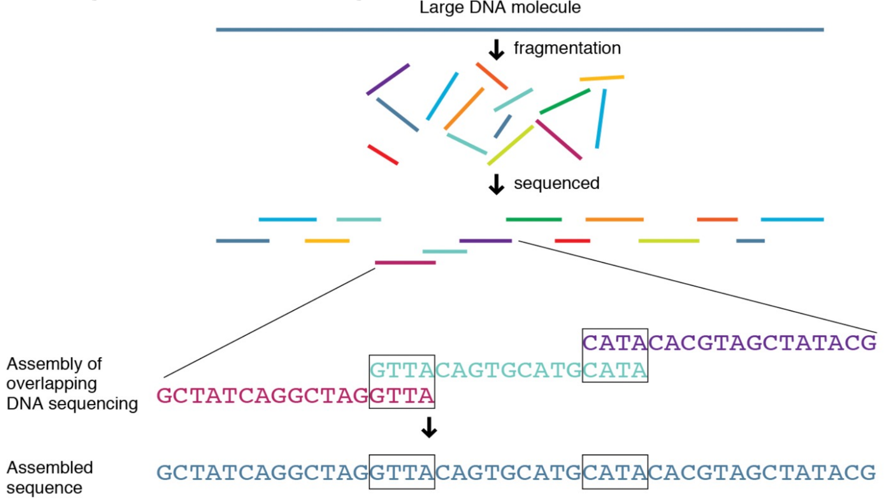
1. Introduction

	16S/ITS Sequencing	Shotgun Sequencing
Bacteria/Fungi Coverage	High	Limited
Cross-Domain Coverage	No	Yes
False Positives	Low Risk	High Risk
Taxonomy Resolution	Genus-Species	Species-Strains
Host DNA Interference	No	Yes
Minimum DNA Input	10 copies of 16S	1 ng
Functional Profiling	No	Yes
Recommended Sample Type	All	Human Microbiome
Cost per Sample	~ $80	~ $200

16s rRNA sequencing



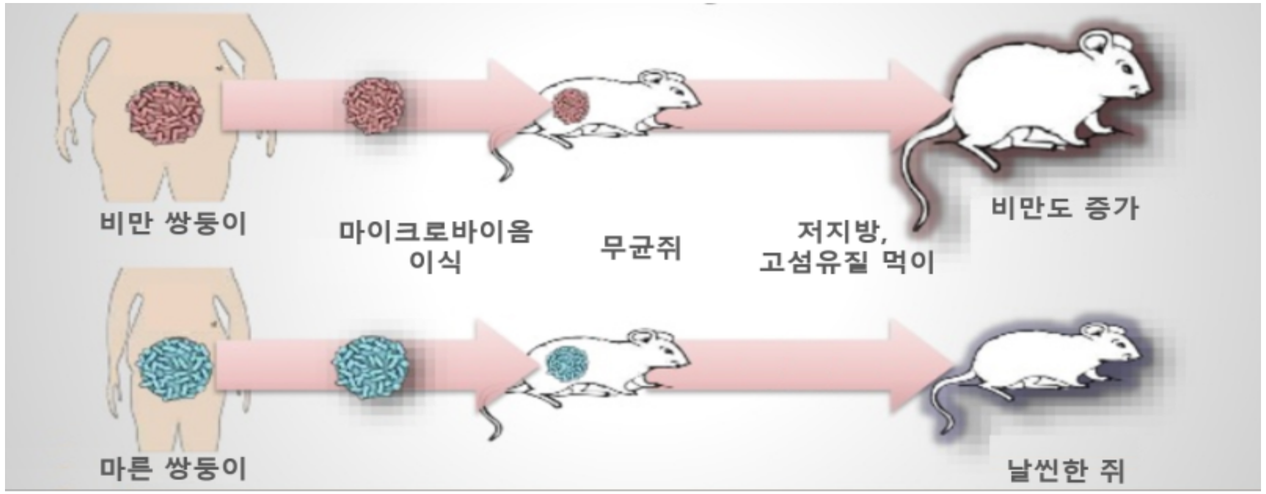
Shotgun sequencing



https://www.abmgood.com/16S-rDNA-Amplicon-Sequencing.html
https://www.zymoresearch.com/blogs/blog/16s-sequencing-vs-shotgun-metagenomic-sequencing
https://www.genome.gov/genetics-glossary/Shotgun-Sequencing

1. Why we look at these metagenomic profiles?

- Profiling the taxonomic and phylogenetic compositions of such communities is critical for understanding their biology and characterizing complex disorders that do not appear to be associated with any individual microbes.
- They populated their guts with intestinal microbes collected from obese women and their lean twin sister.
  - *The mice ate the same diet in equal amounts, yet the animals that received bacteria from an obese twin grew heavier and had more body fat than mice with microbes from a thin twin.*
- The big question in metagenomics is ***who is there (taxonomic profiling)***?



1. What is MetaPhlan2?
: Taxonomic profiling using unique marker genes

- MetaPhlAn2 (metagenomic phylogenetic analysis) is a method for characterizing the taxonomic profiles of whole-metagenome shotgun samples that has been used successfully in large-scale microbial community studies.
- This work complements the original species-level profiling method with a system for eukaryotic and viral quantitation, strain-level identification and strain tracking.
  - *unambiguous taxonomic assignments*
  - *accurate estimation of organismal relative abundance*
  - *species-level resolution for bacteria, archaea, eukaryotes and viruses*
  - *strain identification and tracking*
  - *orders of magnitude speedups compared to existing methods.*

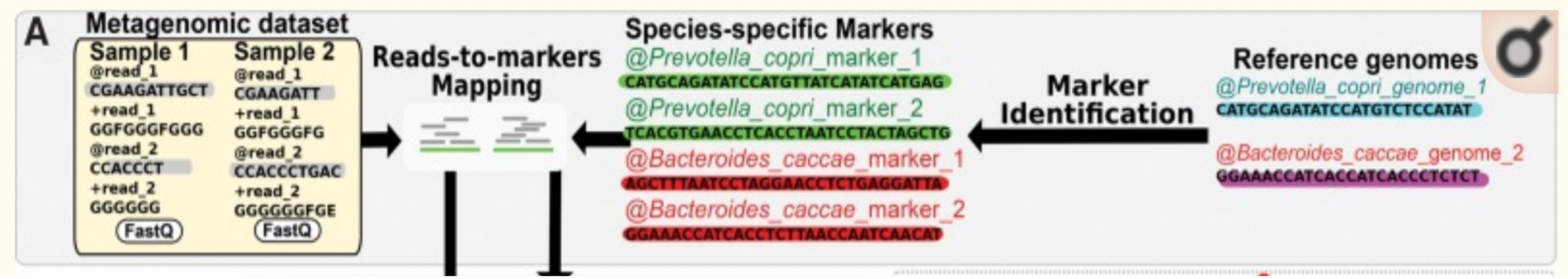
2.
MetaPhlAn Overall Pipeline

SEGATA, Nicola, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods, 2012, 9.8: 811-814.

