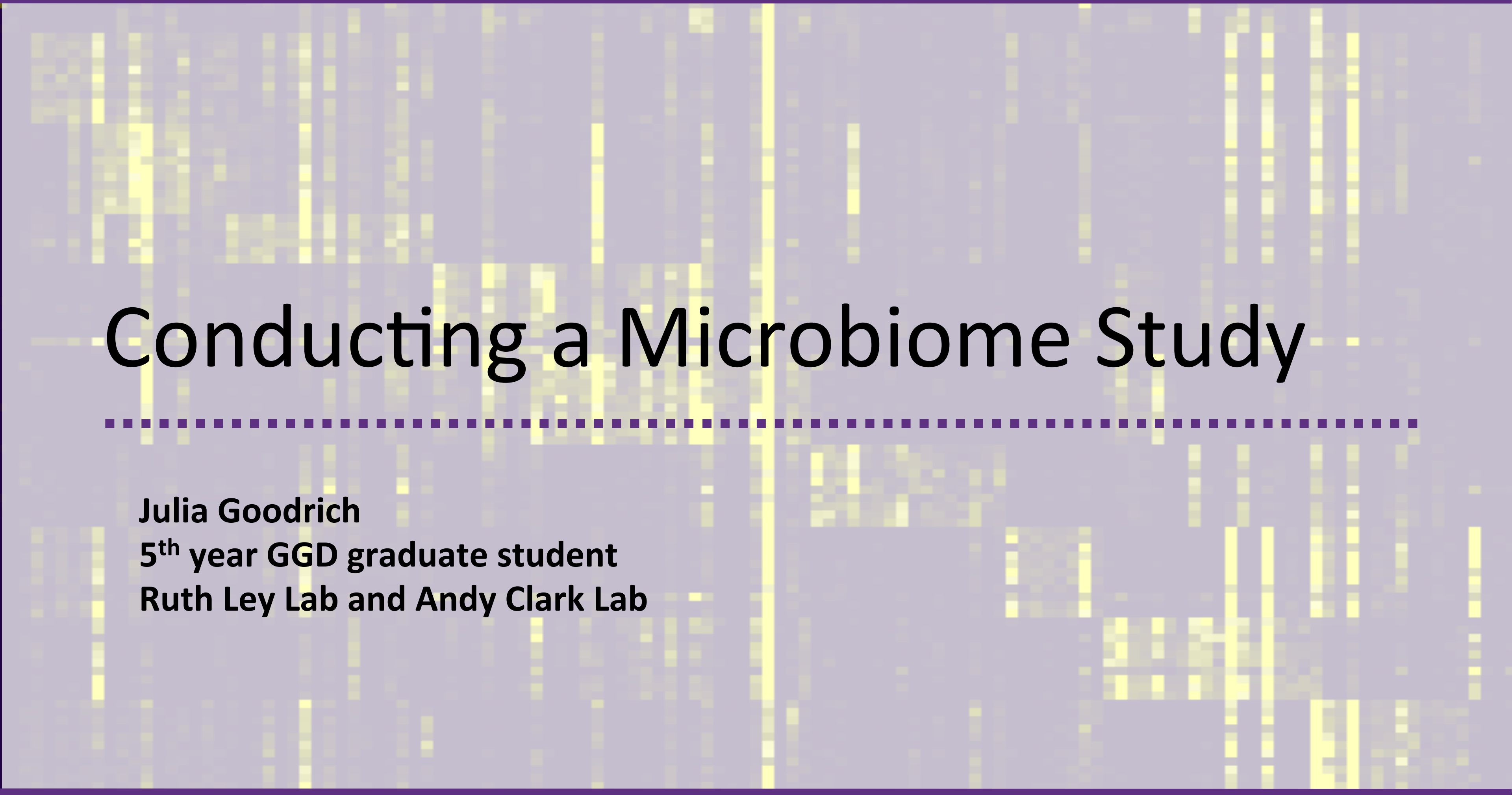
October 13th, 2014



Conducting a Microbiome Study

Julia Goodrich
5th year GGD graduate student
Ruth Ley Lab and Andy Clark Lab


Ruth Ley


Angela Poole


Omry Koren


Tyler Cullender


Antonio Gonzalez

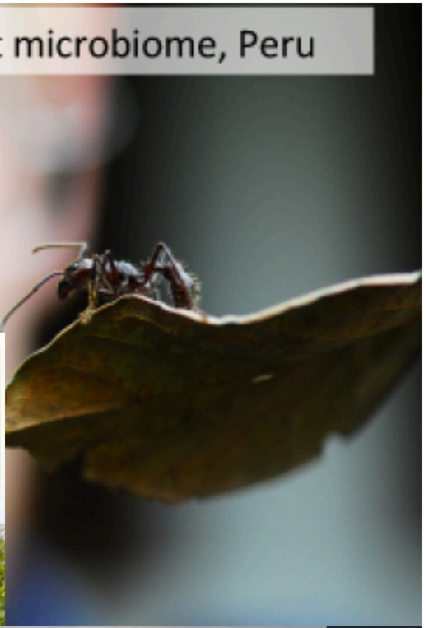

Cell

Leading Edge
Primer

Conducting a Microbiome Study

Julia K. Goodrich,[1,2] Sara C. Di Rienzi,[1,2] Angela C. Poole,[1,2] Omry Koren,[1,2,9] William A. Walters,[3] J. Gregory Caporaso,[4,5] Rob Knight,[6,7,8] and Ruth E. Ley[1,2,*]

Annie Moore & Colleagues sampling cenote gradients, Yucatan, Mexico



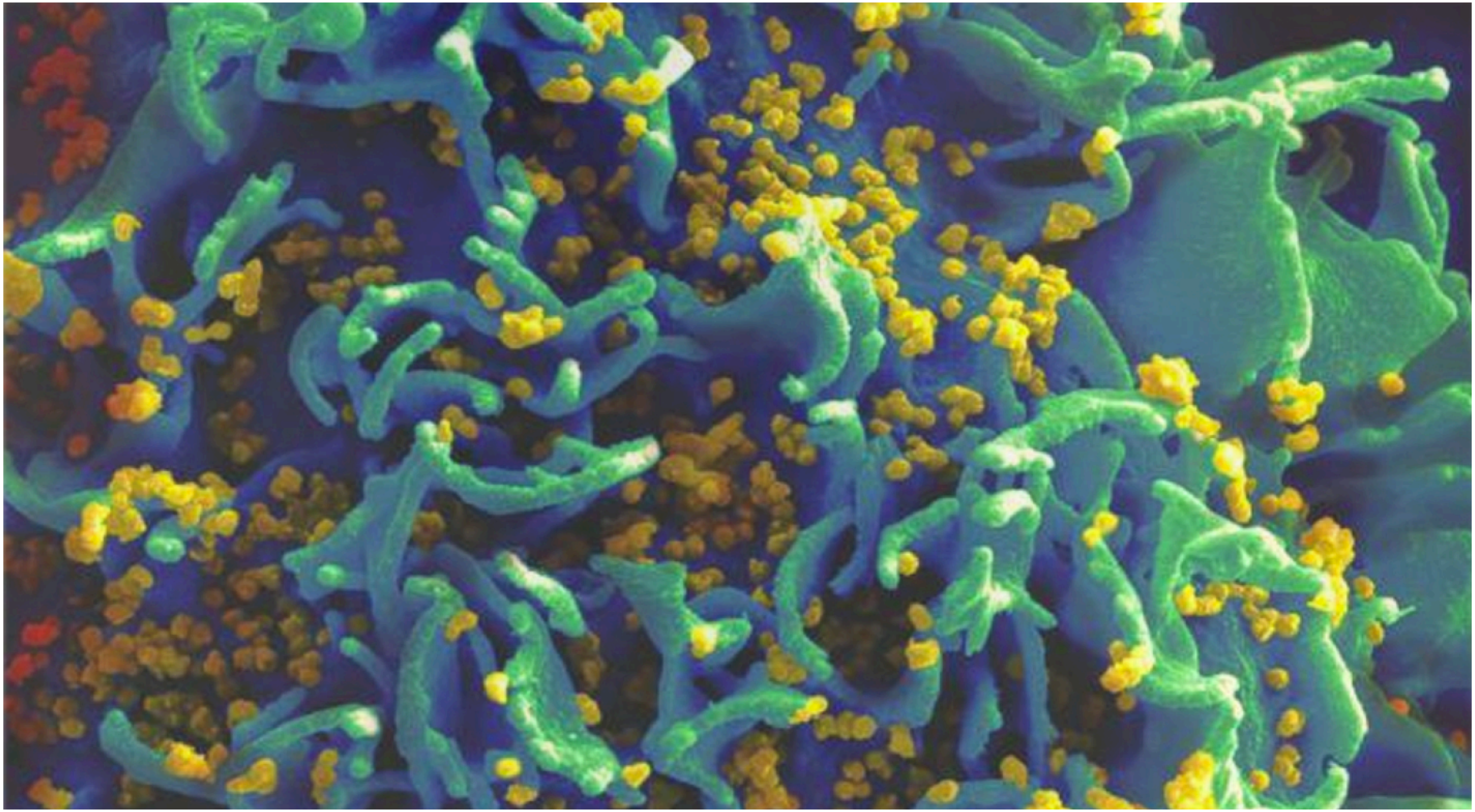
Jon Sanders, The ant microbiome, Peru

Thar Desert (Great Indian Desert)

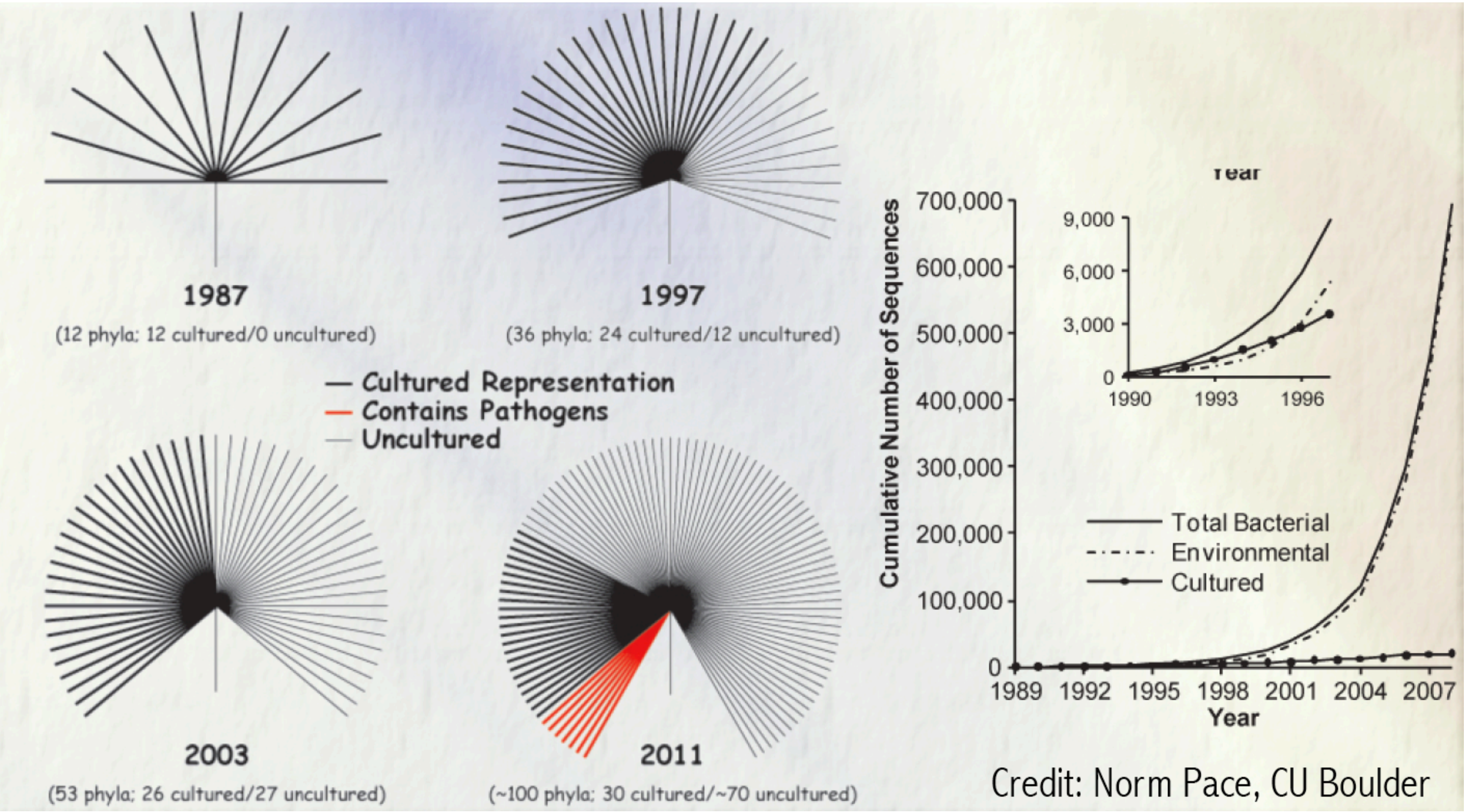
Merlot Microbiome:
High school volunteers
Long Island

Deepwater Horizon oil spill

How can we measure complex communities?



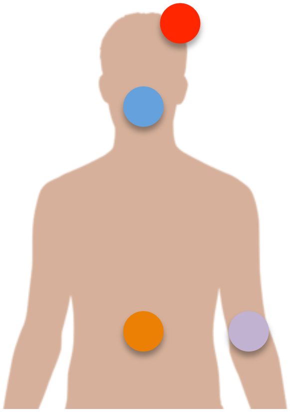
Rapid expansion of our knowledge to uncultured



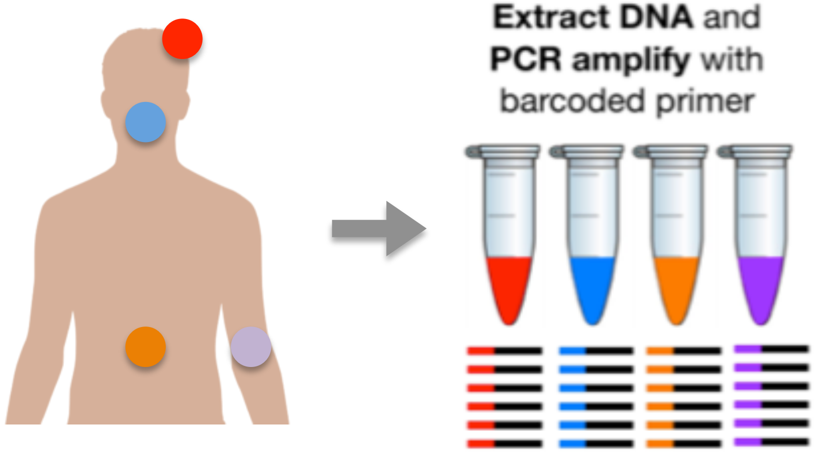
Credit: Norm Pace, CU Boulder


Archaea
Thermoplasmata
Methanomicrobia
Eukaryota
Fungi
Metazoa
Viridiplantae
Heterokonta
Bacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Deltaproteobacteria
Epsilonproteobacteria
Bacteroidetes
Chlorobi
Acidobacteria
Firmicutes
Actinobacteria
Cyanobacteria
Chloroflexi
Deinococcus-Thermus
Who
is in there?

A classic 16S study



A classic 16S study



Extract DNA and
PCR amplify with
barcoded primer

Contamination

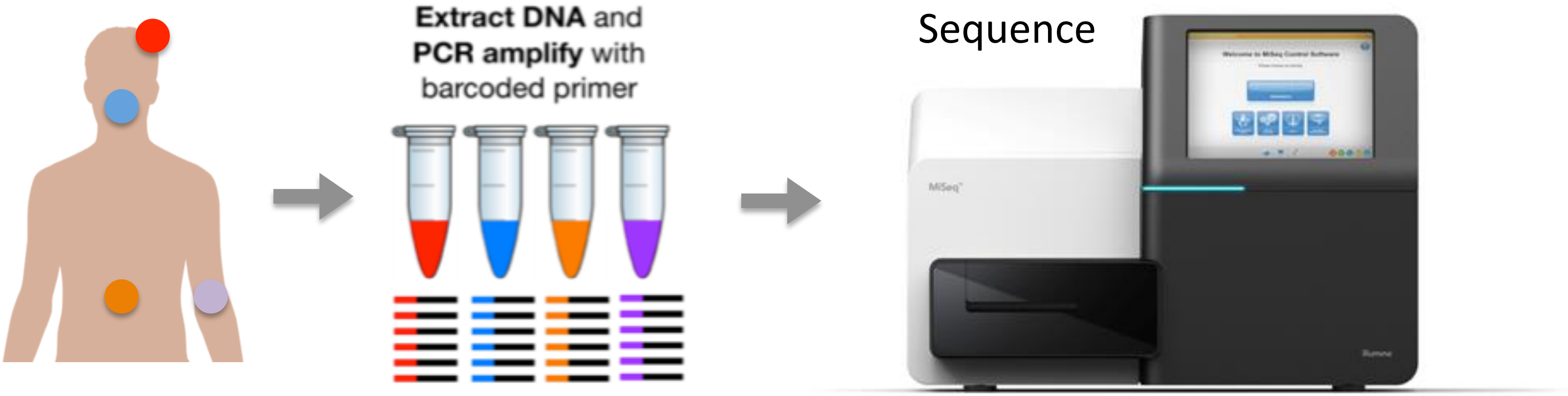
Steps must be taken to minimize contamination from a ‘foreign’ source.

Test kit components and laboratory reagents.

    -DNA free primers*

Can’t subtract contamination sequences during analysis.

A classic 16S study

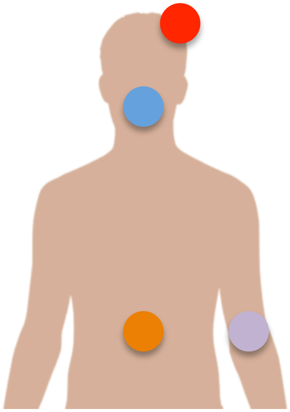


Extract DNA and **PCR amplify** with barcoded primer

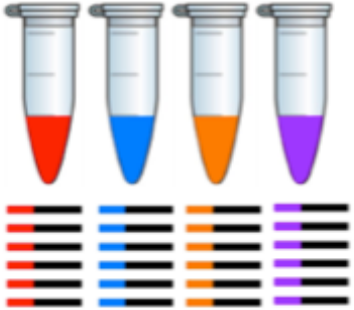
Sequence

~70,000 sequences per sample

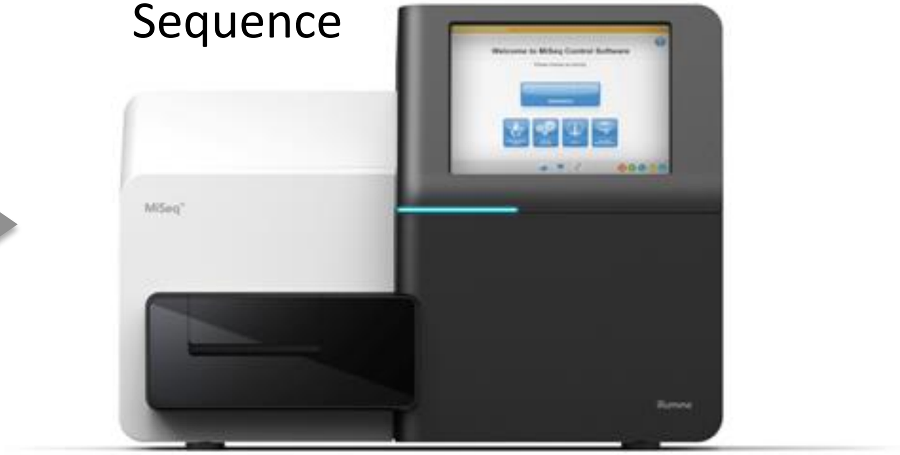
A classic 16S study



Extract DNA and PCR amplify with barcoded primer



Sequence



~70,000 sequences per sample

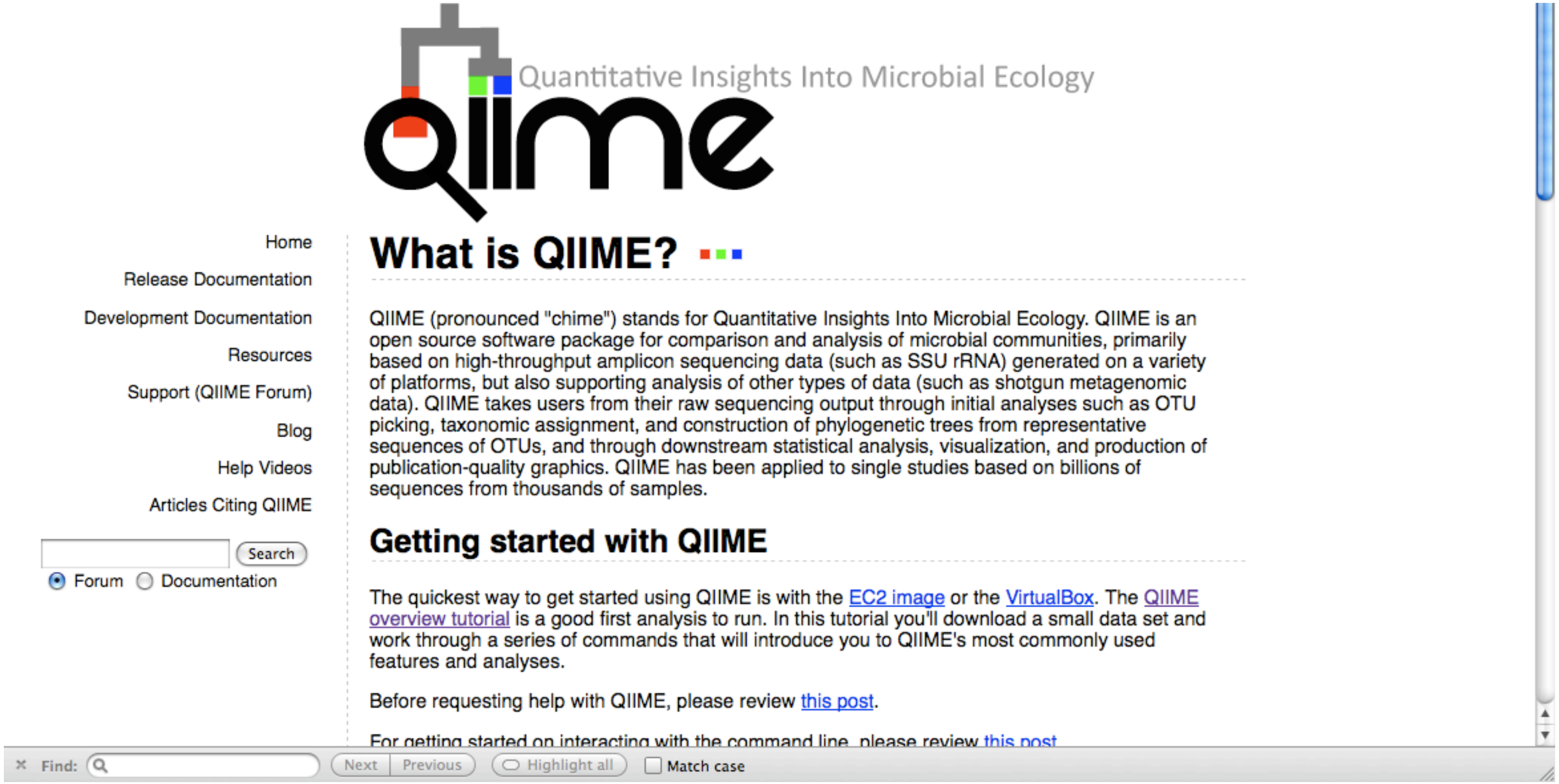
Data Analysis

qiime
Quantitative Insights Into Microbial Ecology

mothur

VAMPS
Visualization and Analysis of Microbial Population Structures

CloVR



Quantitative Insights Into Microbial Ecology

What is QIIME?

QIIME (pronounced "chime") stands for Quantitative Insights Into Microbial Ecology. QIIME is an open source software package for comparison and analysis of microbial communities, primarily based on high-throughput amplicon sequencing data (such as SSU rRNA) generated on a variety of platforms, but also supporting analysis of other types of data (such as shotgun metagenomic data). QIIME takes users from their raw sequencing output through initial analyses such as OTU picking, taxonomic assignment, and construction of phylogenetic trees from representative sequences of OTUs, and through downstream statistical analysis, visualization, and production of publication-quality graphics. QIIME has been applied to single studies based on billions of sequences from thousands of samples.

Getting started with QIIME

The quickest way to get started using QIIME is with the EC2 image or the VirtualBox. The QIIME overview tutorial is a good first analysis to run. In this tutorial you'll download a small data set and work through a series of commands that will introduce you to QIIME's most commonly used features and analyses.

Before requesting help with QIIME, please review this post.

