Inferring community composition and species interaction networks from metagenomic data

Georg Zeller
Team Leader, SCB, EMBL Heidelberg
zeller@embl.de

Slides and course material at:
https://github.com/gezel/microbial_communities_workshop_2018

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

EMBL

Biological diversity of microbial ecosystems




Meta HIT

HMP

An FP7 financed Project
RUMINOMICS

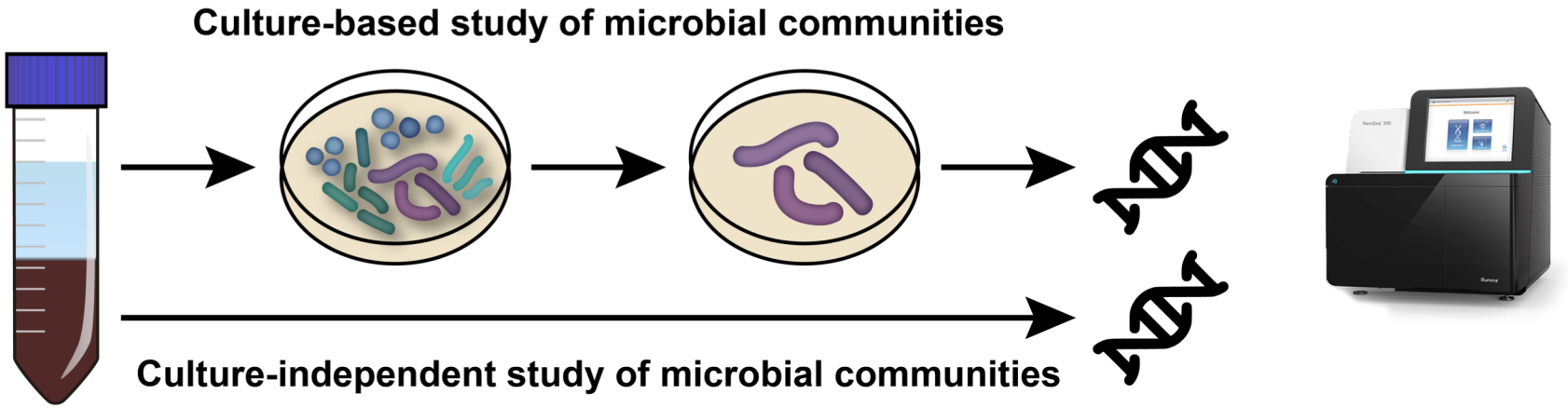
earth microbiome project

TARA OCEANS



The human body is home to complex microbial ecosystems that have co-evolved with the host to form a superorganism

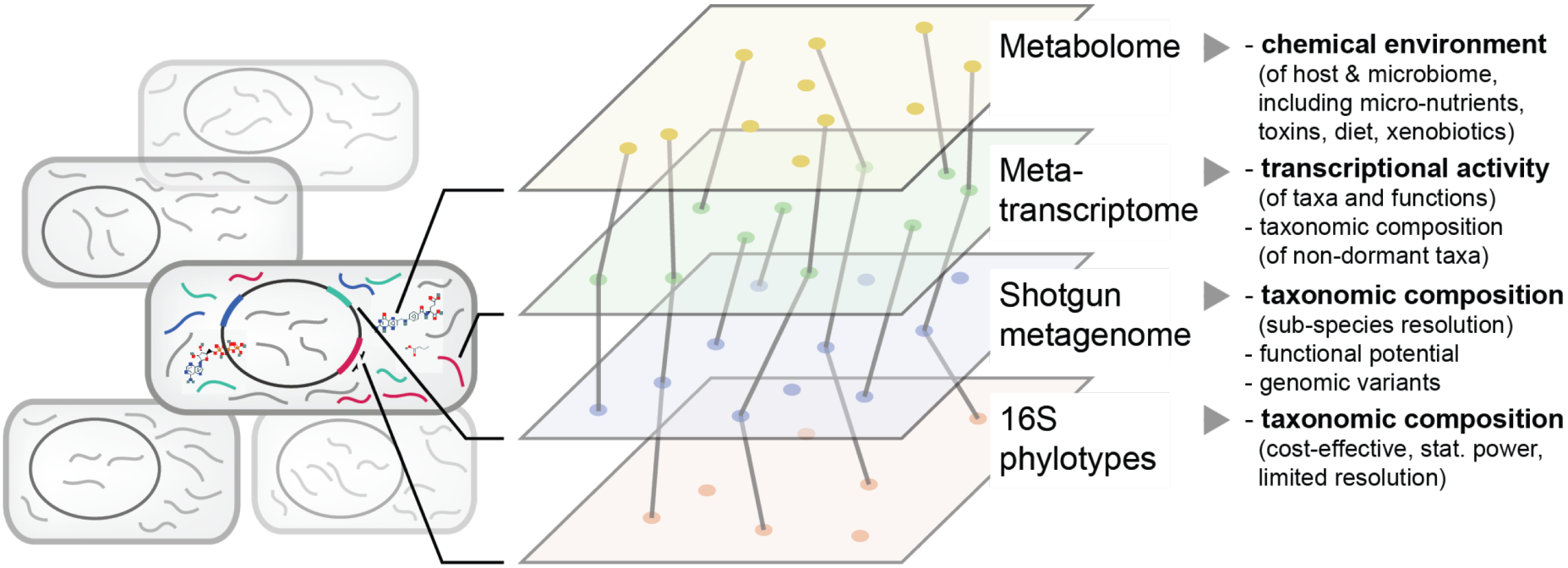
Culture-independent study of microbiomes



Analyzing environmental samples directly avoids bias for cultivable species and potentially allows for quantification of microbes in situ

PART I:
ASSESSING COMMUNITY COMPOSITION USING TAXONOMIC PROFILING

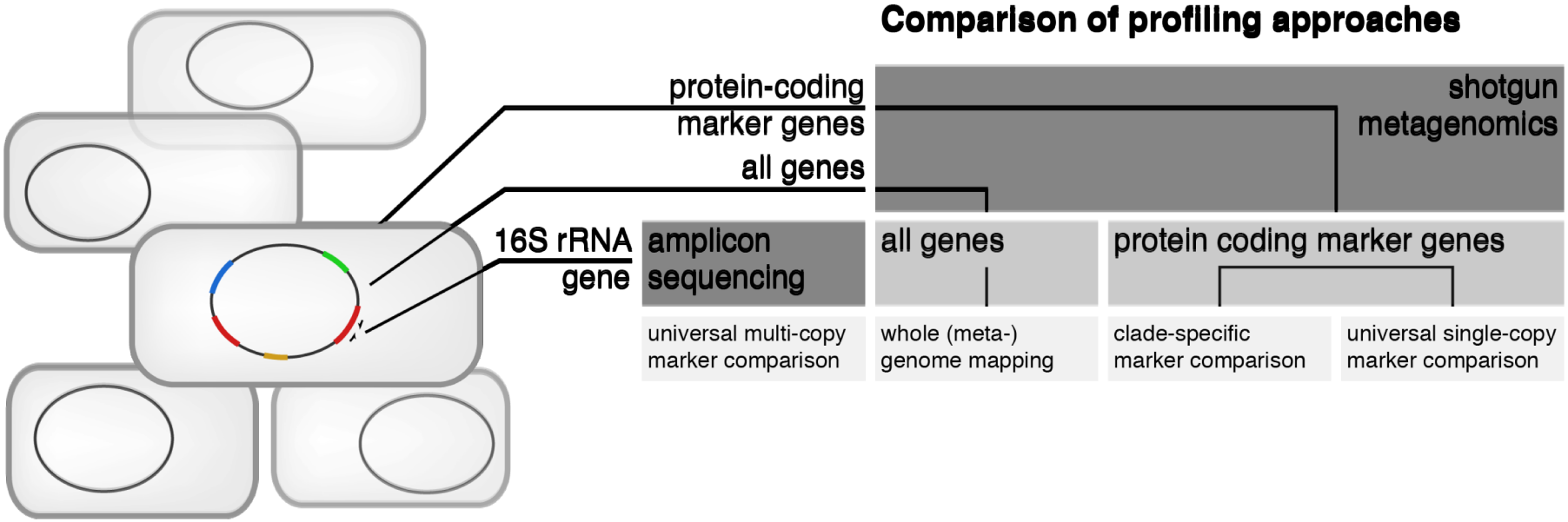
Multi-omics microbiome characterization



- **Who's there?**
- What are they doing?
- **How are they interacting?**

- Taxonomic profiling
- Functional profiling, expression profiling etc.
- Microbe-microbe, host-microbe interactions

Taxonomic profiling approaches for shotgun metagenomics

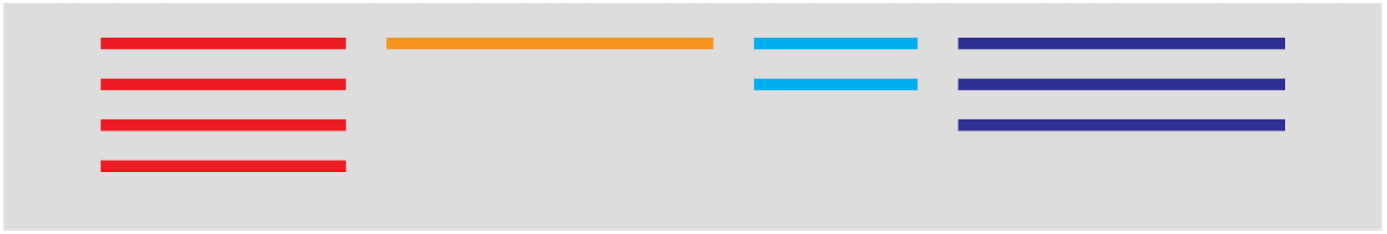


Taxonomic profiling attempts to quantify the abundance (cell counts / sample vol.) of microbial taxa – ideally species – from metagenomic survey data (either 16S rRNA gene sequencing or whole-metagenome shotgun sequencing).

Not to be confused with **taxonomic classification** (assign the taxonomic source of each sequencing read)!