2. Overall pipeline : Taxonomic profiling with MetaPhlAn2

2-1. Acquire reference genome → 2-2. Find clade-specific marker genes. → 2-3. Sequence sample

Input : fastq file

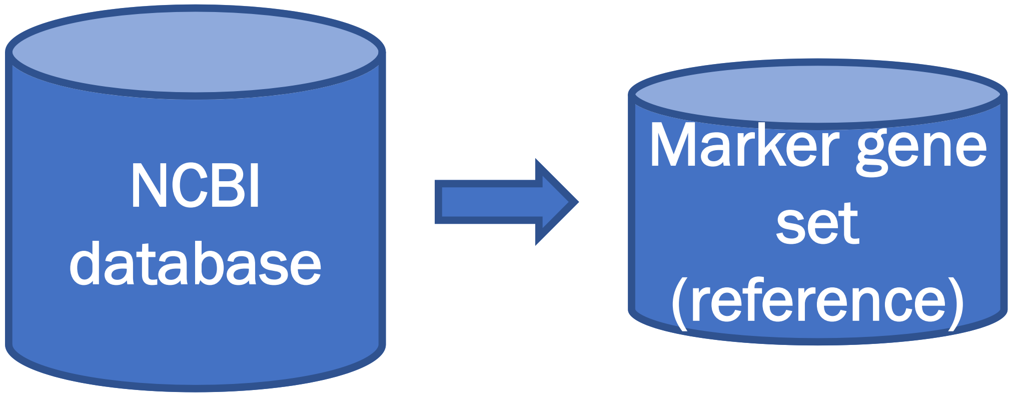

A
Metagenomic dataset
Sample 1
@read_1
CGAAGATTGCT
+read_1
GGFGGGFGGG
@read_2
CCACCCT
+read_2
GGGGGG
FastQ
Sample 2
@read_1
CGAAGATT
+read_1
GGFGGGFG
@read_2
CCACCCTGAC
+read_2
GGGGGGFGE
FastQ
Reads-to-markers Mapping
Species-specific Markers
@Prevotella_copri_marker_1
CATGCAGATATCCATGTTATCATATCATGAG
@Prevotella_copri_marker_2
TCACGTGAACCTCACCTAATCCTACTAGCTG
@Bacteroides_caccae_marker_1
AGCTTTAATCCTAGGAACCTCTGAGGATTA
@Bacteroides_caccae_marker_2
GGAAACCATCACCTCTTAACCAATCAACAT
Marker Identification
Reference genomes
@Prevotella_copri_genome_1
CATGCAGATATCCATGTCTCCATAT
@Bacteroides_caccae_genome_2
GGAAACCATCACCATCACCCTCTCT

2-4. Mapping of metagenomes to the marker gene catalog → 2-5. Estimation of organismal relative abundance. → 2-6. Visualization

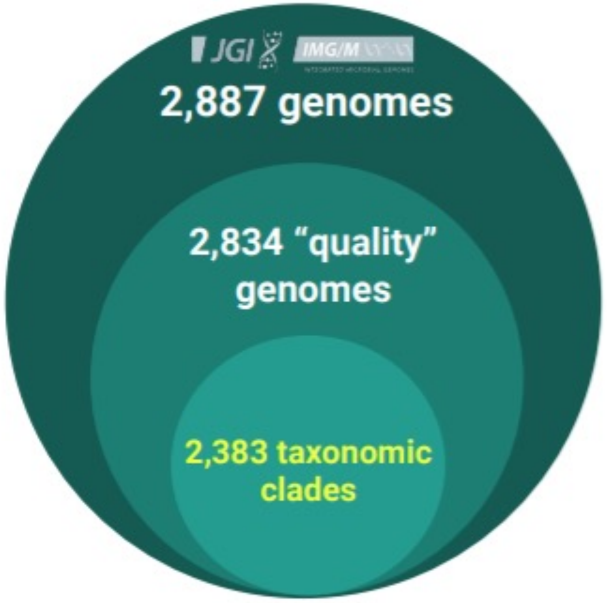
Output is relative abundance at different taxonomic levels

2-1. Acquire reference genome

- Thus, starting from the 2,887 genomes.
- These are screened for minimum length (>50,000 nt), minimum number of CDSs (>50), minimum percentage of coding genome (>75%) and taxonomic label.
- A total of 2,834 genomes pass this quality-control screening, and after a minimal manual curation of the corresponding taxonomy, they span 2 domains, 33 phyla, 66 classes, 130 orders, 278 families, 652 genera and 1,221 species for a total of 2,383 taxonomic clades.



Bacteria, Archaea, Eukaryotes and Viruses



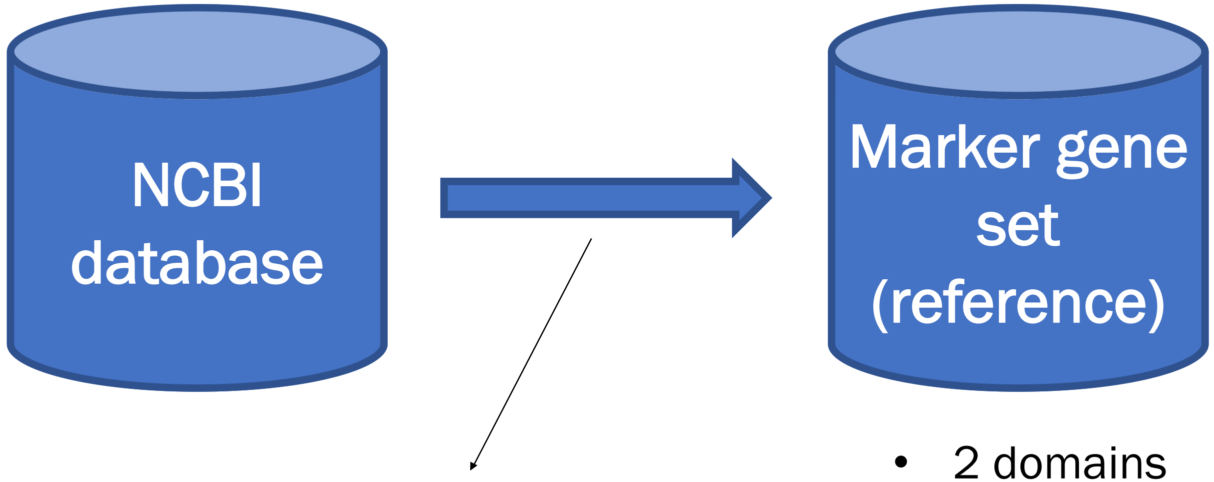
MethPhlAn1(2011)

- 2 domains
- 33 phyla
- 66 classes
- 130 orders
- 278 families
- 652 genera
- 1,221 species

MethPhlAn1(2011)

2-2. Find clade-specific marker genes.

