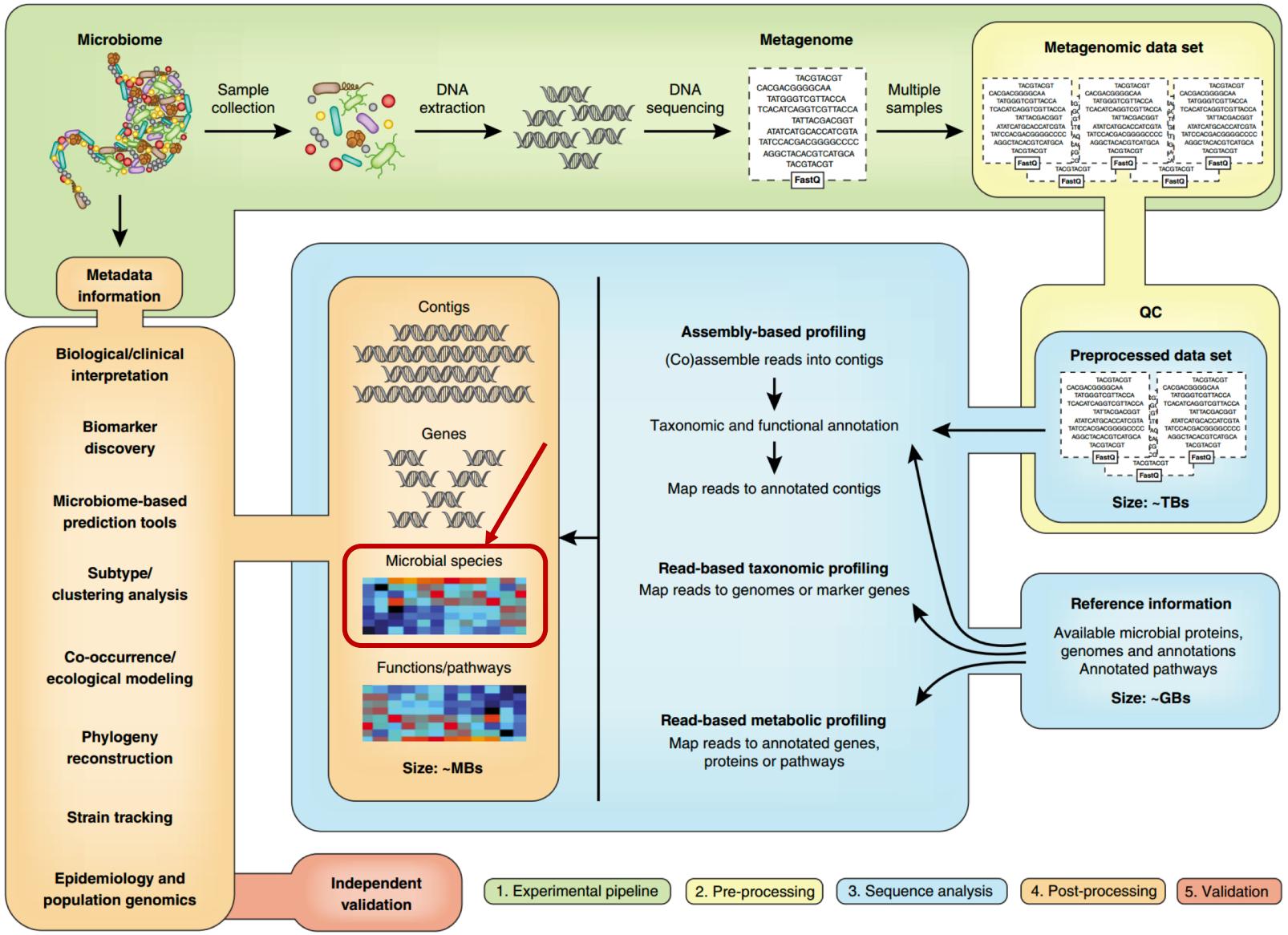
Computational methods for taxonomic profiling of metagenomes

Peng Ye

May-09-2019

Contents

Taxonomic profiling is a key step for most studies



C. Quince et al (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*

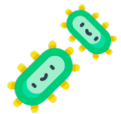
Cast

Species from different genera

A

1. 
2. 

B

1. 
2. 

C

1. 
2. 

Sentences with a subject–verb–object structure

A

1. We constructed and validated a random forest classifier with 28 lipidomic features that effectively discriminated T2D from NGT or prediabetes [1].
2. We constructed a classifier with lipidomic features discriminating T2D from NGT or prediabetes.

B

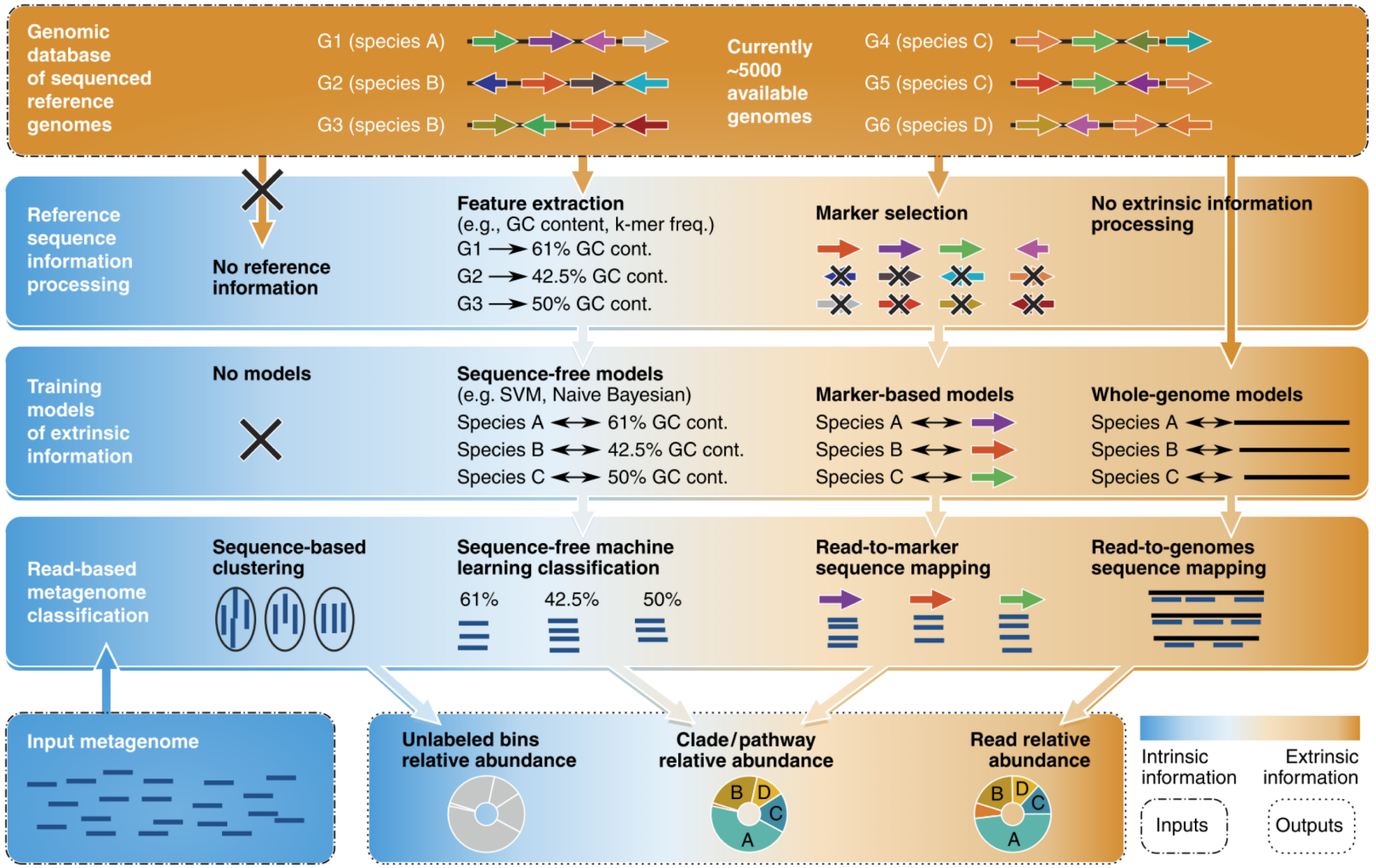
1. MetaPGN accepts genome or metagenome assemblies as input (query assemblies) and requires a reference genome for recruitment of the query assemblies and as the skeleton of the pangenome network [2].
2. MetaPGN accepts assemblies as input and requires a reference genome.

