

Part I: Bacterial Microbiome Analysis

Focus on Role in Human Health and Disease

Scott A. Handley, PhD

Associate Professor

Washington University School of Medicine

The Edison Family Center for Genome Sciences and Systems Biology

What is a Microbiome?

- **Microbiome:** The entire habitat, including the microorganisms (bacteria, archaea, micro-eukaryotes (including fungi, protists) and viruses AND their surrounding conditions
 - If a specific fraction is analyzed a prefix is warranted
 - Bacterial microbiome (bacteriome)
 - Fungal microbiome (mycobiome)
 - Viral microbiome (virome)
 - etc.
- **Metagenome:** The collection of genomes and genes from members of a microbiome



Other Definitions to Keep in Mind

- Targeting specific molecular components of a microbiome
 - **Metatranscriptome:** The suite of expressed RNAs in a microbiome
 - **Metaproteome:** All of the proteins in a microbiome
 - **Metabalome:** All of the metabolites in a microbiome



More Specific Terms

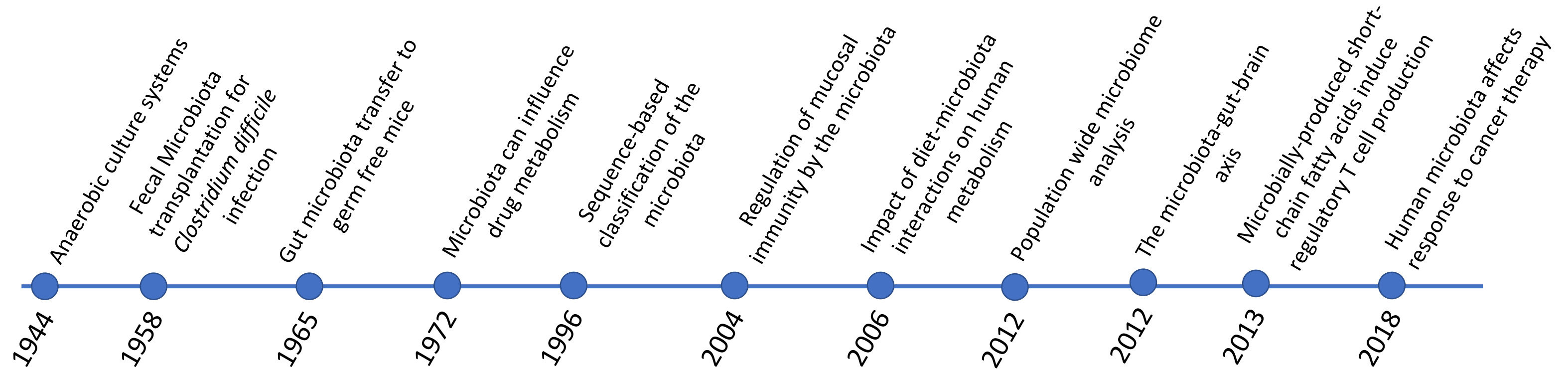
- **Metagenomics:** Random (shotgun) sampling of an entire (all organisms) or fractional (kingdom enriched) component of the genes and genomes of a microbiome
 - Enrichment can take on many forms, from culture, to mechanical filtration
- **Amplicon Surveys:** Analyzing a single variant trackable gene from a set of organisms
 - 16S rRNA surveys for bacteria and archaea
 - ITS surveys for fungi
 - Doesn't work very well for viruses as there is no single shared gene across viral clades
- Will discuss in more detail later

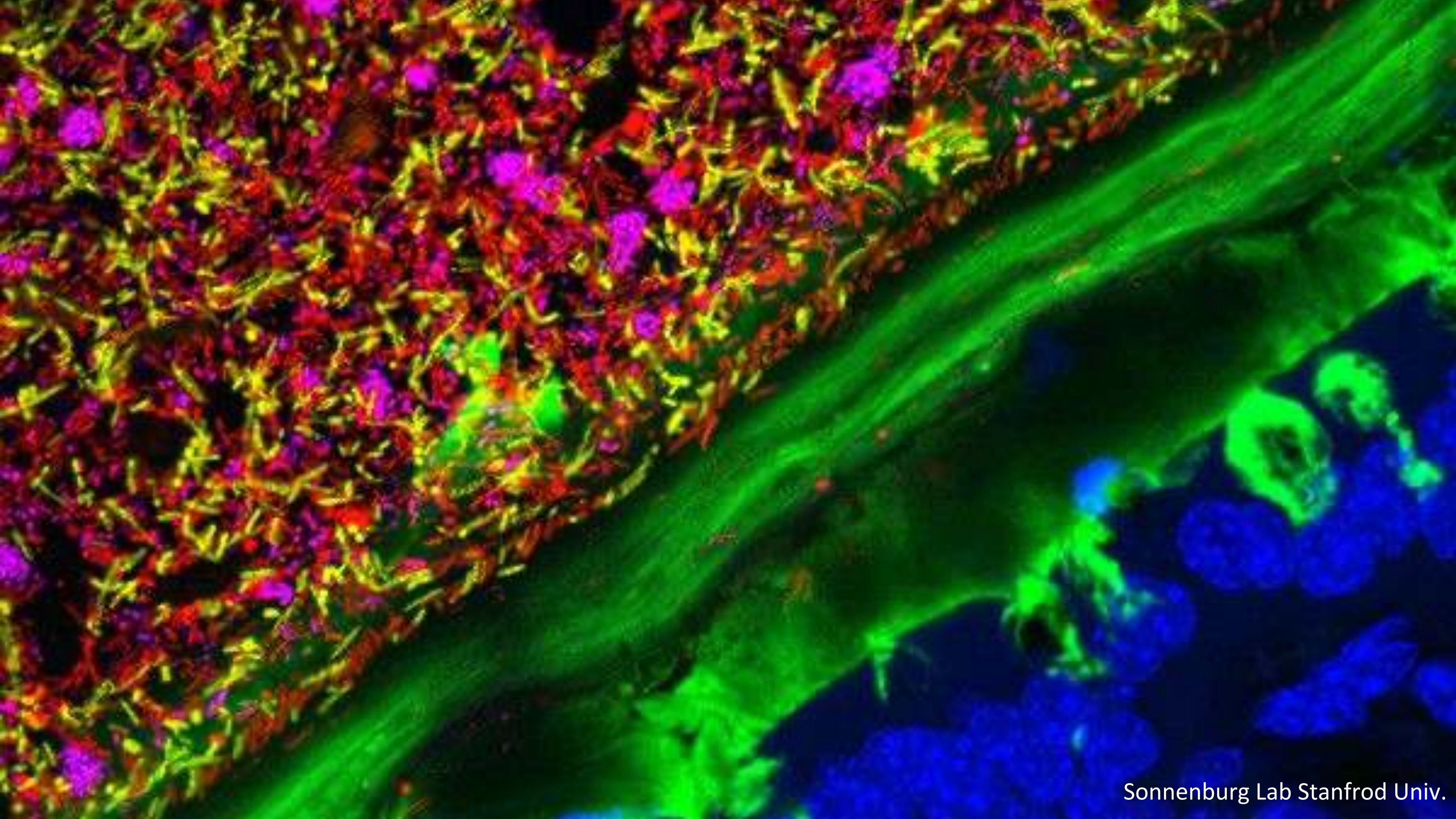


Interactions Between Microbiota and Immunity in Health and Disease

Heavy emphasis on intestinal microbiota-host interactions

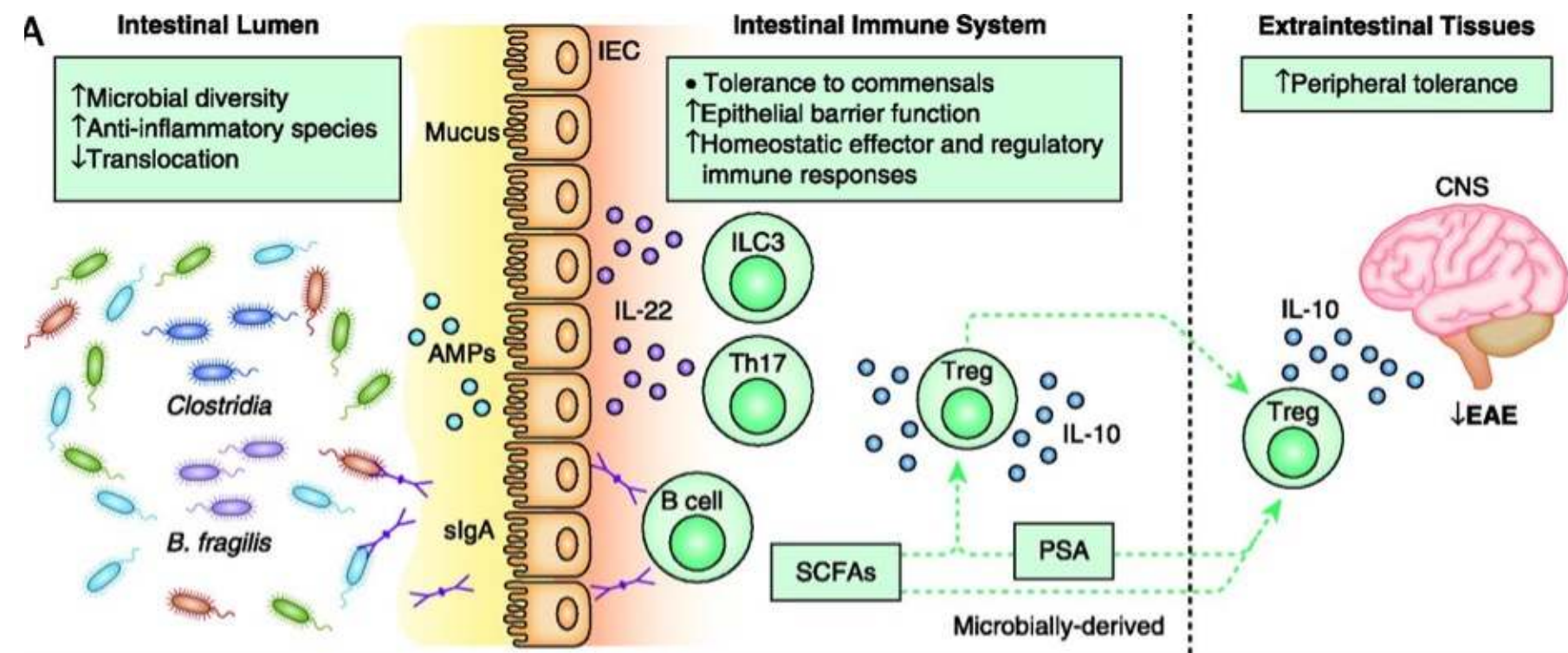
(Very Brief) History of Microbiome Research in Human Health





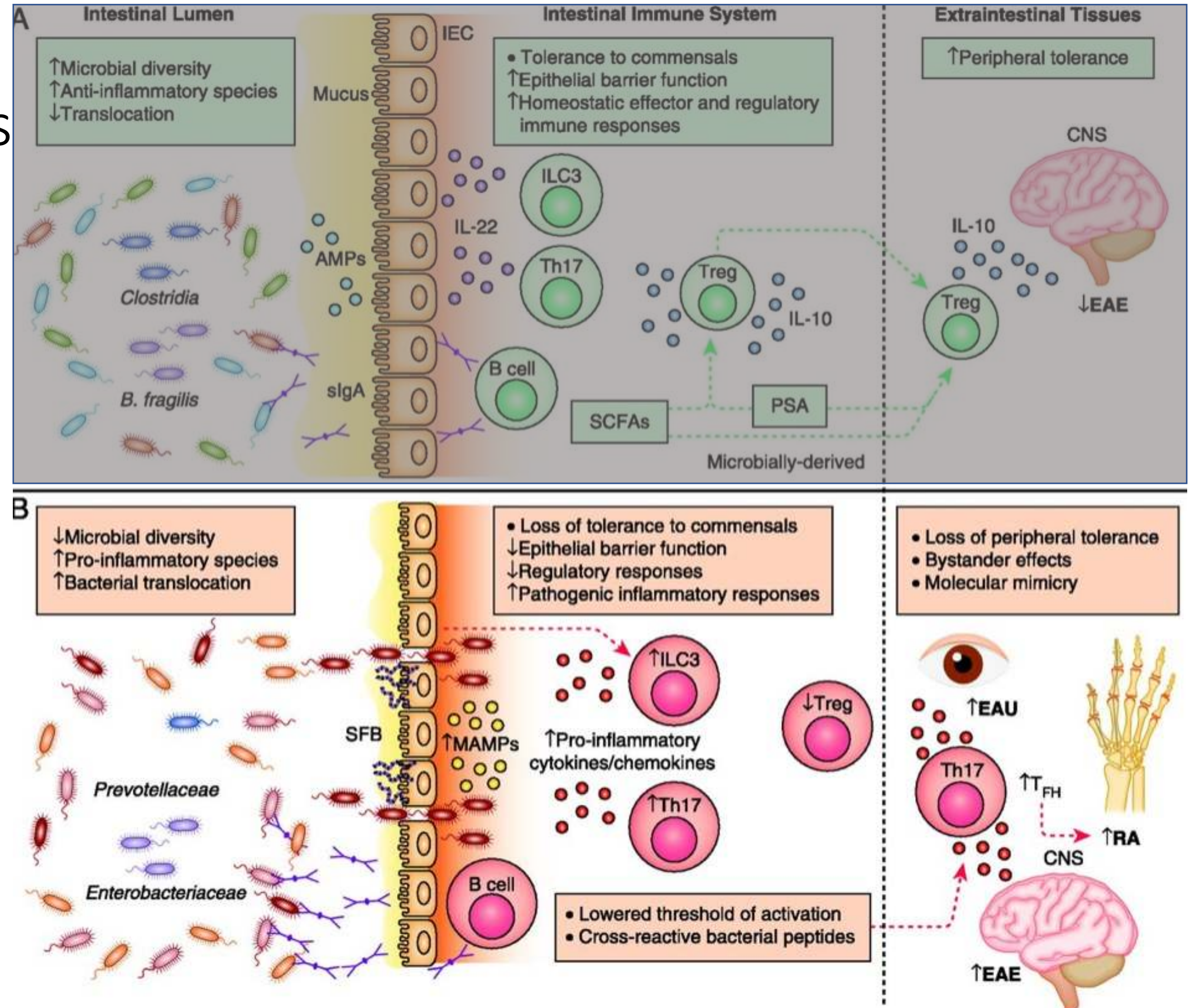
Host-induced Compartmentalization of Intestinal Microbiota

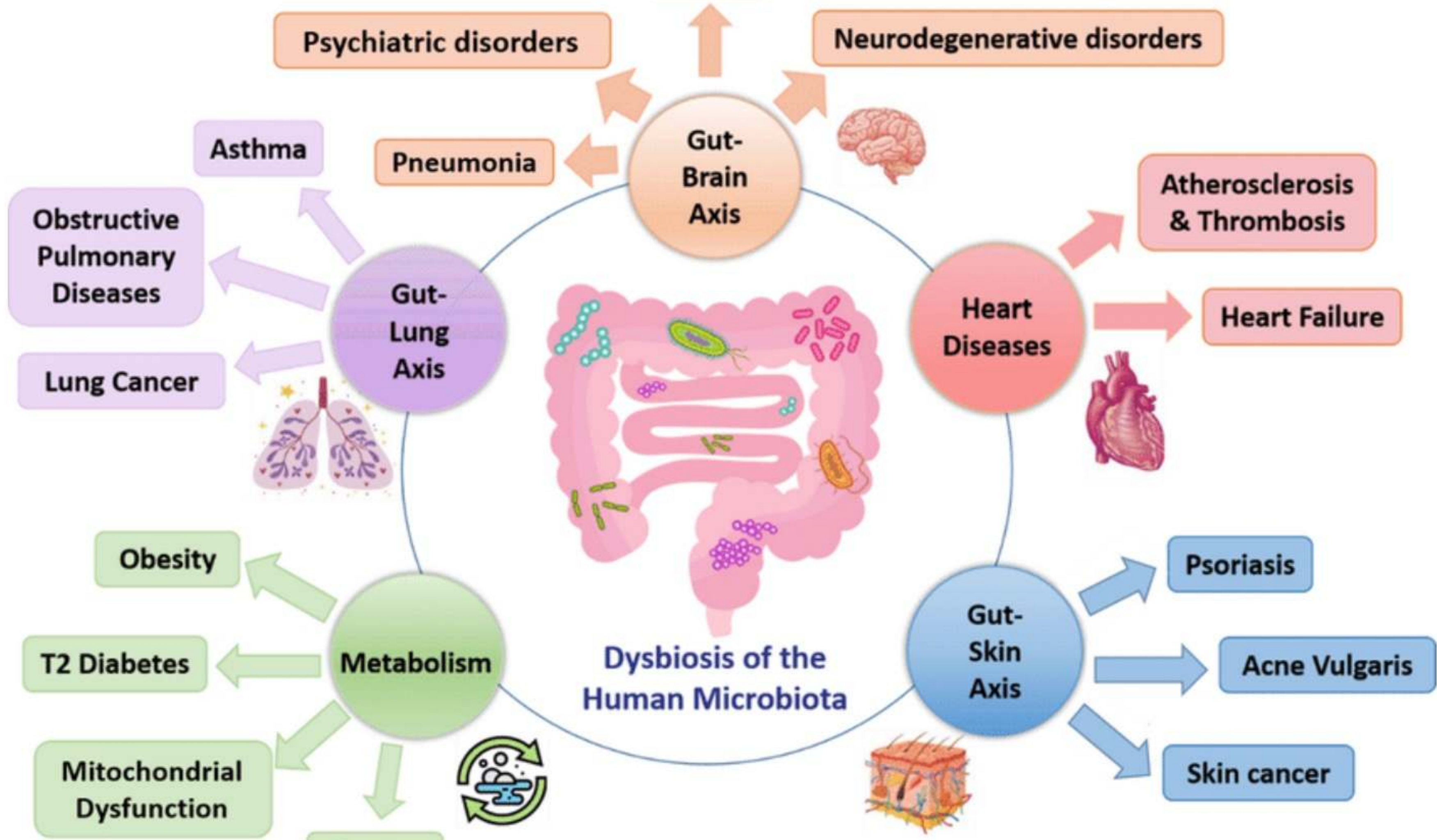
- In a healthy state the host's immune response to the intestinal microbiota is compartmentalized to the mucosal surface
 - Single-layer of epithelium
 - Dense mucous layer
 - Secretory IgA (sIgA) antibodies and antimicrobial peptides (AMPs) also help to maintain barrier function
- “Typically” associated with high-bacterial diversity
 - Regulatory T-cells promoted by bacterial short-chain fatty acids and polysaccharide A
- Microbial product regulation of Regulatory T Cells (Treg) can protect against systemic inflammation
- All designed to maintain homeostatic balance between commensal microbes and their host



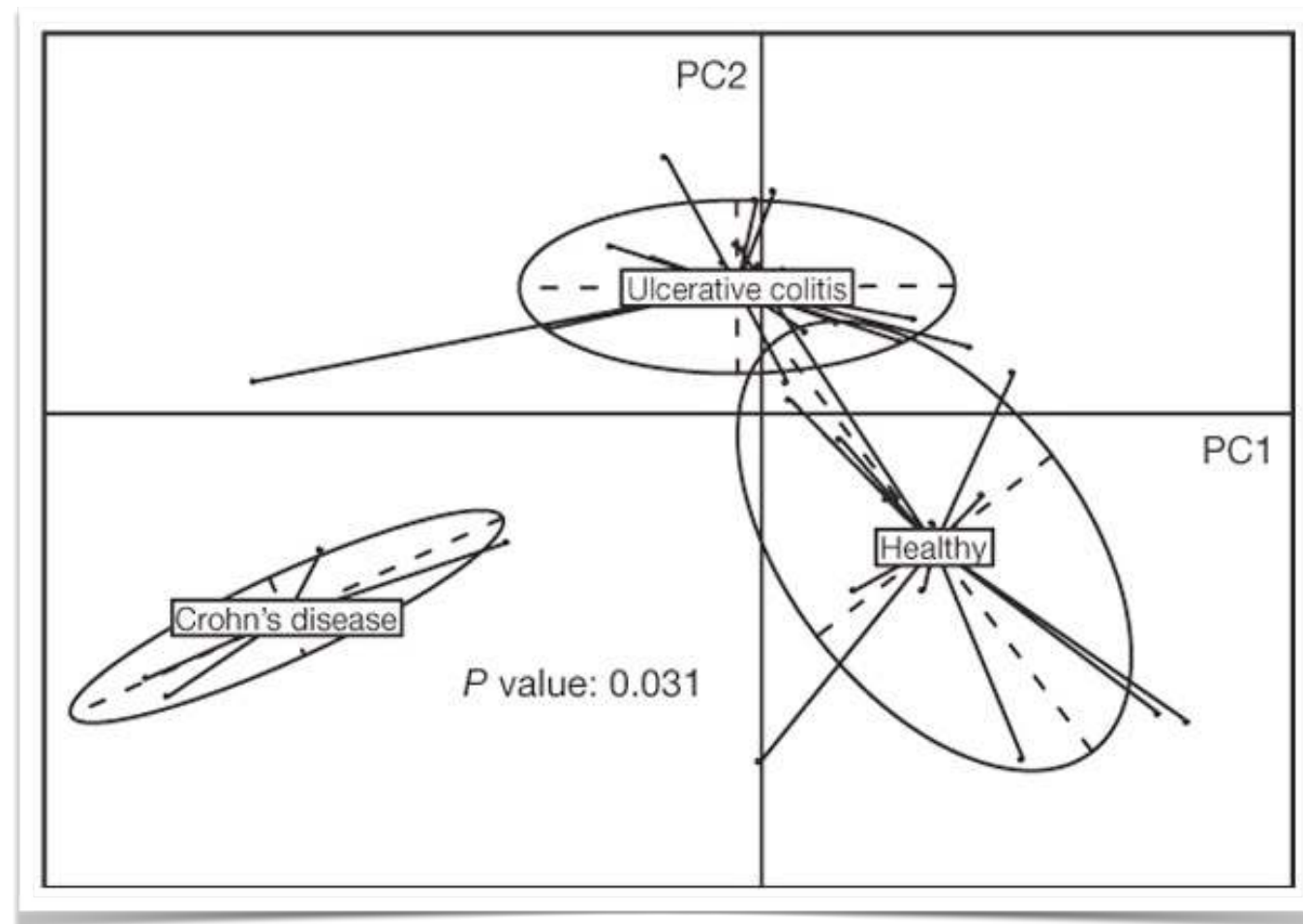
What Happens When Things Go Wrong?

- Loss of barrier function
- Stimulated by the recognition of microbe-associated molecular patterns (MAMPs)
- Pro-inflammatory cytokine and chemokine activation, Th17 and B-cell responses
- Specific bacteria can exacerbate
- Loss of tolerance for self-antigens can occur due to lowered thresholds for activation

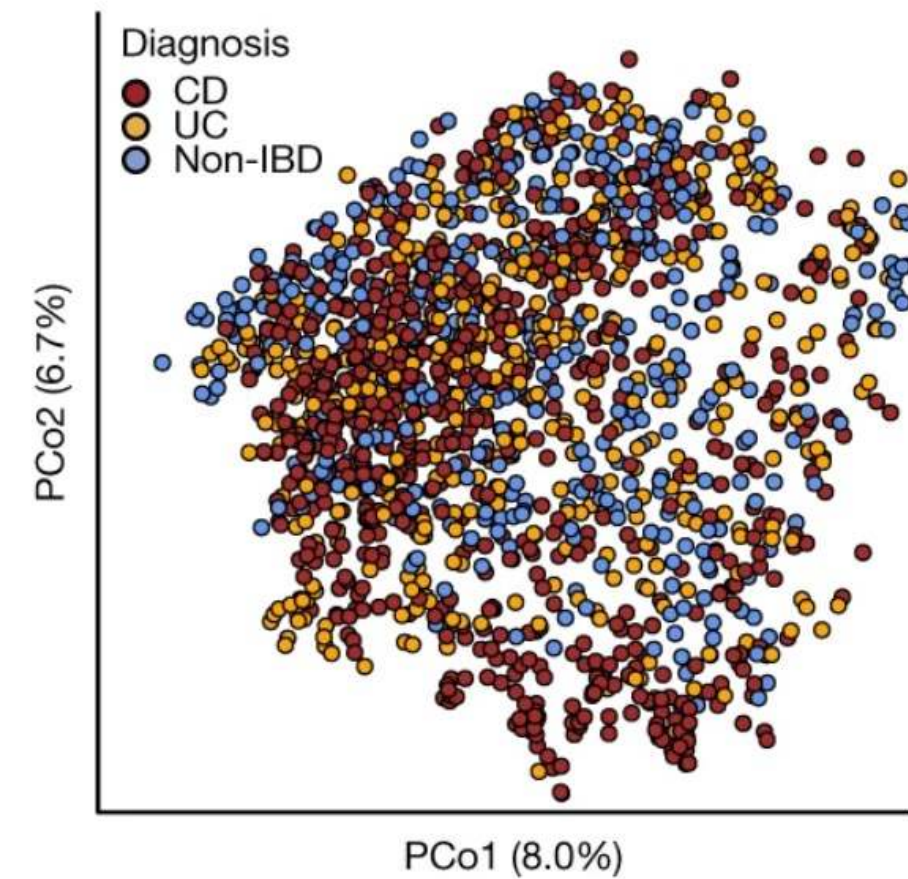




Bacterial 'Dysbiosis' – It's complicated

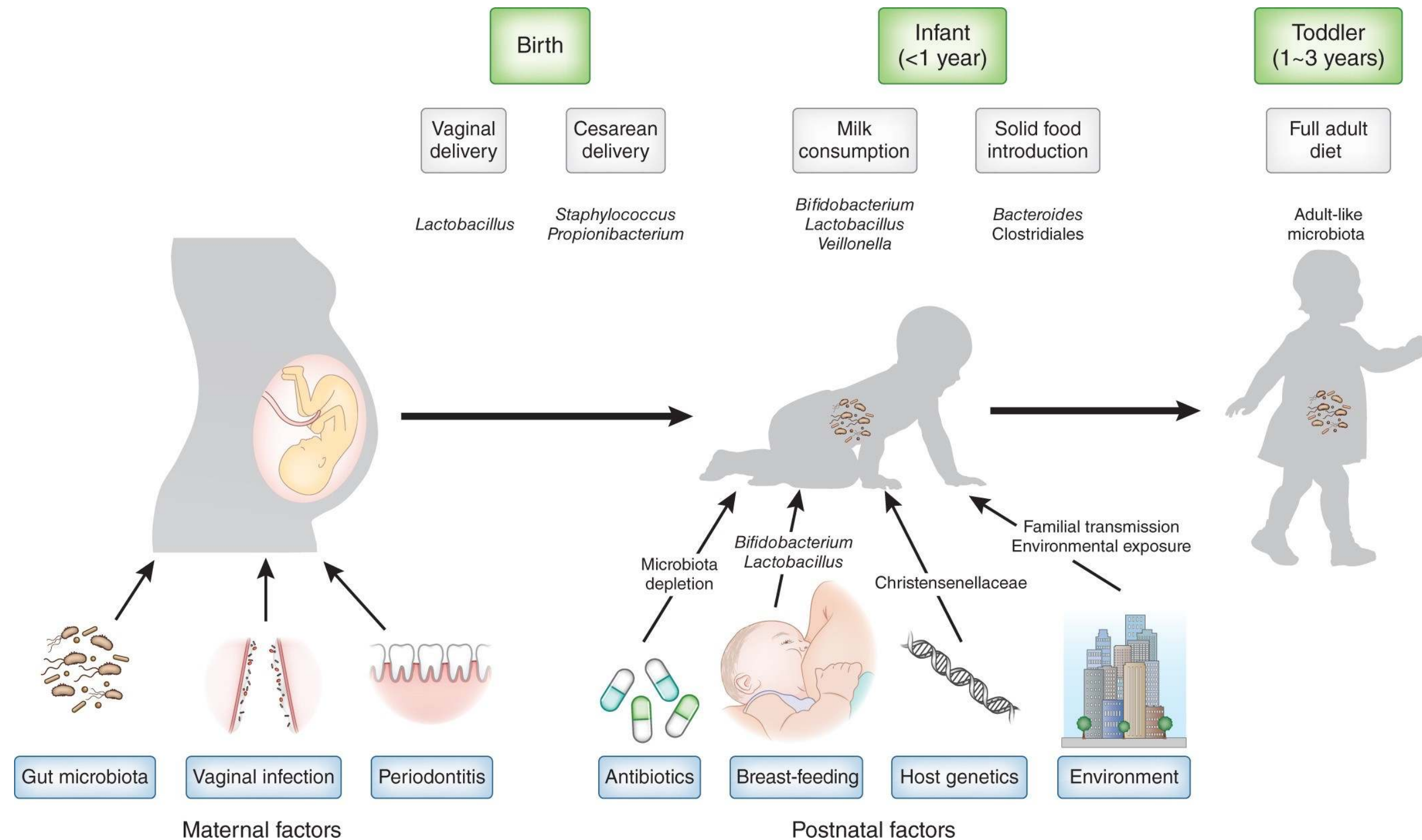


Qin, J. et. al. Nature 464, 59-65 (4 March 2010)



Loyd-Price J et al. Nature. 2019

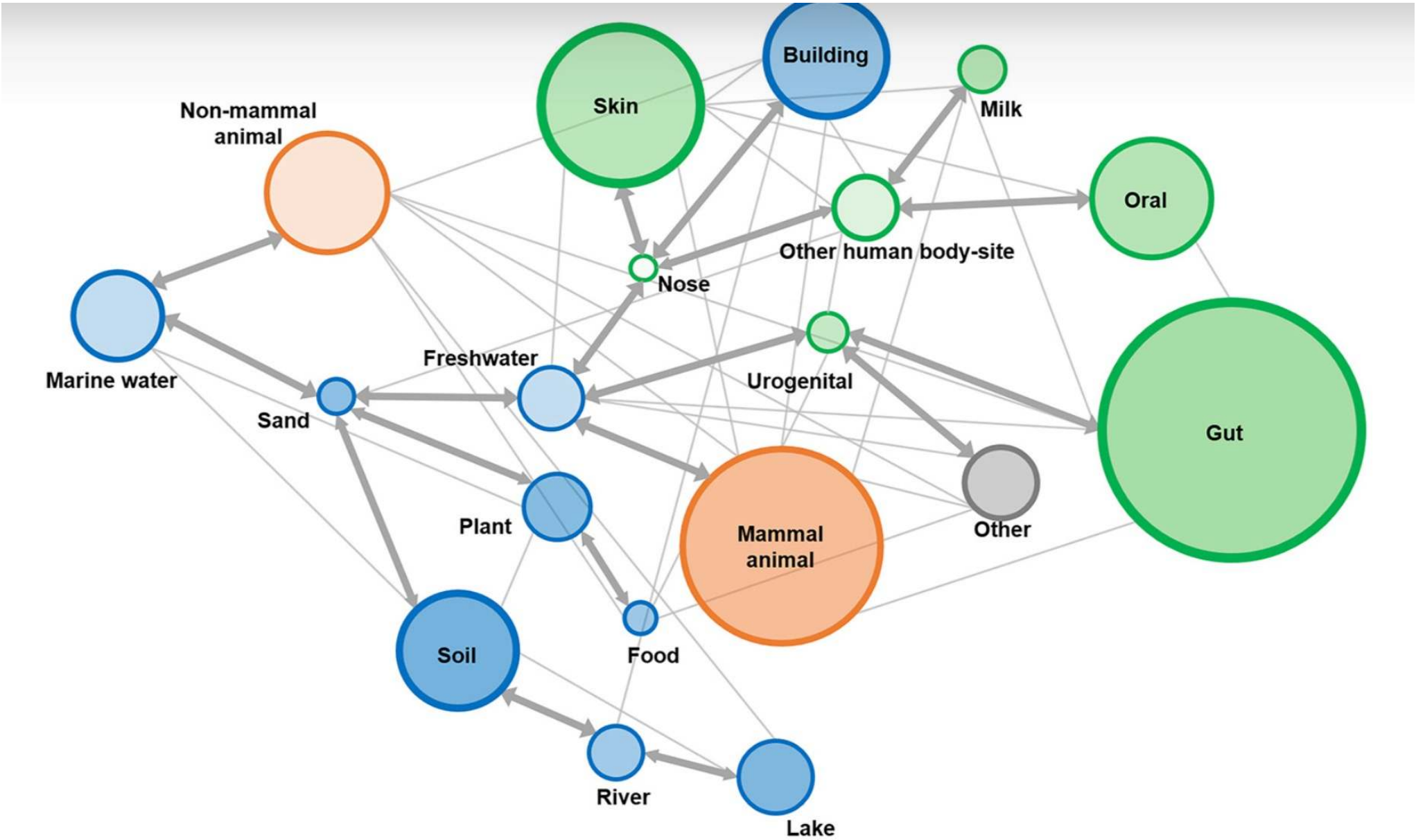
How is our microbiota established / developed?