Acquire reference



- Identify all core genes for all clades.
- Screen core genes for unique marker genes.
- Select most representative marker genes

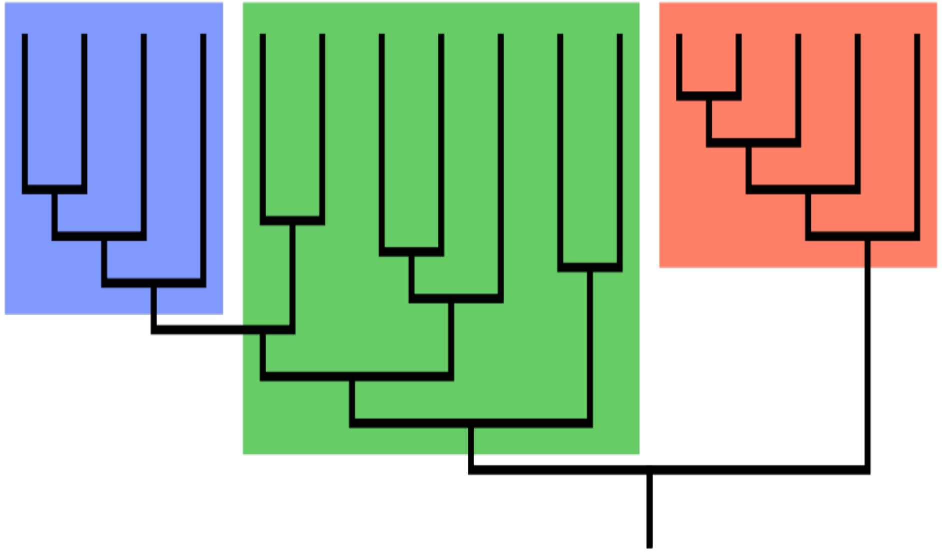
- 2 domains
- 33 phyla
- 66 classes
- 130 orders
- 278 families
- 652 genera
- 1,221 species

MetaPhlAn2(2015)

- ~ 1 million markers from > 7,500 species (*184±45 markers per species*)
- *Profiles all domains of life (Bacteria, viruses, Eukaryote, Archaea)*
- *Quasi-markers used to resolve ambiguity in post-processing*

What are clade-specific marker gene?

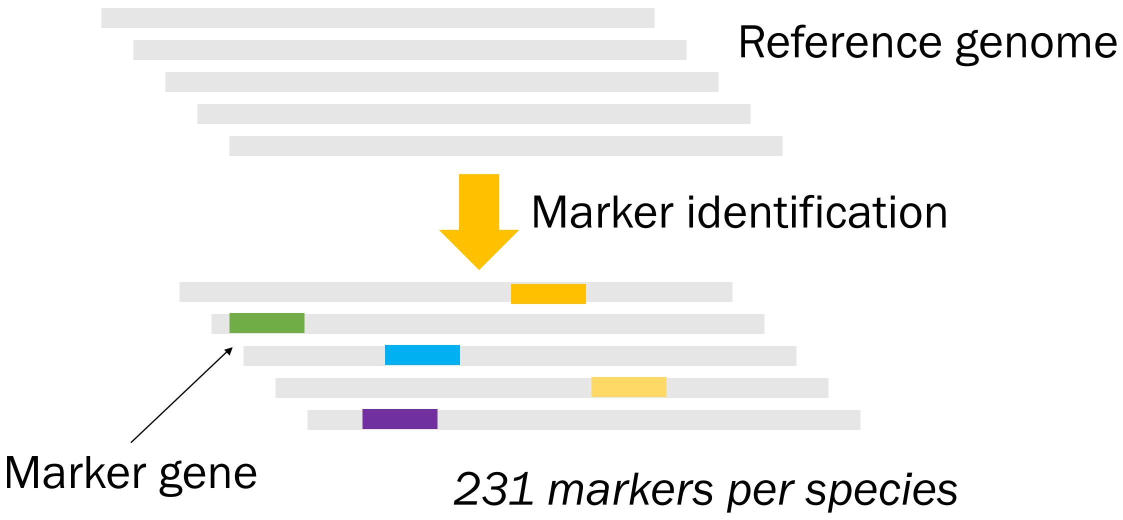
- MetaPhlAn estimates microbial relative abundances by mapping metagenomic reads against a catalog of clade-specific marker sequences currently spanning the bacterial and archaeal phylogenies.
- Clade
  - *Clades are groups of genomes (organisms) that can be as specific as species or as broad as phyla.*



https://en.wikipedia.org/wiki/Clade

What are clade-specific marker gene?

- Clade-specific markers are coding sequence that satisfy the conditions of
  - *1) being strongly conserved within the clade's genomes*
  - *2) not possessing substantial local similarity with any sequence outside the clade*



Reference genome

Marker identification

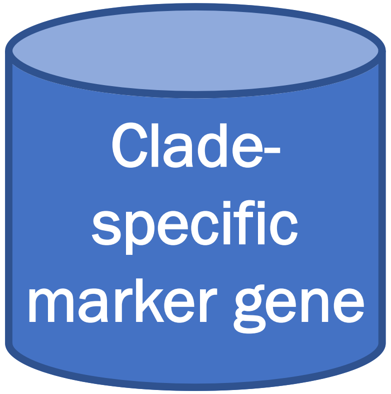
Marker gene

231 markers per species

Taxonomic levels	Number of different clades
Phyla	50
Classes	100
Orders	197
Families	481
Genera	1670
Species	7677
Strains	16903

