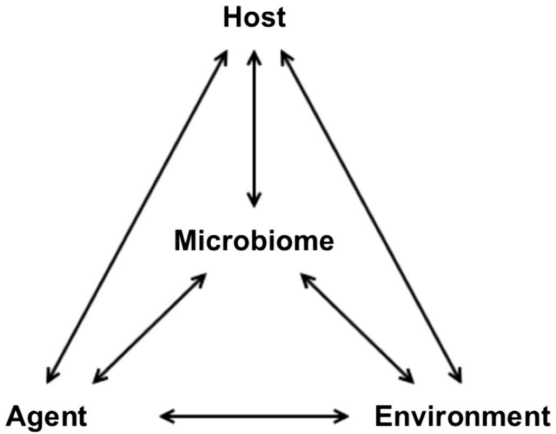
# Analyzing Microbiome Data

## BIOC 6102 Special Topics

## September 6, 2022

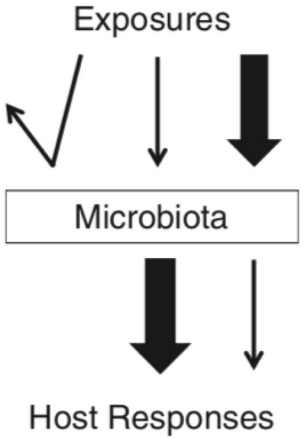
Michael S. Robeson II, Ph.D.
University of Arkansas for Medical Sciences
Department of Biomedical Informatics

# What is Microbiome (Data) Science?




*The study of all facets of the microbiome, e.g. microbial composition, diversity, and function as they interact with the biotic and abiotic features of the environment in which they live, is often referred to as the field of* ***microbiome science****, or microbiomics.*

*The concept of* ***microbiome data science*** *refers to the application of best practices from open data science to microbiome bioinformatics.*



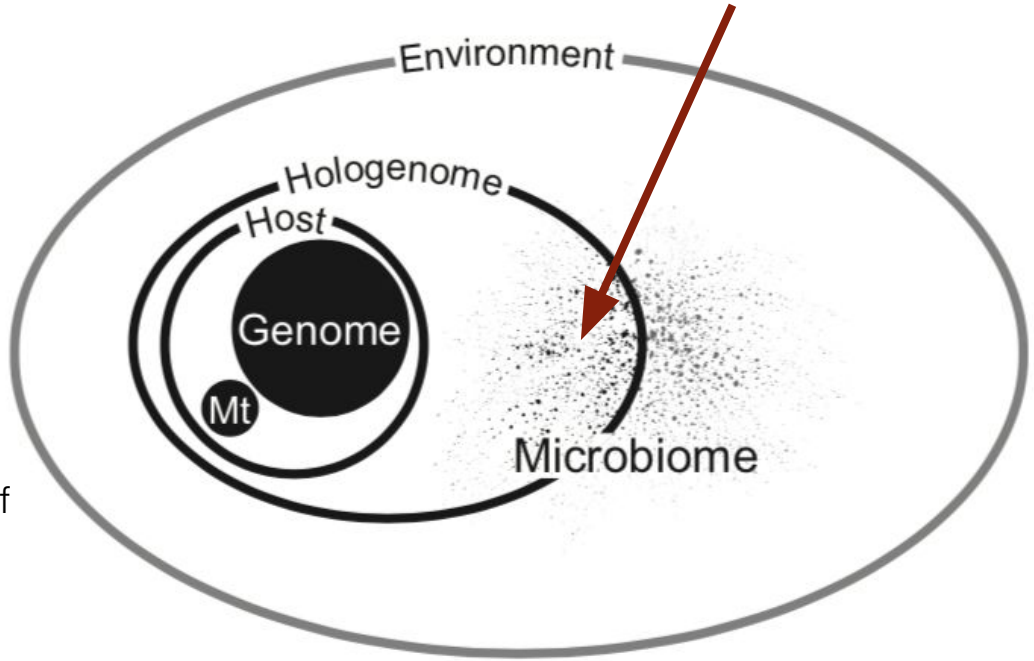


- Bokulich *et al*. 2020. “Measuring the Microbiome: Best Practices for Developing and Benchmarking Microbiomics Methods.” Computational and Structural Biotechnology Journal 18 (January): 4048–62. http://dx.doi.org/10.1016/j.csbj.2020.11.049
- Shetty & Lahti. 2019. “Microbiome Data Science.” Journal of Biosciences 44 (5). 10.1007/s12038-019-9930-2.
- Foxman, B., and Martin, E.T. 2015. Use of the Microbiome in the Practice of Epidemiology: A Primer on -Omic Technologies. American Journal of Epidemiology 182(1): 1–8. doi:10.1093/aje/kwv102.
- Hanson, B.M., and Weinstock, G.M. 2016. The importance of the microbiome in epidemiologic research. Annals of Epidemiology 26(5): 301–305. doi:10.1016/j.annepidem.2016.03.008.

# Holobiont

“The 2nd Genome”
“The Extended Genome”

**Holobiont**:
the host and all of its associated microorganisms.

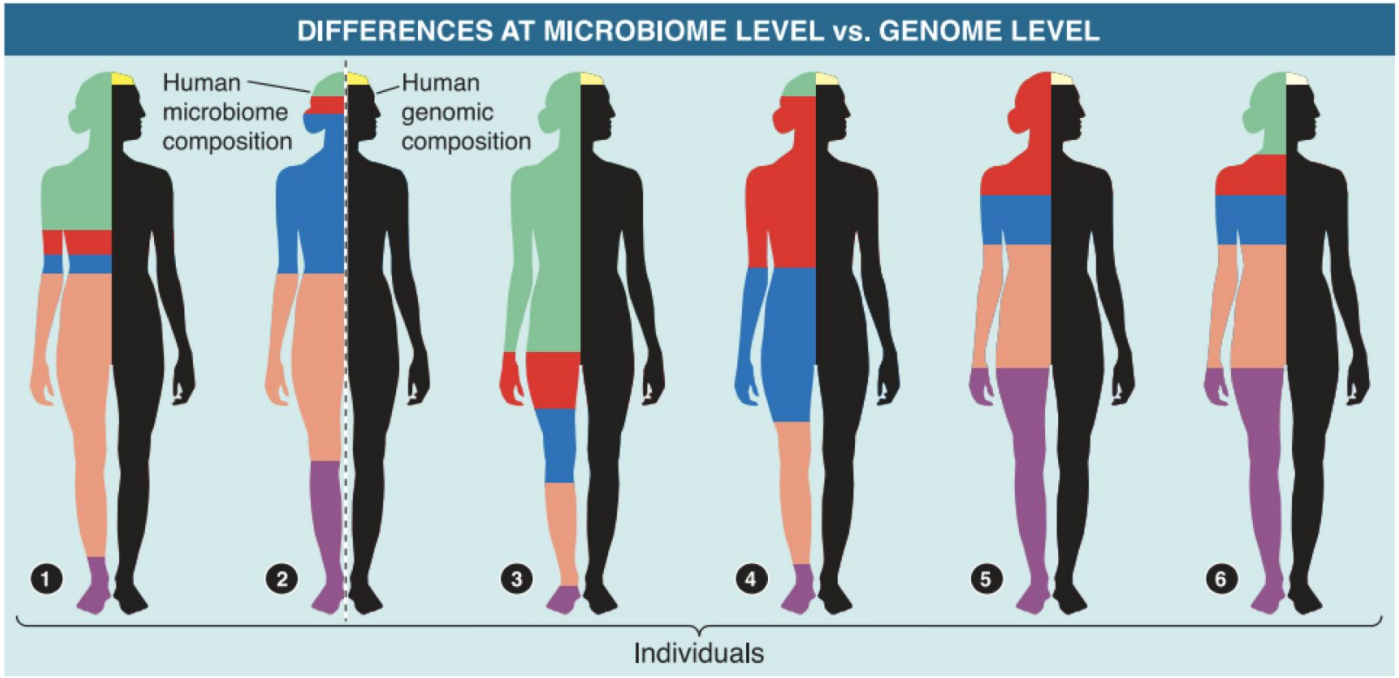



**Hologenome**:
the sum of the genetic information of the host and its microbiota.

Brucker & Bordenstein (2013) “The capacious hologenome.”, Zoology 116(5): 260–261.
Zilber-Rosenberg & Rosenberg (2008) “Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution.”, FEMS Microbiol. Rev. 32(5): 723–735.
O’Hara, Ann M., and Fergus Shanahan. 2006. “The Gut Flora as a Forgotten Organ.” EMBO Reports 7 (7): 688–93.
Relman, D. A., and S. Falkow. 2001. “The Meaning and Impact of the Human Genome Sequence for Microbiology.” Trends in Microbiology 9 (5): 206–8.

# Genetic diversity.

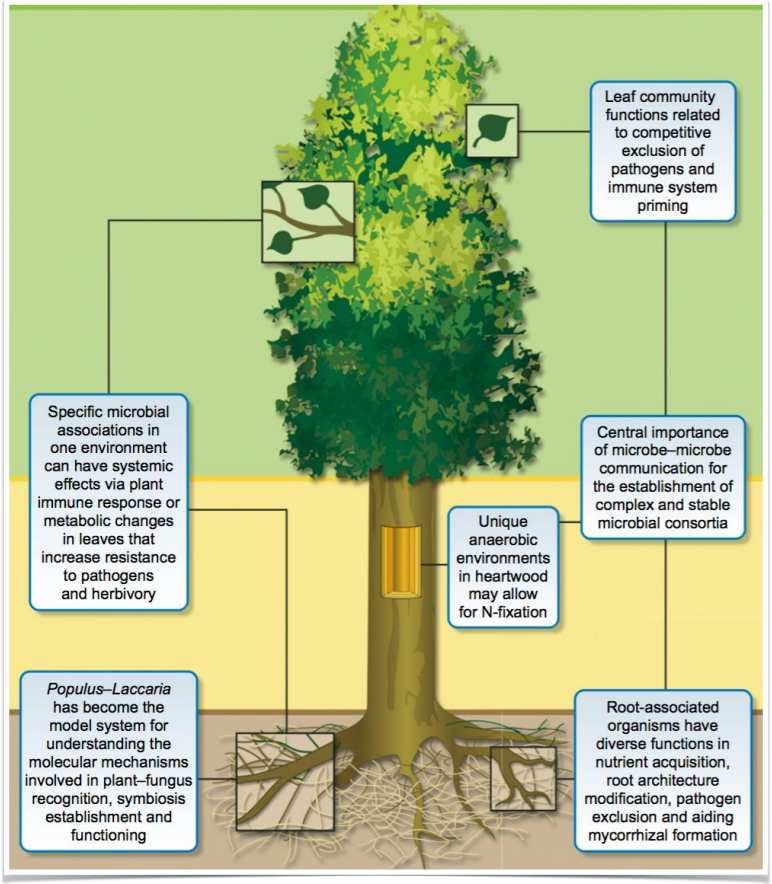
FIGURE 1




**Humans are far more different from each other in their microbiome composition than in their genomic composition. The colors in the left side of each individual represent bacterial phyla, while the colors on the right side indicate host genomic similarity. For the most part, we contain similar phyla living in and on our bodies, but their relative abundances can be drastically different. On the other hand, our genomic composition is nearly identical, with only a small fraction (around 0.1%) differing across individuals.**

Califf et al. (2014) "The Human Microbiome: Getting Personal." Microbe 9: 410–15.

# The Microbiome



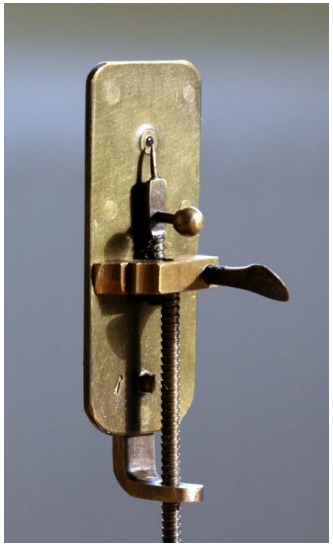

Hacquard, S., & Schadt, C. W. (2015). The New Phytologist, 205(4), 1424–1430.
Zaneveld et al. (2008) Current Opinion in Chemical Biology 12(1): 109–114.

# The Father of Microbiology

*Antonie van Leeuwenhoek*

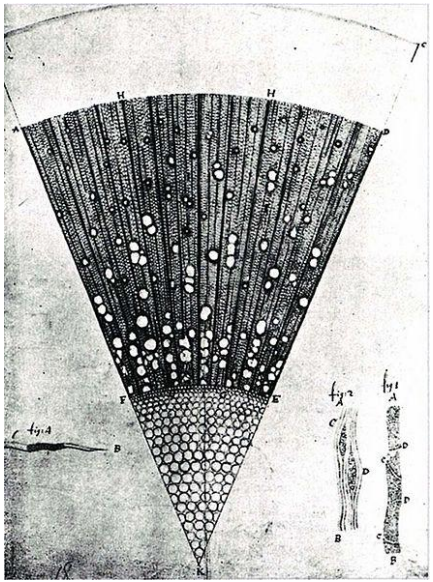
1632-1723

- Helped establish Microscopy / Microbiology as a scientific discipline.
- Used single-lens microscopes



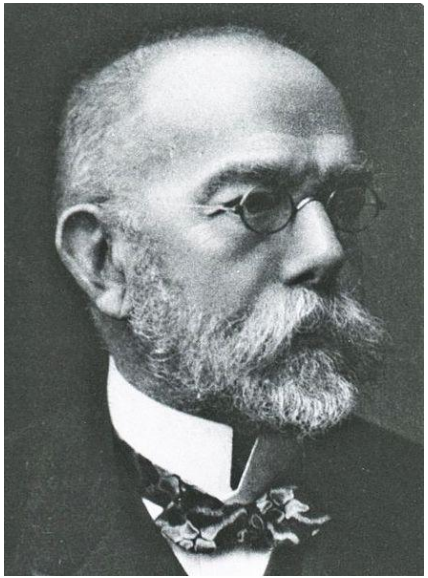
Discoveries:

- Infusoria
- Bacteria (human mouth)
- Vacuole
- Spermatozoa
- Muscle fibers



https://en.wikipedia.org/wiki/Antonie_van_Leeuwenhoek

**Robert Koch**
(1843-1910)



Proposed the initial criteria by which to establish a causative relationship between a microbe and a disease. This had a profound impact on public health.

https://en.wikipedia.org/wiki/Koch%27s_postulates

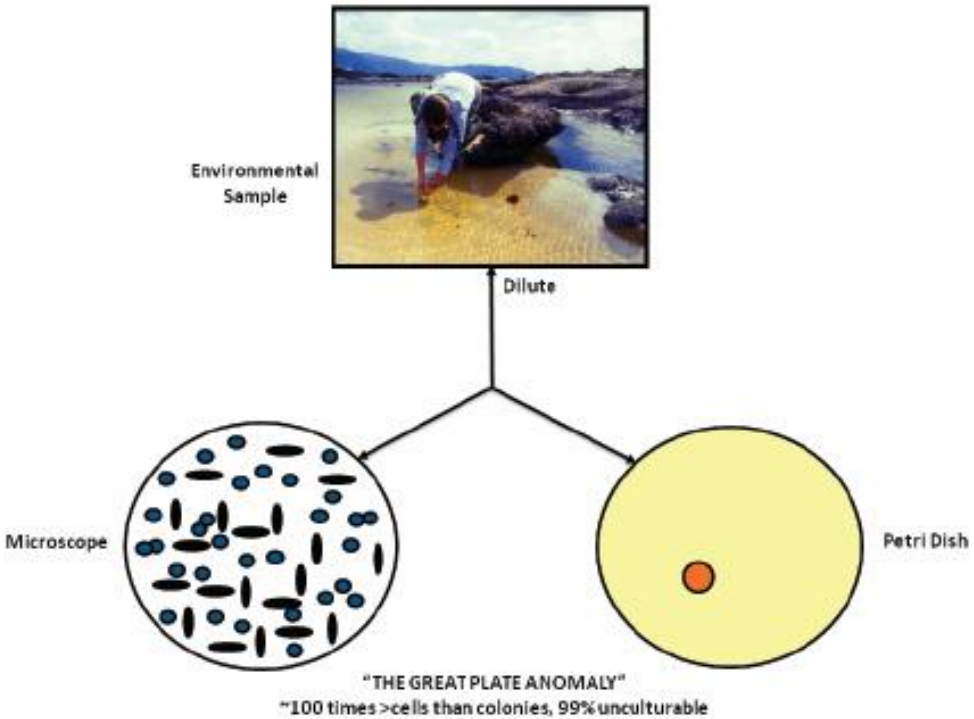
**Julius Richard Petri**
(1852-1921)

Invented petri dish while working as an assistant to Robert Koch

https://en.wikipedia.org/wiki/Julius_Richard_Petri
https://en.wikipedia.org/wiki/Growth_medium

## The Great Plate Count Anomaly.

1. We see millions of various microbes under the microscope, but can only cultivate a small fraction of these microbes.
2. That is, there are many viable but non-culturable (VBNC) microorganisms.
3. We simply do not know how to grow many of them.
4. It can be an onerous process to figure out how to cultivate many members of an entire microbial community.



