Mathematical and Statistical Data Analysis Challenges from the Microbiome

Susan Holmes
http://www-stat.stanford.edu/~susan/

Bio-X and Statistics, Stanford University

Toulouse, 2014 (CIMI) and NIH-R01GM086884

What's happening in Biological Data Analysis?

What's happening in Biological Data Analysis?

1985 1998 2012

What's happening in Biological Data Analysis?

1985 1998 2012

What's happening in Biological Data Analysis?

1985 1998 2012

Challenges for those of us working from the ground up

- Heterogeneity.
- Heteroscedasticity.
- Structured high-dimensionality.
- Graph or Tree integration.
- Validations, all tests significant.
- Reproducibility.

Part I

Heterogeneity

'Homogeneous data are all alike;

all heterogeneous data are heterogeneous

in their own way.'

Heterogeneity of Data

- Status : response/ explanatory.
- Hidden (latent)/measured.
- Types :
  - Continuous
  - Binary, categorical
  - Graphs/ Trees
  - Images
  - Maps/ Spatial Information
  - Rankings
- Amounts of dependency: independent/time series/spatial.
- Different technologies used (454, phyloseq, Illumina, MassSpec, RNA-seq).

Microbiome and other useful `ome'words

Joshua Lederberg:`the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease'

Microbiome Complete collection of genes contained in the genomes of microbes living in a given environment.

Numbers Humans shelter 100 trillion microbes (10^{14}), (we are made of 10 $\times 10^{12}$ cells)
It is estimated that there are more than 1000 `species' of bacteria living in the human gut.

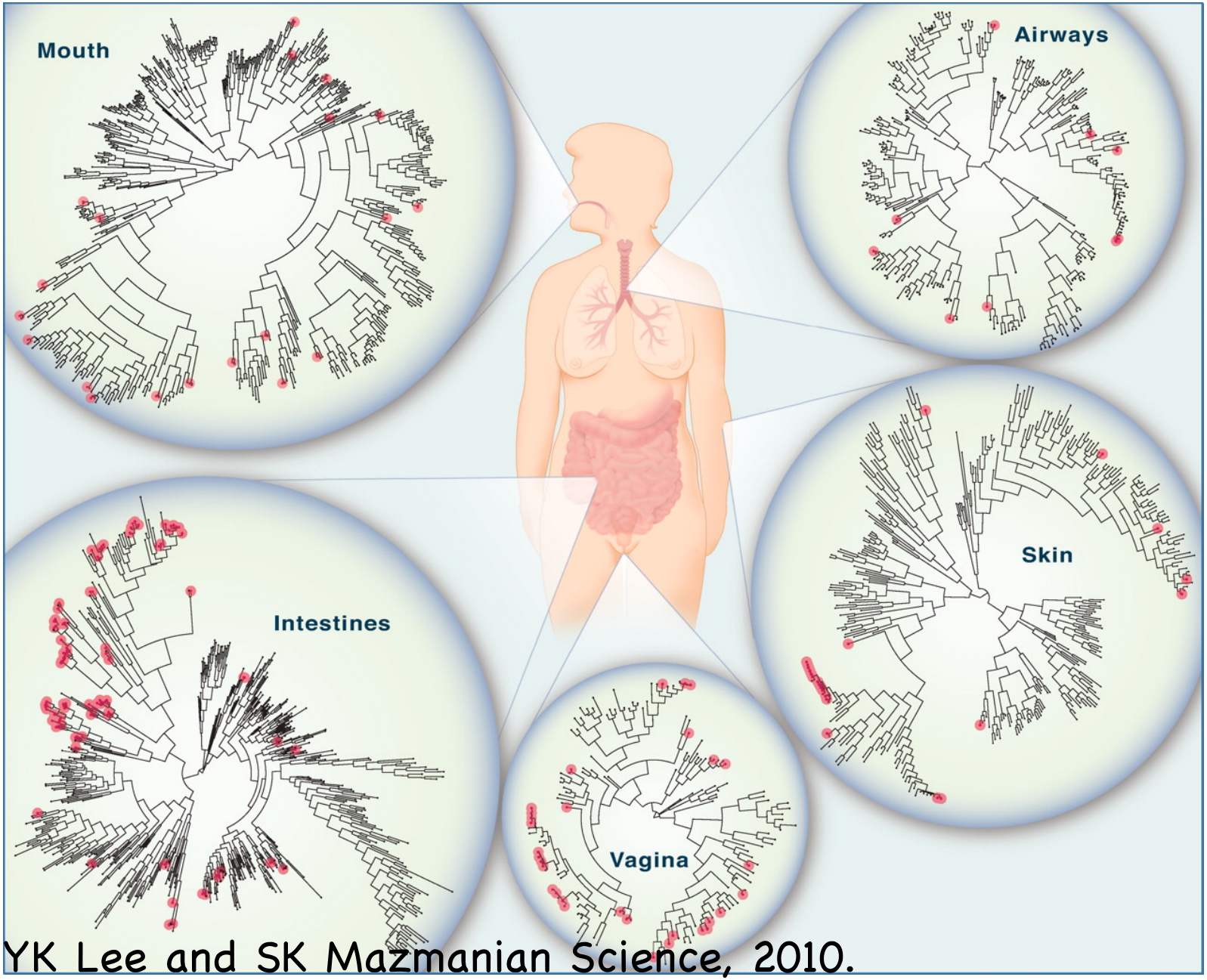
Metagenome Composition of all genes present in an environment (soil, gut, seawater), regardless of species.

Transciptome These are the mRNA transcripts in the cell, it reflects the genes that are being actively expressed at any given time.

Bacteria etc... and Us

The human microbiome or human microbiota is the assemblage of microorganisms that reside on the surface and in deep layers of skin, in the saliva and oral mucosa, in the conjunctiva, and in the gastrointestinal tracts.

- They include bacteria, fungi, and archaea.
- Some of these organisms perform tasks that are useful for the human host. (live in symbiosis)
- Majority have no known beneficial or harmful effect.



YK Lee and SK Mazmanian Science, 2010.

Human Microbiome: What are the data?

DNA The Genomic material present (16sRNA-gene especially).

RNA What genes are being turned on (gene expression), transcriptomics.

Mass Spec Specific signatures of chemical compounds present.

