

Metagenomics

Bioinformatics Applications (PLPTH813)

Sanzhen Liu

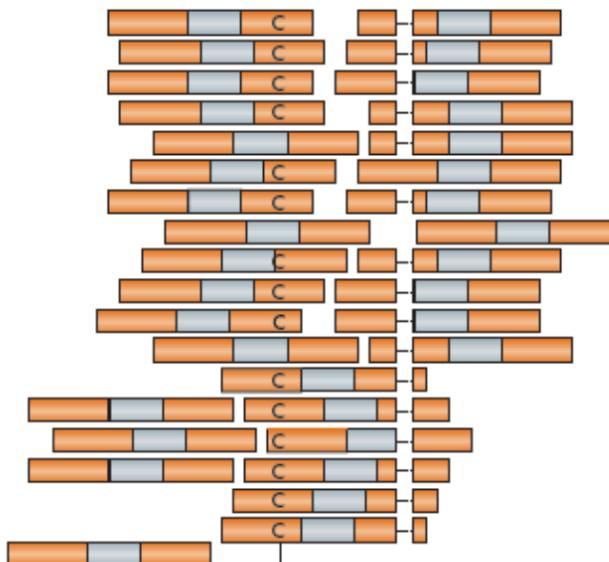
4/16/2019

Variants in sequencing reads

Reference sequence

Chr 1

A



Point mutation

Indel

Homozygous
deletion

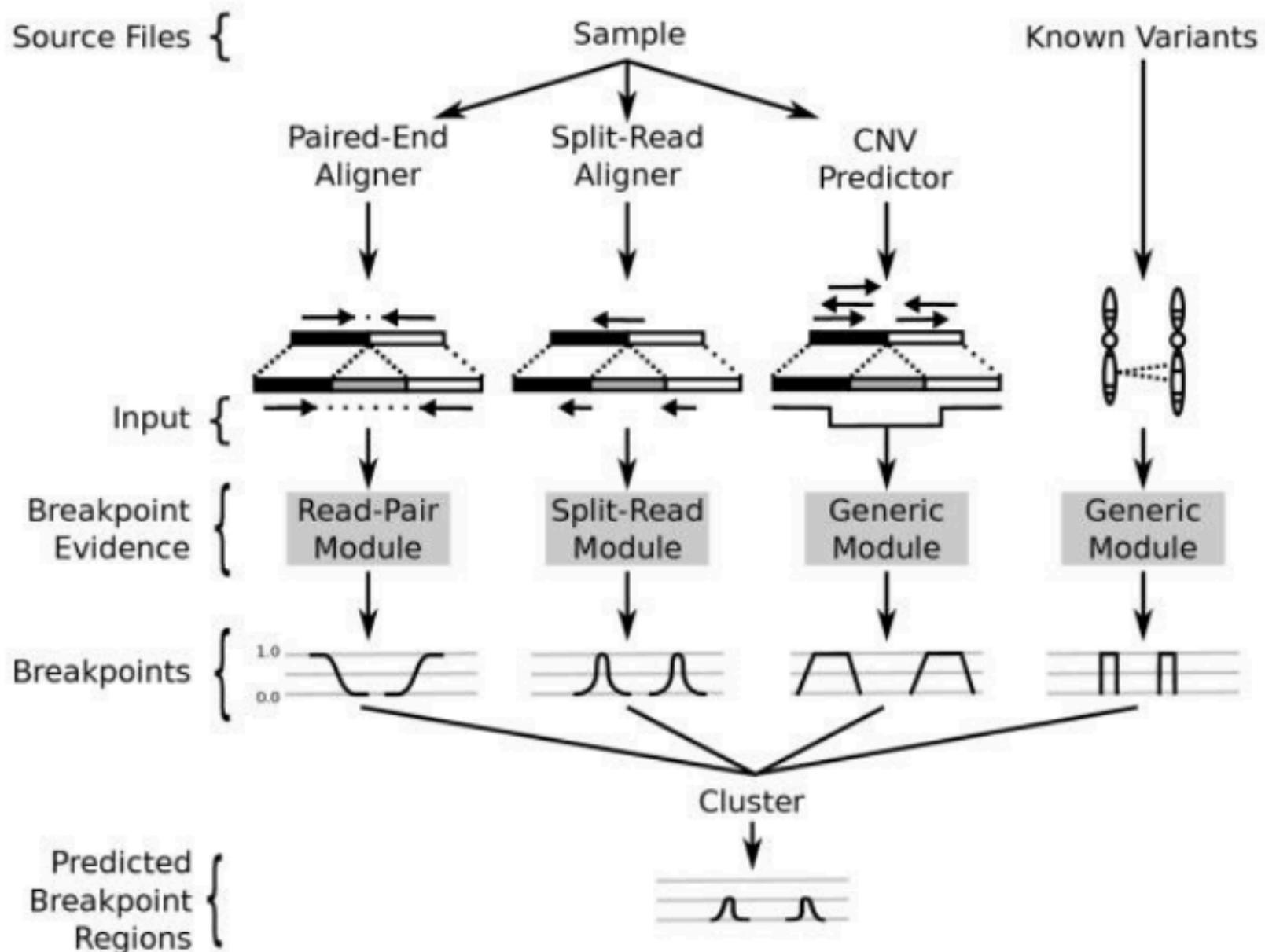
Hemizygous
deletion

Gain

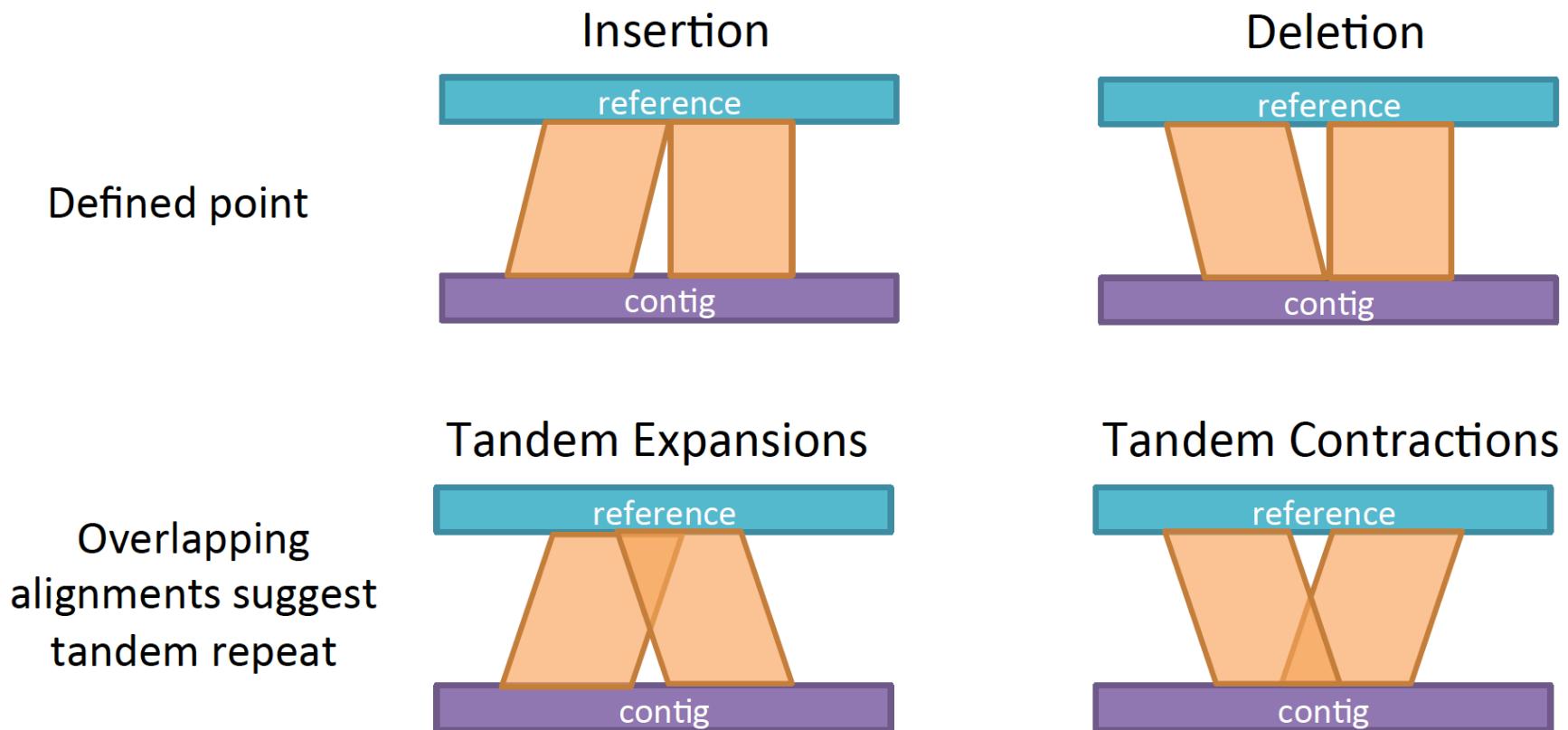
Translocation
breakpoint

Copy number alterations

LUMPY: a integrative framework for SV discovery



Genome assembly and genome-wide SV discovery



Outline

- Introduction of metagenomics
- Marker gene analysis (rRNA)
- Whole genome analysis
- Metatranscriptome analysis

Terms

- **metagenomics**: The application of high-throughput DNA sequencing to profile the genomic composition of a microbial community in a culture-independent manner.
- **microbiomes**: The community composition, biomolecular repertoire and ecology of microorganisms inhabiting particular environments.
- **microbiota**: The collection of microorganisms (of all types: bacteria, archaea, viruses and eukaryotes) inhabiting a particular environment.

metagenomics

Advances in DNA sequencing have enabled culture-independent profiling of microbial community membership and function - the field of metagenomics. (e.g., a study indicated that cultivation methods found <1% of the bacterial & archaeal species in samples.)

