Methods in microbiota research and potential pitfalls

Paulo Czarnewski

Senior Bioinformatician
Scientific Coordinator

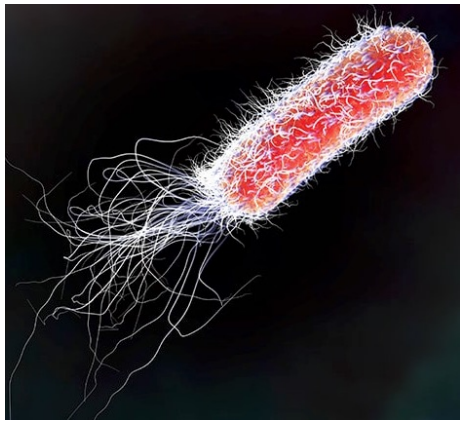
National Bioinformatics Infrastructure Sweden (NBIS)
SciLifeLab, Stockholm University

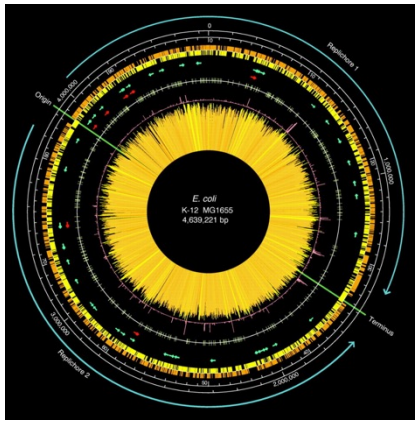
www.czarnewski.com

Introduction

Introduction to metagenomics

Single genome

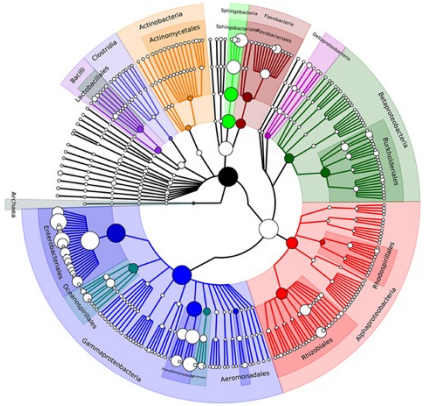




Blattner et al. (1997) *Science*

Metagenomics

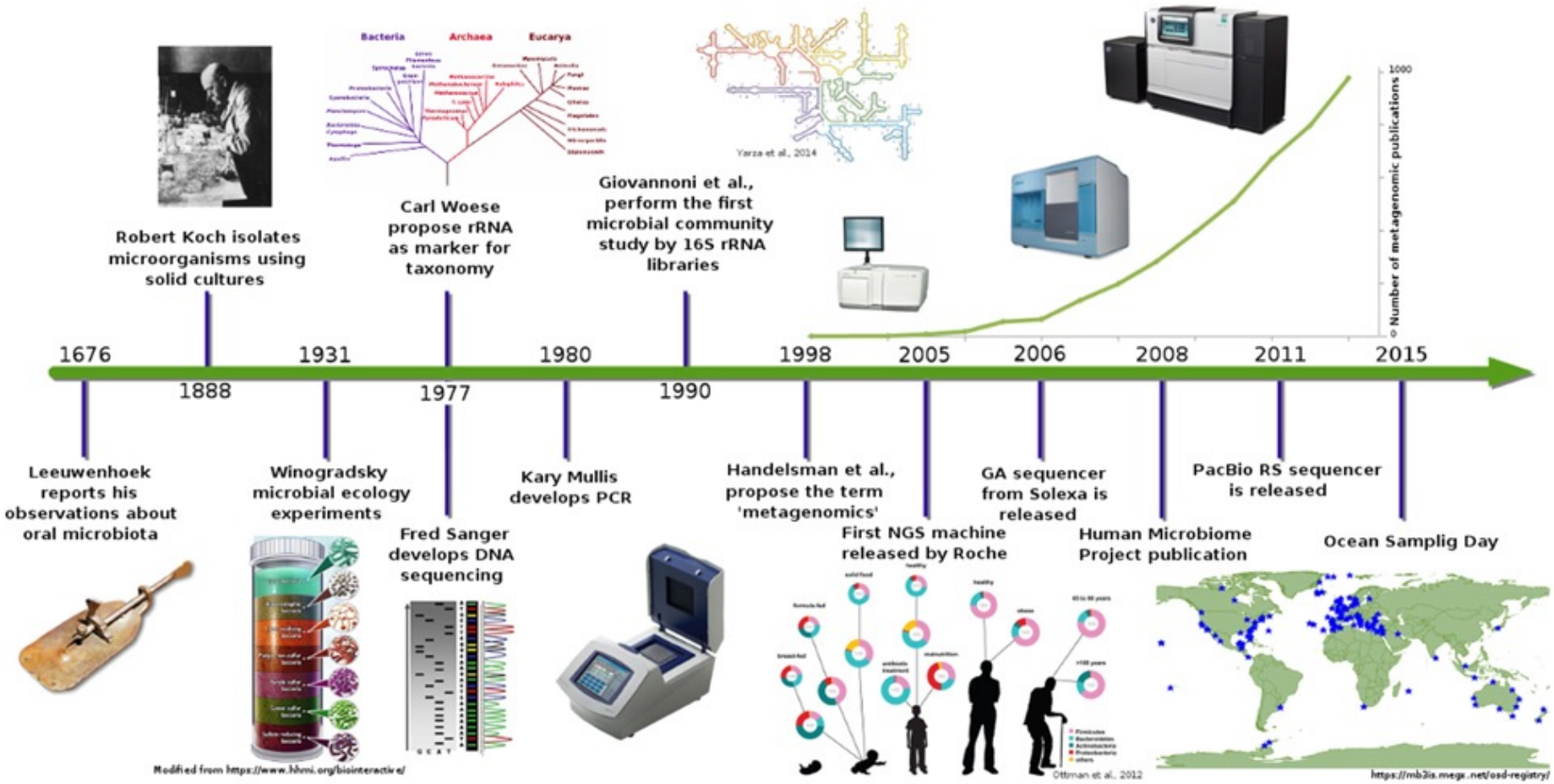
A collection of genomes



Joynson et al. (2017) *Front Microbiol*

http://www.tlc.com/tlcme/how-much-bacteria-are-really-on-your-kids-hands/

The evolution of metagenomics



Escobar-Zepeda et al (2015) *Front Genetics*

The advent of Sequencing

	Roche 454	IonTorrent PGM	Illumina	PacBio RSII[a]
Maximum read length (bp)	1200	400	300[b]	50,000
Output per run (Gb)	1	2	1000[c]	1
Amplification for library construction	Yes	Yes	Yes	No
Cost/Gb (USA Dollar)	$9538.46	$460.00	$29.30	$600
Error kind	Indel	Indel	Substitution	Indel
Error rate (%)	1	~1	~0.1	~13
Run time	20 h	7.3 h	6 days	2 h

Adapted from Glenn, T. 2014 NGS Field Guide—Table 2a—Run time, Reads, Yield|The Molecular Ecologist. Available online at: http://www.molecularecologist.com/next-gen-fieldguide-2014/ (Accessed Aug 17, 2015).

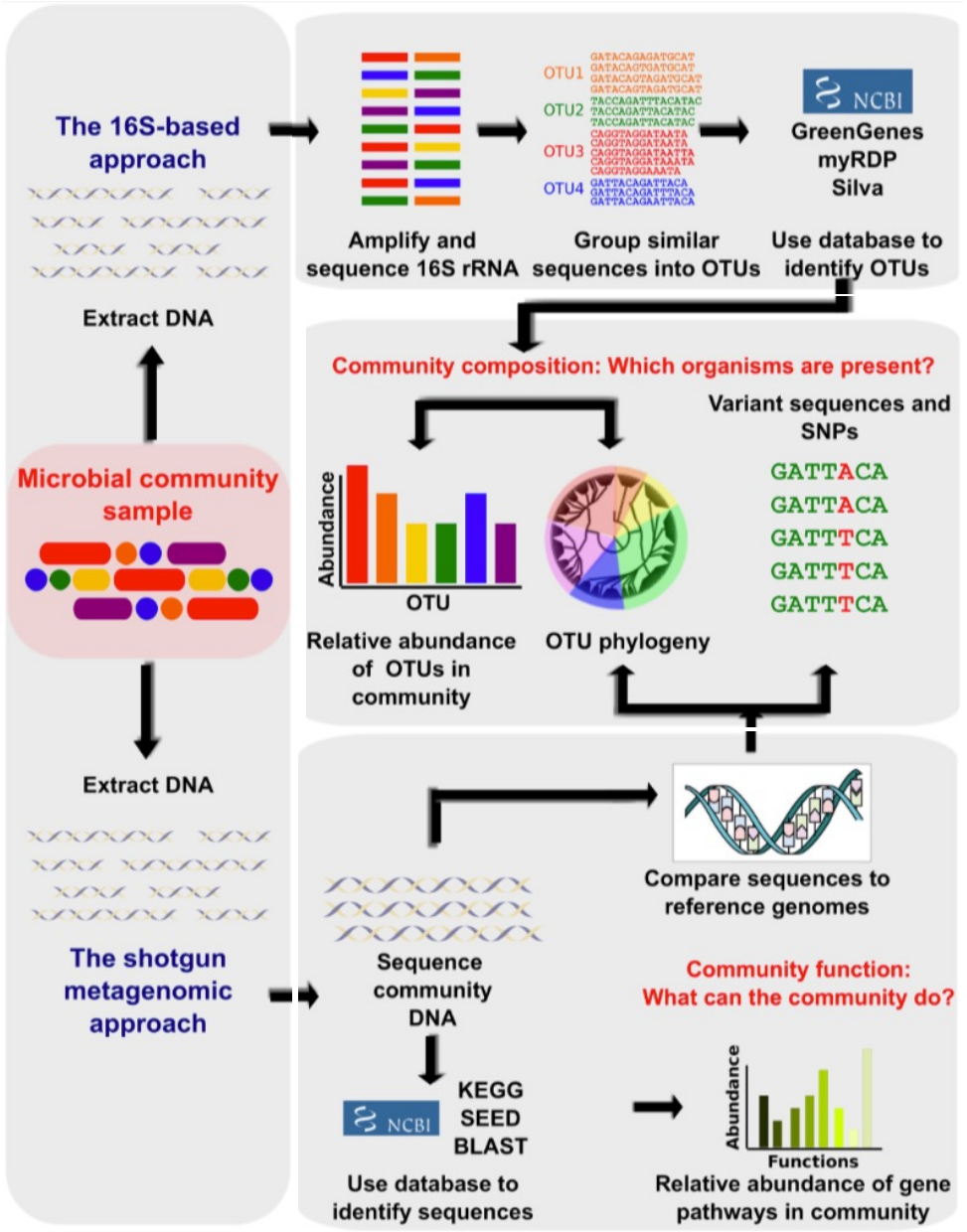
[a] *P6-C4 chemistry.*

[b] *MiSeq read length.*

[c] *Illumina HiSeq 2500 Dual flowcell yield.*

Escobar-Zepeda et al. (2015) *Front Genetics*

A basic microbiome sequencing workflow



Morgan & Huttenhower (2012) *Plos Comp Biol*

Comparison of metagenomic methods

Table 1 Overview of sequencing-based methods to characterise microbiota composition and function

	Full-length 16S rRNA gene sequencing		Targeted 16S rRNA gene amplicon sequencing	Metagenome sequencing	Metatranscriptomics	Single-cell analysis	
Technique	Clade-specific amplification of 16S rRNA gene, vector-based cloning and sequencing	Long-read-based amplification and sequencing of large 16S rRNA gene fragments	PCR-based amplification of target followed by sequencing	Sequencing of entire DNA extracted from samples	Sequencing of entire RNA extracted from samples	Cultivation-based isolation of single bacterial clones	Emulsion or droplet-based single-cell amplification
Target	Entire 16S rRNA gene	Entire 16S rRNA gene	Variable regions, eg, V1-V2; V3-V4; V4; V6	All DNA molecules	All transcribed RNA molecules	Multiple complete genomes	Single bacterial cells
Potential bias	Gene copy-number bias	Amplification bias	Primer/region-specific amplification bias; gene copy-number bias		Transcriptionally more active bacteria are over-represented	Anaerobic and hard-to-culture bacteria are under-represented	More abundant taxa over-represented
Sequencing technology	Sanger	PacBio or Oxford Nanopore	MiSeq	HiSeq/NextSeq	HiSeq/NextSeq	Any NGS technology	HiSeq/NextSeq
Advantages	Species-level resolution	High throughput; species-level resolution	Low cost; high throughput	Information on encoded functional repertoire; high-resolution taxonomic assignment; captures also viruses, fungi and archaea	Information on actively expressed functional content	Reconstruction of multiple complete genomes from complex communities	Reconstruction of multiple complete genomes from complex communities
Disadvantages	No information on functional repertoire; low throughput; much hands-on time	No information on functional repertoire; high error rates; higher cost per base than MiSeq	No information on functional repertoire; low resolution on species level	High cost; high computational burden; large amounts of unannotated data	High cost; high computational burden; large amounts of unannotated data Removal of 16S rRNA required	Tedious culturing and selection approaches (eg, media, oxygen); high cost	High cost

NGS, next generation sequencing.

