16S rRNA Gene Amplicon Survey: Study Design and Case Study

Considerations for a Longitudinal Case Study of Antibiotic Treatment and Virus Infection

Scott A. Handley, PhD
Assistant Professor
Department of Pathology & Immunology
Washington University School of Medicine

Rationale

- 16S amplicon surveys are extensively used to study the mouse bacterial microbiome in a large variety of contexts
    - e.g. disease, nutrition, sociology, neuroscience, etc.
- ***Frequently*** fail due to poor study design
    - Batch effects
        - Cage, paternity/breeding, facility, origin effects
    - Co-housed survival studies (specific example)
    - Statistical considerations
        - Detecting signal from noise
        - Minimize variance
        - Filtering out misbehaved data
- Many of these principles apply to other data types (RNAseq)

Image credit: Davide Bonazzi/@Salmanart

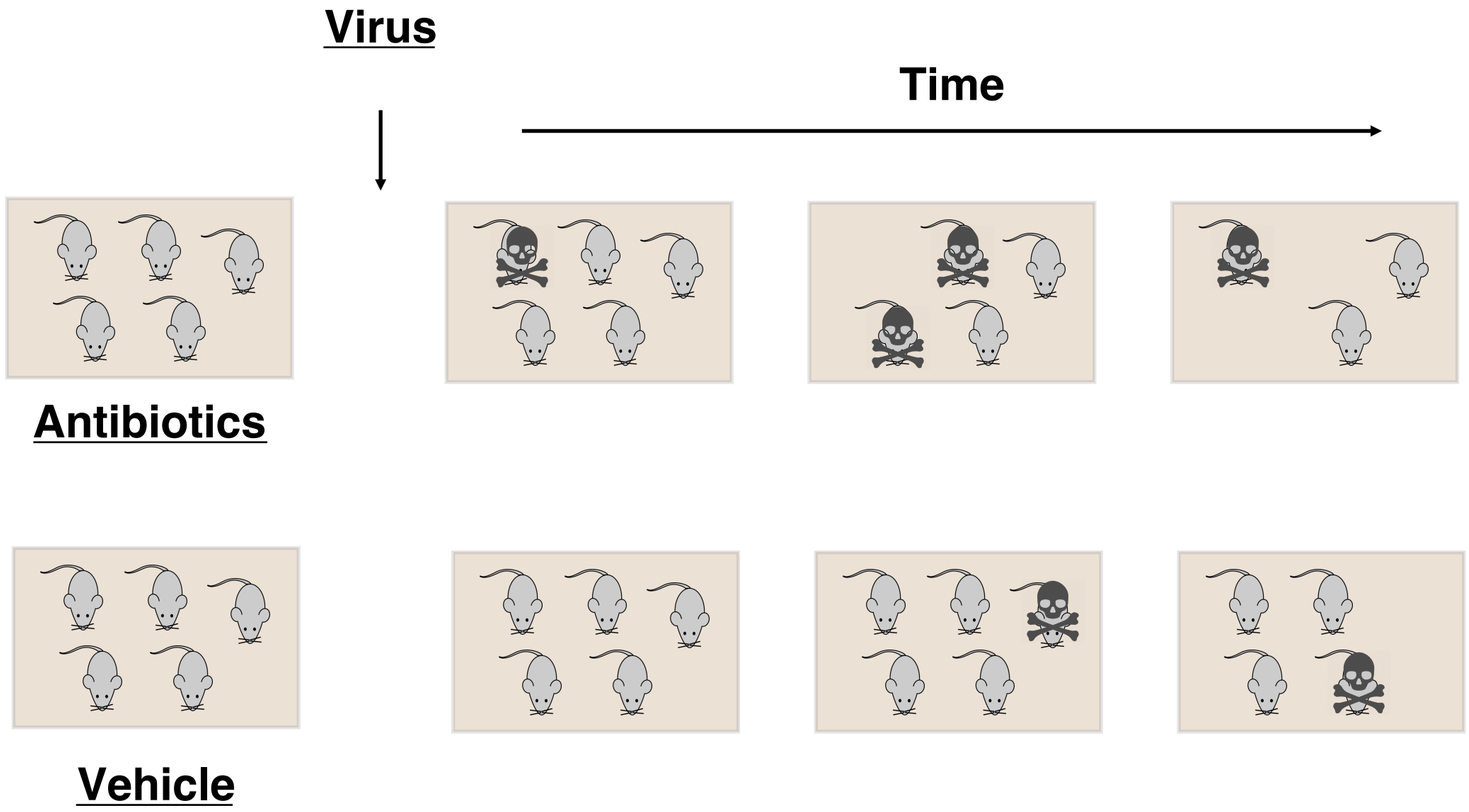
"**Mouse microbes may make scientific studies harder to replicate**" Kelly Servick. Science Aug 16, 2016

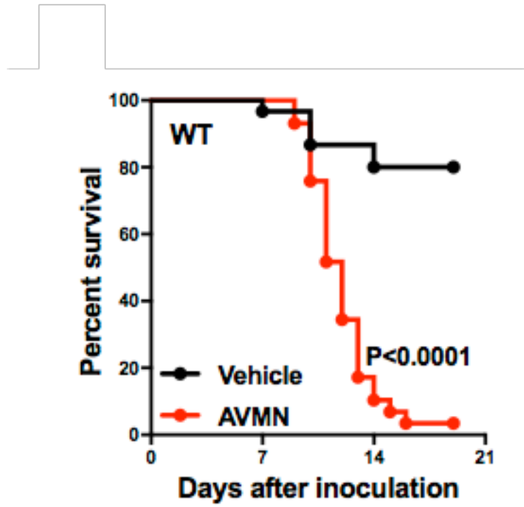
"**Accounting for reciprocal host-microbiome interactions in experimental science**" Stappenbeck, TS and Virgin HW. Nature. 2016 Jun 9;534(7606):191-9

Today's Case Study

Thackray LB, Handley SA, Gorman MJ, Poddar S, Bagadia P, Briseño CG, Theisen DJ, Tan Q, Hykes BL Jr, Lin H, Lucas TM, Desai C, Gordon JI, Murphy KM, Virgin HW, Diamond MS. **Oral Antibiotic Treatment of Mice Exacerbates the Disease Severity of Multiple Flavivirus Infections**. Cell Rep. 2018 Mar 27;22(13):3440-3453.e6. PubMed PMID: 29590614

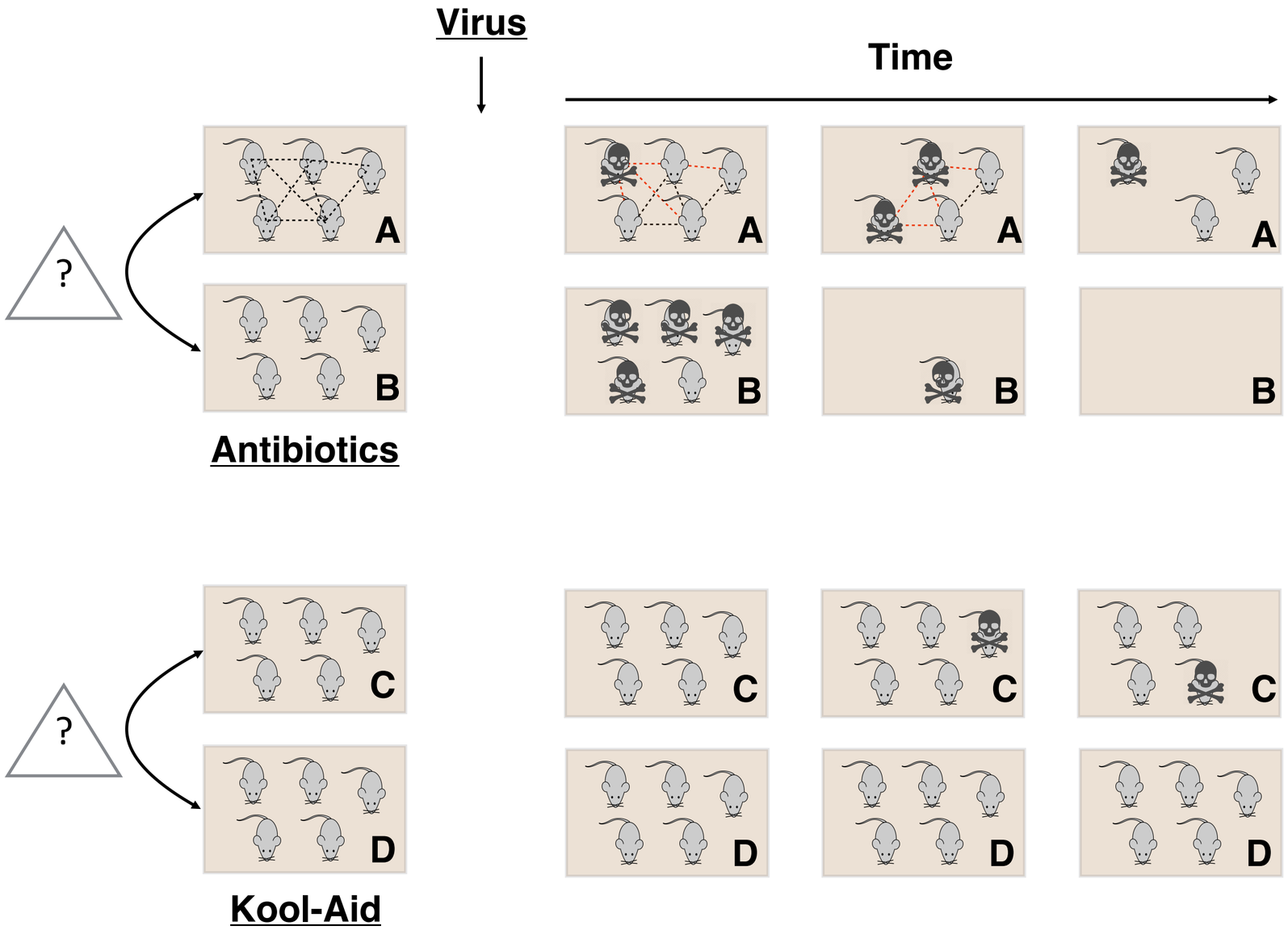
Case Study: Effect of Antibiotics on Viral Pathogenesis





Thackray LB, et al.Cell Rep. 2018 Mar 27;22(13):3440-3453.e6. PubMed PMID: 29590614

Cage and Mouse-to-Mouse Effects

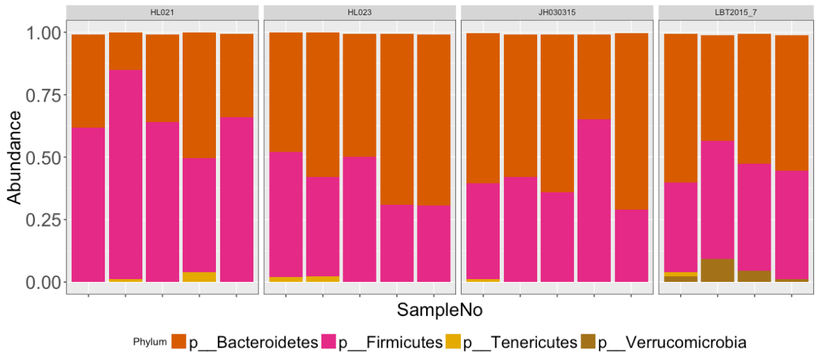


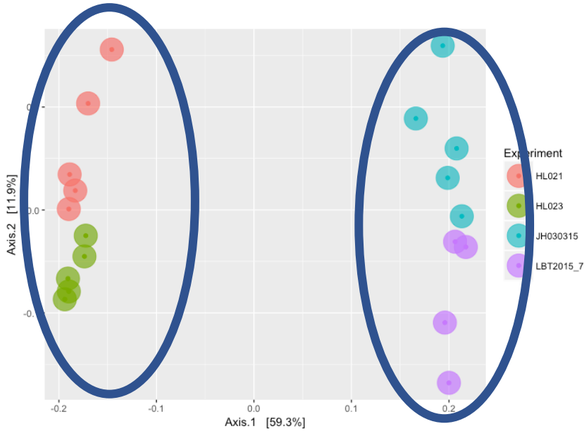
Cage Effects: 14 days post-treatment (pre-infection)