The taxonomic profiling problem – whole-genome mapping

environmental sample

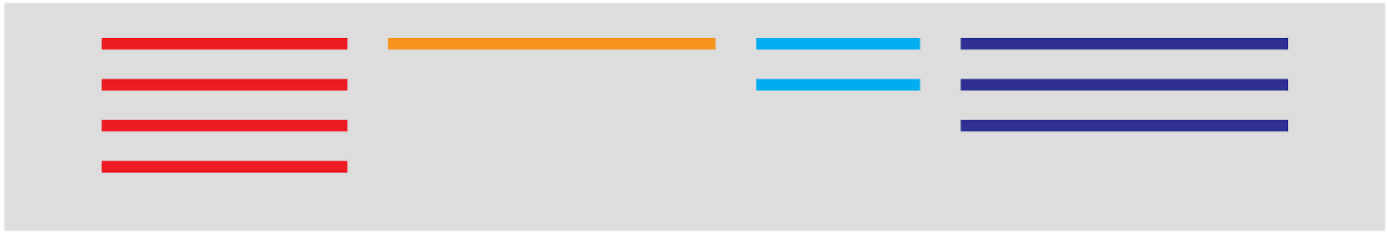


shotgun sequencing

- DNA extraction bias
- sequencing biases
- sampling noise

The taxonomic profiling problem – whole-genome mapping

environmental sample

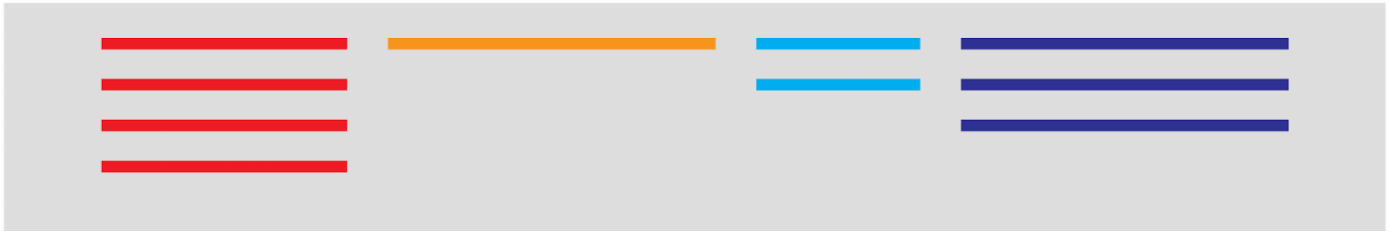


shotgun sequencing

- DNA extraction bias
- sequencing biases
- sampling noise

The taxonomic profiling problem – whole-genome mapping

environmental sample



shotgun sequencing

- DNA extraction bias
- sequencing biases
- sampling noise

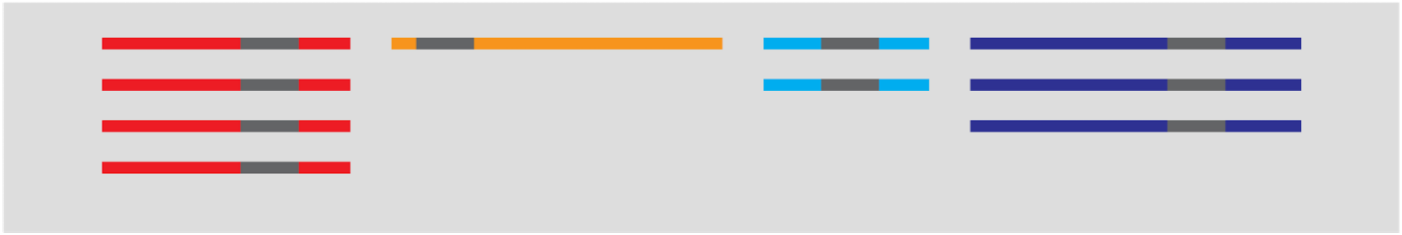
true taxonomic composition

estimated by whole-genome mapping

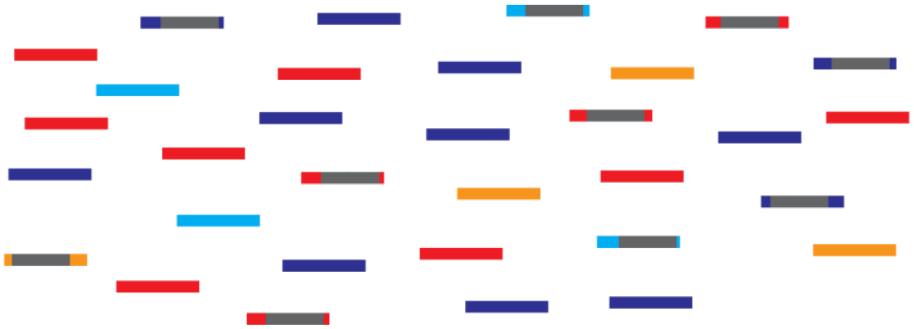
- genome size issue

The taxonomic profiling problem — marker-gene mapping

environmental sample



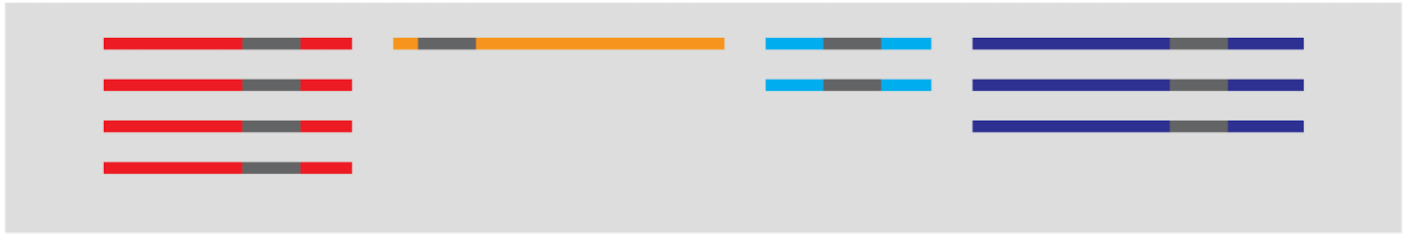
shotgun sequencing



- DNA extraction bias
- sequencing biases
- sampling noise

The taxonomic profiling problem — marker-gene mapping

environmental sample

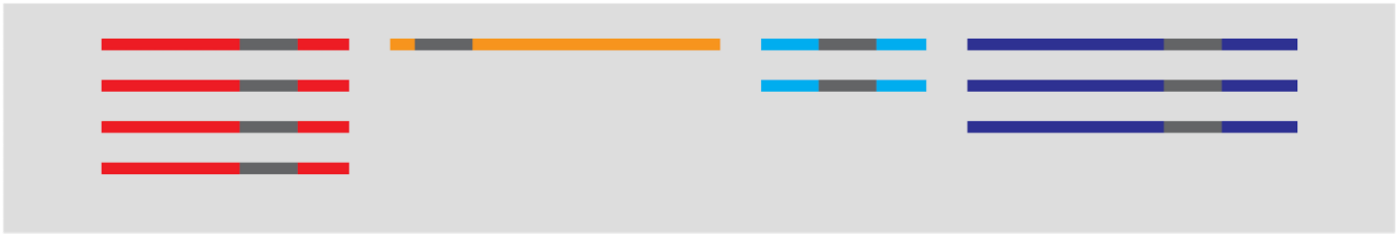


targeted sequencing

- DNA extraction bias
- sequencing biases
- sampling noise