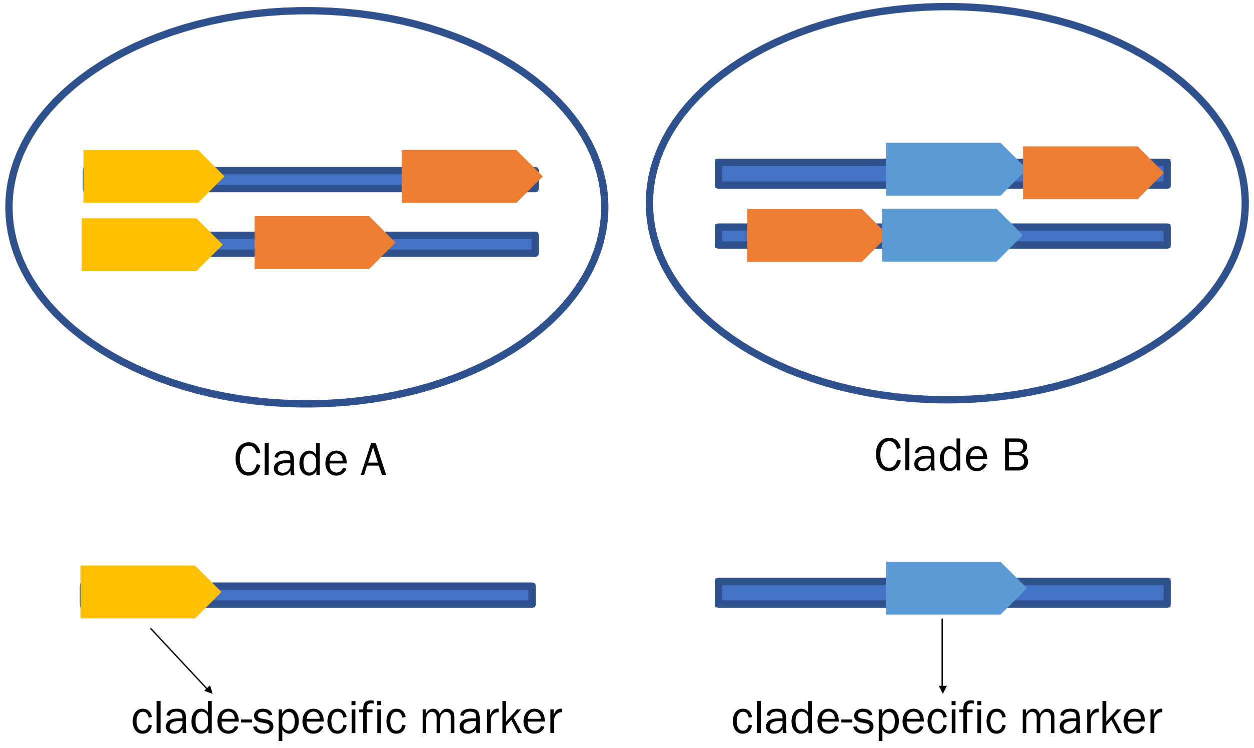
Number of distinct clades at different taxonomic levels considered in the MetaPhlAn2

2-2. Find clade-specific maker genes.



Clade-specific marker gene

MetaPhlAn2
~ 1 million markers from > 7,500 species
(*231 markers per species*)



Clade A

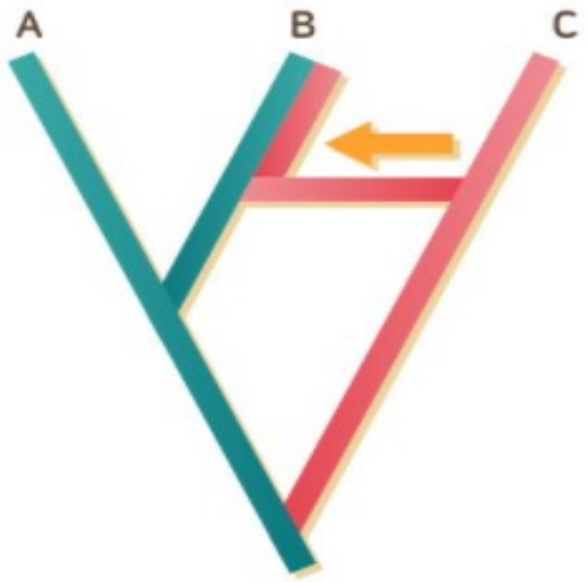
Clade B

clade-specific marker

clade-specific marker

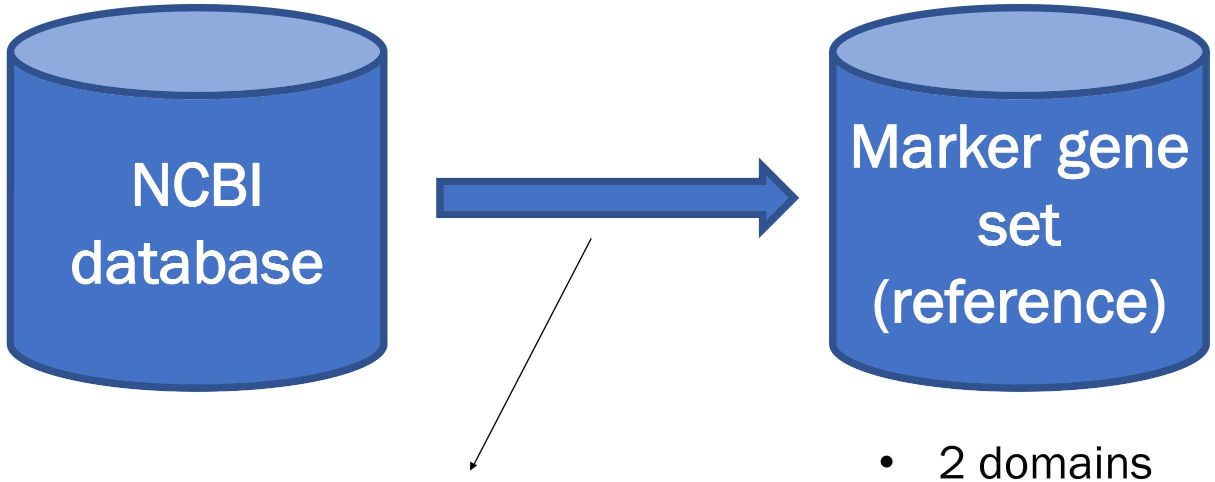
Introduction of the concept of quasi-markers

- Besides, for species with less than 200 markers, MetaPhlAn2 adopts additional quasi-marker sequences that are occasionally present in other genomes
  - *because of vertical conservation or horizontal transfer.*
- At profiling time, if no other markers of the potentially confounding species are detected, the corresponding quasi-local markers are used to improve the quality and accuracy of the profiling.
- Markers and quasi-markers coding sequences that unequivocally identify specific microbial clades at the species level or higher taxonomic levels
  - *markers : specific of the clade*
  - *quasi-markers : show a minimal number of sequence hits in genomes outside the clade*
- Marker and quasi-marker genes -> false positive and false negative rates -> allowing more comprehensive and accurate profiling.



2-2. Find clade-specific marker genes.