Explained in more detail today

Sommer et al (2017) *Gut*

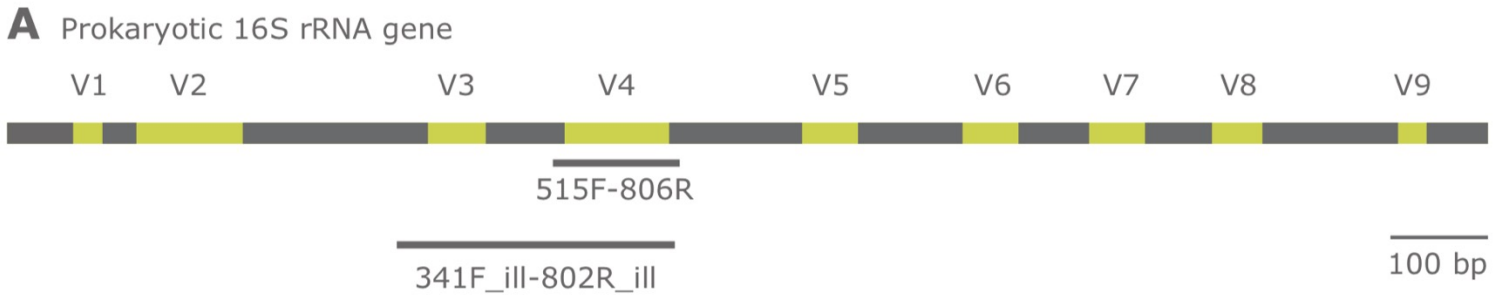
Amplicon Sequencing

Looking at short known pieces

Conserved genes for study of phylogeny

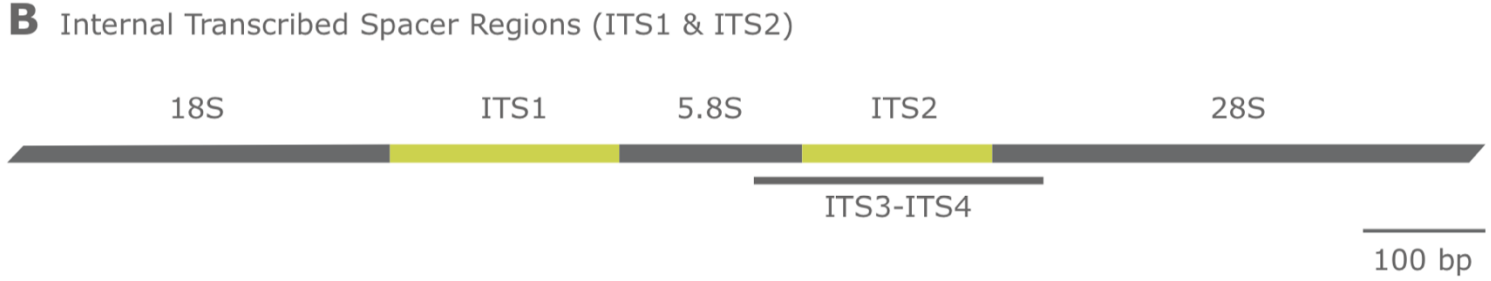
16S

Archaea **Bacteria**

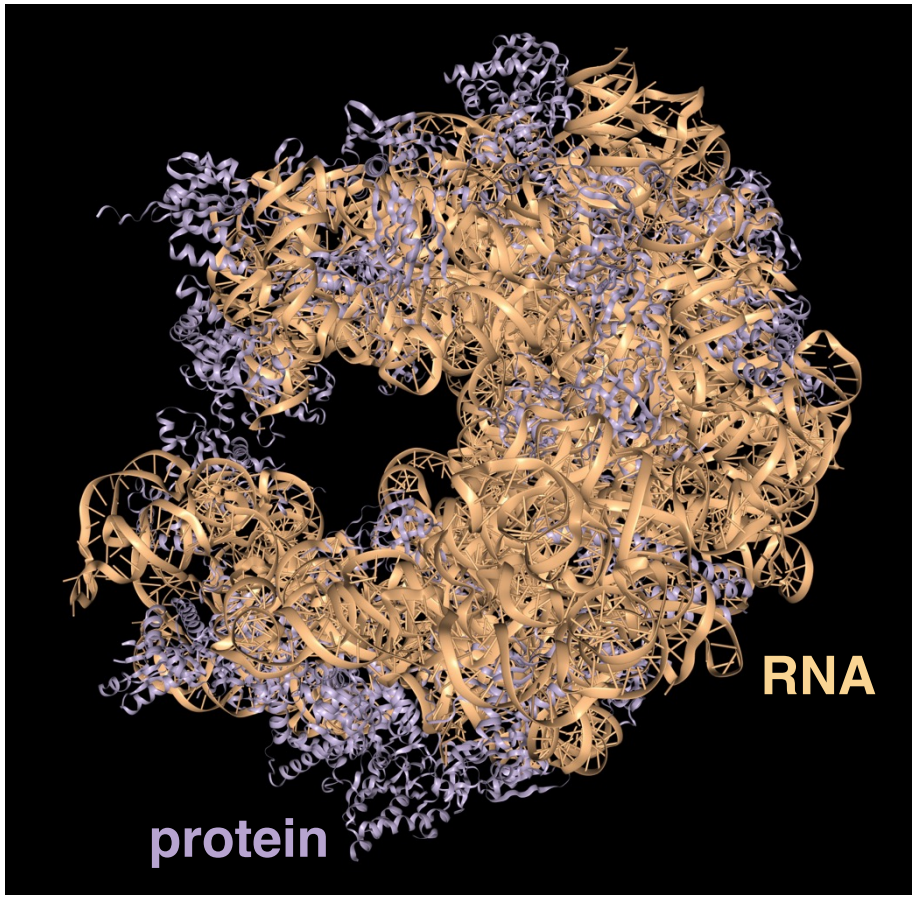


ITS

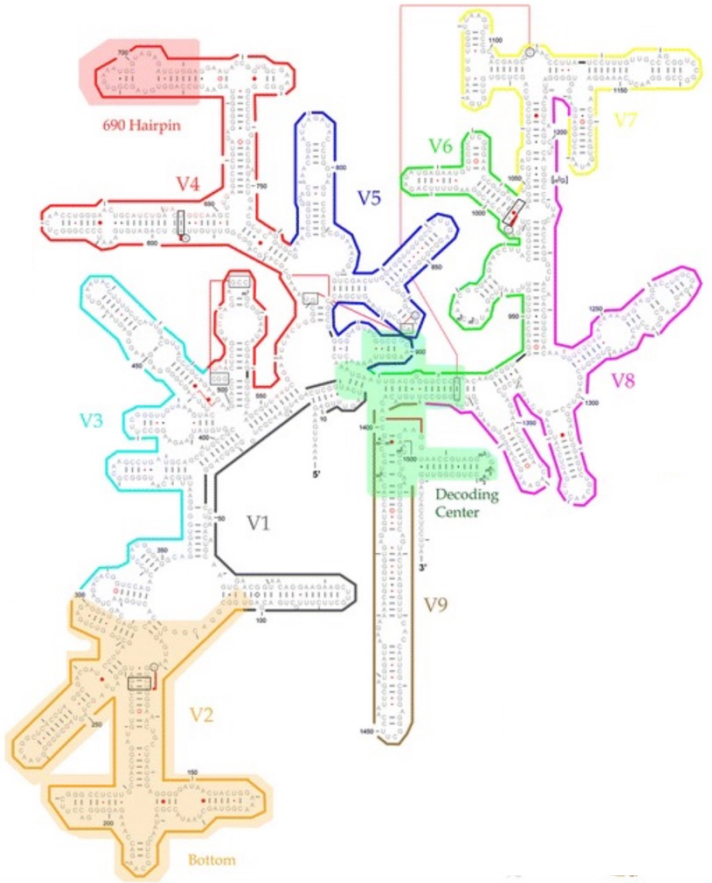
Archaea **Bacteria** **Eukaryotes (Fungi** **Plants** **Mammals)**



Why sequencing ribosomal rRNA gene?



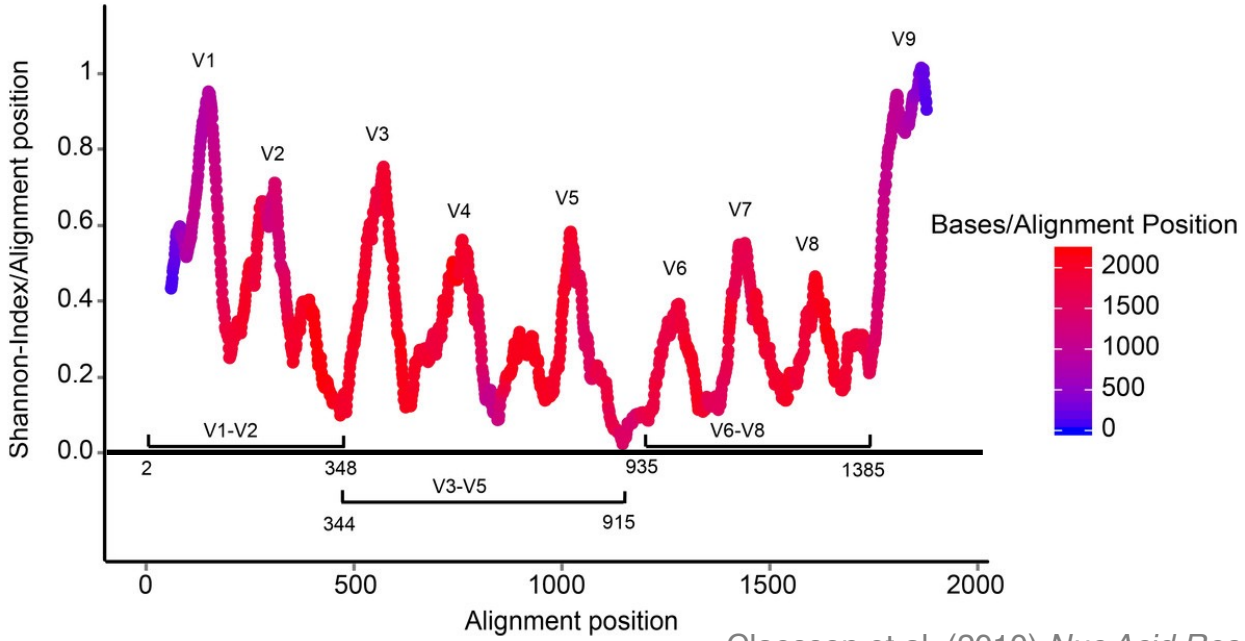
E. coli 16S (PBD Y4BB)



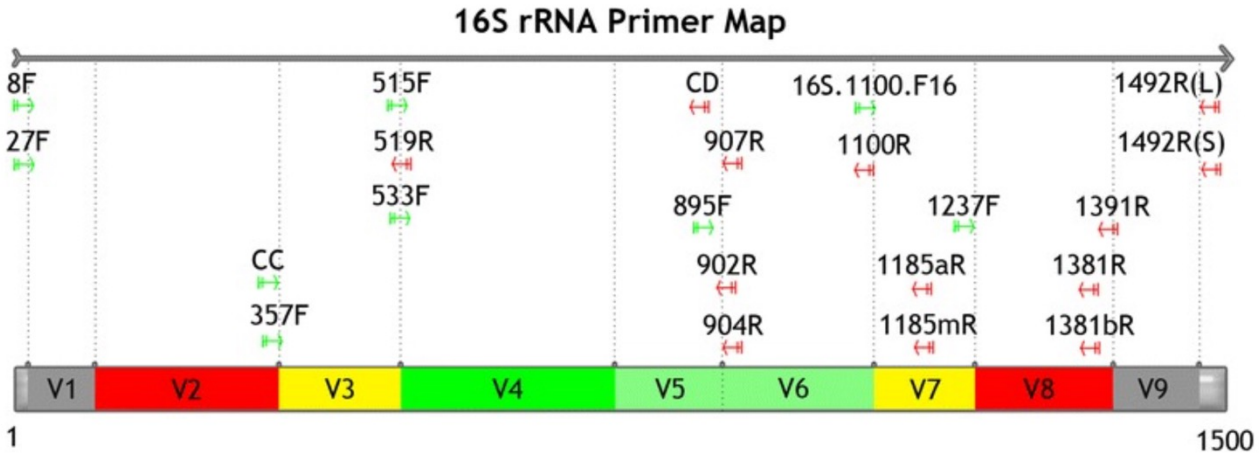
Yang et al (2016) *BMC Bioinformatics*

Present in most species and has conserved/variable regions among species

The variable regions of 16S rRNA

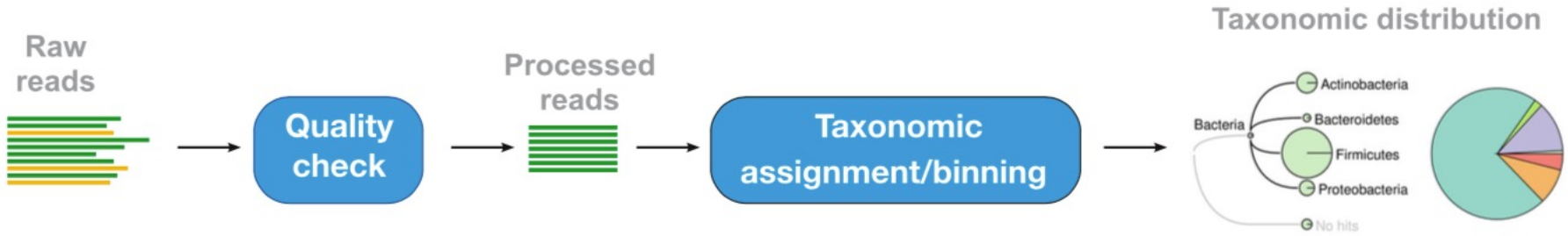


Claesson et al. (2010) *Nuc Acid Res*



Yang (2016) *BMC Bioinformatics*

A basic pipeline for amplicon data analysis



Pre-processing of sequencing reads:
- Read quality filtering
- Removing Chimera sequences