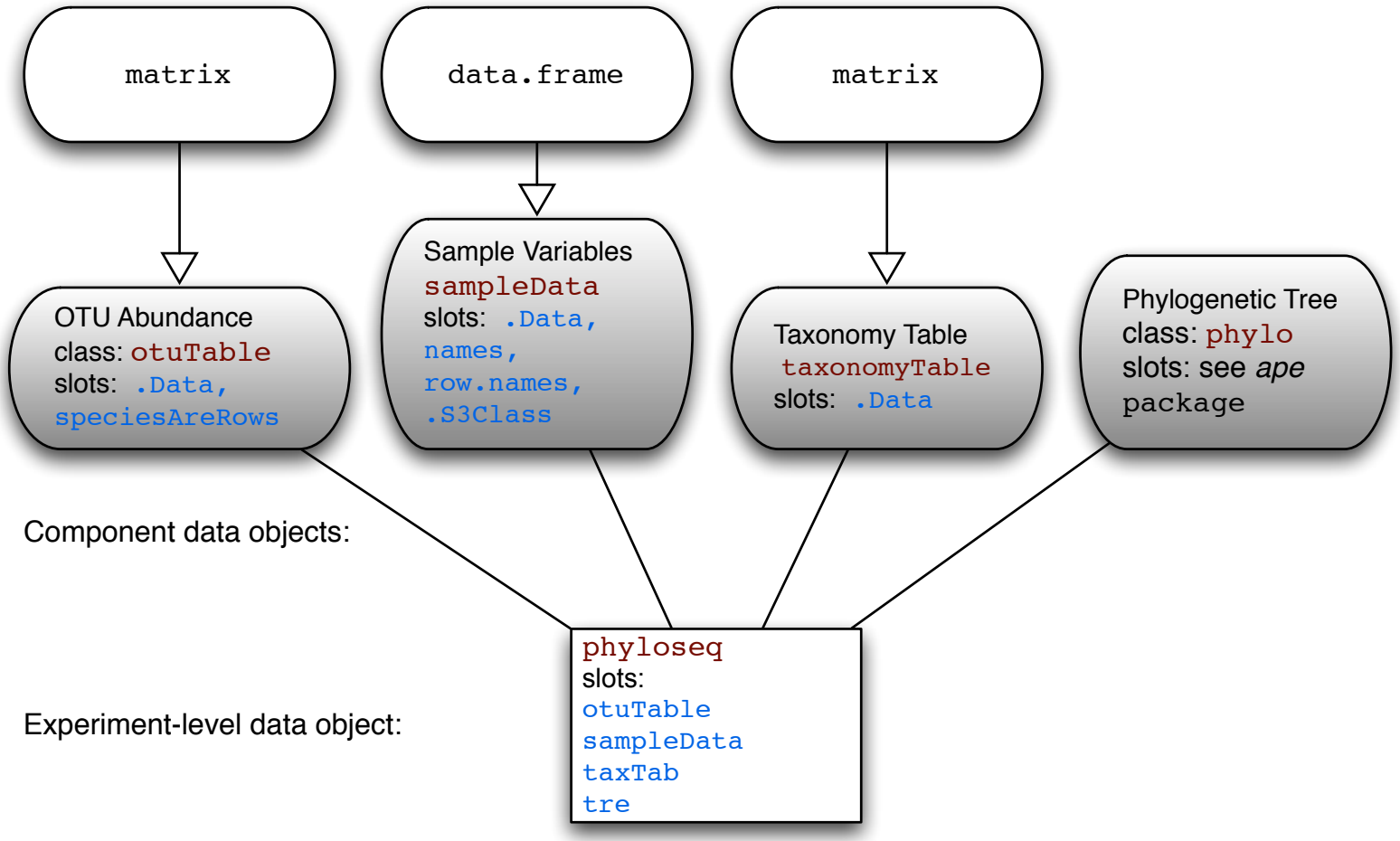
Clinical Multivariate information about patients' clinical status, medication, weight.

Environmental Location, nutrition, time.

Domain Knowledge Metabolic networks, phylogenetic trees, gene ontologies.

Heterogeneous Data Objects

Input and data manipulation with `phyloseq`
(McMurdie and Holmes, 2013, Plos ONE)
As always in R: object oriented data:



Questions from many Disciplines

Biomedical sciences Clinical variables, immune history.

Genetics Host and bacterial genomes and transcriptomes.

Bioinformatics Databases and formats.

Biochemistry Compounds, Metabolic pathways involved.

Ecology Ordination, diversity, environmental influences.

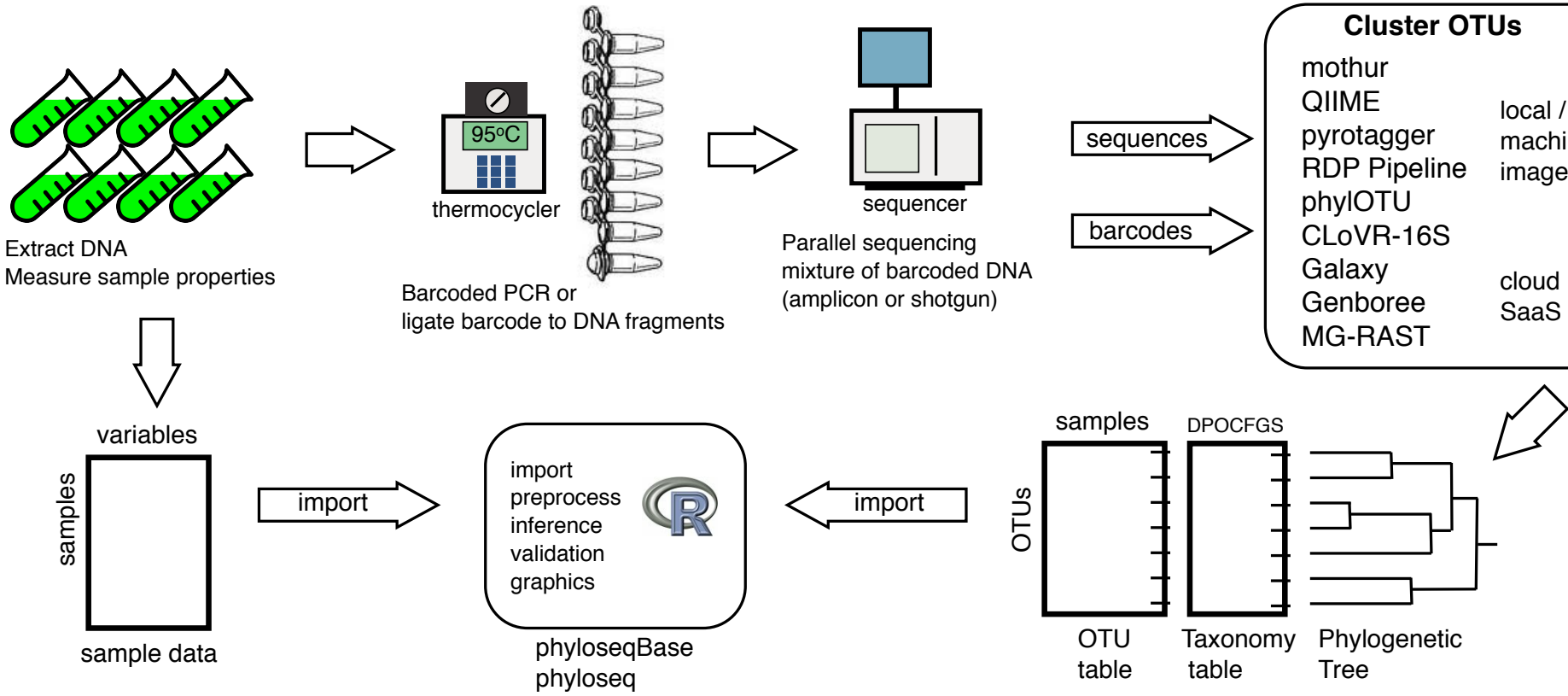
Statistics Decompositions of variability, spatio-temperal analyses, normalizations.

