

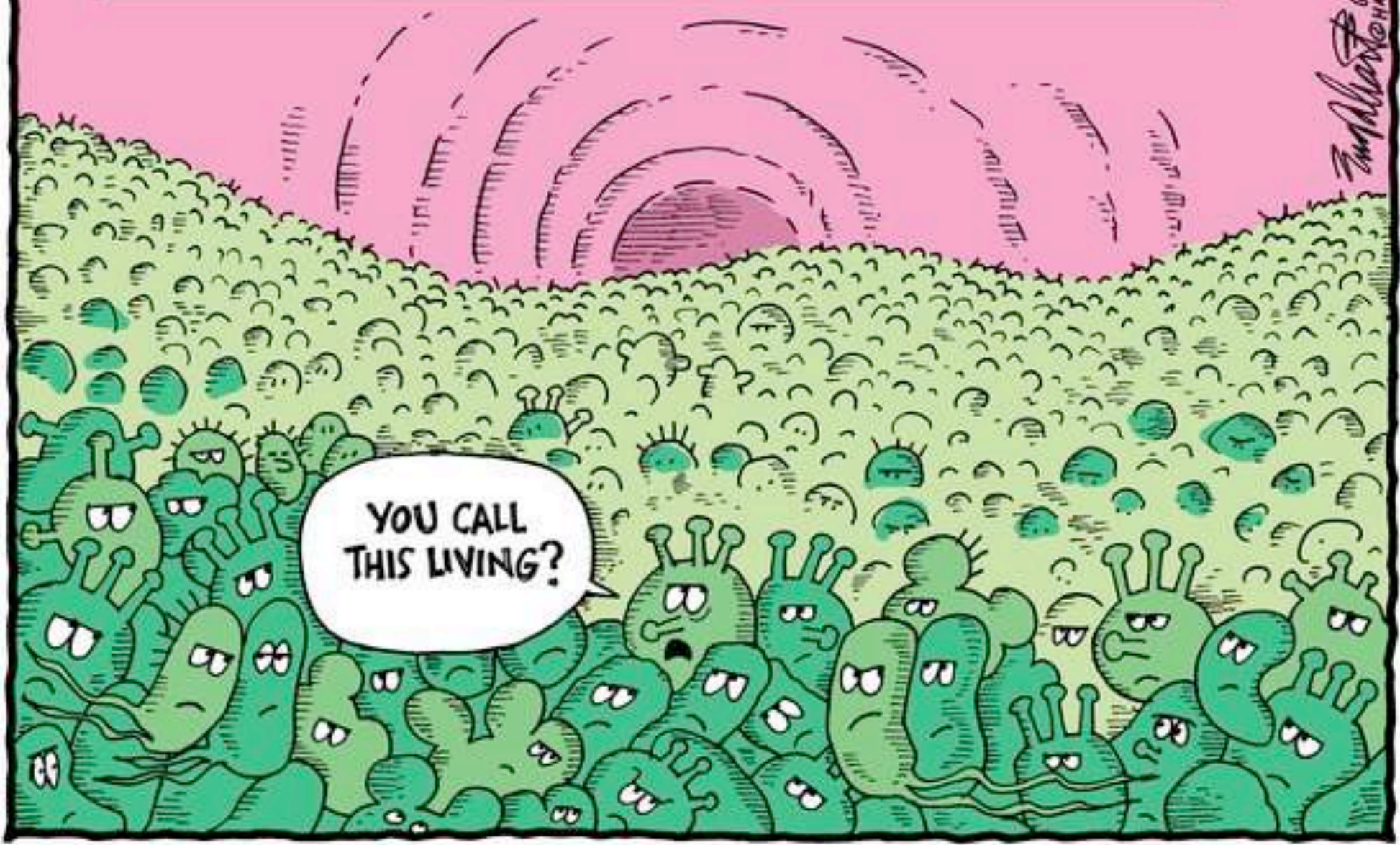
# Amplicon sequencing / Metagenomics

Isheng Jason Tsai

Introduction to NGS Data and Analysis  
v2020



THE HUMAN MICROBIOME PROJECT SAYS THE HUMAN BODY HAS 100 TRILLION MICROSCOPIC LIFE FORMS LIVING IN IT.



6/15/12  
HARTFORD COURANT

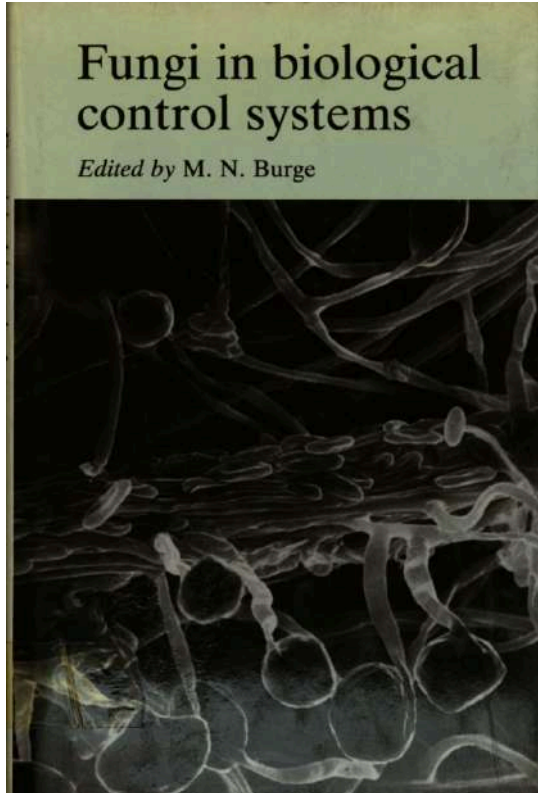
# Lecture outline

- Introduction
- Amplicons
- Diversity measures
- Metagenomics
- Caveats
- Case studies



# What is the microbiome?

Fungi in Biological Control Systems (1988)



A convenient ecological framework in which to examine biocontrol systems is that of the microbiome. This may be defined as a characteristic microbial community occupying a reasonably well defined habitat which has distinct physico-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatres of activity. In relation to fungal diseases of crops and their control, major microbiomes are the phylloplane, spermosphere, rhizosphere and rhizoplane, and numerous kinds of plant residues persisting on or in the soil. Mention should also be made of the wood of standing or felled trees as microbiomes where biocontrol of forest diseases using fungi has been achieved. However, in most cases competitive interactions other than mycoparasitism seem to be of greater importance.

<http://microbe.net/2015/04/08/what-does-the-term-microbiome-mean-and-where-did-it-come-from-a-bit-of-a-surprise/>



# And then what is the metagenome?

Crosstalk R245

## **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products**

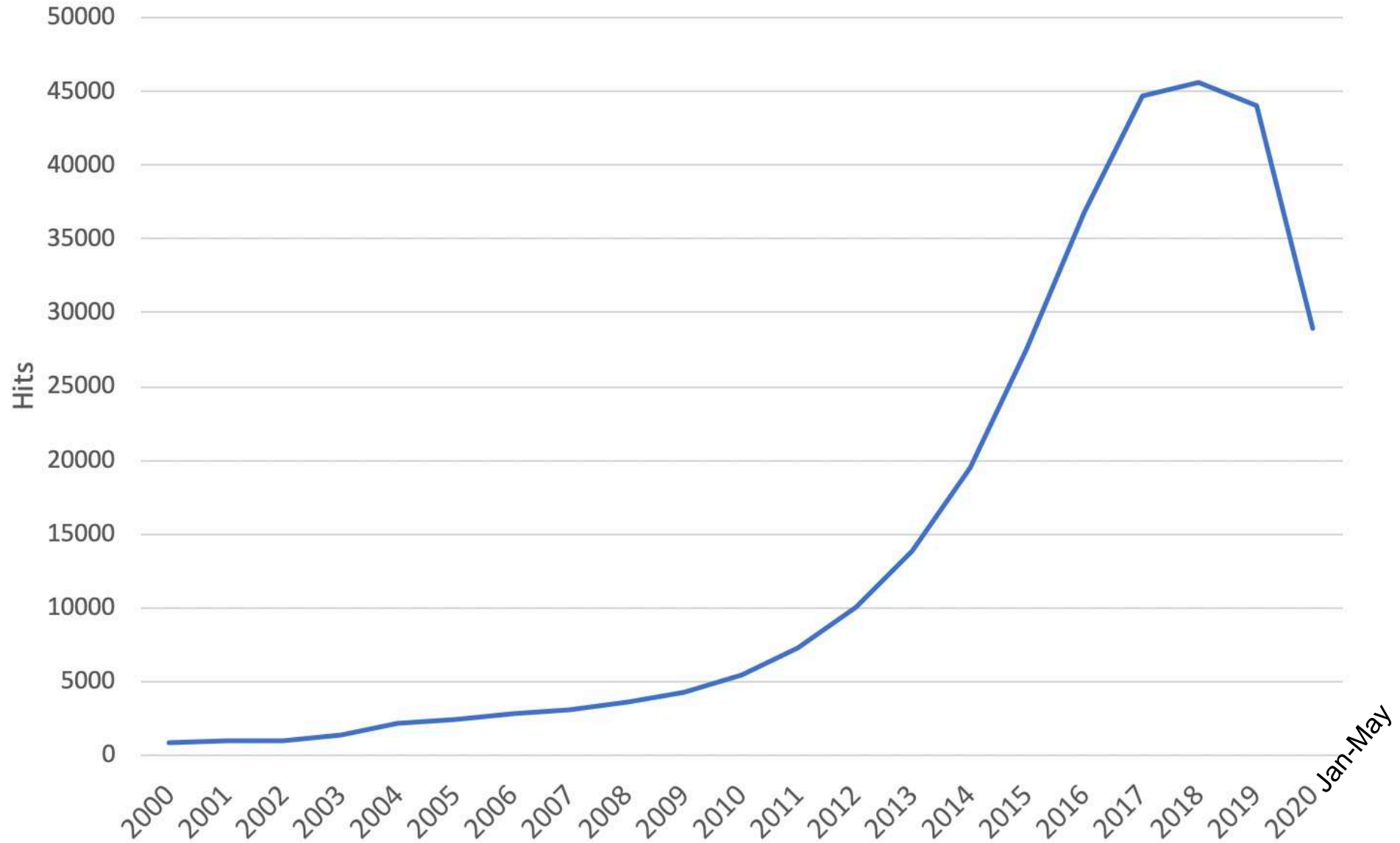
Jo Handelsman<sup>1</sup>, Michelle R Rondon<sup>1</sup>, Sean F Brady<sup>2</sup>, Jon Clardy<sup>2</sup> and Robert M Goodman<sup>1</sup>



**Chemistry & Biology** October 1998, 5:R245–249  
<http://biomednet.com/elecref/10745521005R0245>

**... This approach involves directly accessing the genomes of soil organisms that cannot be, or have not been, cultured by isolating their DNA**

# Google scholar hits for “Microbiome”



# Basic Purpose

Characteristics of (microbial) community

**Who** are they? (species identification ; contain what genes)

**Where** do they come from?

Are their similarities (at what level)

between communities

of different conditions

of similar conditions?

within a community?

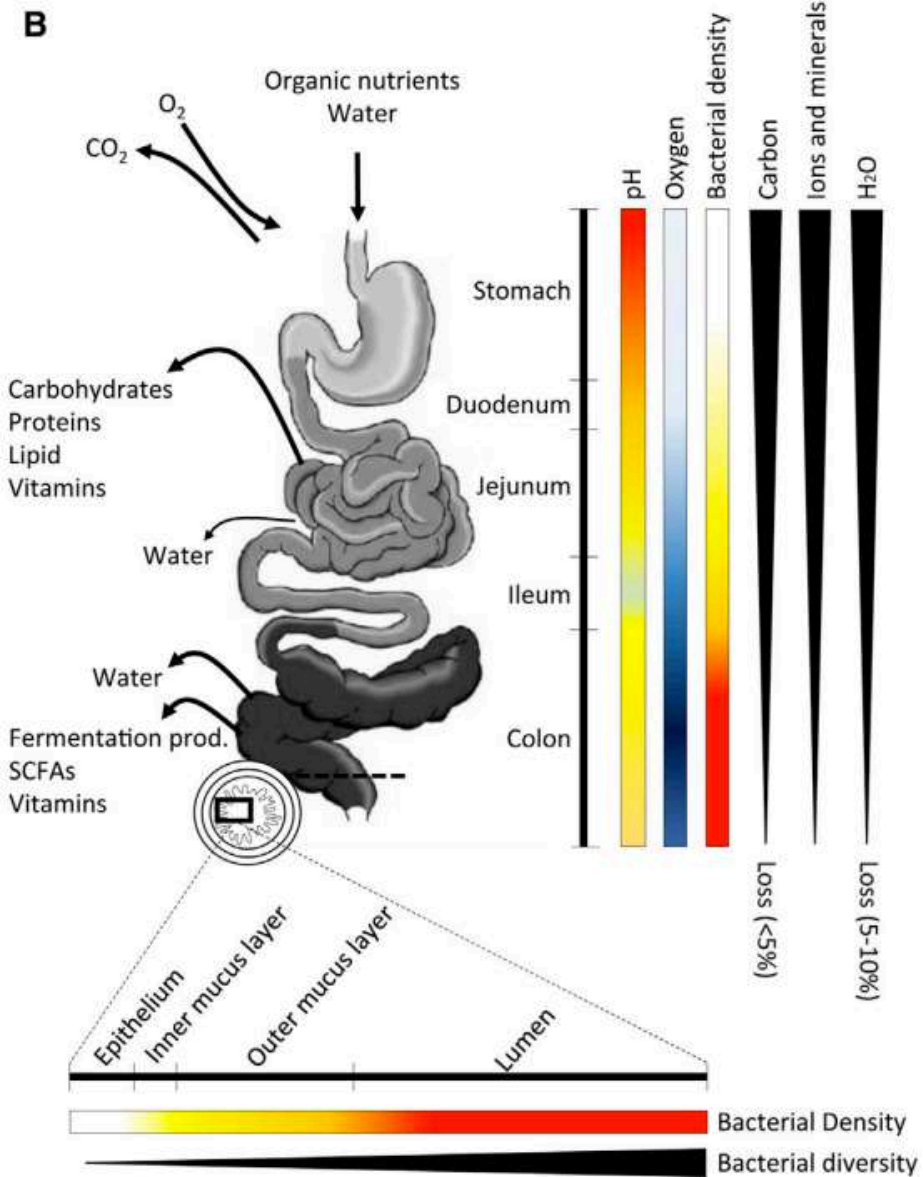
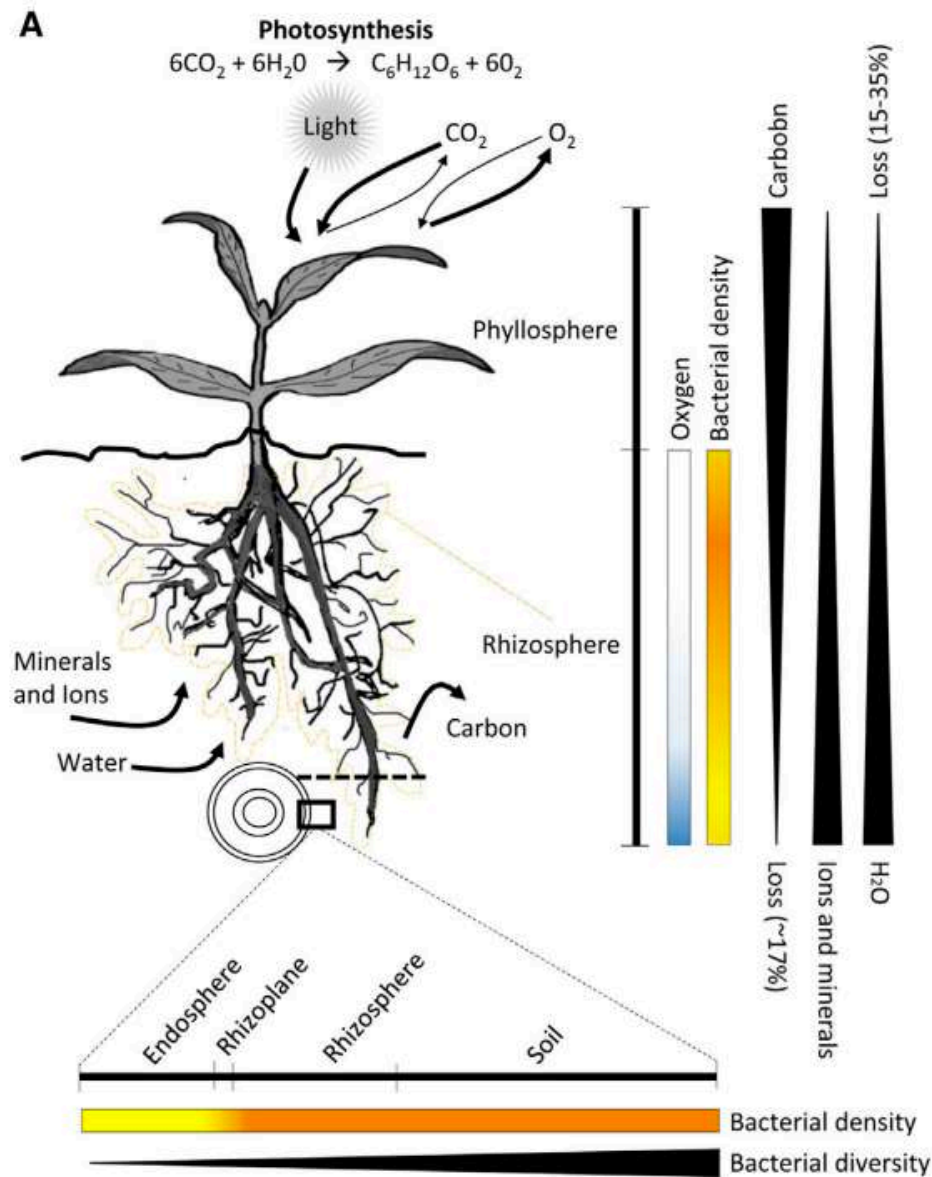
over a time period?

**What** are they doing?

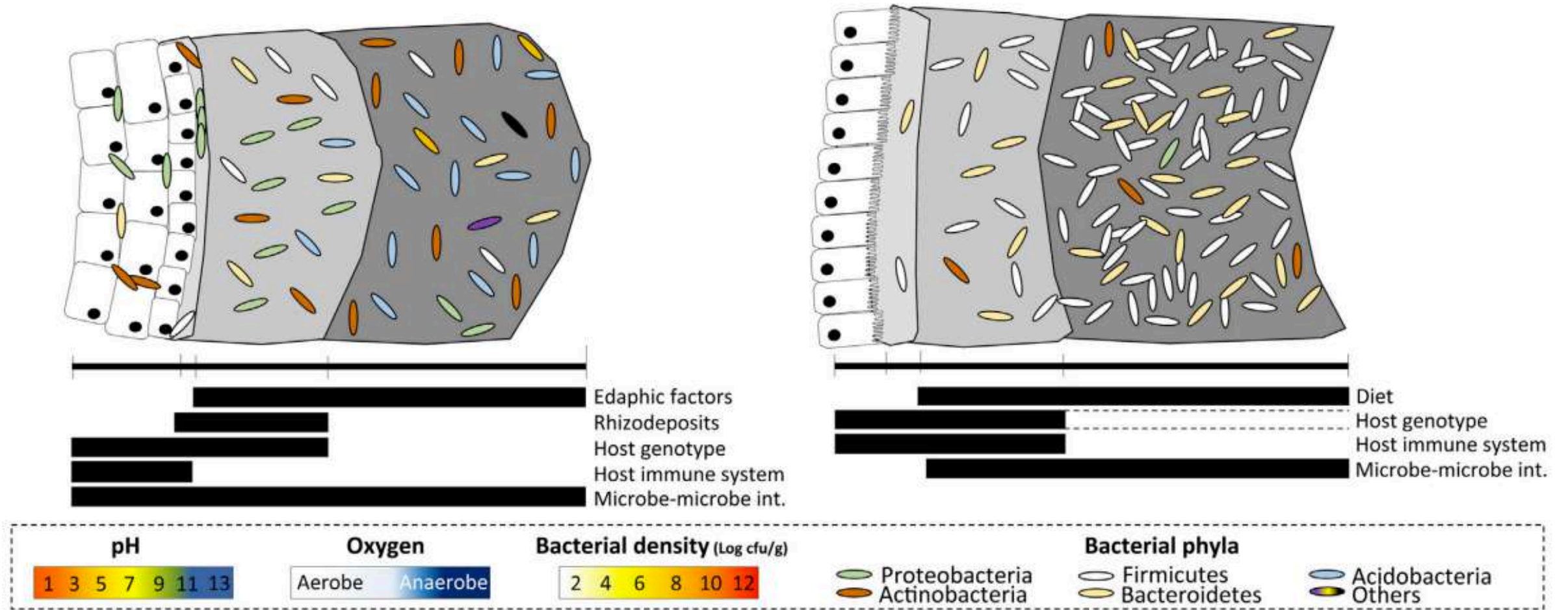
**How** are they doing? (factors influencing the community)



# Two most commonly studied systems



# Two most common systems



# Two most common systems

**Table 1. Percentage of Shotgun Metagenome Reads Assigned to Each Kingdom of Life across Metagenome Studies**

	Cucumber <sup>a</sup>	Wheat <sup>a</sup>	Soybean <sup>b</sup>	Wheat <sup>c</sup>	Oat <sup>c</sup>	Pea <sup>c</sup>	Barley <sup>d</sup>	Gut <sup>e</sup>
Bacteria	99.36	99.45	96	88.5	77.3	73.7	94.04	99.1
Archaea	0.02	0.02	<1	<0.5	<0.5	<0.5	0.054	
Eukaryotes	0.54	0.48	3	3.3	16.6	20.7	5.90	<0.1

<sup>a</sup>Ofek-Lalzar et al. (2014) (metagenomics of rhizoplane samples).

<sup>b</sup>Mendes et al. (2014) (metagenomics of rhizosphere samples).

<sup>c</sup>Turner et al. (2013) (metatranscriptomics of rhizosphere samples).

<sup>d</sup>Bulgarelli et al. (2015) (metagenomics of rhizosphere samples).






<sup>e</sup>Qin et al. (2010) (metagenomics of gut samples).



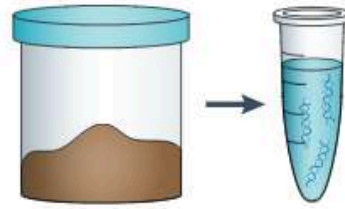
# Useful terminology

Term	Definition	Examples	Reference
Biome	The world's major ecosystems, defined by temperature gradients in latitude and altitude, precipitation and seasonality.	Subtropical, Mediterranean, Polar	Walter and Box (1976)
Microbiome	An assemblage of microorganisms existing in or associated with a habitat; includes active and interacting member as well transient or inactive members.	Human microbiome, Earth microbiome, Lake Erie microbiome, Soil microbiome	Lederberg and McCray (2001)
Core microbiome	Requires qualifiers for locality and habitat of interest. Organisms common across microbiomes hypothesized to play a key role in ecosystem function within a habitat.	To be determined	Turnbaugh et al. (2007)
Habitat	The physical and chemical parameters of an environmental area that determine niche spaces.	Termite hindgut: physical structure, anaerobic conditions, acidity and cellulose availability permit Spirochaeta abundance. Siberian tundra: subarctic temperatures, short summers and low solar energy permit boreal forest predominance.	Whittaker et al. (1973)
Ecosystem	The interactions and dynamics of physical, chemical and biological components of a locality.	Cornfield, Temperate forest, Permafrost, Tidal marsh, Mouse oral cavity	Odum (1953)
Locality	A spatially defined environmental area.	New York City, Lake Michigan, Soil core from a pea field, Vagina of a human subject	Andrewartha and Birch (1984)
Microbial community	Microorganisms that are co-existing and interacting with flanking microbiome members and/or the environment.	Algal mat, Biofilm, Dental plaque	Little et al. (2008)
Niche space	The activity range of a population along the physical and chemical dimensions of a habitat.	Anoxic aquatic or sediment niche spaces for sulfate reducers and methanogens	Hutchinson (1957)
Population	All the organisms belonging to the same species/operational taxonomic unit that live in a locality. For microbes, the species definition will vary by the <b>genes</b> of interest; a strain could be a population, as well as mutant and wild-type organisms.	Grizzly bear in Alaska Sulfolobus islandicus strain in a geothermal pool Wild-type Enterococcus faecalis	Waples and Gaggiotti (2006)
Connectivity	The amount or proportion of interactions within a system. Within a microbiome, this could include interactions among taxa (biotic interactions) or with environmental factors (abiotic) within the locality.	Quorum sensing, Predator-prey dynamics, Resource competition	Gardner and Ashby (1970)
Observation	The sampling unit, includes metadata of locality, habitat, time and/or experimental condition.	Meta-transcriptome of one location in an acid mine drainage site, collected at one time point. 16S tag-sequences from the right palm of one human subject.	
Persistent	Organisms that are consistently detected within a locality through time.	Pseudomonas aeruginosa in Cystic Fibrosis patients Firmicutes in infant guts	

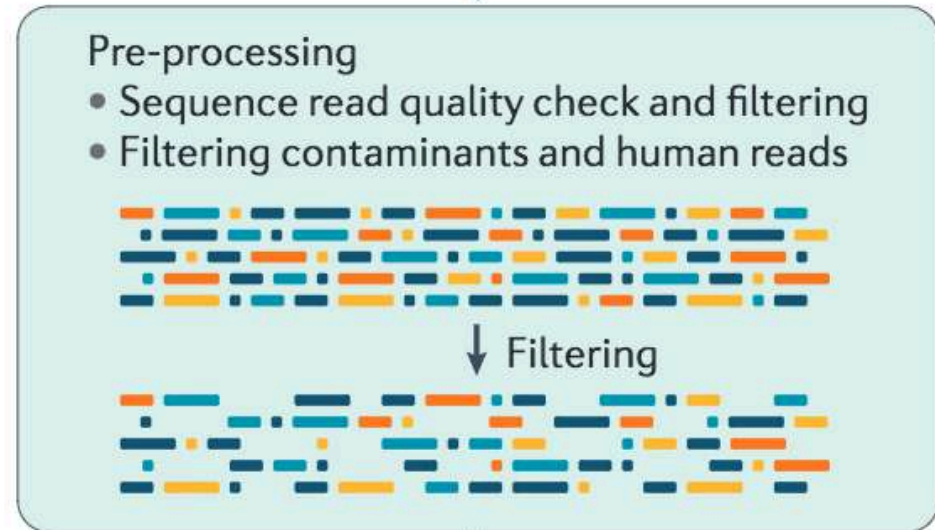
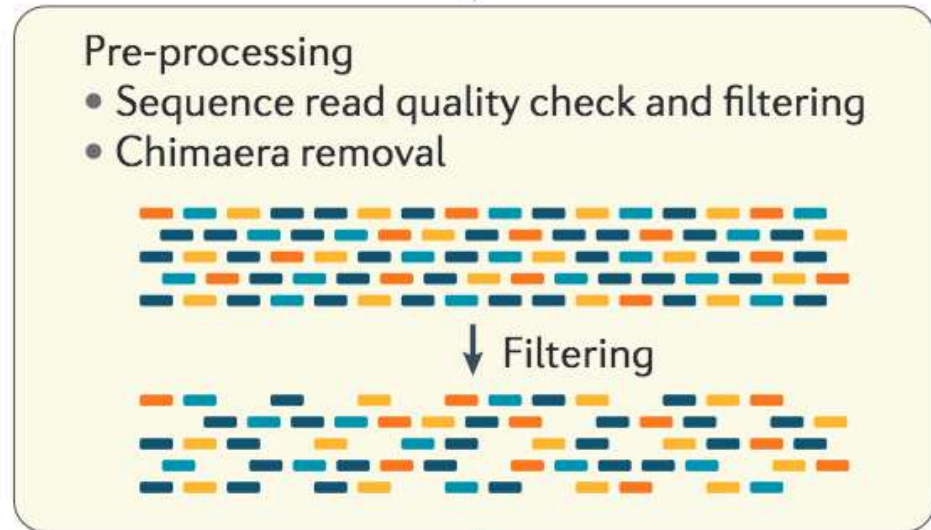
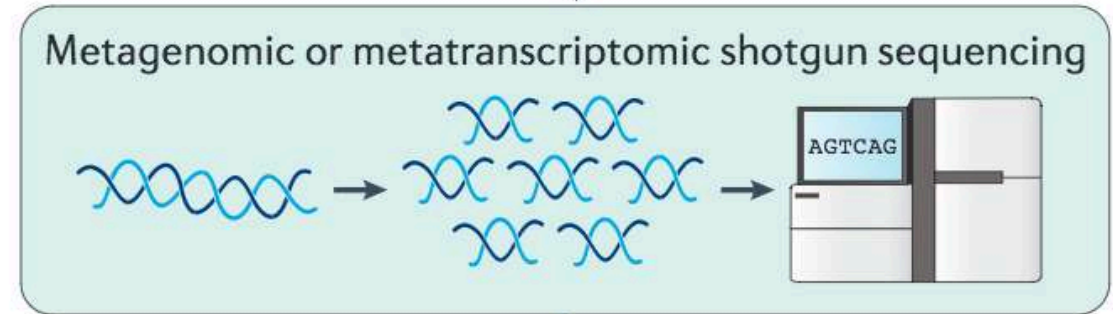
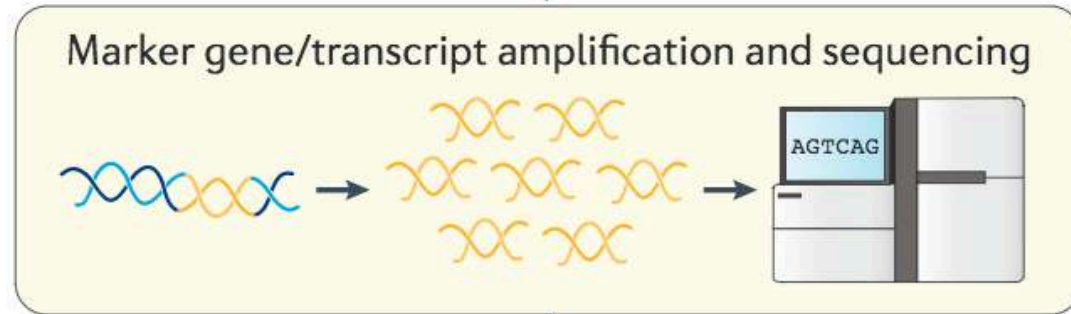
# Things to consider

Study design	In the field	In the laboratory		At the keyboard
 <p>Basic science or applied? (e.g., environmental biomonitoring)</p> <p>What is your study goal?</p> <ul style="list-style-type: none"><li>• presence/absence</li><li>• diversity assessment</li><li>• absolute quantification</li></ul> <p>What taxa will you target?</p> <p>Is the scale of inference for your sample type appropriate to your question?</p> <p>Can you compare complementary data types? (e.g. traditional vs. eDNA)</p> <p>Does your sampling/replication scheme provide good statistical power?</p>	 <p>What type of sample is needed? (water, soil, air)</p> <p>What metadata should you collect?</p> <p>How many replicates will you collect?</p> <p>Does your sampling protocol minimize/control for :</p> <ul style="list-style-type: none"><li>• contamination (e.g., positive and negative controls)</li><li>• any known biases (e.g., inhibitors, sample volume)</li></ul>	 <p><b>Sample Handling Phase</b></p> <p>What extraction method? (physical vs. chemical)</p> <p>How much sample?</p> <p>What locus and primers?</p> <p>Do you need to generate reference sequence data?</p> <p>Are technical replicates needed?</p> <p>What library preparation method will you use?</p> <p>How many samples will you index and pool?</p> <p>What sequence depth is needed per sample ?</p> <p>What read length will you use?</p>	 <p><b>DNA Processing Phase</b></p> <p>What sequencing platform will you use?</p> <p>Do you need paired-end sequencing?</p> <p>Have you included appropriate quality assurances? (e.g., mock community, qPCR, bioanalyser traces)</p> <p>Does your laboratory protocol minimize/control for:</p> <ul style="list-style-type: none"><li>• contamination (e.g., positive and negative controls)</li><li>• any known biases (e.g., primer bias, coverage, taxonomic resolution)</li></ul>	 <p>How complete is the reference database?</p> <p>Do you have adequate sequencing coverage across samples?</p> <p>Are you using appropriate choices for software tools, parameters?</p> <p>Are your biological conclusions upheld using alternative parameters and workflows?</p> <p>Are you including appropriate quality filtering of your data? (see Box 2)</p>

# Workflow

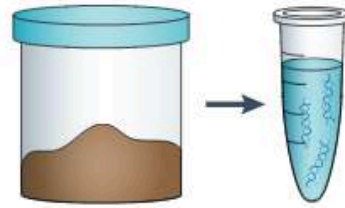


Study design, sample collection, storage and DNA/cDNA/RNA extraction

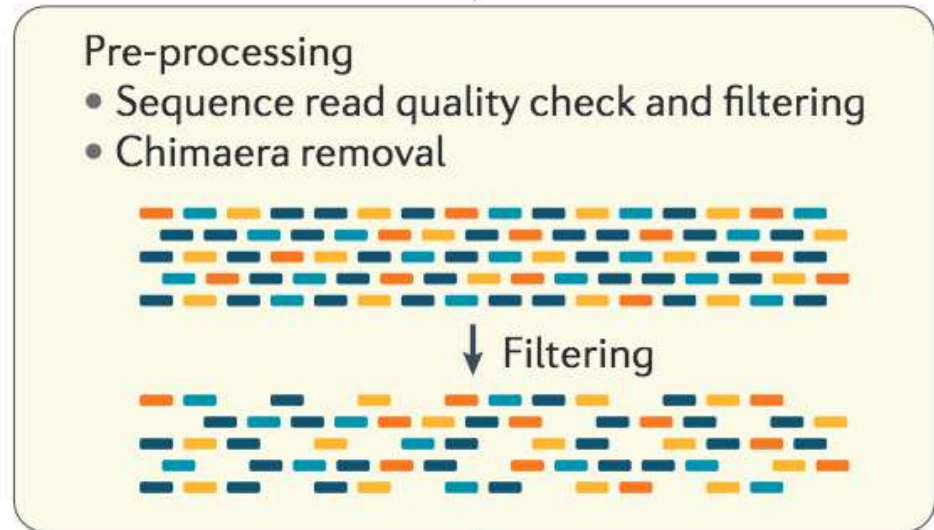
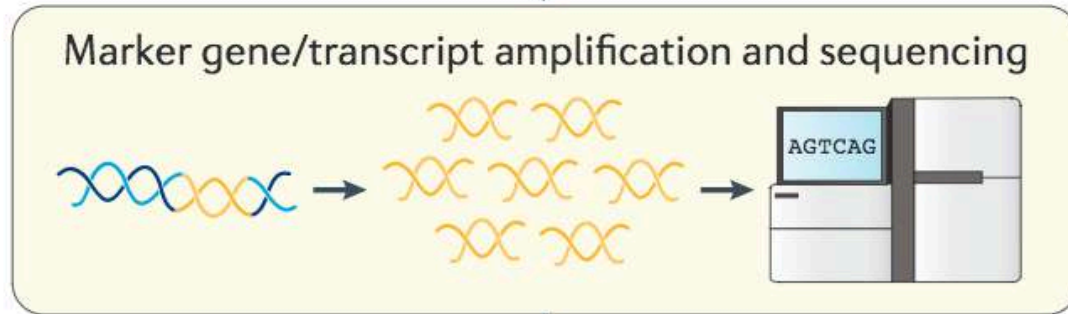




# Workflow



Study design, sample collection, storage and DNA/cDNA/RNA extraction



Before we consider metagenomics, which require a lot of sequencing (more expensive, more challenging to analyse), we may consider a easier but also very useful approach.

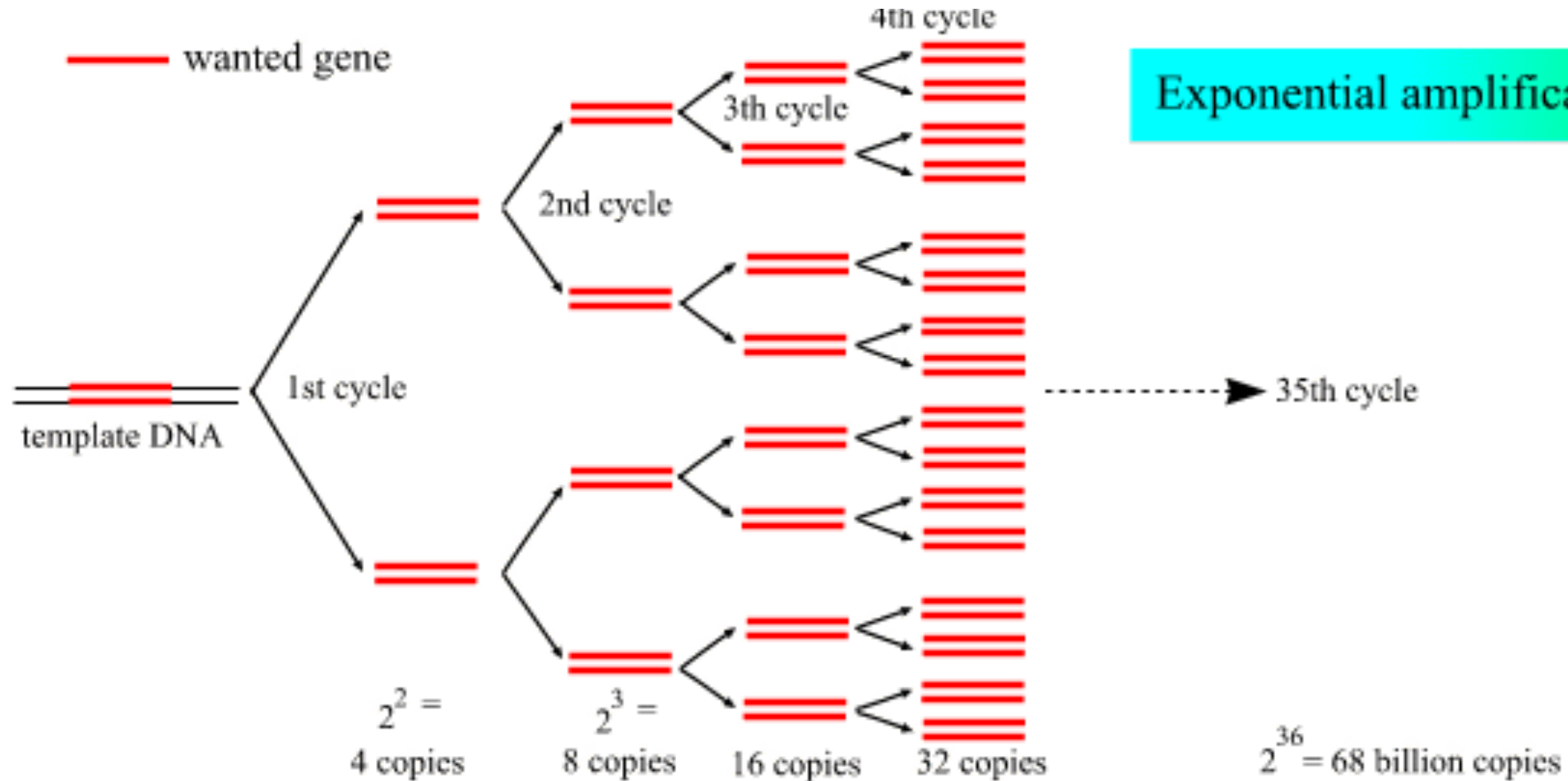
# Amplicon sequencing

Probably the most important point  
of the lecture

Metagenomics  $\neq$  Amplicon sequencing

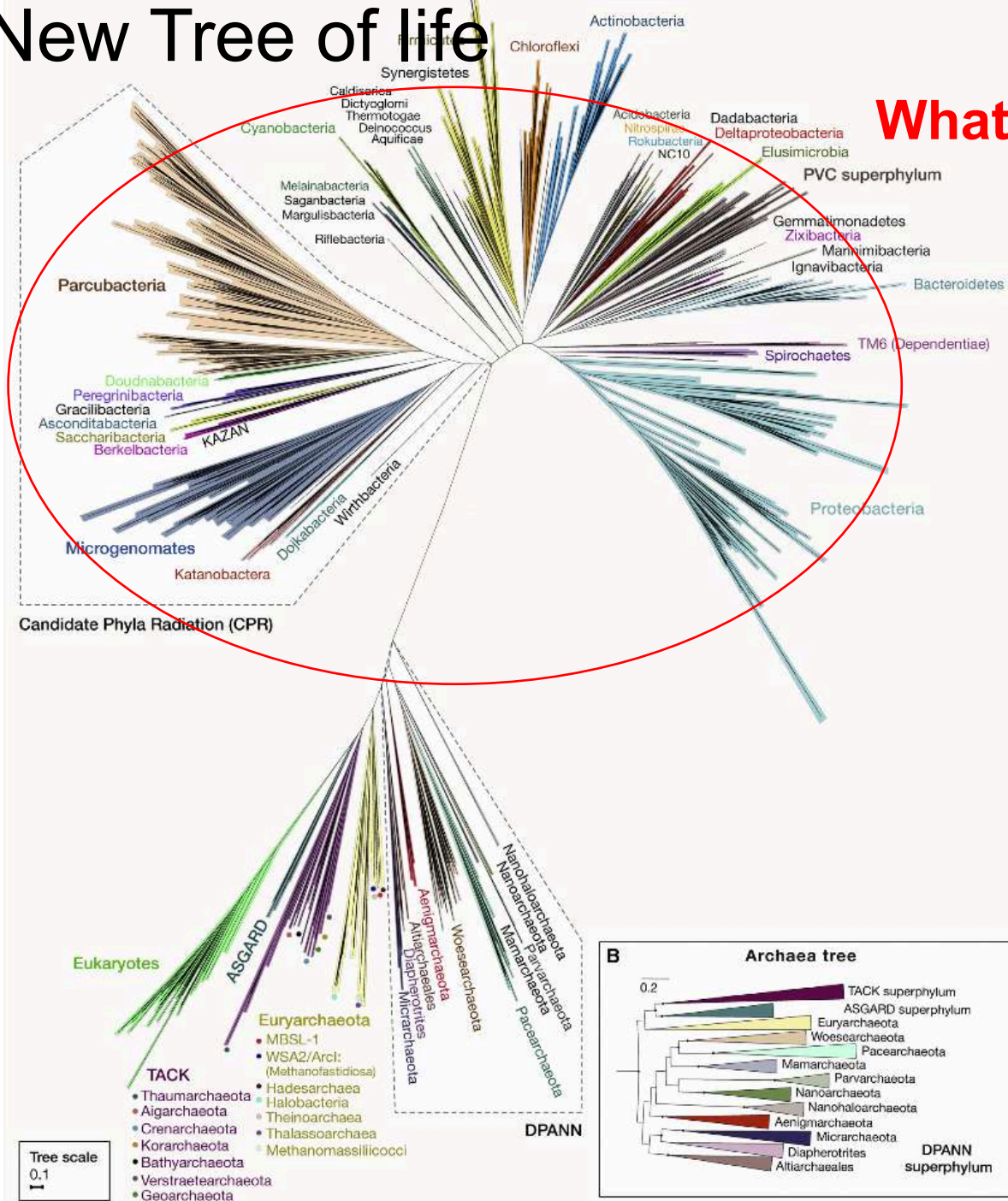
# What is amplicon sequencing?

Anything that requires PCR-based amplification of a specific target gene (locus)





# New Tree of life



What do they have in common?

## Leading Edge Perspective

# Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life

Cindy J. Castelle<sup>1,2,3</sup> and Jillian F. Banfield<sup>1,2,3,4,5,6,\*</sup>

<sup>1</sup>Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA

<sup>2</sup>Innovative Genomics Institute, Berkeley, CA, USA

<sup>3</sup>Chan Zuckerberg Biohub, San Francisco, CA, USA

<sup>4</sup>University of Melbourne, Melbourne, VIC, Australia

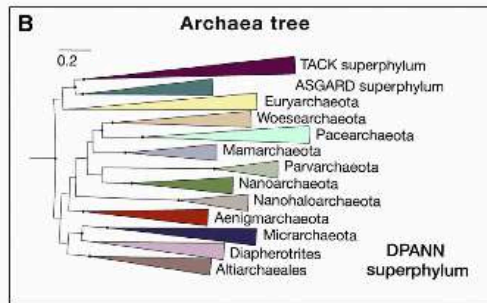
<sup>5</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA

<sup>6</sup>Department of Environmental Science, Policy and Management, University of California, Berkeley, Berkeley, CA, USA

\*Correspondence: [jbanfield@berkeley.edu](mailto:jbanfield@berkeley.edu)

<https://doi.org/10.1016/j.cell.2018.02.016>

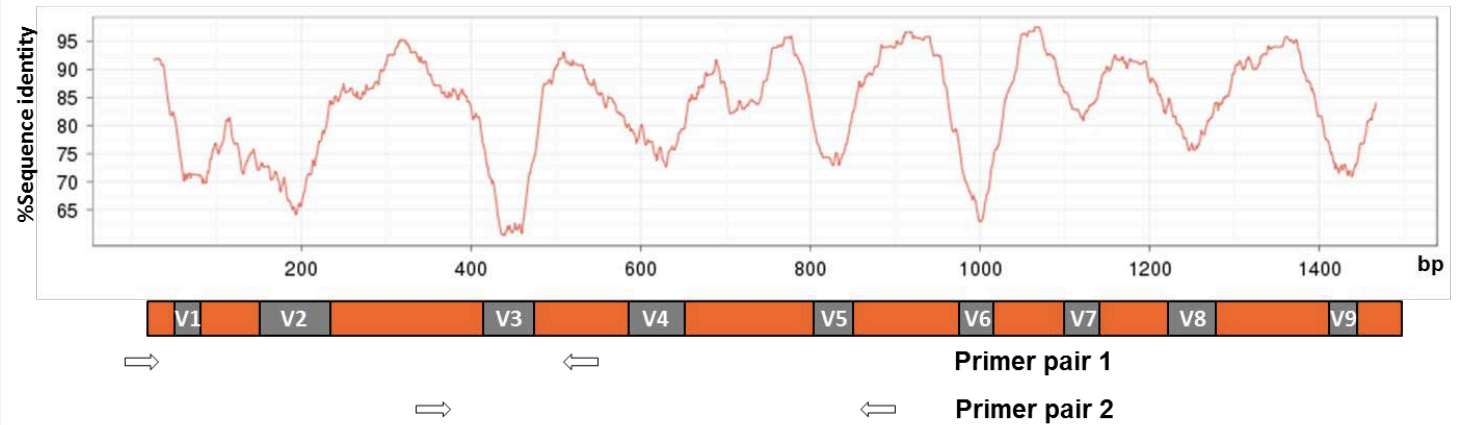
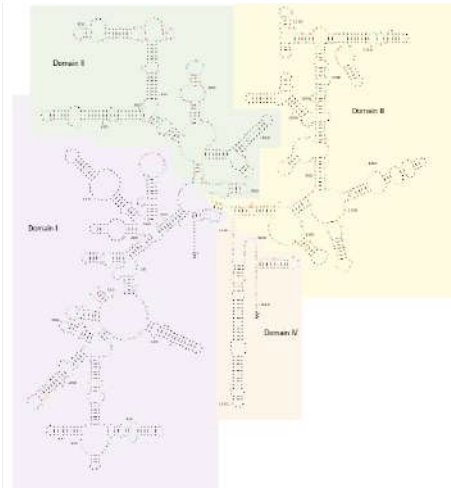
Tree scale  
0.1  
1



- TACK**
- Thaumarchaeota
  - Aigarchaeota
  - Crenarchaeota
  - Korarchaeota
  - Bathyarchaeota
  - Verstraetearchaeota
  - Geoarchaeota
- Euryarchaeota**
- MBSL-1
  - WSA2/Arct: (Methanostadiosa)
  - Hadesarchaea
  - Halobacteria
  - Theinoarchaea
  - Thalassoarchaea
  - Methanomassiliicocci

DPANN

# 16S



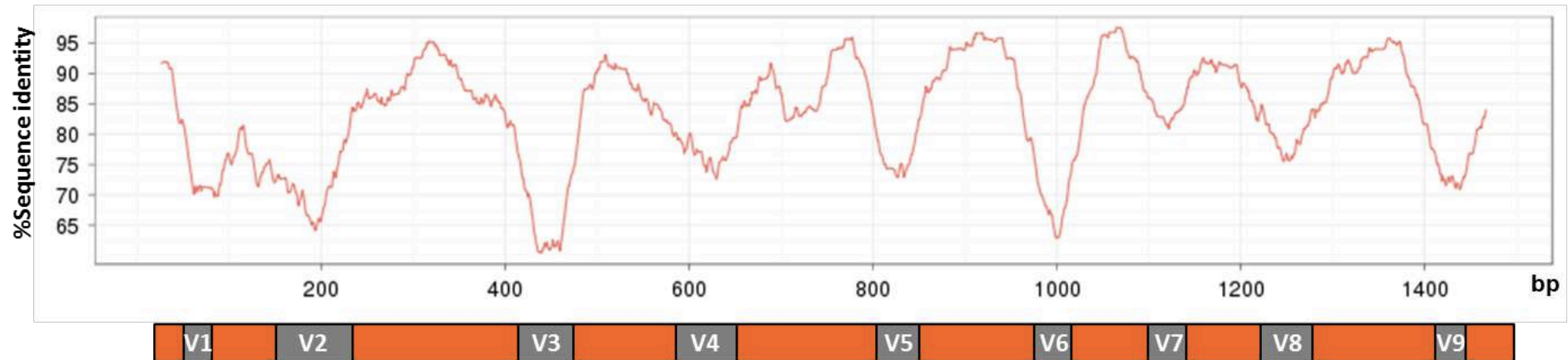
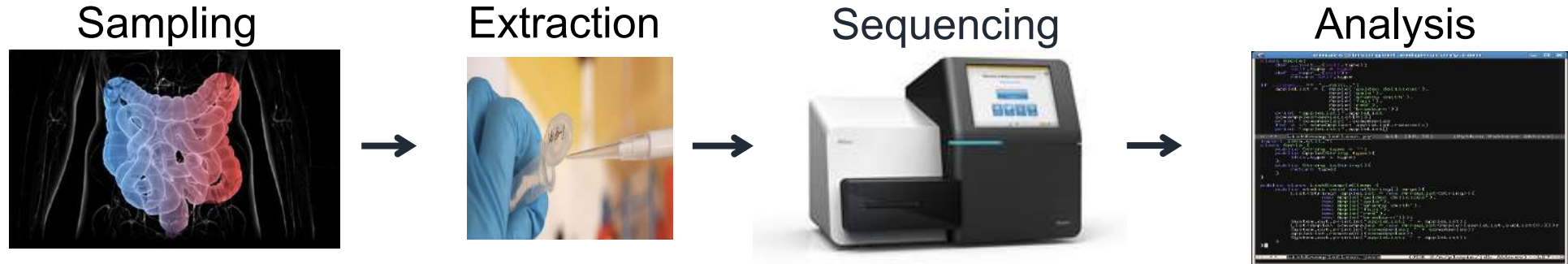
- **Advantages:**

- Universal: Every bacterial and archaea species has this gene
- Conserved regions (for primer design)
- **Variable regions (to distinguish different species)**
- Great databases and alignments (for human related species)
- Mainly used for taxonomical classification

- **Problems:**

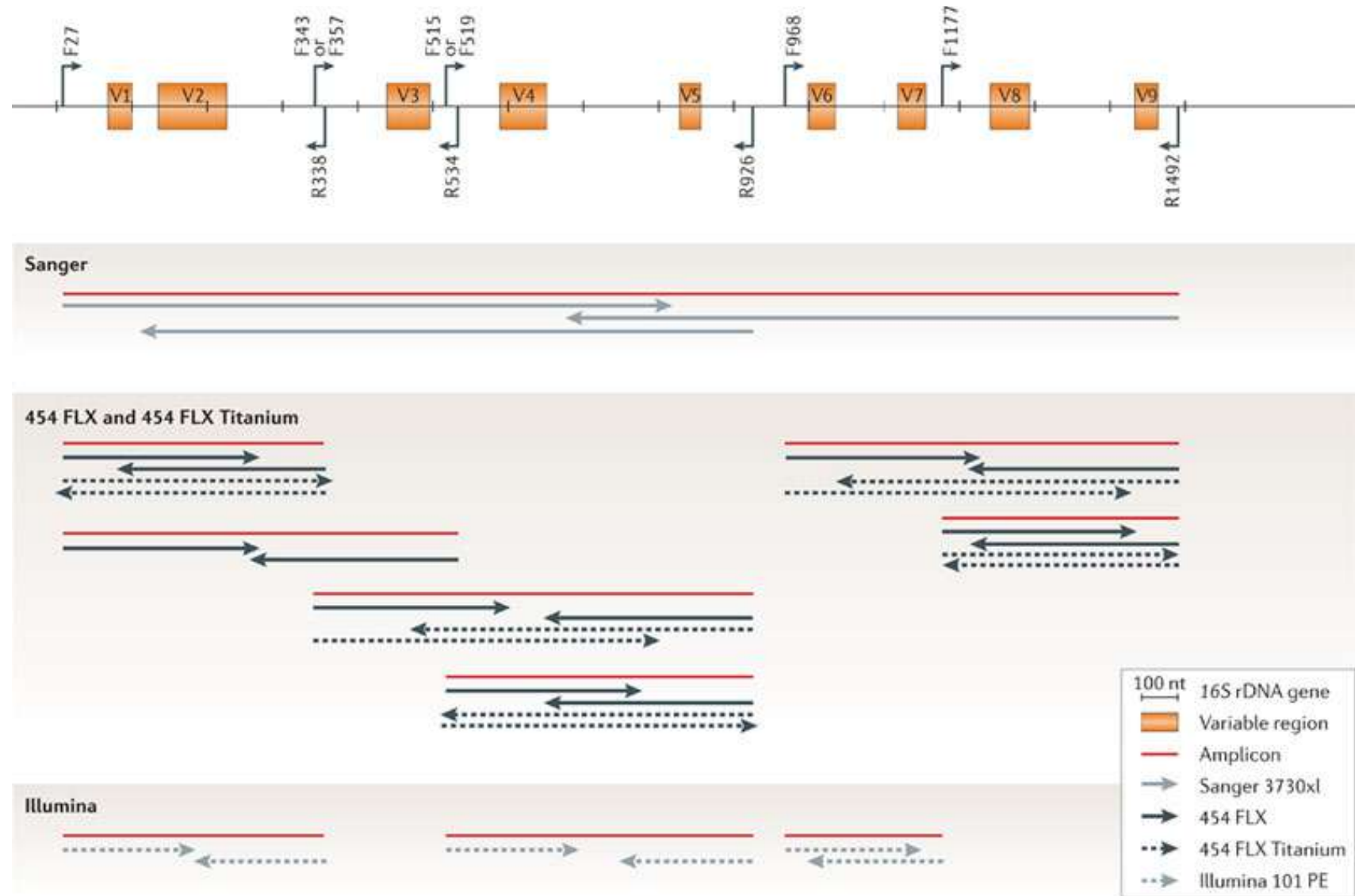
- Variable copy number in each species
- No universal (unbiased) primers
- (Not directly correlated with activity)
- (Lack of functional information)

# Typical workflow



Which region to sequence?