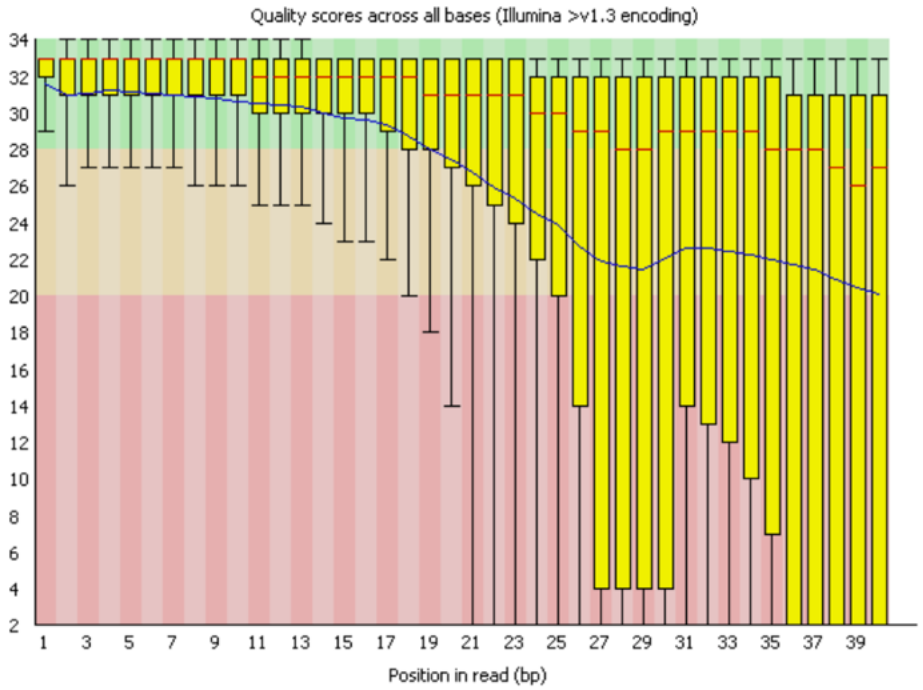
Clustering sequences into OTUs
- Removing Singletons

Classification into taxonomic groups

Calculation of intra- and inter- sample diversity

Prediction of community function using Pycrust

Compare groups from your experiment
This depends on the goal in your experiment



Balvociute & Hason (2017) *BMC Genomics*

Bioinformatics analysis pipelines

NATURE METHODS | VOL.7 NO.5 | MAY 2010

QIIME allows analysis of high-throughput community sequencing data

NATURE METHODS | VOL.10 NO.10 | OCTOBER 2013

UPARSE: highly accurate OTU sequences from microbial amplicon reads

NATURE METHODS | VOL.13 NO.7 | JULY 2016

DADA2: High-resolution sample inference from Illumina amplicon data

NATURE BIOTECHNOLOGY | VOL 37 | AUGUST 2019 | 848-857

Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2

OTUs and ASVs

OUT = Operational Taxonomy Units
ASV = Amplicon Sequence Variants

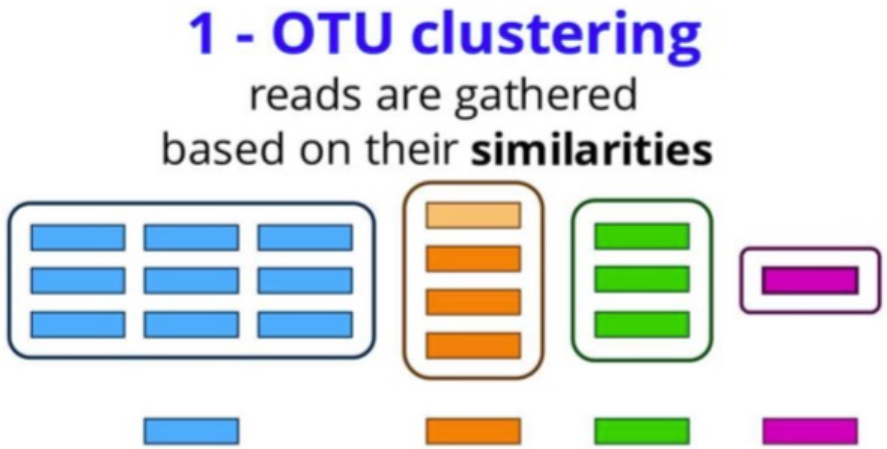
Choose a similarity threshold: **97%** (ASV≈**100%**)

Calculate distance between **sequences**

Define cluster **centroids**
Representative sequence

Continue clustering until all sequences are clustered into **OTUs/ASV**
Some of them might still remain alone (singleton), and should be removed

Count how many sequences is in each OUT/ASV:
OUT/ASV_table.txt



Siegwald et al (2017) *PlotsOne*

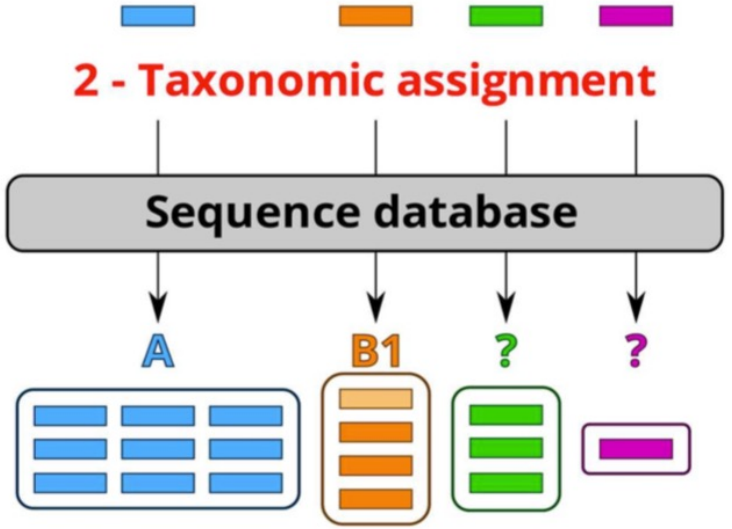
Taxonomic assignment

Many choices …

BLAST
Alignment tool

Phylogenetics placement

Naïve Bayes Classifier
Machine learning approach



Siegwald et al (2017) *PlotsOne*

Most used for 16S profiling

Table 1 Overview of five taxonomic classifications

Taxonomy	Type	No. of nodes	Lowest rank	Latest release
SILVA	Manual	12,117	Genus	Sep 2016
RDP	Semi	6,128	Genus	Sep 2016
Greengenes	Automatic	3,093	Species	May 2013
NCBI	Manual	1,522,150	Species	Today[a]
OTT	Automatic	2,627,066	Species	Sep 2016

[a]For the analyses we have used NCBI taxonomy as published on 5th Oct 2016

Balvociute & Hason (2017) *BMC Genomics*

Within sample “alpha”-diversity

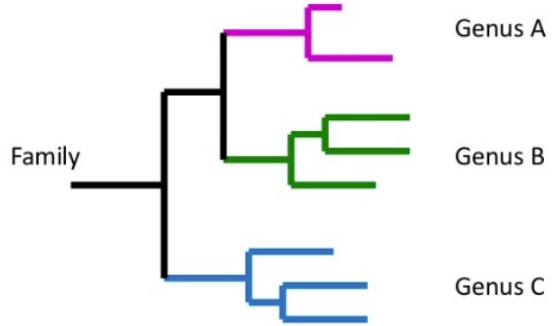
There are many ways and interpretations …

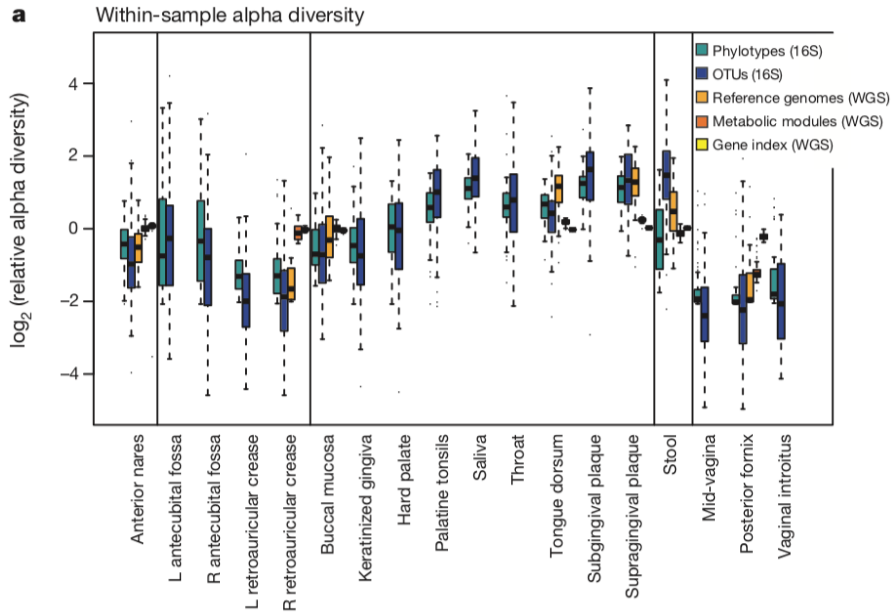
Richness (OTU counts)
How many unique OTUs in the sample

Simpson index
Probability of sampling 2 sequences from the same taxa

Phylogenetic diversity
Sum of branches length in the phylogenetic tree

Other methods:
Shannon, Inverse Simpson, Gini-Simpson, Renyi Entropy





Huttenhower et al (2012) *Nature*

Within sample “beta”-diversity

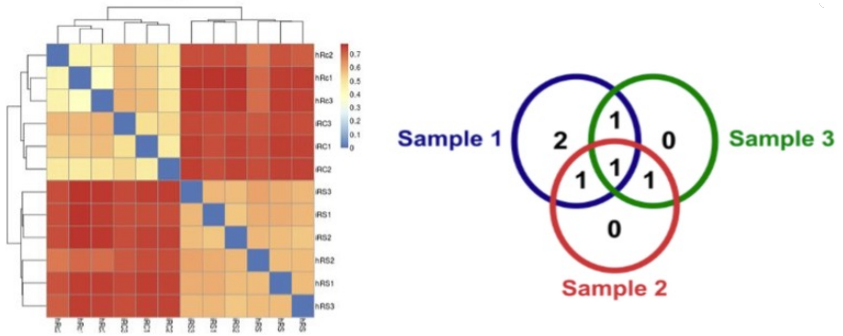
There are many ways and interpretations …

Perform pair-wise comparisons to build a dissimilarity matrix (analog to a correlation matrix)

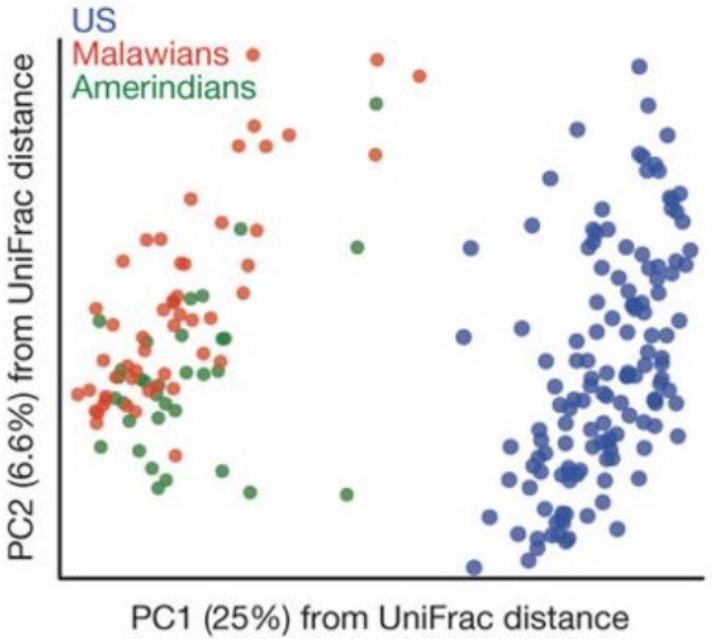
Summarize the matrix using patterns of covariance or hierarchical similarity (Sample Clustering)

