Amplicon / Metagenomics

Isheng Jason Tsai

Introduction to NGS Data and Analysis
Lecture 10

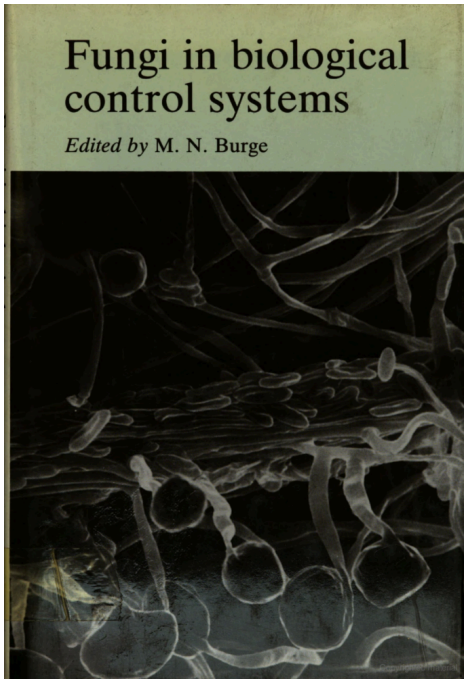
Lecture Outline

- Introduction
  - Amplicon sequencing != metagenomics
  - 16S sequencing
  - Metagenomics

- Concepts
- Amplicon Sequencing
- Metagenomics

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?

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

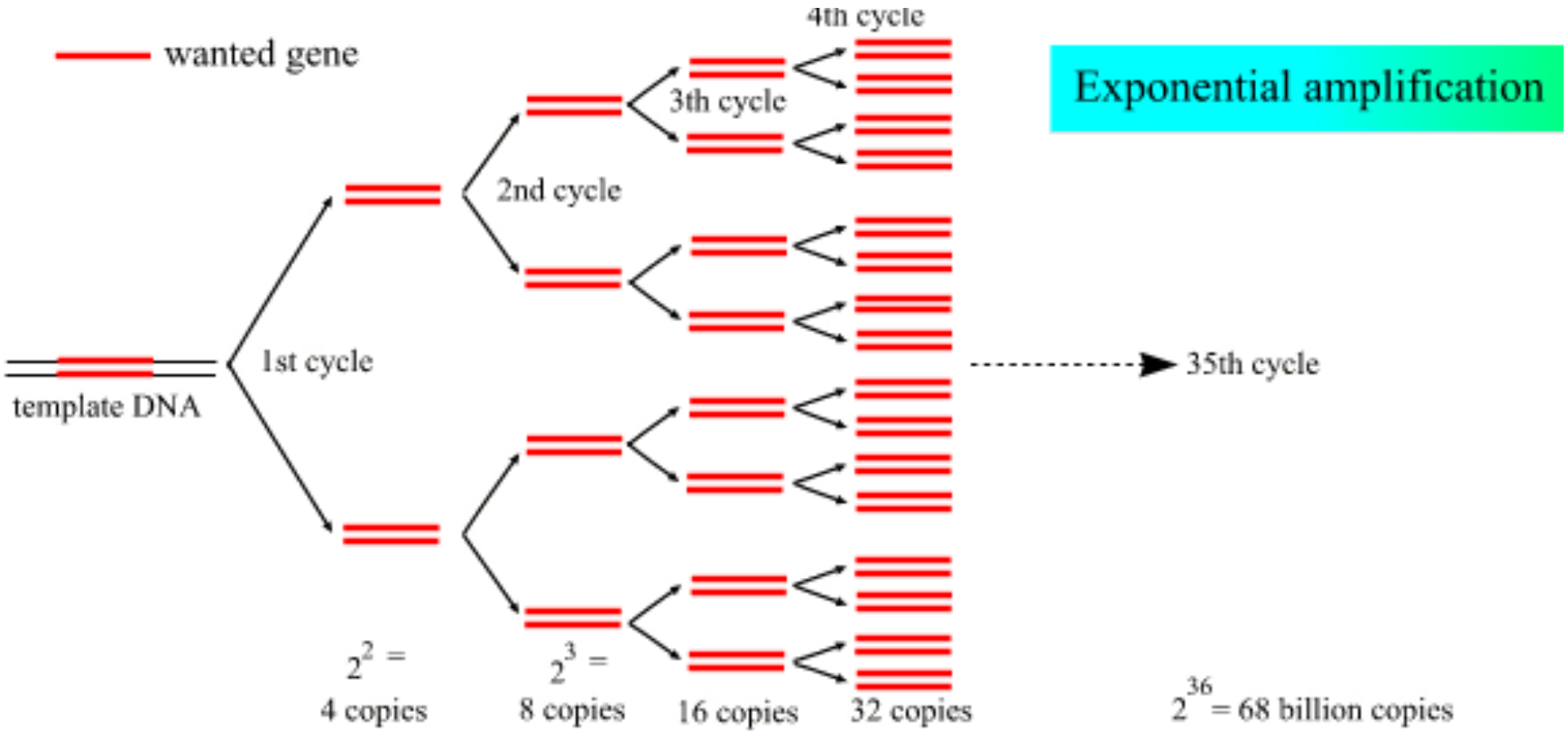
Jo Handelsman[1], Michelle R Rondon[1], Sean F Brady[2], Jon Clardy[2] and Robert M Goodman[1]

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

What is amplicon sequencing?

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



And then what is the metagenome?

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

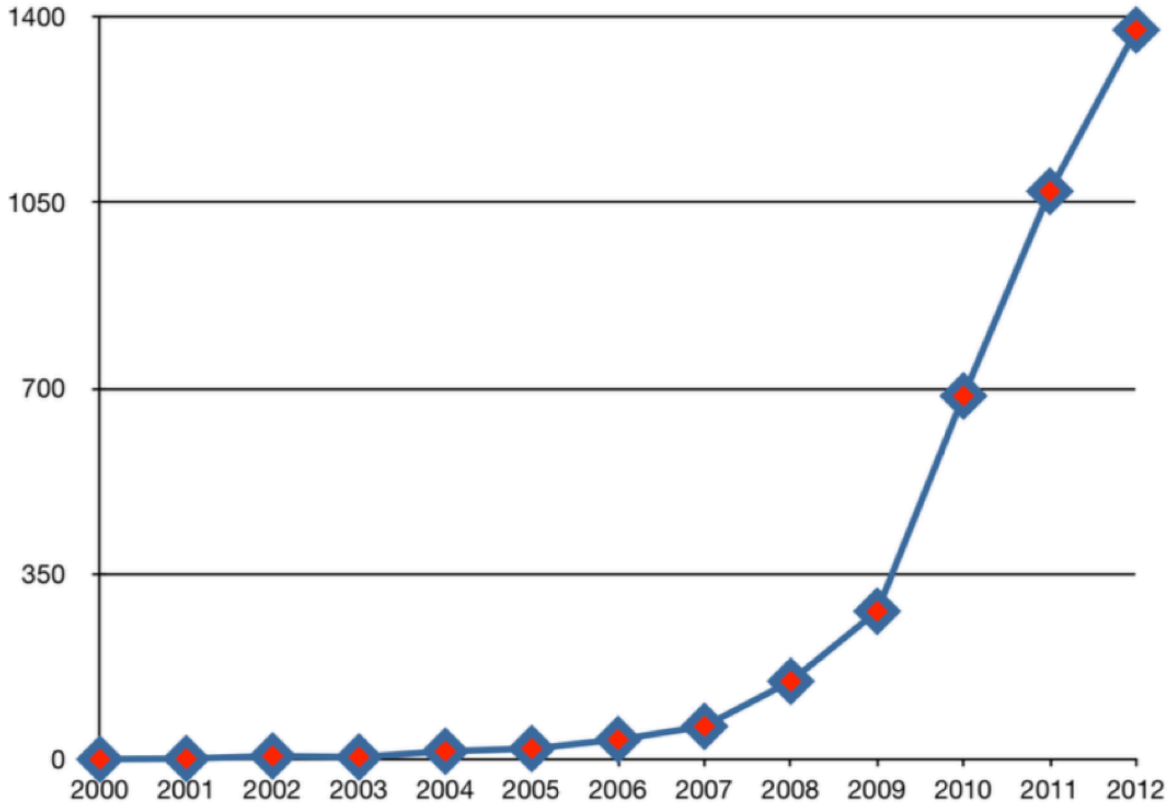
Review

Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities

Kevin Chen*, Lior Pachter*

Metagenomics is the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species. The field has its roots in the culture-independent retrieval of **16S rRNA** genes, pioneered by Pace and colleagues two decades ago.

Pubmed hits for “Microbiome”

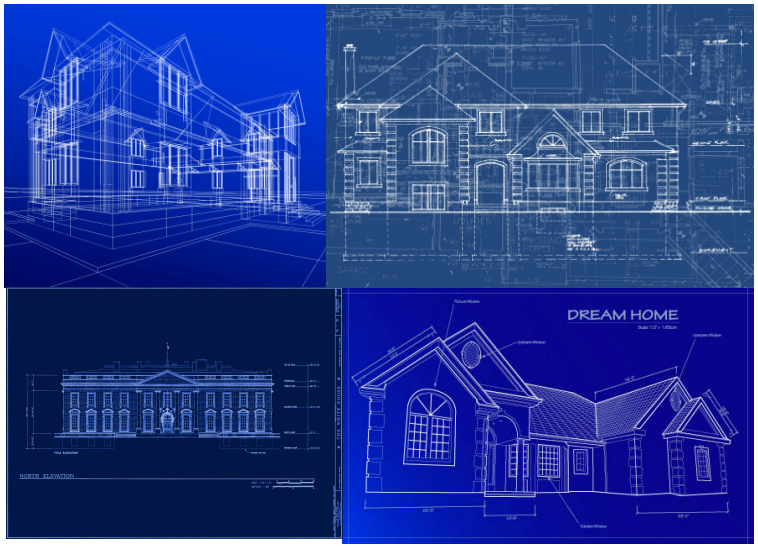


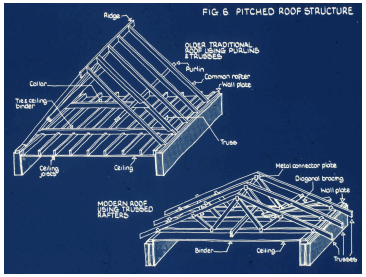
Jonathan Eisen, Slideshare

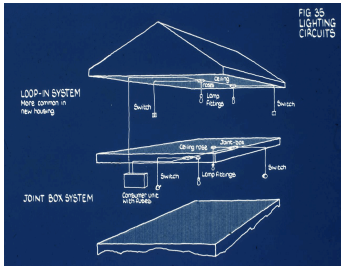
Genomics

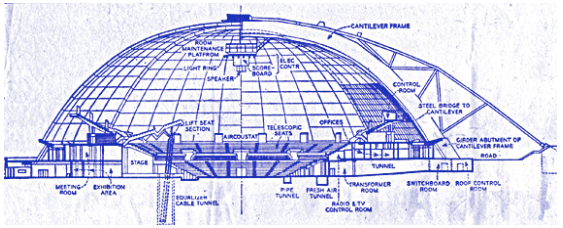
Genome = Parts list of a single genome

Metagenomics and amplicon sequencing









Capture and sequencing of particular gene (16S)

Metagenomics ≠ Amplicon sequencing

Metagenomics is undergoing a crisis

Please don't make things worse ☺

- Crisis 1
  - **The correlation/causation fallacy**. For example….
  - Patients with type II diabetes have a different gut microbiome compared to healthy patients
  - Does the microbiome cause diabetes?
  - Or do they have a different microbiome because they have diabetes? (therefore different diet)
- Crisis 2
  - A lot of people want to do it, but don't know how
  - Errors, bad experimental design, incorrect conclusions

Mick Watson, Slideshare

Basic Purpose

Characteristics of (microbial) community

Who are they?
Where do they come from?
Are their similarities (at what level)
between communities
of different conditions
of similar conditions?
within a community?
What are they doing?
How are they doing?

Useful terminology

Table 1. Terms relevant to core microbiomes.

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. Requires qualifiers for locality and habitat of interest.	Human microbiome, Earth microbiome, Lake Erie microbiome, Soil microbiome	Lederberg and McCray (2001)
Core microbiome	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 **genes** 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	

Shade and Handelsman (2012)

Applications

What have metagenomics been used for?

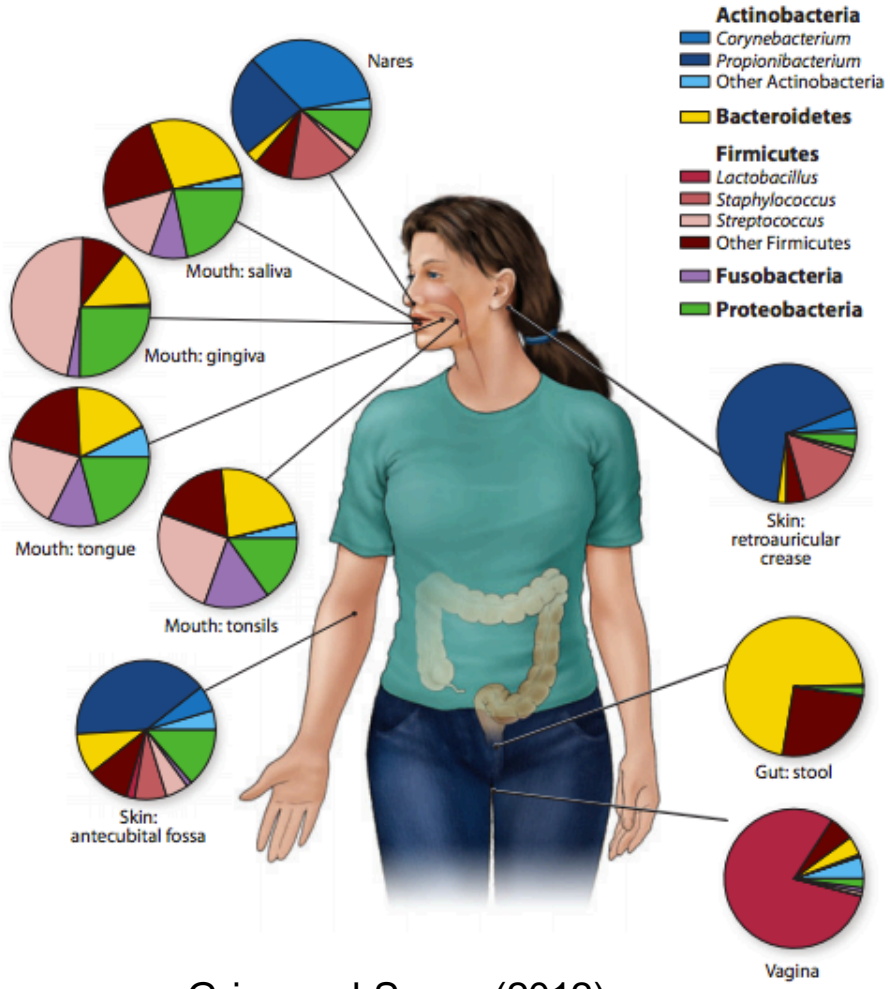
Exploration and categorisation

Rusch *et al.*, 2007 **Plos Biology**

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

Qin *et al.*, 2010 **Nature**

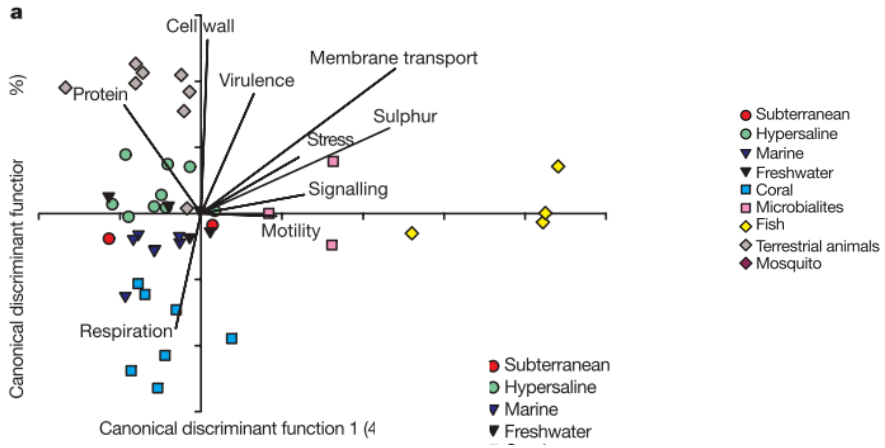
- 127 Human gut metagenomes
- 600 Gbp sequence (200 x Human genomes)
- 3.3 million genes identified
- Minimal gut metagenome defined



Grice and Segre (2012)

What have metagenomics been used for?

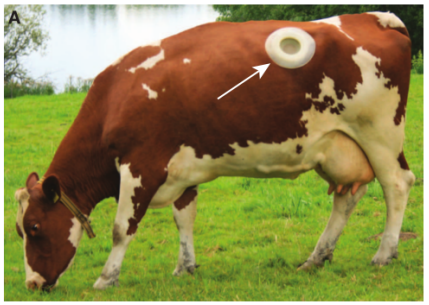
Comparative



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

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

Specific functions

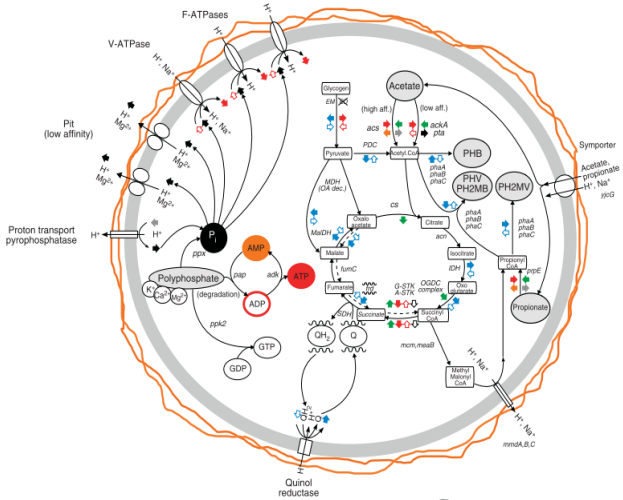


Hess *et al.*, 2011 **Science**

- Identified 27.755 putative carbohydrate-active genes from a cow rumen metagenome
- Expressed 90 candidates of which 57% had enzymatic activity against cellulosic substrates

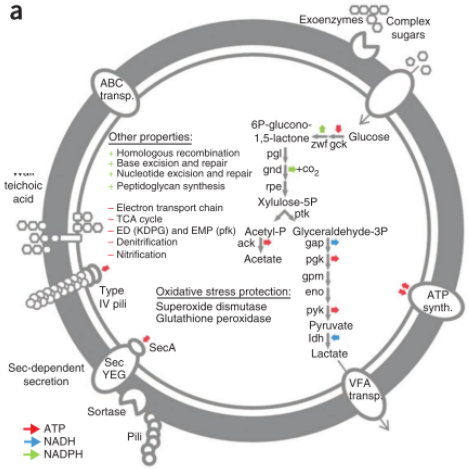
What have metagenomics been used for?

Extracting genomes



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

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



Albertsen *et al*., 2013 **Nat. Biotechnol.**

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

Concept: OTU (Operational Taxonomic Unit)

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

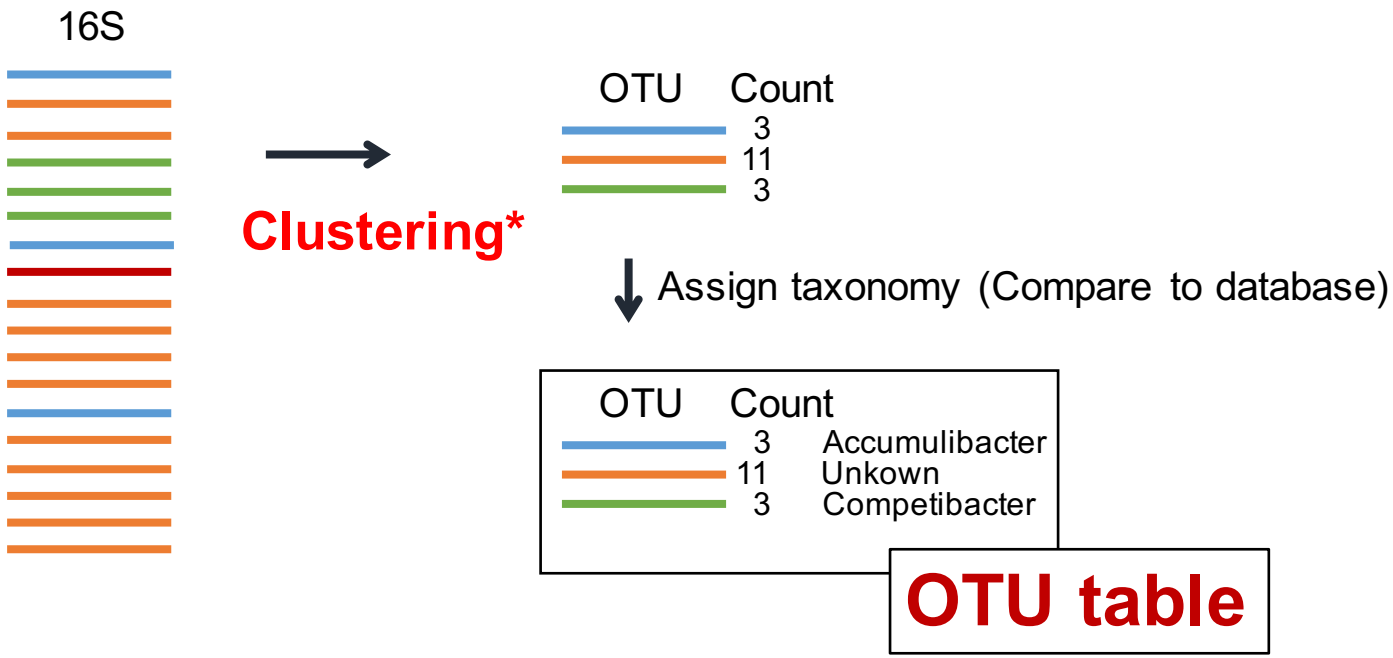
Weaknesses

- Limited resolution
- Logically inconsistent definition

Slide of Aaron Darlin

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

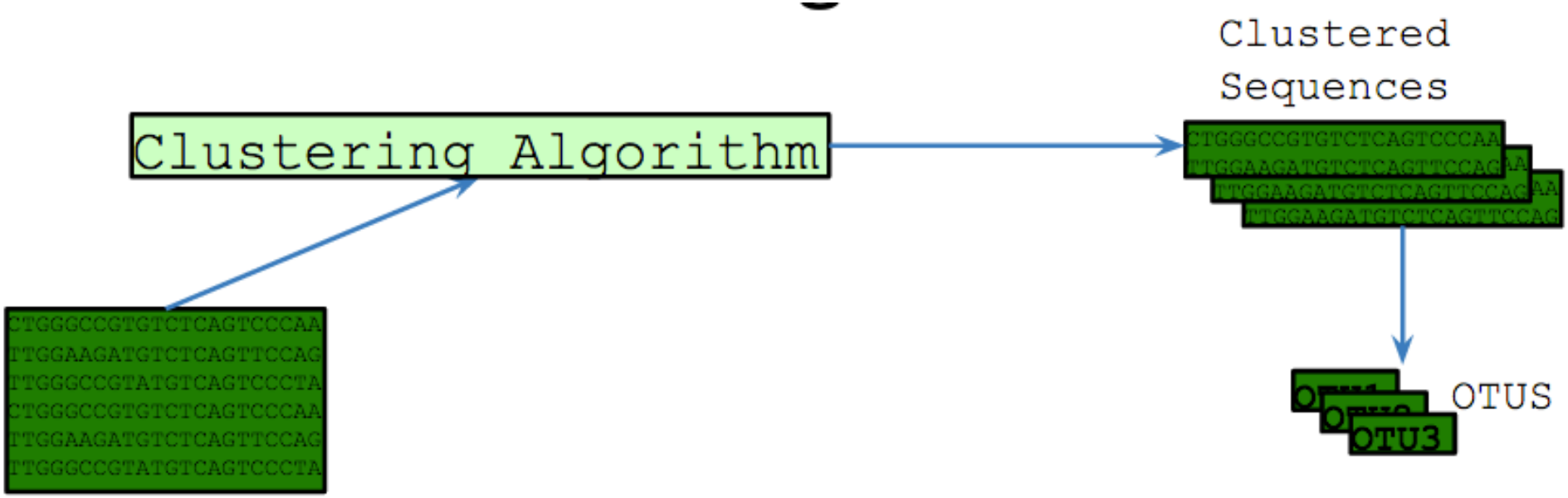
- Reads are clustered based on similarity to one another.