Firmicutes and Bacteroidetes phyla². Culture-independent, genomic approaches have transformed our understanding of the role of the human microbiome in health and many diseases¹. However, owing to the prevailing perception that our indigenous bacteria are largely recalcitrant to culture, many of their functions and phenotypes remain unknown³. Here we describe a novel workflow based on targeted phenotypic culturing linked to large-scale whole-genome sequencing, phylogenetic analysis and computational modelling that demonstrates that a substantial proportion of the intestinal bacteria are culturable. Applying this approach to healthy individuals,

Marker gene analysis

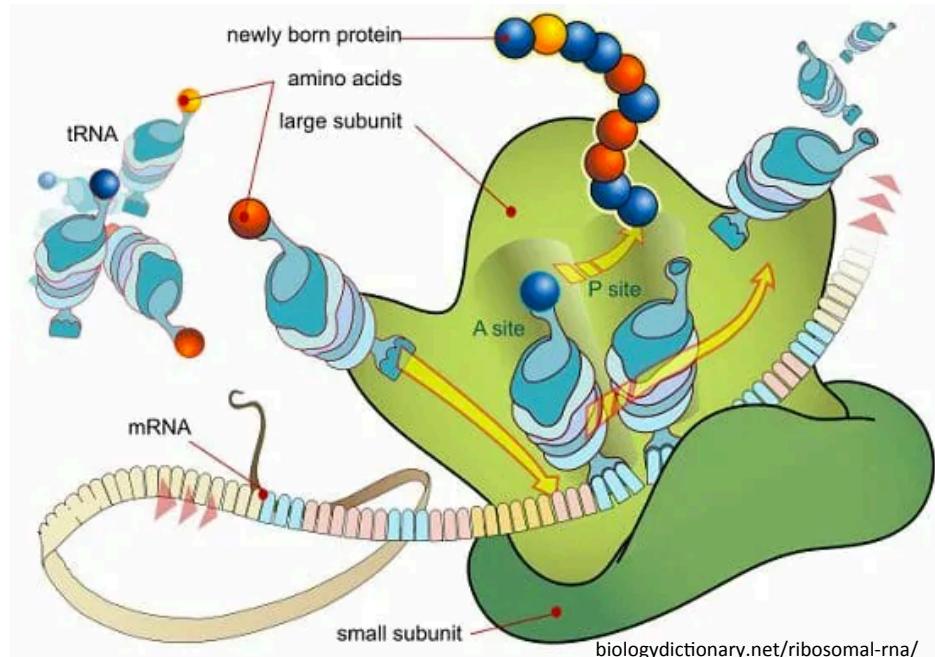
Marker gene methods are fast and low-cost for obtaining a *low-resolution* view of a microbial community.

Marker gene sequencing uses ***primers*** that target a **specific region of interest** to determine microbial community of a sample.



ribosomal RNA

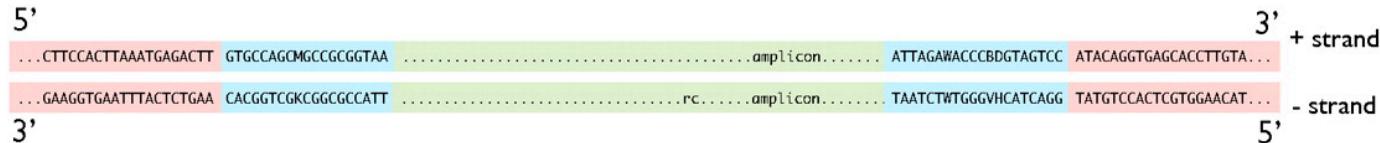
- Conserved sequences within a species
- Generally divergent between species
- Relative short (e.g., 1.5kb 16S rRNA gene)



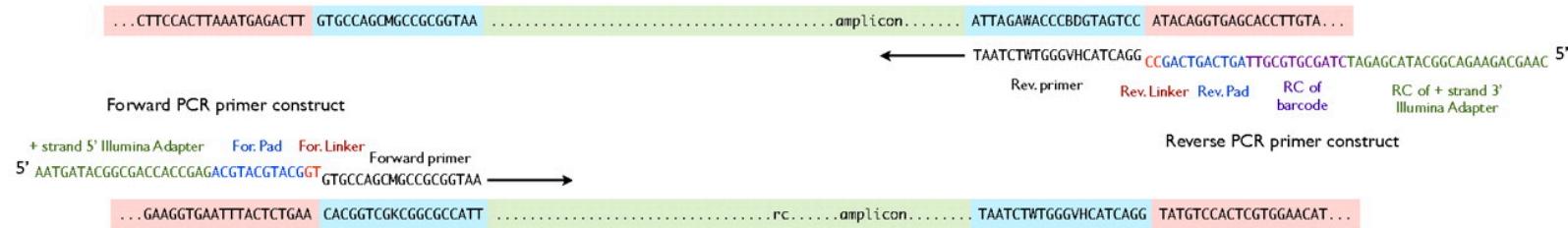
Ribosomal RNA	Eukaryotic	Bacterial
Large subunit	28S rRNA (3354 nt) 5S rRNA (120 nt) 5.8S rRNA (154 nt) 47 proteins	23S rRNA (2839 nt) 5S rRNA (122 nt) 33 proteins
Small subunit	18S rRNA (1753 nt) 32 proteins	16S rRNA (1504 nt) 20 proteins

16S rRNA sequencing

Target gene:



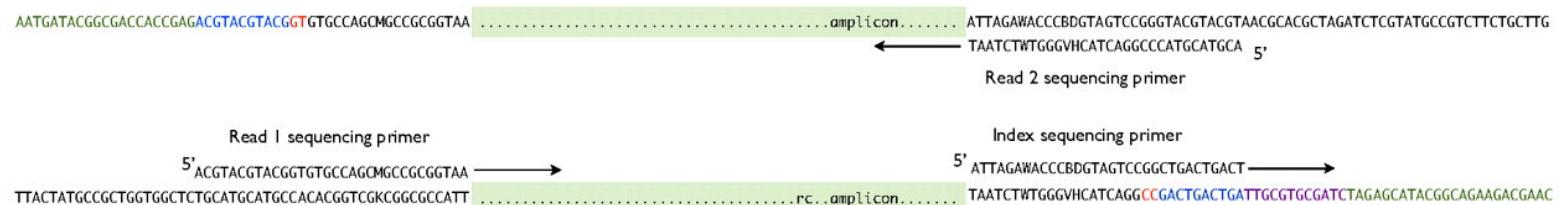
Amplification primers with annealing sites:



Amplification products:



Sequencing primers with annealing sites:



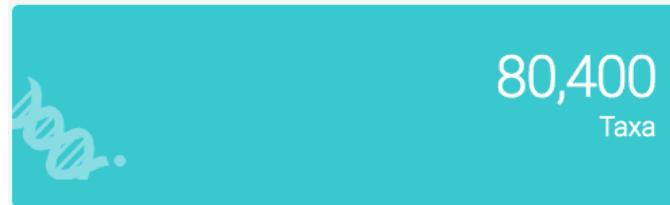
16s rRNA (bacteria)

16S rRNA database: the Ribosomal Database Project, Greengenes, SILVA, EzTaxon

ezbiocloud database (www.ezbiocloud.net)

Dashboard of Prokaryotic Diversity

This dashboard represents EzBioCloud's current microbial taxonomy and genomic diversity.
Feel free to use anything here for your publications, lectures, and presentations.



- bacteria: 60,883
- archaea: 2,843

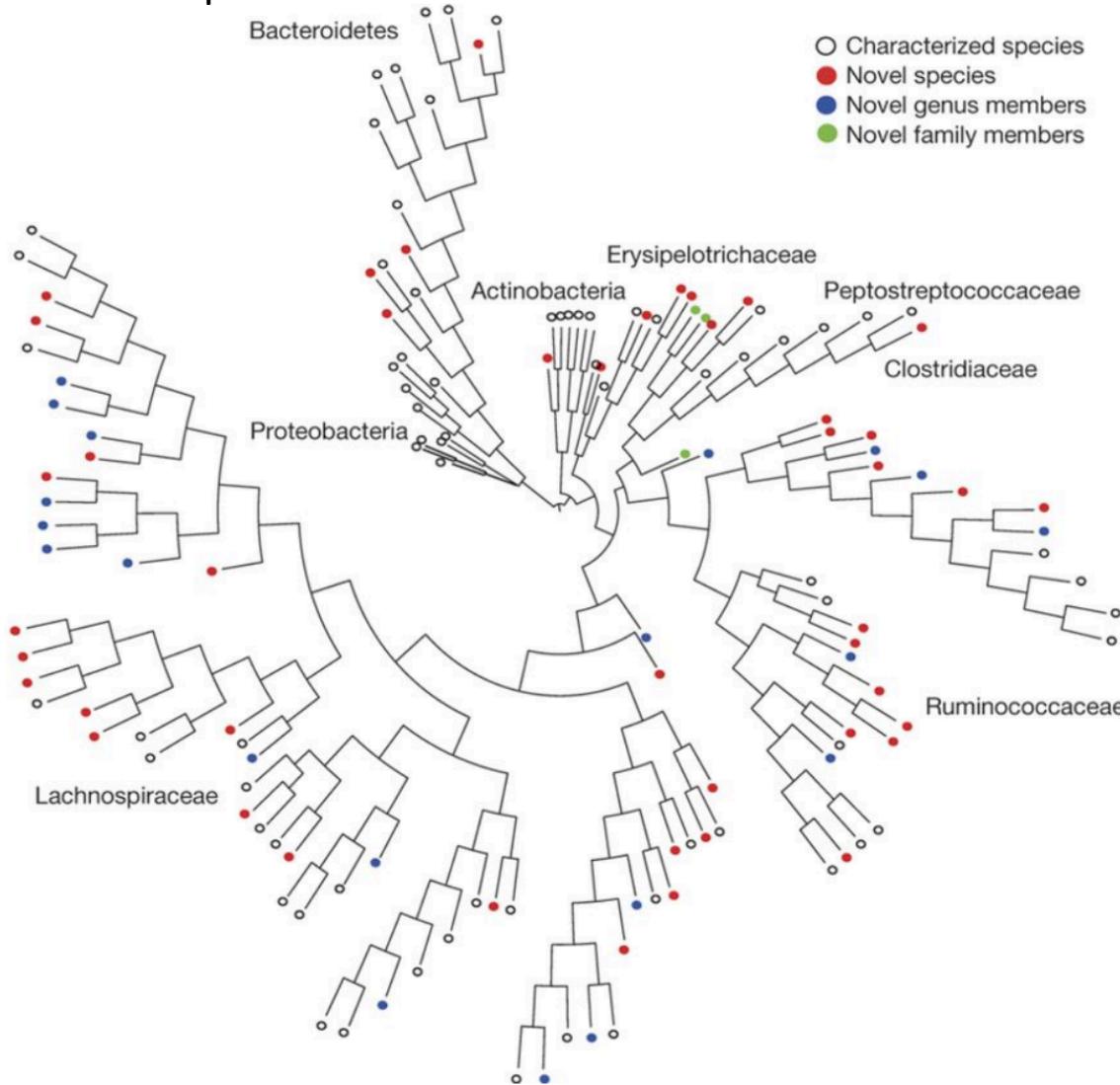
taxonomic classification: to assign taxonomic names to microbial sequences in the data

QIIME

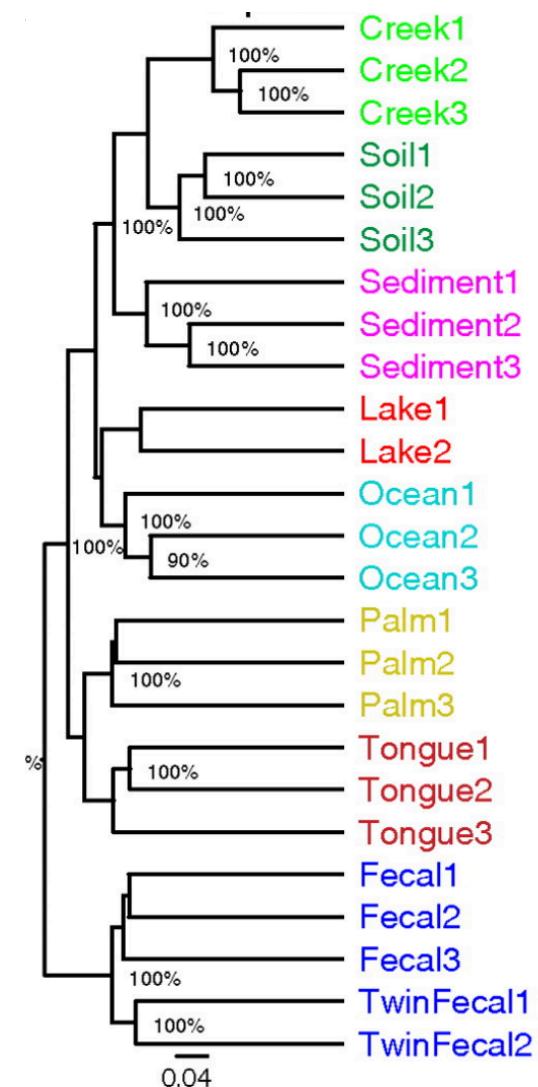
- Quality filtering of reads
- OTU (operational taxonomic unit) picking:
 1. *de novo*: : all-to-all alignments and clustering
 2. closed-reference: OTU aligned to pre-defined cluster centroids in a reference database. Excluded if no matches
- 3. **open-reference**: aligned to the reference & *de novo*
 - Output:
OTU table or the number of times each OTU is observed in each sample.