C

1. Combining sequence similarity-based matching and genetic features-based machine learning classification, we developed a novel scoring system that exhibits higher accuracy than current tools in predicting active prophages on the validation datasets [3].
2. Combining sequence matching and machine learning classification, we developed a scoring system for predicting active prophages.

[1] Zhong et al (2017). Lipidomic profiling reveals distinct differences in plasma lipid composition in healthy, prediabetic, and type 2 diabetic individuals. *GigaScience*
[2] Peng et al (2018). MetaPGN: a pipeline for construction and graphical visualization of annotated pangenome networks. *GigaScience*
[3] Song et al (2019). Prophage Hunter: an integrative hunting tool for active prophages. Nucleic Acids Research (Accepted manuscript)

1. Traditional methods for taxonomic profiling of metagenomes



N. Segata et al (2014). Computational meta'omics for microbial community studies. *Molecular Systems Biology*

1.1 No models

Sequence-based clustering

A rather simple community (50% vs. 50%)

We constructed and validated a random forest classifier with 28 lipidomic features that effectively discriminated T2D from NGT or prediabetes.

MetaPGN accepts assemblies as input and requires a reference genome.

Sequencing reads and assemblies

We constructed and validated
validated a random forest
a random forest classifier
with 28 lipidomic features
features that effectively
that effectively discriminated T2D from NGT or prediabetes

MetaPGN accepts assemblies as
assemblies as input and
input and requires a reference

(Annotation &) quantification

MicrobeGPS (genome-based) [1]
PhyloPhlAn (protein-based) [2]

We constructed and validated a random forest classifier

with 28 lipidomic features that effectively discriminated T2D from NGT or prediabetes

MetaPGN accepts assemblies as input and requires a reference

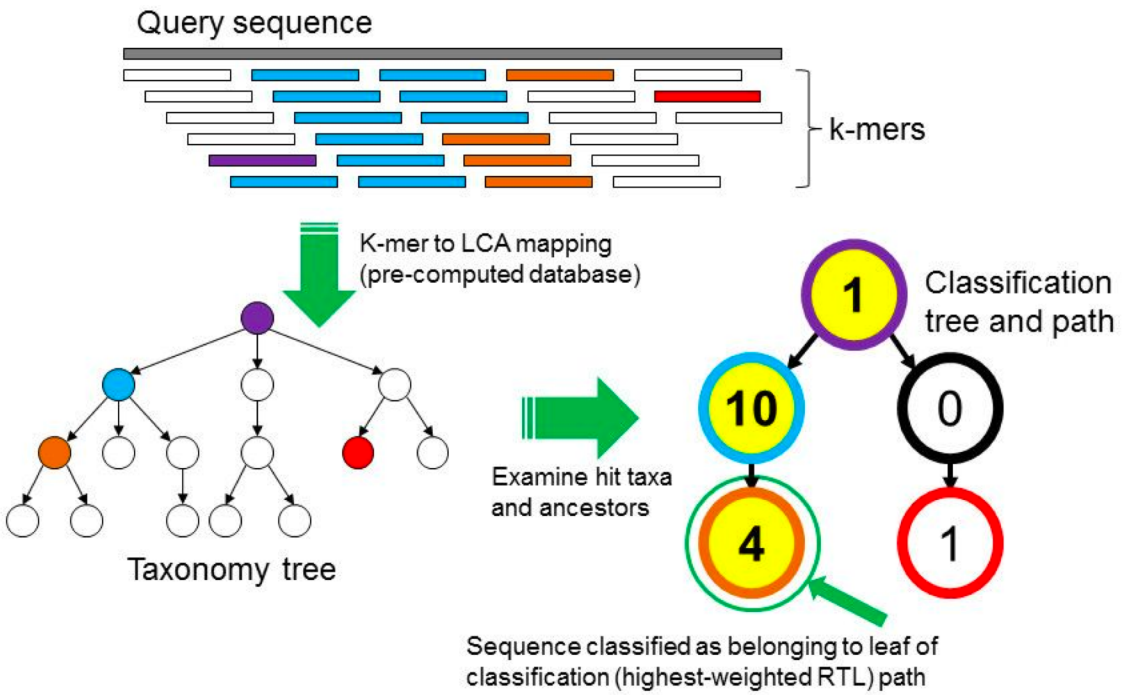
Assembly	Taxon (raw counts)	Taxon (normalized counts [counts/length])
3	6	3
3		
3	3	3

[1] Martin S. Lindner et al (2015). Metagenomic Profiling of Known and Unknown Microbes with MicrobeGPS. *PLOS ONE*
[2] N. Segata et al (2014). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communication*