Phyla

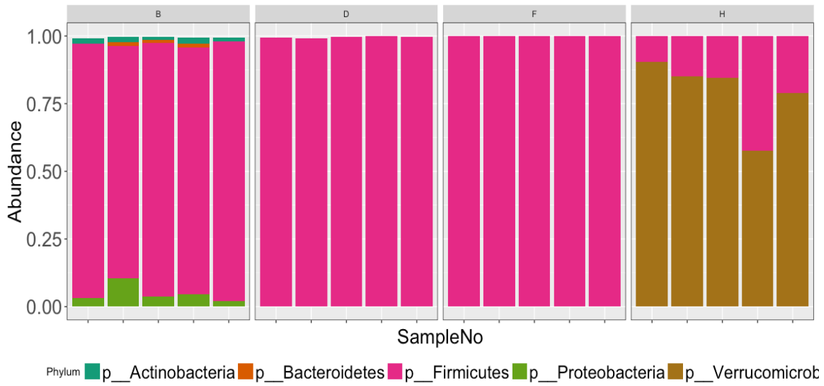
Beta diversity

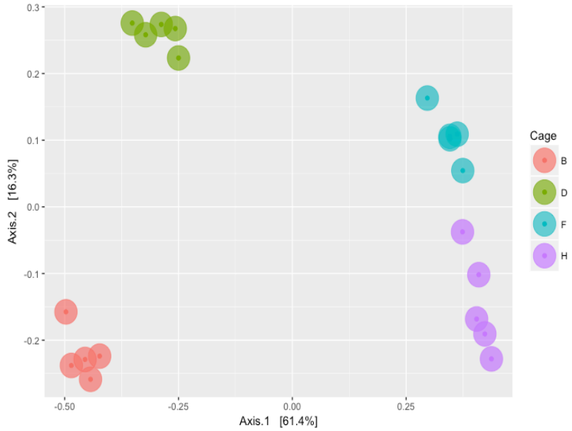
Kool-Aid



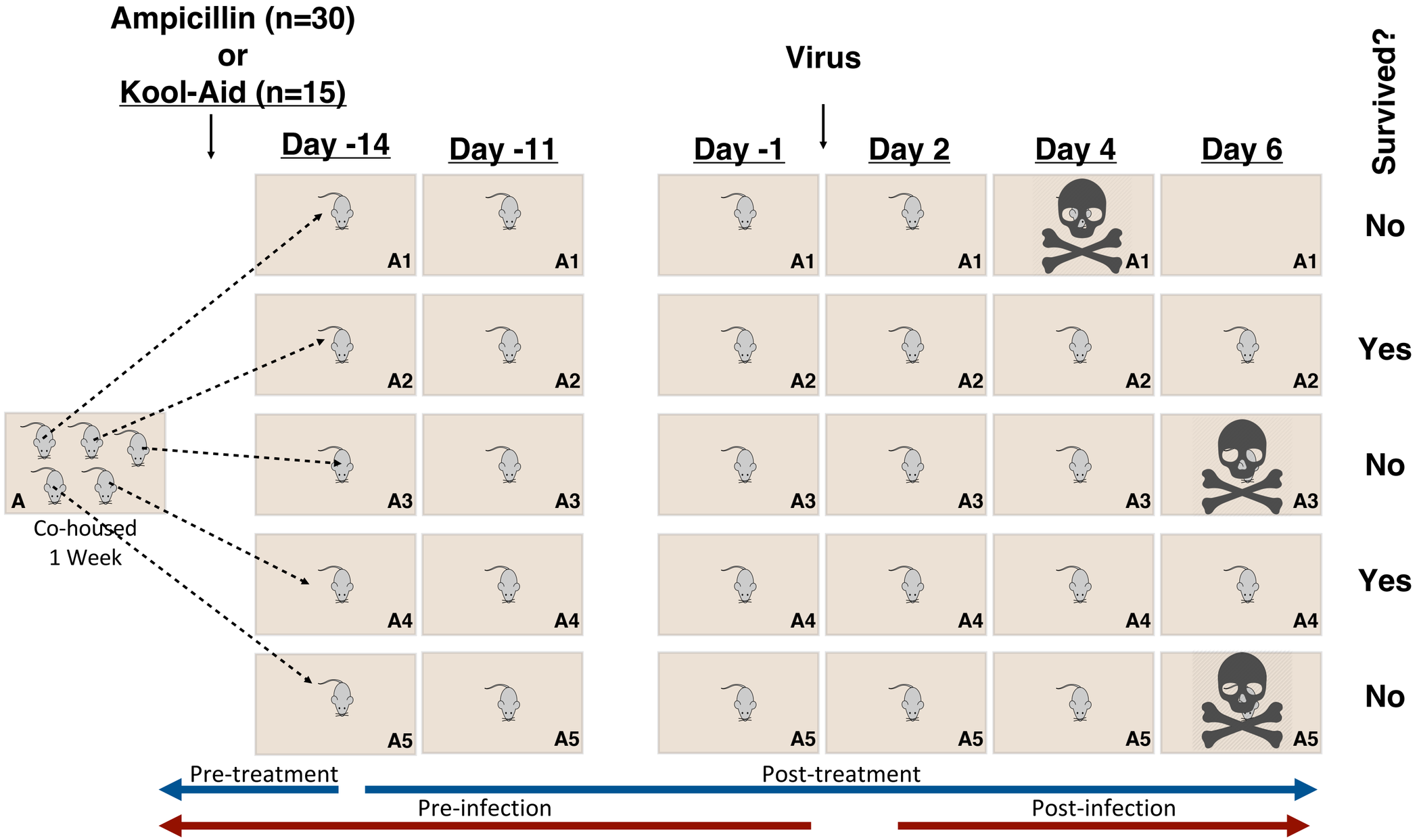


Vancomycin

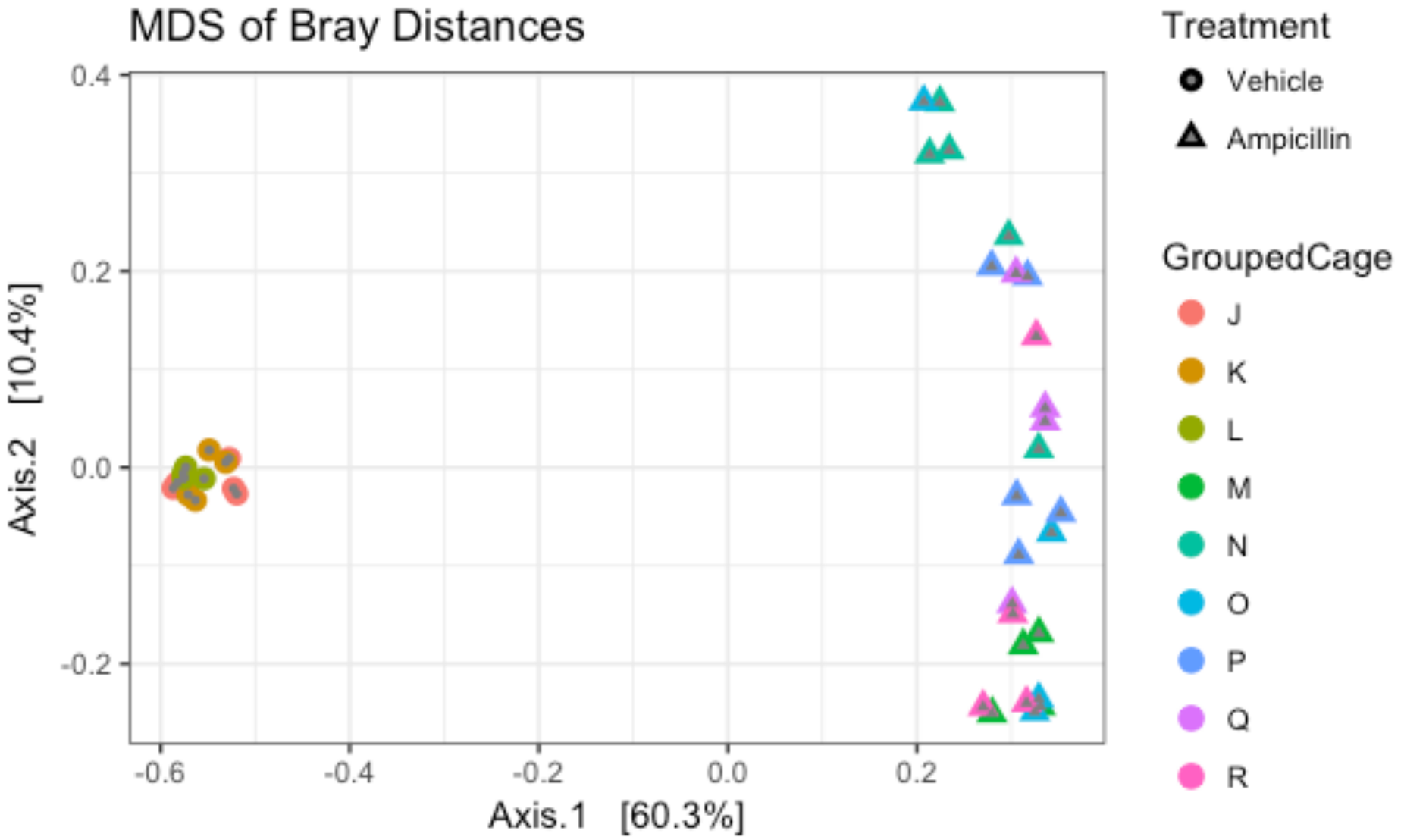




Individual Mouse Isolation Schema



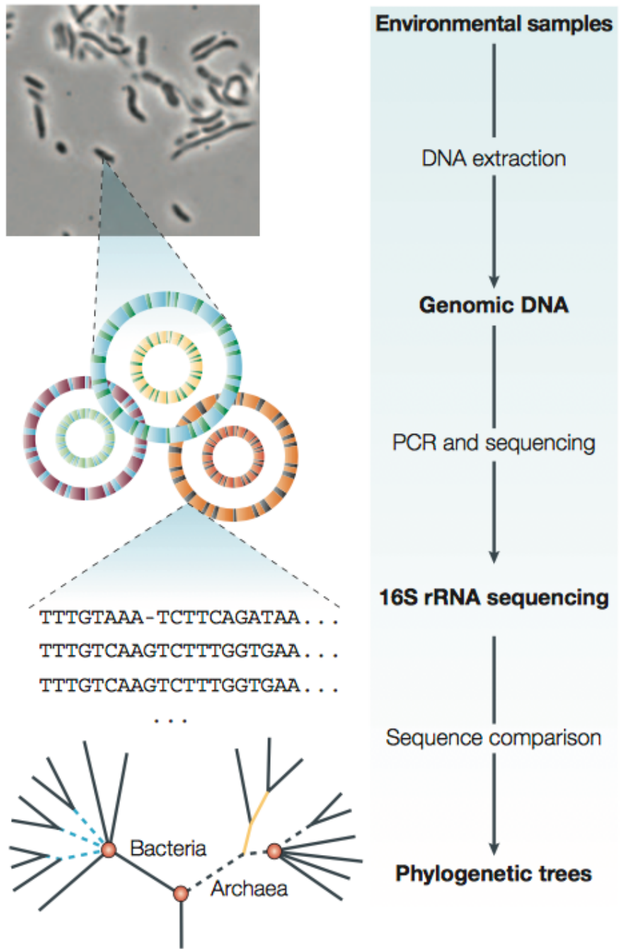
Individual Housing Results



Amplicon Surveys (Highly Opinionated!) Best-practices

It’s the classic garbage in, garbage out all over again …

16S rRNA Amplicon Survey



Study Design

Laboratory

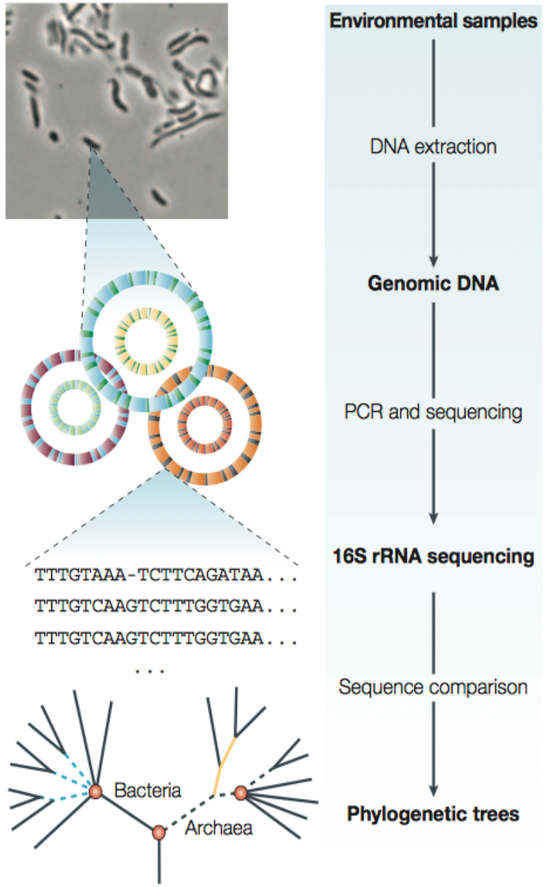
Bioinformatics, Ecological Analysis and Statistics

Tringe, S.G., Rubin, E.M. Nat Rev Genet. 2005 Nov;6(11):805-14

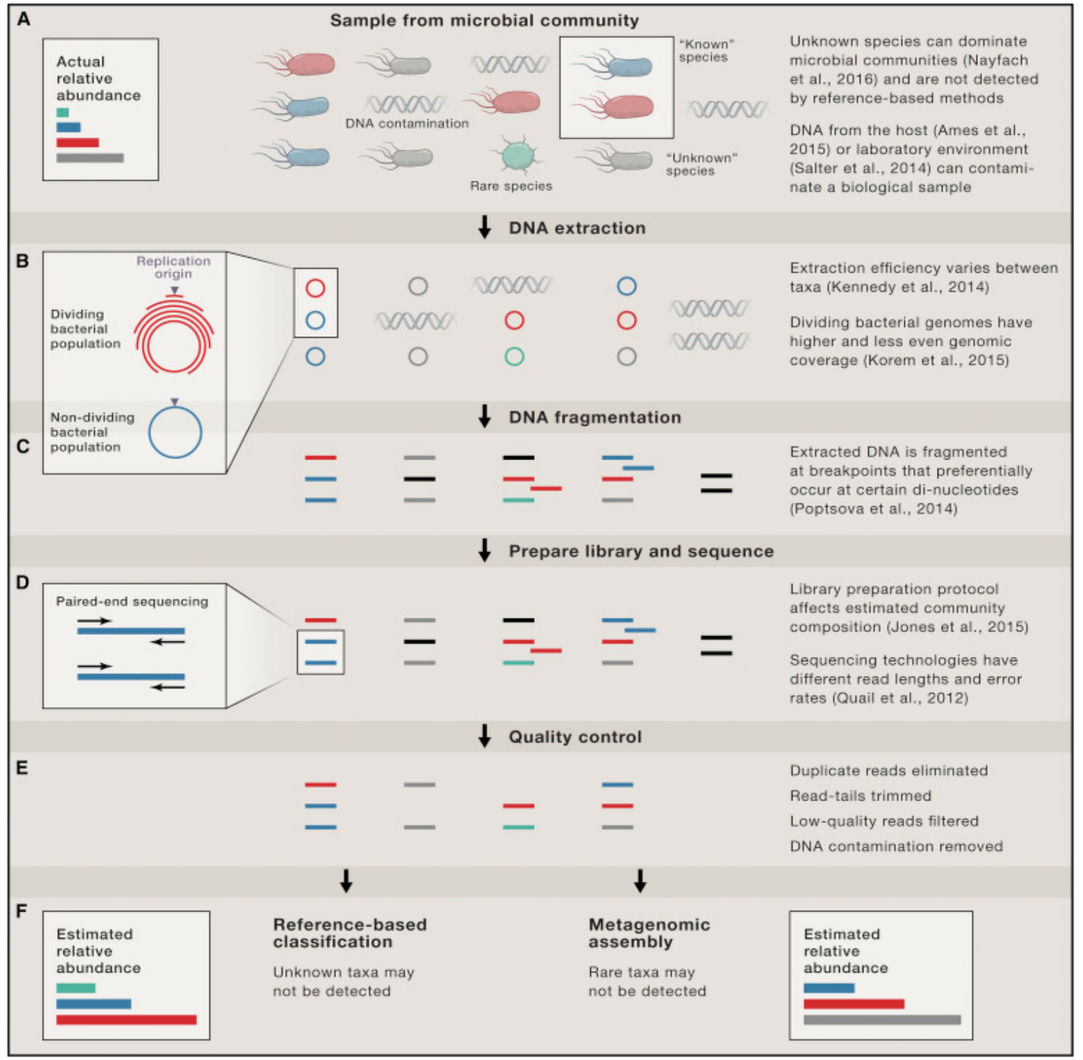
Side note: Amplicon Surveys vs. Metagenomics

Please hold your throwing tomatoes …

16S Amplicon Surveys vs Metagenomics?