Microbial communities

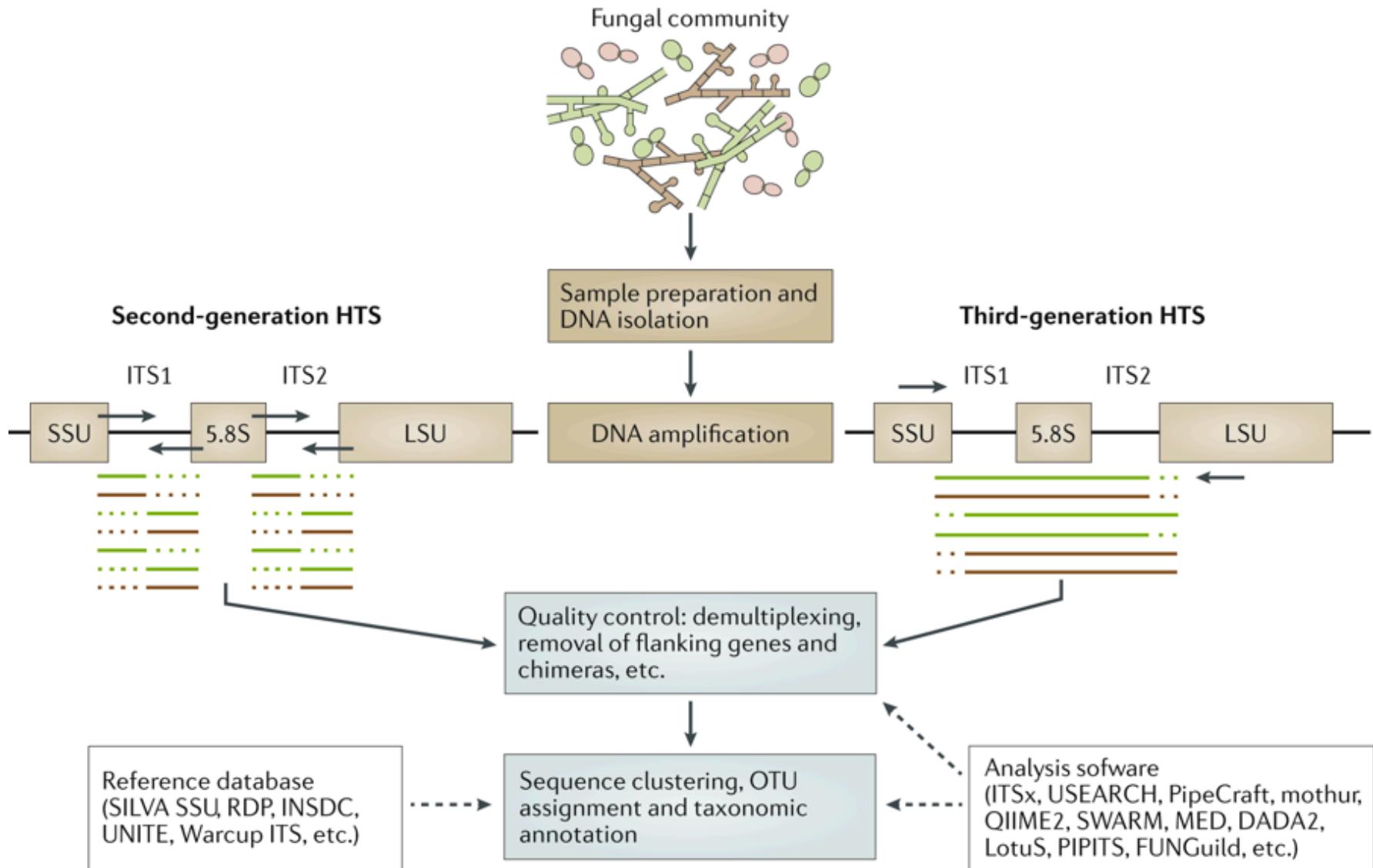
From one sample



From diverse samples



ribosome DNA (fungal example)