Beta-diversity can be:
- Phylogenetic or non-phylogenetic
- Weighted or unweighted

Common methods:
- *Bray-Curtis (weighed, non-phylogenetic)*
- *Jaccard (weighed, non-phylogenetic)*
- *Weighted UniFrac (weighed, phylogenetic)*
- *Unweighted UniFrac (unweighed, phylogenetic)*



Morgan & Huttenhower (2012) *Plos Comp Biol*



Yatsunenko et al (2012) *Nature*

Predicting function from taxomony

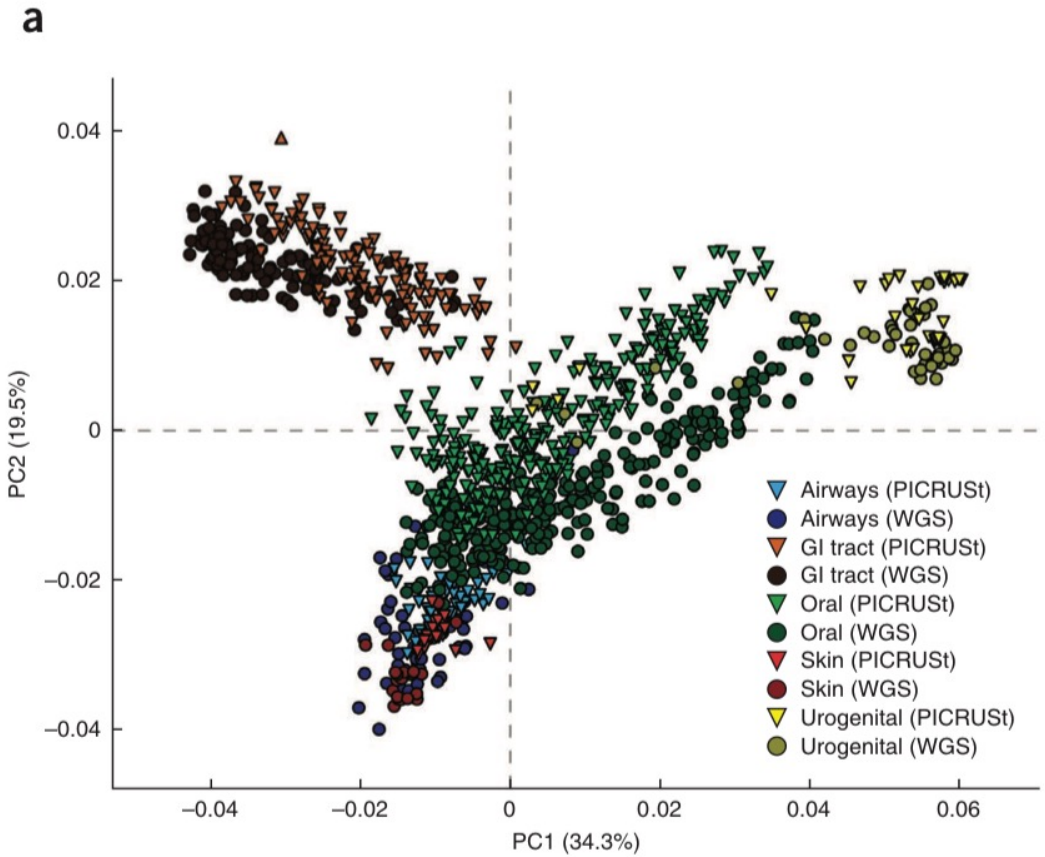
PICRUSt2

16S provides information about which bacterial taxa is in the sample

Many bacterial genomes are available online

PYCRUSt reconstructs a metagenome based on the number of bacteria identified by 16S

Then it performs KEGG enrichment from the reconstructed metagenome

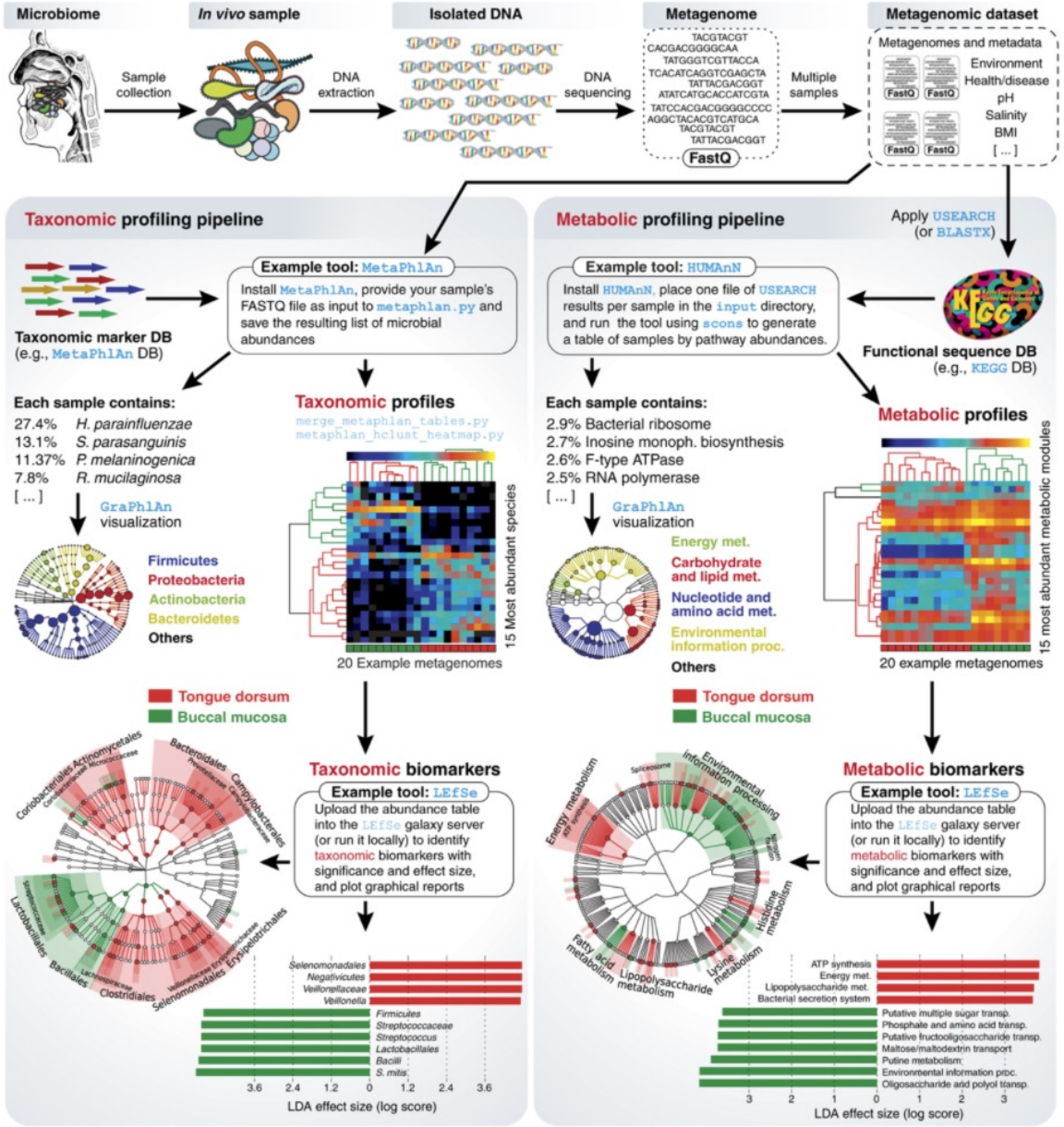


Langille et al (2013) *Nat Biotechnol*

Metagenome Sequencing

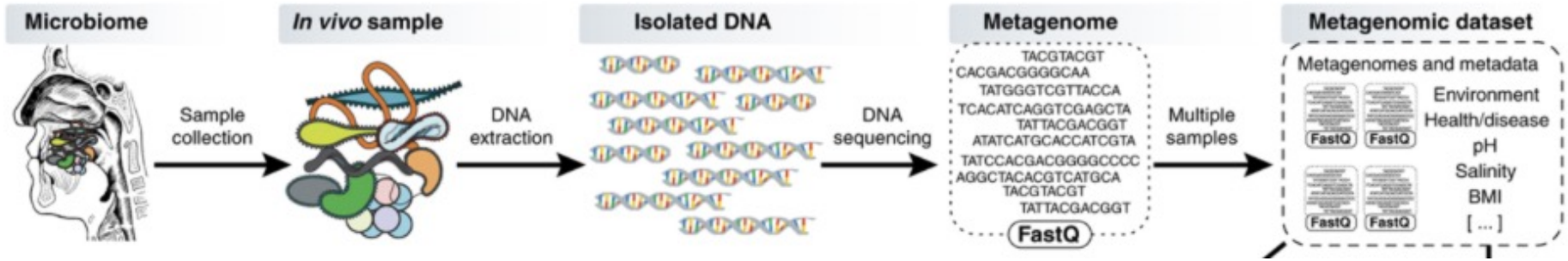
Quantifying the unknown

A basic pipeline for metagenomics data analysis



Segata et al (2013) *Mol Systems Biol*

Common genomes assembling strategies



Segata et al (2013) *Mol Systems Biol*

Nucleic Acids Research, 2012, Vol. 40, No. 20 **e155** Nucleic Acids Research

MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads

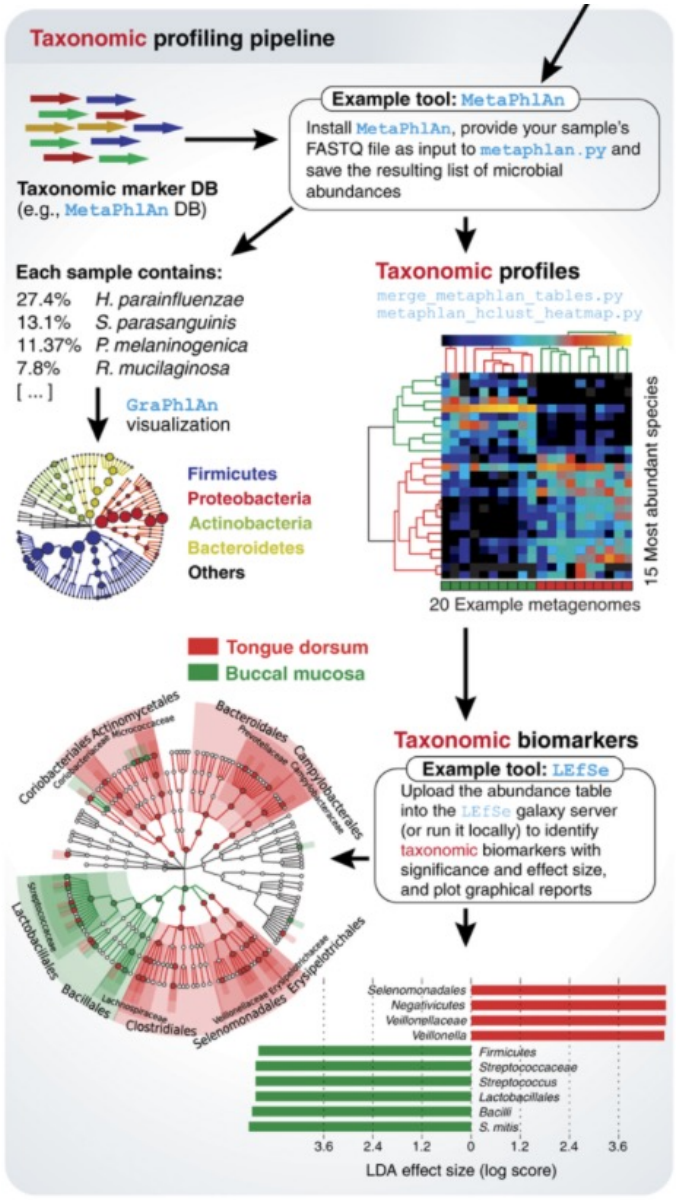
MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph

Bioinformatics, 31(10), 2015, 1674–1676
doi: 10.1093/bioinformatics/btv033
Advance Access Publication Date: 20 January 2015
Applications Note

OXFORD

Other tools:
Eggnog mapper
binning tool - maxbin

Taxonomy identification methods



Segata et al (2013) *Mol Systems Biol*

MetaPhlAn

NATURE METHODS | VOL.9 NO.8 | AUGUST 2012 | 811

Metagenomic microbial community profiling using unique clade-specific marker genes

Kraken
Kraken2
KrakenUniq
Braken

Genome Biology
Wood and Salzberg *Genome Biology* 2014, **15**:R46
METHOD Open Access