For getting started on interacting with the command line, please review this post

Caporaso *et al.* 2010

qiime

Quantitative Insights Into Microbial Ecology

- Install QIIME and every single dependency by hand

QIIME software dependencies

[data-lanemask]
[data-core]
[python]
[setuptools]
[MySQL-python]
[SQLAlchemy]
[pycogent]
[pynast]
[numpy]
[matplotlib]
[mpi4py]
[lxml]
[sphinx]
[raxml]
[fasttree]

[cdbtools]
[chimeraslayer]
[cdhit]
[rdpclassifier]
[blast]
[muscle]
[infernal]
[cytoscape]
[clearcut]
[mothur]
[uclust]
[r]
[ampliconnoise]
[vienna]
[pprospector]

qiime
Quantitative Insights Into Microbial Ecology

qiime

Quantitative Insights Into Microbial Ecology

- Install QIIME and every single dependency by hand
- Use the app-deploy.py script

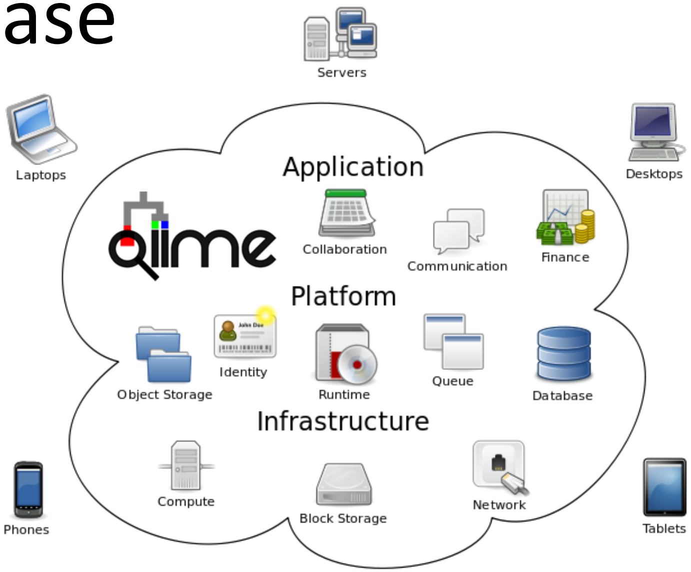
qiime

Quantitative Insights Into Microbial Ecology

- Install QIIME and every single dependency by hand
- Use the app-deploy.py script
- Download the QIIME Virtual Box

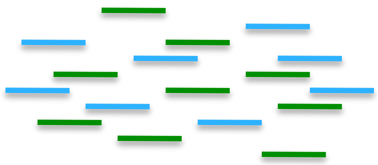
Quantitative Insights Into Microbial Ecology

- Install QIIME and every single dependency by hand
- Use the app-deploy.py script
- Download the QIIME Virtual Box
- Use the latest EC2 QIIME release



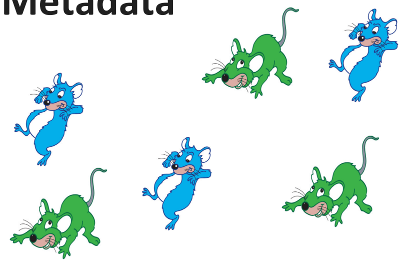
Overview of the analysis pipeline

Sequencing output

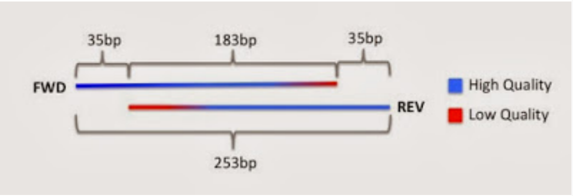


Raw files (fastq format)
Forward
Reverse
Barcodes

Metadata

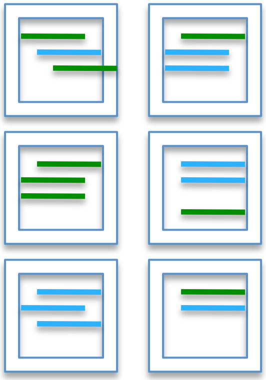


Join pair ends

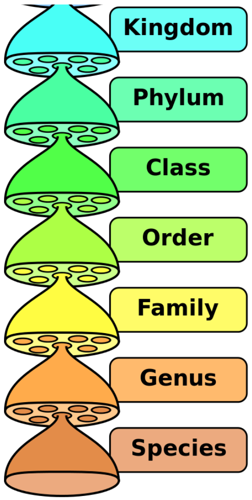


Quality filter and demultiplex

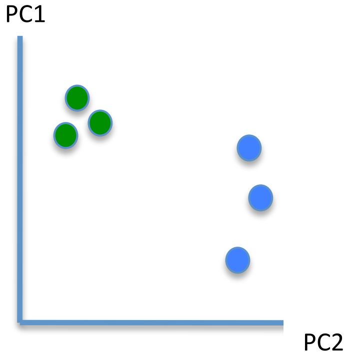
OTU picking



Assign taxonomy



Diversity Analysis



Join paired ends

Example: V4 region of 16S with 250bp paired-end sequencing on Illumina MiSeq. Amplicon length is 253bp

New in QIIME 1.8 Methods available:

1. fastq-join
2. SeqPrep

Important setting! **`--min_overlap`** sets the minimum number of base pairs in the overlapping region.

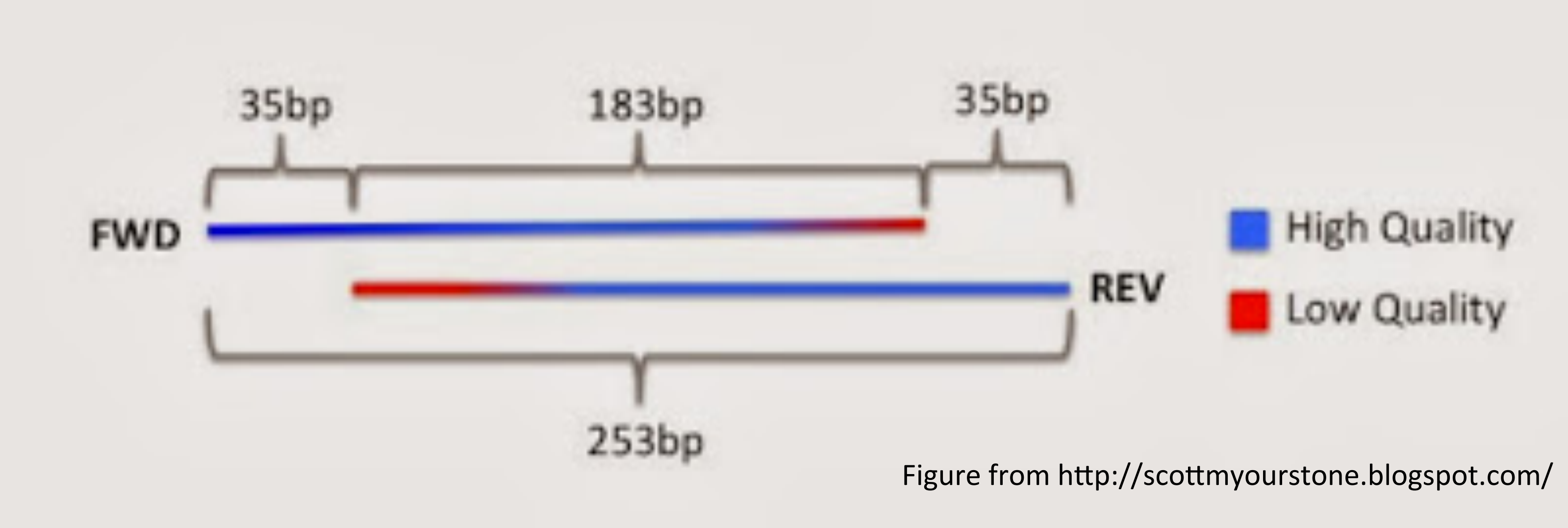
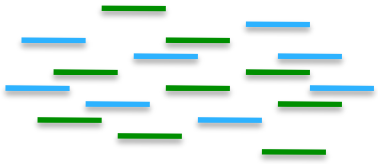


Figure from http://scottmyourstone.blogspot.com/

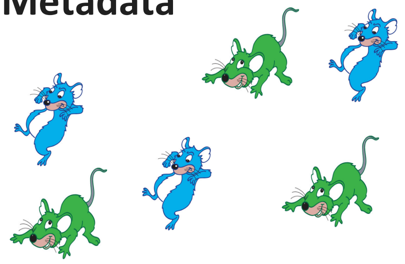
Overview of the analysis pipeline

Sequencing output

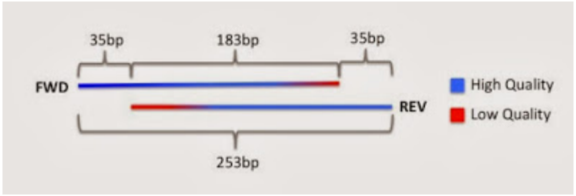


Raw files (fastq format)
Forward
Reverse
Barcodes

Metadata

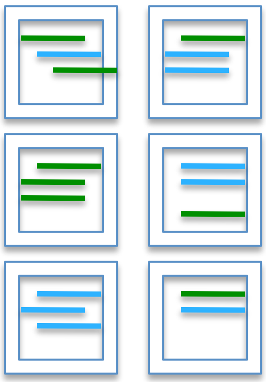


Join pair ends

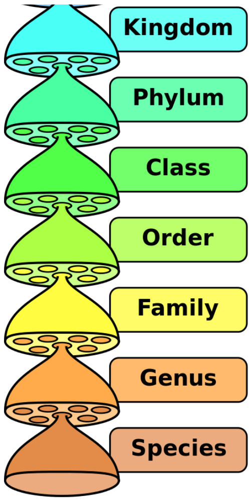


Quality filter and demultiplex

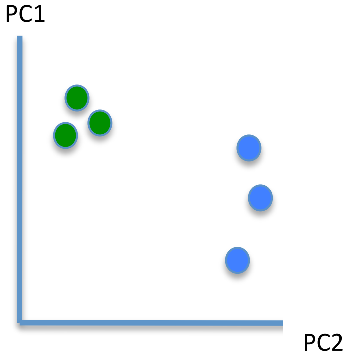
OTU picking



Assign taxonomy

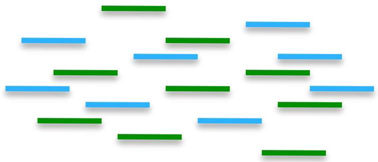


Diversity Analysis



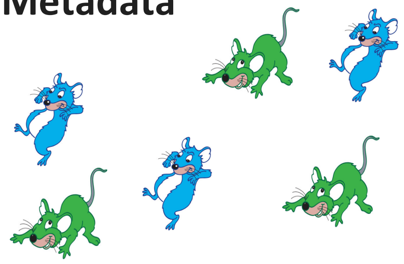
Overview of the analysis pipeline

Sequencing output

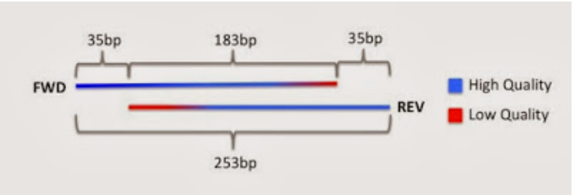


Raw files (fastq format)
Forward
Reverse
Barcodes

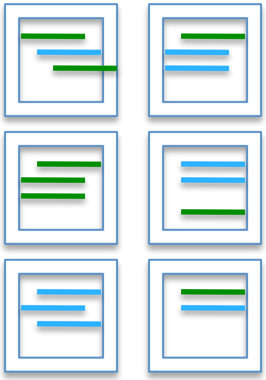
Metadata



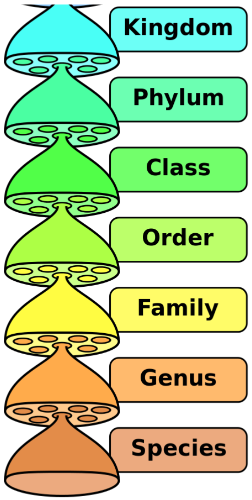
Join pair ends



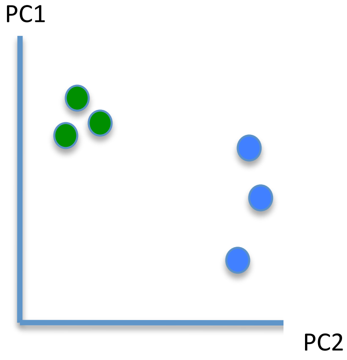
OTU picking



Assign taxonomy

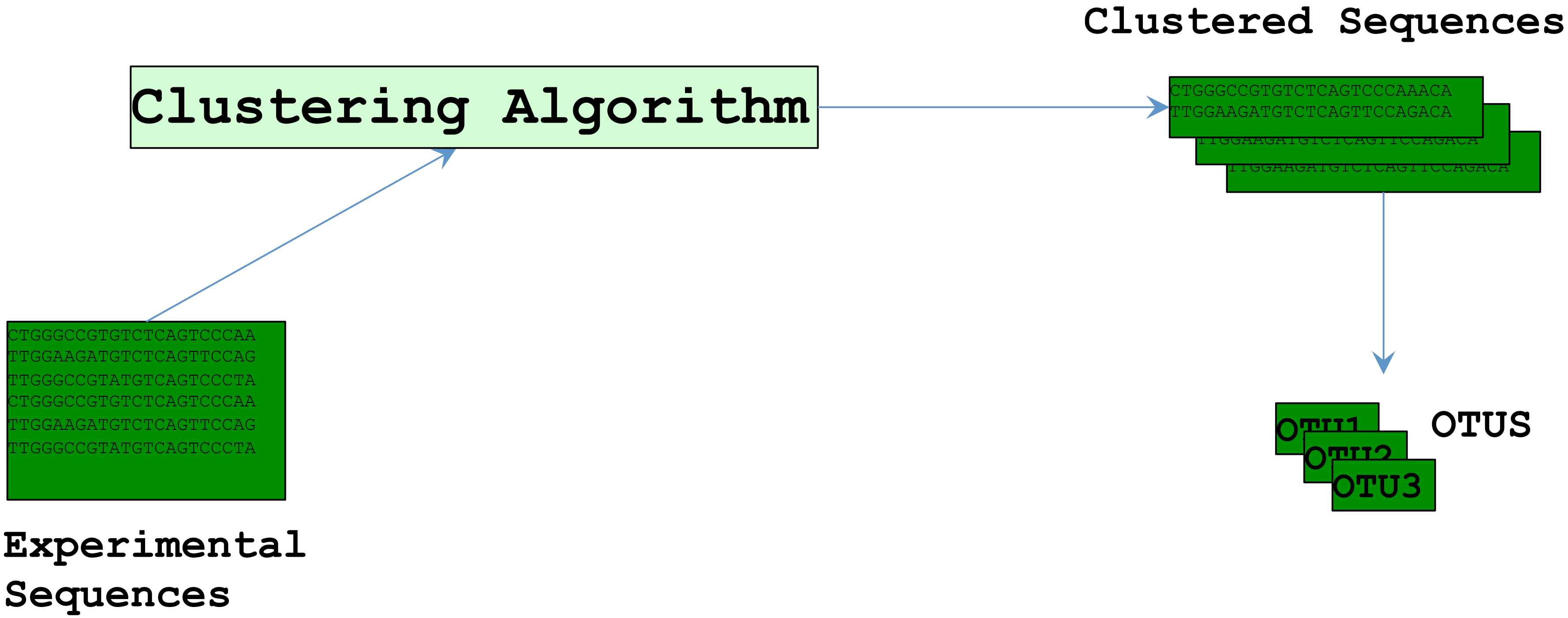


Diversity Analysis



Quality filter and demultiplex

OTU Picking - “de-novo”



Clustered Sequences

Clustering Algorithm

CTGGGCCGTGTCTCAGTCCCAAACA
TTGGAAGATGTCTCAGTTCCAGACA
TTGGAAGATGTCTCAGTTCCAGACA
TTGGAAGATGTCTCAGTTCCAGACA

CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA
CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA

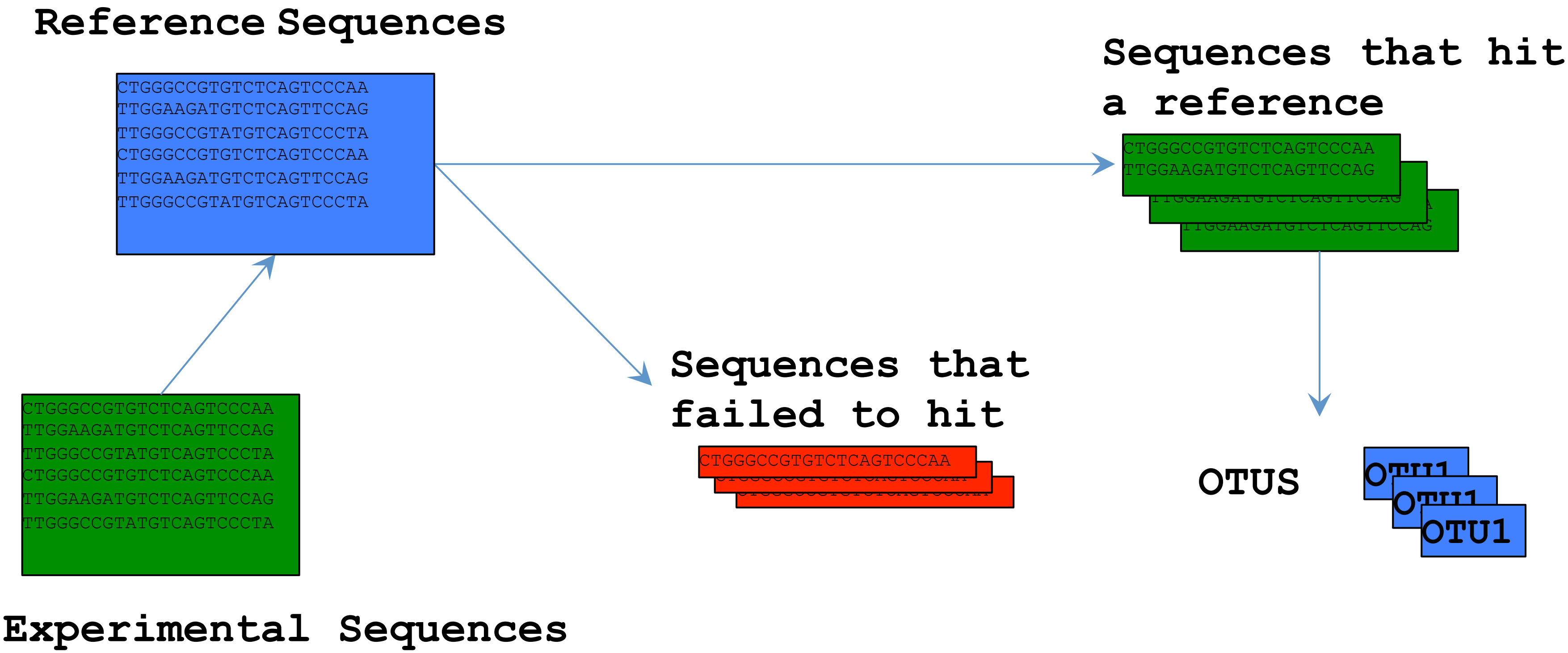
OTU1
OTU2
OTU3
OTUS

Experimental Sequences

- Pros
  - Vast majority of reads are clustered
  - No reference database bias

- Cons
  - Speed; not easily parallelizable
  - Erroneous reads get clustered

OTU Picking - “closed-reference”



Reference Sequences

CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA
CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA

Sequences that hit a reference

CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG

Sequences that failed to hit

CTGGGCCGTGTCTCAGTCCCAA

OTUS

OTU1

Experimental Sequences

CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA
CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA

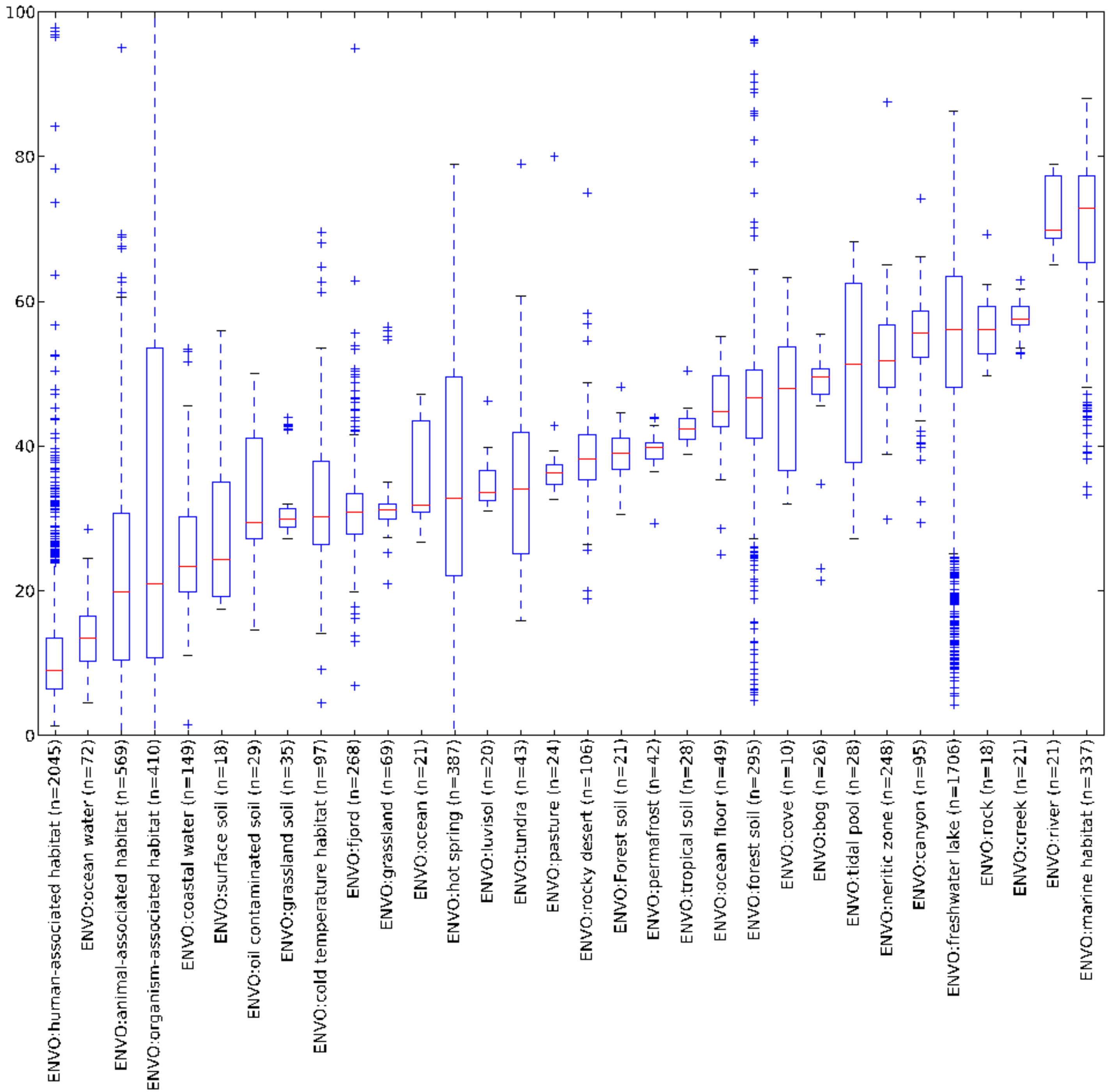
- Pros
  - Reference database is a quality filter
  - Speed; easily parallelizable
- Cons
  - No new OTUs can be observed
  - Reference database bias

Reference database

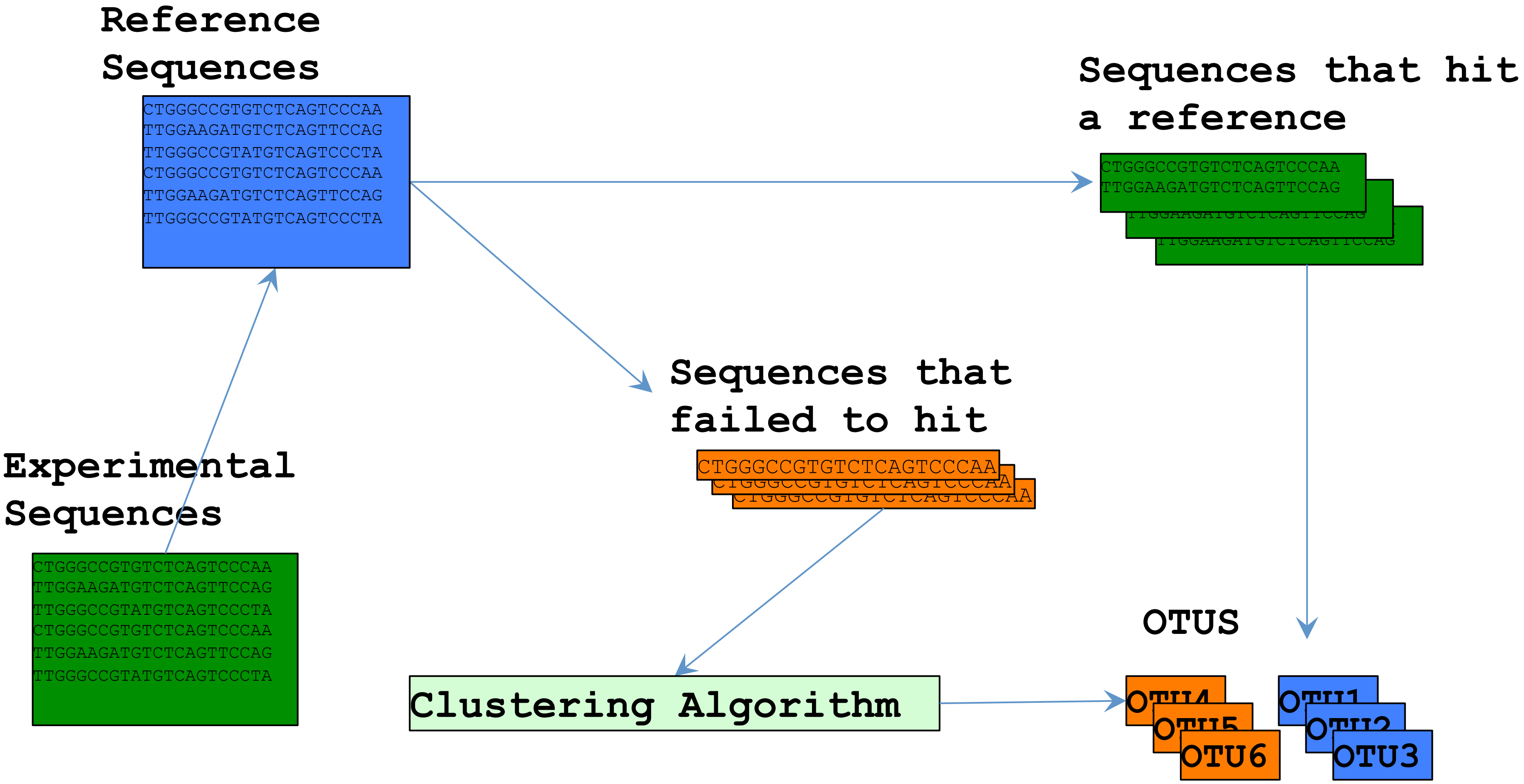
GREENGENES

The 16S rRNA Gene Database and Tools

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



OTU Picking - “open-reference”