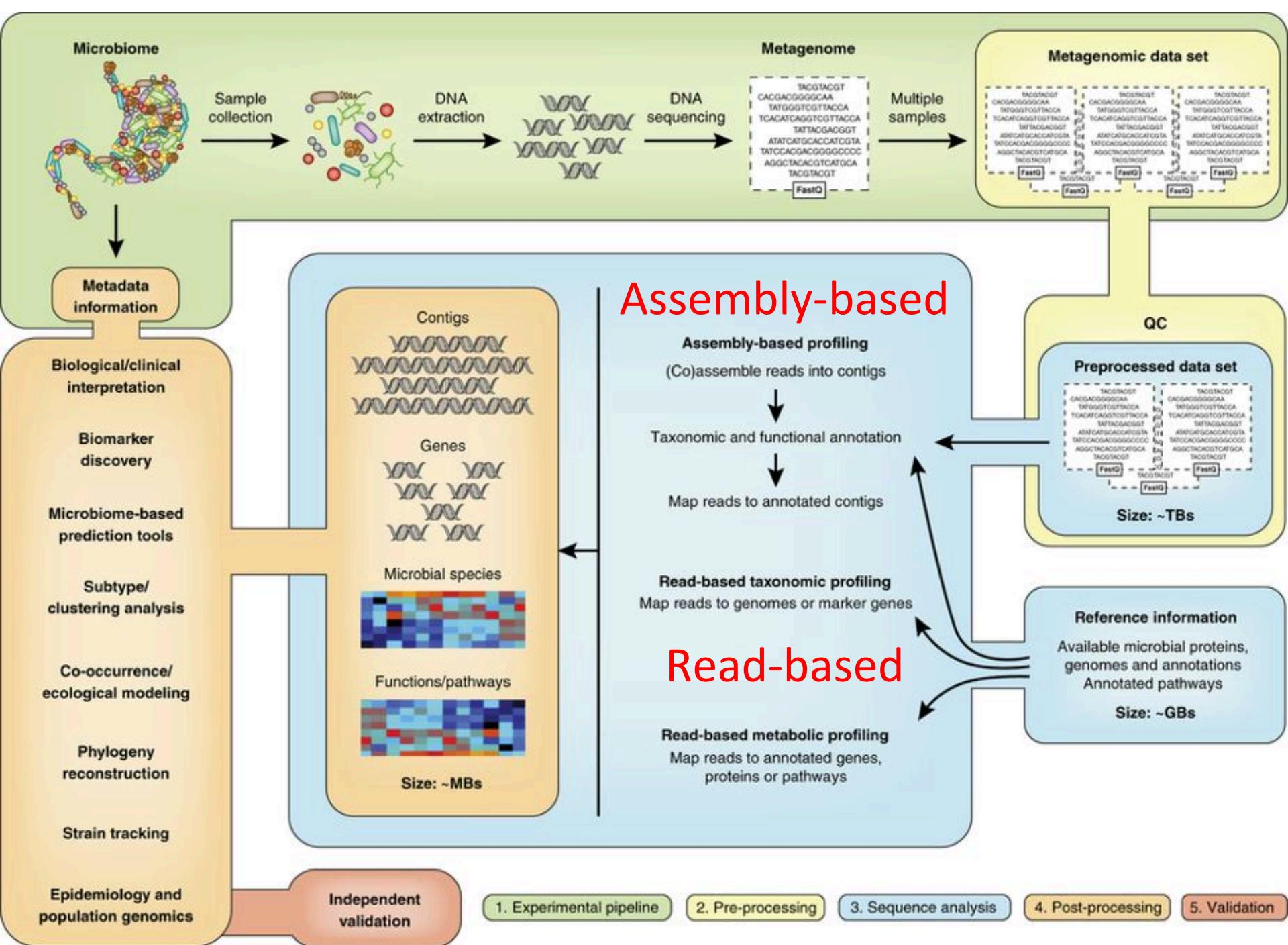


Tools for marker gene analysis

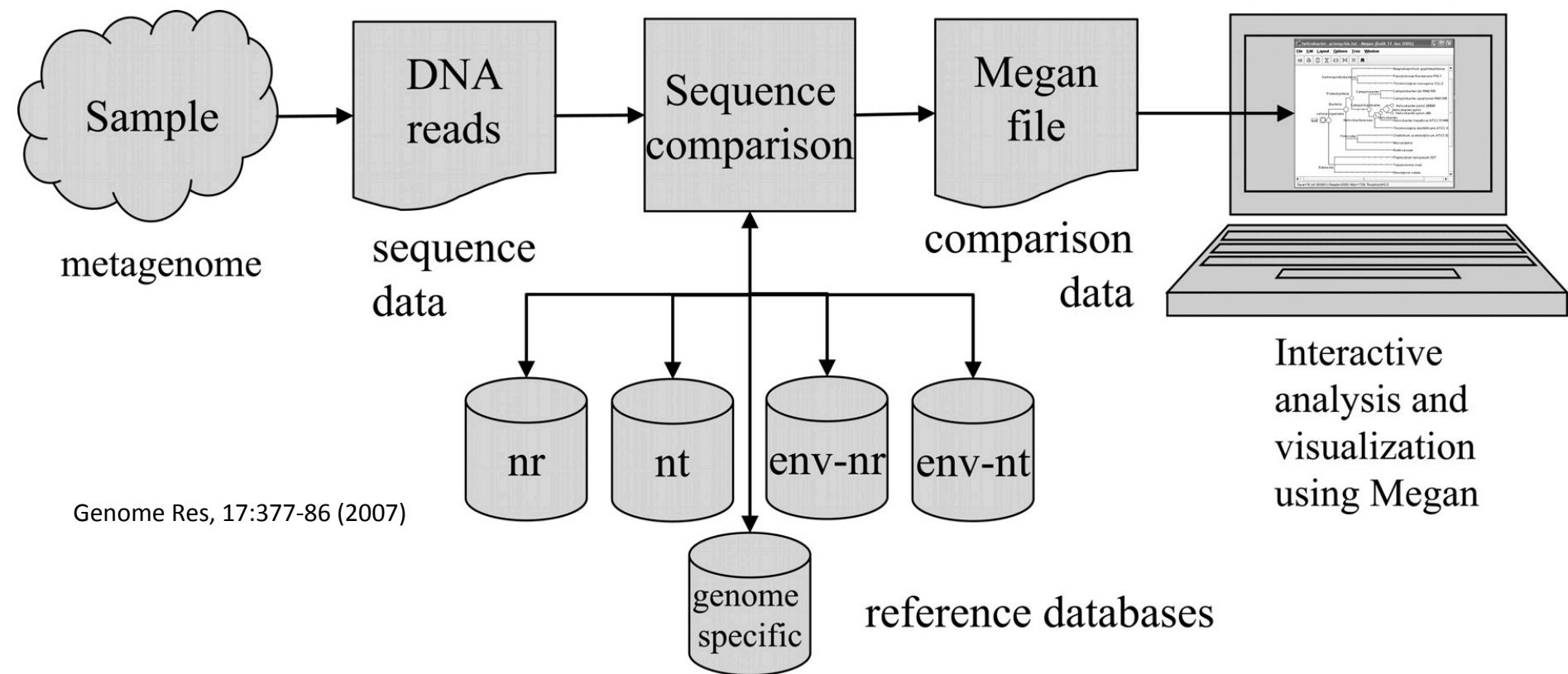
Name	Description and link
DADA2	Amplicon sequence variant analysis pipeline
LotuS	Full pipeline for amplicon data
mothur	Versatile software suite (designed mostly for 16S rRNA)
AMPTk	Full pipeline for amplicon data
OBITools	Versatile software package
PipeCraft	Full pipeline for amplicon data (with graphical user interface)
PIPITS	Full pipeline for fungal ITS amplicon data (only for Illumina data)
QIIME	Full pipeline for amplicon data (designed mostly for 16S rRNA)
SEED2	Full pipeline for amplicon data (with graphical user interface; on Windows)
Microbiology.se	Tools, including ITSx and Metaxa2, for processing ITS, SSU and LSU data
USEARCH	Versatile software package
VSEARCH	Versatile software package

Whole-metagenome analysis

- Sequencing of "all" DNA, including viral and eukaryotic DNA.
- High taxonomic resolution: to species or strain levels
- Assembly of whole microbial genomes
- Relatively expensive

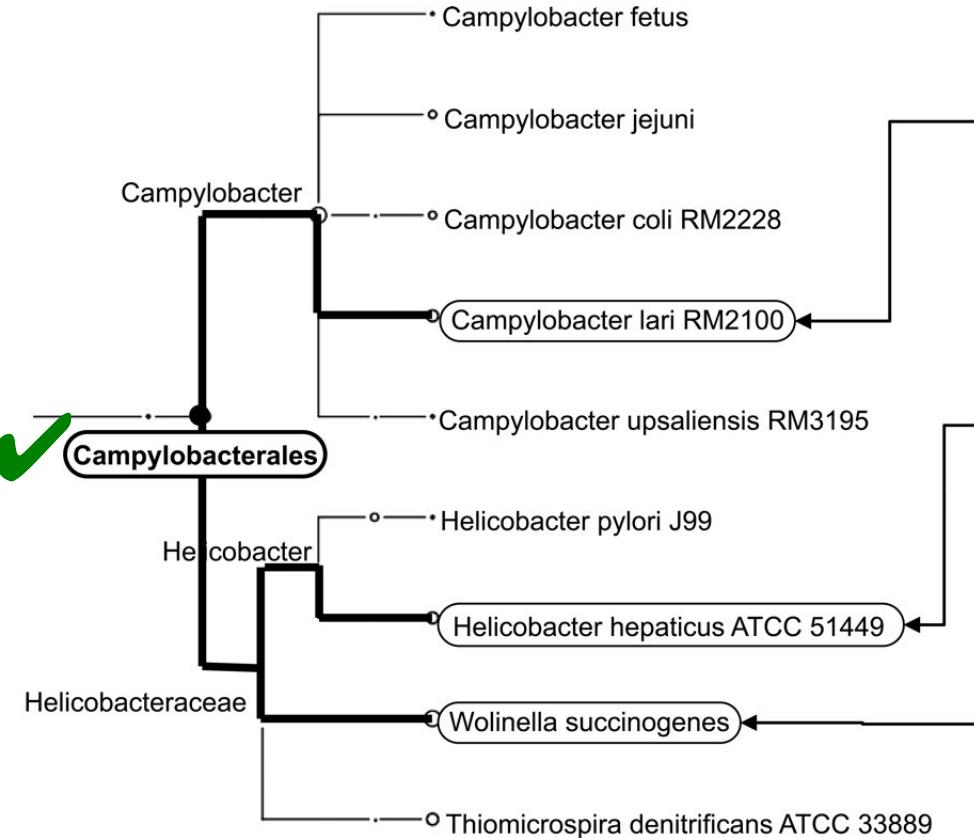


Read-based analysis (Megan)



DNA sequences were aligned against databases of known sequences (e.g., BLAST). MEGAN then assigns taxonomy and summarizes the results.