Systematic Biology Phylogenetic Trees, dynamics of evolution.

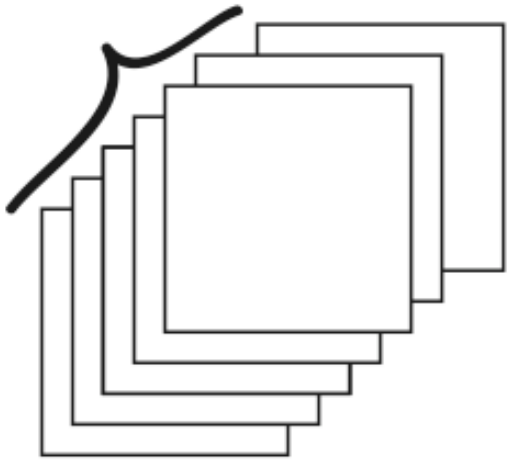
Network science Microbial communities, gene interactions, metabolic pathways,...

Visualization Dimension reduction, rich representations.

Mathematics Metric Geometry, multilinear algebra, probability, topology.


Extract DNA
Measure sample properties
95°C
thermocycler
Barcoded PCR or
ligate barcode to DNA fragments
sequencer
Parallel sequencing
mixture of barcoded DNA
(amplicon or shotgun)
sequences
barcodes
Cluster OTUs
mothur
QIIME
pyrotagger
RDP Pipeline
phylOTU
CLoVR-16S
Galaxy
Genboree
MG-RAST
local /
machi
image
cloud
SaaS
variables
samples
sample data
import
import
preprocess
inference
validation
graphics
phyloseqBase
phyloseq
import
samples
DPOCFGS
OTUs
OTU
table
Taxonomy
table
Phylogenetic
Tree

Useful first order representation: Many Matrices



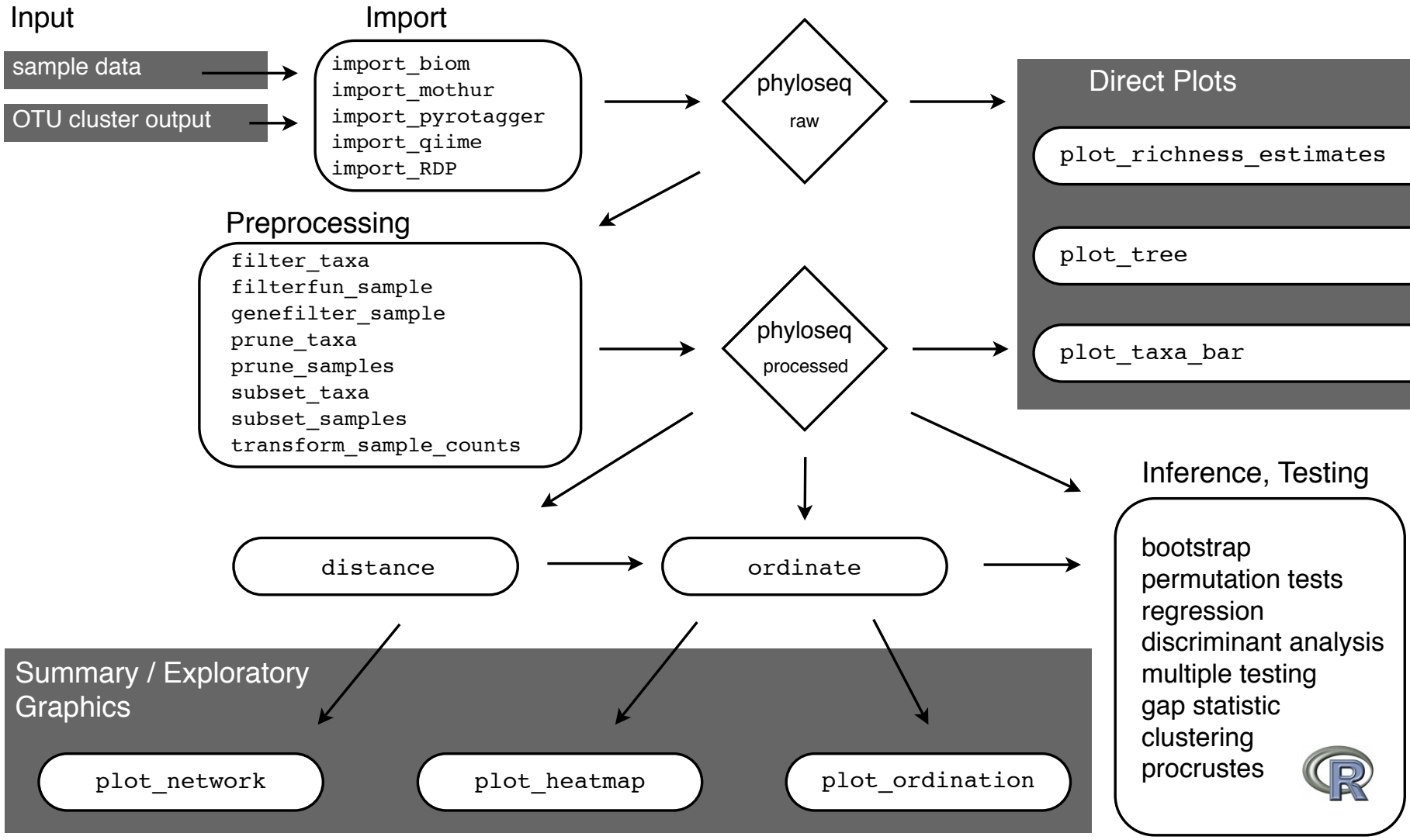
- Time series of abundance matrices.
- Different types of data on same samples (taxa counts, clinical variates, spatial location).
- Networks and trees over time.
- Explanatory (environmental) variables, Response variables.

Holmes (2005), Duality Diagrams.

Getting Started: Data from qiime

Rectangular data + Side Information.

- with a number of taxa or species (often as rows of the table).
- with a certain number of samples/patients recorded as columns of the table.
- Phylogenetic/family relationships between the rows of the table.
- Extra clinical/environmental information about the samples.