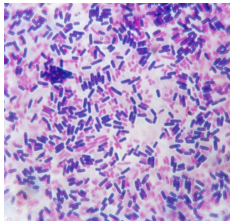

https://www.nap.edu/read/18261/chapter/2

# But how do we identify microbial taxa?

## Analytical Profile Index (API) Strips

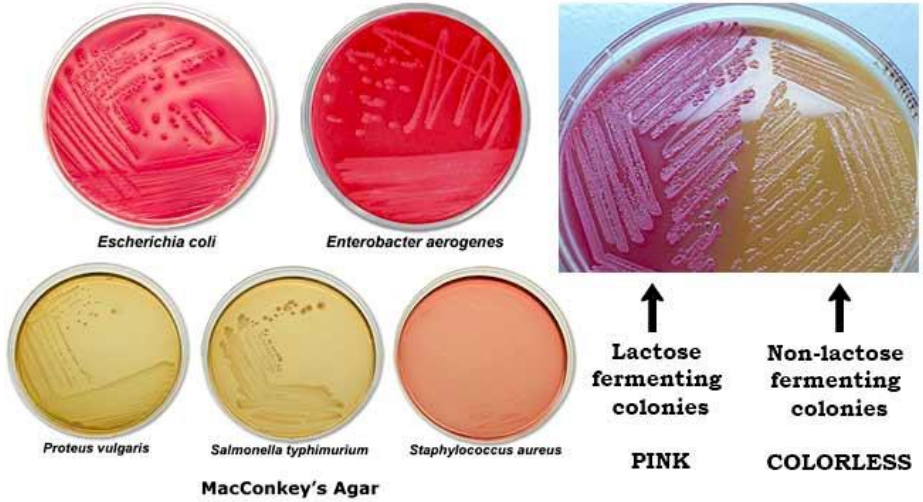

| culture no. | ONPG | ADH | LDC | ODC | CIT | H2S | URE | TDA | IND | VP | GEL | GLU | MAN | INO | SOR | RHA | SAC | MEL | AMY | ARA | identification |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8101 | + | – | + | + | – | – | – | – | + | – | – | + | + | – | + | + | + | + | – | + | *Escherichia coli* |
| 5B | + | – | – | – | + | – | – | – | – | + | – | + | + | + | + | + | – | – | + | + | *Enterobacter agglomerans* |
| 8P44 | – | – | + | + | – | + | – | – | + | – | – | + | + | – | – | – | + | – | – | + | *Edwardsiella hoshinae* |

https://www.jlindquist.com/generalmicro/102bactid2.html
https://en.wikipedia.org/wiki/Analytical_profile_index
https://www.thoughtco.com/gram-stain-procedure-41476
83



## Differential & Selective Media

*MacConkey agar is both*




https://microbiologyinfo.com/macconkey-agar-composition-principle-uses-preparation-and-colony-morphology/

- Inhibits or reduces growth of gram positive bacteria
- Promotes gram negative bacterial growth.
  - Lactose fermenting colonies are pink
  - non-lactose fermenting ones are colorless / same color as the medium.

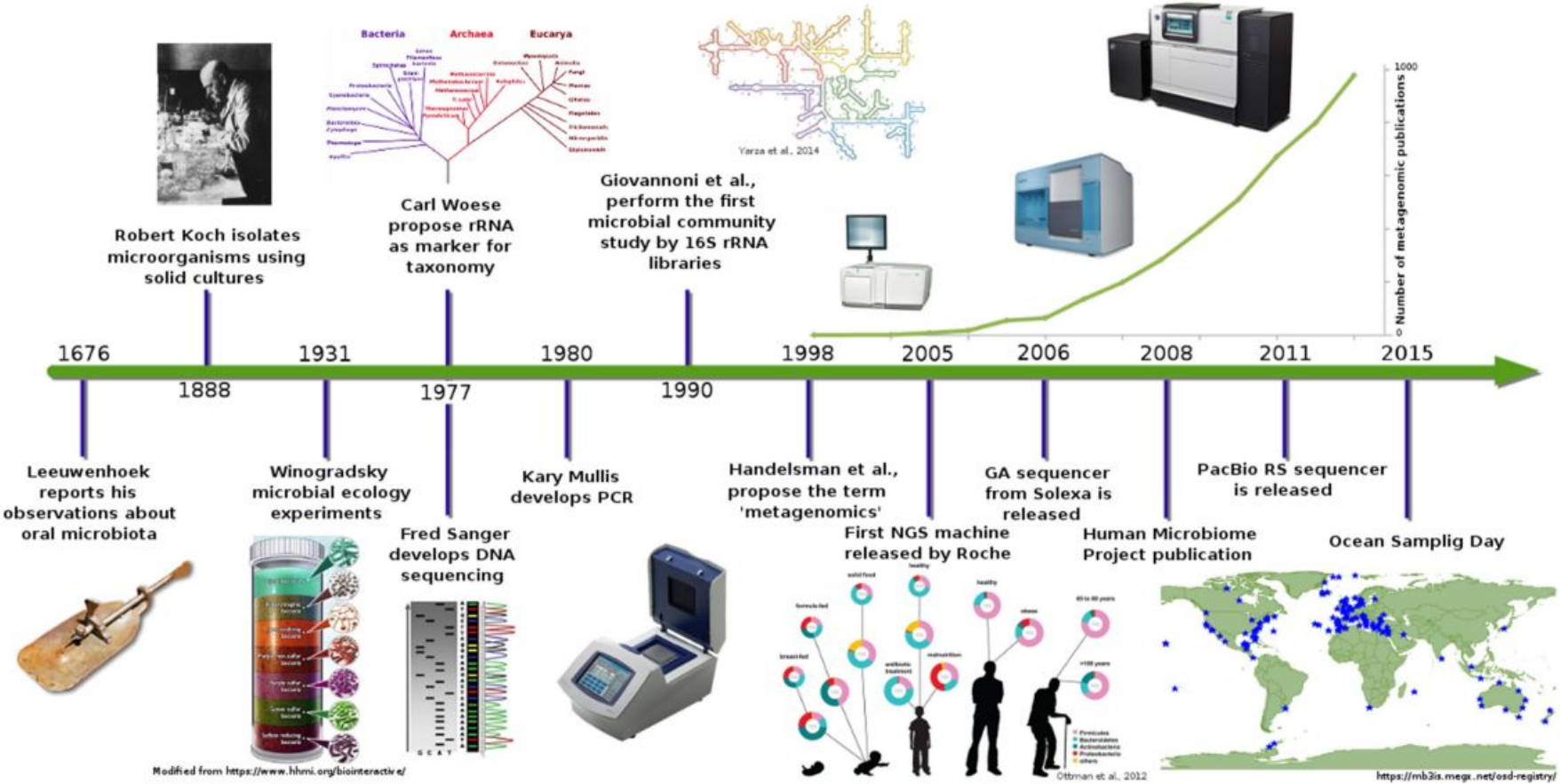
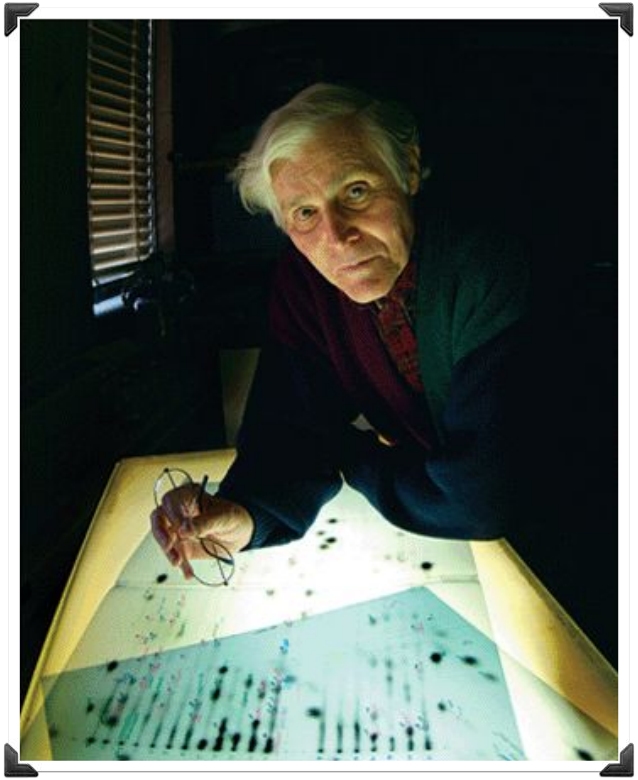


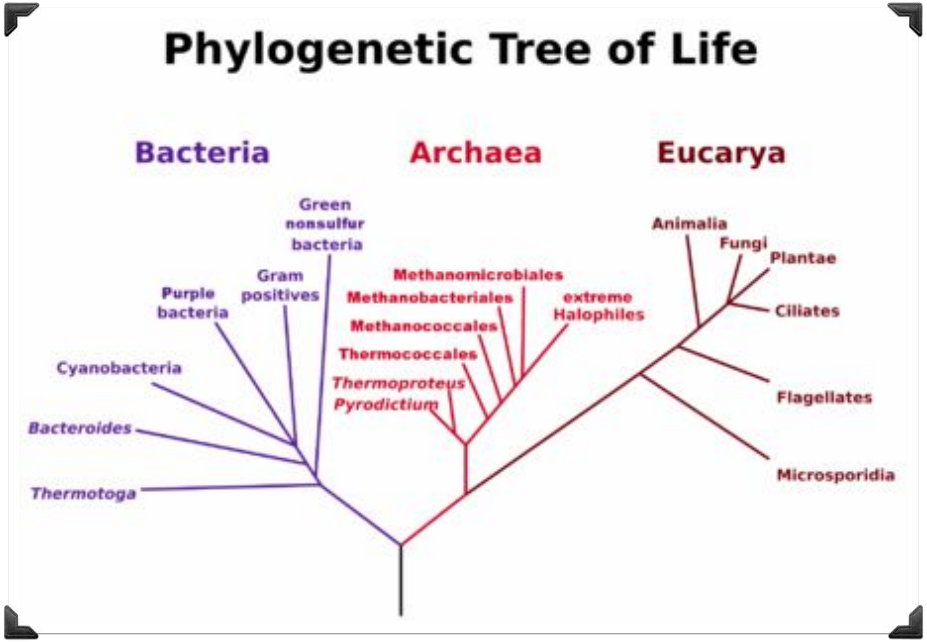

**FIGURE 1 | Metagenomics timeline and milestones.** Timeline showing advances in microbial communities studies from Leeuwenhoek to NGS (Ottman et al., 2012; Yarza et al., 2014).

# Carl Woese

July 15, 1928 – December 30, 2012



https://en.wikipedia.org/wiki/Carl_Woese




http://science.sciencemag.org/content/339/6120/661.full

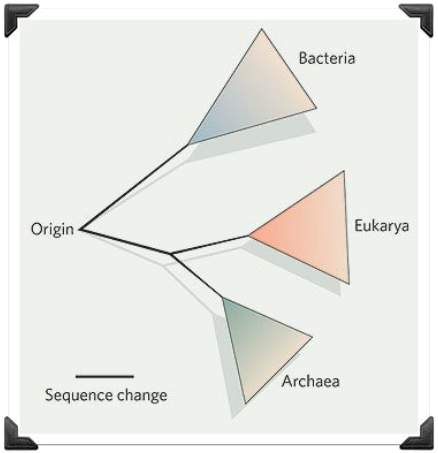

Woese C, Fox G (1977). "Phylogenetic structure of the bacteria domain: the primary kingdoms." Proc Natl Acad Sci USA. 74 (11): 5088–90

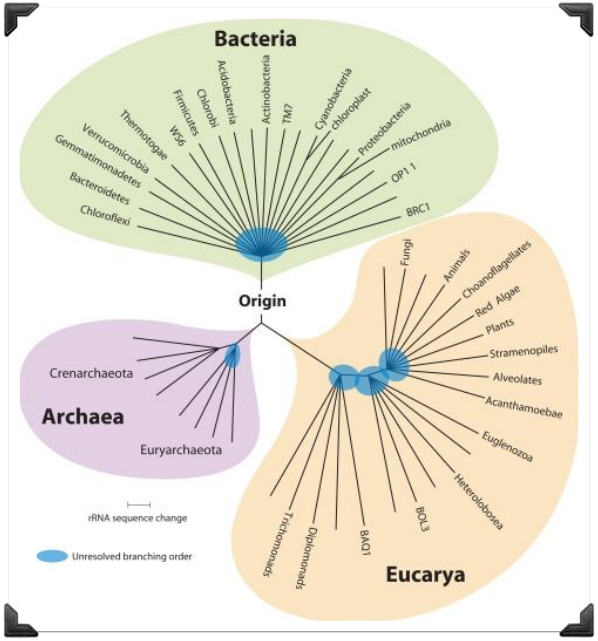
Woese CR, Kandler O, and Wheelis ML. (1990). “Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya.” Proc Natl Acad Sci USA 87 (12): 4576–79.

# Norm Pace

(1942 - )

*"The Man Who Blew The Door Off The Microbial World"*



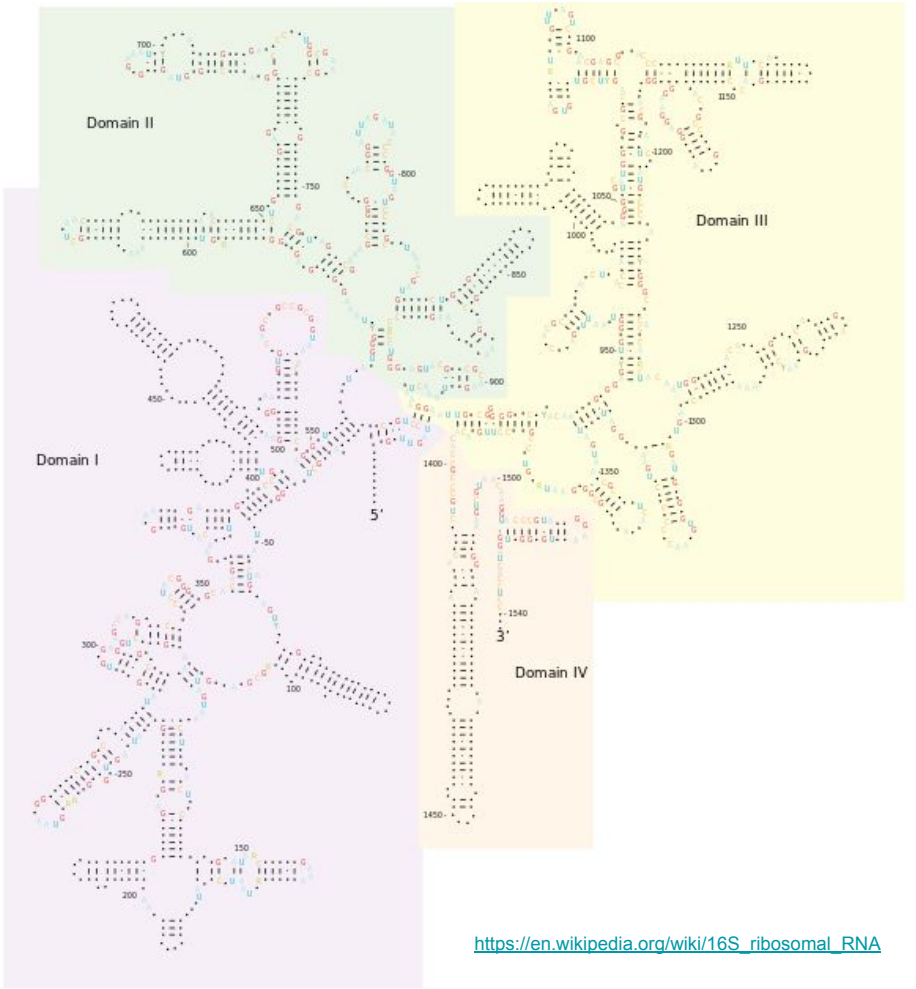






Pace (2006) Nature 441 (7091): 289.
Pace (1986) Adv. Microbial Ecol. 9:1-55.