What about non-mammalian microbiomes?

Global Microbiome Sampling

Habitat	No. of samples
Gut	51,706
Skin	19,455
Oral	10,896
Other human body-site	3,018
Urogenital	1,204
Nose	489
Mammal animal	29,918
Nonmammal animal	11,172
Building	11,248
Soil	10,507
Marine water	6,090
Lake	4,234
Plant	3,456
Freshwater	3,112
River	2,248
Milk	1,636
Sand	968
Food	780
Other	4,074



How is this work done?

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

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

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
- 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

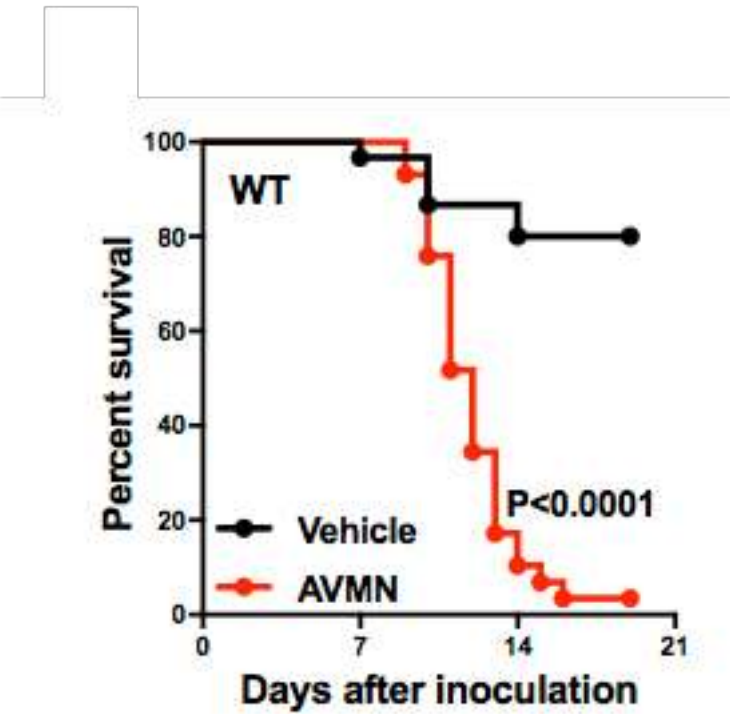
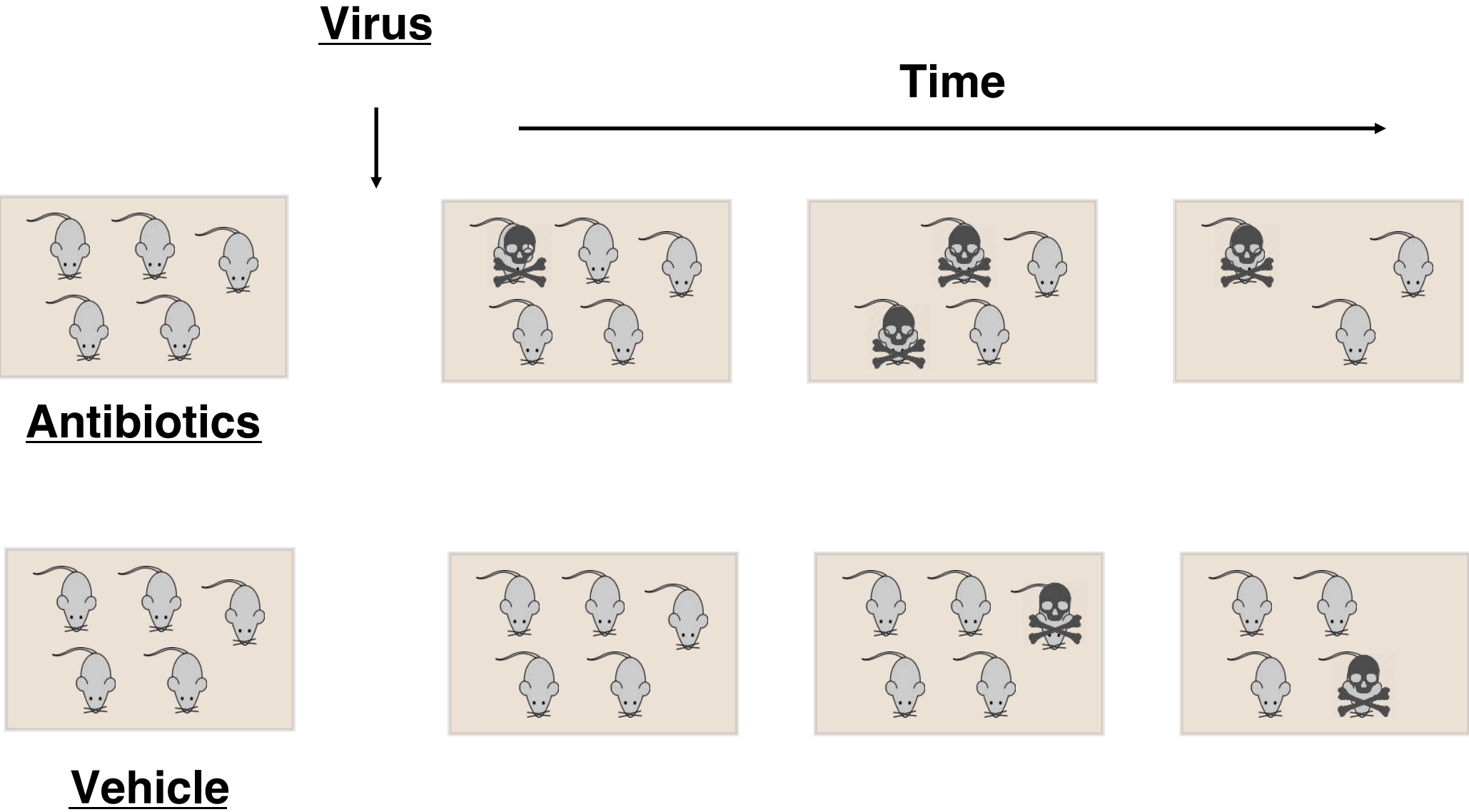
“Accounting for reciprocal host-microbiome interactions in experimental science” Stappenbeck, TS and Virgin HW. Nature. 2016 Jun 9;534(7606):191-9



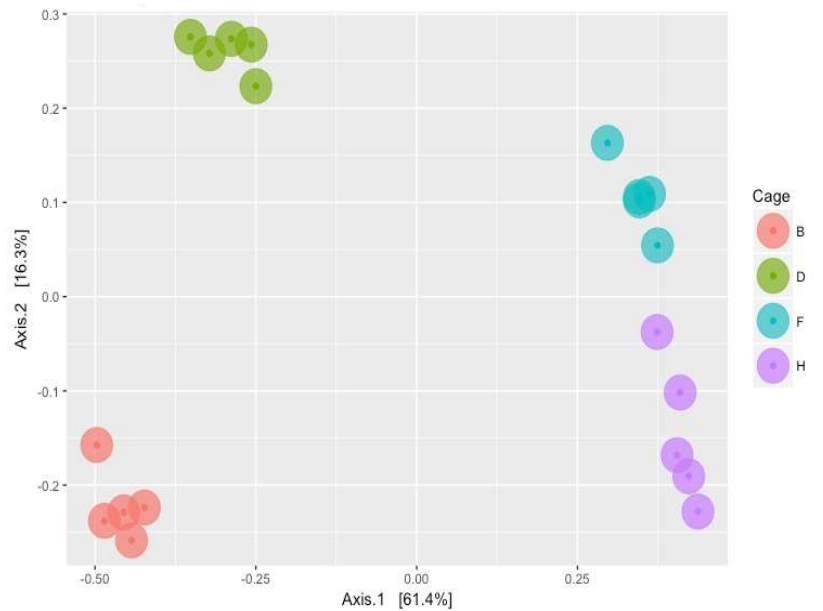
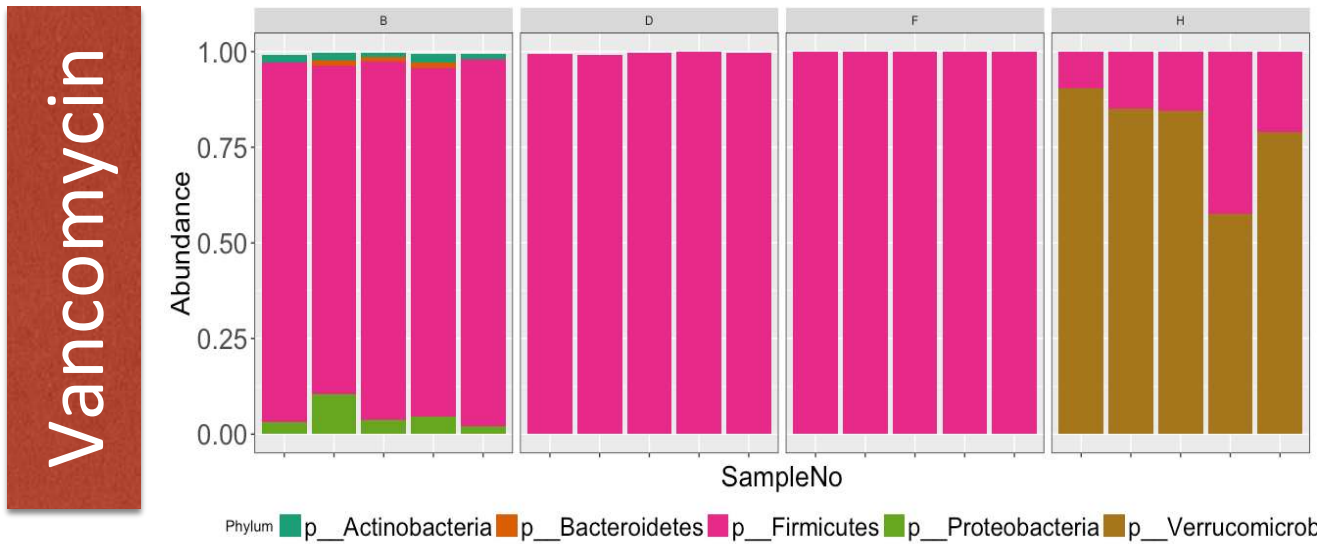
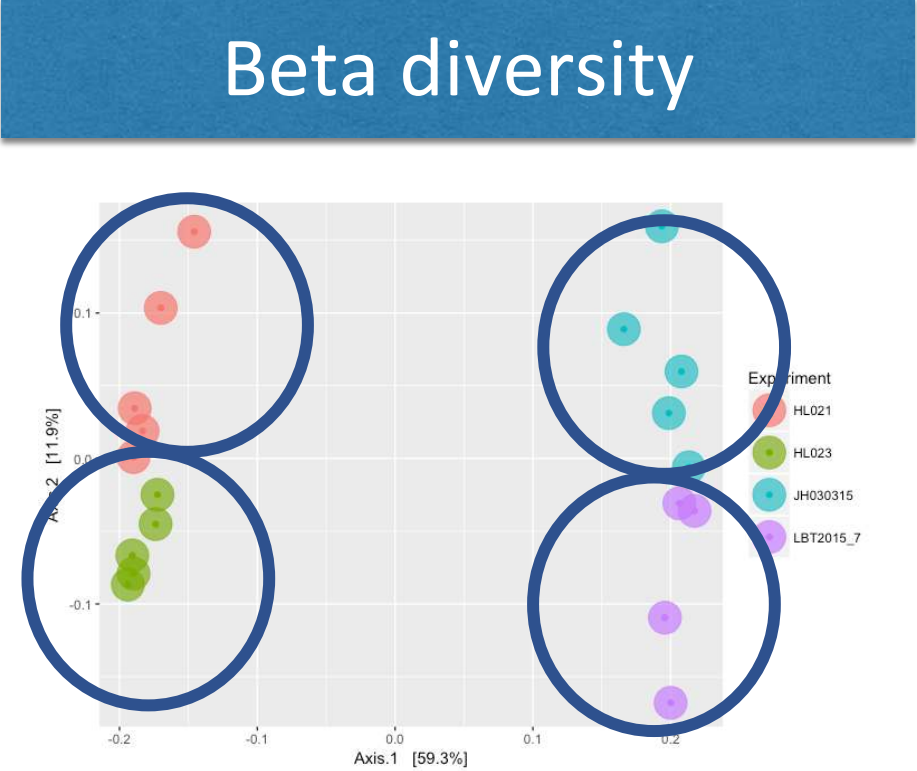
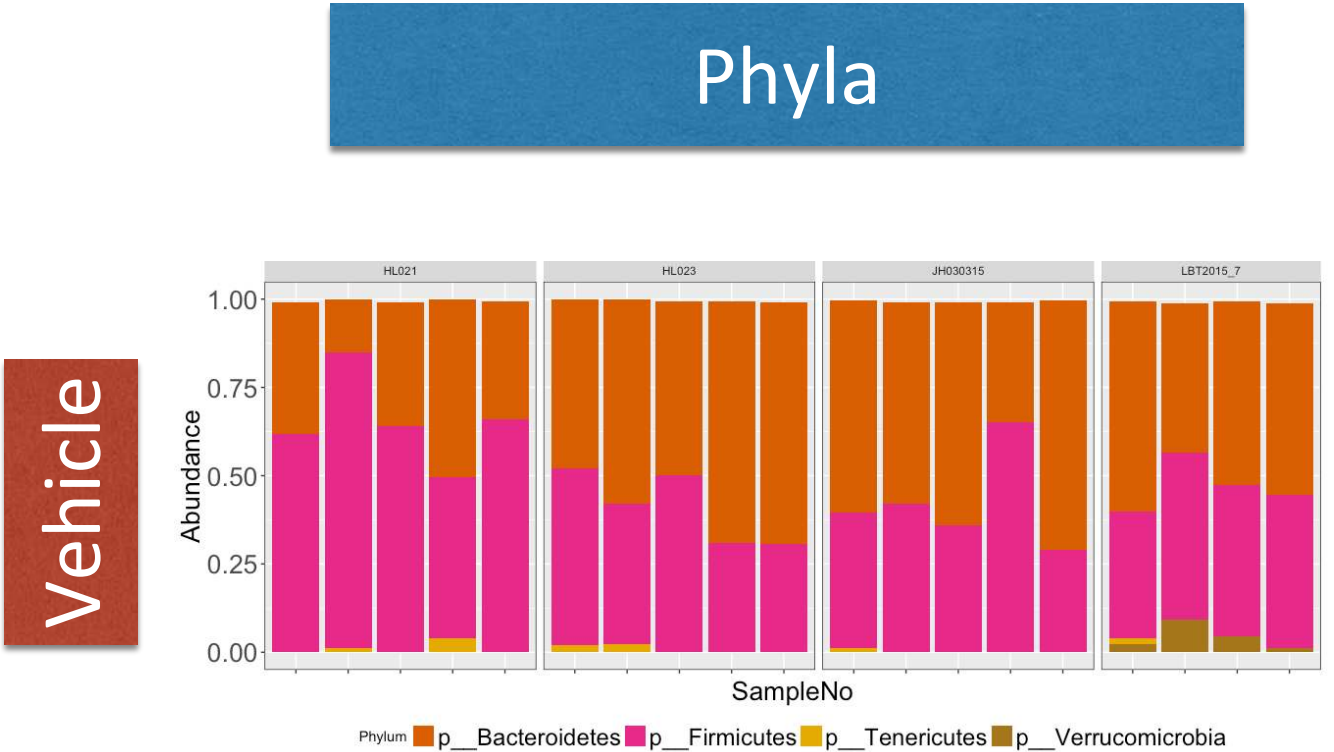
Today's Case Study

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;22(13):3440-3453

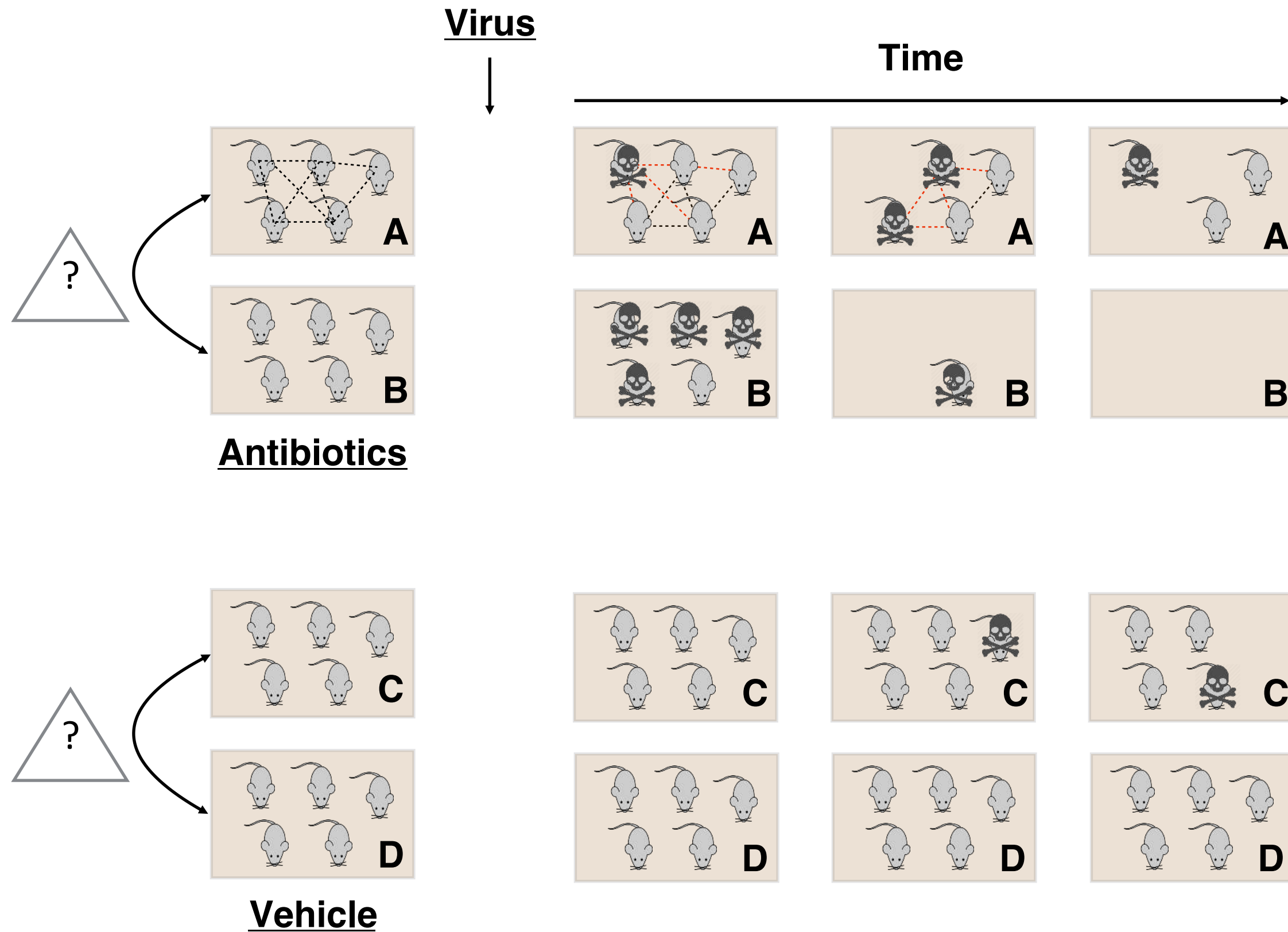
Case Study: Effect of Antibiotics on Flavivirus Pathogenesis



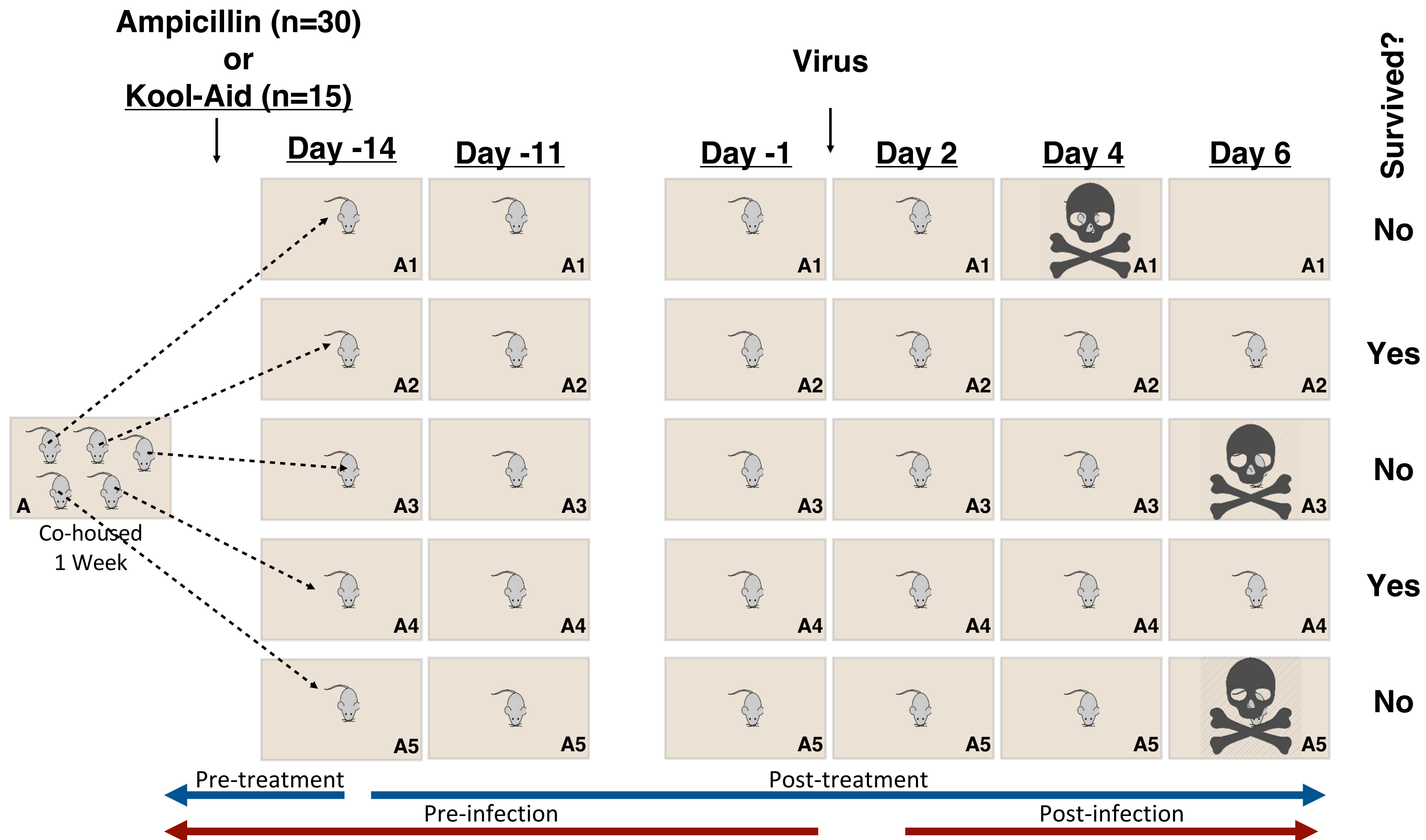
Cage Effects: 16S rRNA Amplicon Survey



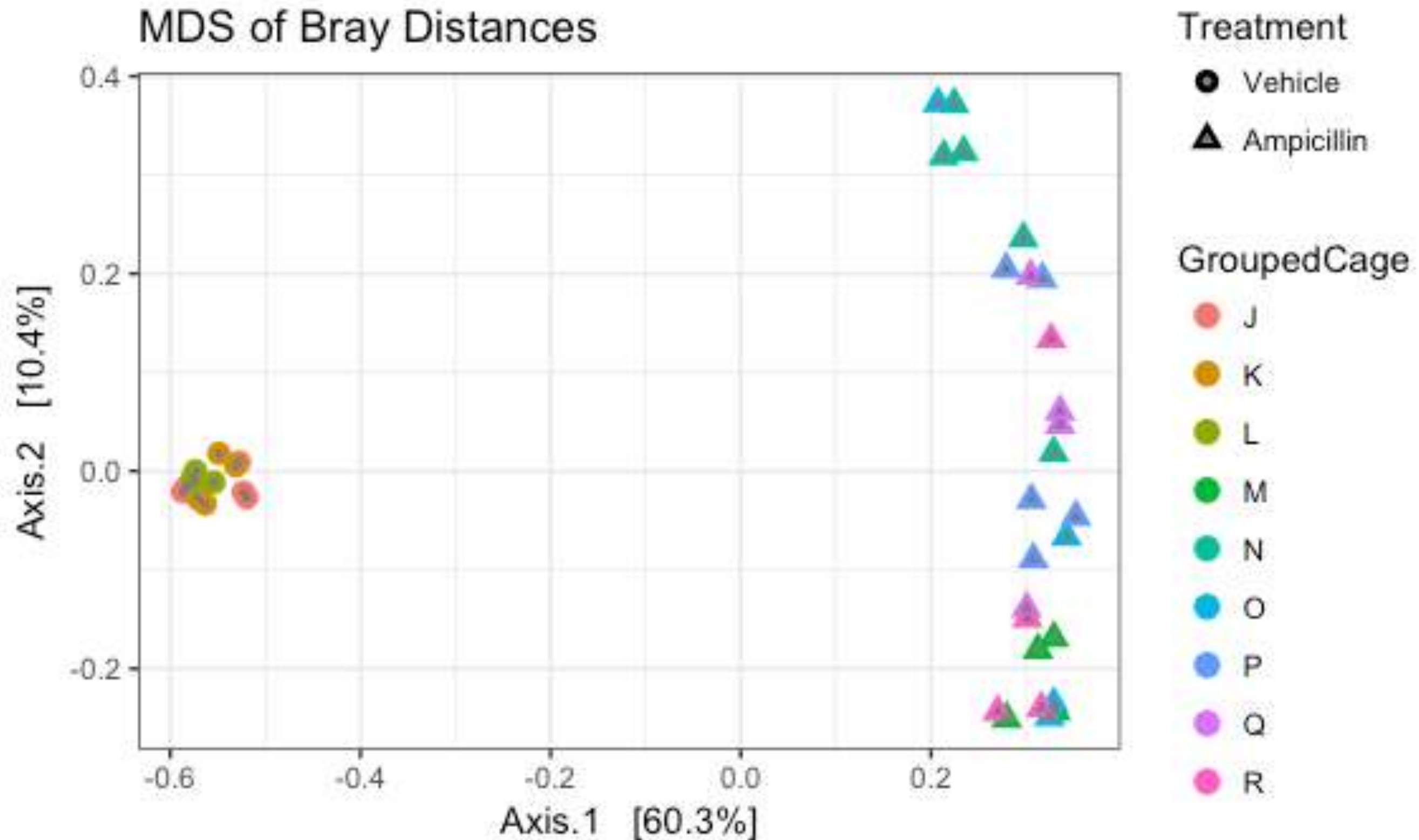
Cage and Mouse-to-Mouse Effects



Individual Mouse Isolation Schema



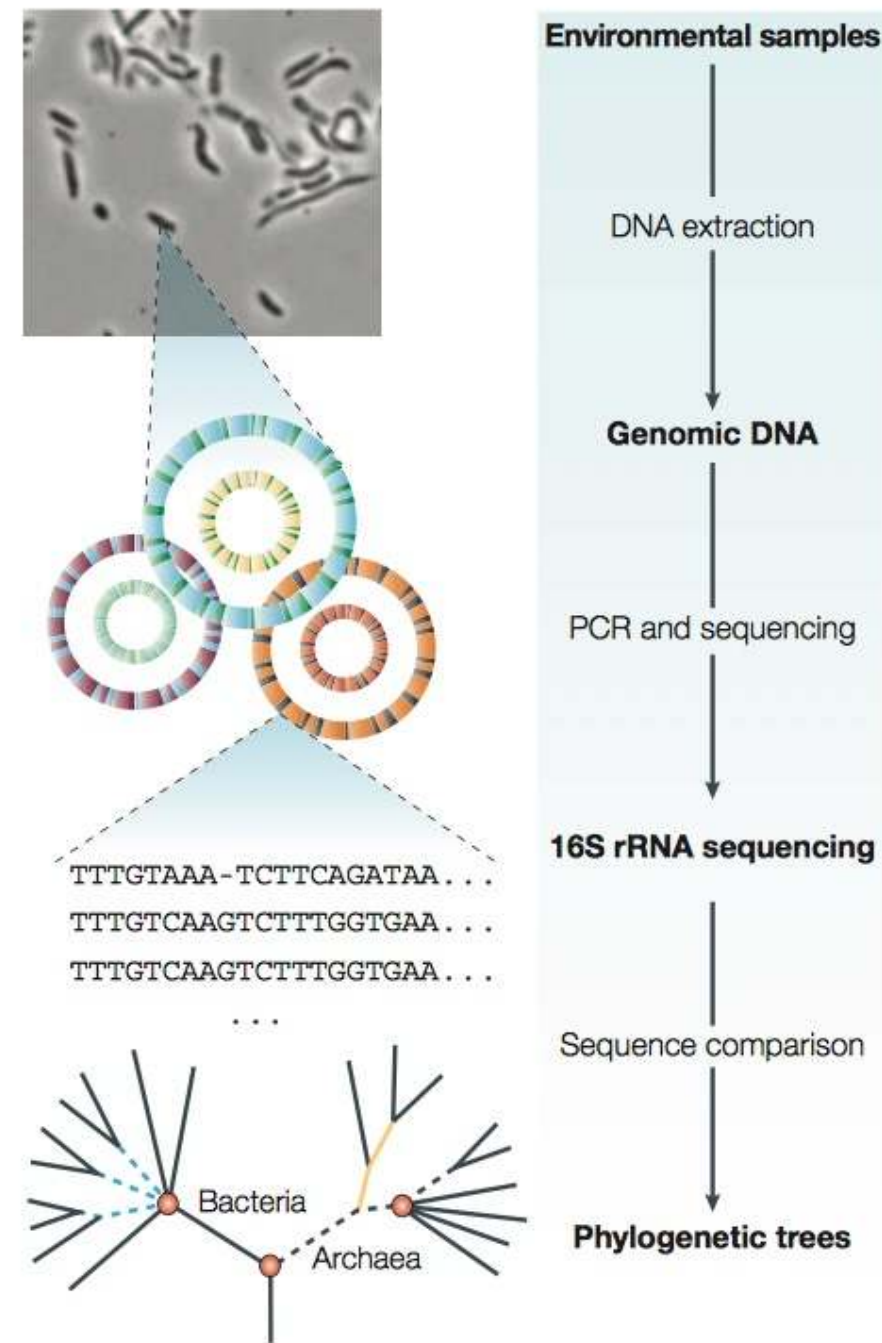
Individual Housing Results



Amplicon Surveys (Highly Opinionated!) Best-practices

It's the classic garbage in, garbage out all over again ...