Tringe, S.G., Rubin, E.M. Nat Rev Genet. 2005 Nov;6(11):805-14



Nayfach S., Pollard KS. Cell. Aug 25;166(5):1103-16

Most of Your Decision Will Boil Down to $$$

- Our labs per sample costs:
  - 16S = $17.50 per sample
  - Metagenome = $225.00 per sample
    - Has been estimated to be as low as $100 per sample
- Study we will discuss today: 270 samples
  - $4,725 vs. $27 - $60,750
- Other considerations:
  - Understanding analytical space
  - Data storage

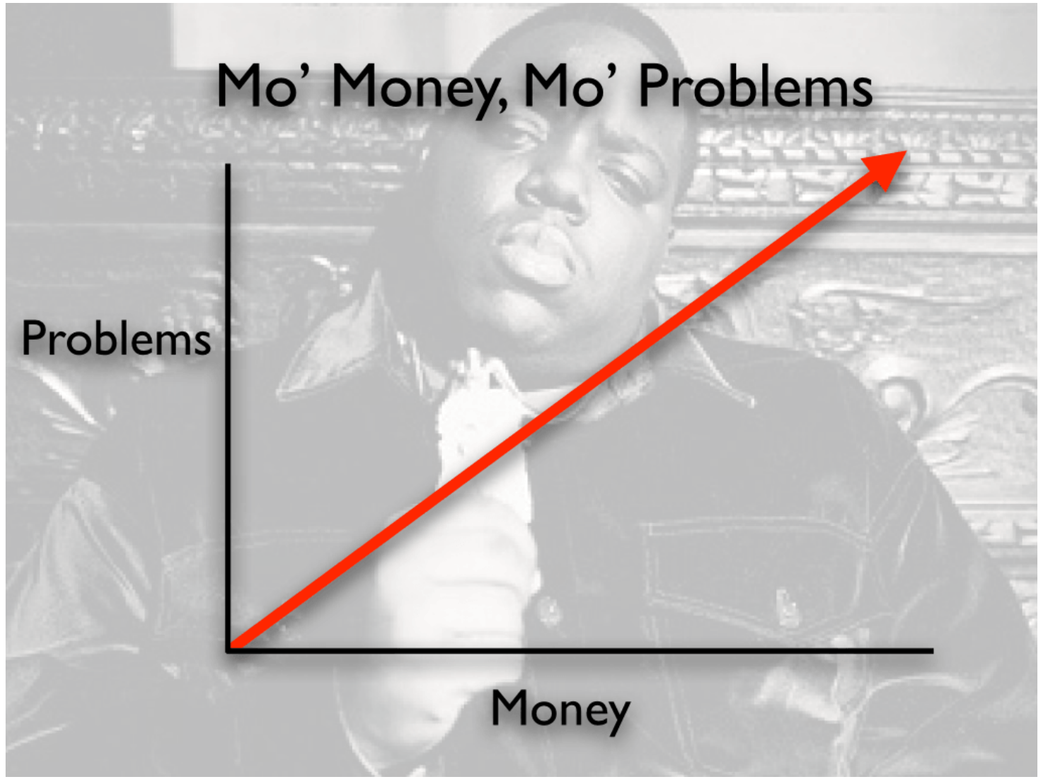
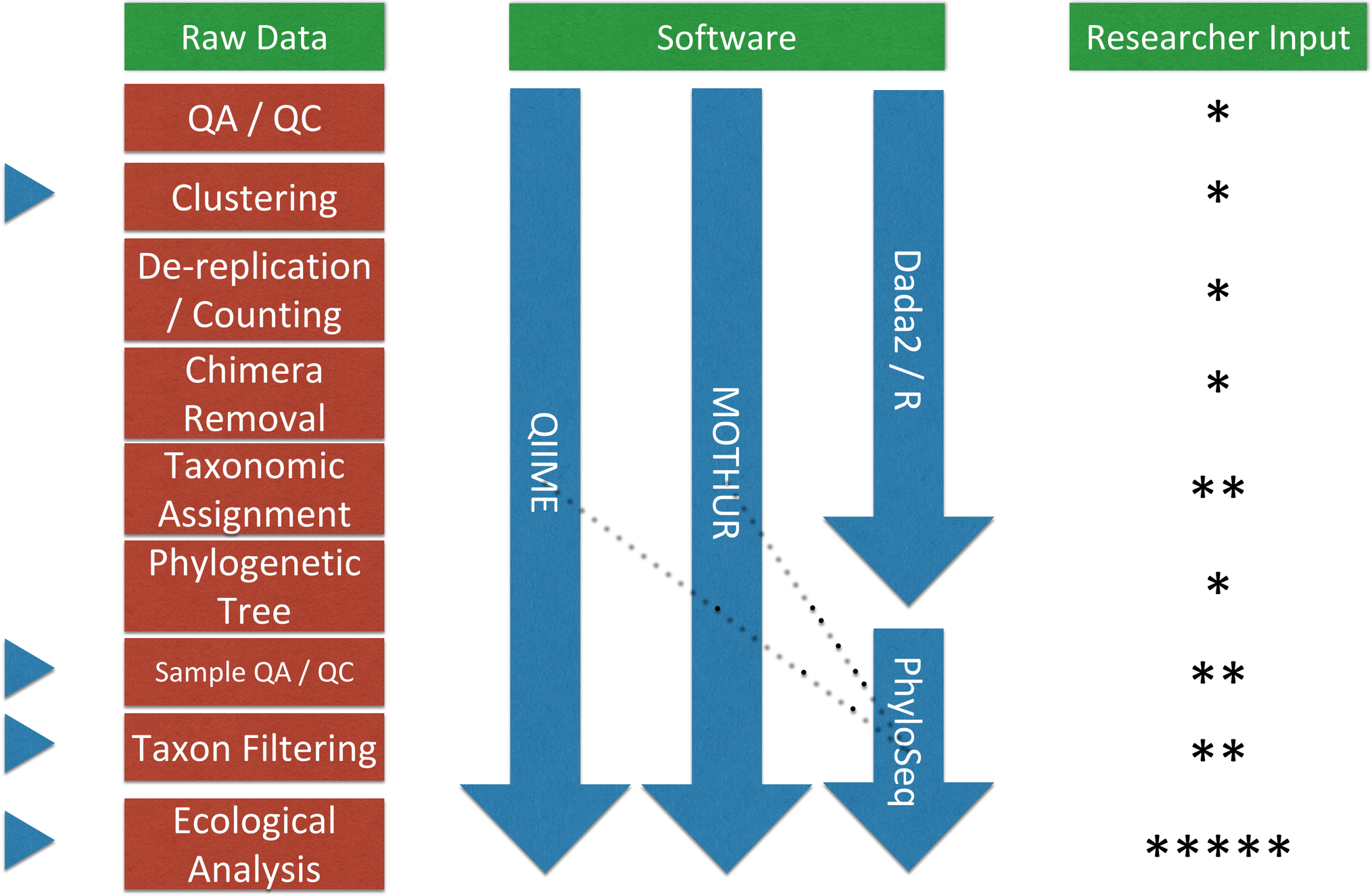


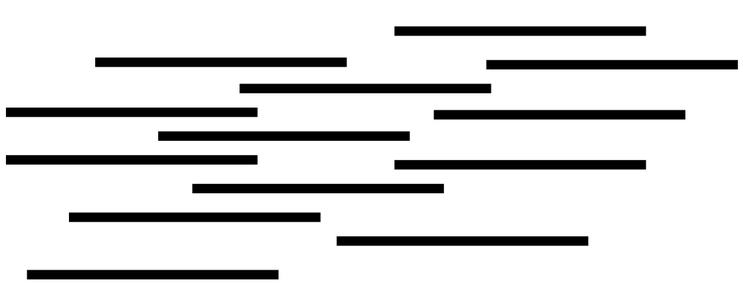
Image credit: The Internet
Quote credit: Notorious B.I.G.

What are the stages of a 16S amplicon computational workflow and how can we create optimal data for analysis?

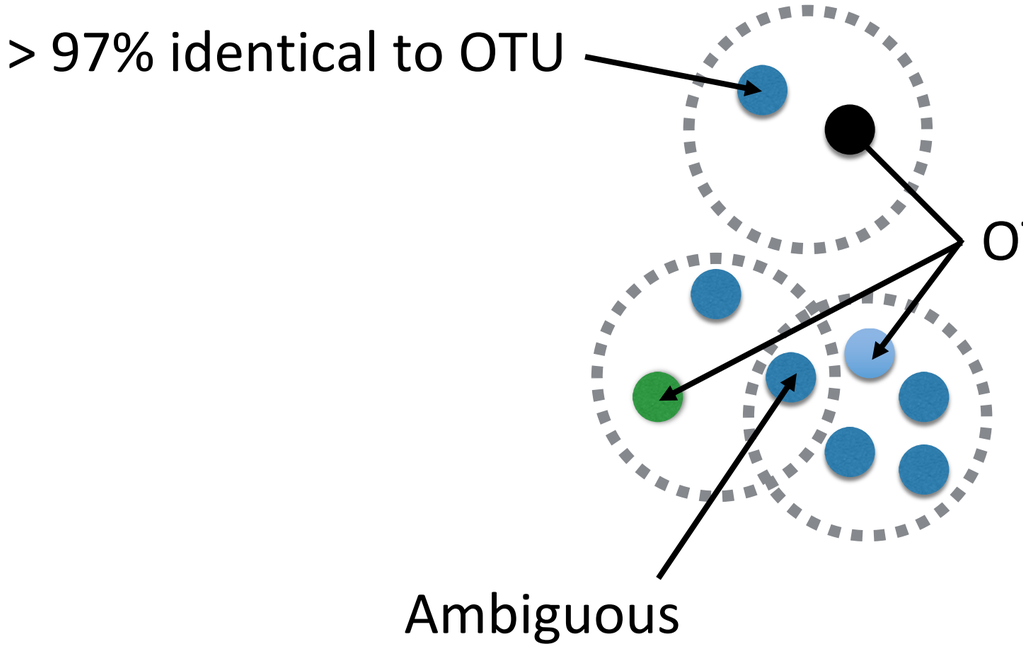

Raw Data
Software
Researcher Input
QA / QC
Clustering
De-replication / Counting
Chimera Removal
Taxonomic Assignment
Phylogenetic Tree
Sample QA / QC
Taxon Filtering
Ecological Analysis
QIIME
MOTHUR
Dada2 / R
PhyloSeq
*
*
*
*
**
*
**
**
*****

Sequence Clustering

16S RNA Amplicons



Amplicon Clusters

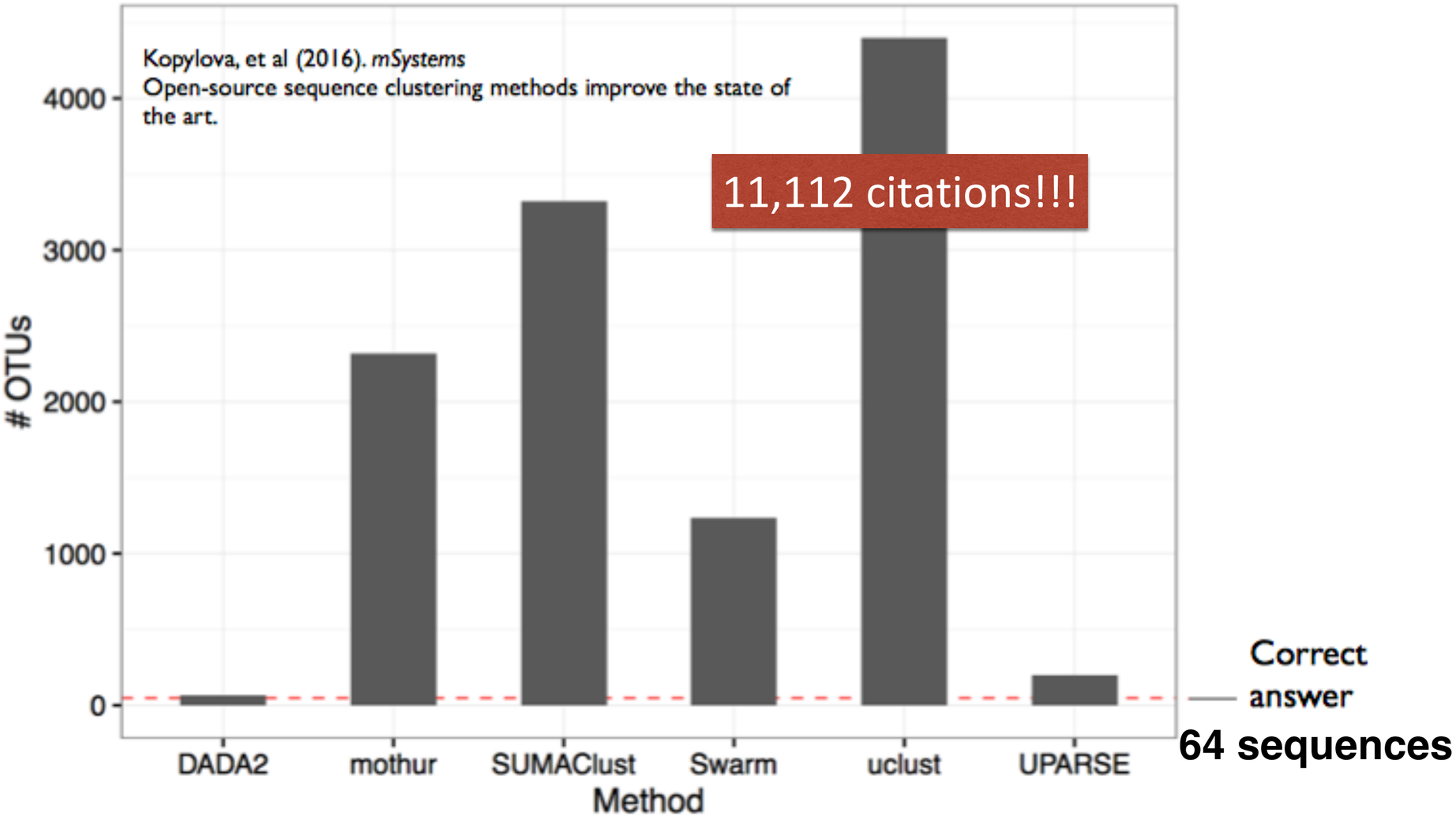


- **UCLUST**
- **UPARSE**
- **SWARM**
- **SUMACLUST**
- ***OTHERS***

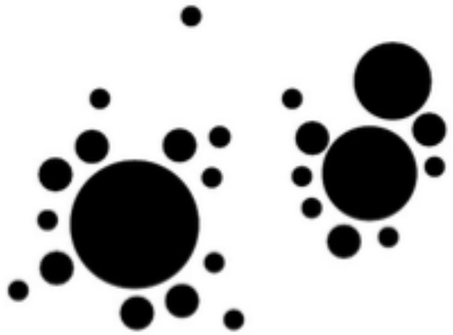
Recognized Problems with Sequence Clustering

- **False-positives:** 1,000s of OTUs when only 10s of sequences are present
  - Due to clustering artifact / noisy sequences
    - Inflates richness (# of species)
    - Sparse matrices
- **Poor taxonomic resolution** defined by arbitrary radius (e.g. 97%)
- **Increased financial cost:** poor data efficiency
- **Increased computational cost:** Clustering is quadratic
- **Unstable:** Sequence and count frequently depend on input order

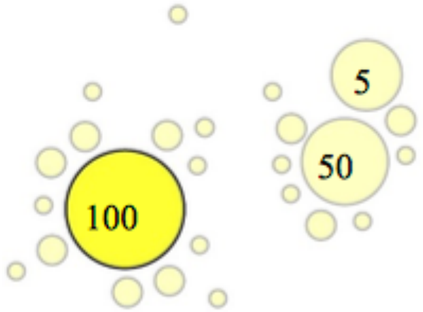
There is some hope



http://benjjneb.github.io/dada2/R/SotA.html



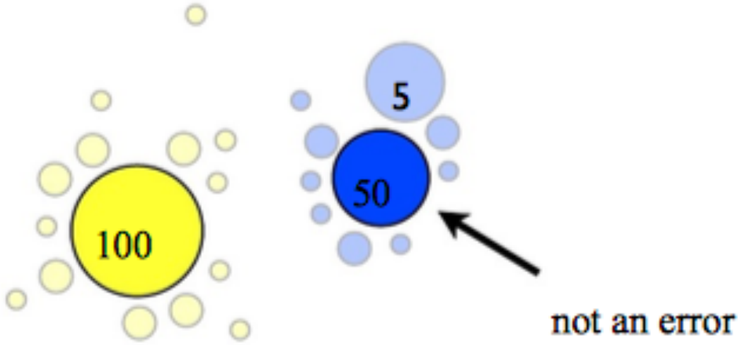
Step 1: Initial guess.
All sequences + errors



Step 2: Initial *error model*

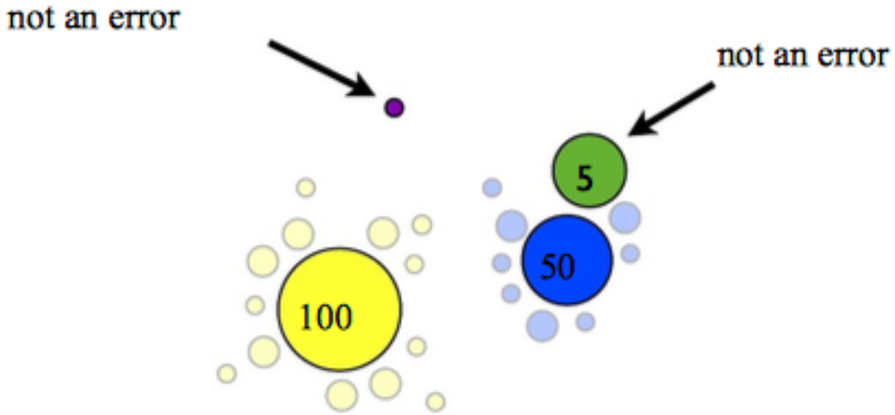
$$\Pr(i \to j) =$$

	A	C	G	T
A	0.97	10_{-2}	10_{-2}	10_{-2}
C	10_{-2}	0.97	10_{-2}	10_{-2}
G	10_{-2}	10_{-2}	0.97	10_{-2}
T	10_{-2}	10_{-2}	10_{-2}	0.97