Reference Sequences

CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA
CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA

Sequences that hit a reference

CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGAAGATGTCTCAGTTCCAG
TTGGAAGATGTCTCAGTTCCAG

Sequences that failed to hit

CTGGGCCGTGTCTCAGTCCCAA
CTGGGCCGTGTCTCAGTCCCAA
CTGGGCCGTGTCTCAGTCCCAA

Experimental Sequences

CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA
CTGGGCCGTGTCTCAGTCCCAA
TTGGAAGATGTCTCAGTTCCAG
TTGGGCCGTATGTCAGTCCCTA

Clustering Algorithm

OTUS

OTU4
OTU5
OTU6
OTU1
OTU2
OTU3

- Pros
  - Best of both worlds
- Cons
  - Downsides of de-novo

OTU table

						
A	**10**	**0**	**9**	**0**	**45**	**8**
B	**1**	**65**	**0**	**88**	**45**	**0**
C	**0**	**0**	**1**	**0**	**0**	**0**

Count table

Feature X Sample table

Samples →

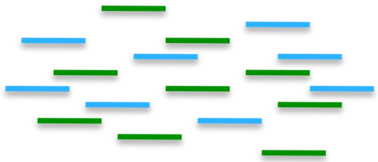
Features ↓

Relative abundance table

						
A	**0.9**	**0**	**0.9**	**0**	**0.5**	**1.0**
B	**0.1**	**1.0**	**0**	**1.0**	**0.5**	**0**
C	**0**	**0**	**0.1**	**0**	**0**	**0**

Overview of the analysis pipeline

Sequencing output



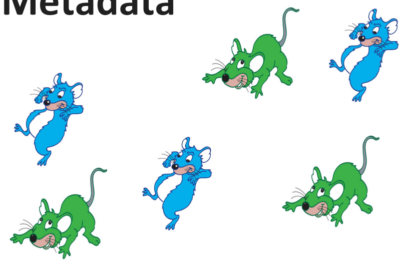
Raw files (fastq format)

Forward

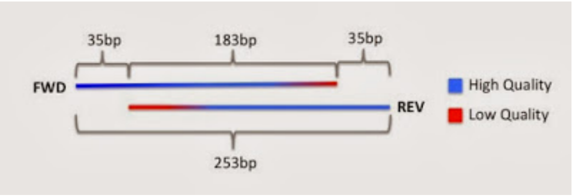
Reverse

Barcodes

Metadata

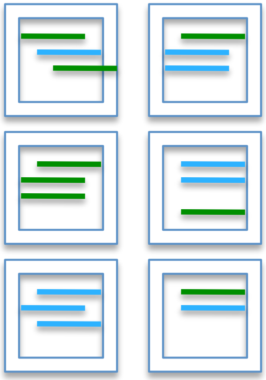


Join pair ends

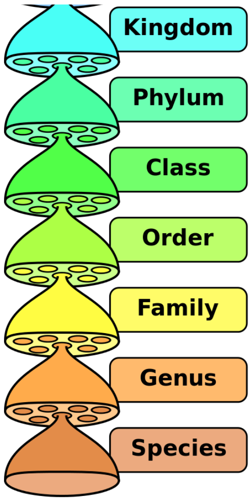


Quality filter and demultiplex

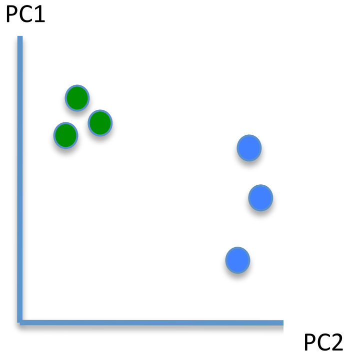
OTU picking



Assign taxonomy



Diversity Analysis



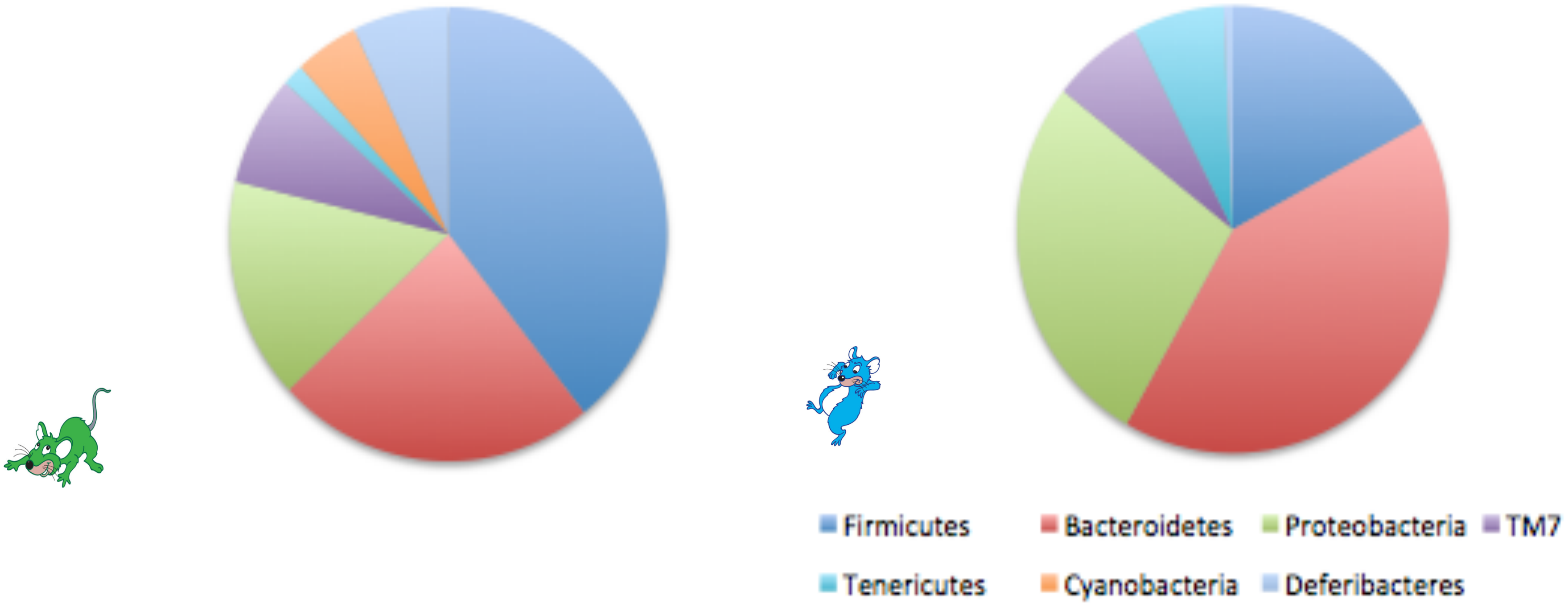
RDP classifier

- Taxonomic Assignment using 16S sequence identity

- Assigns taxonomy by matching sequences segments of length 8 to a database of previously assigned sequences

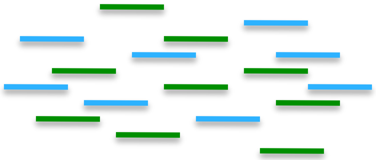
Taxonomic composition

- We can visualize the differences between the abundances at different taxonomic levels and our metadata.



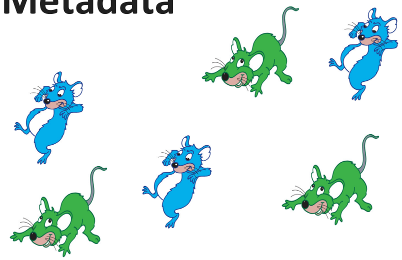
Overview of the analysis pipeline

Sequencing output

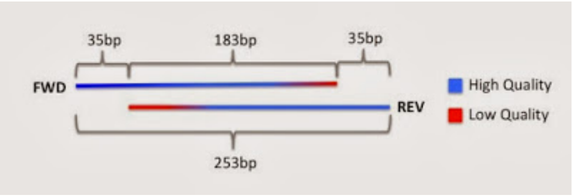


Raw files (fastq format)
Forward
Reverse
Barcodes

Metadata

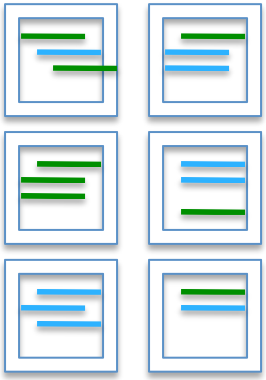


Join pair ends

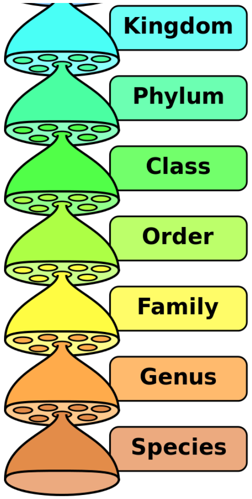


Quality filter and demultiplex

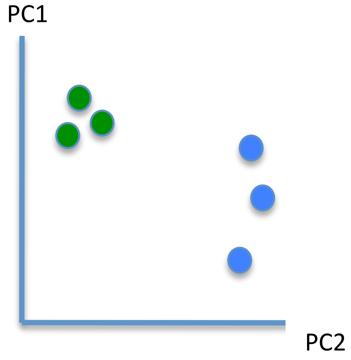
OTU picking



Assign taxonomy

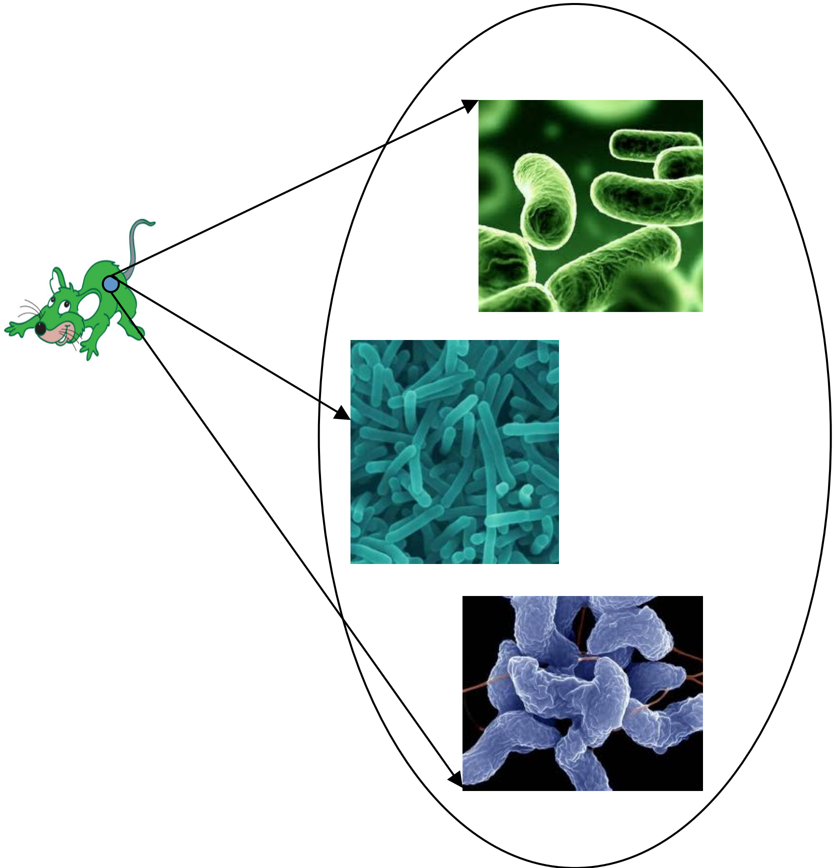


Diversity Analysis

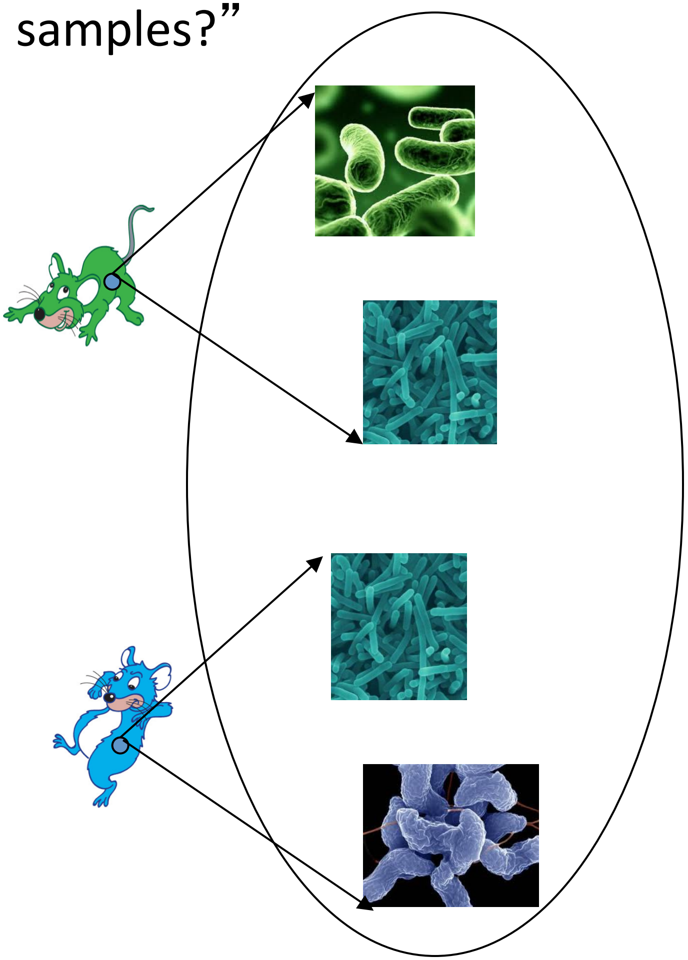


How do we describe and compare diversity?

alpha-diversity:
“How many species are in a sample?”



beta-diversity:
“How many species are shared between samples?”



Quantitative versus Qualitative measures

- Qualitative: Considers presence absence only

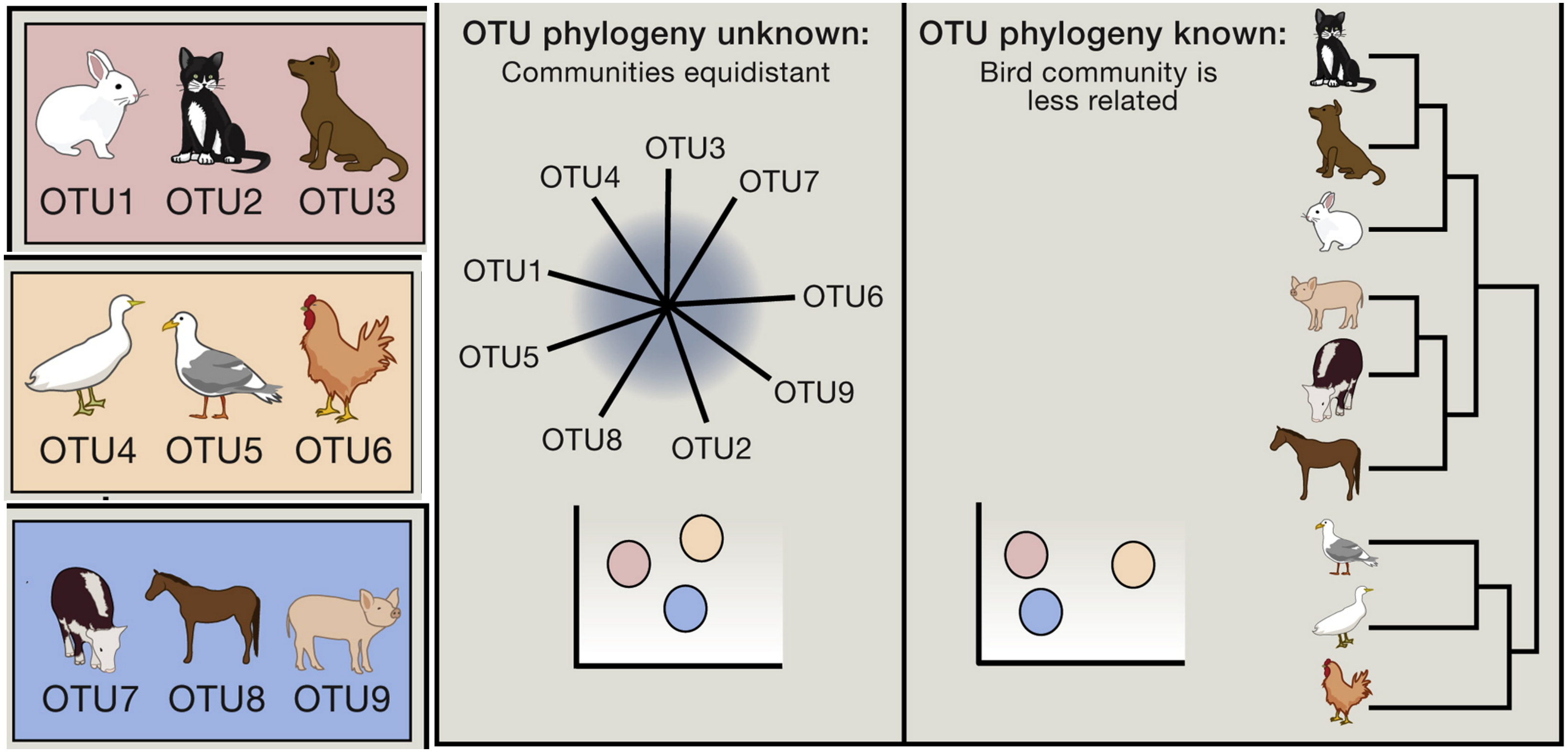
- Quantitative: Also considers relative abundance.

What is a phylogenetic diversity measure?

- α Diversity:
  - Phylogenetic: "How much phylogenetic divergence is in a sample?"
- β Diversity:
  - Phylogenetic: "How much phylogenetic distance is shared between samples?"

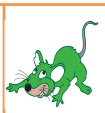
Advantages of phylogenetic techniques

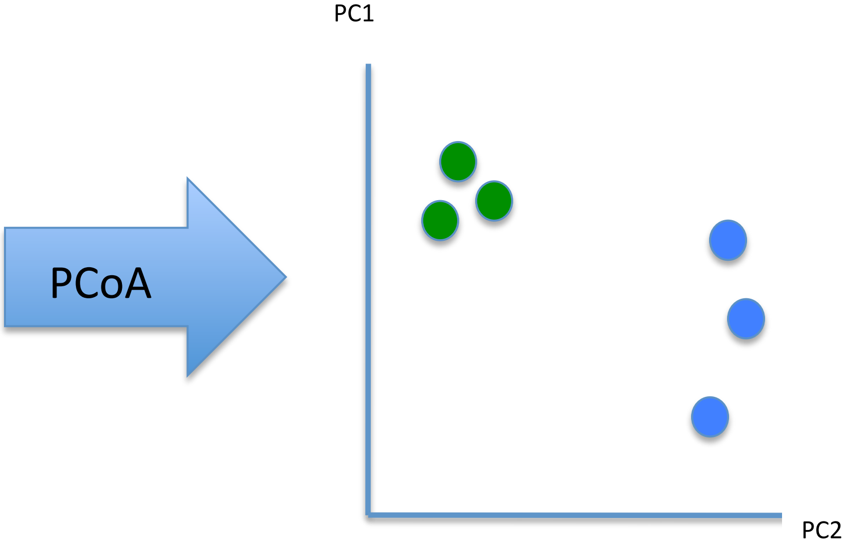
- Phylogenetically related organisms are more likely to have similar roles in a community.
- Taxon-based methods assume a "star phylogeny"



Principal Coordinates Analysis (PCoA)

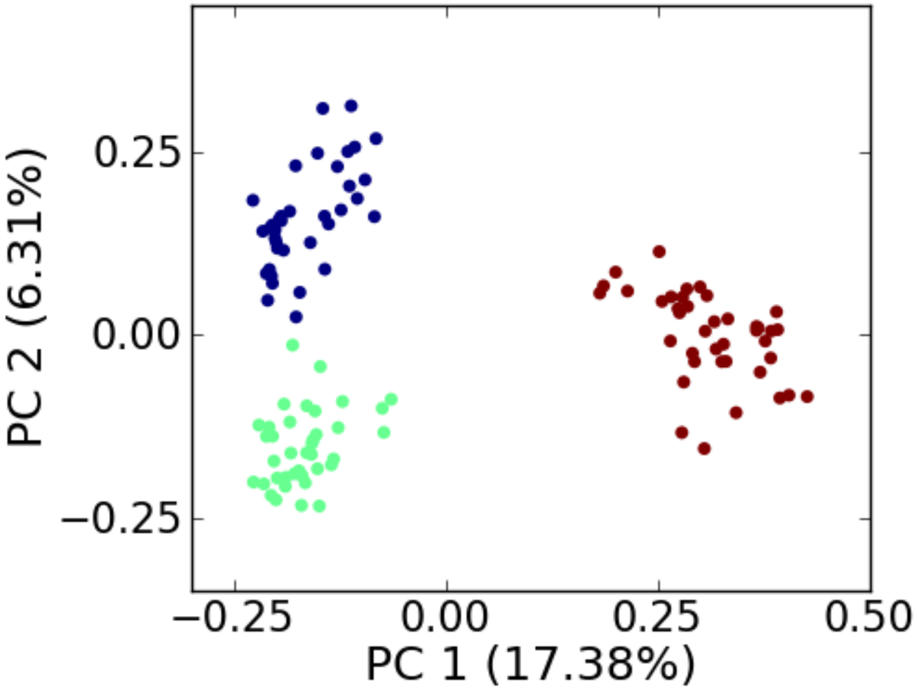
Distance Matrix

						
	**0**	**0.6**	**0.5**	**0.7**	**0.67**	**0.46**
	**0.6**	**0**	**0.74**	**0.54**	**0.37**	**0.7**
	**0.5**	**0.74**	**0**	**0.63**	**.59**	**0.48**
	**0.7**	**0.54**	**0.63**	**0**	**0.46**	**0.75**
	**0.67**	**0.37**	**0.59**	**0.46**	**0**	**0.63**
	**0.46**	**0.7**	**0.48**	**0.75**	**0.63**	**0**

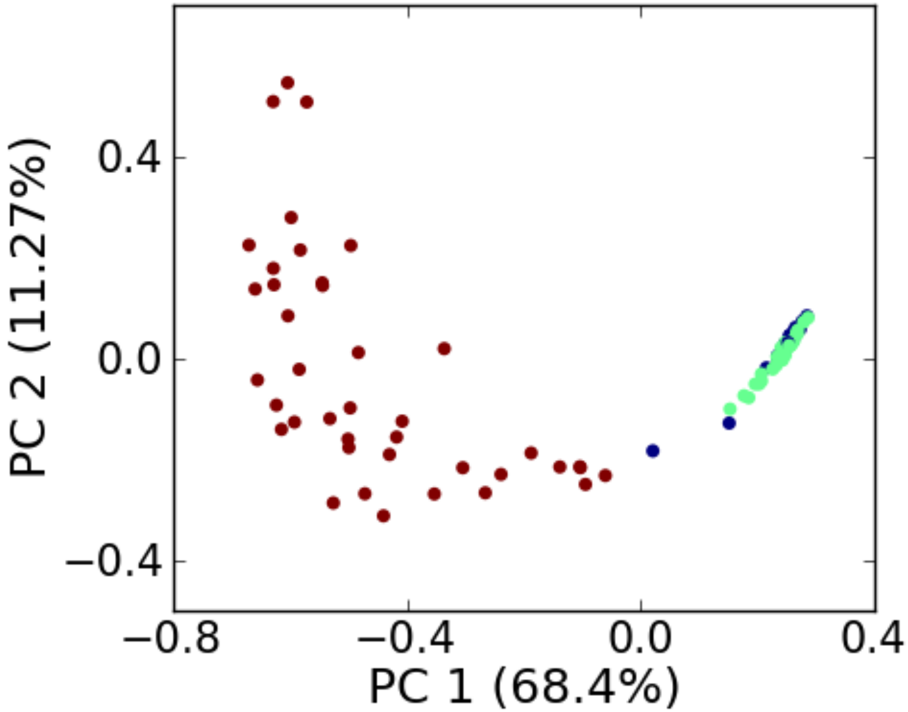


Metric DOES matter

metric 1 (Jaccard)