16S rRNA Amplicon Survey



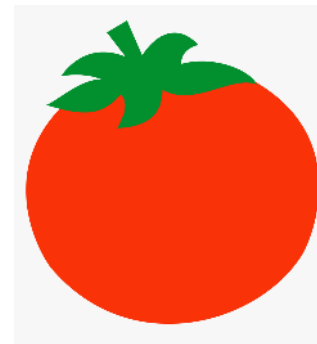
Study Design

Laboratory

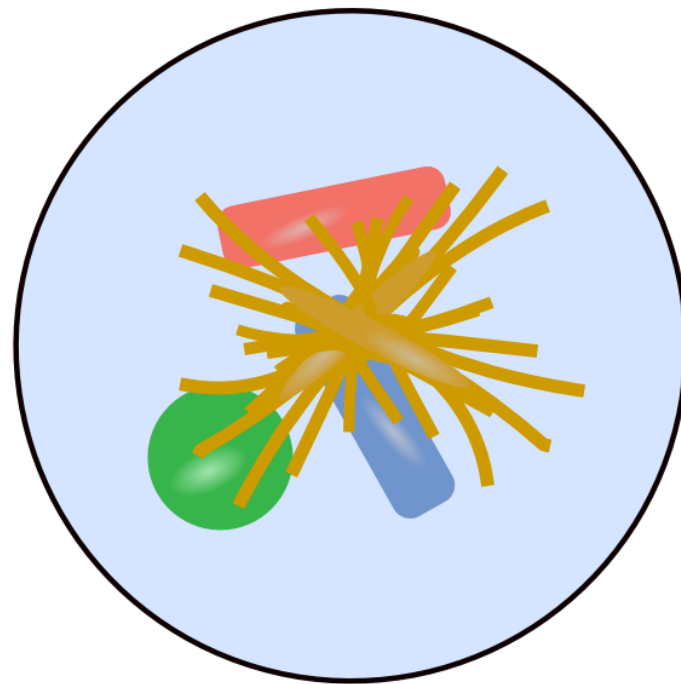
Bioinformatics, Ecological
Analysis and Statistics

Amplicon Surveys vs. Metagenomics

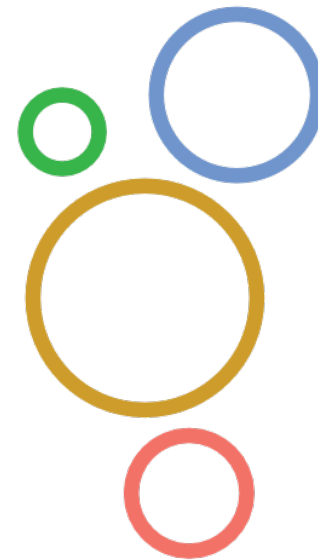
Please hold your throwing tomatoes ...



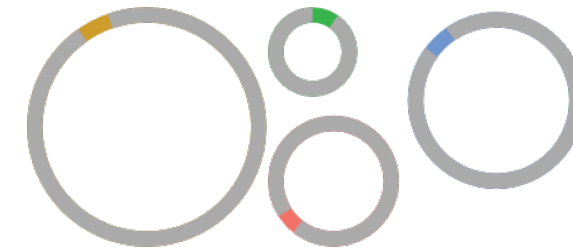
Mixed microbial community



DNA
Extraction



Amplicon sequencing



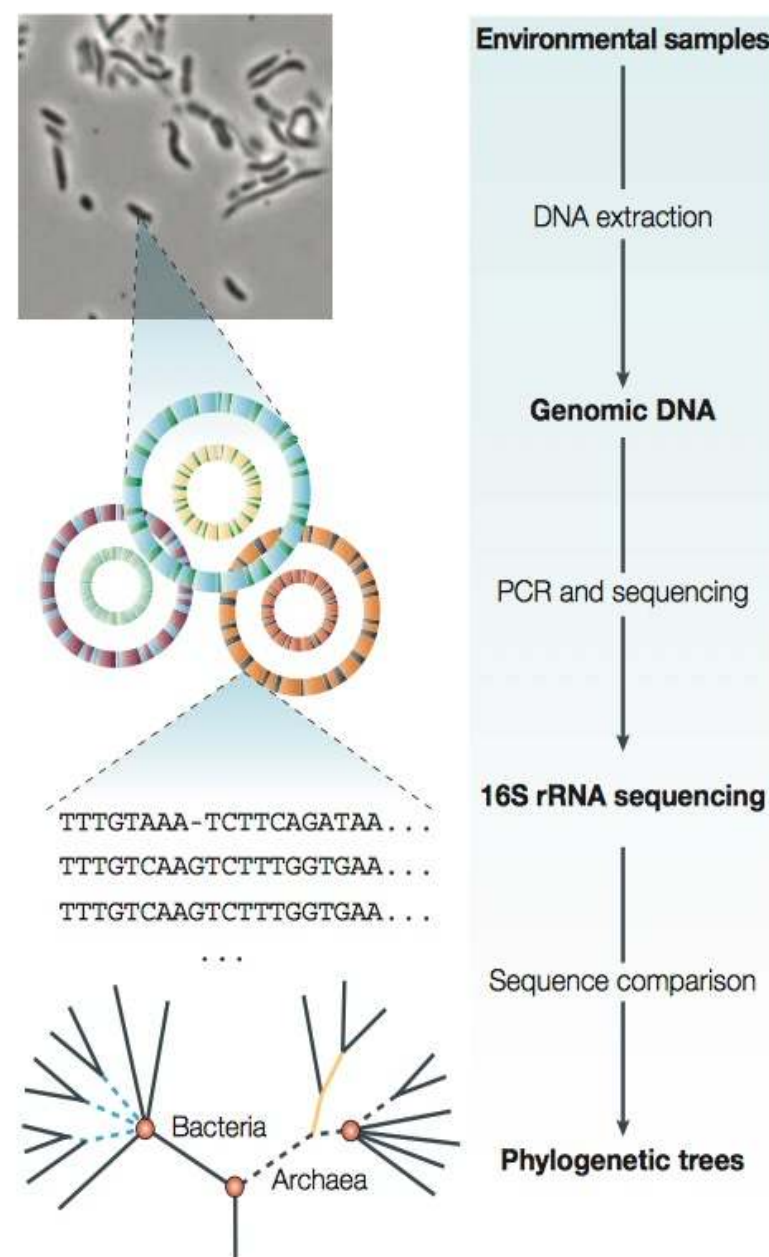
Multiple copies of fragments
from 1 target gene

Metagenomics sequencing

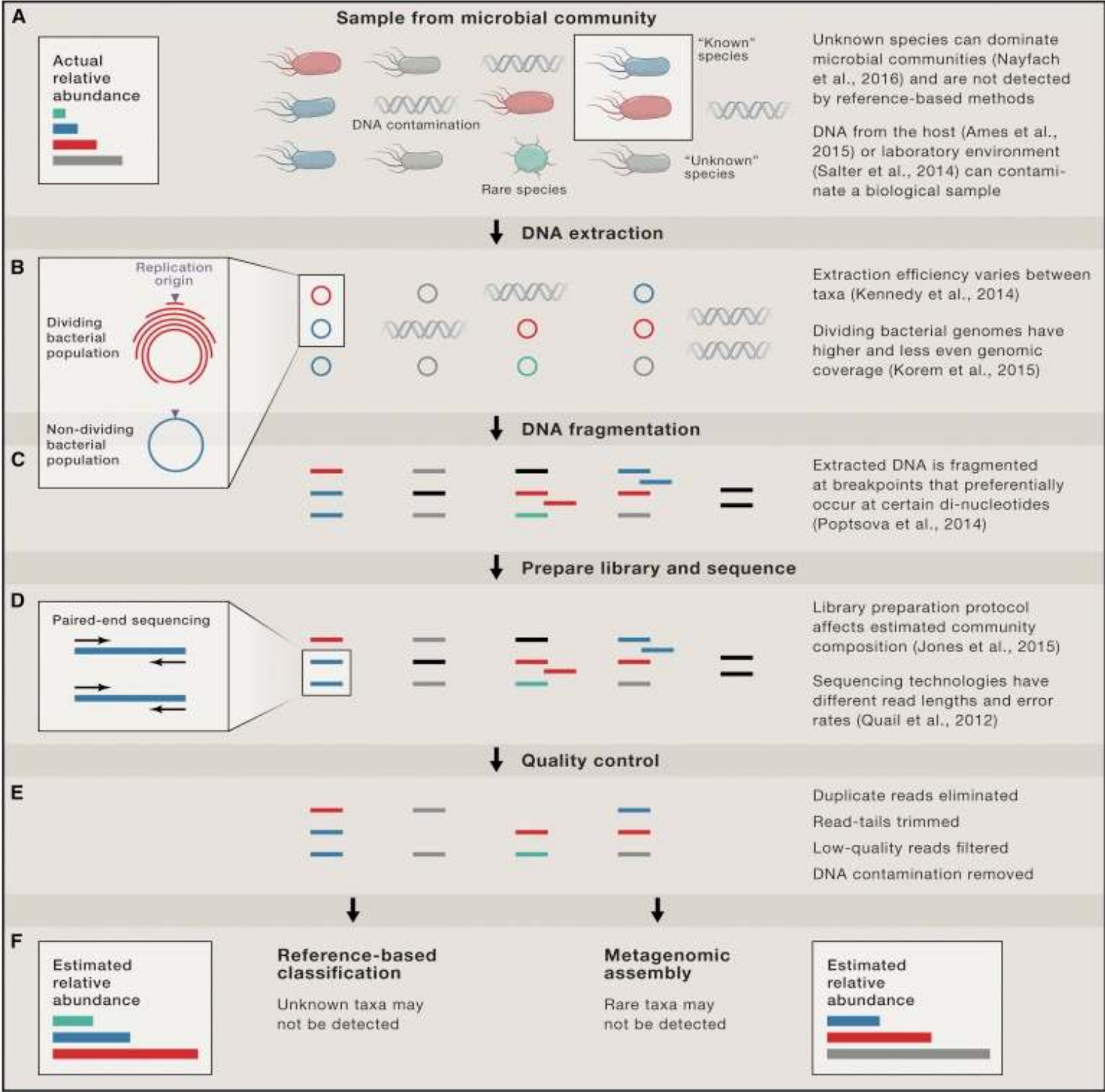


Short sequence
fragments from "all" DNA

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 \$\$\$ and information type

- Our labs per sample costs:

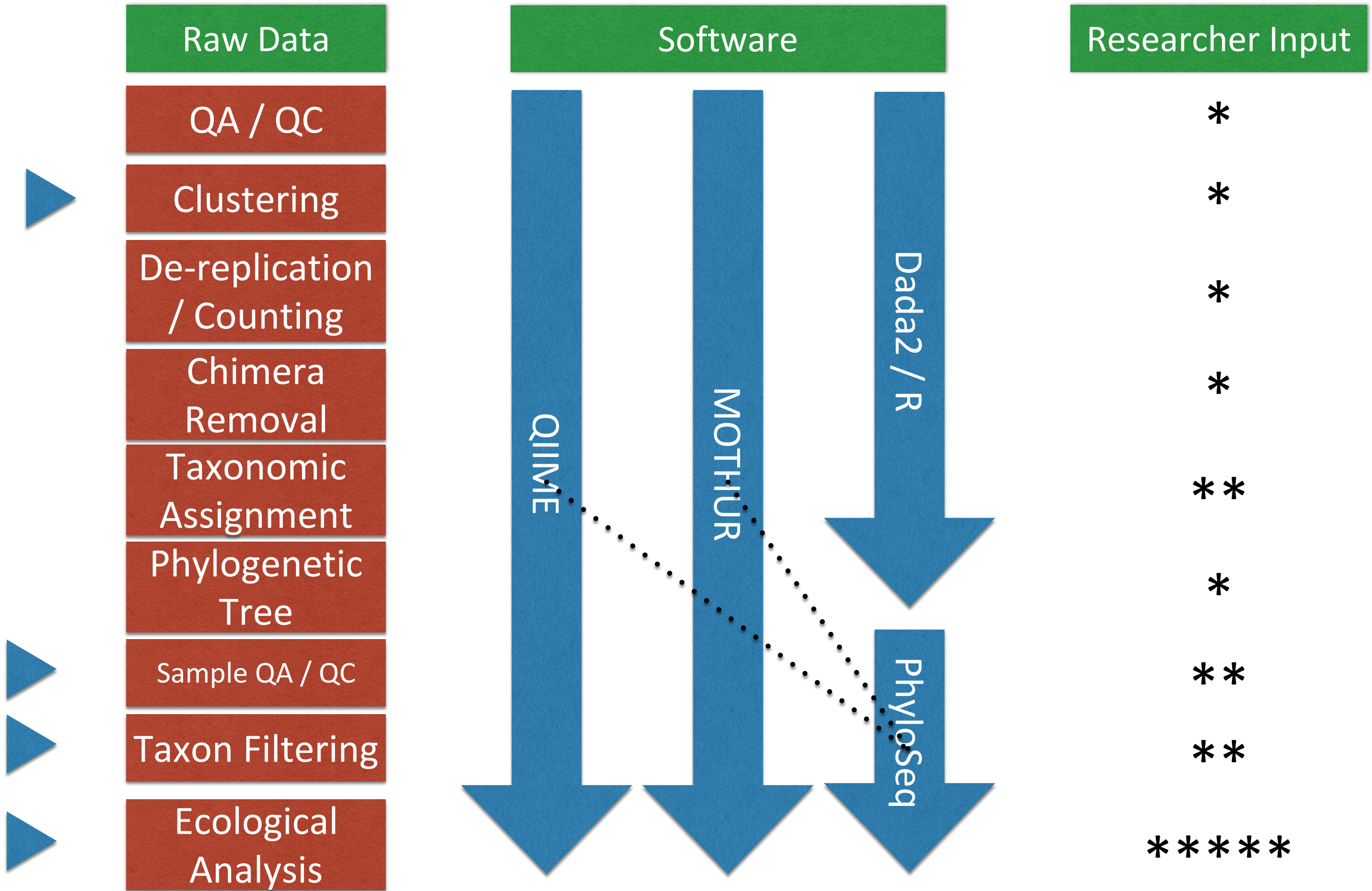
- 16S = \$17.50 per sample
- Metagenome = \$225.00 per sample

- 

- Understanding analytical space
- Data storage
- What type of information do you need? Taxonomic or functional

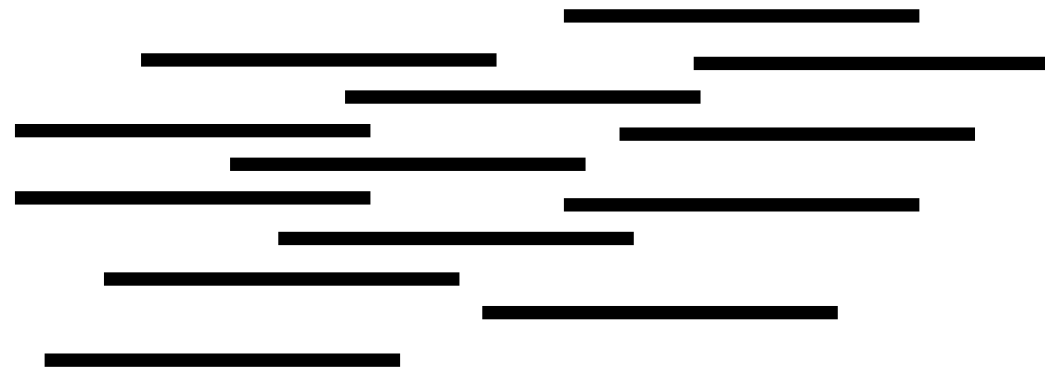
Image credit: The Internet
Quote credit: Notorious B.I.G.

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



Sequence Clustering

16S RNA Amplicons

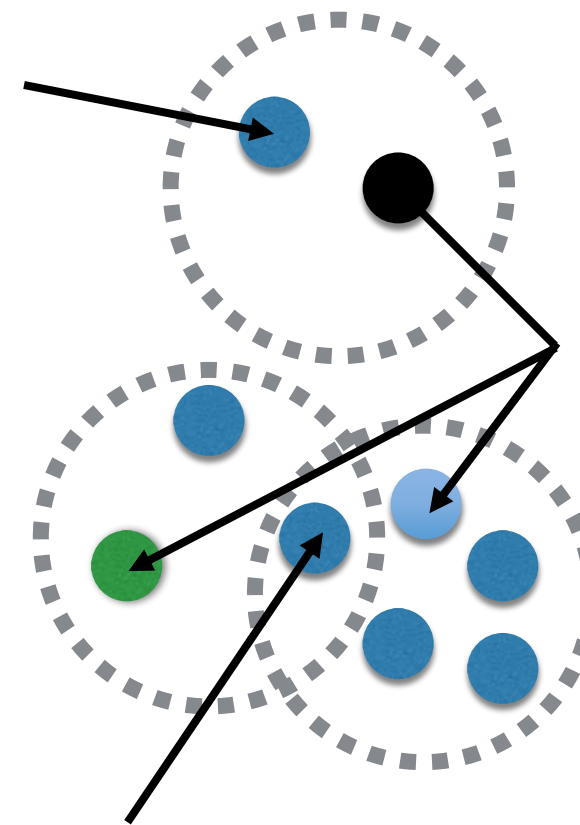


97% Similarity

Amplicon Clusters



> 97% identical



OTU's are 3% different

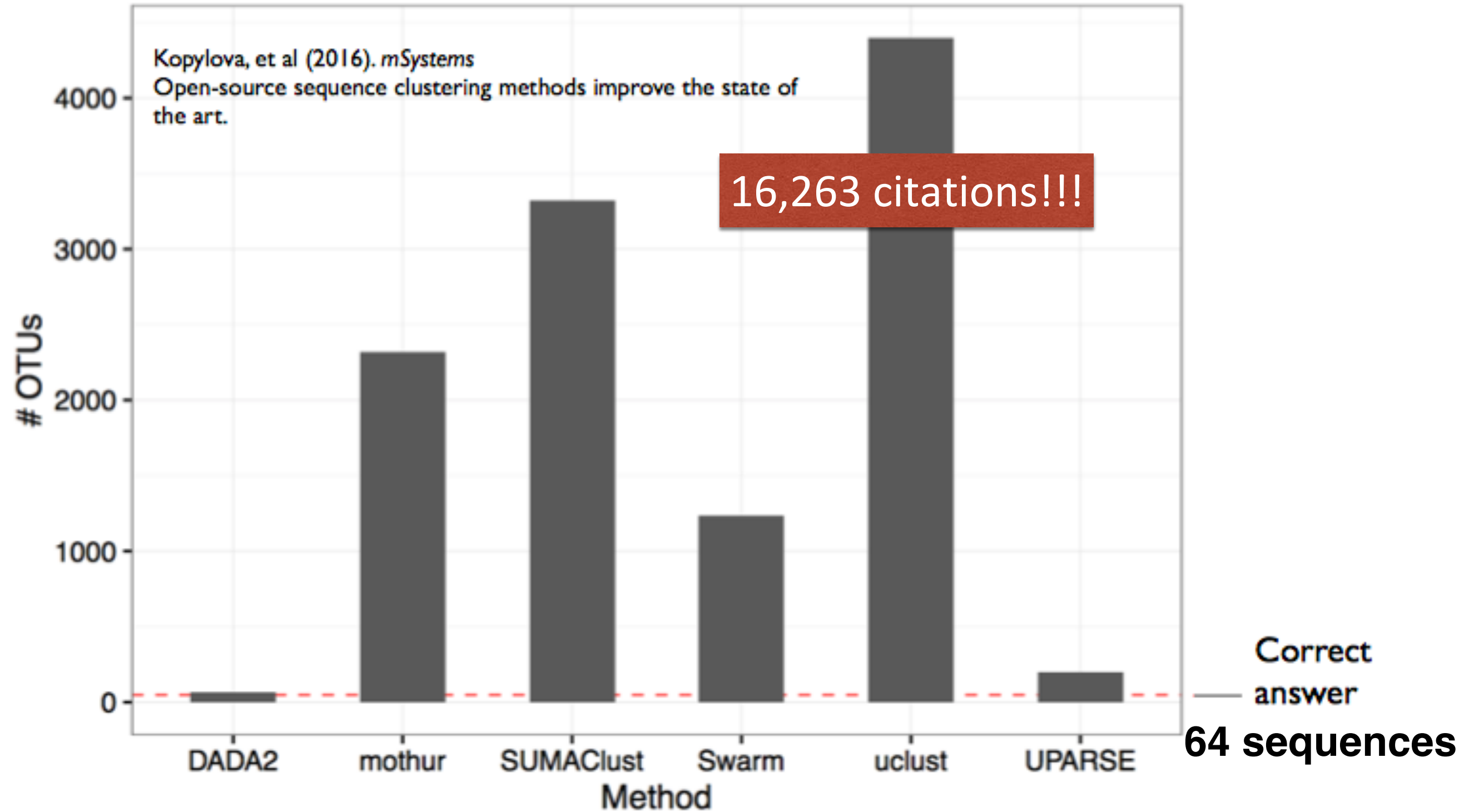
Ambiguous

- **UCLUST**
- **UPARSE**
- **SWARM**
- **SUMACLUSt**
- ***OTHERS***

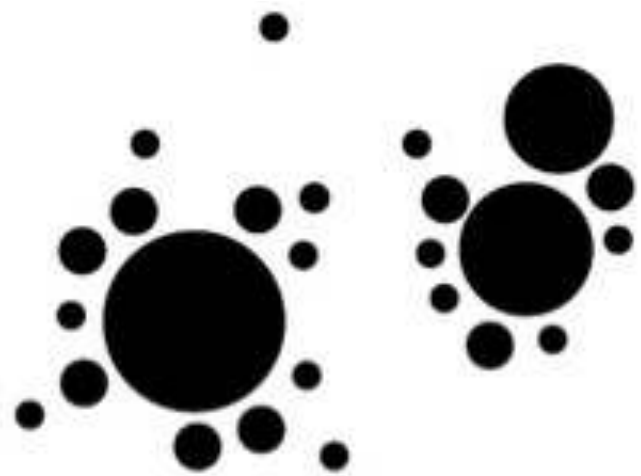
Recognized Problems with Sequence Clustering

- **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
- **Unstable:** Sequence and count frequently depend on input order

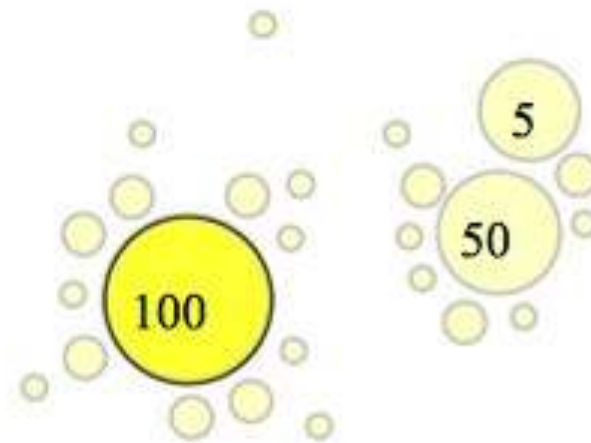
There is some hope



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



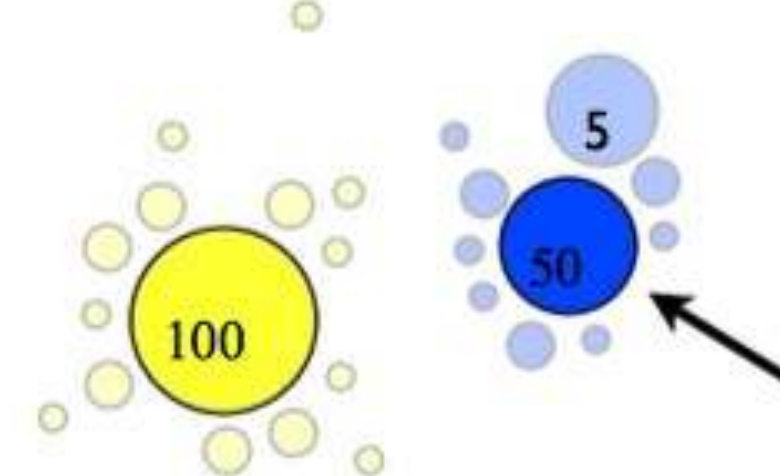
Step 1: Initial guess.
All sequences + errors



Step 2: Initial *error model*

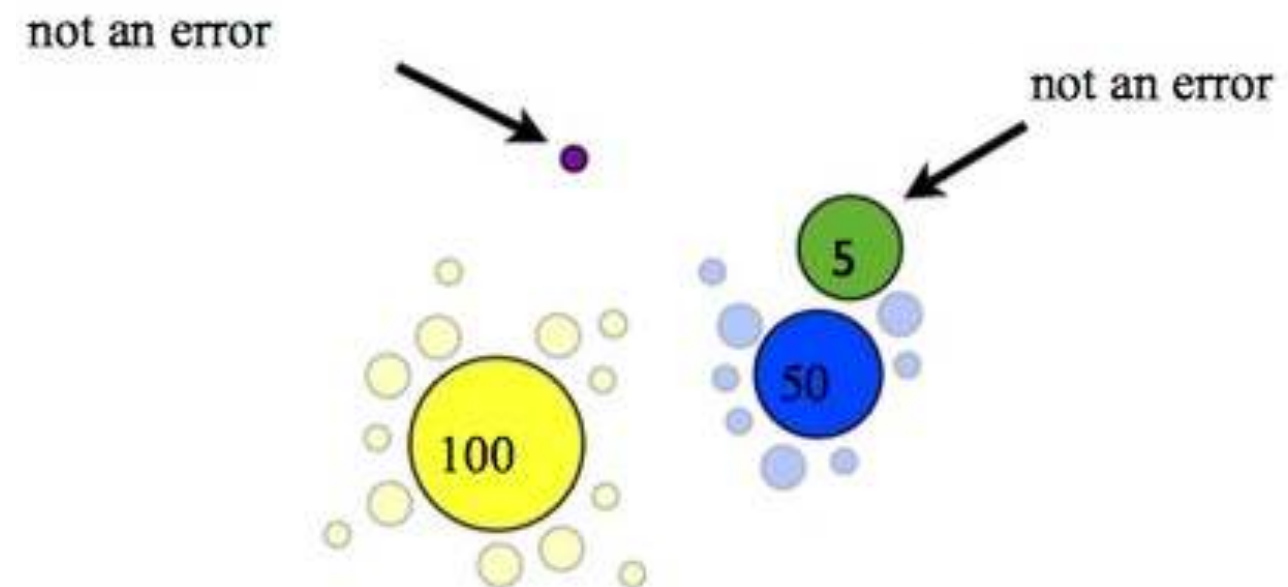
$\text{Pr}(i \rightarrow j) =$

	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

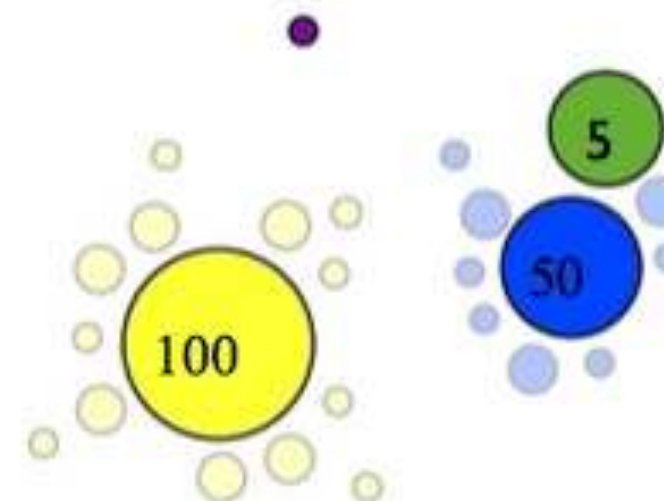


Step 3: Unlikely error under model.
Recruit errors. Update the model

	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

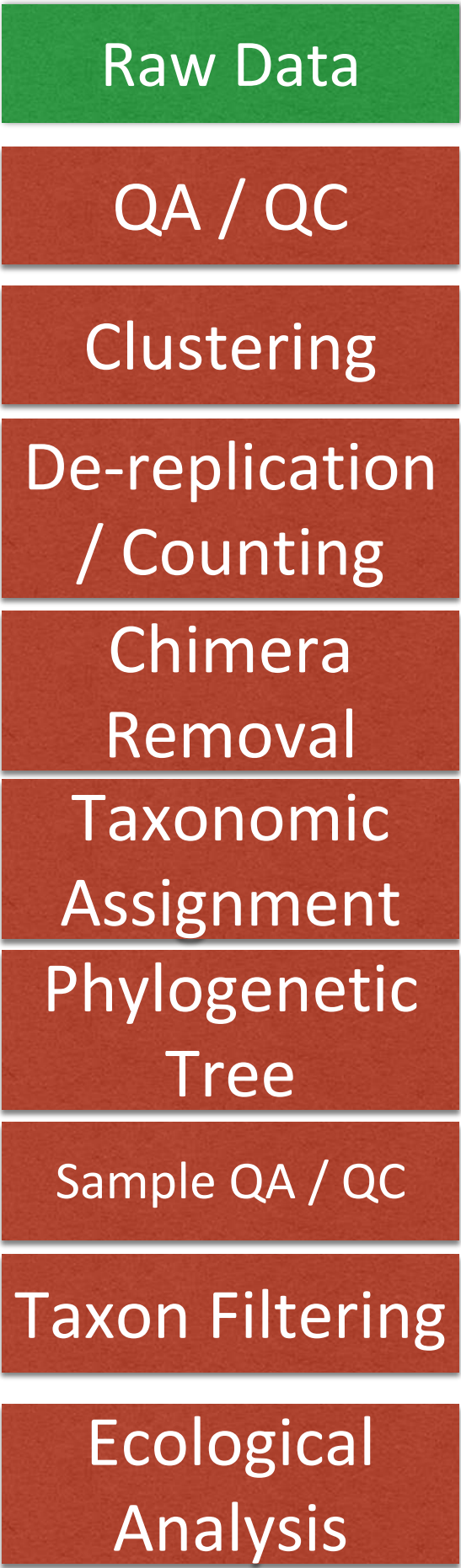


Step 3: Reject more sequences
under new model & update

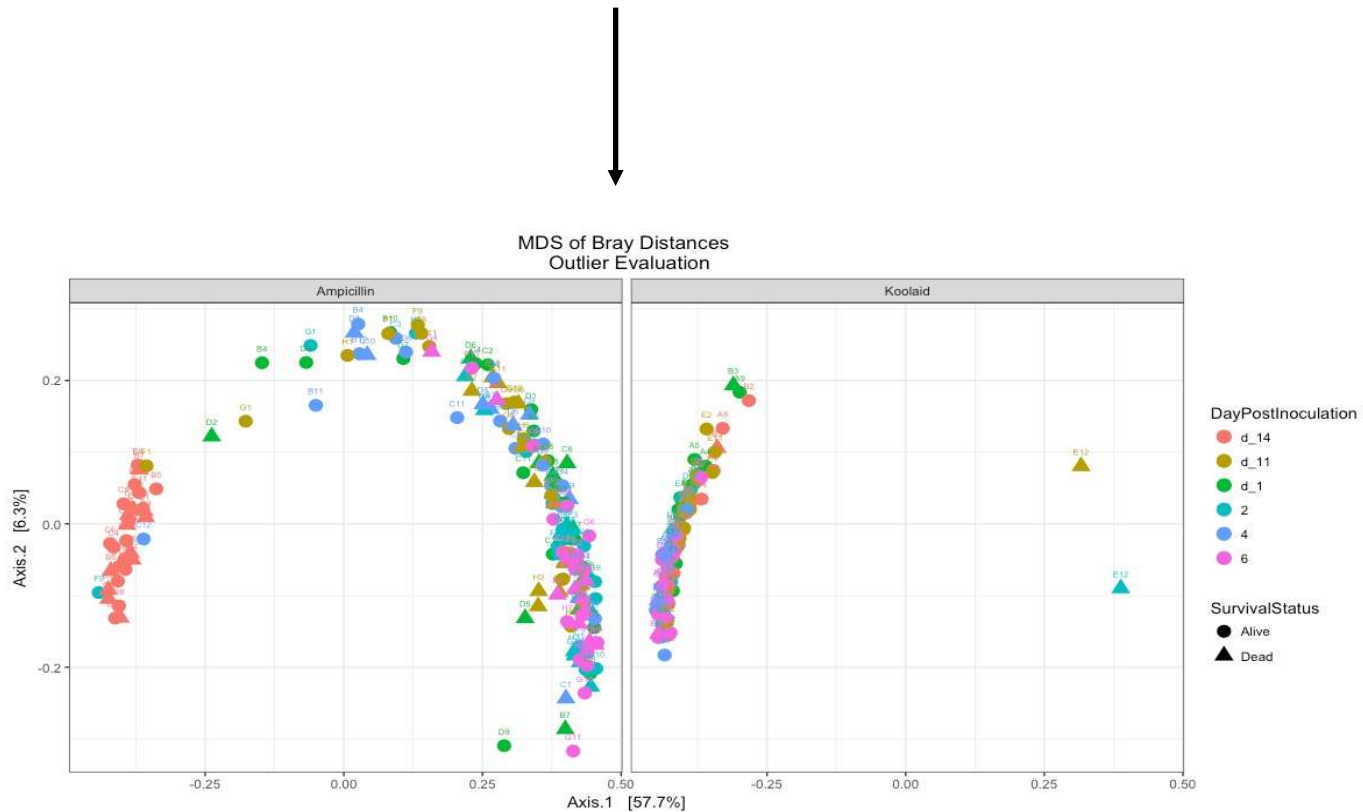


Convergence: All errors are plausible

What does all of this work get you?