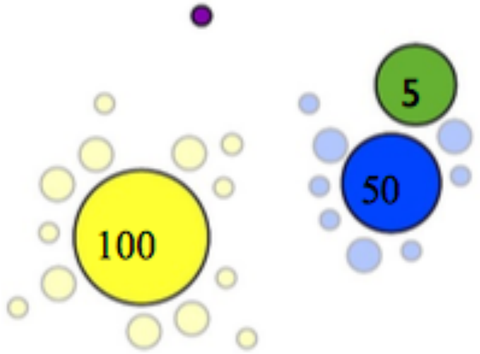


Step 3: Unlikely error under model.
Recruit errors. Update the model

	A	C	G	T
A	0.97	10_{-2}	10_{-2}	10_{-2}
C	10_{-2}	0.97	10_{-2}	10_{-2}
G	10_{-2}	10_{-2}	0.97	10_{-2}
T	10_{-2}	10_{-2}	10_{-2}	0.97



Step 3: Reject more sequences
under new model & update



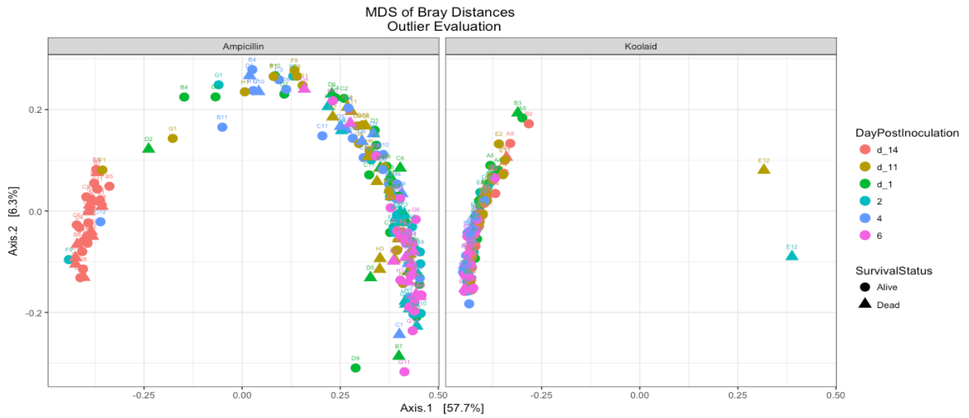
Convergence: All errors are plausible

Dada2: Callahan, BJ et al. Nat Methods. 2016

What does all of this work get you?

- Raw Data
- QA / QC
- Clustering
- De-replication / Counting
- Chimera Removal
- Taxonomic Assignment
- Phylogenetic Tree
- Sample QA / QC
- Taxon Filtering
- Ecological Analysis

ID	Sample 1	Sample 2	Sample 3	Sample 4
ASV 1	0	0	2	0
ASV 2	12	8	8	456
ASV 3	112	101	98	10
ASV 4	435	435	382	3
ASV 5	76	83	68	145



Sparse Matrix
OTU Clustering

ID	Sample 1	Sample 2	Sample 3
OTU 1	0	0	1
OTU 2	1	0	0
OTU 3	1	0	0
OTU 4	1	1	1

Less Sparse Matrix
Sequence Resolution

ID	Sample 1	Sample 2	Sample 3
ASV 1	0	1	1
ASV 2	1	1	0
ASV 3	1	0	1
ASV 4	1	1	1

- More noisy than reality
- Bad for statistical inference
  - Multiple hypothesis testing
  - Poorly defined, difficult to separate distributions

Raw Data

QA / QC

Clustering

De-replication / Counting

Chimera Removal

Taxonomic Assignment

Phylogenetic Tree

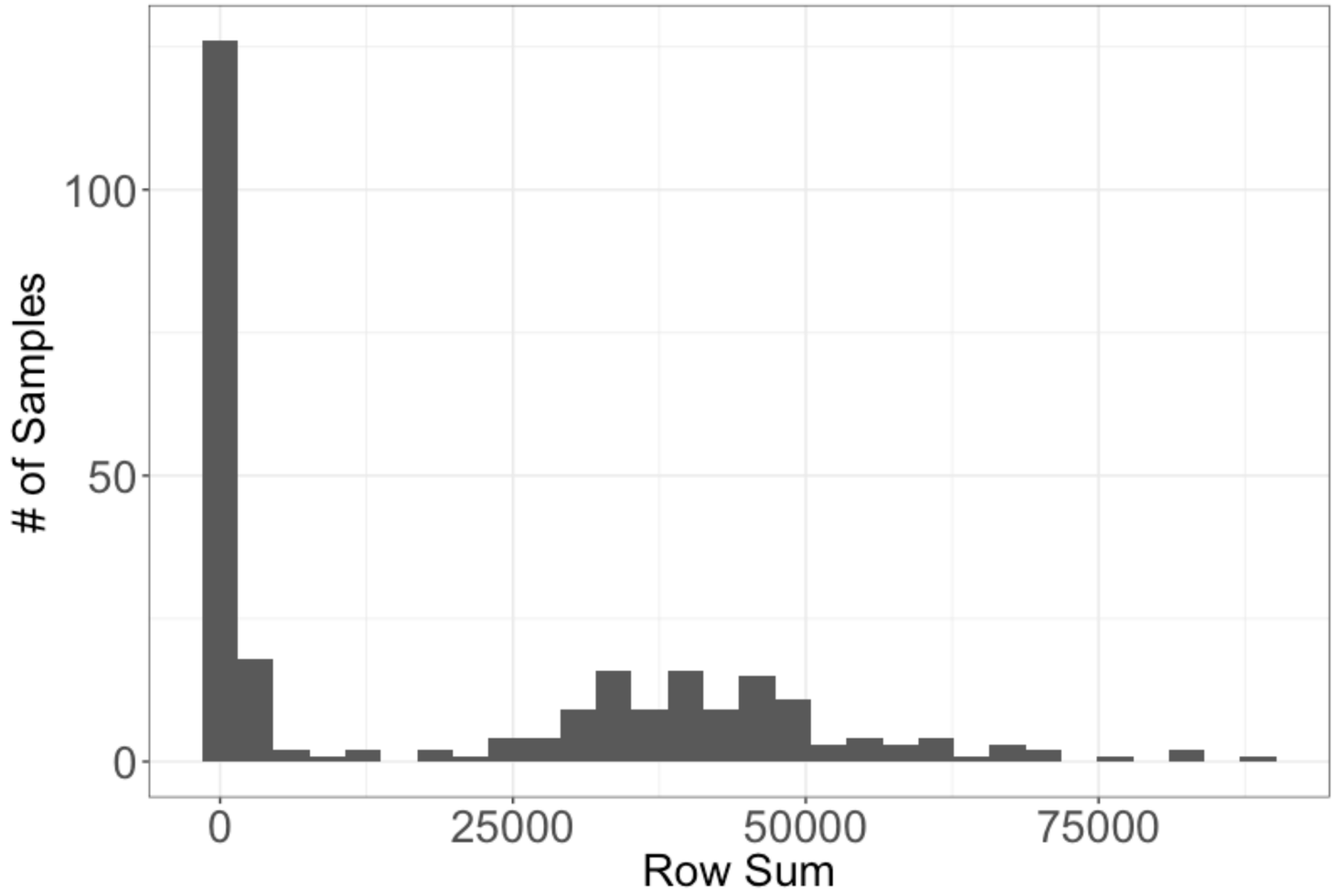
Sample QA / QC

Taxon Filtering

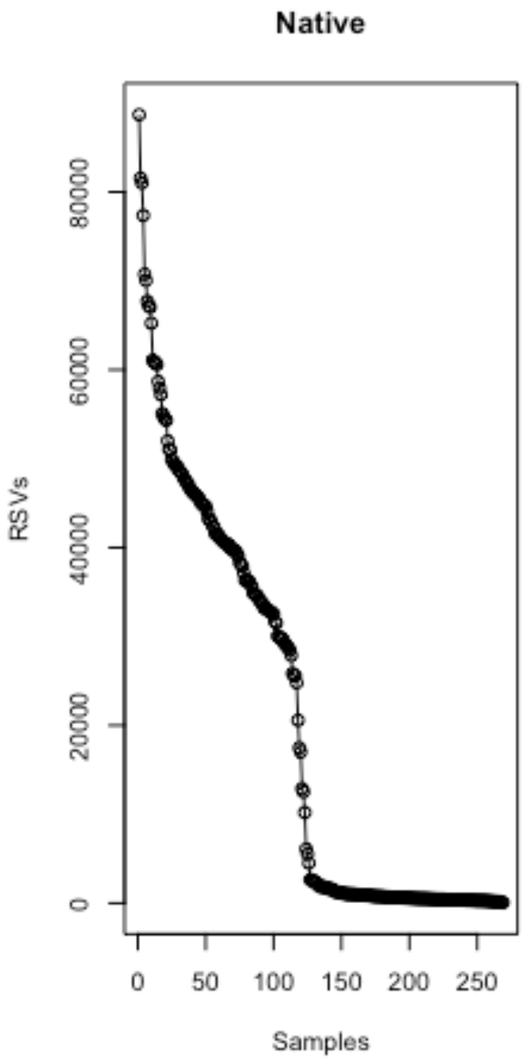
Ecological Analysis

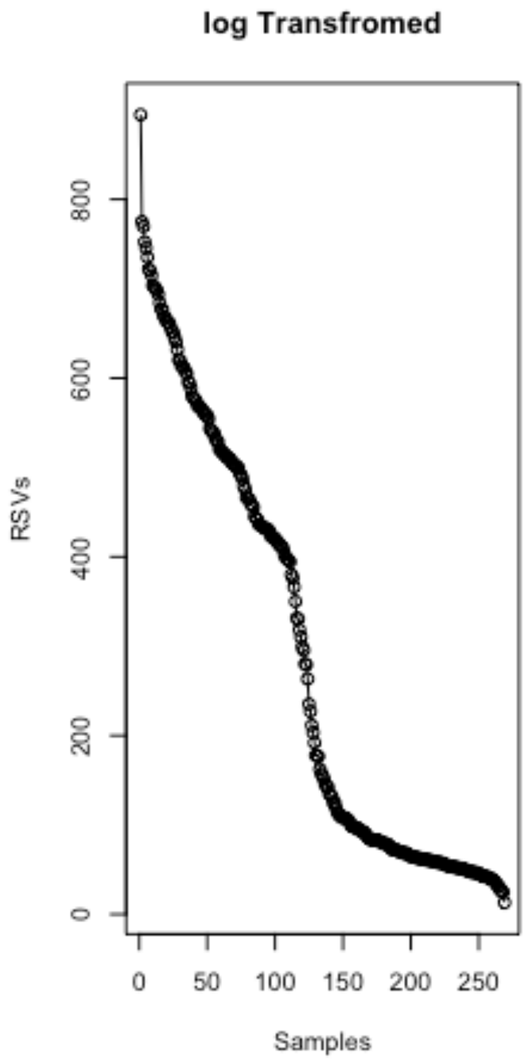
Making Things Normal

Data Transformation

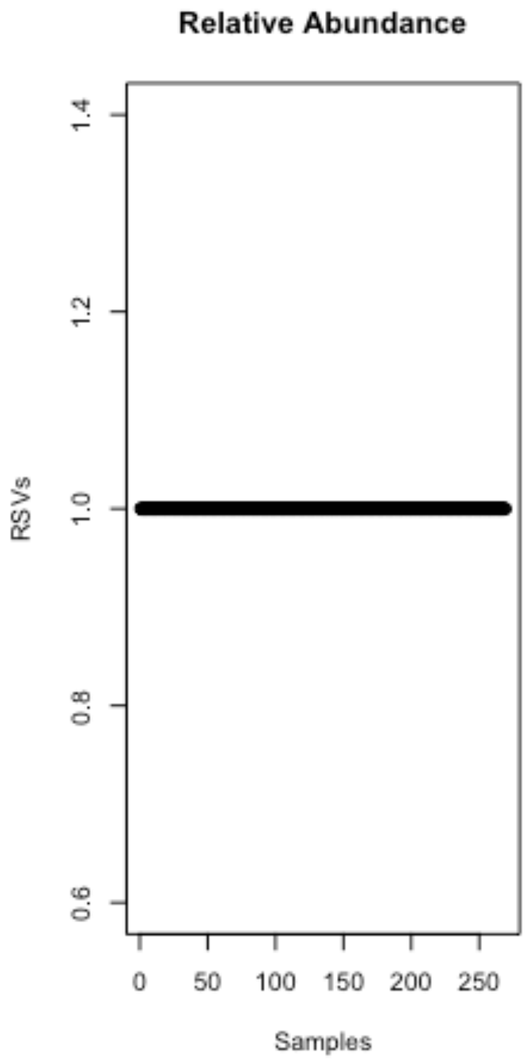


Data Transformation

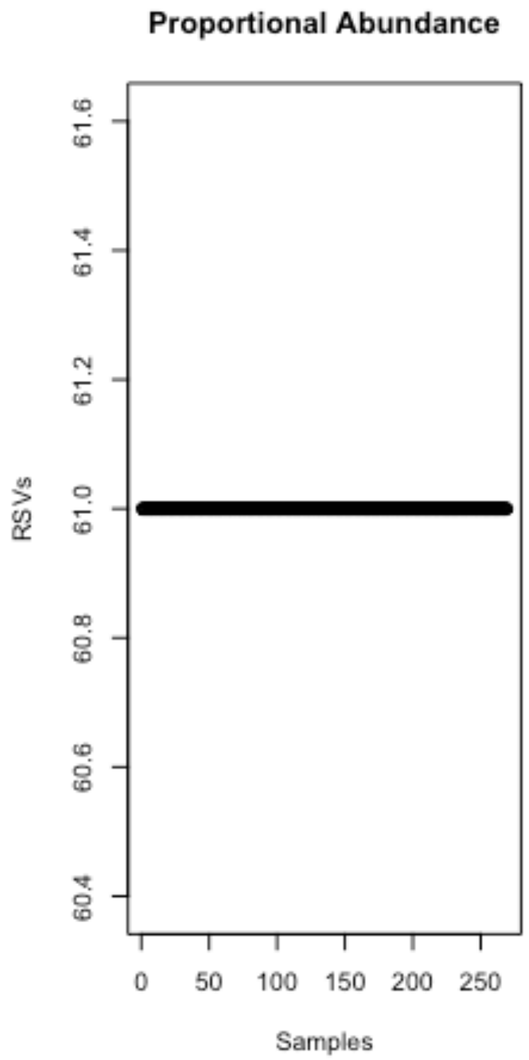




log(1 + x)