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

http://qiime.org/tutorials/otu_picking.html

De novo OTU picking



Pros

- All reads are clustered

Cons

- Not parallelizable
- OTUs may be defined by erroneous reads

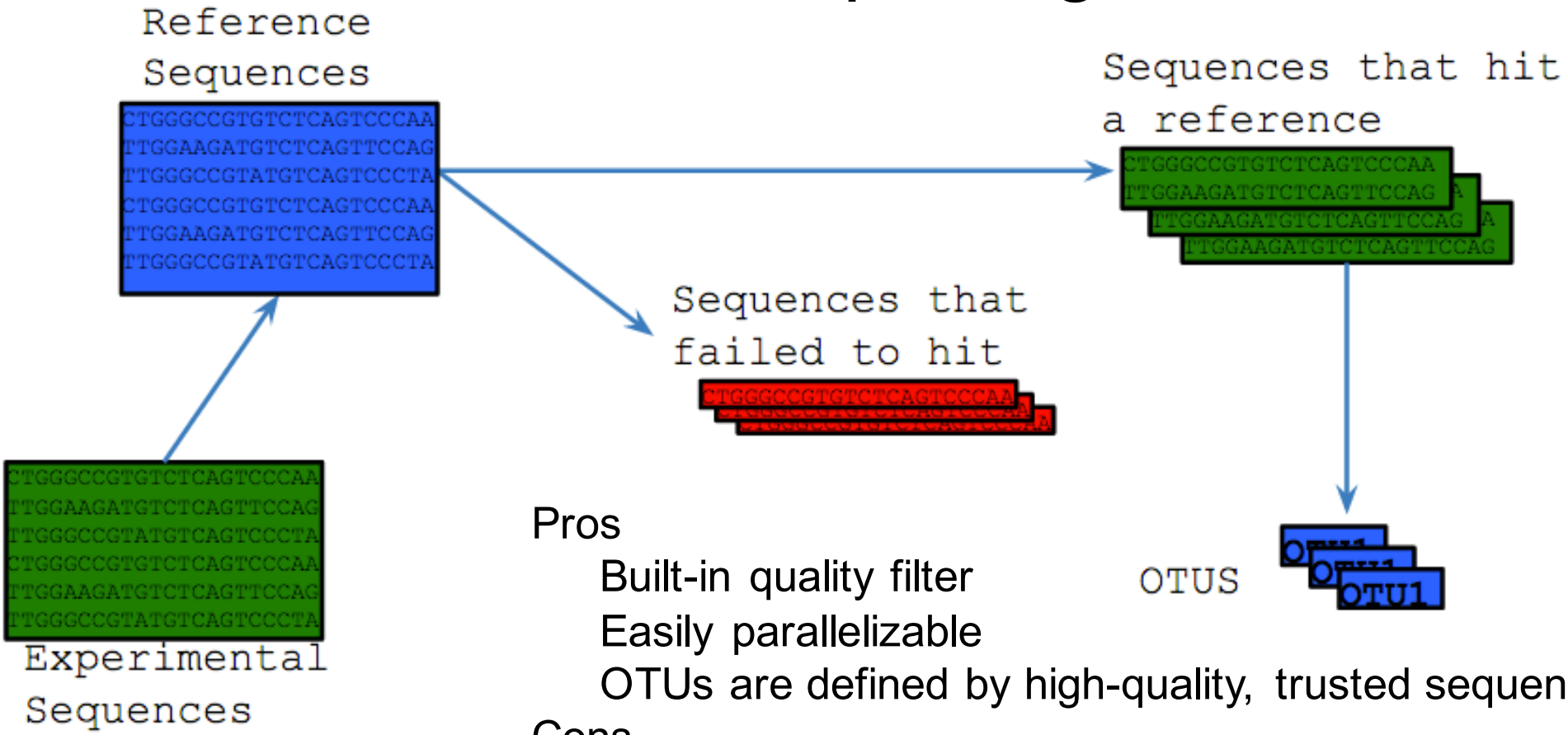
Slide of John Chase

De novo OTU picking

- You **must** use if:
  - You do not have a reference sequence collection to cluster against, for example because you're working with an infrequently used marker gene.
- You **cannot** use if:
  - You are comparing non-overlapping amplicons, such as the V2 and the V4 regions of the 16S rRNA.
  - You working with very large data sets, like a full HiSeq 2000 run. (Technically you can, but it will be *really* slow.)

pick_de_novo_otus.py
http://qiime.org/tutorials/tutorial.html

Closed-reference OTU picking



Pros

- Built-in quality filter
- Easily parallelizable
- OTUs are defined by high-quality, trusted sequences

Cons

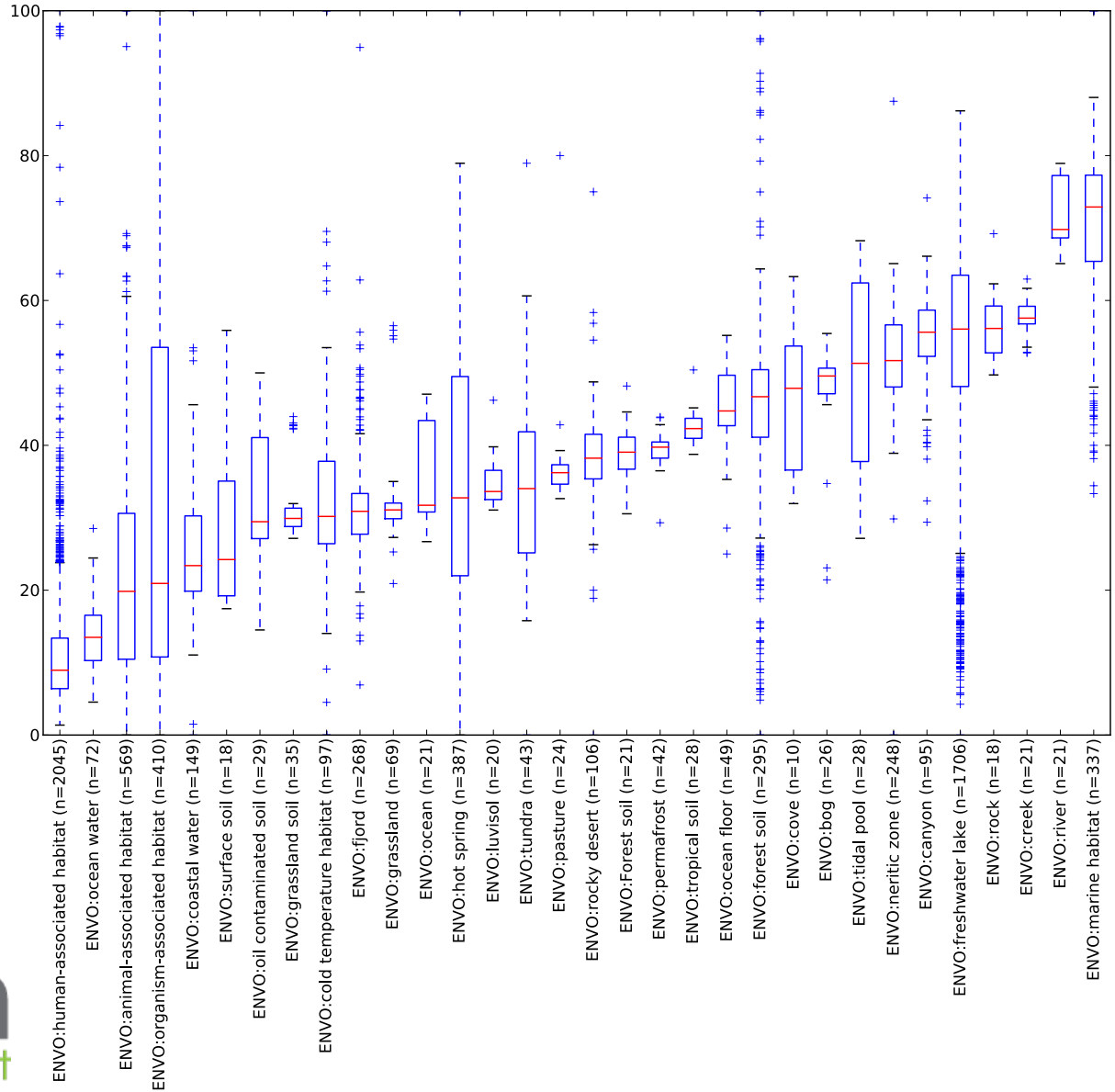
- Reads that don't hit reference dataset are excluded, so you can never observe new OTUs

Slide of John Chase

Closed-reference OTU picking

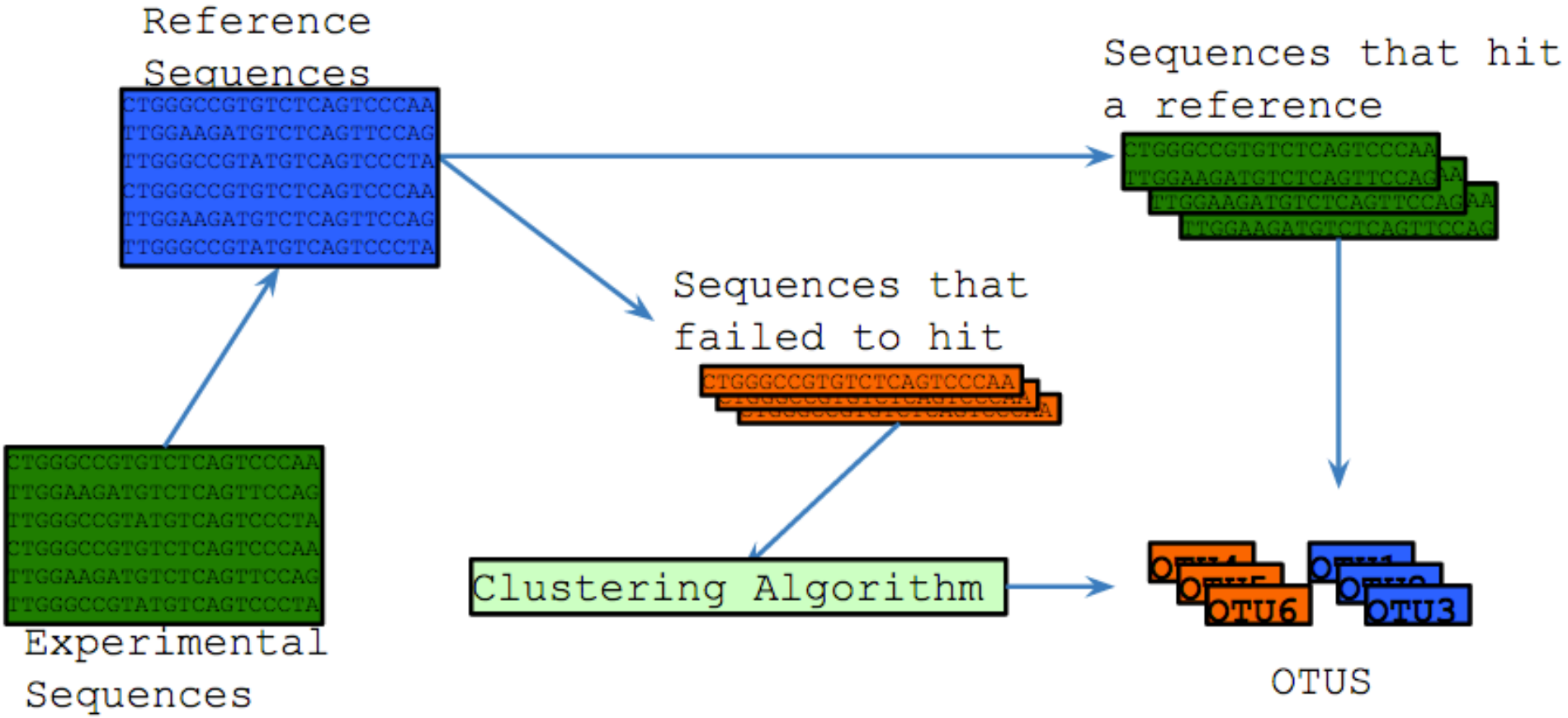
- You **must** use if:
  - You are comparing non-overlapping amplicons, such as the V2 and the V4 regions of the 16S rRNA. Your reference sequences must span both of the regions being sequenced.
- You **cannot** use if:
  - You do not have a reference sequence collection to cluster against, for example because you're working with an infrequently used marker gene.

Percentage of reads that do not hit the reference collection, by environment type.



earth microbiome project

Open-reference OTU picking

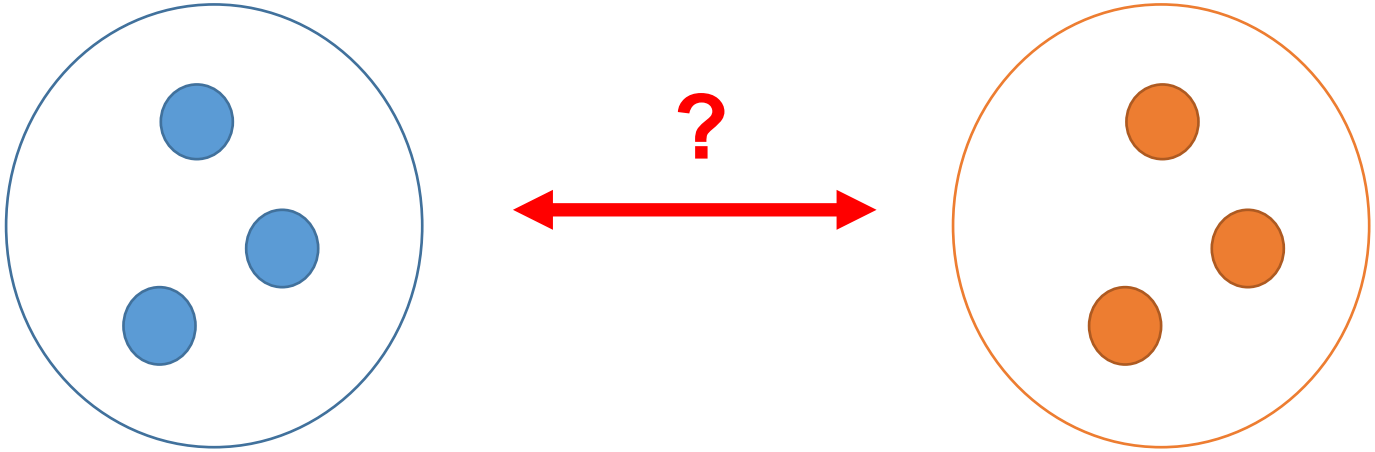


Slide of John Chase

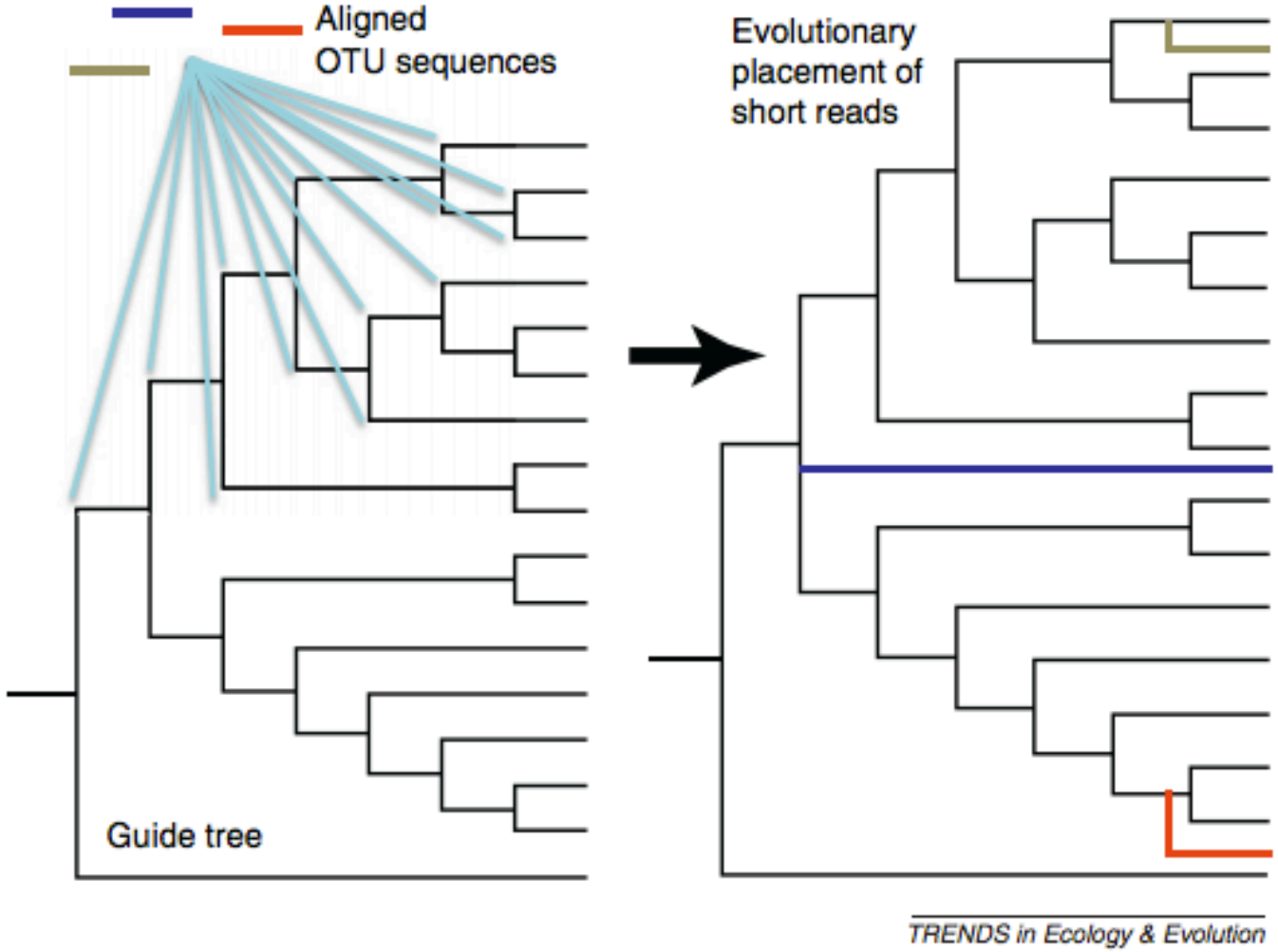
Open-reference OTU picking

- Pros
  - All reads are clustered
  - Partially parallelizable
- Cons
  - Only *partially* parallelizable
  - Mix of high quality sequences defining OTUs (i.e., the database sequences) and possible low quality sequences defining OTUs (i.e., the sequencing reads)

Assigned OTUs -> Loss of information



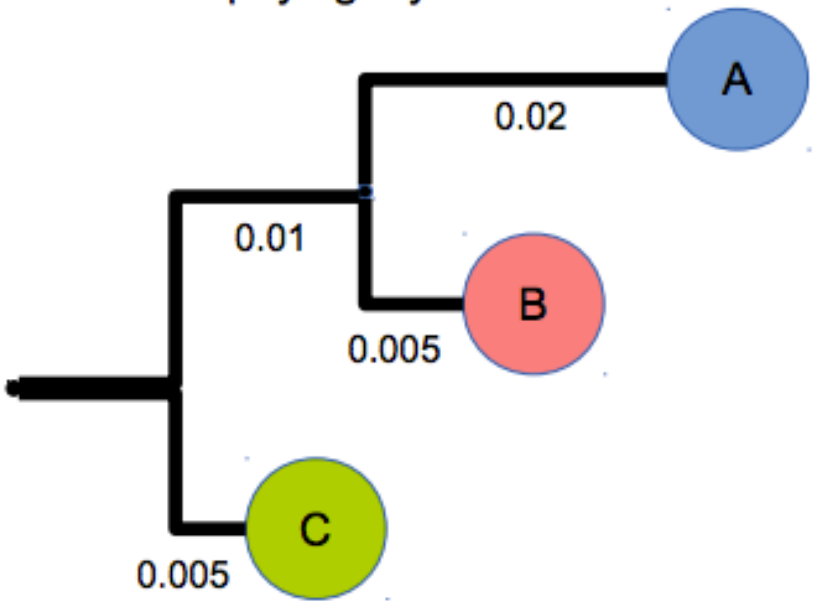
OTU relationship using phylogenetics



Bik et al (2011)

Logical inconsistency: OTUs at 97% ID

Assume the true phylogeny:



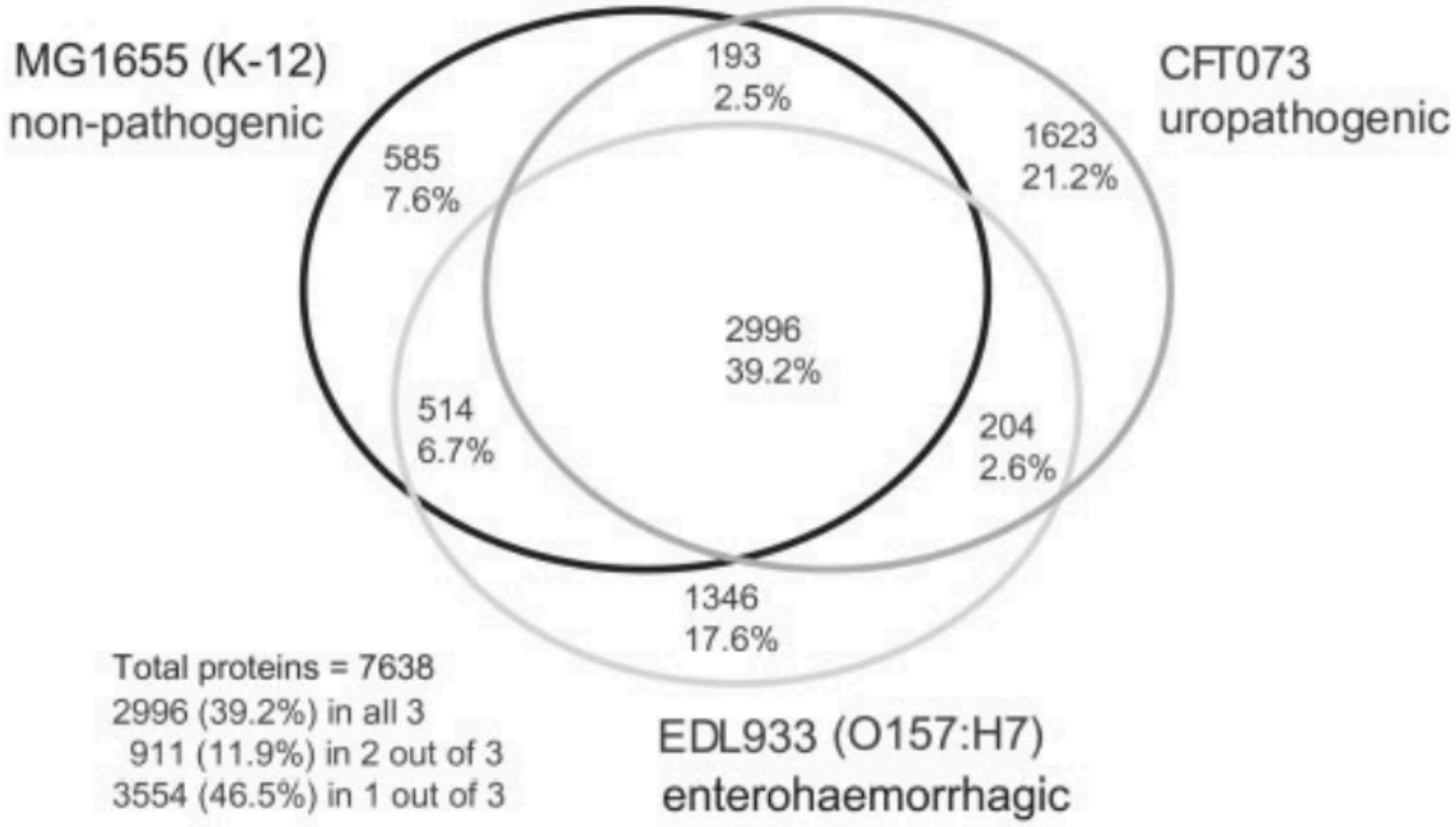
A, B > 97% identity
B, C > 97% identity
A and C not > 97% ID

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??