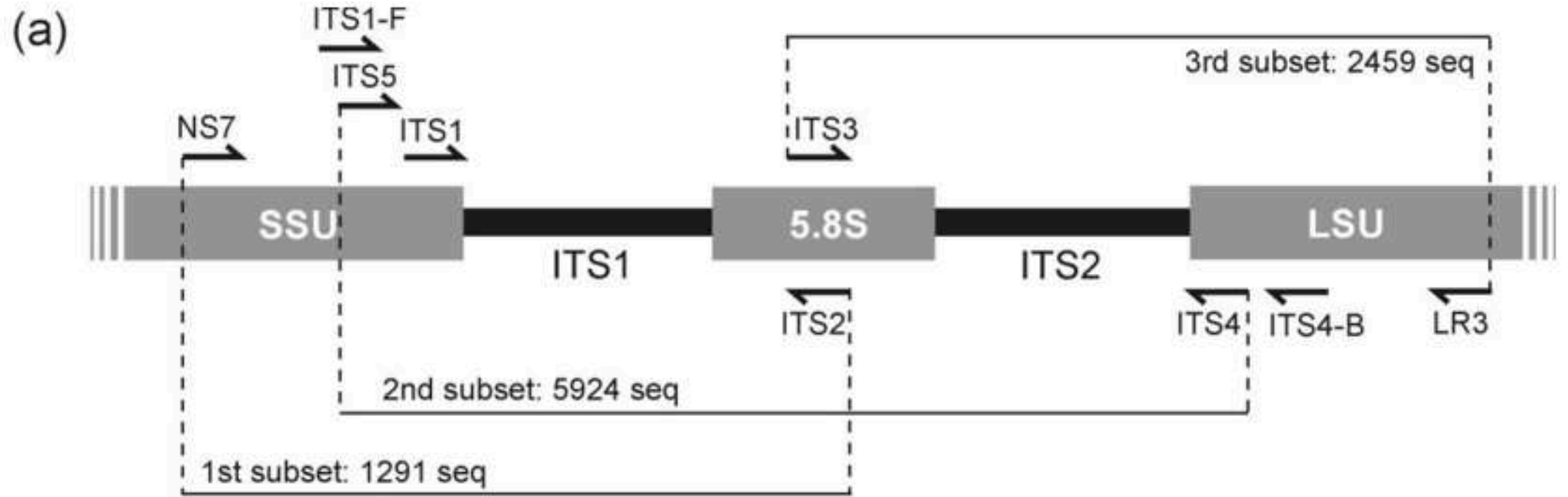
# 16S amplified region



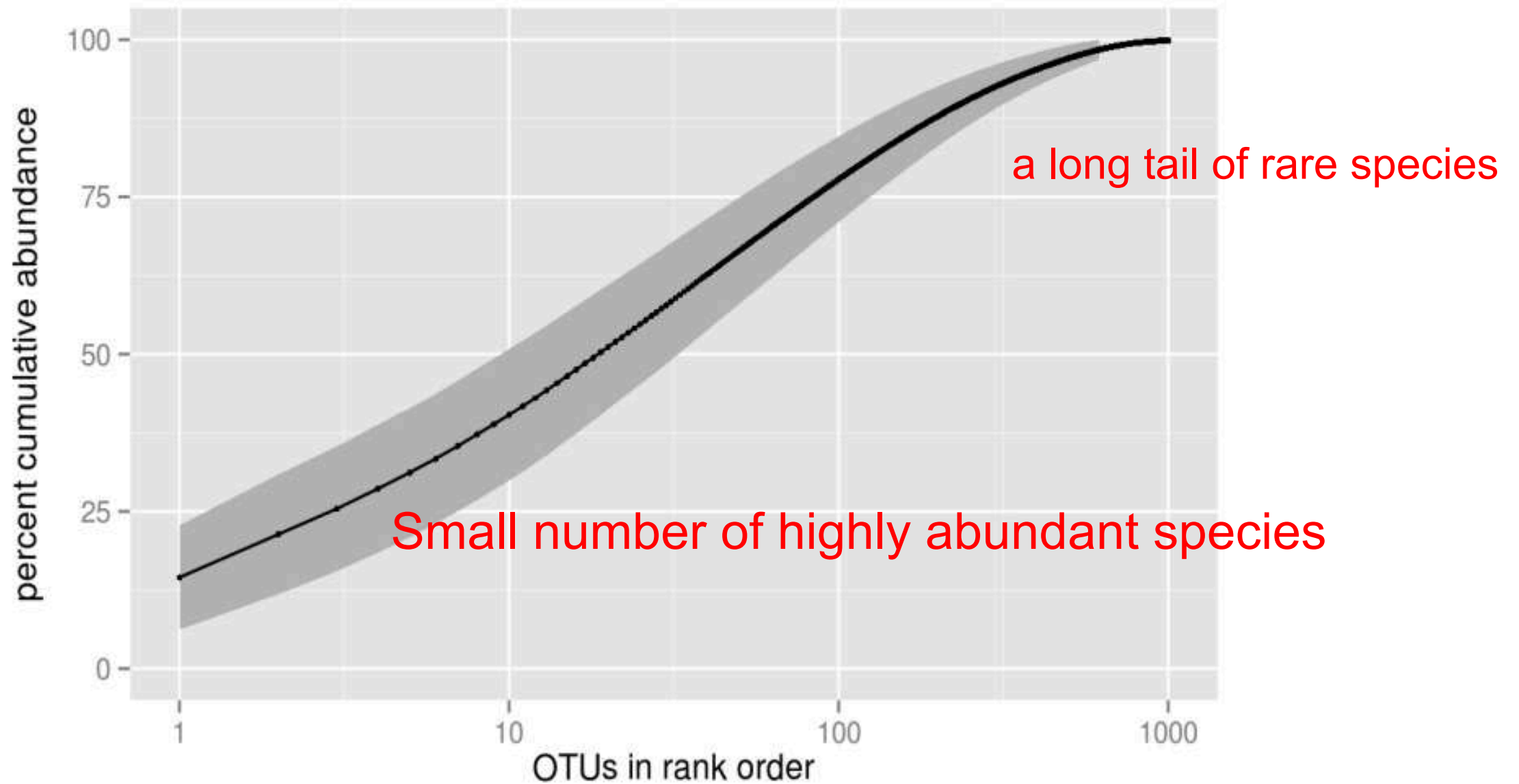
Kuczynski et al (2011)



# ITS for characterization of fungi species



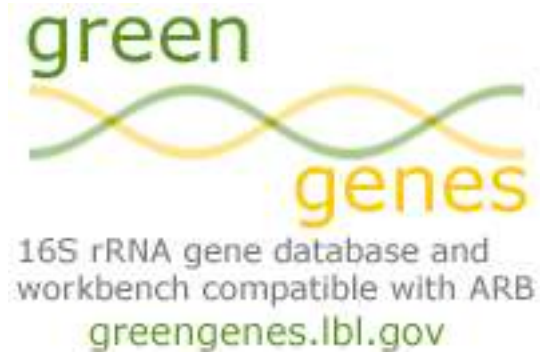
# Typical number of species



# Using a “Classifier” to annotate sequences

Uses an existing phylogeny

Find best unambiguous match to references



<http://greengenes.lbl.gov/cgi-bin/nph-index.cgi>



<http://www.arb-silva.de/>

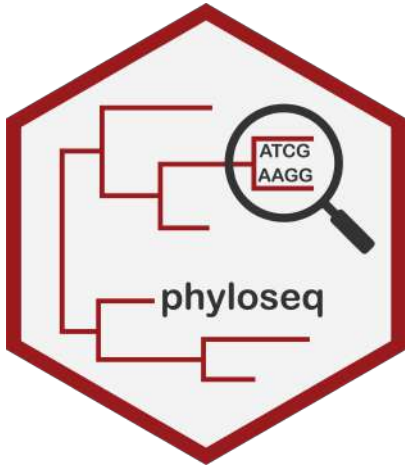


<https://rdp.cme.msu.edu/>

# Analysis Packages



Qiime2 (<https://qiime2.org/> )



Phyloseq ( <https://joey711.github.io/phyloseq/> )

R



Concept: OTU (Operational Taxonomic Unit)

note: being obsoleted soon

# OTU for Ecology

**Operational Taxonomic Unit:** a grouping of similar sequences that can be treated as a single “species”

## Strengths

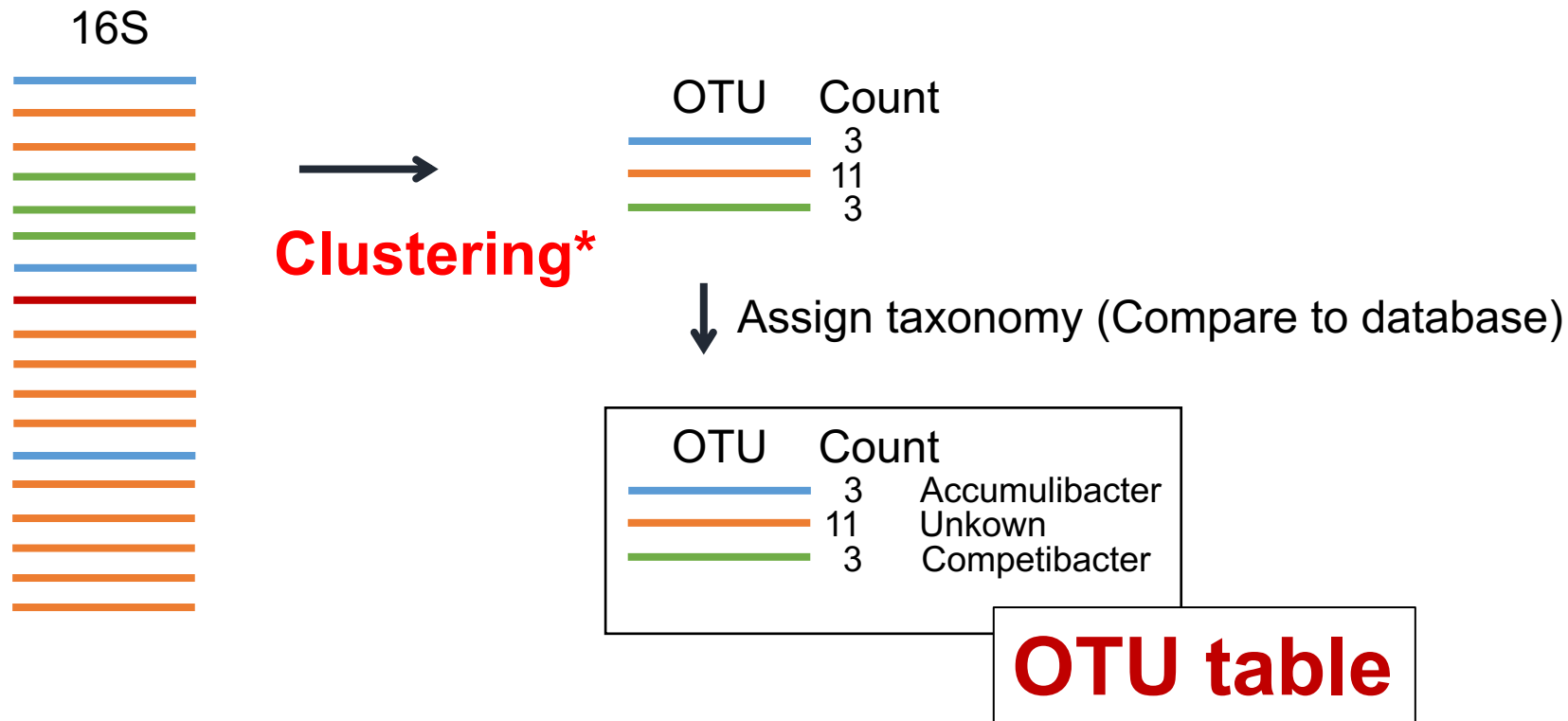
- Conceptually simple
- Mask effect of poor quality data
  - Sequencing error
  - in vitro recombination

## Weaknesses

- Limited resolution
- Logically inconsistent definition

# Assign OTU

- Cluster by their similarity to other sequences in the sample (operations taxonomic units → OTU)
- 95% genus level, **97% species level**, 99% strain level



# OTU “picking”

The process of bin sequences into clusters of OTUs.

## De Novo

Reads are clustered based on similarity to one another.

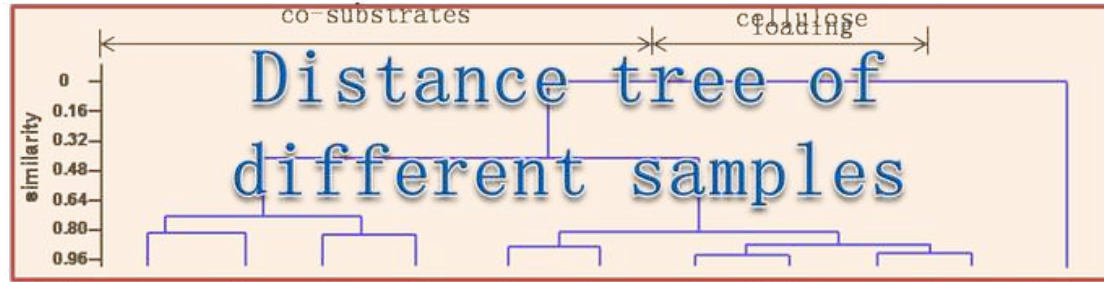
## Reference-based

Closed reference: any reads which don't hit a reference sequence are discarded

Open reference: any reads which don't hit a reference sequence are clustered de novo



# Tree way plot with top OTUs abundance and classification



OUT NO.	G1	X1	G2	X2	S1	S2	High	Medium	Low	Inoculum	Seed
1	16.394%	18.740%	9.072%	10.154%	31.778%	26.680%	32.332%	32.312%	40.820%	39.410%	0.040%
2	7.282%	6.904%	4.352%	4.994%	12.884%	13.464%	13.726%	13.334%	15.630%	13.790%	0.006%
3	7.108%	7.174%	4.070%	4.610%	13.604%	12.230%	13.612%	13.276%	15.874%	14.348%	0.006%
4	5.264%	5.418%	2.988%	3.376%	10.168%	9.328%	10.632%	11.062%	11.600%	10.104%	0.004%
5	0.002%	0.002%	1.136%	0.000%	0.666%	2.178%	0.000%	0.000%	0.006%	0.000%	0.000%
6	0.338%	0.296%	0.382%	0.244%	0.564%	0.866%	0.006%	0.036%	0.070%	0.816%	0.000%
7	2.460%	3.044%	1.586%	2.646%	4.904%	5.052%	2.068%	2.434%	0.708%	1.532%	0.002%
8	1.508%	1.246%	0.980%	1.338%	2.340%	3.154%	1.282%	1.542%	0.460%	0.814%	0.002%
9	1.192%	1.142%	0.784%	1.022%	2.026%	2.516%	1.066%	1.258%	0.386%	0.750%	0.000%
10	1.098%	0.714%	0.626%	0.714%	1.342%	2.322%	0.870%	1.192%	0.336%	0.582%	0.000%
11	3.686%	2.229%	3.910%	1.961%	0.218%	0.910%	0.000%	0.000%	0.000%	0.000%	0.000%
12	1.862%	0.904%	1.356%	0.864%	0.042%	0.052%	0.000%	0.002%	0.004%	0.002%	0.002%
13	1.756%	0.946%	1.302%	0.698%	0.068%	0.078%	0.002%	0.002%	0.000%	0.000%	0.000%
14	5.978%	12.532%	16.236%	16.356%	0.000%	0.000%	0.008%	0.006%	0.004%	0.008%	0.010%
15	10.494%	6.218%	8.710%	4.958%	0.000%	0.000%	0.002%	0.002%	0.010%	0.000%	0.002%
16	3.152%	5.022%	7.460%	7.656%	0.404%	1.214%	0.004%	0.004%	0.006%	0.004%	0.010%
17	8.504%	4.474%	6.390%	3.668%	0.256%	0.332%	0.000%	0.002%	0.008%	0.004%	0.002%
18	2.764%	5.346%	6.940%	7.068%	0.446%	1.230%	0.006%	0.002%	0.008%	0.000%	0.004%
19	4.642%	2.480%	3.866%	2.190%	0.162%	0.210%	0.000%	0.002%	0.002%	0.002%	0.002%
20	2.098%	3.808%	4.594%	4.858%	0.384%	0.986%	0.000%	0.008%	0.002%	0.000%	0.004%
21	1.486%	0.276%	4.892%	0.530%	0.106%	0.396%	0.000%	0.002%	0.002%	0.000%	0.000%
22	0.018%	0.300%	0.042%	3.894%	0.016%	0.064%	0.000%	0.006%	0.000%	0.002%	0.002%
23	0.014%	0.362%	0.064%	5.142%	0.018%	0.134%	0.002%	0.002%	0.000%	0.000%	0.004%
24	0.146%	0.216%	0.084%	0.112%	0.040%	0.018%	1.244%	1.048%	0.662%	1.742%	0.004%
25	0.138%	0.140%	0.062%	0.054%	0.002%	0.006%	0.922%	0.838%	0.536%	1.036%	0.000%
26	0.012%	0.016%	0.002%	0.002%	0.038%	0.020%	0.004%	0.006%	0.004%	0.050%	3.950%
27	0.000%	0.002%	0.000%	0.004%	0.004%	0.004%	0.002%	0.008%	0.002%	0.000%	4.204%
28	0.108%	0.122%	0.018%	0.044%	1.392%	0.316%	0.614%	1.012%	0.880%	0.440%	0.000%
29	0.166%	0.078%	0.014%	0.026%	0.210%	0.102%	1.744%	0.870%	0.914%	0.152%	0.000%
30	0.002%	0.010%	0.008%	0.004%	0.072%	0.016%	0.000%	5.428%	0.128%	0.478%	0.006%
Coverage	89.7%	90.4%	91.1%	89.0%	86.9%	87.6%	80.1%	85.7%	88.9%	86.1%	8.3%

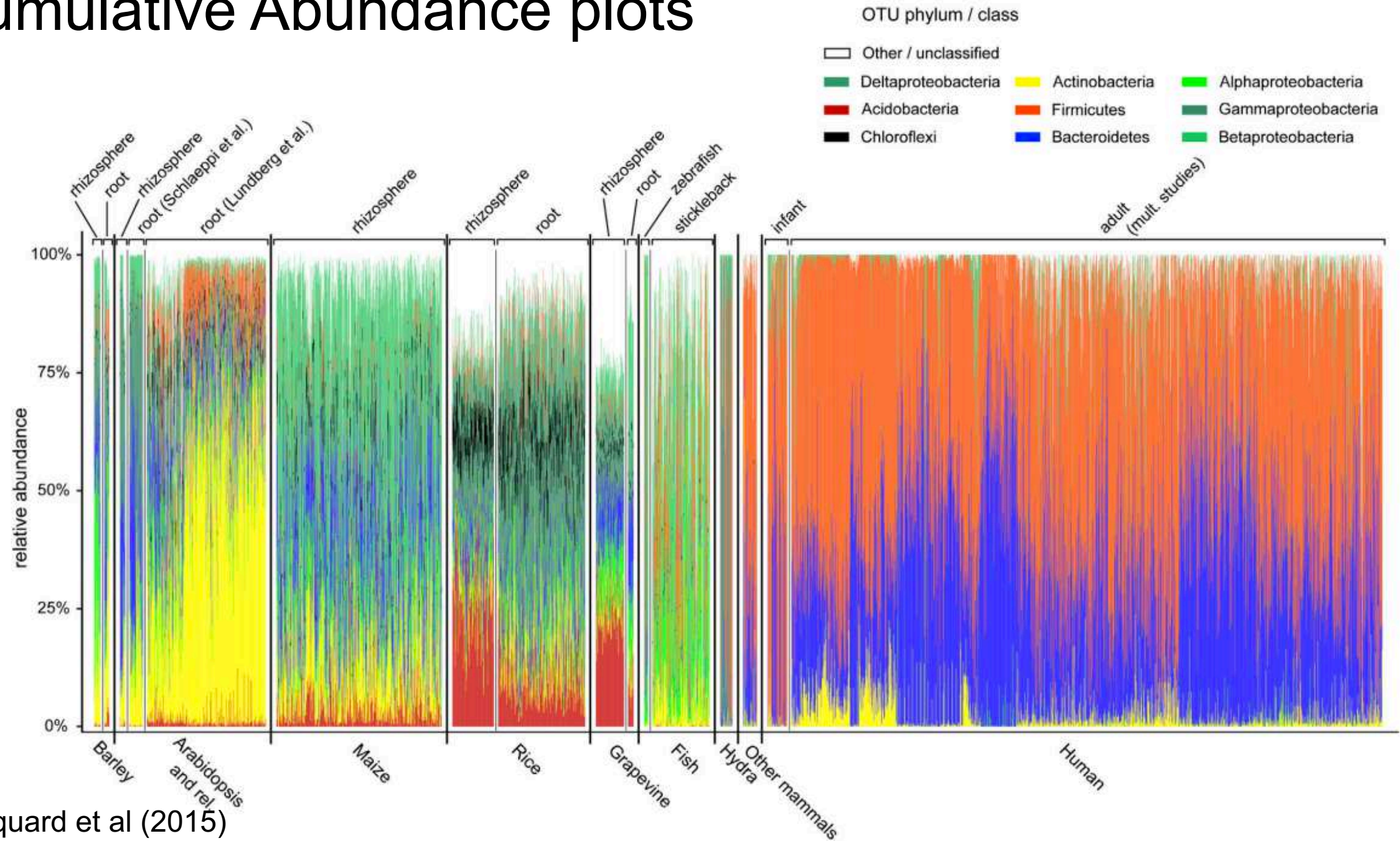
Top OTU abundance matrix

OTU Classification

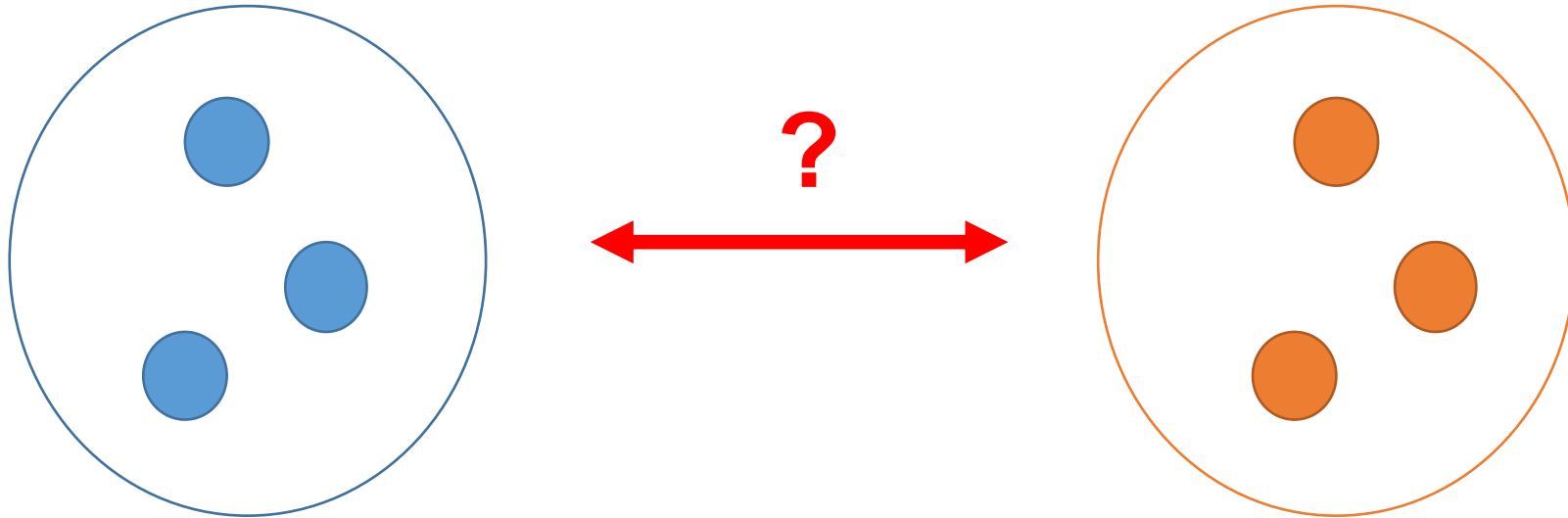
<i>Clostridium</i> sp. 4-2a
<i>Clostridium</i> sp. 4-2a
<i>Clostridium</i> sp. 4-2a
<i>Clostridium</i> sp. 4-2a
<i>Anoxybacillus</i> sp. HT8
<i>Clostridiaceae</i> bacterium 37-7-2Cl
<i>Clostridium</i> sp. YN5
<i>Clostridium</i> sp. YN6
<i>Clostridium</i> sp. YN7
<i>Clostridium</i> sp. YN8
<i>Thermoanaerobacterium xylanolyticum</i> LX-11
<i>Thermoanaerobacterium xylanolyticum</i> LX-11
<i>Thermoanaerobacterium xylanolyticum</i> LX-11
<i>Thermoanaerobacterium aotearoense</i> SCUT27
<i>Thermoanaerobacterium xylanolyticum</i> LX-11
<i>Thermoanaerobacterium aotearoense</i> SCUT27
<i>Thermoanaerobacterium xylanolyticum</i> LX-11
<i>Thermoanaerobacterium aotearoense</i> SCUT27
<i>Thermoanaerobacterium xylanolyticum</i> LX-11
<i>Thermoanaerobacterium aotearoense</i> SCUT27
<i>Thermoanaerobacterium saccharolyticum</i> D22
<i>Thermoanaerobacterium thermosaccharolyticum</i> CECT5853T
<i>Thermohydrogenium kirishiense</i> DSM11055T
<i>Uncultured Bacillus</i> sp. clone 93
<i>Uncultured Bacillus</i> sp. clone 94
<i>Mycobacterium gilvum</i> B4
<i>Nocardioides terrae</i> VA15
<i>Anaerolinea thermolimosa</i> IMO-1
<i>Uncultured bacterium</i> clone 49c
<i>Clostridium cellobioparum</i> JCM 1422



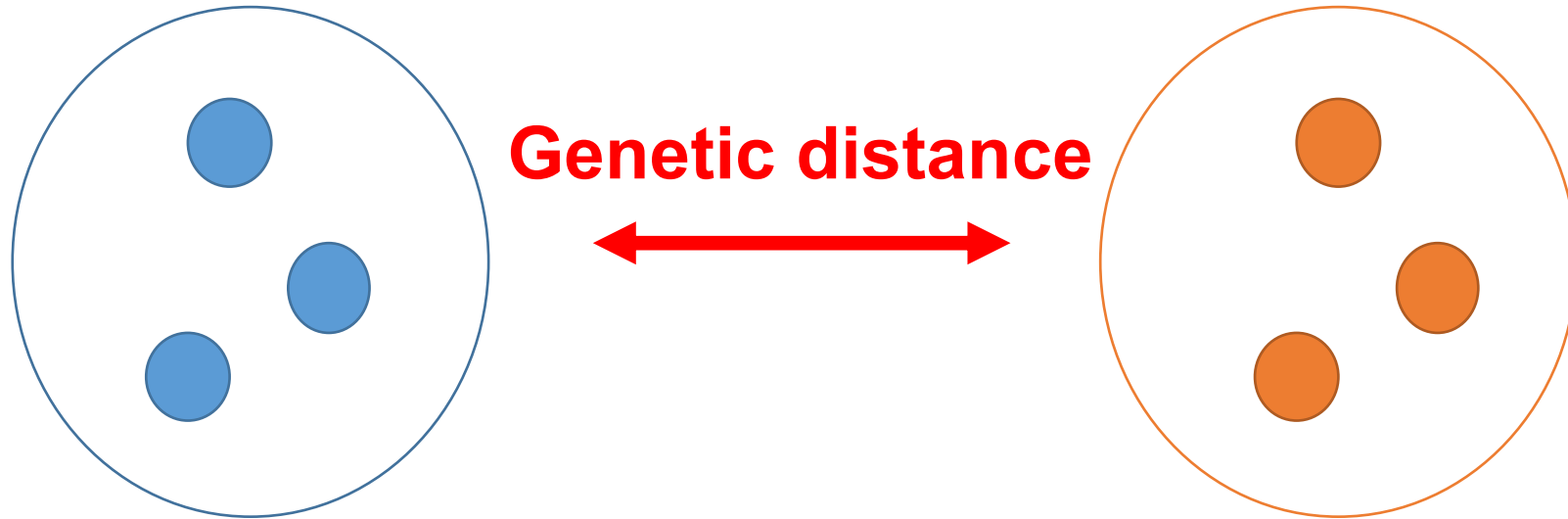
# Cumulative Abundance plots



# Assigned OTUs -> Loss of information

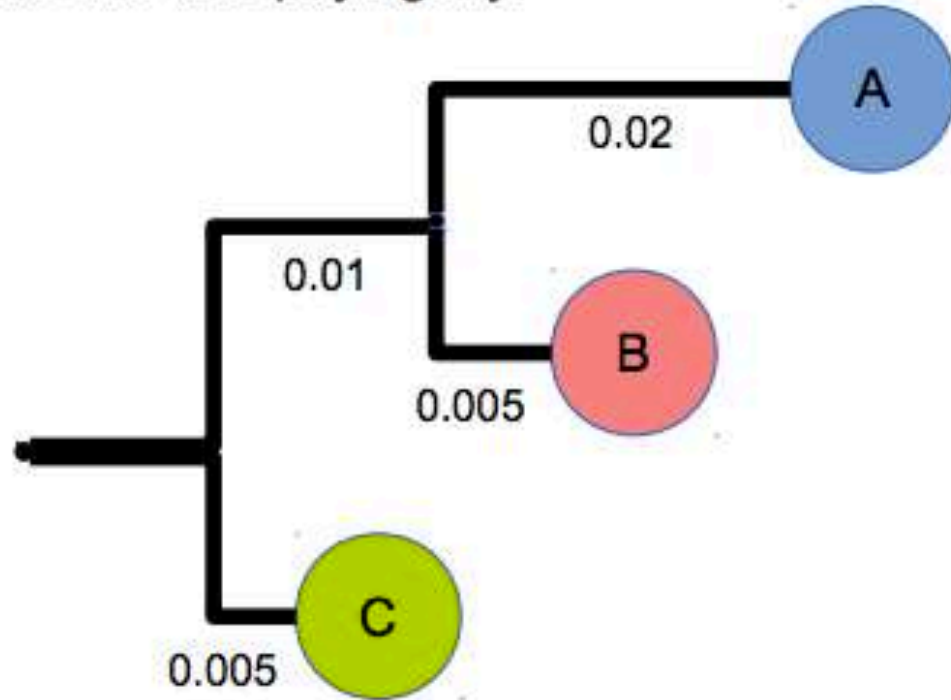


# Assigned OTUs ; assigned distance using phylogeny



# Logical inconsistency: OTUs at 97% ID

Assume the true phylogeny:



A, B > 97% identity  
B, C > 97% identity  
A and C not > 97% ID

## Possible valid OTUs:

AB, C (with A & C centroids)  
A, BC (with A & C centroids)  
ABC (with B centroid)

OTU pipelines will arbitrarily pick one of the three solutions.  
Is this actually a problem??



# New approach: Use of ESV (Exact-) or ASV (Amplicon Sequence Variant)

## DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan<sup>1</sup>, Paul J McMurdie<sup>2</sup>,  
Michael J Rosen<sup>3</sup>, Andrew W Han<sup>2</sup>, Amy Jo A Johnson<sup>2</sup> &  
Susan P Holmes<sup>1</sup>

We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (<https://github.com/benjjneb/dada2>). DADA2 infers sample sequences exactly and resolves differences of as little as 1 nucleotide. In several mock communities, DADA2 identified more real variants and output fewer spurious sequences than other methods. We applied DADA2 to vaginal samples from a cohort of pregnant women, revealing a diversity of previously undetected *Lactobacillus crispatus* variants.

OPEN

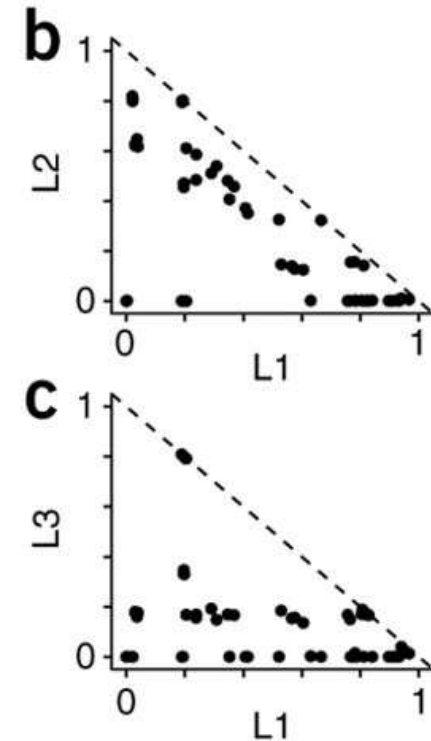
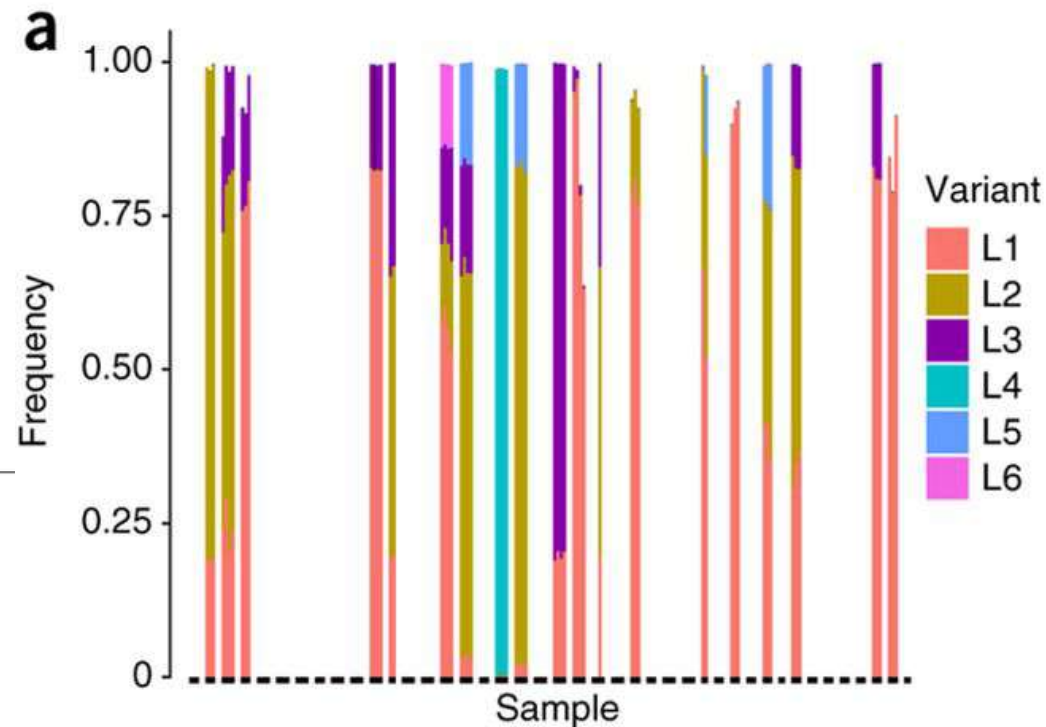
The ISME Journal (2017) 11, 2639–2643  
[www.nature.com/ismej](http://www.nature.com/ismej)

### PERSPECTIVE

## Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan<sup>1</sup>, Paul J McMurdie<sup>2</sup> and Susan P Holmes<sup>3</sup>

<sup>1</sup>Department of Population Health and Pathobiology, NC State University, Raleigh NC, USA; <sup>2</sup>Whole Biome Inc, San Francisco CA, USA and <sup>3</sup>Department of Statistics, Stanford University, Stanford CA, USA



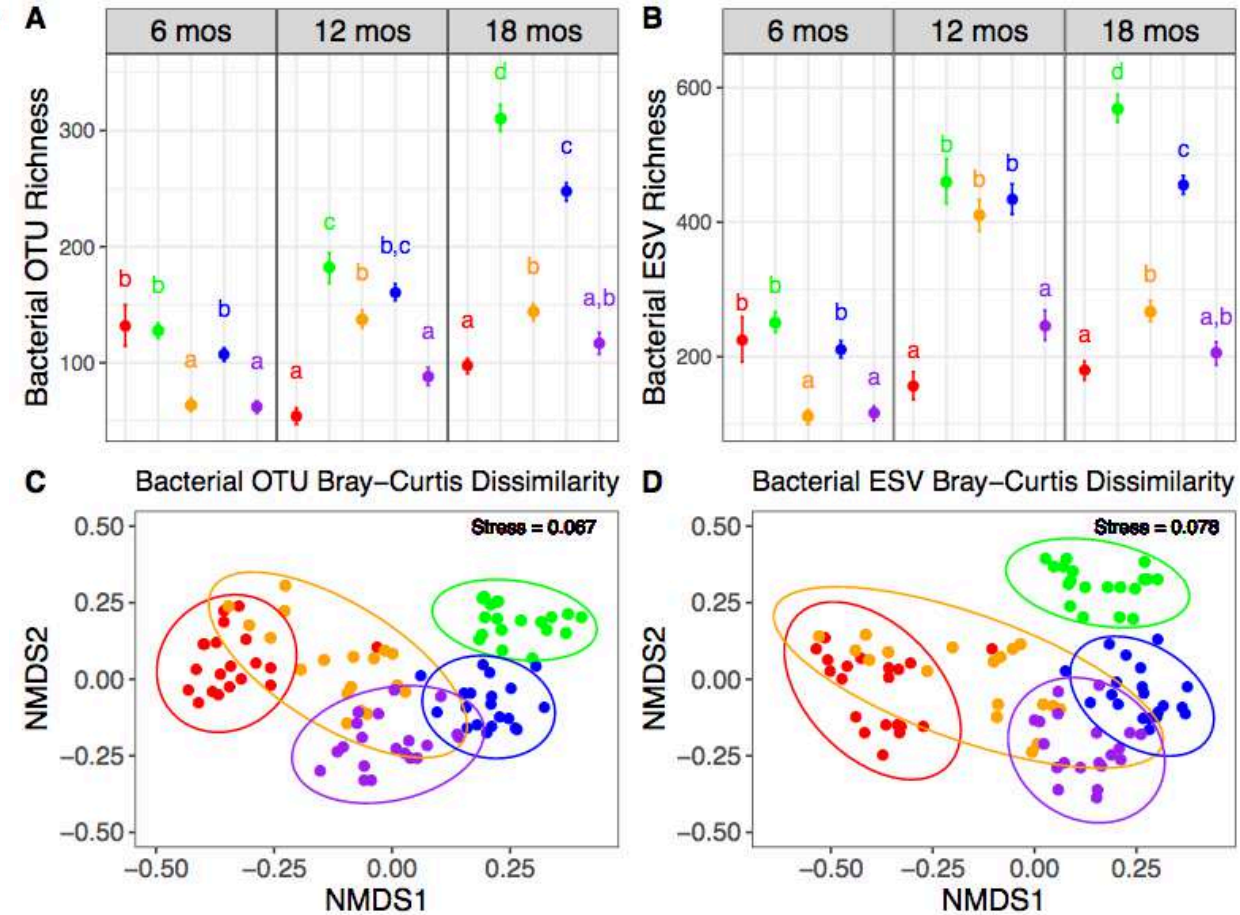
- Need denoising (attempt to correct sequencing errors) first



## Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units

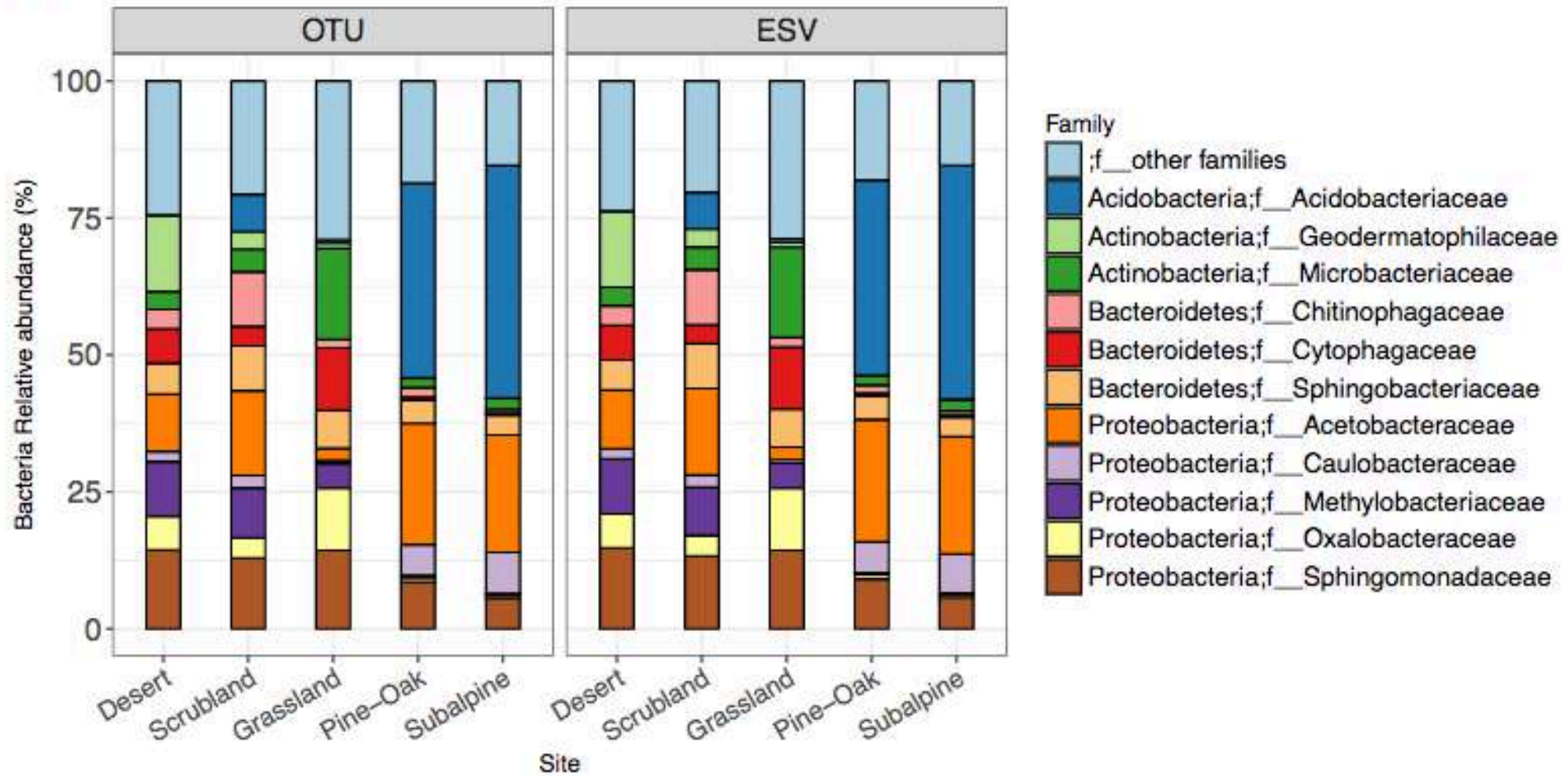
Sydney I. Glassman,<sup>a,b</sup> Jennifer B. H. Martiny<sup>a</sup>

Despite quantitative differences in microbial richness, **we found that all and diversity metrics were highly positively correlated ( $r=0.90$ ) between samples analyzed with both approaches.** Moreover, the community composition of the dominant taxa did not vary between approaches. Consequently, **statistical inferences were nearly indistinguishable.**



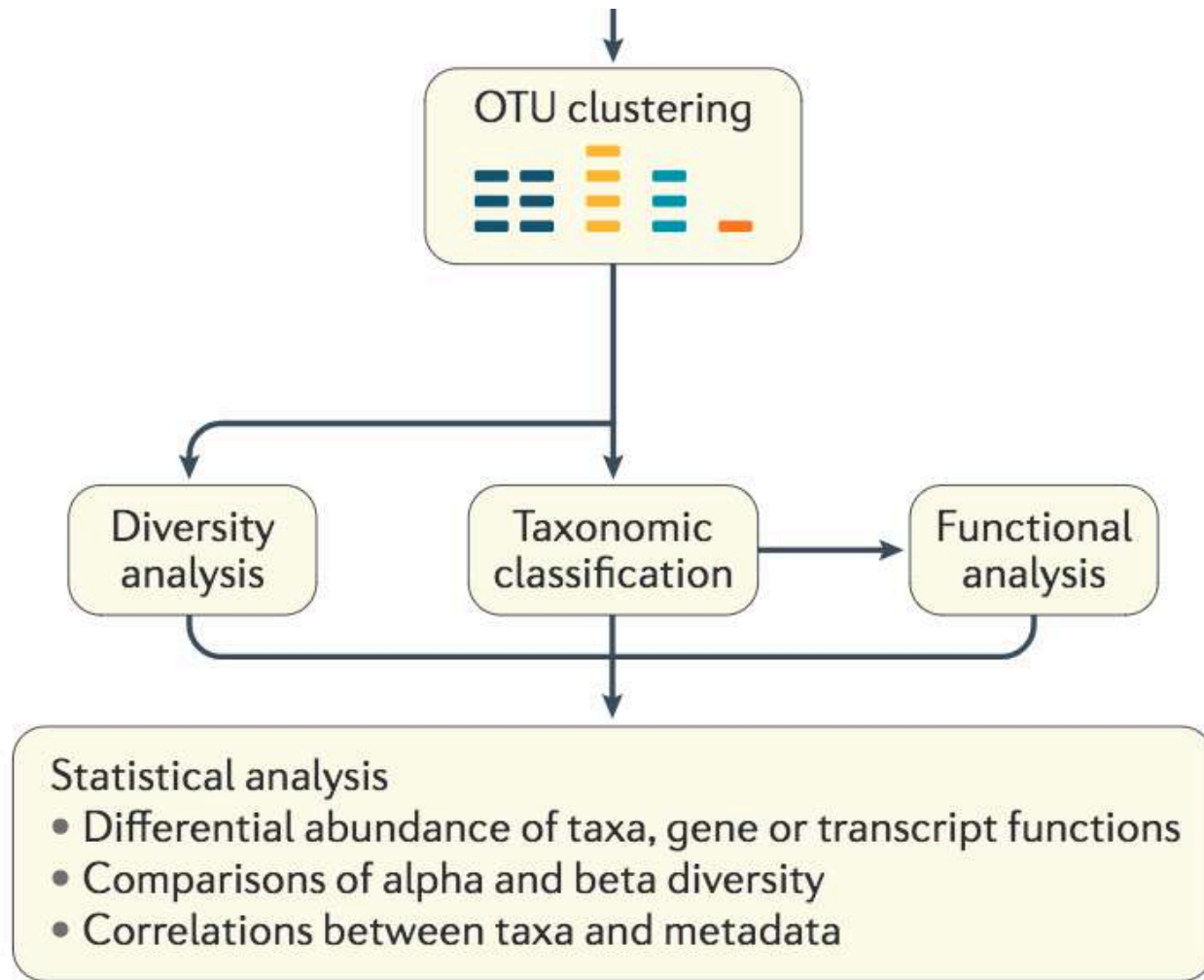
**Figure S2.** Relative abundance of bacterial sequences per sample of the inoculum leaf litter from each site (n=20) for OTUs versus ESVs summarized by A) the 12 most abundant families or B) the 12 most abundant genera.

A)





# Amplicon sequencing: summary



What can we gain from amplicon analyses?  
Powerfulness of amplicon analysis will increase if:

- \* more samples can be achieved with less cost,
- \* better amplicons (less false positives, higher resolution i.e., delineating strain level perhaps)
- \* reproducibility

Intensive research field with long reads

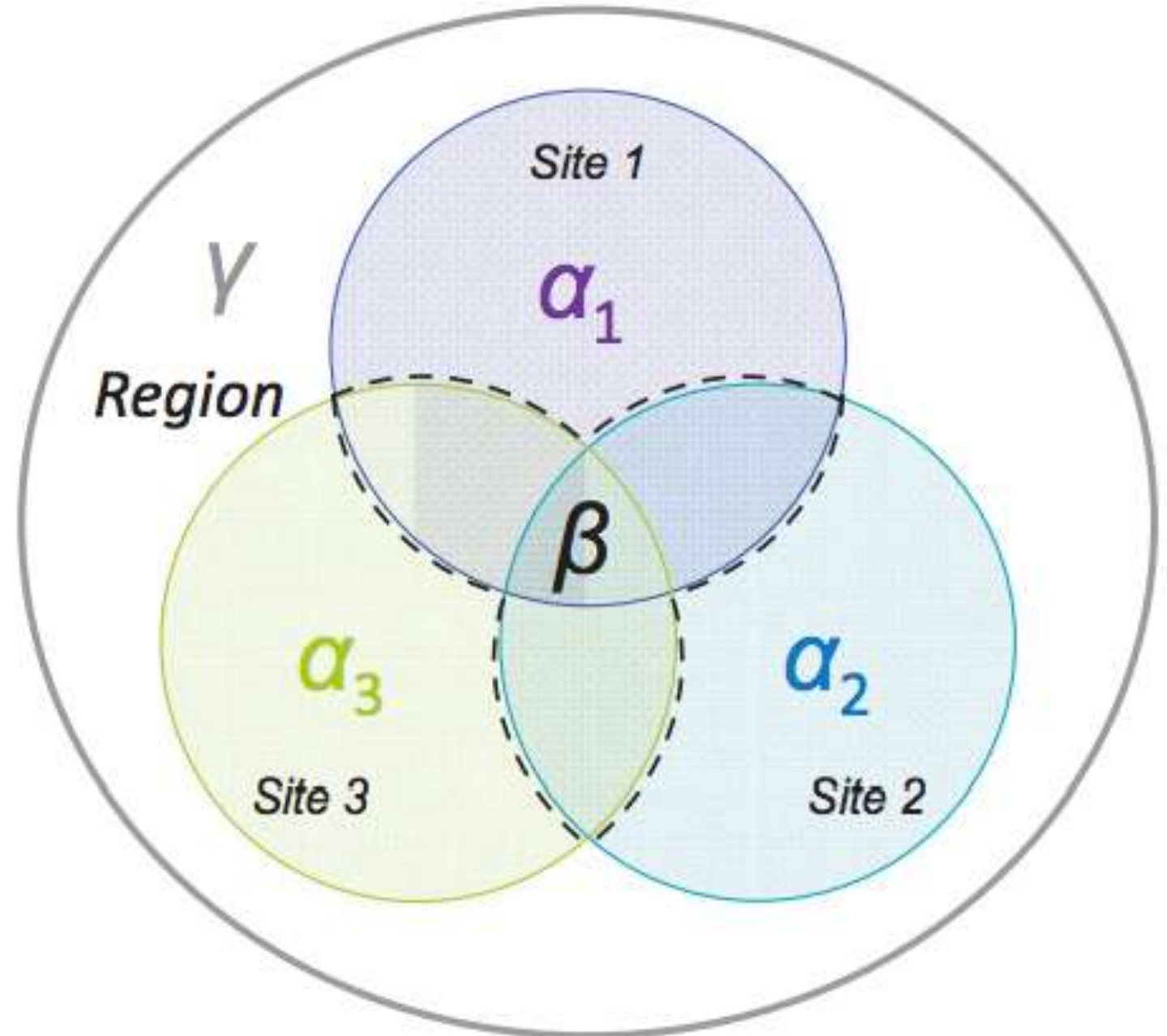
Concept: Diversity measures



# Measures of biodiversity

Zinger et al (2012)

“... measuring biodiversity consists of characterizing the **number**, **composition** and **variation** in taxonomic or functional units (**OTU**) over a wide range of biological organizations”



**Alpha diversity** refers to the diversity within one location or sample. It is often measured as species richness (i.e. number of species), seldom as species evenness (extent of species dominance). Species richness is strongly sensitive to sampling effort, and requires standardized samples, or the use of estimators that corrects undersampling biases, such as Chao1 or ACE. Evenness is less affected by undersampling biases and is usually assessed with Simpson's or Pielou's indices or rank abundance curves (review in Magurran 2004).