- amplification bias

The taxonomic profiling problem — marker-gene mapping

environmental sample



shotgun sequencing

- DNA extraction bias
- sequencing biases
- sampling noise

The taxonomic profiling problem — marker-gene mapping

environmental sample

shotgun sequencing

- DNA extraction bias
- sequencing biases
- sampling noise

true taxonomic composition

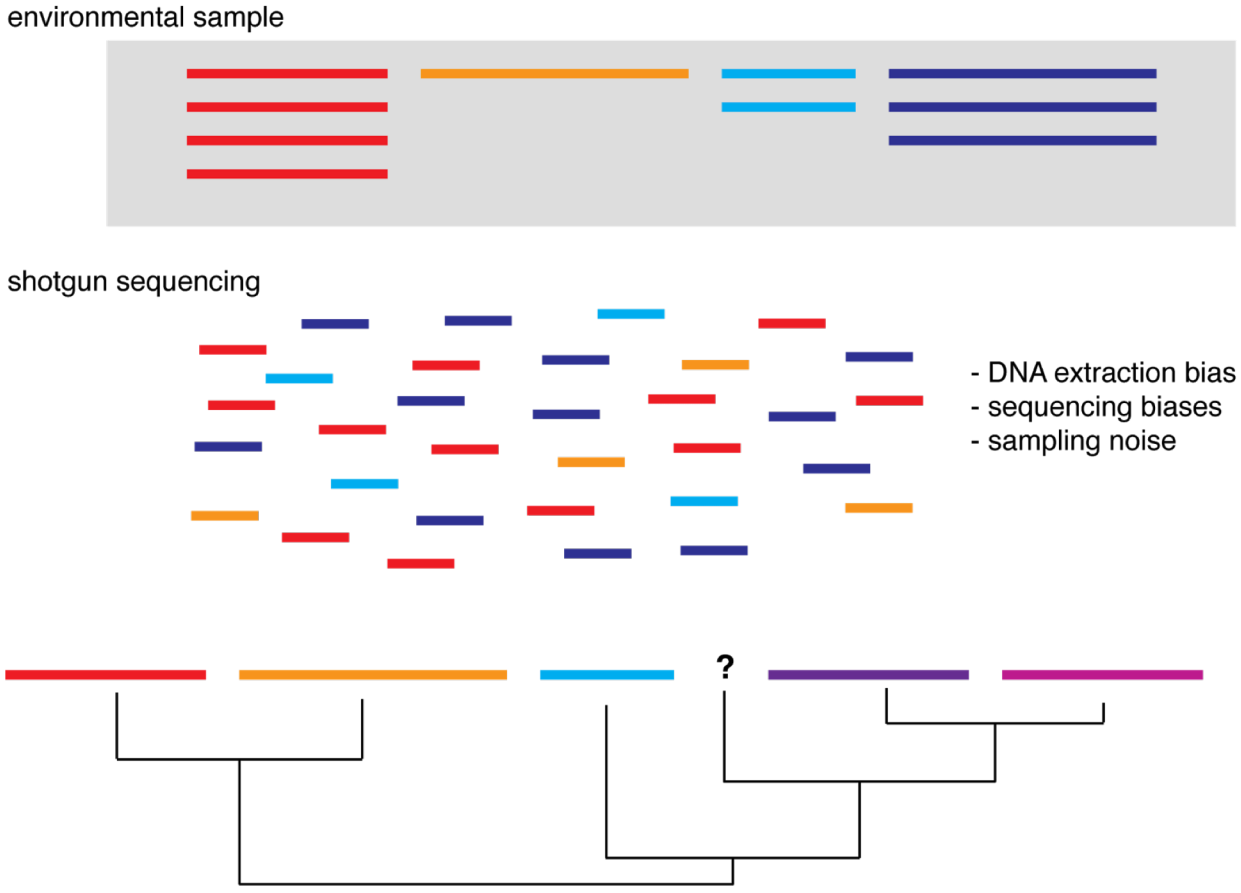
estimated by whole-genome mapping

- genome size issue

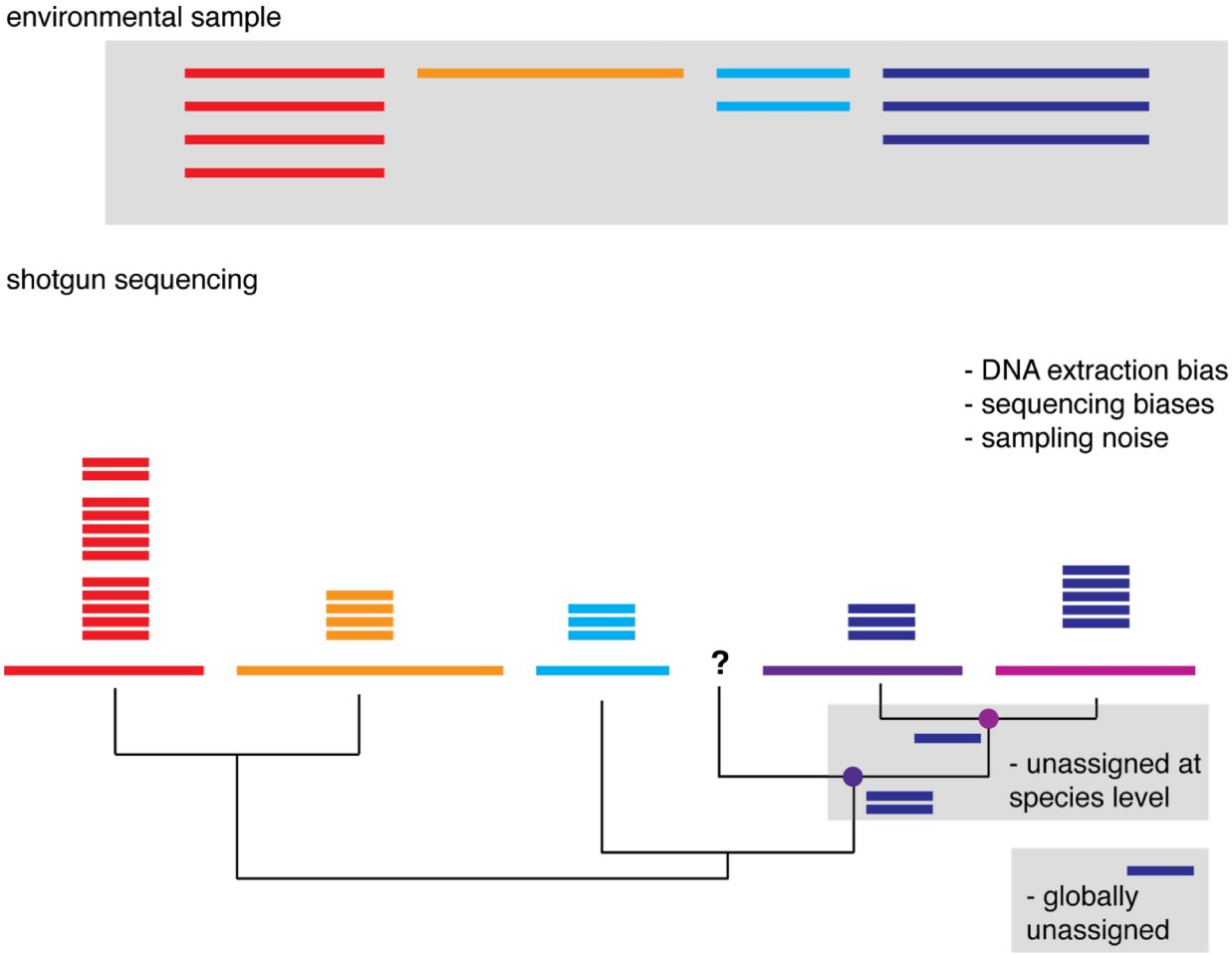
estimated by universal marker

- copy number variation

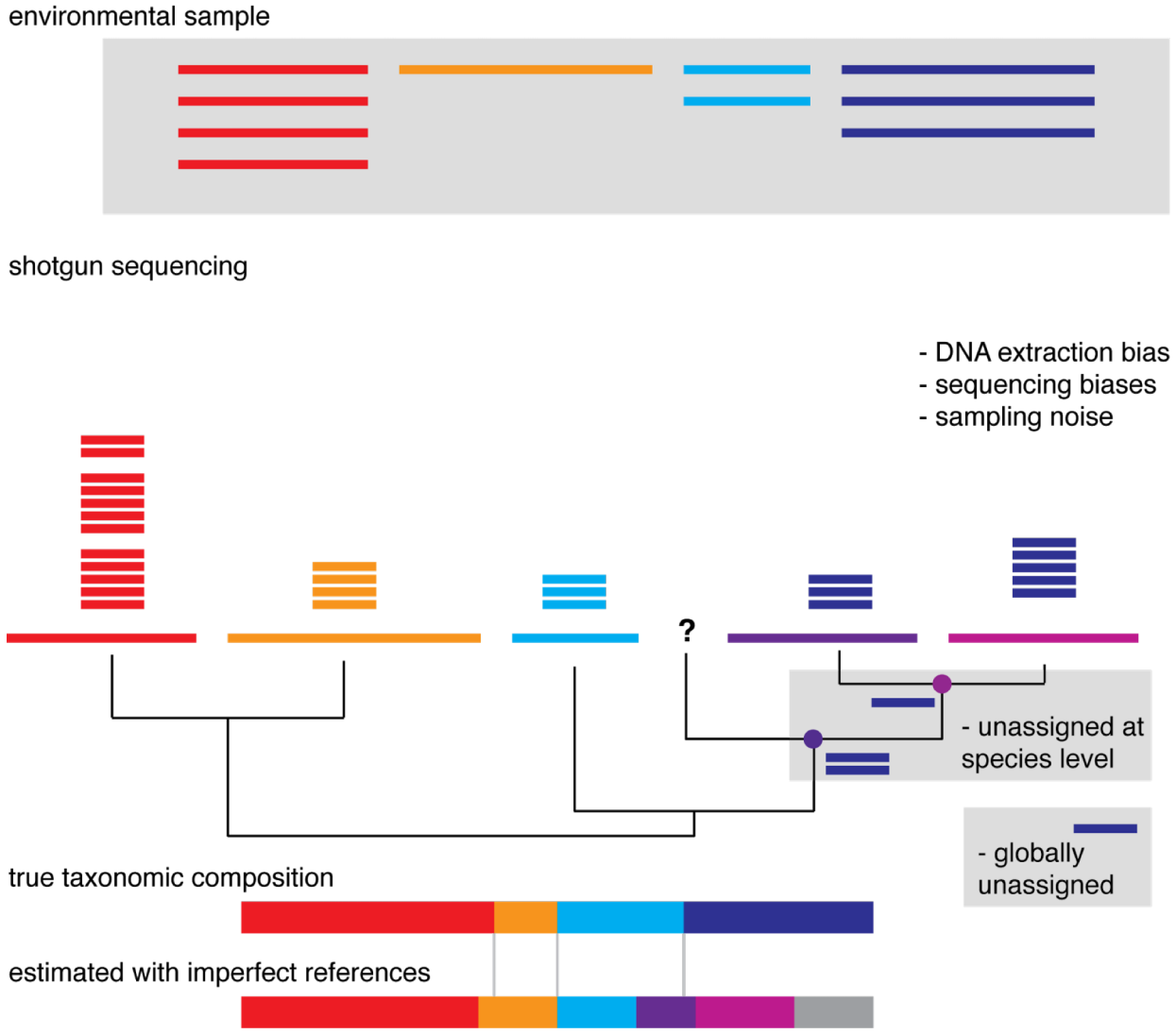
The taxonomic profiling problem – imperfect reference complication



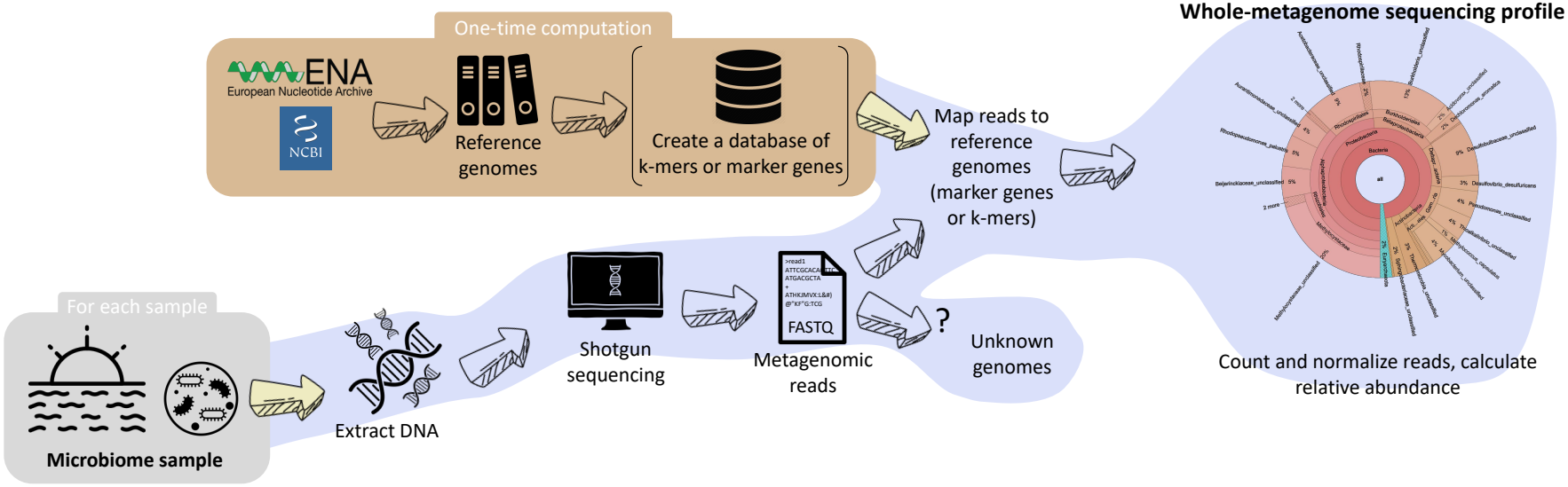
The taxonomic profiling problem – imperfect reference complication

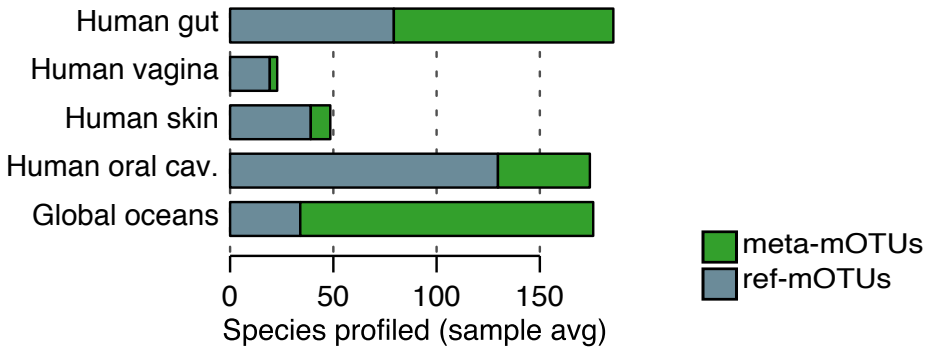


The taxonomic profiling problem – imperfect reference complication



The mOTUs2 profiler

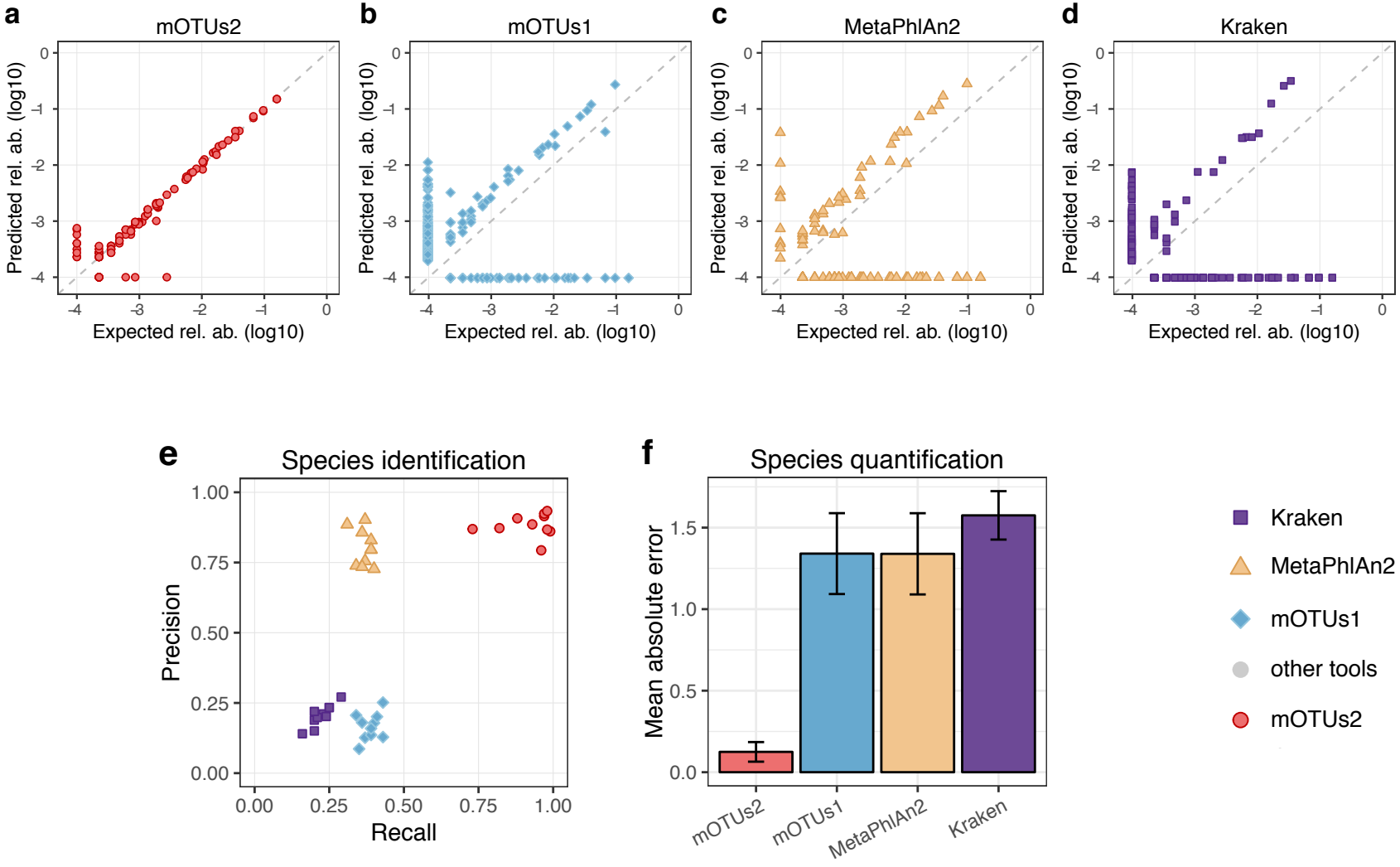




- The mOTUs2 taxonomic profiling tools allows for quantification of species without a sequenced reference genome
- It achieves high precision and recall at species resolution