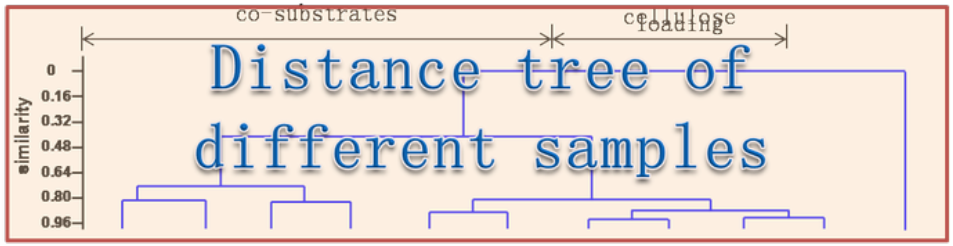
Slide of Aaron Darlin

Same species (16S): Different genomes



Welch et al (2002)

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%	40.620%	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.[illegible]%	0.626%	[illegible]	[illegible]42%	[illegible]22%	0.870%	1.192%	0.336%	0.582%	0.000%
11	3.686%	2.[illegible]%	[illegible]%	[illegible]%	0.218%	[illegible]	[illegible]	[illegible]	[illegible]	0.000%	0.000%
12	1.862%	0.904%	[illegible]%	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%	[illegible]06%	[illegible]51%	0.008%	0.006%	0.004%	0.008%	0.010%
15	10.494%	6.218%	8.710%	4.958%	[illegible]38%	[illegible]56%	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%

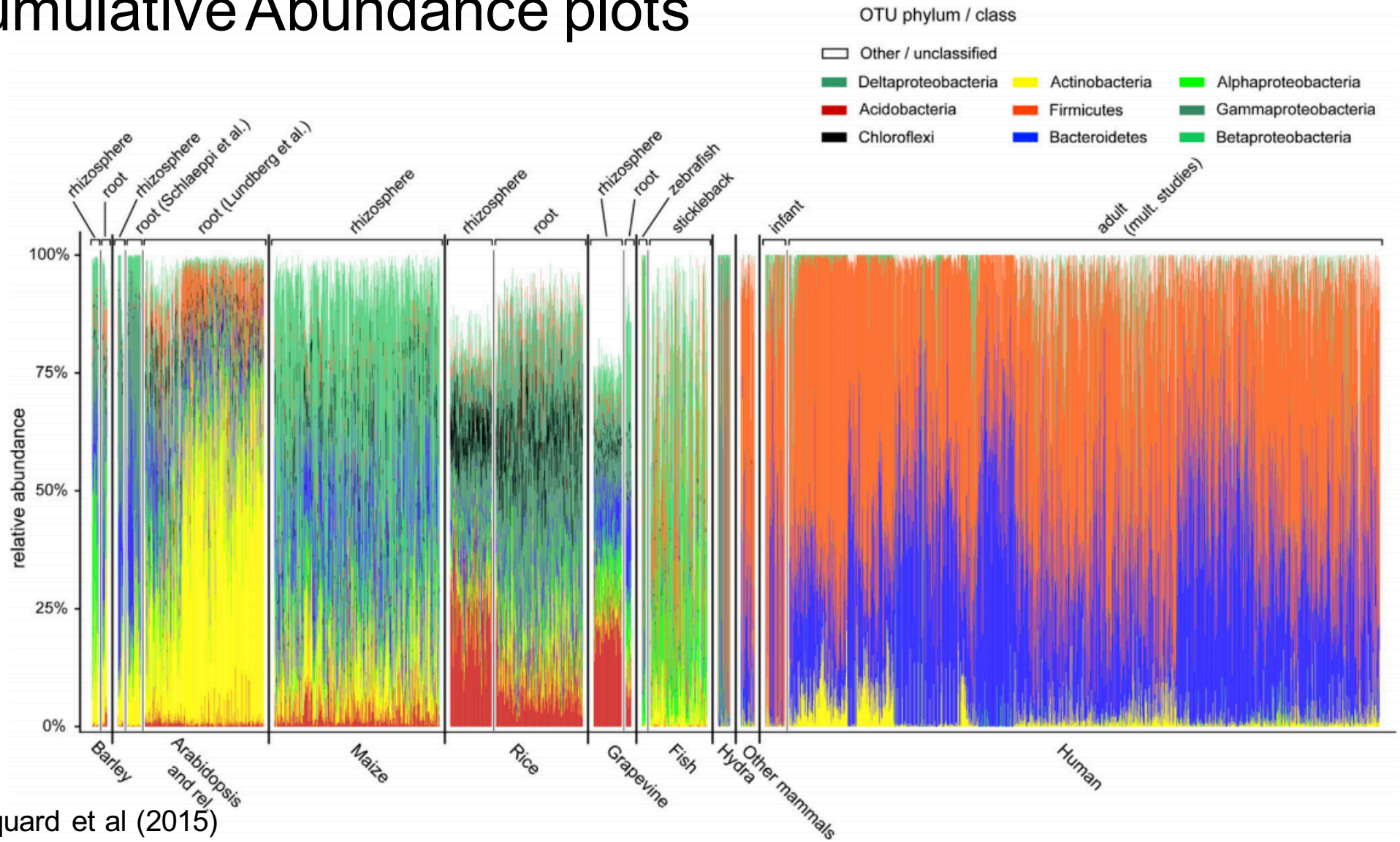
Top OTU abundance matrix

OTU	classification
1	*Clostridium sp.* 4-2a
2	*Clostridium sp.* 4-2a
3	*Clostridium sp.* 4-2a
4	*Clostridium sp.* 4-2a
5	*Anoxybacillus sp.* HT8
6	*Clostridiaceae bacterium* 37-7-2Cl
7	*Clostridium sp.* YN5
8	*Clostridium sp.* YN6
9	*Clostridium sp.* YN7
10	*Clostridium sp.* YN8
11	*Thermoanaerobacterium xylanolyticum* LX-11
12	*Thermoanaerobacterium xylanolyticum* LX-11
13	*Thermoanaerobacterium xylanolyticum* LX-11
14	*Thermoanaerobacterium aotearoense* SCUT27
15	*Thermoanaerobacterium xylanolyticum* LX-11
16	*Thermoanaerobacterium aotearoense* SCUT27
17	*Thermoanaerobacterium xylanolyticum* LX-11
18	*Thermoanaerobacterium aotearoense* SCUT27
19	*Thermoanaerobacterium xylanolyticum* LX-11
20	*Thermoanaerobacterium aotearoense* SCUT27
21	*Thermoanaerobacterium saccharolyticum* D22
22	*Thermoanaerobacterium thermosaccharolyticum* CECT5853T
23	*Thermohydrogenium kirishiense* DSM11055T
24	*Uncultured Bacillus sp.* clone 93
25	*Uncultured Bacillus sp.* clone 94
26	*Mycobacterium gilvum* B4
27	*Nocardioides terrae* VA15
28	*Anaerolinea thermolimosa* IMO-1
29	*Uncultured bacterium* clone 49c
30	*Clostridium cellobioparum* JCM 1422

OTU classification

http://www.shuixia100.com/my-blog/mothur-tutorial-1

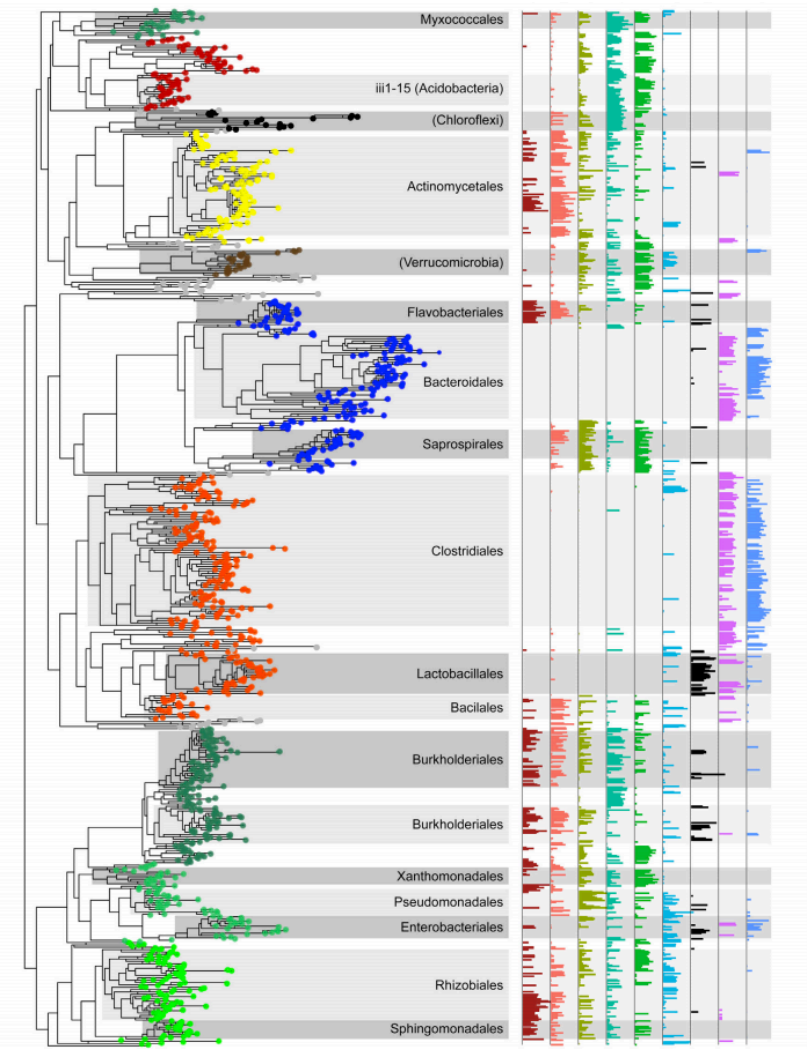
Cumulative Abundance plots



Hacquard et al (2015)

Phylogenetic Analysis of OTU abundances

Relationship between OTUs



How do we compare between different samples?

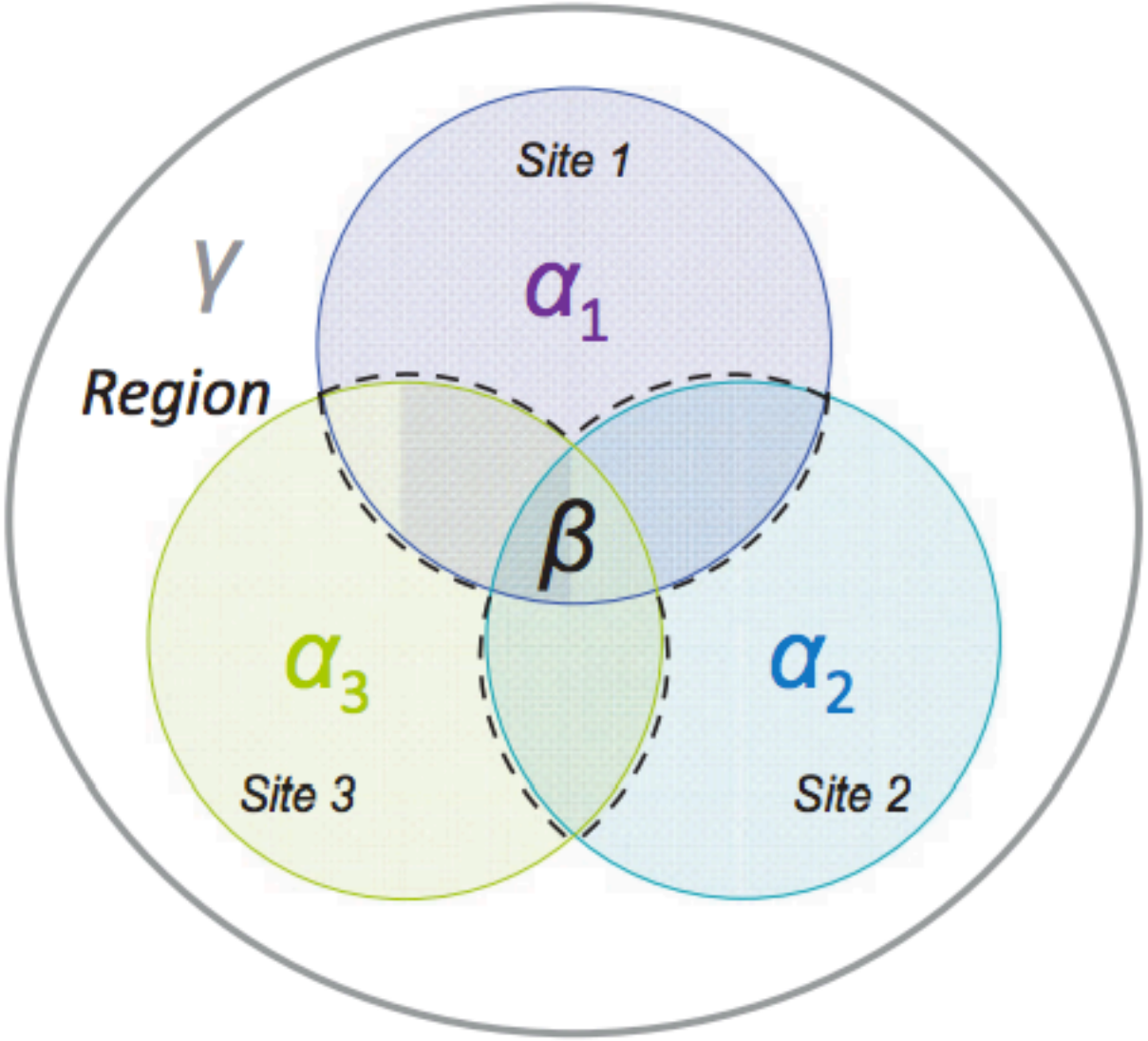
Hacquard et al (2015)

Concept: Diverisity 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"



Measures of biodiversity

Zinger et al (2012)

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).

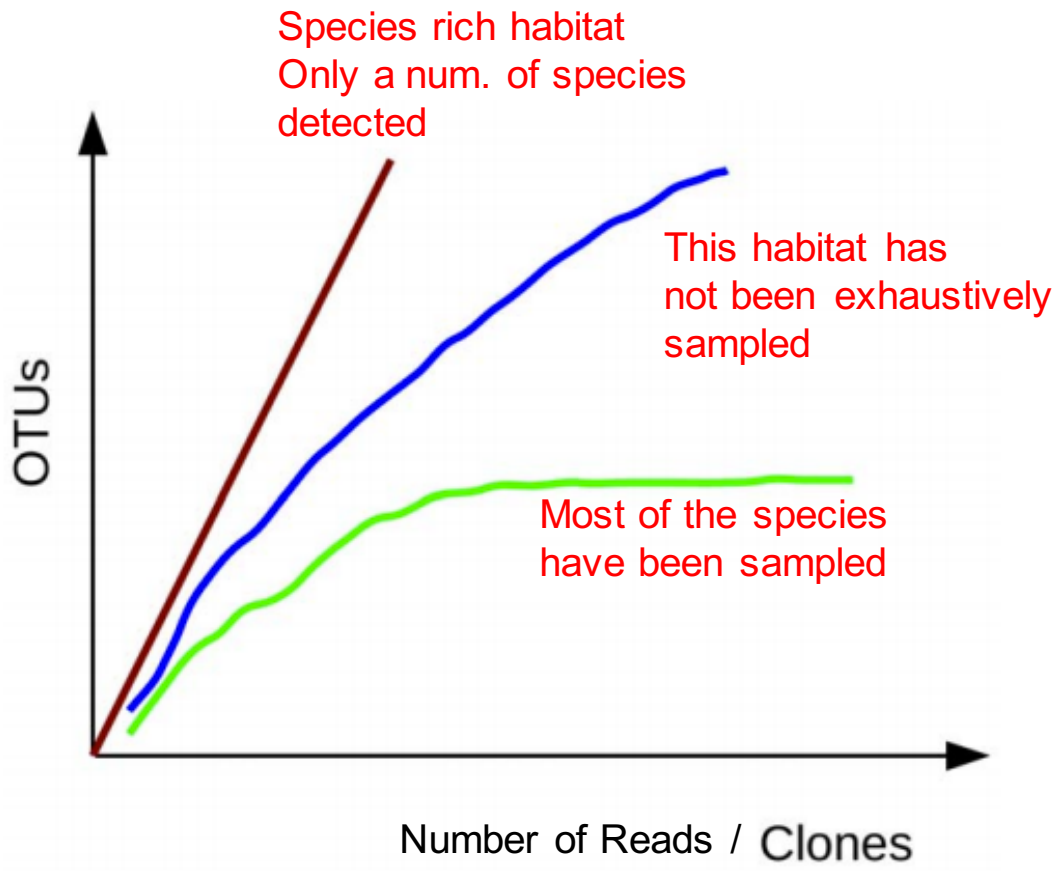
Beta diversity consists in determining the difference in diversity or community composition between two or more locations or samples (i) by considering species composition only, and use incidence data with associated metrics such as Jaccard or Sorensen similarity indices or (ii) by taking species relative 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



Wooley et al (2010)

Alpha diversity

a measure of the diversity within a single sample

Types of alpha diversity

Total # of species = **richness**

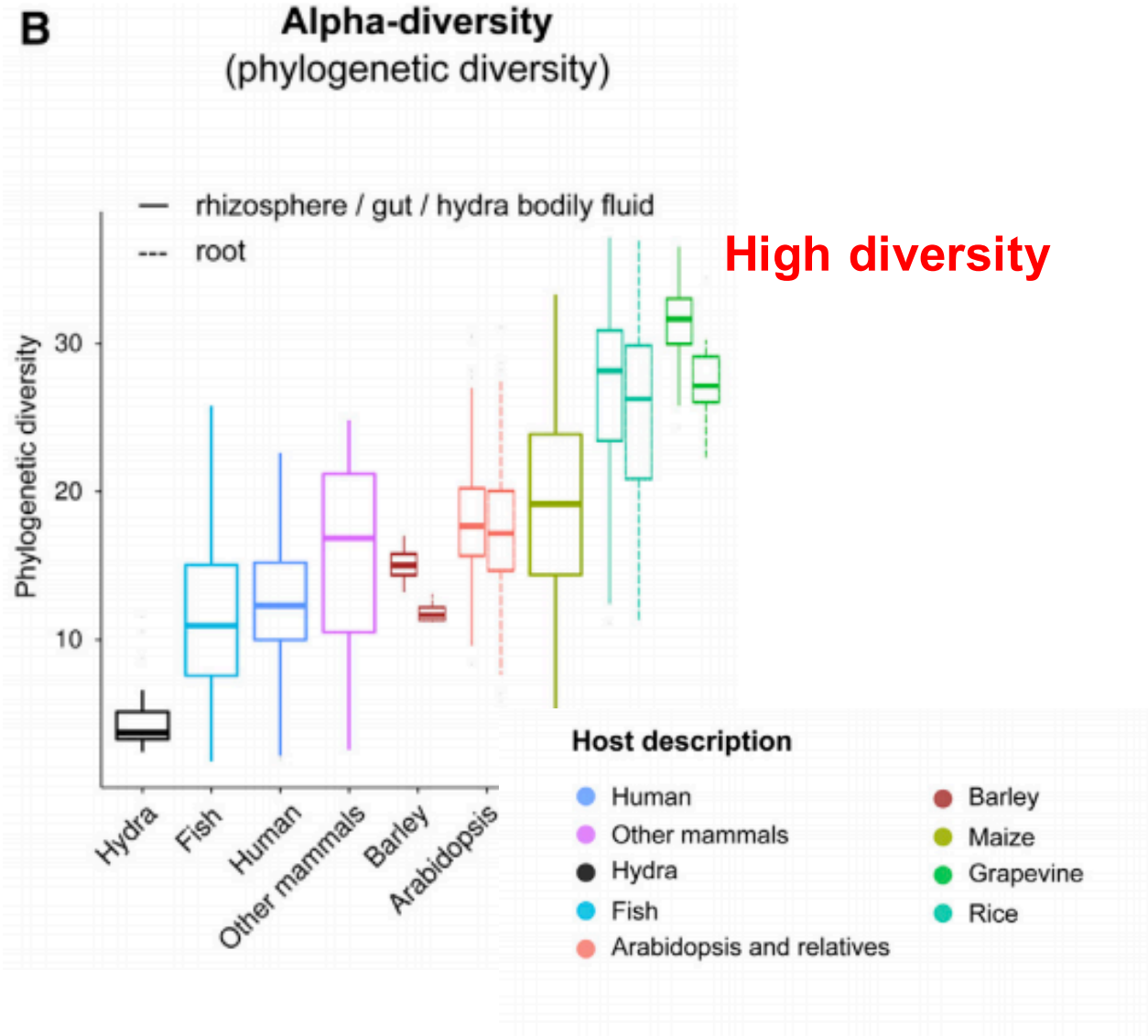
How many OTUs?

Total # of genes = genetic richness

Phylogenetic diversity of genes = genetic PD

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

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



High diversity

Low

Hacquard et al (2015)

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 of PCoA analysis

Beta diversity

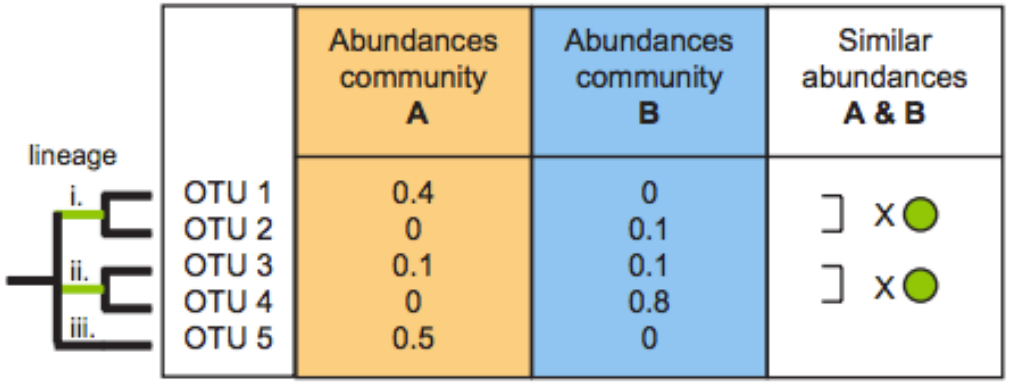
A. Membership:
shared OTU occurences across communities
1 = present, 0 = below detection

List of observed OTUs	Occurences in community **A**	Occurences in community **B**	Shared occurences **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

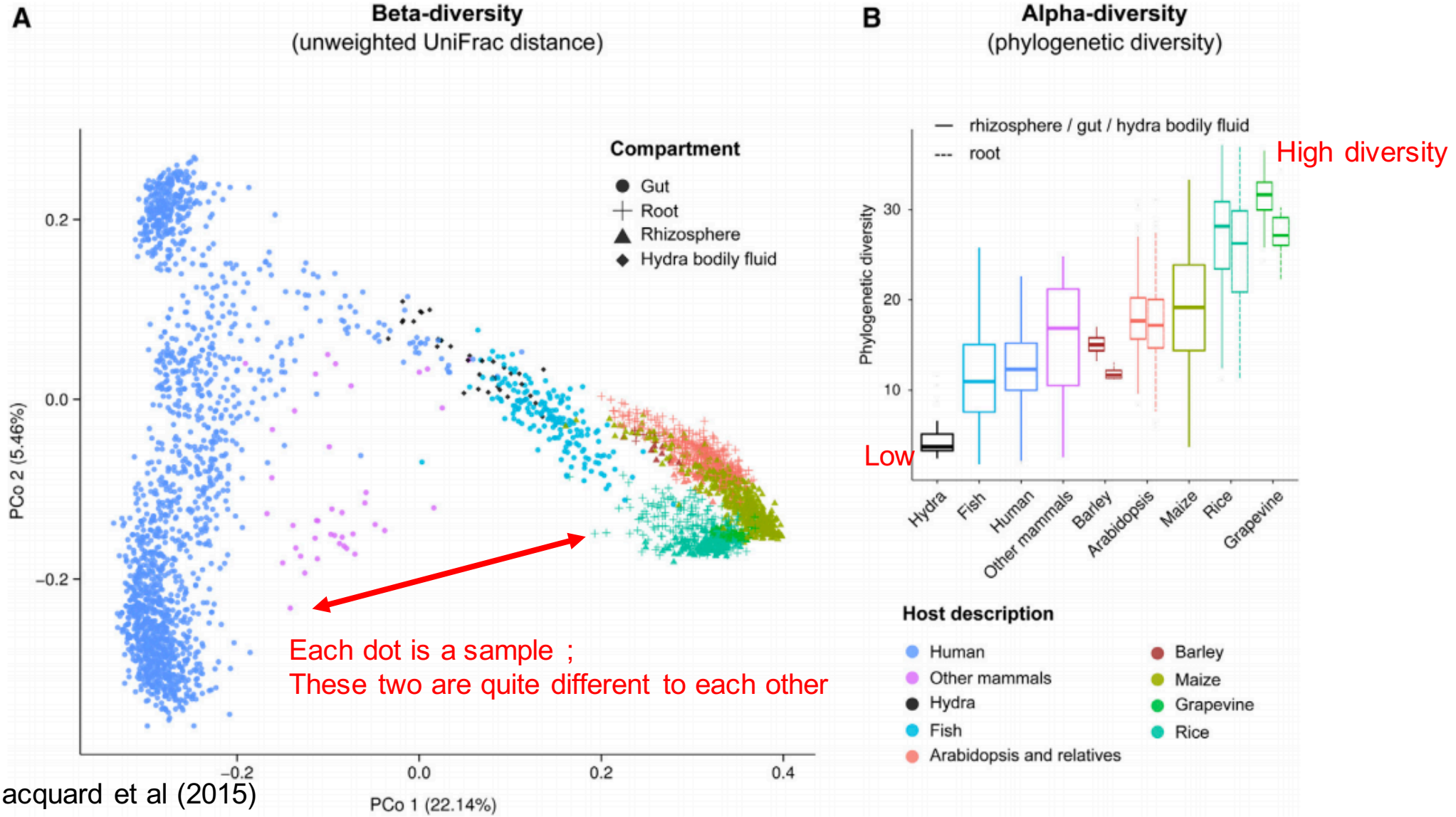
B. Composition:
similar OTU abundances across communities

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



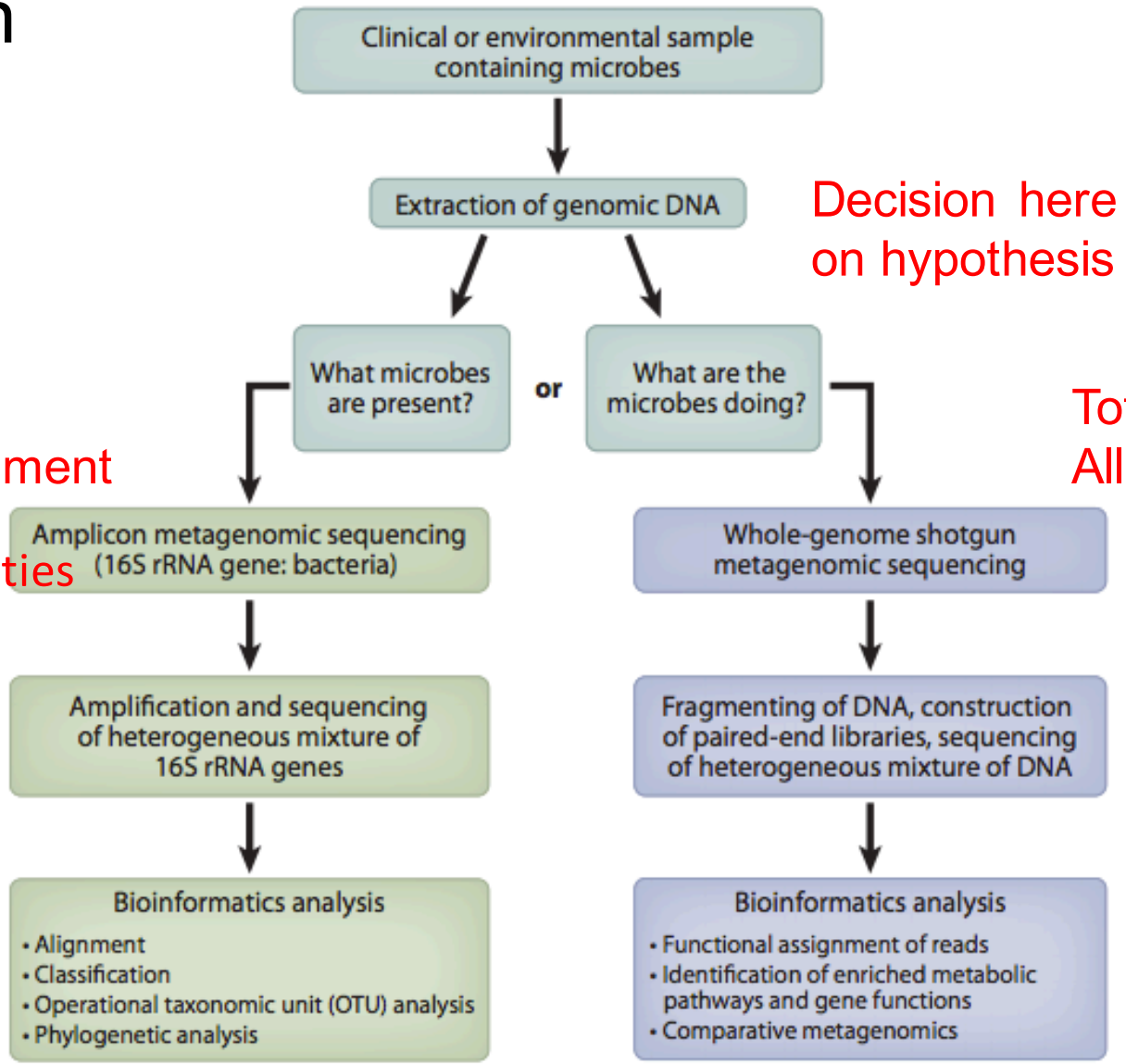
Shade and Handelsman (2012)



Hacquard et al (2015)

Amplicon sequencing or metagenomes?