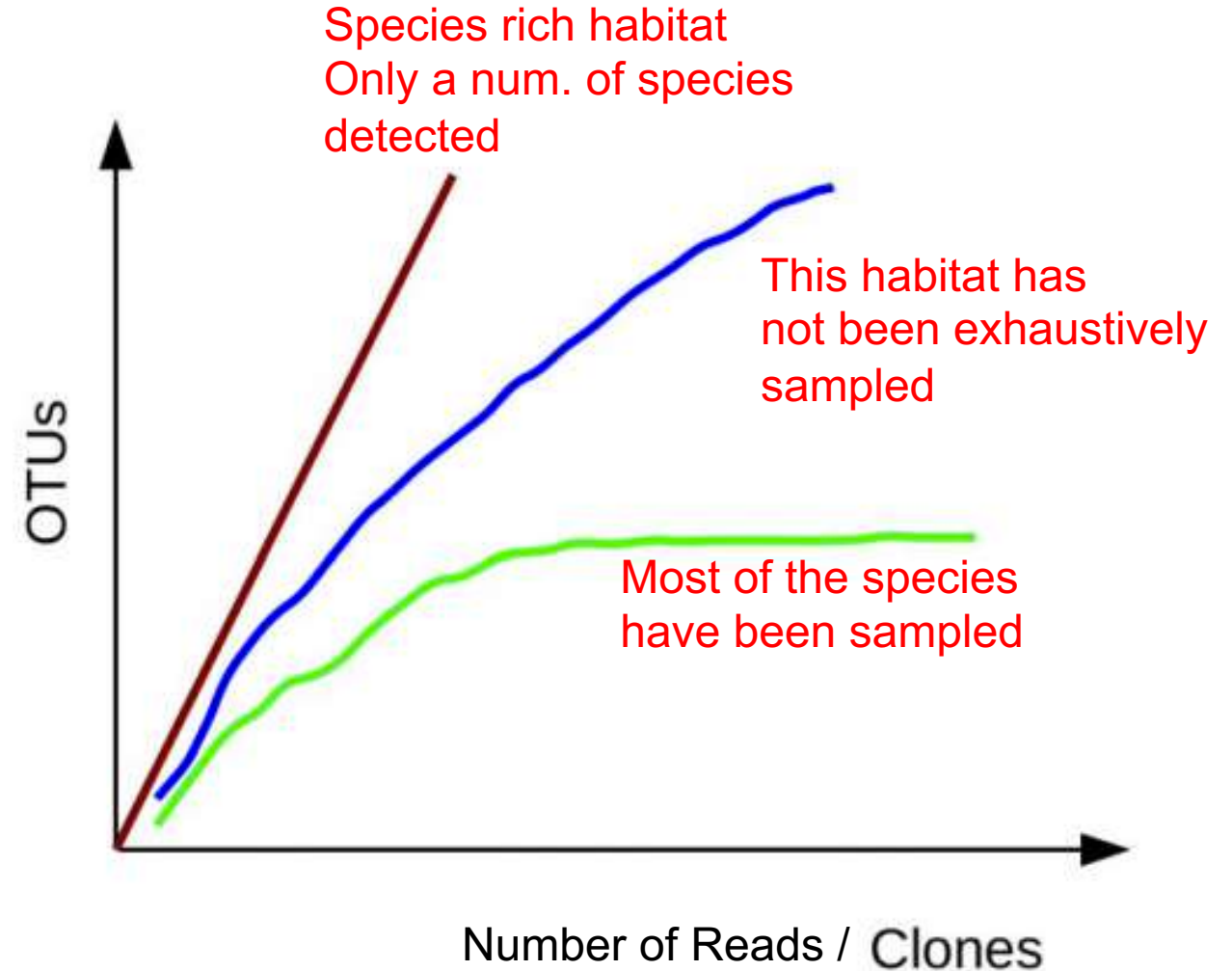
**Beta diversity** consists in determining the difference in diversity or community composition between two or more locations or samples (i) by considering species composition only, and use incidence data with associated metrics such as Jaccard or Sorensen similarity indices or (ii) by taking species relative abundances into account, and use Bray–Curtis or Morisita–Horn dissimilarity measures (Anderson *et al.* 2011). Using abundance data is, however, strongly discussed among microbiologists when dealing with rRNA gene data because of variations in gene copy number among strains (Acinas *et al.* 2004b; Zhu *et al.* 2005) as well as PCR artefacts.

**Gamma diversity**, or regional diversity, is similar to alpha diversity but applies for a larger area that encompasses the units under study.

Finally, the spatial scale of investigation can produce very different results and should be consistent in cross-study comparisons (Magurran 2004).

# Species sampling and Rarefaction

**Rarefaction** allows the calculation of **species richness** for a given number of individual samples, based on the construction of so-called **rarefaction curves**. This curve is a plot of the number of species as a function of the number of samples



# Alpha diversity

a measure of the diversity within a single sample

Types of alpha diversity

Total # of species = **richness**

**How many OTUs?**

Total # of genes = genetic richness

Phylogenetic diversity of genes = genetic PD

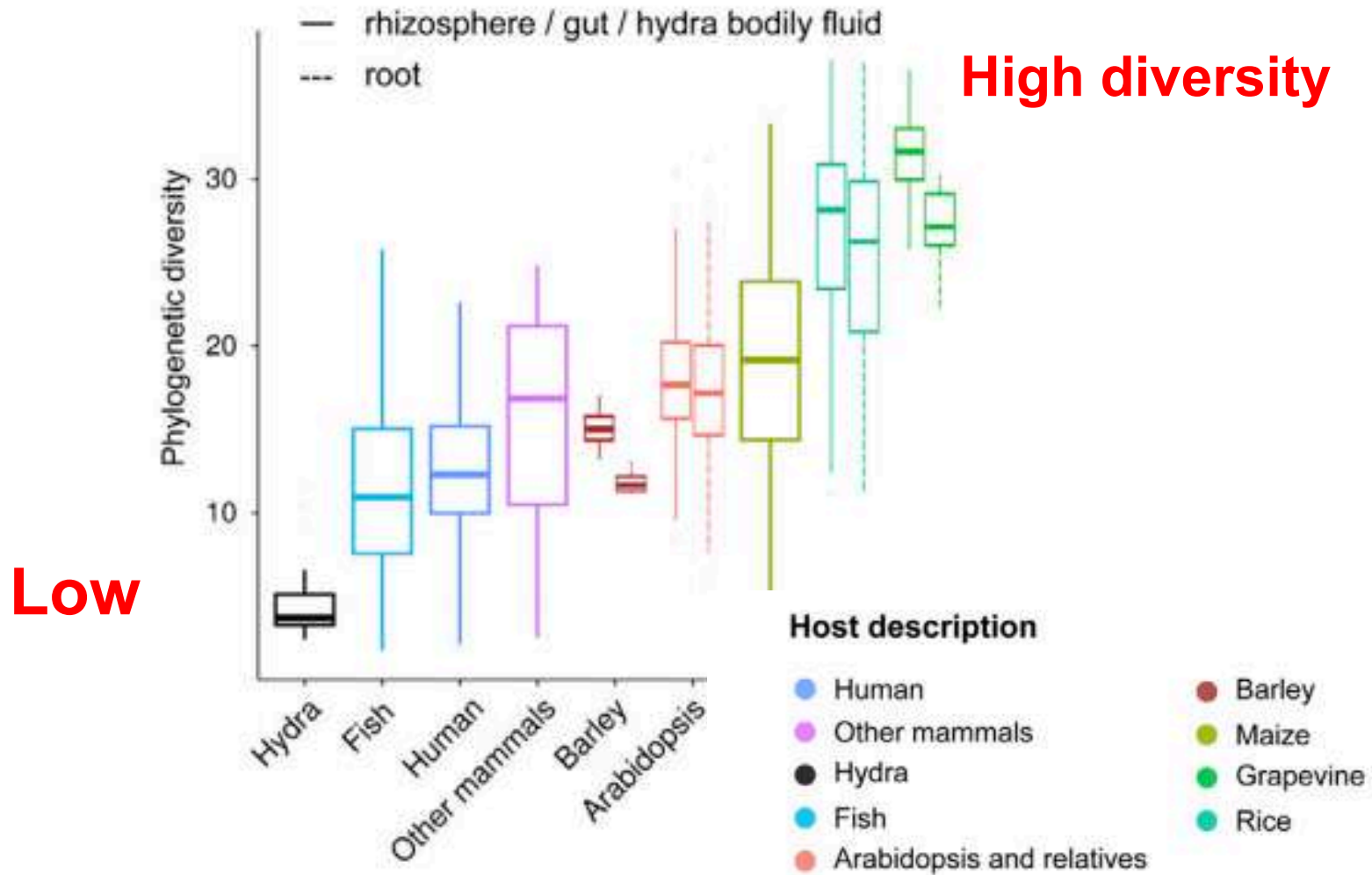
Evenness = What is the distribution of abundance in the community?

**How many OTUs at high abundance and how many OTU at low abundance?**



**B**

### Alpha-diversity (phylogenetic diversity)





# Beta diversity

a measure of **the similarity in diversity between samples**

Types of beta diversity

- Species presence/absence

- Shared phylogenetic diversity

- Gene presence / absence

- Shared phylogenetic diversity of genes

Frequently used as values for PCA or PCoA analysis

# Beta diversity

## A. Membership:

shared OTU occurrences across communities  
 1 = present, 0 = below detection

List of observed OTUs	Occurrences in community A	Occurrences in community B	Shared occurrences A & B
	OTU 1	1	0
OTU 2	0	1	
OTU 3	1	1	X ●
OTU 4	1	1	X ●
OTU 5	1	1	X ●

## B. Composition:

similar OTU abundances across communities

List of observed OTUs	Abundances community A	Abundances community B	Similar abundances A & B
	OTU 1	0.4	0
OTU 2	0	0.1	
OTU 3	0.1	0.1	X ●
OTU 4	0.2	0.5	
OTU 5	0.3	0.3	X ●

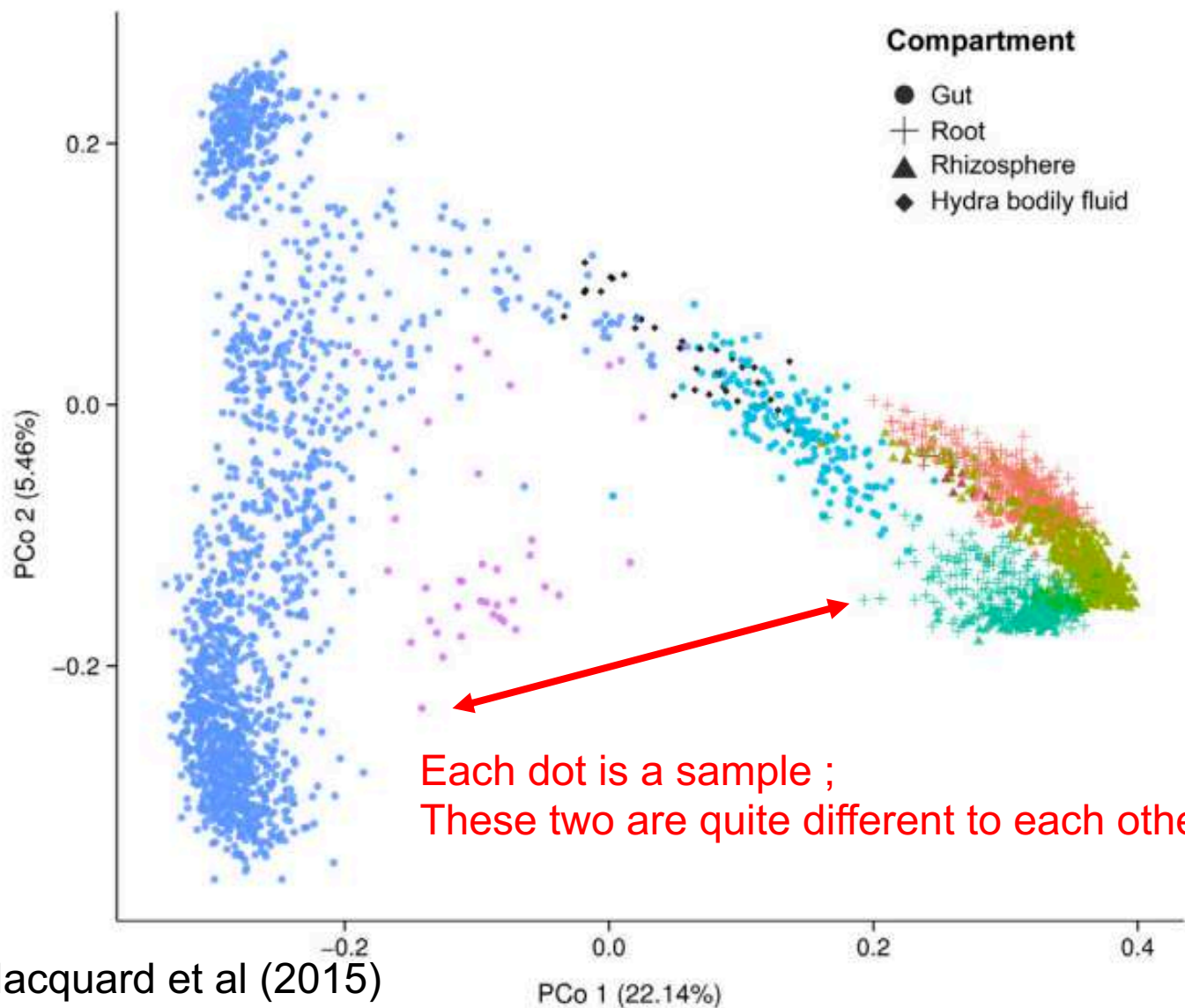
## Phylogeny:

shared OTU lineages across communities

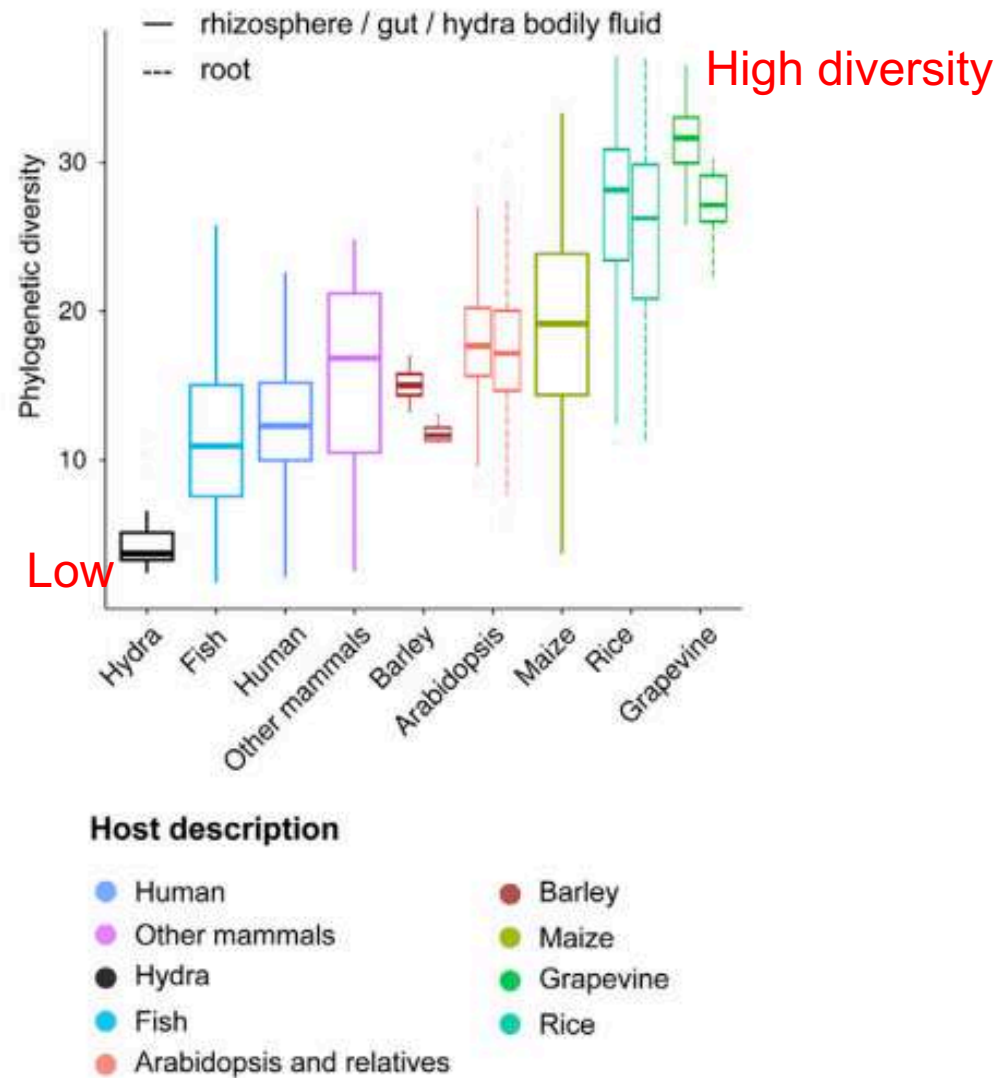
lineage	Abundances community A	Abundances community B	Similar abundances A & B
	OTU 1	0.4	0
OTU 2	0	0.1	
OTU 3	0.1	0.1	] X ●
OTU 4	0	0.8	
OTU 5	0.5	0	

**A**

**Beta-diversity**  
(unweighted UniFrac distance)

**B**

**Alpha-diversity**  
(phylogenetic diversity)



# Metagenomics

Keyword: MAG (metagenome-assembled genomes)

# Advantage of metagenomics approach

**Better classification with Increasing number of complete genomes**

**Focus on whole genome based phylogeny (whole genome phylotyping)**

- Advantages

No amplification bias like in 16S/ITS

Issues

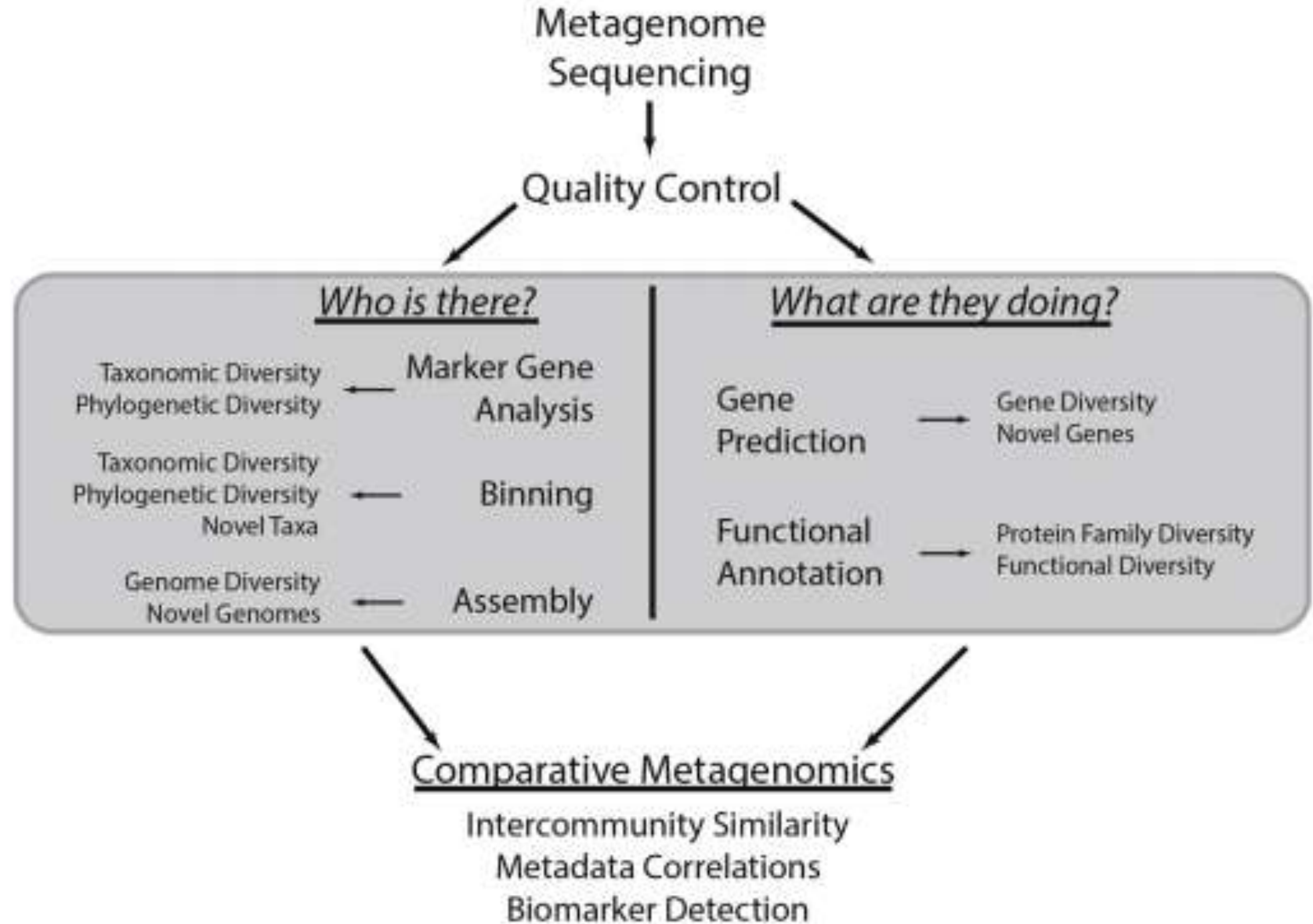
**Poor sampling beyond eukaryotic diversity**

Assembly of metagenomes is **challenging** due to uneven coverage

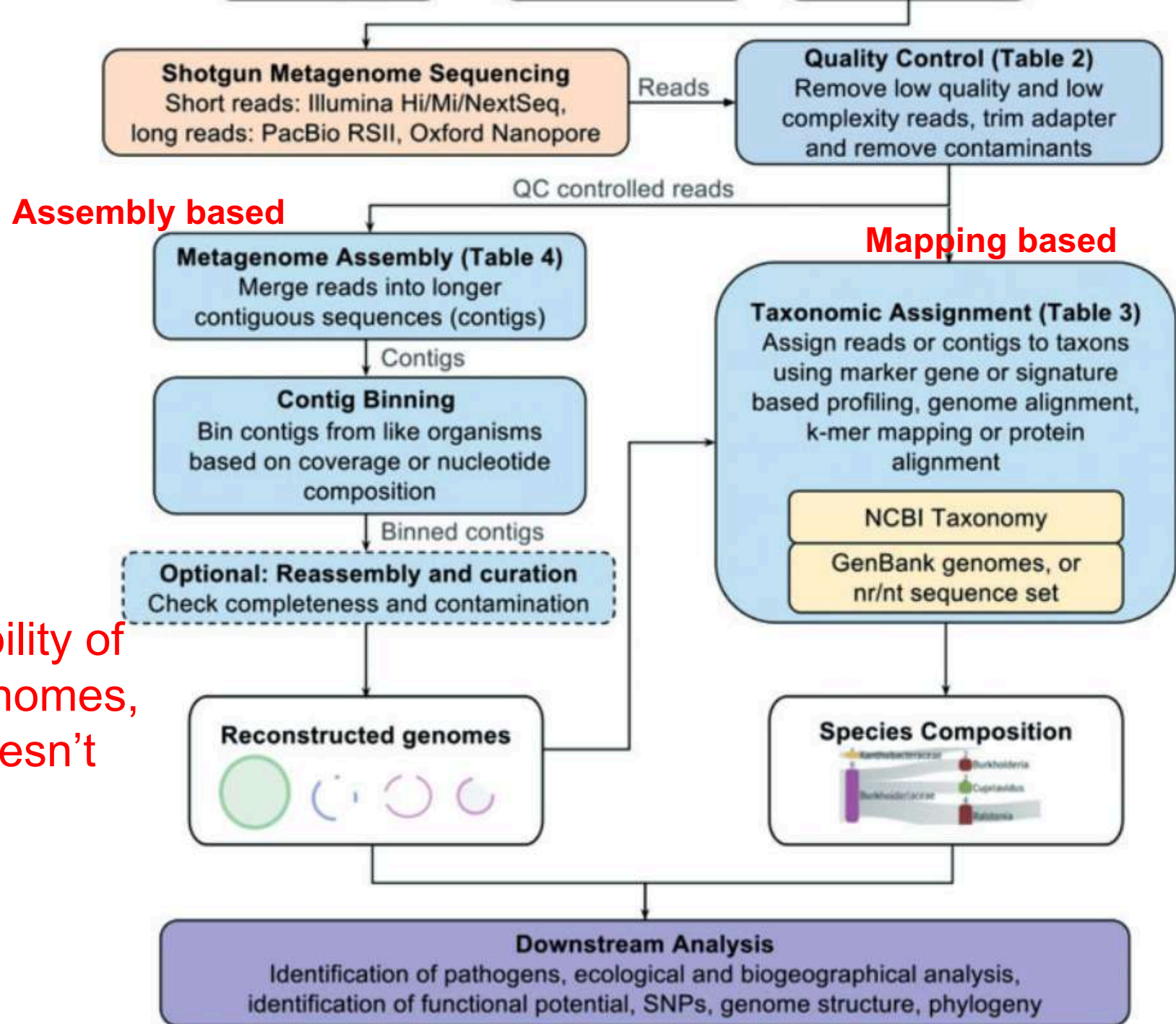
Requires **high** depth of coverage



# Overall workflow

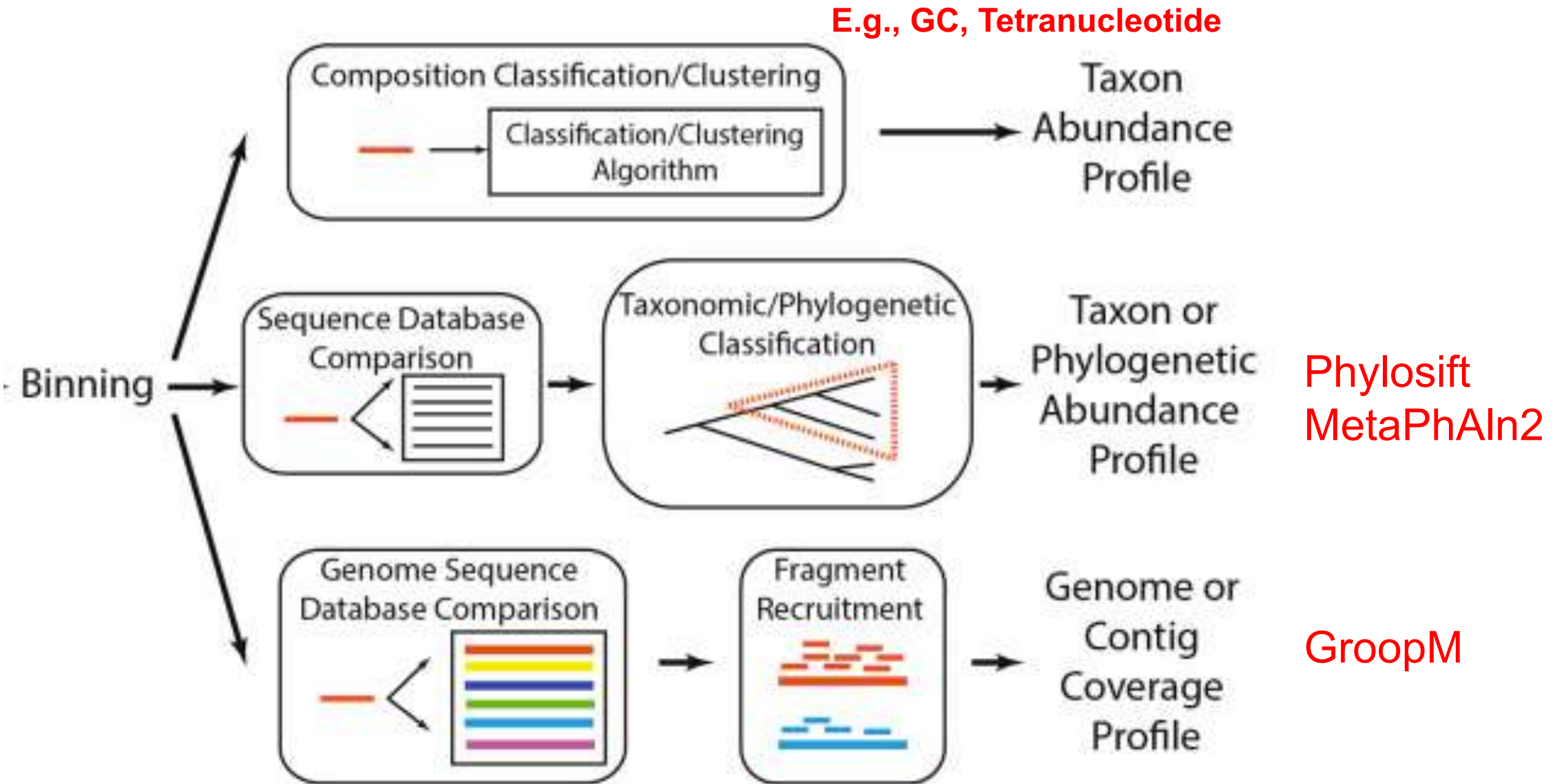


# Overall workflow

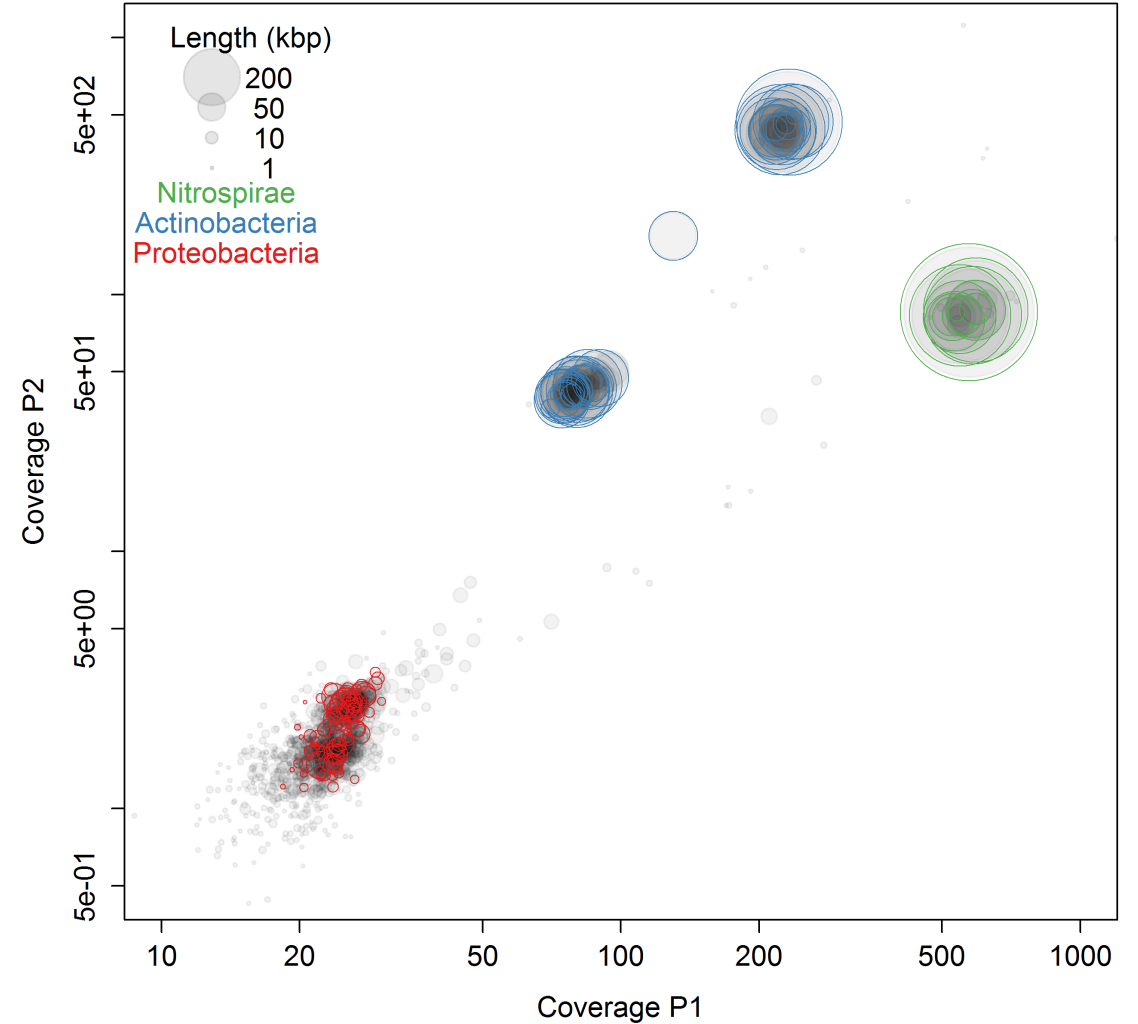
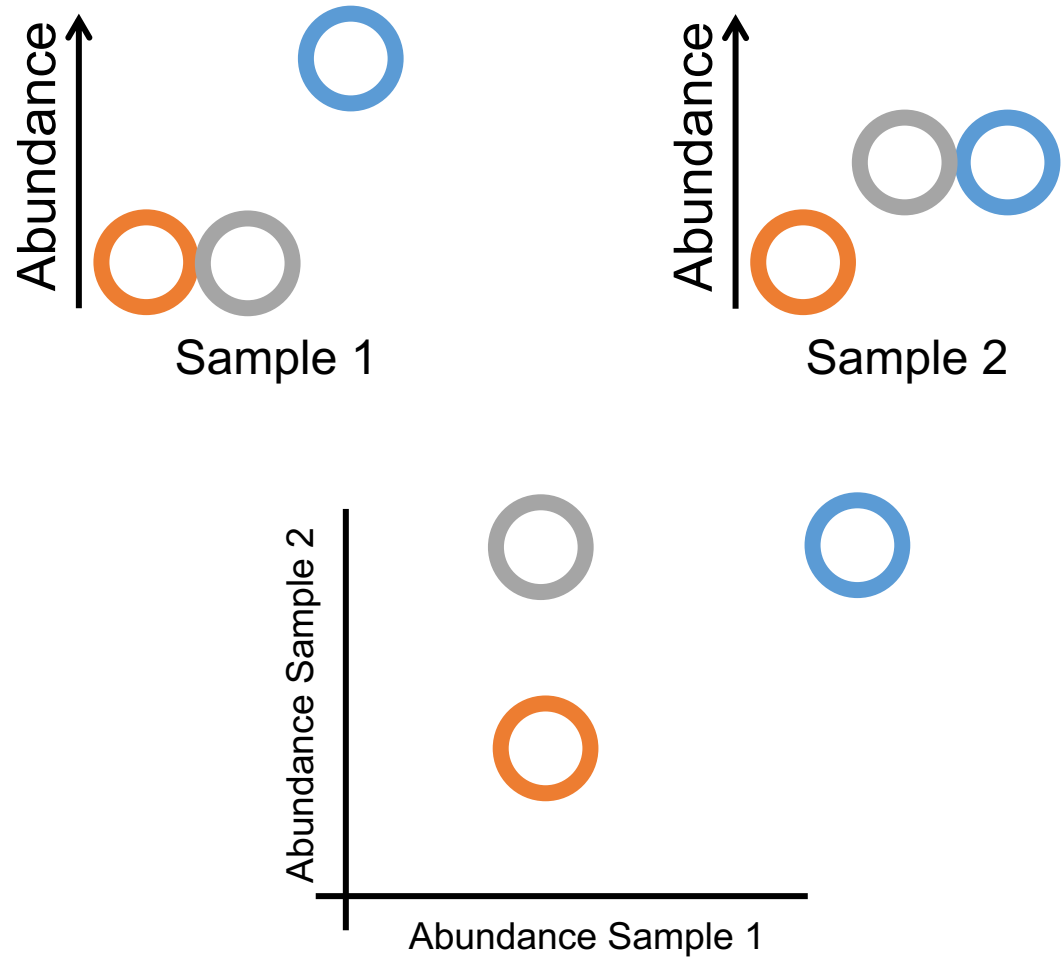


With the increase availability of reference sequenced genomes, probably one day one doesn't require assembly of metagenomes

# Binning methods



# Example of binning based on differential coverage



# Binning methods: A combination of

Classification based on **sequence composition**:

**Advantage** : all reads can be categorised into bins

**Disadvantage**: no taxonomy / function of the bins.

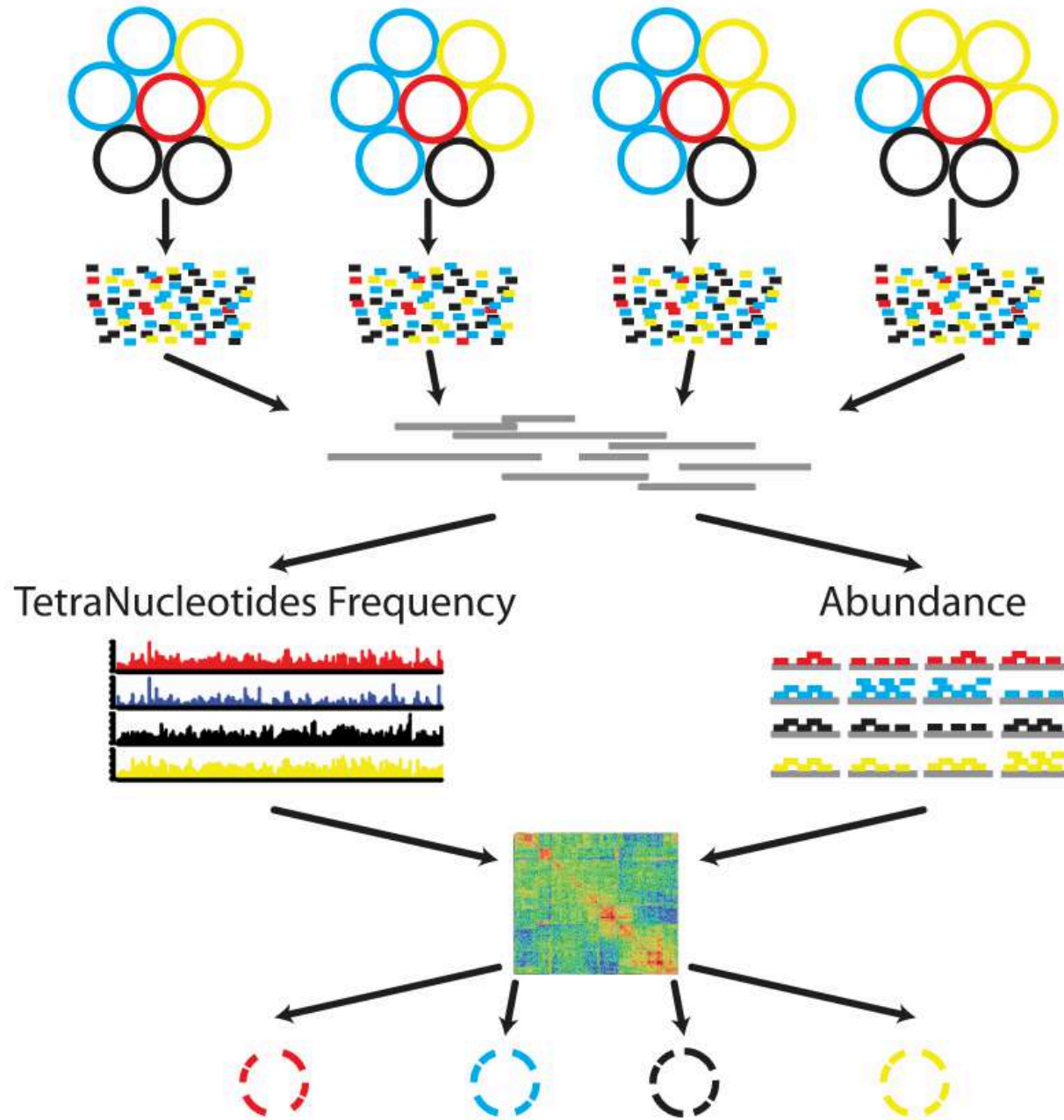
Classification based on **sequence similarity (of known genes)**

**Advantage**: One can determine taxonomy and function of reads.

**Disadvantage**: reads with similarity can not be classified .



# Metabat



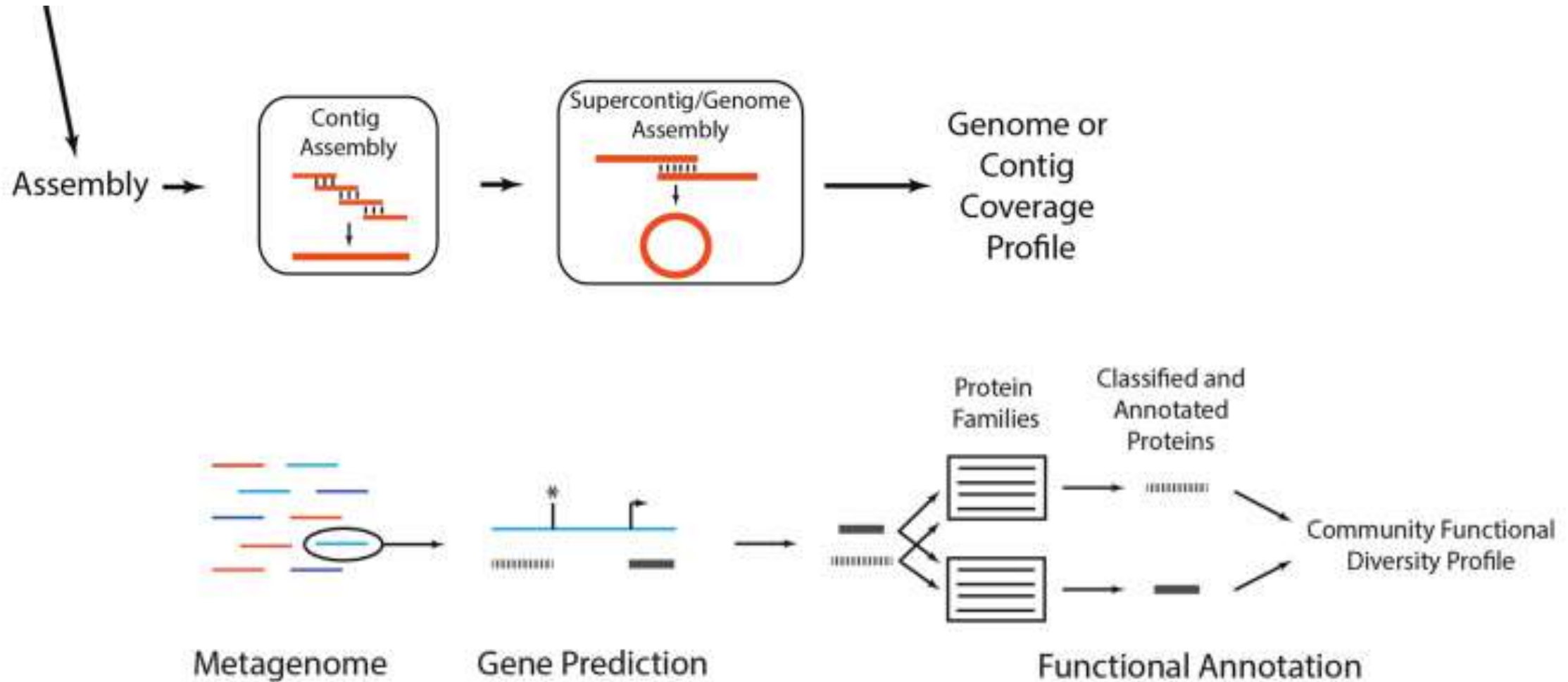
## Preprocessing

- 1 Samples from multiple sites or times
- 2 Metagenome libraries
- 3 Initial de-novo assembly using the combined library

## MetaBAT

- 4 Calculate TNF for each contig
- 5 Calculate Abundance per library for each contig
- 6 Calculate the pairwise distance matrix using pre-trained probabilistic models
- 7 Forming genome bins iteratively




# Actual assembly

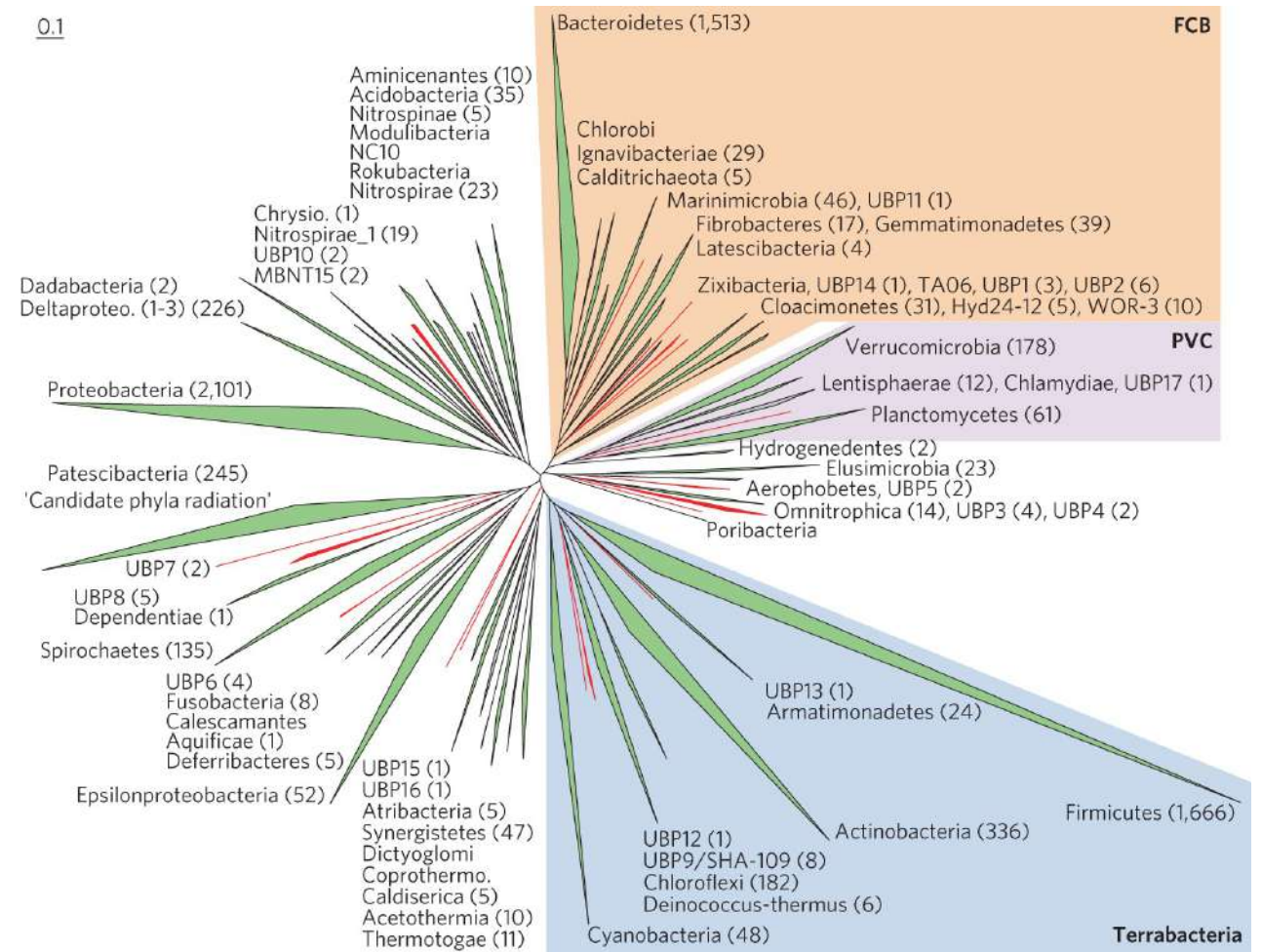


# Algorithm advancements lead to recovery of genomes



## Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks , Christian Rinke , Maria Chuvpochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz \* and Gene W. Tyson\*



The maximum likelihood tree was inferred from the concatenation of 120 proteins and spans a dereplicated set of 5,273 Uncultured Bacterial A and 14,304 NCBI genomes. Phyla containing Uncultivated Bacteria and Archea (UBA) genomes are shown in green with the number of UBA genomes indicated in parentheses. Candidate phyla consisting only of UBA genomes are shown in red and have been named Uncultured Bacterial Phylum 1 to 17 (UBP1–UBP17).