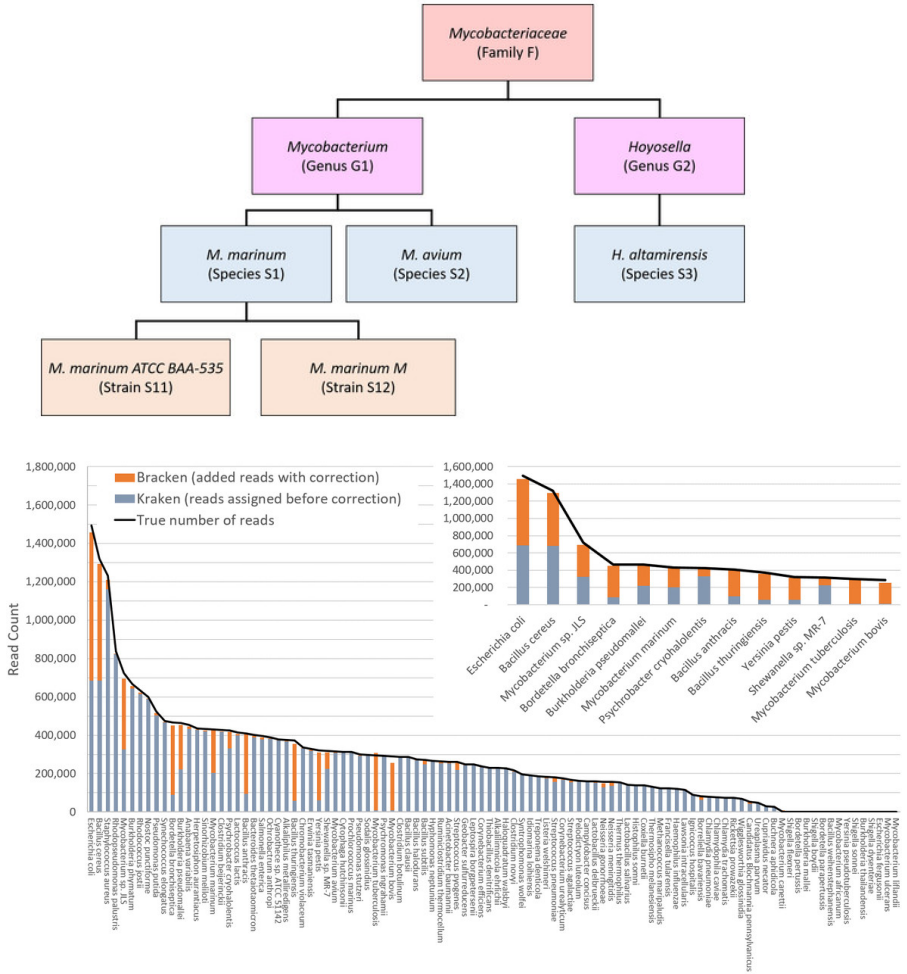
1.2 Sequence-free models

E.g., Kmer-level classification

- Kraken [1]



- Braken, with read reassignment [2]



[1] Derrick E Wood and Steven L Salzberg (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*
[2] Jennifer Lu et al (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ*

1.3 Marker-based models

1.3.1 Single marker-based (e.g. 16S rDNA v4 sequence, ITS2 sequence, RdRp sequence)

Verbs in the sentences (intersection)

A

1. We constructed and validated a random forest classifier with 28 lipidomic features that effectively discriminated T2D from NGT or prediabetes.
2. We constructed a classifier with lipidomic features discriminating T2D from NGT or prediabetes.

B

1. MetaPGN accepts genome or metagenome assemblies as input (query assemblies) and requires a reference genome for recruitment of the query assemblies and as the skeleton of the pangenome network.
2. MetaPGN accepts assemblies as input and requires a reference genome.

C

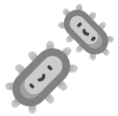
1. Combining sequence similarity-based matching and genetic features-based machine learning classification, we developed a novel scoring system that exhibits higher accuracy than current tools in predicting active prophages on the validation datasets.
2. Combining sequence matching and machine learning classification, we developed a scoring system for predicting active prophages.

Profiling result

A

1.  Species-level
2.  Species-level

B

 Genus-level

C

 Genus-level

1.3 Marker-based models

1.3.2 Single-copy phylogenetic marker genes-based (e.g., 40 genes [1-2])

Verb–object phrases in the sentences (intersection)

A

1. We constructed and validated a random forest classifier with 28 lipidomic features that effectively discriminated T2D from NGT or prediabetes.
2. We constructed a classifier with lipidomic features discriminating T2D from NGT or prediabetes.

B

1. MetaPGN accepts genome or metagenome assemblies as input (query assemblies) and requires a reference genome for recruitment of the query assemblies and as the skeleton of the pangenome network.
2. MetaPGN accepts assemblies as input and requires a reference genome.

C

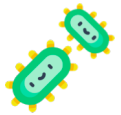
1. Combining sequence similarity-based matching and genetic features-based machine learning classification, we developed a novel scoring system that exhibits higher accuracy than current tools in predicting active prophages on the validation datasets.
2. Combining sequence matching and machine learning classification, we developed a scoring system for predicting active prophages.

Profiling result

A

1.  Species-level
2.  Species-level

B

1.  Species-level
2.  Species-level

C

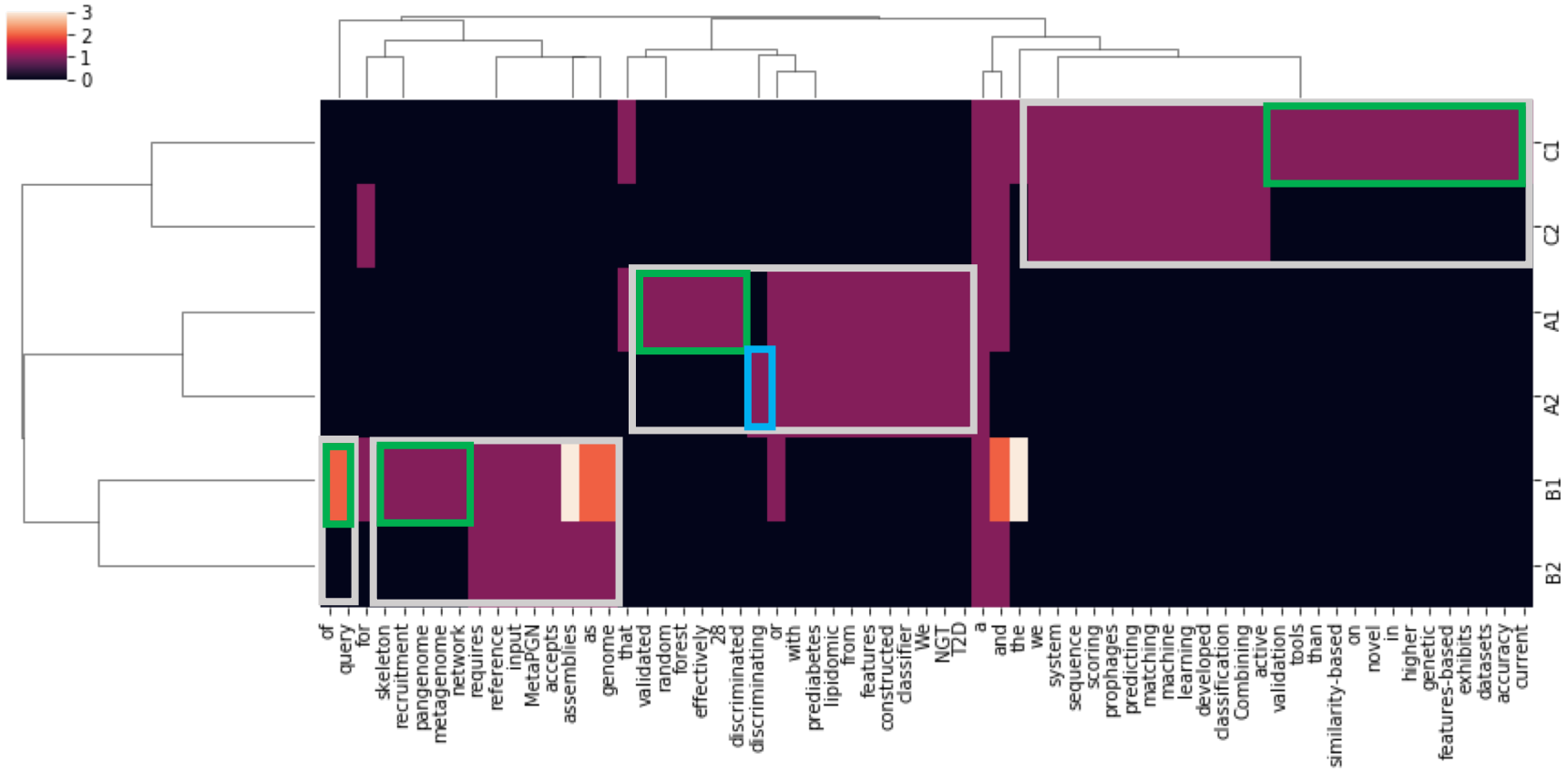
 Genus-level

[1] Daniel R Mende. Lindner et al (2013). Accurate and universal delineation of prokaryotic species. *Nature Methods*
[2] Shinichi Sunagawa et al (2013). Metagenomic species profiling using universal phylogenetic marker genes *Nature Methods*