Acquire reference



- Identify all core genes for all clades.
- Screen core genes for unique marker genes.
- Select most representative marker genes

- 2 domains
- 33 phyla
- 66 classes
- 130 orders
- 278 families
- 652 genera
- 1,221 species

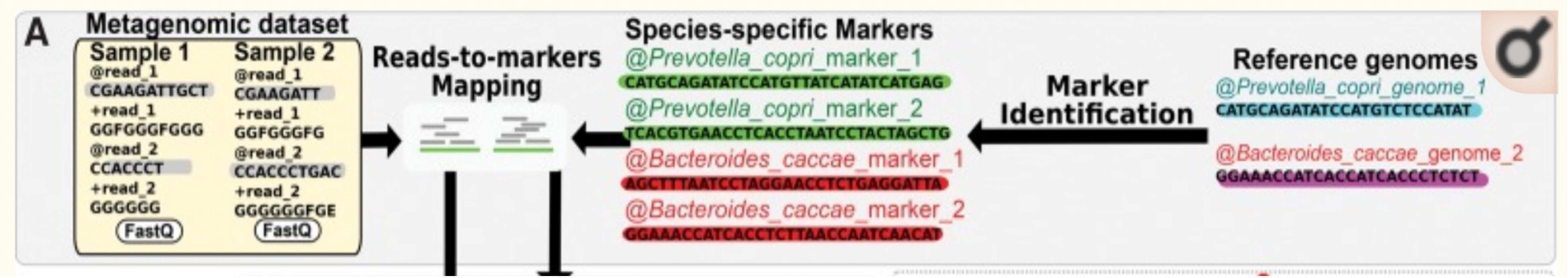
MetaPhlAn1(2011)

MetaPhlAn2(2015)

- ~ 1 million markers from > 7,500 species (*184±45 markers per species*)
- *Profiles all domains of life (Bacteria, viruses, Eukaryote, Archaea)*
- *Quasi-markers used to resolve ambiguity in post-processing*

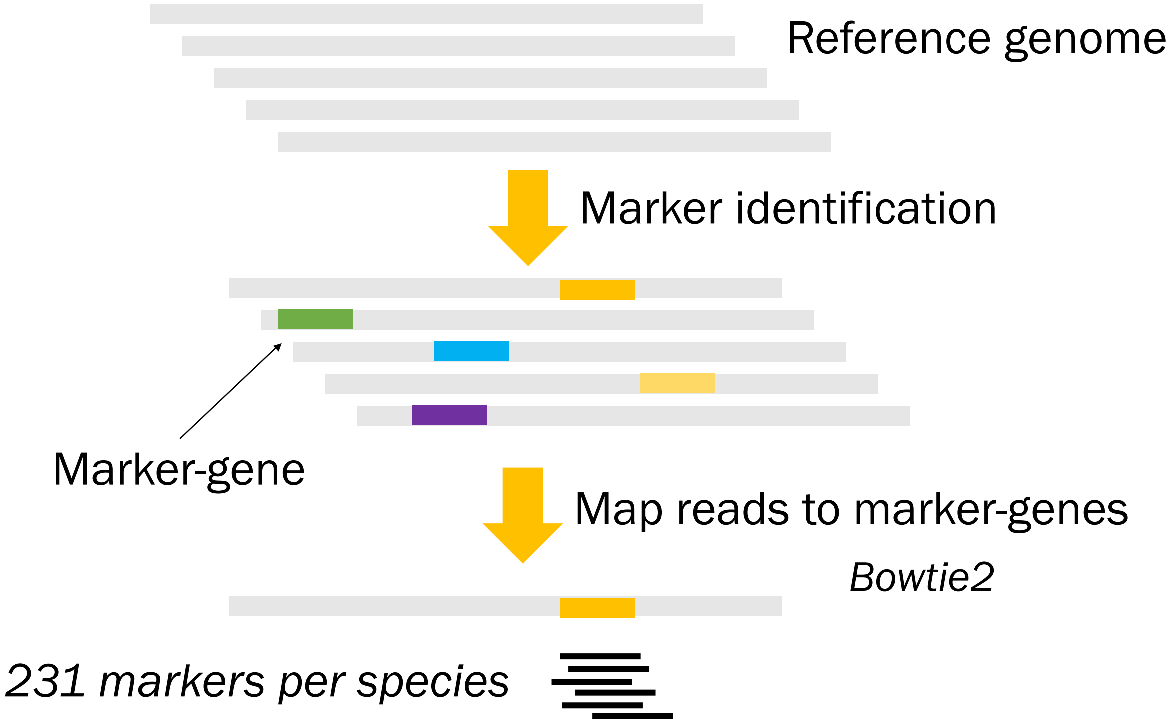
2-4. Mapping of metagenomes to the marker gene catalog

- The selection of the marker genes described above is relatively computationally intensive (typically requiring several CPU-days), but it needs to be performed only once when the set of reference genomes is modified, usually because of the addition of newly sequenced genomes.
- MetaPhlAn users do not need to perform this task, as we provide the most updated reference marker set


A
Metagenomic dataset
Sample 1
@read_1
CGAAGATTGCT
+read_1
GGFGGGFGGG
@read_2
CCACCCT
+read_2
GGGGGG
FastQ
Sample 2
@read_1
CGAAGATT
+read_1
GGFGGGFG
@read_2
CCACCCTGAC
+read_2
GGGGGGFGE
FastQ
Reads-to-markers Mapping
Species-specific Markers
@Prevotella_copri_marker_1
CATGCAGATATCCATGTTATCATATCATGAG
@Prevotella_copri_marker_2
TCACGTGAACCTCACCTAATCCTACTAGCTG
@Bacteroides_caccae_marker_1
AGCTTTAATCCTAGGAACCTCTGAGGATTA
@Bacteroides_caccae_marker_2
GGAAACCATCACCTCTTAACCAATCAACAT
Marker Identification
Reference genomes
@Prevotella_copri_genome_1
CATGCAGATATCCATGTCTCCATAT
@Bacteroides_caccae_genome_2
GGAAACCATCACCATCACCCTCTCT