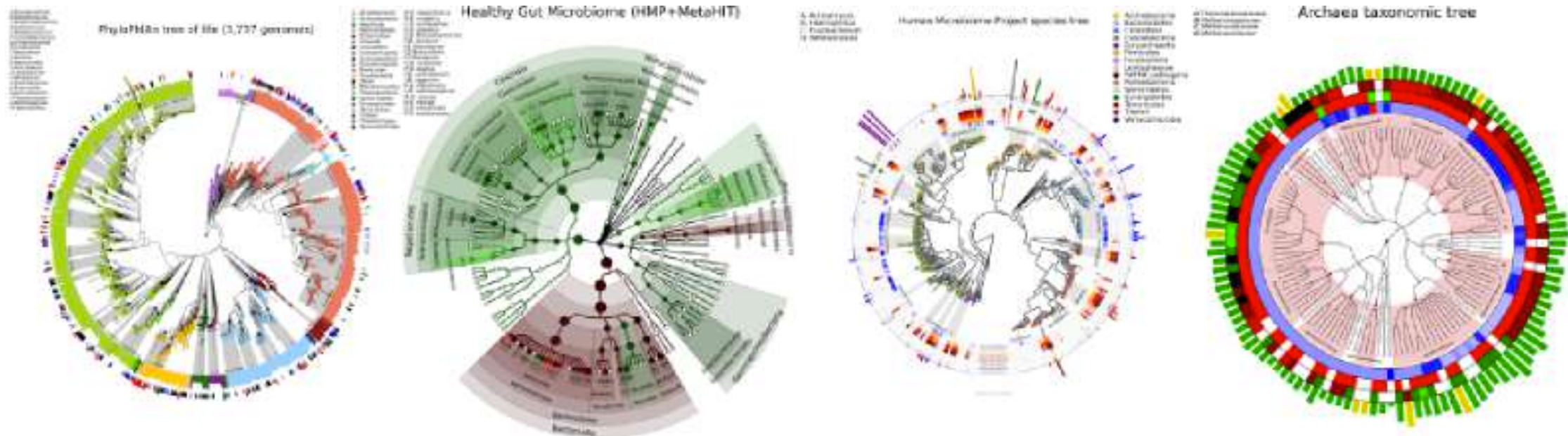
# MetaPhlAn2 – enhanced metagenomic taxonomic profiling

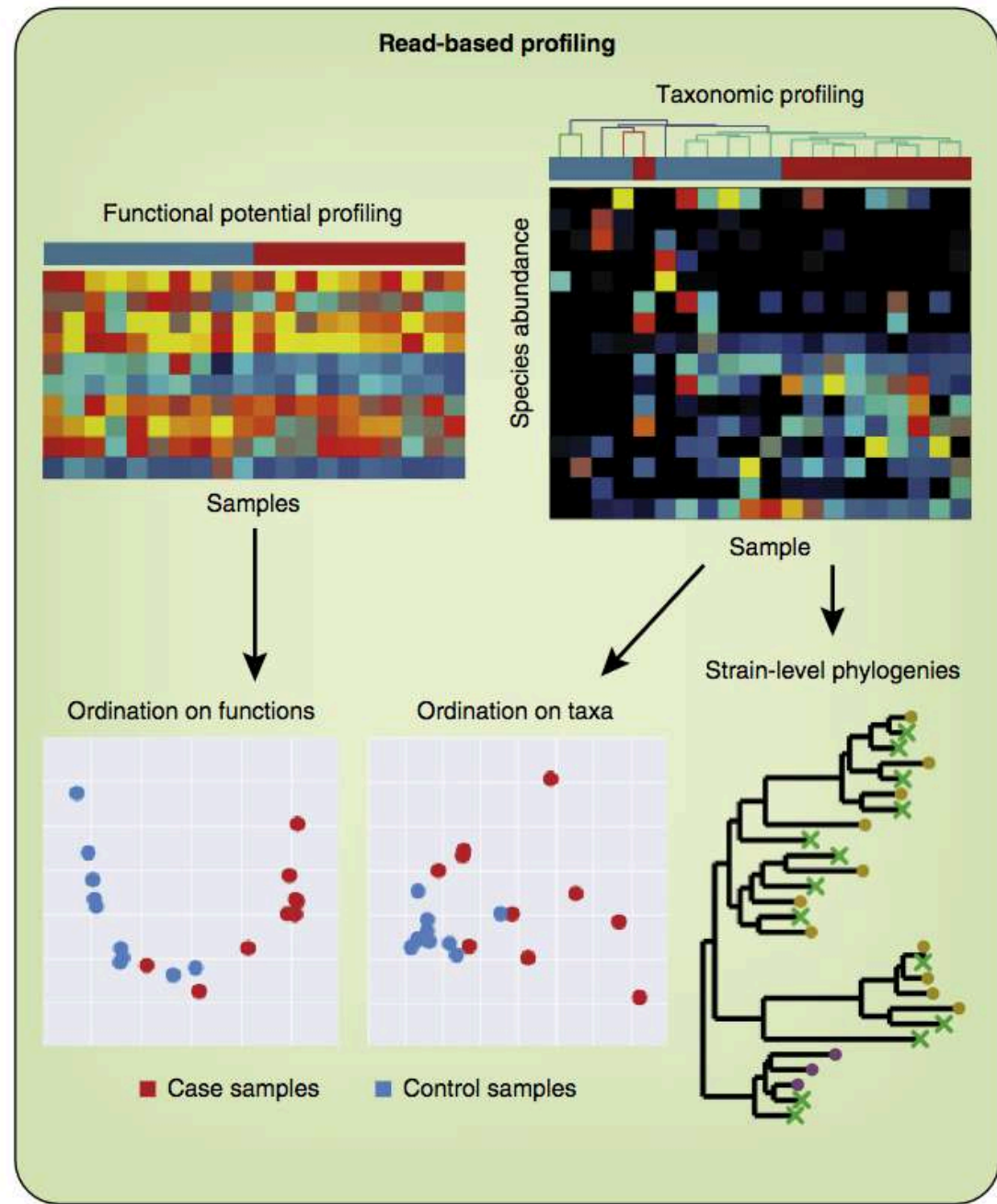
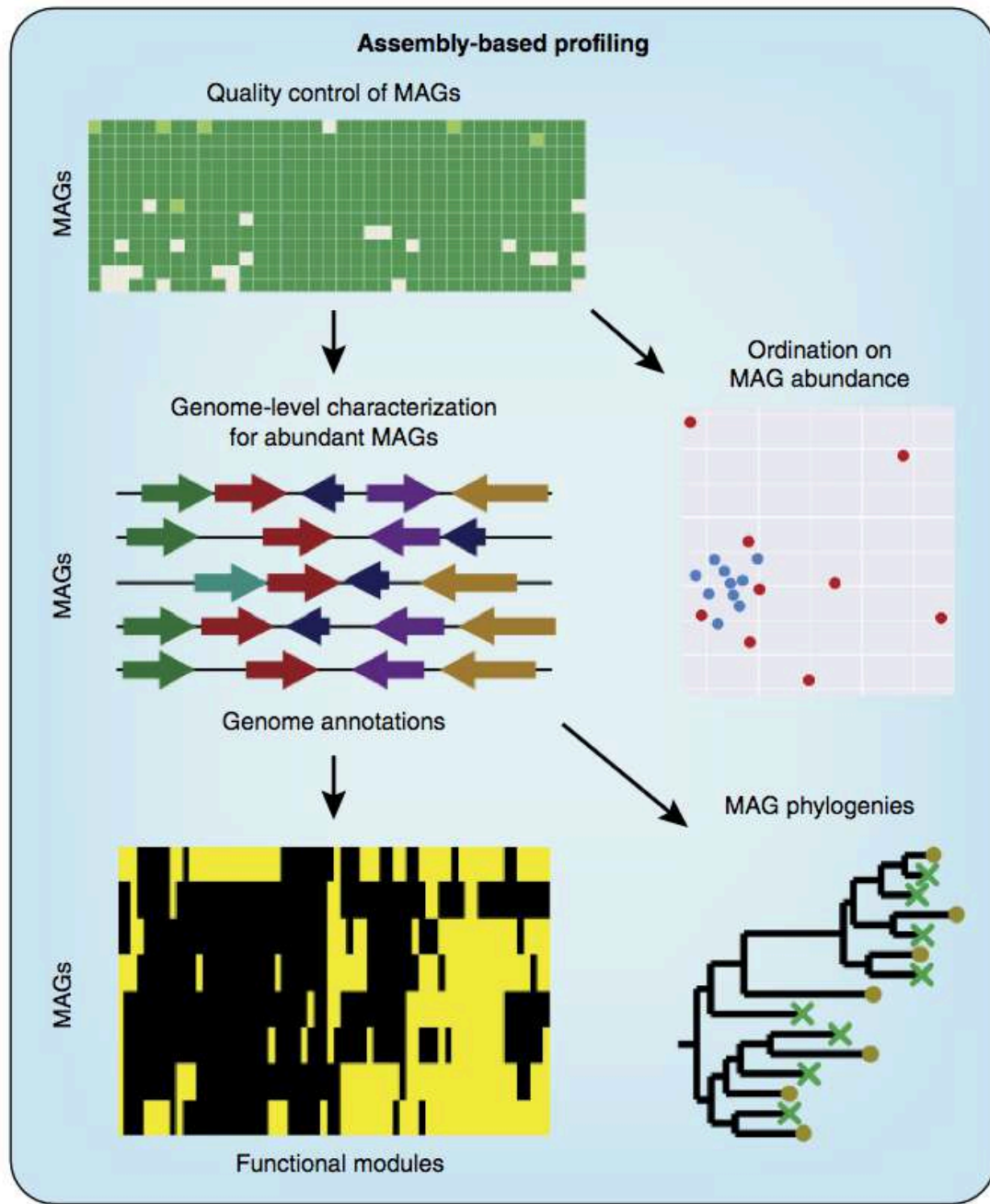
relies on **~1M unique clade-specific marker genes** identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic), allowing:

**Species** level resolution

Good visualisation with **GraphAn**

**(So it's useful with known ecosystems)**





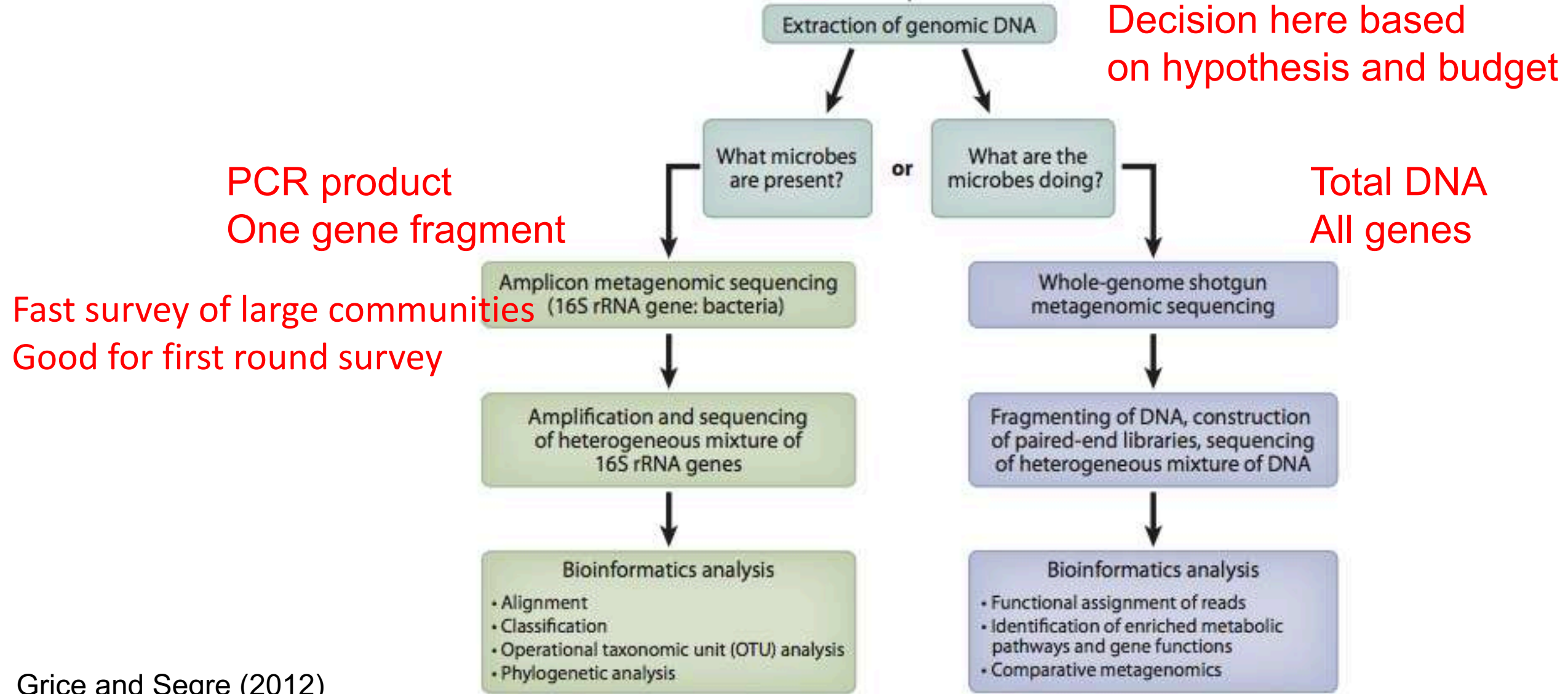


**Table 4 Strengths and weaknesses of assembly-based and read-based analyses for primary analysis of metagenomics data**

	Assembly-based analysis	Read-based analysis ('mapping')
Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned.	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference databases.
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly.	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives.	Cannot resolve organisms for which genomes of close relatives are unknown.
Computational burden	Requires computationally costly assembly, mapping and binning.	Can be performed efficiently, enabling large meta-analyses.
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity.	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes.
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection.	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision.
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates.	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates.

Amplicon sequencing or metagenomes?

# Workflow decision



**Table 1.** Metataxonomics, metagenomics and meta-transcriptomics strategies

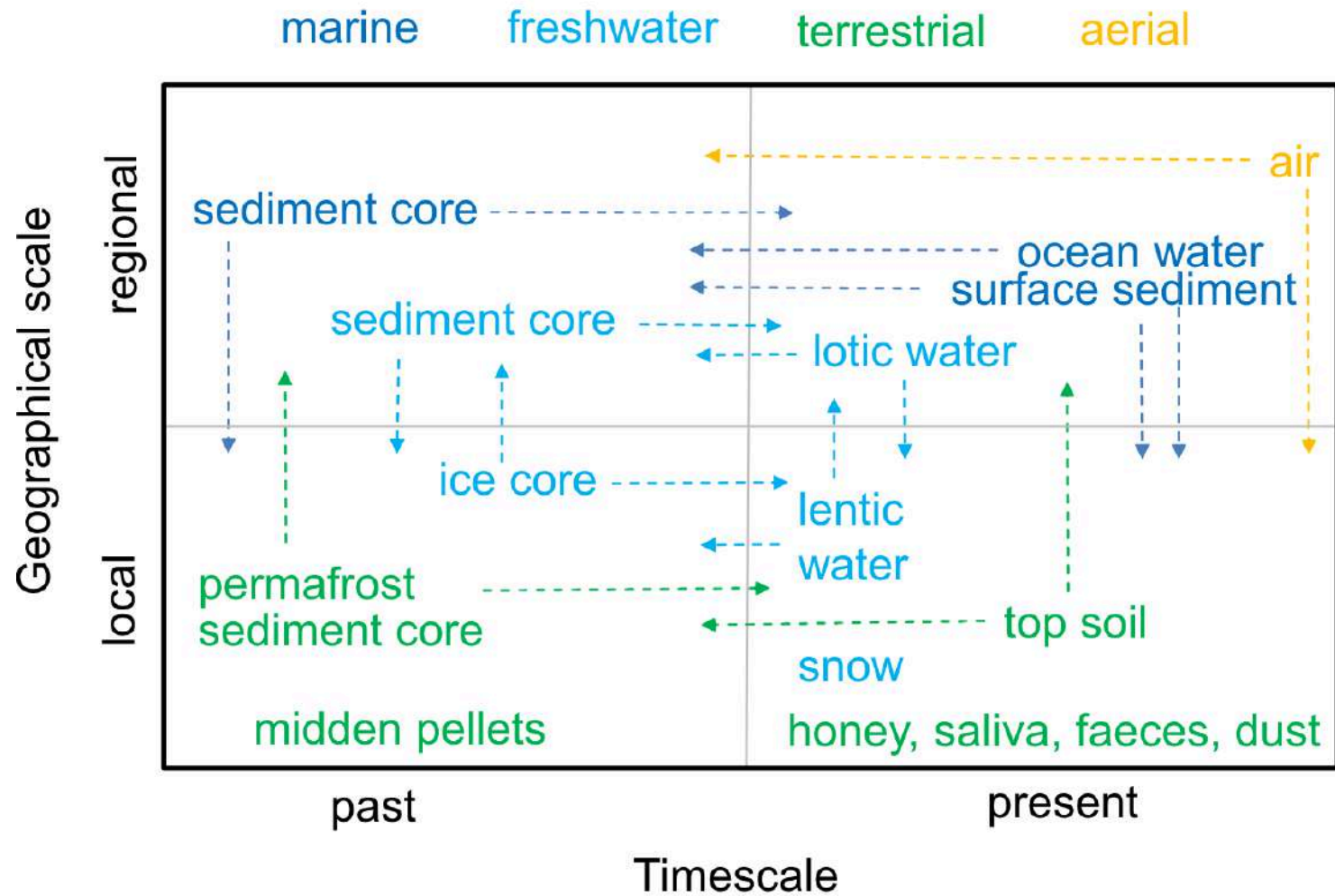
Technique	Advantages and challenges	Main applications
Metataxonomics using amplicon sequencing of the 16S or 18S rRNA gene or ITS	<ul style="list-style-type: none"><li>+ Fast and cost-effective identification of a wide variety of bacteria and eukaryotes</li><li>– Does not capture gene content other than the targeted genes</li><li>– Amplification bias</li><li>– Viruses cannot be captured</li></ul>	<ul style="list-style-type: none"><li>* Profiling of what is present</li><li>* Microbial ecology</li><li>* rRNA-based phylogeny</li></ul>
Metagenomics using random shotgun sequencing of DNA or RNA	<ul style="list-style-type: none"><li>+ No amplification bias</li><li>+ Detects bacteria, archaea, viruses and eukaryotes</li><li>+ Enables <i>de novo</i> assembly of genomes</li><li>– Requires high read count</li><li>– Many reads may be from host</li><li>– Requires reference genomes for classification</li></ul>	<ul style="list-style-type: none"><li>* Profiling of what is present across all domains</li><li>* Functional genome analyses</li><li>* Phylogeny</li><li>* Detection of pathogens</li></ul>
Meta-transcriptomics using sequencing of mRNA	<ul style="list-style-type: none"><li>+ Identifies active genes and pathways</li><li>– mRNA is unstable</li><li>– Multiple purification and amplification steps can lead to more noise</li></ul>	<ul style="list-style-type: none"><li>* Transcriptional profiling of what is active</li></ul>

Every step counts

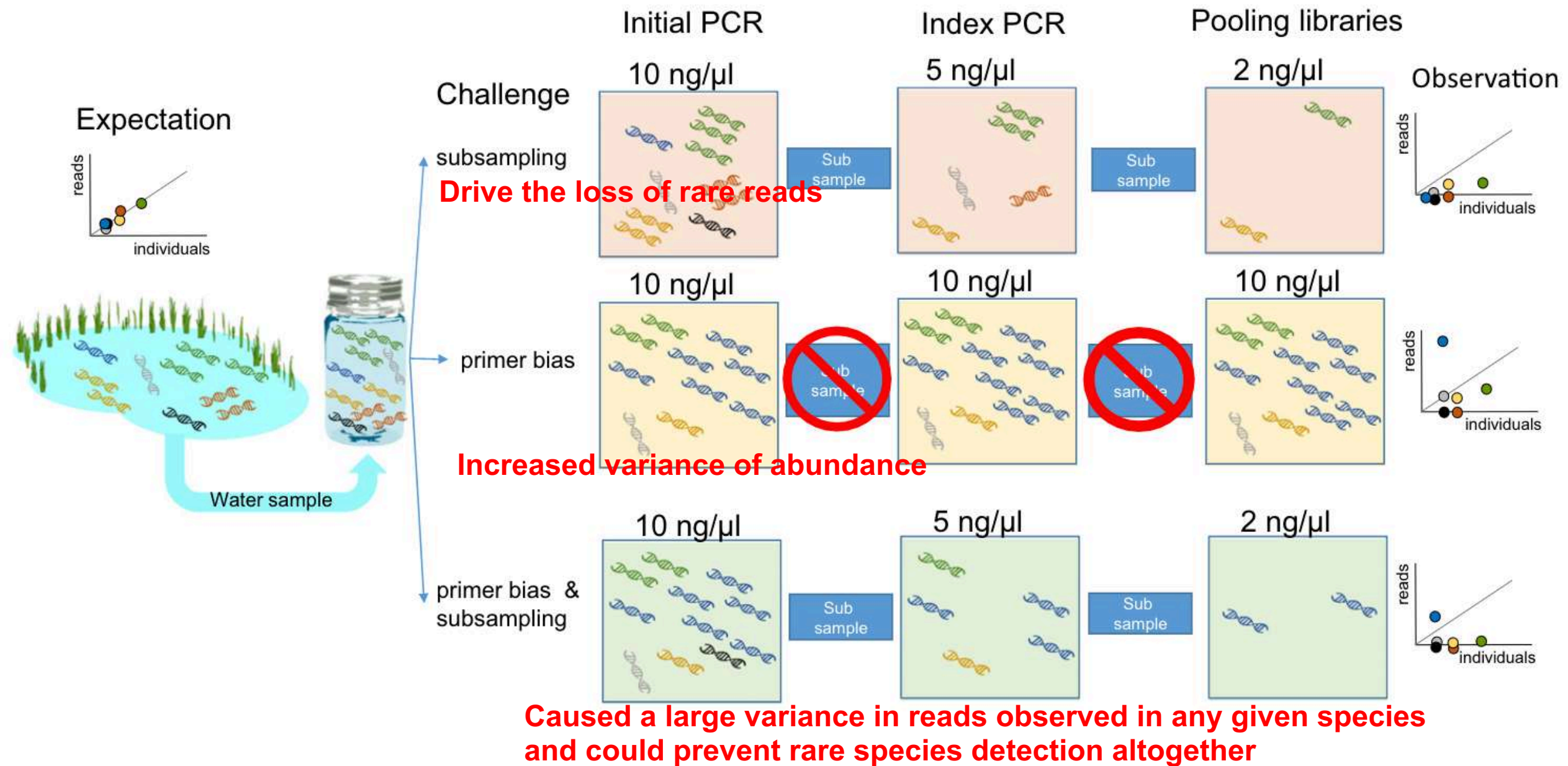


# Is there a consensus to “best practices” for 16S microbiome studies?

- Sample collection
- Sample storage
- Fresh versus Frozen samples
- Use of cryoprotectant
- DNA extraction
- Sequencing strategy
- Mock bacterial communities
- Analysis strategy
- OTU picking methods
- Correcting for gene copy number
- Contamination issues



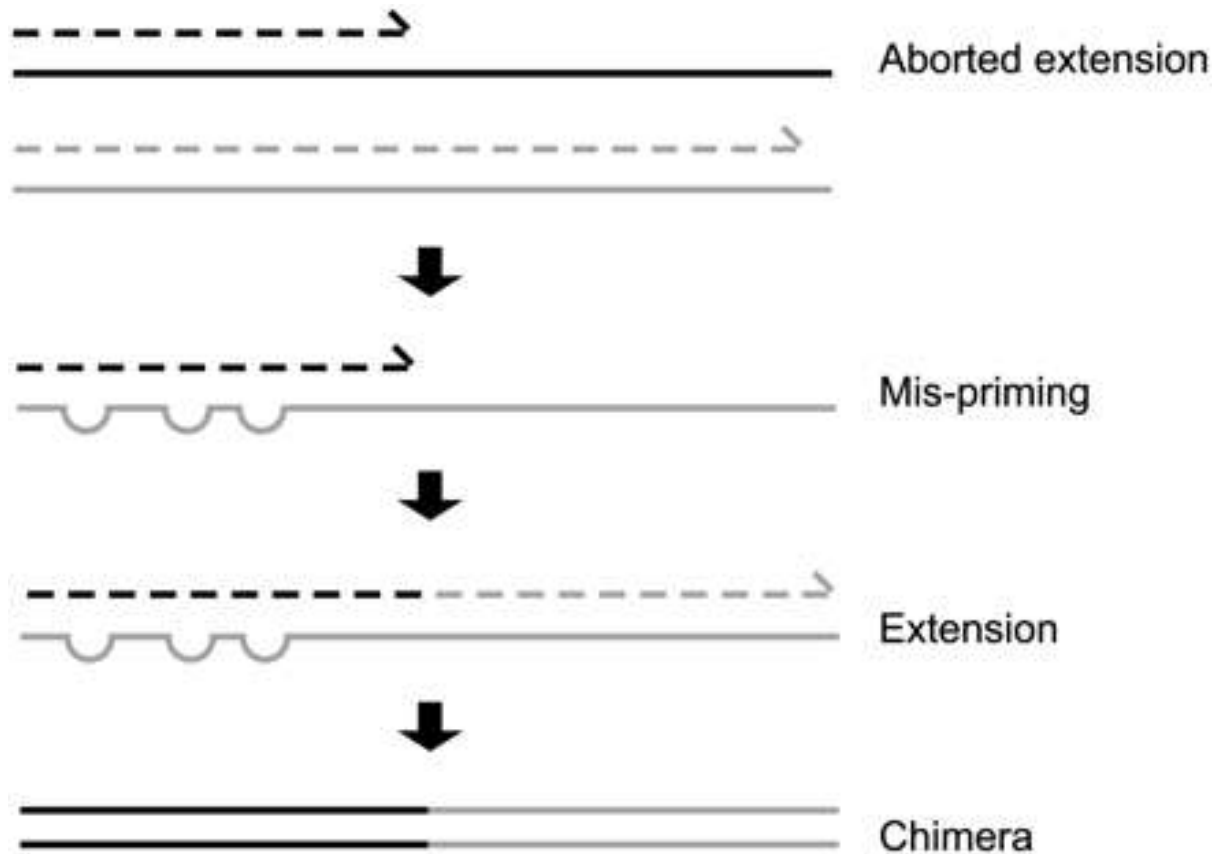
**FIGURE 1** Environmental DNA sample types have different spatial and temporal scopes of inference from different habitats. Consider each sample type as a single sample from that environment. Placement of a sample type in a quadrant is not quantitative, but represents a common scale at which it has been used. Dashed arrows indicate the potential for a sample type to confer information at multiple scales of inference, but additional research to quantify these possibilities is needed



# Potential problems of using amplicons

- Lack of tools for processing ITS/Fungal microbiome data sets
- Amplification bias effects accuracy and replication
- Use of short reads prevents disambiguation of similar strains
- 16S or ITS may not differentiate between similar strains –
  - Clustering is done at 97%
  - Regions may be >99% similar
- Sequencing error inflates number of OTUs
- Chloroplast 16S sequences can get amplified in plant metagenomes

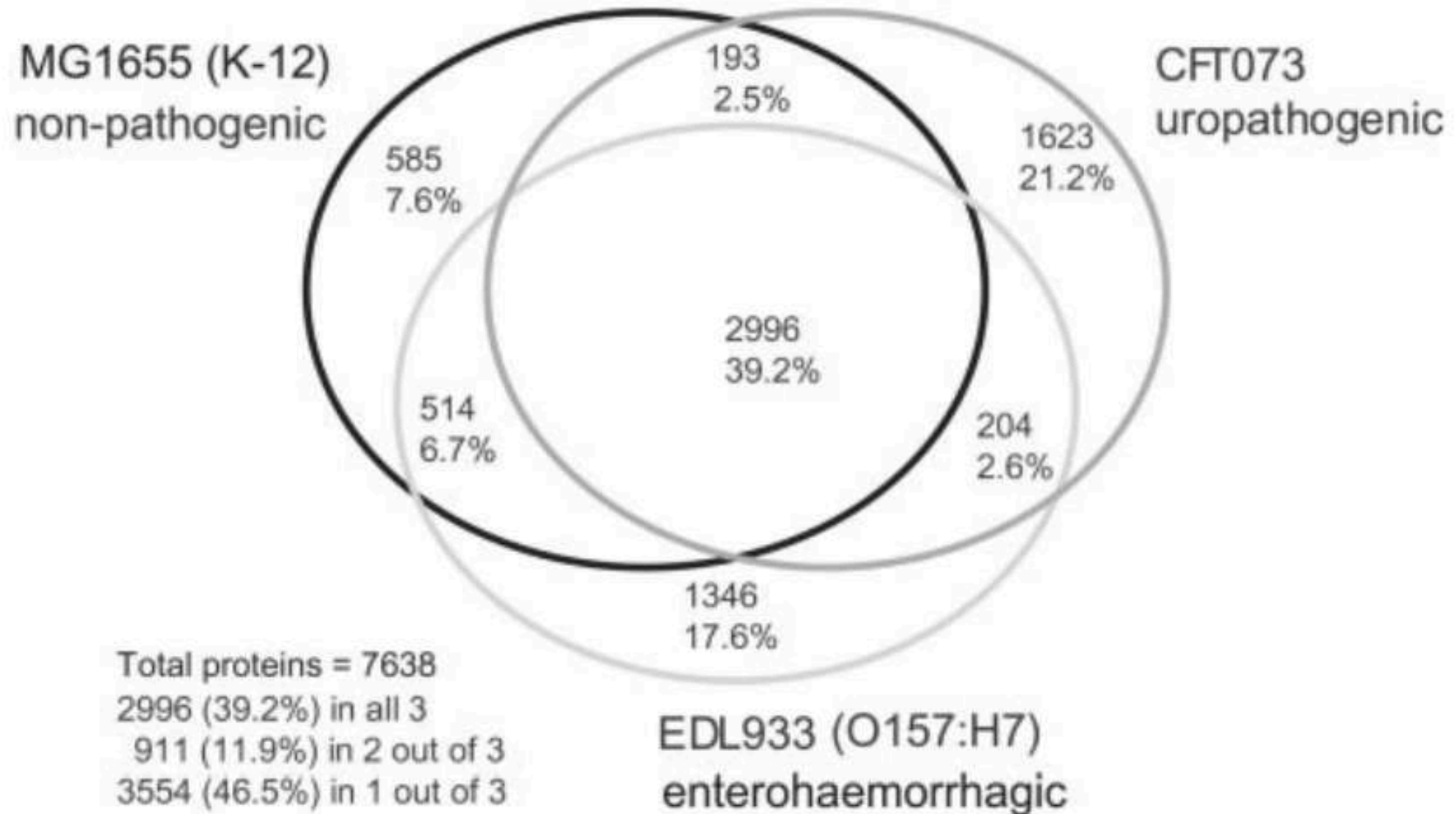
# Chimeric 16S (Artificial sequences formed during PCR amplification)



“Chimeras were found to reproducibly form among independent amplifications and contributed to false perceptions of sample diversity and the false identification of novel taxa, **with less-abundant species exhibiting chimera rates exceeding 70%**”



# Same species (16S): Different genomes



# Database error

## At Least 1 in 20 16S rRNA Sequence Records Currently Held in Public Repositories Is Estimated To Contain Substantial Anomalies

Kevin E. Ashelford,<sup>1\*</sup> Nadia A. Chuzhanova,<sup>3</sup> John C. Fry,<sup>1</sup> Antonia J. Jones,<sup>2</sup>  
and Andrew J. Weightman<sup>1</sup>

*Cardiff School of Biosciences, Cardiff University, Main Building, Park Place, P.O. Box 915, Cardiff CF10 3TL, United Kingdom<sup>1</sup>; Cardiff School of Computer Science, Cardiff University, Queen's Buildings, 5 The Parade, Roath, Cardiff CF24 3AA, United Kingdom<sup>2</sup>; and Biostatistics and Bioinformatics Unit and Institute of Medical Genetics, Cardiff School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, United Kingdom<sup>3</sup>*

Received 7 April 2005/Accepted 28 July 2005

**A new method for detecting chimeras and other anomalies within 16S rRNA sequence records is presented. Using this method, we screened 1,399 sequences from 19 phyla, as defined by the Ribosomal Database Project, release 9, update 22, and found 5.0% to harbor substantial errors. Of these, 64.3% were obvious chimeras, 14.3% were unidentified sequencing errors, and 21.4% were highly degenerate. In all, 11 phyla contained obvious chimeras, accounting for 0.8 to 11% of the records for these phyla. Many chimeras (43.1%) were formed from parental sequences belonging to different phyla. While most comprised two fragments, 13.7% were composed of at least three fragments, often from three different sources. A separate analysis of the *Bacteroidetes* phylum (2,739 sequences) also revealed 5.8% records to be anomalous, of which 65.4% were apparently chimeric. Overall, we conclude that, as a conservative estimate, 1 in every 20 public database records is likely to be corrupt. Our results support concerns recently expressed over the quality of the public repositories. With 16S rRNA sequence data increasingly playing a dominant role in bacterial systematics and environmental biodiversity studies, it is vital that steps be taken to improve screening of sequences prior to submission. To this end, we have implemented our method as a program with a simple-to-use graphic user interface that is capable of running on a range of computer platforms. The program is called Pintail, is released under the terms of the GNU General Public License open source license, and is freely available from our website at <http://www.cardiff.ac.uk/biosi/research/biosoft/>.**



# Extraction protocol matters



Soil Biology & Biochemistry 36 (2004) 1607–1614

Soil Biology &  
Biochemistry

www.elsevier.com/locate/soilbio

## Impact of DNA extraction method on bacterial community composition measured by denaturing gradient gel electrophoresis

Julia R. de Liphay<sup>a,b</sup>, Christiane Enzinger<sup>b,1</sup>, Kaare Johnsen<sup>a,2</sup>, Jens Aamand<sup>a</sup>, Søren J. Sørensen<sup>b,\*</sup>

<sup>a</sup>Department of Geochemistry, Geological Survey of Denmark and Greenland, Øster Voldgade 10, DK-1350 Copenhagen K, Denmark

<sup>b</sup>Department of Microbiology, University of Copenhagen, Sølvgade 83H, DK-1307 Copenhagen K, Denmark

Received 1 September 2003; received in revised form 6 March 2004; accepted 15 March 2004

### Abstract

The impact of DNA extraction protocol on soil DNA yield and bacterial community composition was evaluated. Three different procedures to physically disrupt cells were compared: sonication, grinding–freezing–thawing, and bead beating. The three protocols were applied to three different topsoils. For all soils, we found that each DNA extraction method resulted in unique community patterns as measured by denaturing gradient gel electrophoresis. This indicates the importance of the DNA extraction protocol on data for evaluating soil bacterial diversity. Consistently, the bead-beating procedure gave rise to the highest number of DNA bands, indicating the highest number of bacterial species. Supplementing the bead-beating procedure with additional cell-rupture steps generally did not change the bacterial community profile. The same consistency was not observed when evaluating the efficiency of the different methods on soil DNA yield. This parameter depended on soil type. The DNA size was of highest molecular weight with the sonication and grinding–freezing–thawing procedures (approx. 20 kb). In contrast, the inclusion of bead beating resulted in more sheared DNA (approx. 6–20 kb), and the longer the bead-beating time, the higher the fraction of low-molecular weight DNA. Clearly, the choice of DNA extraction protocol depends on soil type. We found, however, that for the analysis of indigenous soil bacterial communities the bead-beating procedure was appropriate because it is fast, reproducible, and gives very pure DNA of relatively high molecular weight. And very importantly, with this protocol the highest soil bacterial diversity was obtained. We believe that the choice of DNA extraction protocol will influence not only the determined phylogenetic diversity of indigenous microbial communities, but also the obtained functional diversity. This means that the detected presence of a functional gene—and thus the indication of enzyme activity—may depend on the nature of the applied DNA extraction procedure.

“we found that each DNA extraction method resulted in unique community patterns”

Wesolowska-Andersen et al. *Microbiome* 2014, 2:19  
<http://www.microbiomejournal.com/content/2/1/19>



RESEARCH

Open Access

## Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis

Agata Wesolowska-Andersen<sup>1</sup>, Martin Iain Bahi<sup>2</sup>, Vera Carvalho<sup>2</sup>, Karsten Kristiansen<sup>3</sup>, Thomas Sicheritz-Pontén<sup>1</sup>, Ramneek Gupta<sup>1\*</sup> and Tine Rask Licht<sup>2\*</sup>

### Abstract

**Background:** In recent years, studies on the human intestinal microbiota have attracted tremendous attention. Application of next generation sequencing for mapping of bacterial phylogeny and function has opened new doors to this field of research. However, little attention has been given to the effects of choice of methodology on the output resulting from such studies.

**Results:** In this study we conducted a systematic comparison of the DNA extraction methods used by the two major collaborative efforts: The European MetaHIT and the American Human Microbiome Project (HMP). Additionally, effects of homogenizing the samples before extraction were addressed. We observed significant differences in distribution of bacterial taxa depending on the method. While eukaryotic DNA was most efficiently extracted by the MetaHIT protocol, DNA from bacteria within the Bacteroidetes phylum was most efficiently extracted by the HMP protocol.

**Conclusions:** Whereas it is comforting that the inter-individual variation clearly exceeded the variation resulting from choice of extraction method, our data highlight the challenge of comparing data across studies applying different methodologies.

“We observed significant differences in distribution of bacterial taxa depending on the method.”

# Alpha diversity is always overestimated

**Table 1.** Effect of quality filtering and clustering on diversity estimates (OTU number), error rate and data loss of pyrotags amplified from two regions of *E. coli* MG1655 16S rRNA genes.

Read filtering	Number of OTUs at percentage identity thresholds						% errorless reads	% reads used
	100	99	98	97	95	90		
<b>5' forward (V1 and V2)</b>								
Theoretical number	5	4	3	1	1	1		
No quality filtering	643	95	31	16	5	3	68.7	77.9
Reads with N's removed	600	85	29	14	4	3	69.8	76.7
Quality score-based filtering (% per-base error probability)								
3	638	92	31	13	3	3	68.9	77.7
2	632	90	30	14	3	3	69.0	77.6
1	609	79	24	9	3	3	69.1	77.3
0.5	562	66	15	7	3	3	70.7	75.3
0.2	469	30	6	3	3	3	73.2	70.8
0.1	372	26	5	3	3	3	77.8	57.8
<b>3' reverse (V8)</b>								
Theoretical number	1	1	1	1	1	1		
No quality filtering	385	43	13	7	5	4	84.6	94.4
Reads with N's removed	361	40	12	6	4	3	85.3	93.6
Quality score-based filtering (% per-base error probability)								
3	378	40	12	7	5	4	84.8	94.2
2	368	32	10	6	5	4	85.1	93.8
1	342	25	9	6	5	4	85.3	93.3
0.5	310	20	8	6	5	4	87.5	89.5
0.2	236	7	2	2	2	2	89.6	82.1
0.1	196	4	2	2	2	2	90.7	70.6

Diversity estimates should be considered relative to the theoretical number of OTUs from *E. coli*.

Kunin et al (2010)

# Reagent and laboratory contamination

**RESEARCH ARTICLE**

**Open Access**

## Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter<sup>1\*</sup>, Michael J Cox<sup>2</sup>, Elena M Turek<sup>2</sup>, Szymon T Calus<sup>3</sup>, William O Cookson<sup>2</sup>, Miriam F Moffatt<sup>2</sup>, Paul Turner<sup>4,5</sup>, Julian Parkhill<sup>1</sup>, Nicholas J Loman<sup>3</sup> and Alan W Walker<sup>1,6\*</sup>

**RESEARCH HIGHLIGHT**

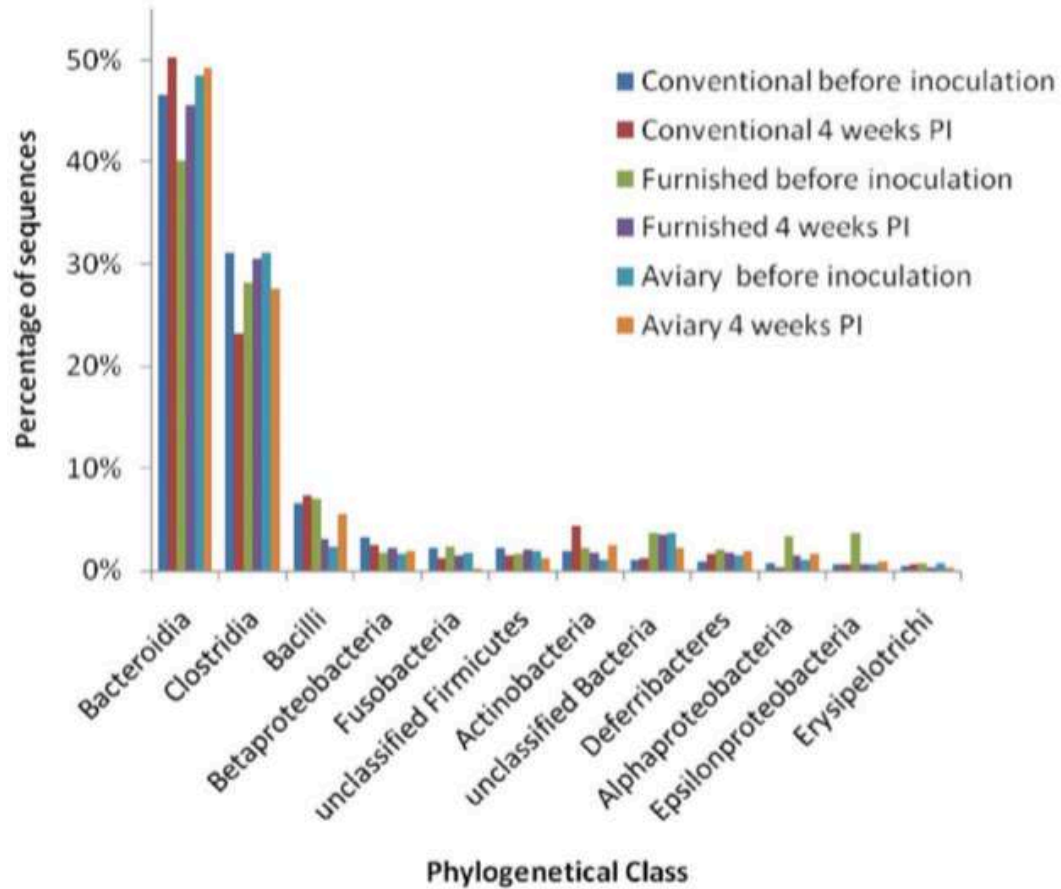
## Tracking down the sources of experimental contamination in microbiome studies

Sophie Weiss<sup>1</sup>, Amnon Amir<sup>2</sup>, Embriette R Hyde<sup>2</sup>, Jessica L Metcalf<sup>2</sup>, Se Jin Song<sup>2</sup> and Rob Knight<sup>2,3,4\*</sup>

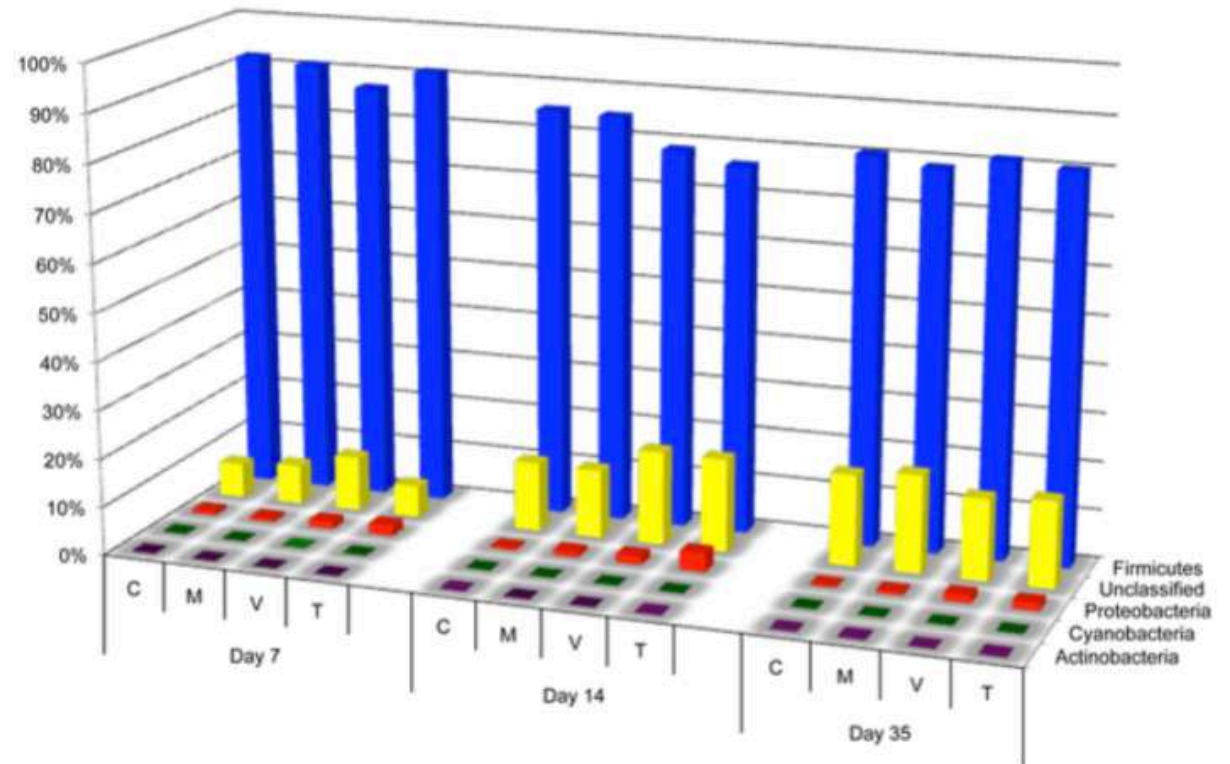


# 2 papers with different results at the same year

Bacteroidetes >>> rest



firmicutes >>> rest > bacteroidetes

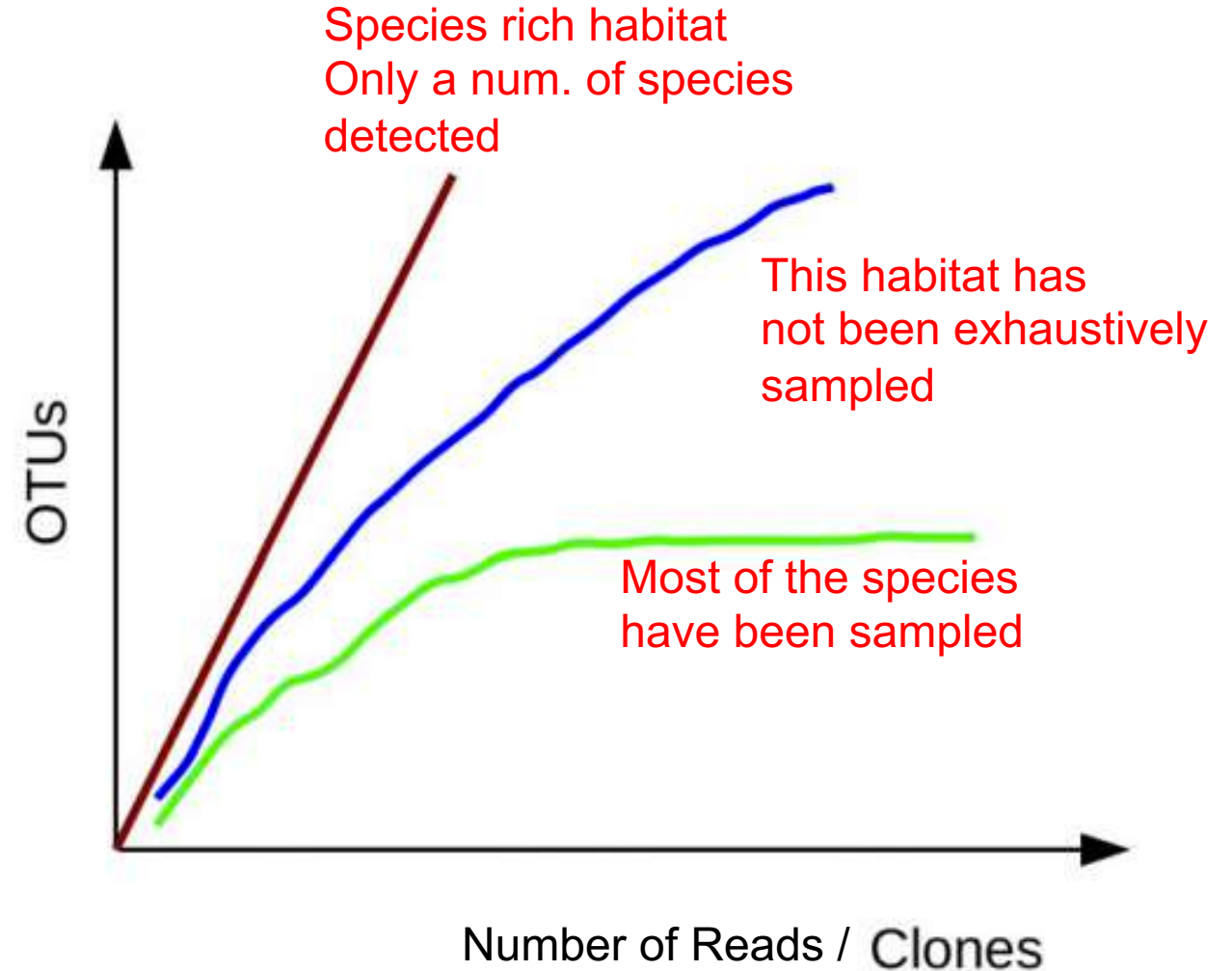


Nordentoft S *et al* (2011) BMC Microbiology

Danzeisen JL *et al* (2011) PLOS one

# Species sampling and Rarefaction

**Rarefaction** allows the calculation of **species richness** for a given number of individual samples, based on the construction of so-called **rarefaction curves**. This curve is a plot of the number of species as a function of the number of samples



# But rarefying microbiome data is wrong

OPEN ACCESS Freely available online

 PLOS | COMPUTATIONAL BIOLOGY

## Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes\*

Statistics Department, Stanford University, Stanford, California, United States of America

Current practice in the normalization of microbiome count data is inefficient in the statistical sense. .... Moreover, specific implementations for DNA sequencing read count data (based on a Negative Binomial model for instance) are already available in RNA-Seq focused R packages such as edgeR and DESeq.... We show how both proportions and rarefied counts result in a high rate of false positives in tests for species that are differentially abundant across sample classes. Regarding microbiome sample-wise clustering, we also show that the rarefying procedure often discards samples that can be accurately clustered by alternative methods. Based on these results and well-established statistical theory, **we advocate that investigators avoid rarefying altogether**. We have provided microbiome-specific extensions to these tools in the R package, phyloseq.