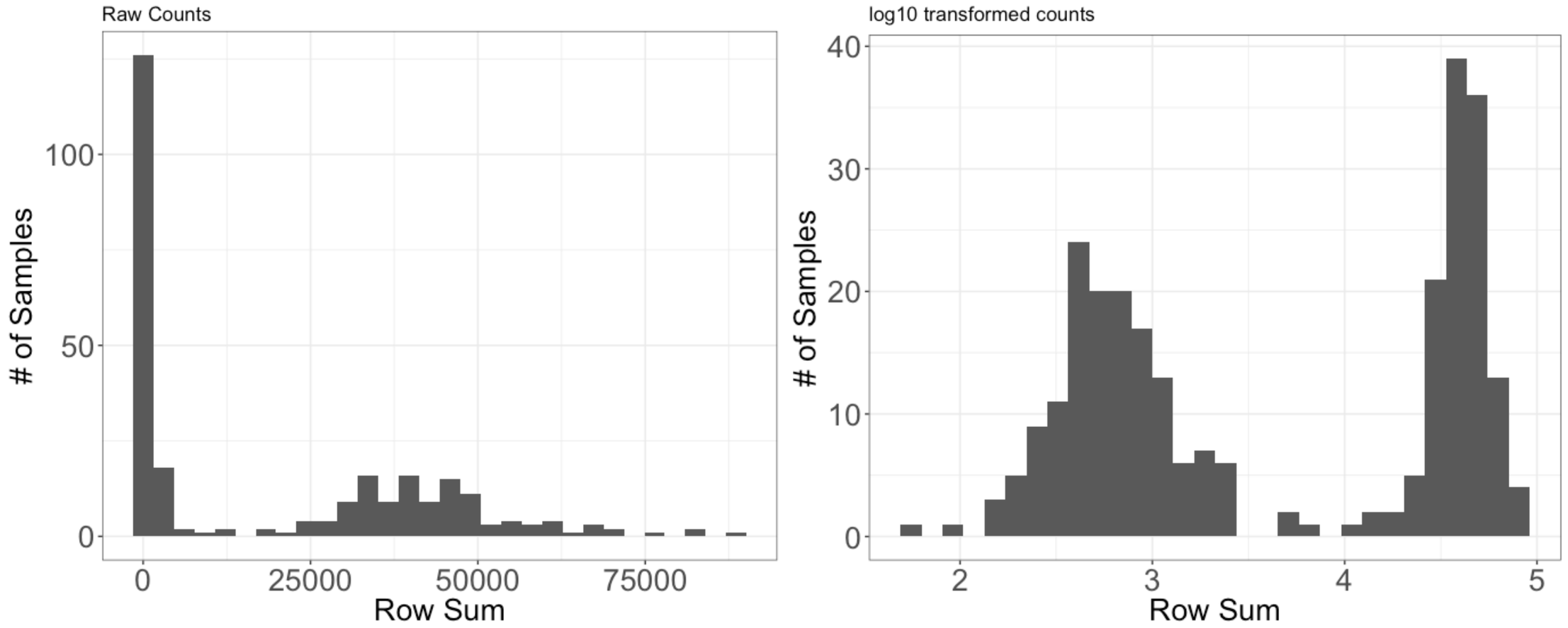


x/sum(x)



min(sample_sum) * x/sum(x)

log Transformation Shifts Towards Normality



Weiss S. et al. Normalization and microbial differential abundance strategies depend on data characteristics. *Microbiome.* 2017

Raw Data

QA / QC

Clustering

De-replication / Counting

Chimera Removal

Taxonomic Assignment

Phylogenetic Tree

Sample QA / QC

Taxon Filtering

Ecological Analysis

Sample Outlier Detection

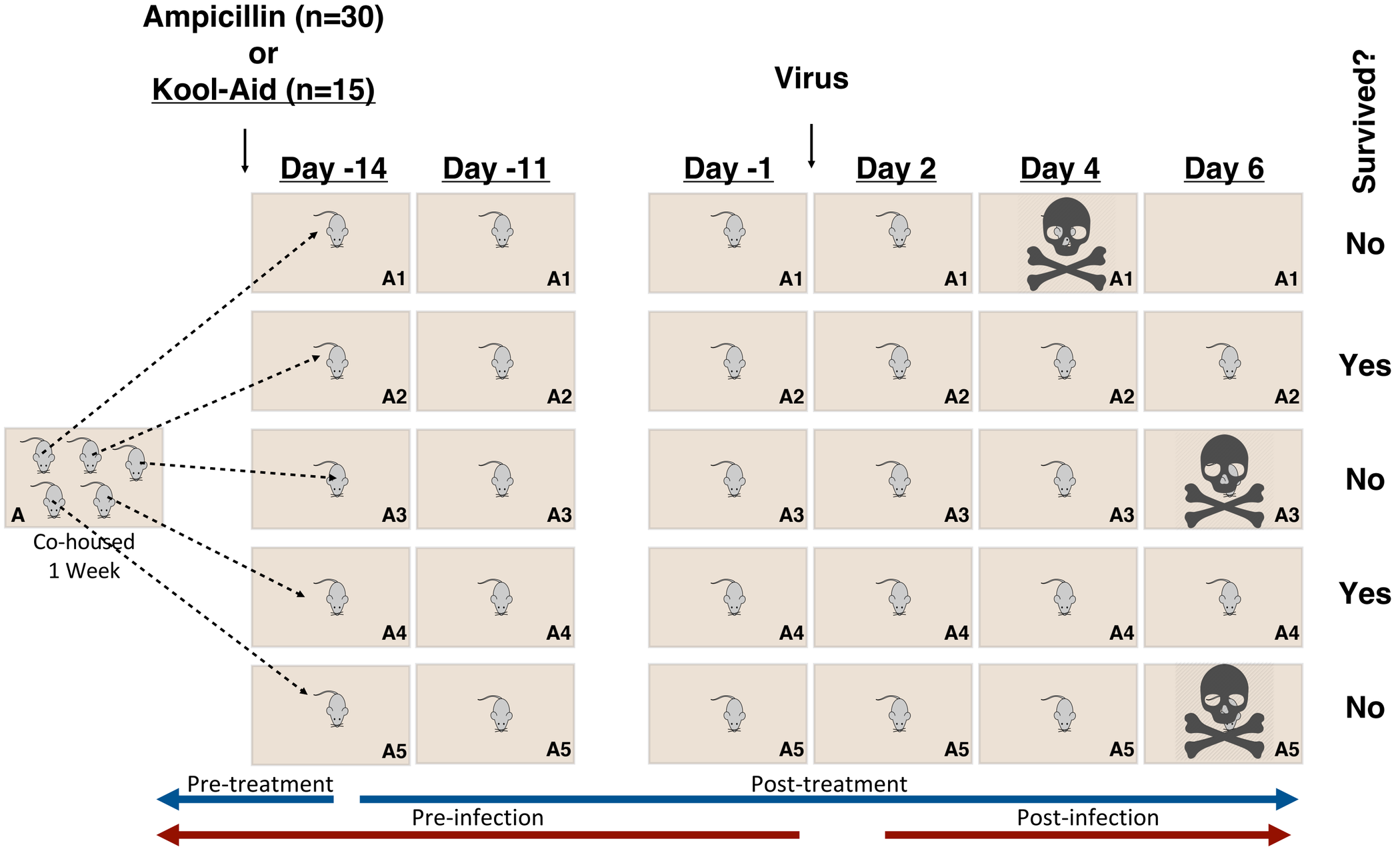
ID	Sample 1	Sample 2	Sample 3	Sample 4
ASV 1	0	0	2	0
ASV 2	12	8	8	456
ASV 3	112	101	98	10
ASV 4	435	435	382	3
ASV 5	76	83	68	145

... n=270

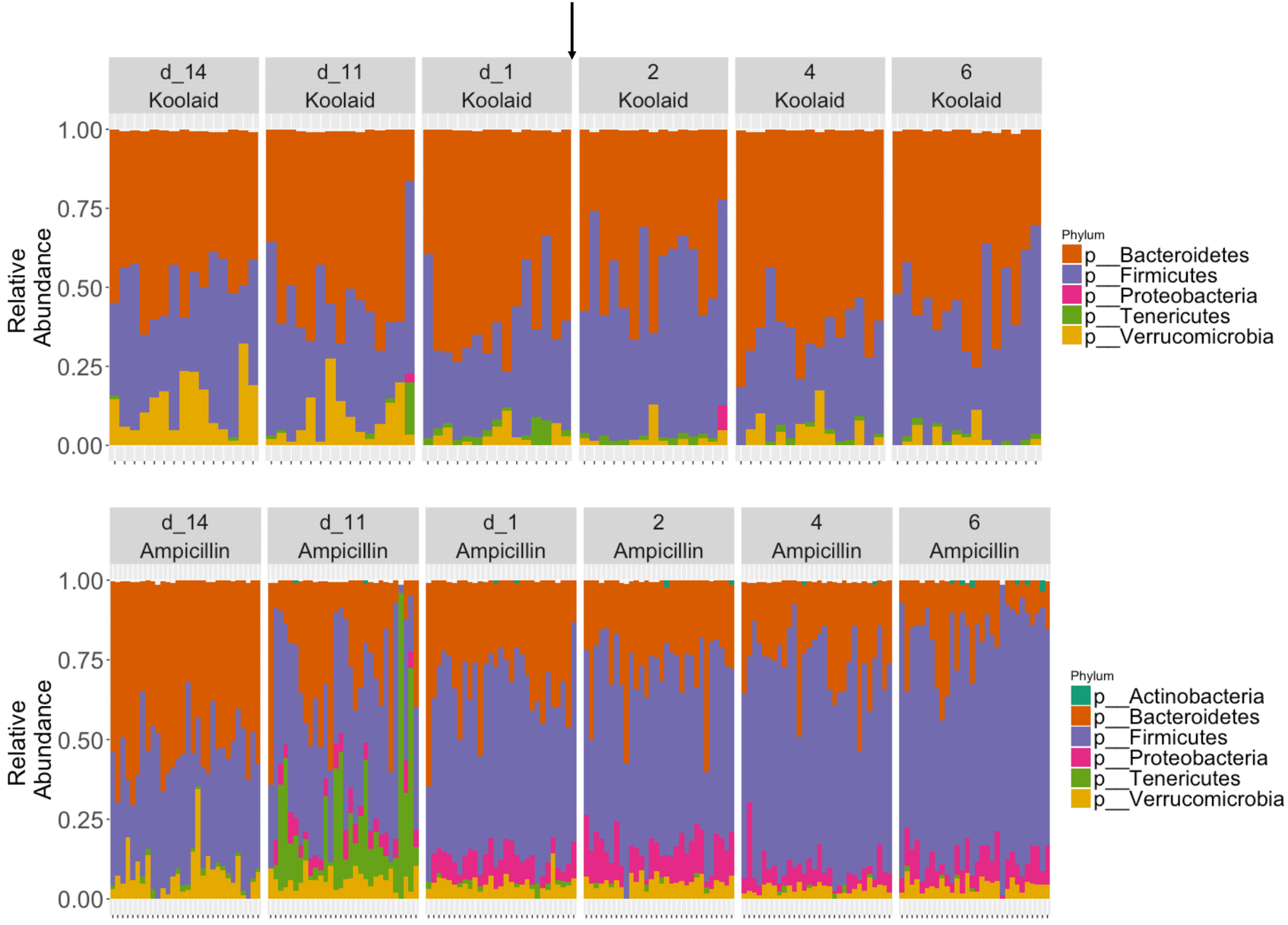
...