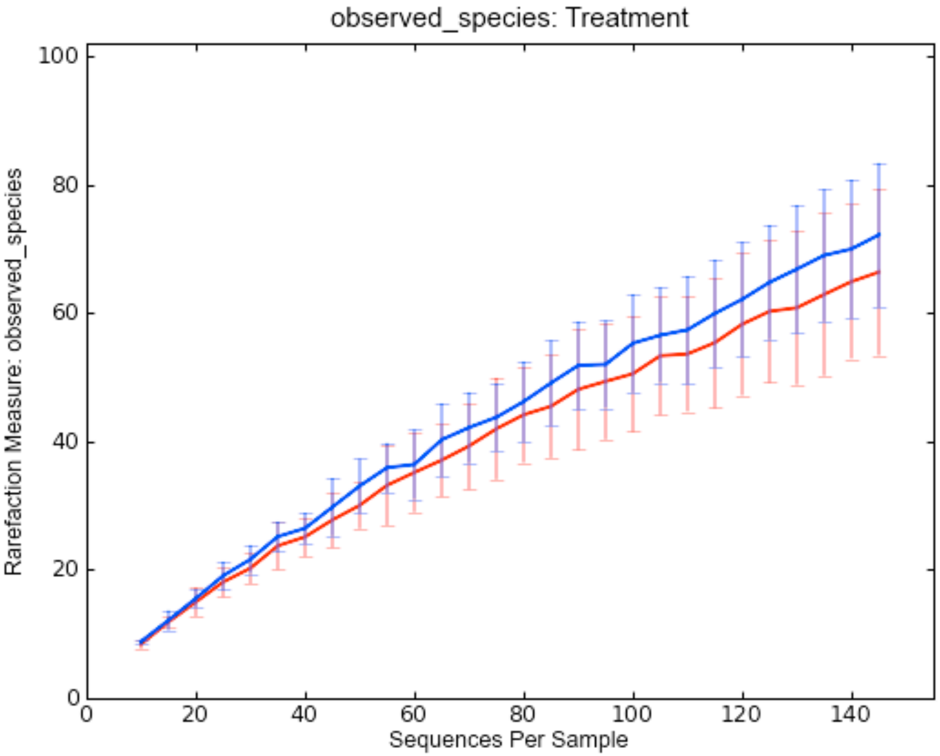


metric 2 (Morisita-Horn)

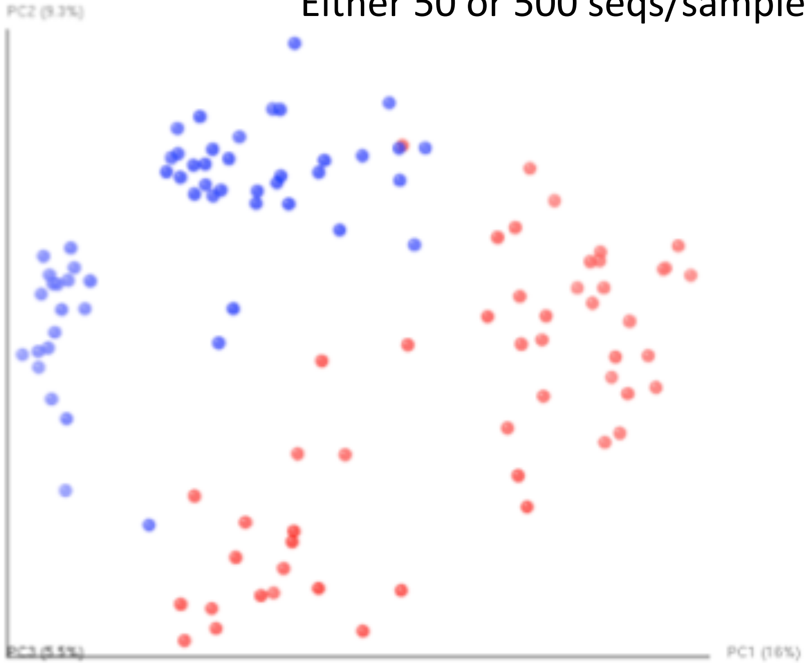


Fierer et al. *PNAS* 2010

Sequencing depth artifacts can affect diversity estimates



Alpha-diversity



Colored by number of
sequences per sample
Either 50 or 500 seqs/sample

Beta-diversity

How to account for uneven sequence counts across samples

- Use frequencies to normalize by total sample sequence count
  - we recommend against this approach, especially for diversity calculations. We have found that it can lead to samples clustering by sequencing depth.
- Rarefaction
  - equal numbers of sequences are randomly selected from each sample

Rarefaction

- How do you pick rarefaction depth?
- Major disadvantage to rarefaction
  - – valuable data from high-sequence count samples are discarded and/or samples discarded
- Rarefaction has recently been shown to introduce errors in analyses

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

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

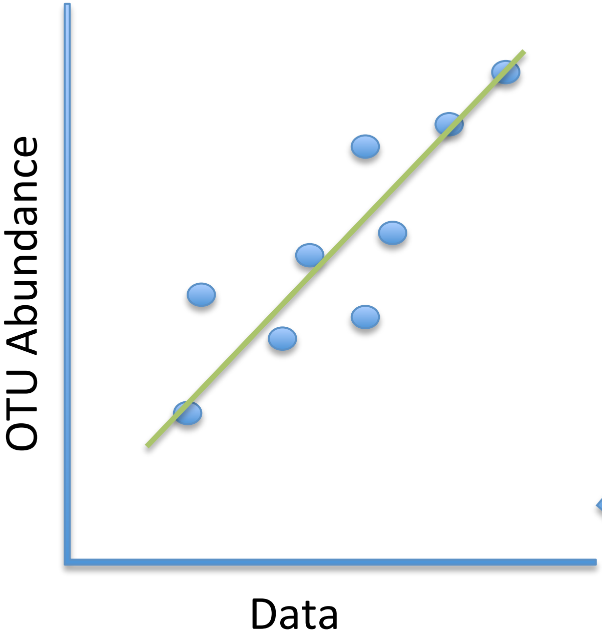
Paul J. McMurdie, Susan Holmes*

http://joey711.github.io/phyloseq/

Relating OTU/Taxonomy abundance to Metadata

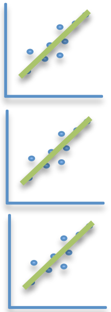
- Correlation
  - Look for correlations between OTU/Taxonomy abundance and continuous metadata.
- Searching for significantly different OTUs
  - Determine if OTU/Taxonomy abundance is significantly different between categorical metadata.
- Clustering
  - Unsupervised clustering of samples given OUT/Taxonomy abundances.
- Classification/prediction
  - Given categorical metadata what OTUs separate the data and how well

Correlation

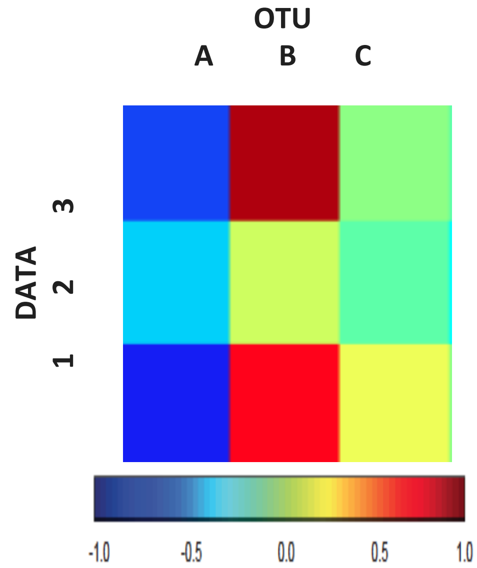


For every OTU

						
A	**10**	**1**	**9**	**0**	**2**	**8**
B	**1**	**65**	**0**	**88**	**90**	**0**
C	**0**	**0**	**1**	**0**	**0**	**0**



For all continuous metadata



Relating OTU/Taxonomy abundance to Metadata

- Correlation
  - Look for correlations between OTU/Taxonomy abundance and continuous metadata.
- Searching for significantly different OTUs
  - Determine if OTU/Taxonomy abundance is significantly different between categorical metadata.
- Clustering
  - Unsupervised clustering of samples given OUT/Taxonomy abundances.
- Classification/prediction
  - Given categorical metadata what OTUs separate the data and how well

Searching for significant OTUs

Which features (OTUs) of your data are most different between sample classes?

Disease

Control

	Sample_1	Sample_2	Sample_3	Sample_4	Sample_5	Sample_6
OTU_1	100	150	1000	250	275	600
OTU_2	345	297	611	35	14	0

Give me the tests!

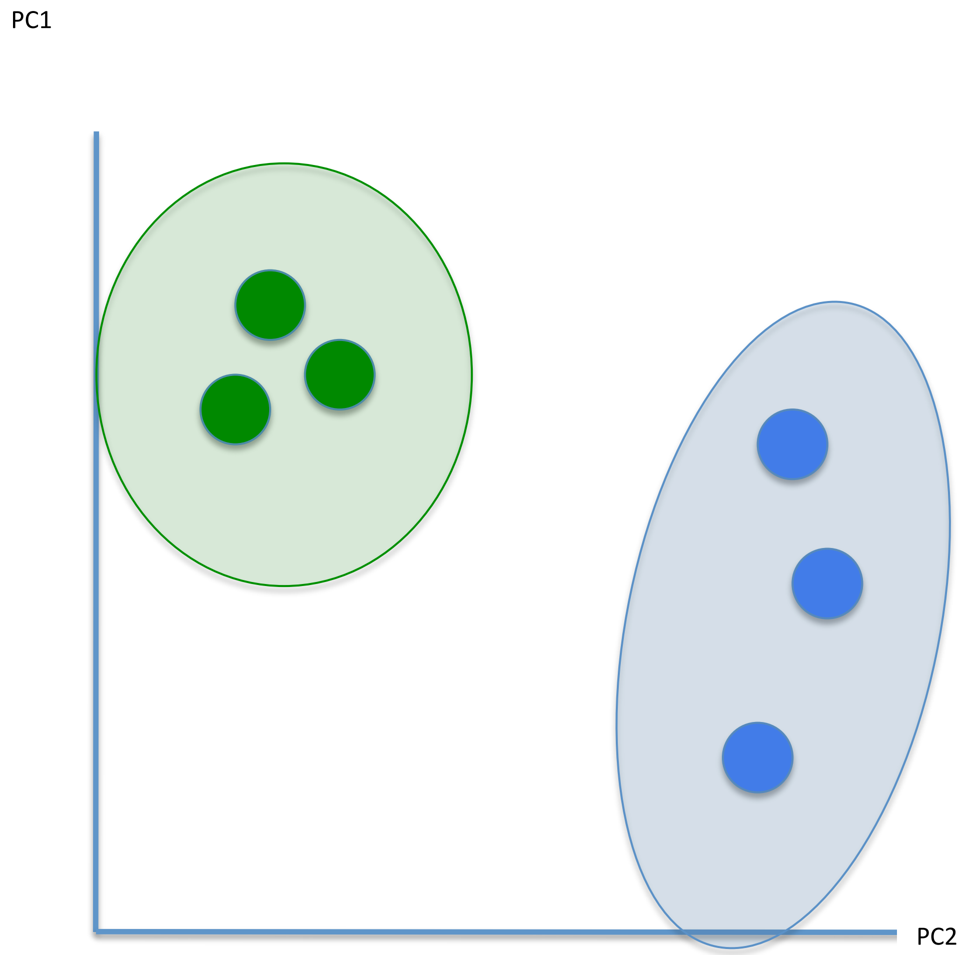
- G-test: the graphical example we saw. Originally developed for single value experiments
- ANOVA: test differences in means
- T-test: ANOVA for 2 groups
- Kruskal-Wallis: Non parametric ANOVA
- Mann-Whitney-U: Kruskal-Wallis for 2 groups
- Bootstrap: Randomizes labels and performs the given test n times and the p-value is = better_or_equal_test_statistic/random_tests

Relating OTU/Taxonomy abundance to Metadata

- Correlation
  - Look for correlations between OTU/Taxonomy abundance and continuous metadata.
- Searching for significantly different OTUs
  - Determine if OTU/Taxonomy abundance is significantly different between categorical metadata.
- Clustering
  - Unsupervised clustering of samples given OUT/Taxonomy abundances.
- Classification/prediction
  - Given categorical metadata what OTUs separate the data and how well

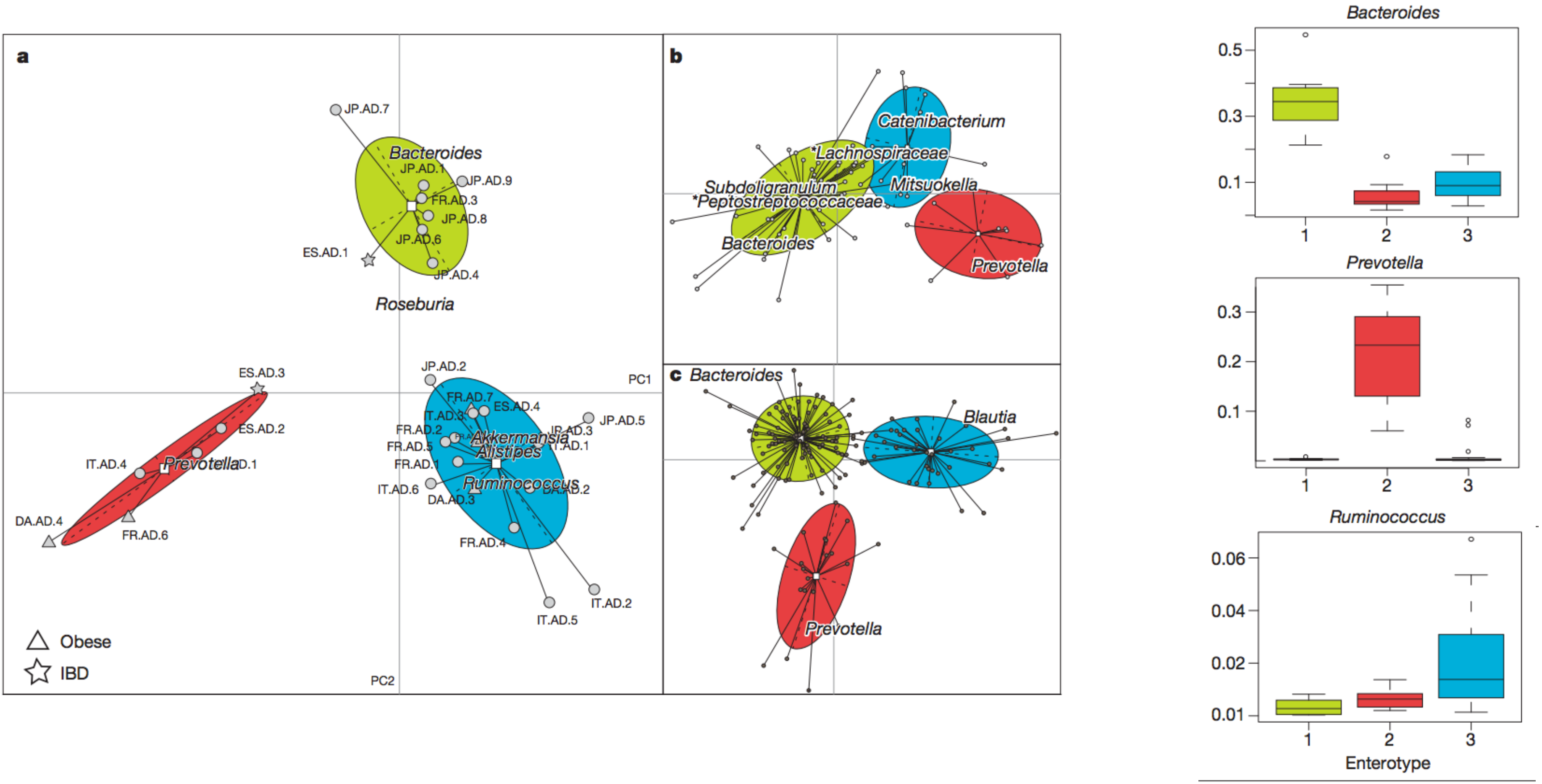
Clustering

- UPGMA
- k-means
- k-medoids
- …many other possibilities

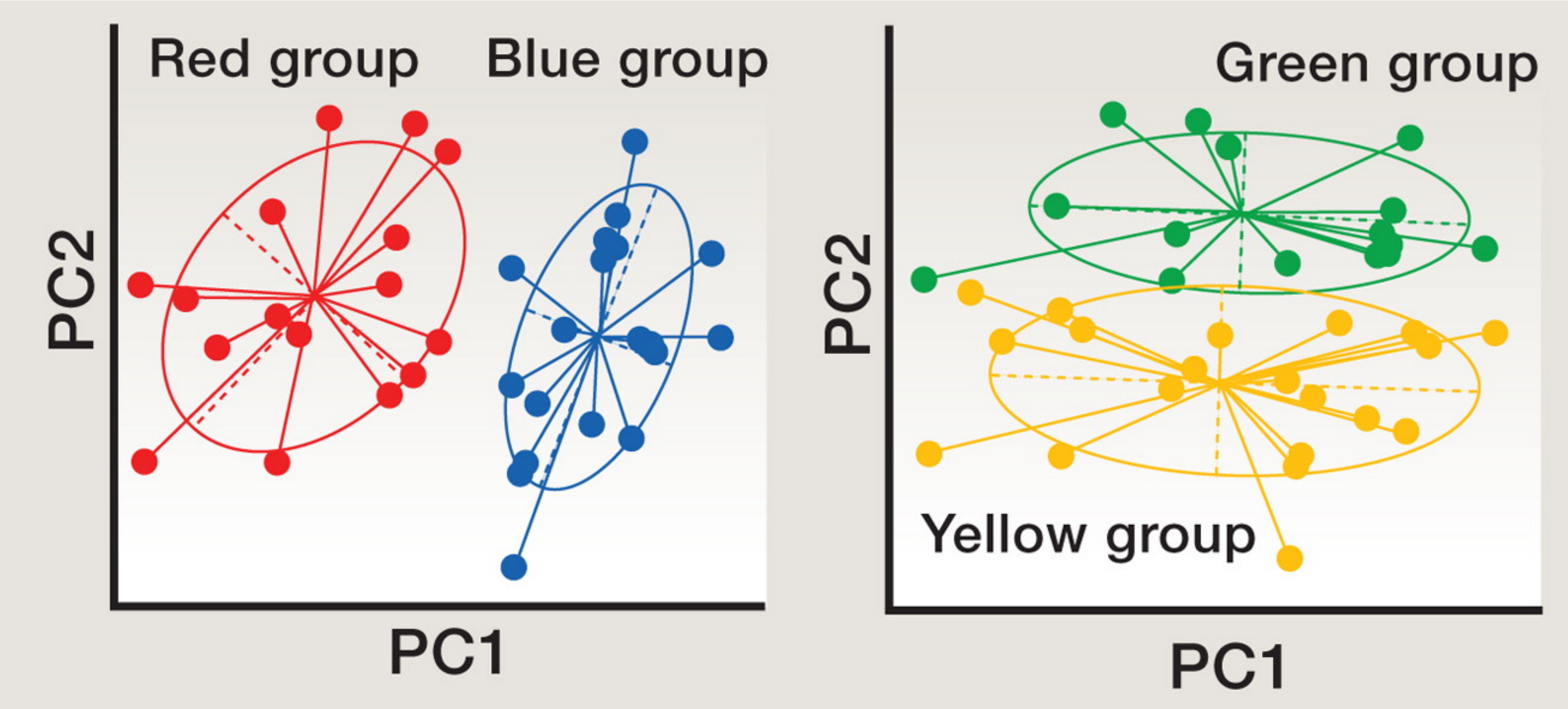


Enterotypes of the human gut microbiome

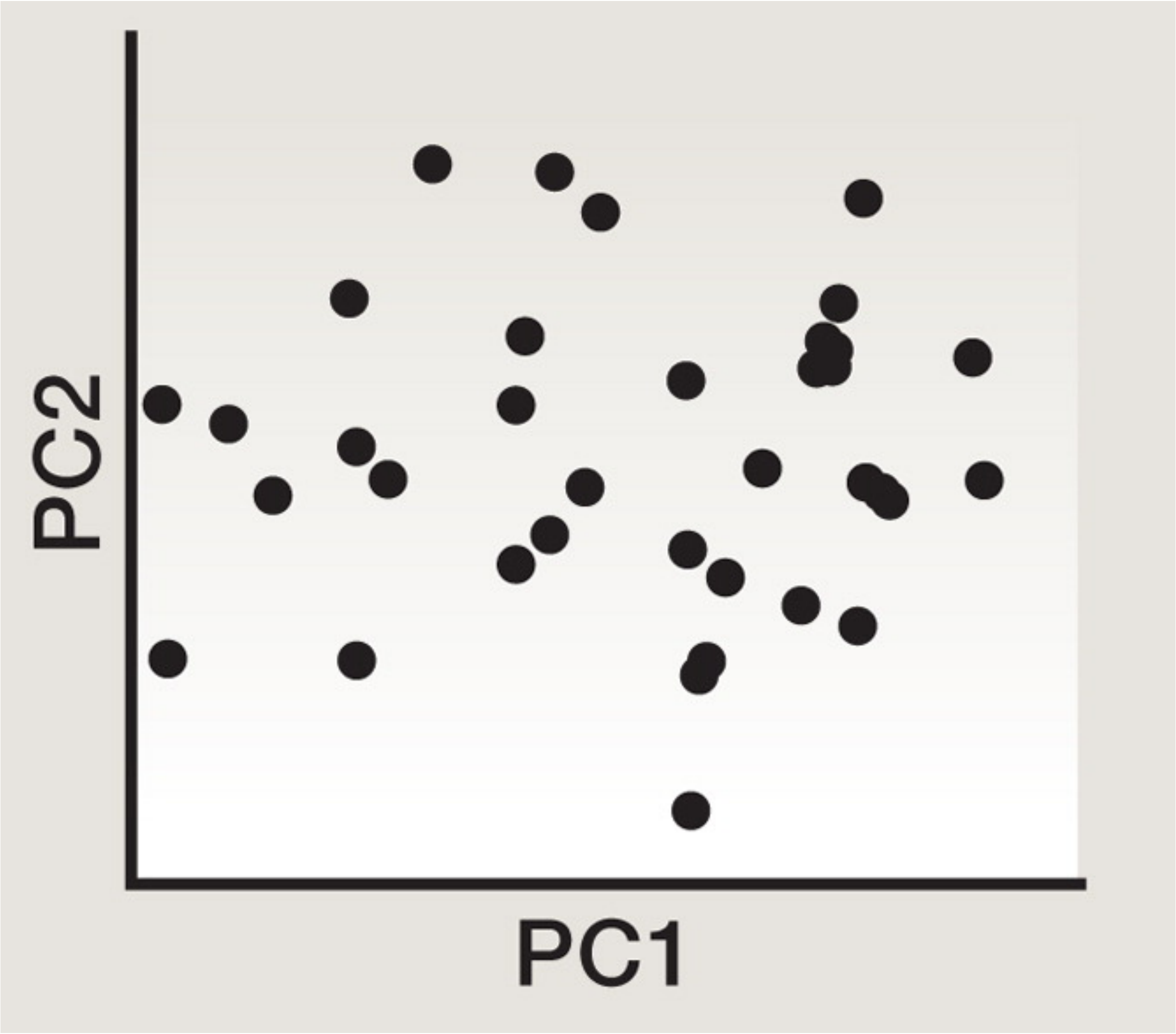
Manimozhiyan Arumugam[1]*, Jeroen Raes[1,2]*, Eric Pelletier[3,4,5], Denis Le Paslier[3,4,5], Takuji Yamada[1], Daniel R. Mende[1], Gabriel R. Fernandes[1,6], Julien Tap[1,7], Thomas Bruls[3,4,5], Jean-Michel Batto[7], Marcelo Bertalan[8], Natalia Borruel[9], Francesc Casellas[9], Leyden Fernandez[10], Laurent Gautier[8], Torben Hansen[11,12], Masahira Hattori[13], Tetsuya Hayashi[14], Michiel Kleerebezem[15], Ken Kurokawa[16], Marion Leclerc[7], Florence Levenez[7], Chaysavanh Manichanh[9], H. Bjørn Nielsen[8], Trine Nielsen[11], Nicolas Pons[7], Julie Poulain[3], Junjie Qin[17], Thomas Sicheritz-Ponten[8,18], Sebastian Tims[15], David Torrents[10,19], Edgardo Ugarte[3], Erwin G. Zoetendal[15], Jun Wang[17,20], Francisco Guarner[9], Oluf Pedersen[11,21,22,23], Willem M. de Vos[15,24], Søren Brunak[8], Joel Doré[7], MetaHIT Consortium†, Jean Weissenbach[3,4,5], S. Dusko Ehrlich[7] & Peer Bork[1,25]



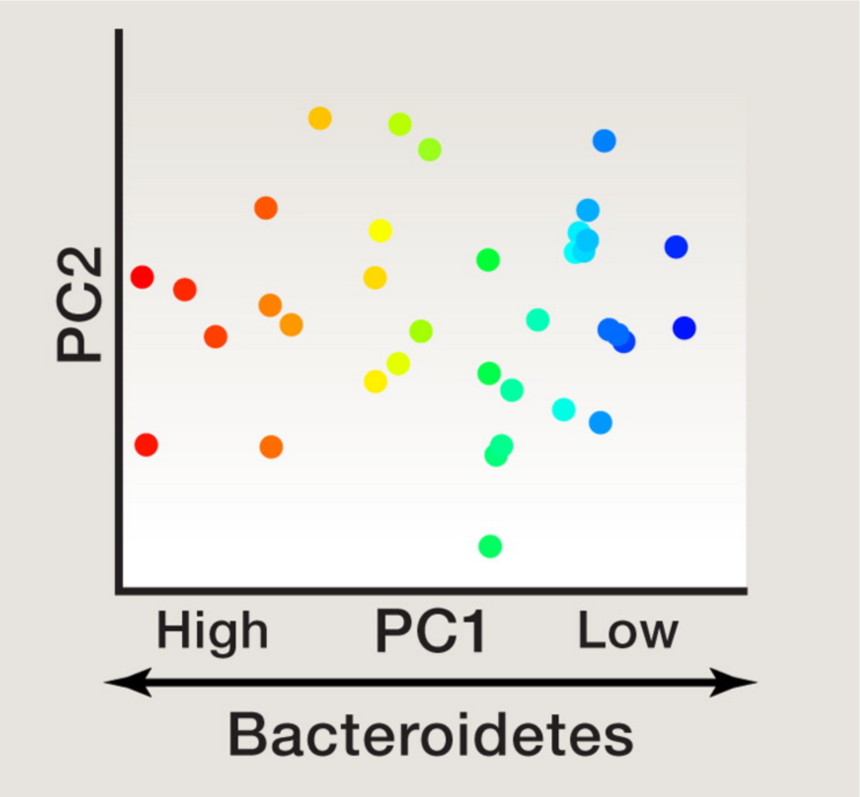
You can always find clusters….