Original Abundance			Rarefied Abundance		
	A	B		A	B
OTU1	62	500	OTU1	62	50
OTU2	38	500	OTU2	38	50
<b>Total</b>	<b>100</b>	<b>1000</b>		<b>100</b>	<b>100</b>

### Standard Tests for Difference

	P-value	chi-2	Prop	Fisher
Original		0.0290	0.0290	0.0272
Rarefied		0.1171	0.1171	0.1169

**Figure 1. A minimal example of the effect of rarefying on statistical power.** Hypothetical abundance data in its original (Top-Left) and rarefied (Top-Right) form, with corresponding formal test results for differentiation (Bottom).

doi:10.1371/journal.pcbi.1003531.g001

# Is it time to revisit bacterial taxonomy?

## RESOURCE

nature  
biotechnology

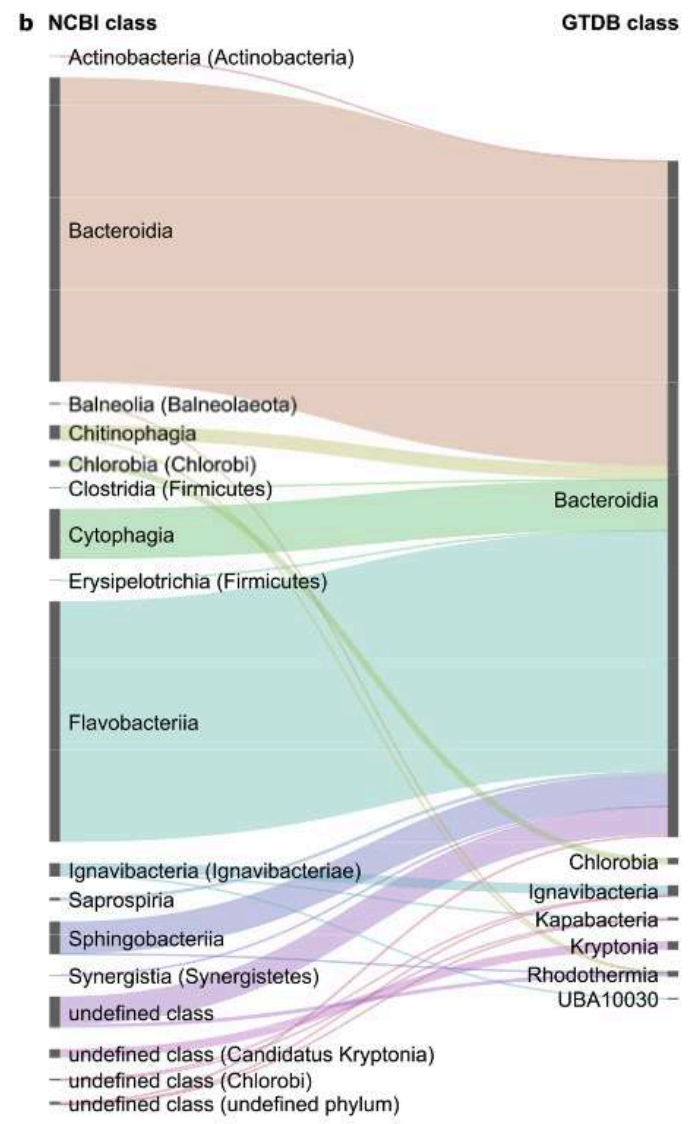
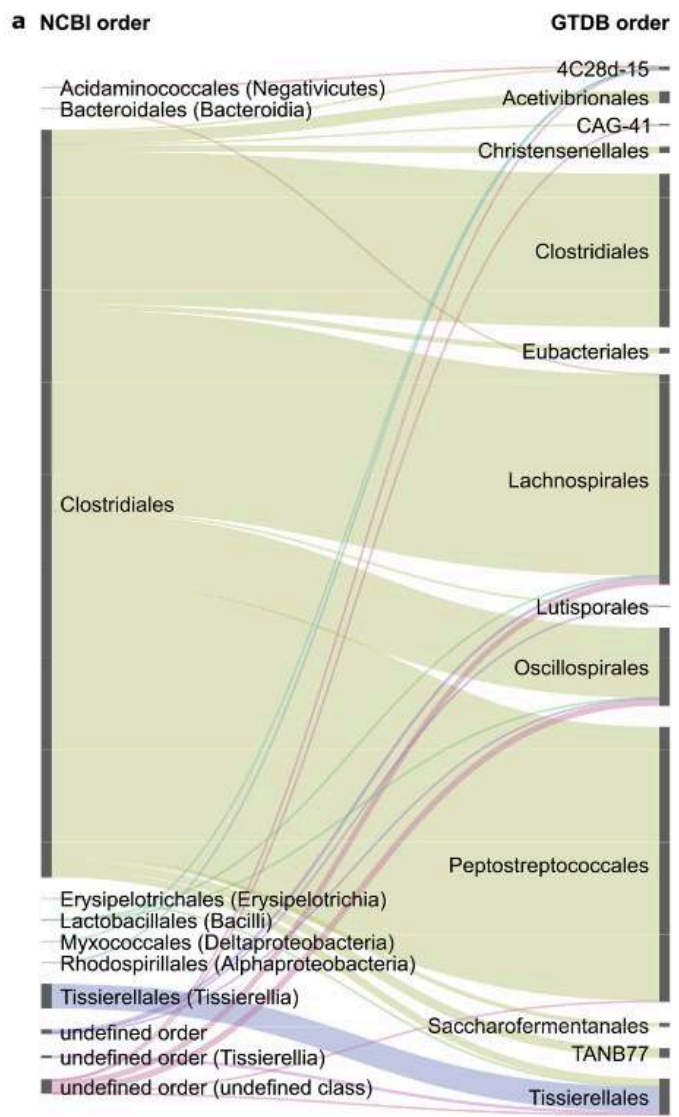
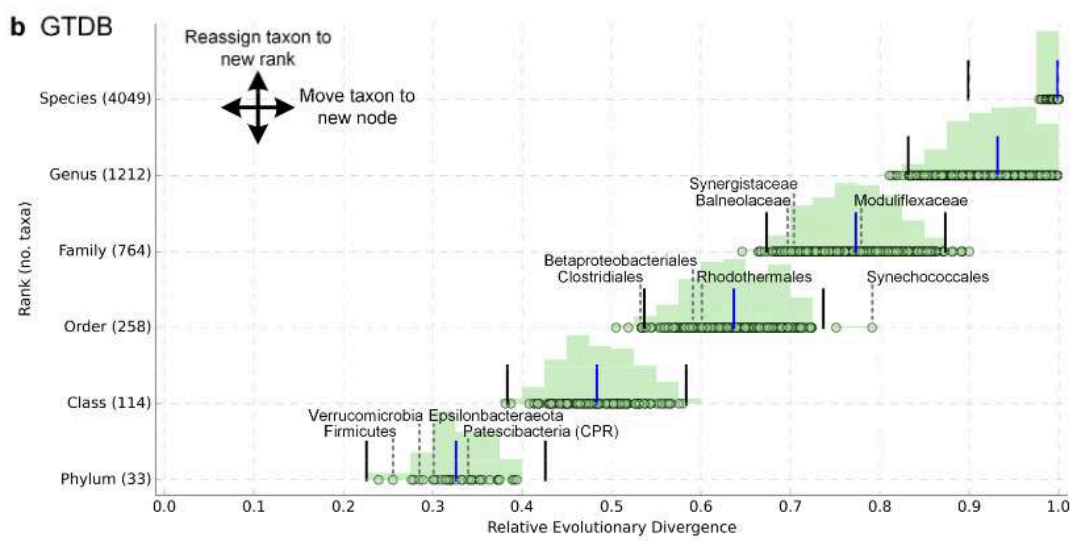
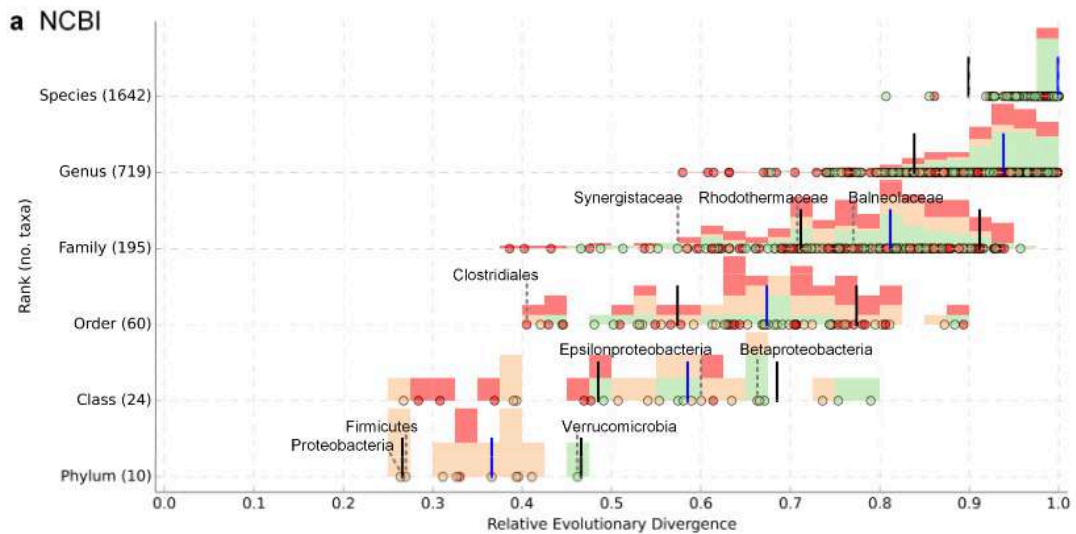
A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life

Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke<sup>id</sup>, Adam Skarshewski, Pierre-Alain Chaumeil & Philip Hugenholtz<sup>id</sup>

Under this approach, **58% of the 94,759 genomes comprising the Genome Taxonomy Database had changes to their existing taxonomy.** This result includes the description of 99 phyla, including six major monophyletic units from the subdivision of the Proteobacteria, and amalgamation of the Candidate Phyla Radiation into a single phylum. Our taxonomy should enable improved classification of uncultured bacteria and provide a sound basis for ecological and evolutionary studies.

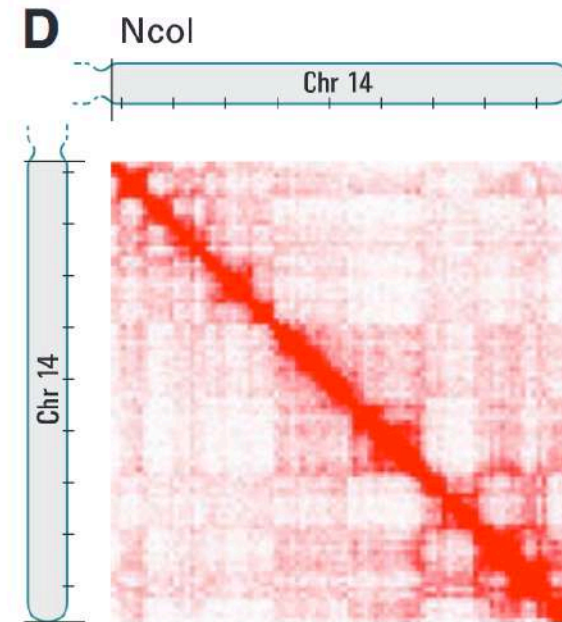
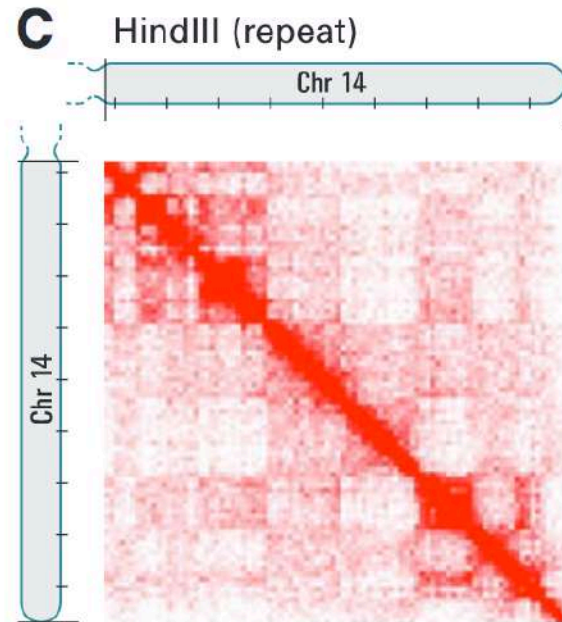
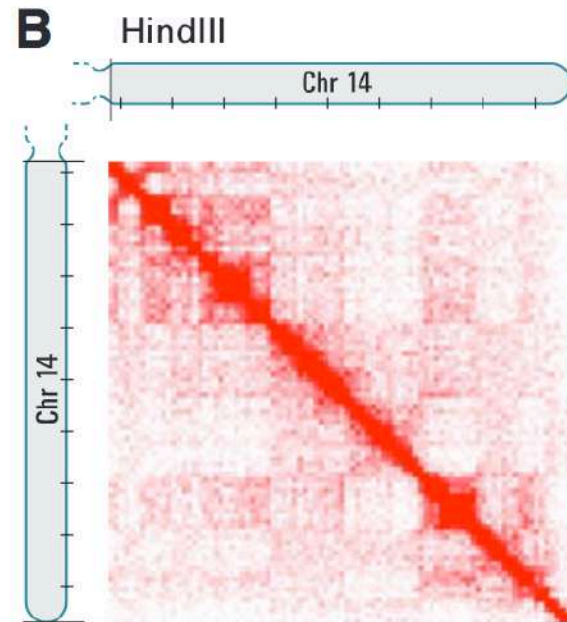
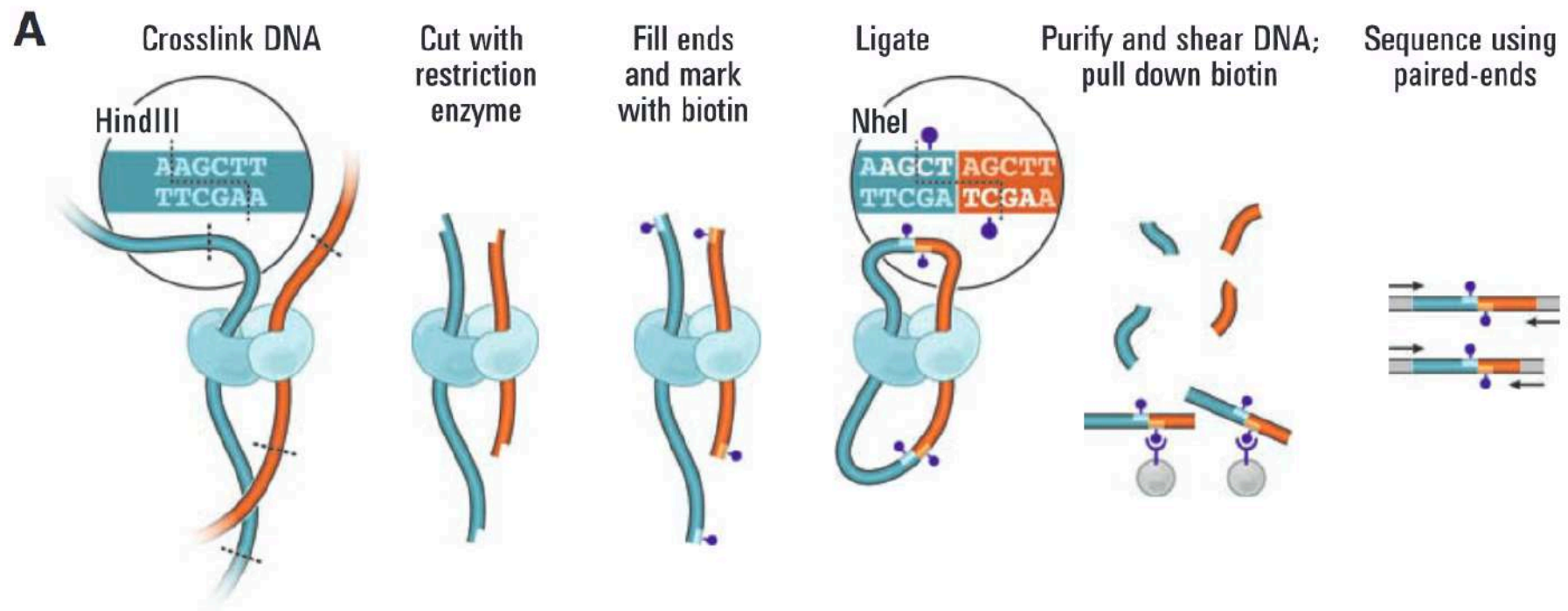


# Is it time to revisit bacterial taxonomy?

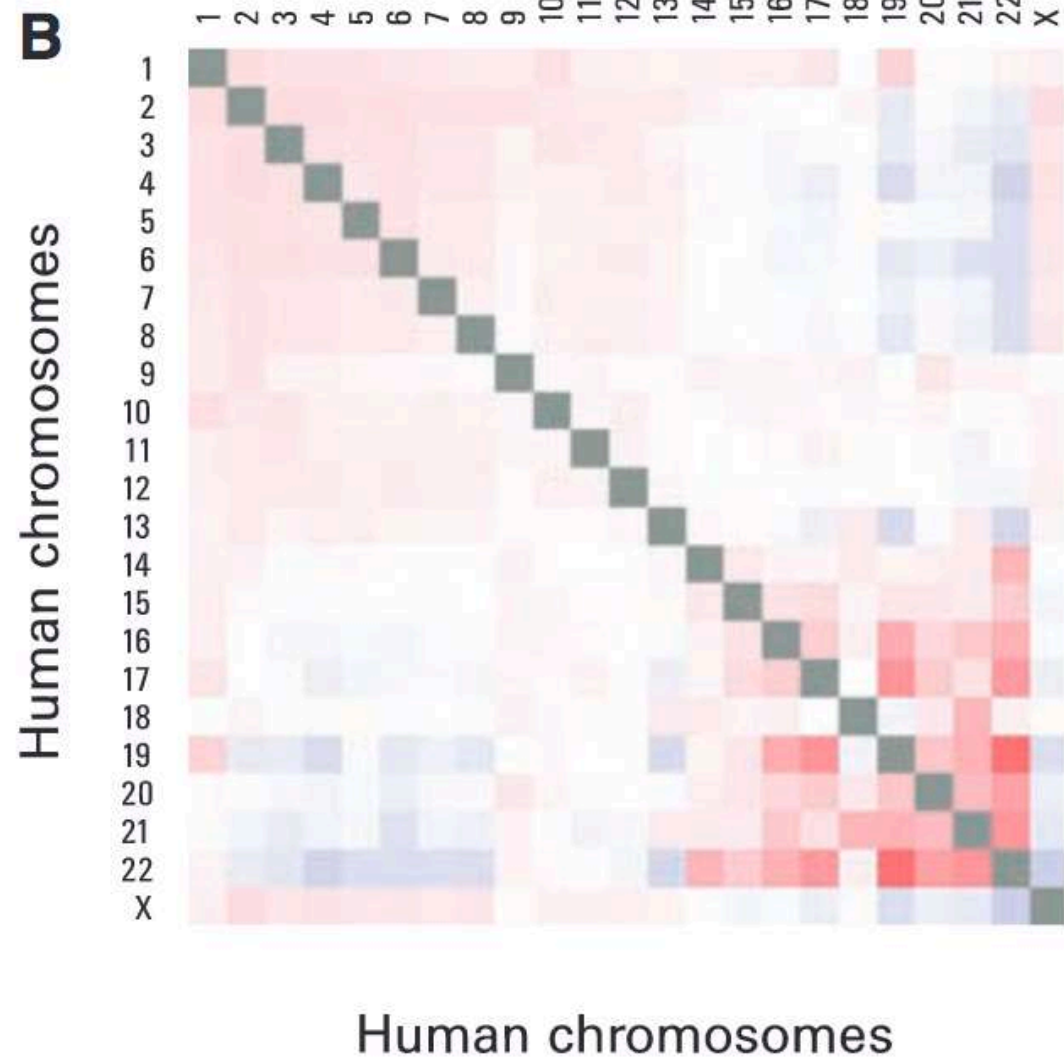
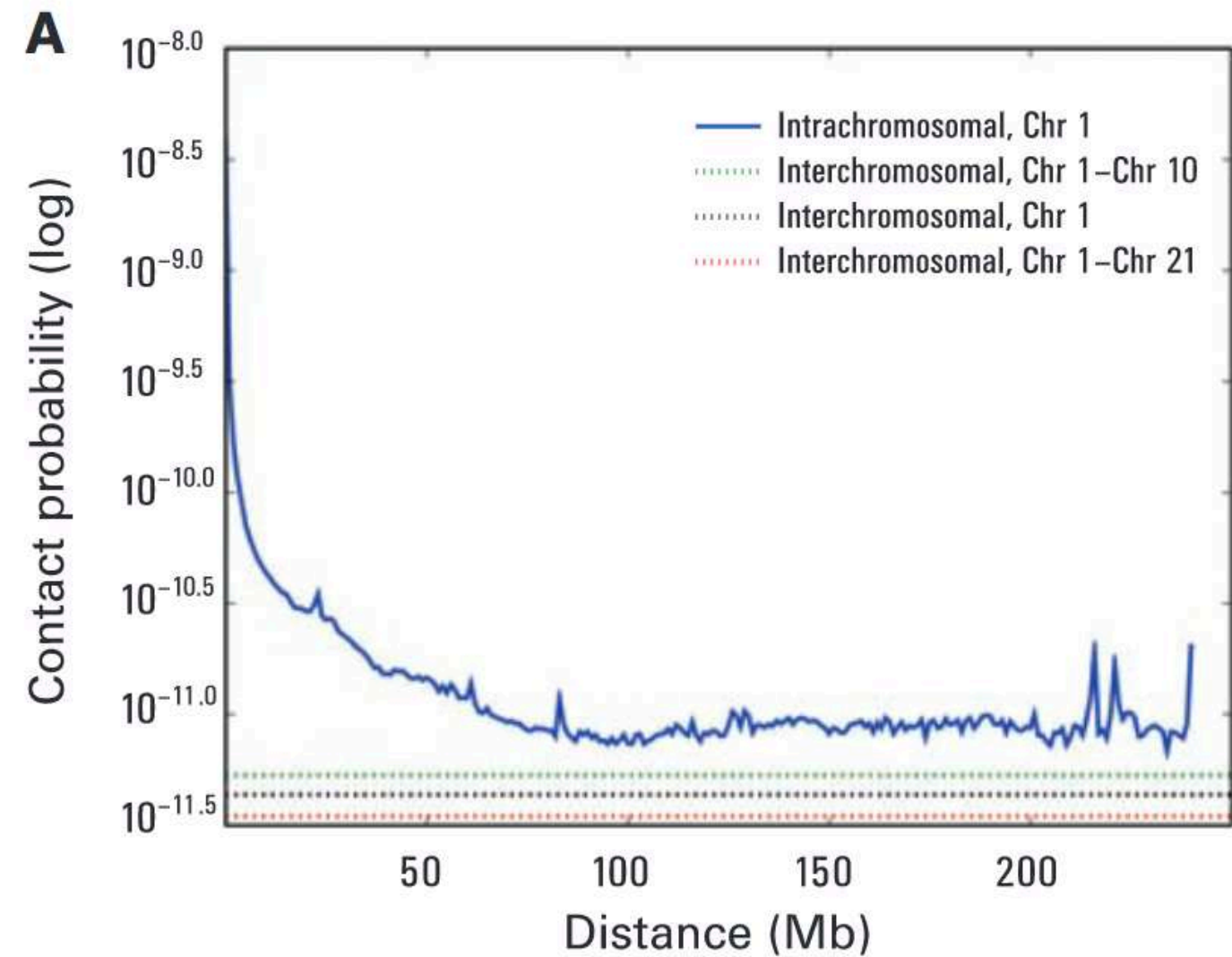


~~New~~ techniques that will change metagenomics

[1] Chromosome conformation capture and [2] long reads

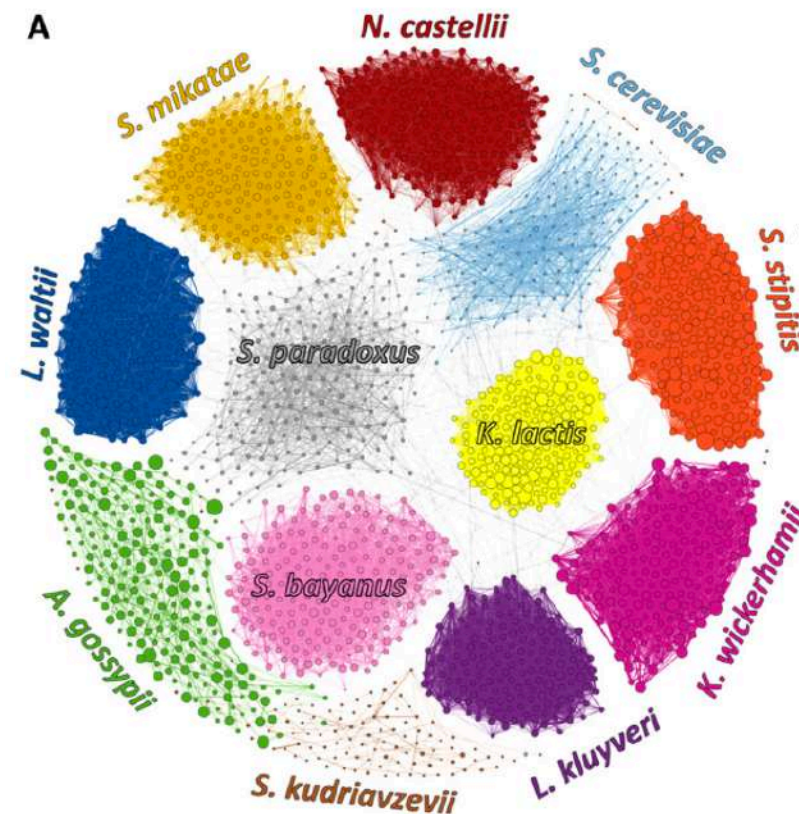
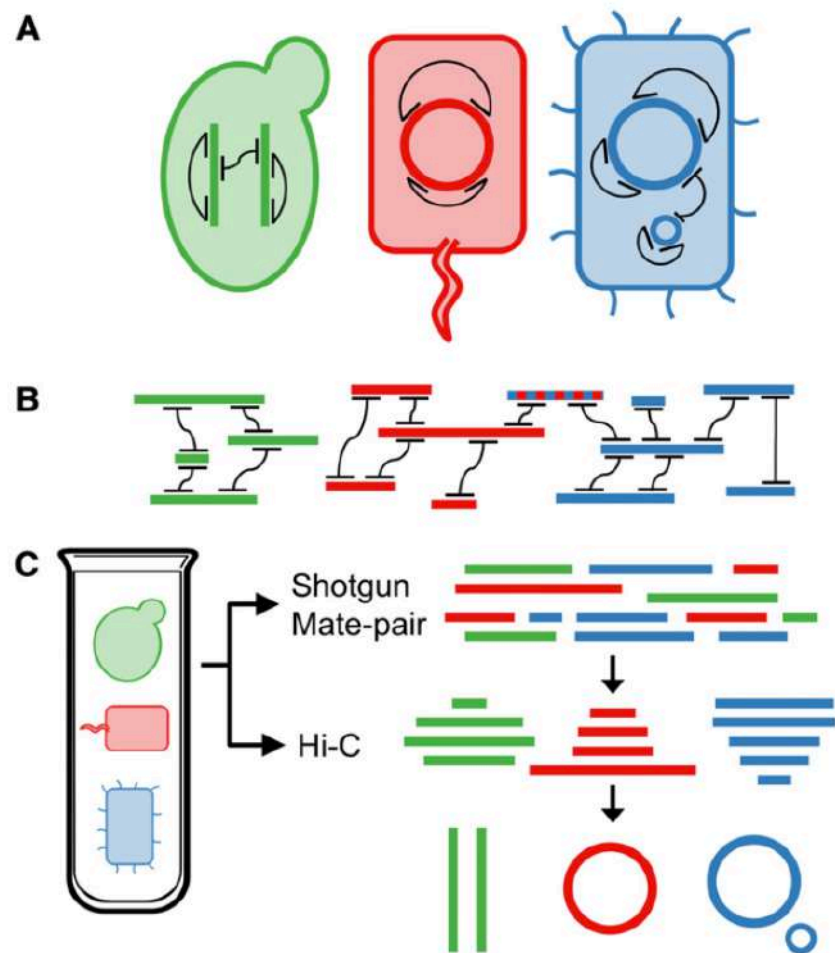






# Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps

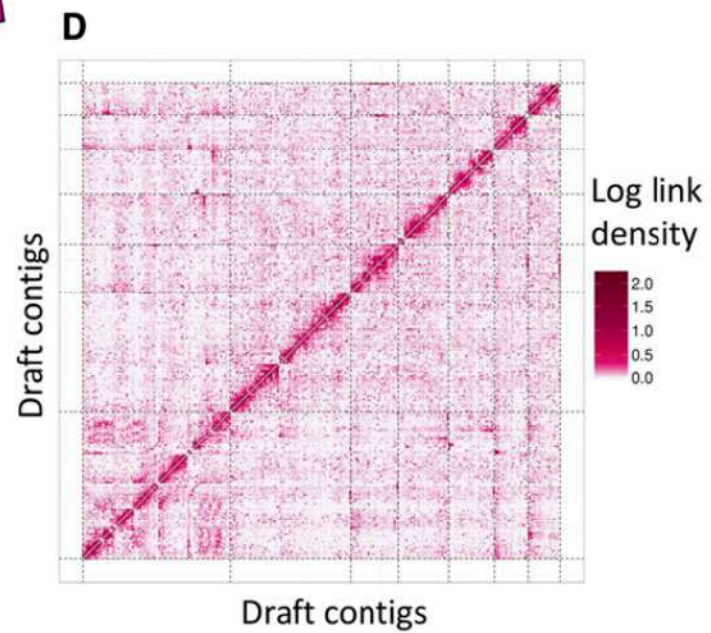
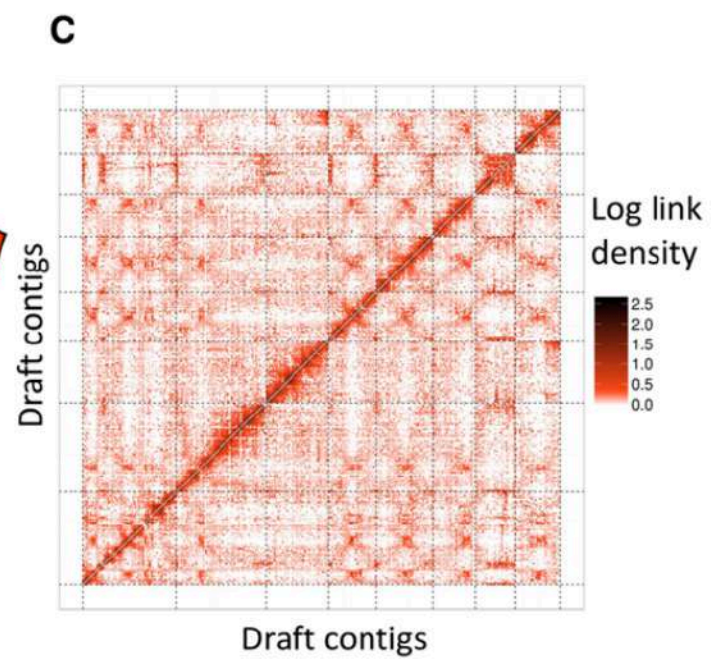
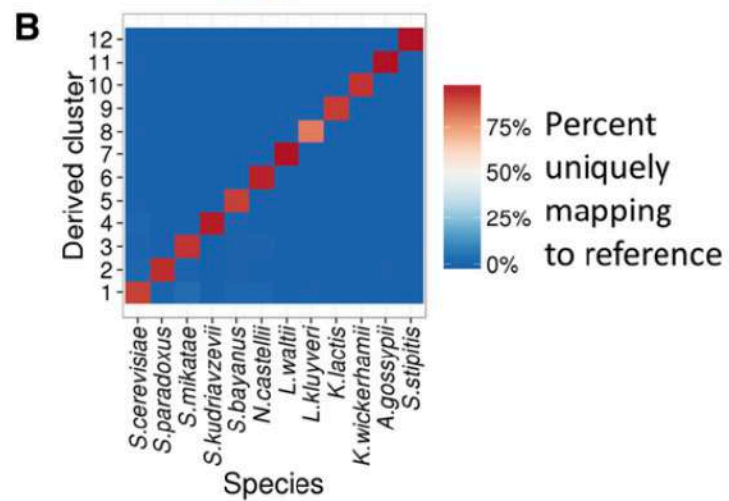
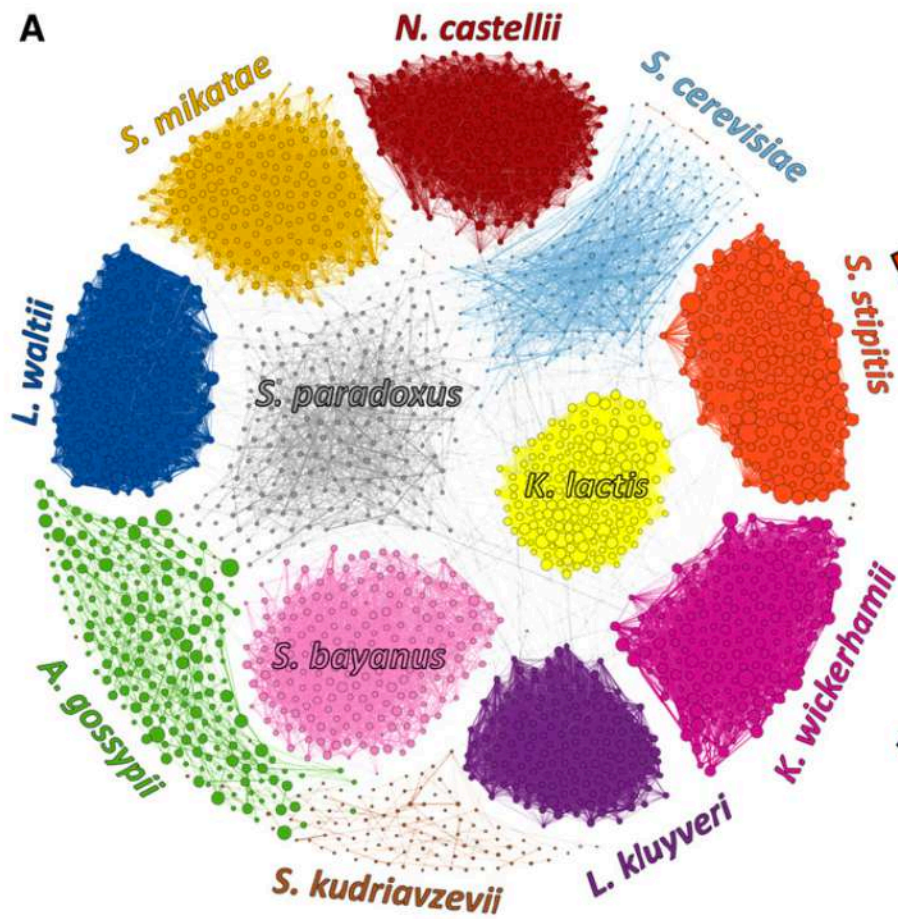
Joshua N. Burton,<sup>1</sup> Ivan Liachko,<sup>1</sup> Maitreya J. Dunham,<sup>2</sup> and Jay Shendure<sup>2</sup>  
 Department of Genome Sciences, University of Washington, Seattle, Washington 98195-5065



■ Table 2 Sequencing libraries used in MetaPhase analyses

Sample	Library Type	Read Length, bp	Read Pairs, millions
M-Y	Shotgun	101	85.7
	Mate-pair	100	9.2
	Hi-C	100	81.0





# Applications

# Exploration and categorisation (early 2010s)



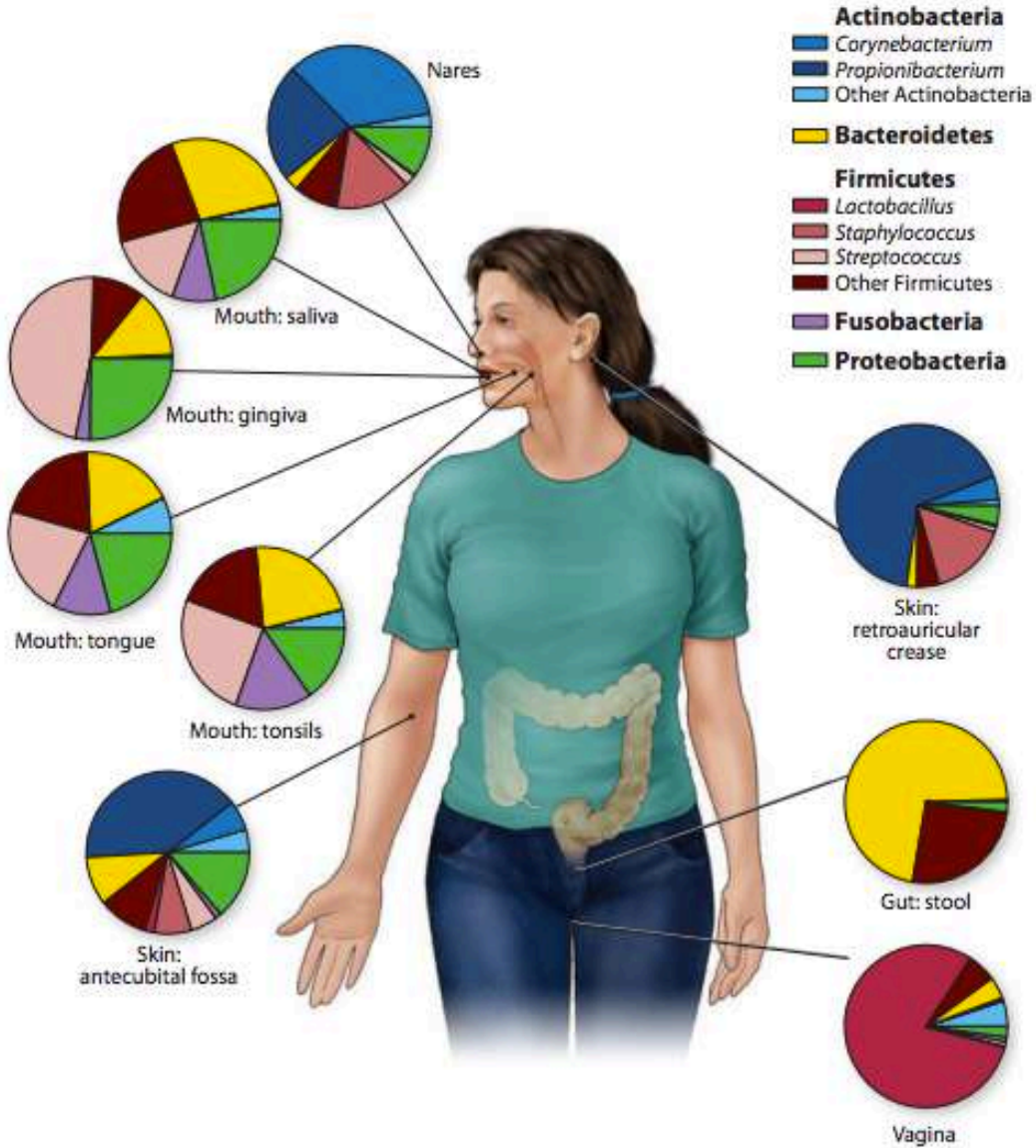
Rusch *et al.*, 2007 Plos Biology



Qin *et al.*, 2010 Nature

- 6.3 Gbp of sequence (2x Human genomes, 2000 x Bacterial genomes)
- Most sequences were novel compared to the databases

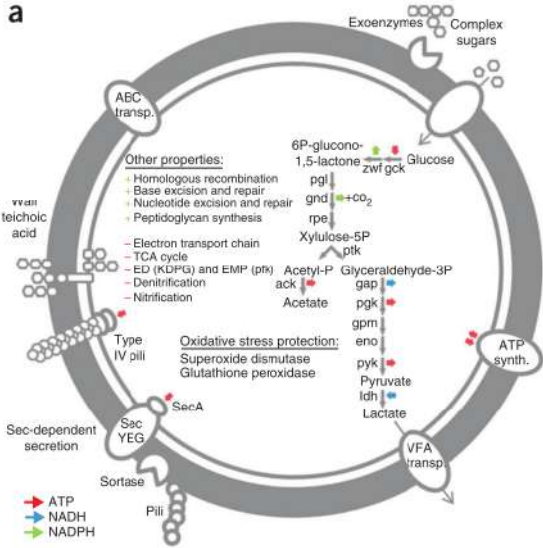
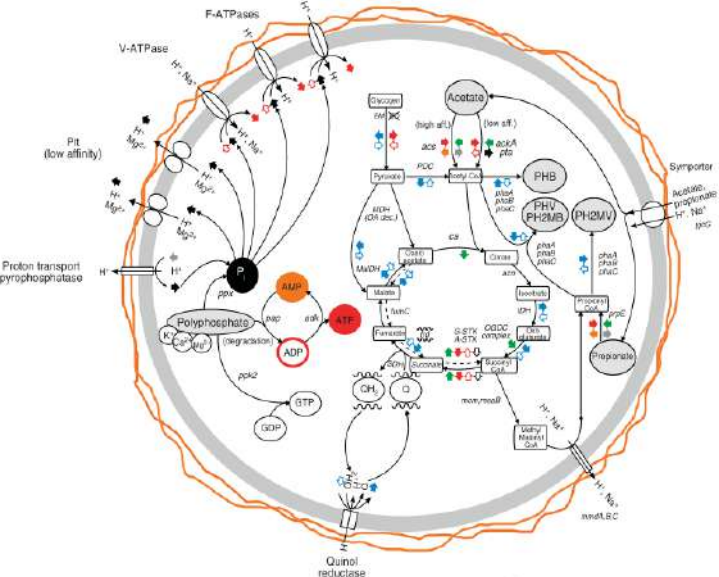
- 127 Human gut metagenomes
- 600 Gbp sequence (200 x Human genomes)
- 3.3 million genes identified
- Minimal gut metagenome defined



Grice and Segre (2012)



# Extracting genomes

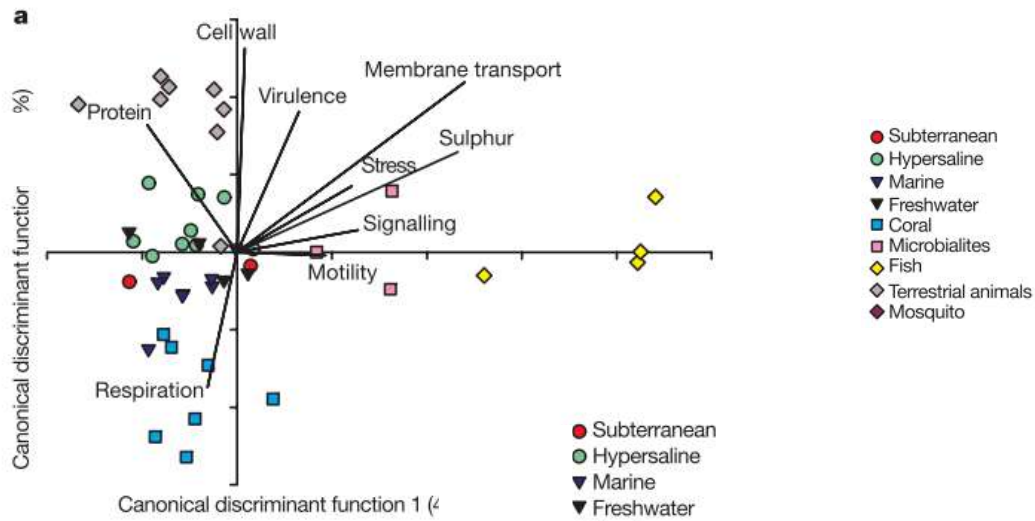


Garcia Martin *et al.*, 2006 **Nat. Biotechnol.**    Albertsen *et al.*, 2013 **Nat. Biotechnol.**

- Genome extraction from low complexity metagenome
- *Candidatus Accumulibacter phosphatis*
- The first genome of a polyphosphate accumulating organism (PAO) with a major role in enhanced biological phosphorus removal

- Genome extraction of low abundant species (< 0.1%) from metagenomes
- First complete TM7 genome
- Access to genomes of the "uncultured majority"

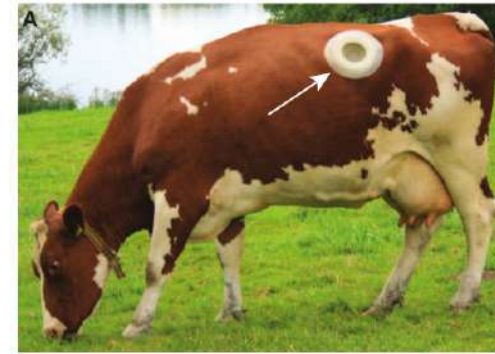
# Comparative



## Dinsdale *et al.*, 2008 **Nature**

- A characteristic microbial fingerprint for each of the nine different ecosystem types

# Specific functions



## Hess *et al.*, 2011 **Science**

- Identified 27,755 putative carbohydrate-active genes from a cow rumen metagenome
- Expressed 90 candidates of which 57% had enzymatic activity against cellulosic substrates



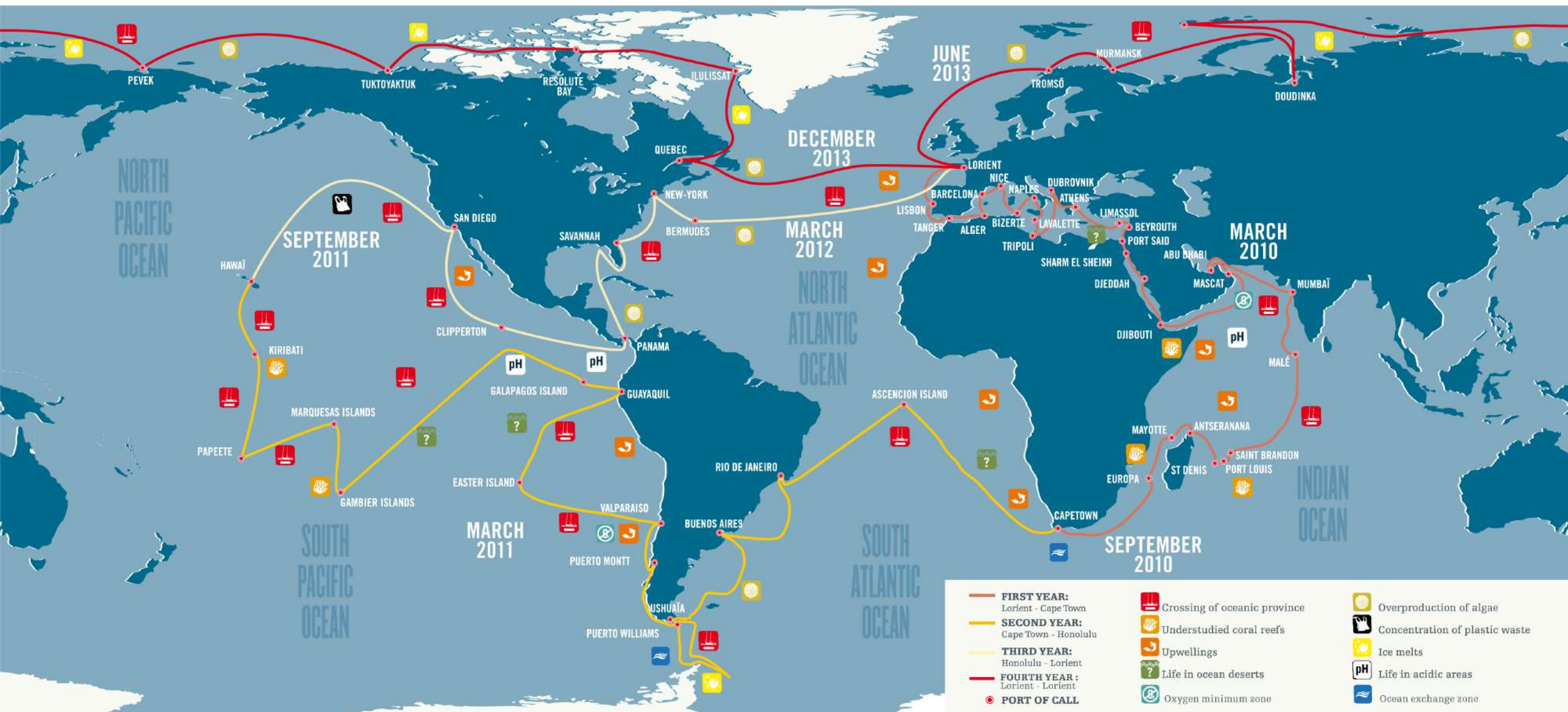
# TARA STUDYING TAIWAN'S BIODIVERSITY



27 March 2018

**For 4 days Tara successfully continued scientific research, then weighed anchor and left behind Orchid Island and Green Island, off the eastern coast of Taiwan.**

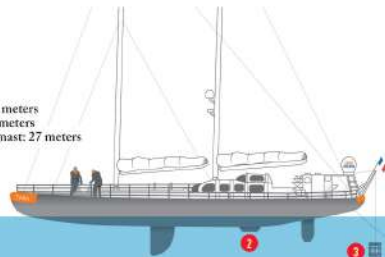
Tara Pacific's 3 target species were found and collected on site. According to Emilie Boissin, scientific coordinator of this mission (CRIOBE), *"these sites are interesting because they represent the northern boundary of the distribution area for these tropical species"*.



<https://oceans.taraexpeditions.org/wp-content/uploads/2014/05/TARAOCEANS-CARTE.jpg>



**TARA**  
length: 36 meters  
width: 10 meters  
height of mast: 27 meters



# THE TARA OCEANS EXPEDITION 2009-2013

REVIVING THE TRADITION OF THE GREAT EXPEDITIONS OF THE 19<sup>TH</sup> CENTURY, TARA SAILED THE WORLD'S OCEANS FOR THREE AND A HALF YEARS.

FOR THE FIRST TIME, MARINE PLANKTON IN ITS ENTIRETY WAS COLLECTED AND STUDIED - FROM VIRUSES AND BACTERIA TO FISH LARVAE AND JELLYFISH.

## WHY THIS EXPEDITION?

**OXYGEN**  
↑  
**CARBON**  
↓

THE OCEANS regulate the climate and atmosphere of our planet. Plankton produce half of the oxygen generated globally each year by photosynthesis, and absorb atmospheric CO<sub>2</sub>. Affected by pollution, over-fishing, and rising temperatures, will plankton continue to efficiently absorb carbon and regulate the climate?

**PLANKTON** designates all the organisms drifting with the currents. These microscopic organisms are the foundation of the marine food chain, ensuring the survival of fish, marine mammals, and billions of humans beings. They react quickly to climate changes and to ocean acidification. We must learn more about this complex, dynamic ecosystem and its role in global equilibrium.

**CORAL REEFS** are privileged places for aquatic biodiversity, but they are suffering from climate change, marine pollution, and over-fishing. Tara was the ideal platform for exploring 5 rarely-studied coral sites: Djibouti, Saint-Brandon, Mayotte, and the islands of Gambier and Kiribati.

## A CONCENTRATION OF HIGH TECH

A unique space for microscopic imagery set up aboard Tara - the dry lab - where researchers characterize the organisms collected, their functional diversity and their complexity.

THE UNDERWATER VISION PROFILER observes plankton during sampling.

THE FLOWCAM is used to count and identify organisms as they pass through a laser beam at high speed.



## THREE METHODS OF COLLECTION AND OBSERVATION. MORE THAN 35,000 SAMPLES

**1 NETS**  
Tara deployed 7 types of nets (mesh sizes from 5 to 690 microns), immersed between the surface and 1,000 meters deep. The specialized Manta net is used for collecting plastic on the surface.

**2 PERISTALTIC PUMP**  
Water is pumped from a depth of 10 to 120 meters, then passes through a series of strainers and filters to separate organisms by size.

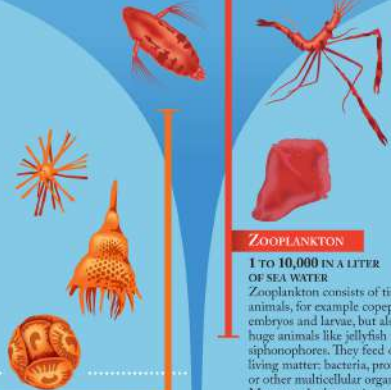
**3 THE "ROSETTE" CTD AND THE UVP**  
This apparatus contains 10 Niskin bottles to collect water from different depths, as well as instruments to characterize many parameters including pressure, temperature, conductivity, nitrogen, oxygen, fluorescence, etc. The bottles are programmed to collect water at different depths. The UVP (Underwater Vision Profiler) deployed down to a depth of 2,000 meters allowed us to record about 20 physico-chemical parameters, and image particles and organisms.

## VOYAGE OF THE SAMPLES

**PORTS-OF-CALL**  
At stopovers every 6 to 8 weeks, the samples - conserved with liquid nitrogen, alcohol and fixatives - were sent to partner laboratories.

FROM VIRUSES  
TO SAMPLE ZOOPLANKTON:

TO FISH LARVAE  
WE NEED 1 000 000 LITERS OF SEA WATER.



**ZOOPLANKTON**  
1 TO 10,000 IN A LITER OF SEA WATER  
Zooplankton consists of tiny animals, for example copepods, embryos and larvae, but also huge animals like jellyfish and siphonophores. They feed on living matter: bacteria, protists, or other multicellular organisms. Most zooplankton migrates to the surface, or to great depths to feed and protect themselves from predators during the night.

**PROTISTS, INCLUDING PHYTOPLANKTON**  
1 TO 100 MILLION IN A LITER OF SEA WATER  
The principal ocean biodiversity consists of multitudes of species of unicellular organisms with a nucleus: the protists. Certain of them (diatoms, dinoflagellates, etc.) are photosynthetic. Along with cyanobacteria, they constitute phytoplankton, and are the base of the food chain. Phytoplankton produces half of the oxygen on the planet and absorbs half of atmospheric carbon, thus acting as a major regulator of climate.

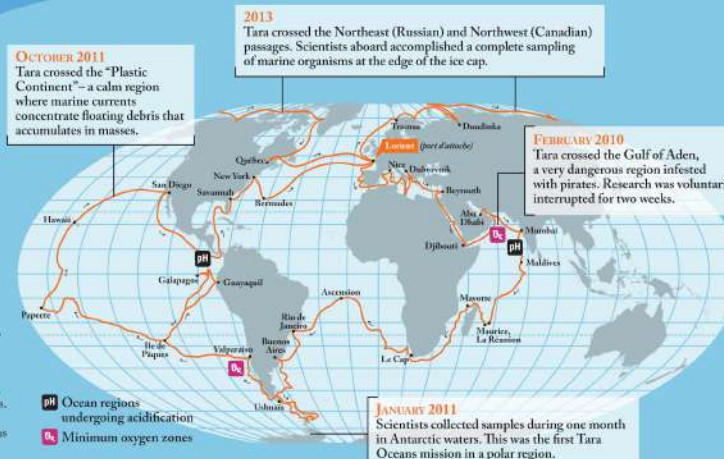
**BACTERIA**  
1 TO 10 BILLION IN A LITER OF SEA WATER  
Bacteria are prokaryotes: cells without nuclei. Certain species - the cyanobacteria - can perform photosynthesis. They are a food for protists and certain zooplankton. Bacteria are responsible for a wide array of metabolic functions in the ocean.

**VIRUSES**  
10 TO 100 BILLION IN A LITER OF WATER  
The marine virosphere is immense, and includes the phages (viruses of bacteria) and giant viruses (gigaviruses). Viruses play an essential role in recirculating living matter.