LCA: lowest common ancestor



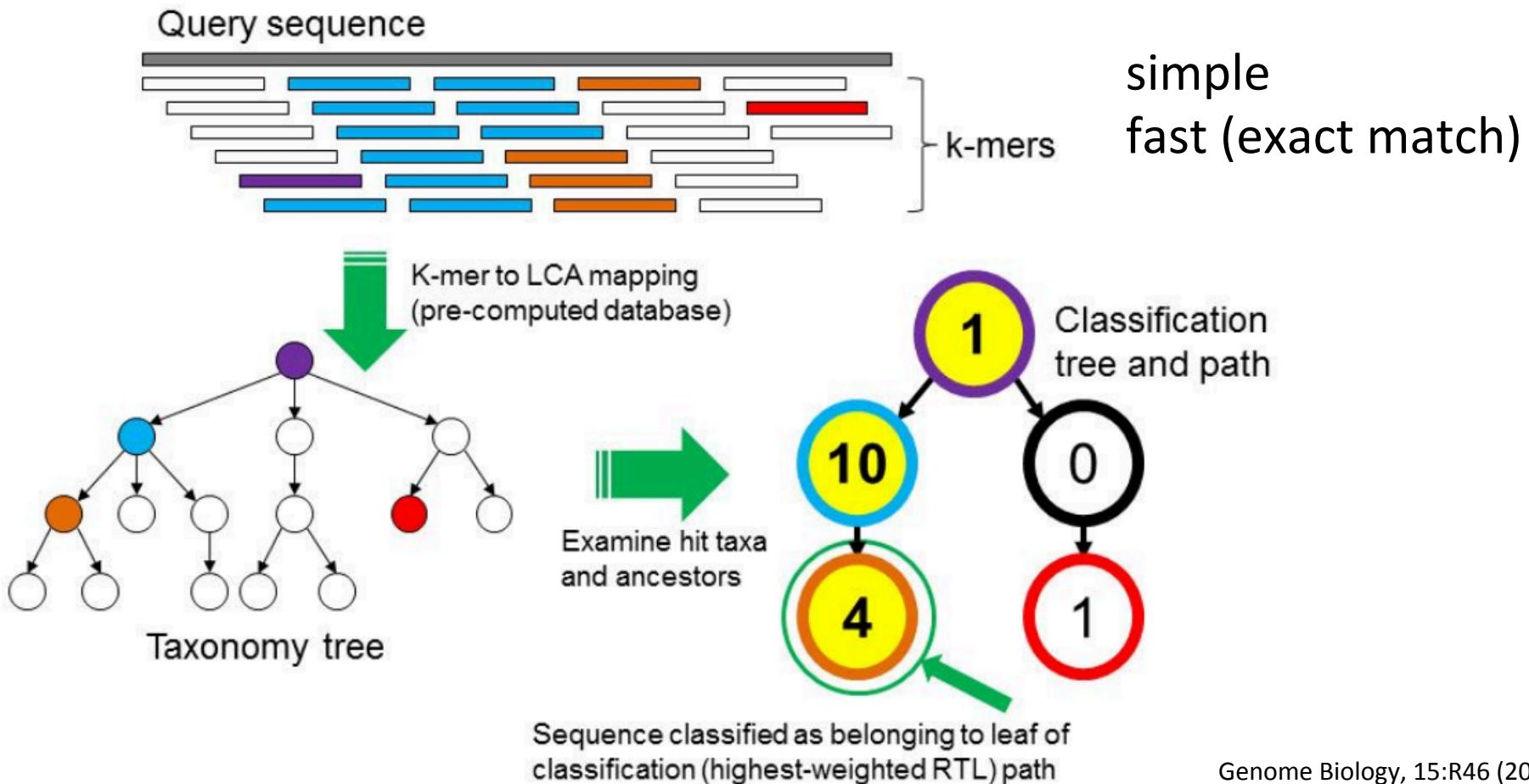
>gi|57241447|ref|ZP_00369393.1| flagellar motor switch protein FliG
[Campylobacter lari RM2100]
Score = 33.9 bits (76), Expect = 1.8
Identities = 13/26 (50%), Positives = 19/26 (73%)
Query: 79 LMVFDDLATVEENGIREIINRADKK 2
LMF FDD++ + N IRE++ ADK+
Sbjct: 243 LMFTFDDISQLSTNAIREVLKAADKR 268

>gi|32262158|gb|AAP77207.1| flagellar motor switch protein FliG
[Helicobacter hepaticus ATCC 51449]
Score = 33.5 bits (75), Expect = 2.4
Identities = 13/26 (50%), Positives = 20/26 (76%)
Query: 79 LMVFDDLATVEENGIREIINRADKK 2
+MF F+D++ ++ N IREI+ ADKK
Sbjct: 244 MMFTFEDISKLDNNAIREILKIAADKK 269

>gi|34484004|emb|CAE11000.1| FLAGELLAR MOTOR SWITCH PROTEIN FLIG
[Wolinella succinogenes]
Score = 32.7 bits (73), Expect = 4.1
Identities = 13/26 (50%), Positives = 19/26 (73%)
Query: 79 LMVFDDLATVEENGIREIINRADKK 2
+MF F+D+ ++ N IREI+ ADKK
Sbjct: 242 MMFTFEDIEKLDNNAIREILKVADKK 267

A simple LCA algorithm assigns reads to taxa
(e.g., the read was assigned to *Campylobacterales*)

Read-based analysis (Kraken: k-mer analysis)



Genome Biology, 15:R46 (2014)

Kraken has a database that contains records consisting of a k-mer and the LCA of all organisms whose genomes contain that k-mer.

Assembly-based analysis

Metagenome *de novo* assembly is conceptually similar to whole-genome assembly (de Bruijn graph).

Challenges:

1. Uneven sequencing coverage of each genome
2. Multiple strains from a species

Resulting in fragmented assemblies (million of contigs)

metagenomic assemblers: MEGAHIT and metaSPAdes

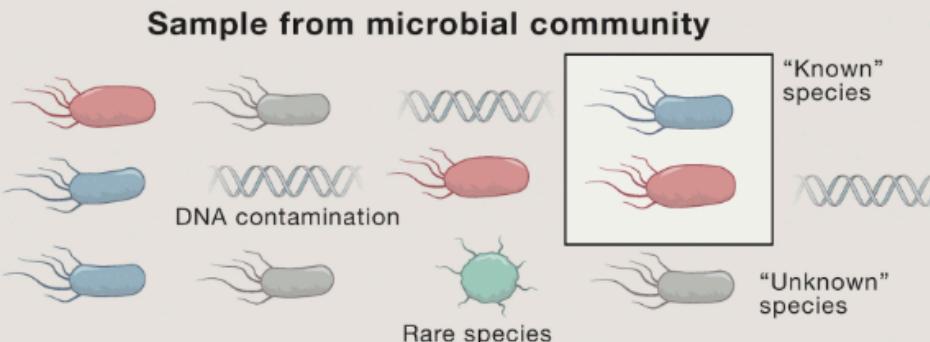
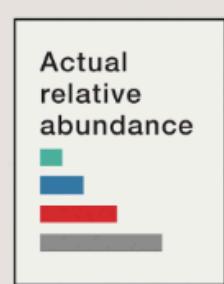
Contig binning: to group contigs into species

Number of genomes and what genomes (species)

- Taxonomic classification: contig aligned against known genomes
- Sequence composition are used for binning contigs from unknown genomes
 1. Different genomes have particular combinations of bases (tetramers)
 2. Sequencing coverage: sequences from the same genome have similar sequencing depth
- Contigs are reconstructed into metagenomic assembled genomes (MAGs)

Example of bacterial whole-genome metagenomics (I)