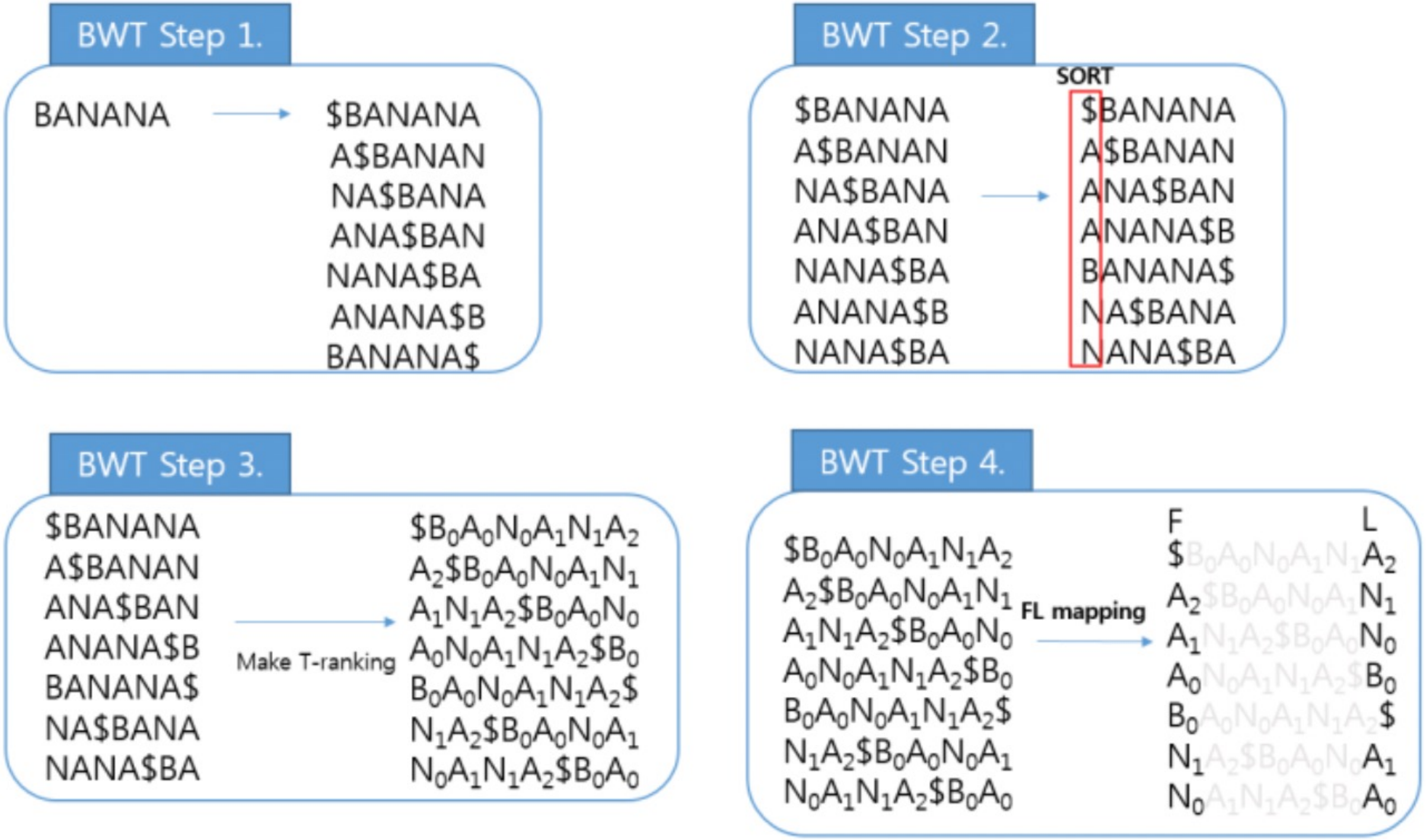
TRUONG, Duy Tin, et al. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research*, 2017, 27.4: 626-638.

2-4. Mapping of metagenomes to the marker gene catalog



- Each sample are mapped the markers using Bowtie2.
- The MetaPhlAn classifier compares each metagenomic read from a sample to this marker catalog to identify high-confidence matches.
- We used Metaphlan2, which attempts to find reads corresponding to clade-specific genes to assign the corresponding read to the target clade.

2-4. Mapping of metagenomes to the marker gene catalog



BWT Step 1.

BANANA → $BANANA, A$BANAN, NA$BANA, ANA$BAN, NANA$BA, ANANA$B, BANANA$

BWT Step 2.

$BANANA, A$BANAN, NA$BANA, ANA$BAN, NANA$BA, ANANA$B, NANA$BA → SORT → $BANANA, A$BANAN, ANA$BAN, ANANA$B, BANANA$, NA$BANA, NANA$BA

BWT Step 3.

$BANANA, A$BANAN, ANA$BAN, ANANA$B, BANANA$, NA$BANA, NANA$BA → Make T-ranking → $\$B_0A_0N_0A_1N_1A_2$, $A_2\$B_0A_0N_0A_1N_1$, $A_1N_1A_2\$B_0A_0N_0$, $A_0N_0A_1N_1A_2\$B_0$, $B_0A_0N_0A_1N_1A_2\$$, $N_1A_2\$B_0A_0N_0A_1$, $N_0A_1N_1A_2\$B_0A_0$

BWT Step 4.

$\$B_0A_0N_0A_1N_1A_2$, $A_2\$B_0A_0N_0A_1N_1$, $A_1N_1A_2\$B_0A_0N_0$, $A_0N_0A_1N_1A_2\$B_0$, $B_0A_0N_0A_1N_1A_2\$$, $N_1A_2\$B_0A_0N_0A_1$, $N_0A_1N_1A_2\$B_0A_0$ → FL mapping →

F		L
$	$B_0A_0N_0A_1N_1$	A_2
A_2	$\$B_0A_0N_0A_1$	N_1
A_1	$N_1A_2\$B_0A_0$	N_0
A_0	$N_0A_1N_1A_2\$$	B_0
B_0	$A_0N_0A_1N_1A_2$	$
N_1	$A_2\$B_0A_0N_0$	A_1
N_0	$A_1N_1A_2\$B_0$	A_0

- NGS sequence alignment tool : Bowtie2
- The algorithm of bowtie is Burrows-Wheeler Transform.

https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=ossiriand&logNo=220656153907

2-5. Estimation of organismal relative abundance.