*Size proportions of the micro-organisms are not respected in these drawings.*

**WORLD COURIER**  
This international specialist in shipping sensitive products expedited the precious samples collected aboard Tara to Heidelberg (Germany), then redistributed them to partner laboratories around the world.

september 2009 - december 2013  
60 STOPOVERS, 35 COUNTRIES  
140,000 KILOMETERS AROUND THE WORLD



## PARTNER LABORATORIES

23 LABS AND SCIENTIFIC INSTITUTIONS

- 8 in France
- 5 in the United States
- 2 in Germany
- 2 in Italy
- 1 in Belgium
- 1 in Ireland
- 1 in Spain
- 1 in Canada
- 1 in Saudi Arabia
- 1 in Russia

## THE "TARANAUTES"

TAKING TURNS ON BOARD:

- 90 crew members, artists, and journalists
- 160 researchers
- 40 nationalities

## SCIENTIFIC RESULTS

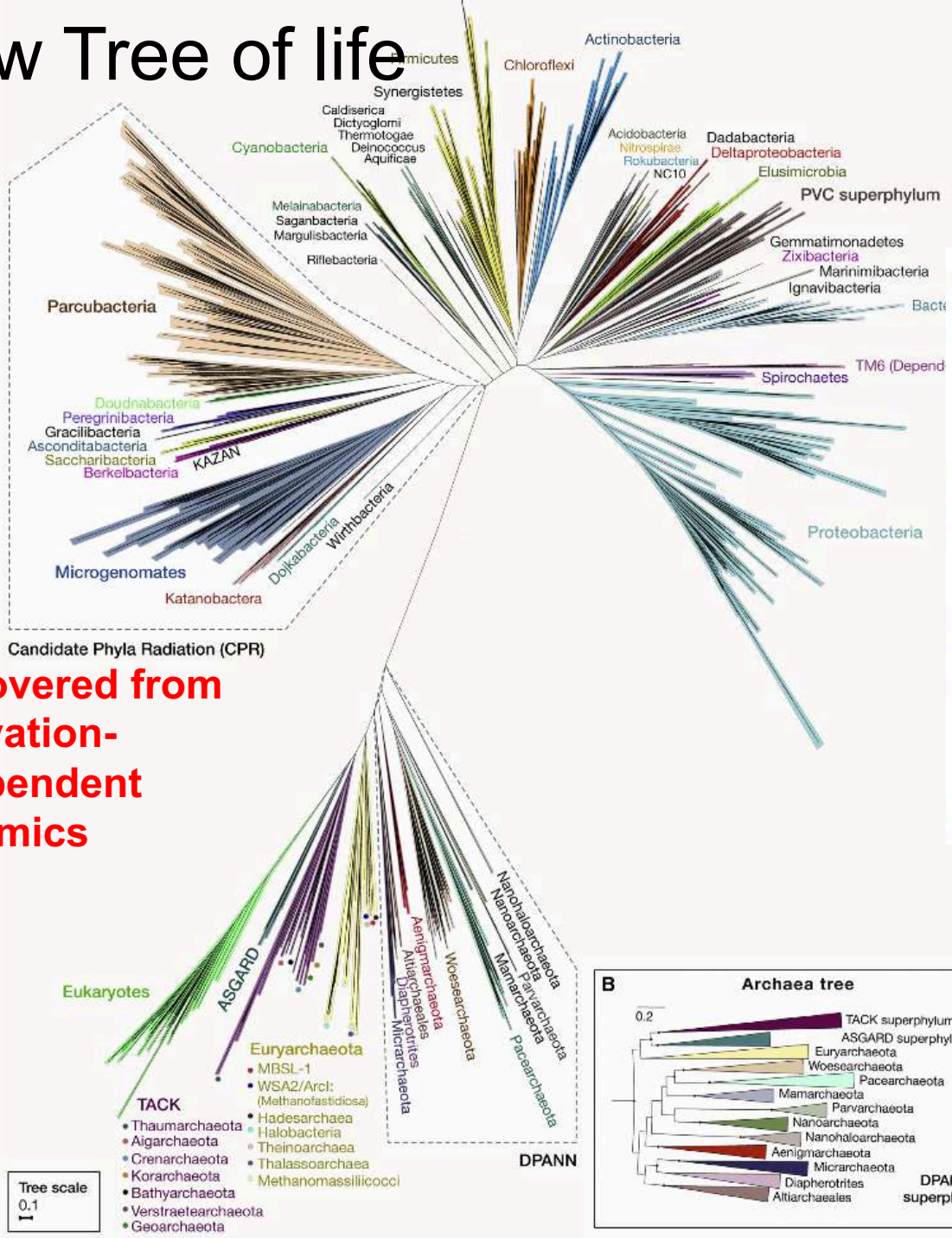
Based on the data from Tara Oceans, many scientific articles detailing the planktonic ecosystem and its dynamics have been published, or are on the way to being published in international journals. Ongoing analysis of this data, thanks to the Oceanomics project\* will help establish a reference for ocean ecosystems, and set up a method for predicting and following the evolution of these ecosystems in relation to climate change.

\*Oceanomics (an International Oceanographic Commission project) aims to promote national and sustainable use of marine plankton, one of the planet's most important ecosystems in terms of biodiversity, bio-resources, and global ecological change.





# New Tree of life



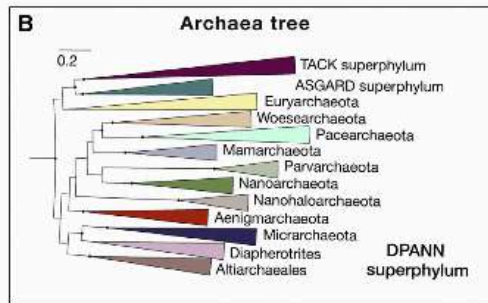
## Leading Edge Perspective

# Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life

Cindy J. Castelle<sup>1,2,3</sup> and Jillian F. Banfield<sup>1,2,3,4,5,6,\*</sup>

- <sup>1</sup>Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA
- <sup>2</sup>Innovative Genomics Institute, Berkeley, CA, USA
- <sup>3</sup>Chan Zuckerberg Biohub, San Francisco, CA, USA
- <sup>4</sup>University of Melbourne, Melbourne, VIC, Australia
- <sup>5</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA
- <sup>6</sup>Department of Environmental Science, Policy and Management, University of California, Berkeley, Berkeley, CA, USA
- \*Correspondence: [jbanfield@berkeley.edu](mailto:jbanfield@berkeley.edu)
- <https://doi.org/10.1016/j.cell.2018.02.016>

Discovered from cultivation-independent genomics



Tree scale  
0.1



# Case studies – human microbiome

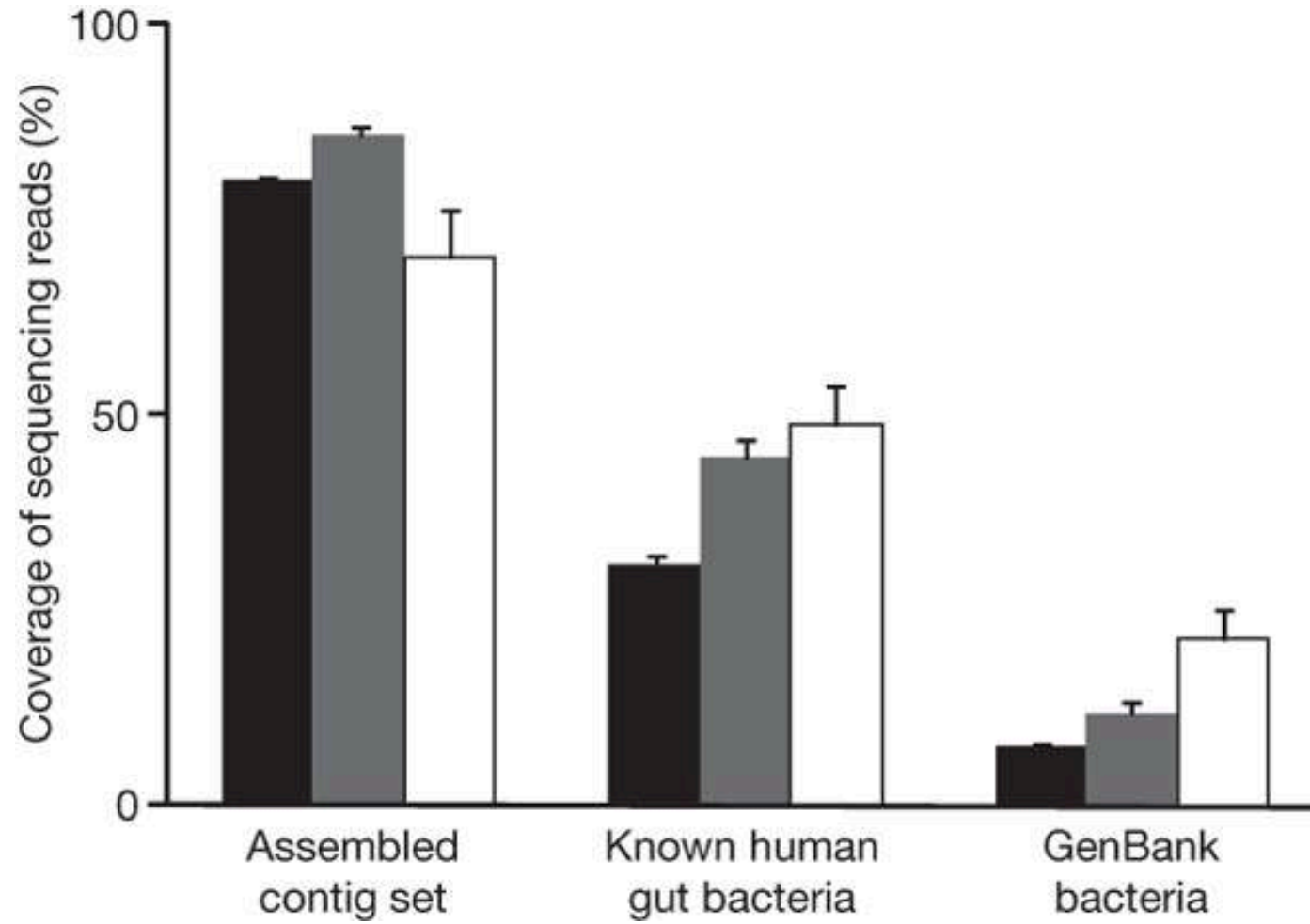
## ARTICLES

---

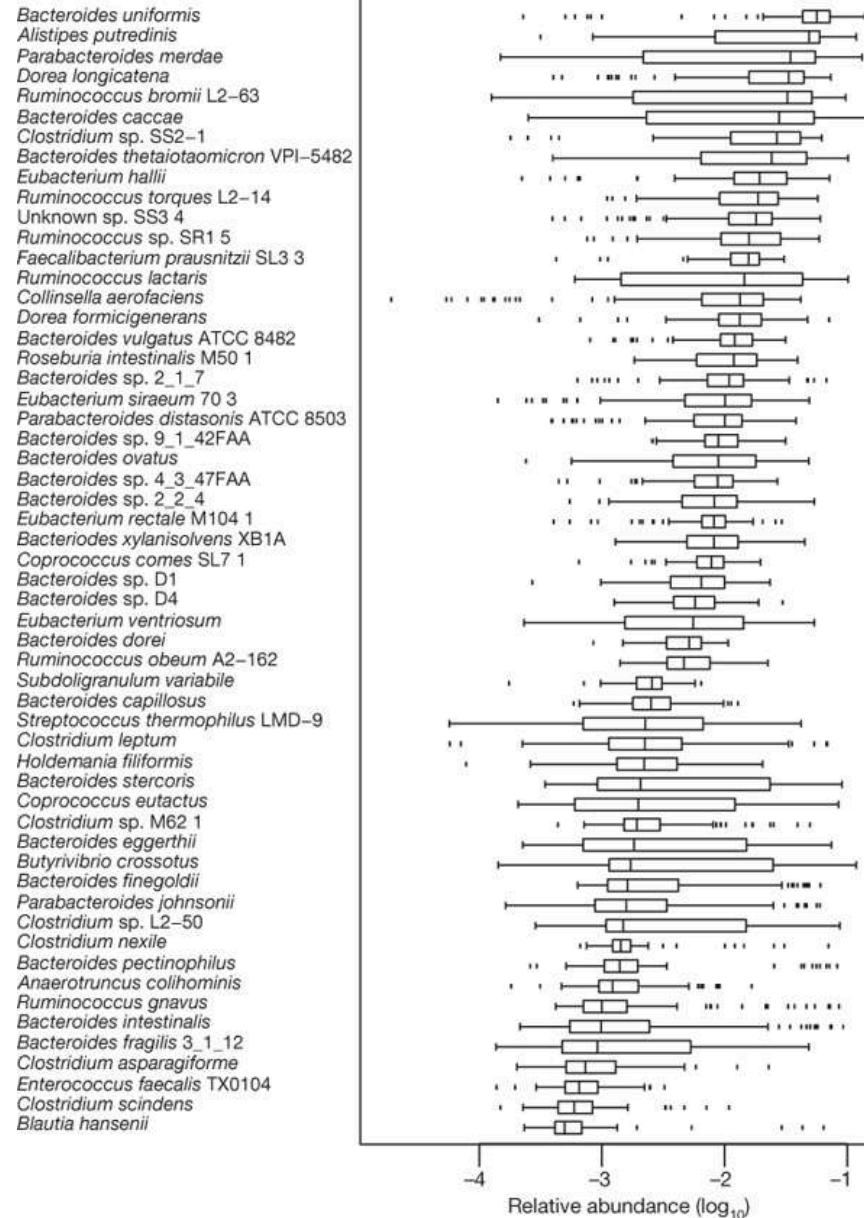
# A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin<sup>1\*</sup>, Ruiqiang Li<sup>1\*</sup>, Jeroen Raes<sup>2,3</sup>, Manimozhiyan Arumugam<sup>2</sup>, Kristoffer Solvsten Burgdorf<sup>4</sup>, Chaysavanh Manichanh<sup>5</sup>, Trine Nielsen<sup>4</sup>, Nicolas Pons<sup>6</sup>, Florence Levenez<sup>6</sup>, Takuji Yamada<sup>2</sup>, Daniel R. Mende<sup>2</sup>, Junhua Li<sup>1,7</sup>, Junming Xu<sup>1</sup>, Shaochuan Li<sup>1</sup>, Dongfang Li<sup>1,8</sup>, Jianjun Cao<sup>1</sup>, Bo Wang<sup>1</sup>, Huiqing Liang<sup>1</sup>, Huisong Zheng<sup>1</sup>, Yinlong Xie<sup>1,7</sup>, Julien Tap<sup>6</sup>, Patricia Lepage<sup>6</sup>, Marcelo Bertalan<sup>9</sup>, Jean-Michel Batto<sup>6</sup>, Torben Hansen<sup>4</sup>, Denis Le Paslier<sup>10</sup>, Allan Linneberg<sup>11</sup>, H. Bjørn Nielsen<sup>9</sup>, Eric Pelletier<sup>10</sup>, Pierre Renault<sup>6</sup>, Thomas Sicheritz-Ponten<sup>9</sup>, Keith Turner<sup>12</sup>, Hongmei Zhu<sup>1</sup>, Chang Yu<sup>1</sup>, Shengting Li<sup>1</sup>, Min Jian<sup>1</sup>, Yan Zhou<sup>1</sup>, Yingrui Li<sup>1</sup>, Xiuqing Zhang<sup>1</sup>, Songgang Li<sup>1</sup>, Nan Qin<sup>1</sup>, Huanming Yang<sup>1</sup>, Jian Wang<sup>1</sup>, Søren Brunak<sup>9</sup>, Joel Doré<sup>6</sup>, Francisco Guarner<sup>5</sup>, Karsten Kristiansen<sup>13</sup>, Oluf Pedersen<sup>4,14</sup>, Julian Parkhill<sup>12</sup>, Jean Weissenbach<sup>10</sup>, MetaHIT Consortium†, Peer Bork<sup>2</sup>, S. Dusko Ehrlich<sup>6</sup> & Jun Wang<sup>1,13</sup>

# Human gut microbiome

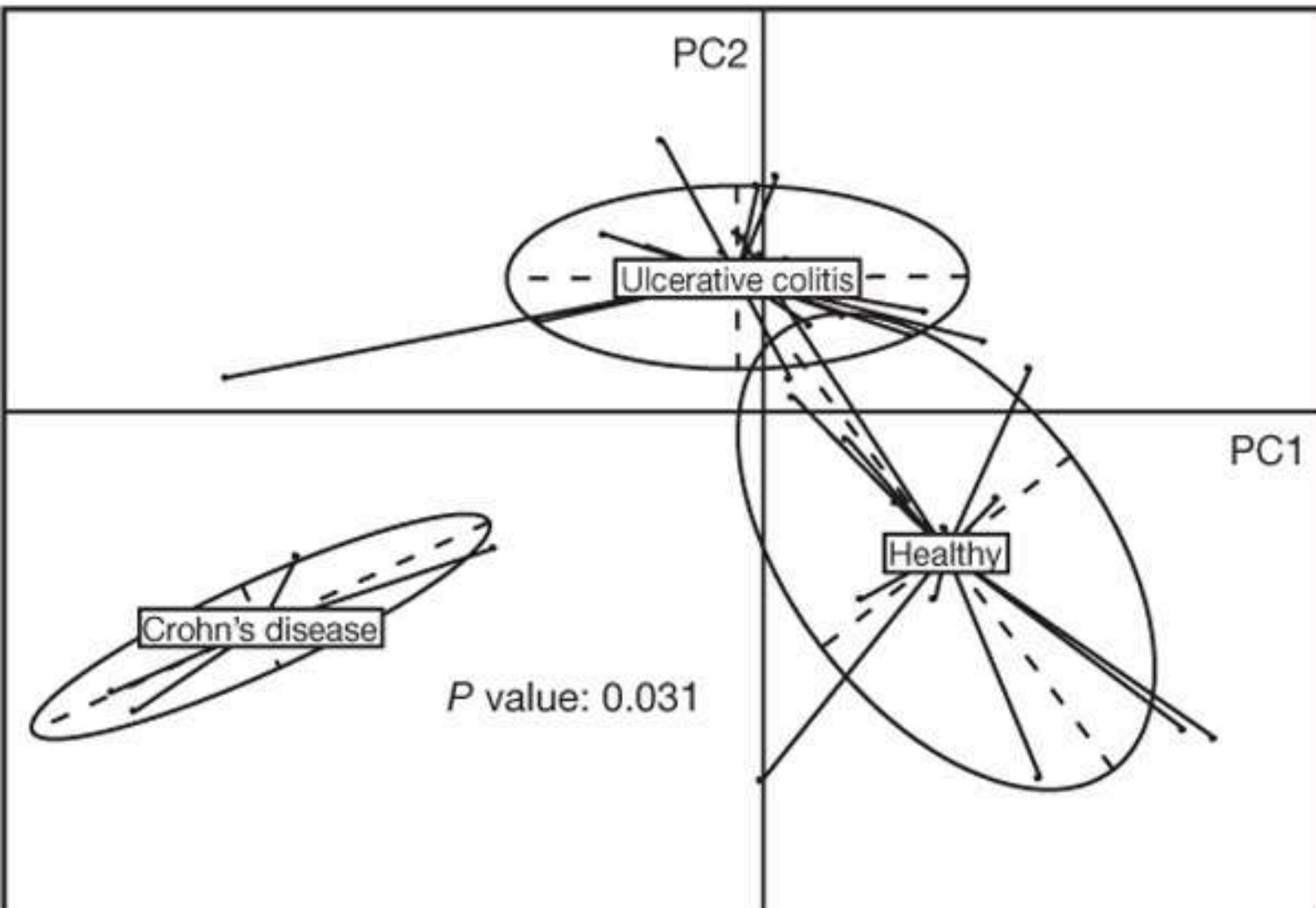


# Human gut microbiome





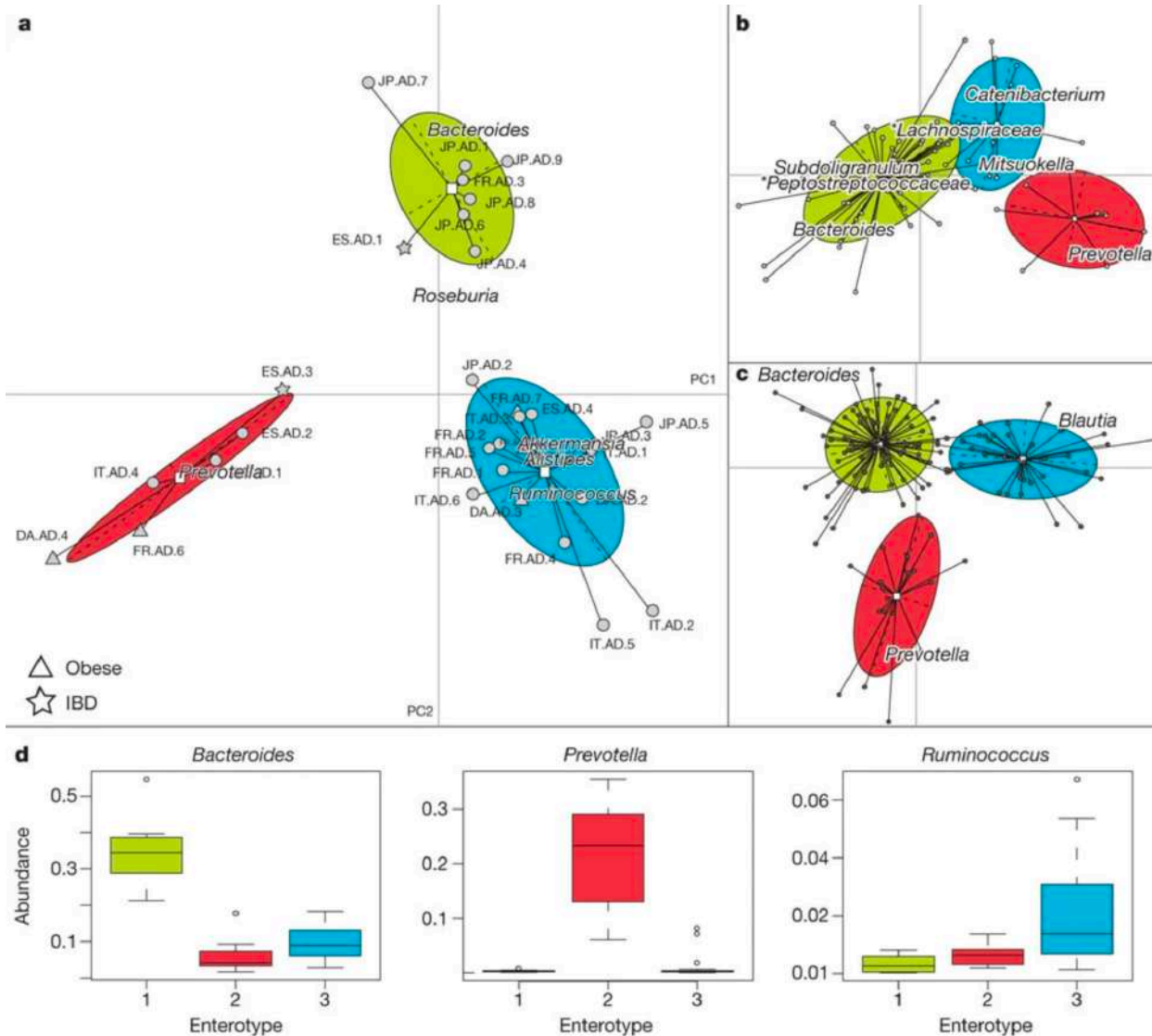
# Human gut microbiome



We can check which OTUs constitute the clustering (and separation) patterns

- > Biology
- > Biomarkers

# Human gut microbiome - enterotypes



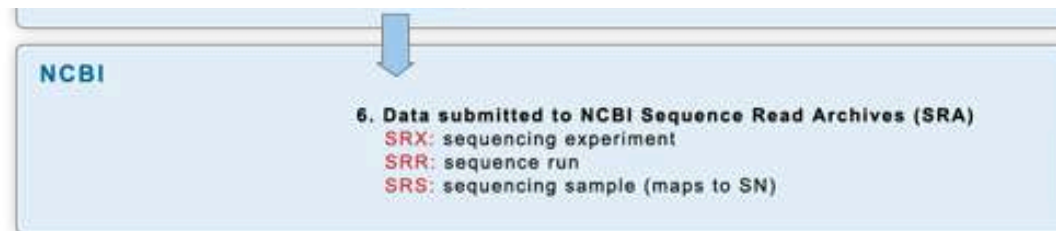
By combining 22 newly sequenced faecal metagenomes of individuals from four countries with previously published data sets, here we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific.

The enterotypes are mostly driven by species composition, but abundant molecular functions are not necessarily provided by abundant species, highlighting the importance of a functional analysis to understand microbial communities.



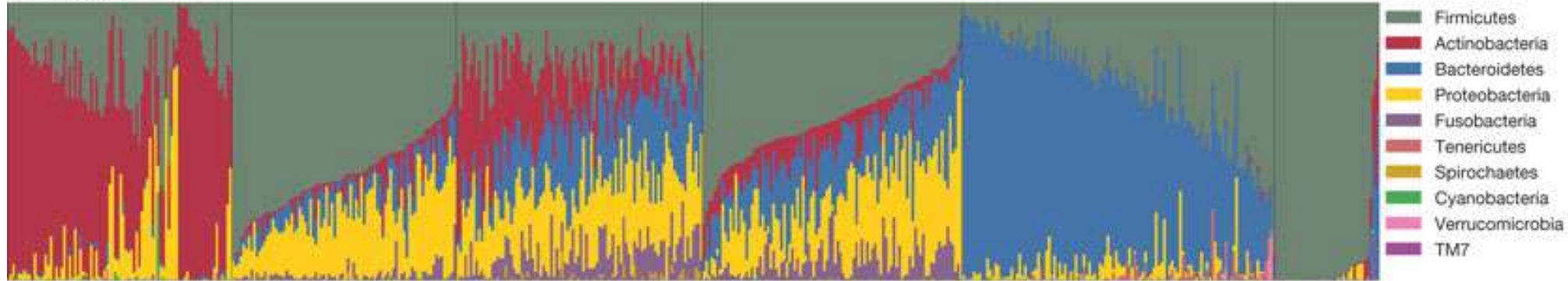
**Table 1 | HMP donor samples examined by 16S and WGS**

Body region	Body site	Total samples	Total 16S samples	V13 samples	V13 read depth (M)*	V35 samples	V35 read depth (M)*	Samples V13 and V35	Total WGS samples	Total read depth (G)†	Filtered reads (%)‡	Human reads (%)§	Remaining read depth (G)†	Samples 16S and WGS
Gut	Stool	352	337	193	1.4	328	2.4	184	136	1,720.7	15	1	1,450.6	124
Oral cavity	Buccal mucosa	346	330	184	1.3	314	1.7	168	107	1,438.0	9	82	136.7	91
	Hard palate	325	325	179	1.2	310	1.7	164	1	10.9	20	25	5.9	1
	Keratinized gingiva	335	329	183	1.3	319	1.7	173	6	72.3	5	47	34.4	0
	Palatine tonsils	337	332	189	1.2	315	1.9	172	6	74.8	2	80	13.5	1
	Saliva	315	310	166	0.9	292	1.5	148	5	55.7	1	91	4.2	0
	Subgingival plaque	334	328	186	1.2	314	1.8	172	7	92.1	5	79	15.3	1
	Supragingival plaque	345	331	192	1.3	316	1.9	177	115	1,500.7	15	40	674.8	101
	Throat	331	325	176	1.0	312	1.7	163	7	78.8	4	79	13.6	1
	Tongue dorsum	348	332	193	1.3	320	2.0	181	122	1,620.1	15	19	1,084.3	106
Airway	Anterior nares	316	302	169	1.0	283	1.2	150	84	1,129.9	3	96	14.3	70
Skin	Left antecubital fossa	269	269	158	0.7	221	0.5	110	0	NA	NA	NA	0	NA
	Left retroauricular crease	313	312	188	1.6	295	1.5	171	9	126.3	9	73	22.1	8
	Right antecubital fossa	274	274	158	0.7	229	0.5	113	0	NA	NA	NA	0	NA
	Right retroauricular crease	319	316	190	1.4	304	1.6	178	15	181.9	18	59	42.4	12
Vagina	Mid-vagina	145	143	91	0.6	140	1.0	88	2	22.6	0	99	0.2	0
	Posterior fornix	152	142	89	0.6	136	1.0	83	53	702.1	6	90	25.2	43
	Vaginal introitus	142	140	87	0.6	131	0.9	78	3	36.5	1	98	0.6	1
Total		5,298	5,177	2,971	19	4,879	26.3	2,673	681	8,863.3	11	49	3,538.1	560

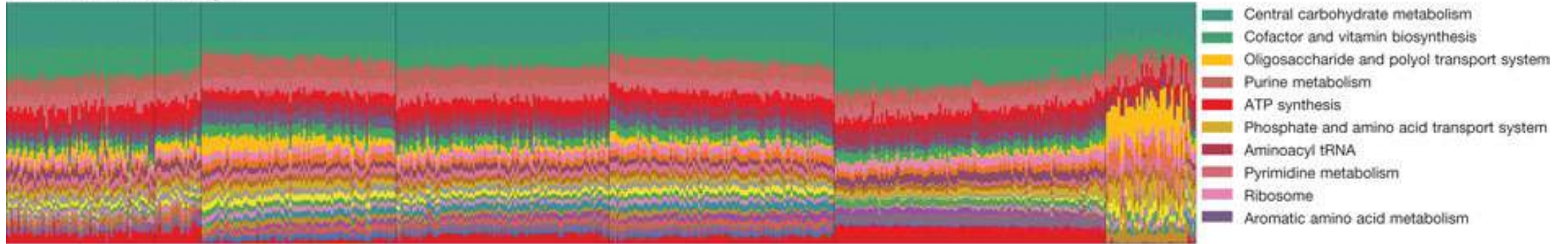


# Human microbiome

**a** Phyla



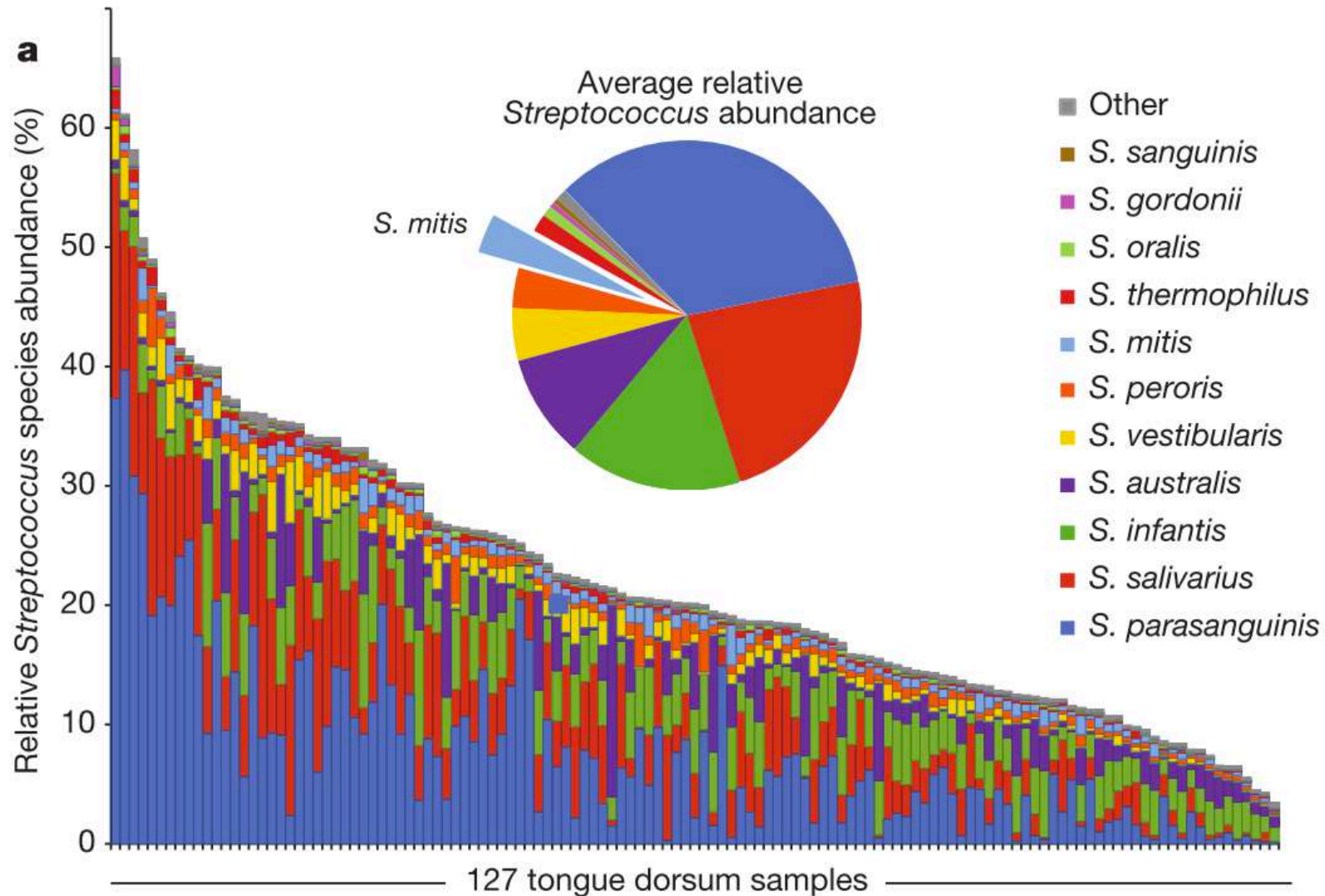
**b** Metabolic pathways



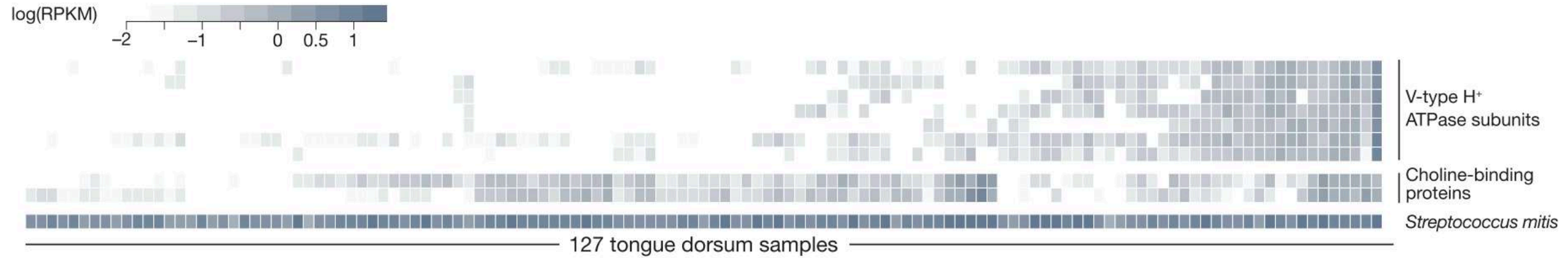
Anterior nares RC Buccal mucosa Supragingival plaque Tongue dorsum Stool Posterior fornix



# Inter-individual variation in the microbiome proved to be specific, functionally relevant and personalized



# Gene loss & Structural variants are common

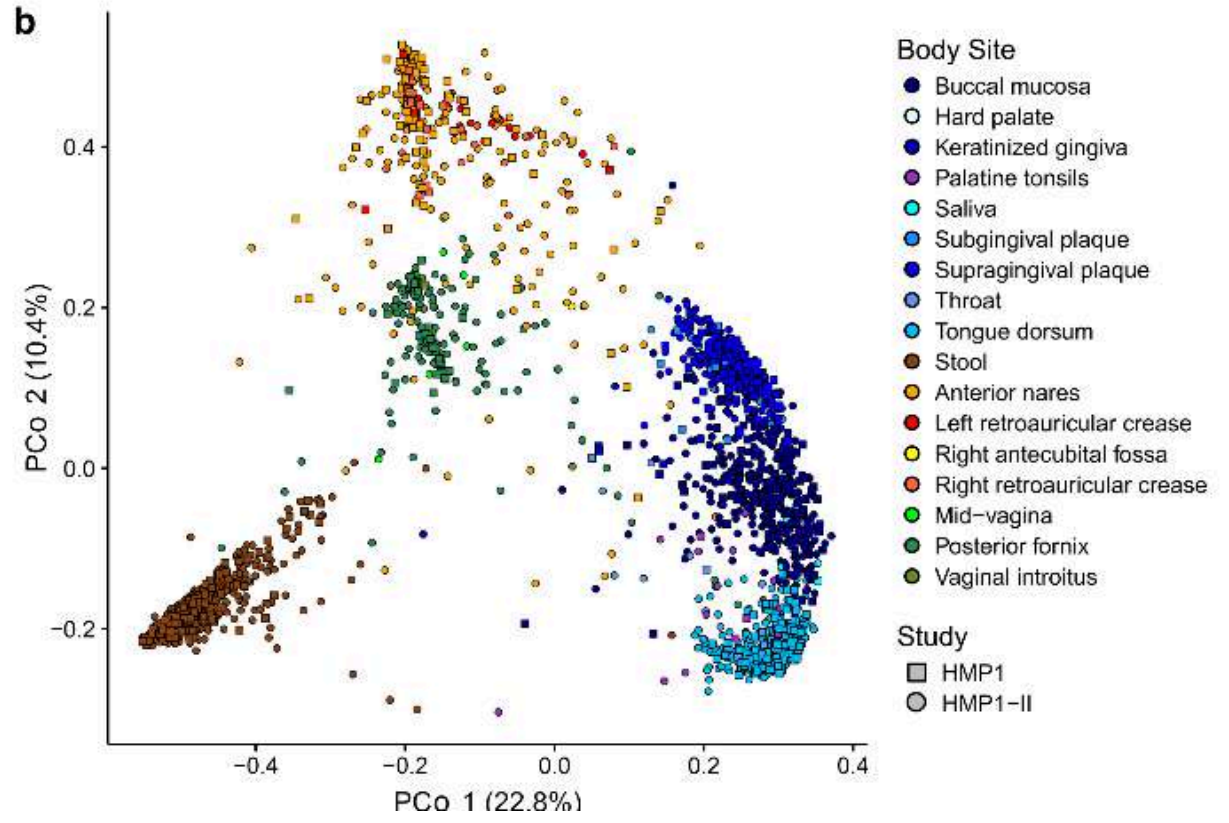
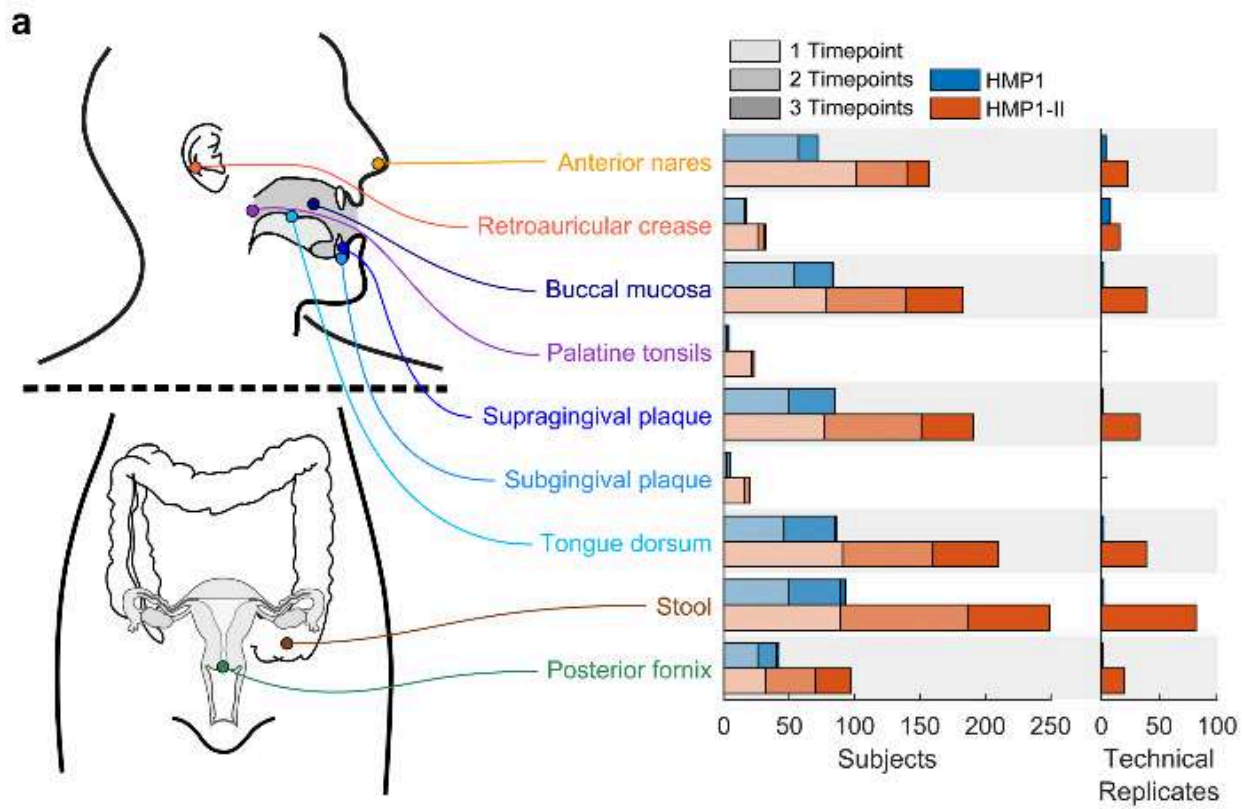


# Strains, functions and dynamics in the expanded Human Microbiome Project

Jason Lloyd-Price<sup>1,2\*</sup>, Anup Mahurkar<sup>3\*</sup>, Gholamali Rahnavard<sup>1,2</sup>, Jonathan Crabtree<sup>3</sup>, Joshua Orvis<sup>3</sup>, A. Brantley Hall<sup>2</sup>, Arthur Brady<sup>3</sup>, Heather H. Creasy<sup>3</sup>, Carrie McCracken<sup>3</sup>, Michelle G. Giglio<sup>3</sup>, Daniel McDonald<sup>4</sup>, Eric A. Franzosa<sup>1,2</sup>, Rob Knight<sup>4,5</sup>, Owen White<sup>3</sup> & Curtis Huttenhower<sup>1,2</sup>

**The characterization of baseline microbial and functional diversity in the human microbiome has enabled studies of microbiome-related disease, diversity, biogeography, and molecular function. The National Institutes of Health Human Microbiome Project has provided one of the broadest such characterizations so far. Here we introduce a second wave of data from the study, comprising 1,631 new metagenomes (2,355 total) targeting diverse body sites with multiple time points in 265 individuals. We applied updated profiling and assembly methods to provide new characterizations of microbiome personalization. Strain identification revealed subspecies clades specific to body sites; it also quantified species with phylogenetic diversity under-represented in isolate genomes. Body-wide functional profiling classified pathways into universal, human-enriched, and body site-enriched subsets. Finally, temporal analysis decomposed microbial variation into rapidly variable, moderately variable, and stable subsets. This study furthers our knowledge of baseline human microbial diversity and enables an understanding of personalized microbiome function and dynamics.**

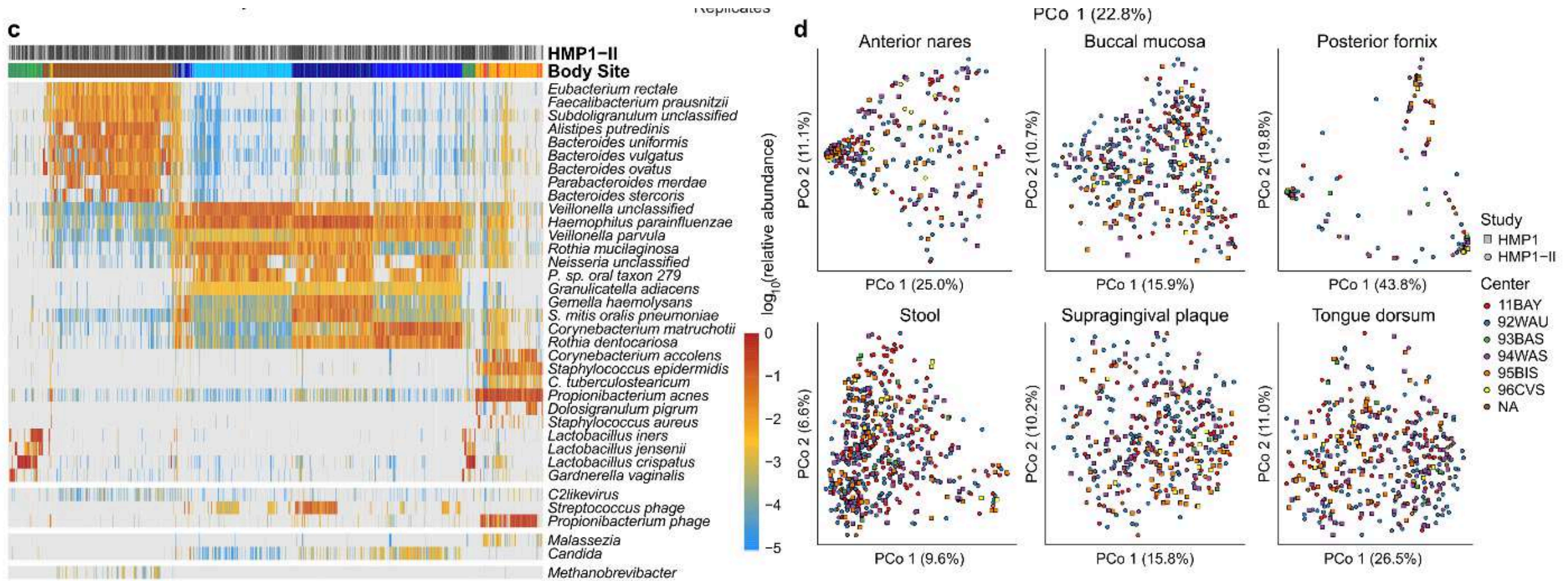




**Extended Data Figure 1 | Extended body-wide metagenomic taxonomic profiles in HMP1-II.** **a**, The combined HMP1-II datasets include a total of 2,355 metagenomes (724 previously published and 1,631 new, including 252 technical replicates). These span the project's six targeted body sites (anterior nares, buccal mucosa, supragingival plaque, tongue dorsum, stool, and posterior fornix) in addition to at least 20 samples each from 3 additional sites, of the 18 total sampled sites: retroauricular crease, palatine tonsils, and subgingival plaque. Metagenomes are now available for at least one body site for a total of 265 individuals. **b**, PCoA using Bray-Curtis

distances among all microbes at the species level. **c**, Relative abundances of the most prevalent and abundant microbes (bacterial, viral, eukaryotic, and archaeal) among all body sites, as profiled by MetaPhlAn2<sup>20</sup>. Prevalent eukaryotic microbes are shown at the genus level. **d**, Taxonomic profiles do not vary more between sequencing centres, batches, or clinical centres than they do among individuals within body sites. Ordinations show Bray-Curtis principal coordinates of species-level abundances at each body site. Within-site ecological structure is as expected<sup>1</sup>, with no divergence associated with technical variables along the first two ordination axes.