n=724

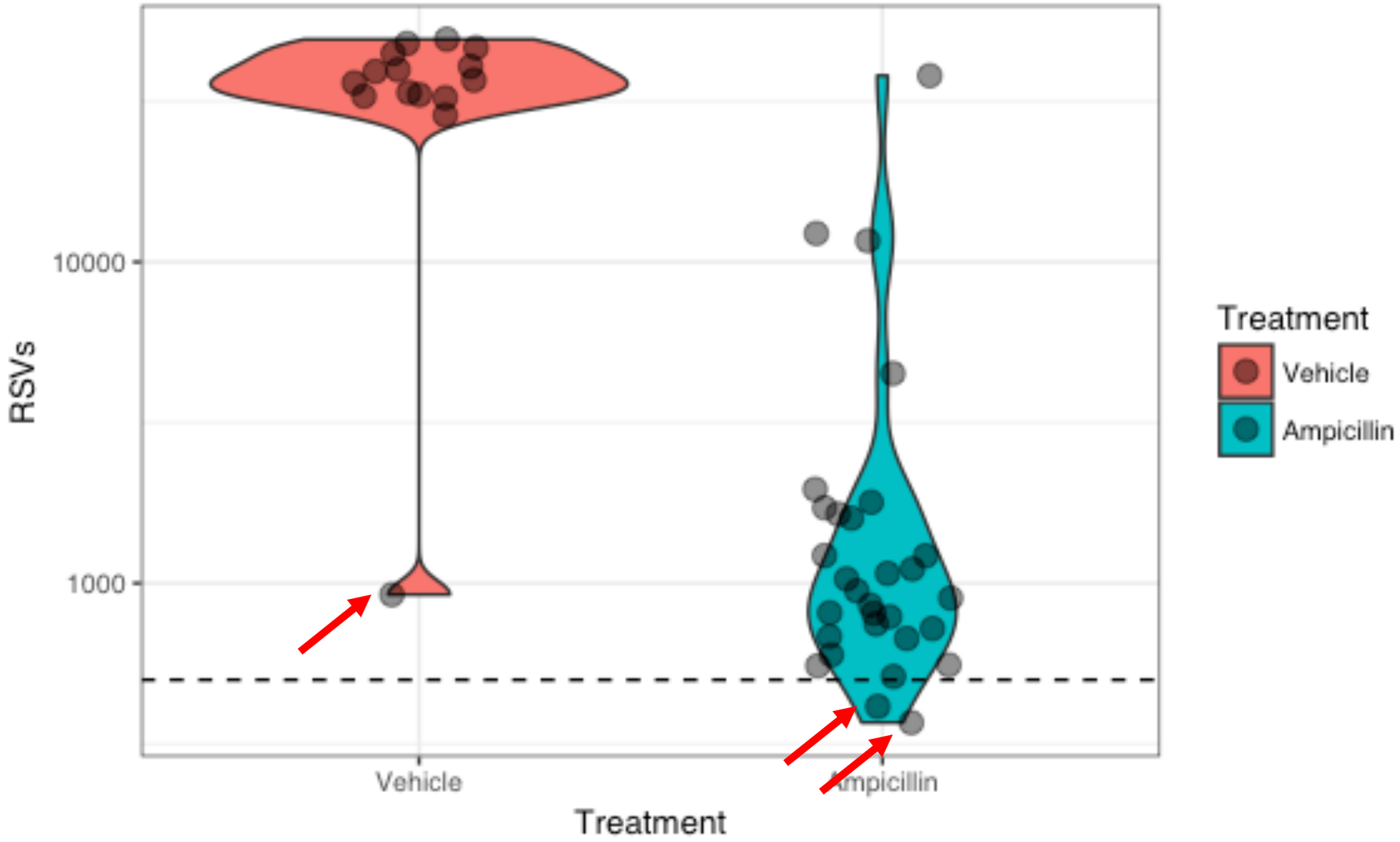
Individual Mouse Isolation Schema



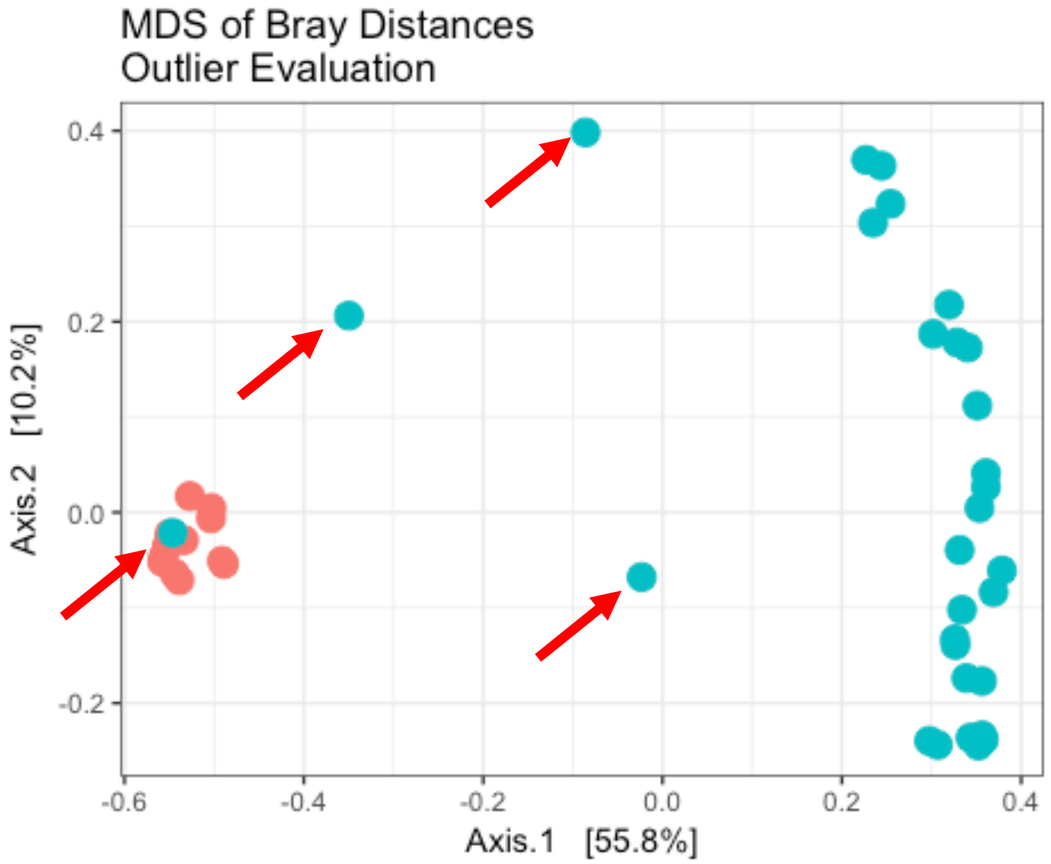
Virus

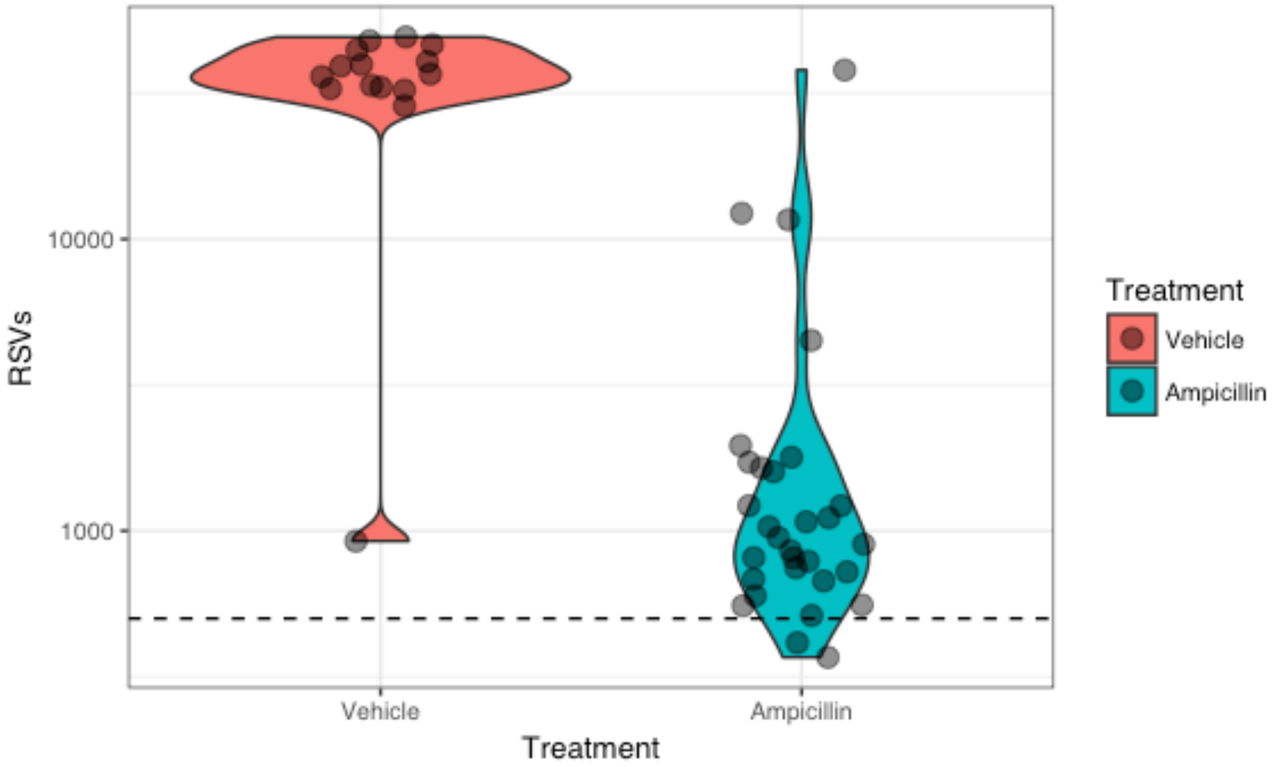

d_14 Koolaid
d_11 Koolaid
d_1 Koolaid
2 Koolaid
4 Koolaid
6 Koolaid
Relative Abundance
1.00
0.75
0.50
0.25
0.00
Phylum
p__Bacteroidetes
p__Firmicutes
p__Proteobacteria
p__Tenericutes
p__Verrucomicrobia
d_14 Ampicillin
d_11 Ampicillin
d_1 Ampicillin
2 Ampicillin
4 Ampicillin
6 Ampicillin
Relative Abundance
1.00
0.75
0.50
0.25
0.00
Phylum
p__Actinobacteria
p__Bacteroidetes
p__Firmicutes
p__Proteobacteria
p__Tenericutes
p__Verrucomicrobia

Sample Outlier Detection – Unexpectedly Low # of Sequences



Samples that “perform” unexpectedly

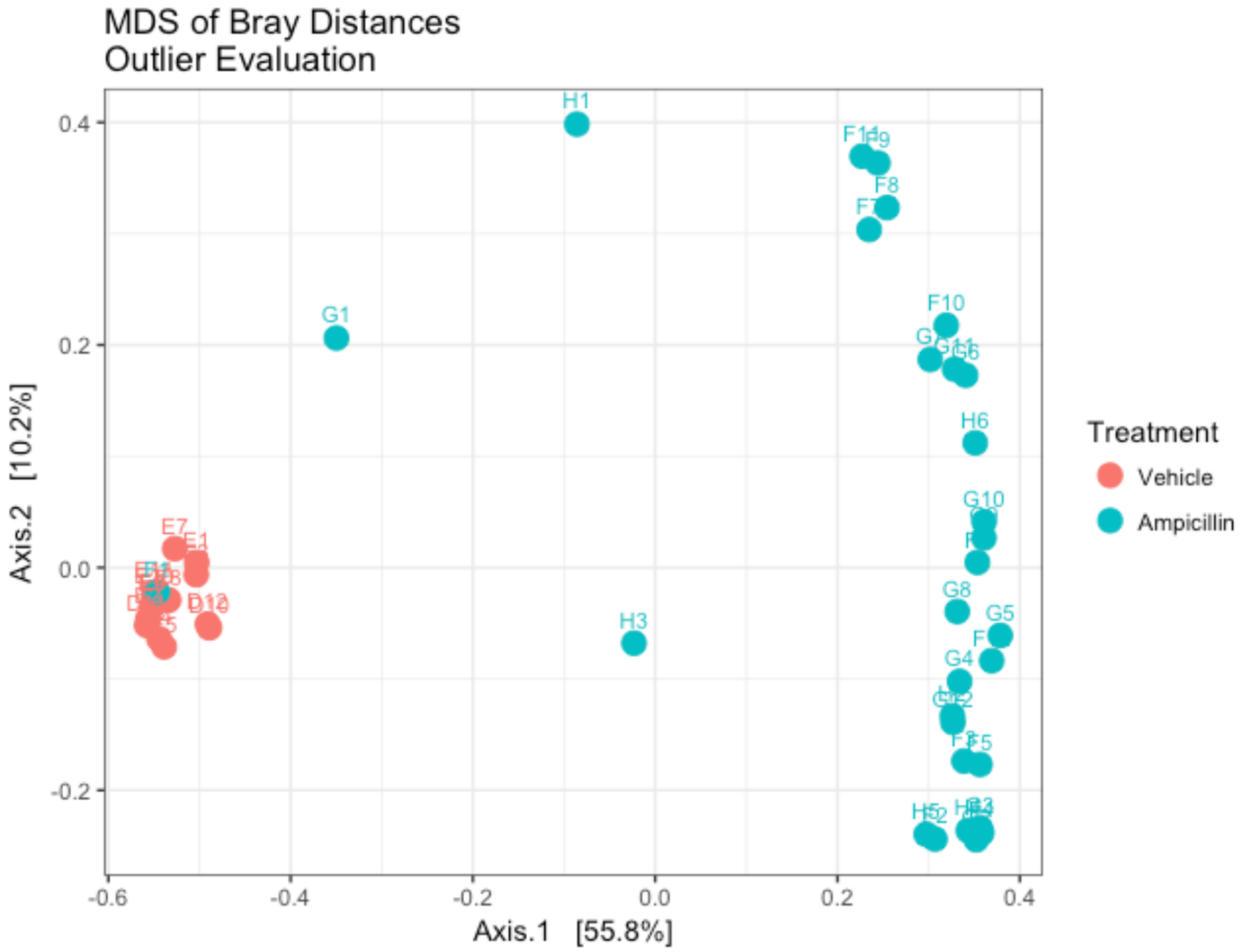


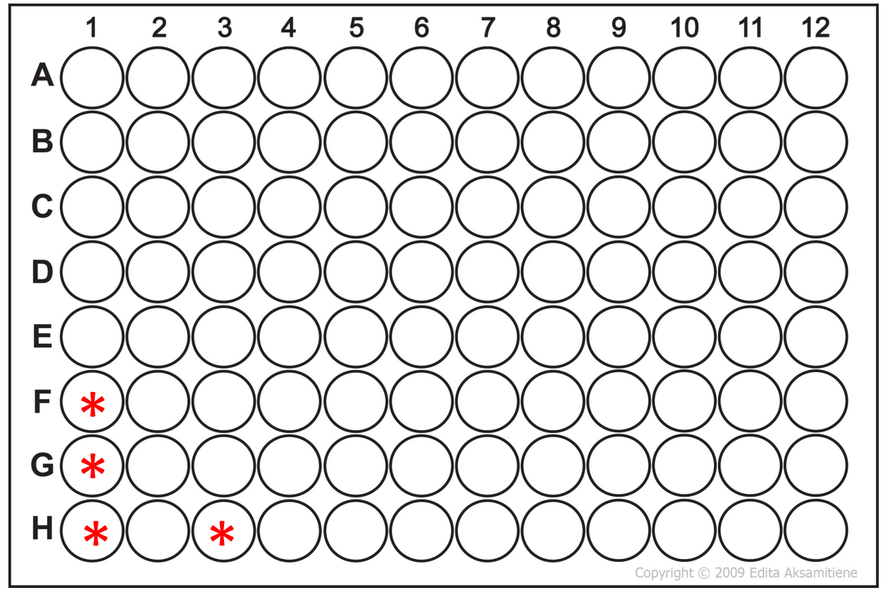


Rules of Thumb for Sample Detection and Removal

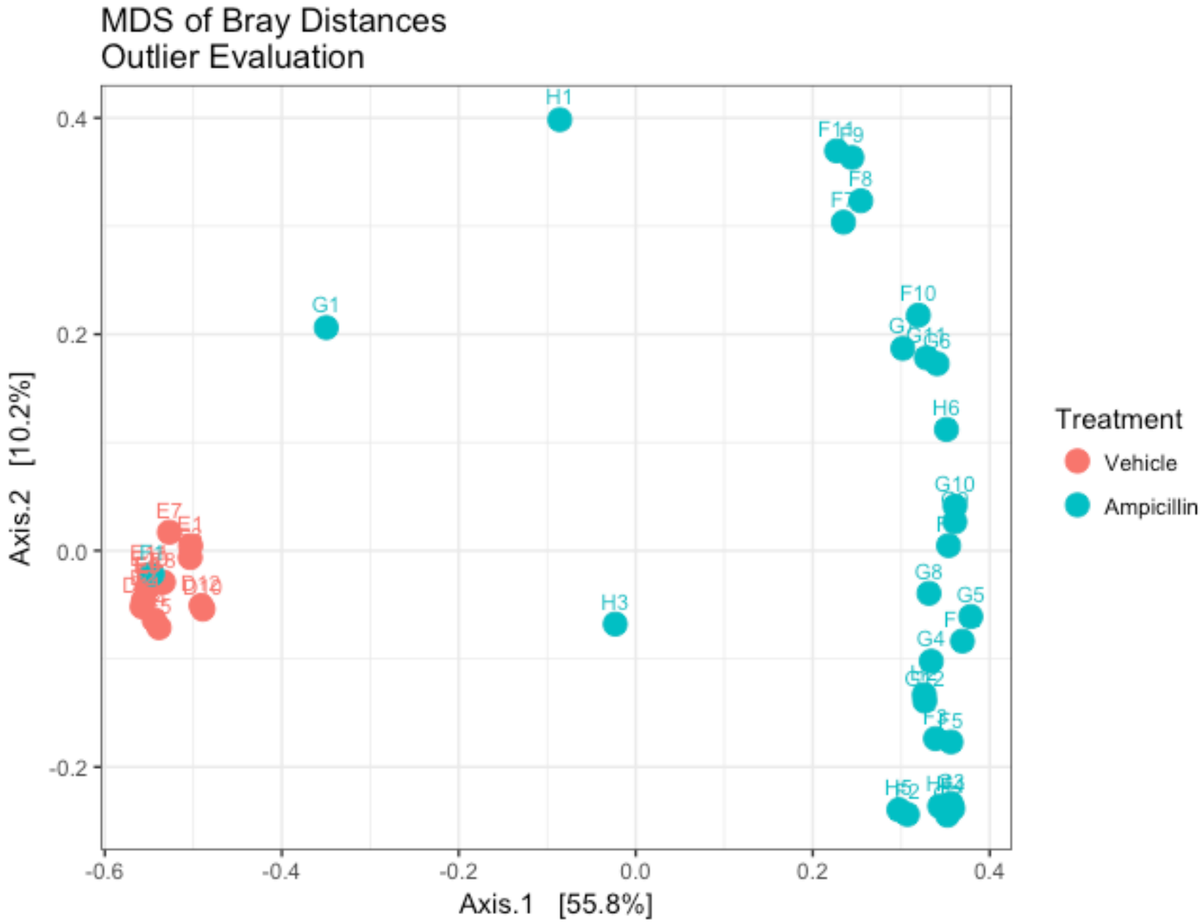
- **Justify and document!!!**
- Except in extreme cases, test how sample removal alters your downstream results. Do the experiment!
- Know your data. When are you comfortable removing a sample based on your knowledge of the system
- Explore using multiple plot types
- Include enough detail to make analysis interpretable and reproducible