Workflow decision



Decision here based on hypothesis and budget

PCR product
One gene fragment

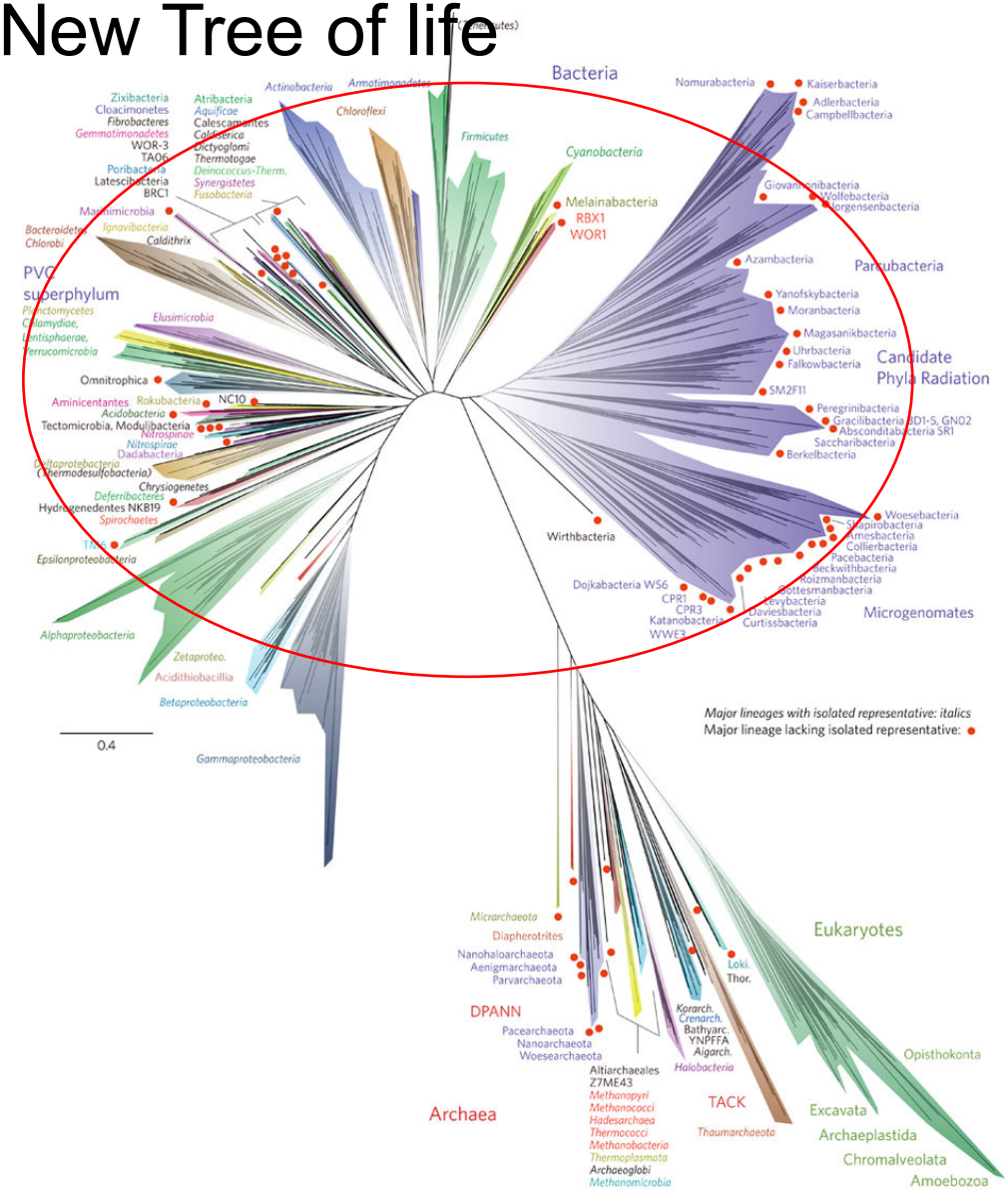
Total DNA
All genes

Fast survey of large communities
Good for first round survey

Grice and Segre (2012)

Amplicon sequencing

New Tree of life

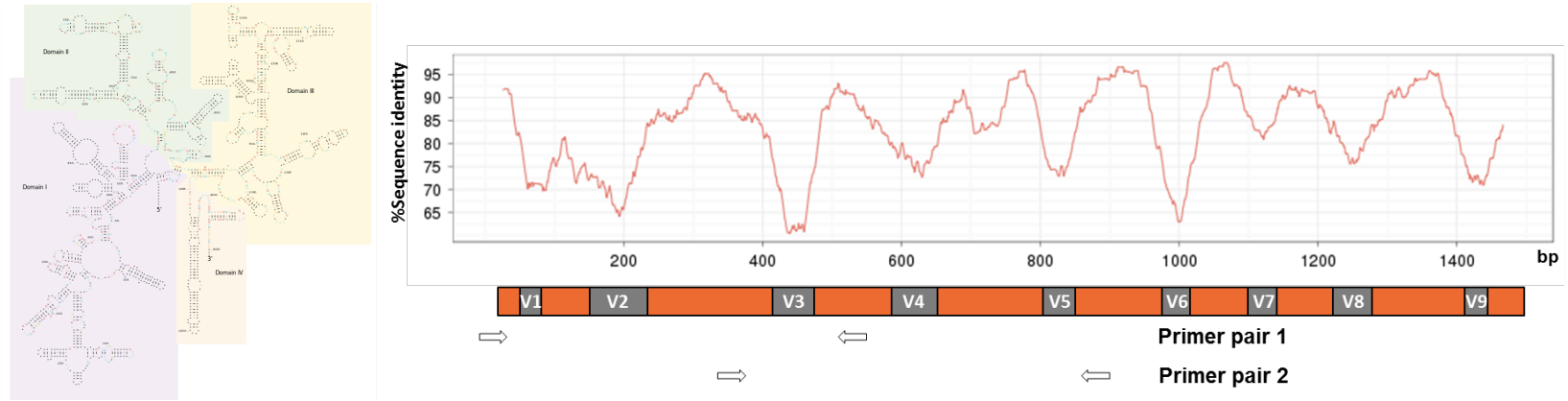


What do they have in common?

Hug et al (2016)

http://www.nytimes.com/2016/04/12/science/scientists-unveil-new-tree-of-life.html?_r=0

16S



- **Advantages**:
  - Universal: Every bacterial and archea 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

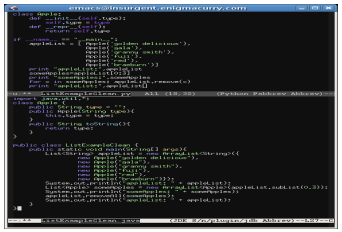
- Problems:
  - Variable copy number in each species
  - No universal (unbiased) primers
  - (Not directly correlated with activity)
  - (Lack of functional information)

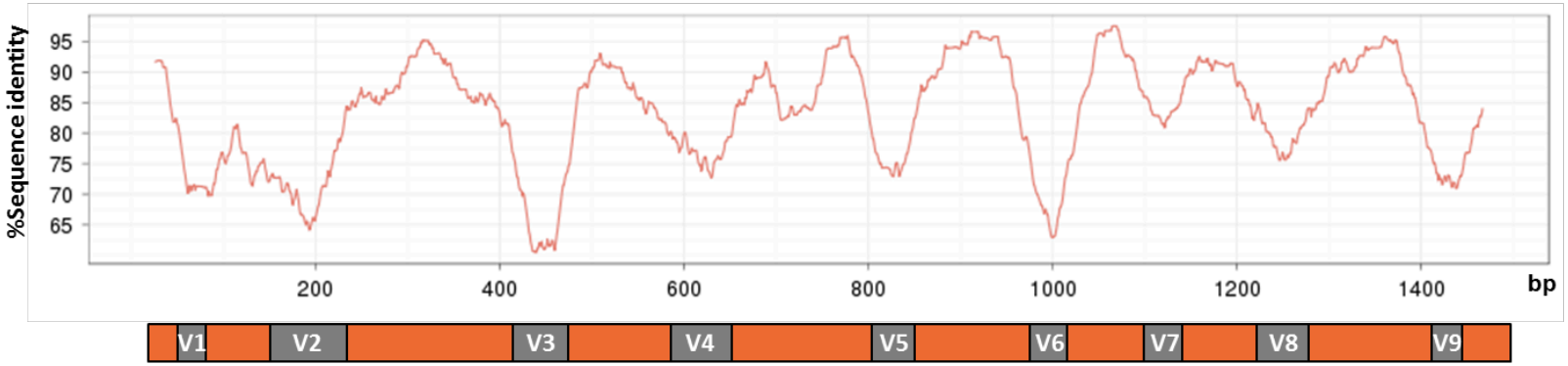
http://en.wikipedia.org/wiki/16S_ribosomal_RNA

Typical workflow

Sampling → Extraction → Sequencing → Analysis

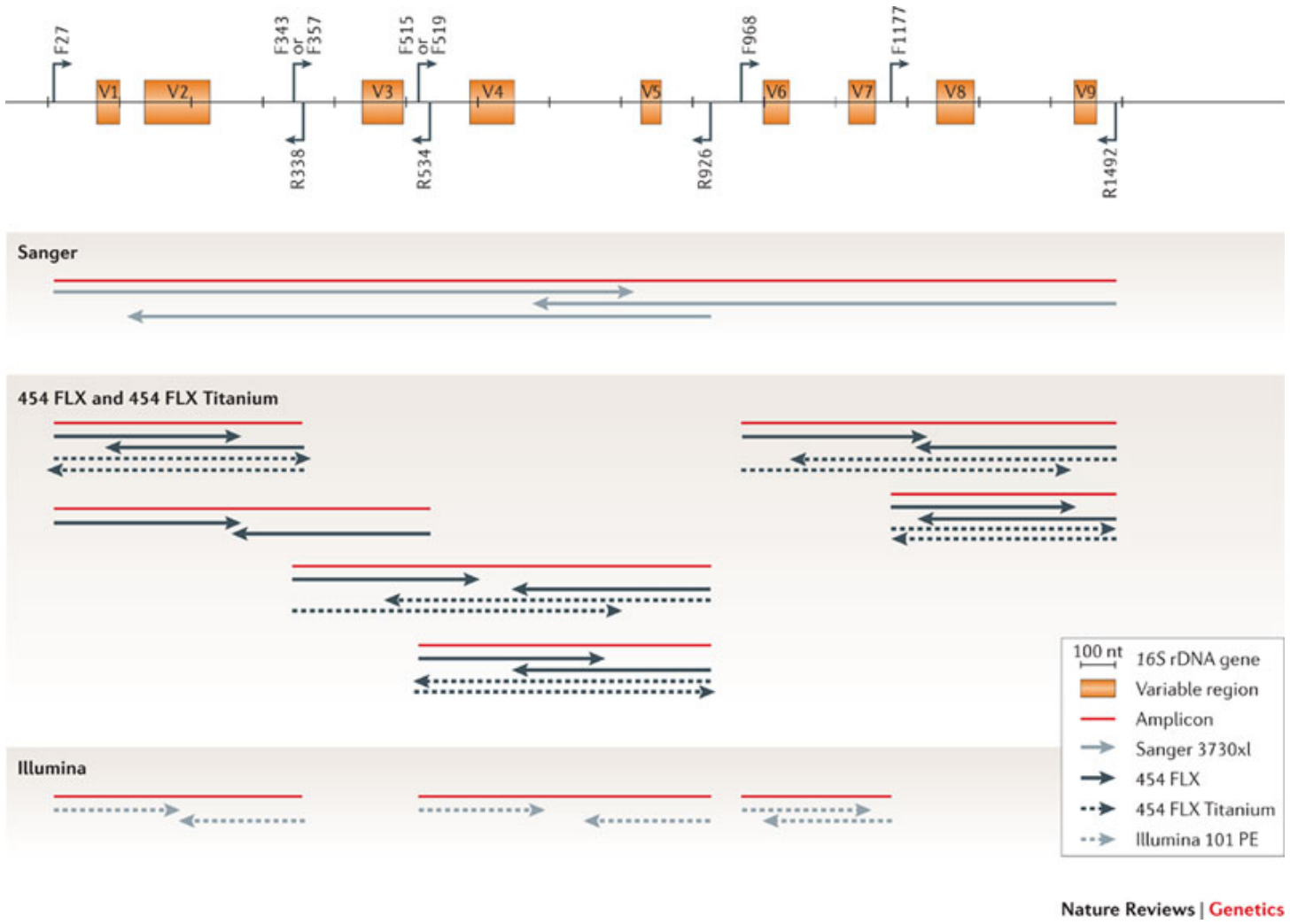




Which region to sequence?

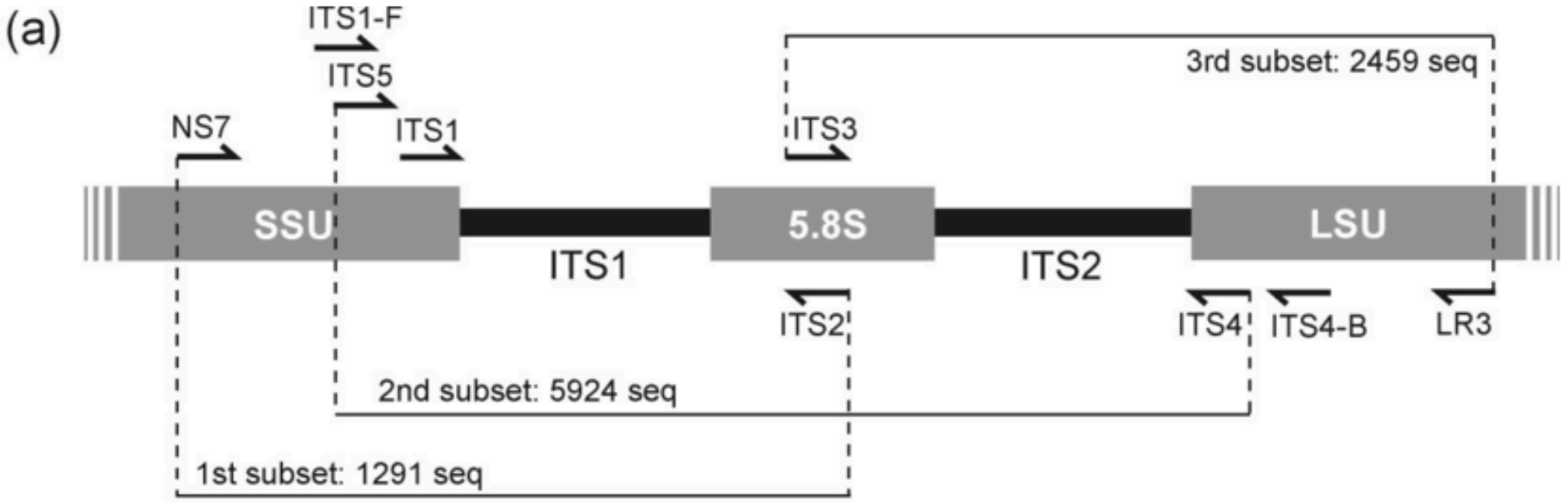
16S amplified region



Kuczynski et al (2011)

Nature Reviews | Genetics

ITS for characterization of fungi species



Bellemain et al (2010)

Workflow

Filter for contaminants and low quality reads

Assemble overlapping reads

Reduce datasets (clustering)

Perform taxonomic classification and compute diversity metrics

- Quality plots and read trimming
  - FastQC

    http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - FASTX

    http://hannonlab.cshl.edu/fastx_toolkit/
- Chimera removal
  - AmpliconNoise

    http://code.google.com/p/ampliconnoise/
  - UCHIME

    http://www.drive5.com/uchime/

Slide of Surya Saha (slideshare)

Workflow

Filter for contaminants and low quality reads

Assemble overlapping reads

Reduce datasets (clustering)

Perform taxonomic classification and compute diversity metrics

Because the read length (2x250 or 2x300) will be longer than the actual DNA fragment

- Merge overlapping paired end reads
  - FLASH

    http://www.genomics.jhu.edu/software/FLASH/index.shtml
  - FastqJoin

    http://code.google.com/p/ea-utils/wiki/FastqJoin
  - CD-HIT read-linker

    http://weizhong-lab.ucsd.edu/cd-hit/wiki/doku.php?id=cd-hit-auxtools-manual

Slide of Surya Saha (slideshare)

Workflow

Filter for contaminants and low quality reads

Assemble overlapping reads

Reduce datasets (clustering)

Perform taxonomic classification and compute diversity metrics

- Clustering with high stringency
  - UCLUST/USEARCH (16S only)

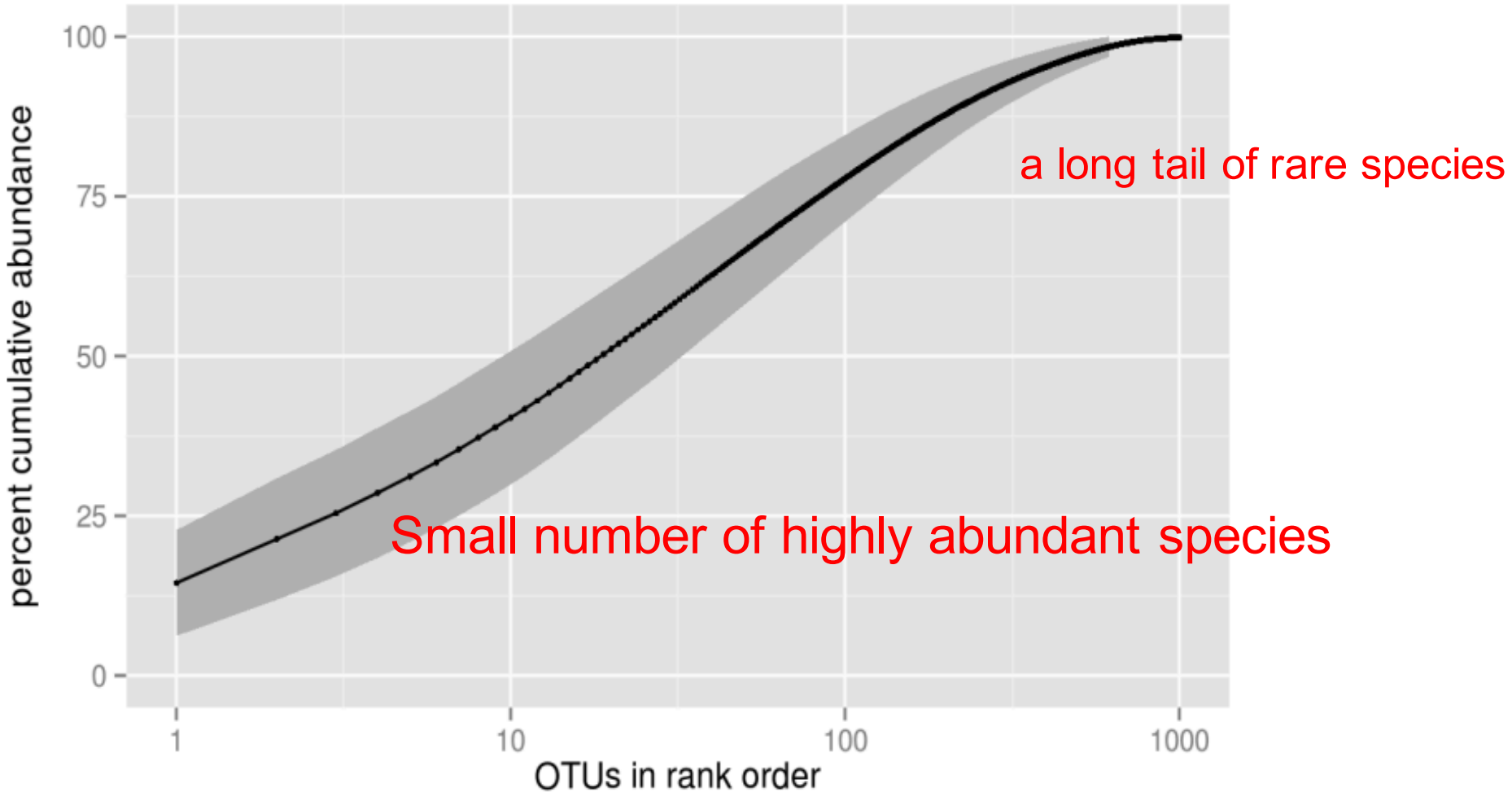
    http://www.drive5.com/usearch/
  - CD-HIT-OTU (16S only)

    http://weizhong-lab.ucsd.edu/cd-hit-otu/
  - phylOTU (16S only)

    https://github.com/sharpton/PhylOTU

Slide of Surya Saha (slideshare)

Typical number of OTUs



Workflow

Filter for contaminants and low quality reads

Assemble overlapping reads

Reduce datasets (clustering)

Perform taxonomic classification and compute diversity metrics

- Composition based classifiers
  - RDP database + classifier

    http://rdp.cme.msu.edu/classifier/classifier.jsp
- Homology based classifiers
  - ARB + Silva database (16S only)

    http://www.arb-home.de/
  - GreenGenes database (16S only)

    http://greengenes.lbl.gov/cgi-bin/nph-index.cgi
  - UNITE database (ITS only)

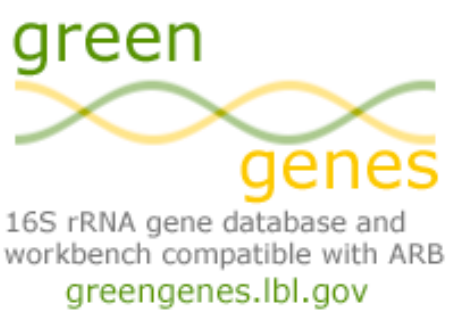
    http://unite.ut.ee/
  - FungalITSPipeline (ITS only)

    http://www.emerencia.org/fungalitspipeline.html

Slide of Surya Saha (slideshare)

Using a “Classifier” to annotate OTUs

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

Qiime (qiime.org)

Mothur (mothur.org)

R

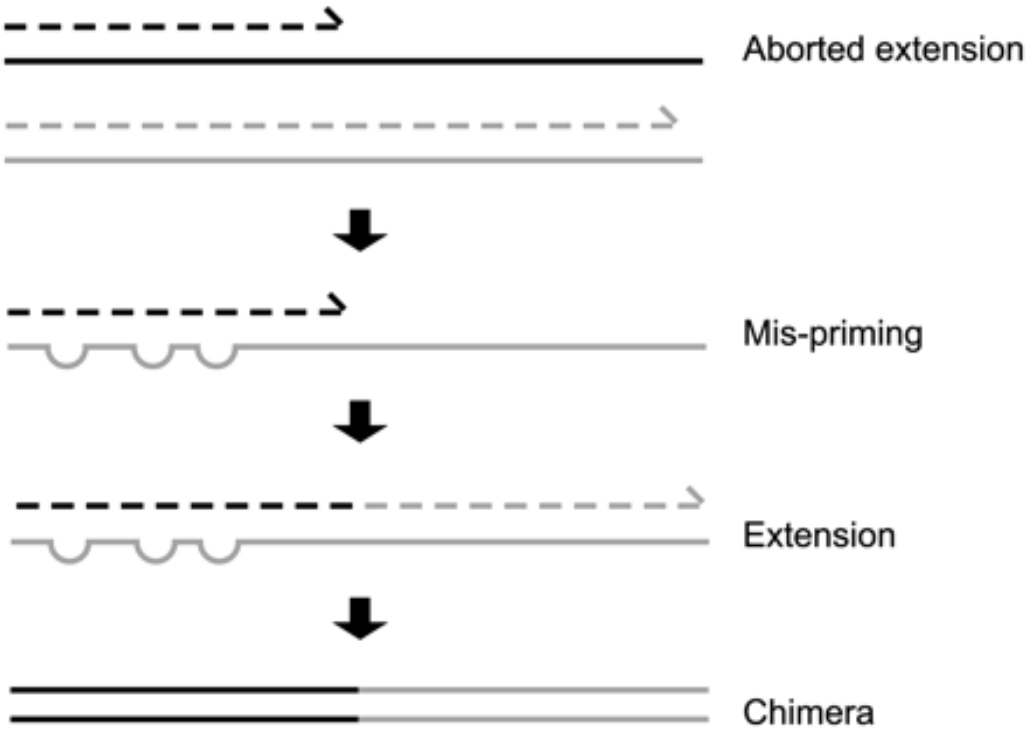
Phyloseq （ https://joey711.github.io/phyloseq/ ）