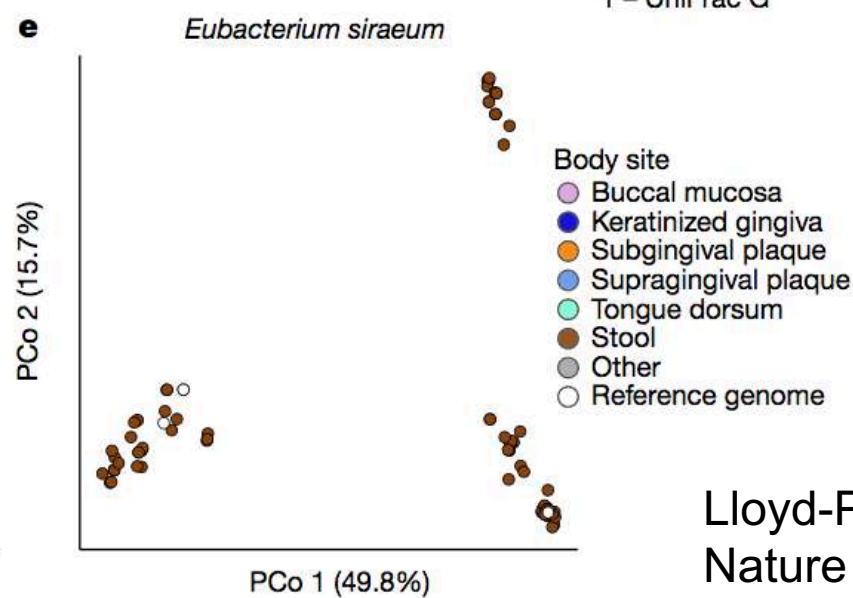
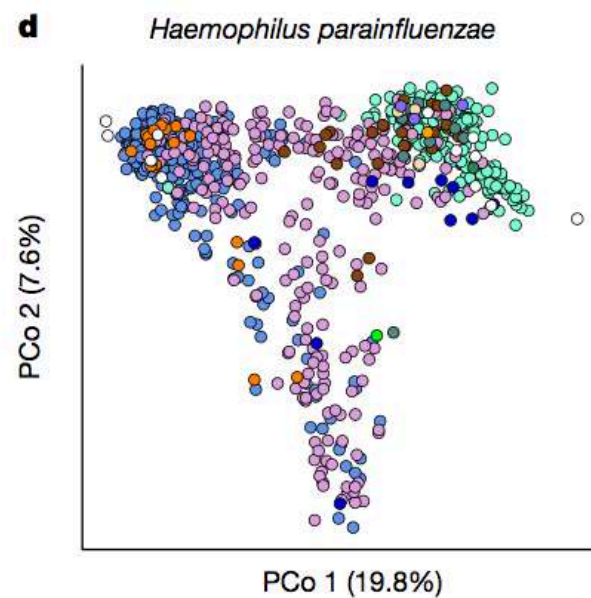
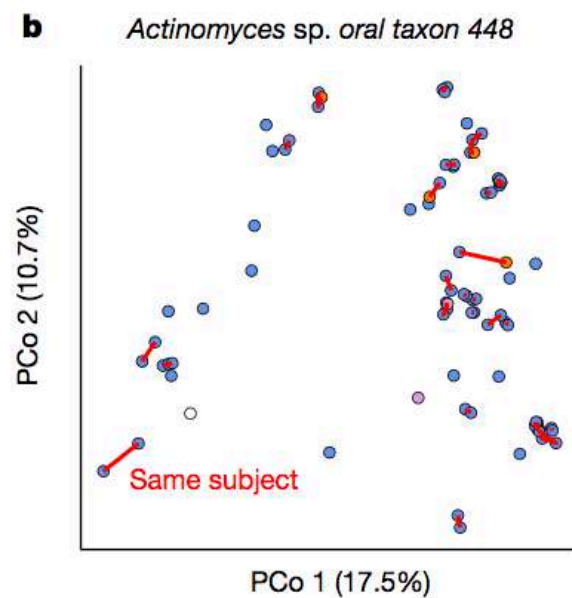
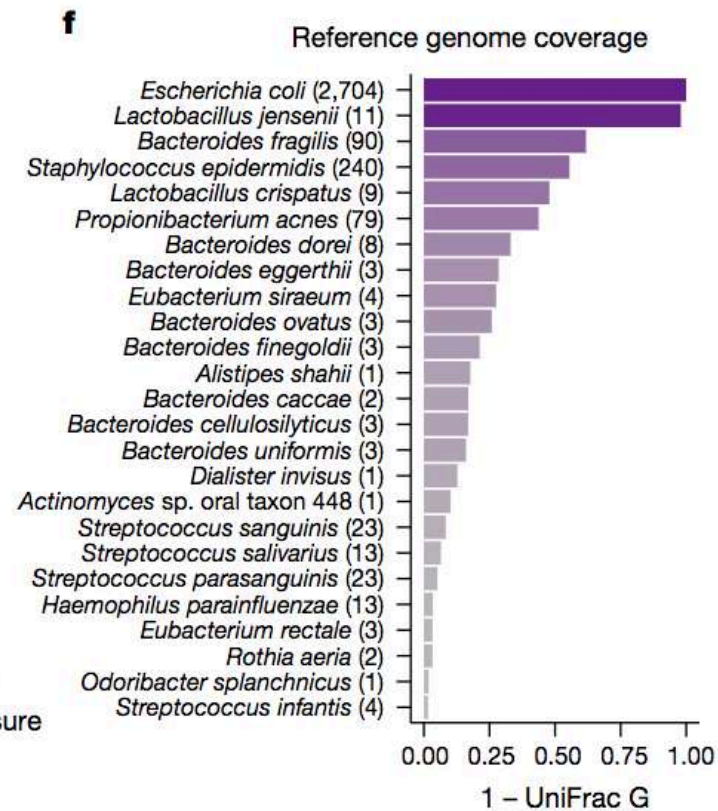
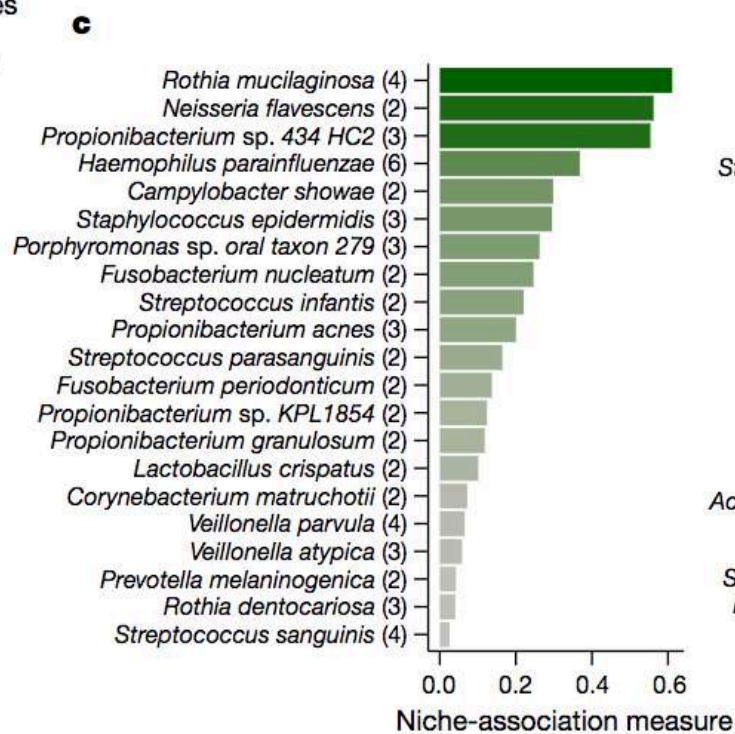
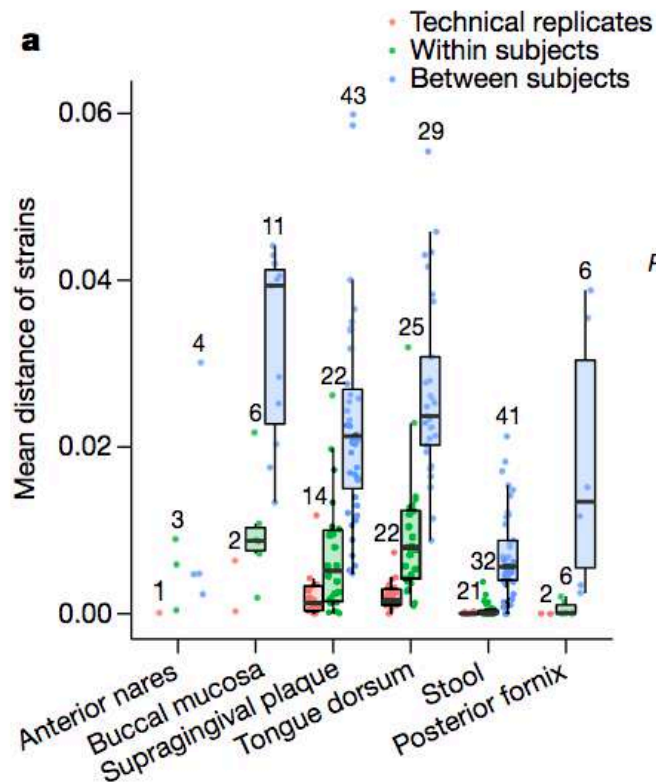




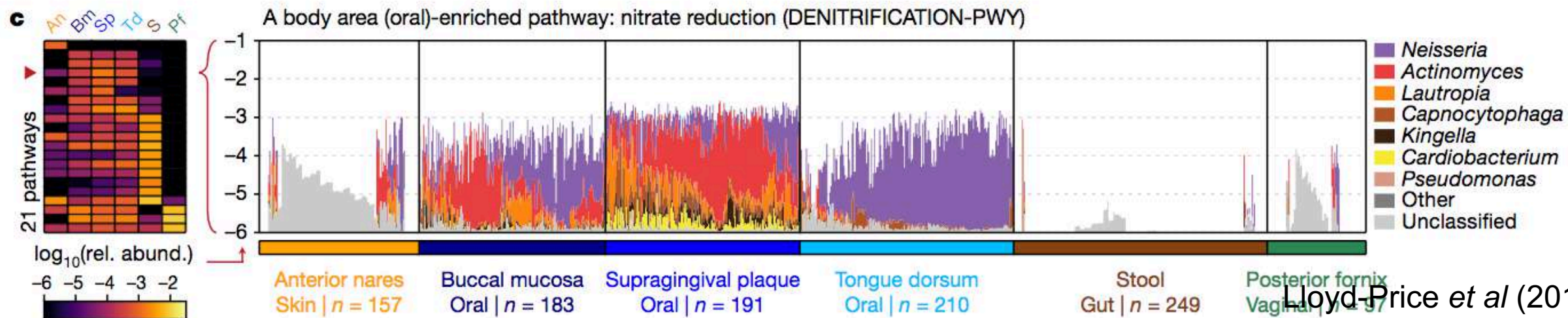
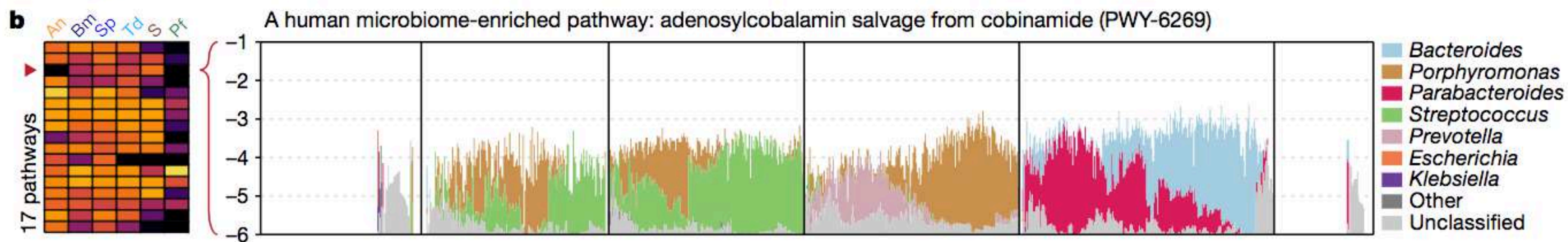
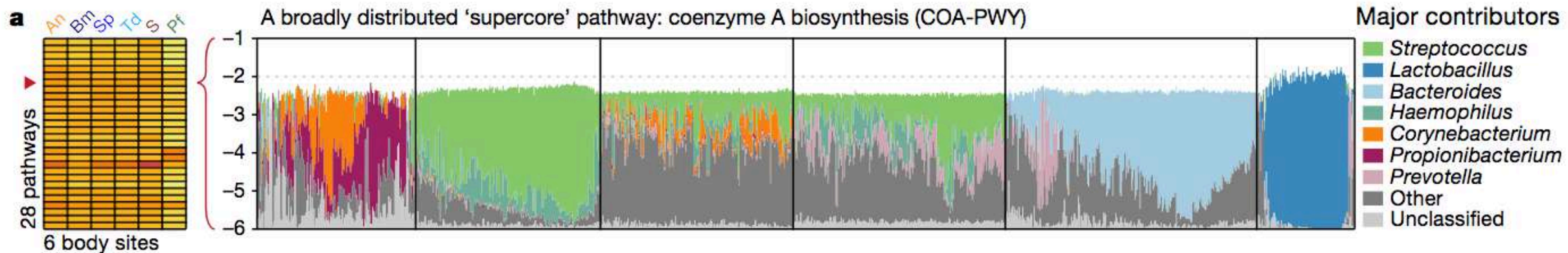
**Extended Data Figure 1 | Extended body-wide metagenomic taxonomic profiles in HMP1-II.** **a**, The combined HMP1-II datasets include a total of 2,355 metagenomes (724 previously published and 1,631 new, including 252 technical replicates). These span the project's six targeted body sites (anterior nares, buccal mucosa, supragingival plaque, tongue dorsum, stool, and posterior fornix) in addition to at least 20 samples each from 3 additional sites, of the 18 total sampled sites: retroauricular crease, palatine tonsils, and subgingival plaque. Metagenomes are now available for at least one body site for a total of 265 individuals. **b**, PCoA using Bray-Curtis

distances among all microbes at the species level. **c**, Relative abundances of the most prevalent and abundant microbes (bacterial, viral, eukaryotic, and archaeal) among all body sites, as profiled by MetaPhlan2<sup>20</sup>. Prevalent eukaryotic microbes are shown at the genus level. **d**, Taxonomic profiles do not vary more between sequencing centres, batches, or clinical centres than they do among individuals within body sites. Ordinations show Bray-Curtis principal coordinates of species-level abundances at each body site. Within-site ecological structure is as expected<sup>1</sup>, with no divergence associated with technical variables along the first two ordination axes.











# Environment dominates over host genetics in shaping human gut microbiota

Daphna Rothschild<sup>1,2\*</sup>, Omer Weissbrod<sup>1,2\*</sup>, Elad Barkan<sup>1,2\*</sup>, Alexander Kurilshikov<sup>3</sup>, Tal Korem<sup>1,2</sup>, David Zeevi<sup>1,2</sup>, Paul I. Costea<sup>1,2</sup>, Anastasia Godneva<sup>1,2</sup>, Iris N. Kalka<sup>1,2</sup>, Noam Bar<sup>1,2</sup>, Smadar Shilo<sup>1,2</sup>, Dar Lador<sup>1,2</sup>, Arnau Vich Vila<sup>3,4</sup>, Niv Zmora<sup>5,6,7</sup>, Meirav Pevsner-Fischer<sup>5</sup>, David Israeli<sup>8</sup>, Noa Kosower<sup>1,2</sup>, Gal Malka<sup>1,2</sup>, Bat Chen Wolf<sup>1,2</sup>, Tali Avnit-Sagi<sup>1,2</sup>, Maya Lotan-Pompan<sup>1,2</sup>, Adina Weinberger<sup>1,2</sup>, Zamir Halpern<sup>7,9</sup>, Shai Carmi<sup>10</sup>, Jingyuan Fu<sup>3,11</sup>, Cisca Wijmenga<sup>3,12</sup>, Alexandra Zhernakova<sup>3</sup>, Eran Elinav<sup>5§</sup> & Eran Segal<sup>1,2§</sup>

**Human gut microbiome composition is shaped by multiple factors but the relative contribution of host genetics remains elusive. Here we examine genotype and microbiome data from 1,046 healthy individuals with several distinct ancestral origins who share a relatively common environment, and demonstrate that the gut microbiome is not significantly associated with genetic ancestry, and that host genetics have a minor role in determining microbiome composition. We show that, by contrast, there are significant similarities in the compositions of the microbiomes of genetically unrelated individuals who share a household, and that over 20% of the inter-person microbiome variability is associated with factors related to diet, drugs and anthropometric measurements. We further demonstrate that microbiome data significantly improve the prediction accuracy for many human traits, such as glucose and obesity measures, compared to models that use only host genetic and environmental data. These results suggest that microbiome alterations aimed at improving clinical outcomes may be carried out across diverse genetic backgrounds.**

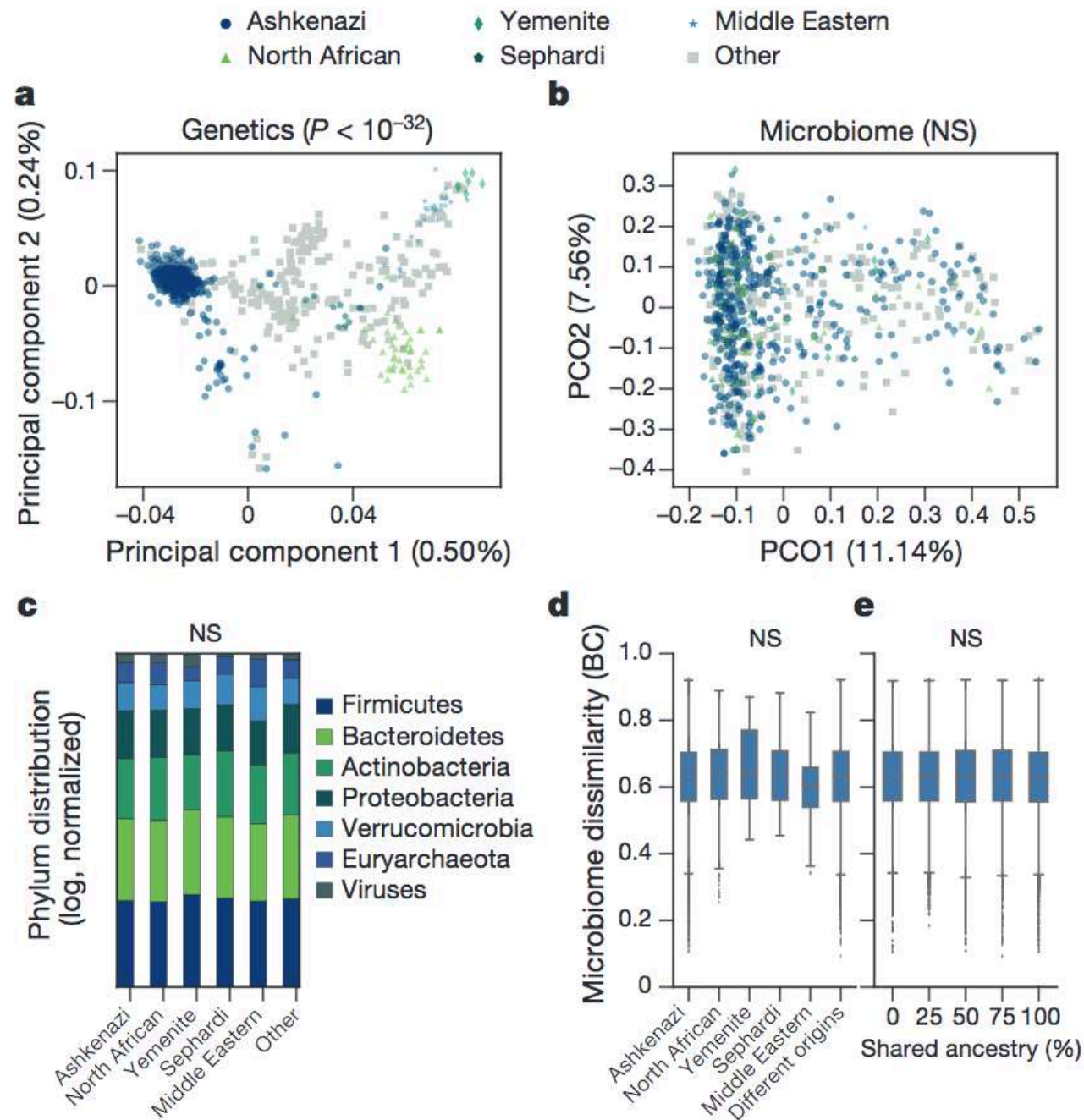
Received 7 June 2017; accepted 16 January 2018.

Published online 28 February 2018.



- 1,046 healthy Israeli adults
- 16S rRNA + metagenomics
- Genotyping 712,540 SNPs
- Questionnaires

**Figure 1 | Genetic ancestry is not significantly associated with microbiome composition.** **a**, Genetic principal components are strongly associated with self-reported ancestry, with Ashkenazi ( $n = 345$ ), North African ( $n = 42$ ), Middle Eastern ( $n = 24$ ), Sephardi ( $n = 10$ ), Yemenite ( $n = 8$ ) and admixed/other (other) ( $n = 286$ ) ancestries ( $P < 10^{-32}$ ; Kruskal–Wallis). **b**, As in **a**, but for microbiome principal coordinate analysis ( $P > 0.08$ ; Kruskal–Wallis). **c**, The distribution of average phylum abundance among 582 non-admixed individuals (in log scale, normalized to sum to 1.0) is not associated with ancestry ( $P > 0.05$ ; Kruskal–Wallis). NS, not significant. **d**, Box plots of Bray–Curtis (BC) dissimilarities across all pairs of 737 individuals for whom the ancestries of all grandparents are known, demonstrating that microbiome composition is not associated with ancestry ( $P > 0.06$ ; Kruskal–Wallis test for the top five Bray–Curtis PCOs).  $n = 105,570$  (Ashkenazi), 1,711 (North African), 528 (Middle Eastern), 136 (Sephardi) and 78 (Yemenite) same ancestry pairs;  $n = 61,048$  different ancestry pairs. The lower and upper limits of the boxes represent the 25% and 75% percentiles, respectively, and the top and bottom whiskers represent the 5% and 95% percentiles, respectively. **e**, Box plots of Bray–Curtis dissimilarities across pairs of 946 individuals (including admixed individuals), organized according to shared ancestry fraction (the fraction of grandparents of the same ancestry), for pairs with 0% ( $n = 167,618$ ), 25% ( $n = 33,119$ ), 50% ( $n = 100,163$ ), 75% ( $n = 34,187$ ) and 100% ( $n = 111,898$ ) shared ancestry fractions. The lower and upper limits of the boxes represent the 25% and 75% percentiles, respectively, and the top and bottom whiskers represent the 5% and 95% percentiles, respectively. The figure demonstrates that microbiome similarity is not associated with ancestral similarity ( $P = 0.73$ ; Mantel test).

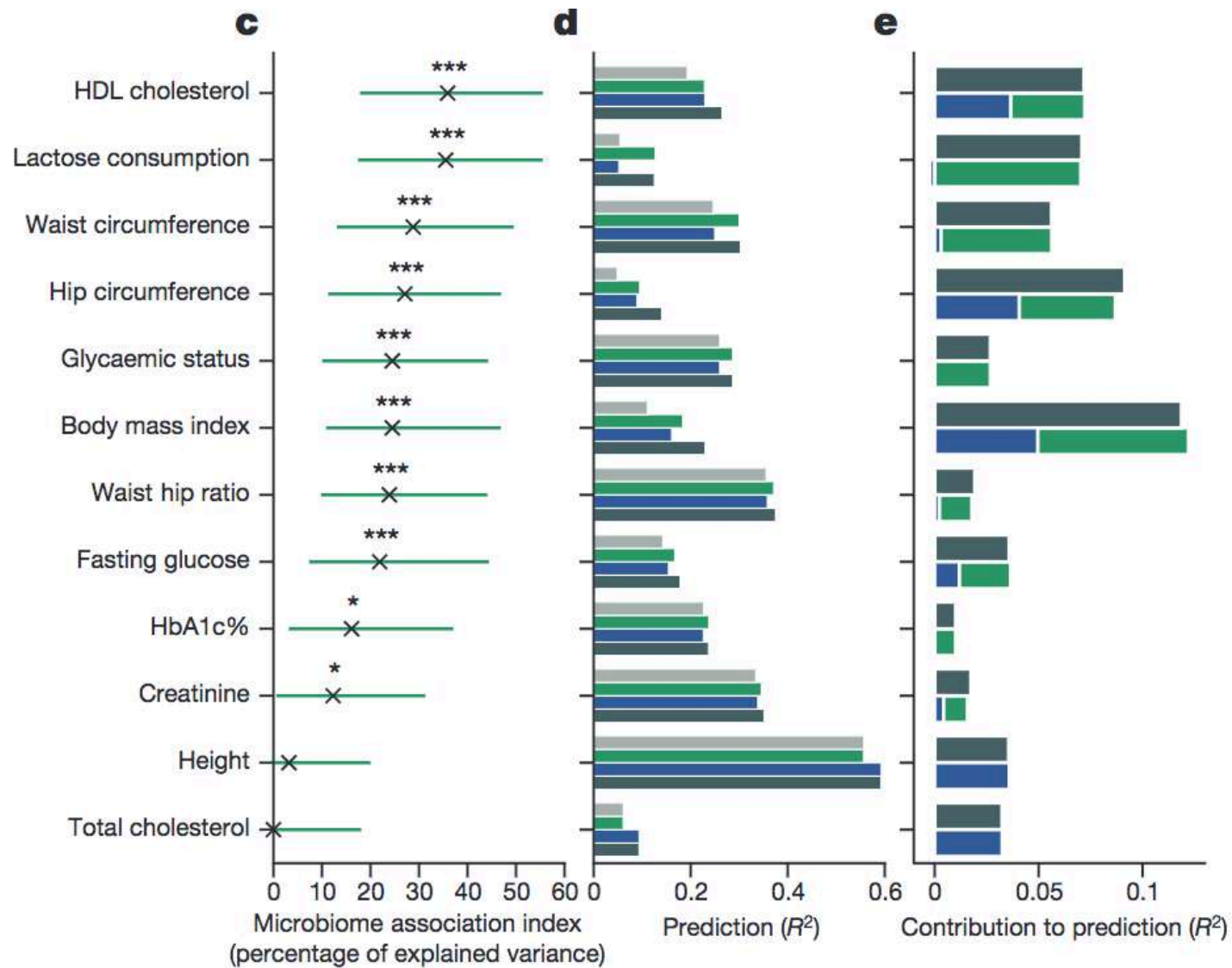


**b**

Phenotype	Microbiome association index		Genetic heritability (literature)
	Israeli cohort	LLD cohort	
HDL	35.9%***	27.9%***	23.9%–48%
Lactose cons.	35.5%***	N/A	N/A
Waist circ.	28.8%***	26%***	15%–24%
Hip circ.	27.1%***	28%***	10.6%–27%
Glycaemic status	24.5%***	N/A	N/A
BMI	24.5%***	27.8%***	14%–32%
WHR	23.9%***	6.9%*	12%–14%
Fasting glucose	21.9%***	8%**	9%–33%
HbA1c%	16.1%*	8.4%	21%–32%
Creatinine	12.3%*	6.7%	19%–25%
Height	3.2%	25.9%***	33%–68%
Total cholesterol	0%	13.5%	14%–53%

indicate a greater confidence in the estimation. **b**,  $b^2$  estimates from the analysis of 715 individuals with measured genotyped and gut microbiomes from the Israeli cohort (left column) and of 836 individuals from the LLD cohort (middle column) are comparable to previous genetic heritability estimates<sup>27–34</sup> (right column). \*FDR < 0.05, \*\*FDR < 0.01 and \*\*\*FDR < 0.001. Cons., consumption, circ., circumference. **c**,  $b^2$  estimates





\*\*\*FDR < 0.001. Cons., consumption, circ., circumference. **c**,  $b^2$  estimates of several human phenotypes and their 95% confidence intervals, evaluated using 715 individuals. \*FDR < 0.05, \*\*FDR < 0.01 and \*\*\*FDR < 0.001. **d**, Phenotype prediction accuracy for 715 individuals, evaluated using a LMM under different sets of predictive features (measured using coefficient of determination ( $R^2$ )), using four different models for each phenotype: (i) 'Basic', age, gender and diet features; (ii) 'Basic + microbiome', basic features and relative abundances of bacterial genes; (iii) 'Basic + genetics', basic features and host genotypes; and (iv) 'Basic + genetics + microbiome': basic features, relative abundances of bacterial genes and host genotypes. **e**, The additive contribution of microbiome and genetics to prediction performance evaluated using a LMM across 715 individuals, over a model that includes only basic features. The joint contribution of microbiome and genetics is similar to the sum of the individual contributions, suggesting these are independent contributions.

Basic: Age + gender + calories

- Basic
- Basic + microbiome
- Basic + genetics
- Basic + genetics + microbiome
- Microbiome
- Genetics
- Genetics + microbiome

# Case studies – cow rumen metagenomics



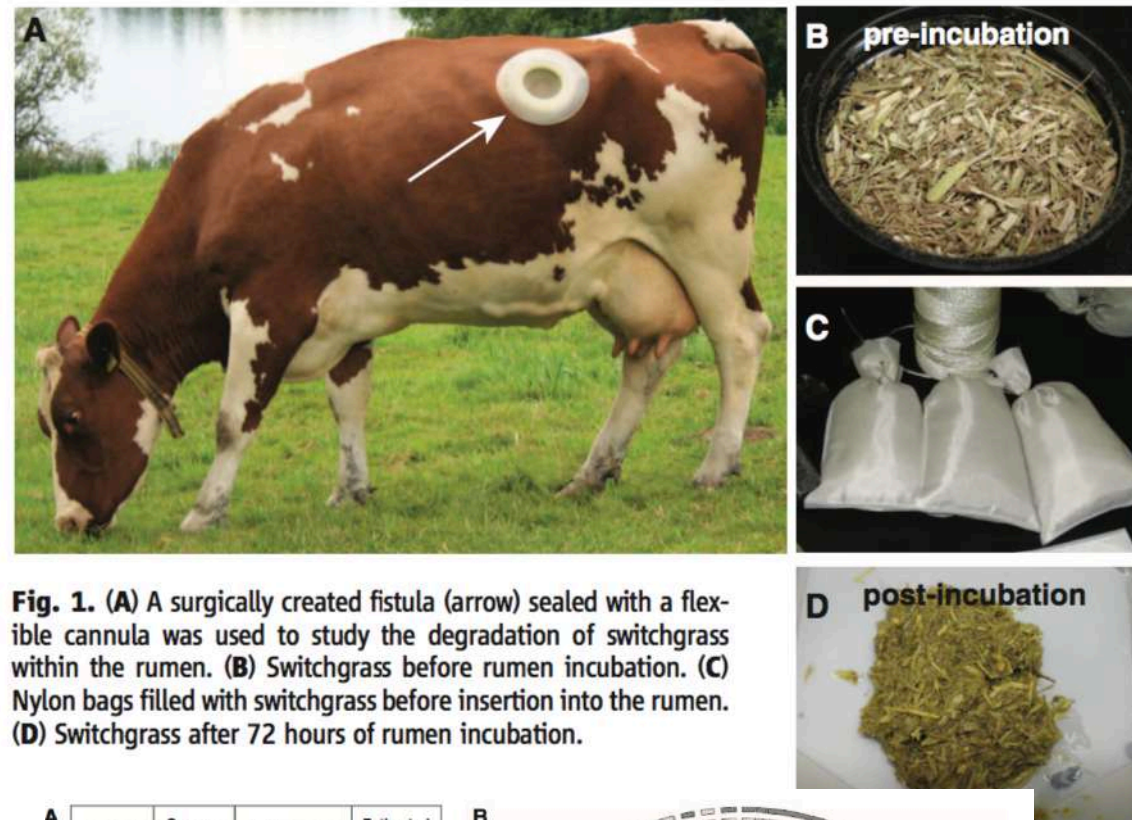
## Example of metagenomics

# Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen

Matthias Hess,<sup>1,2\*</sup> Alexander Sczyrba,<sup>1,2\*</sup> Rob Egan,<sup>1,2</sup> Tae-Wan Kim,<sup>3</sup> Harshal Chokhawala,<sup>3</sup> Gary Schroth,<sup>4</sup> Shujun Luo,<sup>4</sup> Douglas S. Clark,<sup>3,5</sup> Feng Chen,<sup>1,2</sup> Tao Zhang,<sup>1,2</sup> Roderick I. Mackie,<sup>6</sup> Len A. Pennacchio,<sup>1,2</sup> Susannah G. Tringe,<sup>1,2</sup> Axel Visel,<sup>1,2</sup> Tanja Woyke,<sup>1,2</sup> Zhong Wang,<sup>1,2</sup> Edward M. Rubin<sup>1,2†</sup>

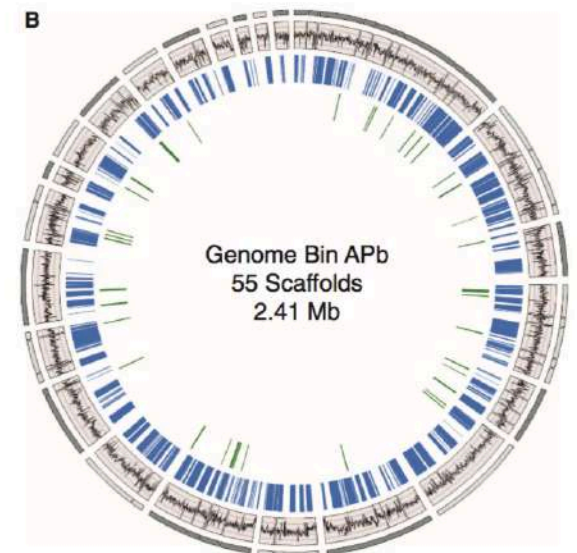
- 268Gb of metagenomics data
- Identified 27,755 putative carbohydrate-active genes from a cow rumen metagenome
- Expressed 90 candidates of which 57% had enzymatic activity against cellulosic substrates
- Assembled 15 uncultured microbial genomes

Hess *et al.*, 2011 **Science**



**Fig. 1.** (A) A surgically created fistula (arrow) sealed with a flexible cannula was used to study the degradation of switchgrass within the rumen. (B) Switchgrass before rumen incubation. (C) Nylon bags filled with switchgrass before insertion into the rumen. (D) Switchgrass after 72 hours of rumen incubation.

Genome Bin	Genome Size (Mb)	Phylogenetic Order	Estimated Completeness
AFa	2.87	Spirochaetales	92.98%
AMa	2.21	Spirochaetales	91.23%
Ala	2.53	Clostridiales	90.10%
AGa	3.08	Bacteroidales	89.77%
AN	2.02	Clostridiales	78.50%
AJ	2.24	Bacteroidales	75.96%
AC2a	2.07	Bacteroidales	75.96%
AWa	2.02	Clostridiales	75.77%
AH	2.52	Bacteroidales	75.45%
AQ	1.91	Bacteroidales	71.36%
AS1a	1.75	Clostridiales	70.99%
APb	2.41	Clostridiales	64.85%
BOa	1.67	Clostridiales	64.16%
ADa	2.99	Myxococcales	62.13%
ATa	1.87	Clostridiales	60.41%

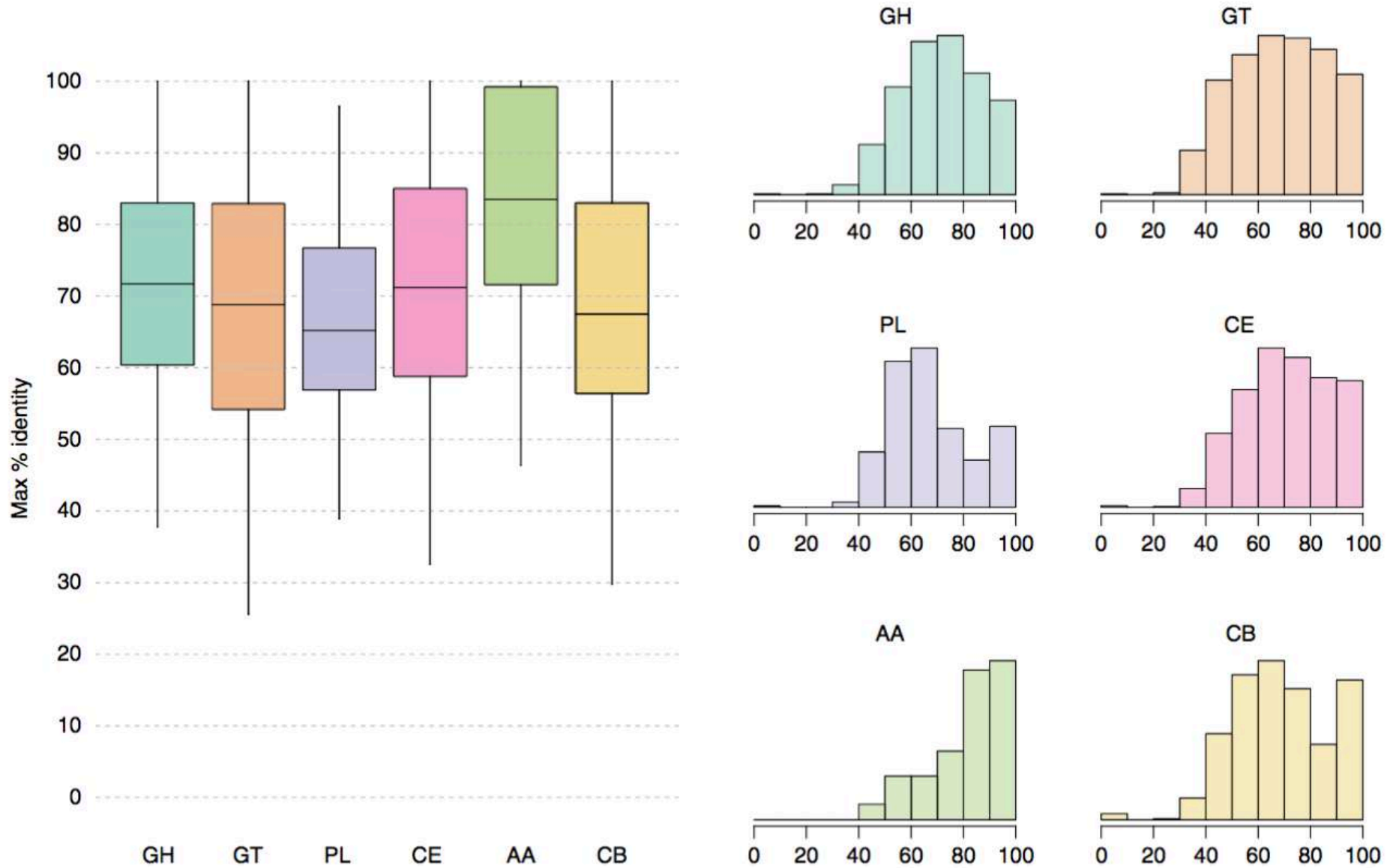


enome; innermost circle (green tick marks), location of glycoside hydrolase genes on draft genome.

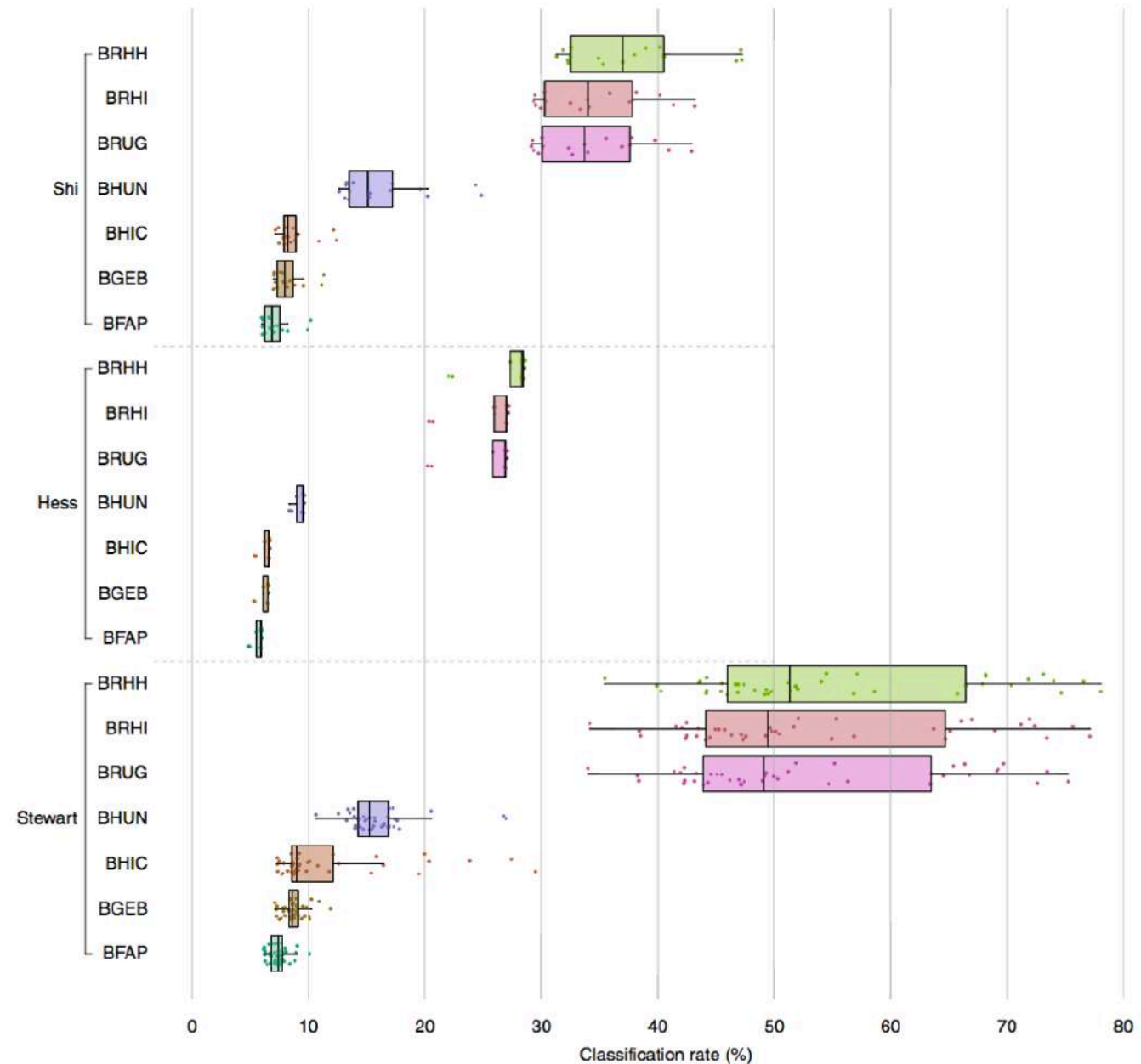




- The draft genomes contain over 69,000 proteins predicted to be involved in carbohydrate metabolism, over 90% of which do not have a good match in public databases.



- Inclusion of the 913 genomes presented here improves metagenomic read classification by sevenfold against the study's own data, and by fivefold against other publicly available rumen datasets.
- dataset substantially improves the coverage of rumen microbial genomes in the public databases and represents a valuable resource for biomass-degrading enzyme discovery and studies of the rumen microbiome



**Fig. 4** Classification rate for three datasets against various Kraken databases. BFAP bacterial, archaeal, fungal and protozoan genomes from RefSeq, BGEB BFAP + 1003 GEBA genomes, BHIC BFAP + 63 hRUG genomes, BHUN BFAP + 410 genomes from the Hungate 1000 project, BRUG BFAP + 850 RUG MAGs, BRHI BFAP + all 913 genomes from this study, BRHH BFAP + 913 RUGs + 410 Hungate 1000 genomes. Addition of rumen-specific RUGs or Hungate 1000 genomes has the most dramatic effect

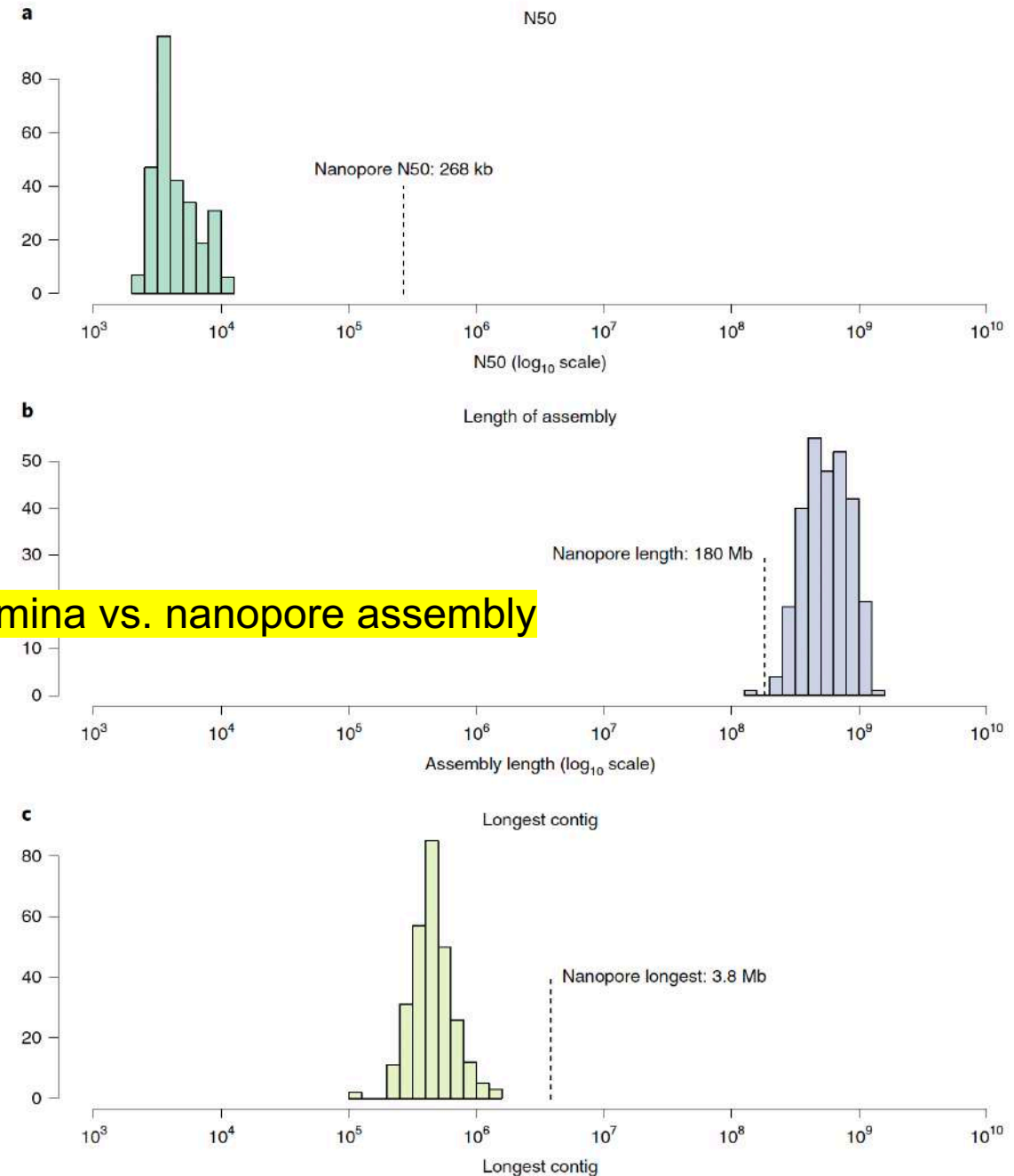


# Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery

Robert D. Stewart<sup>1</sup>, Marc D. Auffret<sup>2</sup>, Amanda Warr<sup>1</sup>, Alan W. Walker<sup>3</sup>, Rainer Roehe<sup>2</sup> and Mick Watson<sup>1\*</sup>

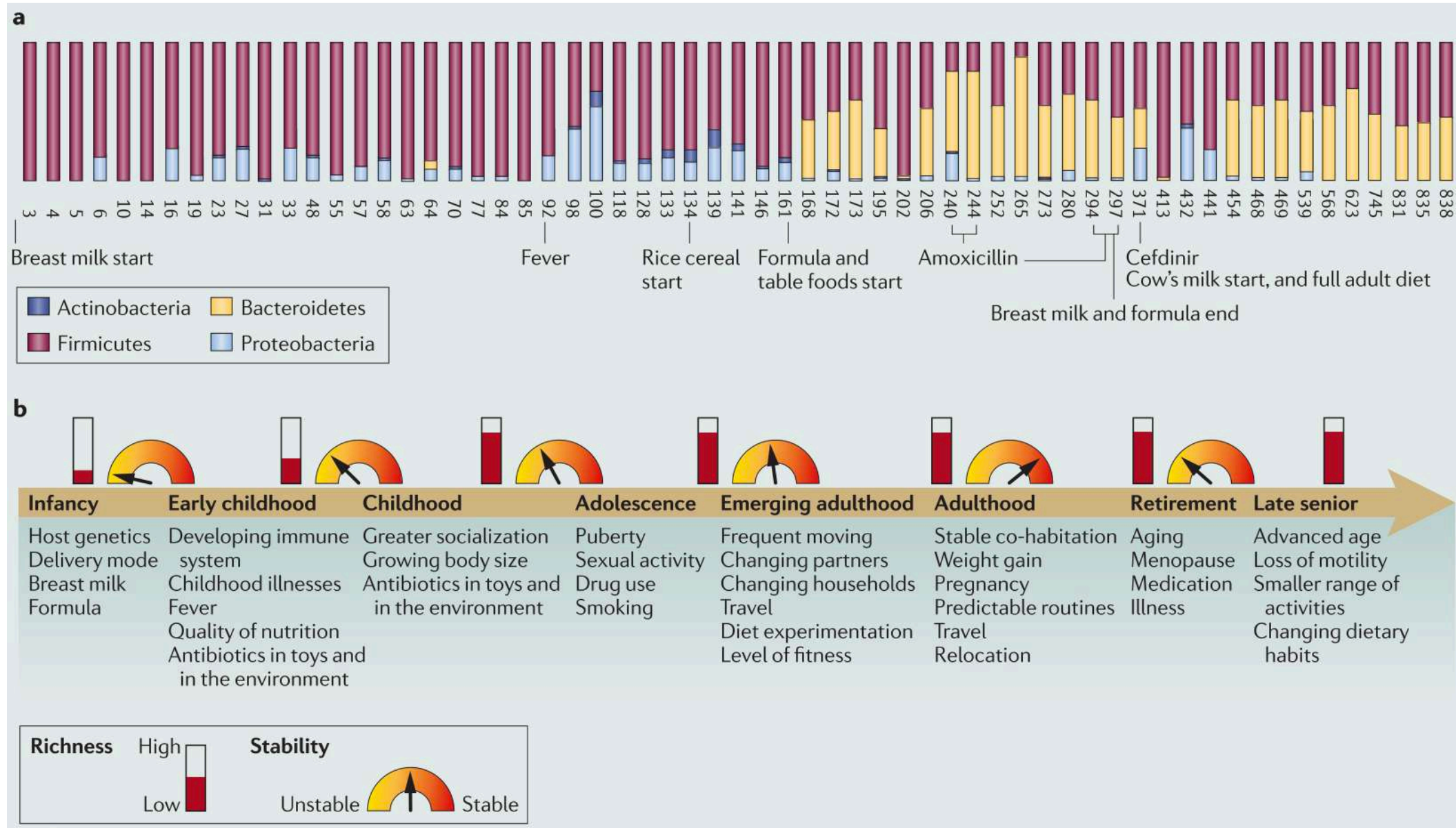
- **6.5 Tb** of sequence data derived from **283** ruminant cattles
- Using metagenomic binning and Hi-C techniques
- Assembly of **4,941** draft bacterial and archaeal genomes
- Long read is being used: “We also present a metagenomic assembly of nanopore (MinION) sequencing data (from one rumen sample) that contains at least three whole bacterial chromosomes as single contigs”

## 282 Illumina vs. nanopore assembly



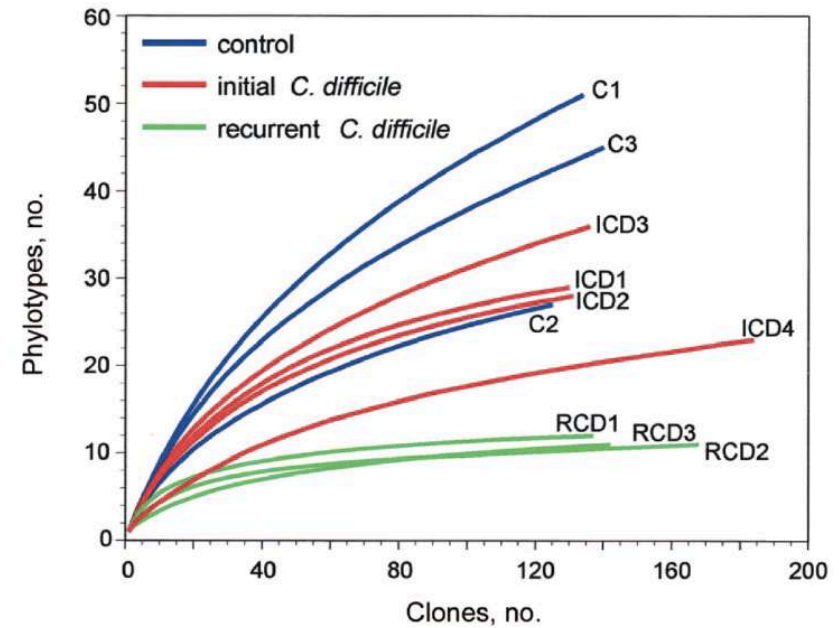
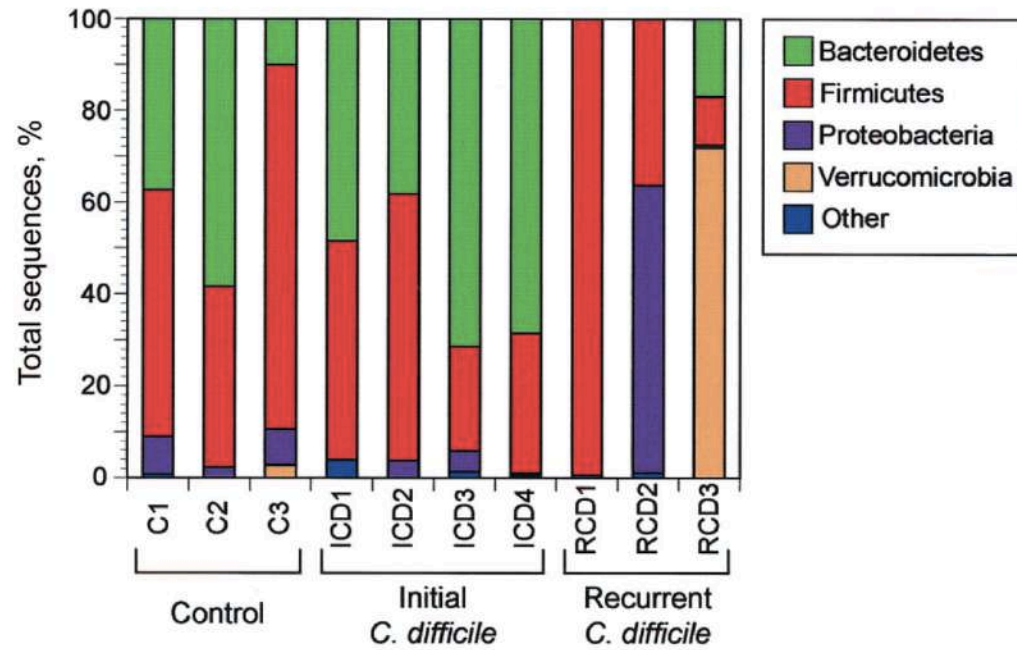
Case studies – others

# The gut microbiome during life



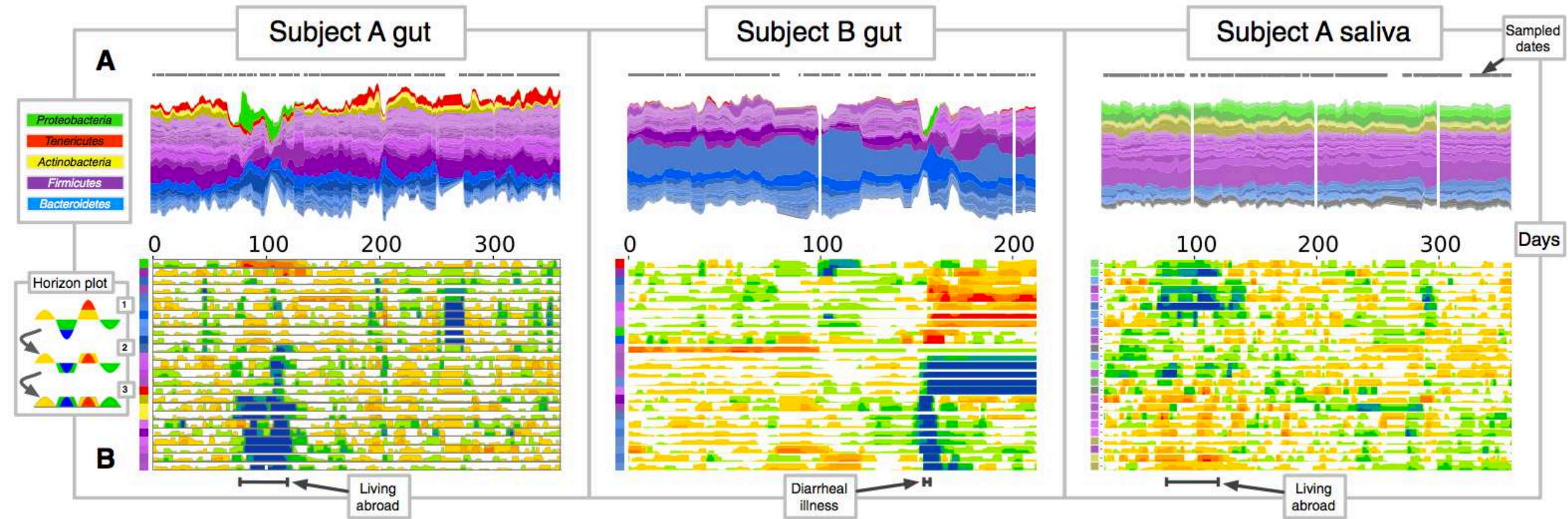
# Decreased diversity with *Clostridium difficile* – associated diarrhea