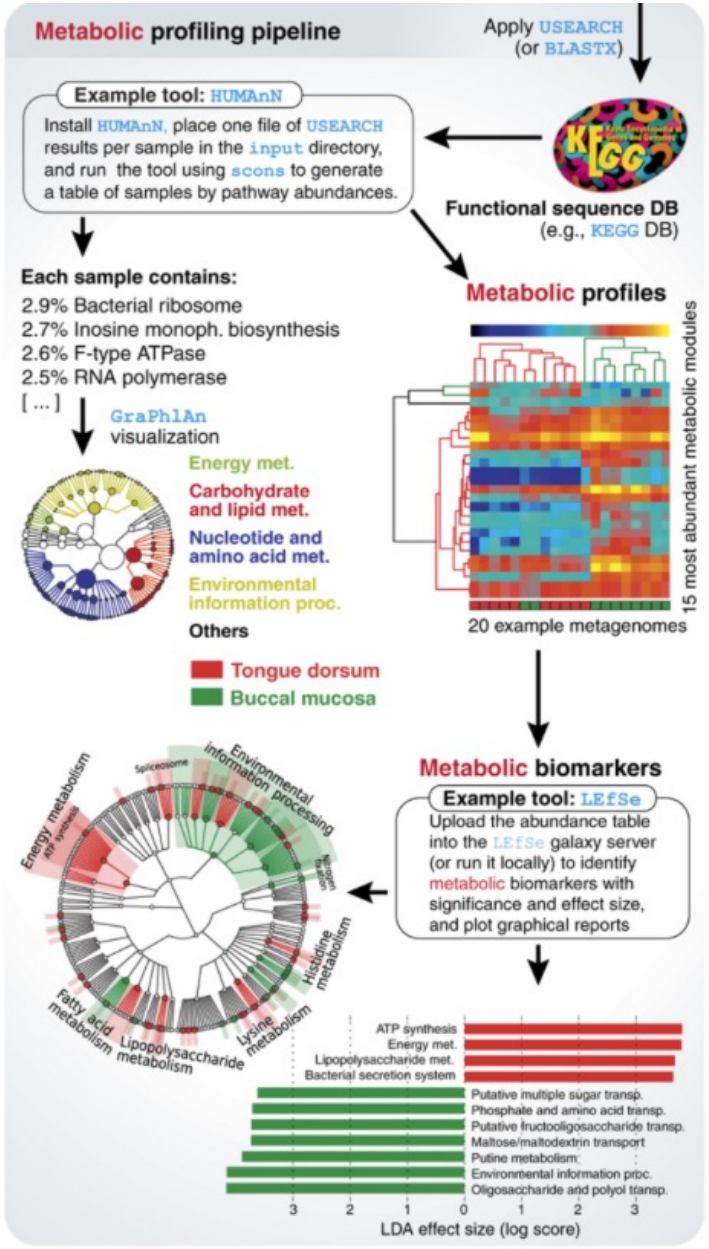
Kraken: ultrafast metagenomic sequence classification using exact alignments