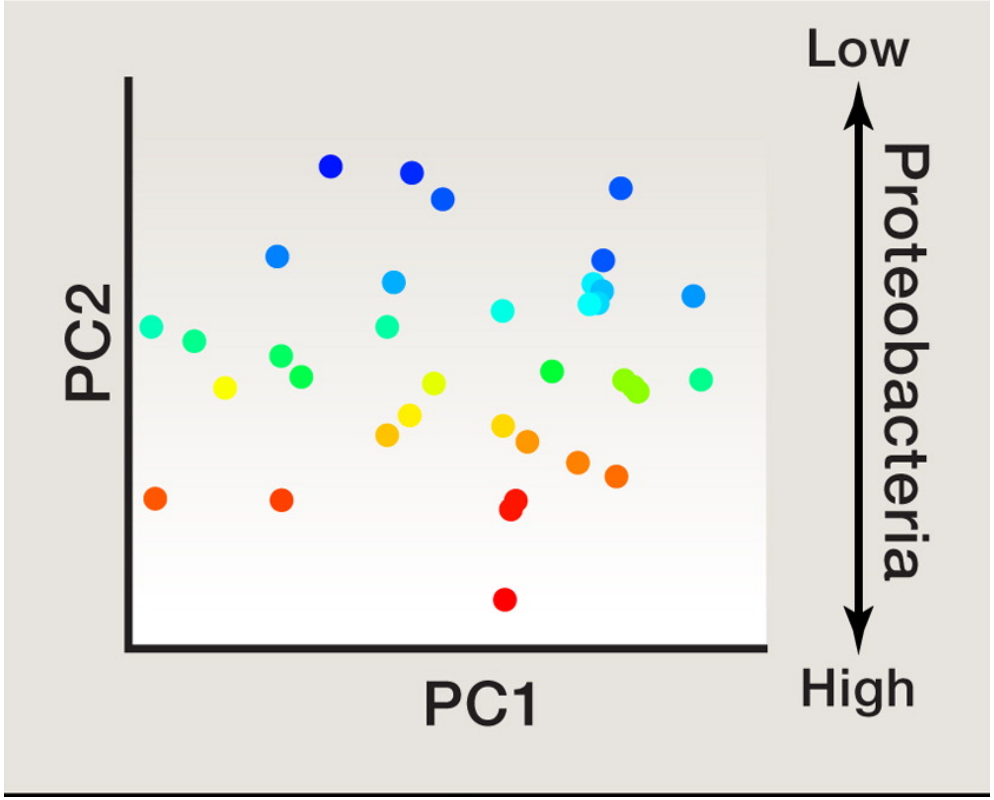


…but are they actually clusters…

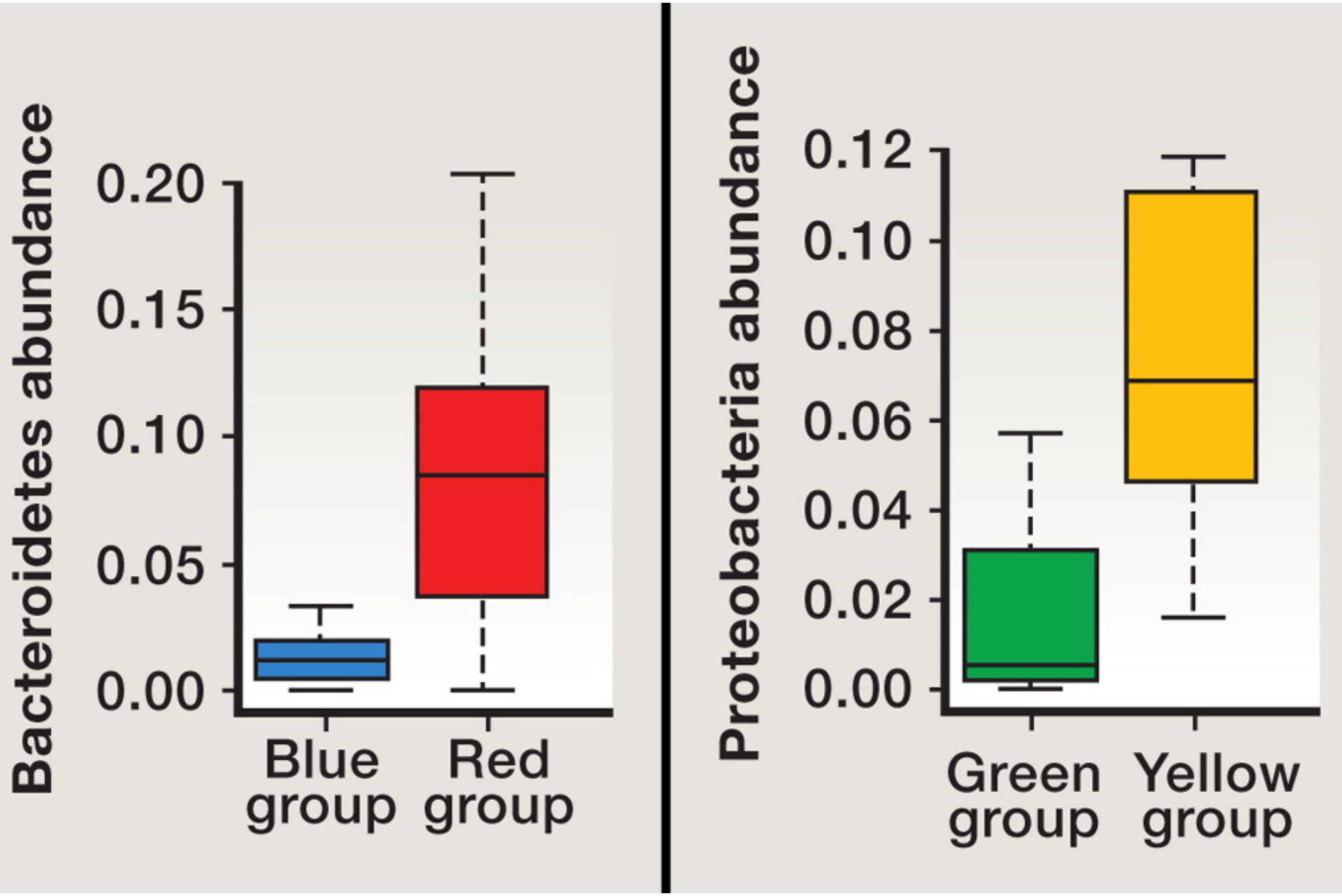


…or just a gradient.





Plotting mean values of the abundances of taxa that drive the gradients in the PCoA plots does not constitute a validation of the clustering patterns



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets

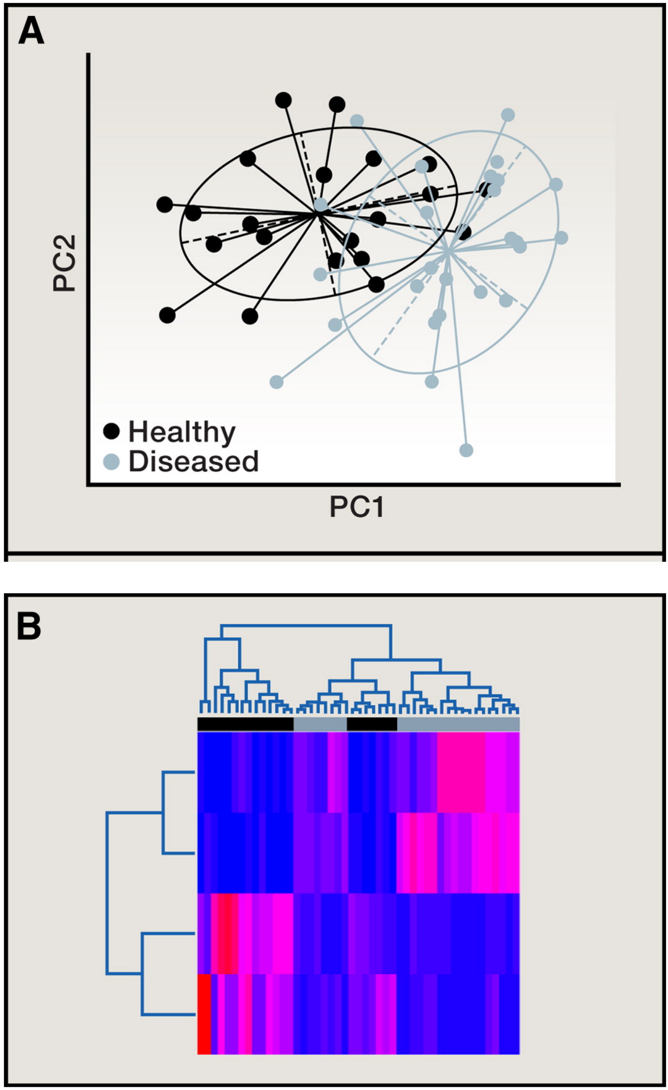
Omry Koren[1☯], Dan Knights[2☯], Antonio Gonzalez[2☯], Levi Waldron[3,4], Nicola Segata[3], Rob Knight[5,6], Curtis Huttenhower[3], Ruth E. Ley[1*]

Relating OTU/Taxonomy abundance to Metadata

- Correlation
  - Look for correlations between OTU/Taxonomy abundance and continuous metadata.
- ANOVA
  - Determine if OTU/Taxonomy abundance is significantly different between categorical metadata.
- Clustering
  - Unsupervised clustering of samples given OTU/Taxonomy abundances.
- Classification/prediction
  - Given categorical metadata what OTUs separate the data and how well.

Classification

- Supervised machine learning
- What does it tell you?
  - Whether or not your samples are separated by a group of features
  - What features best separate your groups

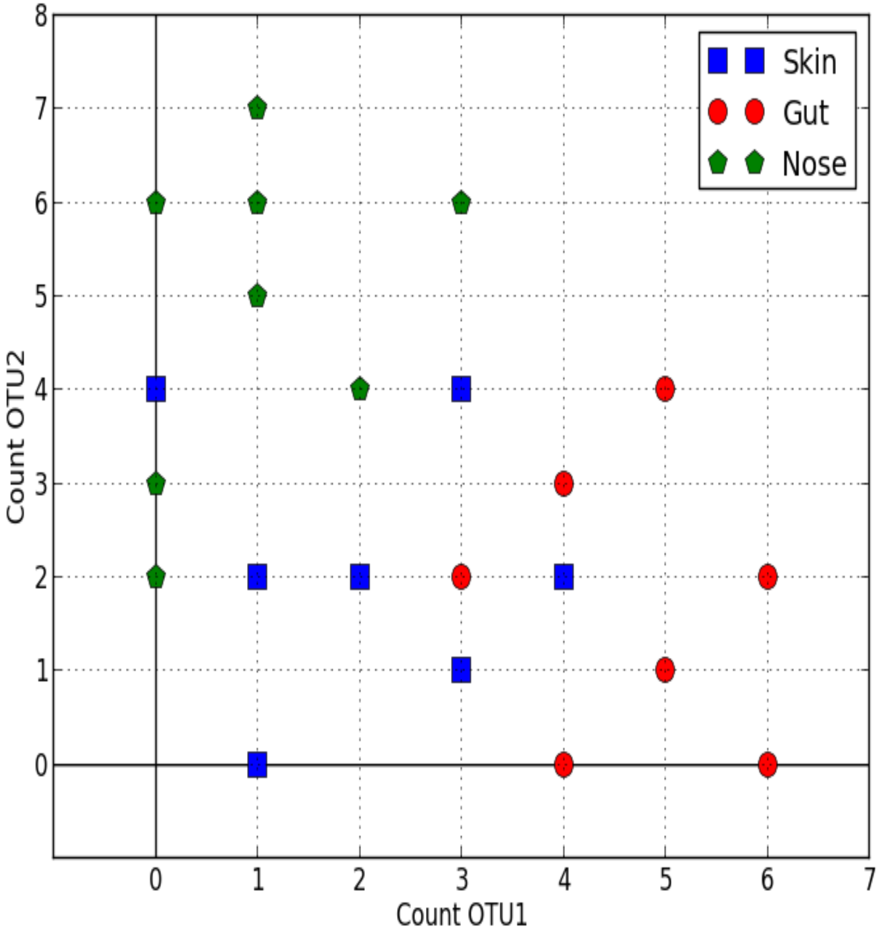


Classification

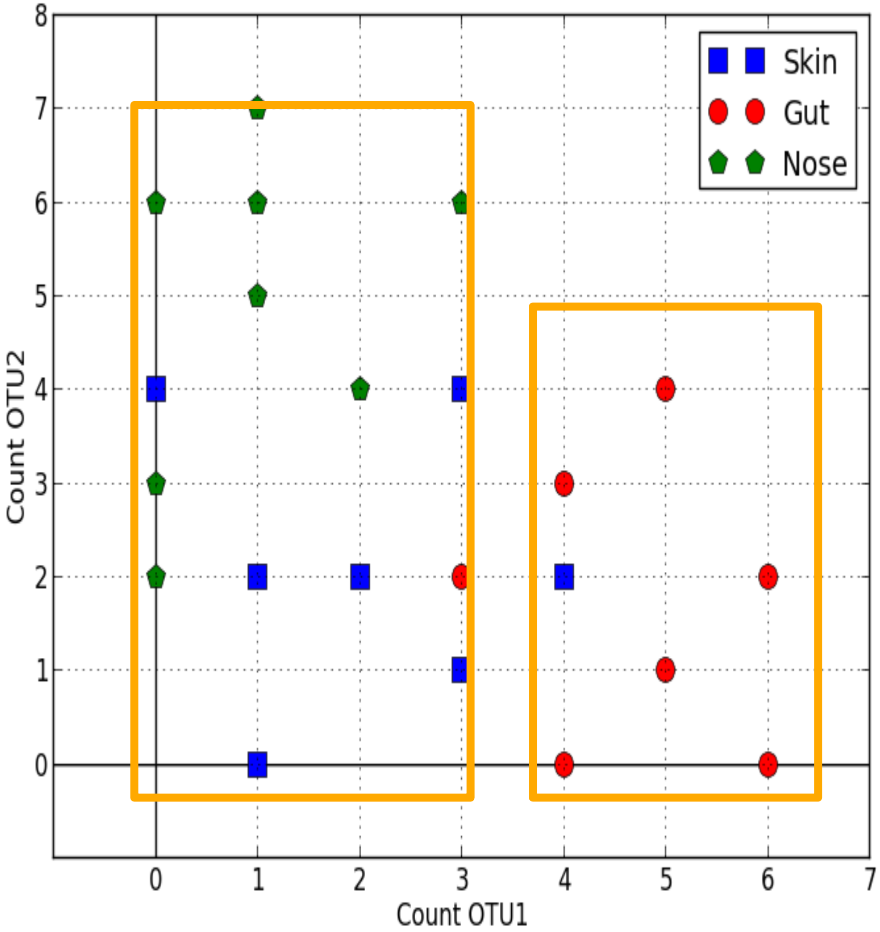
- QIIME
  - Only implements random forests classification

Examples of commonly employed supervised classification methods are described in Knights et al. (2011))

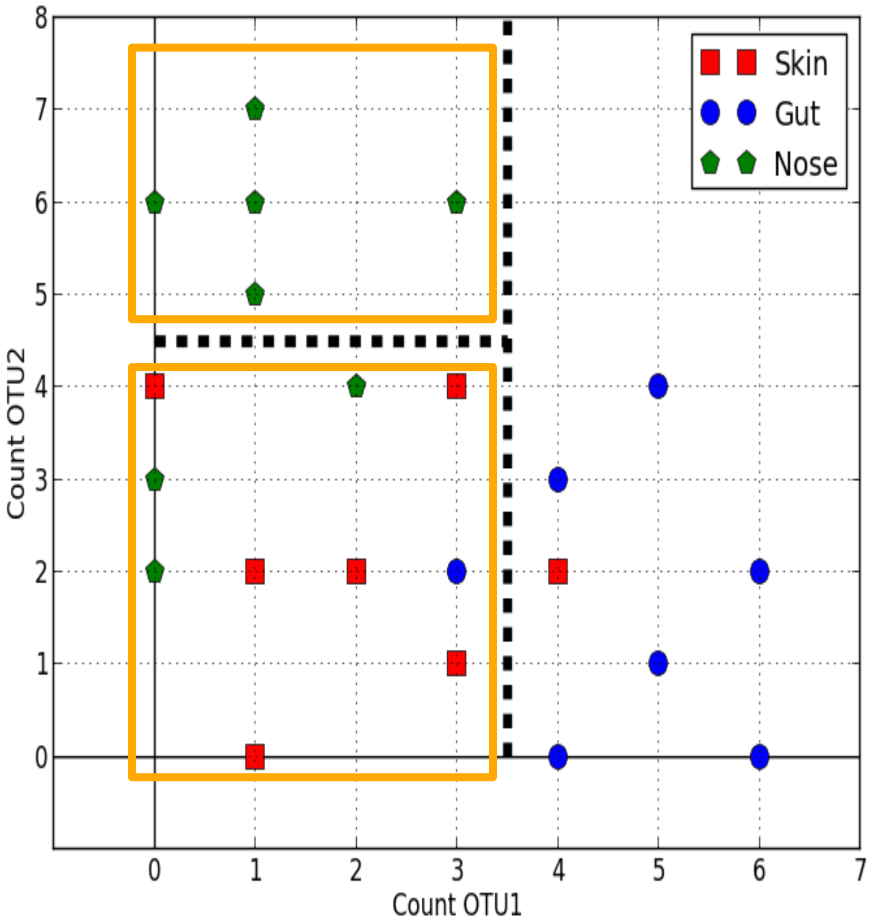
How does it work?

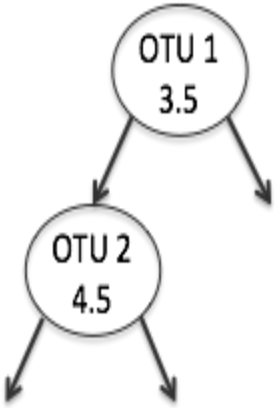


How does it work?

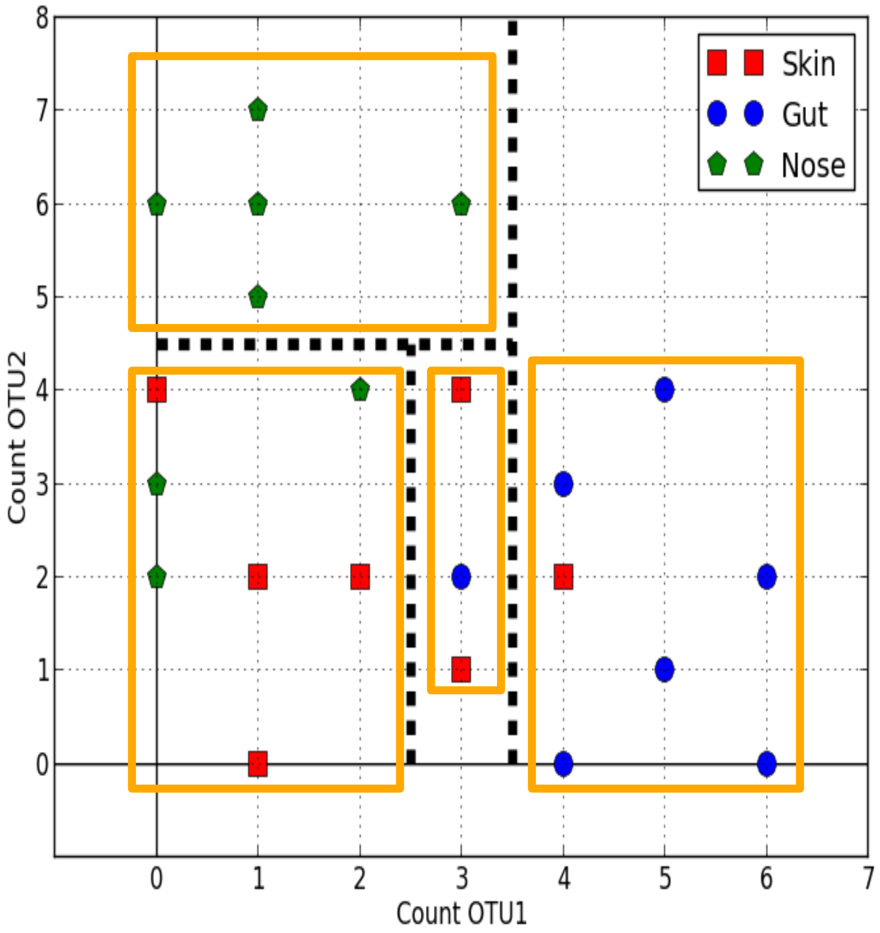


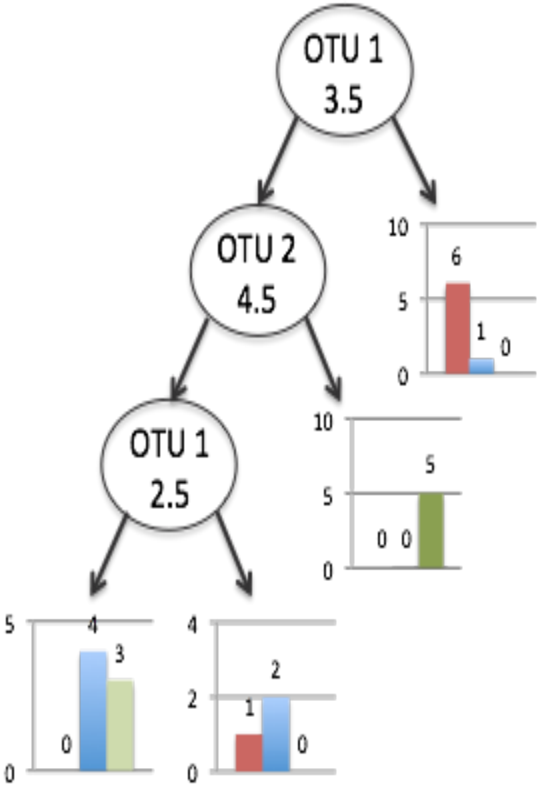
How does it work?



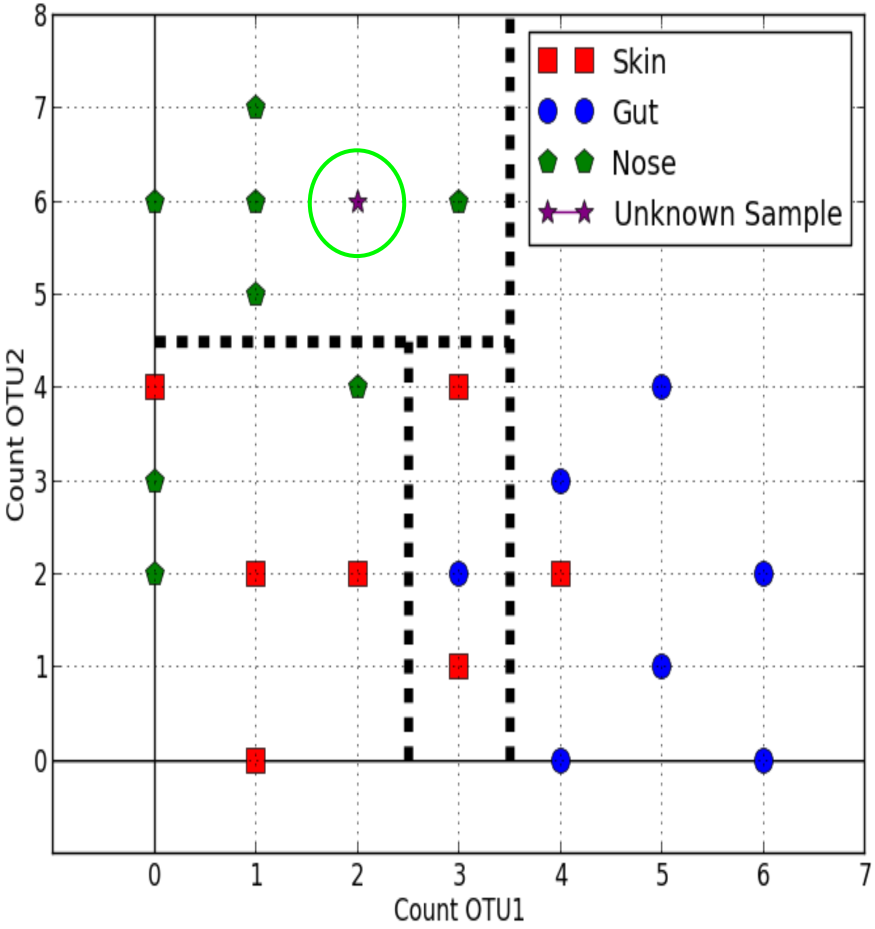


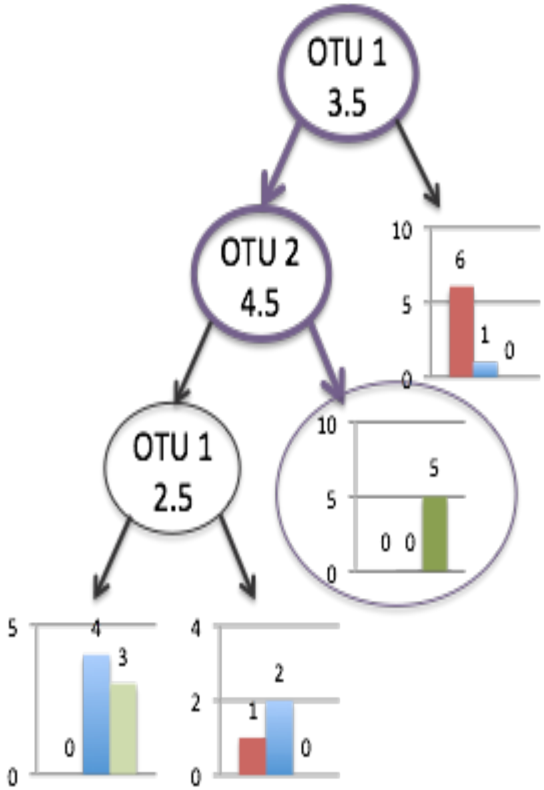
How does it work?