Potential problem

- 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

Slide of Surya Saha

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%**"

Haas et al (2011)

Metagenomics

Advantage of metagenomics approach

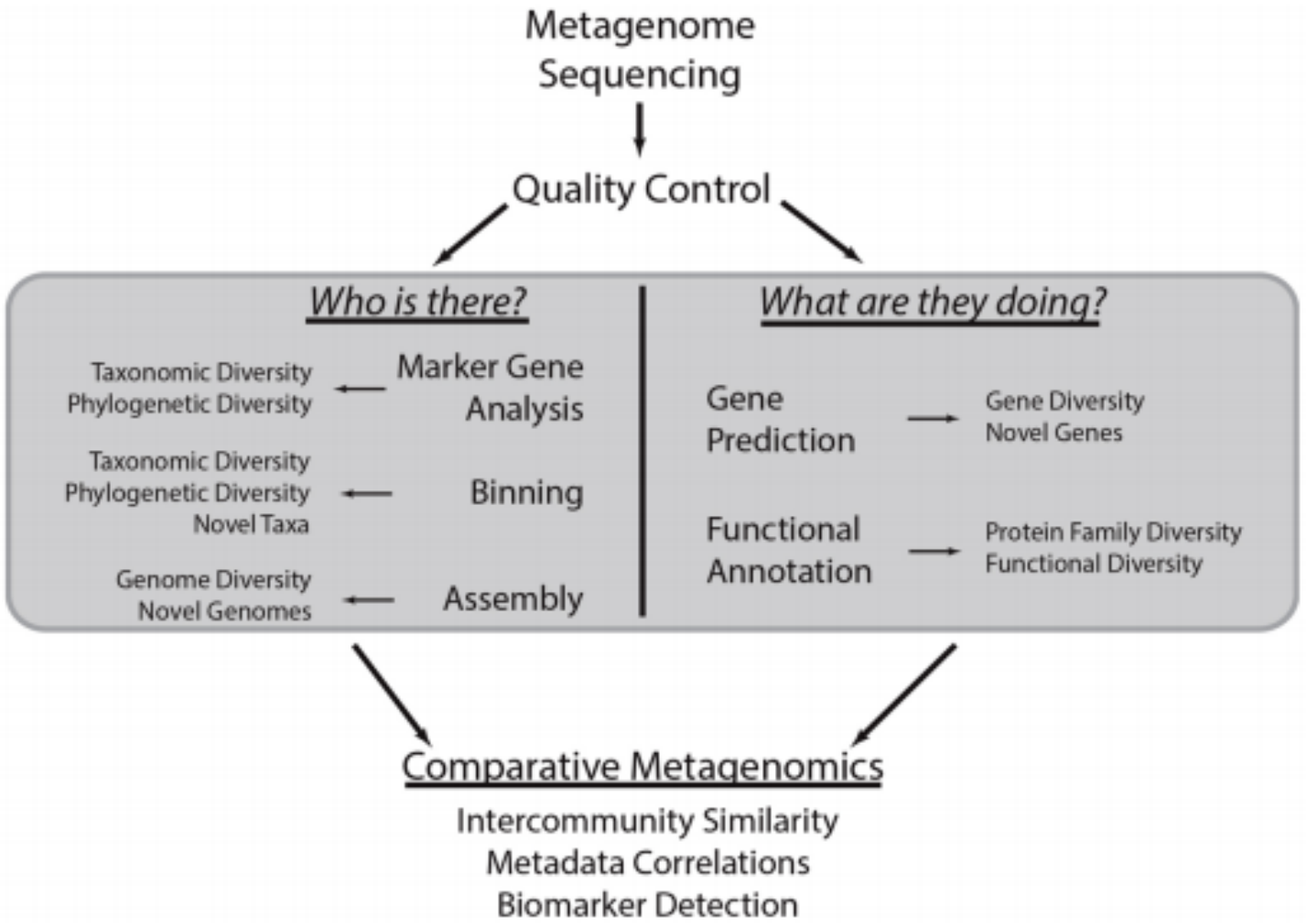
Better classification with Increasing number of complete genomes
Focus on whole genome based phylogeny (whole genome phylotyping)

•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

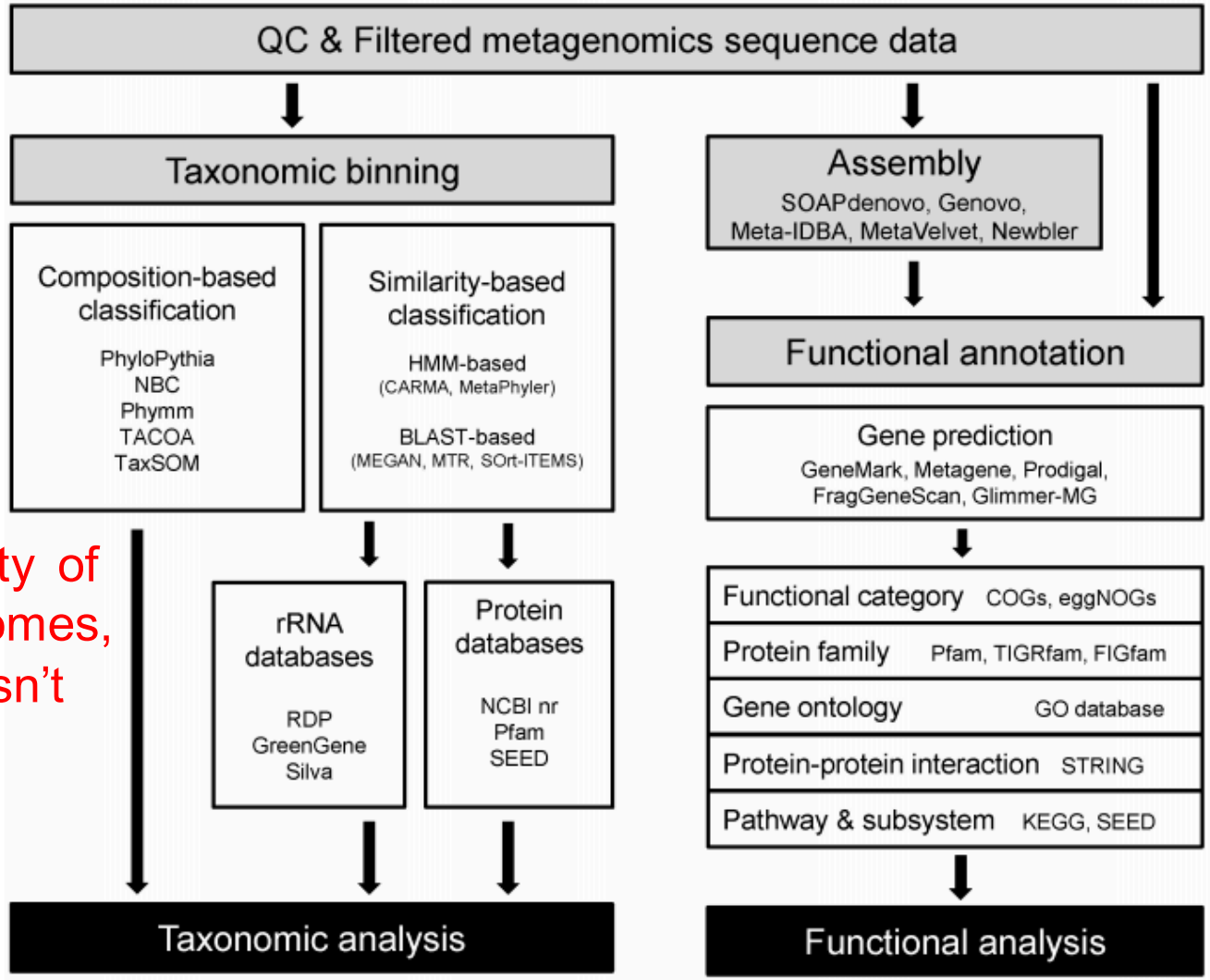
Slide of Surya Saha (slideshare)

Overall workflow



Sharpton (2014)

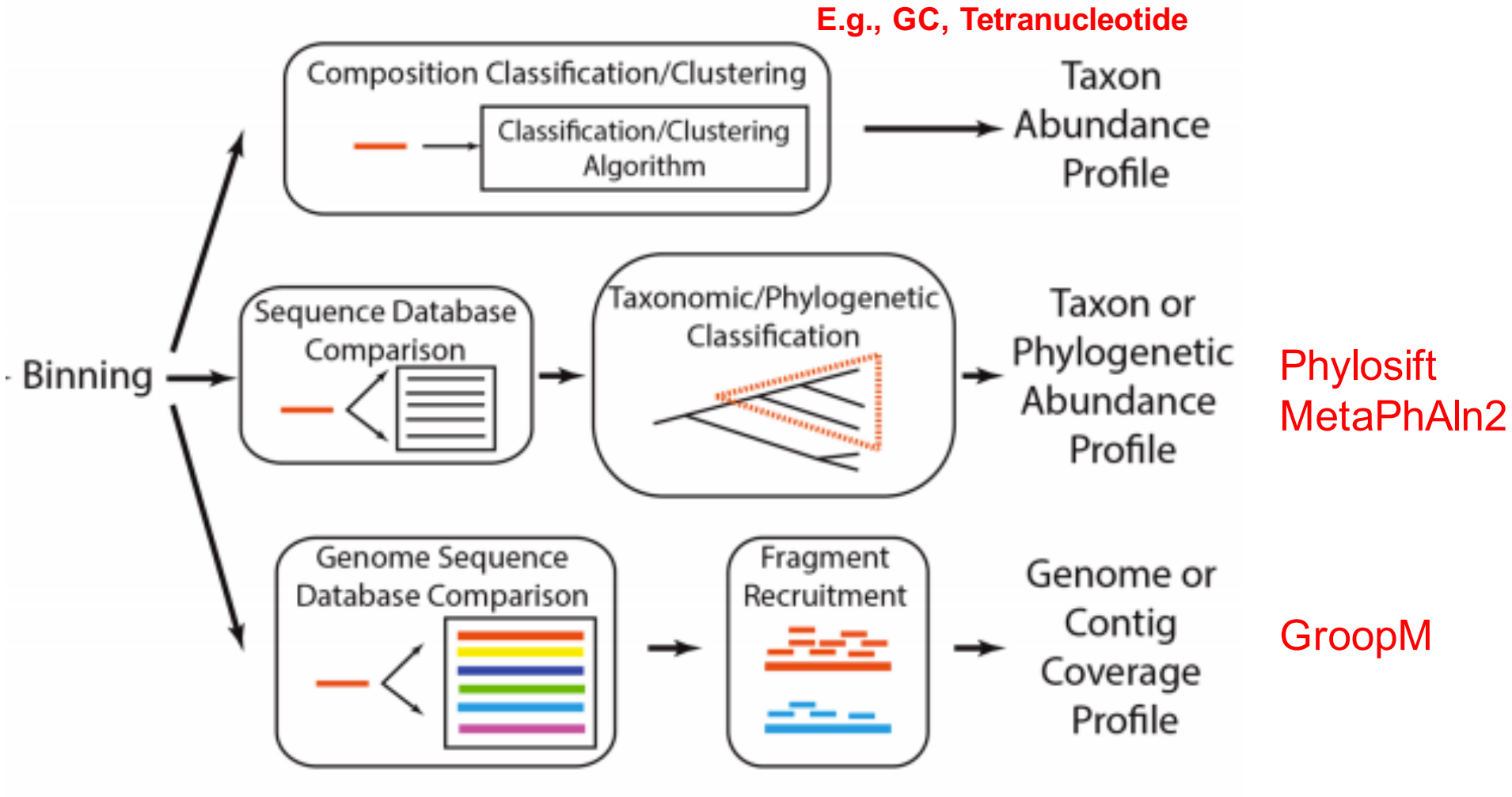
Overall workflow



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

Kim et al (2013)

Binning methods



Sharpton (2014)

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 .

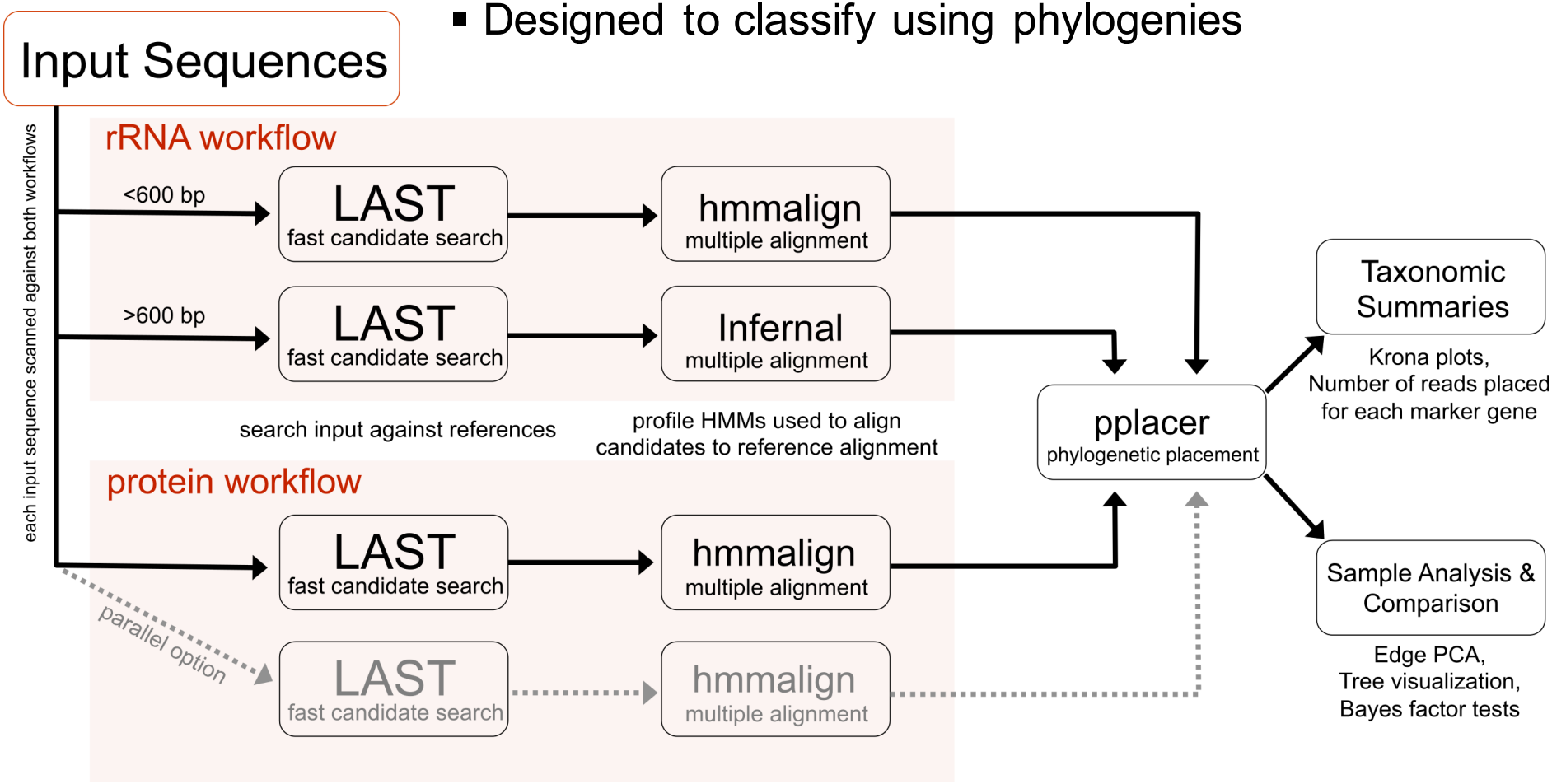
Sharpton (2014)

PhyloSift

mining the global metagenome

https://phylosift.wordpress.com/

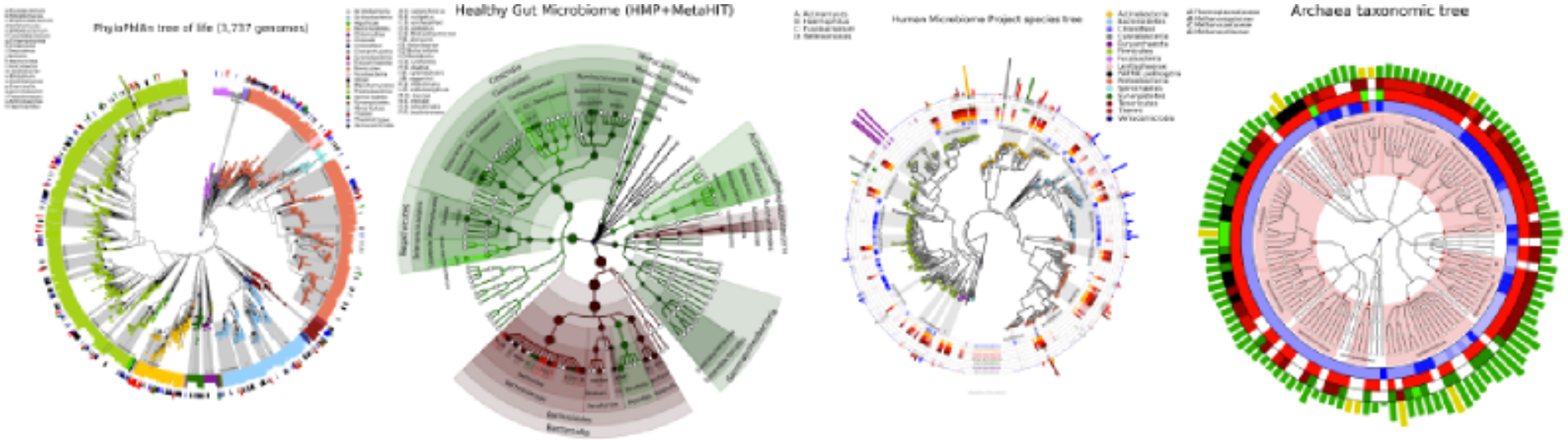
- Uses a database of 37 universal proteins & rRNA genes.
- Designed to classify using phylogenies



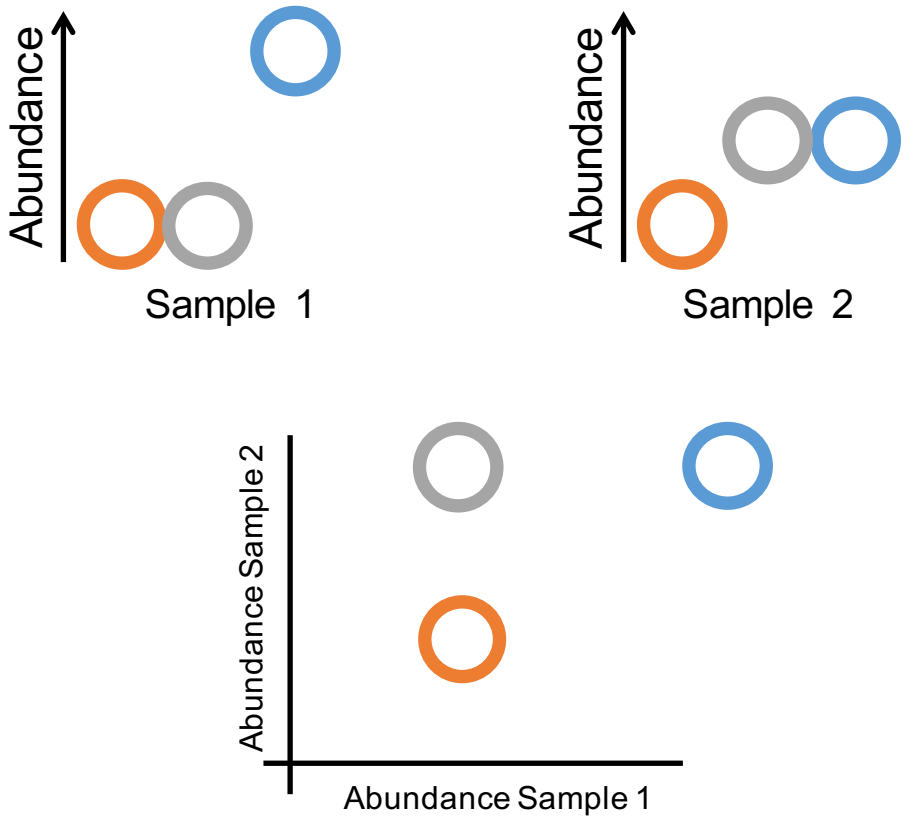
MetaPhAln2 – enhanced metagenomic taxonomic profiling

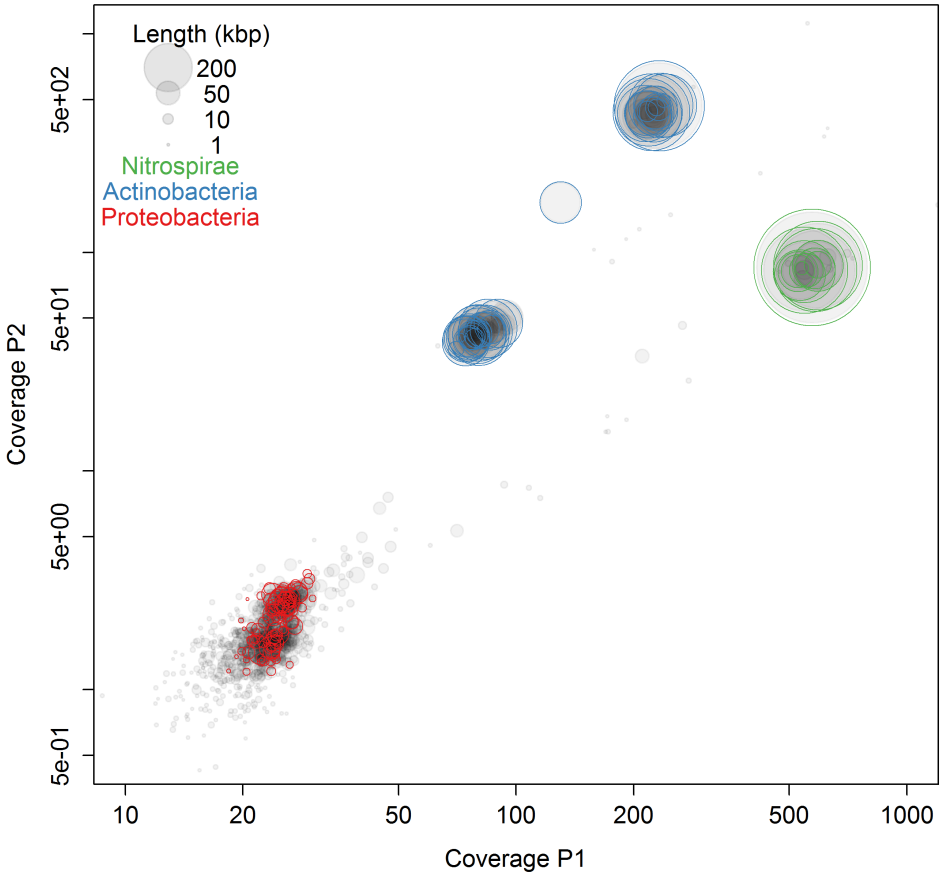
relies **on ~1M unique clade-specific marker genes** identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic), allowing:

Species level resolution
Good visualisation with **GraphAln**
(So it's useful with known ecosystems)



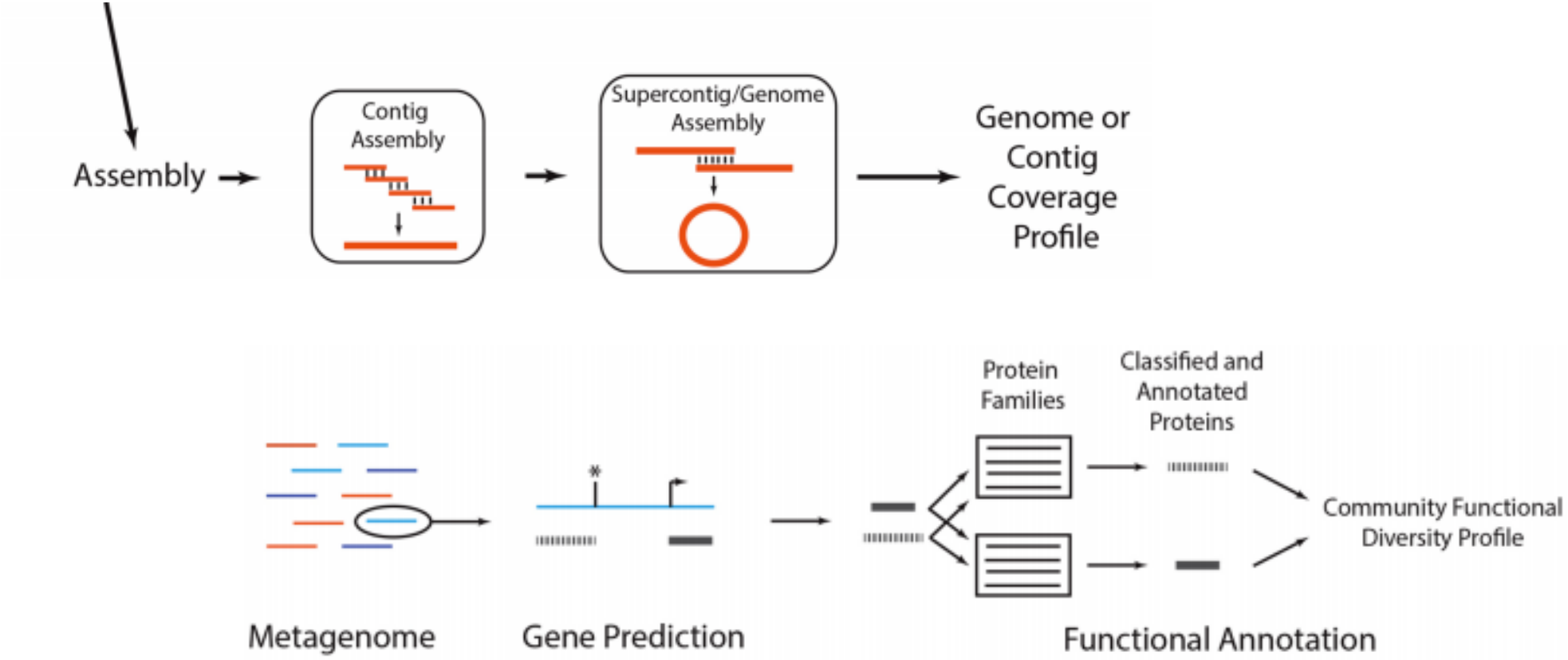
Example of binning based on differential coverage