- Calculation of the relative abundance of each taxonomic unit priority to markers
  - *Sum the total reads mapped to clade markers*
  - *Divide by marker's total length*
  - *Abundances in every clade-level sum up to 100%*
  - *Relative abundances are estimated by weighting read counts assigned using the direct method with the total nucleotide size of all the markers in the clade and normalizing by the sum of all directly estimated weighted read counts.*

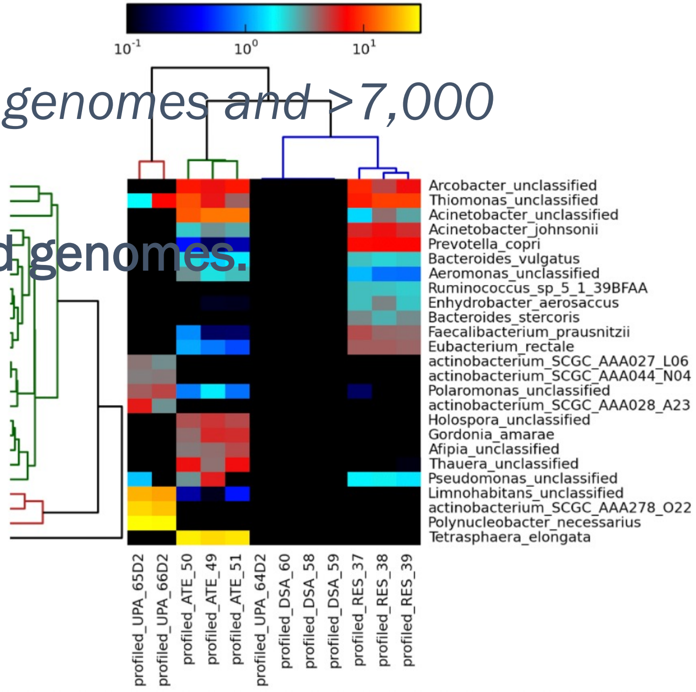
3.
MetaPhlAn2 is more enhanced metagenomic taxonomic profiling compared to MetaPhlAn1

3-1. Description of the main MethPhlAn2 additions compared to MetaPhlAn1

- 1. Profiling of all domains of life.
- 2. A 6-fold increase in the number of considered species.
- 3. Strain-level identification for organisms with sequenced genomes.
- 4. Introduction of the concept of quasi-markers, allowing more comprehensive and accurate profiling.
- 5. Integration of MetaPhlAn with post-processing and visualization tools.
- 6. Parallelization, Python3.

3-1. Description of the main MethPhlAn2 additions compared to MetaPhlAn1

- 1. **Profiling of all domains of life.**
  - *Bacteria and Archaea -> + viruses and Eukaryotic microbes (Fungi, Protozoa)*
- 2. **A 6-fold increase in the number of considered species.**
  - *Markers are now identified from >16,000 reference genomes and >7,000 unique species.*
- 3. **Strain-level identification for organisms with sequenced genomes.**
- 4. **Visualization, Parallelization, Python3.**

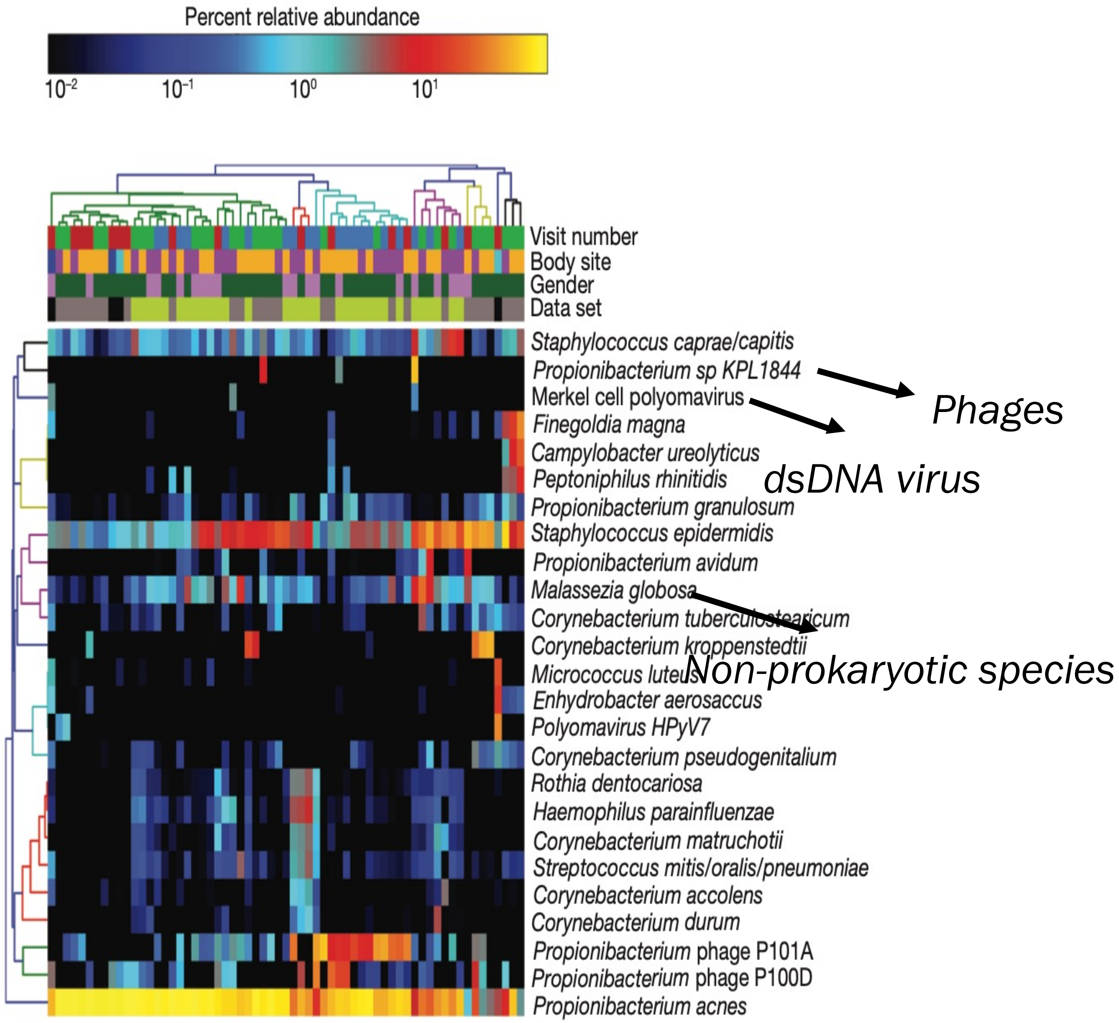


3-1. Description of the main MethPhlAn2 additions compared to MetaPhlAn1

- 4. **Introduction of the concept of quasi-markers, allowing more comprehensive and accurate profiling.**
  - *Marker and quasi-marker genes -> false positive and false negative rates*
- Markers and quasi-markers coding sequences that unequivocally identify specific microbial clades at the species level or higher taxonomic levels
  - *markers : specific of the clade*
  - *quasi-markers : show a minimal number of sequence hits in genomes outside the clade*
- Quasi-markers are added only if the number of (strict) markers is < 200