1.3 Marker-based models

1.3.3 Unique clade-specific marker genes-based (~1M markers from >7,500 species [1-2])

Copy number of words in the sentences (subtraction)



Profiling result

Species-level

Genus-level

Species-level

Species-level

Species-level

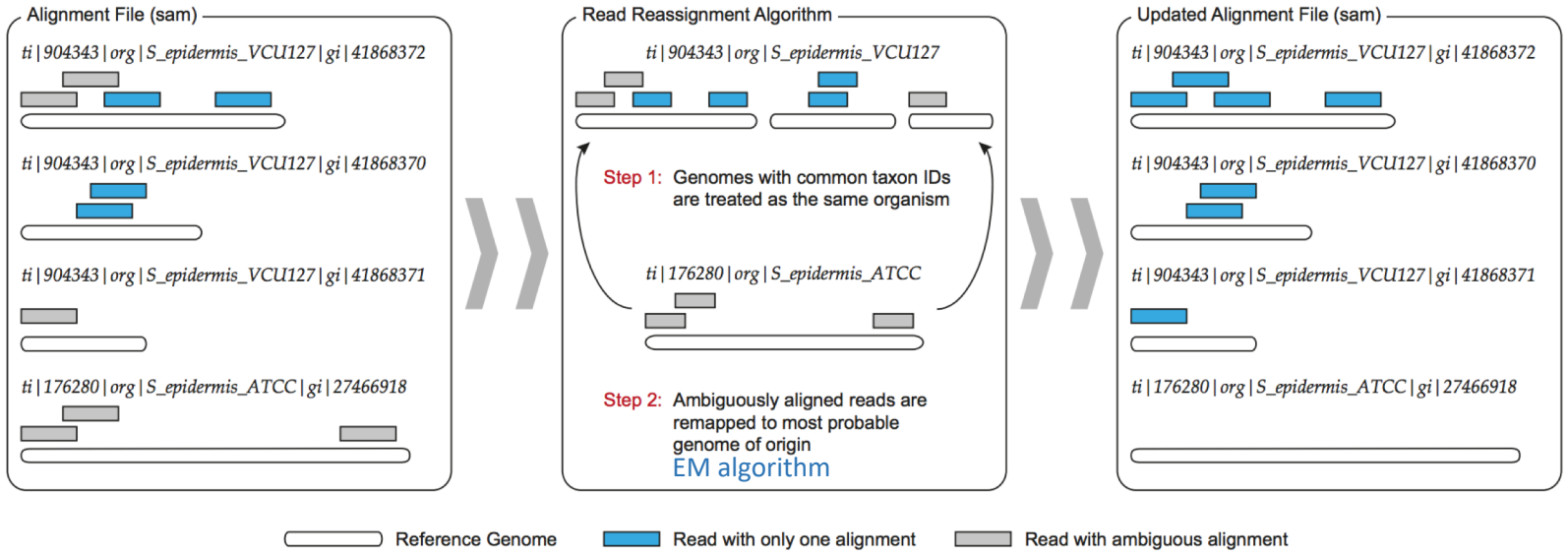
Genus-level

[1] N. Segata et al (2013). metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*
[2] Duy Tin Truong et al (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*

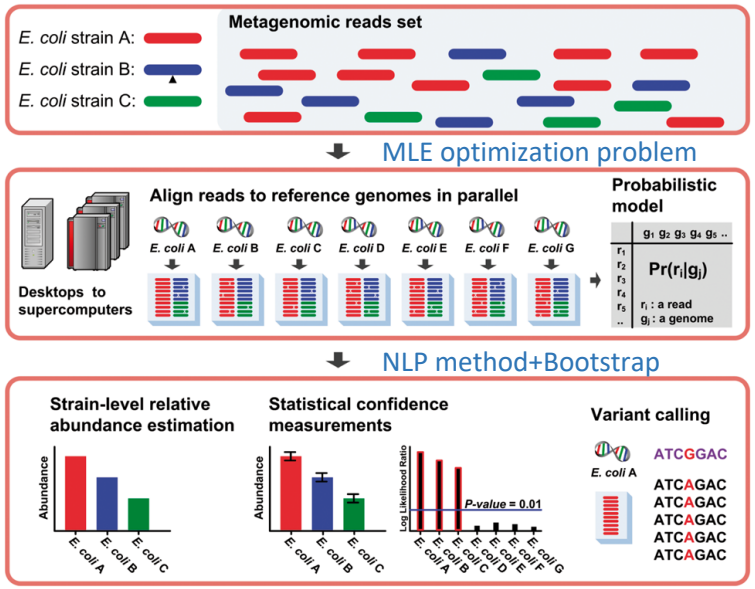
1.4 Whole-genome models

Read-level classification

- With read reassignment [1-2]