How does it work?





Classification

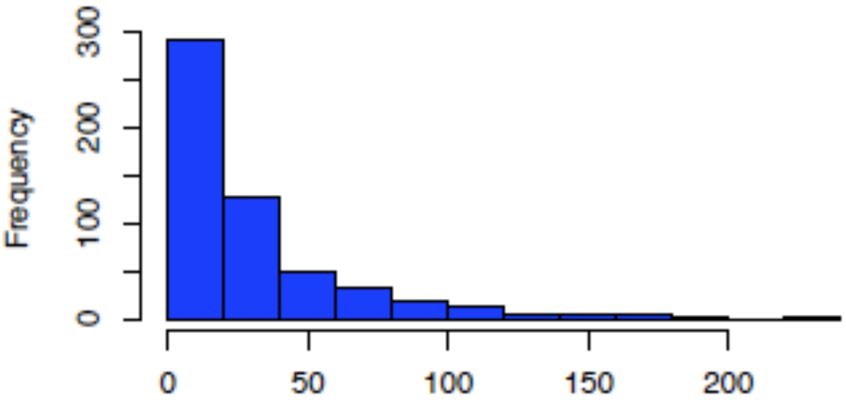
- QIIME
  - Only implements random forests classification
- Many other resources available outside of QIIME…
  - Several R packages
  - Scikit-learn: machine learning in python

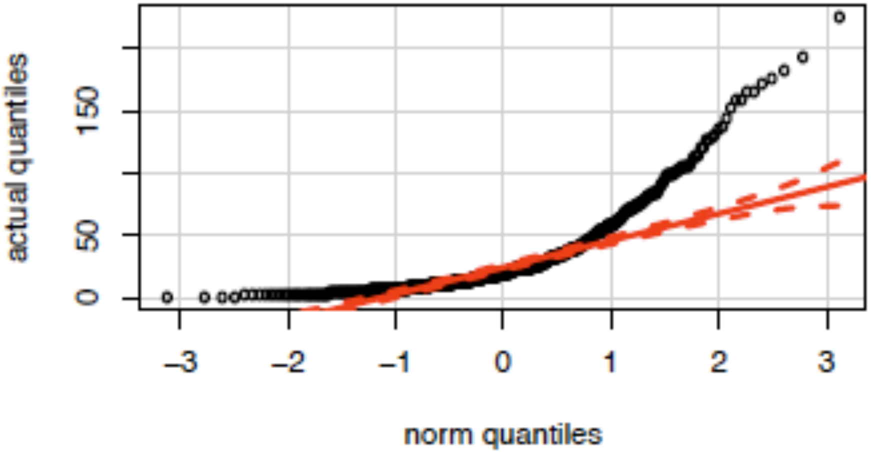
Examples of commonly employed supervised classification methods are described in Knights et al. (2011))

Keep in mind…

- Distribution of OTU/Taxonomy abundances
  - OTU abundances are not normally distributed, but pearson correlation, ANOVA and many other statistical tests in QIIME assume normality.
- Multiple testing correction
  - If you test all OTUs/Taxonomy against all metadata you are performing thousands of tests.

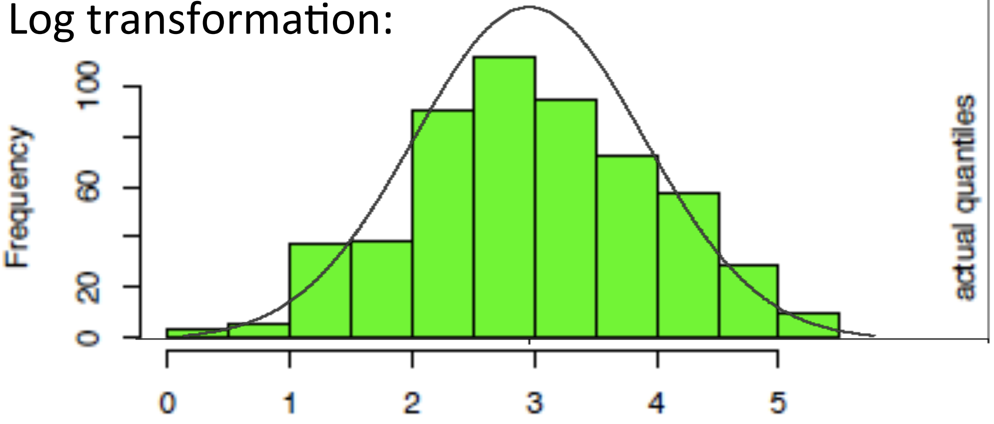
Data Normalization

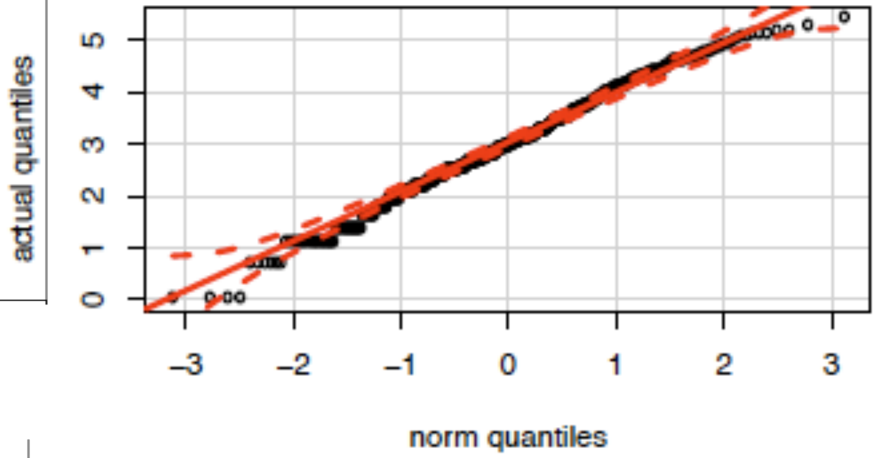




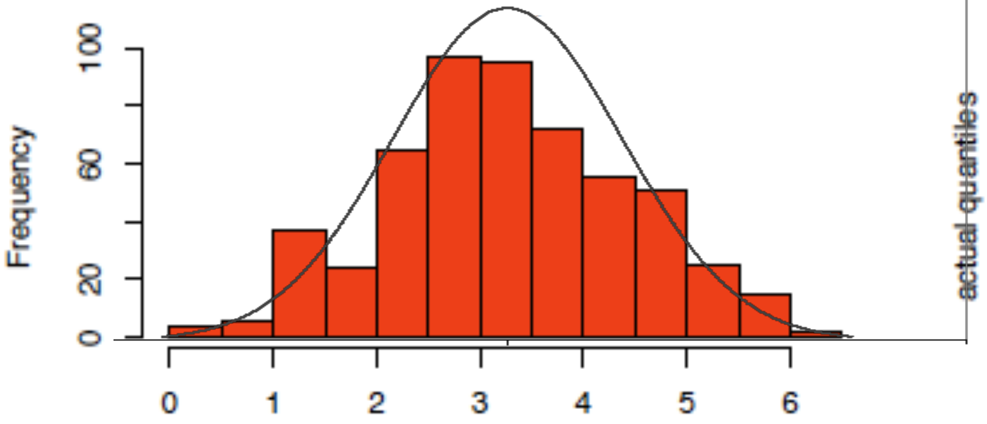
Data Normalization

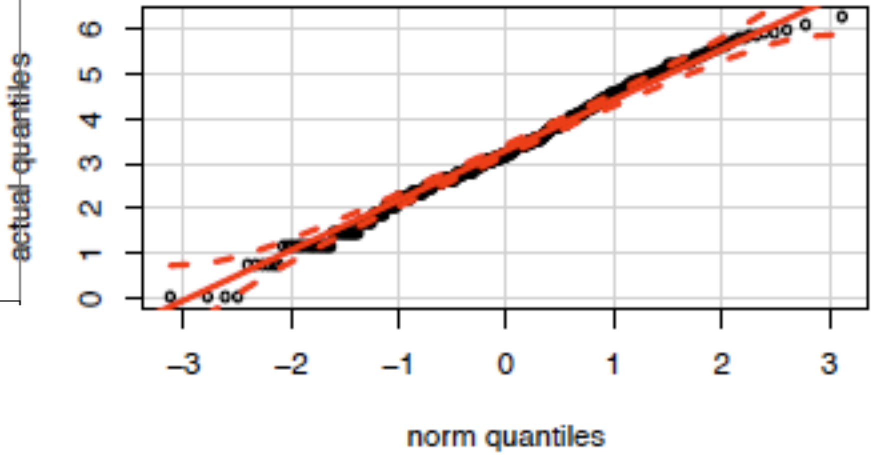
Log transformation:



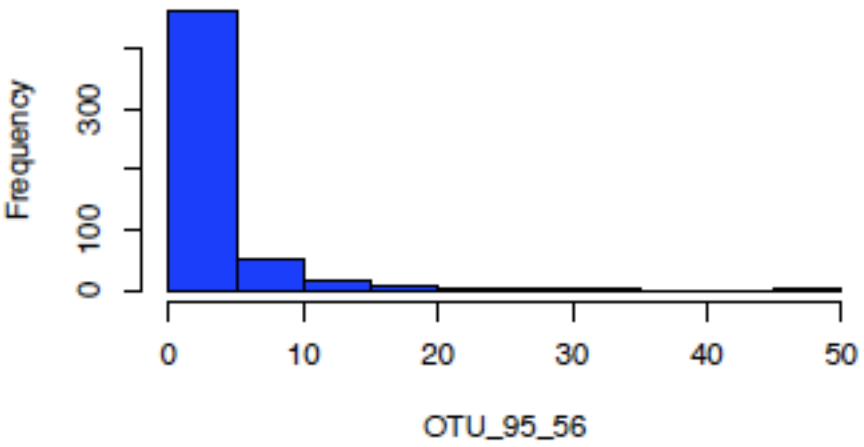


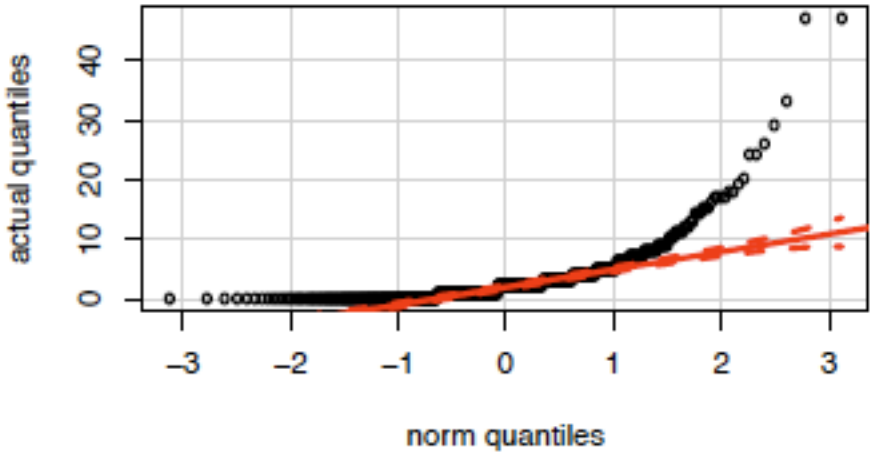
Box-Cox transformation:





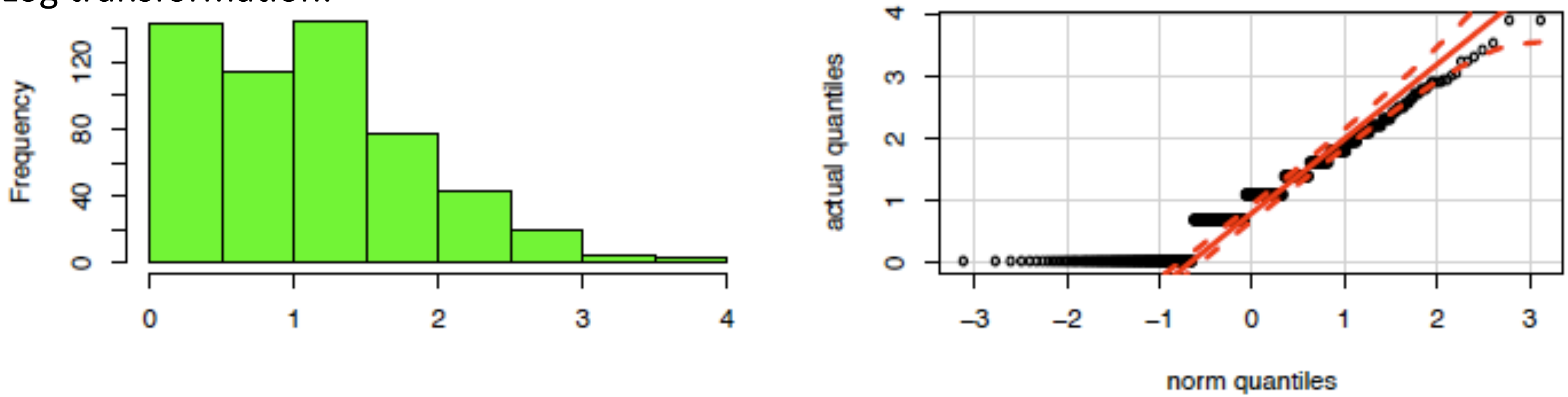
Zero-inflated OTUs



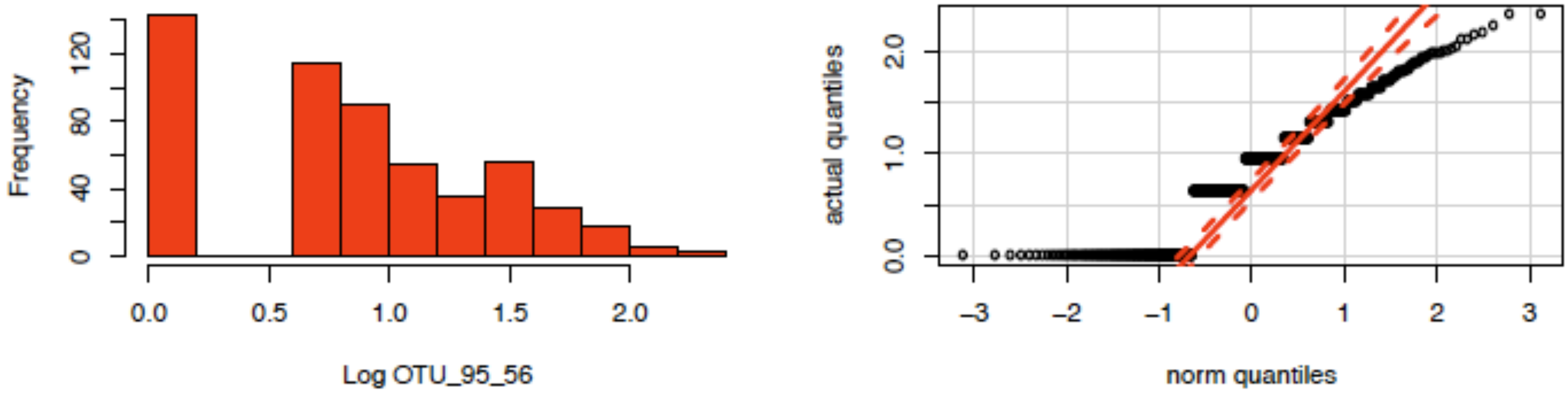


Zero-inflated OTUs

Log transformation:



Box-Cox transformation:



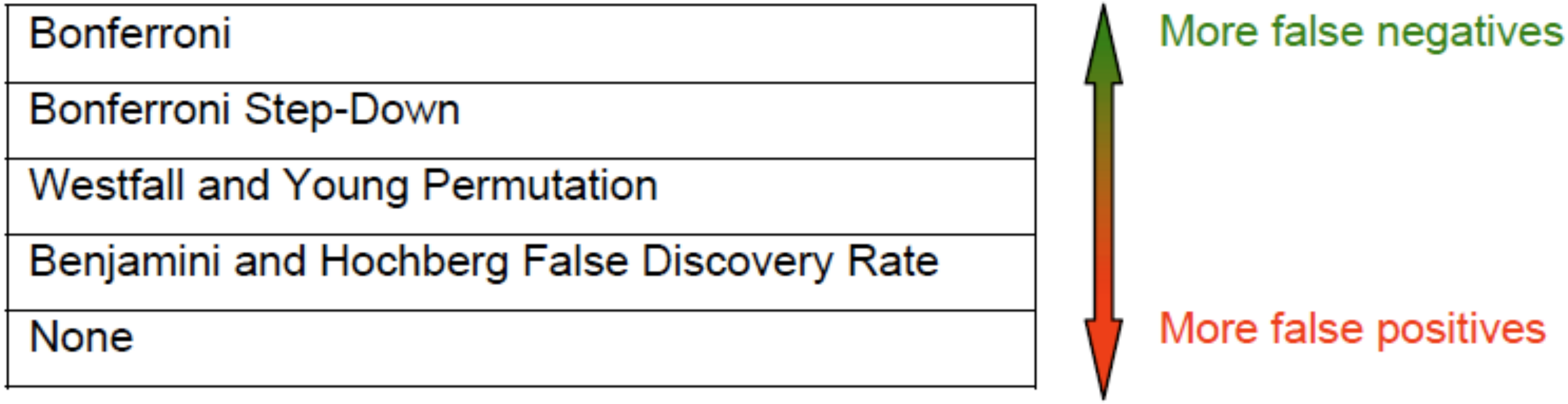
Keep in mind…

- Distribution of OTU/Taxonomy abundances
  - OTU abundances are not normally distributed, but pearson correlation, ANOVA and many other statistical tests in QIIME assume normality.

- Multiple testing correction
  - If you test all OTUs/Taxonomy against all metadata you are performing thousands of tests.

Multiple Testing

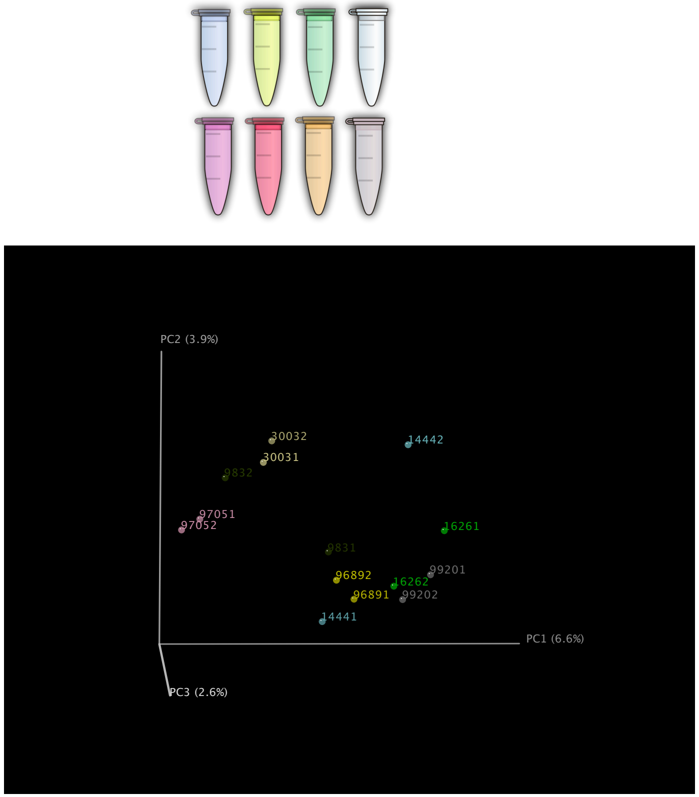
Number of genes tested (N)	False positives incidence	Probability of calling 1 or more false positives by chance ($100(1-0.95^N)$)
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

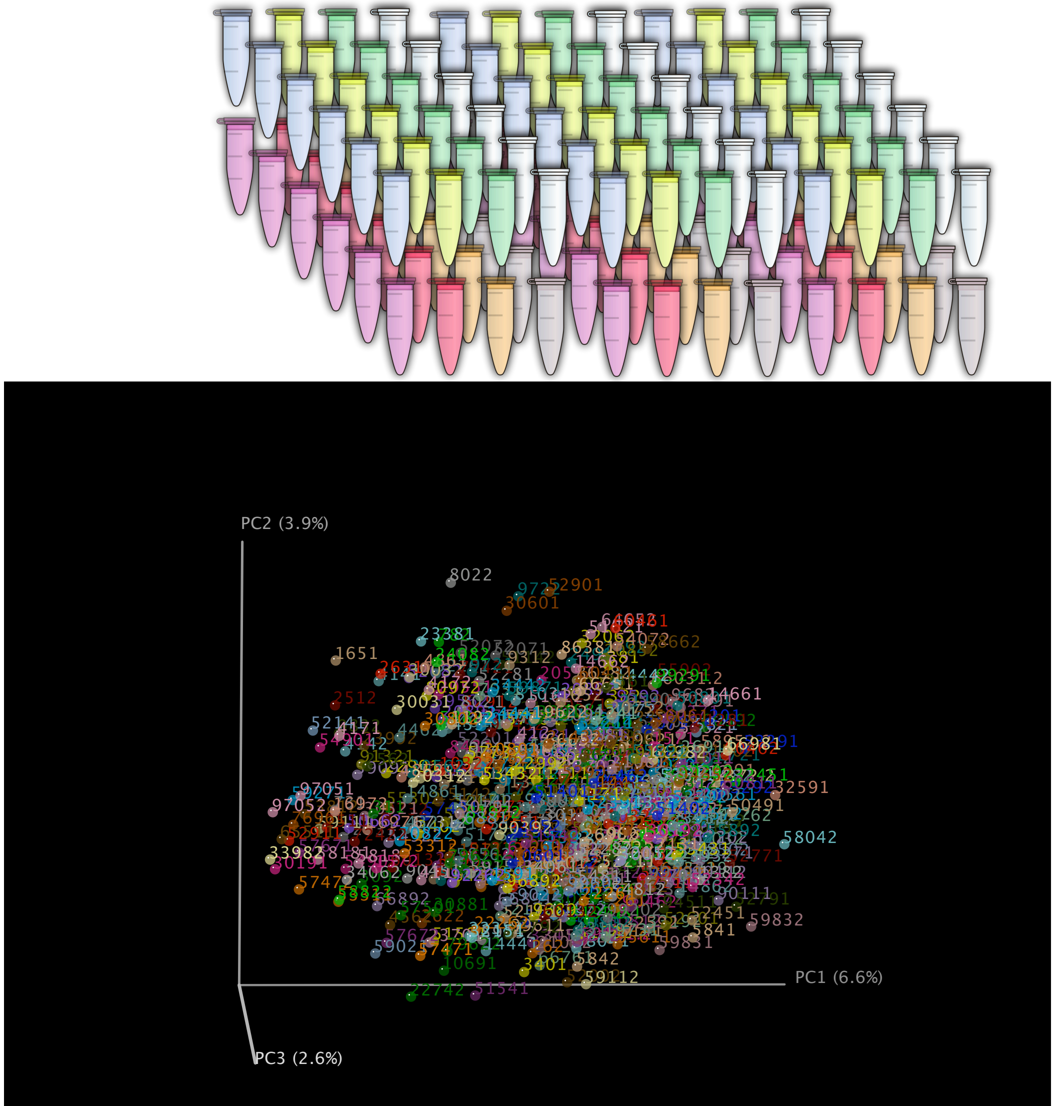


Bonferroni
Bonferroni Step-Down
Westfall and Young Permutation
Benjamini and Hochberg False Discovery Rate
None

More false negatives
More false positives

Potential Issues When Doing Larger Studies

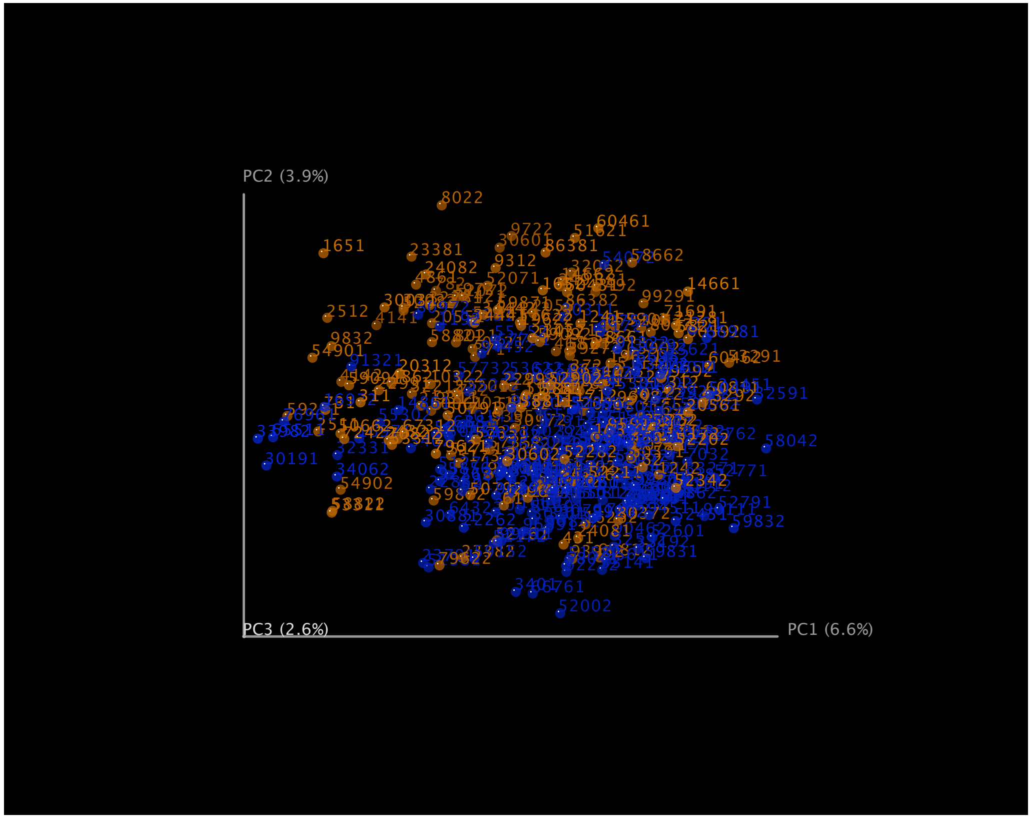




Potential Issues When Doing Larger Studies

- Different sequencing runs
- Different collection times
- Different processing times
- Change of protocol/personnel
- Multiple collection sites/sequencing sites
- BE CONSISTENT!