A

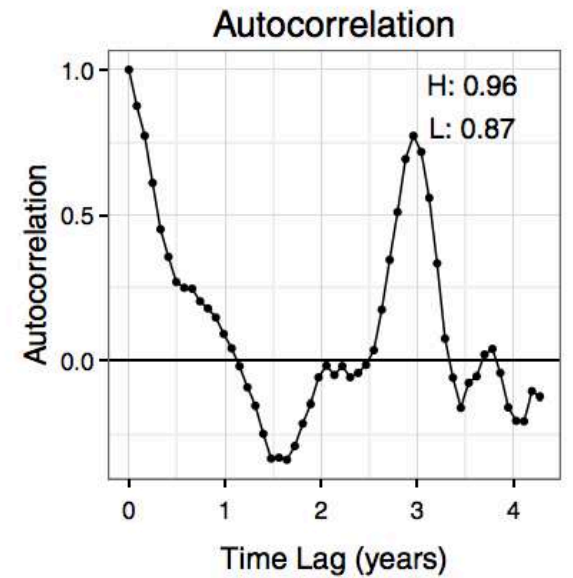
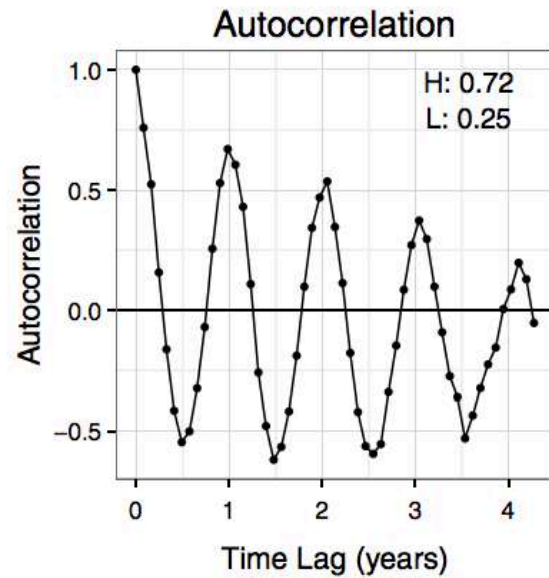
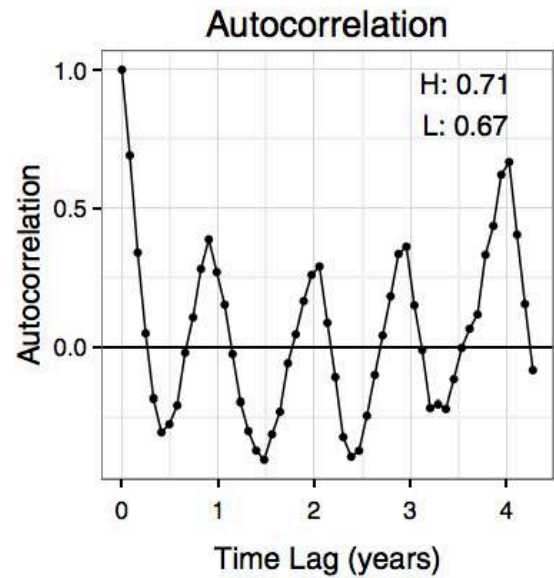
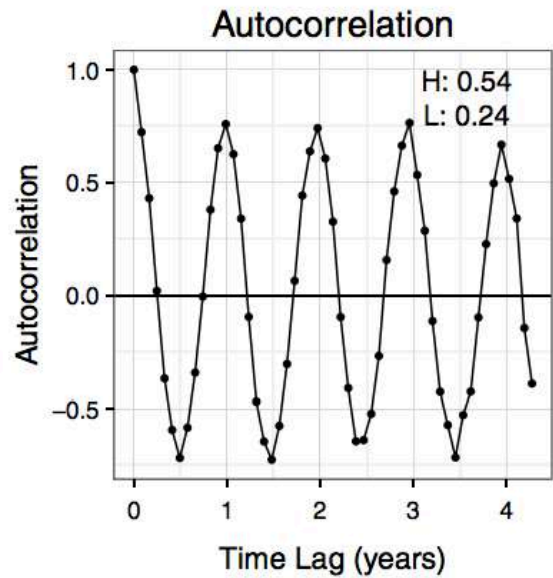
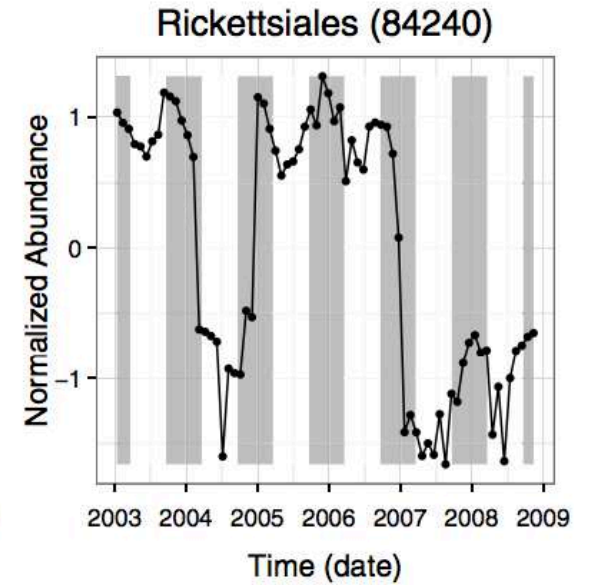
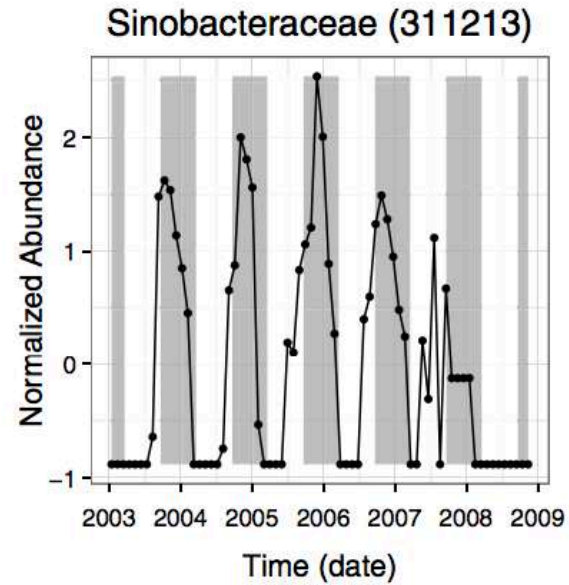
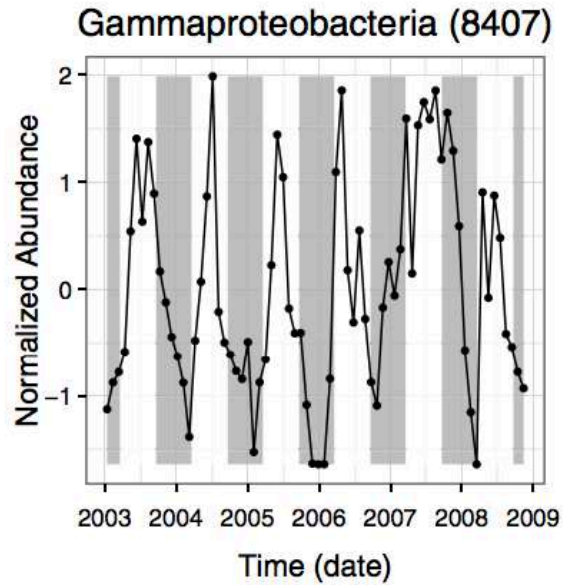
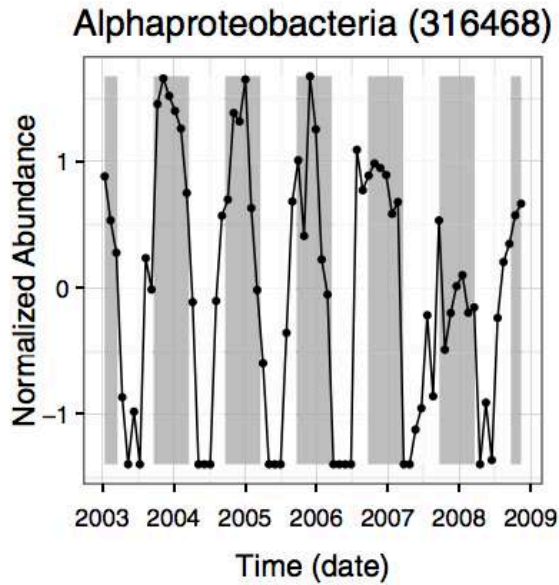




# Tracking microbiome on a daily scale

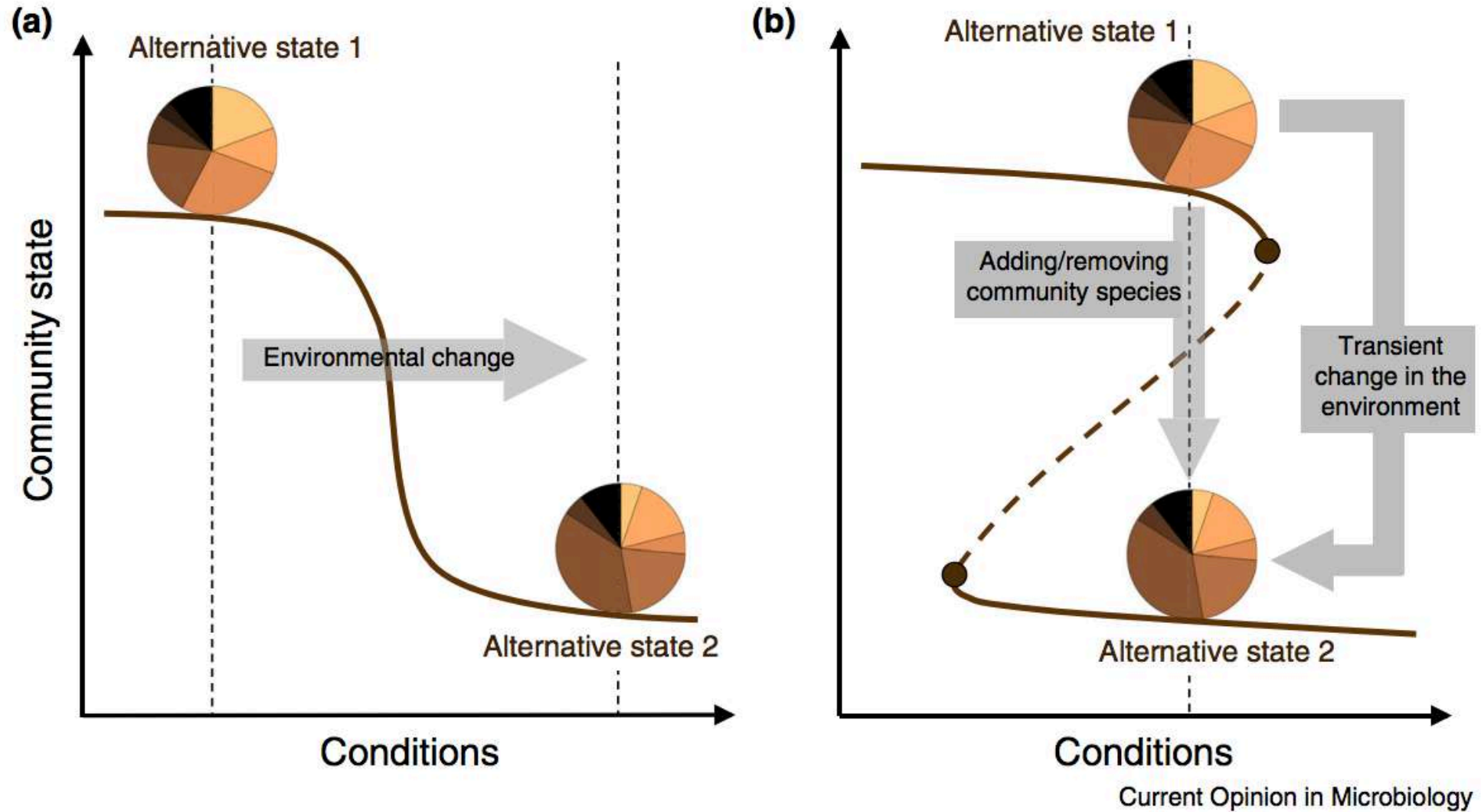


# Tracking microbiome spanning 6 years





# Tracking microbiome on a daily scale



# Question: What community gets reset and what don't?

A. Shade, J.S. Read, N.D. Youngblut, N. Fierer, R. Knight, T.K. Kratz, N.R. Lottig, E.E. Roden, E.H. Stanley, J. Stombaugh, et al.

Lake microbial communities are resilient after a whole-ecosystem disturbance **Yes**  
ISME J, 6 (2012), pp. 2153–2167

L. Dethlefsen, D.A. Relman

Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation

Proc Natl Acad Sci U S A, 108 (2011), pp. 4554–4561 **No**

L.A. David, A.C. Materna, J. Friedman, M.I. Campos-Baptista, M.C. Blackburn, A. Perrotta, S.E. Erdman, E.J. Alm

Host lifestyle affects human microbiota on daily timescales **Yes and No**  
Genome Biol, 15 (2014), p. R89



# Priority effect

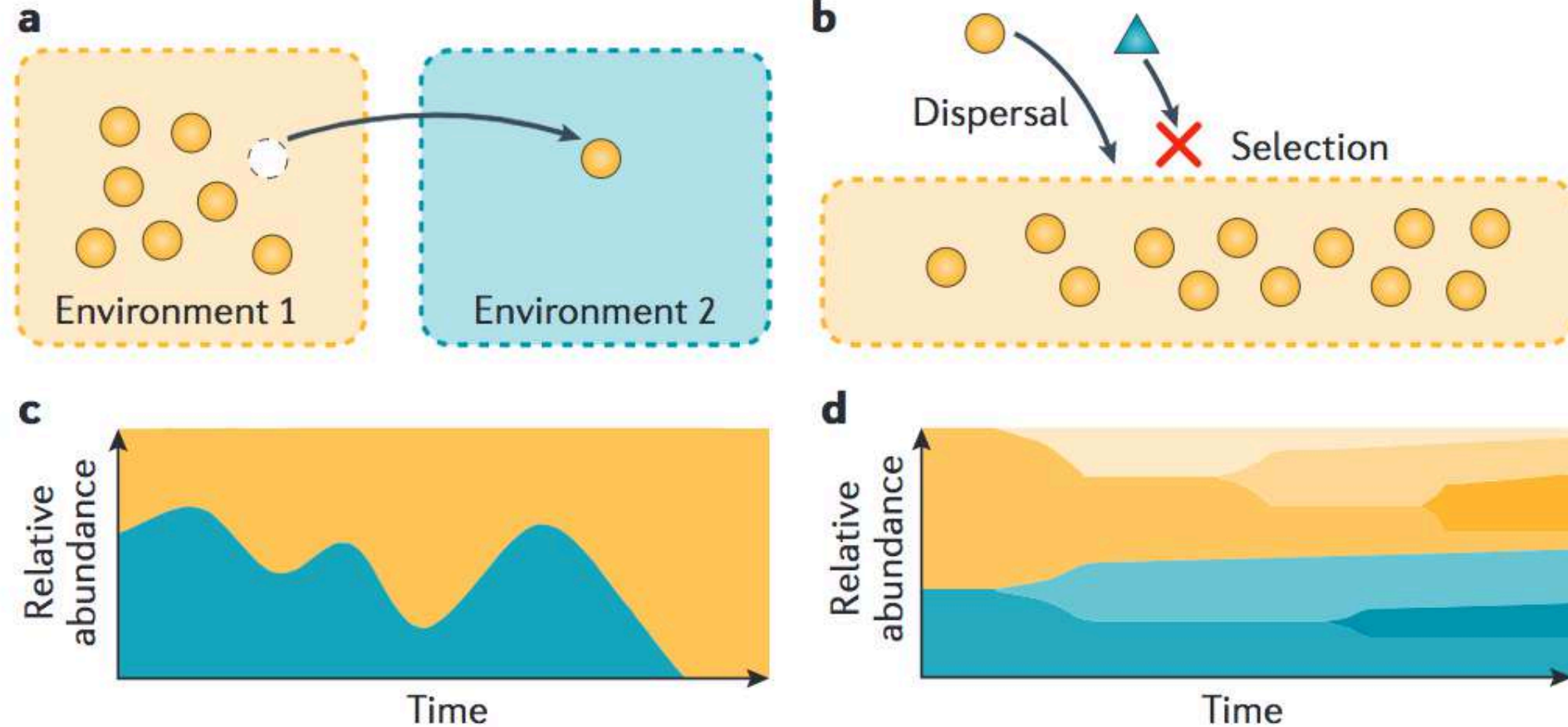
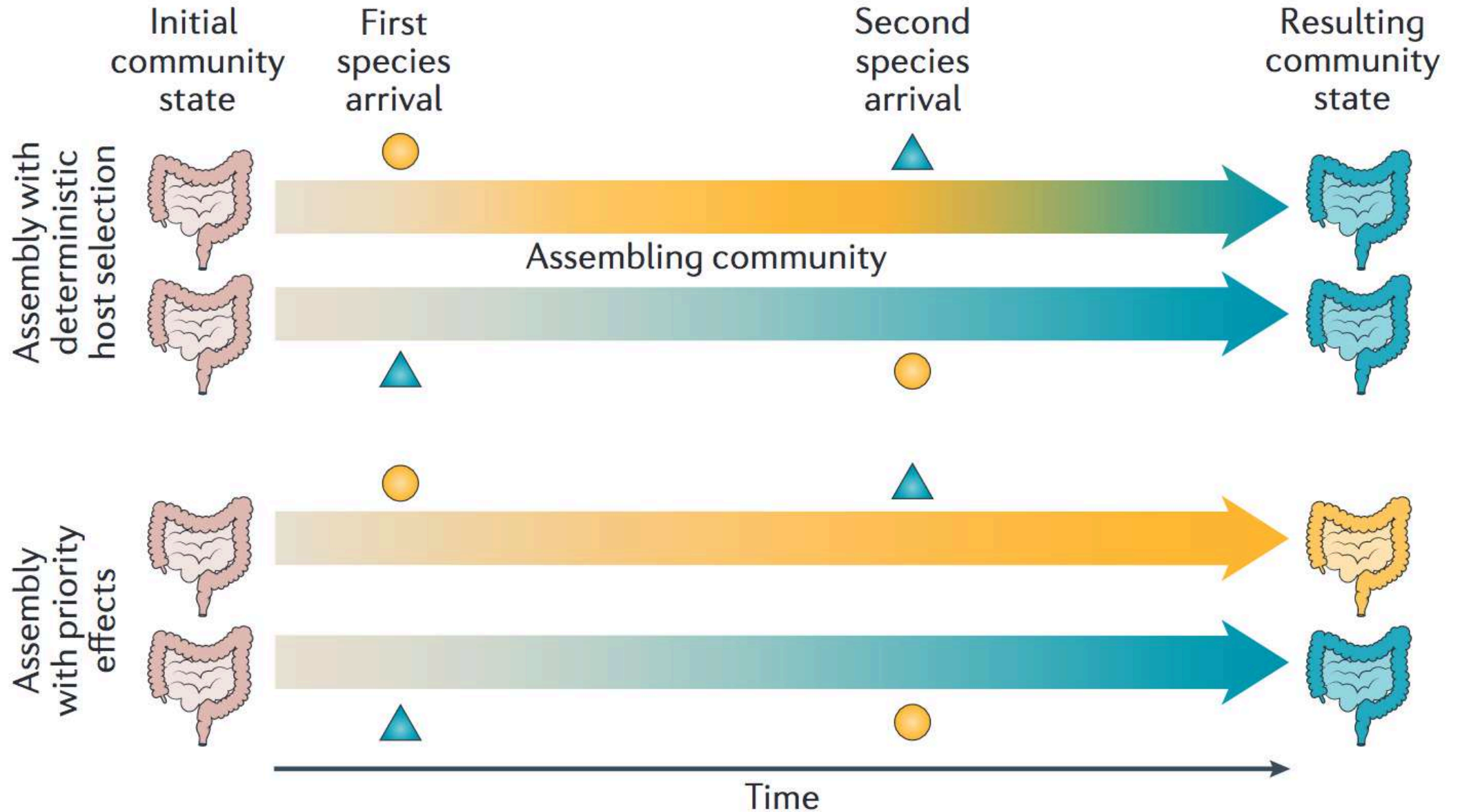


Figure 1 | **Four processes that affect ecological communities.** **a** | The arrow represents dispersal of an organism (orange circle) from Environment 1 (orange shading) to Environment 2 (blue shading). **b** | Deterministic fitness differences between two species (orange circle, blue triangle) cause the orange environment to select for one (orange circle) and against the other (blue triangle). **c** | Stochastic changes in the relative abundances of two species (orange area and blue area) result in changes in community structure within one environment through time. As a result, one population (blue) has gone locally extinct by the end of the time period. **d** | Mutation and/or recombination within a population (blue and orange areas) results in new genetic variation through time, leading to new strains (as denoted by different shades).

# Priority effect





# A communal catalogue reveals Earth's multiscale microbial diversity

Luke R. Thompson<sup>1,2,3</sup>, Jon G. Sanders<sup>1</sup>, Daniel McDonald<sup>1</sup>, Amnon Amir<sup>1</sup>, Joshua Ladau<sup>4</sup>, Kenneth J. Locey<sup>5</sup>, Robert J. Prill<sup>6</sup>, Anupriya Tripathi<sup>1,7,8</sup>, Sean M. Gibbons<sup>9,10</sup>, Gail Ackermann<sup>1</sup>, Jose A. Navas-Molina<sup>1,11</sup>, Stefan Janssen<sup>1</sup>, Evguenia Kopylova<sup>1</sup>, Yoshiki Vázquez-Baeza<sup>1,11</sup>, Antonio González<sup>1</sup>, James T. Morton<sup>1,11</sup>, Siavash Mirarab<sup>12</sup>, Zhenjiang Zech Xu<sup>1</sup>, Lingjing Jiang<sup>1,13</sup>, Mohamed F. Haroon<sup>14</sup>, Jad Kanbar<sup>1</sup>, Qiyun Zhu<sup>1</sup>, Se Jin Song<sup>1</sup>, Tomasz Kosciolk<sup>1</sup>, Nicholas A. Bokulich<sup>15</sup>, Joshua Lefler<sup>1</sup>, Colin J. Brislawn<sup>16</sup>, Gregory Humphrey<sup>1</sup>, Sarah M. Owens<sup>17</sup>, Jarrad Hampton-Marcell<sup>17,18</sup>, Donna Berg-Lyons<sup>19</sup>, Valerie McKenzie<sup>20</sup>, Noah Fierer<sup>20,21</sup>, Jed A. Fuhrman<sup>22</sup>, Aaron Clauset<sup>19,23</sup>, Rick L. Stevens<sup>24,25</sup>, Ashley Shade<sup>26,27,28</sup>, Katherine S. Pollard<sup>4</sup>, Kelly D. Goodwin<sup>3</sup>, Janet K. Jansson<sup>16</sup>, Jack A. Gilbert<sup>17,29</sup>, Rob Knight<sup>1,11,30</sup> & The Earth Microbiome Project Consortium\*

Our growing awareness of the microbial world's importance and diversity contrasts starkly with our limited understanding of its fundamental structure. Despite recent advances in DNA sequencing, a lack of standardized protocols and common analytical frameworks impedes comparisons among studies, hindering the development of global inferences about microbial life on Earth. Here we present a meta-analysis of microbial community samples collected by hundreds of researchers for the Earth Microbiome Project. Coordinated protocols and new analytical methods, particularly the use of exact sequences instead of clustered operational taxonomic units, enable bacterial and archaeal ribosomal RNA gene sequences to be followed across multiple studies and allow us to explore patterns of diversity at an unprecedented scale. The result is both a reference database giving global context to DNA sequence data and a framework for incorporating data from future studies, fostering increasingly complete characterization of Earth's microbial diversity.

earth  
microbiomeproject

BY THE  
NUMBERS

27,751

samples

7 continents

43 countries

2,212,796,183

total DNA sequences

307,572

unique DNA sequences  
(approx. species)

50+

peer-reviewed  
publications

500+

scientists



92 environmental  
features

66 animal host  
species

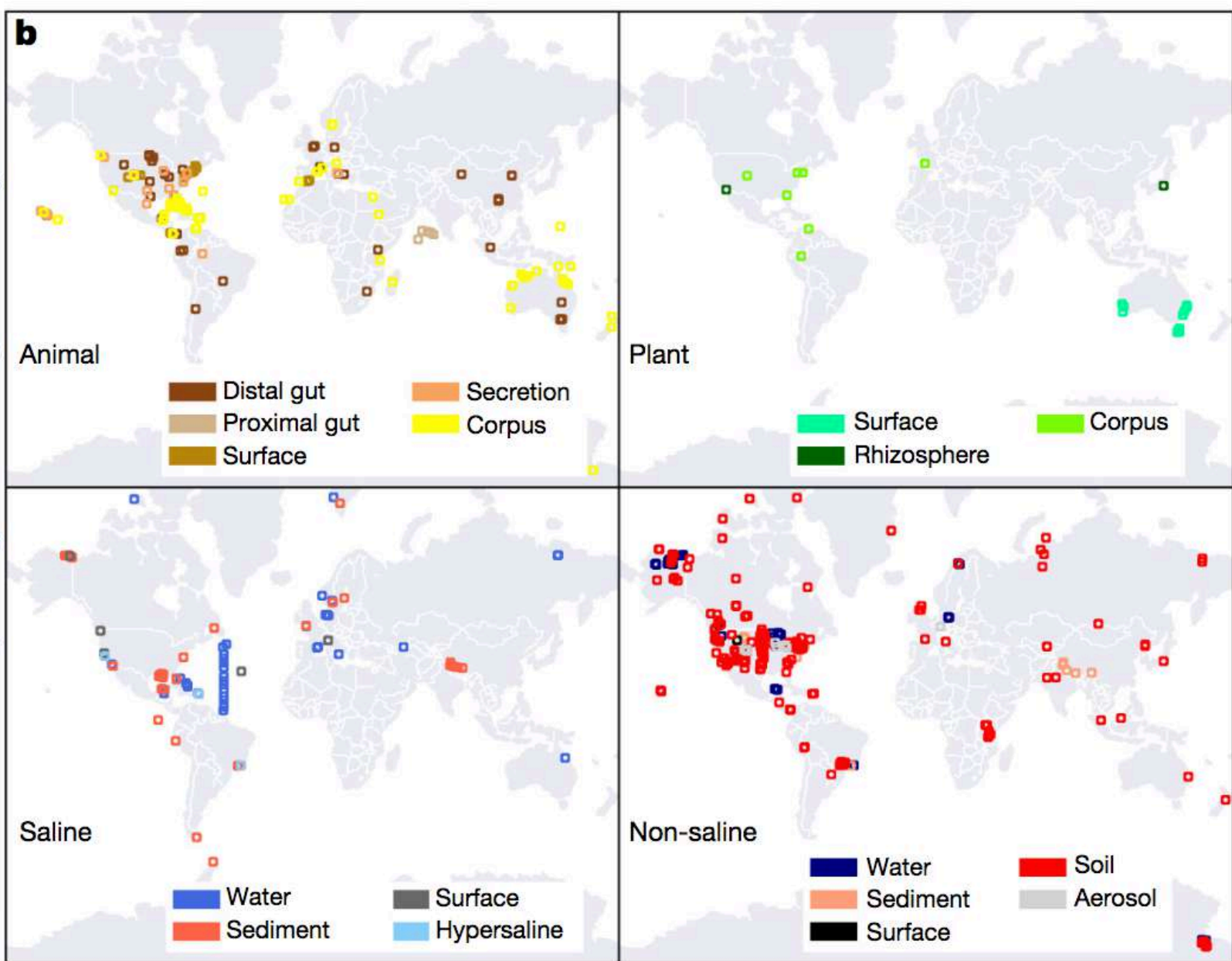
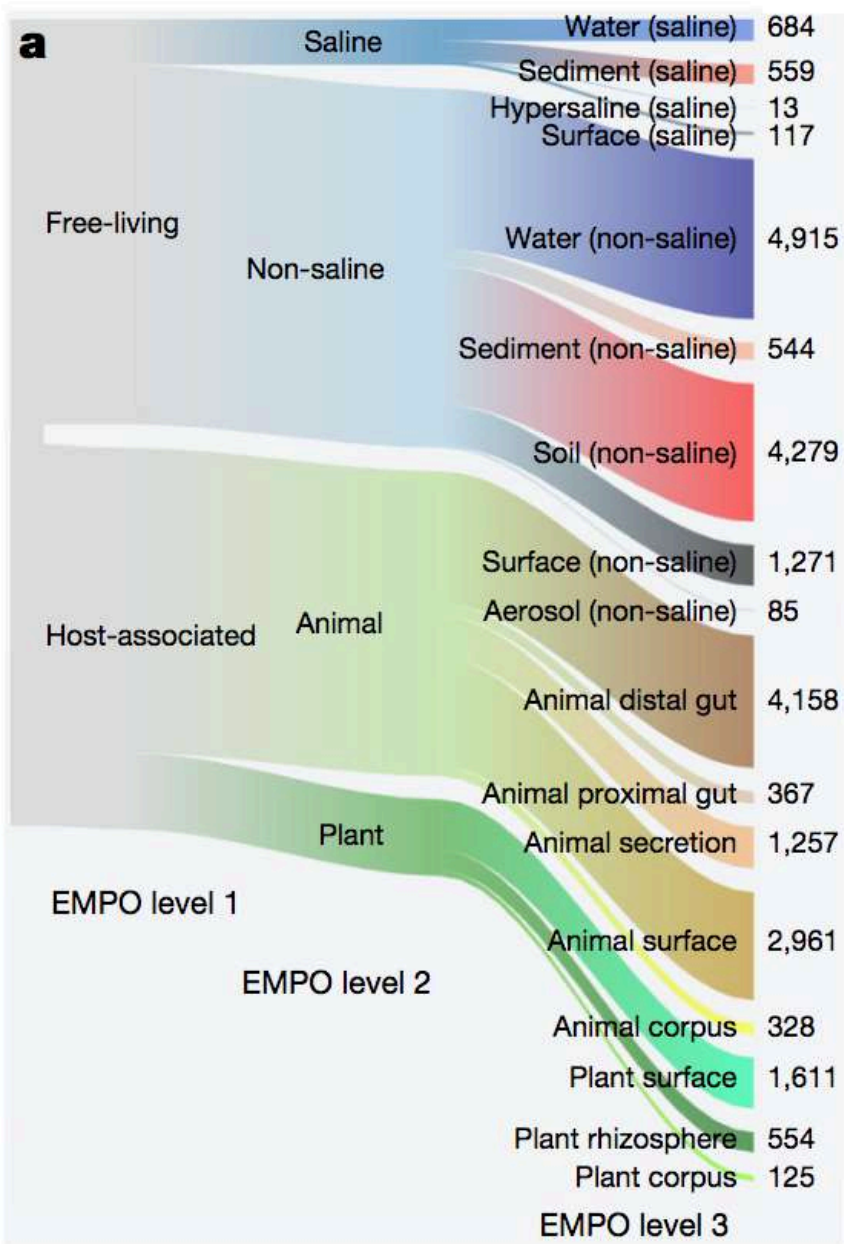
2 – 12 pH range  
(stomach acid to household  
ammonia)

78.9 °N –  
78.2 °S

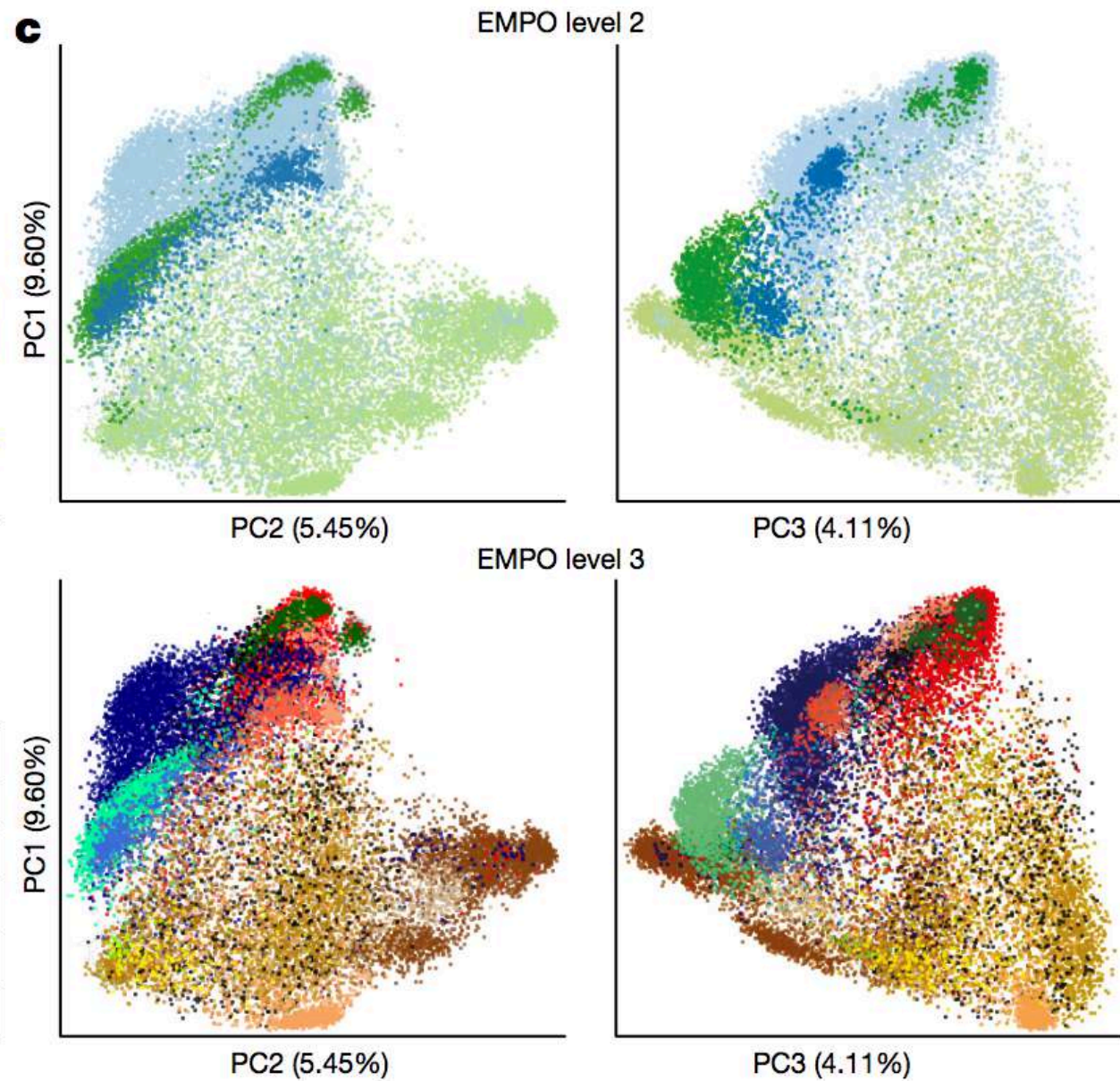
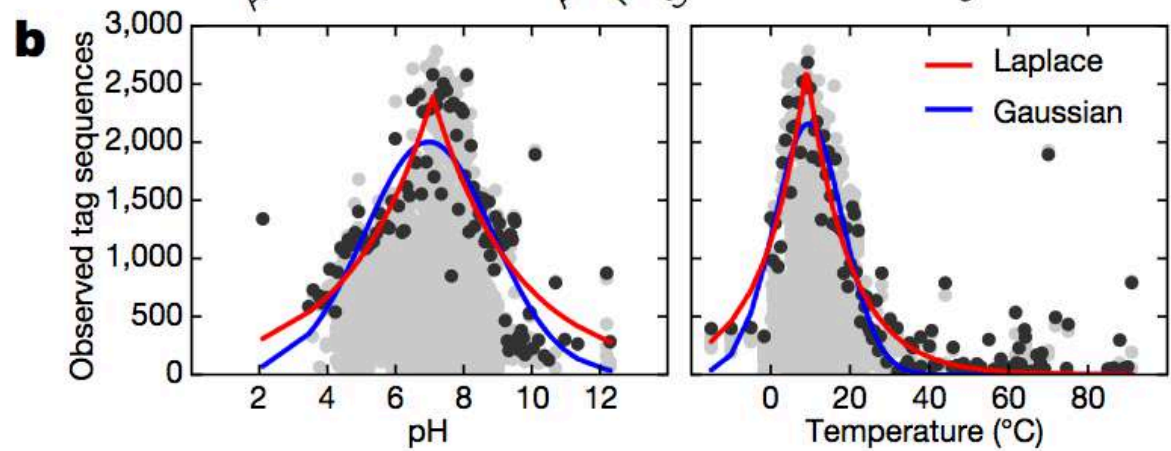
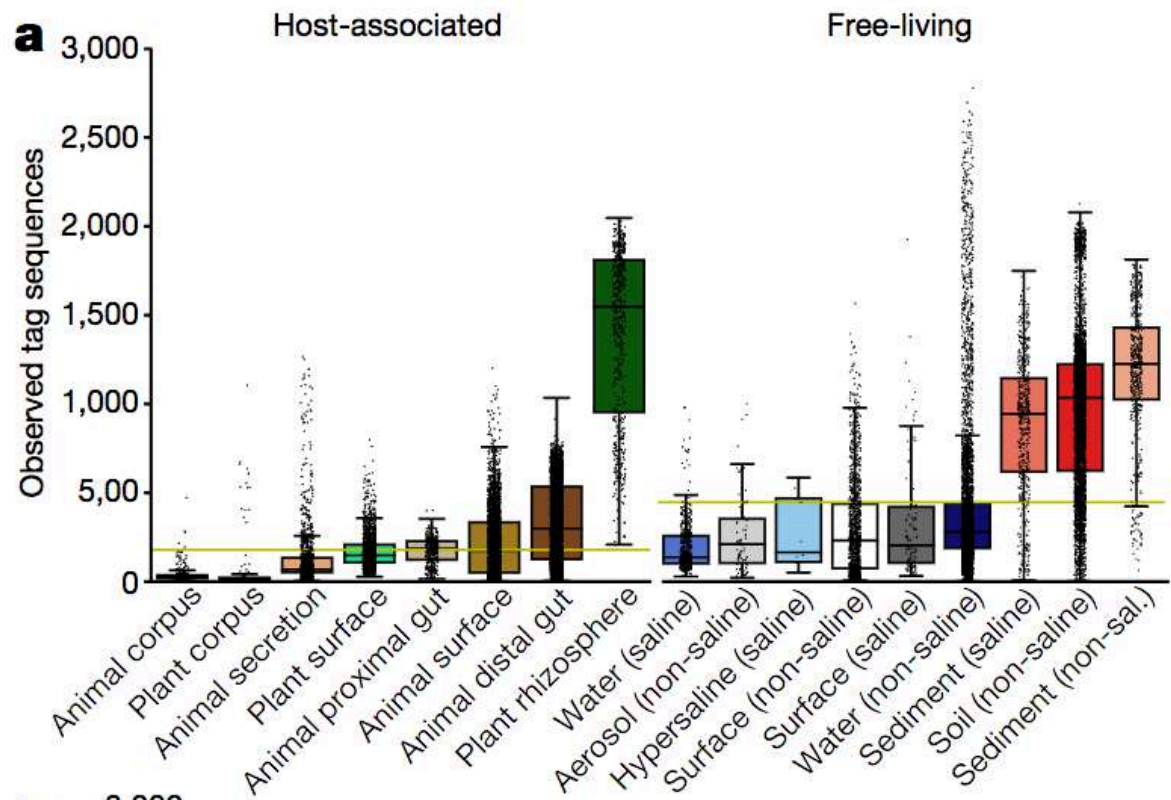
latitude range (Arctic  
Circle to Antarctica)

1 reference database of  
bacteria that reside on  
Planet Earth









Case studies – a really bad example

RESEARCH ARTICLE

# Microbiome restoration diet improves digestion, cognition and physical and emotional wellbeing

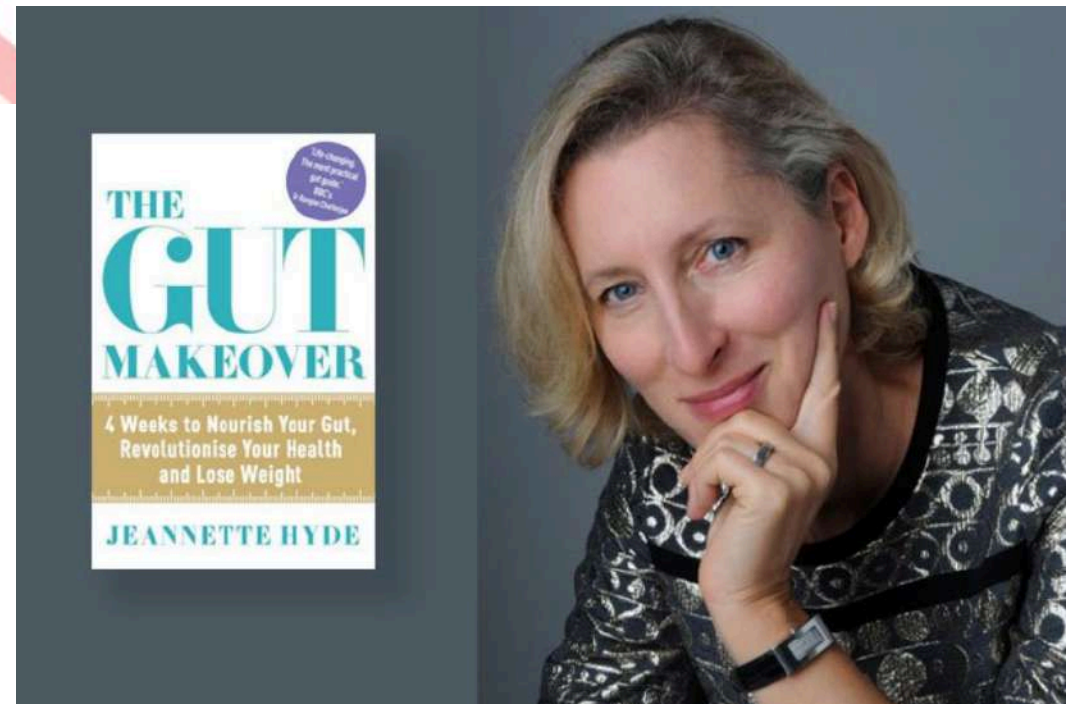
Kate Lawrence<sup>1\*</sup>, Jeannette Hyde<sup>2</sup>

**1** Department of Psychology, St Mary's University, Twickenham, United Kingdom, **2** Independent researcher, registered nutritional therapist BSc (Hons) Nutritional Therapy CNHC mBANT, Twickenham, United Kingdom

\* [Kate.lawrence@stmarys.ac.uk](mailto:Kate.lawrence@stmarys.ac.uk)



[https://www.buzzfeed.com/stephaniemlee/gut-makeover-microbiome-diet-retraction?utm\\_term=.haVkkqdo1Xx#.gaGqjrQPRw](https://www.buzzfeed.com/stephaniemlee/gut-makeover-microbiome-diet-retraction?utm_term=.haVkkqdo1Xx#.gaGqjrQPRw)





Therefore in summary, this "research article" has the following attributes:

1. **No objective data on health outcomes was collected**; the study presents only participant-reported subjective data,
2. **No objective data on treatment compliance was collected**; we do not know if the participants followed the diet nor to what extent,
3. **No objective data on treatment effect/mechanisms**: **The authors claim that the intervention changes the gut microbiome but failed to measure even a single parameter, microbe, molecule, or metabolite.**
4. **Participants were a positively self-selected "convenience sample" ripe and ready for a placebo response given their demonstrated positive expectations.**
5. **Impossible attribution, especially to the gut microbiome**: With no control group, no-one knows if the supposed "improvements" were due to the psychosocial intervention, the diet, the season, the natural history the non-disease being non-studied, chance; the attribution of supposed benefit to a mechanism involving the gut microbiome is not supported by any data in this publication.
6. **Short duration with no durability of effect**: No demonstrated durability to the supposed benefits; the study was of notably short duration (4 weeks),
7. **Wild attribution without any shred of evidence**: The treatment included 1) diet intervention and 2) psychosocial support and then **the authors attributed (without any supporting data whatsoever) the subjective/undocumented/purported benefits to 3) changes in the gut microbial composition.**



# Summary

## Amplicon sequencing

- inexpensive but very effective
- moving away from OTU to Amplicon Sequence Variants (ASV)
- longer amplicons with more resolution (strain level!) are coming
- Every step from sample collection to data deposit matters

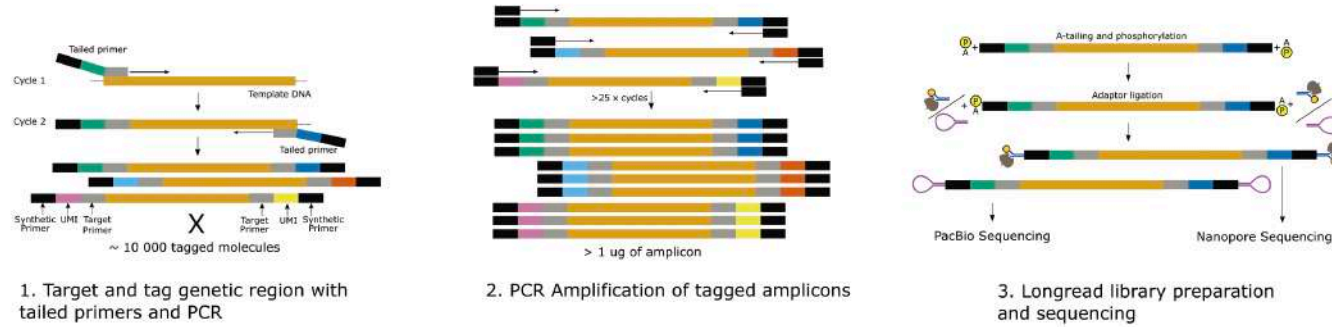
## Metagenomics

- expensive but has all the information you want (or not want) ; extremely powerful
- Metagenomics assembled genomes are being more complete
- **Metagenomics + HiC + Long reads : LOTS of resolved genomes!**
- Integration with other data is key to breakthrough
- Tremendous potential in this field ; but please do not oversell it

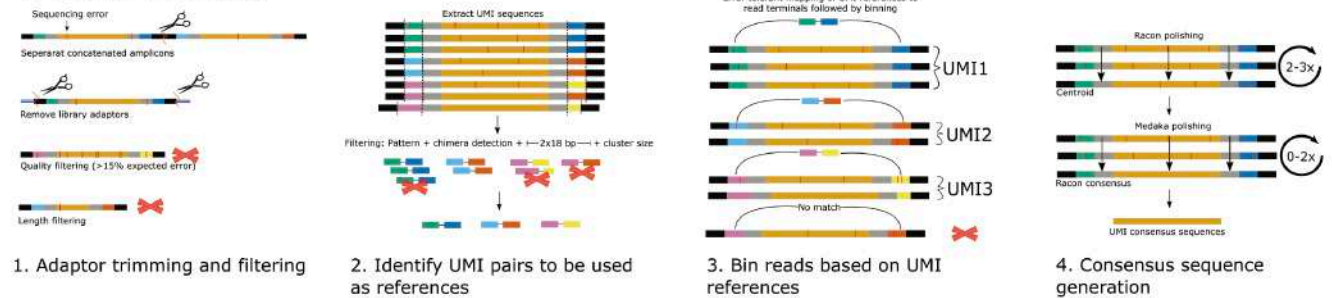
# Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing

Søren M. Karst<sup>1,\*</sup>, Ryan M. Ziels<sup>2,\*</sup>, Rasmus H. Kirkegaard<sup>1</sup>, Emil A. Sørensen<sup>1</sup>, Daniel McDonald<sup>3</sup>, Qiyun Zhu<sup>3</sup>, Rob Knight<sup>3,4,5,6</sup> and Mads Albertsen<sup>1</sup>

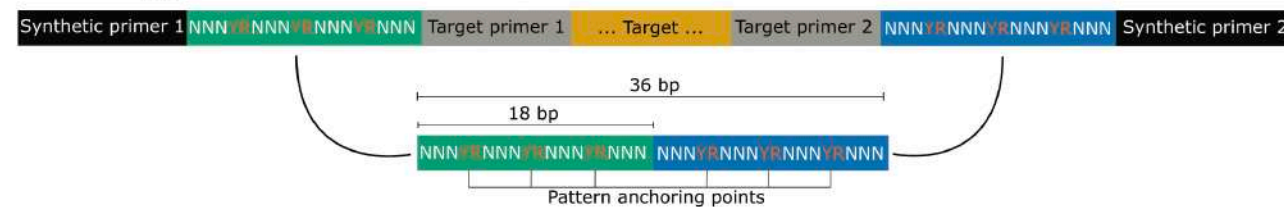
## a DNA Library Preparation and Sequencing



## b Data Processing



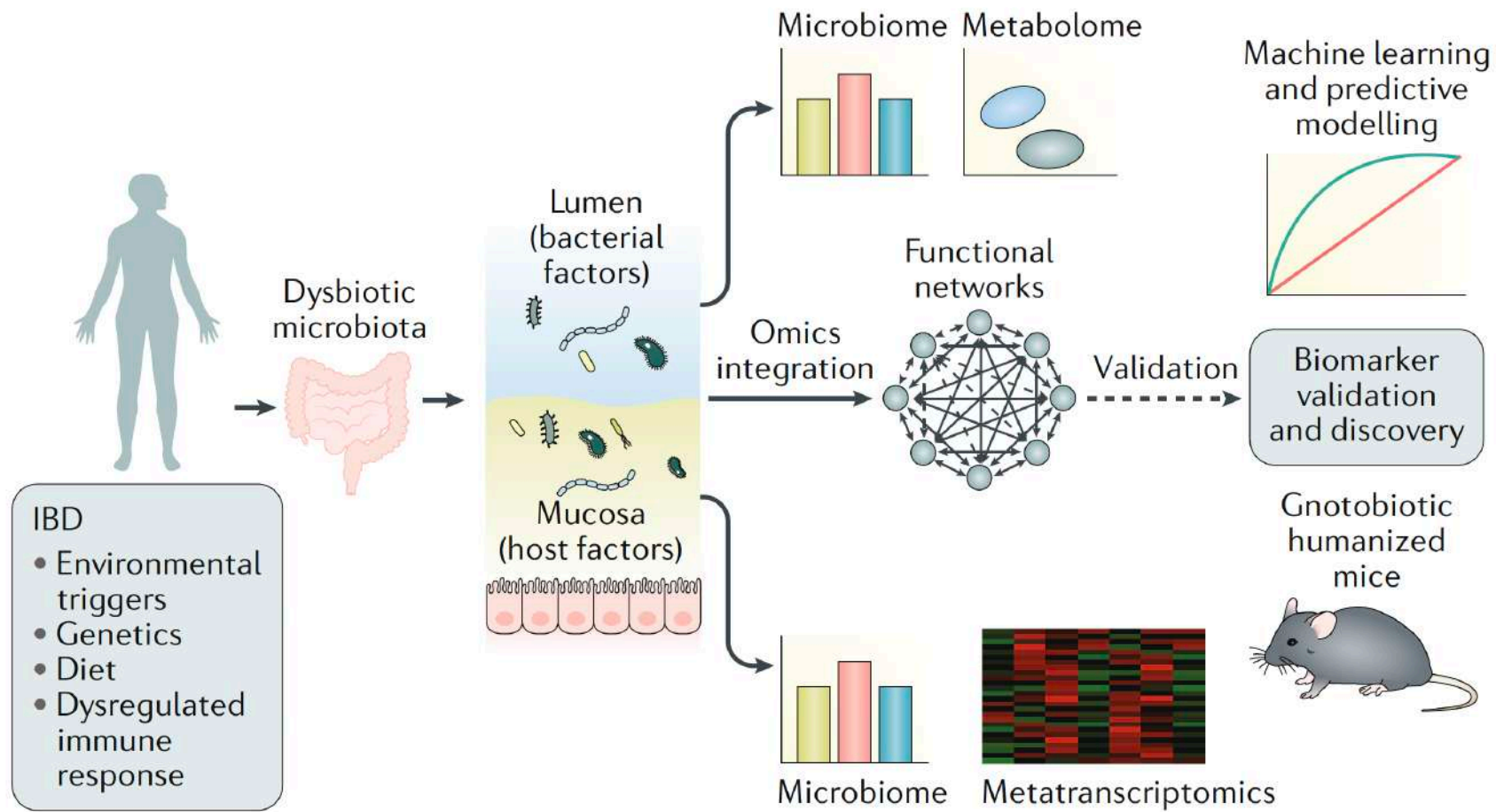
## c UMI tagged molecule



	Raw error rate (%)	Consensus (>Q40) error rate (%)	Chimera rate (%)
ONT UMI	11.63 (± 5.70)	0.0049 (± 0.0113)	0.017
PB CCS	13.01 (± 3.84)	0.0080 (± 0.0210)	1.940
PB UMI	0.46 (± 1.19)	0.0006 (± 0.0052)	0.022

	Deletion (>Q40) error rate (%)		Insertion (>Q40) error rate (%)		Mismatch (>Q40) error rate (%)	
	hp-	hp+	hp-	hp+	hp-	hp+
ONT UMI	0.0002	0.0093	0.0028	0.0006	0.0007	0.0010
PB CCS	0.0005	0.0229	0.0016	0.0011	0.0020	0.0021
PB UMI	0.0000	0.0003	0.0001	0.0001	0.0005	0.0006



**Fig. 1 | Conceptual model for multi-omics integration towards mechanistic biomarker discovery.** Multi-omics (including metabolome, microbiome and transcriptome data) are collected from patients with IBD and integrated to identify personalised functional signatures using complex and comprehensive network analysis. Validation of biomarker signature for precision medicine could be achieved through machine-learning approaches and studies in gnotobiotic humanized mice.

# References

<https://www.notion.so/References-papers-links-in-start-learning-genomics-b7e57b28e9194bb29a02f483e0b894ad>