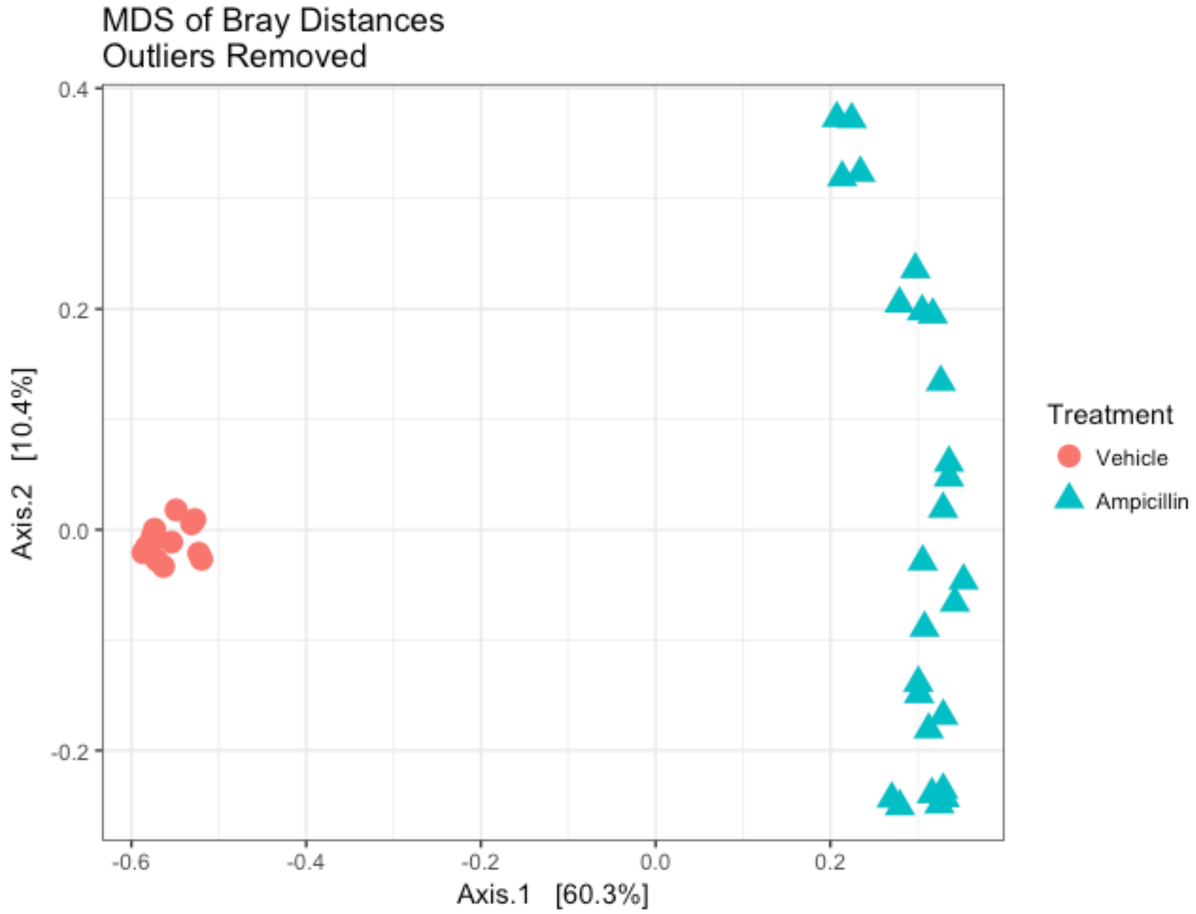
Understand your data better





Cleaned Data

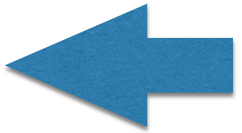




Feature Outlier Detection

ID	Sample 1	Sample 2	Sample 3	Sample 4
ASV 1	0	0	2	0
ASV 2	12	8	8	456
ASV 3	112	101	98	10
ASV 4	435	435	382	3
ASV 5	76	83	68	145

… n=270



…

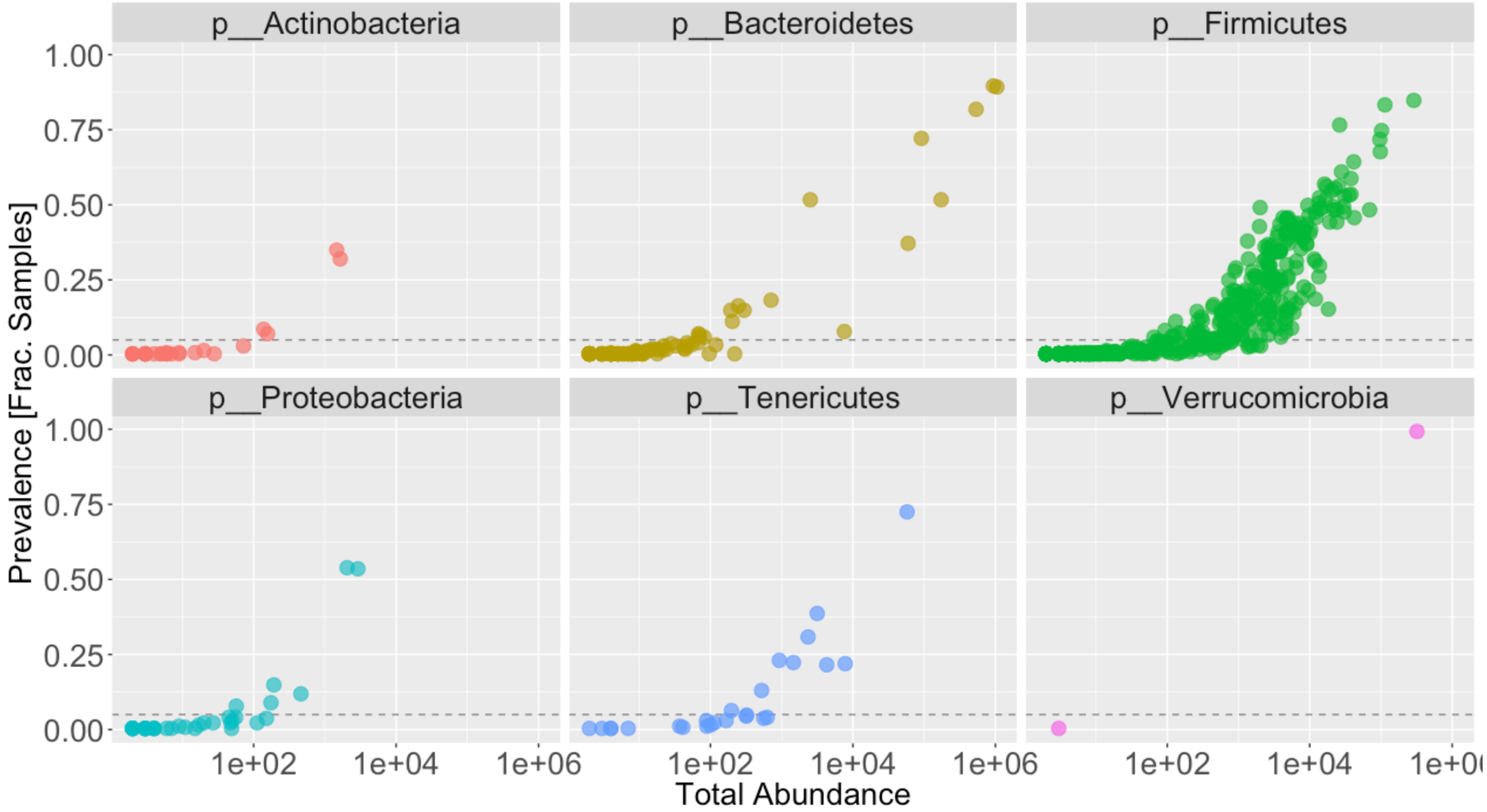
n=724

Low-abundant feature removal is commonplace

- *“We removed all taxa that were under 1% relative abundance and present in less than 3% of all samples.”*

Sequence/Taxa Outlier Detection

Filtering out low impact information

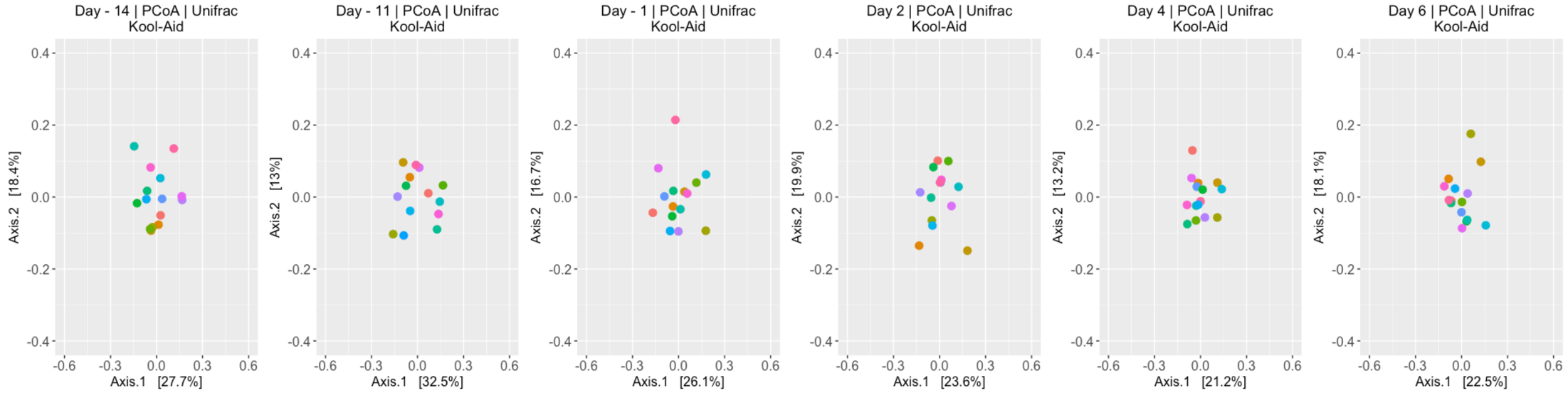


Rules of Thumb for Feature Detection and Removal

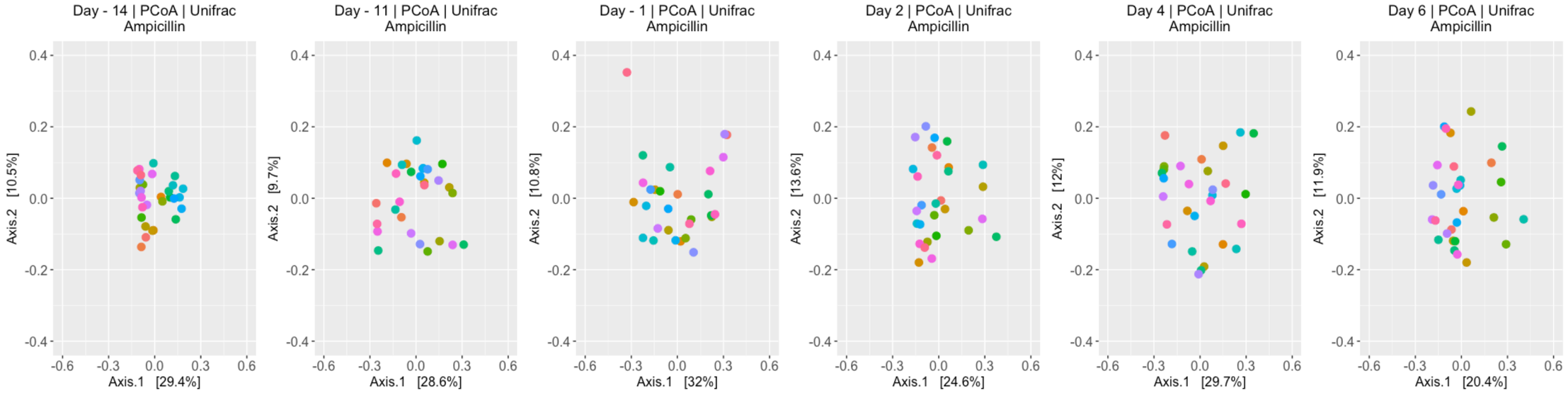
- **Justify and document!!!**
- Except in extreme cases, test how feature removal alters your downstream results. Do the experiment!
- Know your data. When are you comfortable removing a feature based on your knowledge of the system
- Explore using multiple plot types
- Include enough detail to make analysis interpretable and reproducible

Beta Diversity Throughout the Course of the Experiment
Colored by Cage

Kool-Aid



Ampicillin



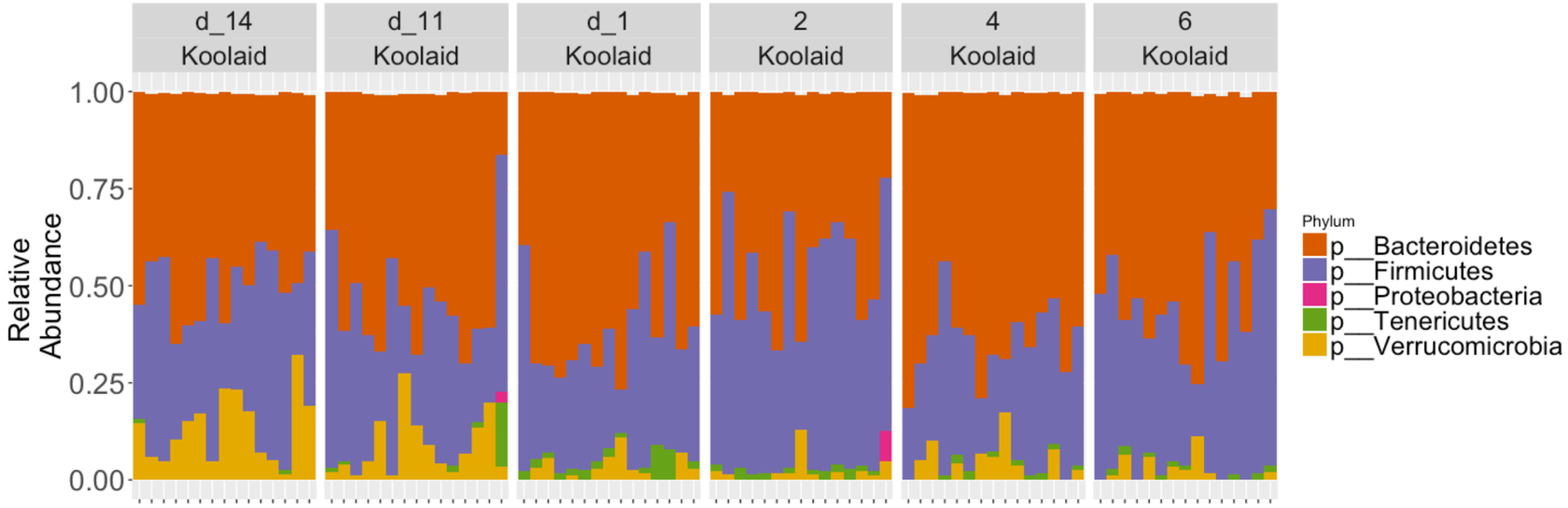
Summary

- Explore -> Document -> Test
- Does any of this really matter?
  - Sometimes?
    - Less so for community ecology measurements
    - More so for detection of differentially abundant taxa
  - Detailed exploration provides more opportunities for insights
  - Don’t publish garbage data