Batch Effect/Run Effect

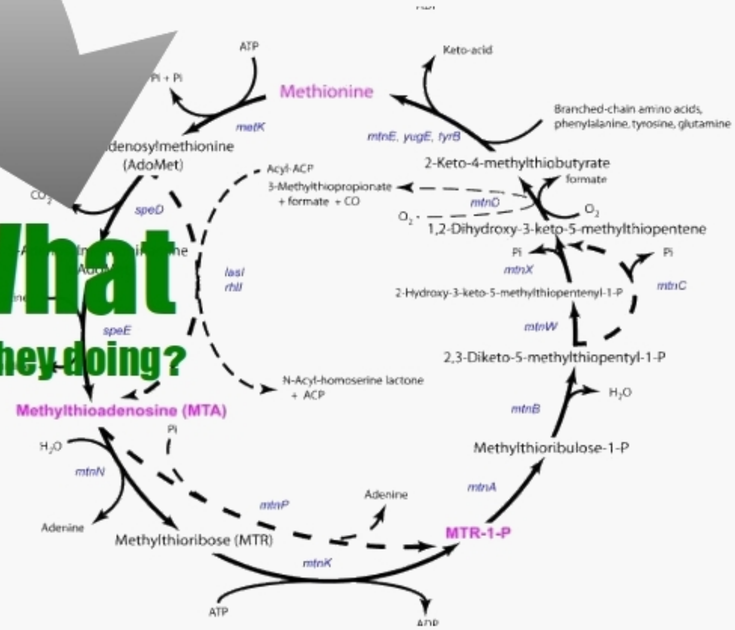


Who

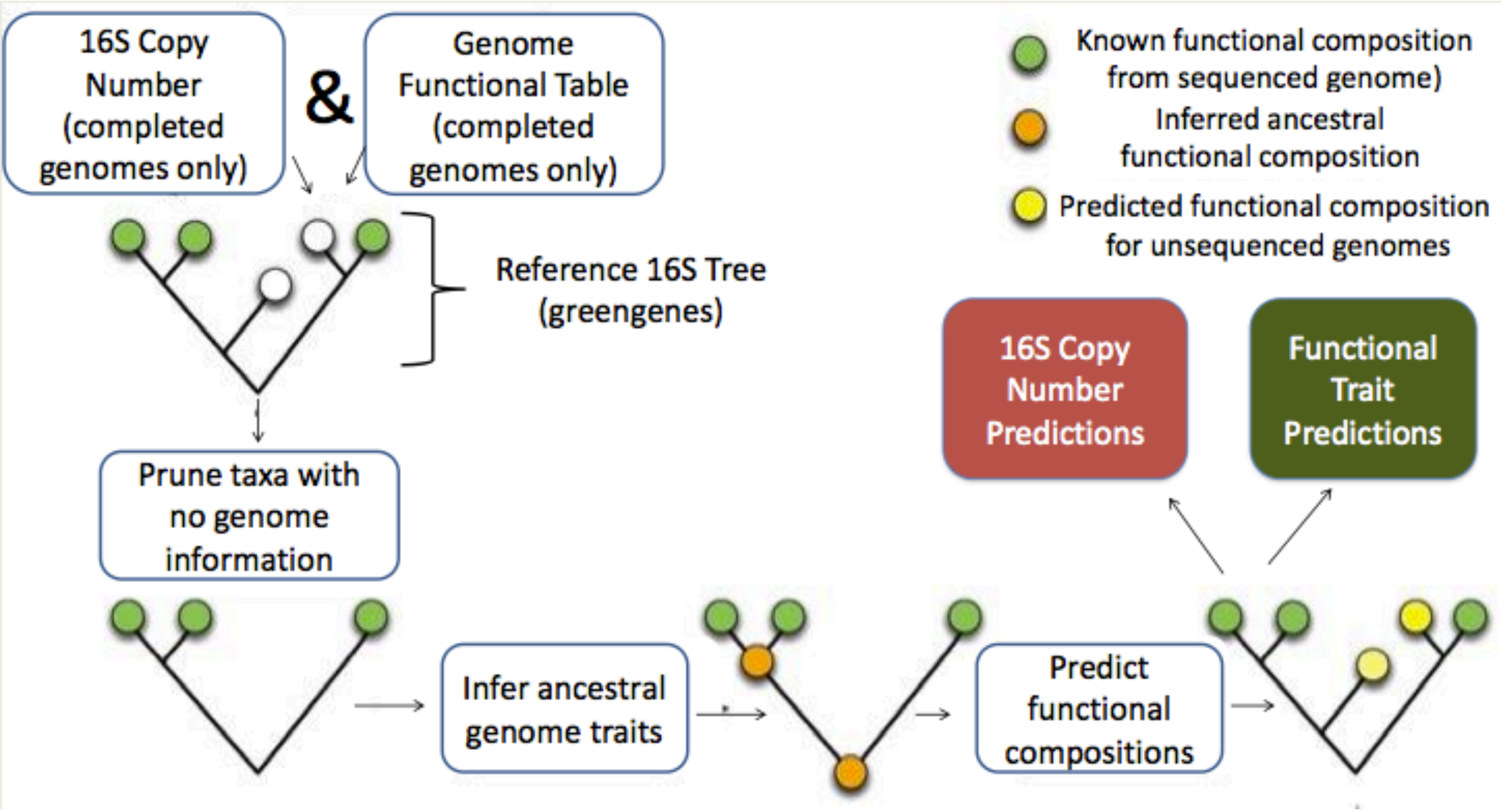
is in there?

What

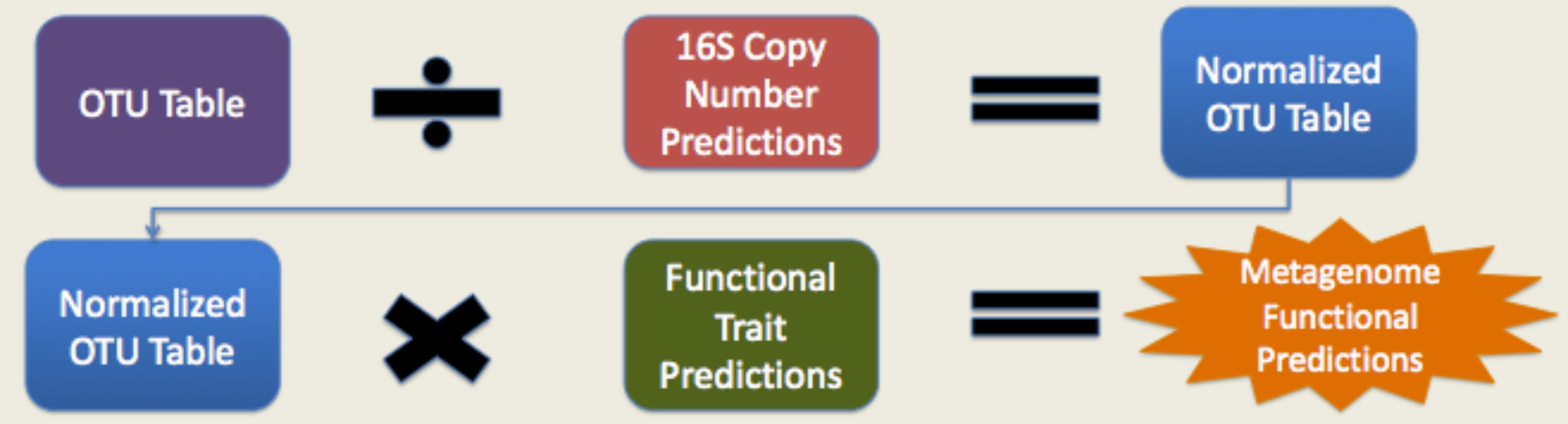
are they doing?



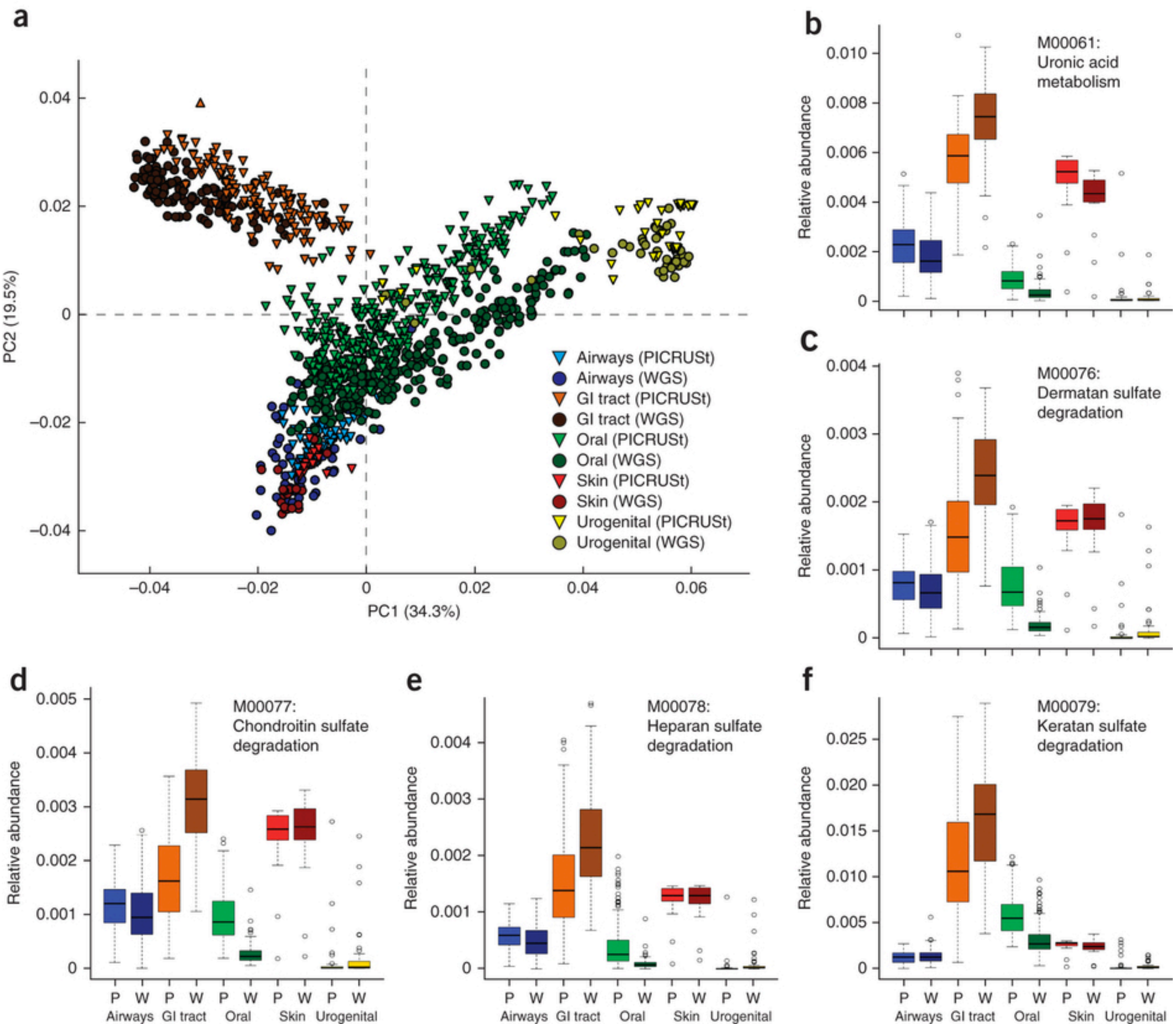
Metagenomic prediction from 16S - PICRUSt

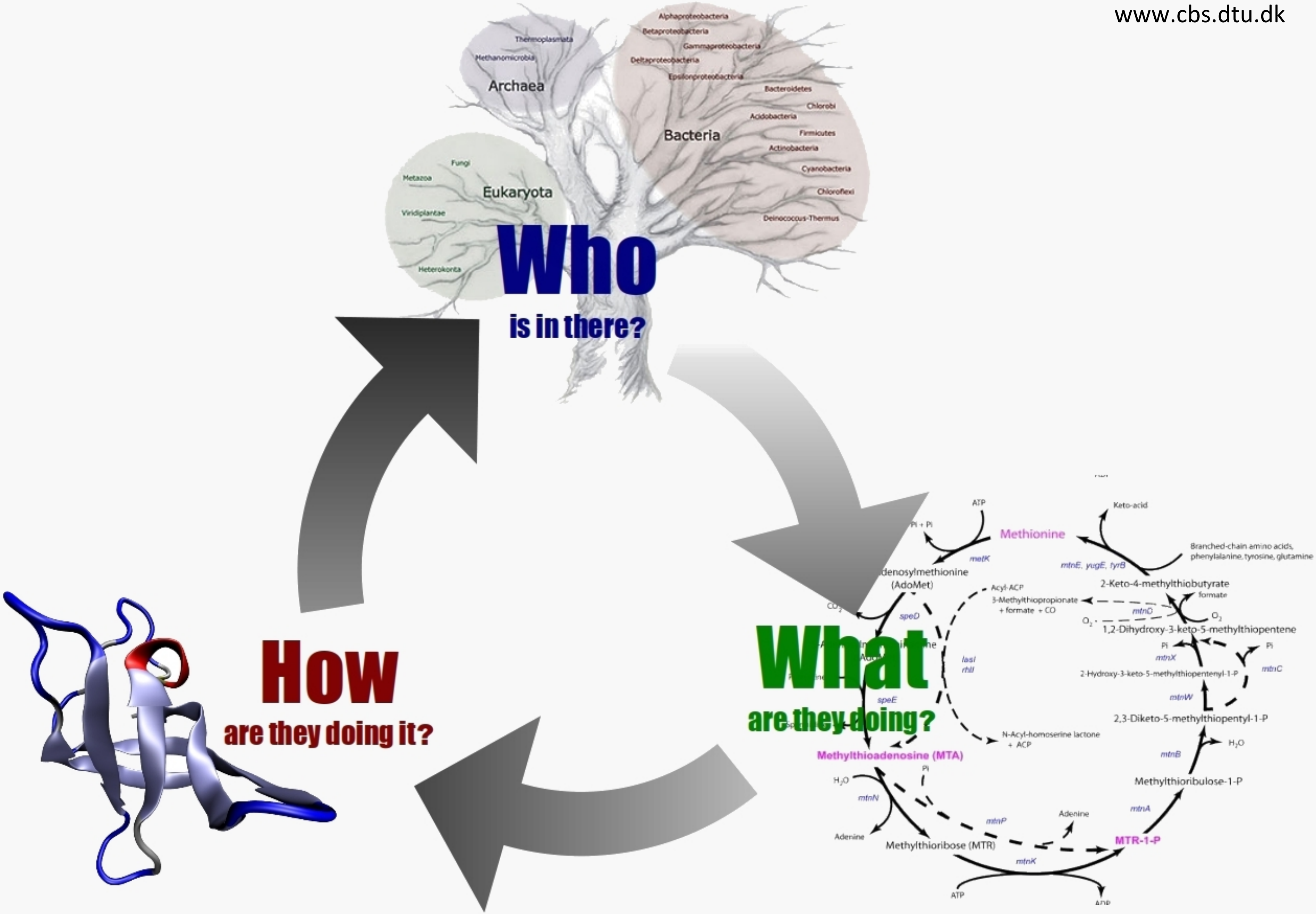


Metagenomic prediction from 16S - PICRUSt



http://picrust.github.io/picrust/


a
PC2 (19.5%)
PC1 (34.3%)
0.04
0.02
0
−0.02
−0.04
−0.04
−0.02
0
0.02
0.04
0.06
Airways (PICRUSt)
Airways (WGS)
GI tract (PICRUSt)
GI tract (WGS)
Oral (PICRUSt)
Oral (WGS)
Skin (PICRUSt)
Skin (WGS)
Urogenital (PICRUSt)
Urogenital (WGS)
b
M00061:
Uronic acid
metabolism
Relative abundance
0.010
0.008
0.006
0.004
0.002
0
c
M00076:
Dermatan sulfate
degradation
Relative abundance
0.004
0.003
0.002
0.001
0
d
M00077:
Chondroitin sulfate
degradation
Relative abundance
0.005
0.004
0.003
0.002
0.001
0
P W P W P W P W P W
Airways GI tract Oral Skin Urogenital
e
M00078:
Heparan sulfate
degradation
Relative abundance
0.004
0.003
0.002
0.001
0
P W P W P W P W P W
Airways GI tract Oral Skin Urogenital
f
M00079:
Keratan sulfate
degradation
Relative abundance
0.025
0.020
0.015
0.010
0.005
0
P W P W P W P W P W
Airways GI tract Oral Skin Urogenital

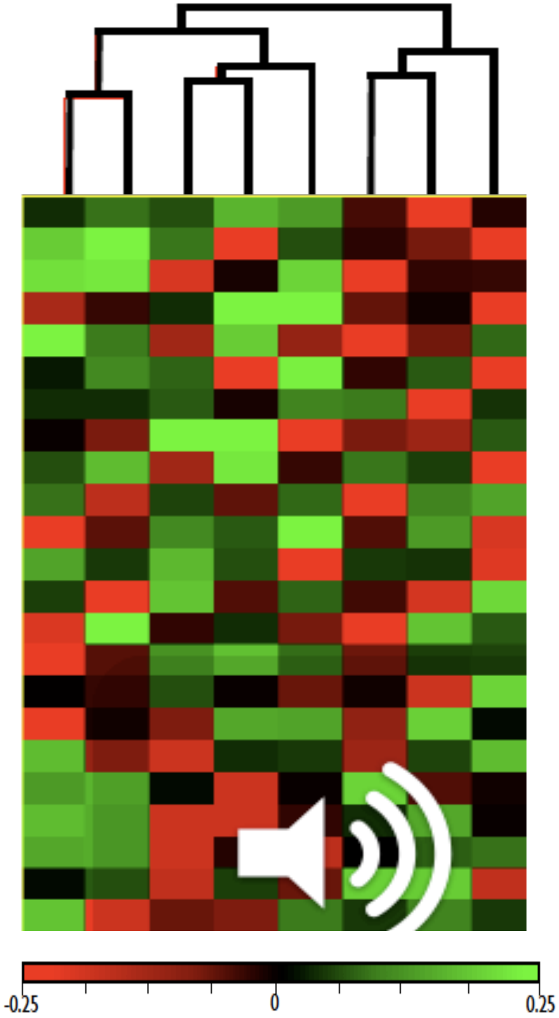


Who
is in there?

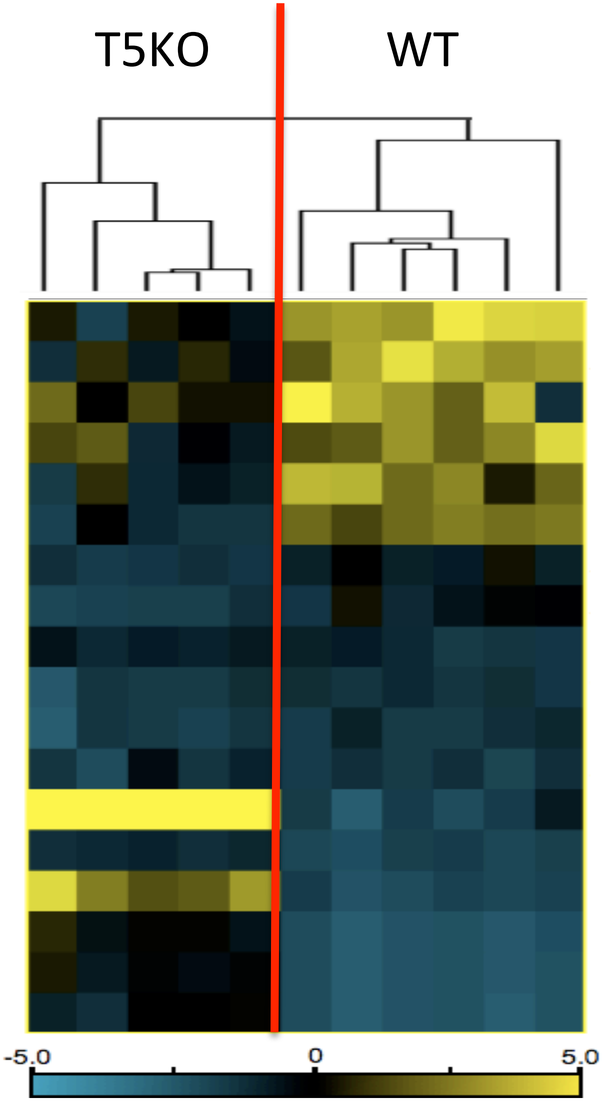
What
are they doing?

How
are they doing it?