H. Daims & C. Dorninger, DOME, University of Vienna

Actual assembly



Sharpton (2014)

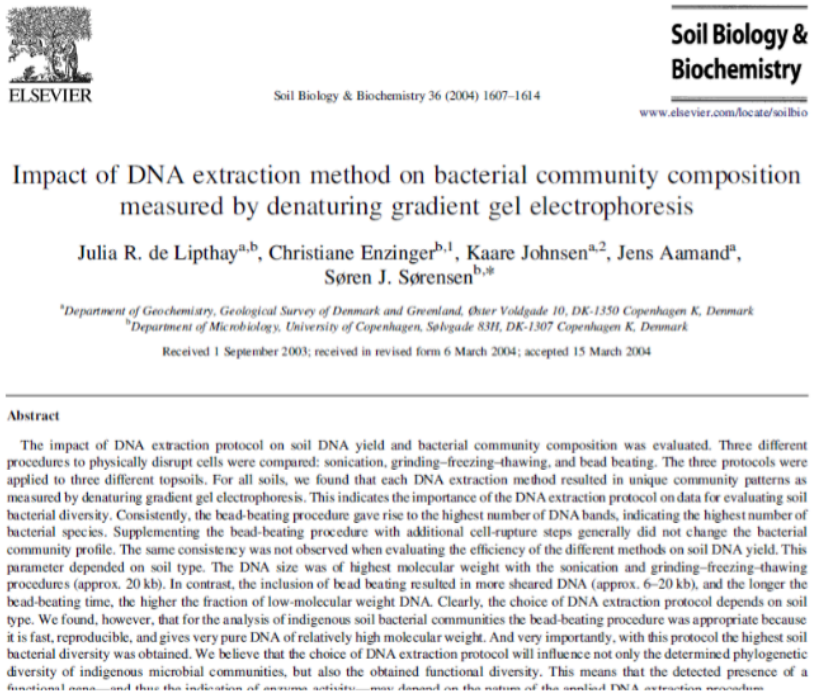
Available tools

data	platform/software	description
Pretreatment	MG-RAST	DMP, QC, DR
	CLC bio	ALR, DMP, QC, DR, OL
	IMG/M system	QC, DR
	PRINSEQ	QC, DR, summary statistics
	NGS QC Toolkit	ALR, QC
	DeconSeq	DNA contamination removal
	FASTX-Toolkit	ALR, DMP, QC
Assembly	Velvet	genome assembler
	ABySS	
	SOAPdenovo2	
	CLC bio	genome/metagenome assembler
	IDBA-UD	metagenome assembler
	MetaVelvet	
	Ray Meta	
	Omega	
	MEGAHIT	

Ju and Zhang (2016)

Biases

Extraction protocol matters



ELSEVIER

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 Lipthay[a,b], Christiane Enzinger[b,1], Kaare Johnsen[a,2], Jens Aamand[a], Søren J. Sørensen[b,*]

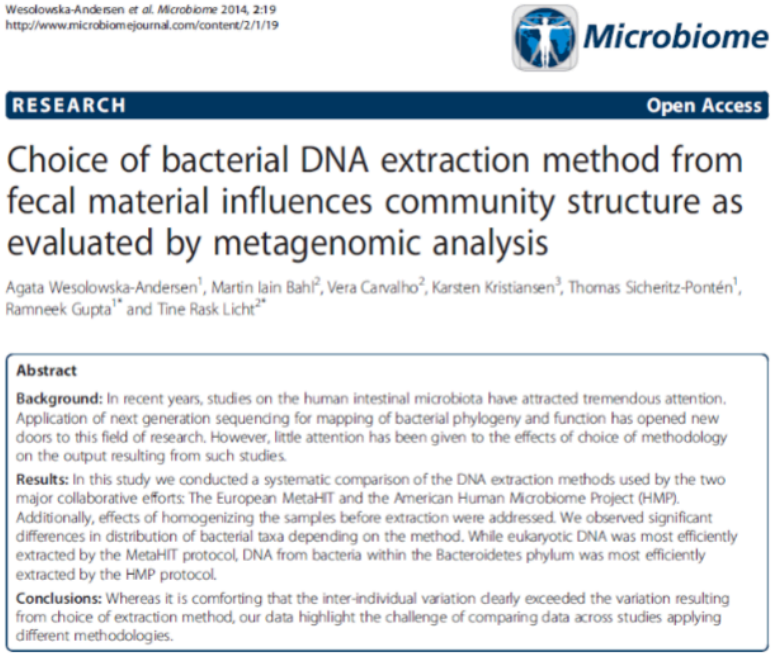
[a]Department of Geochemistry, Geological Survey of Denmark and Greenland, Øster Voldgade 10, DK-1350 Copenhagen K, Denmark
[b]Department of Microbiology, University of Copenhagen, Sølvgade 83H, DK-1307 Copenhagen K, Denmark

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

Microbiome

RESEARCH Open Access

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

Agata Wesolowska-Andersen[1], Martin Iain Bahl[2], Vera Carvalho[2], Karsten Kristiansen[3], Thomas Sicheritz-Pontén[1], Ramneek Gupta[1*] and Tine Rask Licht[2*]

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.

	Number of OTUs at percentage identity thresholds							
Read filtering	100	99	98	97	95	90	% errorless reads	% reads used
5′ forward (V1 and V2)								
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
3′ reverse (V8)								
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[1*], Michael J Cox[2], Elena M Turek[2], Szymon T Calus[3], William O Cookson[2], Miriam F Moffatt[2], Paul Turner[4,5], Julian Parkhill[1], Nicholas J Loman[3] and Alan W Walker[1,6*]

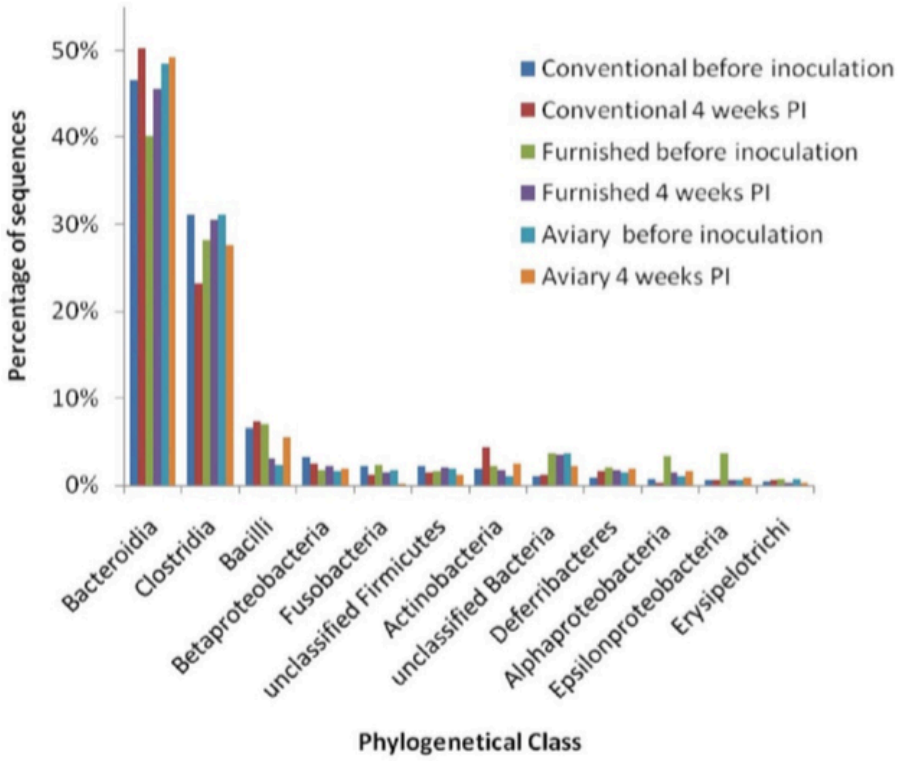
RESEARCH HIGHLIGHT

Tracking down the sources of experimental contamination in microbiome studies

Sophie Weiss[1], Amnon Amir[2], Embriette R Hyde[2], Jessica L Metcalf[2], Se Jin Song[2] and Rob Knight[2,3,4*]

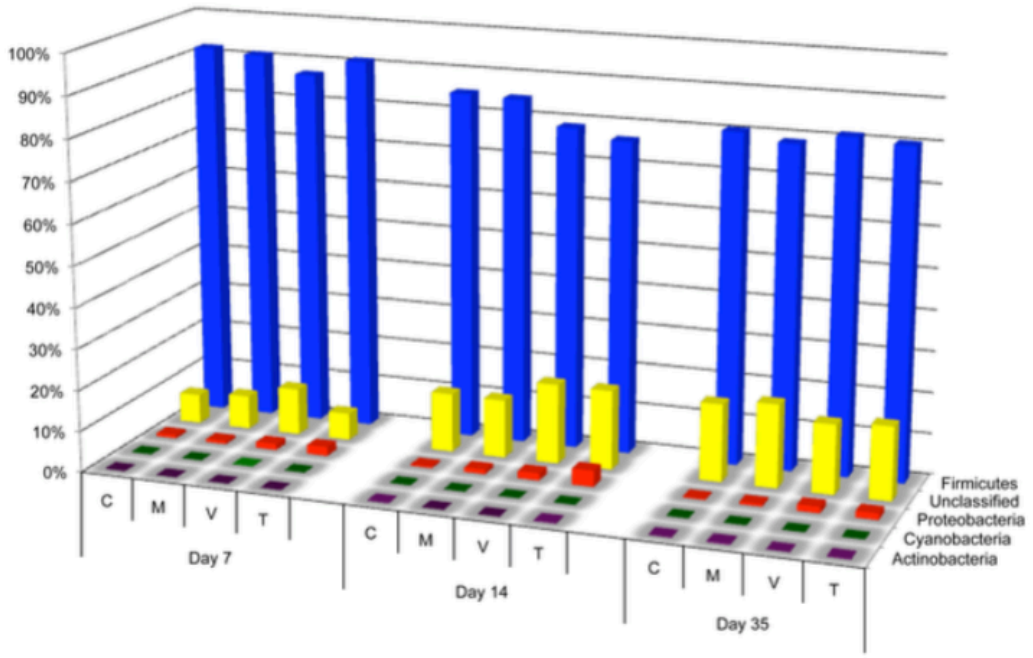
2 papers with different results at the same year

Bacteroidetes >>> rest



Nordentoft S *et al* (2011) BMC Microbiology

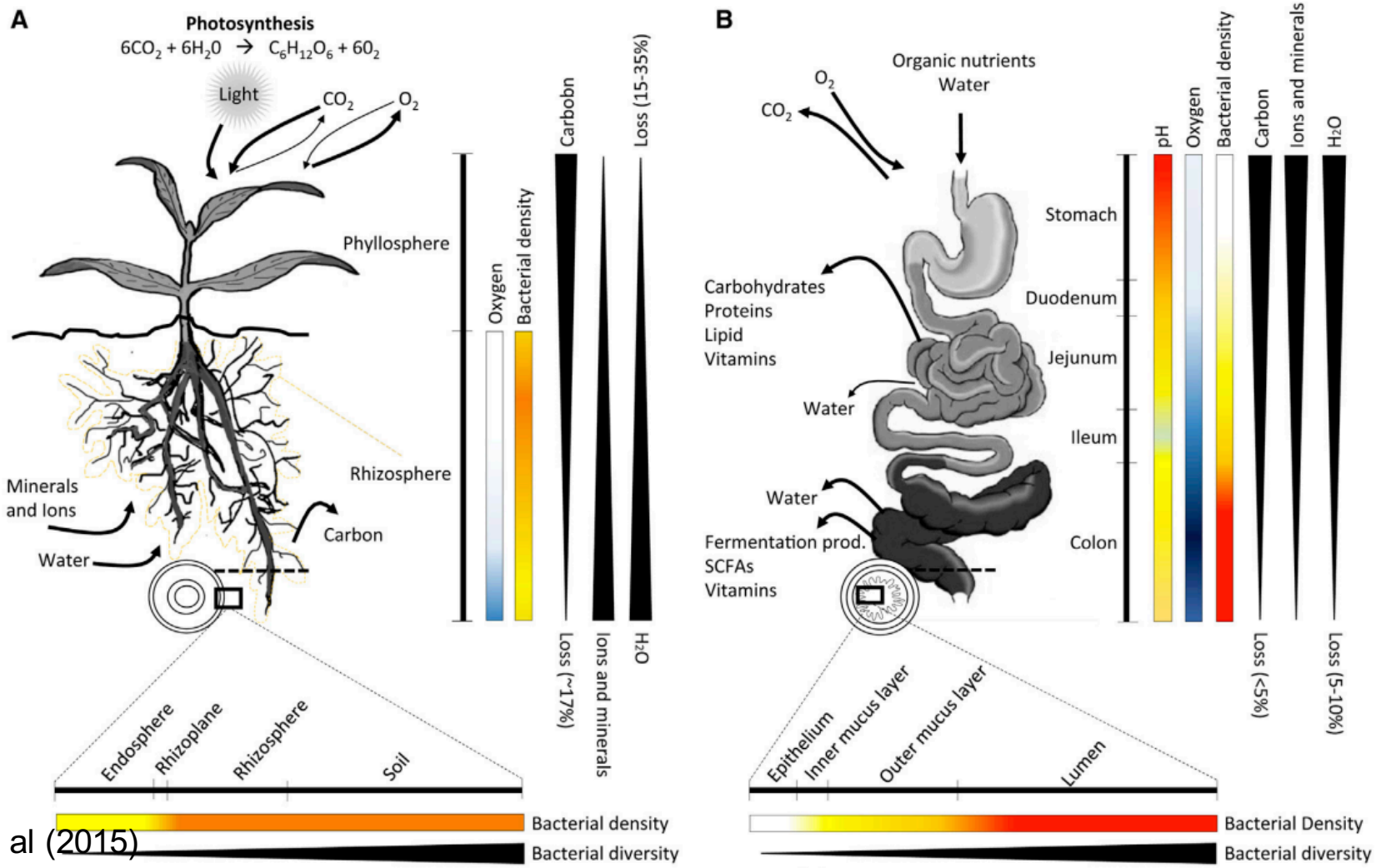
firmicutes >>> rest > bacteroidetes



Danzeisen JL *et al* (2011) PLOS one

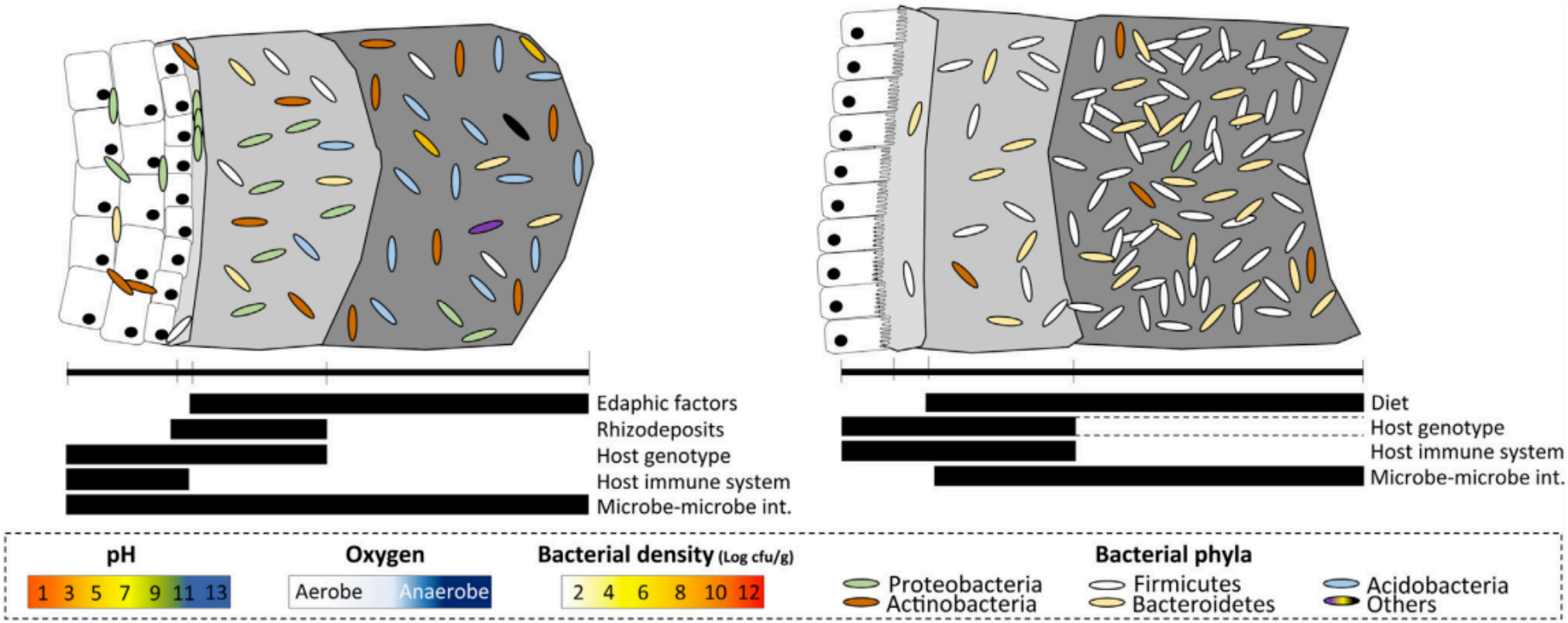
Case studies

Two most common systems



Hacquard et al (2015)

Two most common systems



Hacquard et al (2015)

Two most common systems

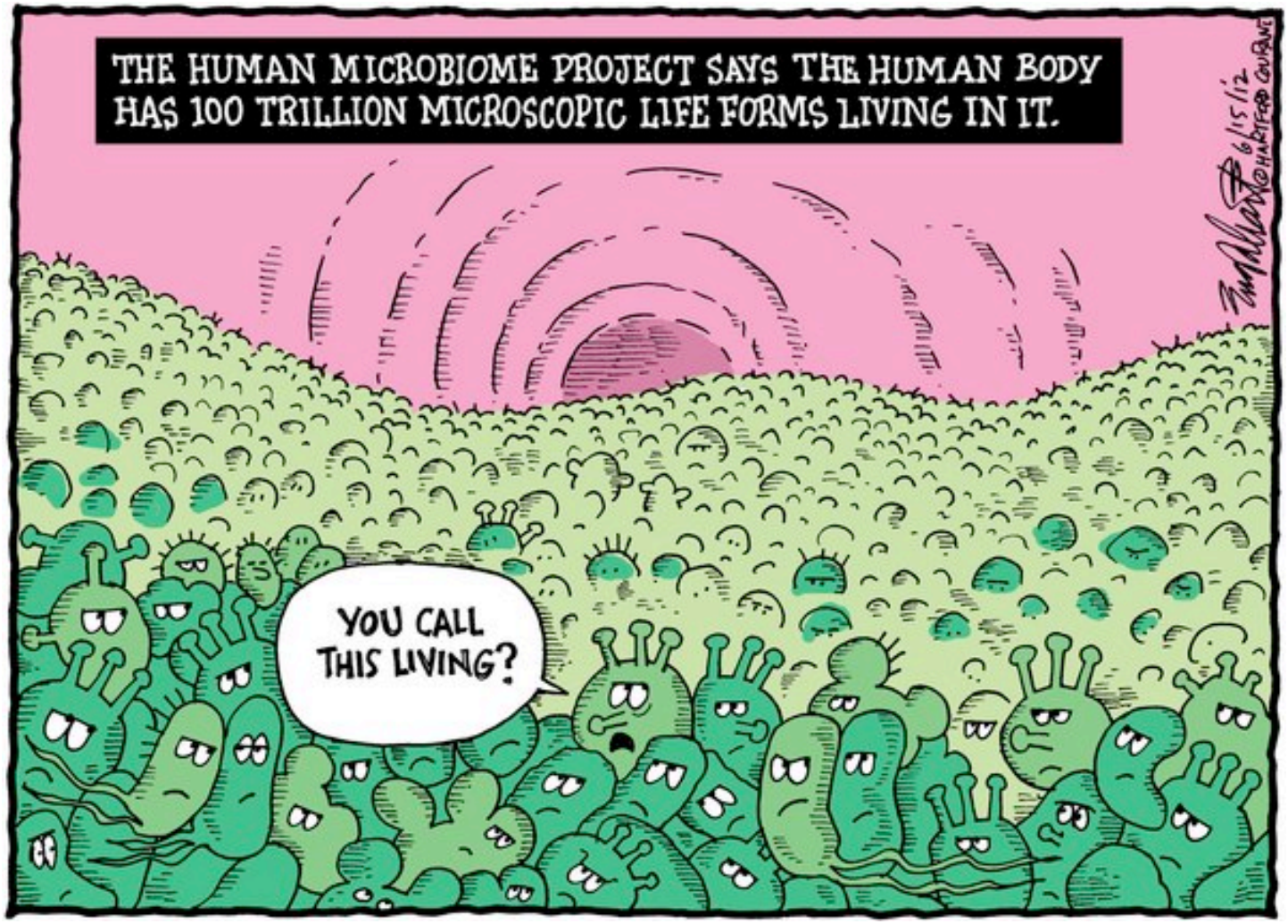
Table 1. Percentage of Shotgun Metagenome Reads Assigned to Each Kingdom of Life across Metagenome Studies

	Cucumber[a]	Wheat[a]	Soybean[b]	Wheat[c]	Oat[c]	Pea[c]	Barley[d]	Gut[e]
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

[a]Ofek-Lalzar et al. (2014) (metagenomics of rhizoplane samples).
[b]Mendes et al. (2014) (metagenomics of rhizosphere samples).
[c]Turner et al. (2013) (metatranscriptomics of rhizosphere samples).
[d]Bulgarelli et al. (2015) (metagenomics of rhizosphere samples).
[e]Qin et al. (2010) (metagenomics of gut samples).