# 16S rRNA gene / amplicon markers

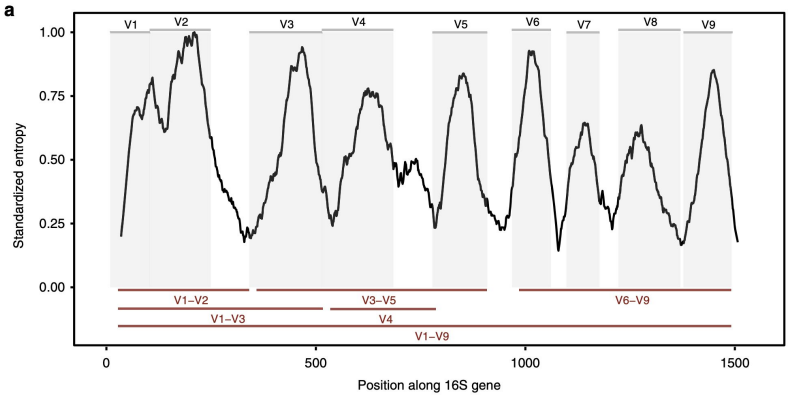
- low cost enables sequencing of hundreds to thousands of samples simultaneously
  - useful for tracking microbial compositions in large surveys
- low complexity of amplicon sequence data make it much more computationally tractable for a wide range of research applications
- short reads are less reliable for species-level identification and
  - though sequence variant information can be used to differentiate “phylotypes”

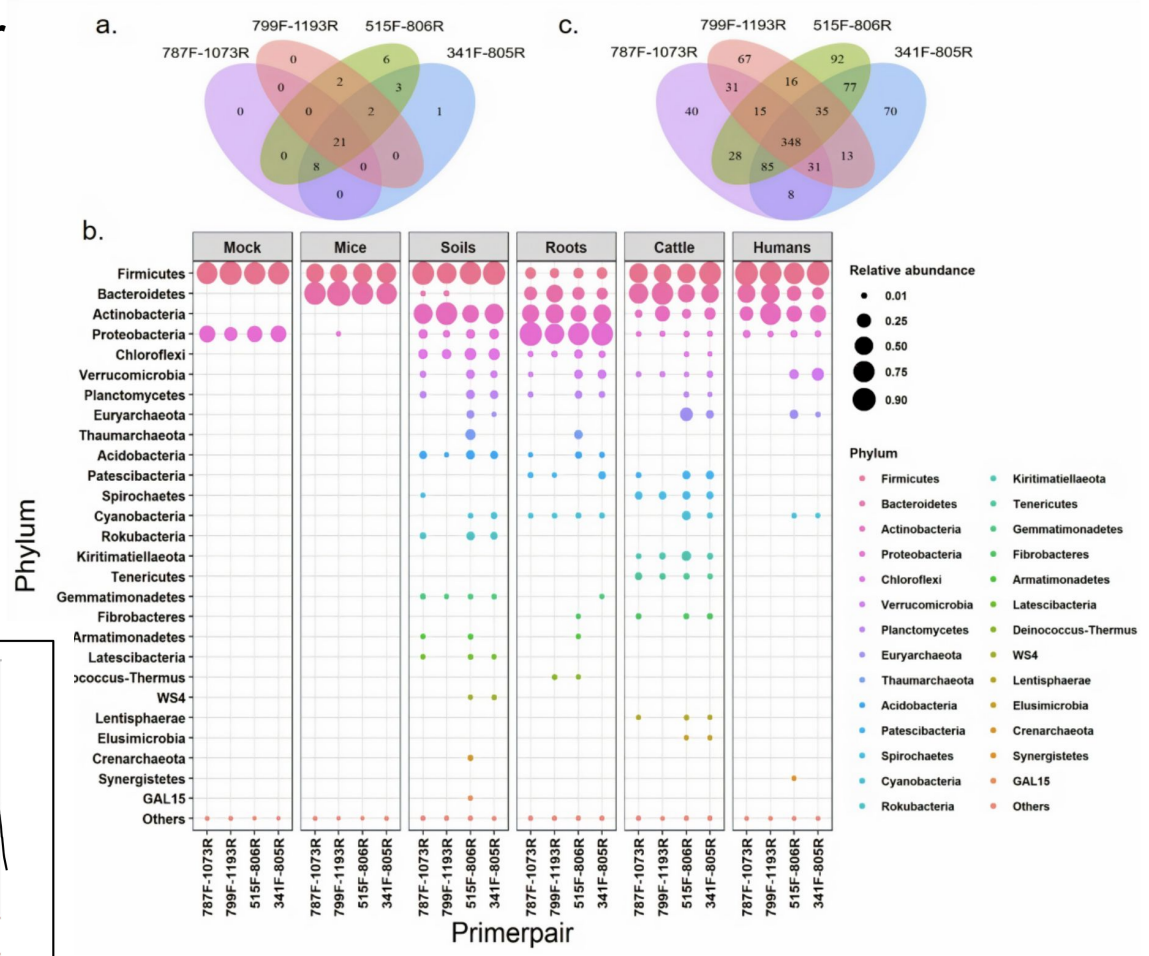



https://en.wikipedia.org/wiki/16S_ribosomal_RNA

# Different variable regions vary in their ability to discriminate among taxa.

V1-V3 : *skin*
V3-V5
V4 (515F-806R) : *EMP*
V3V4 (341-805R)
V4V5 : *EMP*
V5V6 (787F-1073R)
V5V6V7 (799F-1193R)




Uddin et al.. 2020. “Evaluation of Primer Pairs for Microbiome Profiling from Soils to Humans within the One Health Framework.” *Molecular Ecology Resources*, June. https://doi.org/10.1111/1755-0998.13215.

**How do you define:**

1) Microbiota?
2) Metagenome?
3) Microbiome?

Marchesi & Ravel (2015) “The vocabulary of microbiome research: a proposal.”, Microbiome 3:31. doi:10.1186/s40168-015-0094-5.

# Microbiota

"The assemblage of microorganisms present in a defined environment."

A "microbial census" is predominantly performed by 16S rRNA gene sequencing.

*Amplicon sequencing surveys are often, and incorrectly, referred as metagenomics surveys!*

Marchesi & Ravel (2015) "The vocabulary of microbiome research: a proposal.", Microbiome 3:31. doi:10.1186/s40168-015-0094-5.

# Metagenome

“The collection of genomes and genes from the members of a microbiota.”

Often obtained through shotgun sequencing of DNA extracted from a sample.

Marchesi & Ravel (2015) “The vocabulary of microbiome research: a proposal.”, Microbiome 3:31. doi:10.1186/s40168-015-0094-5.

# Microbiome

"This term refers to the entire habitat, including the micro-organisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions."

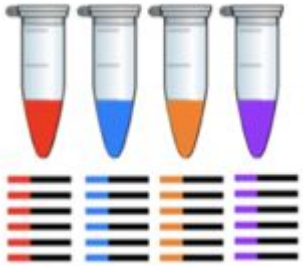
*Based on the term "biome". That is, the biotic and abiotic factors of given environments.*

Marchesi & Ravel (2015) "The vocabulary of microbiome research: a proposal.", Microbiome 3:31. doi:10.1186/s40168-015-0094-5.

**PSA:**
Sometimes the best way to study microbial communities may require neither DNA sequencing nor “-omcis”!

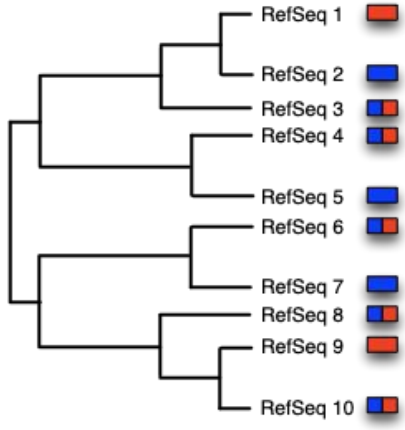
FISH, qPCR, etc., may be all you need!

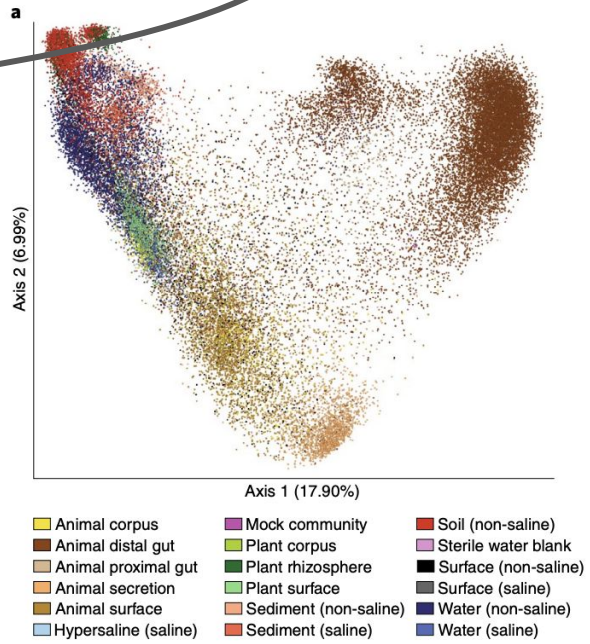
*Use the right tool for the question or hypothesis at hand!*

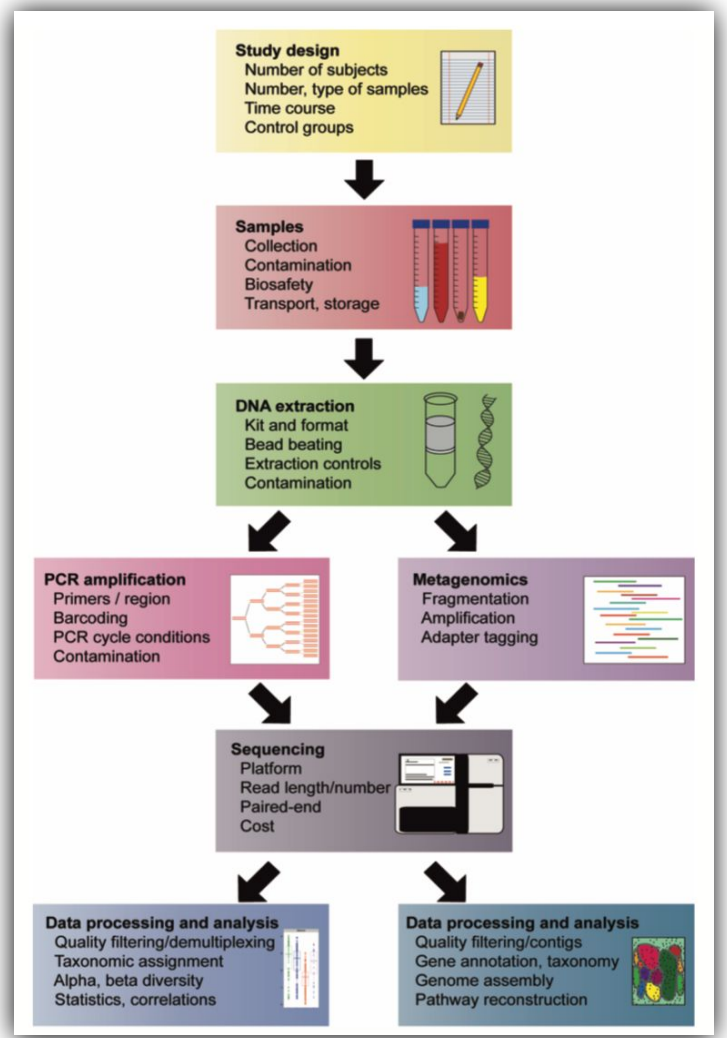


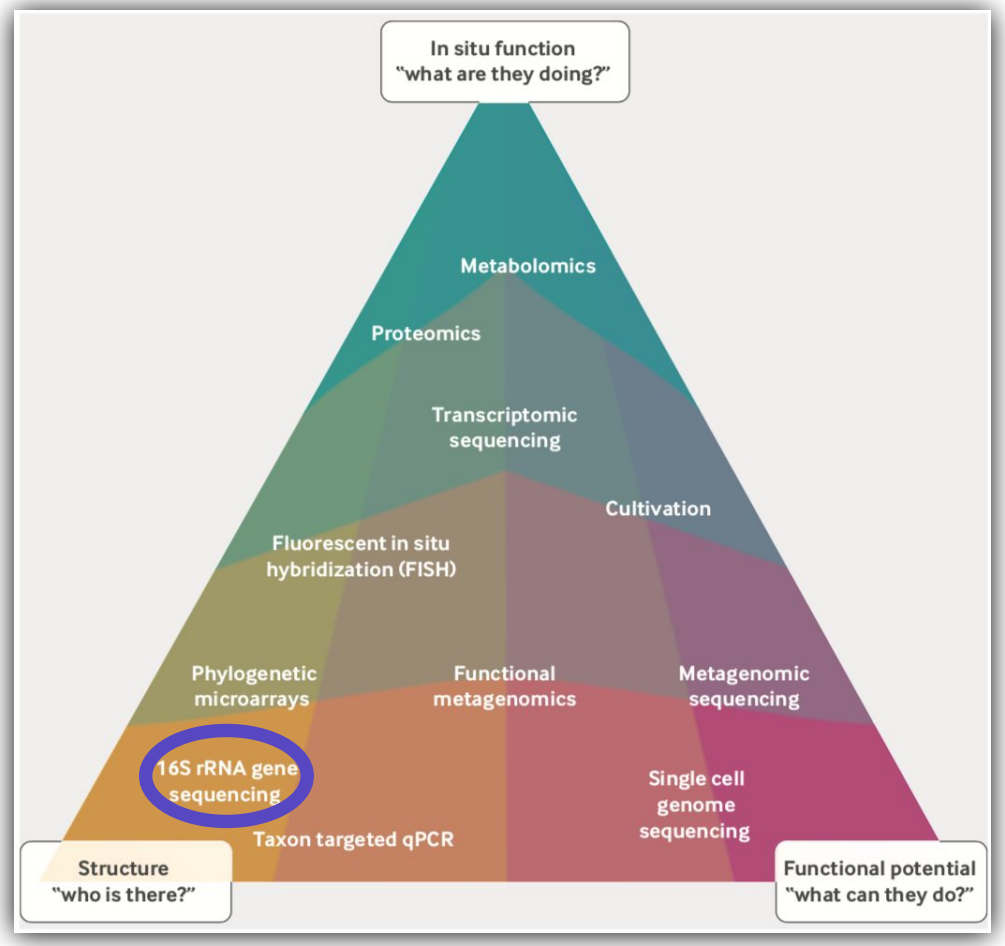
qiime2

>GCACCTGAGGACAGGCATGAGGAA…
>GCACCTGAGGACAGGGGAGGAGGA…
>TCACATGAACCTAGGCAGGACGAA…
>CTACCGGAGGACAGGCATGAGGAT…
>TCACATGAACCTAGGCAGGAGGAA…
>GCACCTGAGGACACGCAGGACGAC…
>CTACCGGAGGACAGGCAGGAGGAA…
>CTACCGGAGGACACACAGGAGGAA…
>GAACCTTCACATAGGCAGGAGGAT…
>TCACATGAACCTAGGGGCAAGGAA…
>GCACCTGAGGACAGGCAGGAGGAA…



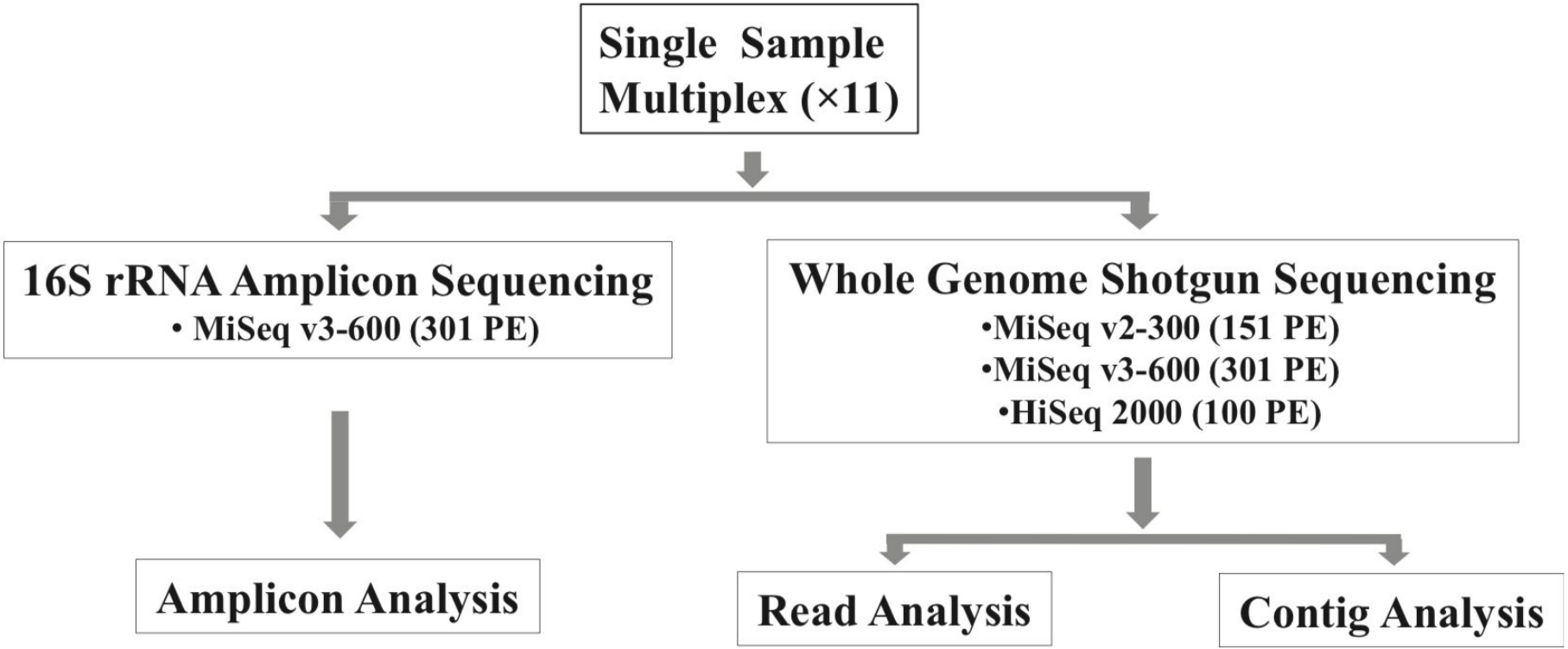




- Bik, E.M. 2016. The Hoops, Hopes, and Hypes of Human Microbiome Research. The Yale Journal of Biology and Medicine 89(3): 363–373. Yale Journal of Biology and Medicine.
- Young, V.B. 2017. The role of the microbiome in human health and disease: an introduction for clinicians. BMJ 356: j831. British Medical Journal Publishing Group. doi:10.1136/bmj.j831.