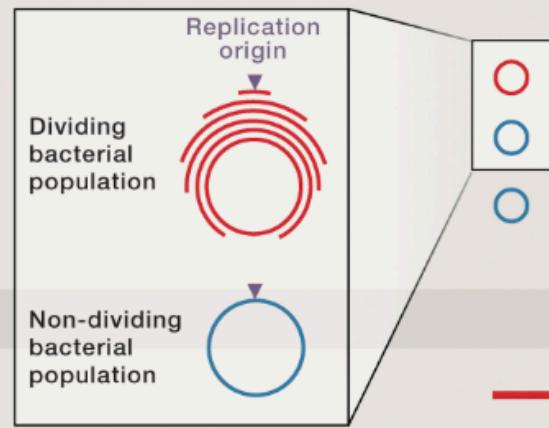
A



Unknown species can dominate microbial communities (Nayfach et al., 2016) and are not detected by reference-based methods

DNA from the host (Ames et al., 2015) or laboratory environment (Salter et al., 2014) can contaminate a biological sample

↓ DNA extraction



Extraction efficiency varies between taxa (Kennedy et al., 2014)

Dividing bacterial genomes have higher and less even genomic coverage (Korem et al., 2015)

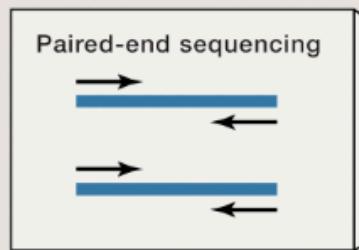
↓ DNA fragmentation



Extracted DNA is fragmented at breakpoints that preferentially occur at certain di-nucleotides (Poptsova et al., 2014)

Example of bacterial whole-genome metagenomics (II)

D



↓ Prepare library and sequence

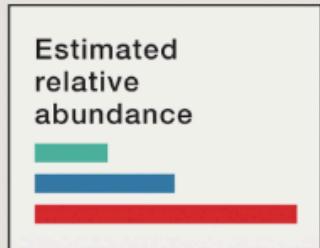
Library preparation protocol affects estimated community composition (Jones et al., 2015)
Sequencing technologies have different read lengths and error rates (Quail et al., 2012)

E



Duplicate reads eliminated
Read-tails trimmed
Low-quality reads filtered
DNA contamination removed

F

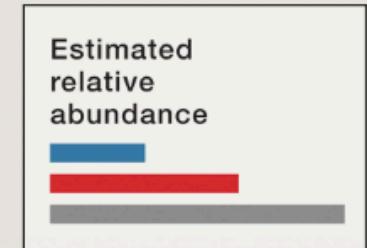


Reference-based classification

Unknown taxa may not be detected

Metagenomic assembly

Rare taxa may not be detected



Metatranscriptome

Metatranscriptomics uses RNA sequencing to profile transcription in microbiomes, providing information on active functional outputs of the microbiome.

1. To acquire high-quality RNA might be challenging (e.g., feces)
2. rRNA contamination

Metatranscriptome analysis

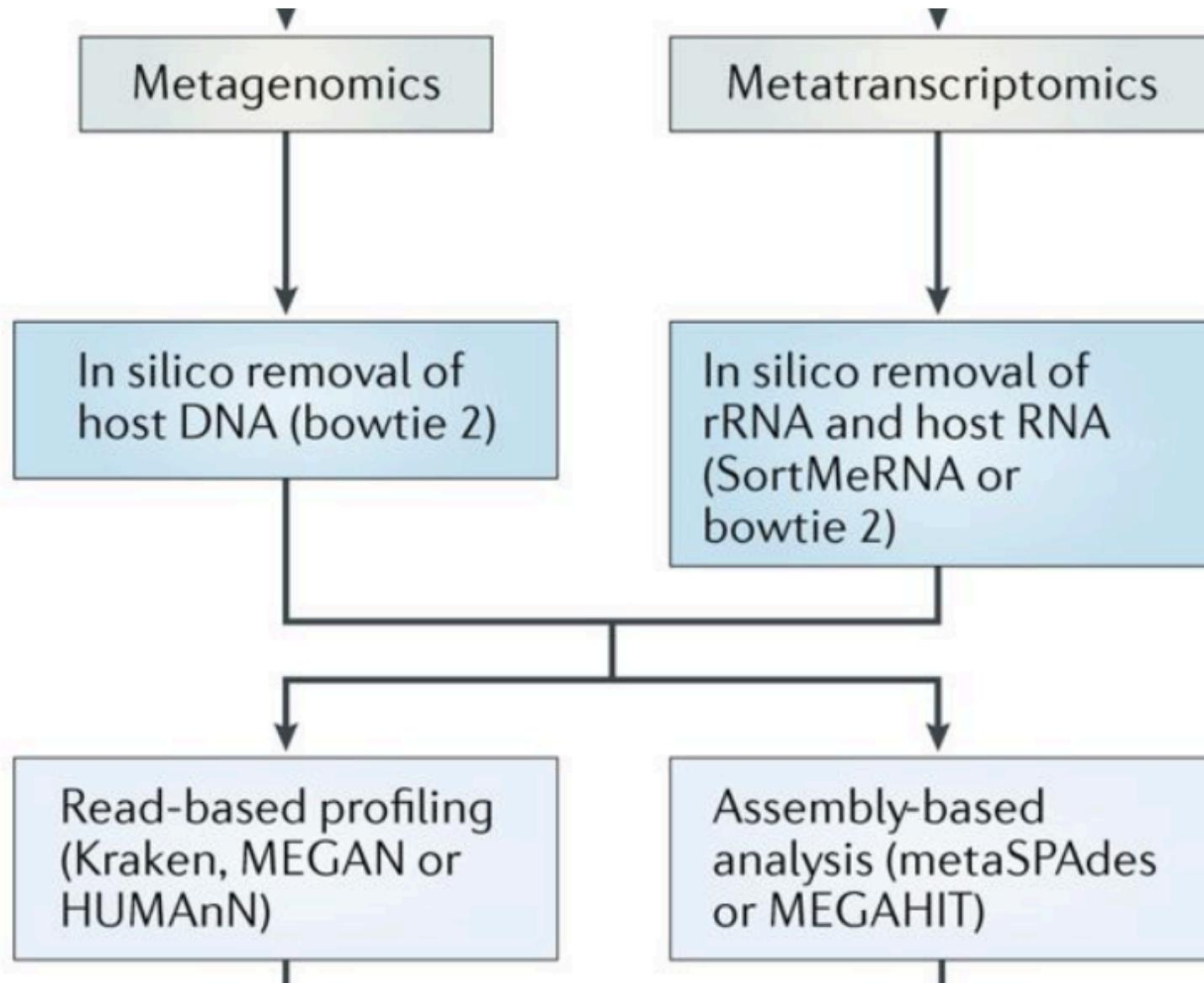
1. mRNA or gene expression is the focus
2. rRNA removal but not completely

Generally, the following reads are discarded:

- non-mRNA reads
- reads derived from hosts

RNA-Seq analyzing tools for assembly and differential expression are useful for metatranscriptome analysis.