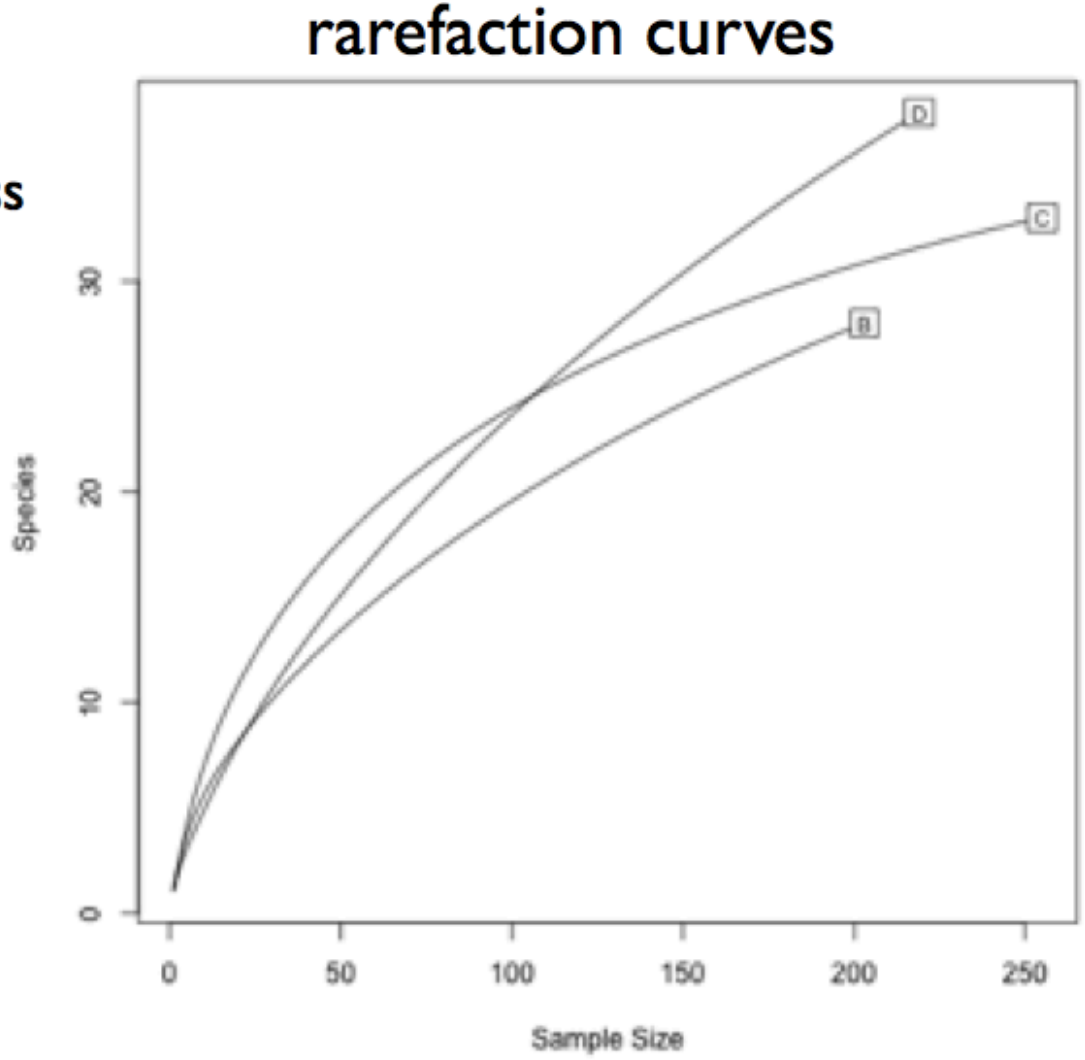

Input
sample data
OTU cluster output
Import
import_biom
import_mothur
import_pyrotagger
import_qiime
import_RDP
phyloseq
raw
Direct Plots
plot_richness_estimates
plot_tree
plot_taxa_bar
Preprocessing
filter_taxa
filterfun_sample
genefilter_sample
prune_taxa
prune_samples
subset_taxa
subset_samples
transform_sample_counts
phyloseq
processed
Inference, Testing
bootstrap
permutation tests
regression
discriminant analysis
multiple testing
gap statistic
clustering
procrustes
distance
ordinate
Summary / Exploratory Graphics
plot_network
plot_heatmap
plot_ordination

Part II

Heteroscedasticity: Mixtures and how to Normalize them

How to deal with different numbers of reads?

- Sanders 1968
- non-parametric richness
- estimate coverage
- Normalize?



Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*

Current Method: Rarefying

Ad hoc library size normalization by random subsampling without replacement.

1. Select a minimum library size, $N_{L,min}$. This has also been called the rarefaction level though we will not use the term here.
2. Discard libraries (microbiome samples) that have fewer reads than $N_{L,min}$.
3. Subsample the remaining libraries without replacement such that they all have size $N_{L,min}$.

Often $N_{L,min}$ is chosen to be equal to the size of the smallest library that is not considered defective, and the process of identifying defective samples comes with a risk of subjectivity and bias. In many cases researchers have also failed to repeat the random subsampling step (3) or record the pseudorandom number generation seed/process --- both of which are essential for reproducibility.

Aim of the studies: Differential Abundance

Like differentially expressed genes, a species/OTU is considered differentially abundant if its mean proportion is significantly different between two or more sample classes in the experimental design.

Rarefaction is Inadmissible

Unfortunately, rarefying biological count data is unjustified despite its current ubiquity in microbiome analyses.
The following is a minimal example to explain why rarefying is statistically inadmissible, especially with regards to variance stabilization.
Suppose we want to compare two different samples, called A and B, comprised of 100 and 1000 reads, respectively. In these hypothetical communities only two types of microbes have been observed, OTU1 and OTU2

According to Table 1, Left.

Table : **A minimal example of the effect of rarefying on power.**

Original Abundance		
	A	B
OTU1	62	500
OTU2	38	500
Total	100	1000

Rarefied Abundance		
	A	B
OTU1	62	50
OTU2	38	50
	100	100

Standard Tests for Difference			
P-value	χ^2	Prop	Fisher
Original	0.0290	0.0290	0.0272
Rarefied	0.1171	0.1171	0.1169

Hypothetical abundance data in its original (Top-Left) and rarefied (Top-Right) form, with corresponding formal test results for differentiation (Bottom).

Formally comparing the two proportions according to a standard test is done either using a χ^2 test (equivalent to a two sample proportion test here) or a Fisher exact test. By rarefying (Table 1, top-right) so that both samples have the same number of counts, we are no longer able to differentiate between them.

This loss of power is completely attributable to reducing the size of B by a factor of 10, which also increases the confidence intervals corresponding to each proportion such that they are no longer distinguishable from those in A, even though they are distinguishable in the original data.

The variance of the proportion's estimate $\hat{p}$ is multiplied by 10 when the total count is divided by 10.

Equalization of variances

In this binomial example the variance of the proportion estimate is $\mathrm{Var}(\frac{X}{n}) = \frac{pq}{n} = \frac{q}{n}\mathrm{E}(\frac{X}{n})$, a function of the mean.
This is a common occurrence and one that is traditionally dealt with in statistics by applying variance-stabilizing transformations.
However, in order to find the right transformation, we need a good model for the error.

Two parameterizations of the negative binomial

In classical probability, the negative binomial is often introduced as the distribution of the number of successes in a sequence of Bernoulli trials with probability of success p before the number r failures occur. Thus with the two parameters r and p, the probability distribution for the negative binomial is given as

$$X \sim NB(r; p)$$

$$\begin{aligned} P(X = k) &= \binom{k + r - 1}{k} (1 - p)^r p^k \\ &= \frac{\Gamma(k + r)}{k!\Gamma(r)} (1 - p)^r p^k \end{aligned}$$

The mean of the distribution is $m = \frac{pr}{1-r}$ and the variance $Var(X) = \frac{pr}{(1-p)^2}$.