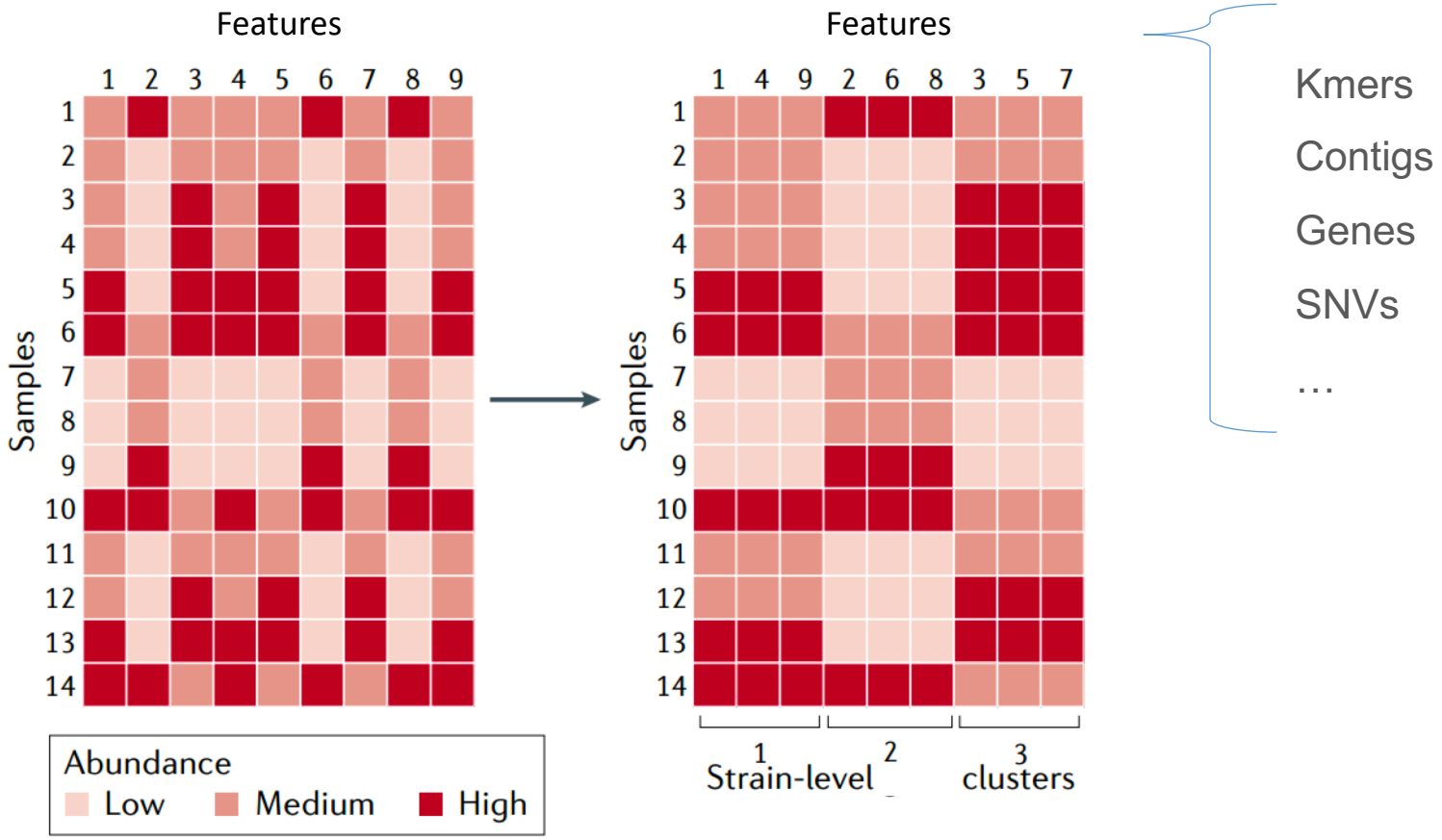


- With confidence interval estimation [3]



[1] Owen E. Francis et al (2013). Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Research*
[2] Hong et al (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*
[3] Tae-Hyuk Ahn et al (2015). Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics.*

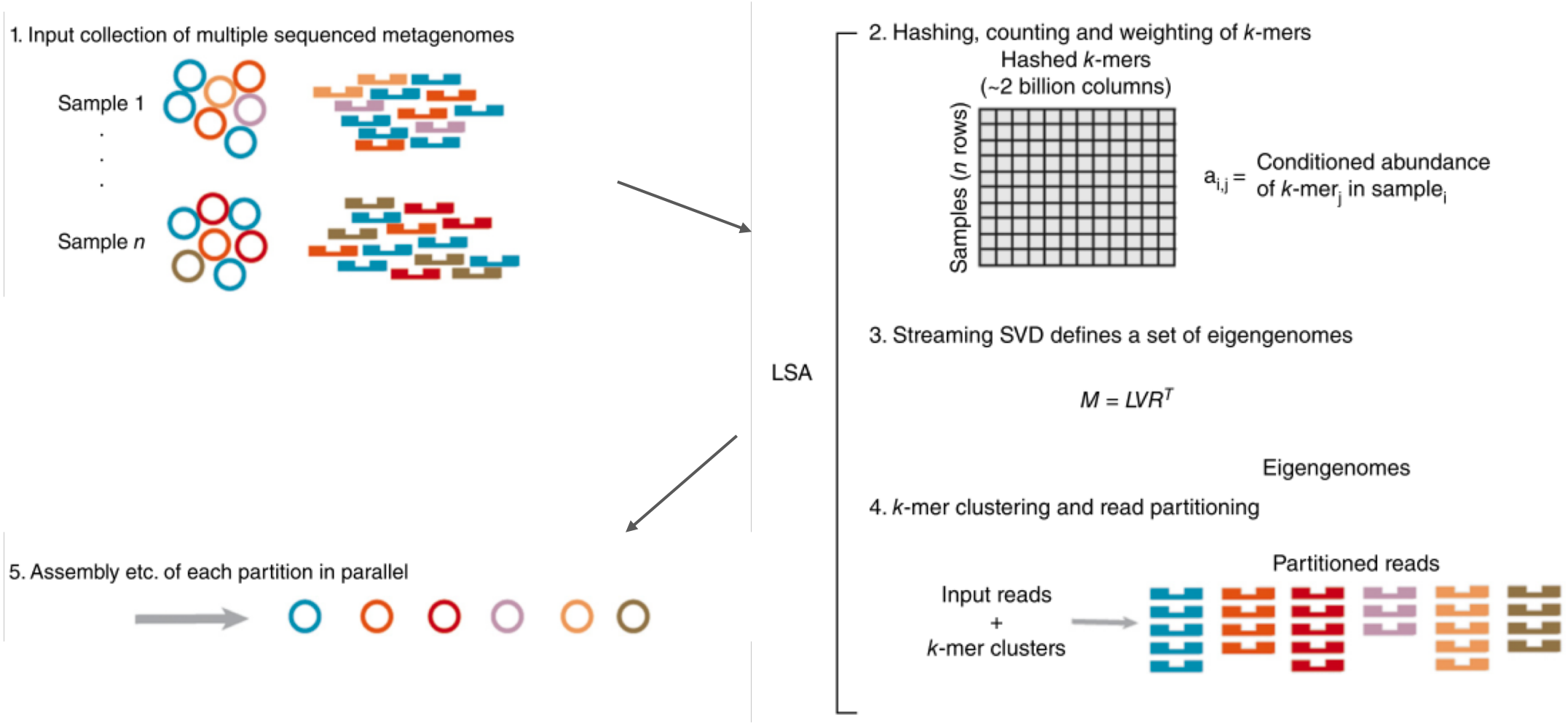
2. Advanced methods for taxonomic profiling of metagenomes



Adapted from: Jun Wang and Huijue Jia (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology*