Centrifuge

Genome Research 1
Method
CSH PRESS GENOME RESEARCH

Centrifuge: rapid and sensitive classification of metagenomic sequences

Metagenomic Metabolic profiling

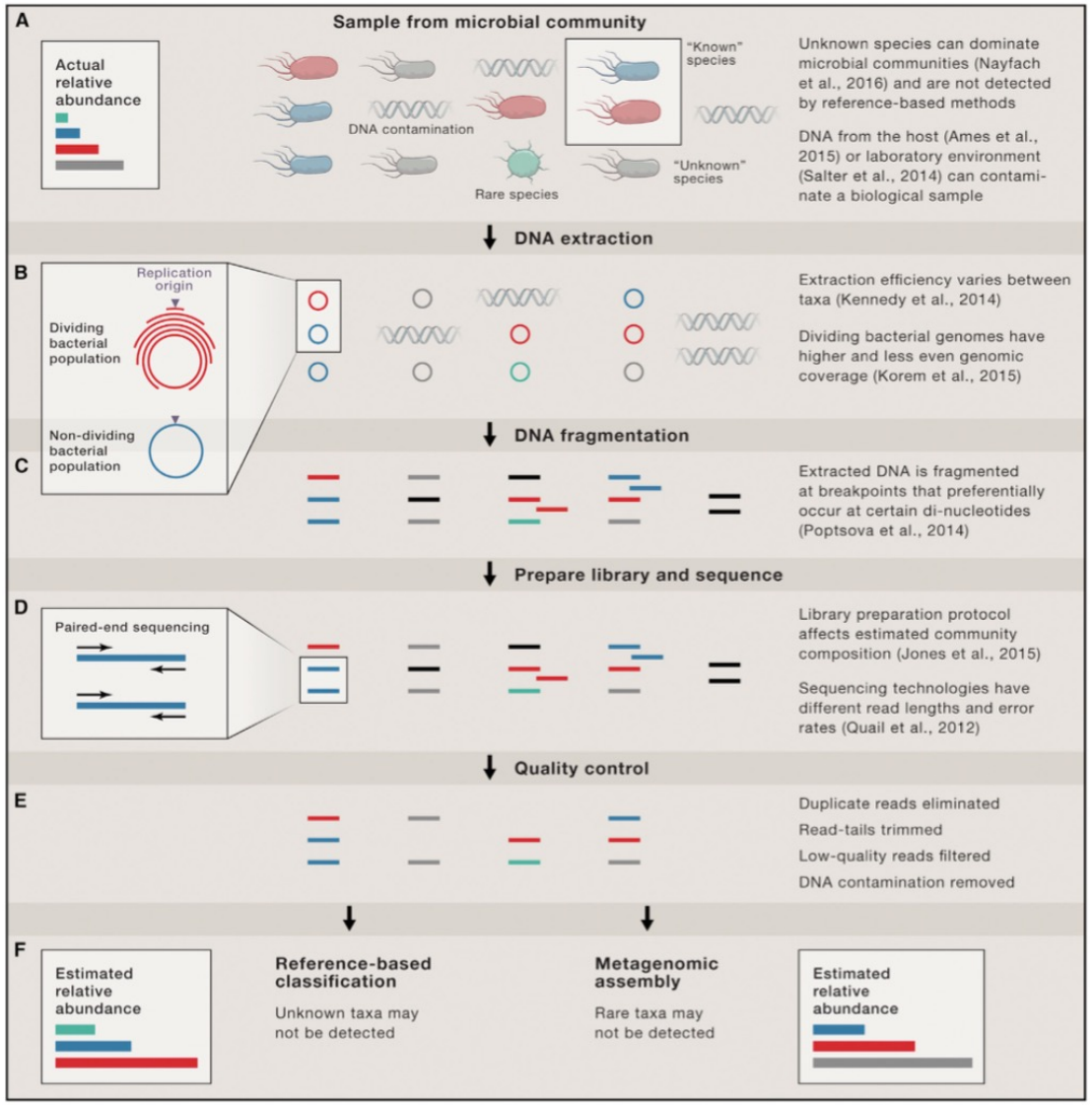


Segata et al (2013) *Mol Systems Biol*

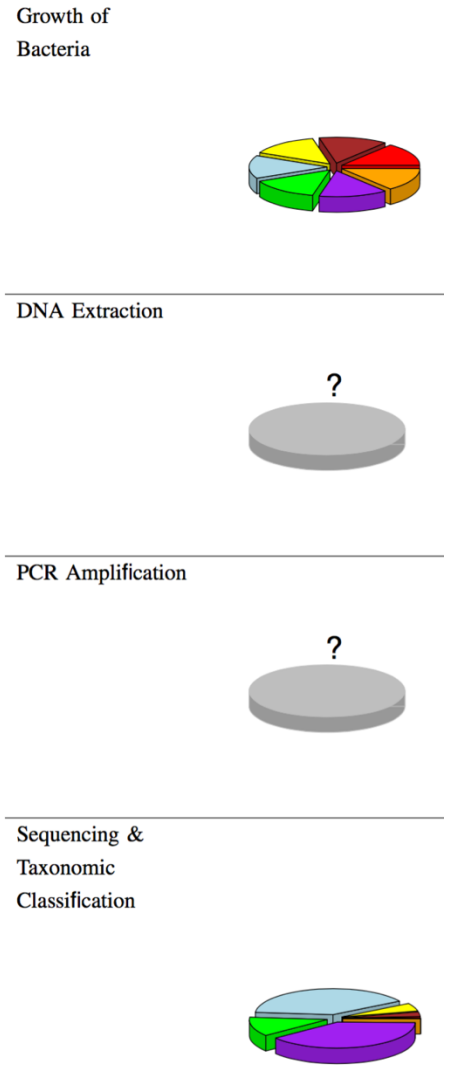
Potential pitfalls and misconceptions

The impact of experimental design

What you quantify is not exactly what is there

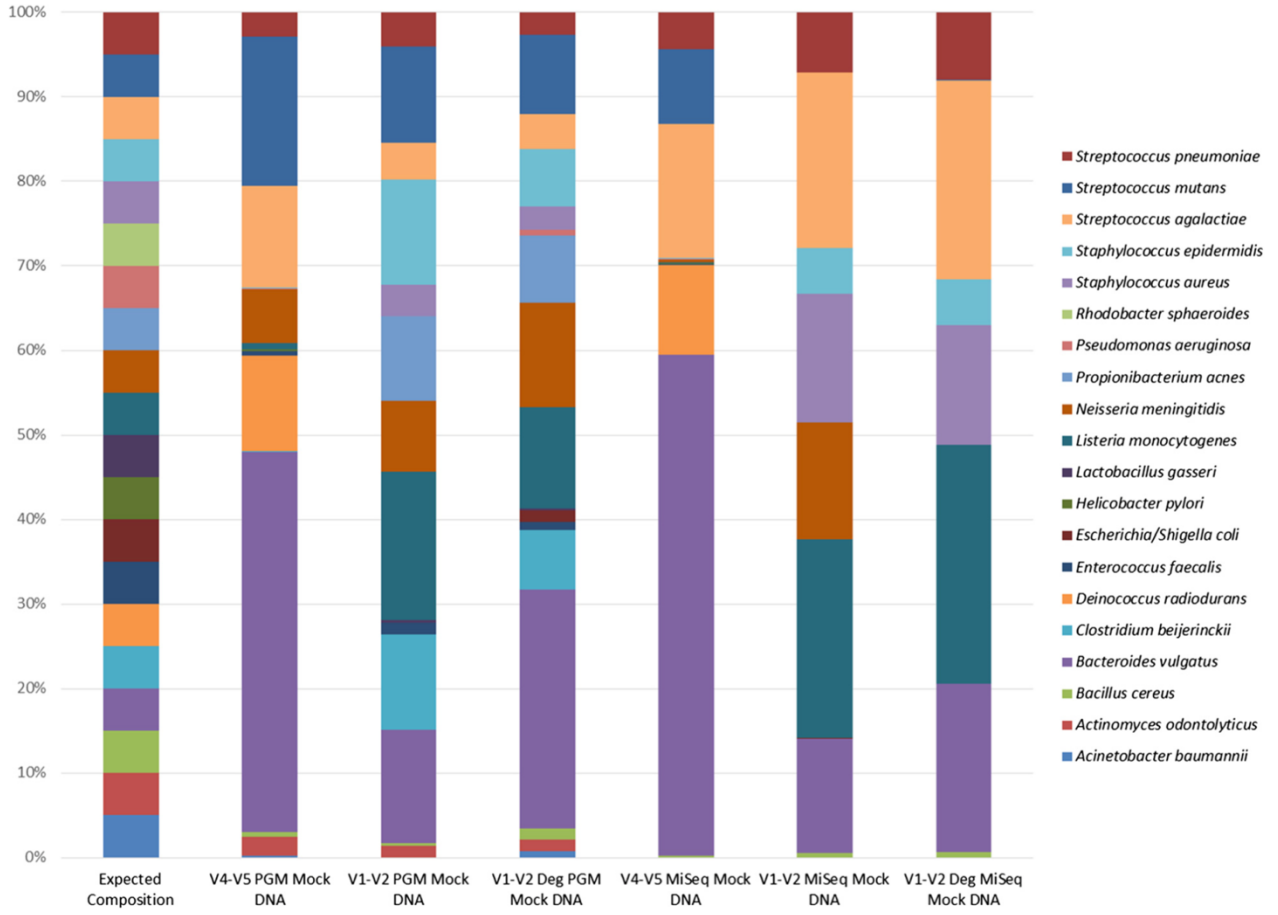


Nayfach, Pollard (2018) *Cell*



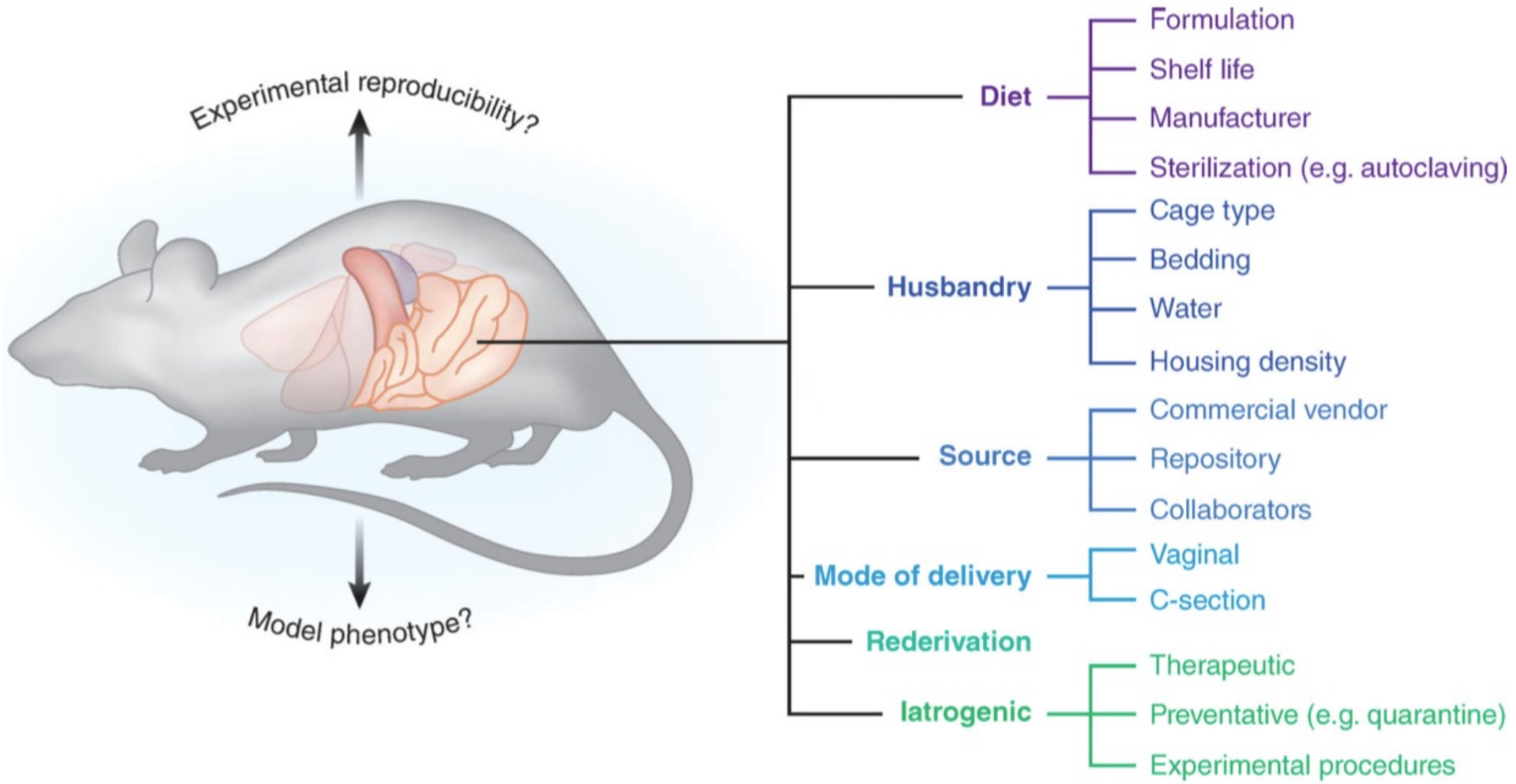
Brooks et al (2015) *BMC Microbiol*

The extraction method impact on your “microbiota"



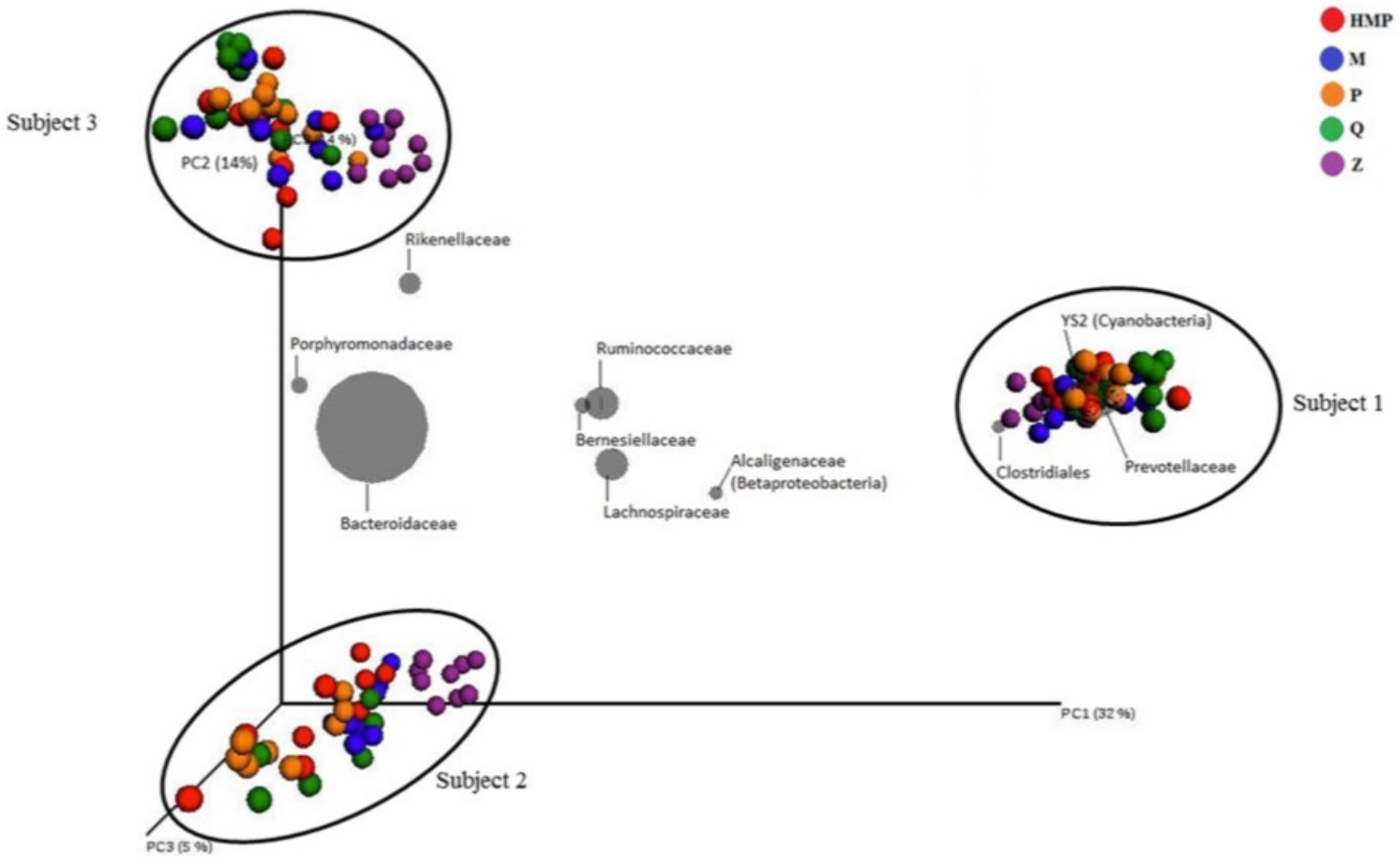
Fouthy et al (2016) *BMC Microbiol*

The microbiota is highly sensitive to batches



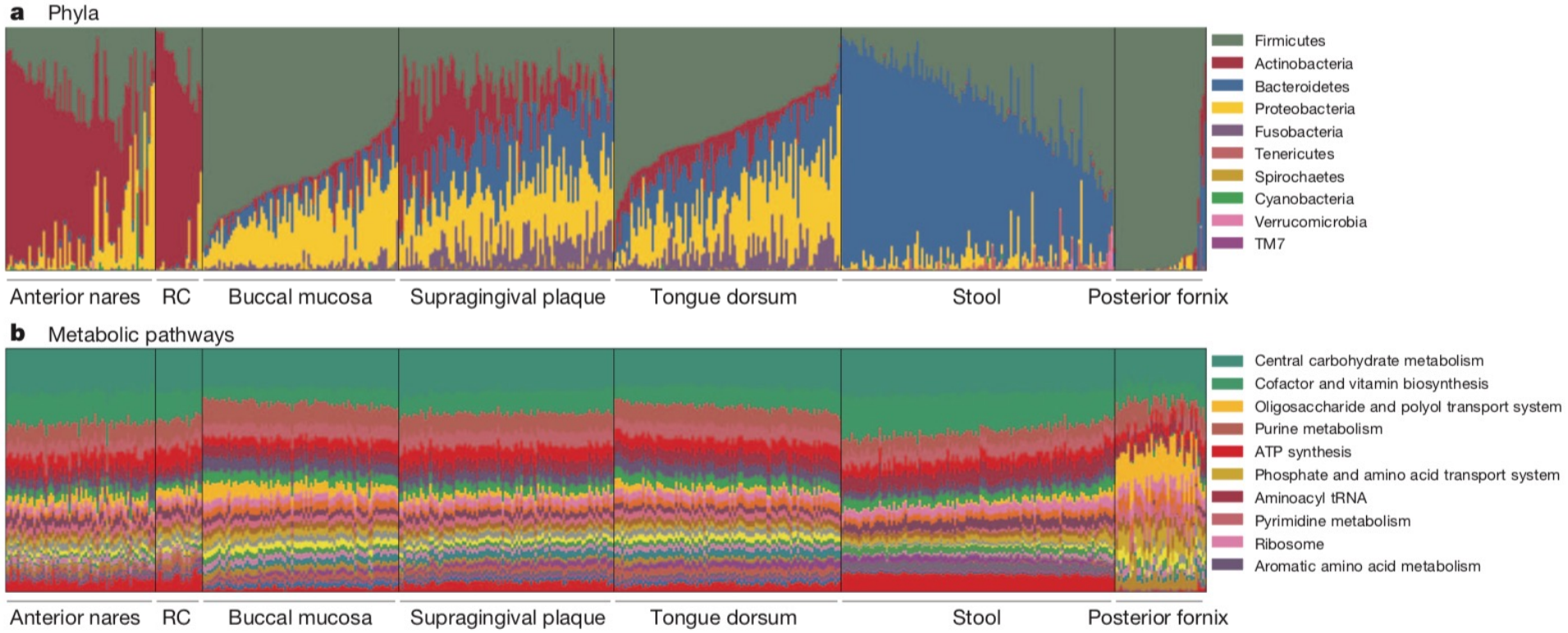
Franklin & Ericsson (2017) *Lab Anim*
Ubeda et al (2012) *J Exp Med*

Microbiome data is highly variable

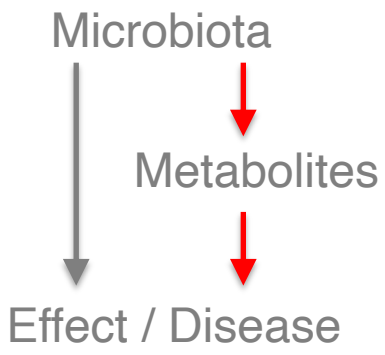


Mackenzie et al (2015) *Front Microbiol*

Different microbiota can have similar function



Huttenhower et al (2012) *Nature*



Big changes in microbial community might have little effect on downstream processes

Importance of testing causality!

Ni et al (2017) *Nat Rev*

Be aware for cofounding factors

E.g.: stool consistency, diet habits, age, gender, antibiotics, etc …

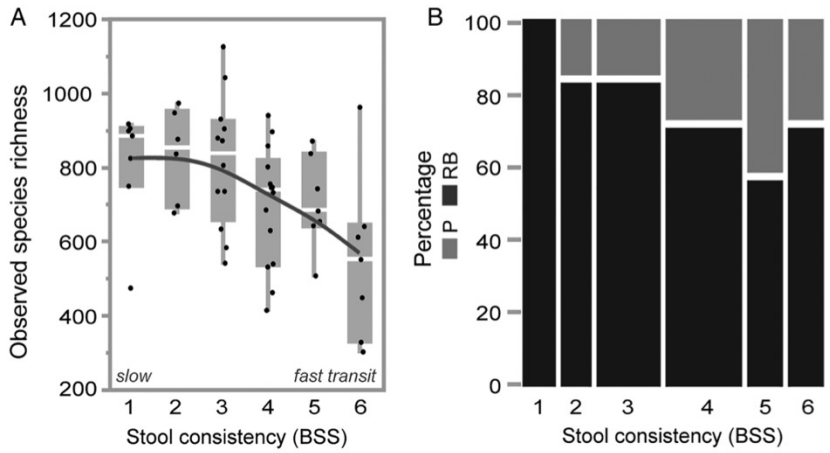


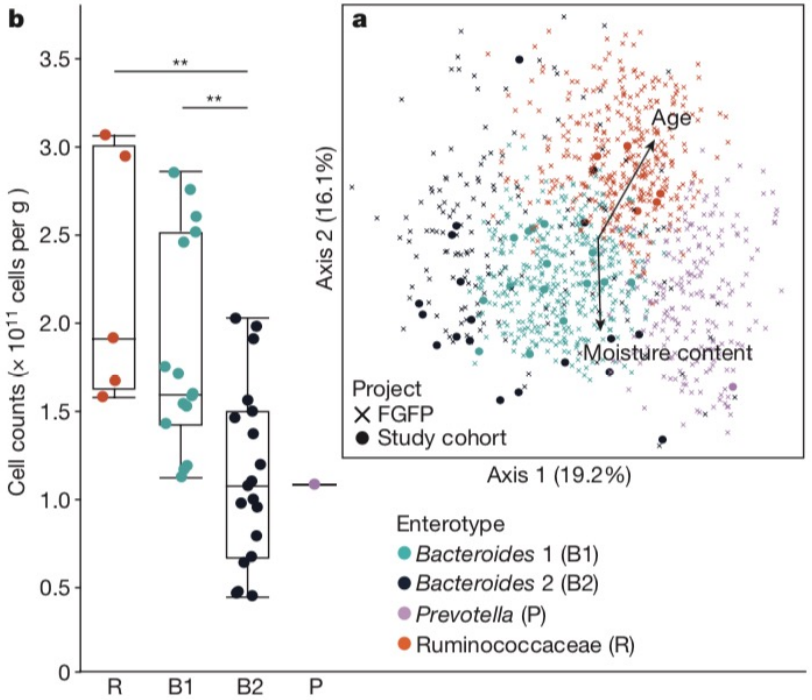
Table 1 Genera abundances significantly correlated with stool consistency

	Total data set		RB enterotype		P enterotype	
Genus	**r**	**q Value**	**r**	**q Value**	**r**	**q Value**
Akkermansia	−0.504	0.0722	−0.528	0.0342	−0.078	0.6696
Bacteroides	0.177	0.6134	0.460	0.0718	−0.048	0.7803
Butyricimonas	−0.348	0.0722	−0.406	0.0718	−0.264	0.8620
Methanobrevibacter	−0.126	0.0722	−0.095	0.0718	−0.134	0.7978
Methanosphaera	−0.305	0.0966	−0.307	0.1318	NA	NA
Odoribacter	−0.094	0.0966	−0.048	0.2232	−0.284	0.8551
Oxalobacter	−0.424	0.0722	−0.460	0.0350	−0.051	0.6127

Genera abundances significantly correlated with stool consistency (BSS) ($q<0.1$) in the total data set, the RB enterotype, or the P enterotype. Spearman's rank order correlation with Benjamini-Hochberg false discovery rate correction.
BSS, Bristol Stool Scale; NA, not assigned; RB, Ruminococcaceae-Bacteroides.