Sometimes the distribution is given a different parameterization which we use here. This takes as the two parameters: the mean m and $r = \frac{1-p}{p}m$, then the probability mass distribution is rewritten:

$$X \sim NB(m; r)$$

$$\begin{aligned} P(X = k) &= \binom{k + r - 1}{k} \left(\frac{r}{r+m}\right)^r \left(\frac{m}{r+m}\right)^k \\ &= \frac{\Gamma(k+r)}{k!\Gamma(r)} \left(\frac{r}{r+m}\right)^r \left(\frac{m}{r+m}\right)^k \end{aligned}$$

The variance is $Var(X) = \frac{m(m+r)}{r} = m + \frac{m^2}{r}$, we will also use $\phi = \frac{1}{r}$ and call this the overdispersion parameter, giving $Var(X) = m + \phi m^2$. When $\phi = 0$ the distribution of X will be Poisson(m). This is the (mean=m,overdispersion=ϕ) parametrization we will use from now on.

Negative Binomial as a hierarchical mixture for read counts

To address this, we take the means of the Poisson variables to be random variables themselves having a Gamma distribution with (hyper)parameters shape r and scale $p/(1-p)$. We first generate a random mean, λ, for the Poisson from the Gamma, and then a random variable, k, from the Poisson(λ). The marginal distribution is:

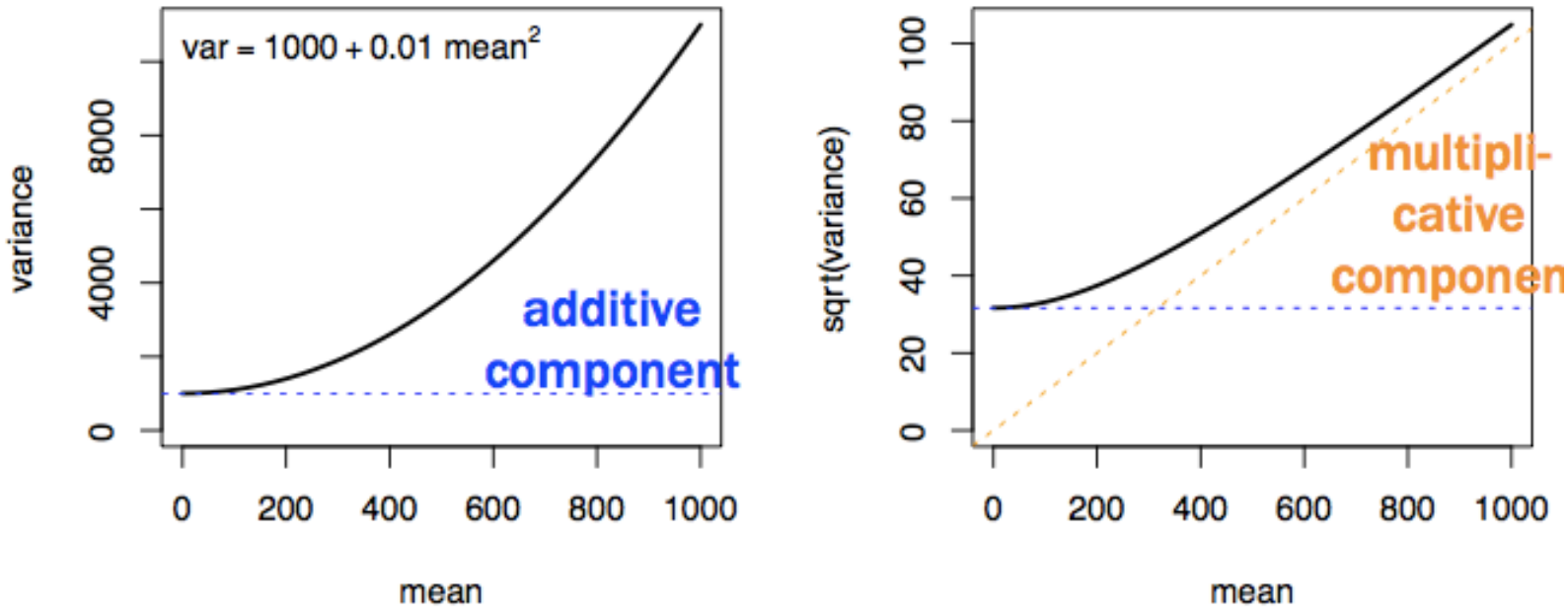
$$
\begin{aligned}
P(X=k) &= \int_0^\infty Po_\lambda(k) \times \gamma_{(r,\frac{p}{1-p})} d\lambda = \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} \times \frac{\lambda^{r-1} e^{-\lambda\frac{1-p}{p}}}{(\frac{p}{1-p})^r \Gamma(r)} d\lambda \\
&= \frac{(1-p)^r}{p^r k! \Gamma(r)} \int_0^\infty \lambda^{r+k-1} e^{-\lambda/p} d\lambda \\
&= \frac{(1-p)^r}{p^r k! \Gamma(r)} p^{r+k} \Gamma(r+k) \\
&= \frac{\Gamma(r+k)}{k! \Gamma(r)} p^k (1-p)^r
\end{aligned}
$$

Variance Stabilization

Prefer to deal with errors across samples which are independent and identically distributed.
In particular homoscedasticity (equal variances) across all the noise levels.
This is not the case when we have unequal sample sizes and variations in the accuracy across instruments.
A standard way of dealing with heteroscedastic noise is to try to decompose the sources of heterogeneity and apply transformations that make the noise variance almost constant.
These are called variance stabilizing transformations.