# Experimental Strategy




Ranjan *et al*. (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem. Biophys. Res. Commun. 469(4): 967–977


qiime2

# Quantitative Insights Into Microbial Ecology 2

**Popularity primarily due to**:

- Supportive and interactive community.
- Cutting edge functionality quickly added via plugins.

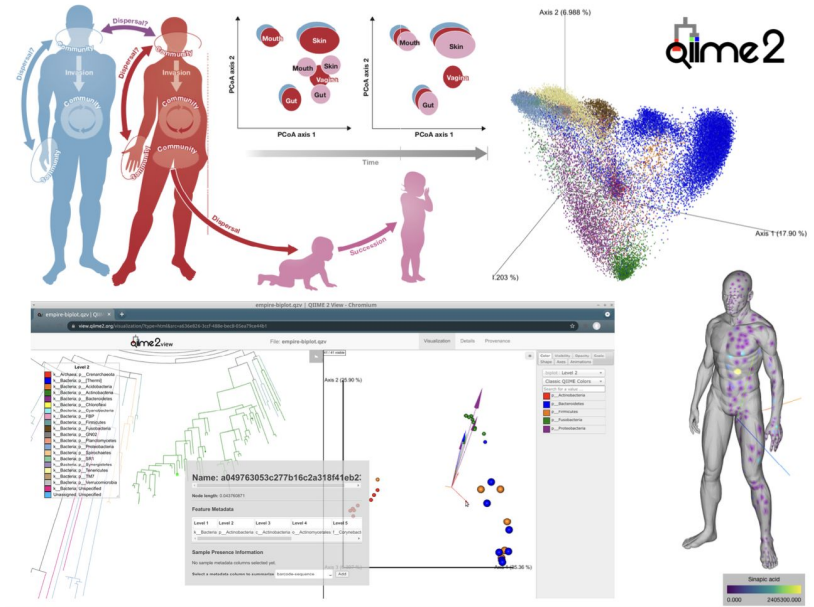
Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 — 7134 * — 2019
E Bolyen, JR Rideout, MR Dillon, NA Bokulich, CC Abnet, GA Al-Ghalith, ... Paperpile
Nature biotechnology 37 (8), 852-857

Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 — 6413
E Bolyen, JR Rideout, MR Dillon, NA Bokulich, CC Abnet, GA Al-Ghalith, ...
Nature biotechnology 37 (8), 852-857, 2019

QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science — 794 *
E Bolyen, JR Rideout, MR Dillon, NA Bokulich, C Abnet, GA Al-Ghalith, ...
PeerJ Preprints, 2018

**Currently uploading QIIME 2 workshop videos each week to YouTube: https://www.youtube.com/c/QIIME2**

## Foundations of the Human Microbiome
## BMIG 6202 - Spring 2022




### Course Description:

This graduate course will provide students with foundational knowledge and practical analytical skills required for analyzing microbiome data.

Students will explore the microbial inhabitants of the human body, with an emphasis on how microbial communities affect human health and disease progression.

Students will learn how to perform current leading-edge microbiome analysis directly from a QIIME 2 developer!

✓No course prerequisites.
✓Review foundational research.
✓Discuss host-microbiome interactions.
✓Learn how to plan a microbiome study.
✓Learn how to analyze microbiome data.

**Location:** TDB

**Times:**
Tu 9:00 - 10:00 AM (Lecture)
Thu 9:00 - 11:00 AM (Computer Lab)

**Contact:** MRobeson@uams.edu
SJun@uams.edu



NSF Award: 1565100

**Low-level features**

- **Decentralized provenance tracking** automates bioinformatics record keeping facilitating reproducibility.
- **Multiple user interfaces**. The same functionality is accessible through graphical interface, command line interface, and API, which target different types of users.
- **Plugin architecture** allows the software to keep pace with the field. Any developer can create and distribute a QIIME 2 plugin.

# “What did you do 5 months ago?”

## notes.txt

```
echo "core_diversity_analyses.py -i
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/otu_table_mc2_w_tax_no_pynast_failures.
biom -o /home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/cd_16662/ -e 16662 -m
/home/caporaso/analysis/atacama-7may2013/map.txt -aO 26 -t
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/rep_set.tre -c
SiteName,Depth,ExtractGroupNo,TransectID,Vegetation" | qsub -keo -N ata-cd
```

The above failed during OTU category significance (see the log file for the error - maybe one of these categories has a value that is observed only once?), so re-running without that step for now...

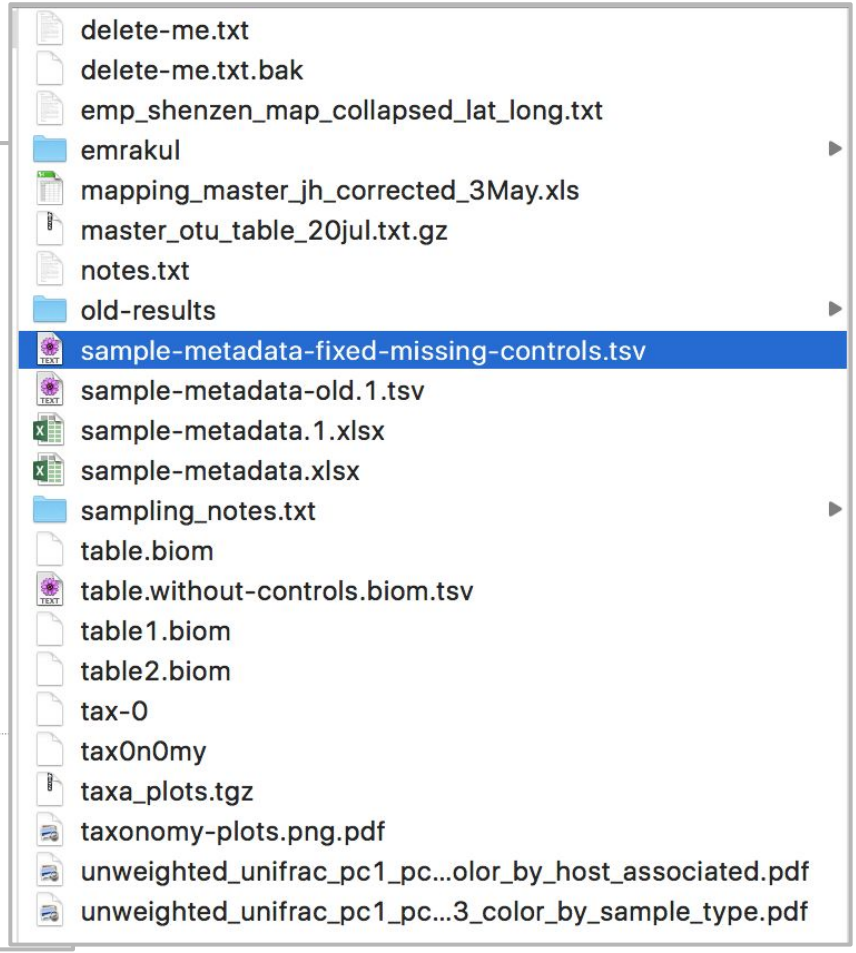
```
echo "core_diversity_analyses.py -i
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/otu_table_mc2_w_tax_no_pynast_failures.
biom -o /home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/cd_16662/ -e 16662 -m
/home/caporaso/analysis/atacama-7may2013/map.txt -aO 26 -t
/home/caporaso/analysis/atacama-7may2013/slout_r1/or_otus/rep_set.tre -c
SiteName,Depth,ExtractGroupNo,TransectID,Vegetation --suppress_otu_category_significance
--recover_from_failure" | qsub -keo -N ata-cd

echo "pick_open_reference_otus.py -i /home/caporaso/analysis/atacama-7may2013/slout_r2/seqs.fna
-r /data/gg_13_5_otus/rep_set/97_otus.fasta -o
/home/caporaso/analysis/atacama-7may2013/slout_r2/or_otus/ -aO 28 -p
/home/caporaso/analysis/atacama-7may2013/uc_fast_params.txt" | qsub -keo -N or-3
```

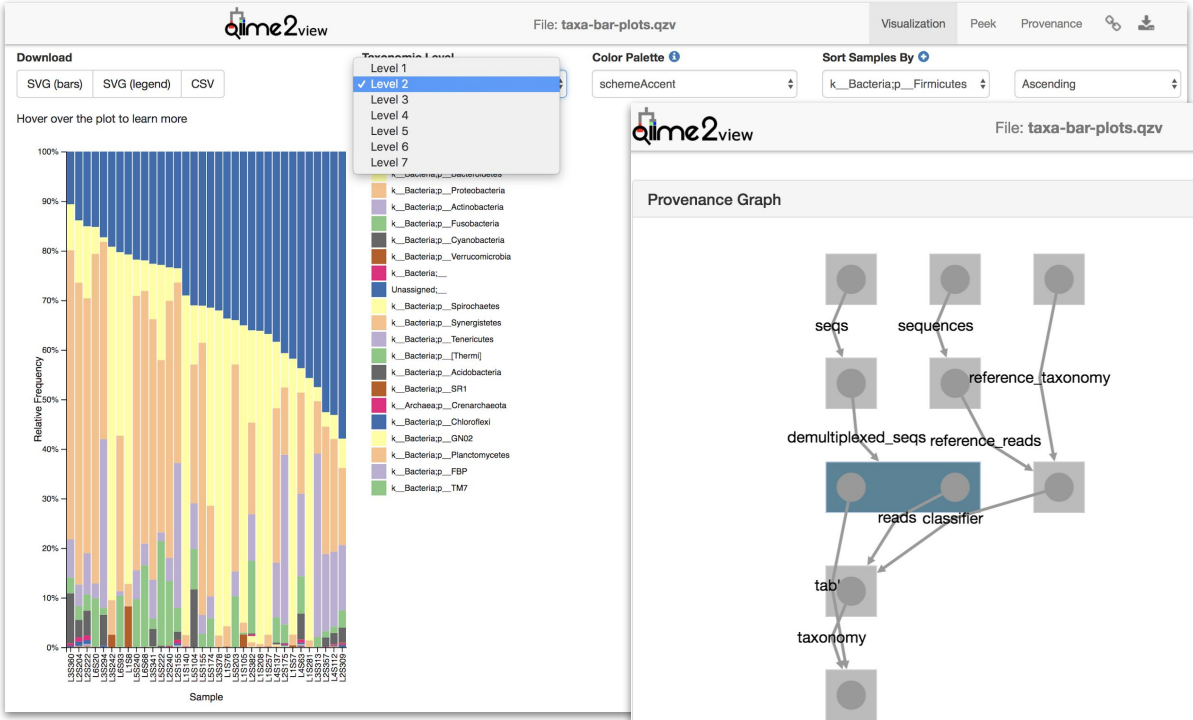
## Katy’s analysis on first and second sequencing runs

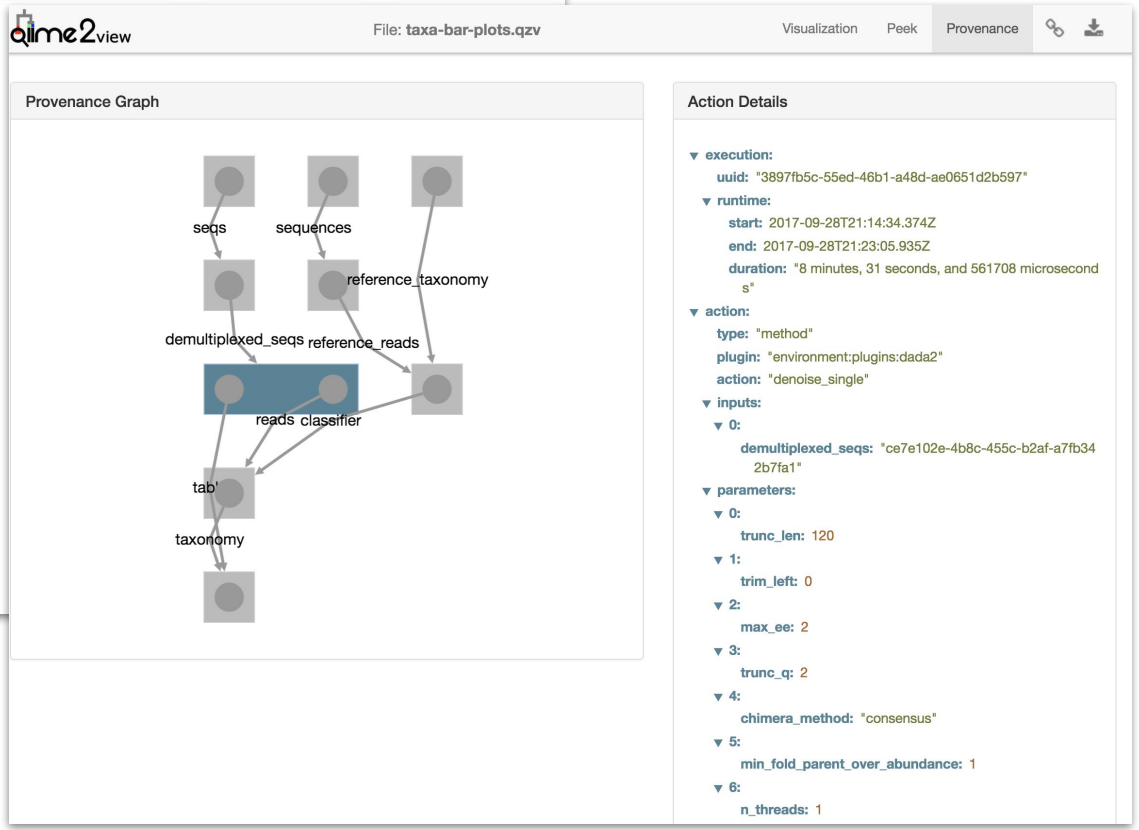
You have read access to all files in `/home/caporaso/analysis/atacama-7may2013`.
The `split_libraries_fastq.py` input files you need are:

sequences:
`/home/caporaso/analysis/atacama-7may2013/2014.04.30/Undetermined_S0_L001_R1_001.fastq.gz`

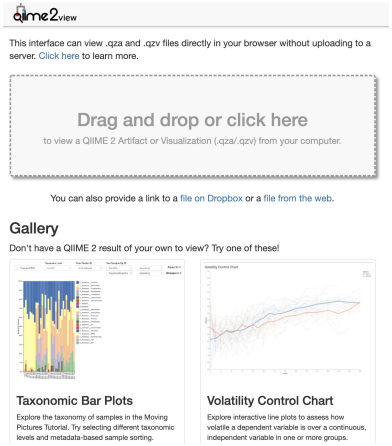


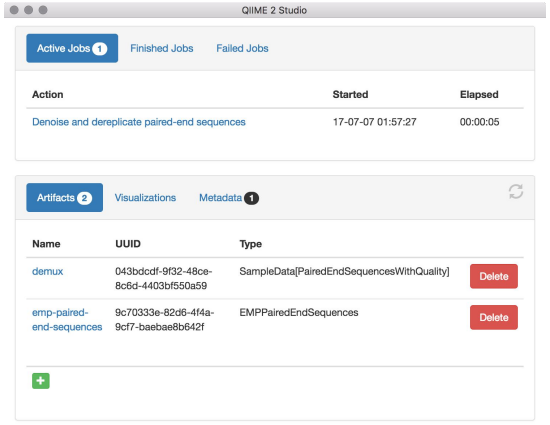
# QIIME 2 integrated data provenance ensures reproducibility

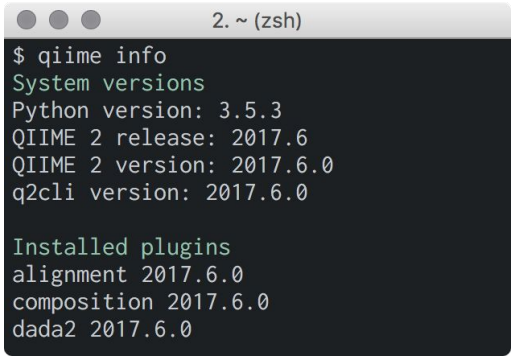


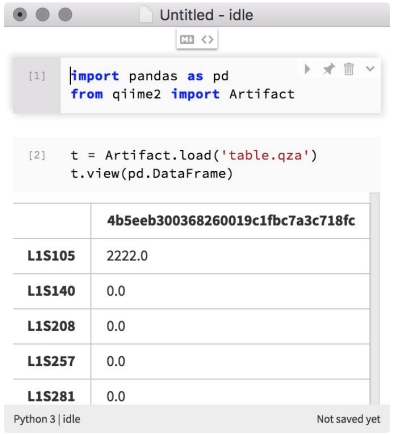



# Use an interface based on your needs.




Galaxy
Cyverse

Galaxy
EBI Metagenomics
Portal
QIITA
NIH Nephele
Cyverse

mothur
QIIME 1

phyloseq

**Related Software**

**Computational Sophistication**

**Data analyst (clinician, policy maker, research subject)**

**Cancer researchers and other domain scientists**

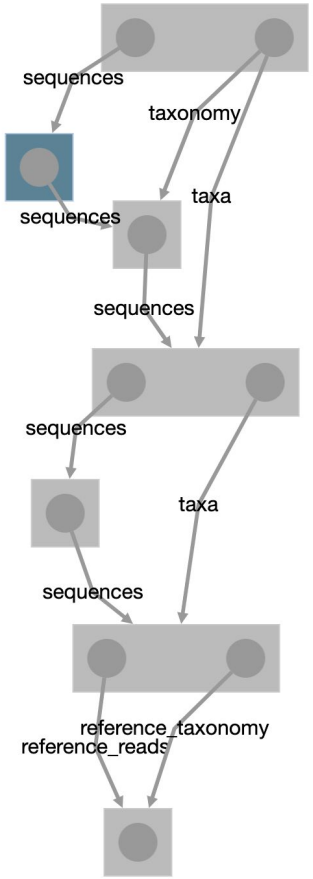
**Power users**

**Data Scientists**

# Plugins

***Developing tools to democratize database acquisition, formatting, and curation.***

```
▼ parameters:
  ▼ 0:
      num_degenerates: 5
  ▼ 1:
      homopolymer_length: 8
  output-name: "clean_sequences"
```