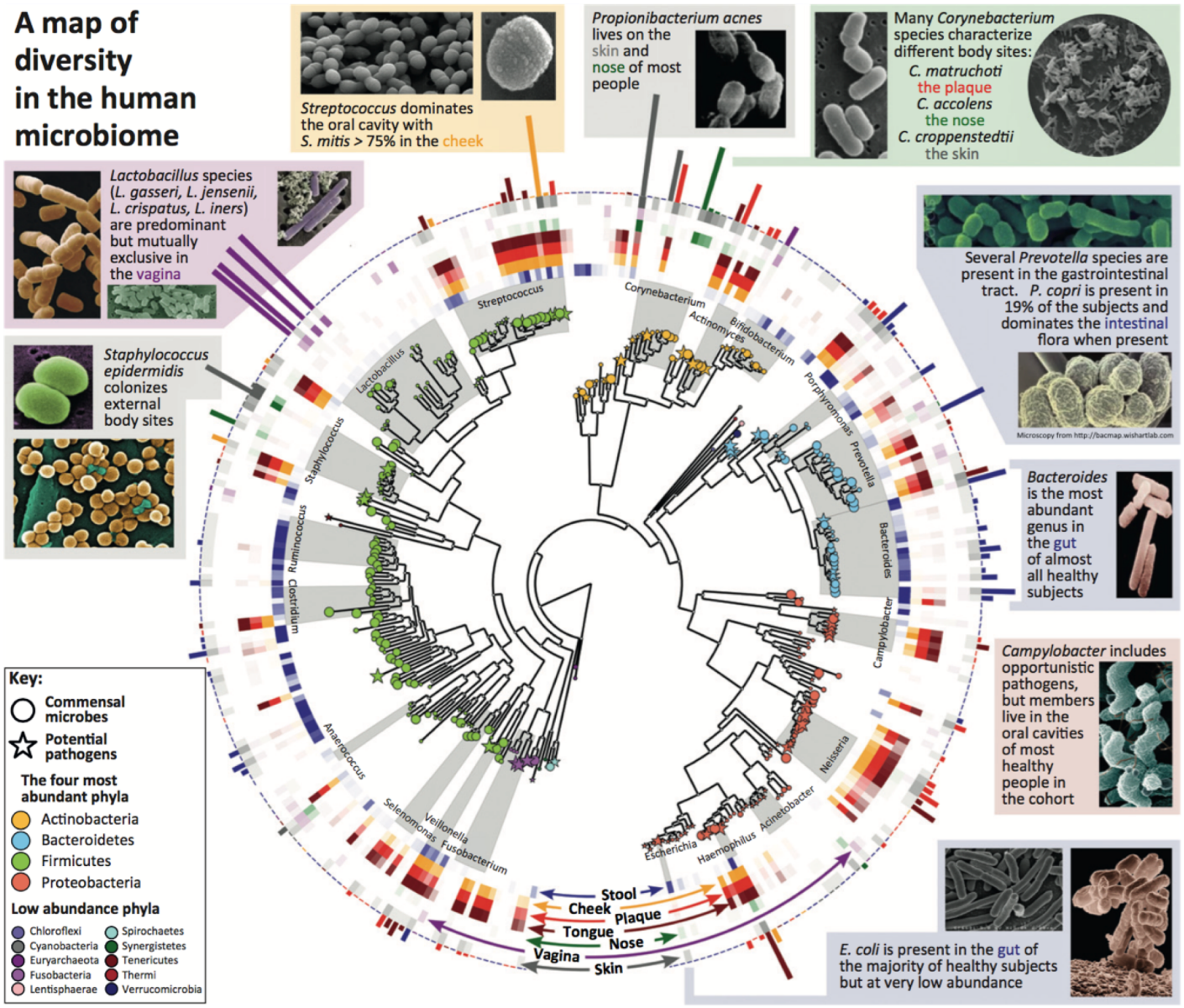
[1] Milanese A. et al., in revision at *Nat Comm.*

Evaluation of the mOTUs2 profiler I



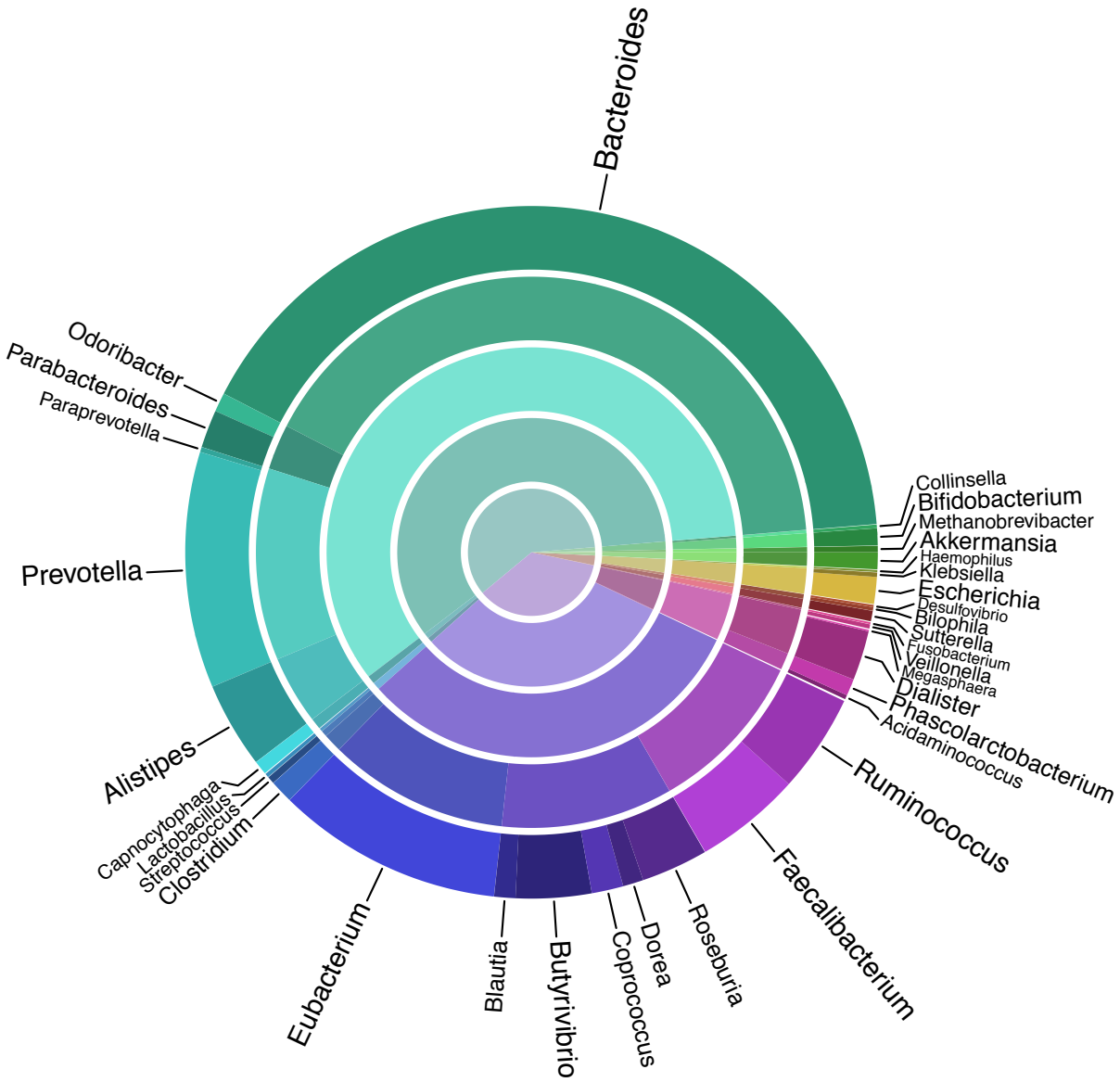
[1] Milanese A. et al., in revision at *Nat Comm.*
[2] Sunagawa S. et al., *Nat Methods* 2013
[3] Truong D.T. et al., *Nat Methods* 2015
[4] Wood D.E. et al., *Genome Biol.* 2014

The human body is home to microbial ecosystems



[1] Morgan X. et al., *Trends Genet*. 2013

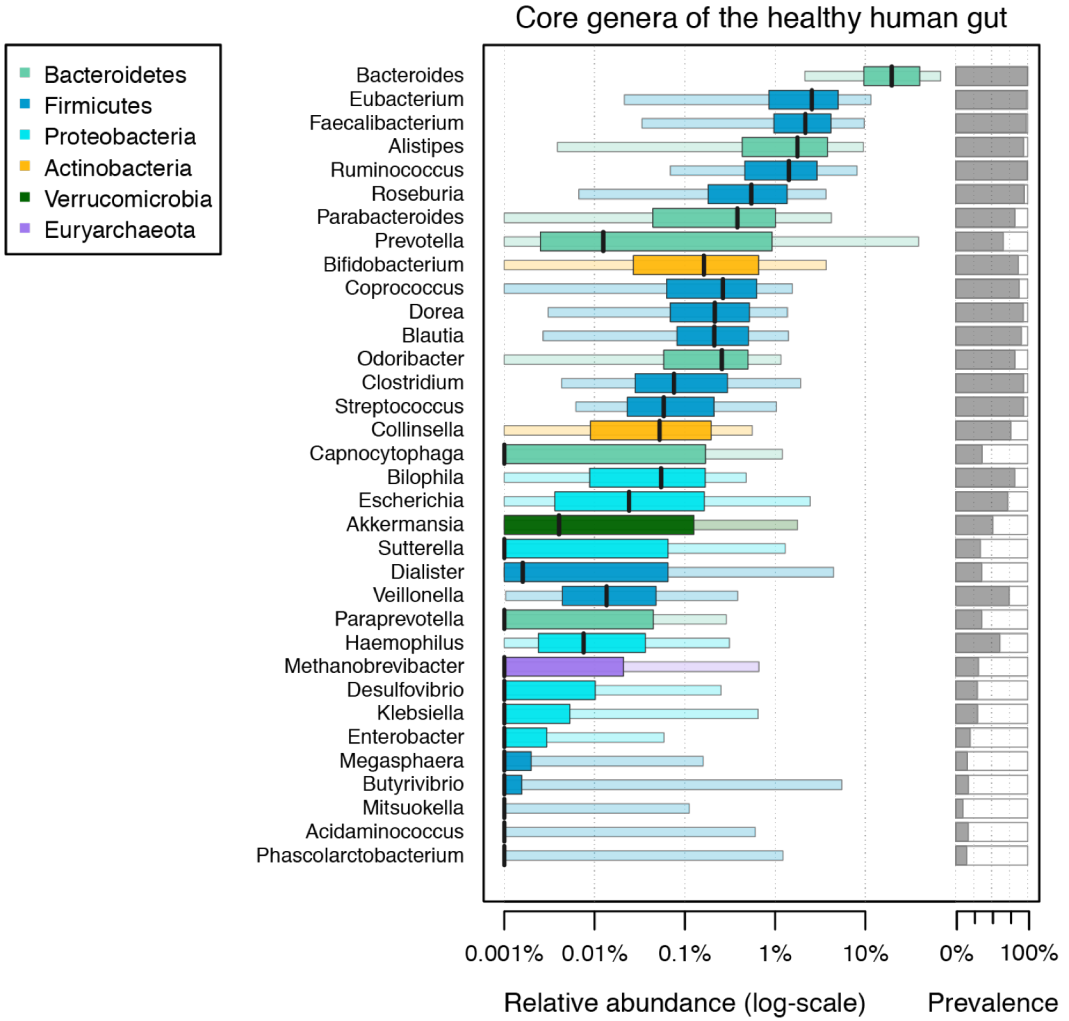
Understanding microbial abundance data



Average gut microbiome composition from phyla (inner ring) to genera (outer ring)