2.1 Kmer-based model

Eigengenomes-based binning

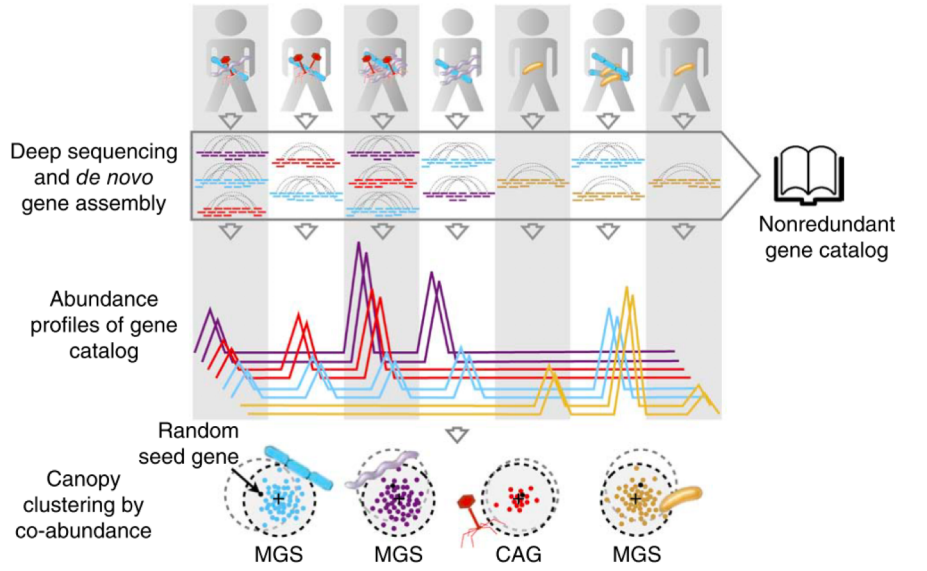


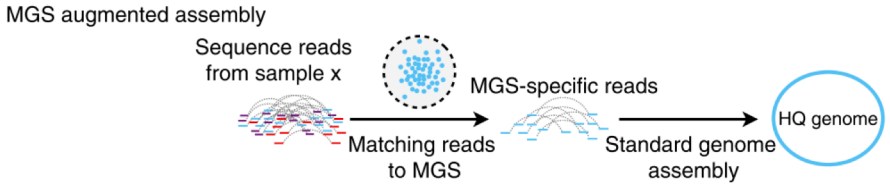
B. Cleary et al (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology*

2.2 Gene abundance-based models

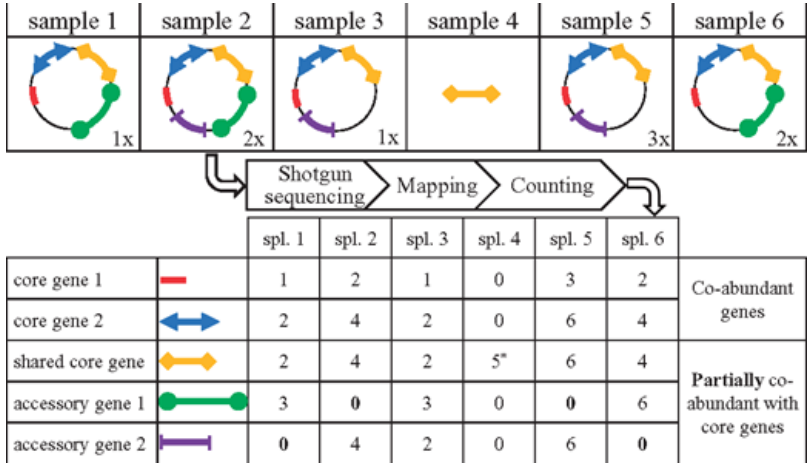
Co-abundance binning

- Co-abundance gene groups (CAGs)
  -> Metagenomic species (MGSs) [1]

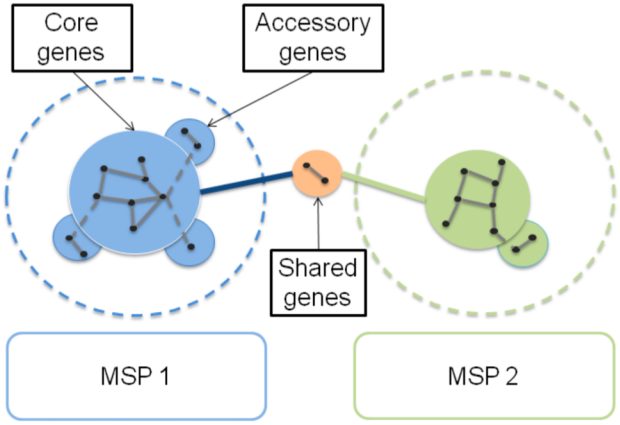




- Metagenomic Species Pan-genomes (MSPs) [2]



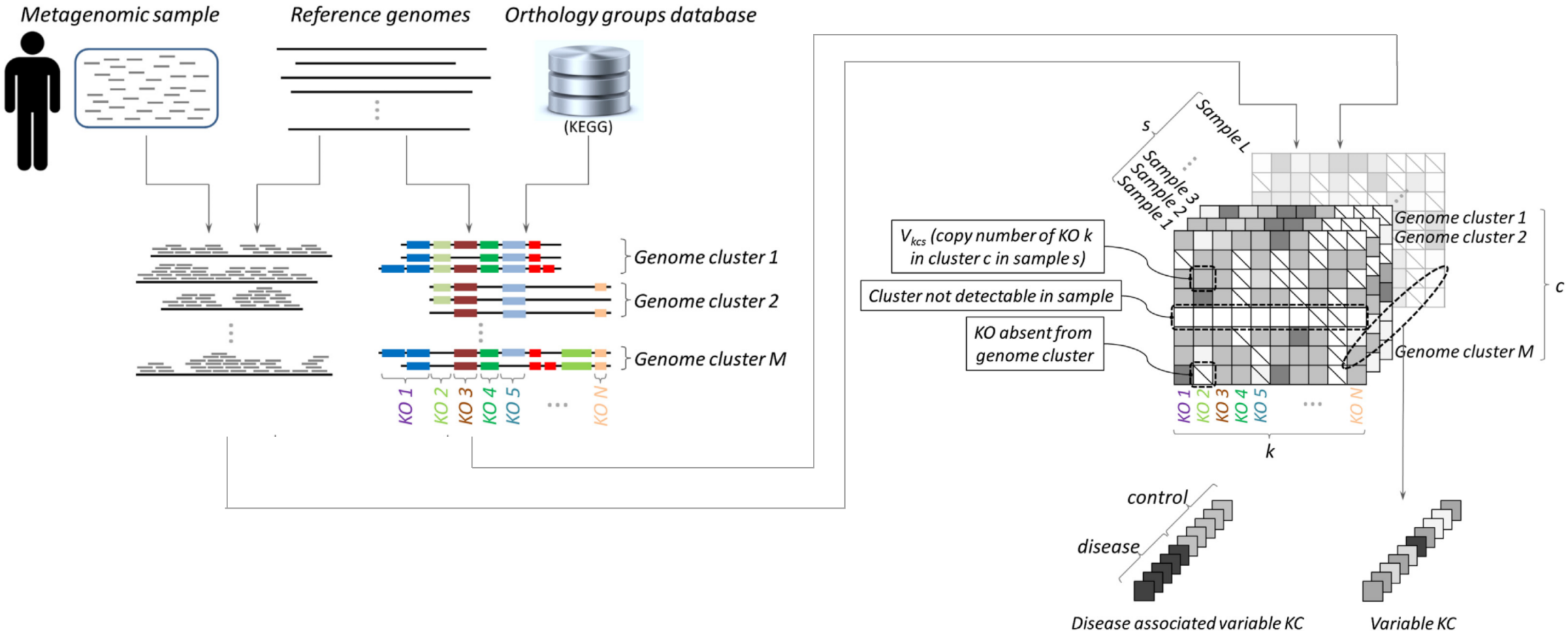
		spl. 1	spl. 2	spl. 3	spl. 4	spl. 5	spl. 6	
core gene 1		1	2	1	0	3	2	Co-abundant genes
core gene 2		2	4	2	0	6	4	
shared core gene		2	4	2	5*	6	4	**Partially** co-abundant with core genes
accessory gene 1		3	**0**	3	0	**0**	6	
accessory gene 2		**0**	4	2	0	6	**0**	