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

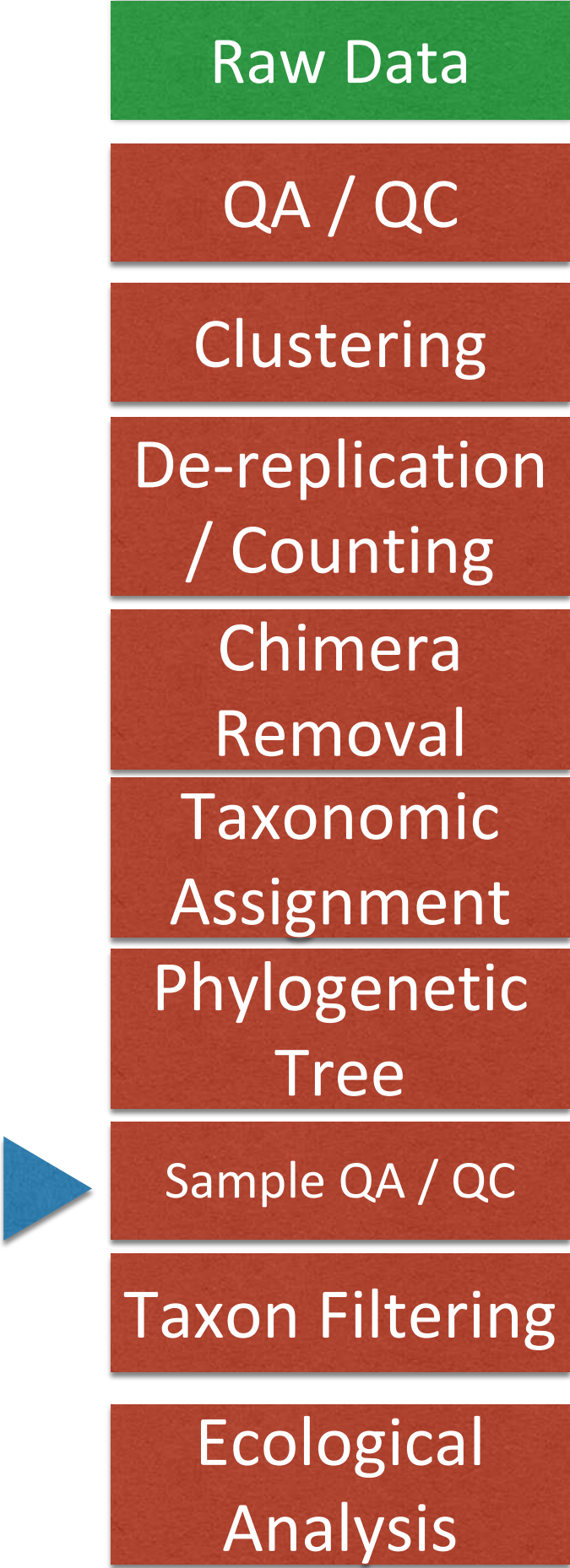
ID	Sample 1	Sample 2	Sample 3
OTU 1	0	0	1
OTU 2	1	0	0
OTU 3	1	0	0
OTU 4	1	1	1

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

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

Sample Outlier Detection



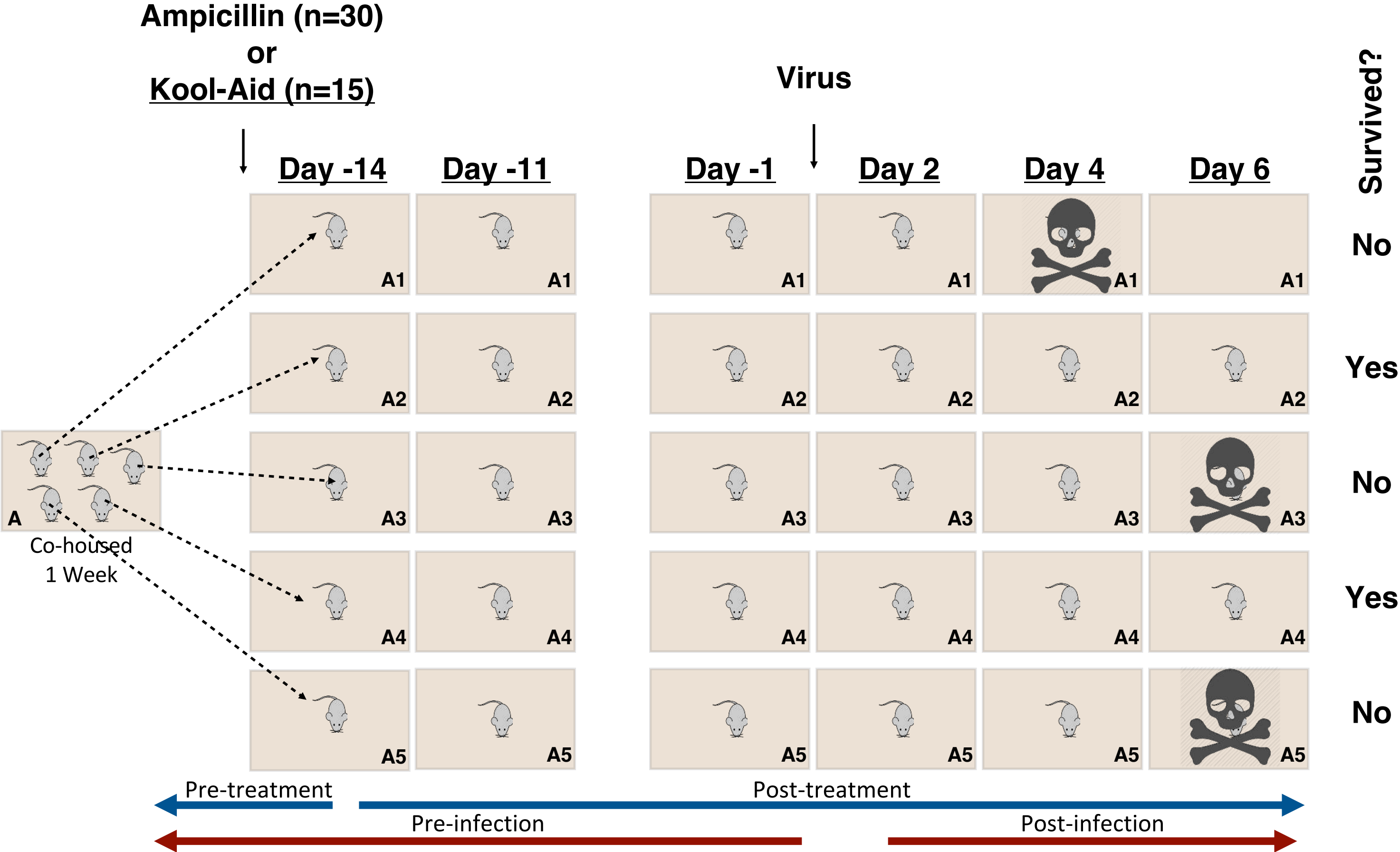
ID	Sample 1	Sample 2	Sample 3	Sample 4
ASV 1	0	0	2	0
ASV 2	12	8	8	456
ASV 3	112	101	98	10
ASV 4	435	435	382	3
ASV 5	76	83	68	145

...
n=724

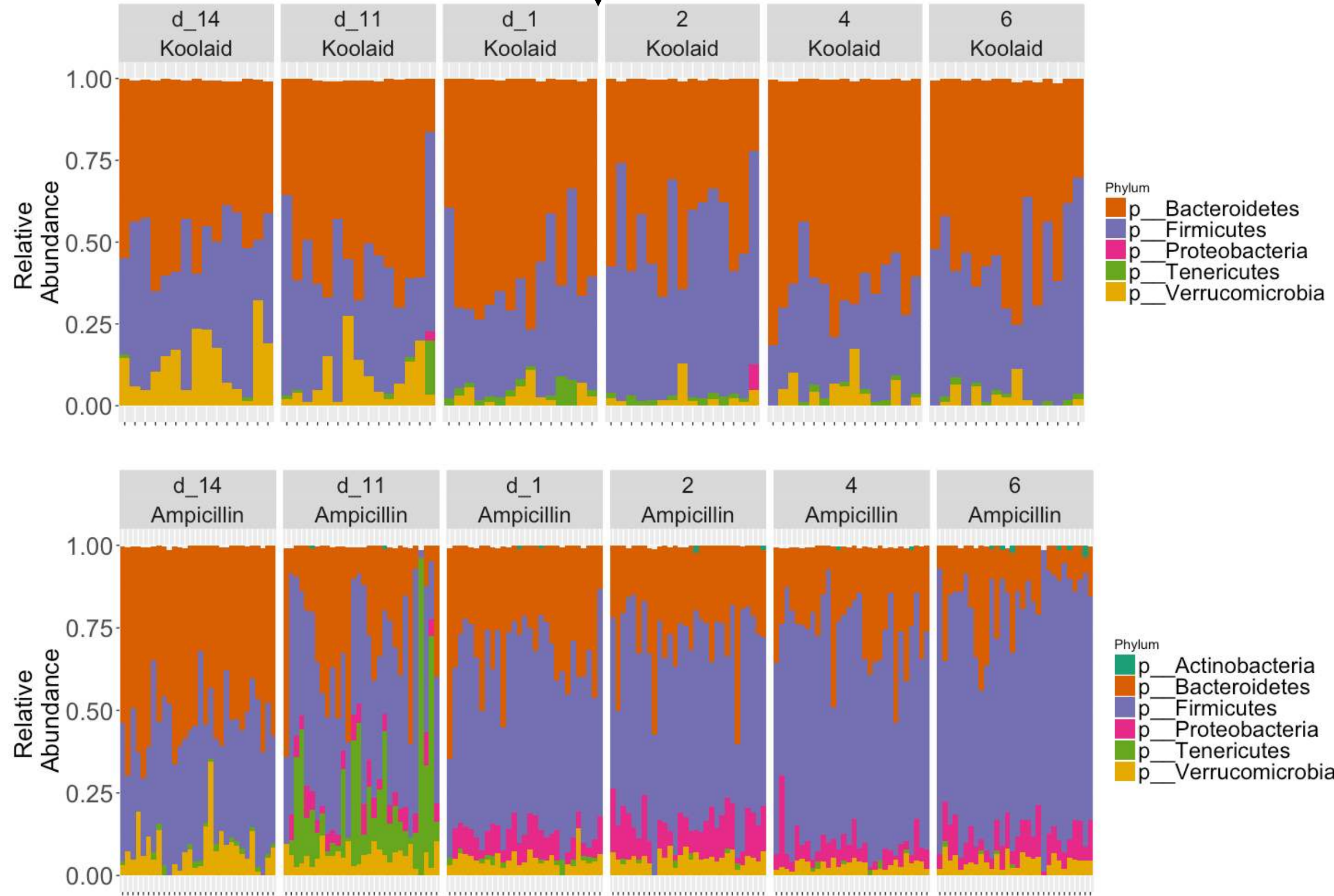


... n=270

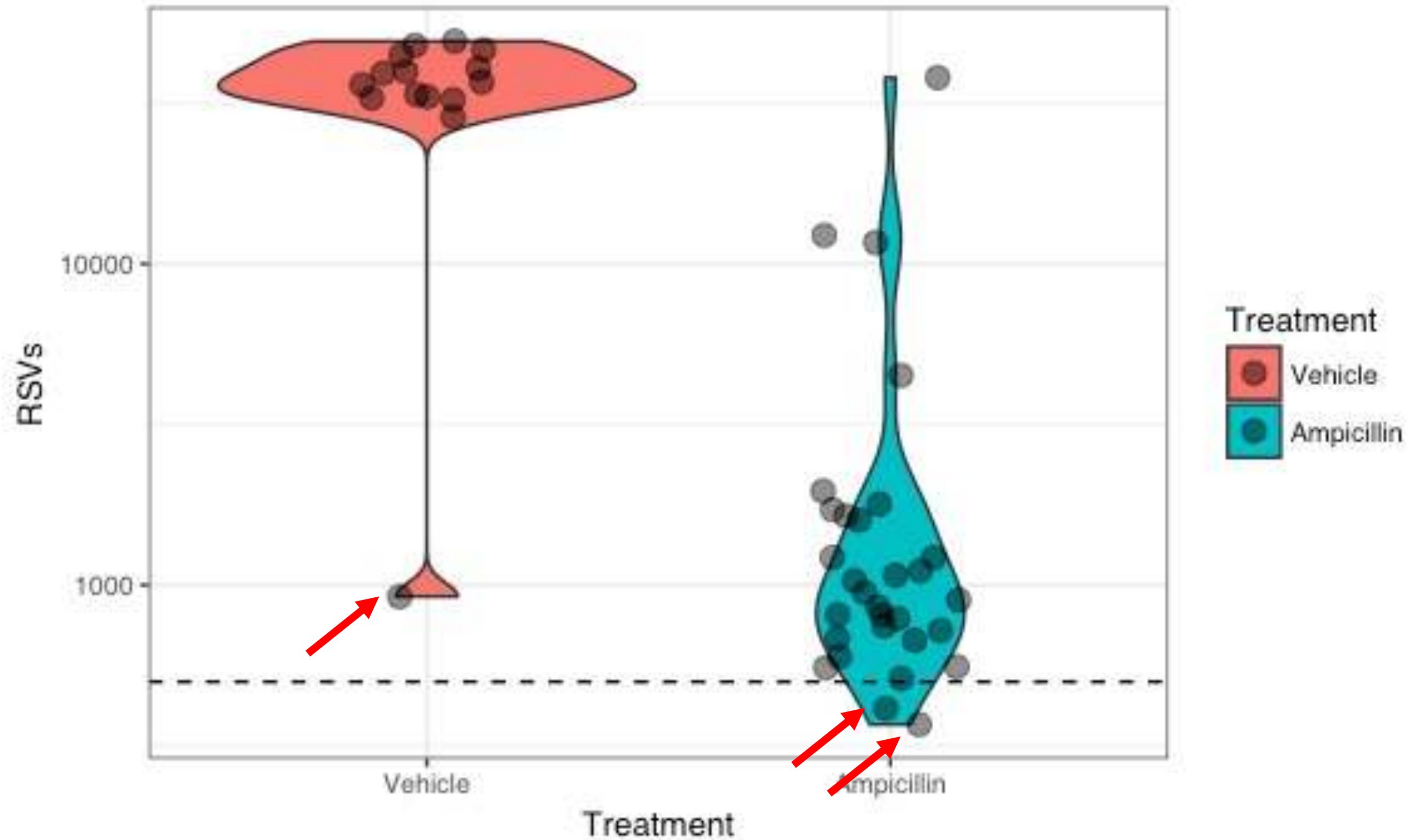
Individual Mouse Isolation Schema



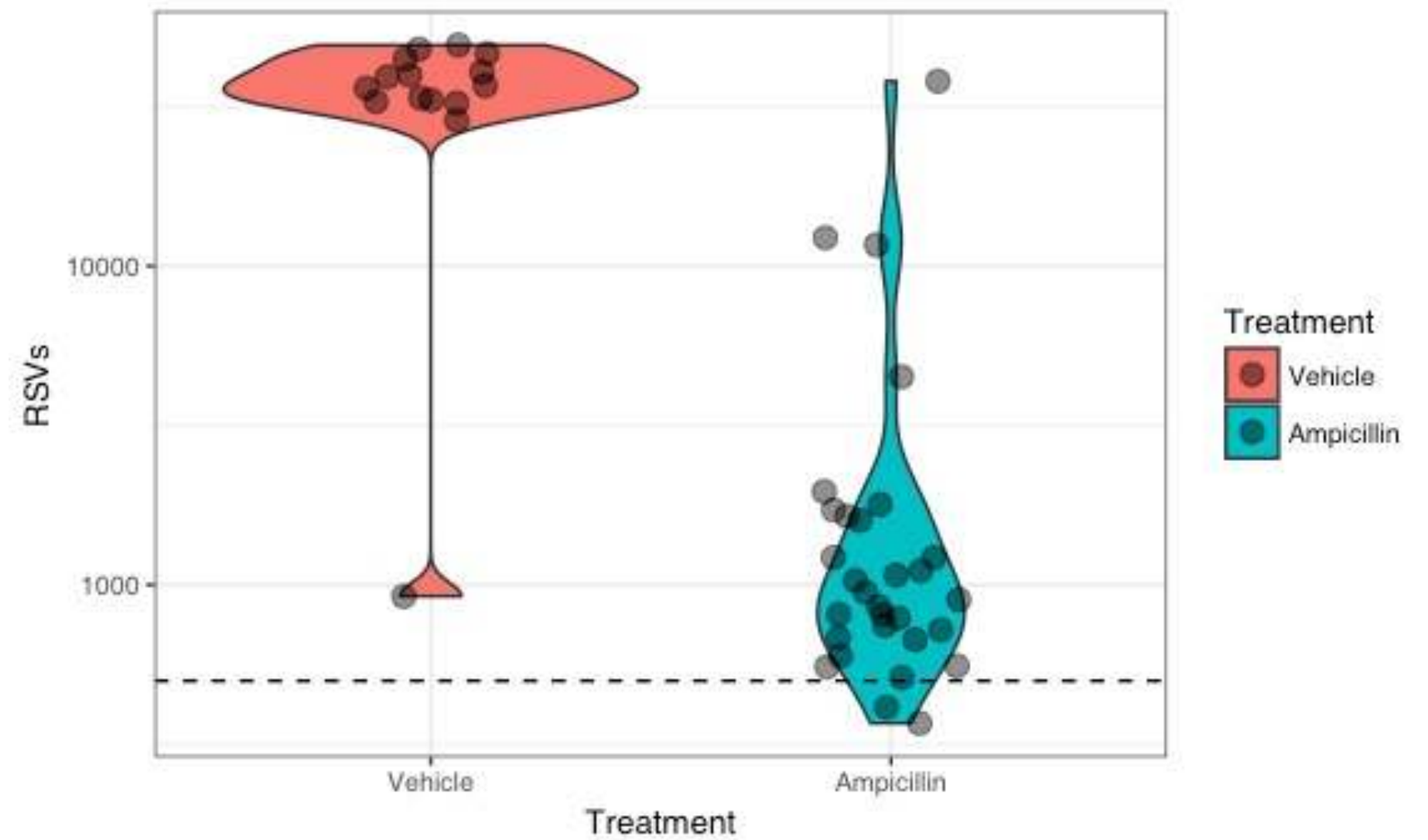
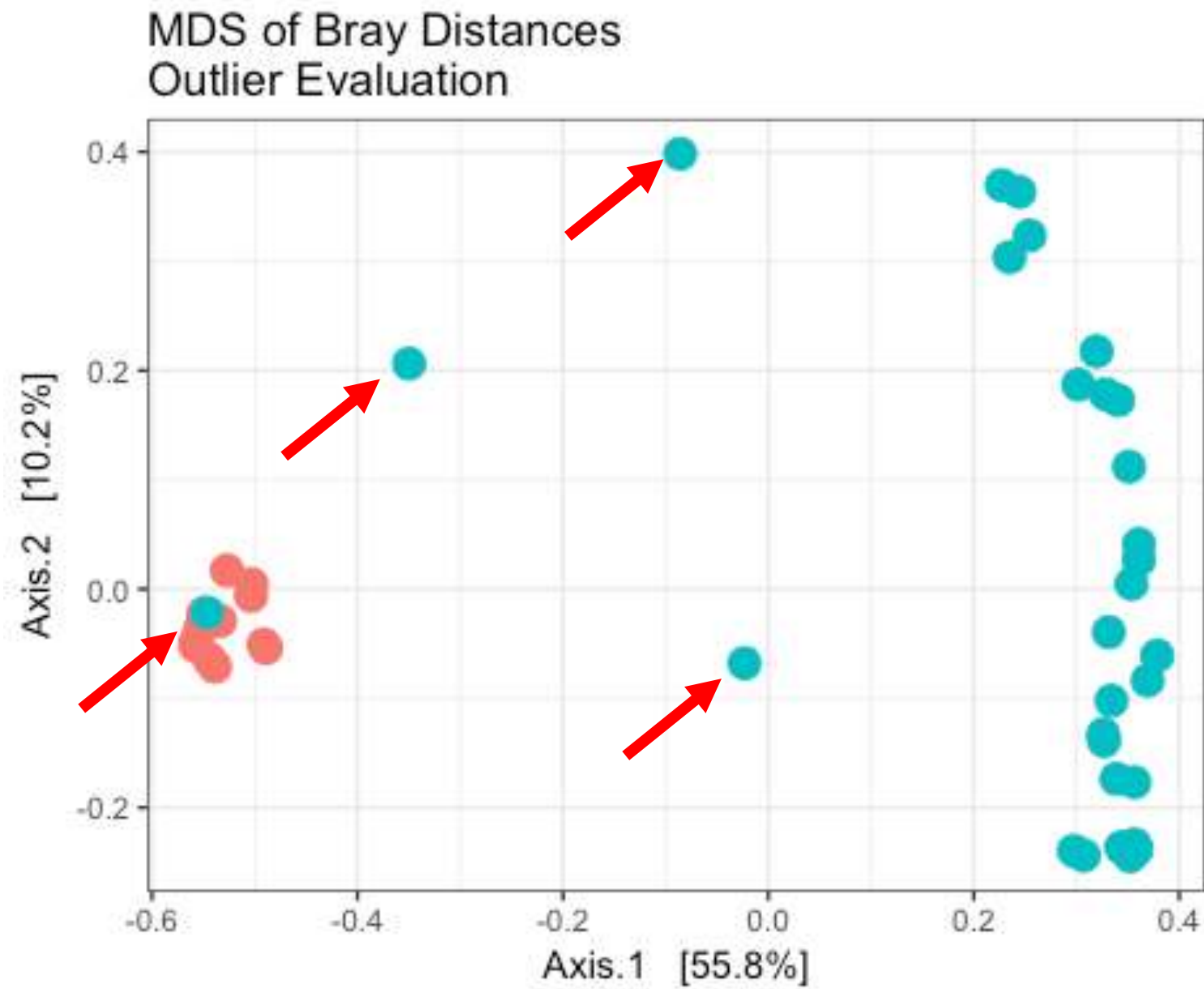
Virus



Sample Outlier Detection – Unexpectedly Low # of Sequences



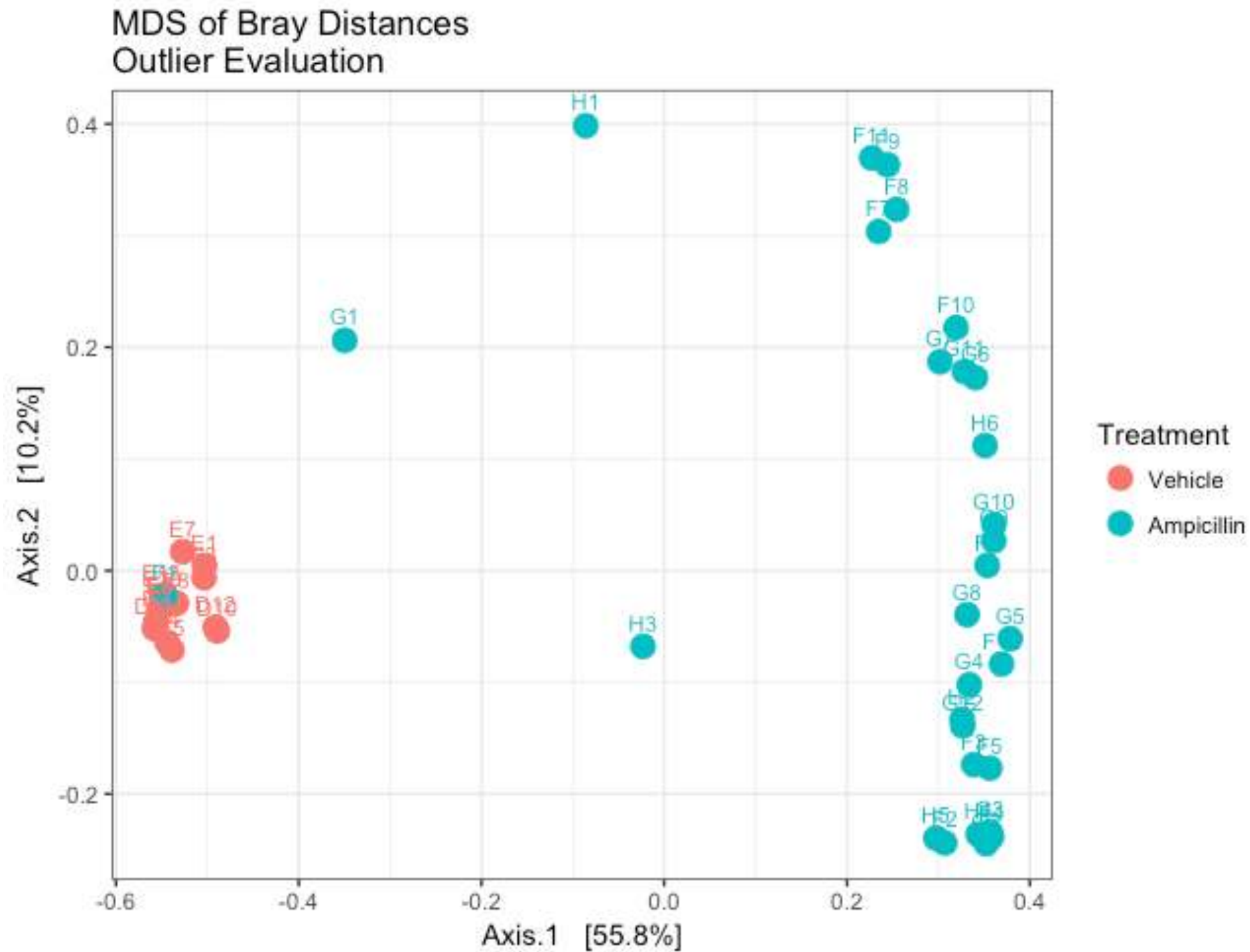
Samples that “perform” unexpectedly



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

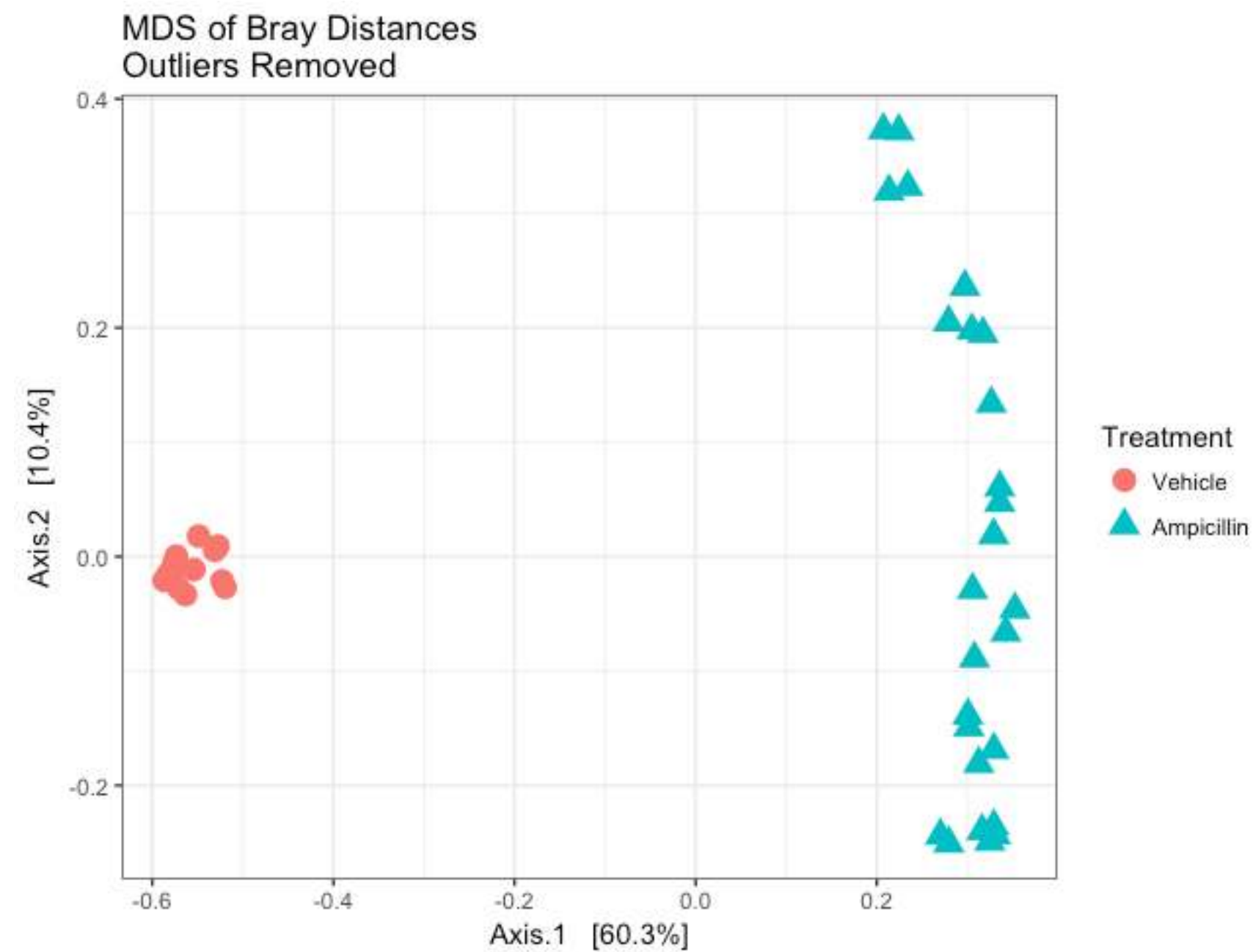
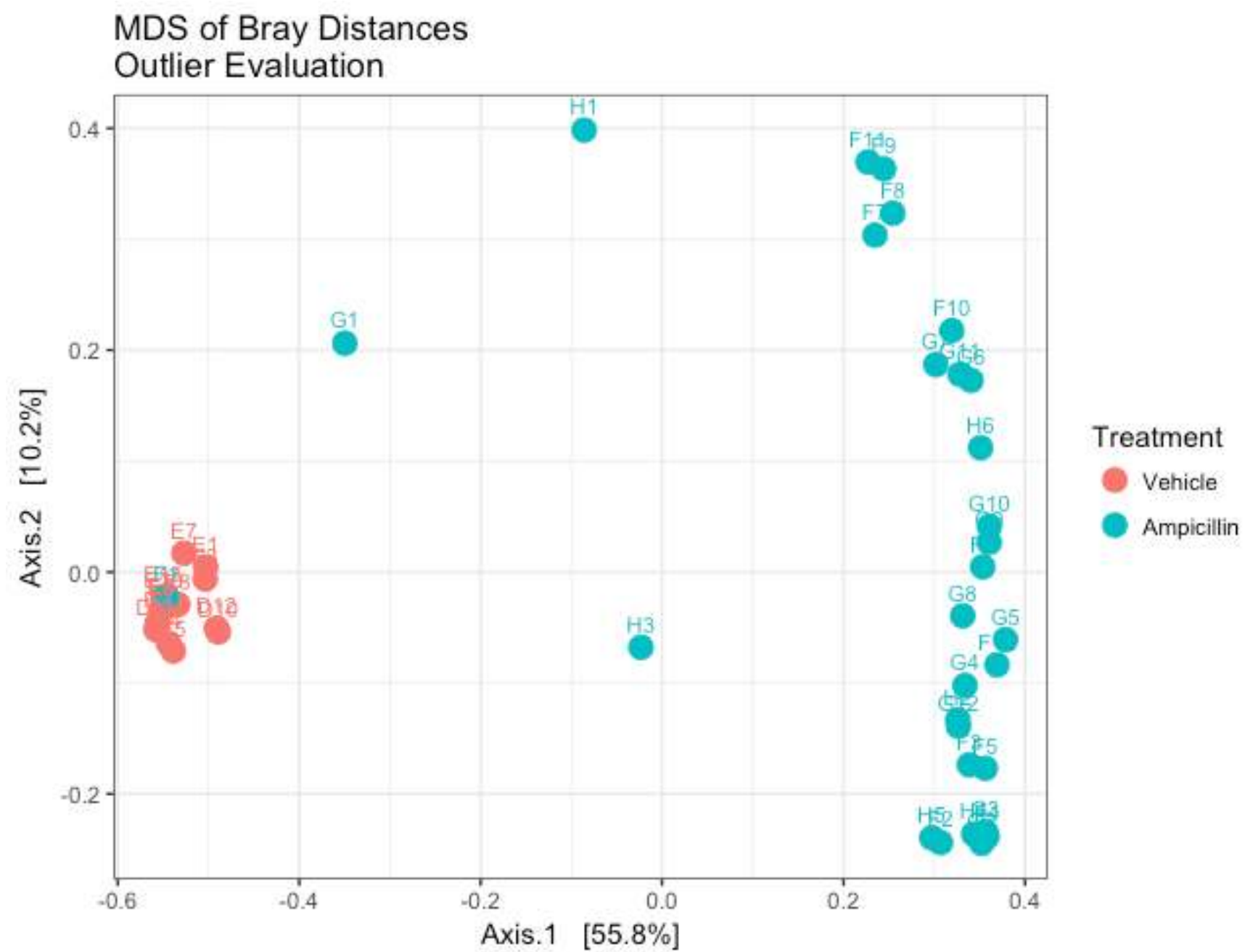
Understand your data better