Individuality of gut microbiome composition

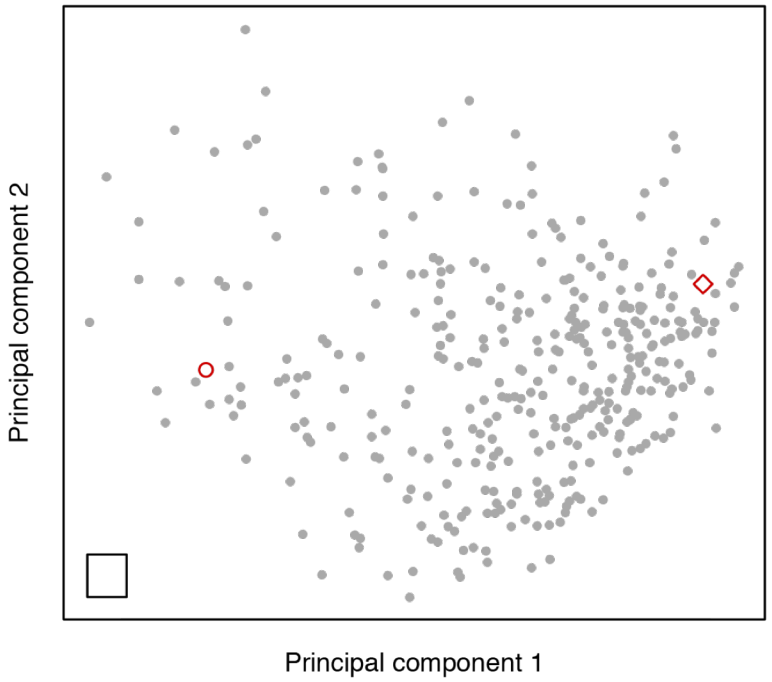
Based on 95 most abundant and prevalent gut species from 364 healthy individuals (stool samples, 3 continents)



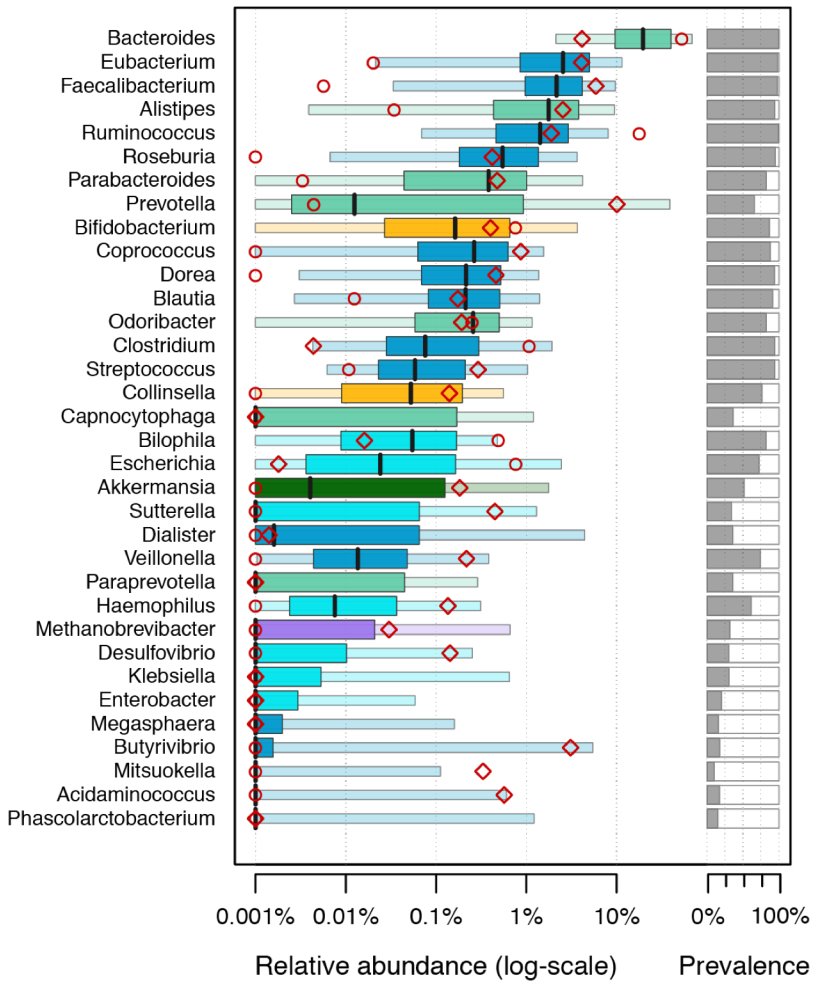
Individuality of gut microbiome composition

Based on 95 most abundant and prevalent gut species from 364 healthy individuals (stool samples, 3 continents)

Ordination of individual samples

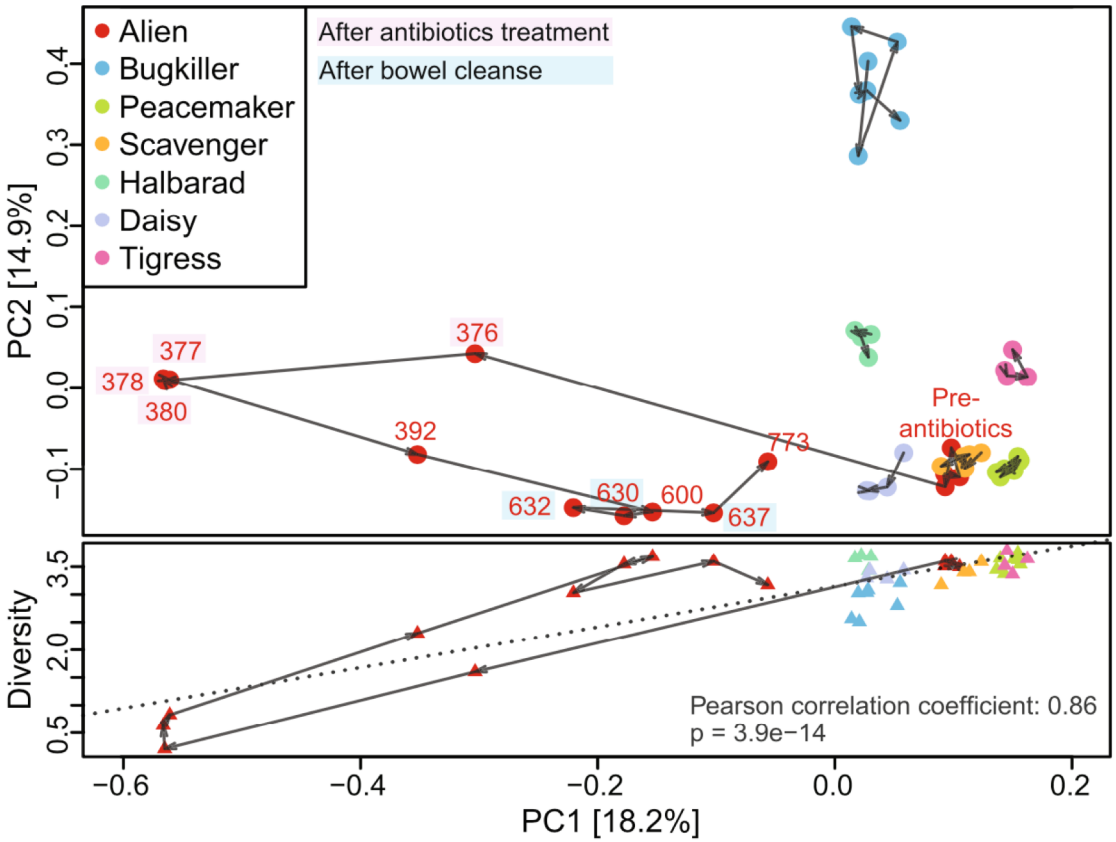


Core genera of the healthy human gut



Taxonomic abundances can vary extremely from subject to subject in many host-associated microbial communities

Temporal stability of community composition



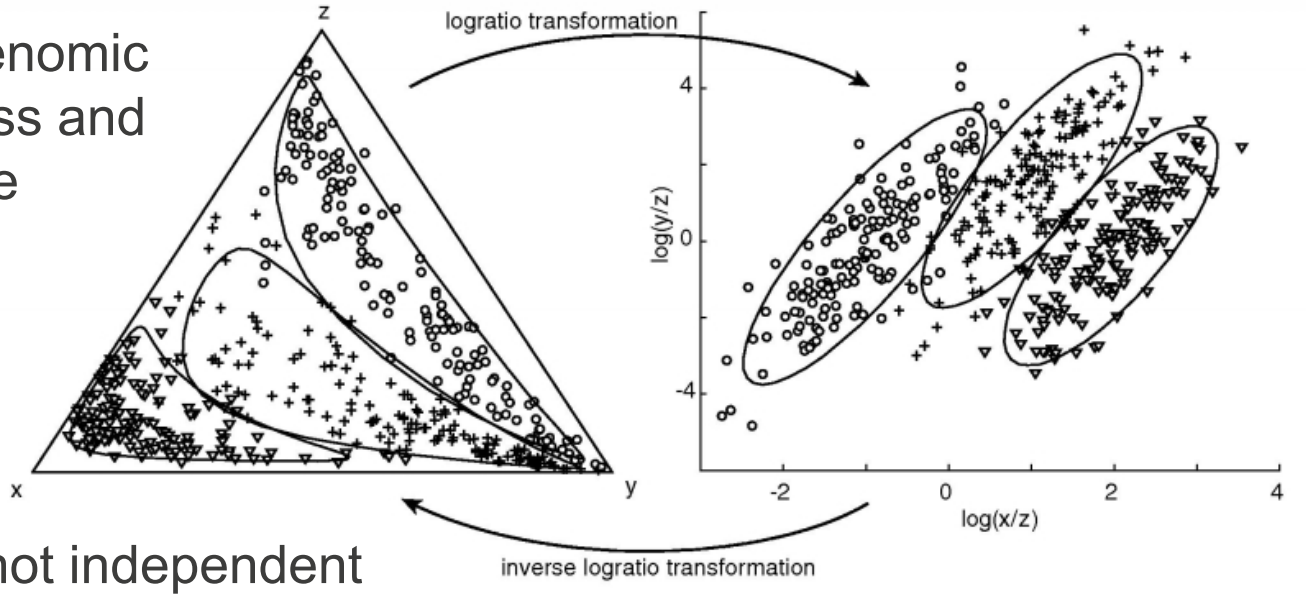
Gut microbiome composition is generally stable over time within an individual.

Changes in response to perturbations can be observed e.g. for antibiotics (as for “Alien” – red dots).