Hacquard et al (2015)

Human gut microbiome



Human gut microbiome

Vol 464|4 March 2010|doi:10.1038/nature08821 nature

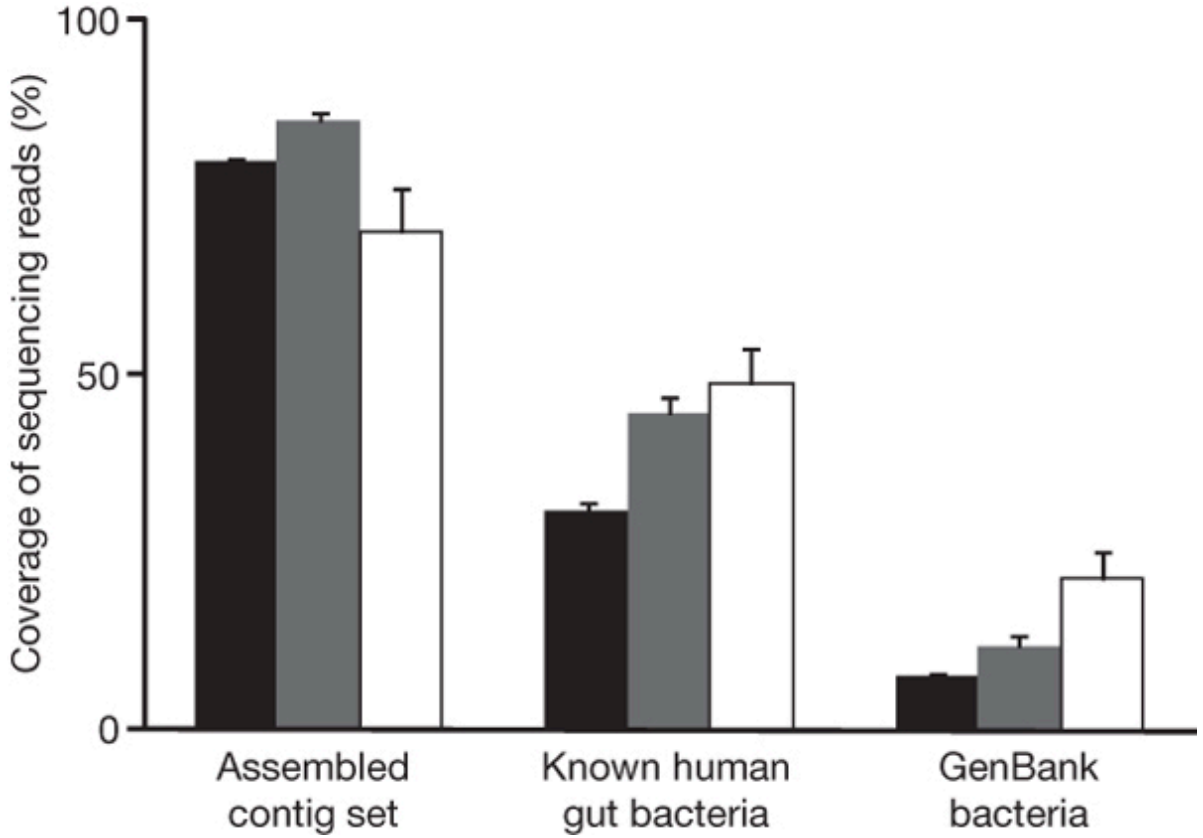
ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

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

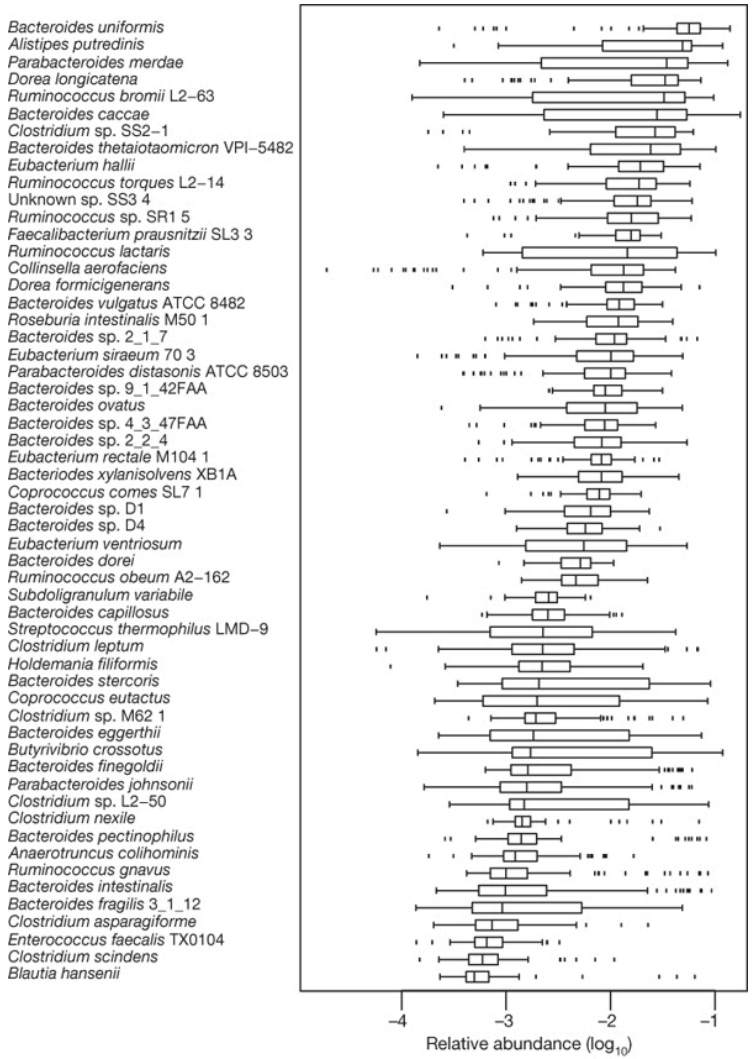
doi:10.1038/nature08821

Human gut microbiome



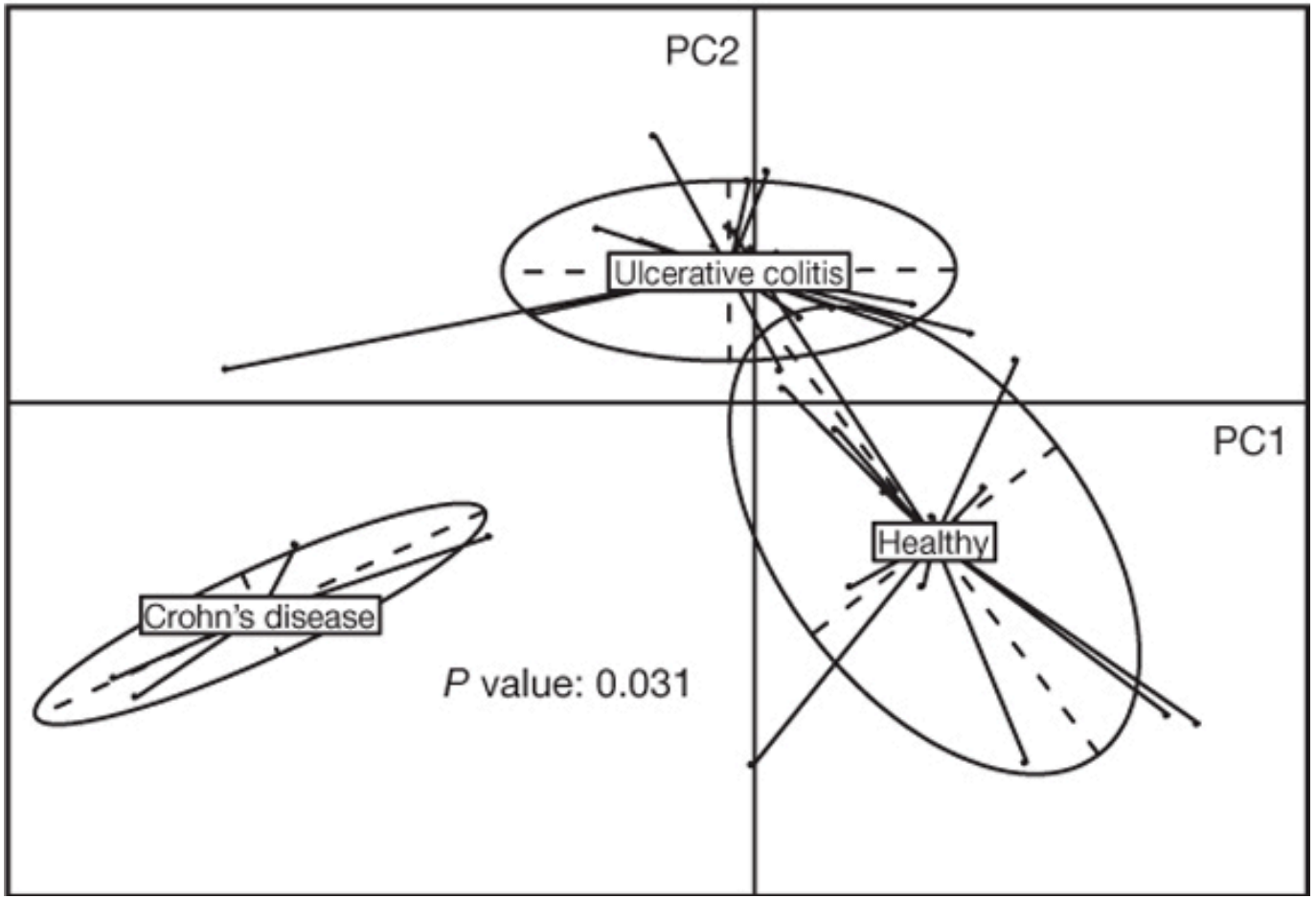
doi:10.1038/nature08821

Human gut microbiome



doi:10.1038/nature08821

Human gut microbiome

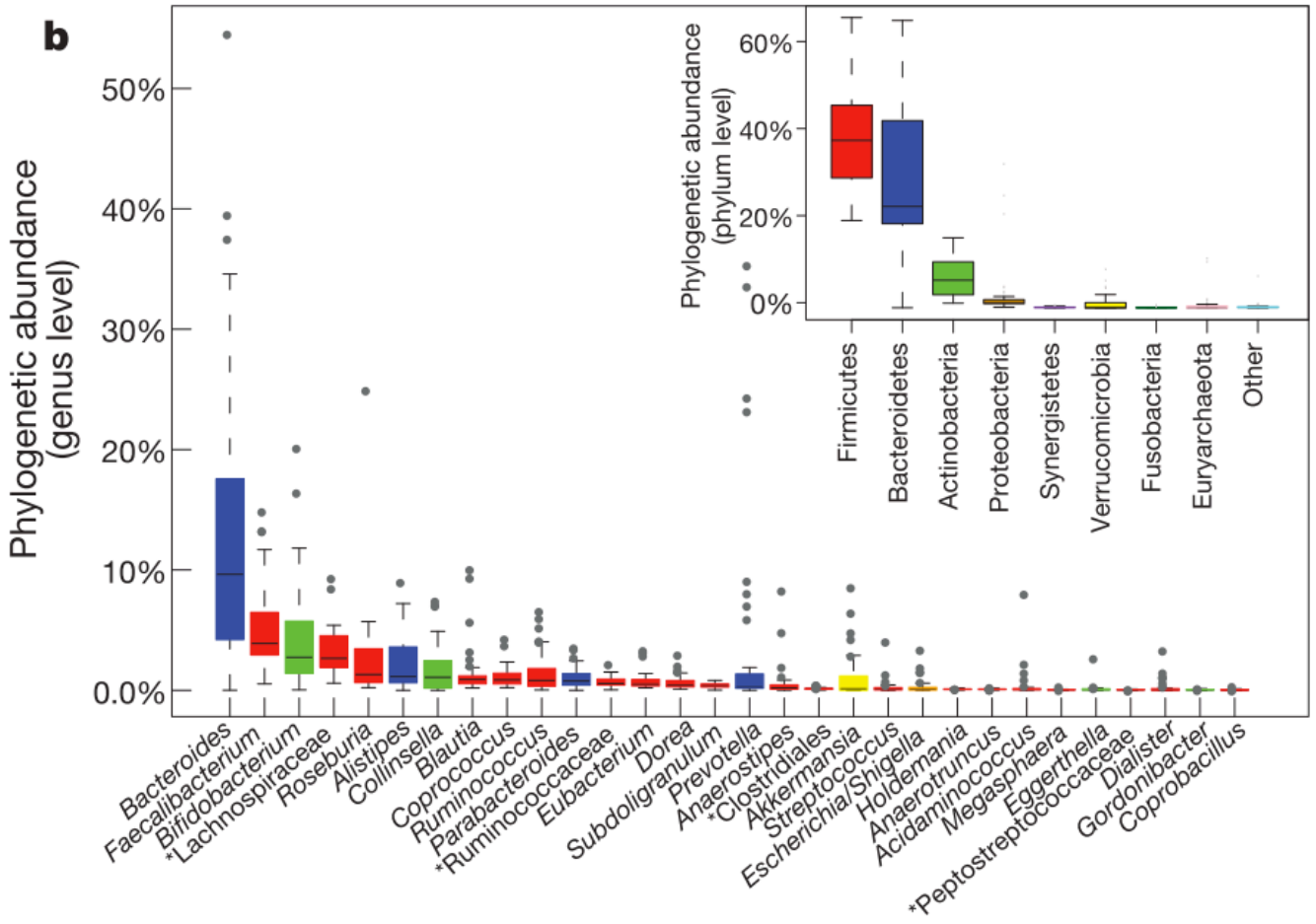


We can check which OTUs constitute the clustering (and separation) patterns

-> Biology
-> Biomarkers

doi:10.1038/nature08821

Human gut microbiome



doi:10.1038/nature09944

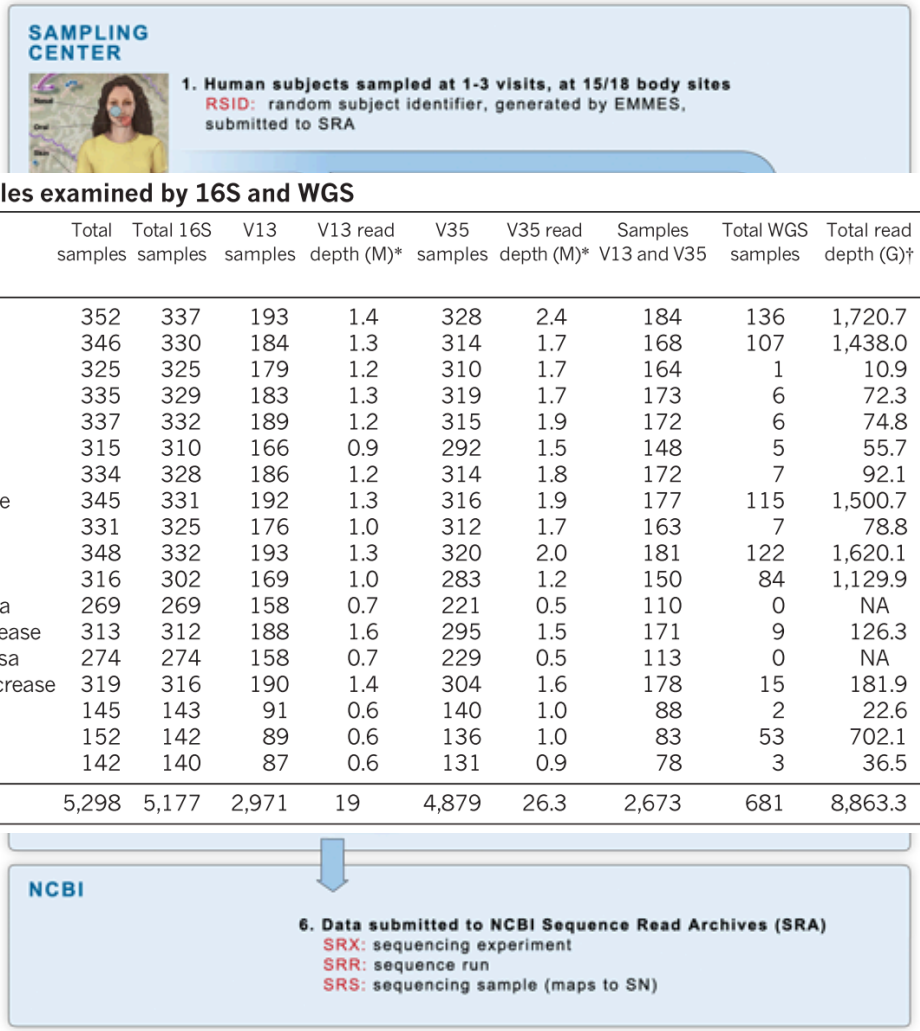
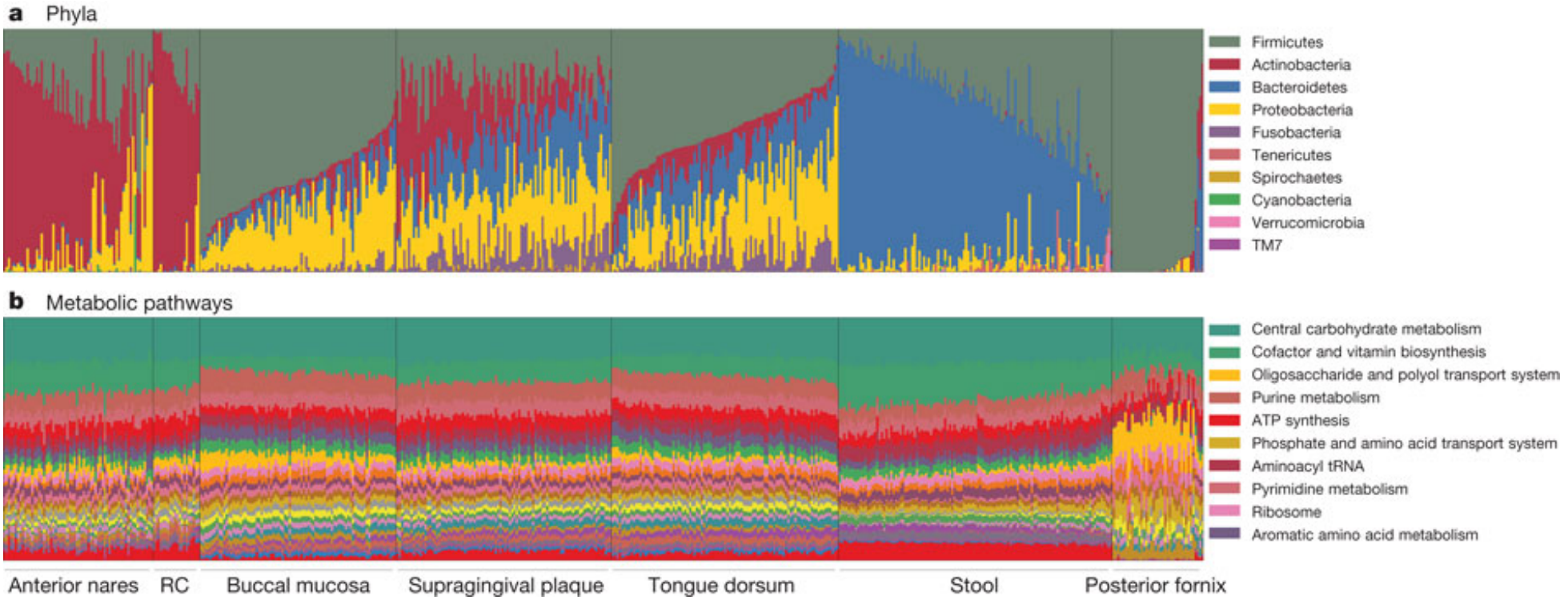