	1	2	3	4	5	6	7	8	9	10	11	12
A												
B												
C												
D												
E												
F	*											
G	*											
H	*		*									

Copyright © 2009 Edita Aksamitiene

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

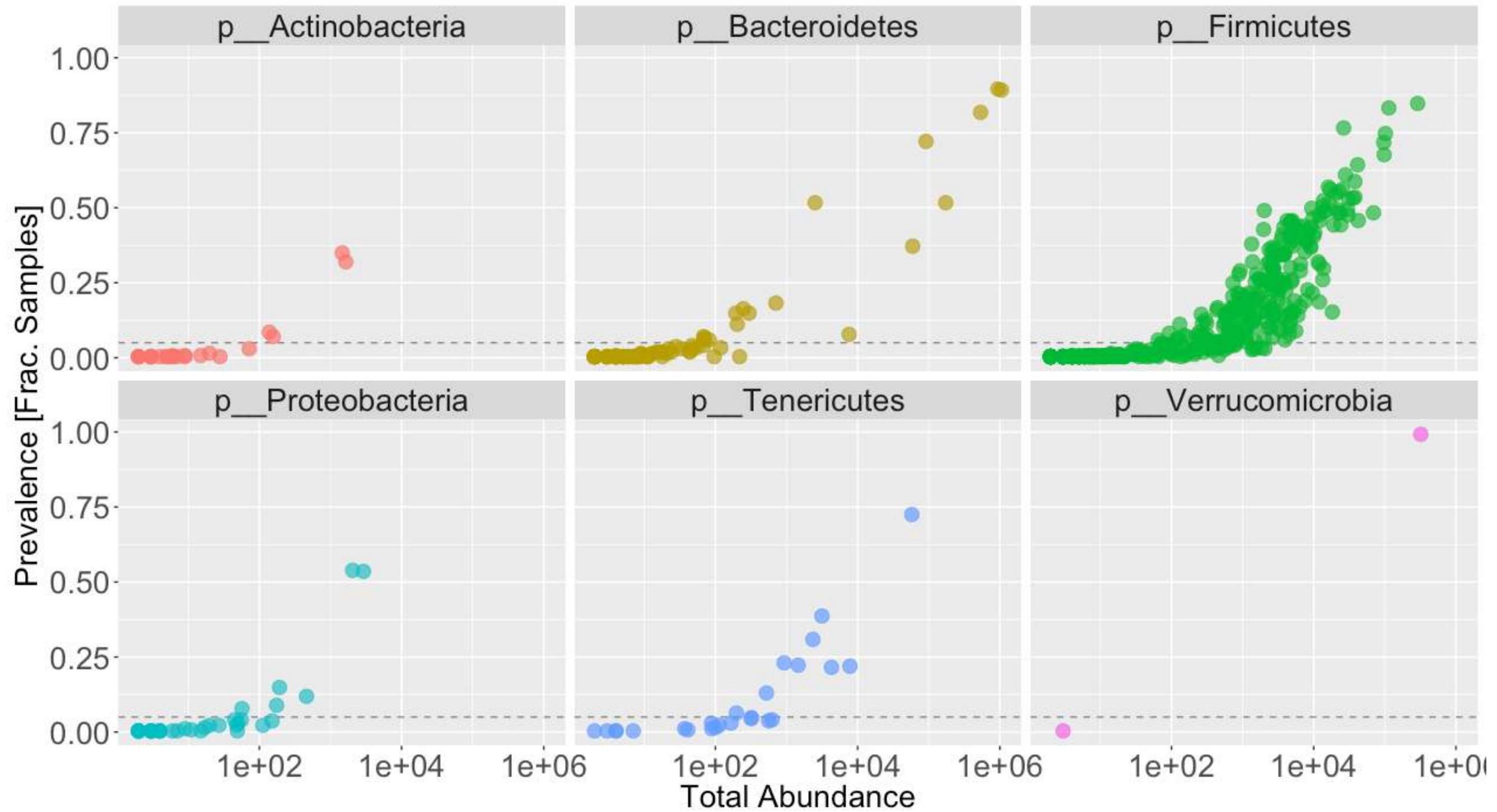

...
n=724

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



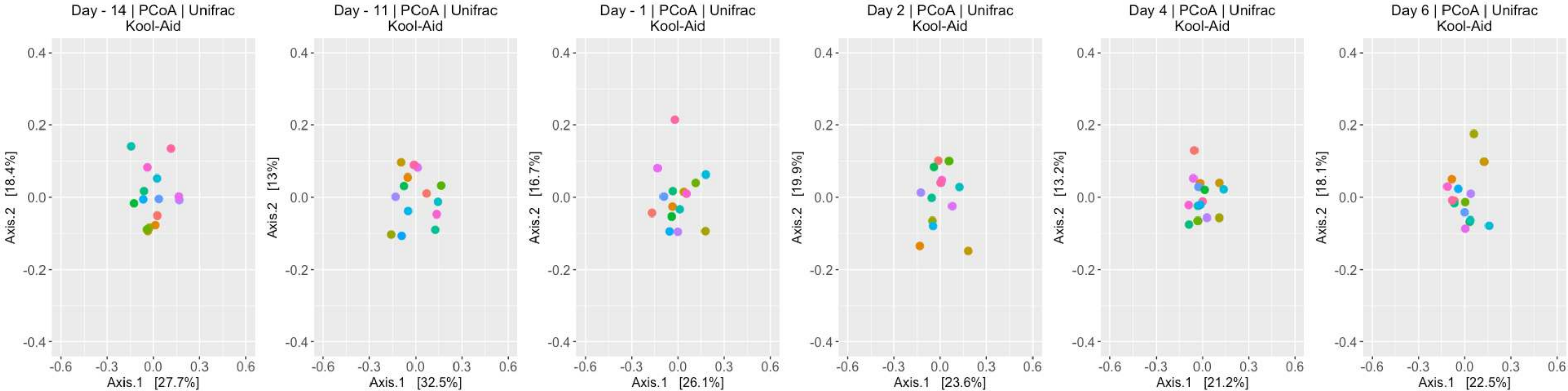
Rules of Thumb for Feature Detection and Removal

- **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

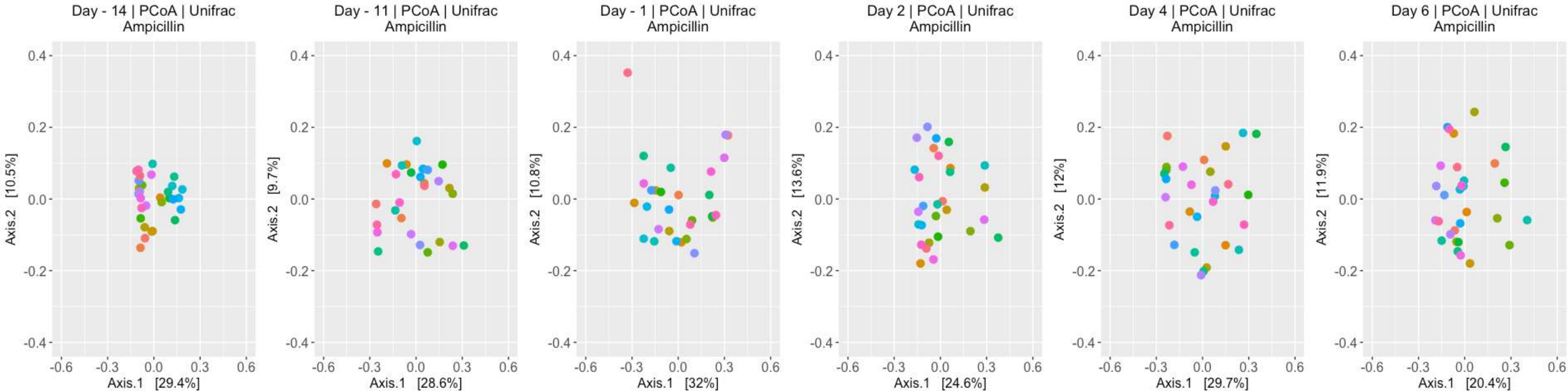
Beta Diversity Throughout the Course of the Experiment

Colored by Cage

Kool-Aid



Ampicillin



Summary

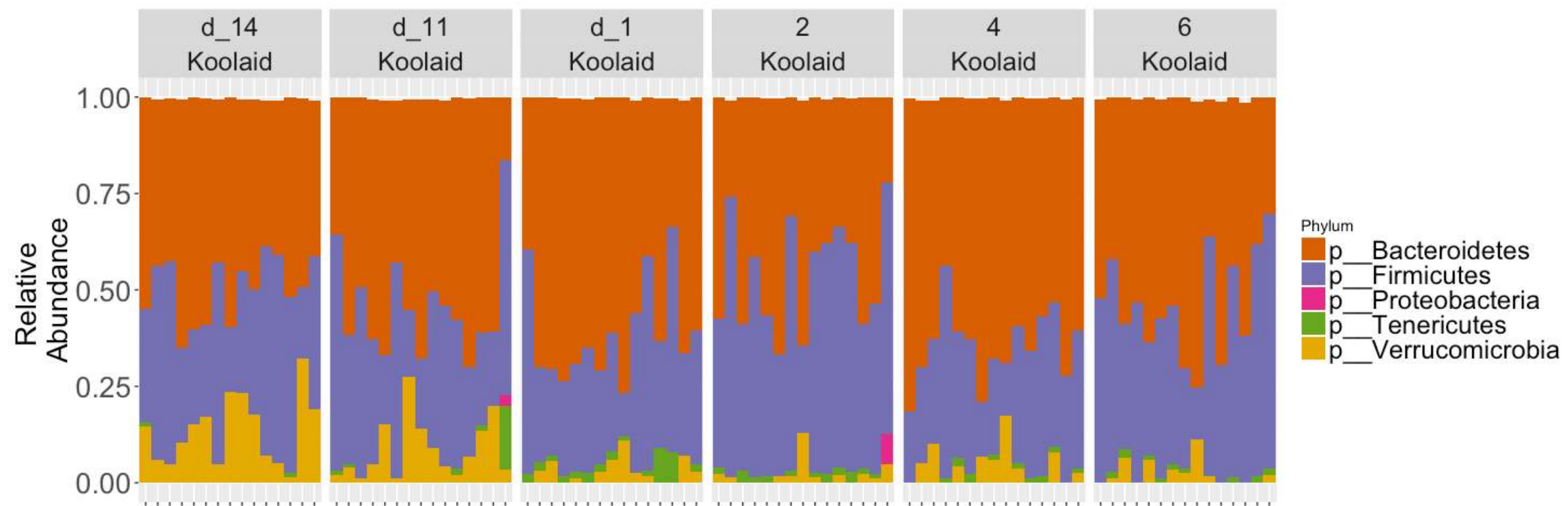
- Explore -> Document -> Test
- Does any of this really matter?
 - Sometimes?
 - Less so for community ecology measurements
 - More so for detection of differentially abundant taxa
 - Detailed exploration provides more opportunities for insights
 - Don't publish garbage data

Frequently Used 16S rRNA Analysis Techniques

Also used for other techniques (ITS, 18S, metagenomics, etc.)

Community Composition

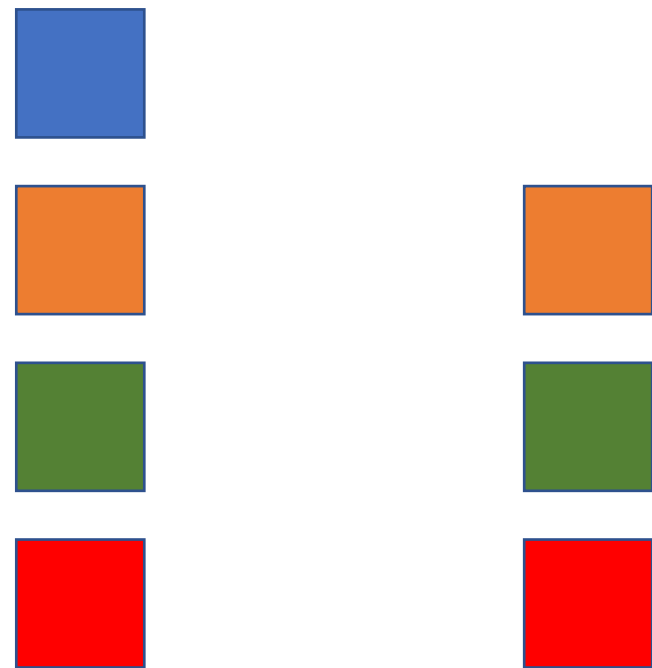
- Broad overview
- Nothing statistical



Alpha Diversity: Within Sample Diveristy

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

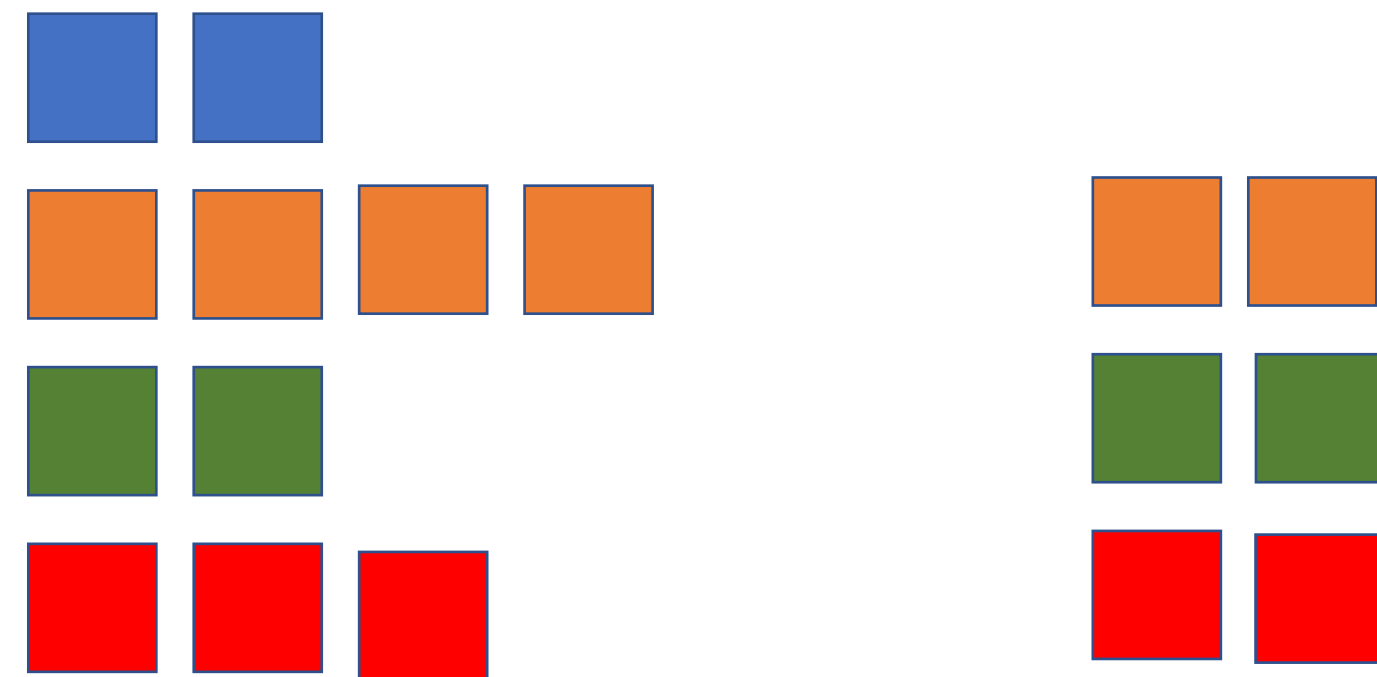
Richness



Sample 1

Sample 2

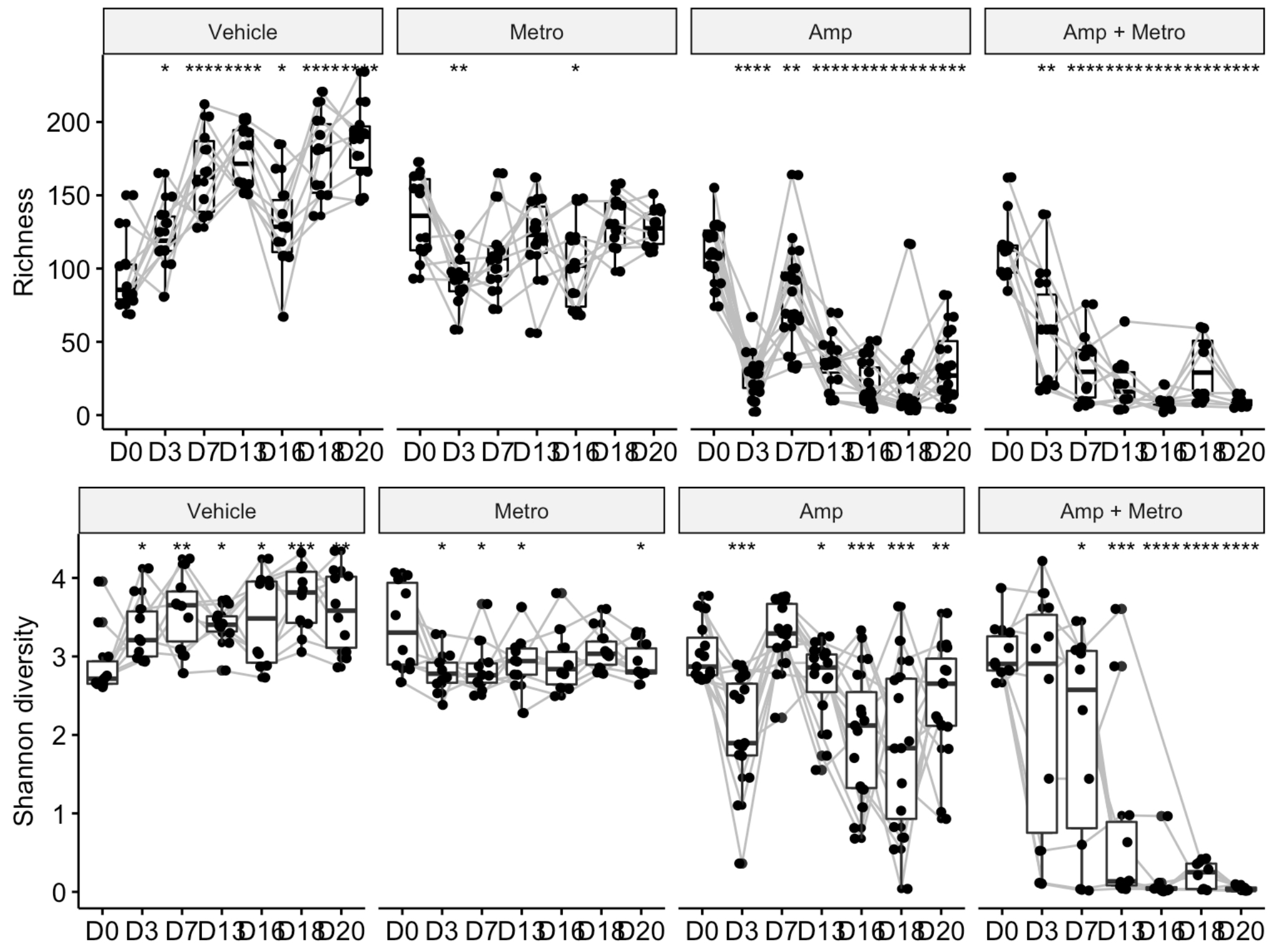
Diversity



Sample 1

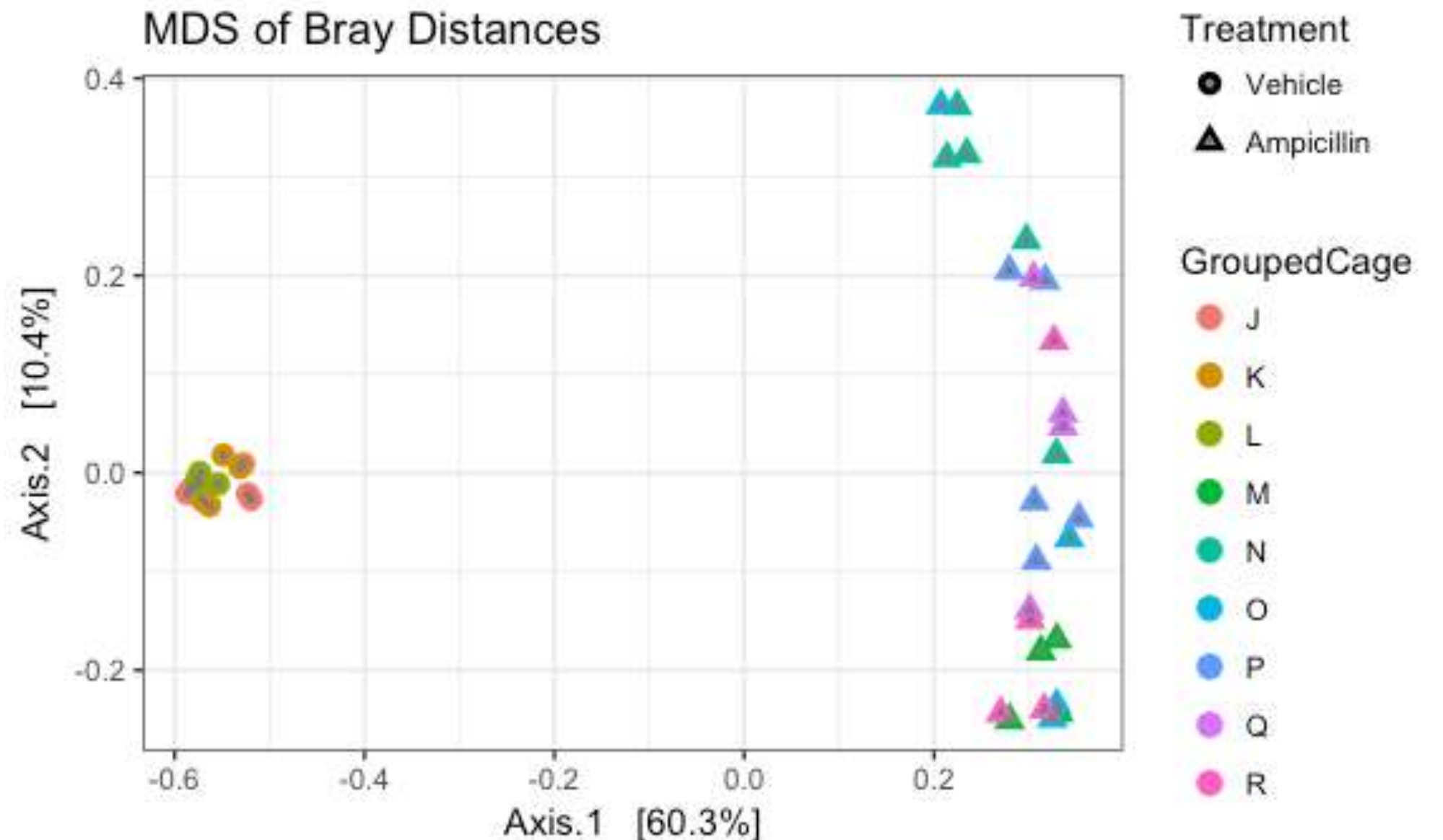
Sample 2

Richness Example



Beta Diversity: Between Sample Similarity

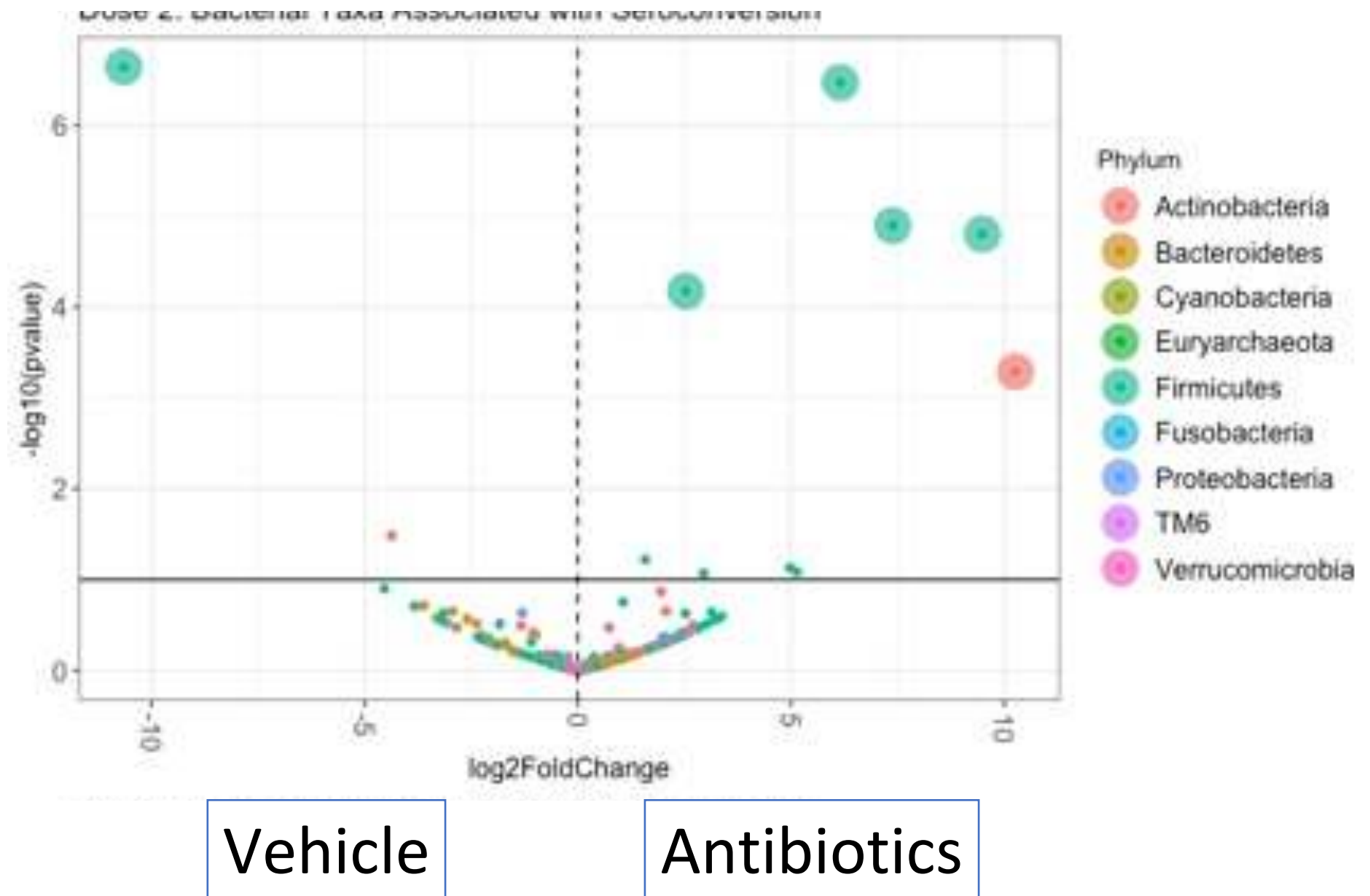
- Distance between one sample to all other samples
- Multivariate
- Can incorporate relative abundances or not
- Can incorporate phylogenetic relatedness or not
- Most frequently displayed in an ordination plot



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

Differential Abundance Analysis

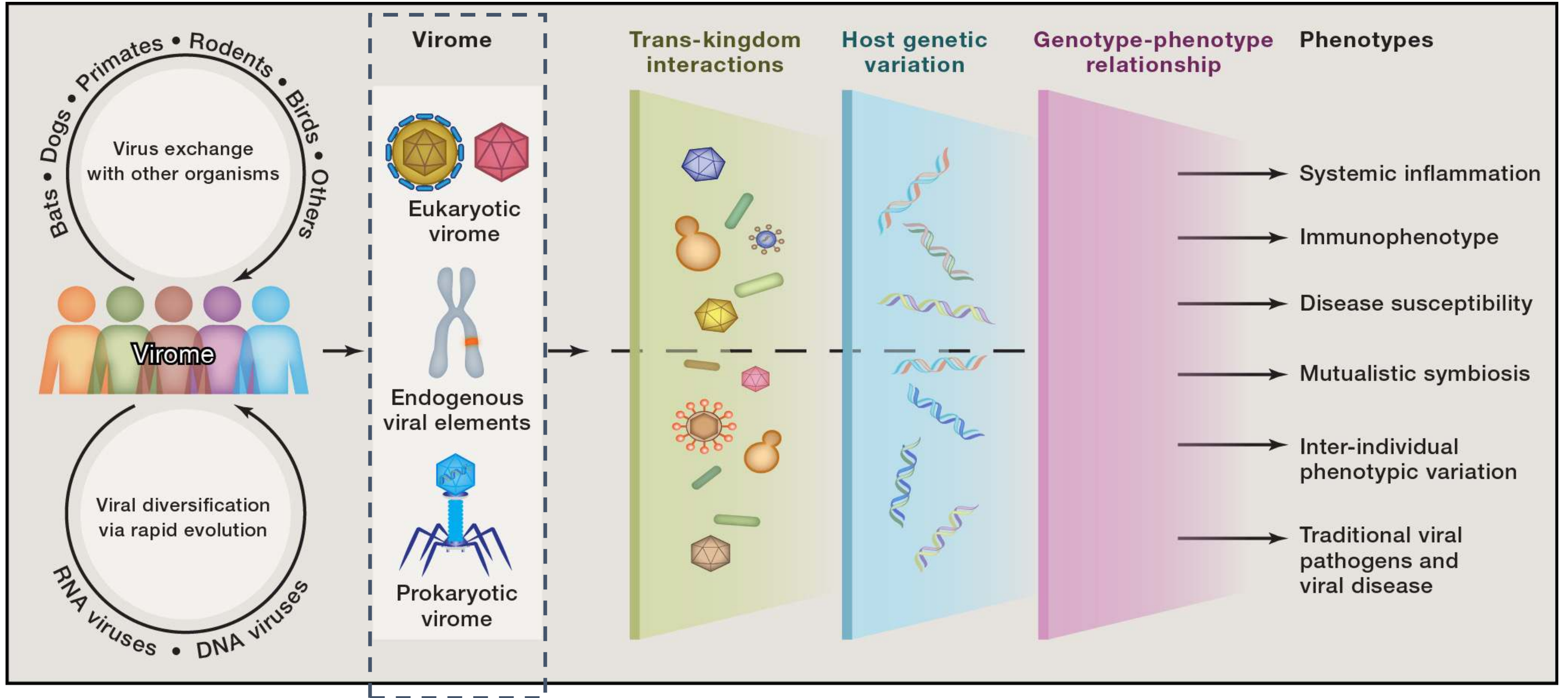
- What specific taxa are different between study groups?
 - Lots of methods
 - DeSeq2
 - Random Forest
 - LeFse
 - ANCOM
 - Gneiss
 - ...