[1] Voigt A.Y. et al., *Genome Biol* 2015

Metagenomic data is compositional in nature

Due to the metagenomic sequencing process and relative abundance normalization for library size differences, the relative abundance of one species is not independent from those of other species.

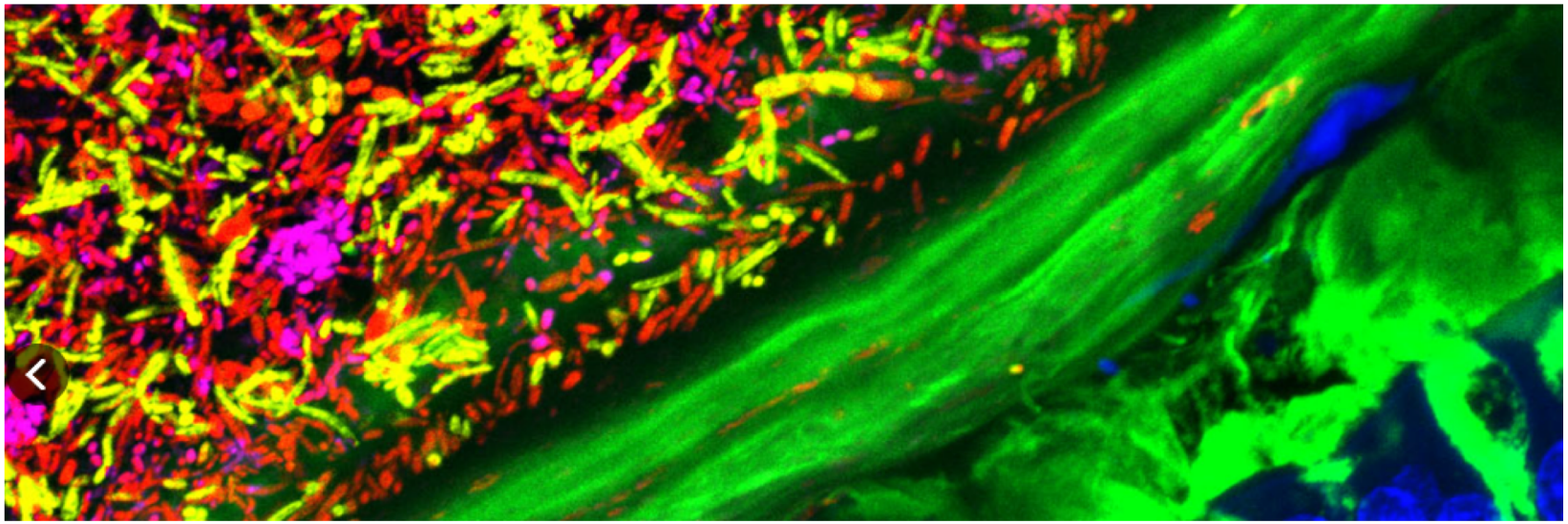


For example, if only the most abundant species doubles, the relative abundance of all others will be half.

Summary:
characteristics of human microbiome abundance data

- High inter-individual, low temporal variability
- Sparsity (many species are absent in a large proportion of subjects)
- Based on read-count statistics – *not* Gaussian
- Compositional data violates independence assumptions

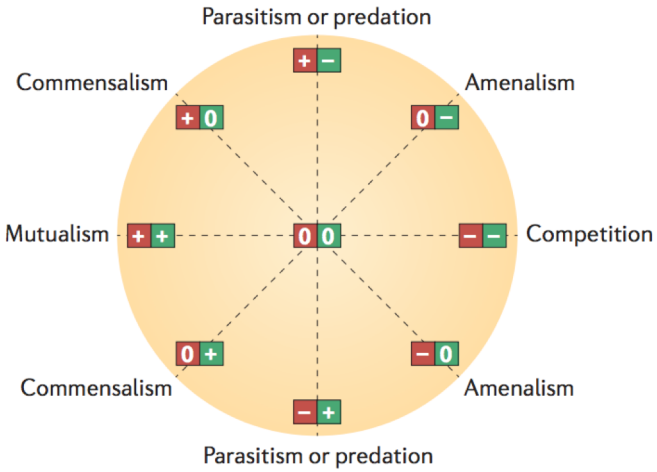
https://github.com/gezel/microbial_communities_workshop_2018



PART II:

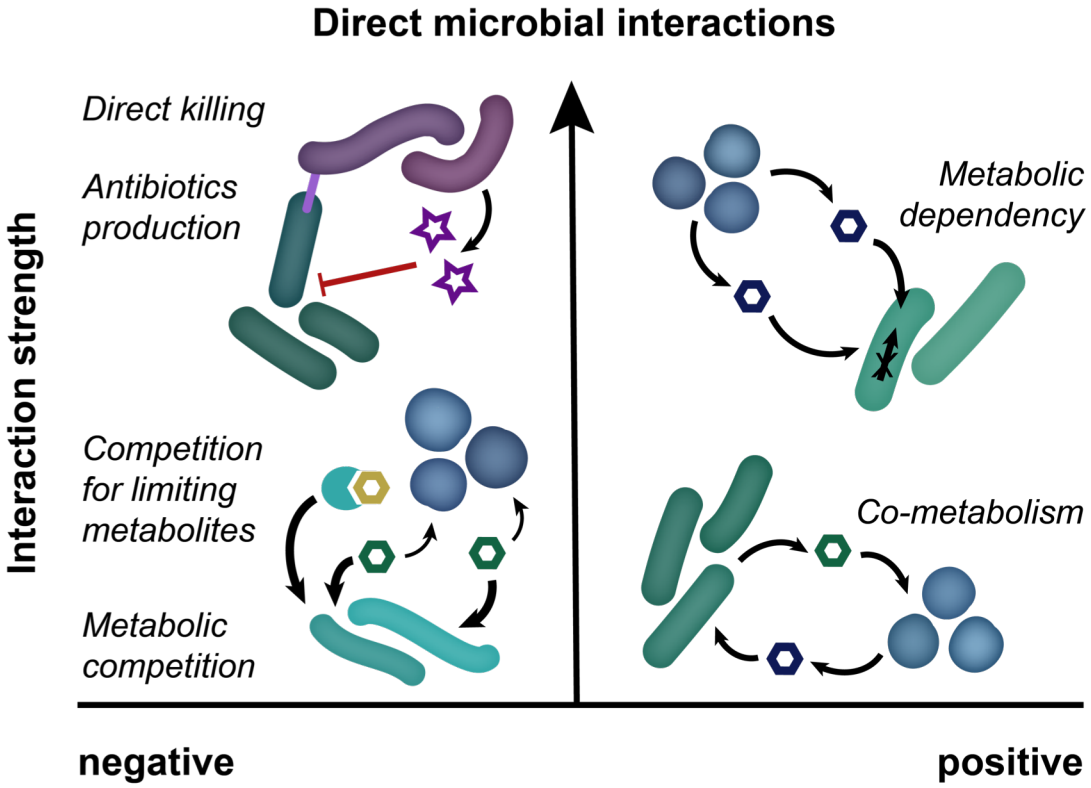
INFERRING SPECIES INTERACTION NETWORKS FROM CO-ABUNDANCE DATA

Types of species-species interactions in microbial communities



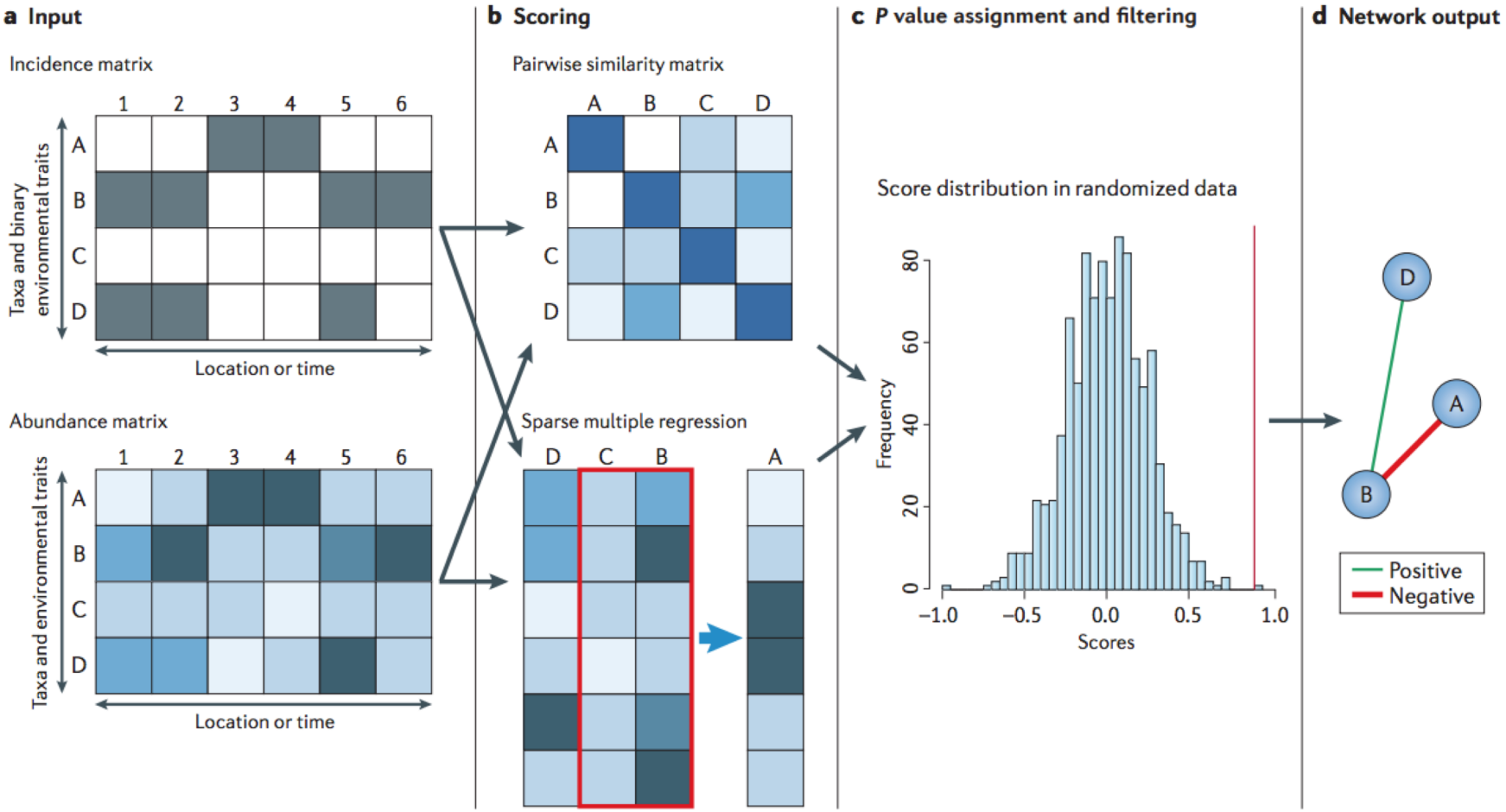
positive negative interactions

From co-abundance data, one can only infer static interactions!