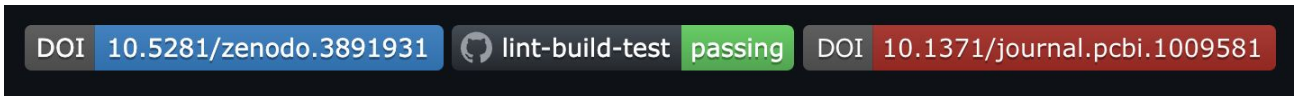



RESEARCH ARTICLE

## RESCRIPt: Reproducible sequence taxonomy reference database management

**Michael S. Robeson, II[1], Devon R. O'Rourke[2], Benjamin D. Kaehler[3], Michal Ziemski[4], Matthew R. Dillon[2], Jeffrey T. Foster[2], Nicholas A. Bokulich[4]***

**1** University of Arkansas for Medical Sciences, Department of Biomedical Informatics, Little Rock, Arkansas, United States of America, **2** Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, United States of America, **3** School of Science, University of New South Wales, Canberra, Australia, **4** Laboratory of Food Systems Biotechnology, Institute of Food, Nutrition, and Health, ETH Zürich, Switzerland


RESCRIPt: Reproducible sequence taxonomy reference database management 74 2021

MS Robeson, DR O'Rourke, BD Kaehler, M Ziemski, MR Dillon, JT Foster, ... Paperpile

PLoS computational biology 17 (11), e1009581

- Bolyen *et al.* 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57.
- Robeson *et al.* 2021. "RESCRIPt: Reproducible Sequence Taxonomy Reference Database Management." *PLoS Computational Biology* 17 (11): e1009581. http://dx.doi.org/10.1371/journal.pcbi.1009581

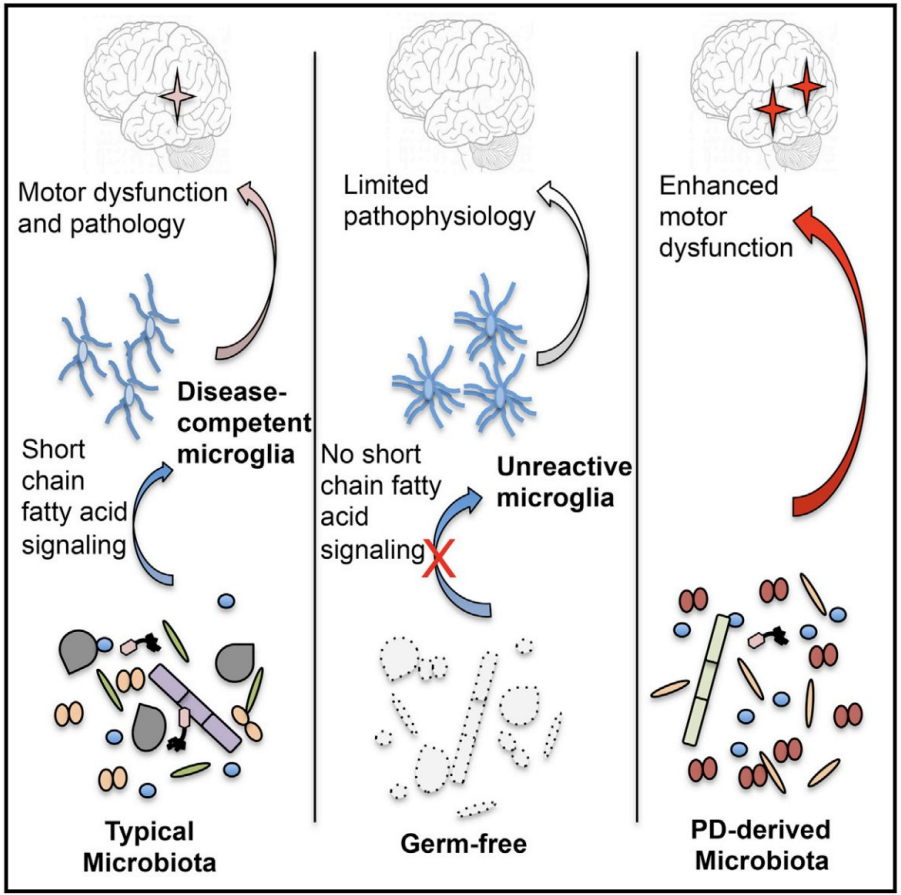
# Parkinson's Mouse QIIME 2 Tutorial

https://docs.qiime2.org/2020.11/tutorials/pd-mice/

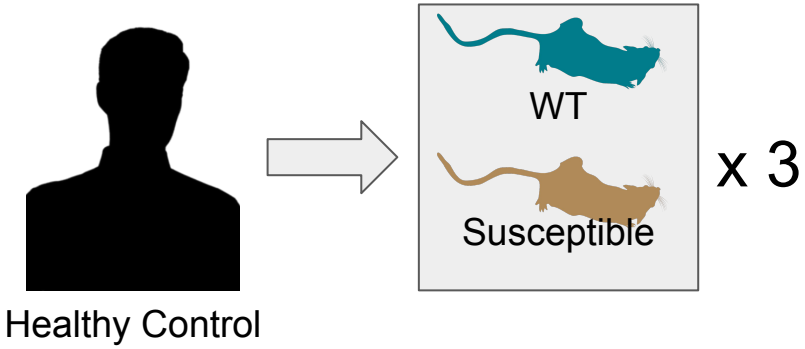
## Gut Microbiota Regulate Motor Deficits and Neuroinflammation in a Model of Parkinson's Disease

Timothy R. Sampson,[1,*] Justine W. Debelius,[2] Taren Thron,[1] Stefan Janssen,[2] Gauri G. Shastri,[1] Zehra Esra Ilhan,[3] Collin Challis,[1] Catherine E. Schretter,[1] Sandra Rocha,[4] Viviana Gradinaru,[1] Marie-Francoise Chesselet,[5] Ali Keshavarzian,[6] Kathleen M. Shannon,[7,9] Rosa Krajmalnik-Brown,[3] Pernilla Wittung-Stafshede,[4] Rob Knight,[2,8] and Sarkis K. Mazmanian[1,10,*]

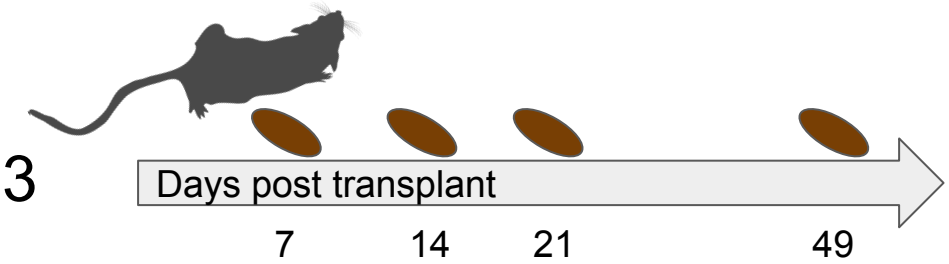
http://dx.doi.org/10.1016/j.cell.2016.11.018



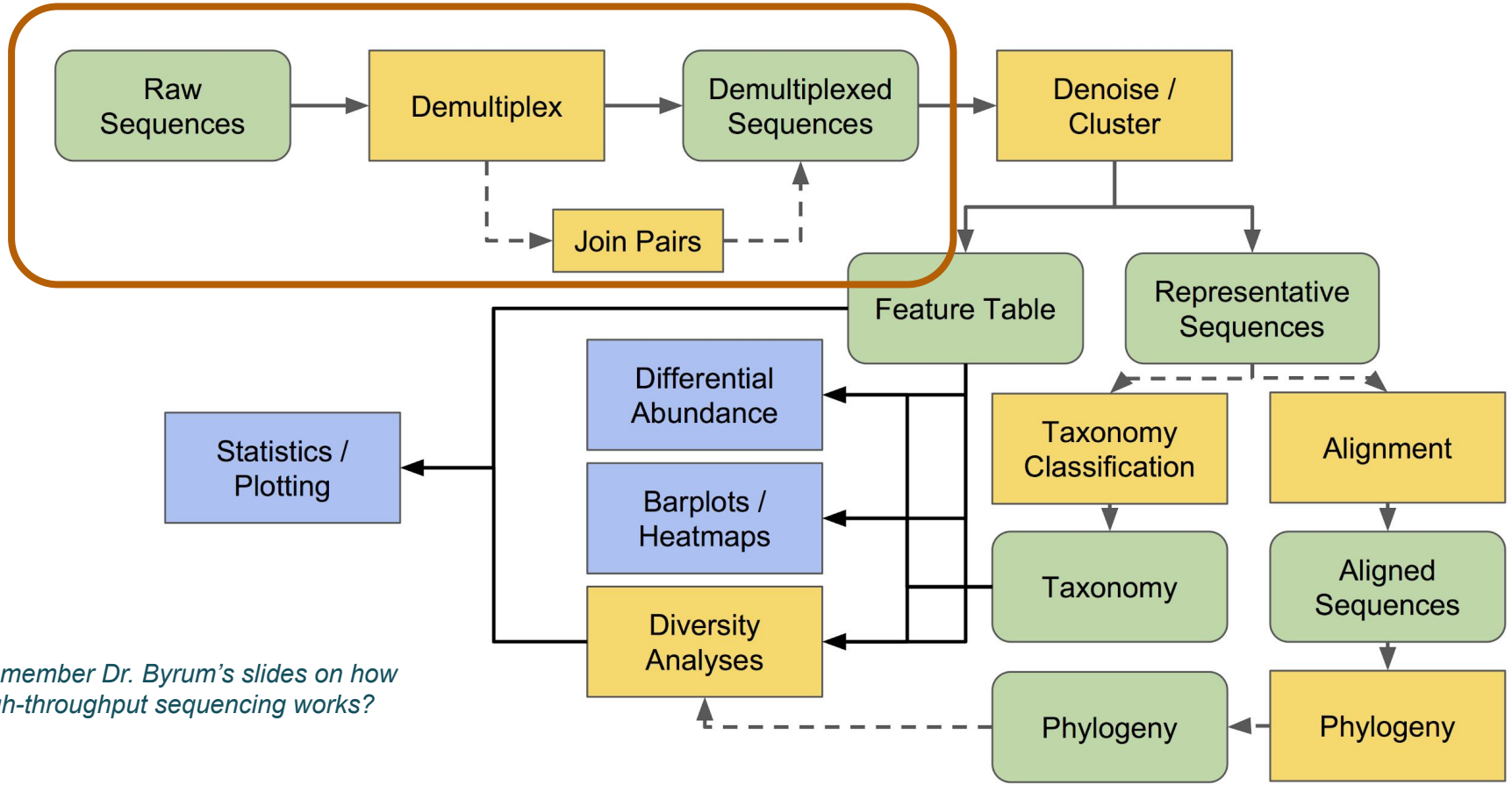
# Humanized Mice Experiment




- 12 donors:
  - 6 healthy
  - with parkinson's disease
- 2 genotypes:
  - Susceptible
  - PD
- 3 cages per donor with mixed mice
- 4 time points per mouse



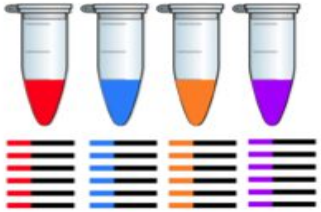
## Import and demultiplex data




*Remember Dr. Byrum's slides on how high-throughput sequencing works?*

Extract DNA, isolate and amplify the rRNA from all samples using barcoded PCR, and sequence.

Barcoded per-sample rRNA



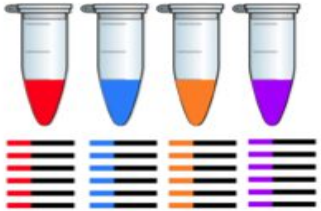
Track per-sample barcodes (e.g., in spreadsheet)

sample-metadata.tsv

| SampleID | BarcodeSequence |
|---|---|
| 4ac2 | AACGCAC |
| e375 | AAGAGAT |
| 4gd8 | ACAGCAG |
| 9872 | ACAGCTA |

# Extract DNA, isolate and amplify the rRNA from all samples using barcoded PCR, and sequence.

Barcoded per-sample rRNA



Pool and sequence samples

Track per-sample barcodes (e.g., in spreadsheet)

sample-metadata.tsv

| SampleID | BarcodeSequence |
|---|---|
| 4ac2 | AACGCAC |
| e375 | AAGAGAT |
| 4gd8 | ACAGCAG |
| 9872 | ACAGCTA |

sequences.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCCATGGTGAAATCCCTCGGCTCAACCGAGGAACTG
+
abaaaaa`^`a_]^\``\``a`^`]]]^^`a[VXGX``Z_\\\_^\a^SYOZVV
SVYGYVDXOZVT\TITBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-6X_9267:1:1:25:267
TACGTATGGGGCAAGCGTTATCCGGAATTATTGCGCGTAAAGACTGCGTAGGTG
GTGGCTTAAGCGCAGGGTTTAAGGCAAT
+
aa^^[__^^^_^_^[^^[^^____ZZ[^
XUWWURZUYY]XXRZRNVTRTNTWUUU^
@HWI-6X_9267:1:1:25:609
TACGTAGGGGGCAAGCGTTATCCGGATT
GATGGACAAGTCTGATGTGAAAGGCTGG
+
aaab`aaa`aaaaaaaaaaaaaaa^aa
[][I^^aZZ^WW^_^`ZZ_T]XY^^\^Z
@HWI-6X_9267:1:1:25:519
GACGGAGGATGCAAGTGTTATCCGGAAT
GTTTACTAAGTCAACTGTTAAATCTTGA
+
abaaaaaa`aaaaaa\aaaaaaa```aa
WY]]_Z_XX\\[]]]^`[\XTVX]`T_V
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAAT
GTTAGGTAAGTCAGATGTGAAAGCCCCG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`
^RVH_PHOWZM[PTRPTRYUBBBBBBBB
```

barcodes.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:267
AAGAGAT
+
bbbbbbb
@HWI-6X_9267:1:1:25:609
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:519
ACAGCAG
+
bbbbbbb
@HWI-6X_9267:1:1:25:1109
ACAGCTA
+
bbbbbbb
@HWI-6X_9267:1:1:25:434
ACACGAG
+
```

Demultiplexed sequence data

Multiplexed sequence data

4ac2.fastq(.gz)

e375.fastq(.gz)

4gd8.fastq(.gz)

9872.fastq(.gz)

```
@HWI-6X 9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAA
AGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAAAGCCC
CGGGCTCCACCTGGGAATGG
+
aaaba^`a^N `\ ``a a]Zaa^^\Z`[M]a`[VYa^ X
^ Z]NZ\`]TY\] ^RVH PHOWZM[PTRPTRYUBBBBBB
BBBBBBBBBBBBBBBBBBB
```

*or*

sample-metadata.tsv

| SampleID | BarcodeSequence |
|---|---|
| 4ac2 | AACGCAC |
| e375 | |
| 4gd8 | |
| 9872 | |