Break For Part 1

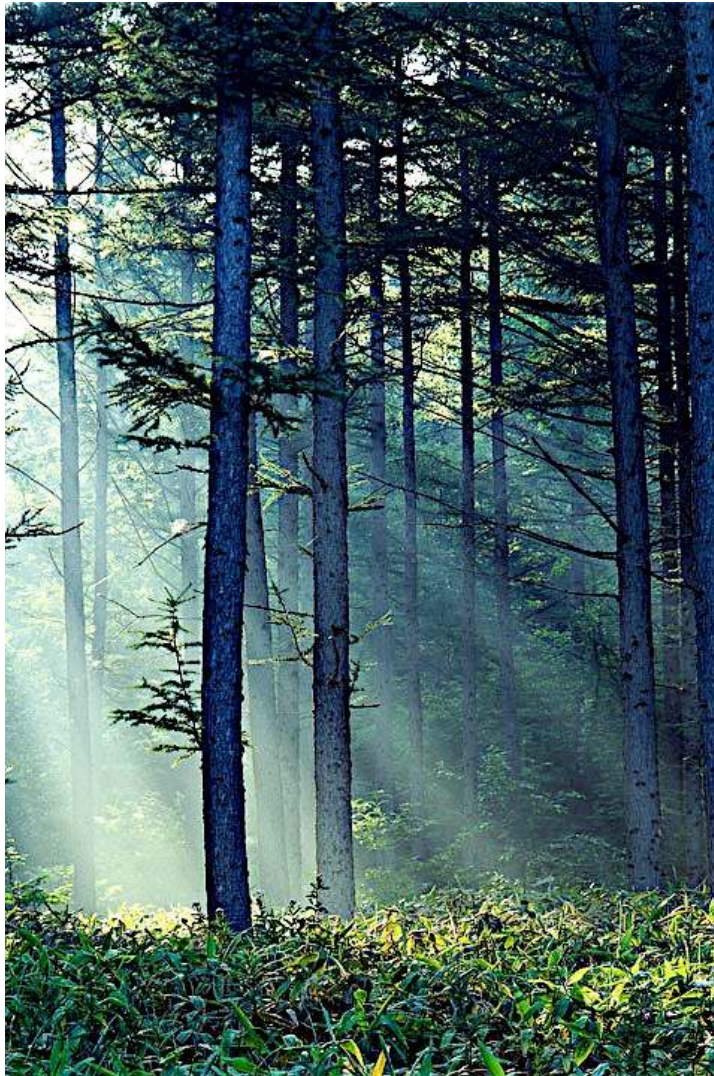
Part 2: Virome Analysis

What is the Virome?

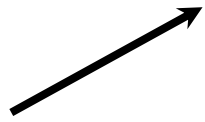


Virome vs. Infection

VIROME



Bacteriophage



Health or Disease

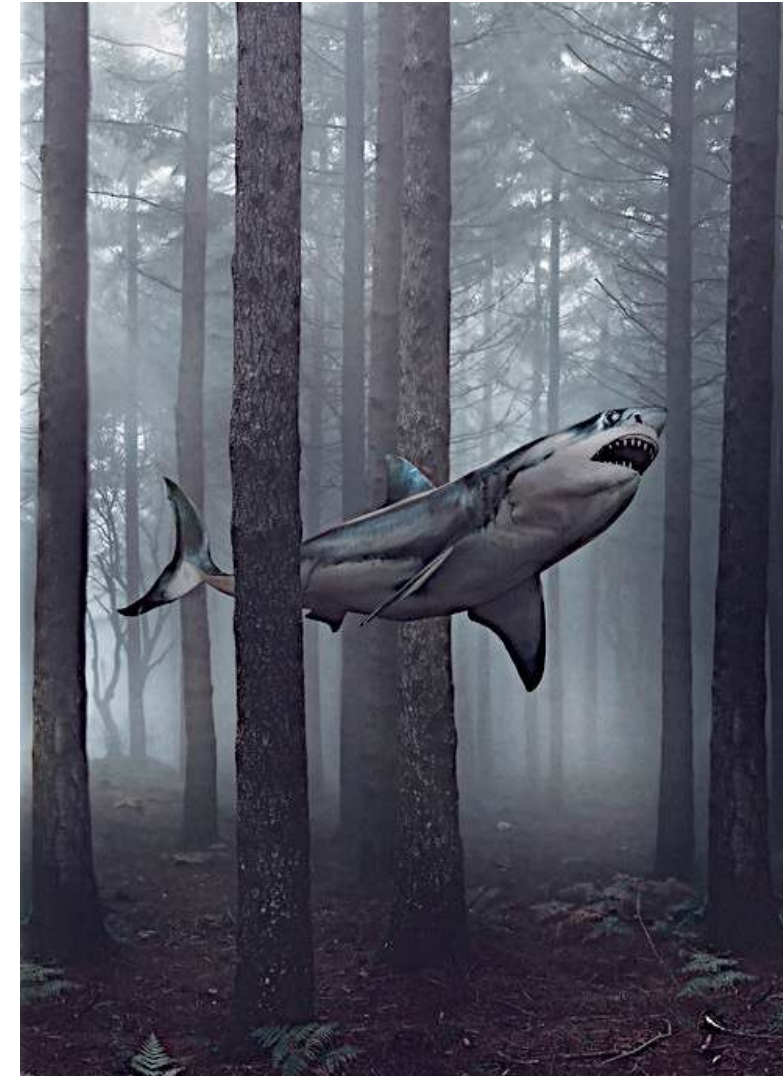


Who is there?

Are they "normally" there?

Why are they there?

INFECTION



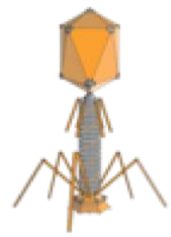
Novel or Known
Pathogen Detection /
Discovery



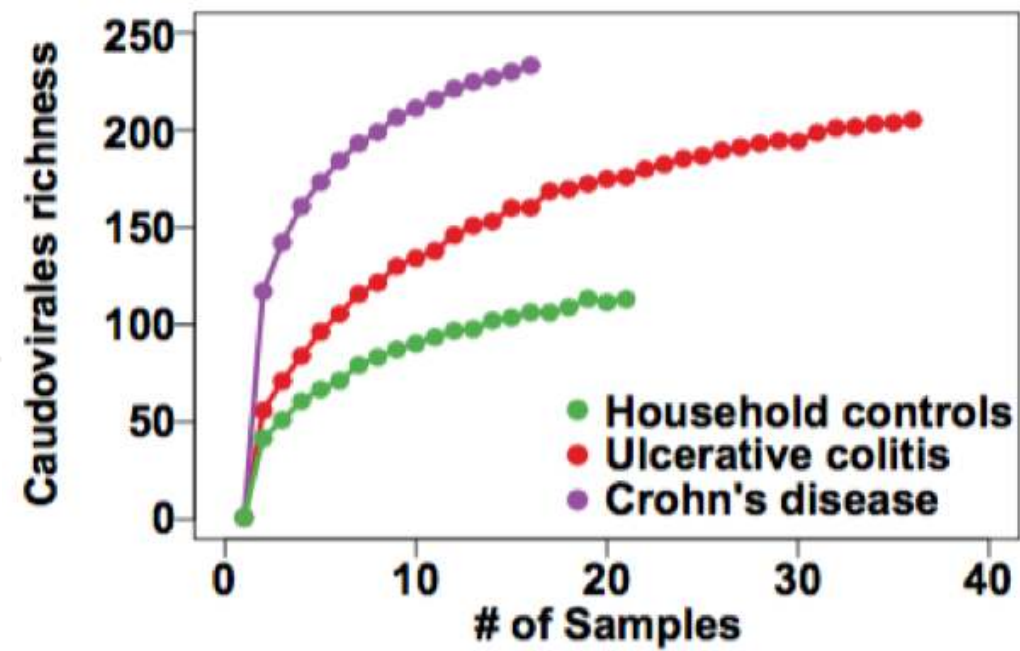
Health or Disease



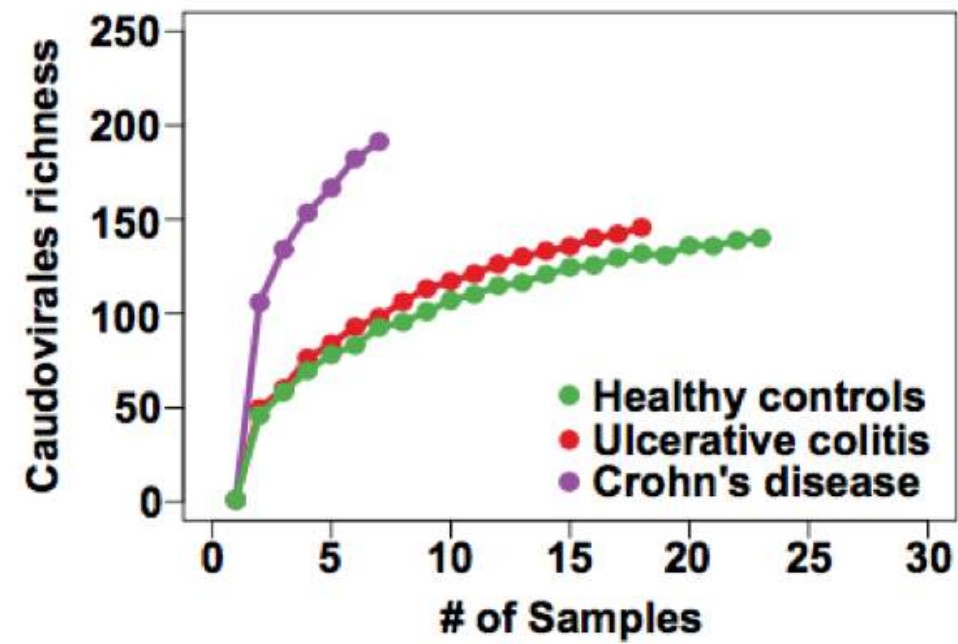
Virome Association with Health / Disease



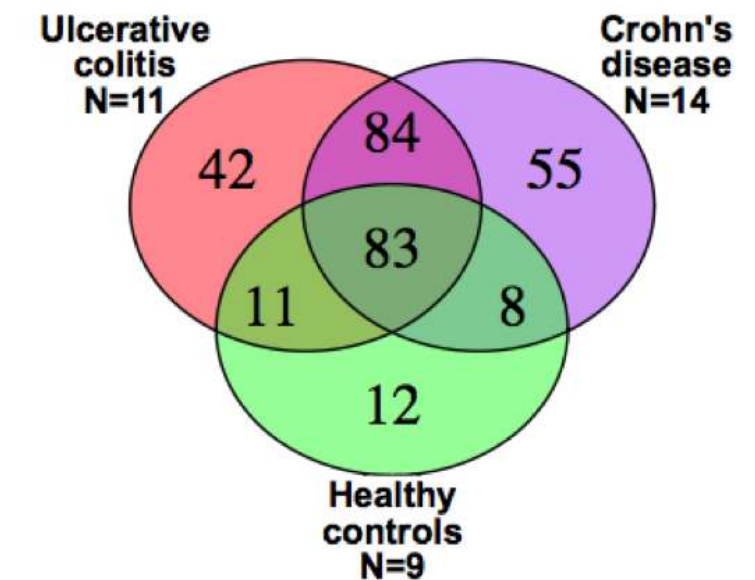
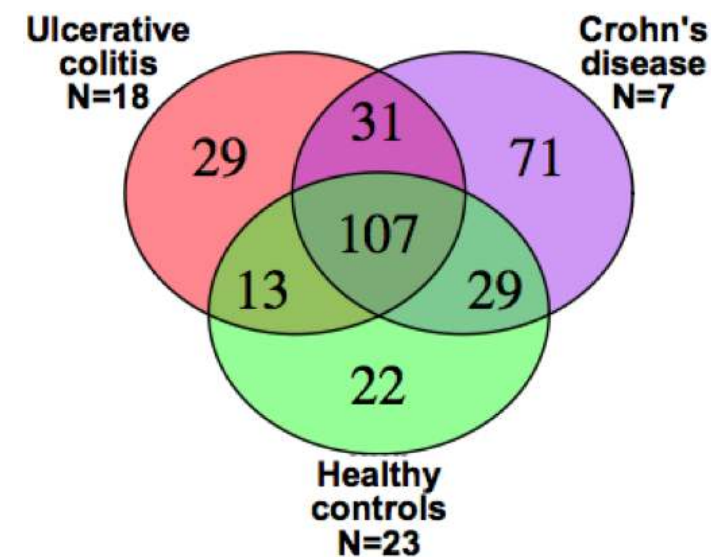
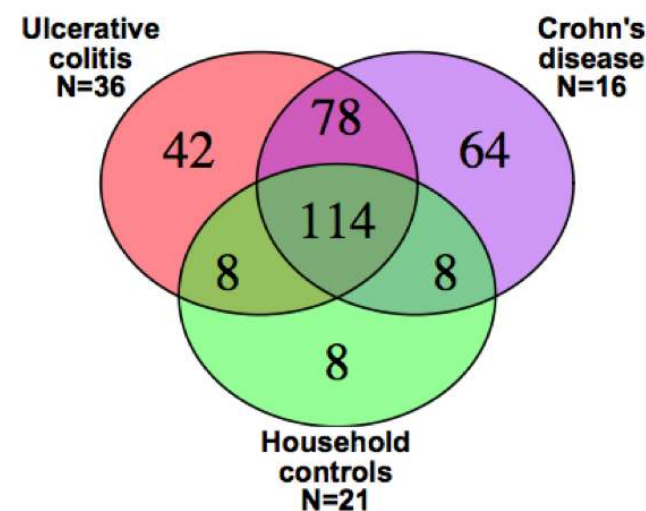
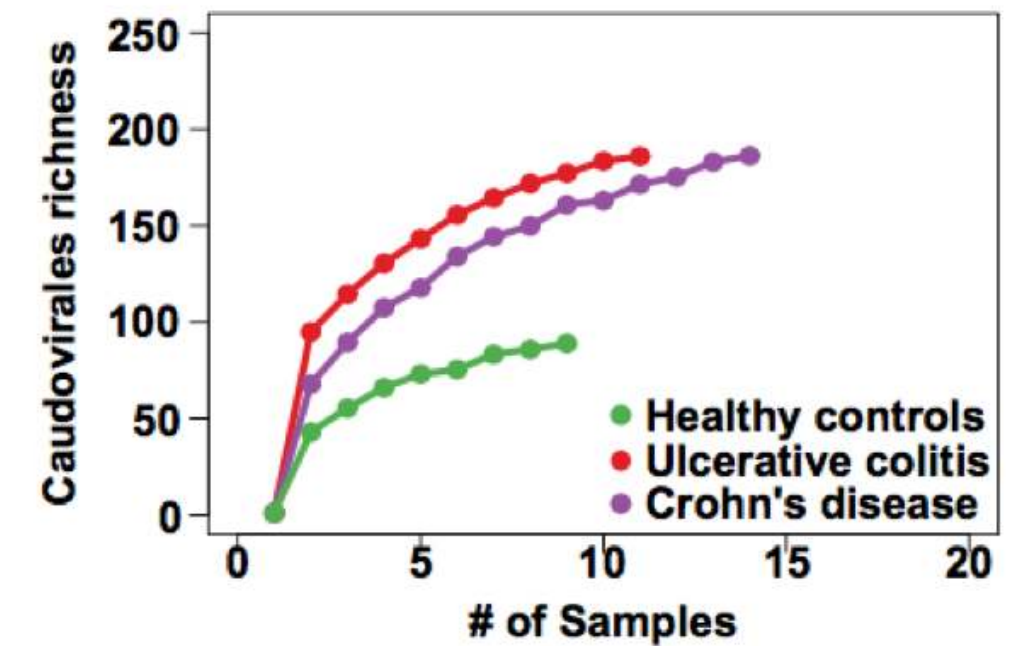
United Kingdom



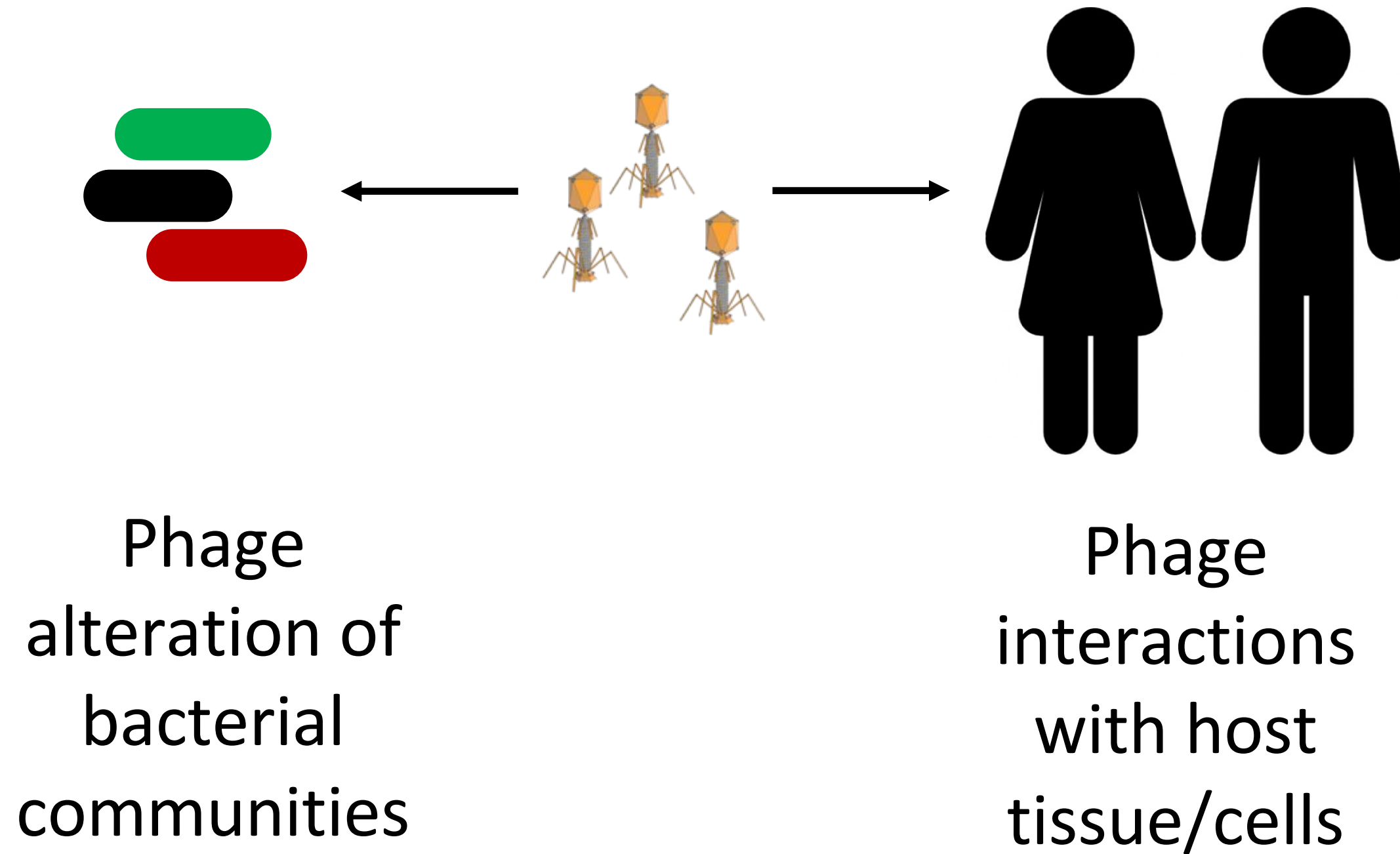
Chicago



Boston

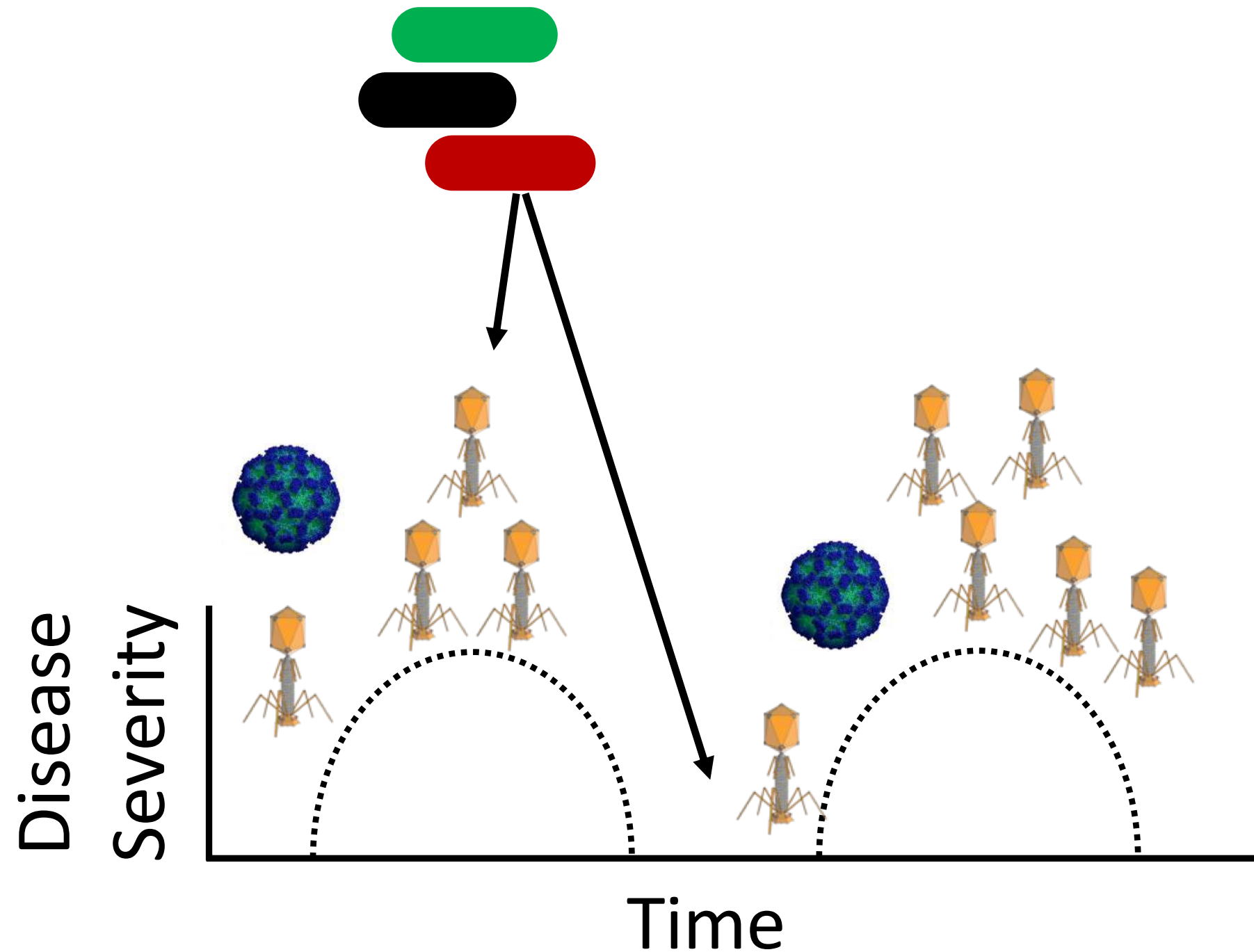


Impact of Enteric Phage Expansion?

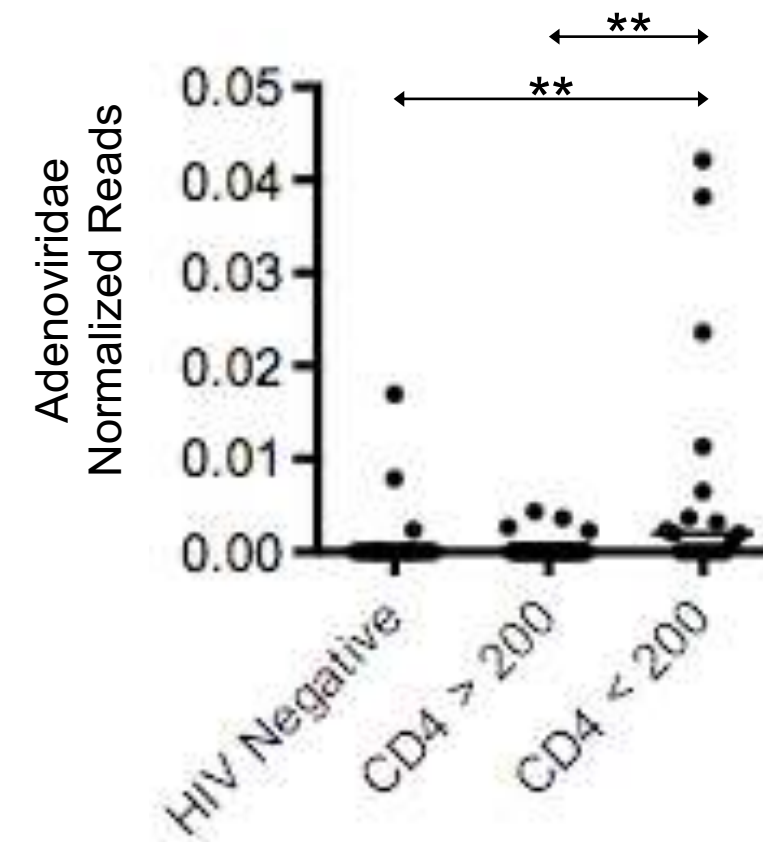
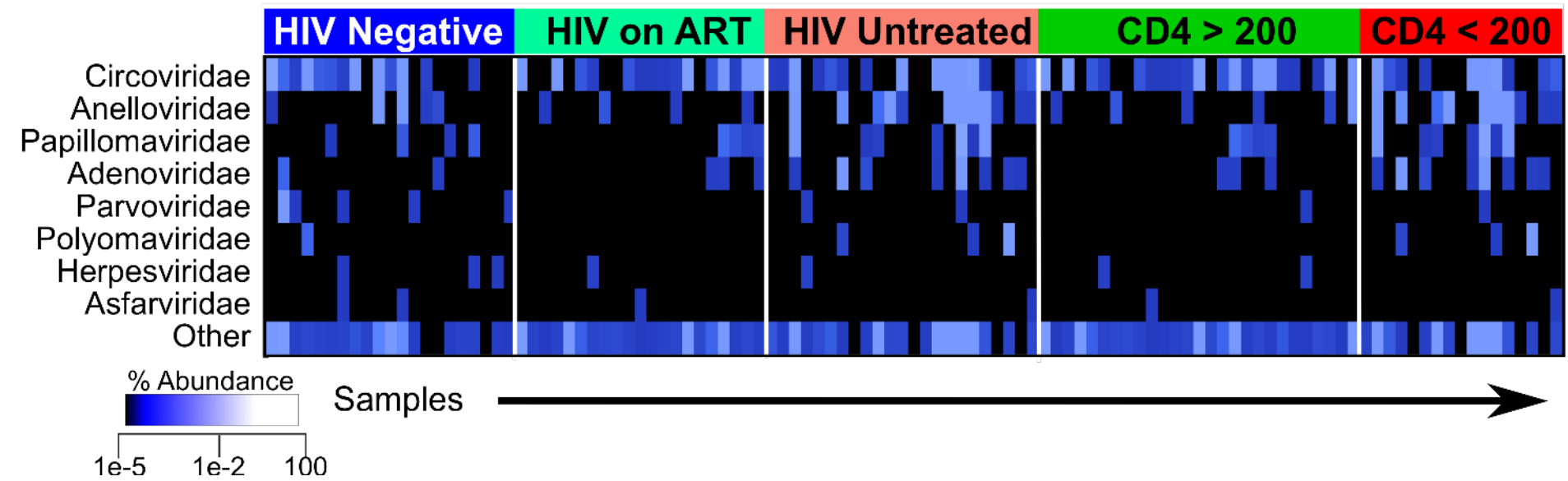
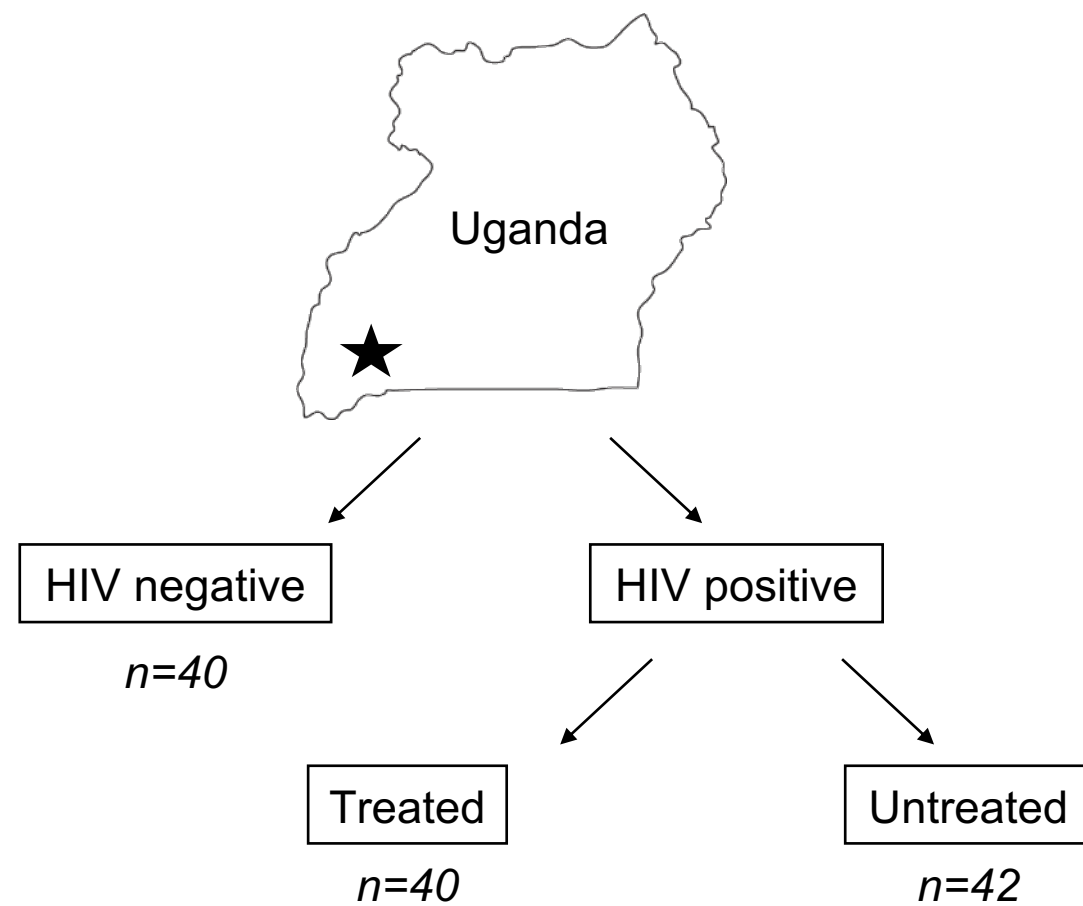


Impact of Enteric Phage Expansion?