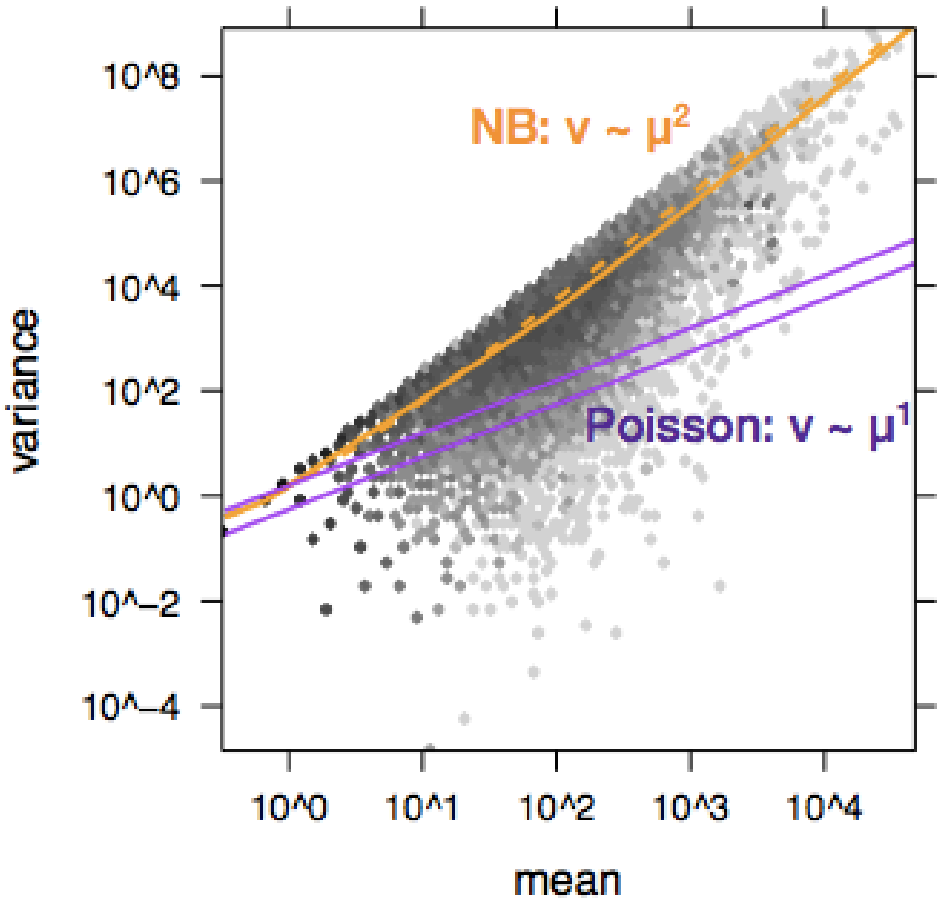
Take for instance different Poisson variables with mean μ_i. Their variances are all different if the μ_i are different. However, if the square root transformation is applied to each of the variables, then the transformed variables will have approximately constant variance. Actually if we take the transformation $x \longrightarrow 2\sqrt{x}$ we obtain a variance approximately equal to 1..

The additive-multiplicative error model



Trey Ideker et al.: JCB (2000)
David Rocke and Blythe Durbin: JCB (2001), Bioinformatics (2002)
For robust affine regression normalisation: W. Huber et al. Bioinformatics (2002)
For background correction in RMA: R. Irizarry et al., Biostatistics (2003)

Two component error models



Microarrays

$\text{var}(\mu) = b + c \cdot \mu^2$

b: background

c: asymptotic coefficient of variation

Sequencing counts

early edgeR:

$\text{var}(\mu) = \mu + \alpha \cdot \mu^2$

μ: from Poisson

α: dispersion

DESeq

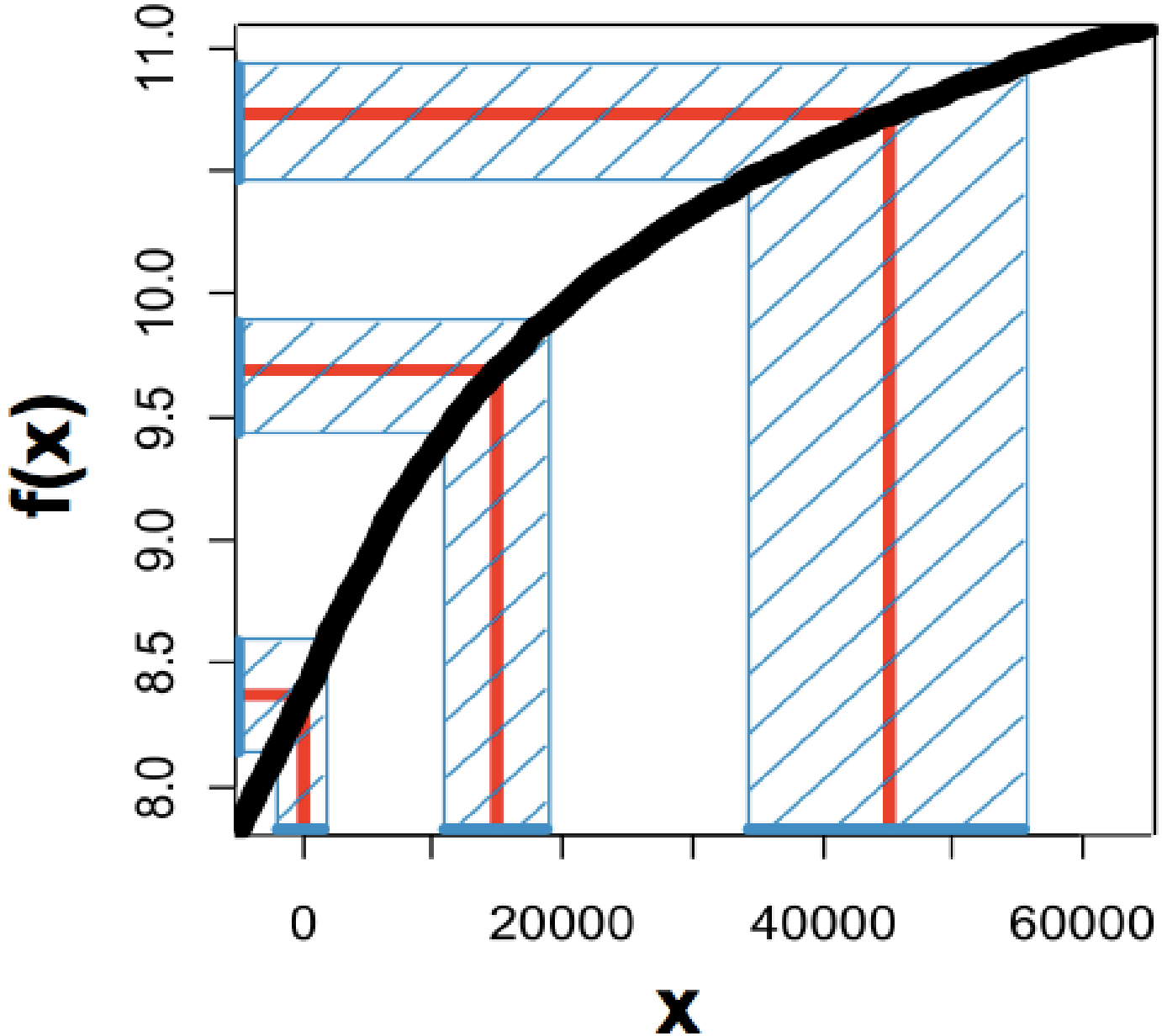
$\text{var}(\mu) = \mu + \alpha(\mu) \cdot \mu^2$

DESeq parametric option

$\alpha(\mu) = a_1/\mu + a_0 \quad \Leftrightarrow$

$\text{var}(\mu) = \mu + a_1 \cdot \mu + a_0 \cdot \mu^2$

variance stabilizing transformation



Modeling read counts

If technical replicates have same number of reads: s_j,
Poisson variation with mean $\mu = s_j u_i$.
Taxa i incidence proportion u_i.
Number of reads for the sample j and taxa i would be

$$K_{ij} \sim \text{Poisson}\,(s_j u_i)$$

Modeling Counts

For biological replicates within the same group -- such as treatment or control groups or the same environments -- the proportions u_i will be variable between samples.
Call the two parameters r_i and $\frac{p_i}{1-p_i}$.
So that U_{ij} the proportion of taxa i in sample j is distributed according to Gamma$(r_i, \frac{p_i}{1-p_i})$.
K_{ij} have a Poisson-Gamma mixture of different Poisson variables.
This gives the Negative Binomial with parameters $(m = u_i s_j)$ and ϕ_i as a satisfactory model of the variability.