[1] H. Nielson et al (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature Biotechnology
[2] F. plaza onate et al (2019). MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*

2.3 CNV-based model

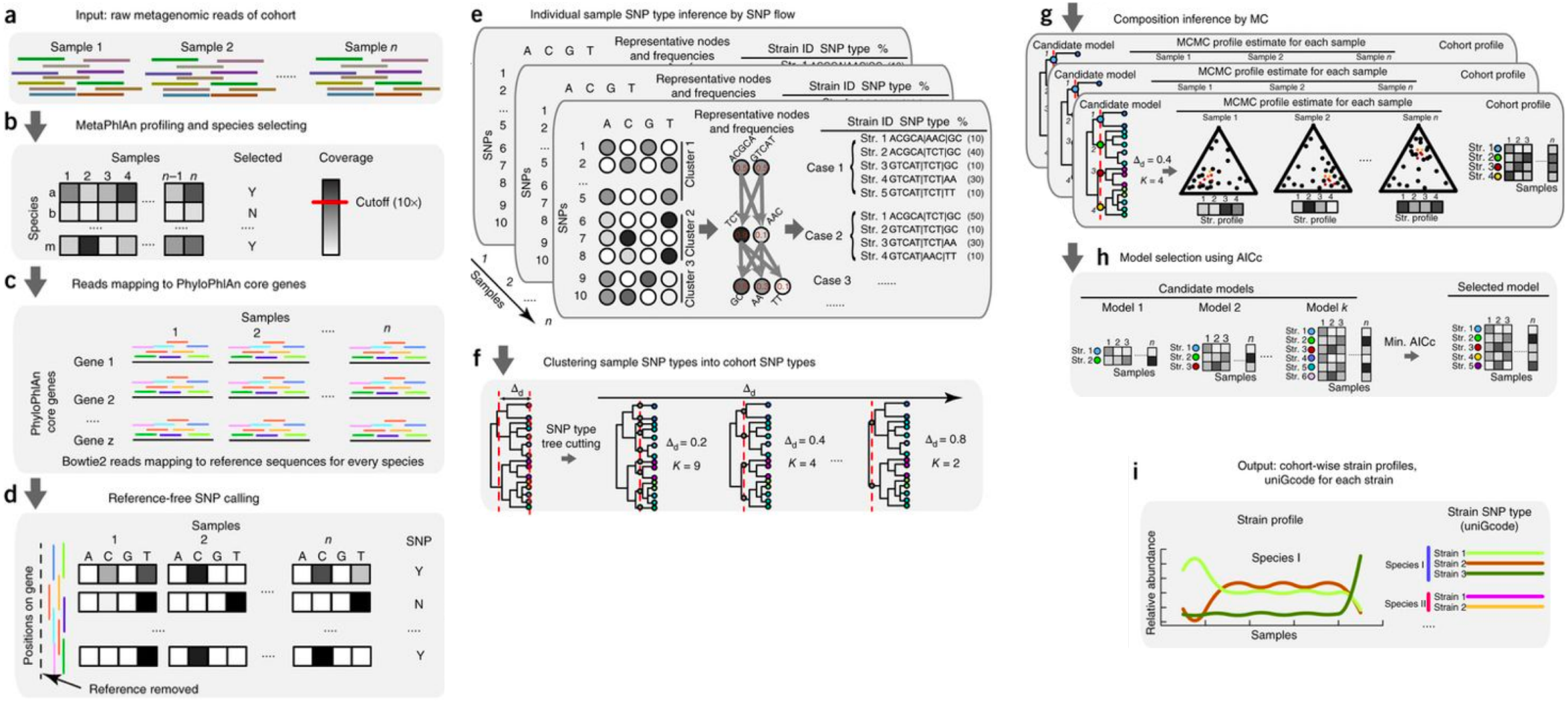
CNV-based clustering



S. Greenblum et al (2015). Extensive strain-level copy-number variation across human gut microbiome species. *Cell*

2.4 SNV-based model

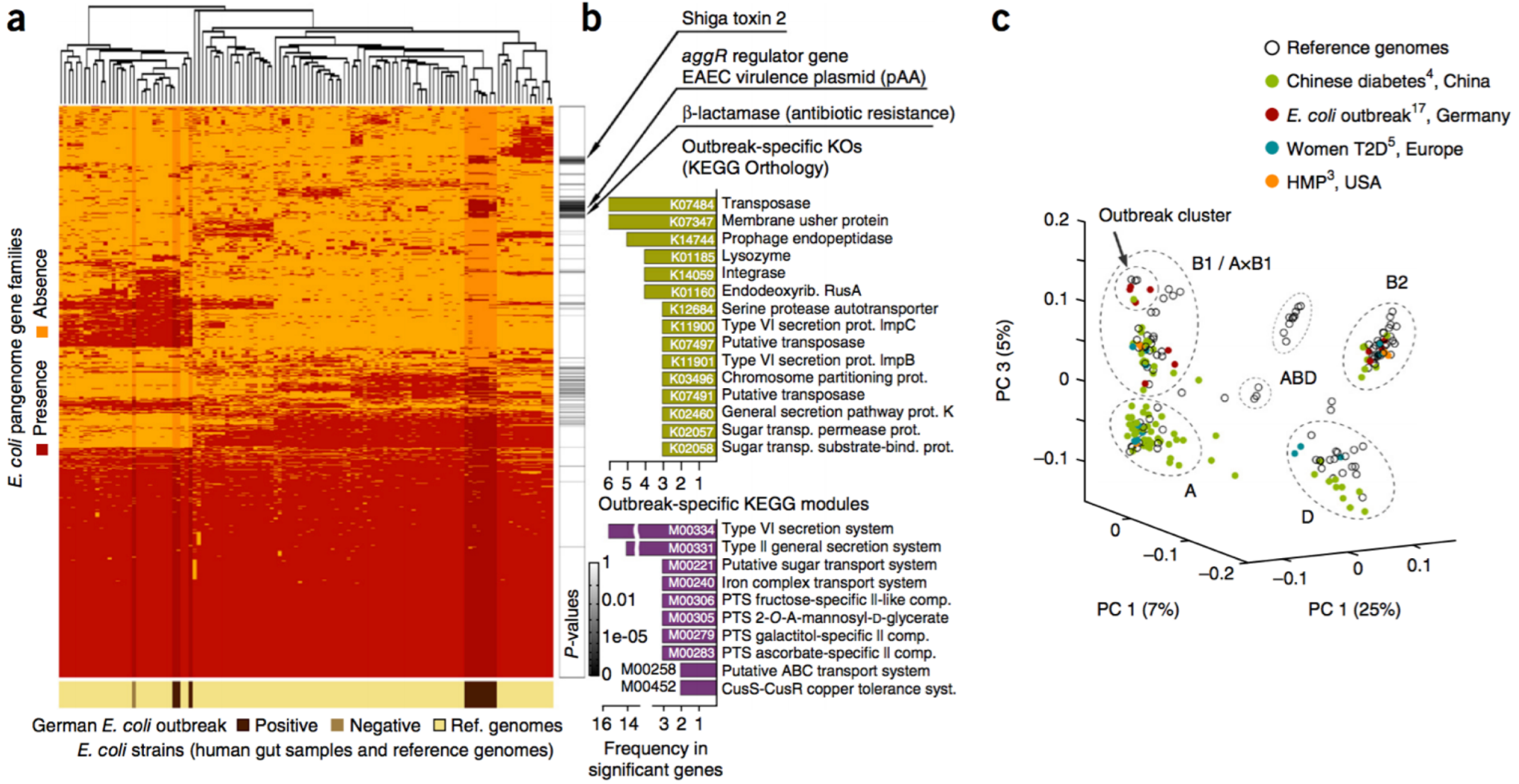
SNV frequency-based clustering



C. Luo et al (2015). ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology*

3. Analysis of metagenomes with a targeted pangenome

Pangenome-based phylogenomic analysis

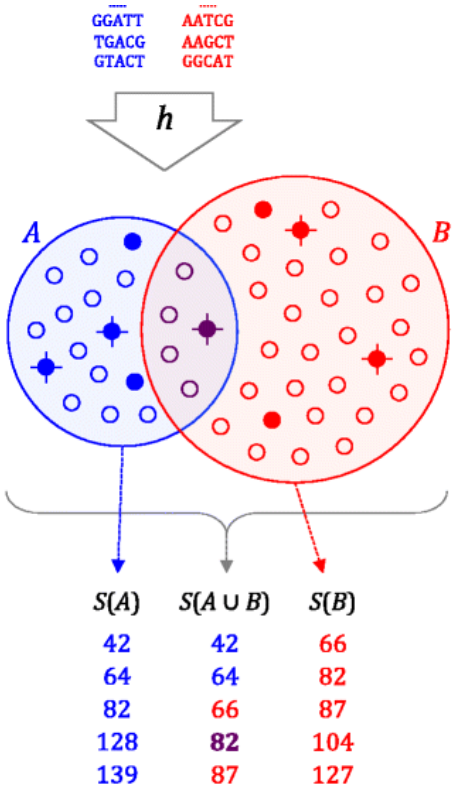


Matthias Scholz et al (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*

4. Comparison of metagenomes without taxonomic annotation

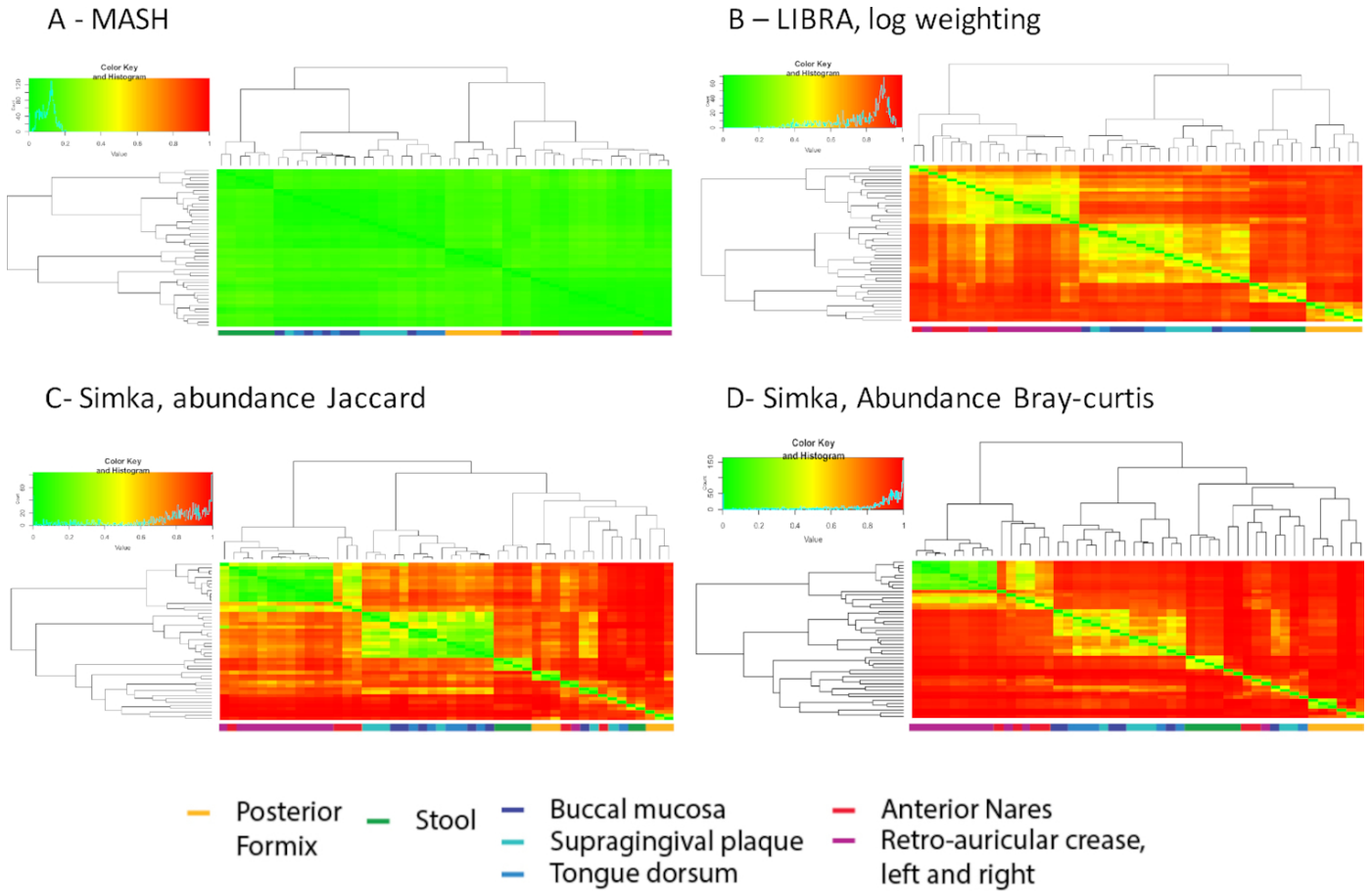
Kmer-based clustering of metagenomic samples

Concept (an example) [1]



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Application in HMP 16S rRNA datasets [2]



[1] Brian D. Ondov et al (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*
[2] Illyoung Choi et al (2018). Libra: scalable k-mer–based tool for massive all-vs-all metagenome comparisons. *GigaScience*

Summary

- Traditional methods
  - Profile microbial communities sample by sample
  - Mainly include
    - Methods without models
    - Sequence-free models
    - Marker-based models
    - Whole-genome models
- Advanced methods
  - Utilize bioinformation in multiple samples
  - Mainly include
    - Kmer-based models
    - Gene abundance-based models
    - CNV-based models
    - SNV-based models
- We may
  - Consider pangenome-based analysis if there are targets
  - Directly compare metagenomes without taxonomic annotation

Thanks!

Any questions you may have?

pengye21@163.com
https://github.com/peng-ye/Intro-taxonomic_profiling_of_metagenomes.git