Table 1 | HMP donor samples examined by 16S and WGS

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

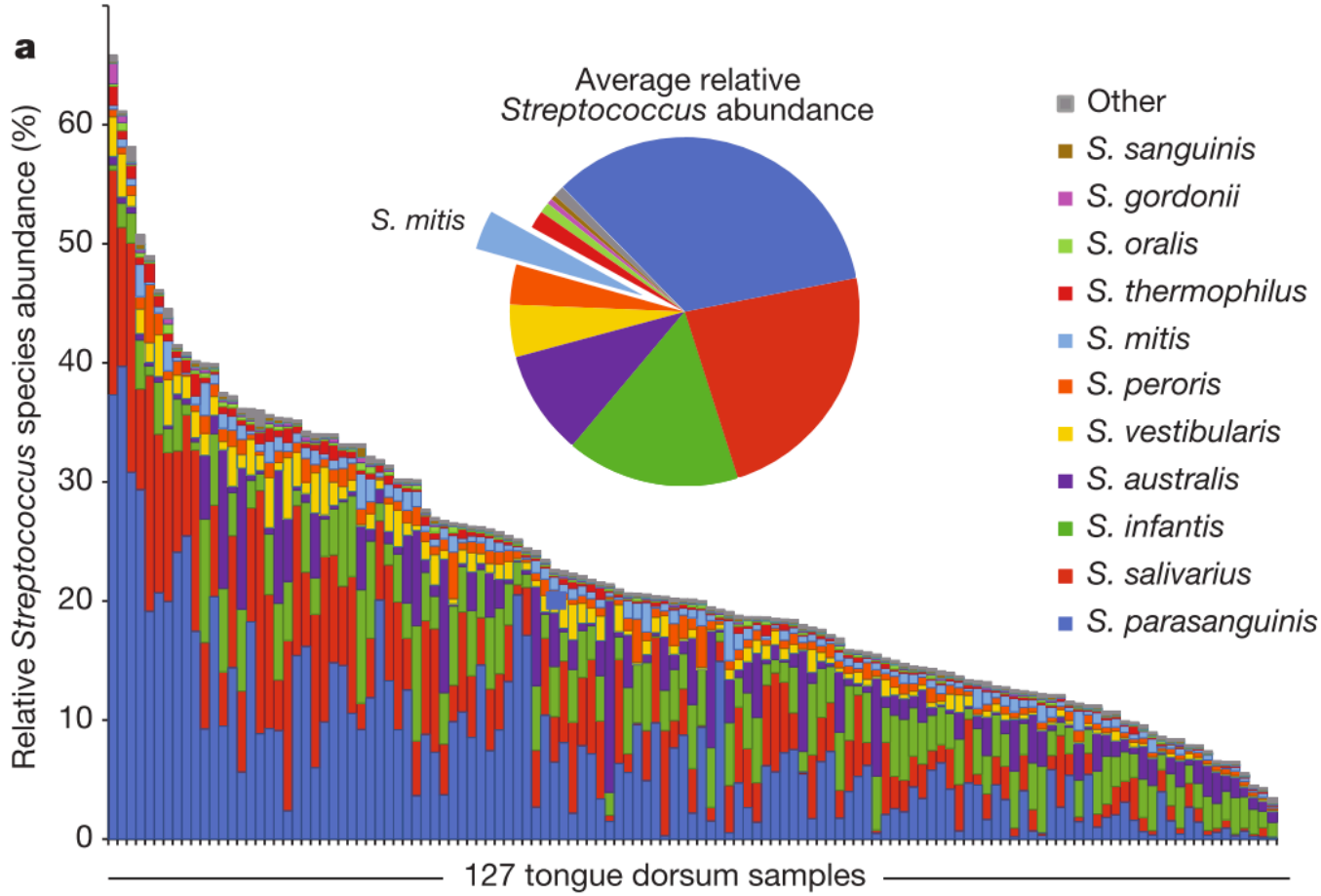
doi:10.1038/nature11209

Human microbiome



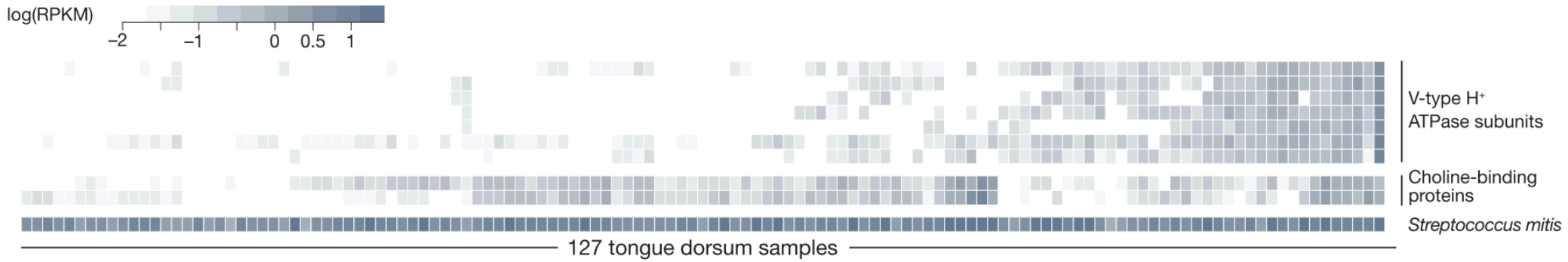
doi:10.1038/nature11234

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



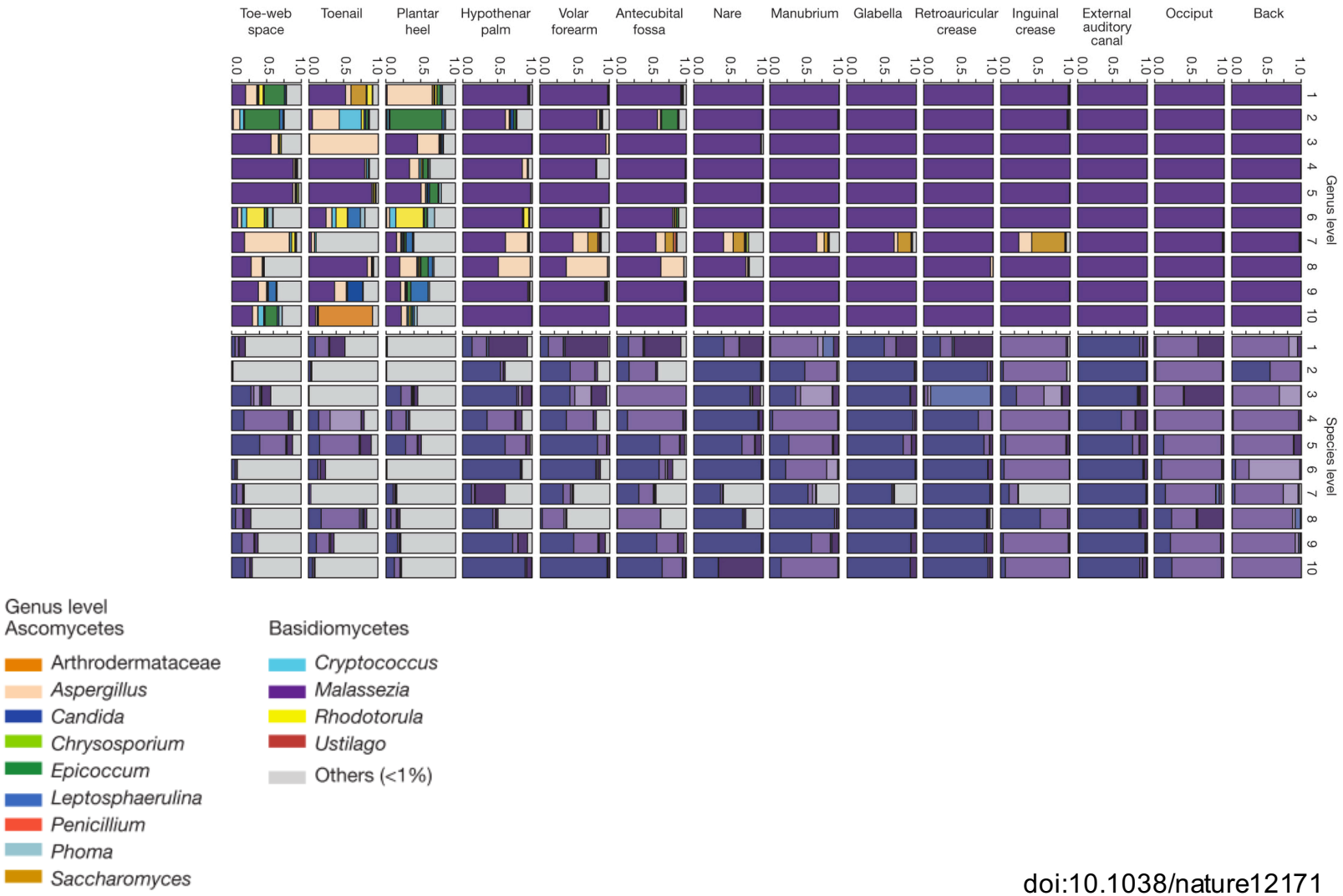
doi:10.1038/nature11234

Gene loss & Structural variants are common



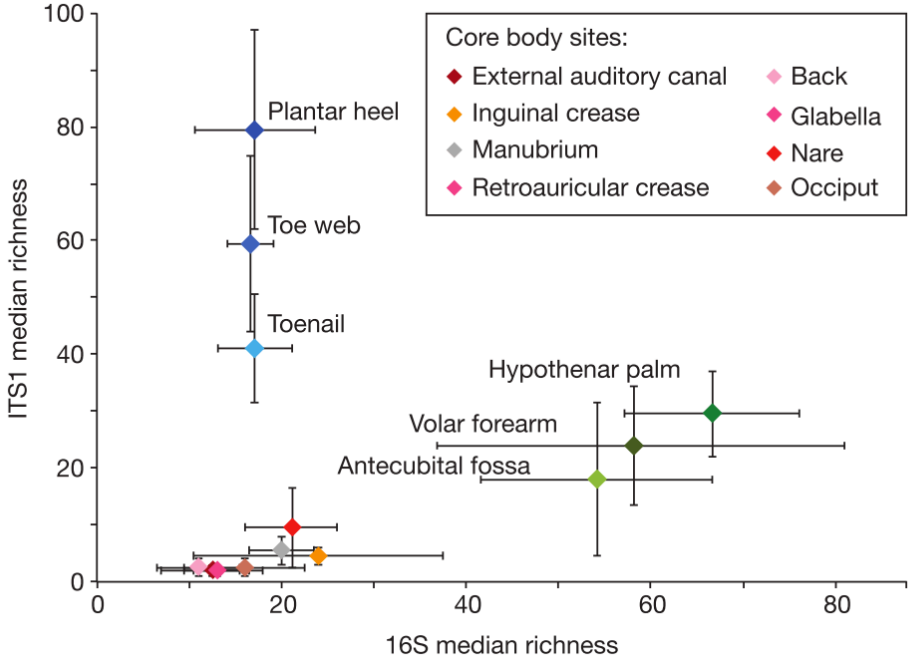
doi:10.1038/nature11234

Skins



doi:10.1038/nature12171

Skins



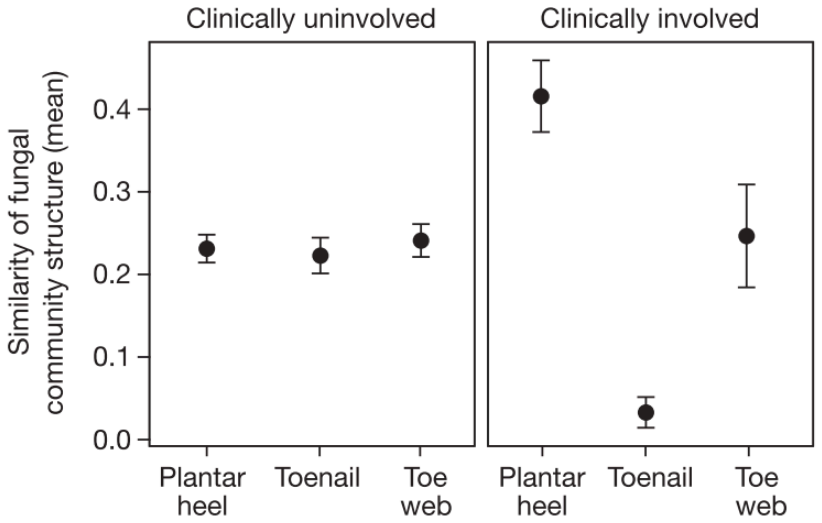
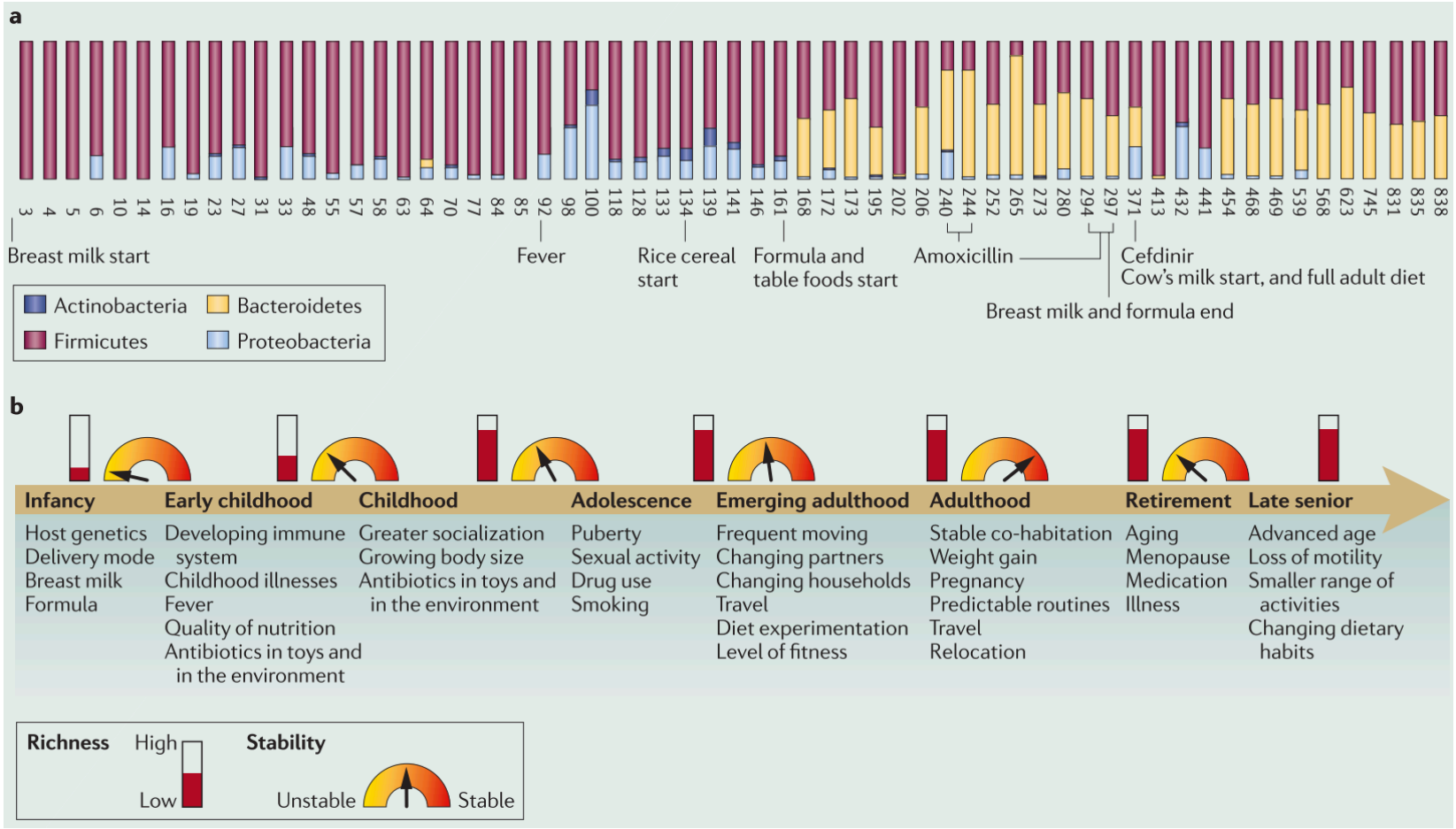


Figure 4 | Clinical involvement alters shared fungal community structure.

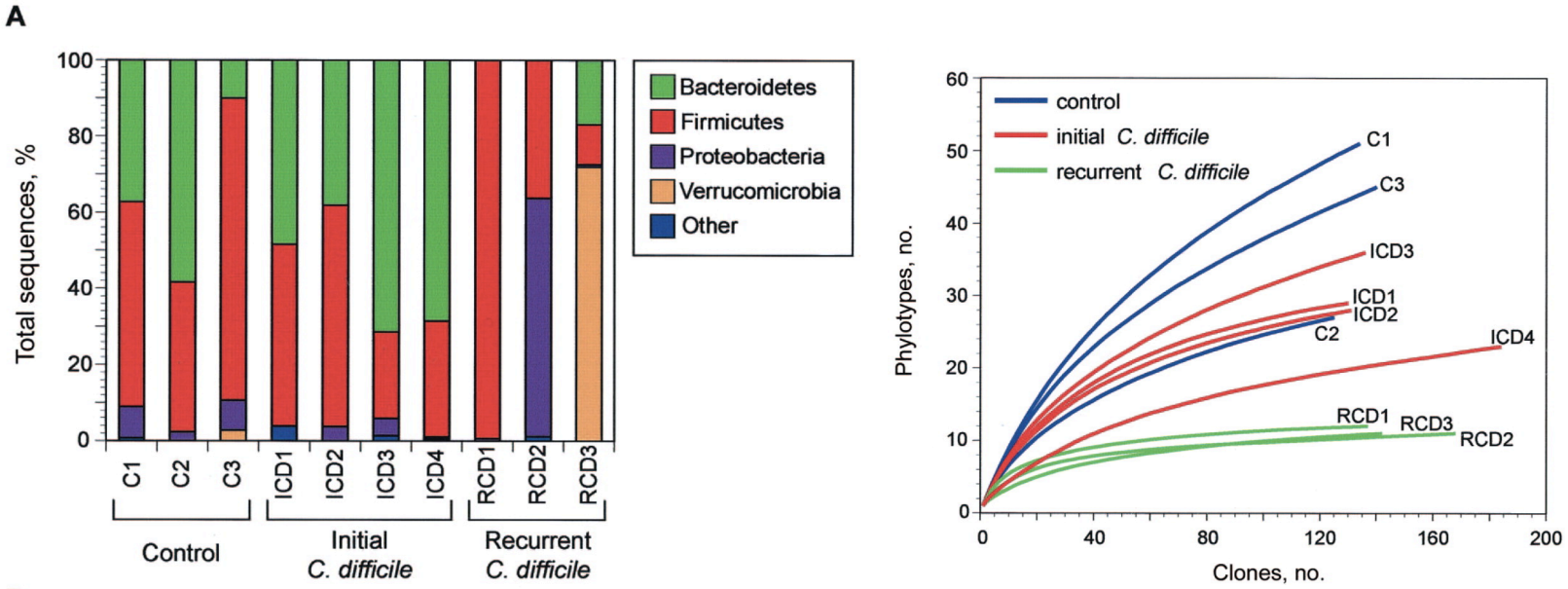
doi:10.1038/nature12171

The gut microbiome during life



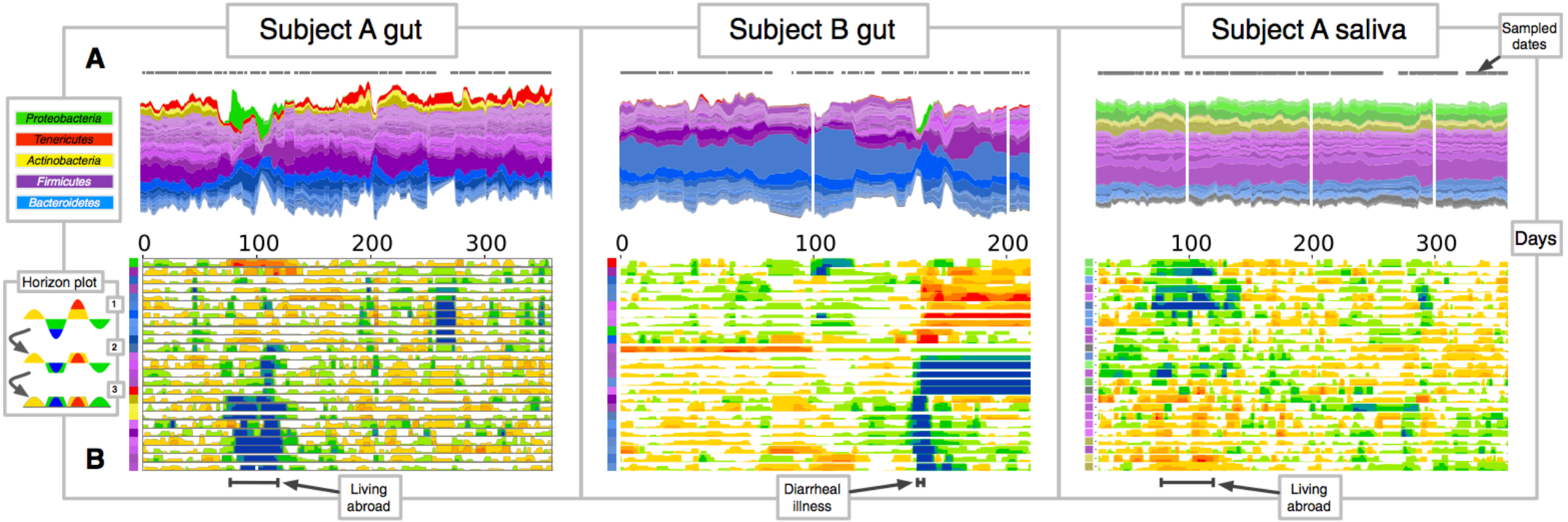
doi:10.1038/nrmicro2540

Decreased diversity with *Clostridium difficile* – assciated diarrhea



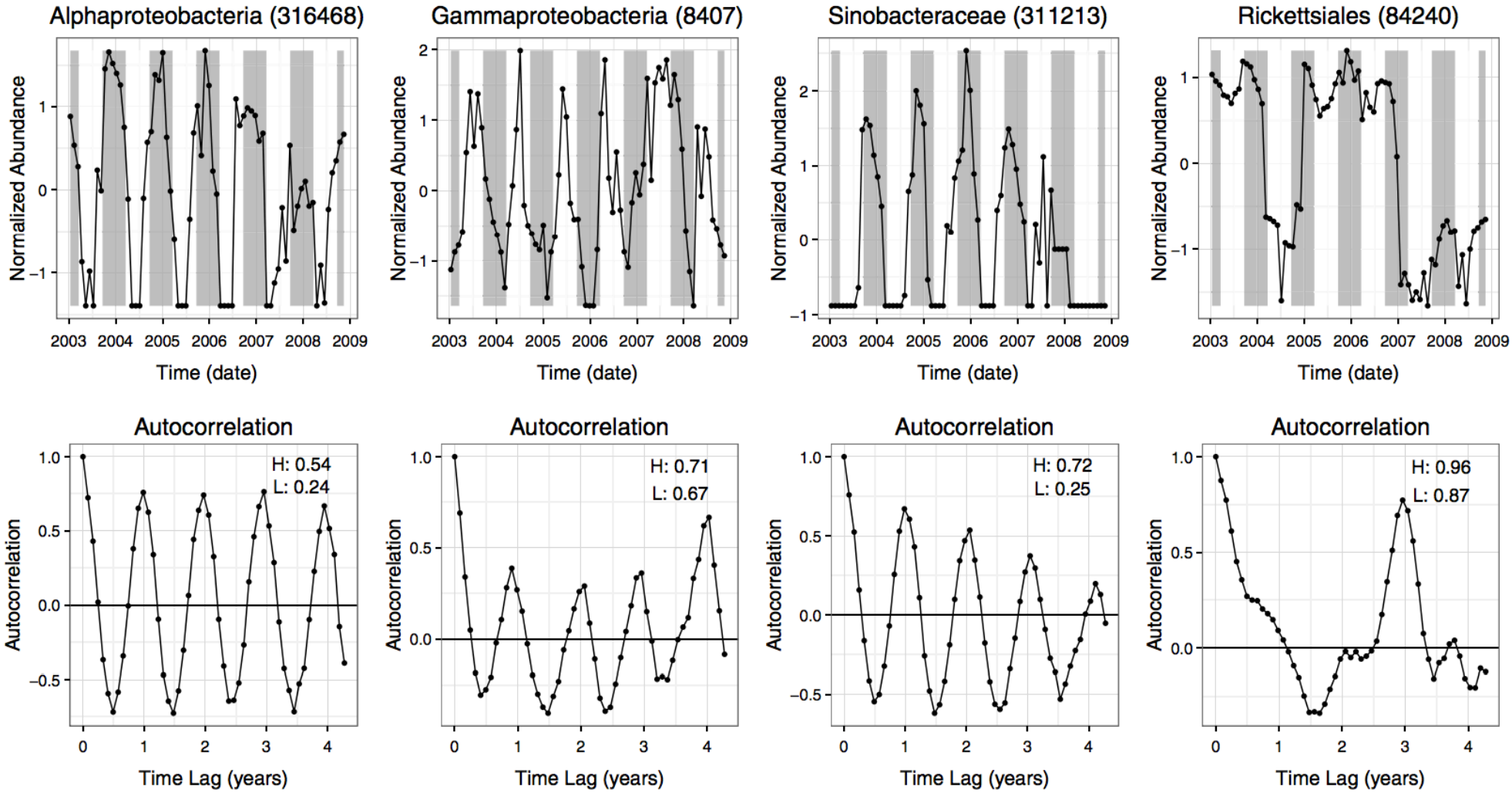
doi:10.1086/525047

Tracking microbiome on a daily scale



David et al. Genome Biology 2014, 15:R89

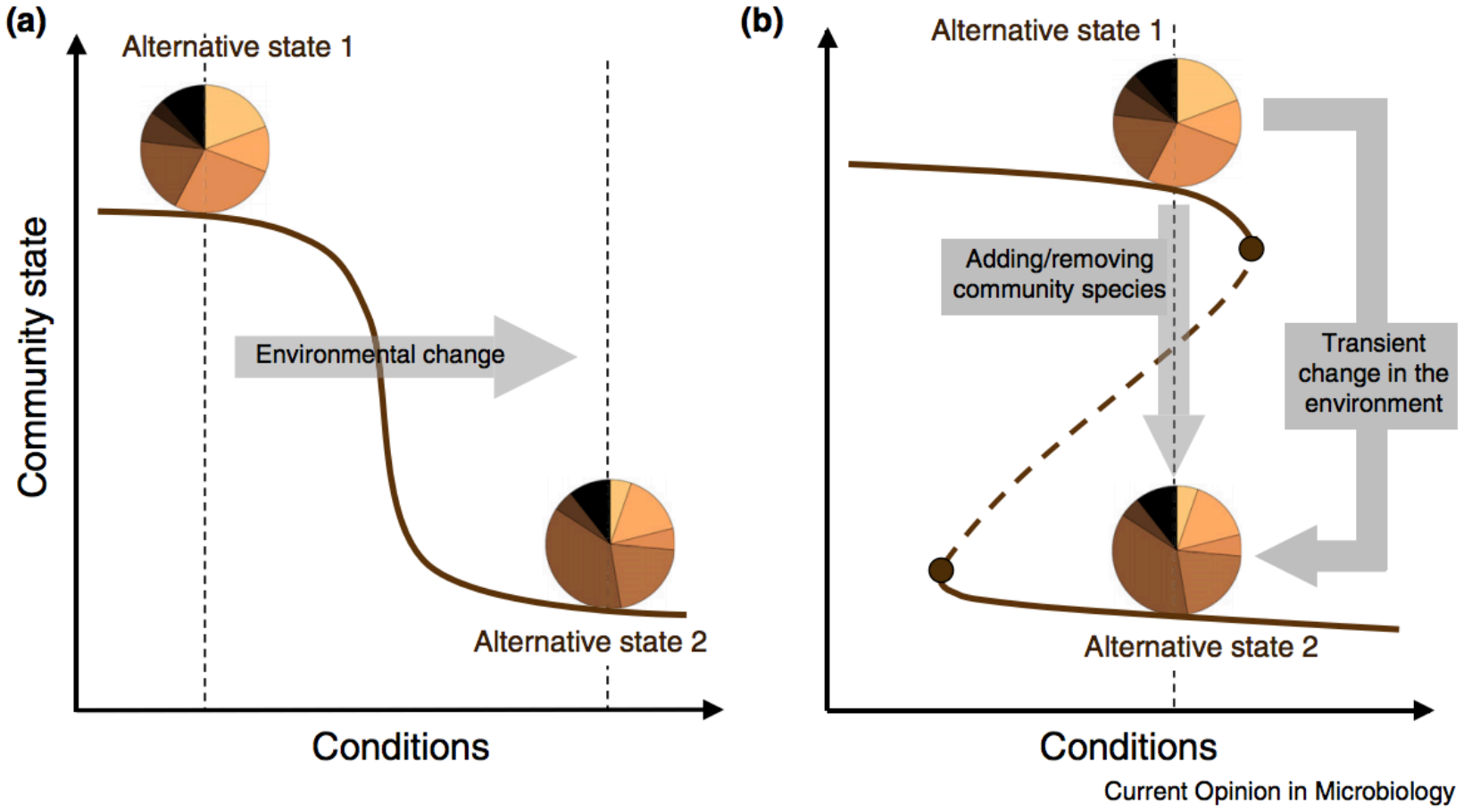
Tracking microbiome spanning 6 years



Faust et al 2015

Current Opinion in Microbiology

Tracking microbiome on a daily scale



Faust et al 2015

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

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
ISME J, 6 (2012), pp. 2153–2167

Yes

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

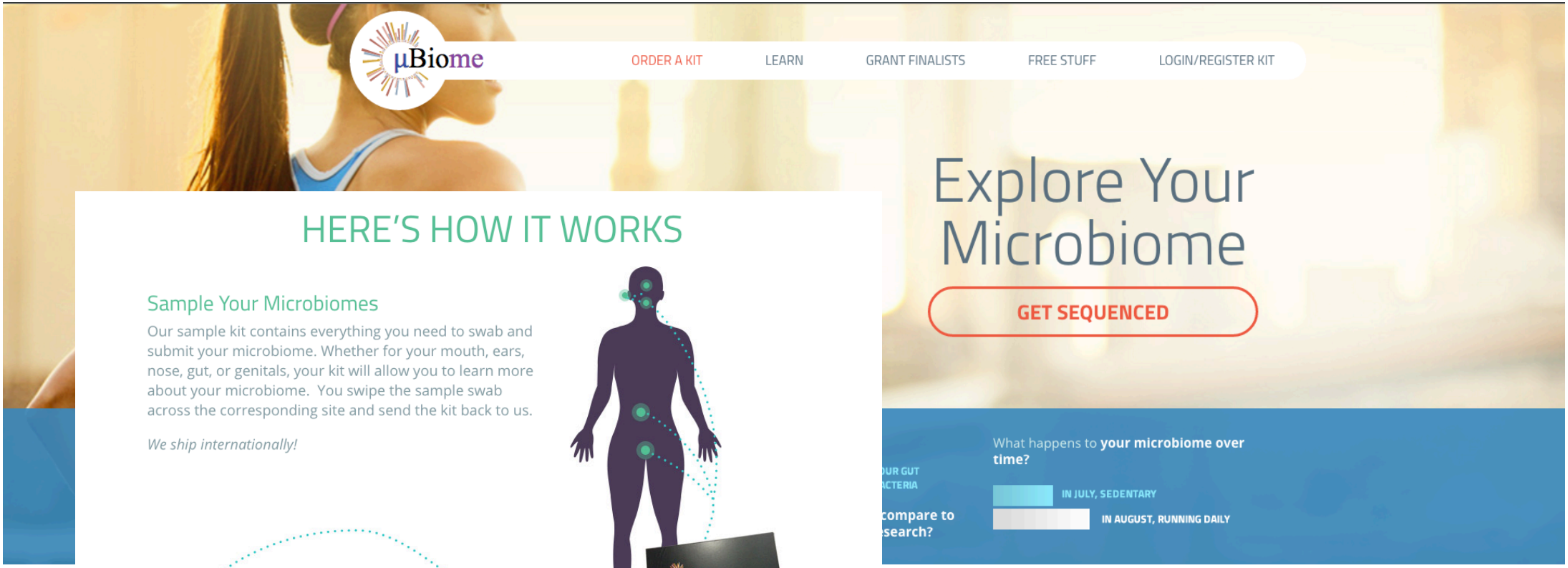
No

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
Genome Biol, 15 (2014), p. R89

Yes and No

Faust et al 2015

Future?



HERE'S HOW IT WORKS

Sample Your Microbiomes

Our sample kit contains everything you need to swab and submit your microbiome. Whether for your mouth, ears, nose, gut, or genitals, your kit will allow you to learn more about your microbiome. You swipe the sample swab across the corresponding site and send the kit back to us.

We ship internationally!

Answer a Simple Survey

Our online survey lets you ask and answer questions about your health and lifestyle. Then, **you can see how you correlate with others**. We tell you how you correlate to other users, to existing peer-reviewed studies, and to your previous samples.

Box 1 | **Ten areas of microbiome inquiry that should be pursued**

- Understanding microbiome characteristics in relation to families: which features are inherited and which are not?*
- Understanding secular trends in microbiome composition: which taxonomic groups have been lost or gained?‡
- For diseases that have changed markedly in incidence in recent decades, do changes in the microbiome have a role? Notable examples include childhood-onset asthma, food allergies, type 1 diabetes, obesity, inflammatory bowel disease and autism.*‡
- Do particular signatures of the metagenome predict risks for specific human cancers and other diseases that are associated with ageing? Can these signatures be pursued to better understand oncogenesis? (Work on *Helicobacter pylori* provides a clear example of this.)*
- How do antibiotics perturb the microbiome, both in the short-term and long-term? Does the route of administration matter?*
- How does the microbiome affect the pharmacology of medications? Can we 'micro-type' people to improve pharmacokinetics and/or reduce toxicity? Can we manipulate the microbiome to improve pharmacokinetic stability?*‡
- Can we harness knowledge of microbiomes to improve diagnostics for disease status and susceptibility?*
- Can we harness the close mechanistic interactions between the microbiome and the host to provide hints for the development of new drugs?‡
- Specifically, can we harness the microbiome to develop new narrow-spectrum antibiotics?‡
- Can we use knowledge of the microbiota to develop true probiotics (and prebiotics)?*‡

*Areas currently under investigation. ‡Proposed areas for investigation.

doi:10.1038/nrg3182

Warnings

Too much and noisy data
Bad data design everywhere
Reproducibility

A lot of potential but competitive field

But please don’t oversell the microbiome

Nature **490**, 55–60 (04 October 2012) | doi:10.1038/nature11450
Received 30 August 2011 | Accepted 27 July 2012 | Published online 26 September 2012

Don’t forget…

earth microbiome project

Additional references

A good introductory popular science video about microbiome
http://www.youtube.com/watch?v=5DTrENdWvvM

Nature series on human microbiome
http://www.nature.com/nature/focus/humanmicrobiota/

Website with lots of reading materials
http://www.slideshare.net/