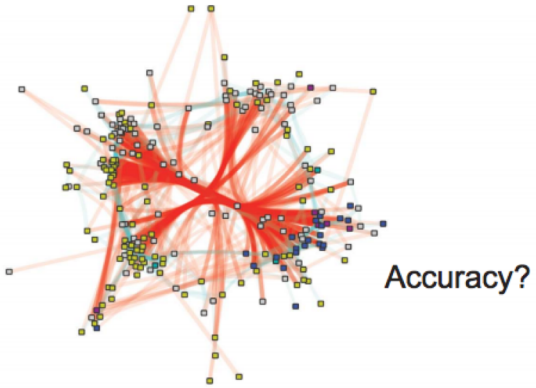
[1] Faust K. and Raes J., *Nat Rev Microbiol* 2012

Inference of microbial interactions from co-abundance data



[1] Faust K. and Raes J., *Nat Rev Microbiol* 2012

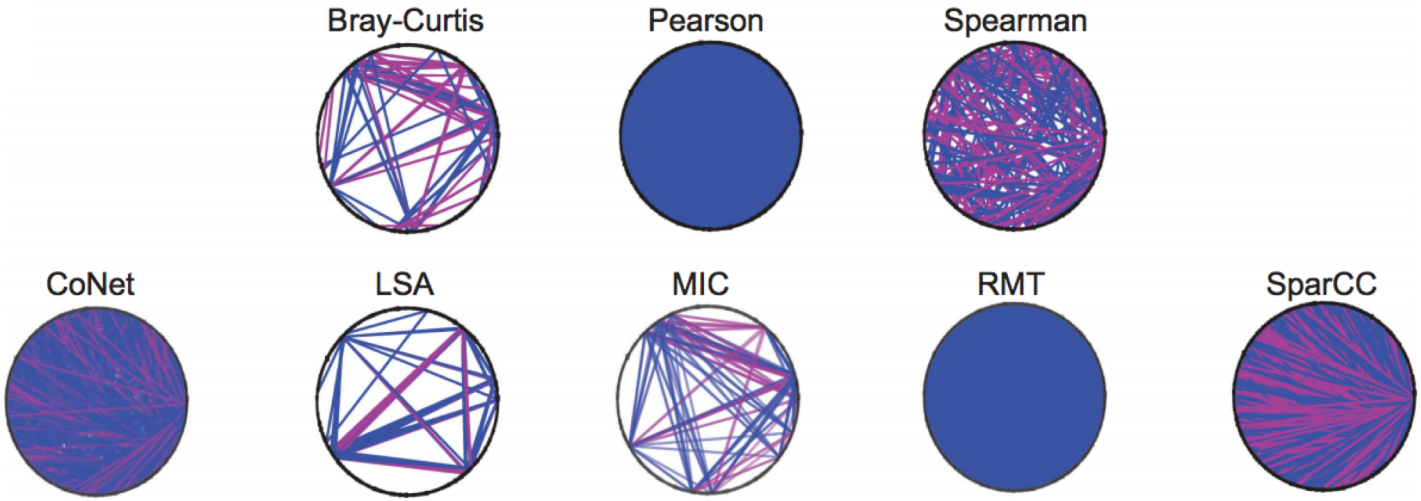
Why not use simple correlation networks?



Simple correlation coefficients yield very dense networks – likely containing many false positive interactions!

In particular they suffer from microbiome abundance data being

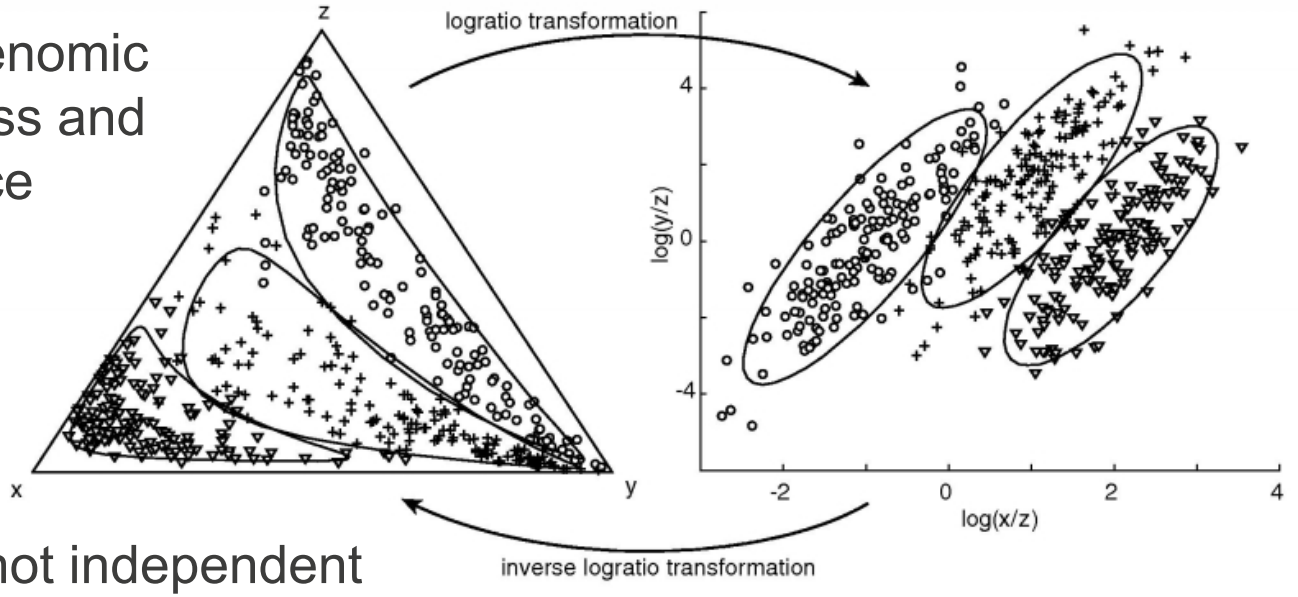
- non-Gaussian and
- compositional



[1] Weiss S. et al., *ISME J* 2016

Correlations in compositional in nature

Due to the metagenomic sequencing process and relative abundance normalization for library size differences, the relative abundance of one species is not independent from those of other species.

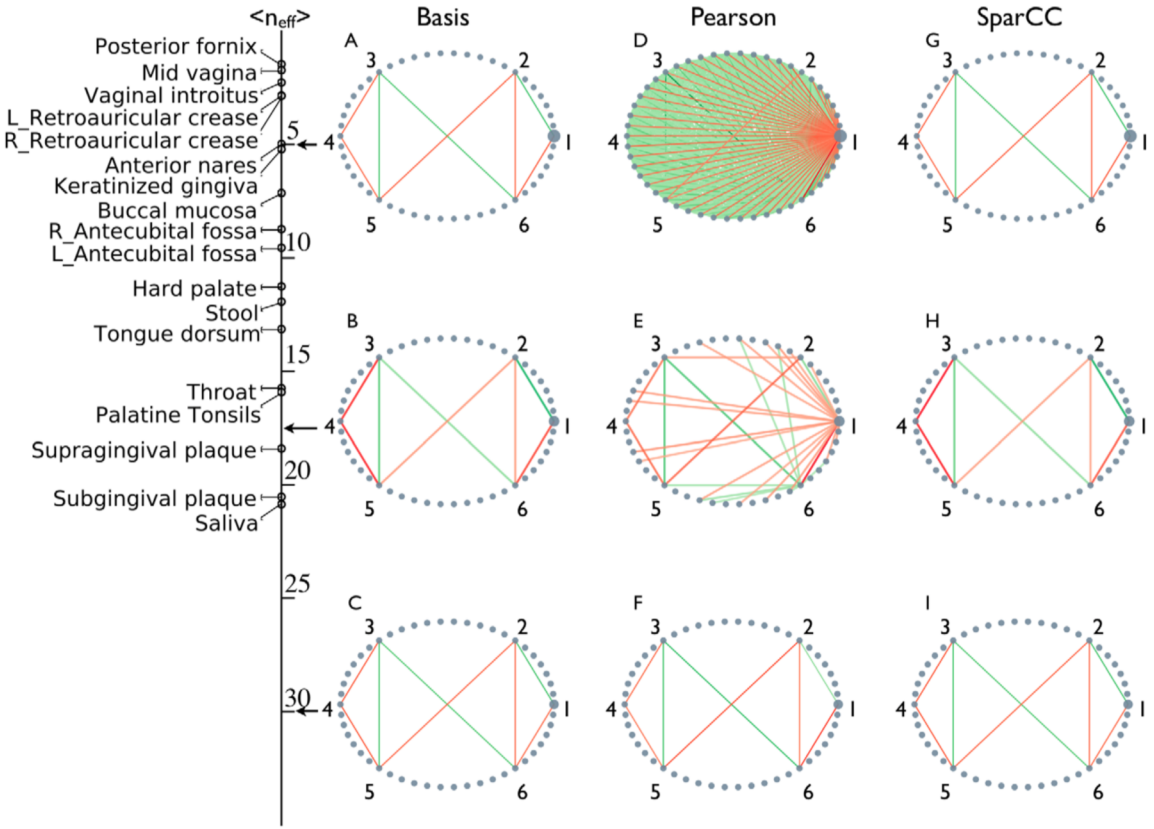


For example, if only the most abundant species doubles, the relative abundance of all others will be half. **The most abundant species varying across samples can be enough to cause the relative abundance of others to become correlated.**

- (Centered) log-ratio transformations of relative abundances can be used instead to compute meaningful correlations

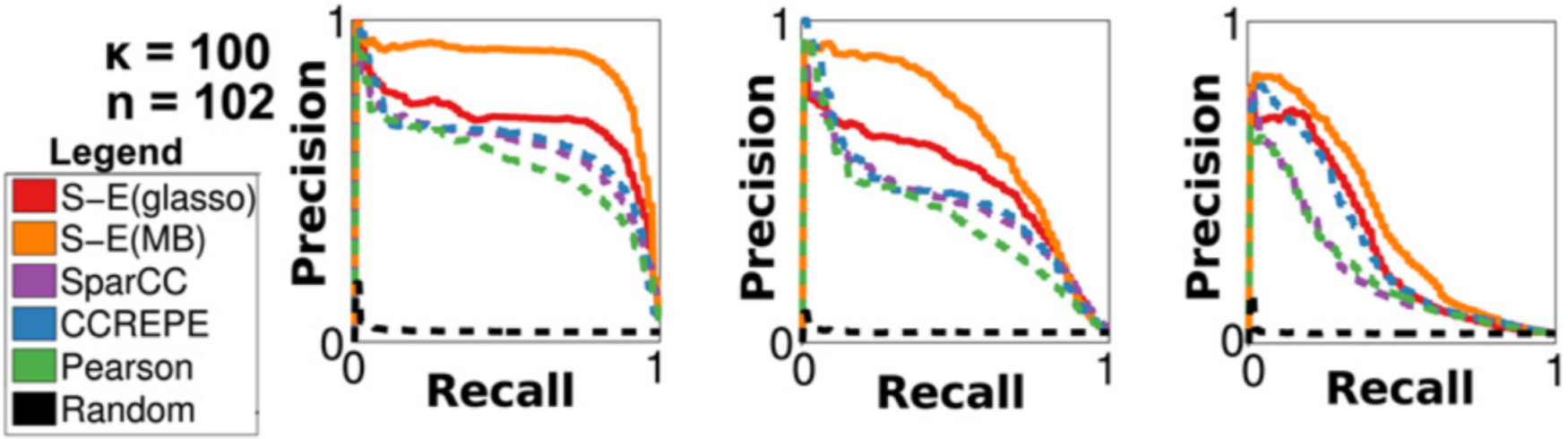
Inferring correlations while accounting for compositionality – SparCC

- Applies log-ratio transform to address compositionality
- Difference to naïve methods strongest in low-diversity communities (e.g. in the vagina) where compositionality effects are most pronounced.