Vandeputte et al (2015) *Gut*

Analysis of healthy controls



Vandeputte et al (2018) *Nature*

Cofounding factors: the IBD case

Microbial diversity is reduced in IBD patients

IBD patients tend to have more loose stool

Highlighted are the top bacteria associated with loose stool

Vandeputte et al (2015) *Gut*

Table 1 | Differential abundance in specific taxa according to disease phenotype comparisons (DESeq2).

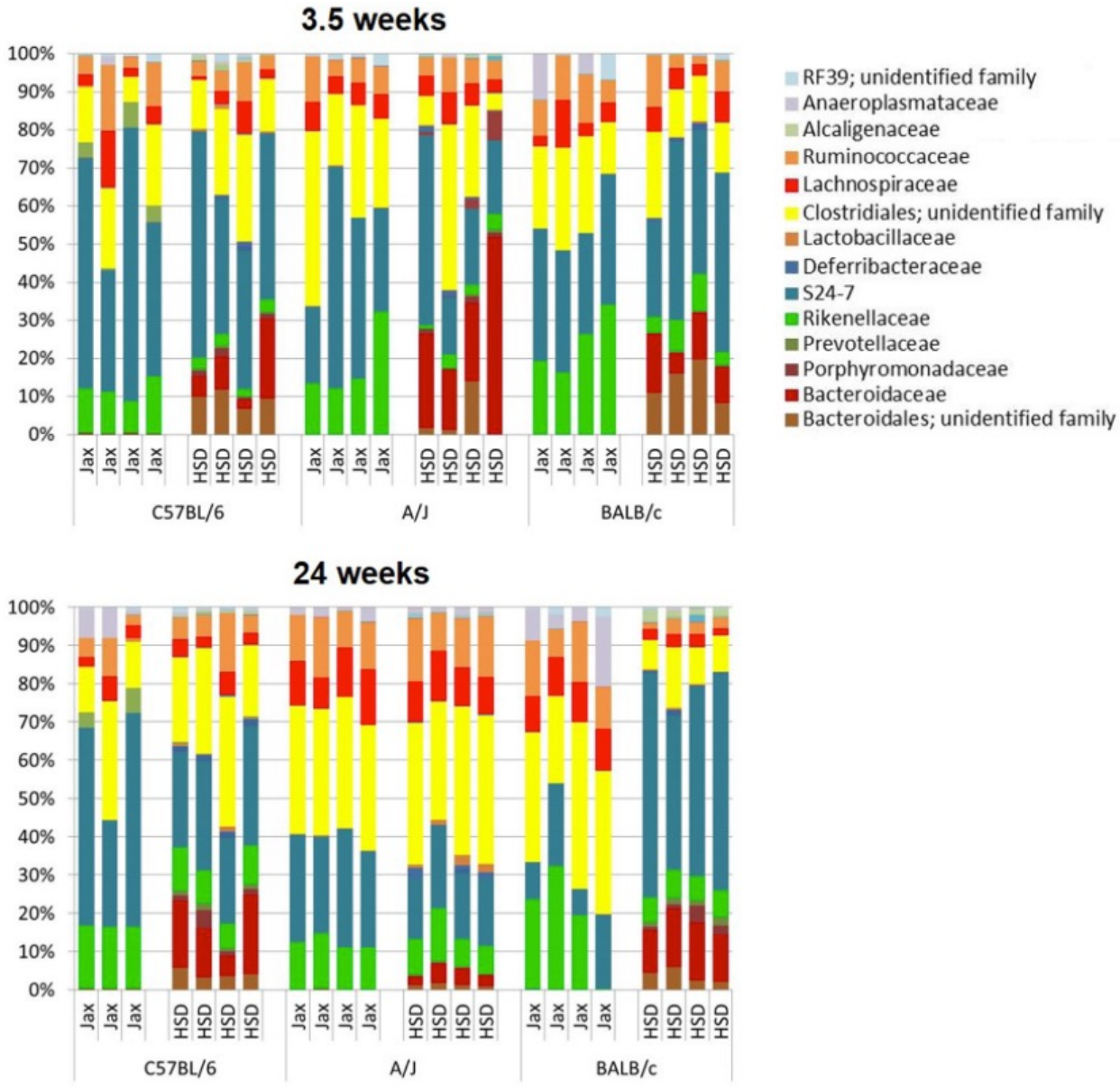
Groups compared	BaseMean	log2(fold change)	padj	Taxonomic annotation
ICD-r over ICD-nr	13.32	−7.05	0.0000000	*Faecalibacterium prausnitzii*
	4.08	−5.57	0.0000011	Lachnospiraceae
	3.32	−5.08	0.0000001	Ruminococcaceae
	6.49	−5.40	0.0000009	Ruminococcaceae
	19.85	−6.11	0.0000005	Ruminococcaceae
	94.87	−5.87	0.0000001	Ruminococcaceae
	53.59	−8.10	0.0000012	Ruminococcaceae *Ruminococcus*
	72.13	−5.14	0.0000241	Clostridiales
ICD-r over HC	14.91	7.20	0.0000000	Alteromonadales [Chromatiaceae]
	13.32	−7.22	0.0000000	*Faecalibacterium prausnitzii*
	2.94	−5.34	0.0000007	Ruminococcaceae
	2.33	−5.62	0.0000000	Clostridiales
	10.41	−7.47	0.0000000	Lachnospiraceae
	15.12	−7.62	0.0000001	Lachnospiraceae *Coprococcus*
	4.13	−8.43	0.0000000	Lachnospiraceae
	19.85	−7.15	0.0000000	Ruminococcaceae
	5.43	−8.72	0.0000000	Ruminococcaceae
	7.16	−6.98	0.0000151	Clostridiales
	3.78	−6.69	0.0000249	Clostridiales
	2.10	−6.53	0.0000000	Ruminococcaceae
	5.76	−7.85	0.0000000	Ruminococcaceae
	53.59	−8.64	0.0000000	Ruminococcaceae *Ruminococcus*
	72.13	−5.71	0.0000000	Clostridiales
	121.13	−9.95	0.0000000	*Prevotella copri*
	5.87	−7.58	0.0000000	*Methanobrevibacter*
ICD-nr over HC	14.91	6.47	0.0000185	Alteromonadales [Chromatiaceae]
	2.94	−7.00	0.0000756	Ruminococcaceae
	5.43	−8.17	0.0000682	Ruminococcaceae
	121.13	−7.82	0.0000185	*Prevotella copri*
CCD over HC	5.43	−8.65	0.0000000	Ruminococcaceae
	121.13	−7.94	0.0000000	*Prevotella copri*
UC over HC	6.53	6.32	0.0000978	*Alistipes massiliensis*

Criteria for inclusion: BaseMean > 1 and padj < 0.0001. Brackets indicate putative taxonomy based upon phylogenetic placement as given in the Greengenes taxonomy. BaseMean is the mean of normalized counts for all samples; padj is the Benjamini–Hochberg adjusted *P* value.

Alteromonadales [Chromatiaceae]
Faecalibacterium prausnitzii
Ruminococcaceae
Clostridiales
Lachnospiraceae
Lachnospiraceae *Coprococcus*
Lachnospiraceae
Ruminococcaceae
Ruminococcaceae
Clostridiales
Clostridiales
Ruminococcaceae
Ruminococcaceae
Ruminococcaceae *Ruminococcus*
Clostridiales
Prevotella copri
Methanobrevibacter
Alteromonadales [Chromatiaceae]
Ruminococcaceae
Ruminococcaceae
Prevotella copri
Ruminococcaceae
Prevotella copri
Alistipes massiliensis