4.
Example and
Comparison with existing approaches

4-1. Example : MetaPhlAn2 characterization of all skin shotgun metagenomes



Phages

dsDNA virus

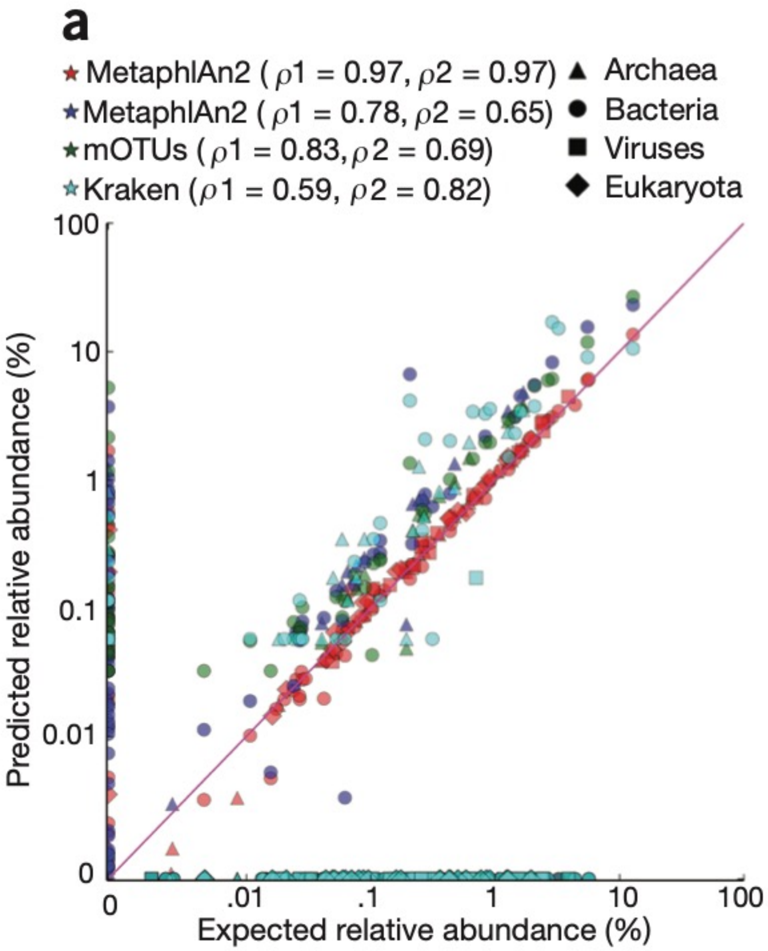
Non-prokaryotic species

- We applied MetaPhlAn2 to four elbow-skin samples that we sequenced from three subjects.
- Our data showed that *Propionibacterium acnes* and *Staphylococcus epidermidis* dom- inated these sites, in agreement with expected* genus-level results.

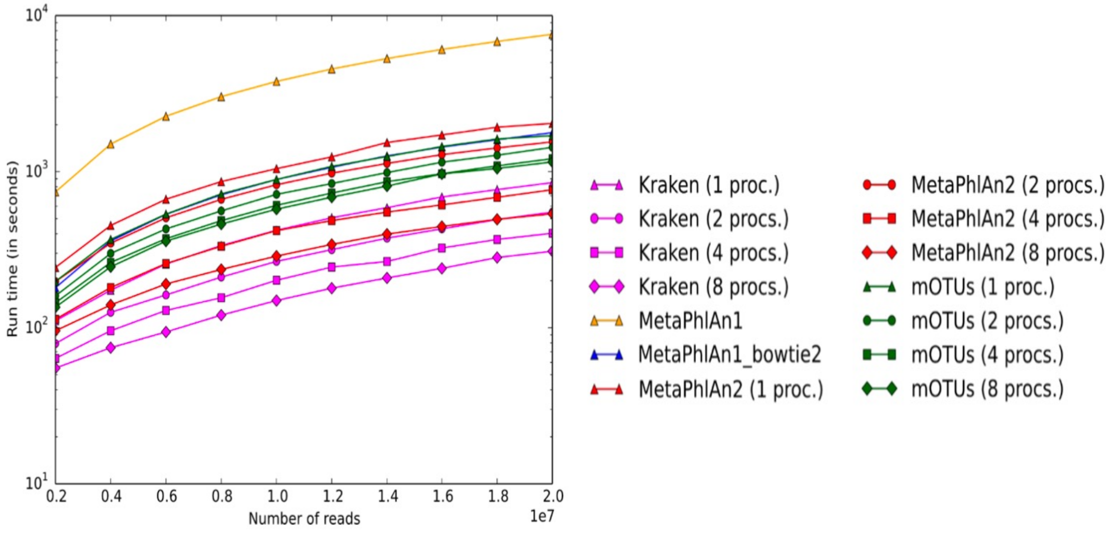
(*Grice, E.A. *et al. Science* **324**, 1190–1192 (2009)).

- Phages and double-stranded DNA viruses of the *Polyomavirus* genus were also consistently detected.

4-2. Evaluation taxonomic profilers using synthetic metagenomes



- MetaPhlAn2 proved more accurate than mOTU and Kraken.
- With the adoption of the BowTie2 fast mapper and support for parallelism, MetaPhlAn2 is more than ten times faster than MetaPhlAn1.



Supplementary Fig. 9. Run-time comparison between the validated methods. The original implementation of MetaPhlAn14 was based on Blastn[16], but we evaluate here also its extension based on BowTie23. MetaPhlAn2, mOTUS, and Kraken are evaluated at increasing number of processors (from 1 to 8)

Discussion

- Metagenomic shotgun sequencing data can identify microbes populating a microbial community and their proportions, but existing taxonomic profiling methods are inefficient for increasing large data sets.
- Shotgun metagenomic data are rapidly decreasing in cost to a per-sample level comparable to that of 16S gene survey.
- MetaPhlAn provides a further advantage over 16S rRNA based investigations.
  - *Species level*
  - *Statistical support(~10^8 reads per sample vs ~10^4 reads per sample)*
  - *Amplification step*
  - *Accuracy*
- MetaPhlAn is a method for characterizing the taxonomic profiles of whole-metagenome shotgun (WMS) samples that has been used successfully in large-scale microbial community studies.
- This work complements the original species-level profiling method with a system for eukaryotic and viral quantitation, strain-level identification and strain tracking.

5.
Google Co-lab Notebook

GOOGLE COLAB NOTEBOOK

- Below link is a google colab notebook link which will be used in metaphlan2 presentation in Advanced Bioinformatics 1 lecture.
  - *Link*

  https://colab.research.google.com/drive/1QzMMe8AogsBi7iuhhXVNbv1kK4jekBby?usp=sharing

Reference

- TRUONG, Duy Tin, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature methods*, 2015, 12.10: 902-903.
- SEGATA, Nicola, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 2012, 9.8: 811-814.
- TRUONG, Duy Tin, et al. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome research*, 2017, 27.4: 626-638.

THANK YOU
ANY QUESTIONS?