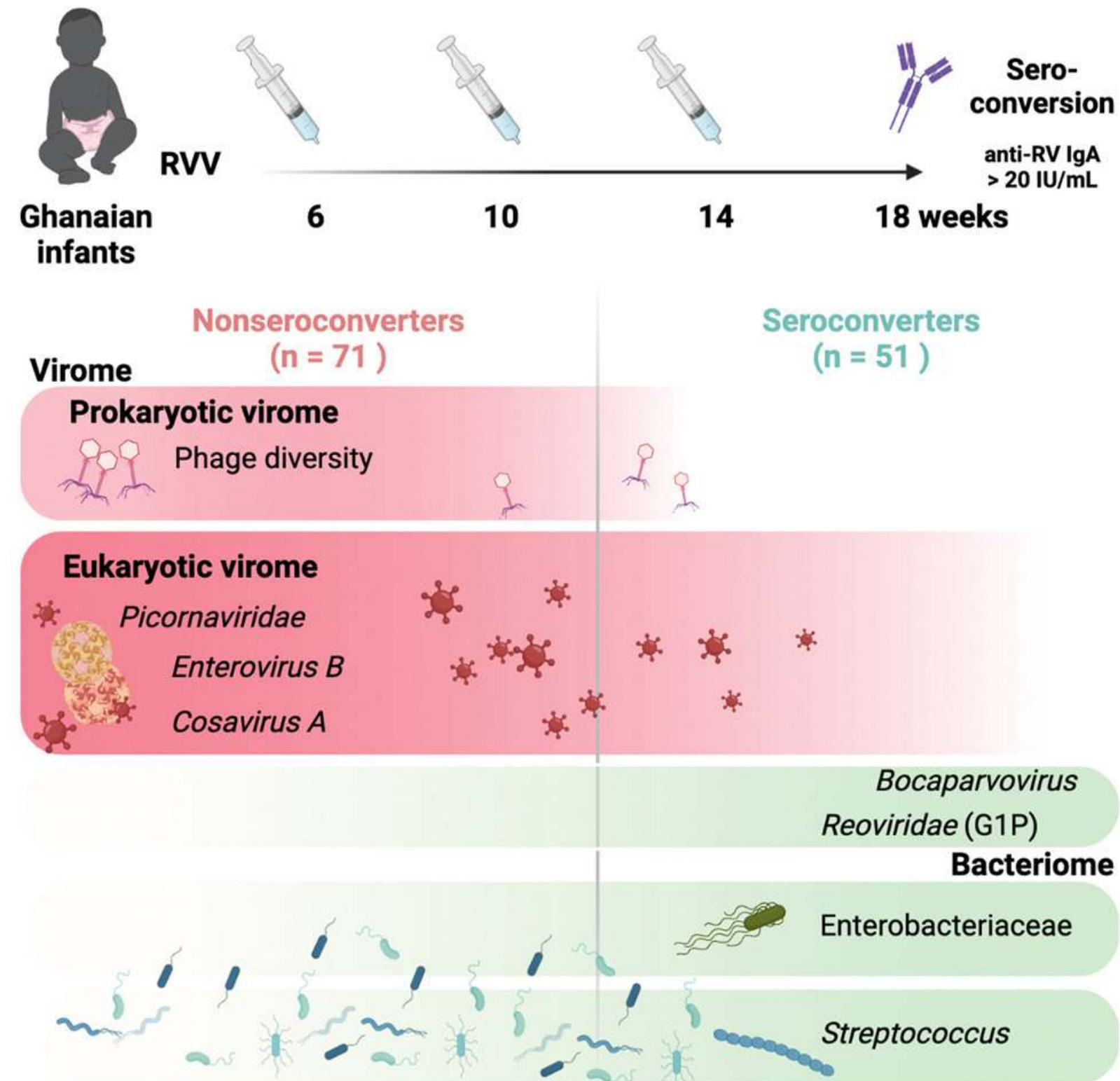
Probiotic or FMT Engraftment



Tracking Viral Infection



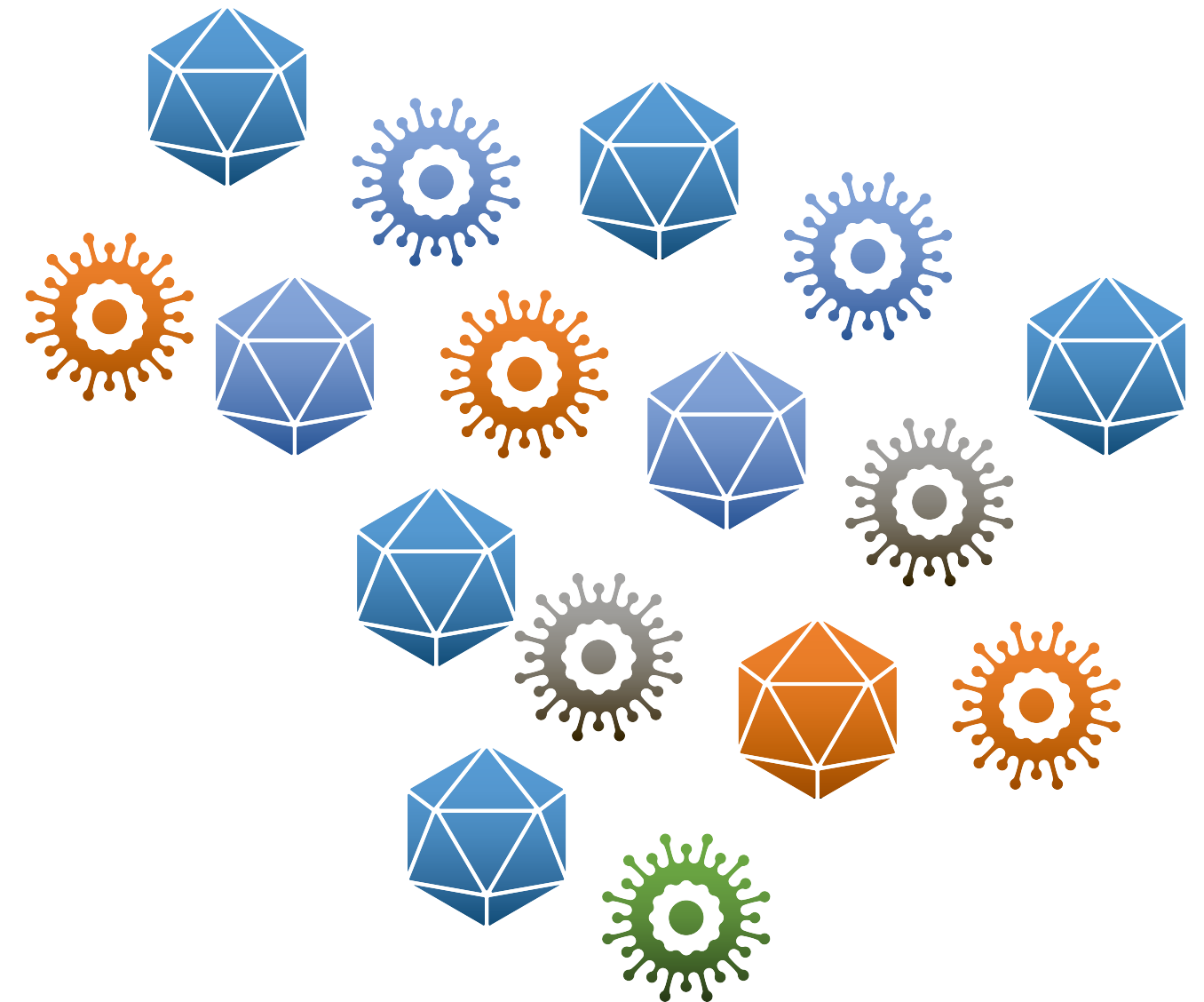
Virome and Infection!



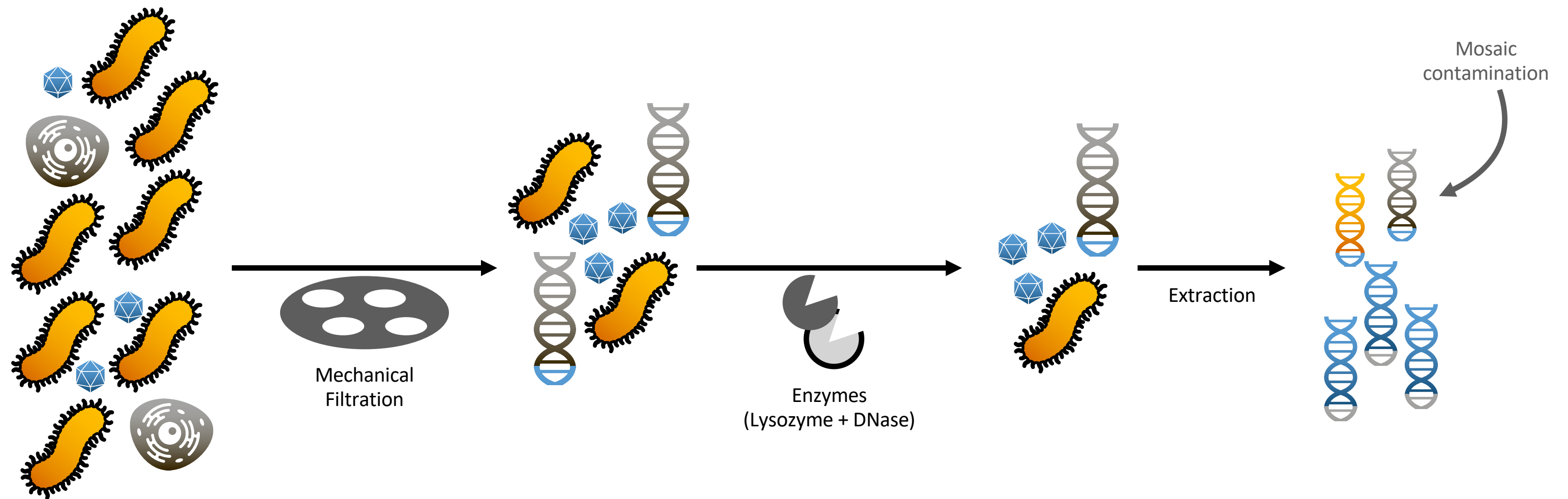
How do we study viromes?

Considerations:

- Viral enrichment (VLP) vs. no enrichment
 - Viral fraction is typically low-biomass, VLP enrichment focuses sequencing effort
- Metagenomic Sequencing
 - Lack of evolutionary conserved sequence like 16S, 18S or ITS
- Analysis
 - Long evolutionary distances due to rapid viral evolution
 - Under-represented databases
 - Standardized tools and databases are improving, but still lacking
 - False-positive classifications



Enrichment $\neq$ Purification



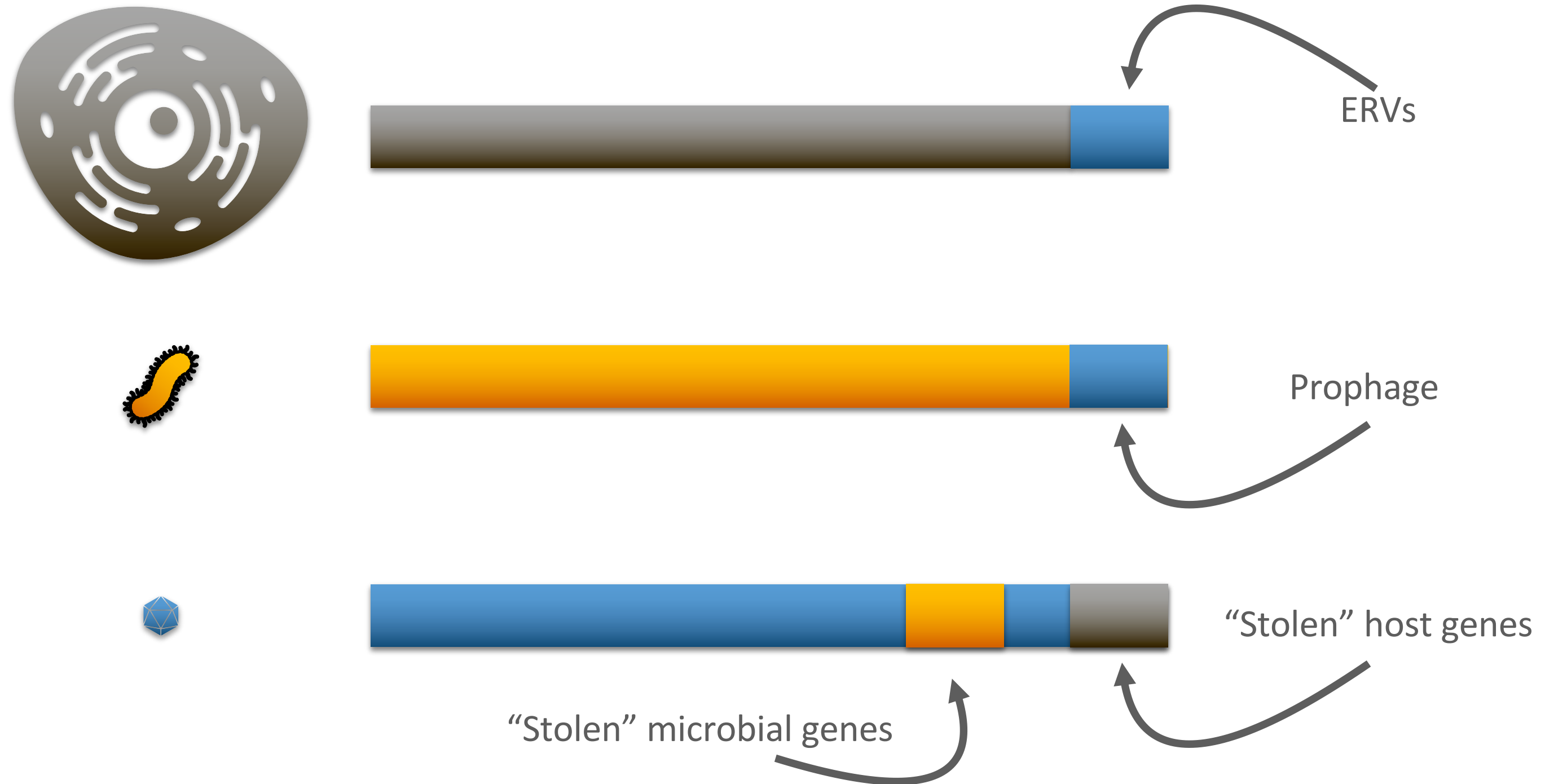
Total Metagenome

Primary VLP Enrichment

Final VLP Enrichment

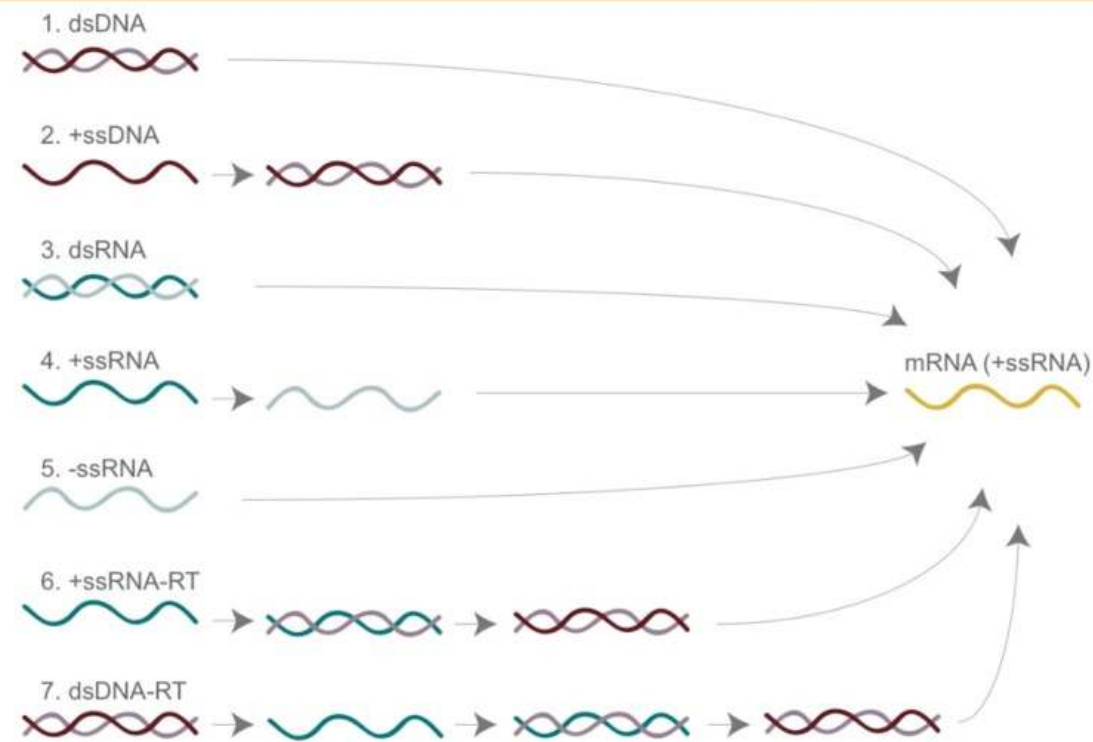
Total Nucleic Acid

Genetic Mosaicism

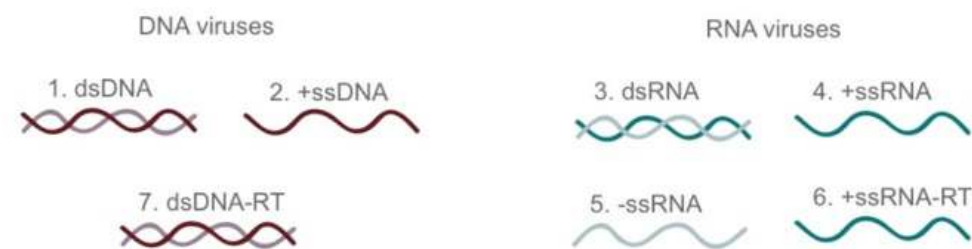


Viral Genome Architectures

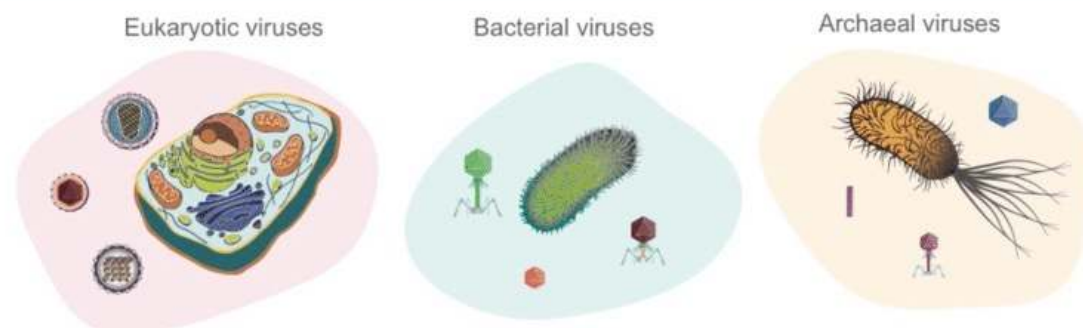
A. Baltimore Classification



B. Nucleotide Type Classification



C. Host-Domain Classification



We don't actually start with this

1st strand synthesis

cDNA synthesis

Illumina sequencing requires dsDNA

We start with this

Hecatomb

<https://github.com/shandley/hecatomb>

Development

Washington University (USA)



Kathie Mihindukulasuriya



Leran Wang



Barry Hykes



Chandni Desai

Flinders University (Australia)



Michael Roach



Rob Edwards



LM [redacted]
To: Handley, Scott

Thursday, Nov 28, 2019, 9:26 AM

Dear Scott,

I am running BlastX but so far it has been an hecatomb of the 300 sequences ran, only 18 were identified as viruses. I'd be lucky if I'll have 30 viruses out of 700 initially identified.
Just to be sure, the pipeline I have done was:

1. blastn of the reads and discard those that had lower e-value for the human genome
2. get all the reads for each patient/tissue that mapped on a specific virus then generate a cluster as
 - a. if the reads were overlapping, merge them into a contig using a consensus generated with EMBOSS cons from a clustalX alignment
 - b. reads that did not overlap were given as a separate contig
 - c. the contigs mapping on the same virus were concatenated with an NNNNN string in between
3. run blastX and retrieve the top 10 hits
4. those that have all hits as bacteria are discarded (which are alarmingly about 97% of the hits!)
5. manually check all the others (since they are few, I can do that)

Is this pipeline acceptable? is this failure rate normal or is there something weird in the data?
Thank you

[redacted]

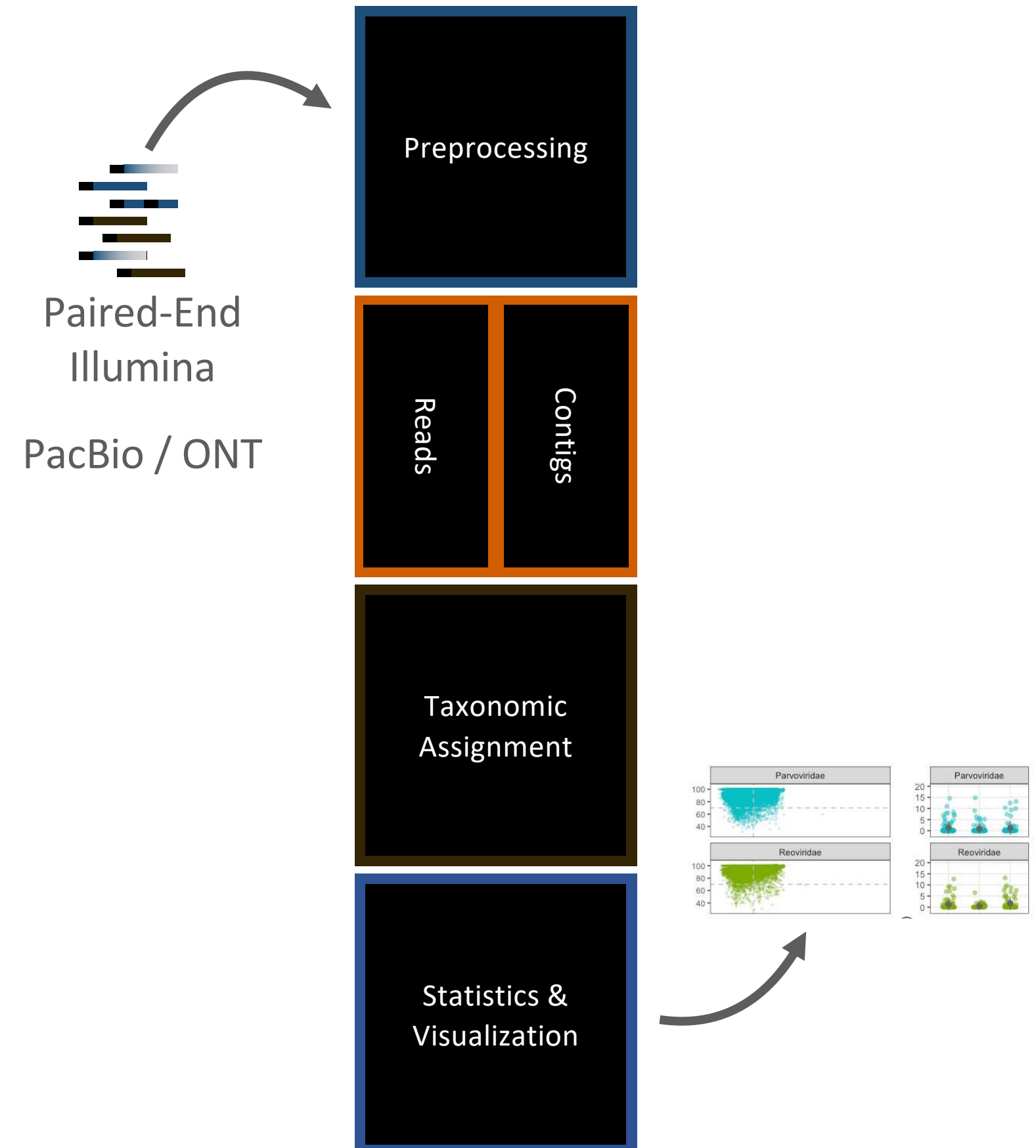
Other Software

- **VirusSeeker** - <https://github.com/guoyanzhao/VirusSeeker-Virome>
 - Gold standard for removal of false-positives
 - Challenging to run under different compute architectures
 - Results are difficult to integrate with other data types
- **IdSeq** - <https://idseq.net>
 - Cloud-based
 - “All” microorganisms, just not virus
 - No phage analysis
 - ‘Complicated’ terms-of-service
- **VirScan / VirFinder / DeepVirFinder / cenotetaker2**
 - Viral contig annotation only



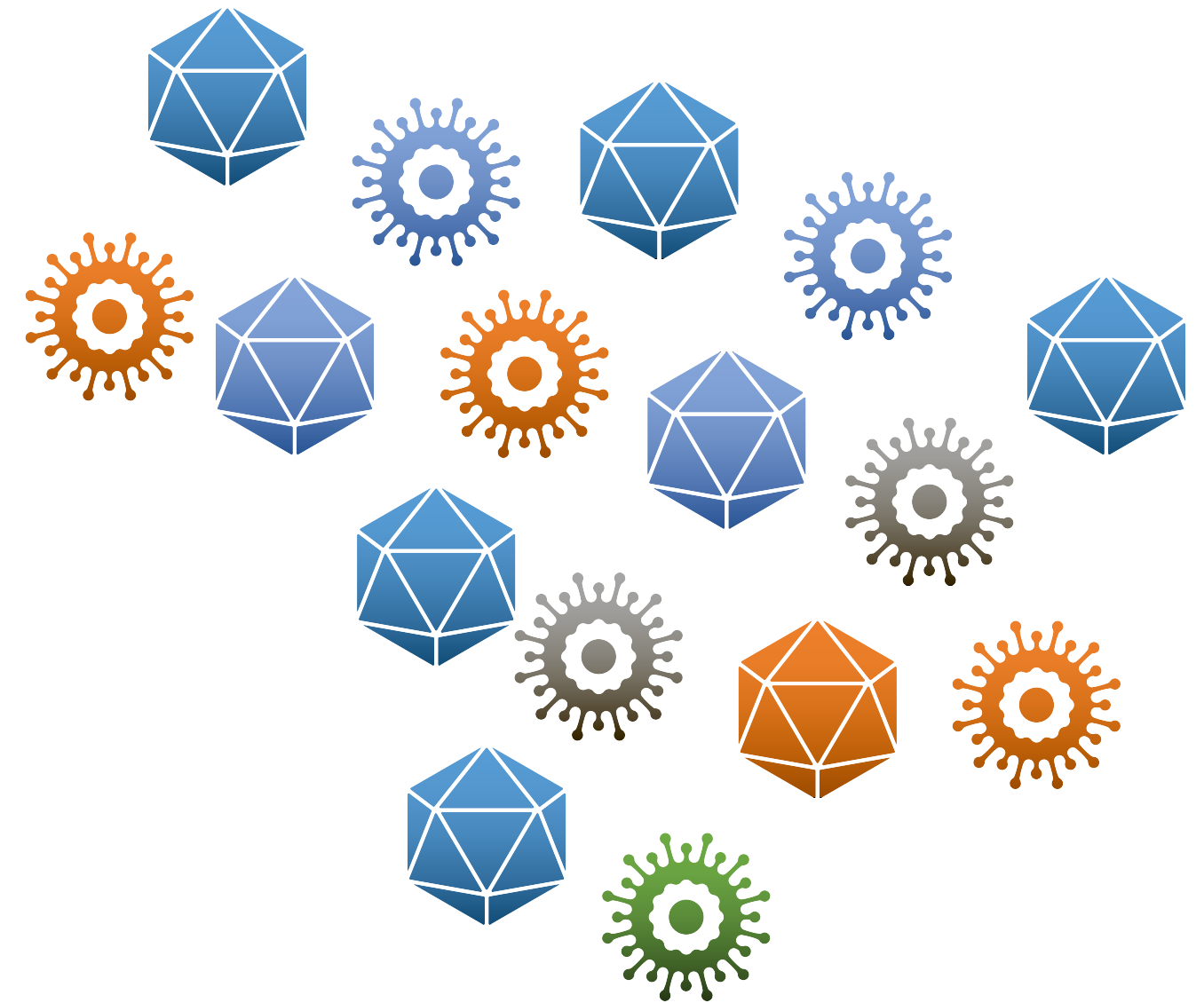
What hecatomb *is*?

- Broadly:
 - Virome analysis software
- Specifically:
 - Computational workflow to detect and annotate viral sequences from metagenomic sequences
 - Can detect and analyze both phage and eukaryotic viral sequences
 - Works on individual reads and contigs
 - Integrates taxonomy, counts, sample data and external data sources into a single R object
 - Workflow management with [Snakemake](#)
 - Dependency management with [Conda](#)
 - Recognizes **resource imperfection** and balances it with **data integration** and **investigator tools**



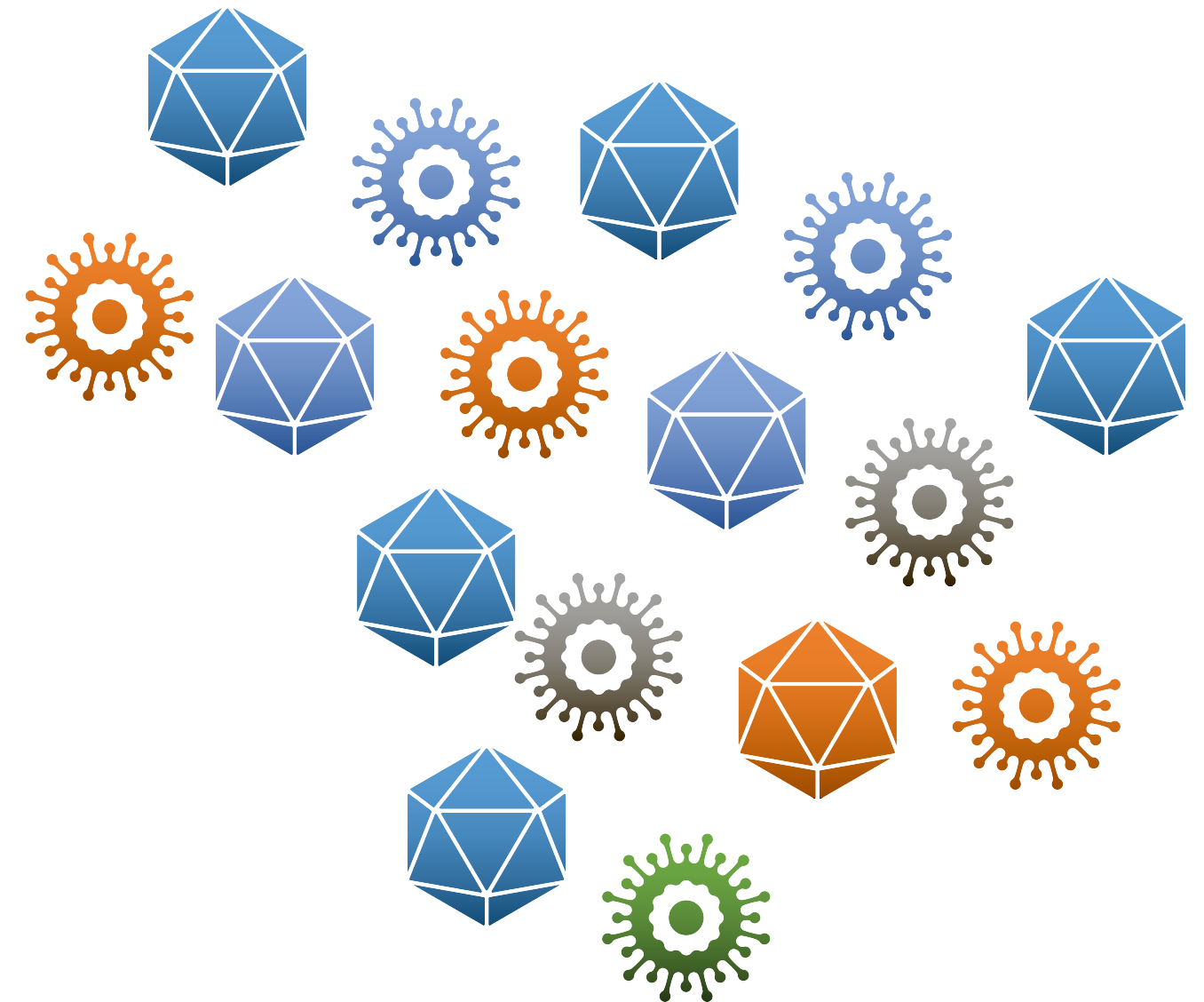
What hecatomb is not?