Frequently Used 16S Analysis Techniques

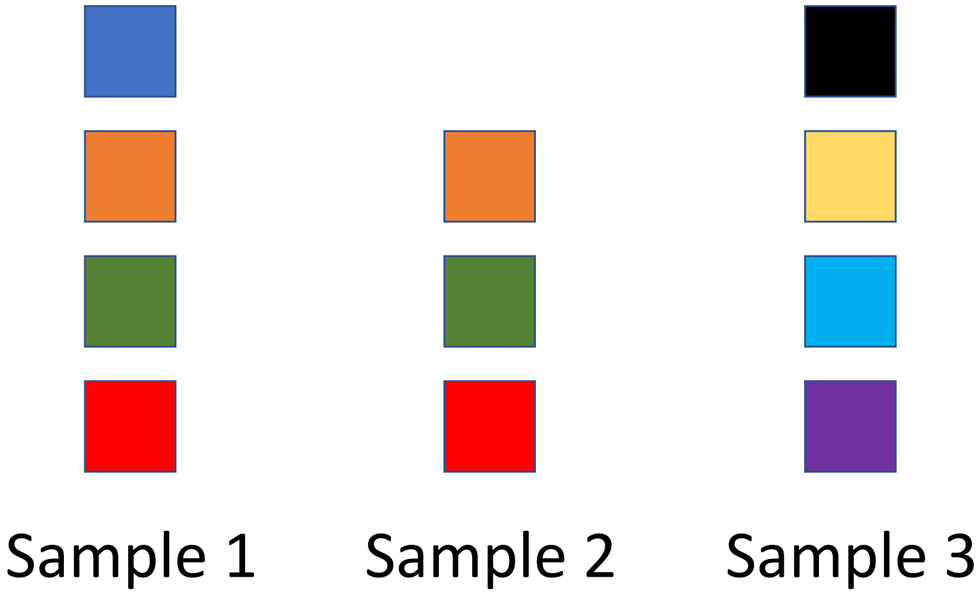
Community Composition

- Broad overview
- Nothing statistical

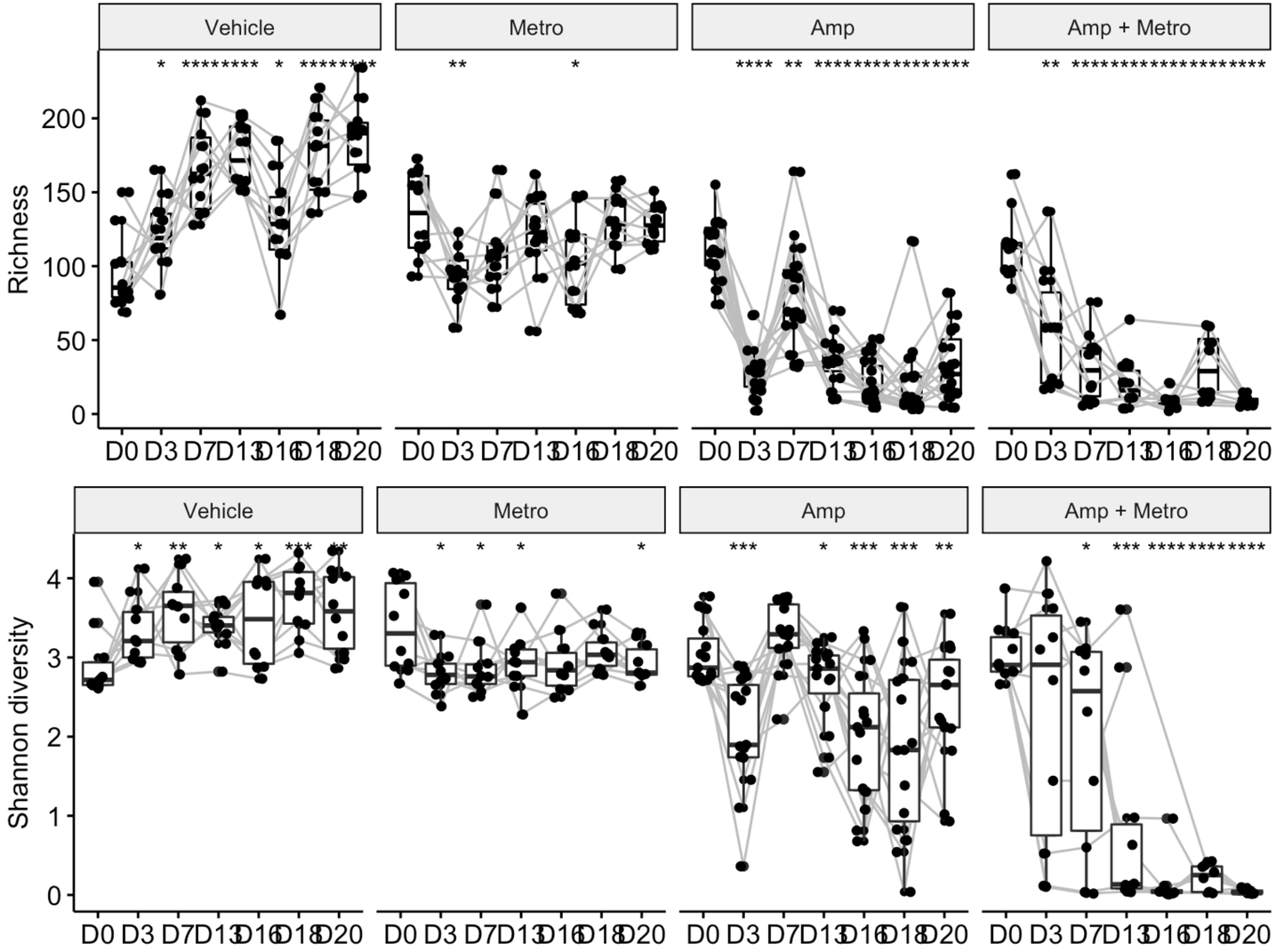


Alpha Diversity: Richness

- Richness: Number of unique taxa (ASVs) that are observed in a sample
  - Taxonomy independent
  - Abundance independent (presence / absence)
- Loads of other Alpha diversity measures (Chao1, Shannon, Simpsons, etc.)

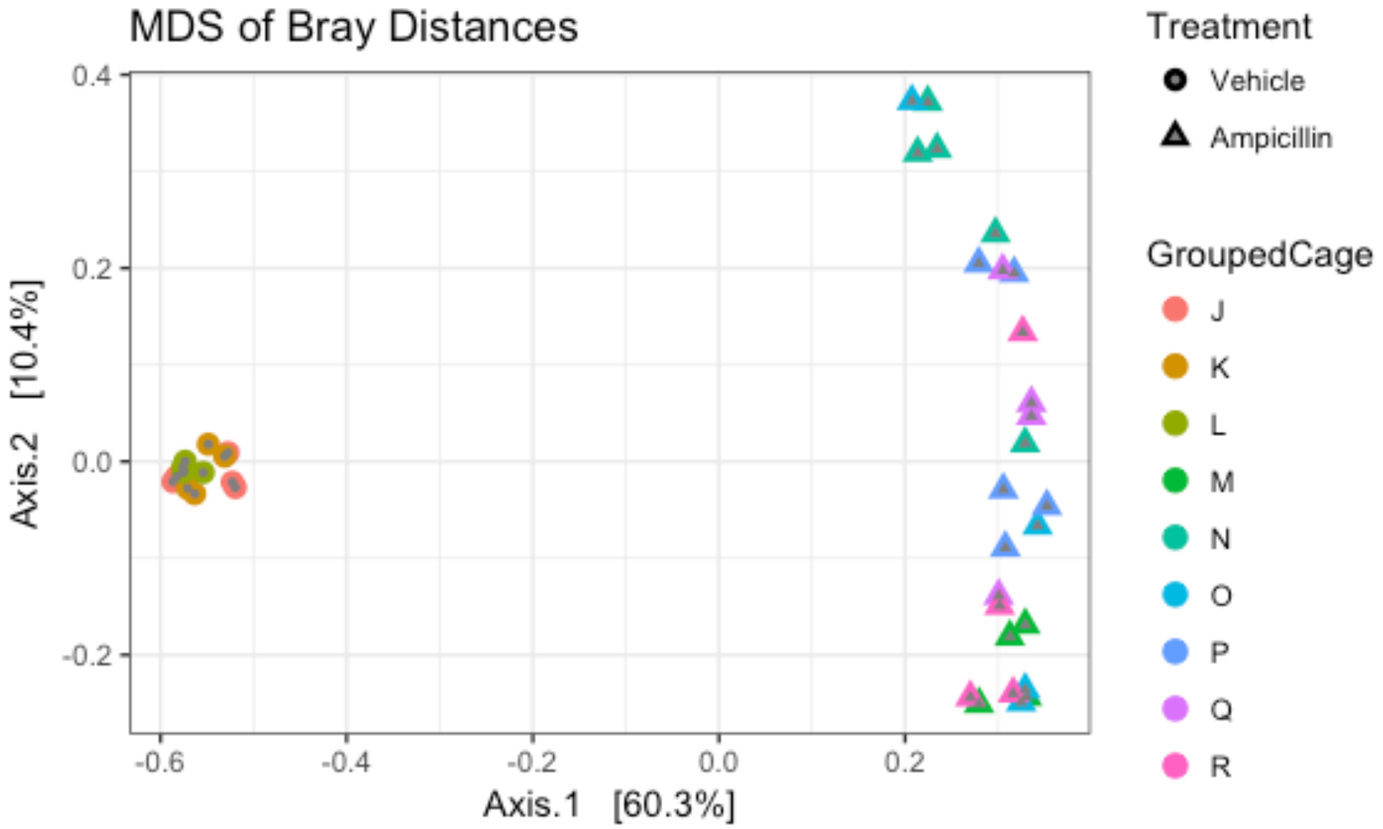


Richness Example



Beta Diversity

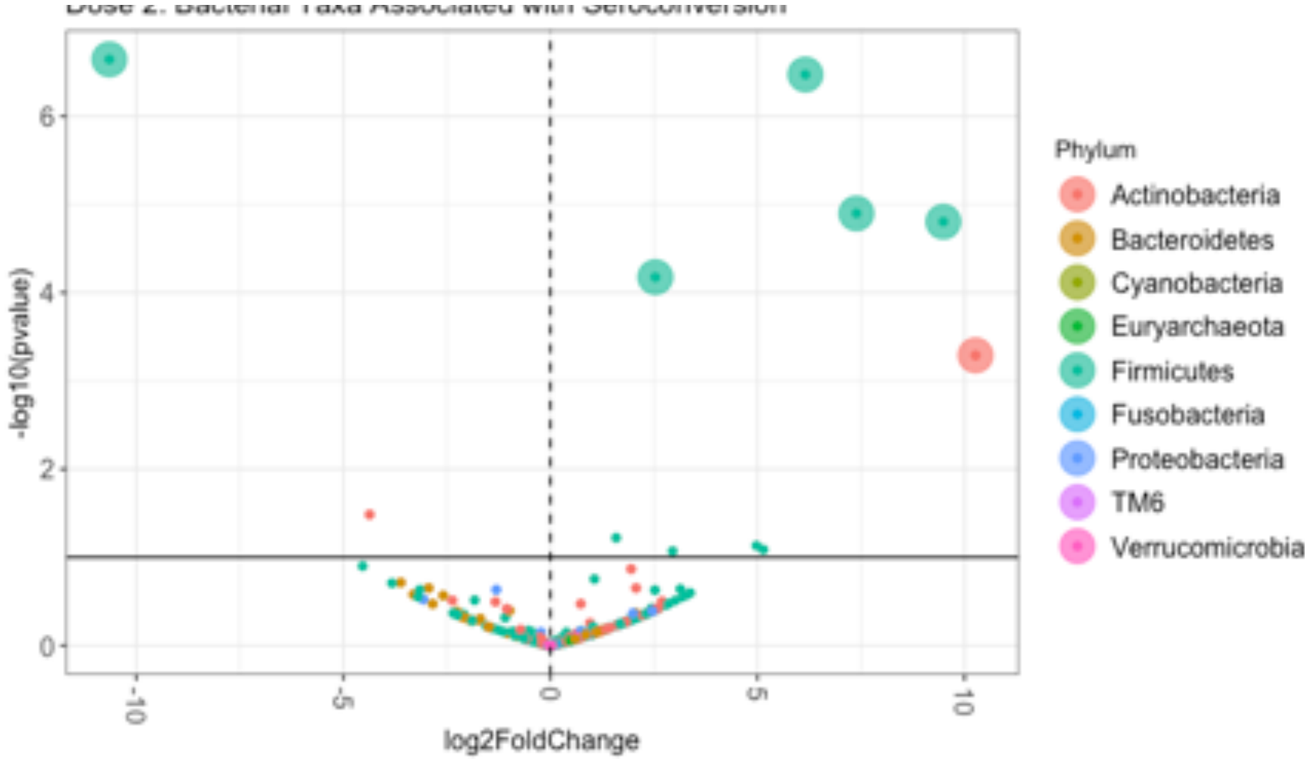
- Between sample similarity
  - Distance between one sample to all other samples
  - Multivariant
  - Can incorporate relative abundances or not
  - Can incorporate phylogenetic information or not
  - Most frequently displayed in an ordination plot



To learn about distance measures and ordination:
https://sites.google.com/site/mb3gustame/home

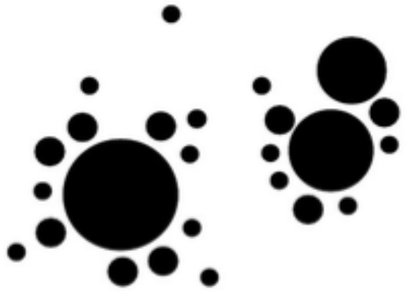
Differential Abundance Analysis

- What specific taxa are different between study groups?
  - Lots of methods
    - DeSeq2
    - Random Forest
    - LeFse
    - ANCOM
    - Gneiss
    - …

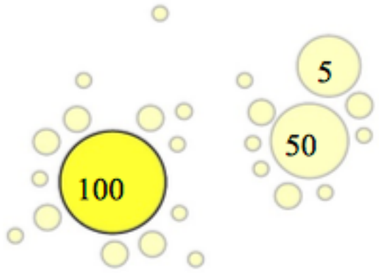


Rest of today

- Morning: Resolve sequence variants with dada2
- Afternoon: Analyze antibiotic treated mice case study



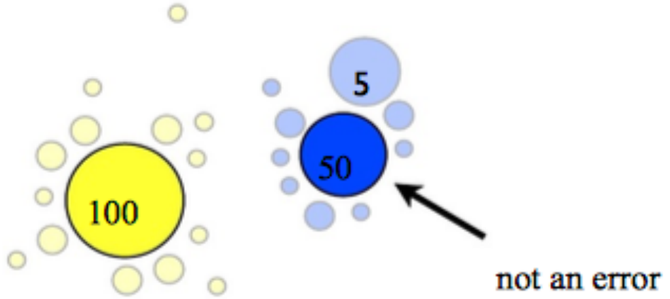
Step 1: Initial guess.
All sequences + errors



Step 2: Initial *error model*

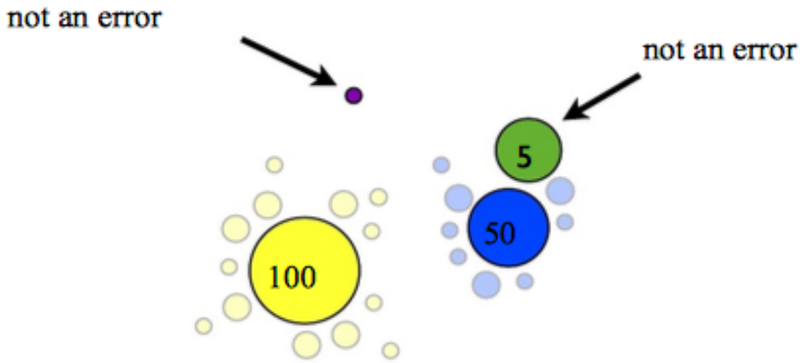
$Pr(i \rightarrow j) =$

	A	C	G	T
A	0.97	10_{-2}	10_{-2}	10_{-2}
C	10_{-2}	0.97	10_{-2}	10_{-2}
G	10_{-2}	10_{-2}	0.97	10_{-2}
T	10_{-2}	10_{-2}	10_{-2}	0.97

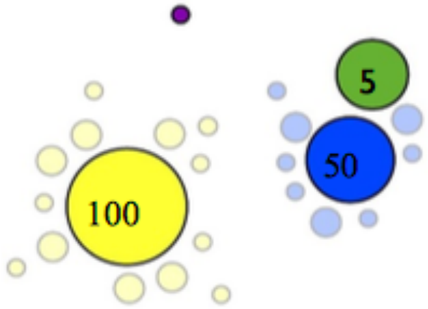


Step 3: Unlikely error under model.
Recruit errors. Update the model

	A	C	G	T
A	0.97	10_{-2}	10_{-2}	10_{-2}
C	10_{-2}	0.97	10_{-2}	10_{-2}
G	10_{-2}	10_{-2}	0.97	10_{-2}
T	10_{-2}	10_{-2}	10_{-2}	0.97



Step 3: Reject more sequences
under new model & update



Convergence: All errors are plausible

Dada2: Callahan, BJ et al. Nat Methods. 2016

Dada2 workflow