[1] Friedman J. and Alm E., *PLoS Comput Biol* 2012

Sparse network inference – SPIEC-EASI



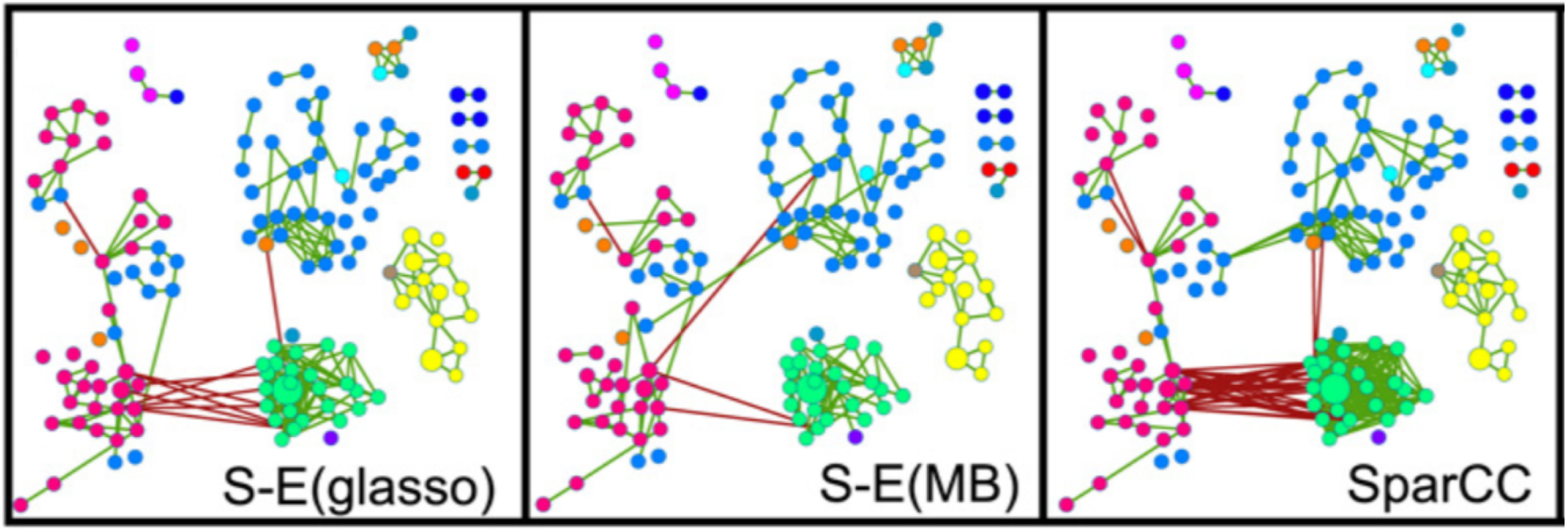
- Applies centered log-ratio transform.
- Infers a sparse network using a conditional independence approach and Graphical model inference.
- It models dependencies between species that cannot be explained by indirect correlations using sparse regression.
- Outperforms other methods on simulated data.

[1] Kurtz Z.D., Mueller C.L. et al., *PLoS Comput Biol* 2015

Network visualization

- Determine a threshold for correlation based method to visualize as edges those correlations with an absolute value above the threshold.
- Embed species as nodes in a 2D plot so that species connected by an edge are close to each other using graph layout algorithms.

Comparison of human gut networks inferred with SPIEC-EASI (S-E) and SparCC



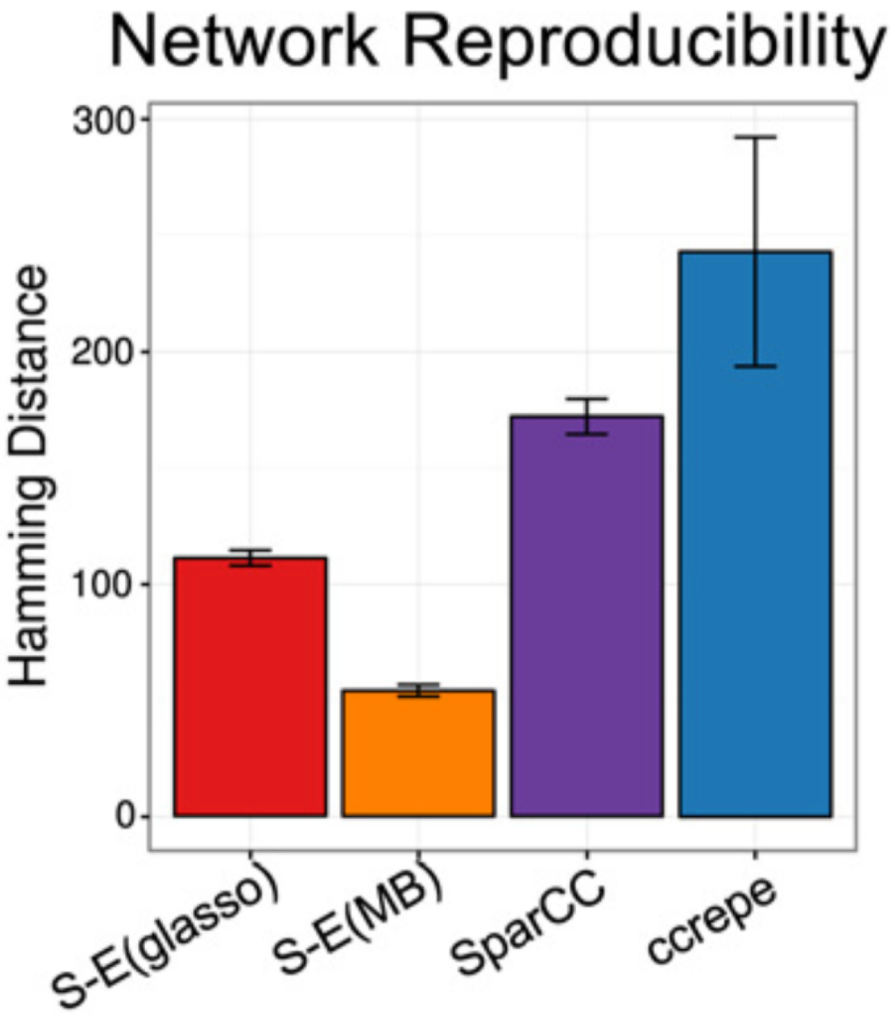
S-E networks for the human gut microbiome are sparser than those inferred by SparCC (and other methods not shown here)

[1] Kurtz Z.D., Mueller C.L. et al. *PLoS Comput Biol* 2015
[2] Friedman J. and Alm E., *PLoS Comput Biol* 2012

Assessing network robustness

An elegant way of assessing network robustness is based on splitting a data set and comparing the networks that were independently computed on partial data sets (in the figure by a distance on the edges).

SPIEC-EASI (S-E) network inference appears more reproducible on American Gut data subsets than SparCC (lower Hamming distance on edges)



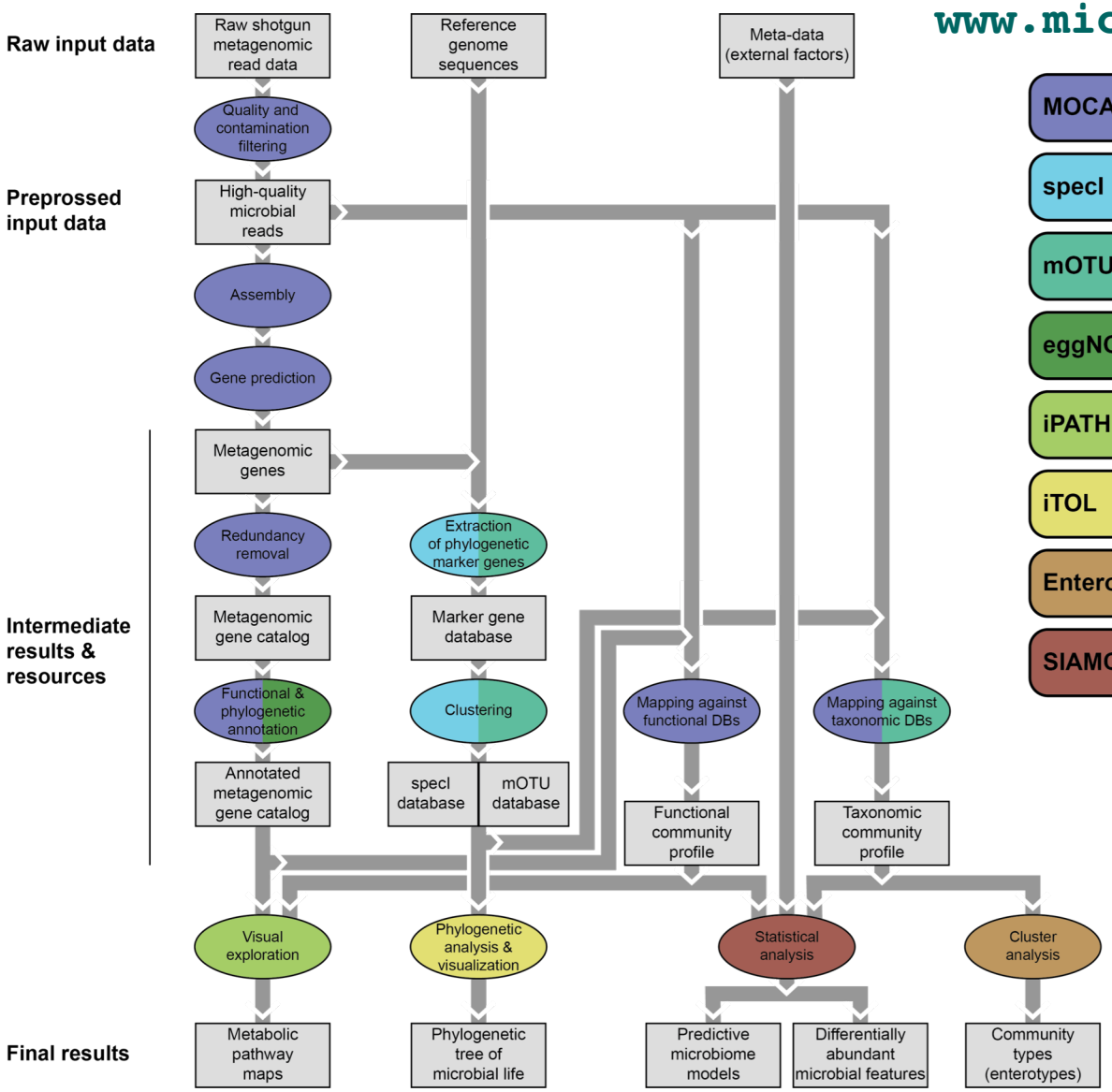
[1] Kurtz Z.D., Mueller C.L. et al., *PLoS Comput Biol* 2015

Summary of pitfalls in network inference and interpretation

- Having **more species and samples** (p>n) precludes some types of analysis
- **Non-Gaussian** data (careful with Pearson cor. / Euclidean dist.)
- **Sparse** data (co-absence can inflate correlation measures)
- **Non-linear** relationships (not easily captured by many methods)
- **Compositionality** (can lead to spurious correlations if not accounted for)
- Determining **significance** usually relies on permutation-based null models and thus **depends on the exact permutation approach**
- Massive **multiple hypothesis testing** needs to be corrected for
- Distinguishing direct from **indirect interactions** is difficult
- **Reproducible/robust network inference** is an issue for many methods
- There's still **a long way between network construction and interpretation** in terms of microbial ecology, co-metabolism or chemical warfare

[1] Faust K. and Raes J., *Nat Rev Microbiol* 2012
[2] Weiss S. et al., *ISME J* 2016
[3] Friedman J. and Alm E., *PLoS Comput Biol* 2012
[4] Kurtz Z.D., Mueller C.L. et al., *PLoS Comput Biol* 2015

Bioinformatics tools for microbiomics



www.microbiome-tools.embl.de

- **MOCAT** — Metagenomics profiling pipeline
- **specI** — Prokaryotic species identification
- **mOTUs** — Metagenomic species profiling
- **eggNOG** — Resource for orthologous genes
- **iPATH** — Cellular pathway mapping tool
- **iTOL** — Interactive tree visualization
- **Enterotyping** — Gut microbiome community typing
- **SIAMCAT** — Statistical analysis of microbiome data

We are constantly improving usability and interoperability within the framework of

de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE

THANK YOU

https://github.com/gezel/microbial_communities_workshop_2018

zeller@embl.de