- A bacterial, fungal or other organism analysis tool
- Overly-opinionated
 - Settings are typically set to *annotate* instead of *remove/filter* data
- A ‘push-button’ tool
 - Data production (e.g. quality-control, taxonomic assignment) is relatively well-automated
 - Data analysis is meant to be interactive and managed by an invested researcher



Hecatomb Philosophy

- There is no ***perfect*** search strategy or ***perfect*** reference database
 - There will always be true/false positives/negatives
- Hecatomb's approach is to do a ***good enough*** (quite good actually) job at data cleaning, assembly, taxonomic assignment
 - Does not require:
 - monolithic databases (nr/nt)
 - Super computing
- Instead, hecatomb is designed to provide maximum information for investigator decision making, statistics and visualization
 - This off-sets the need for perfect/exhaustive resources



How Hecatomb Works

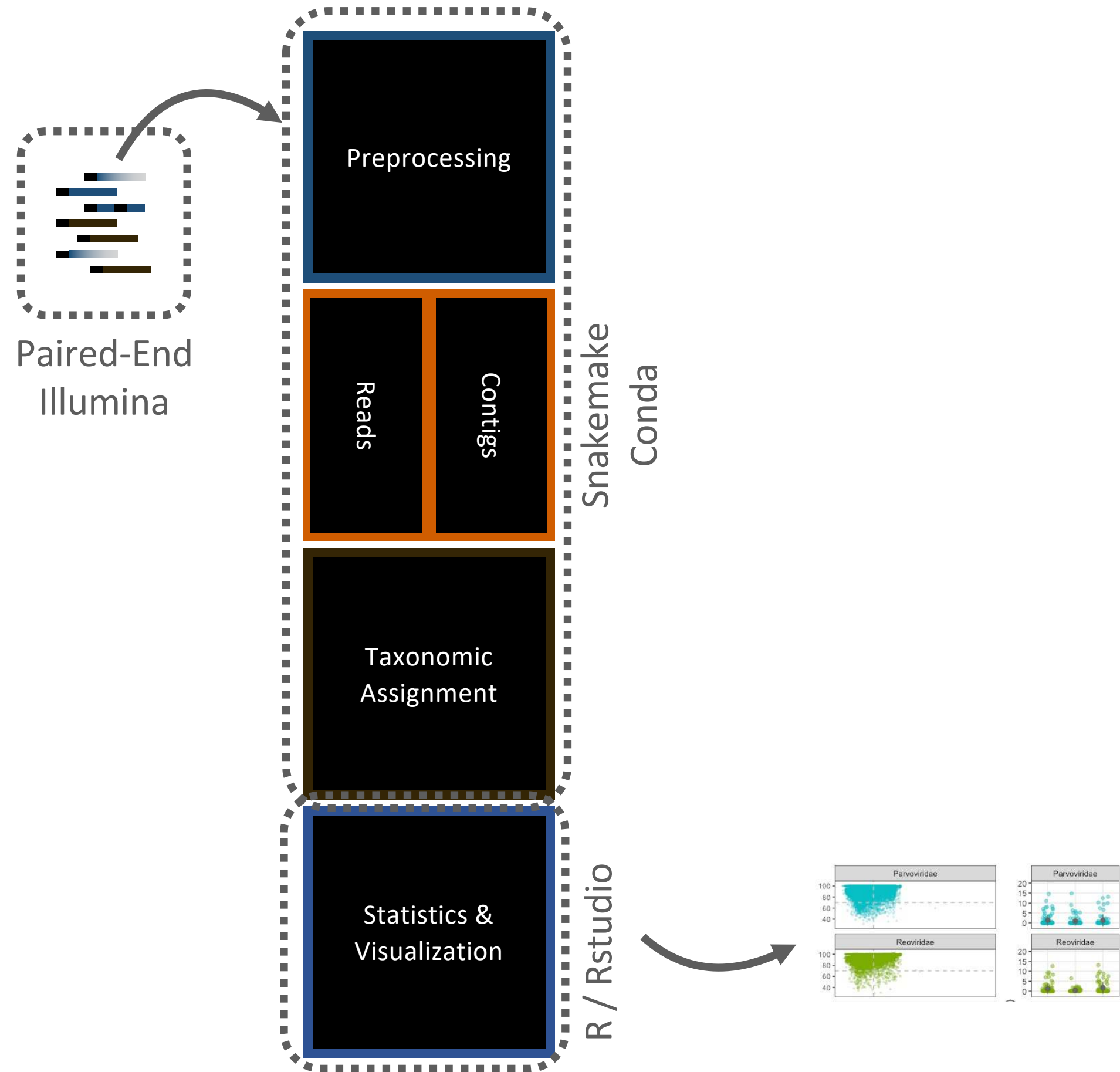
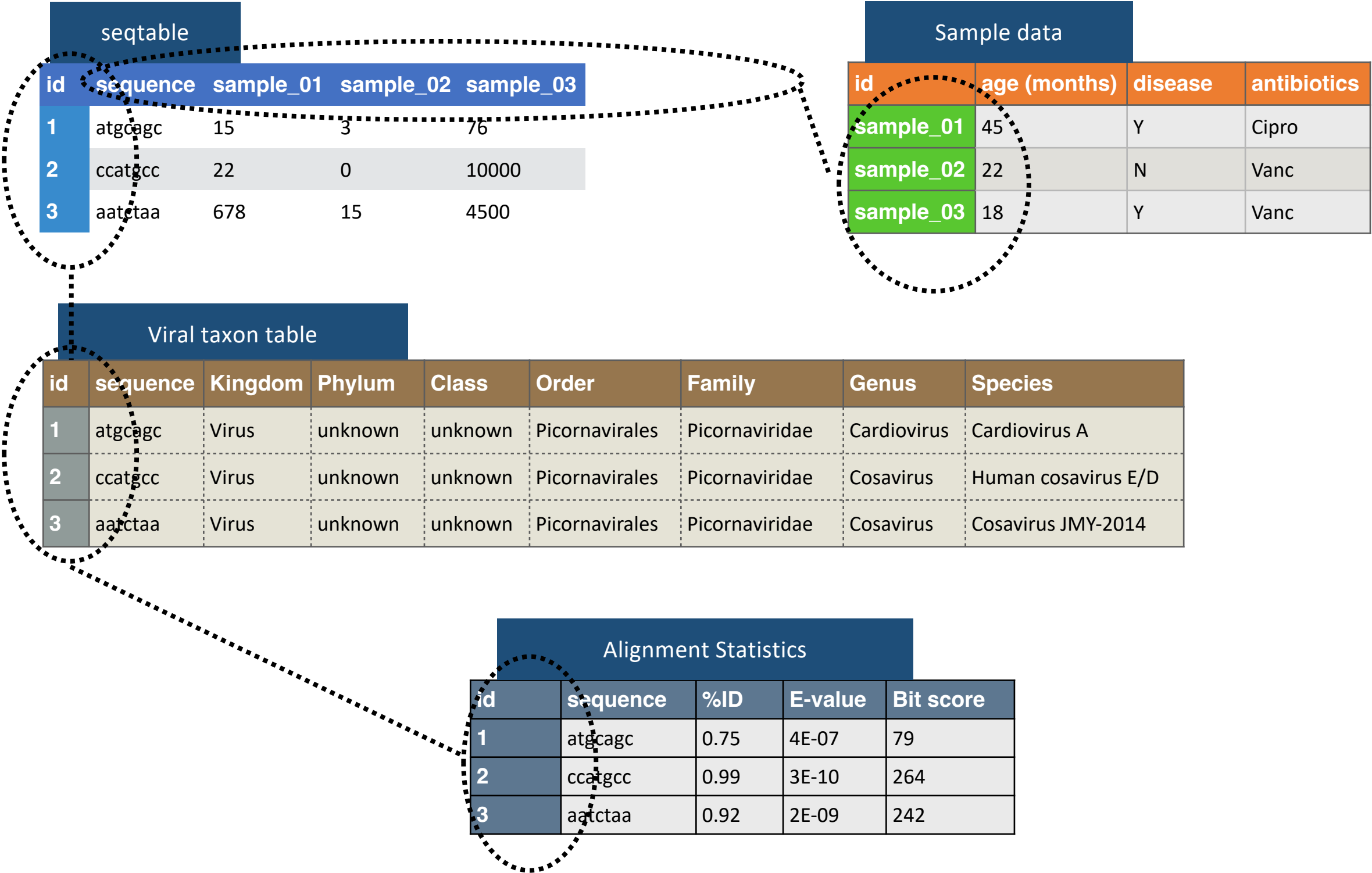
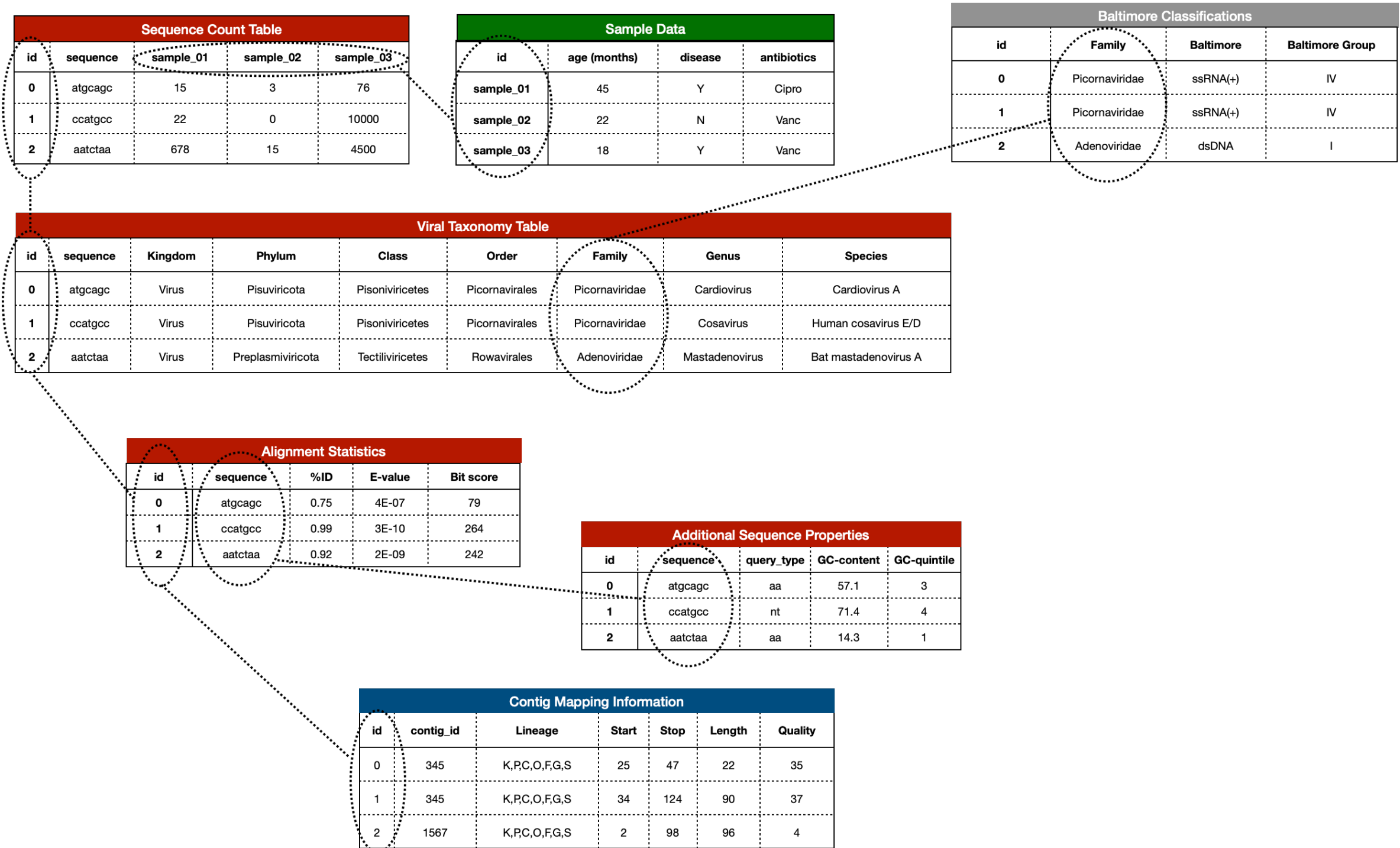


Table Building

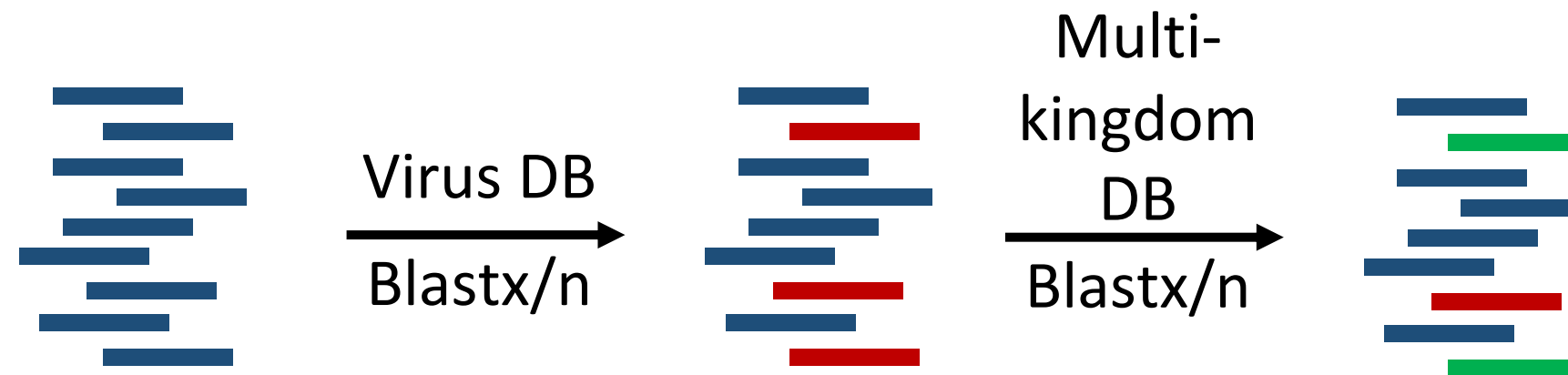


More tables!

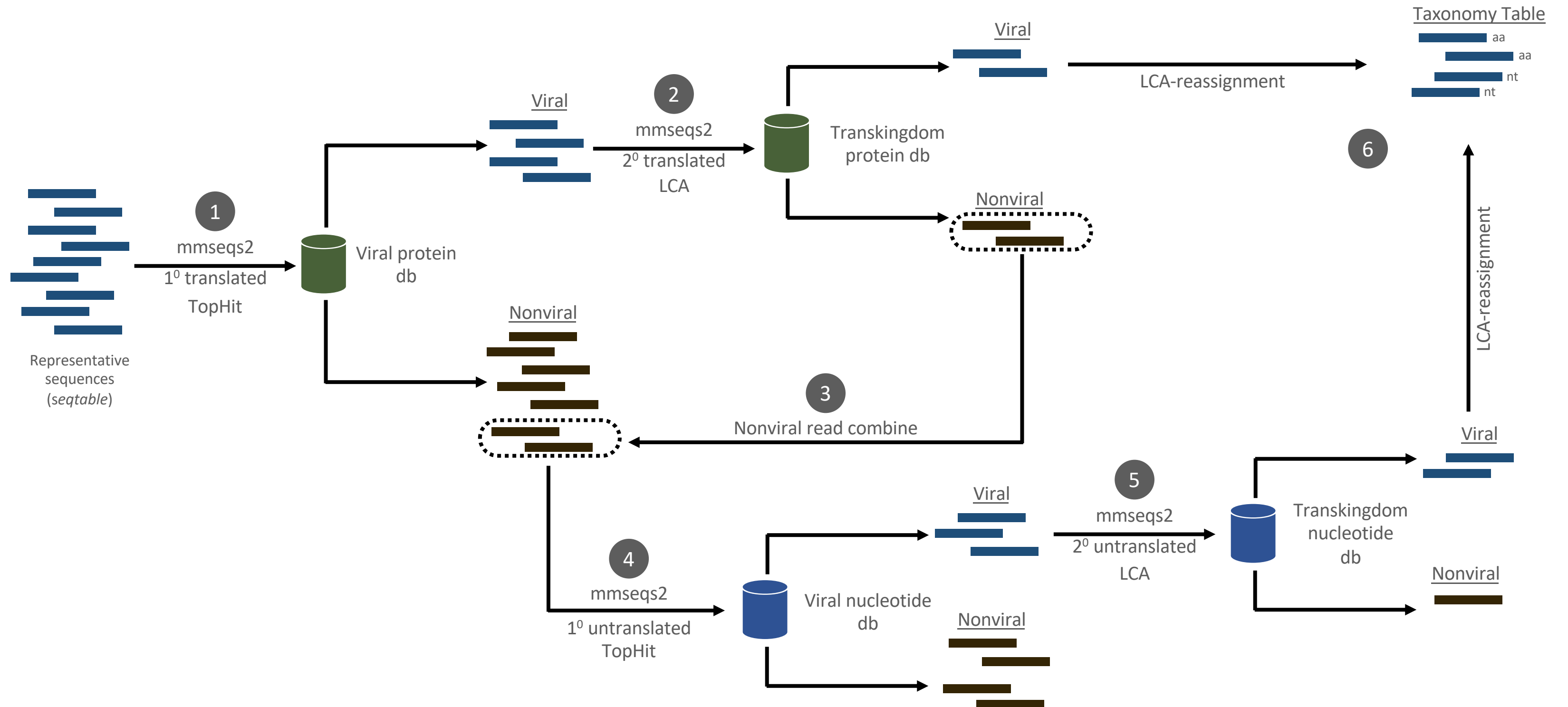


Basic False Positive Challenge

Balancing computation time with specificity

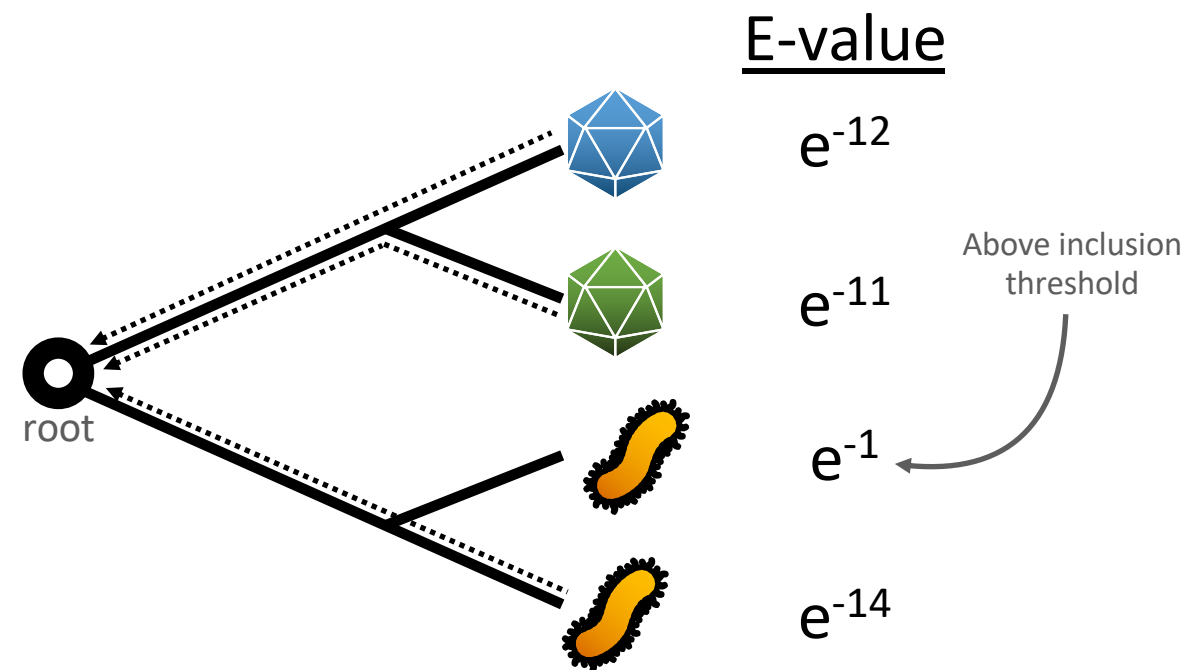


Taxonomic Assignment



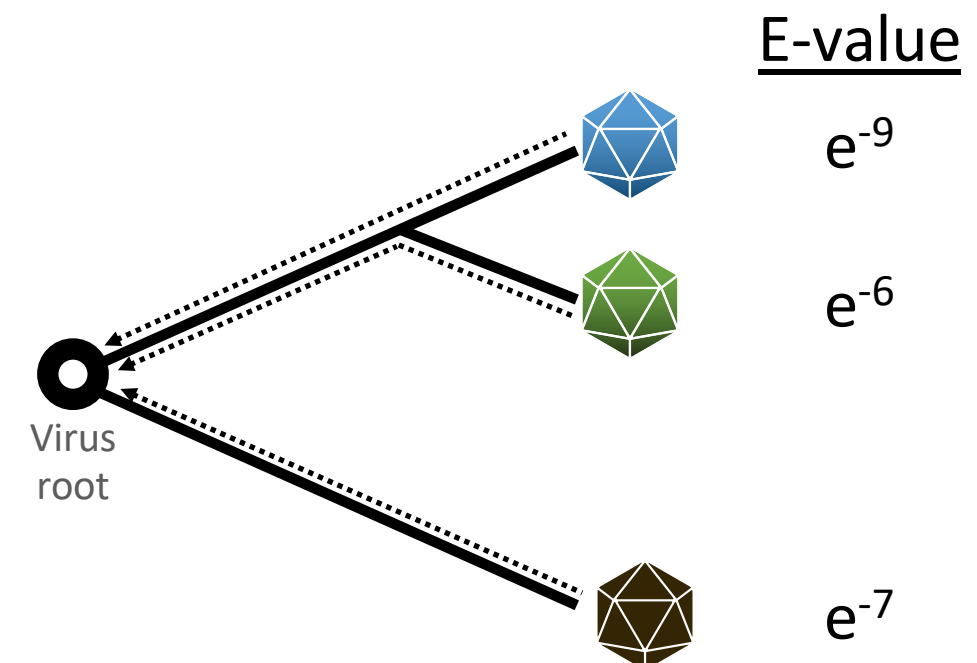
Lowest-Common Ancestor Challenges

Transkingdom Rooting



- Transkingdom LCA will assign '*root*' taxonomy
- Requires tree and some inclusion threshold

Virus Rooting



- Transviral LCA will assign '*virus root*' taxonomy
- Requires tree and some inclusion threshold

Running Hecatomb

Preprocessing

Reads

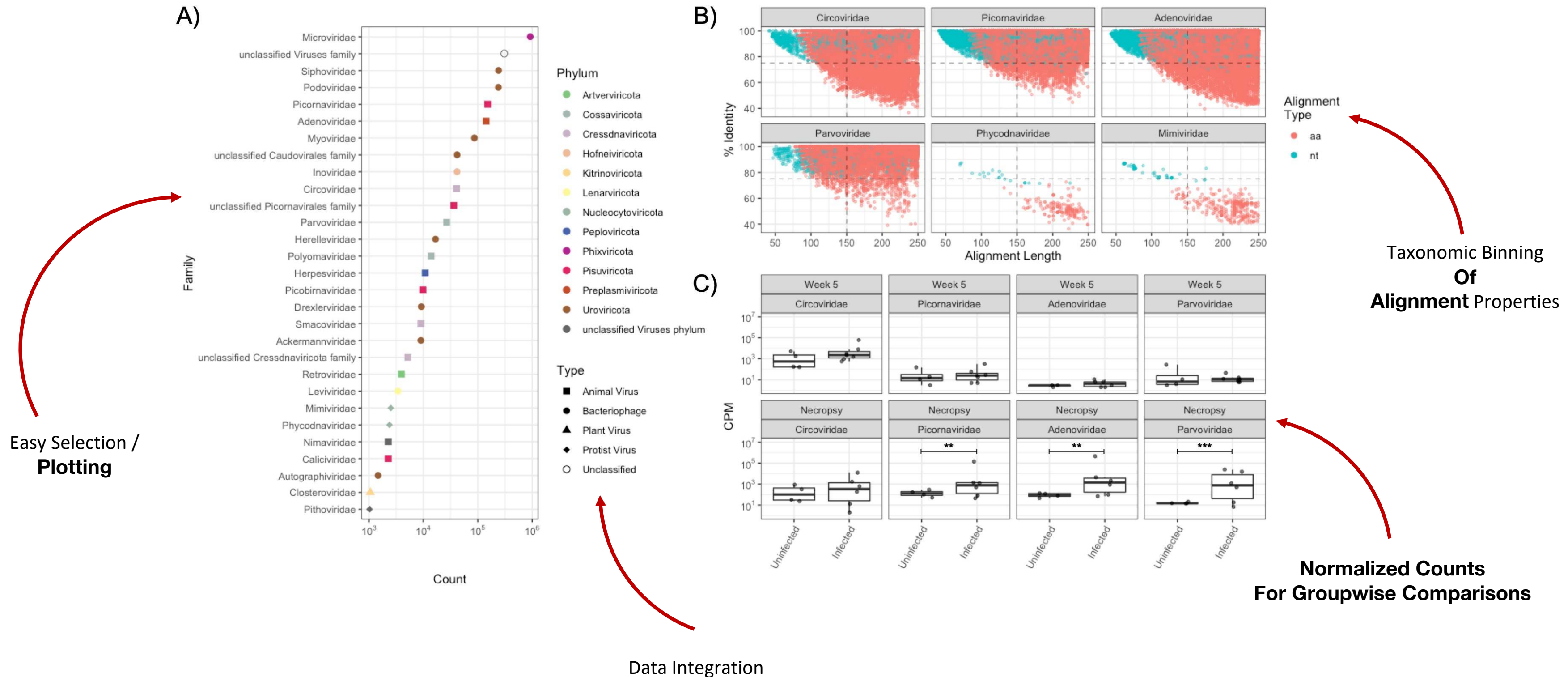
Contigs

Taxonomic
Assignment

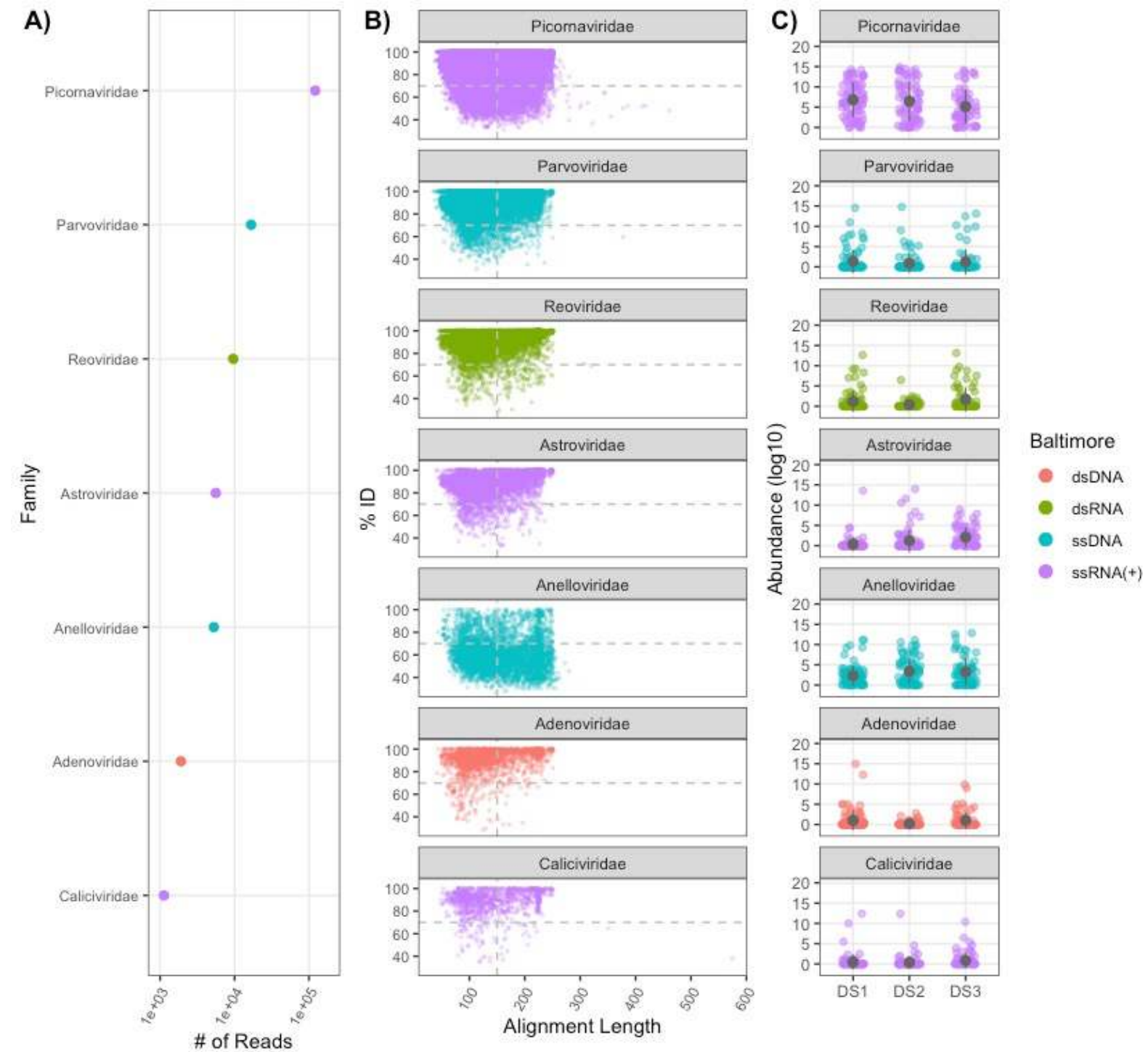
Statistics &
Visualization

- <https://github.com/shandley/hecatomb>
 - More information on the wiki: <https://github.com/shandley/hecatomb/wiki>
- Dependencies
 - Snakemake
 - Conda
 - R
 - RStudio (technically not necessary, but very helpful)
- Run
 - `snakemake --snakefile ./Snakefile --configfile ../config/my_config.yaml --resources mem_mb=100000 --cores 64 --use-conda --conda-frontend mamba`

Example Analysis



Example Analysis



Acknowledgements

Handley Lab

- Leran Wang
- Kathie Mihindukulasuriya
- Ana Jung
- Lindsay Droit
- Leah Heath
- Luis Alberto Chicas Cardenas
- Andrew Kim
- Yuhao Li

Washington University

- Dave Wang
- Mike Diamond
- Megan Baldrige
- Jeff Gordon
- Skip Virgin (Vir Biotech)

Flinders University

- Rob Edwards
- Mike Roach
- Elizabeth Dinsdale

San Diego State University

- Anca Segall

Universidad de los Andes

- Alejandro Reyes

IBD

- Miles Parkes
(Addenbrook Hospital,
Cambridge)
- Kate Jeffrey (Moderna)

HIV

- Doug Kwon (Harvard)
- Dan Barouch (Harvard)

Rotavirus

- Vanessa Harris
(Amsterdam Institute
for Global Health)