Halfvarson et al (2015) *Nat Microbiol*

The mouse vendors impact on the microbiota

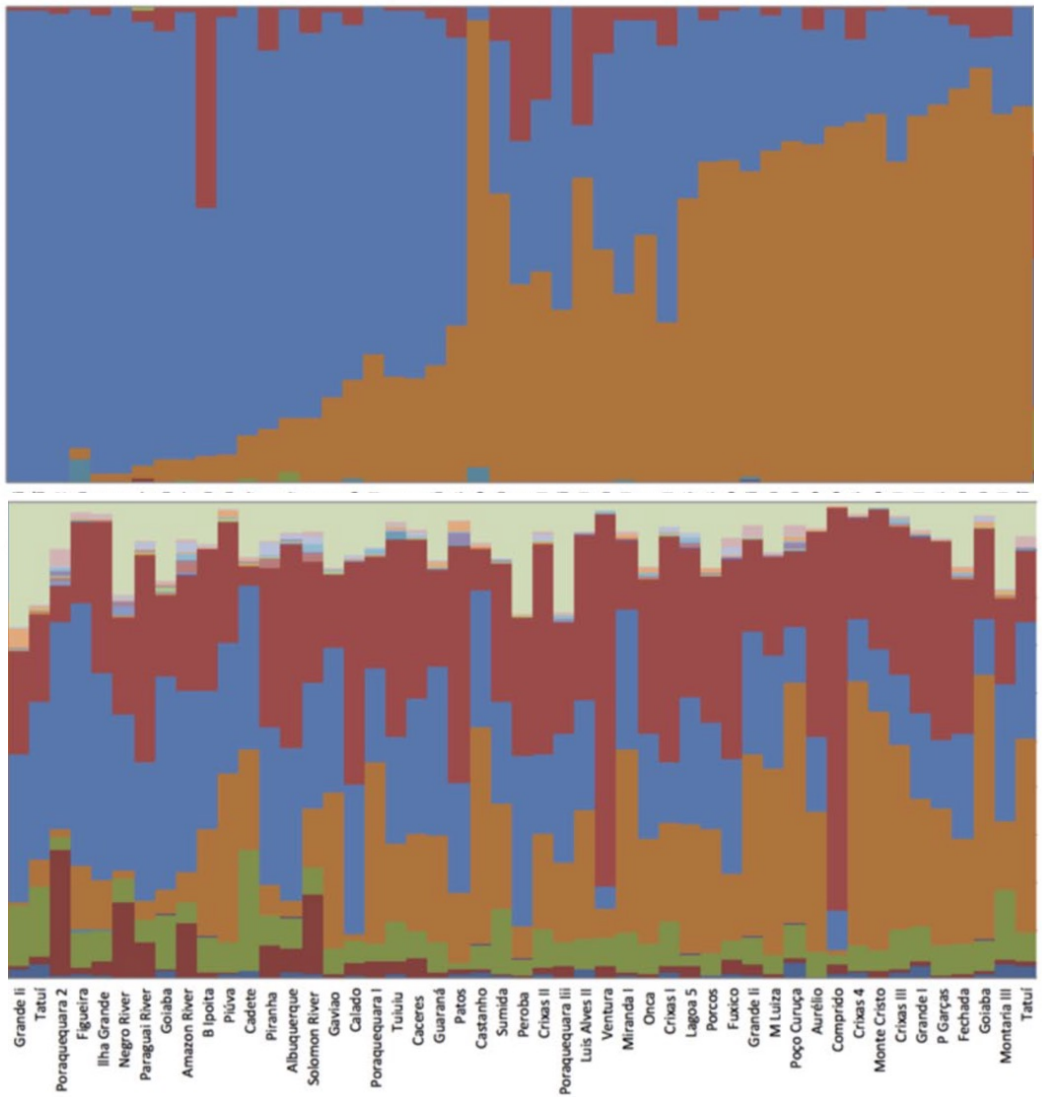


Ericsson et al (2015) *Plos One*

The 16S analysis has limitations and biases

Shotgun sequencing

Amplicon sequencing

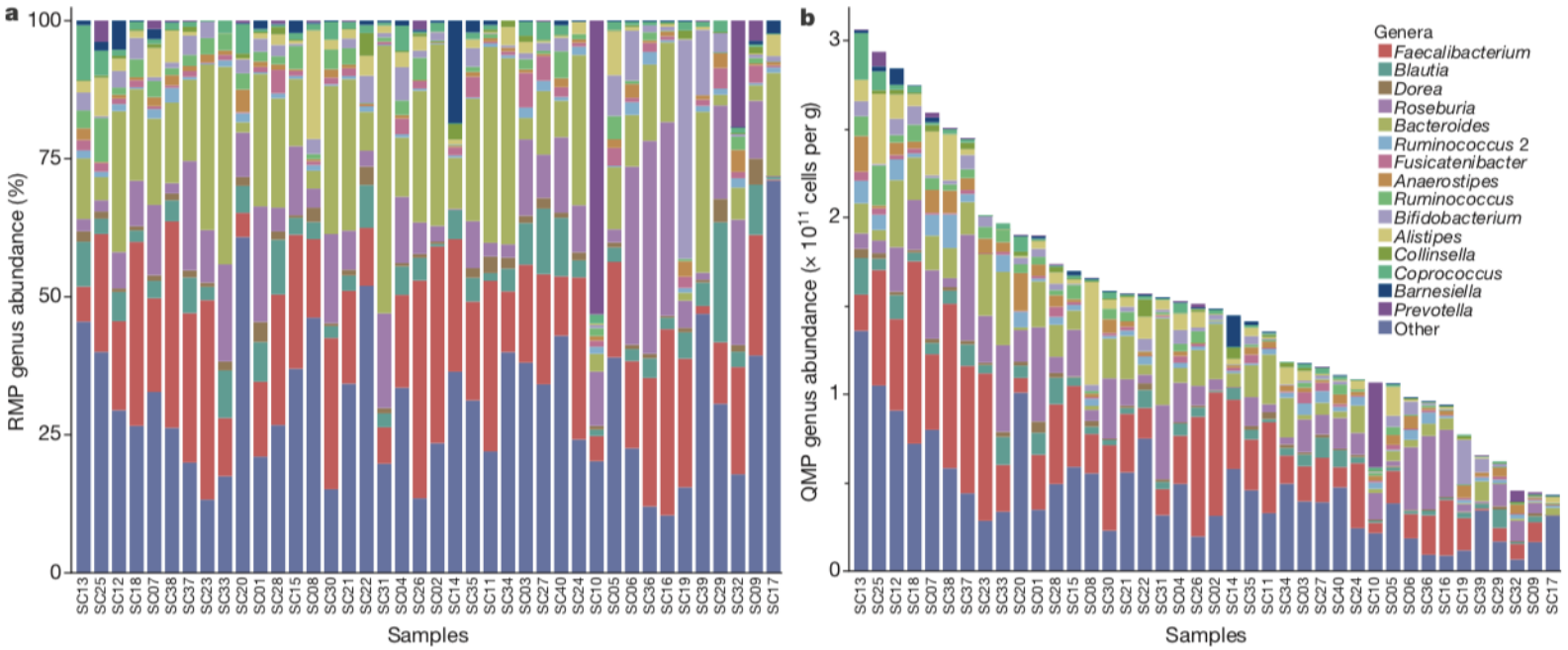


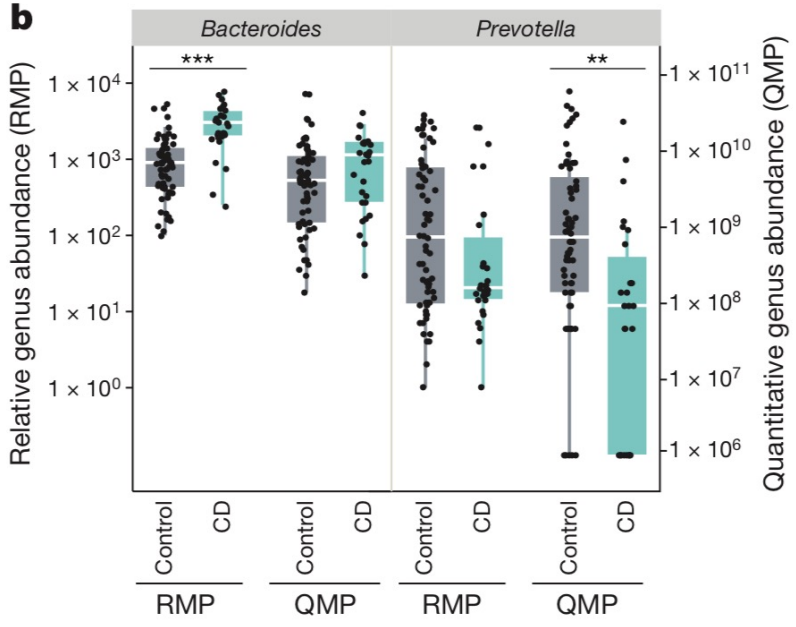
- Firmicutes
- Acidobacteria
- Actinobacteria
- DeinococcusThermus
- Parcubacteria
- Planctomycetes
- Bacteroidetes
- Euryarchaeota
- Aminicenantes
- Fusobacteria
- Latescibacteria
- Candidatus Saccharibacteria
- Chlorobi
- Cyanobacteria
- Chlamydiae
- Armatimonadetes
- Gemmatimonadetes
- Chloroflexi
- Proteobacteria
- Nitrospirae
- Verrucomicrobia

Tessler et al (2017) *Sci Rep*

The bias of relative quantification

Current 16S and metagenome analysis are based on relative quantification





Vandeputte et al (2018) *Nature*

Recommendations

It all starts on the study design …

Some Recommendations

1. **Collect as much information as possible about cofounding factors**
   And account for them in the analysis, if possible

2. **Your experimental / extraction batches should not match your sample groups**
   Have balanced batch with at least two samples from each group

3. **Always include:**
   - Negative Extraction control (extract with no samples)
   - Negative Elution control (final elution sample buffer)

4. **Remember: 16S data is not always comparable between studies**
   It has lots of methodological biases

5. **Go for metagenomics instead of amplicon-based sequencing**
   If the budget allows

6. **Move towards quantitative approaches for absolute microbe quantification**
   By using spike-in controls

7. **Discuss your experiment with an experienced bioinformatician that will do the analysis**
   Before starting the experiment!

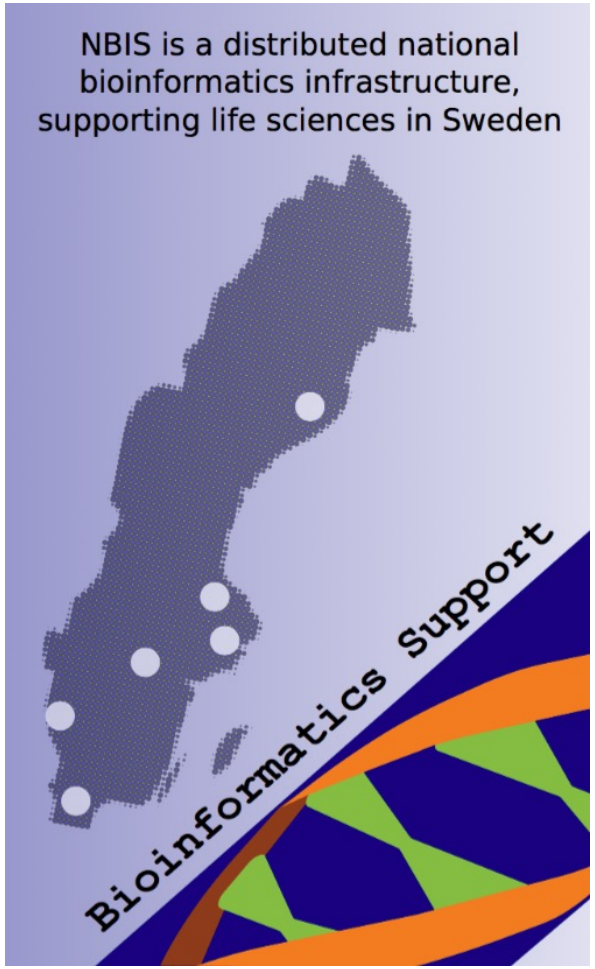
8. **Ask for bioinformatics help!**

Bioinformatics Help

For further support …

About NBIS

NBIS is the bioinformatics infrastructure at SciLifeLab



www.nbis.se

We offer a diversified project-tailored support

General support services

Bioinformatics drop-in sessions

Weekly drop-in sessions in Lund, Göteborg, Linköping, Stockholm, Uppsala, and Umeå.
Read more

Consultation meetings (free)

Study design consultation meetings with one of our experts.
Read more | Apply

Short- and Medium-term Support (user fee)

This track is mainly targeting projects with a reasonably well-defined bioinformatics problem, but also longer and more open-ended projects are welcome to apply. We offer support in a wide range of bioinformatics areas, including NGS, proteomics, metabolomics, biostatistics, systems development and data management. Project applications are reviewed every second week, and projects are selected based on available expertise, and feasibility. Time and outcome estimations are done individually for each project before contracting.
Read more | Apply

Long-term Support (free, competitive peer-review)

Bioinformatics support to a limited set of scientifically outstanding projects. Applications are submitted in open calls three times per year (February/June/October), and are selected based on scientific peer-review by a national committee. This support has been enabled primarily by a generous grant from the Knut and Alice Wallenberg Foundation.
Read more | Apply

Bioinformatics Drop-in

Bioinformatics Drop-in is available weekly to anyone in the Swedish research community wanting to discuss bioinformatics questions with experts from NBIS.

If you have questions about bioinformatics, feel free to join us over Zoom, Tuesdays 14:00-15:00, by following this link: https://meet.nbis.se/dropin.

Note that there are, typically, no drop-ins on public holidays; see further in the schedule below.

www.nbis.se

References

For further learning …

References

- Blattner et al. (1997) *Science* - The Complete Genome Sequence of *Escherichia coli* K-12
- Tessler et al (2017) *Sci Rep*
- Yang et al (2016) *BMC Bioinformatics*
- Callahan et al (2016) Nat Methods - DADA2: high-resolution sample inference from illumina amplicon data
- Edgar (2010) *Nat Methods* - UPARSE: highly accurate otu sequences from microbial amplicon reads
- Caporaso et al (2010) *Nat Methods* - QIIME allows analysis of high-throughput community sequencing data
- Vandeputte et al (2015) *Gut* - Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates
- Mackenzie et al (2015) *Front Microbiol* - Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences
- Nayfach, Pollard (2018) *Cell* - Toward Accurate and Quantitative Comparative Metagenomics
- Langille et al (2013) *Nat Biotechnol* - Predictive functional profiling of microbial communities using 16s rRNA marker gene sequences
- Vandeputte et al (2018) *Nature* - Quantitative microbiome profiling links gut community variation to microbial load
- Sommer et al (2017) *Gut* - Microbiomarkers in inflammatory bowel diseases: caveats come with caviar
- Ubeda et al (2012) *J Exp Med* - Familial transmission rather than defective innate immunity shapes the distinct intestinal microbiota of TLR-deficient mice
- Franklin & Ericsson (2017) Lab Anim - Microbiota and reproducibility of rodent models
- Ericsson et al (2015) *Plos One* - Effects of Vendor and Genetic Background on the Composition of the Fecal Microbiota of Inbred Mice
- Ni et al (2017) *Nat Rev* - Gut microbiota and IBD: causation or correlation?
- Escobar-Zepeda et al. (2015) *Front Genetics* - The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics

References

- Nicola Segata et al (2013) Mol Systems Biol - Computational meta'omics for microbial community studies
- Brooks et al (2015) *BMC Microbiol* - The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies.
- Fouthy et al (2016) *BMC Microbiol* - 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform
- Morgan & Huttenhower (2012) *Plos Comp Biol - Chapter 12: Human Microbiome Analysis*
- Joynson et al. (2017) *Front Microbiol* - Metagenomic Analysis of the Gut Microbiome of the Common Black Slug Arion ater in Search of Novel Lignocellulose Degrading Enzymes
- Siegwald et al (2017) *PlotsOne* - Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics
- Balvociute & Hason (2017) *BMC Genomics* - SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?
- Yatsunenko et al (2012) *Nature* - Human gut microbiome viewed across age and geography
- Li et al (2015) *Bioinformatics* - MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph
- Wood & Salzberg (2014) *Genome Biol* - Kraken: ultrafast metagenomic sequence classification using exact alignments
- Abubucker et al (2012) *Plos Comp Biol* - Metabolic reconstruction for metagenomic data and its application to the human microbiome.
- Segata et al (2012) *Nat Methods* - Metagenomic microbial community profiling using unique clade-specific marker genes
- Kanehisa et al (2016) Nuc Acid Res - KEGG as a reference resource for gene and protein annotation

Other Resources

- https://slideplayer.com/slide/11914858/
- https://www.slideshare.net/beiko/ccbc-tutorial-beiko

Thank you!

Paulo Czarnewski

Senior Bioinformatician
Scientific Coordinator

National Bioinformatics Infrastructure Sweden (NBIS)
SciLifeLab, Stockholm University

www.czarnewski.com