The TLR5-/- Microbiome



Metagenome



Metatranscriptome

Thanks!

scikit-bio

- Generic biological data analysis in python
  - Replacement for PyCogent
- Thoroughly tested and documented
  - http://scikit-bio.org

IP[y]: IPython
Interactive Computing

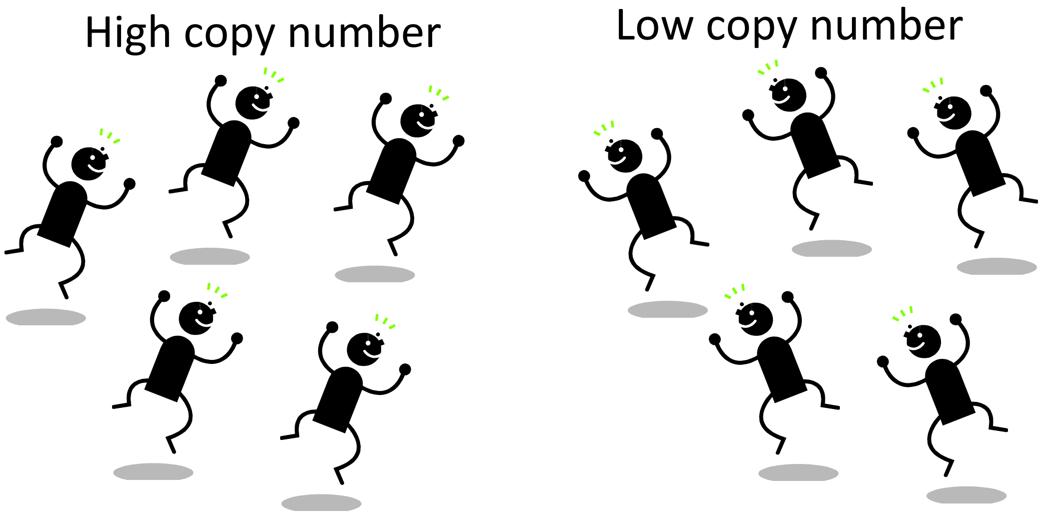
NumPy

qiita (pronounced cheetah)

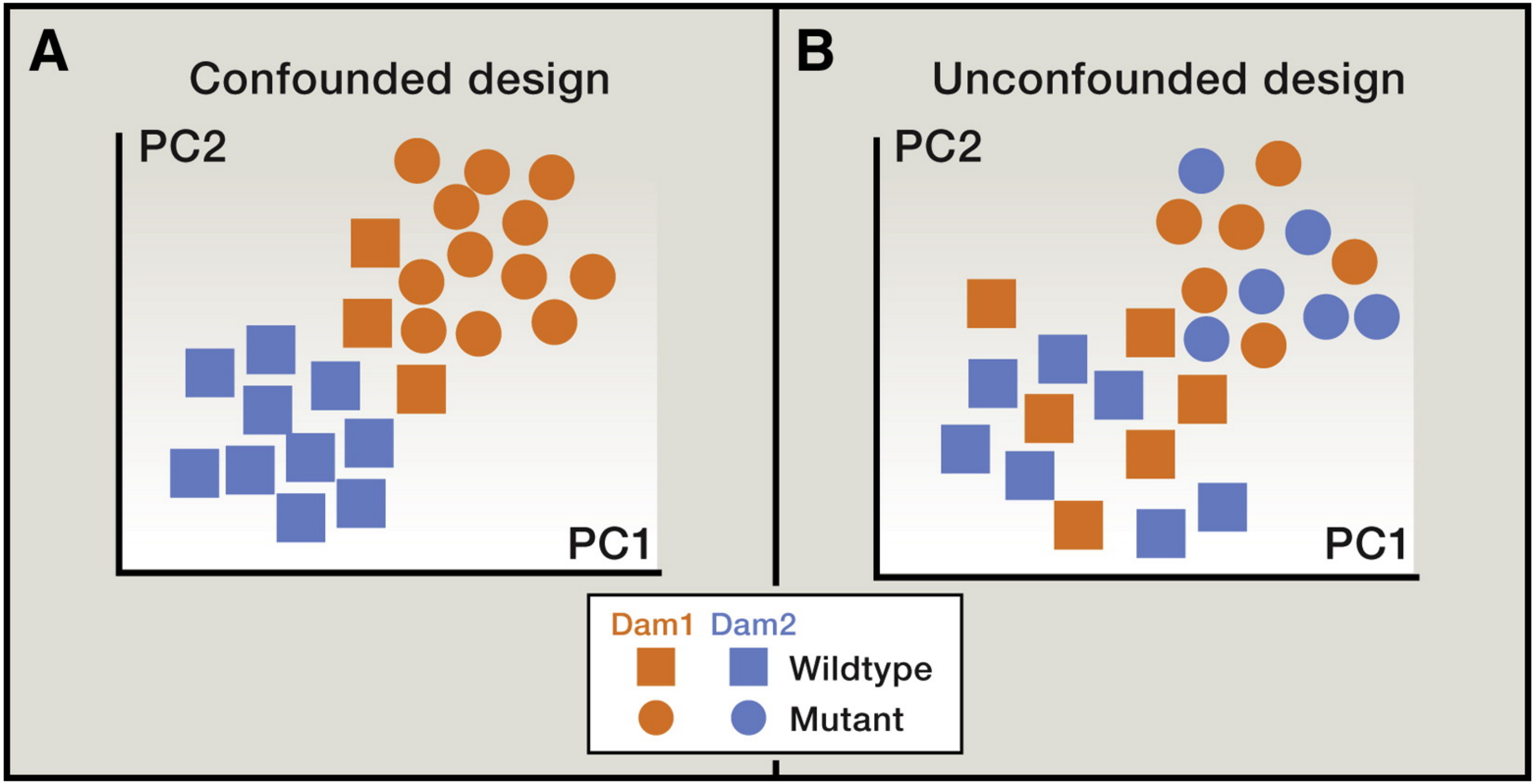
- Web interface to create and share meta-analyses.
- MIxS compliant metadata (GSC).
- Submit datasets to EBI

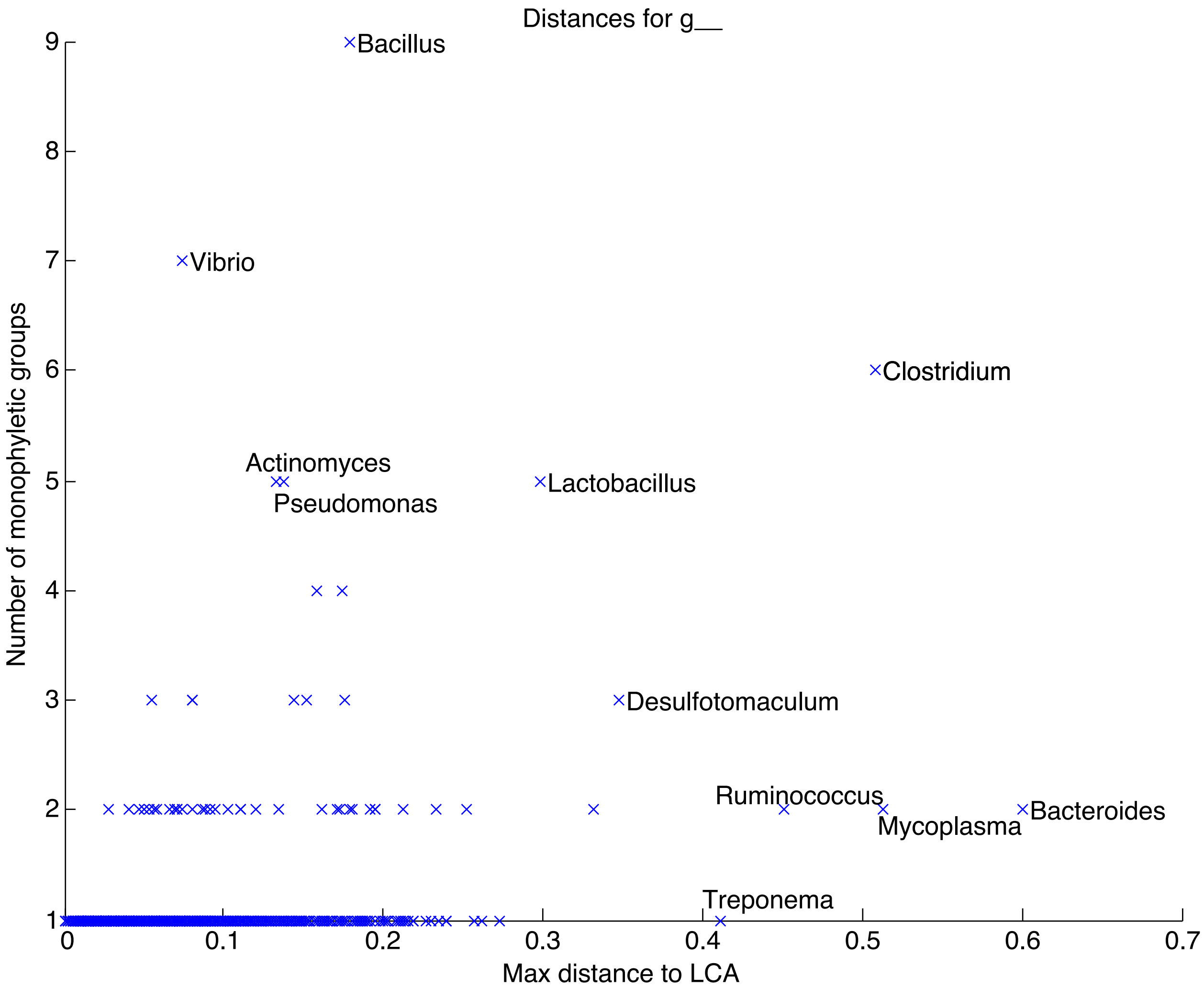
Do amylase levels affect the microbiome?

Human Subjects



- Collect metadata
- Balance genders
- Determine BMI range
- Consider controlling diet
- Weight gain or loss could become a variable
- Determine types of samples to be collected
- Collect samples at multiple timepoints
- Determine sample processing and storage conditions


A
Confounded design
PC2
PC1
B
Unconfounded design
PC2
PC1
Dam1
Dam2
Wildtype
Mutant


Distances for g__
Bacillus
Vibrio
Clostridium
Actinomyces
Pseudomonas
Lactobacillus
Desulfotomaculum
Ruminococcus
Mycoplasma
Bacteroides
Treponema
Number of monophyletic groups
Max distance to LCA
1
2
3
4
5
6
7
8
9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

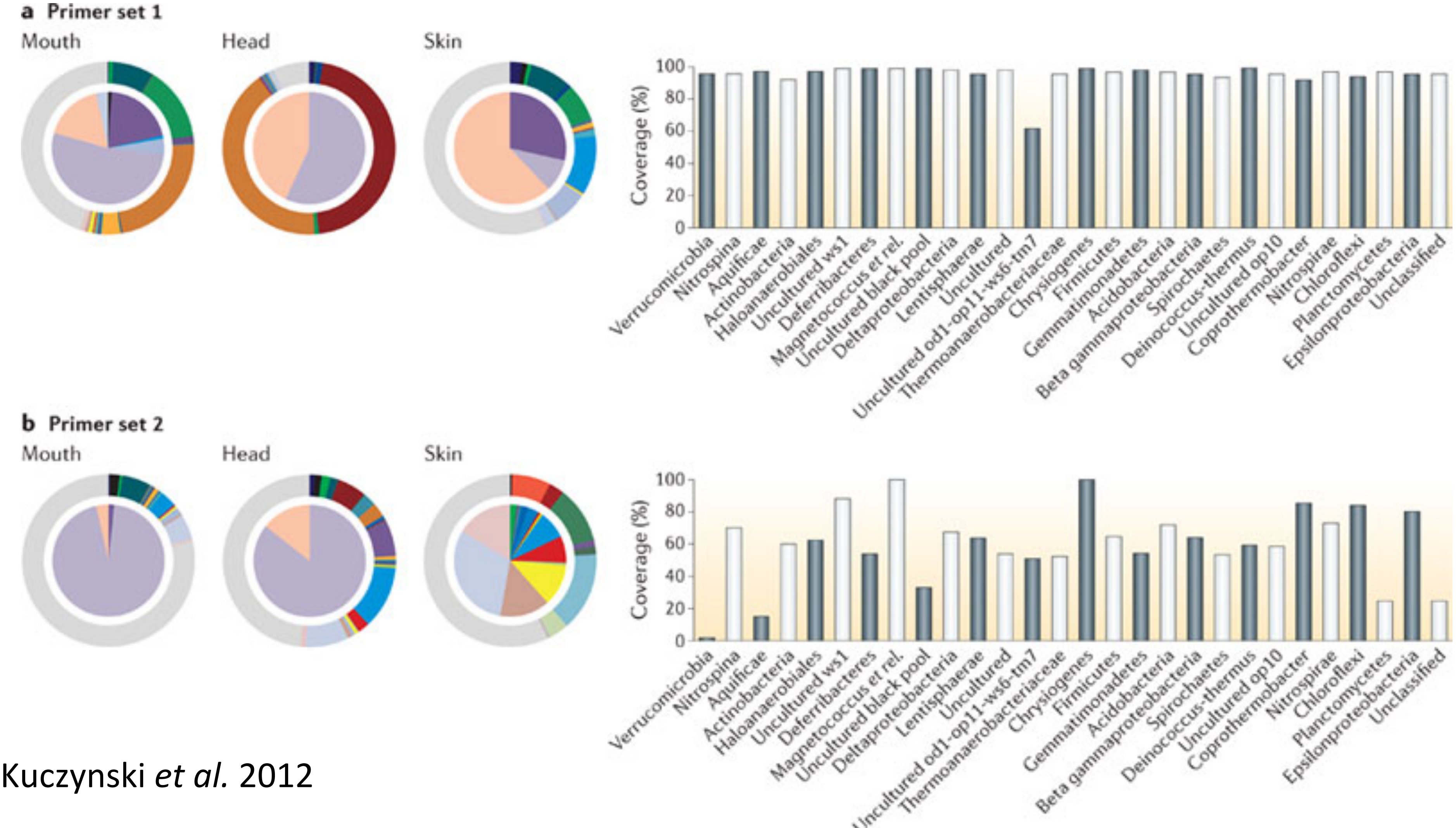
Fecal sample processing and storage conditions

- Fresh
- Frozen and ground under liquid nitrogen
- Freeze dried and milled

Be consistent.

Bahl et al. (Feb. 2012) Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. FEMS Microbiology Letters. Epub ahead of print.

Effects of primer choice in targeted amplicon sequencing



Kuczynski *et al.* 2012

Choice of universal PCR primer

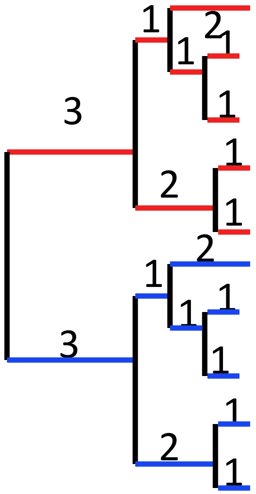
- F27–R338 is highly specific for bacteria (as opposed to archaea and eukaryotes) but lacks sensitivity for taxa such as *Bifidobacterium*, which is an important member of the gut microbiota.

- F515–R806 amplifies a broad range of bacterial and archaeal phyla (good for soil) but poorly amplifies *Propionibacterium*(bad for skin)

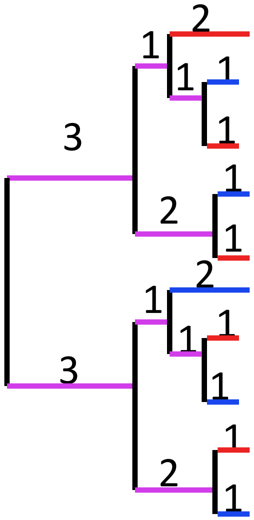
UniFrac

The UniFrac metric measures the difference between two environments in terms of the branch length that is unique to one environment or the other.

D = 1



D = 0.5



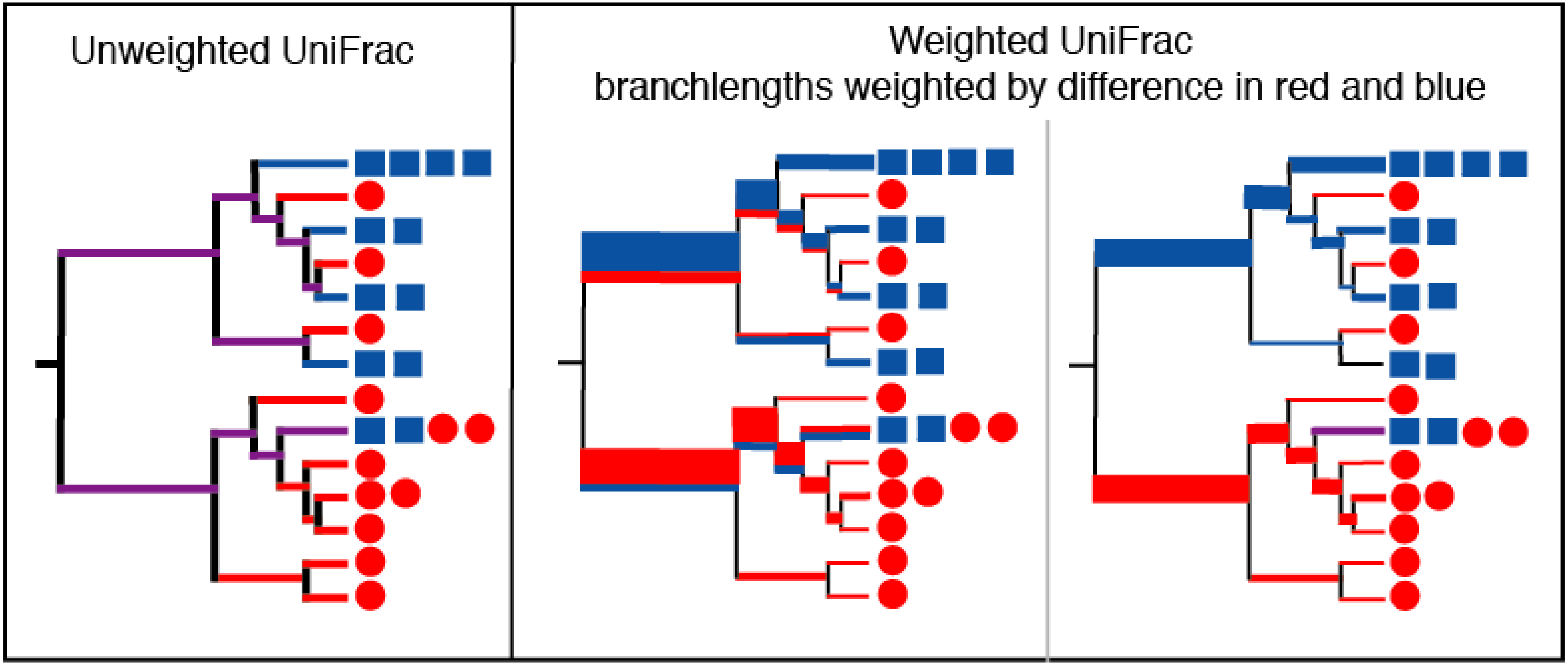
In the tree above all of the branch length is unique to one environment or the other.

In the tree above there is about as much branch length unique to each environment as shared between environments.

Lozupone and Knight, 2005

Weighted UniFrac

Qualitative **Quantitative**



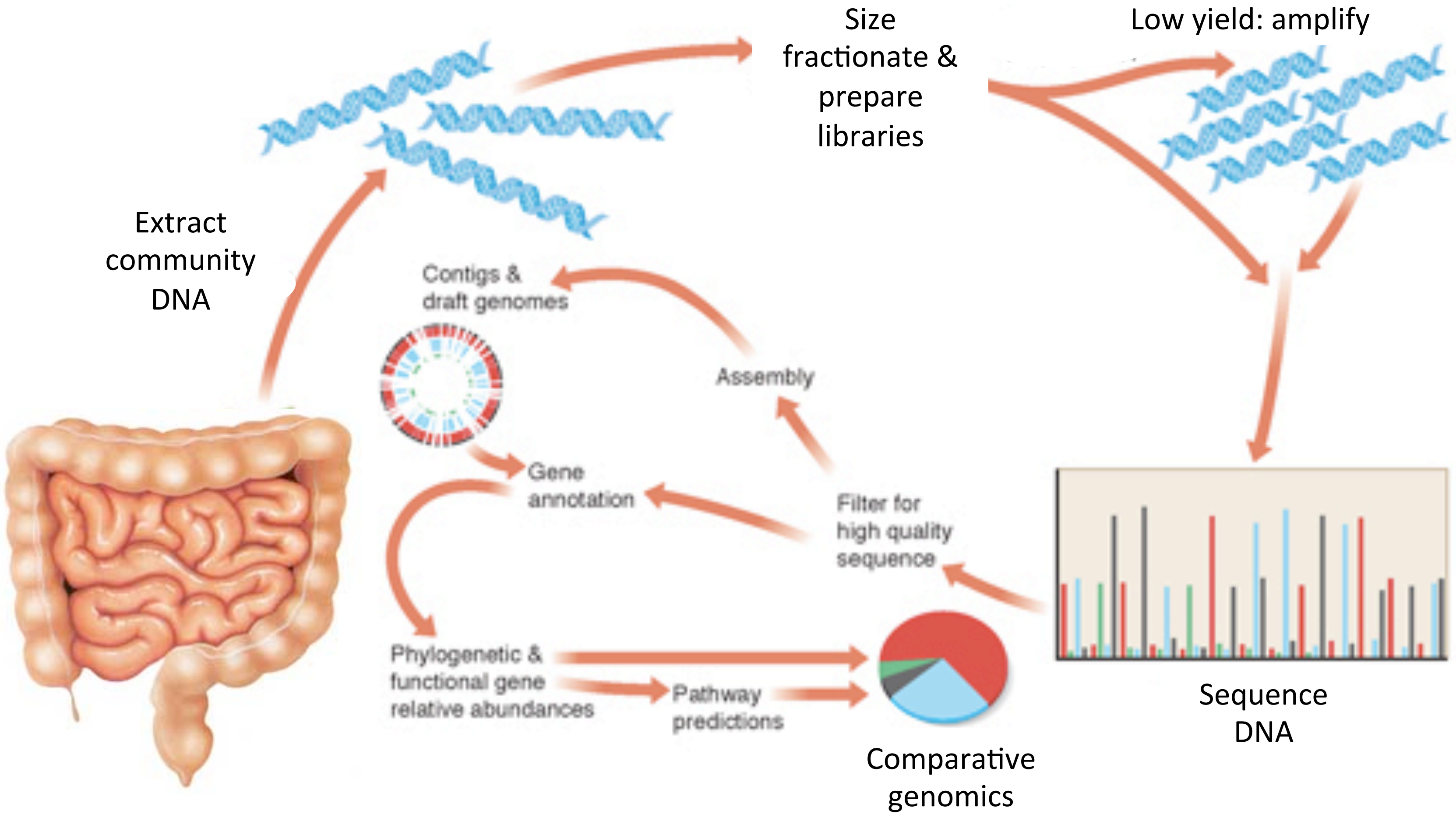
Lozupone et al., 2007. Appl Environ Microbiol 73:1576

Metagenomics

- Provides information about genes with the potential for being expressed, but cannot determine which ones are functional

- Also, because we sequence total DNA, it is not possible to distinguish genes from actively growing cells from those in dormant or dead cells

The Metagenomic Pipeline



Size
fractionate &
prepare
libraries

Low yield: amplify

Extract
community
DNA

Sequence
DNA

Comparative
genomics

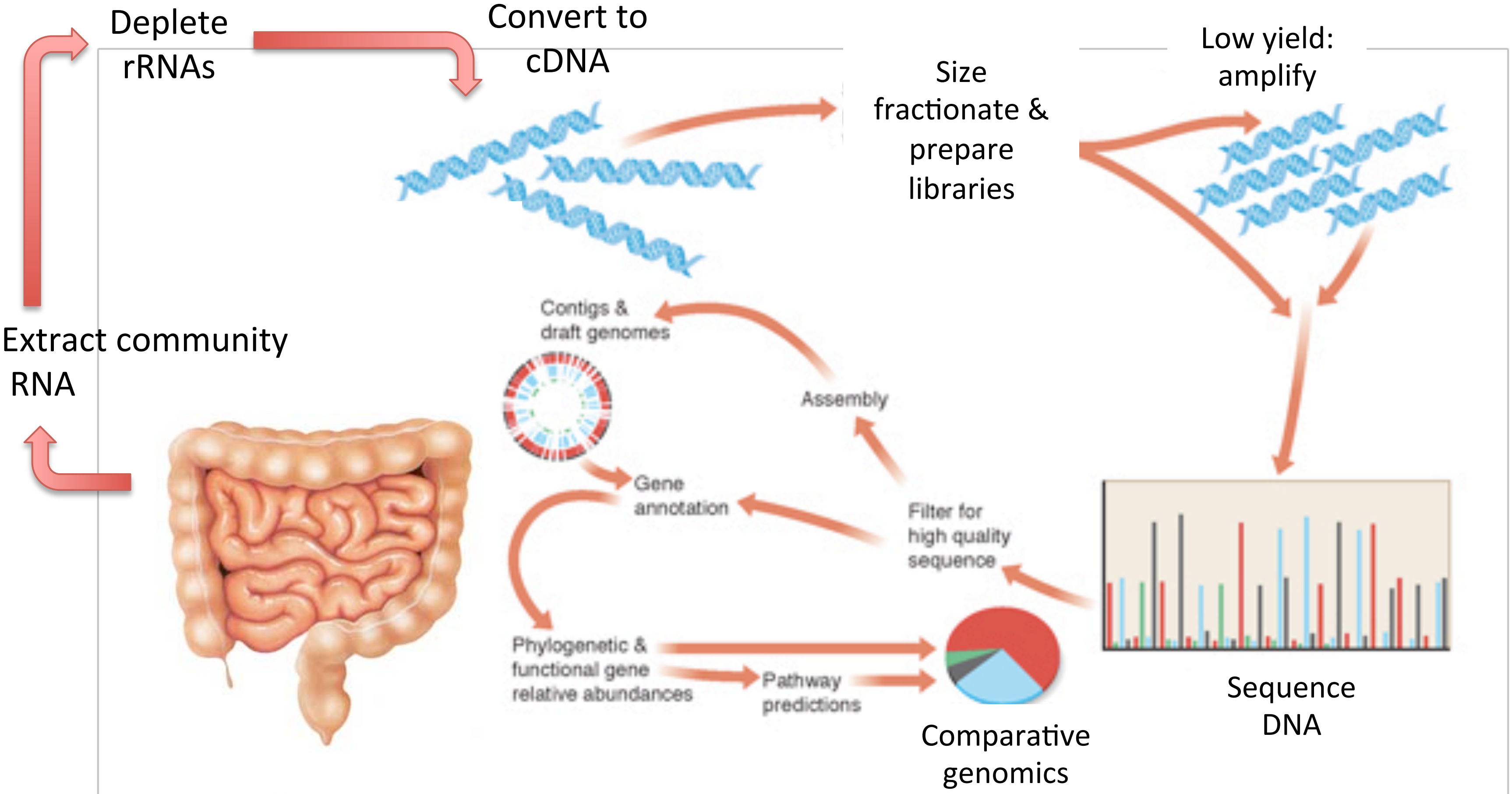
MICROBE
Issue: July 2011, Dr. Jansson
Penumbra Design, Inc. 06/09/11
Fig.#01

Metatranscriptomics

- The study of RNA molecules present in a cell at any given time

- Neither primers nor probes are needed, so there is no need to anticipate important genes beforehand and transcripts from microbial assemblages are sequenced with little bias.

The Metatranscriptomic Pipeline



Extract community RNA

Deplete rRNAs

Convert to cDNA

Size fractionate & prepare libraries

Low yield: amplify

Sequence DNA

Comparative genomics

MICROBE
Issue: July 2011, Dr. Jansson
Penumbra Design, Inc. 06/09/11
Fig.#01