Different Conditions

Samples belong to different conditions such as treatment and control or different environments.
Estimate the values of the parameters separately for each of the different biological replicate conditions/classes.
Use the index c for the different conditions, we then have the counts for the taxa i and sample j in condition c having a Negative Binomial distribution with $m_c = u_{ic}s_j$ and ϕ_{ic} so that the variance is written

$$u_{ic}s_j + \phi_{ic}s_j^2 u_{ic}^2 \tag{1}$$

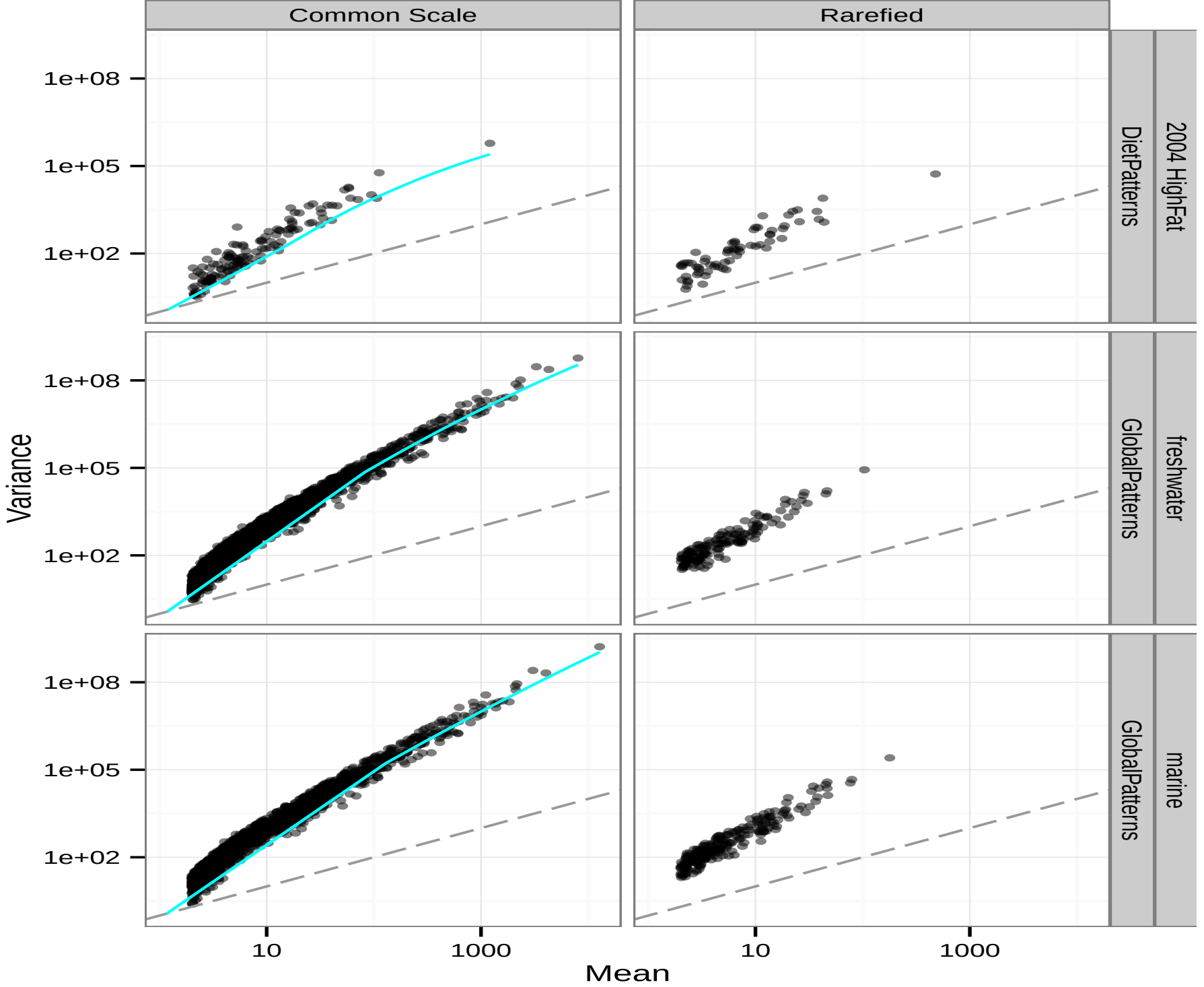
Estimate the parameters u_{ic} and ϕ_{ic} from the data for each OTU and sample condition.
The end result provides a variance stabilizing transformation of the data that allows a statistically efficient comparisons between conditions.
This application of a hierarchical mixture model is very similar to the random effects models used in the context of analysis of variance.

Using RNA-seq implementation : DESeq2

McMurdie and Holmes (2014) ``Waste Not, Want Not: Why rarefying microbiome data is inadmissible", to appear PLOS Computational Biology, Methods.

Examples of Overdispersion in Microbiome Data.