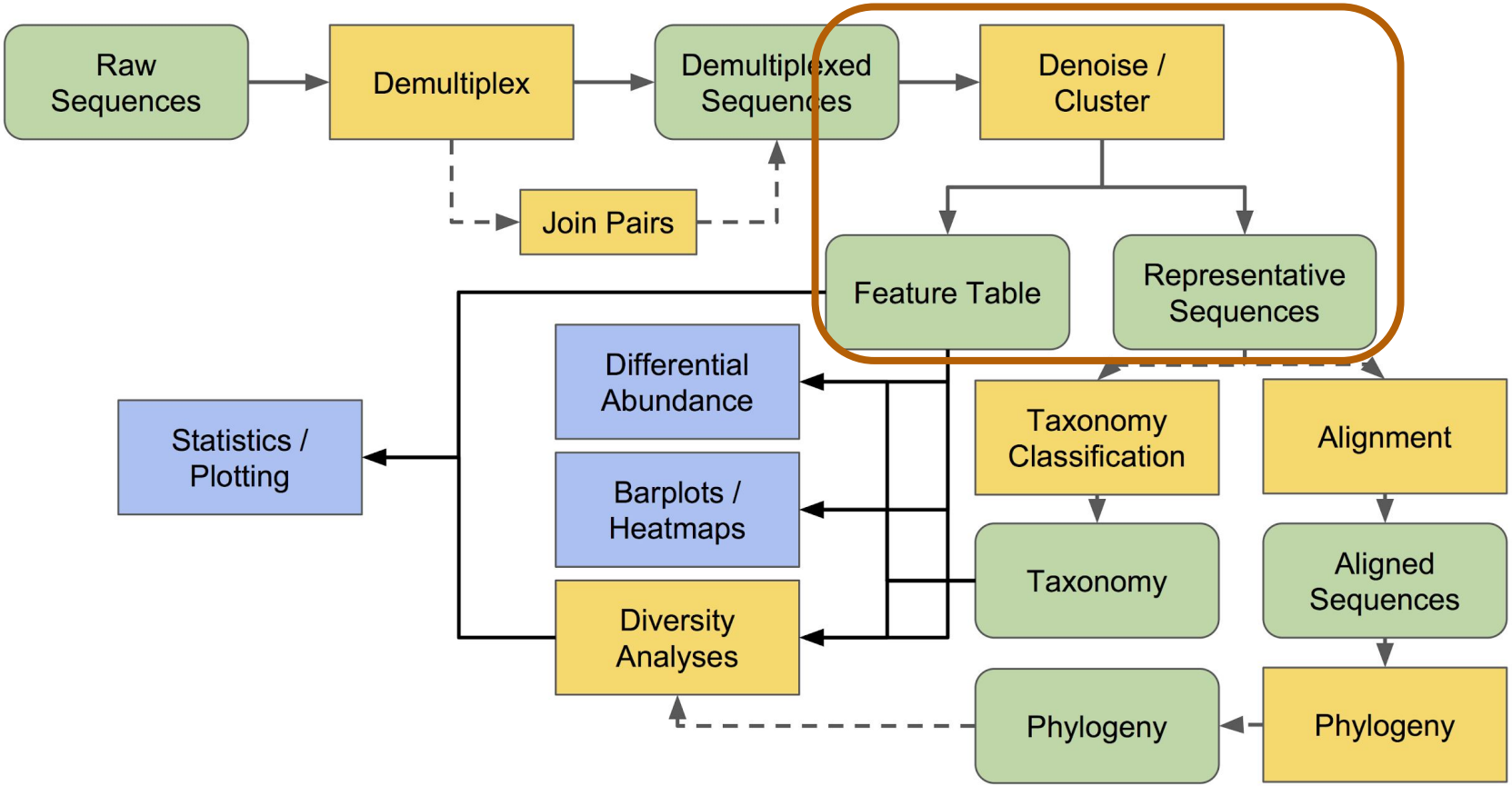
sequences.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCCA                                        TG
+
abaaaaa`^`a_]^
SVYGYVDXOZVT\T
@HWI-6X_9267:1
TACGTATGGGGCAA
GTGGCTTAAGCGCA
+
aa^^[__^^^_^_^
XUWWURZUYY]XXR
@HWI-6X_9267:1
TACGTAGGGGGCAA
GATGGACAAGTCTG
+
```

barcodes.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
AACGCAC
+
bbbbbbb
@HWI-6X_9267:1:1:25:267
AAGAGAT
+
bbbbbbb
@HWI-6X_9267:1:1:25:609
AACGCAC
+
```

## Denoise


Raw Sequences
Demultiplex
Demultiplexed Sequences
Denoise / Cluster
Join Pairs
Feature Table
Representative Sequences
Differential Abundance
Statistics / Plotting
Barplots / Heatmaps
Diversity Analyses
Taxonomy Classification
Alignment
Taxonomy
Aligned Sequences
Phylogeny
Phylogeny

- Clustering
  - Typical threshold, i.e. 97% similarity, “within species variation”.
  - Remove noisy sequences and reduce the amount of sequences to process
  - There are different methods (closed or open reference) and algorithms (sortmerna, vclust)

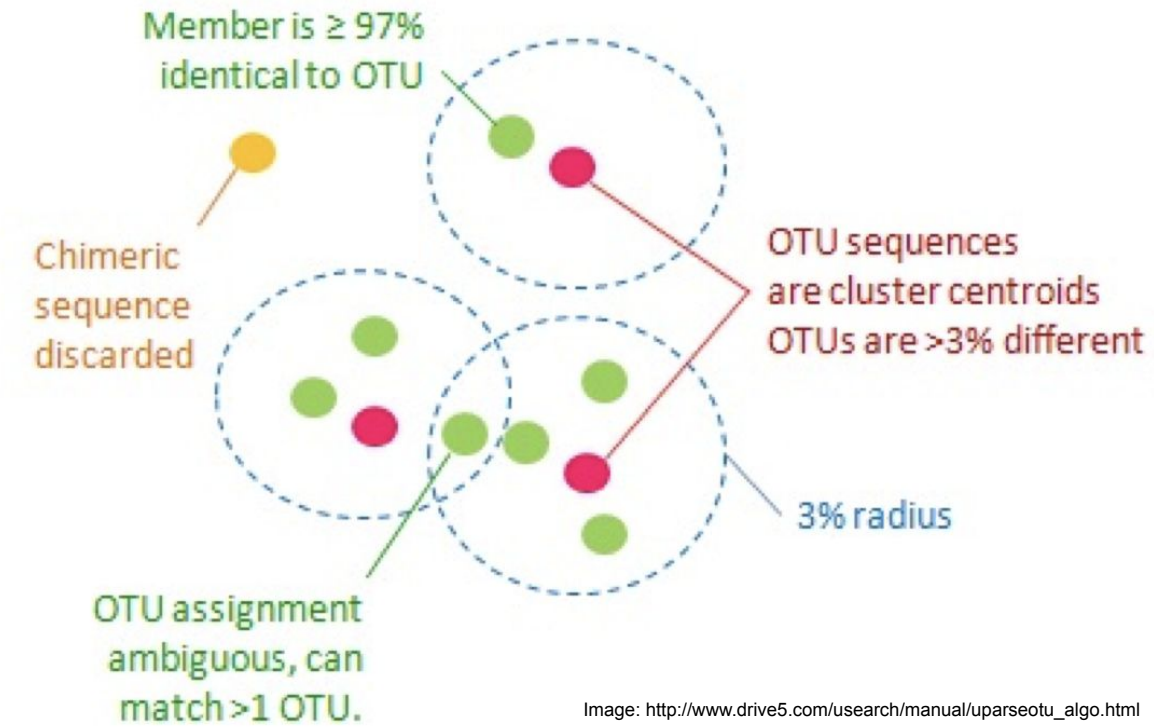
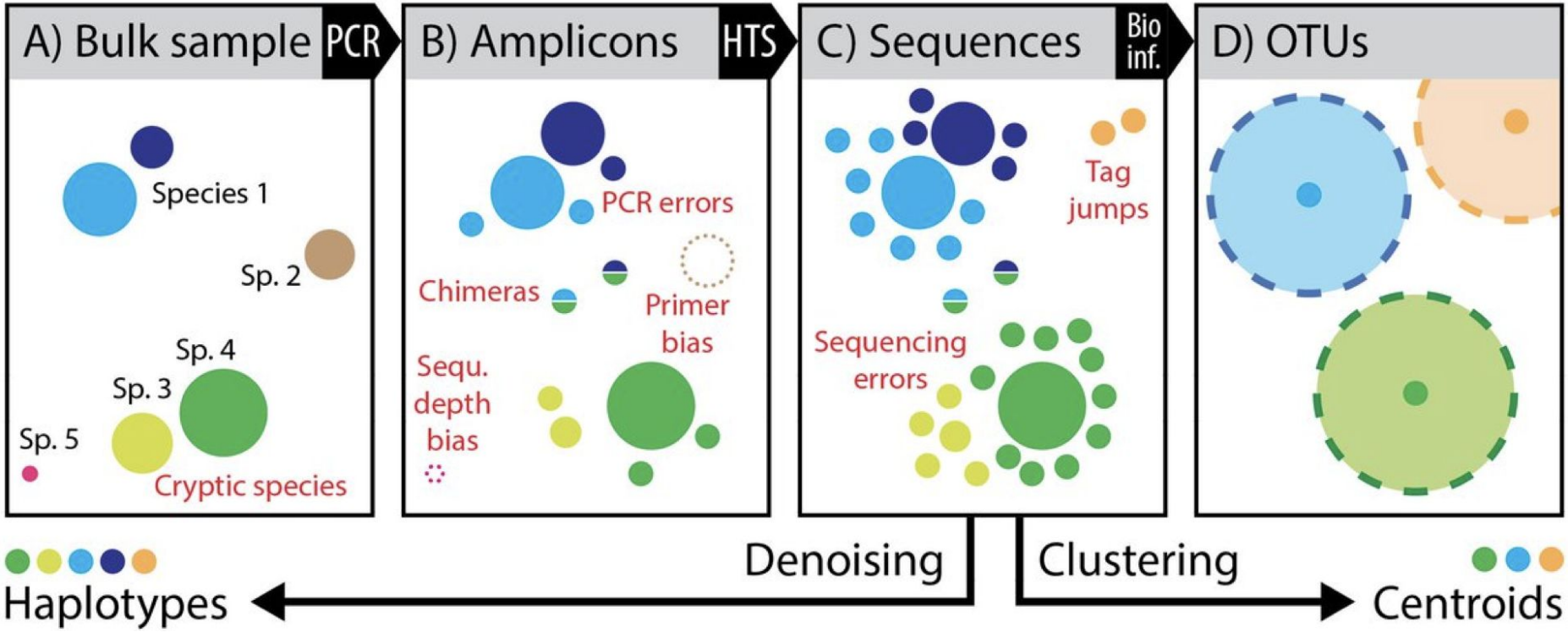


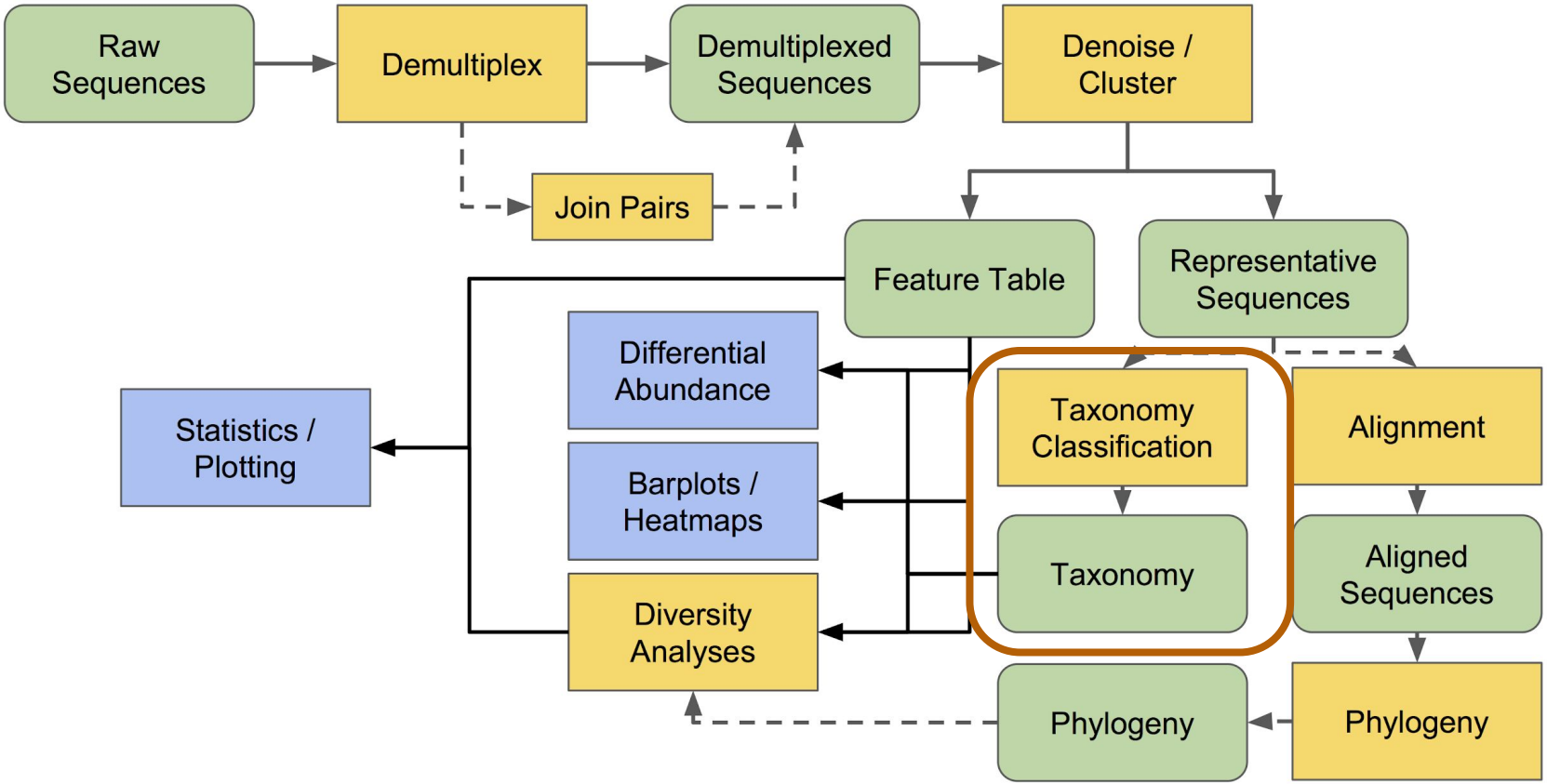

Image: http://www.drive5.com/usearch/manual/uparseotu_algo.html



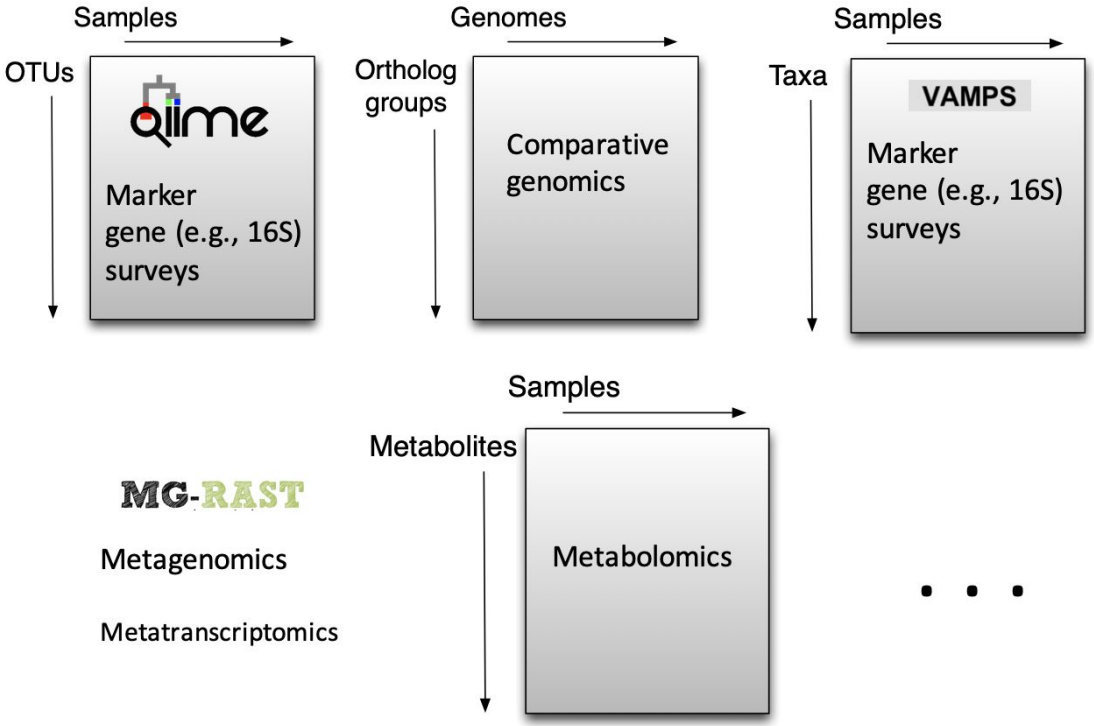

- Remove noise
  - Find the cleanest sequence
  - Correct and/or discard super noisy sequences
  - Examples are: DADA2 and Deblur

Credit: Mehrbod Estaki & QIIME 2 devs.