Metagenome and metatranscriptome analysis



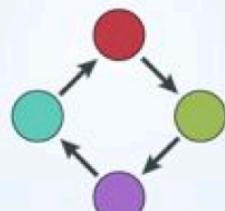
Strengths and weaknesses of metagenomic studies

Strengths

Inclusion of a large number of samples



Integrative workflows

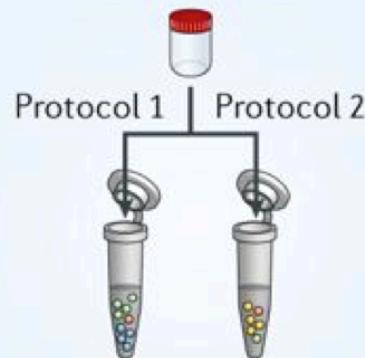


Time to result



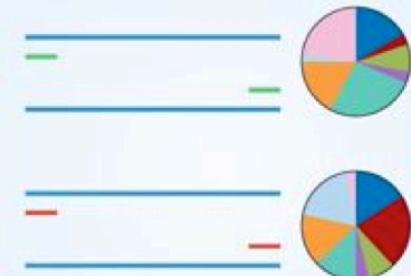
Weaknesses

Extraction bias



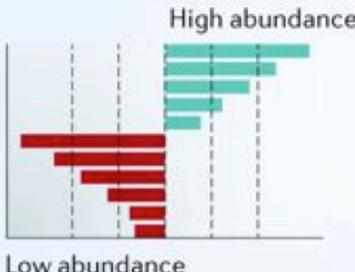
The bacterial diversity that is captured is a function of the extraction protocol

Primer bias



Species identification is dependent on the targeted hypervariable region

Detection of differentially abundant taxa



Detection of uncultured bacteria

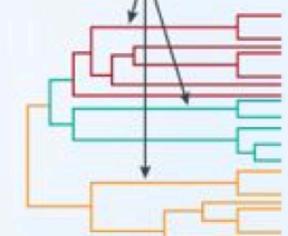


Bioinformatics biases

Variations in methodology

Discrimination between species is difficult owing to amplicon length

ATGGAAGTCGAACGGAGAGAATGCTAGCTTGC
TAATAATTCTCGTGGCCGCCACGGGAGAGTA
GTGAGTAACCTGCGCGCCTCGGAAC



biases (cont.)

Technical biases from DNA extraction, primers, PCR, and sequencing procedure

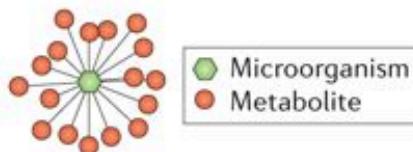
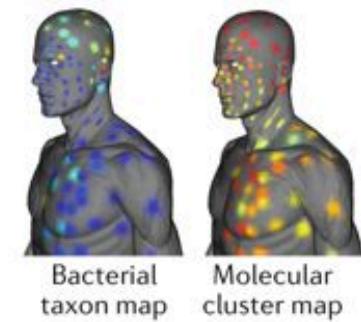
Index (barcode) switching: chimeric products during PCR
solution: less PCR cycles and **chimaera filtering?**

Clustering Issue: satellite OTUs introduced by sequencing errors could lead to the overestimation of richness of the microbial community.

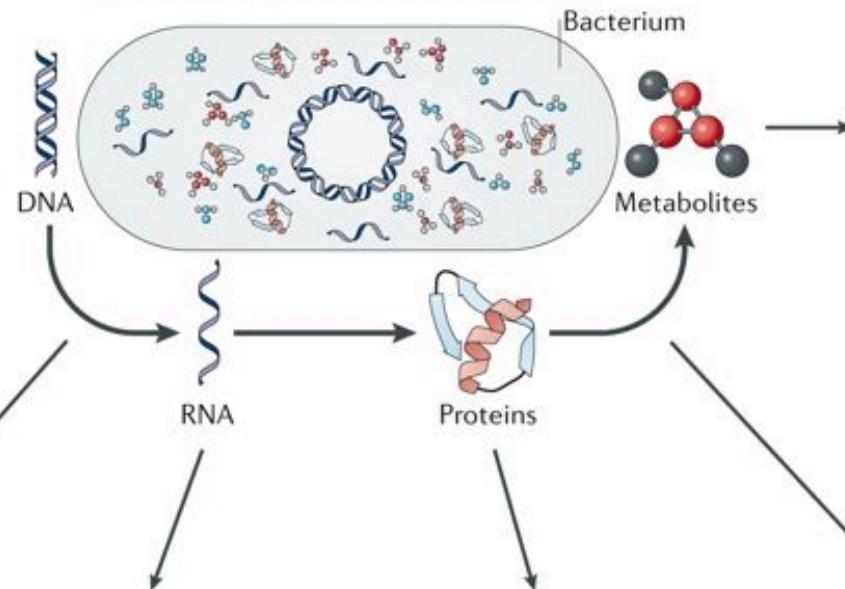
solution: removal of rare OTUs and multiple rounds of clustering

High-level analyses

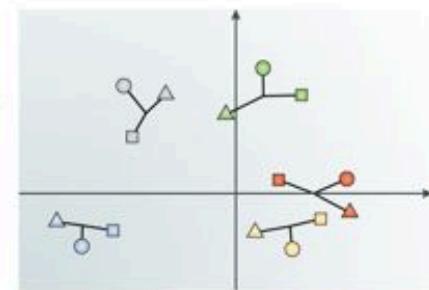
a Spatial correlation analysis



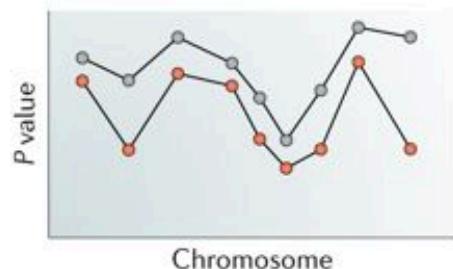
Central dogma of molecular biology



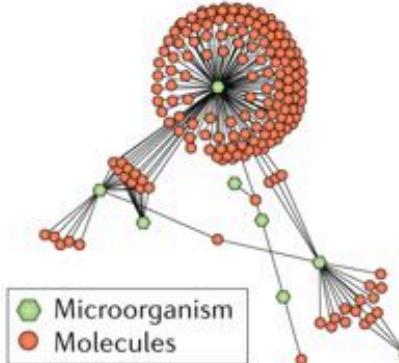
f MCIA



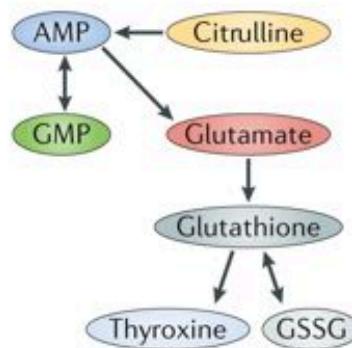
b Sparse canonical correlation



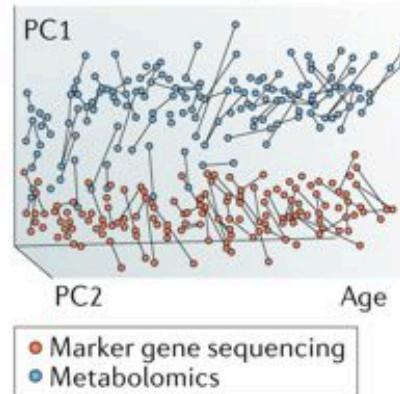
c Correlation networks



d Metabolic activity networks



e Procrustes analysis



References

1. Caporaso, J. G. et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4516–4522 (2011).
2. Browne, H. P. et al. Culturing of ‘unculturable’ human microbiota reveals novel taxa and extensive sporulation. *Nature* 533, 543–546 (2016).
3. Nayfach, S. & Pollard, K. S. Toward Accurate and Quantitative Comparative Metagenomics. *Cell* 166, 1103–1116 (2016).
4. **Knight, R. et al. Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422 (2018).**
5. Nilsson, R. H. et al. Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat. Rev. Microbiol.* 17, 95–109 (2019).
6. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* 38, 525–552 (2004).
7. Bashiardes, S., Zilberman-Schapira, G. & Elinav, E. Use of Metatranscriptomics in Microbiome Research. *Bioinform. Biol. Insights* 10, 19–25 (2016).
8. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. doi:10.1101/gr.5969107