# Taxonomy Classification



# *sample x observation contingency matrix*




Credit: Bob Bowers

# Taxonomic assignment of observed sequences.

**Reference Database**
Silva, Greengenes, etc.

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCCATGGTGAA
ATCCCTCGGCTCAACCGAGGAACTG
>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAA
AGGCTGGGGCTCAACCCCGGGACGG
>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGCGCAGGGTTTA
AGGCAATGGCTTAACTATTGTTCTC
>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGTCAACTGTTAA
ATCTTGAGGCTCAACCTCGAAATCG
>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGTCAGATGTGAA
AGCCCCGGGCTCCACCTGGGAATGG
```

FeatureData[Sequence]

```
>reference-sequence-1
TTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGTCAACATGGT
GACTCAACCGAGGAACTGAATTGAAGGTGGGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTG
GCTTGGTAAGTCAACATGGTGACTCAACCGAGGAACTGAA
>reference-sequence-2
AACGTAGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGTCTGATGTGAAAGG
```

FeatureData[Taxonomy]

| | |
|---|---|
| reference-sequence-1 | Bacteria; Proteobacteria; Gammaproteobact |
| reference-sequence-2 | Bacteria; Bacteroidetes; Flavobacteria; F |
| reference-sequence-3 | Bacteria; Proteobacteria; Deltaproteobact |
| reference-sequence-4 | Archaea; Euryarchaeota; DSEG; 104A5 |

Compare observed sequences to annotated reference sequences to make taxonomic assignments.

FeatureData[Taxonomy]

| | |
|---|---|
| feature5 | Bacteria; Proteobacteria |
| feature4 | Bacteria; Proteobacteria |
| feature2 | Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales |
| feature1 | Bacteria; Proteobacteria |
| feature3 | Bacteria; Proteobacteria; Deltaproteobacteria |

???

# Classify Taxonomies

`RDP Classifier (naive Bayes)`

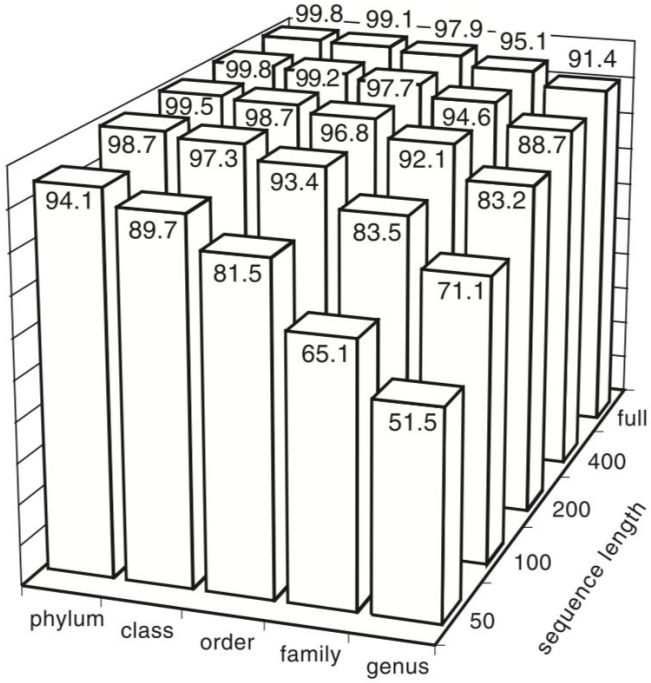
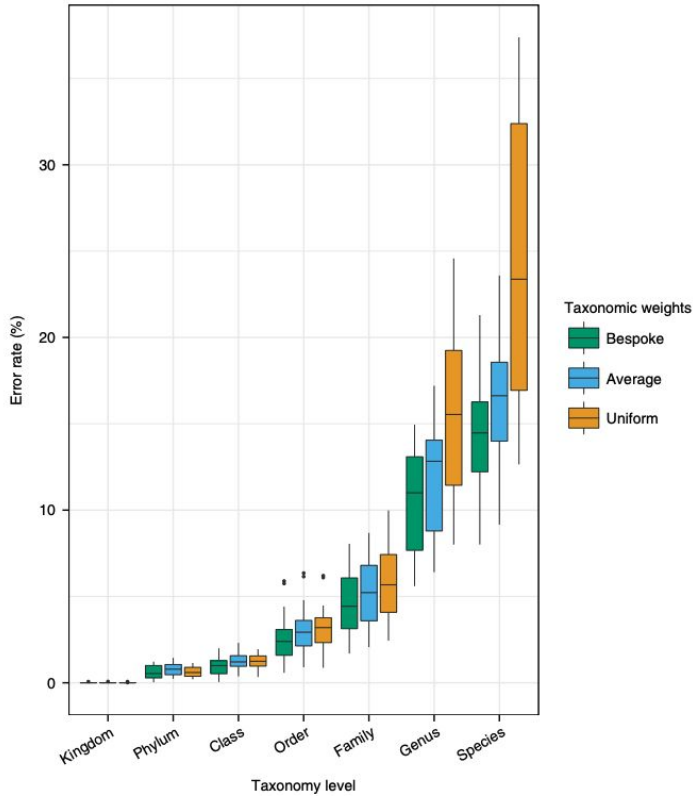



FIG. 1. Overall classification accuracy by query size (exhaustive leave-one-out testing using the Bergey corpus). Numbers are percentages of tests correctly classified.

Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Wang et al. 2007. Applied and Environmental Microbiology.

# Habitat Aware Classification

`qiime clawback ...`

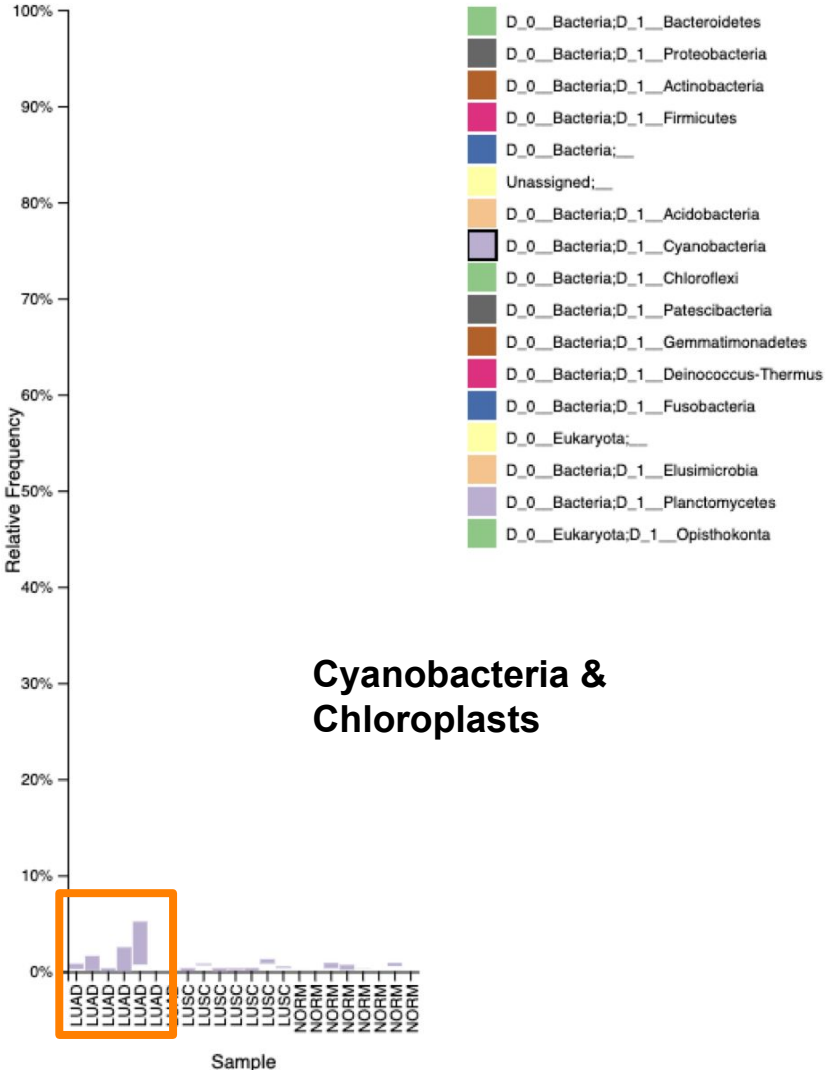



Kaehler et al. (2019) “Species Abundance Information Improves Sequence Taxonomy Classification Accuracy.” Nature Communications 10 (1): 4643.

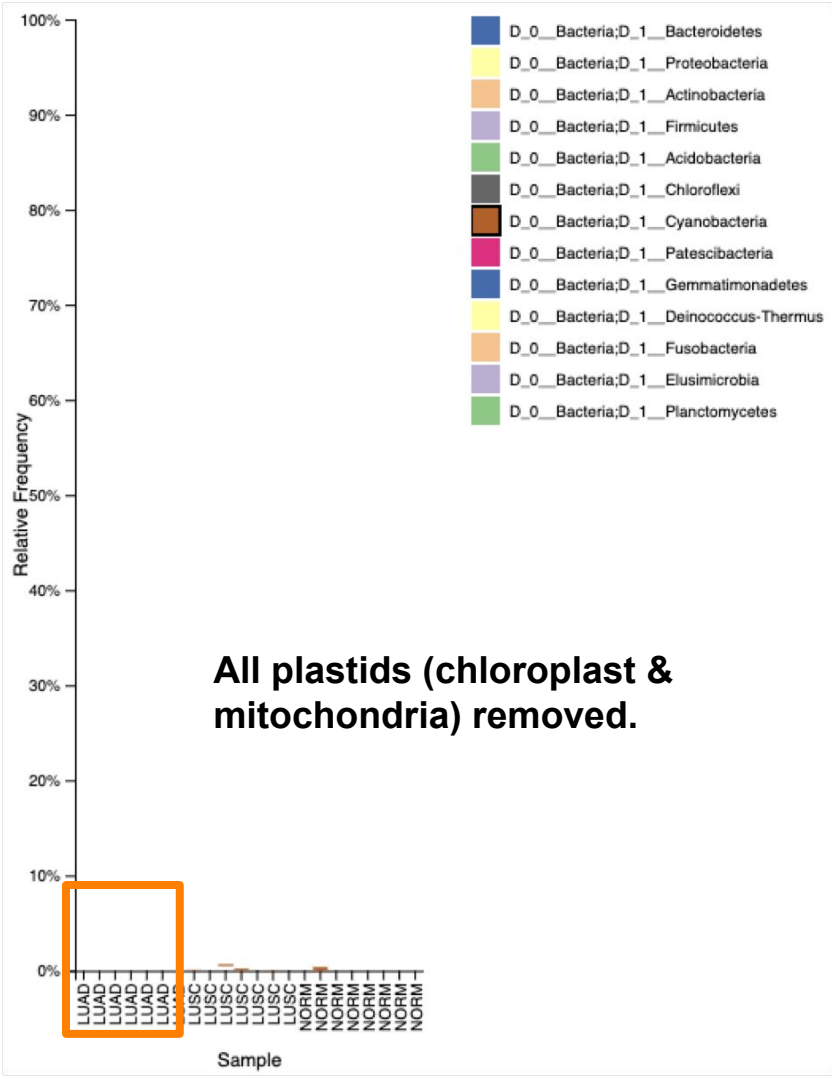
## New SILVA taxonomy

d__Bacteria; **p__Proteobacteria;** c__Alphaproteobacteria; o__Rickettsiales; **f__Mitochondria; g__Mitochondria**
d__Bacteria; **p__Proteobacteria;** c__Alphaproteobacteria; o__Rickettsiales; **f__Mitochondria; g__Mitochondria; s__Aspergillus_alliaceus**

d__Bacteria; **p__Cyanobacteria;** c__Cyanobacteriia; **o__Chloroplast; f__Chloroplast; g__Chloroplast**
d__Bacteria; **p__Cyanobacteria;** c__Cyanobacteriia; **o__Chloroplast; f__Chloroplast; g__Chloroplast; s__Phytophthora_lateralis**

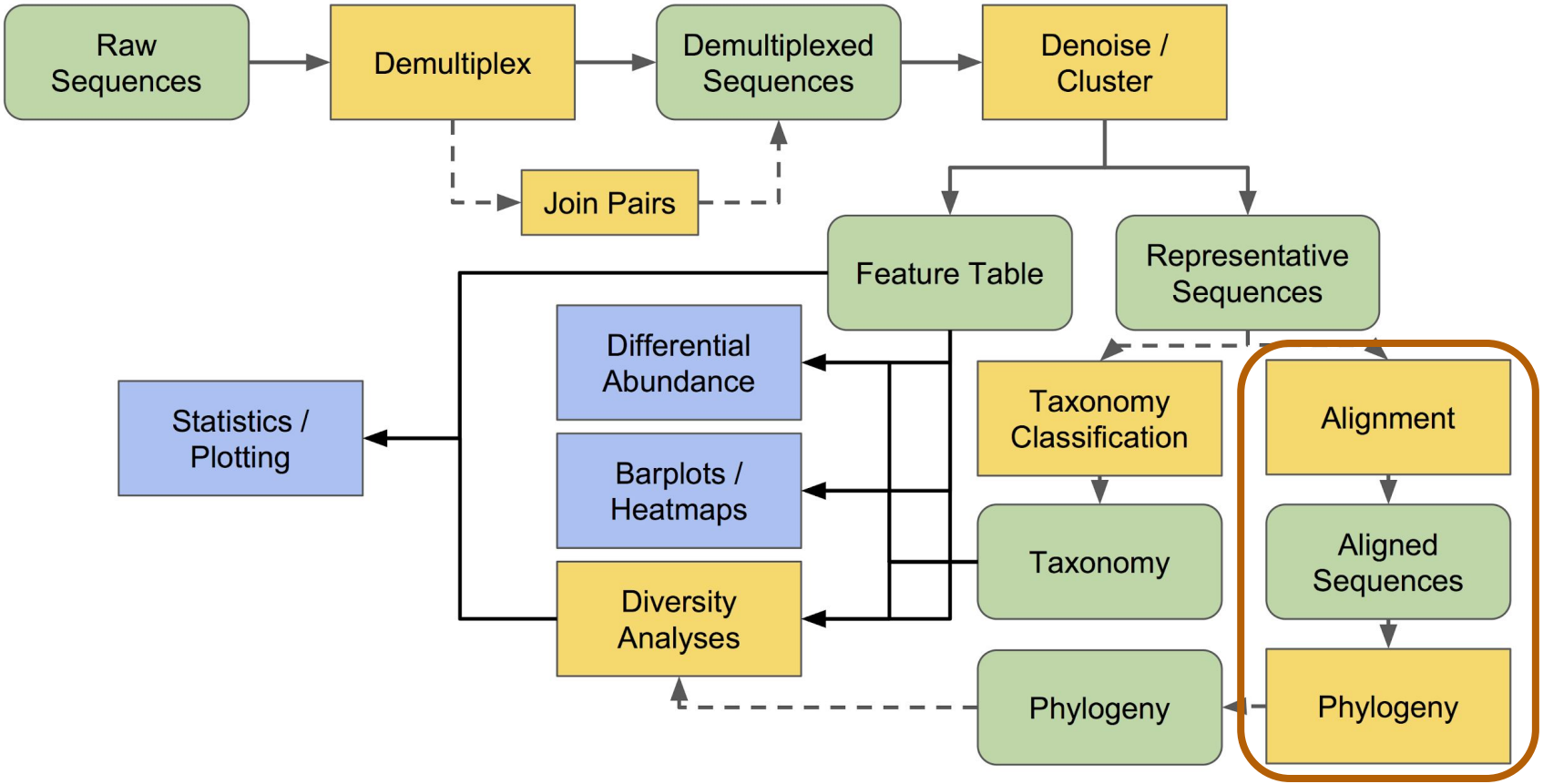



**Cyanobacteria & Chloroplasts**



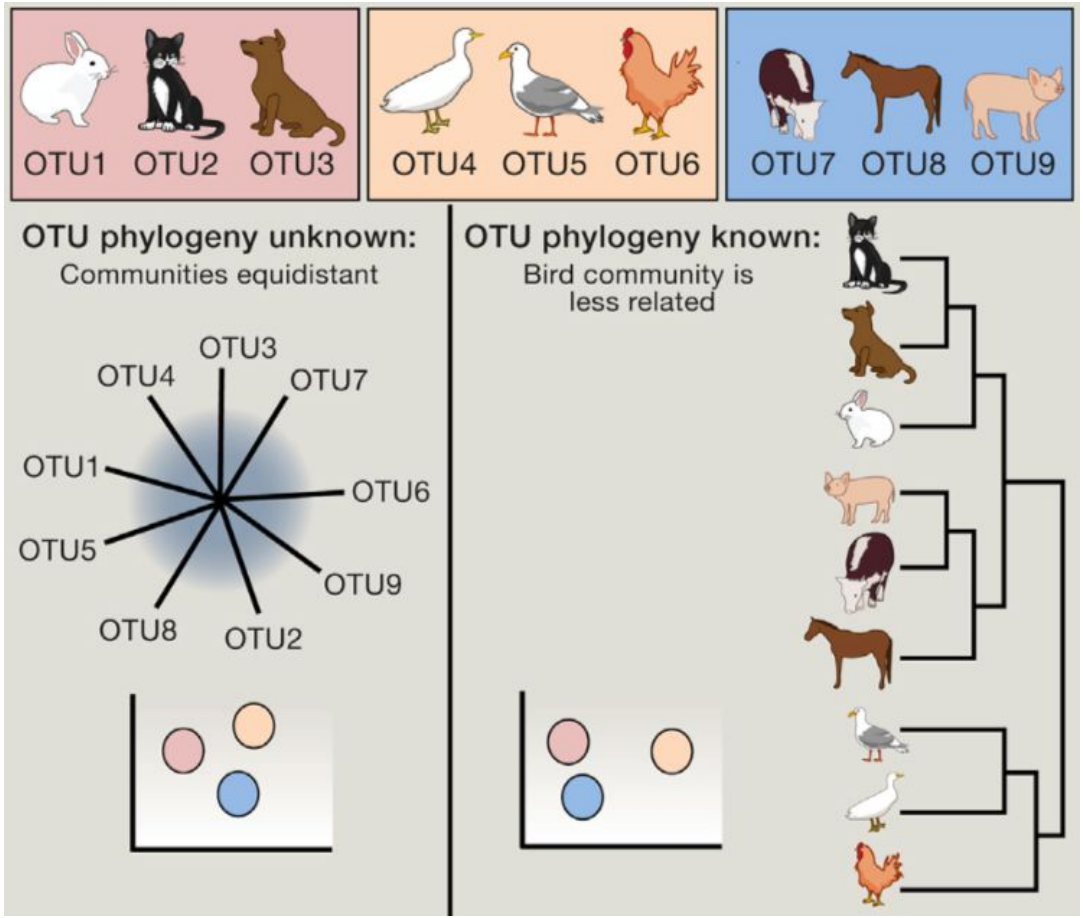

**All plastids (chloroplast & mitochondria) removed.**

## Phylogenetics



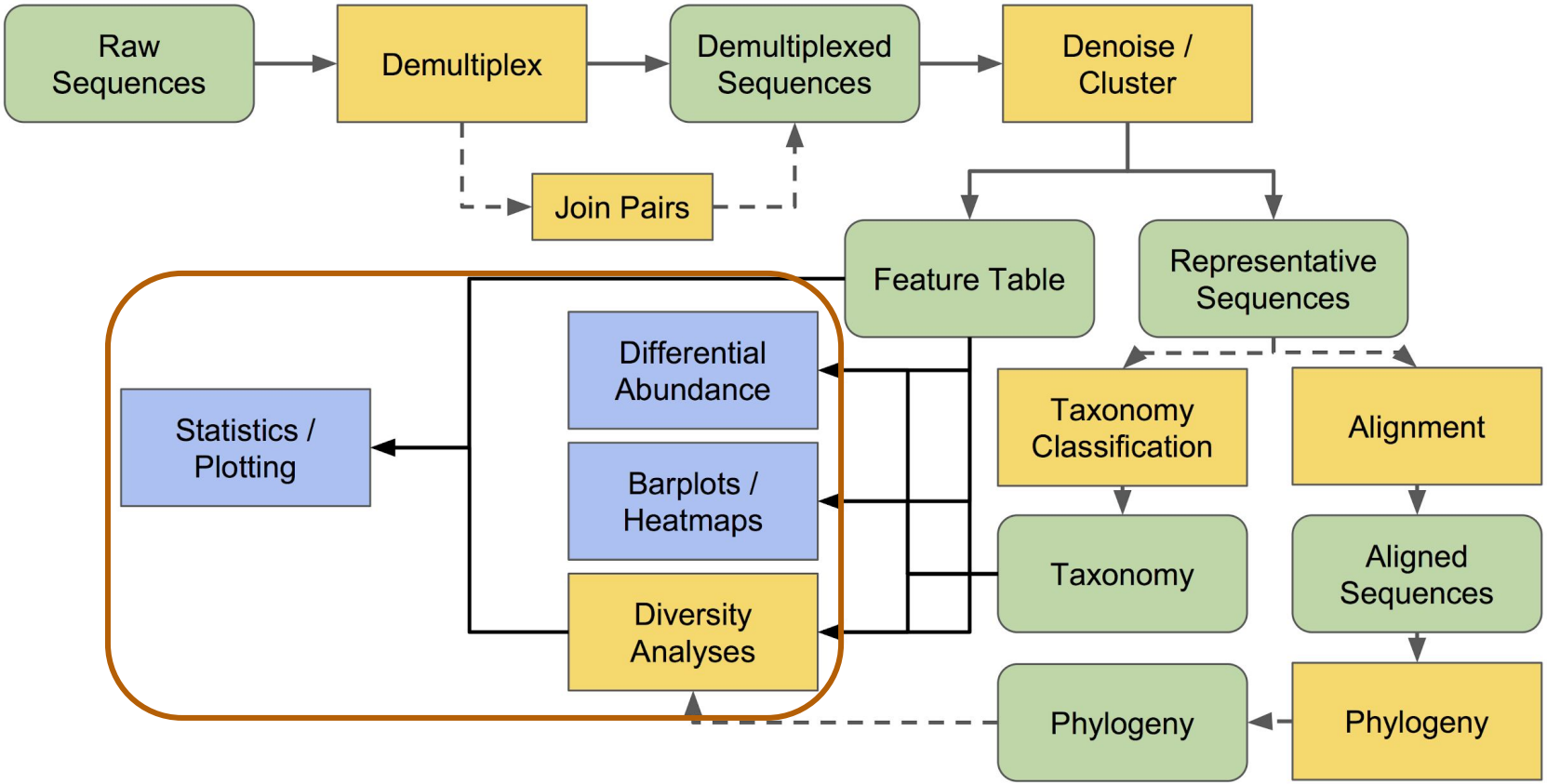
**Non-phylogenetic diversity metrics** assume that all taxa are equally related, so don't make assumptions about evolutionary relationships.

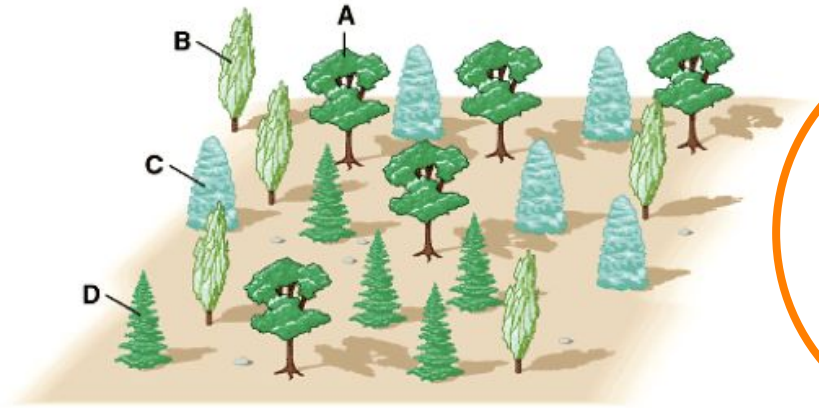
**Phylogenetic diversity metrics** incorporate evolutionary relationships between taxa, but assume that we know what those relationships are.




Goodrich et al. (2014) Conducting a Microbiome Study. Cell 158(2): 250–262

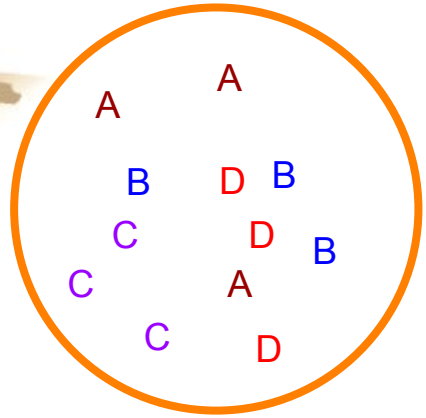
## Analysis


Raw Sequences
Demultiplex
Demultiplexed Sequences
Denoise / Cluster
Join Pairs
Feature Table
Representative Sequences
Differential Abundance
Statistics / Plotting
Barplots / Heatmaps
Diversity Analyses
Taxonomy Classification
Alignment
Taxonomy
Aligned Sequences
Phylogeny
Phylogeny




**Community 1**

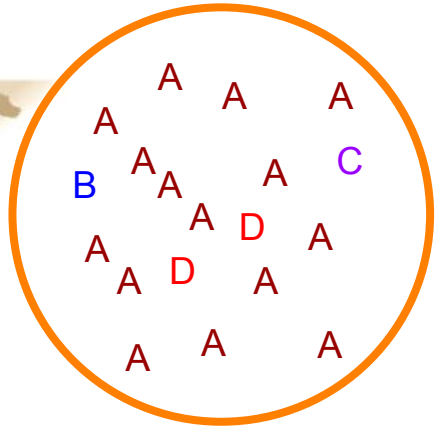
**A: 25%  B: 25%  C: 25%  D: 25%**




- Evenness.
- Approx. equal representation of taxa within the community.

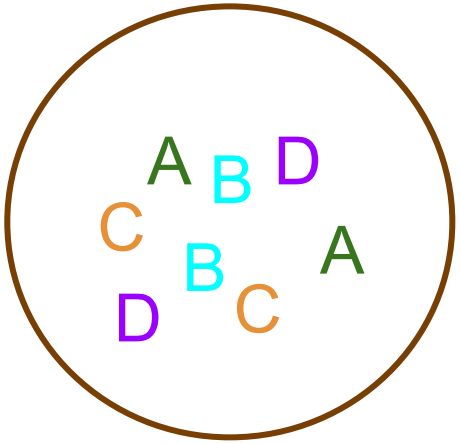
**Community 2**

**A: 80%  B: 5%  C: 5%  D: 10%**

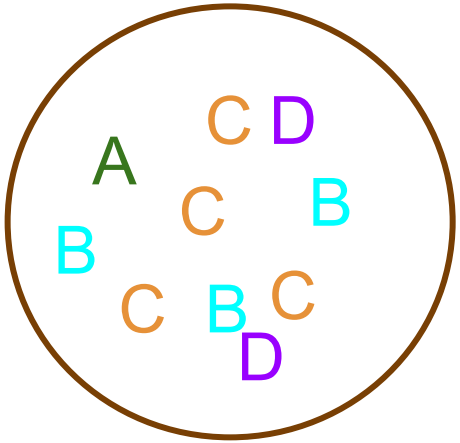



- Not-even / dominance.
- Unequal distribution of taxa.

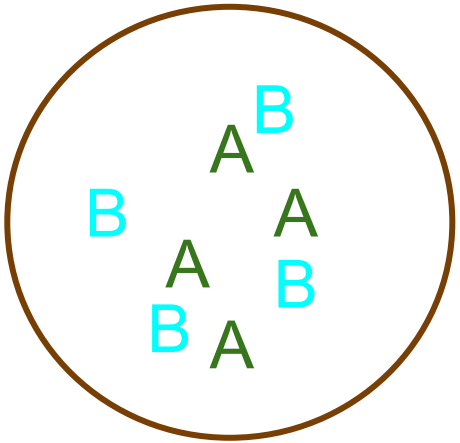
(A) Rich and even



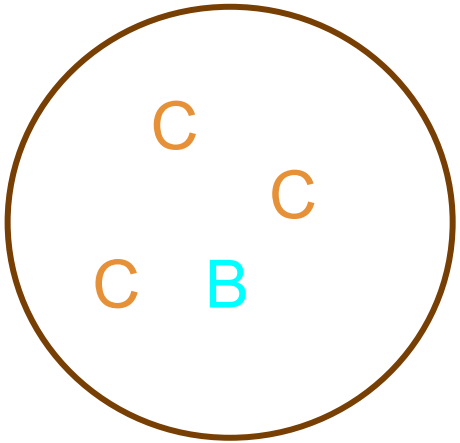

(B) Rich but not even

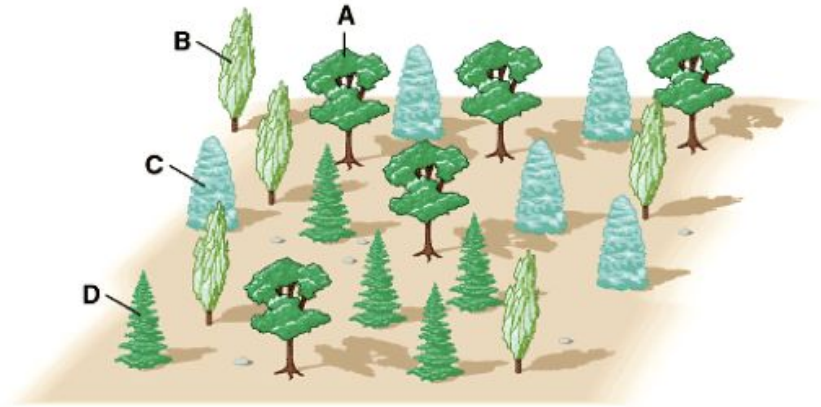



(C) Not rich but even




(D) Not rich and not even






Community 1

A: 25% B: 25% C: 25% D: 25%

Community 2

A: 80% B: 5% C: 5% D: 10%



**Communities 1 & 2 can be said to have**

**A. same richness**

- **a. both contain same 4 species (presence / absence)**
- **b. qualitative metrics**

**B. different abundance / evenness**

- **a. both contain same 4 species, with varying counts**
- **b. quantitative metrics**

Lozupone *et al*. 2007. “Quantitative and Qualitative Beta Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities.” *Applied and Environmental Microbiology* 73 (5): 1576–85.

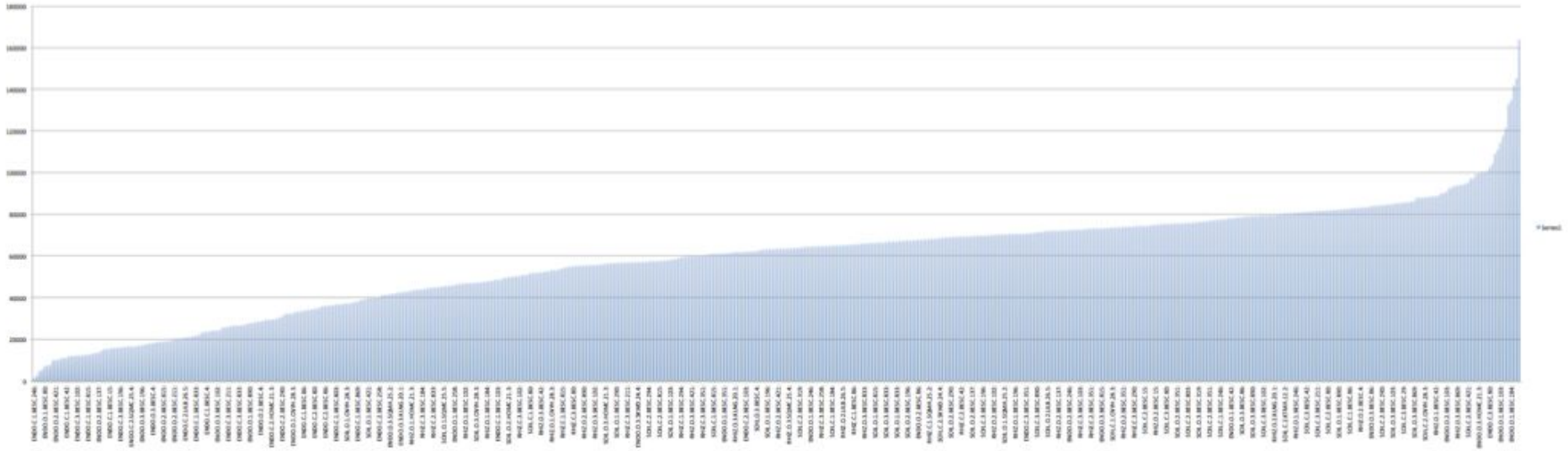
**Table 1**. Categories of diversity measurements

| | Measurement of diversity within a single community (α-diversity) | Measurement of diversity shared among communities (β-diversity) |
|---|---|---|
| Only presence/absence of taxa considered | *Qualitative α-diversity* (Richness)<br>*Species-based:*<br>Chao 1<br>ACE<br>Rarefaction<br>*Divergence-based:*<br>Phylogenetic diversity (PD) | *Qualitative β-diversity*<br>Species-based:<br>Sörensen index<br>Jaccard index<br>*Divergence-based:*<br>Unweighted UNIFRAC<br>Taxonomic similarity ($\Delta_S$) |
| Additionally accounts for the number of times that each taxon was observed | *Quantitative α-diversity* (Richness and/or Evenness)<br>*Species-based:*<br>Shannon's index<br>Simpson's index<br>*Divergence-based:*<br>θ | *Quantitative β diversity*<br>*Species-based:*<br>Sörensen quantitative index<br>Morisita-Horn measure<br>*Divergence-based:*<br>Weighted UNIFRAC<br>$F_{ST}$<br>DPCoA |

Also see this wonderful community post on diversity measures: https://forum.qiime2.org/t/alpha-and-beta-diversity-explanations-and-commands/2282

Lozupone & Knight. 2008. "Species Divergence and the Measurement of Microbial Diversity." *FEMS Microbiology Reviews* 32 (4): 557–78.

# Variation in sequencing depth can be a problem.

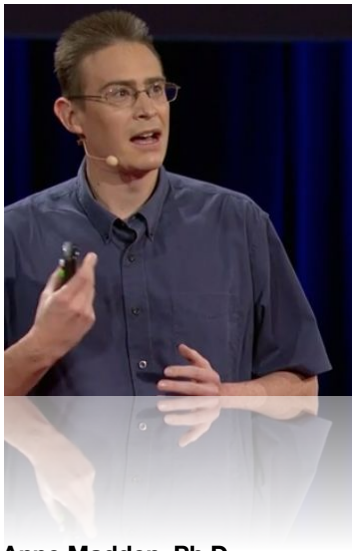



https://qiime2.org/

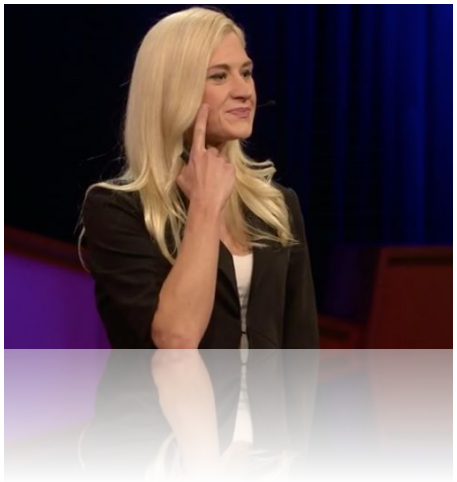
https://www.youtube.com/c/QIIME2

# TED Talks about the microbiome

**Jonathan Eisen, Ph.D.**:
https://www.ted.com/talks/jonathan_eisen_meet_your_microbes?language=en



**Rob Knight, Ph.D.**:
https://www.ted.com/talks/rob_knight_how_our_microbes_make_us_who_we_are?language=en



**Anne Madden, Ph.D.**:
https://www.ted.com/talks/anne_madden_meet_the_microscopic_life_in_your_home_and_on_your_face?language=en

# Some great reading!