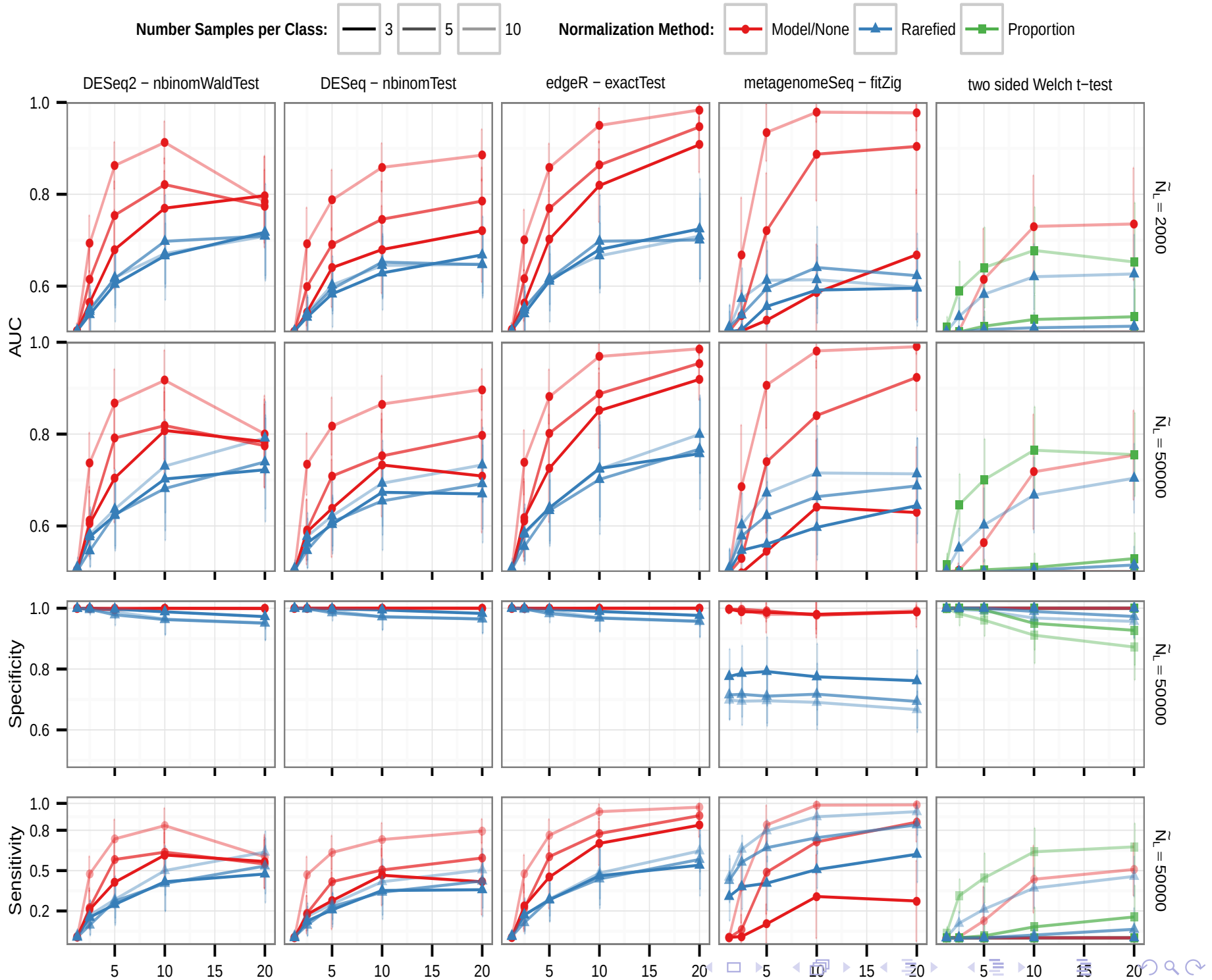
Common-Scale Variance versus Mean for Microbiome Data. Each point in each panel represents a different OTU's mean/variance estimate for a biological replicate and study. The data in this figure come from the Global Patterns surveyand the Long-Term Dietary Patterns study(Right) Variance versus mean abundance for rarefied counts. (Left) Common-scale variances and common-scale means, estimated according to the DESeq package. The dashed gray line denotes the $\sigma^2 = \mu$ case (Poisson; $\phi = 0$). The cyan curve denotes the fitted variance estimate using DESeq, with `method=`pooled', sharingMode=`fit-only', fitType=`local'`.


Common Scale
Rarefied
2004 HighFat
freshwater
marine
DietPatterns
GlobalPatterns
GlobalPatterns
Variance
Mean
1e+08
1e+05
1e+02
10
1000

Improvement in Power and FDR

Performance of differential abundance detection with and without rarefying summarized by "Area Under the Curve" (AUC) metric of a Receiver Operator Curve (ROC) (vertical axis).
Briefly, the AUC value varies from 0.5 (random) to 1.0 (perfect).
The horizontal axis indicates the effect size, shown as the factor applied to OTU counts to simulate a differential abundance.
Each curve traces the respective normalization method's mean performance of that panel, with a vertical bar indicating a standard deviation in performance across all replicates and microbiome templates.

The right-hand side of the panel rows indicates the median library size, N, while the darkness of line shading indicates the number of samples per simulated experiment.
Color shade and shape indicate the normalization method.
Detection among multiple tests was defined using a False Discovery Rate (Benjamini-Hochberg) significance threshold of 0.05.


Number Samples per Class: 3 5 10
Normalization Method: Model/None Rarefied Proportion
DESeq2 − nbinomWaldTest
DESeq − nbinomTest
edgeR − exactTest
metagenomeSeq − fitZig
two sided Welch t−test
AUC
Specificity
Sensitivity
$\tilde{N}_L$ = 2000
$\tilde{N}_L$ = 50000
$\tilde{N}_L$ = 50000
$\tilde{N}_L$ = 50000

Prinicipal Component Analysis: Dimension Reduction

PCA seeks to replace the original (centered) matrix X by a matrix of lower rank, this can be solved using the singular value decomposition of X:

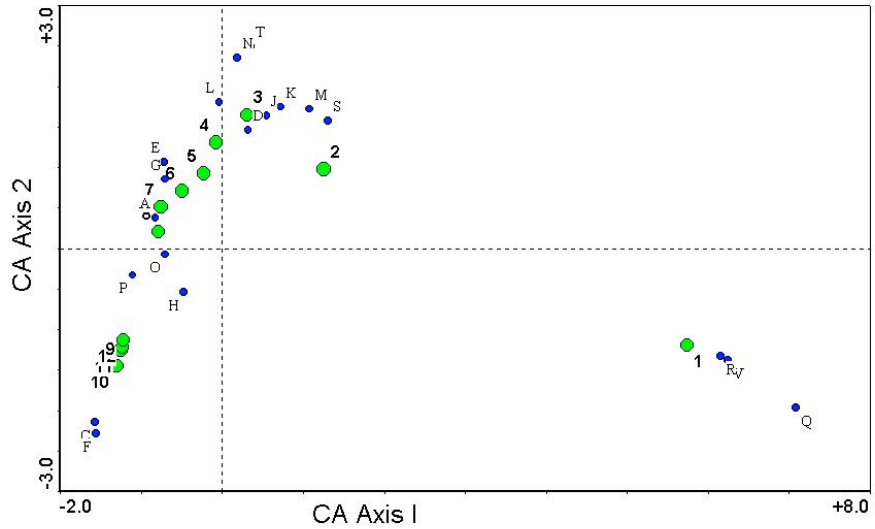
$$X = USV', \text{ with } U'DU = I_n \text{ and } V'QV = I_p \text{ and Sdiagonal}$$

$$XX' = US^2U', \text{ with } U'DU = I_n \text{ and } S^2 = \Lambda$$

PCA is a linear nonparametric multivariate method for dimension reduction. D and Q are the relevant metrics on the dual row and column spaces of n samples and p variables.

Ordination for Ecology built from Distances

The Boomlake plant data:



Biplot representing both species and locations
Blue circles with letters are species scores
Sampling locations are green circles with numbers.
Sample 1 is actually in the lake, and sample 12 is far away. Species are located closely to the samples they occur in. If you looked carefully into the data matrix, you would find that species R and Q are strictly aquatic, while species F is a

Multidimensional Scaling (MDS) also called PCoA

Simple classical multidimensional scaling.

- Square D elementwise $D^{(2)} = D_2$.
- Compute $\frac{-1}{2} H D_2 H = B$.
- Diagonalize B to find the principal coordinates SV'.
- Choose a number of dimensions by inspecting the eigenvalue's screeplot.

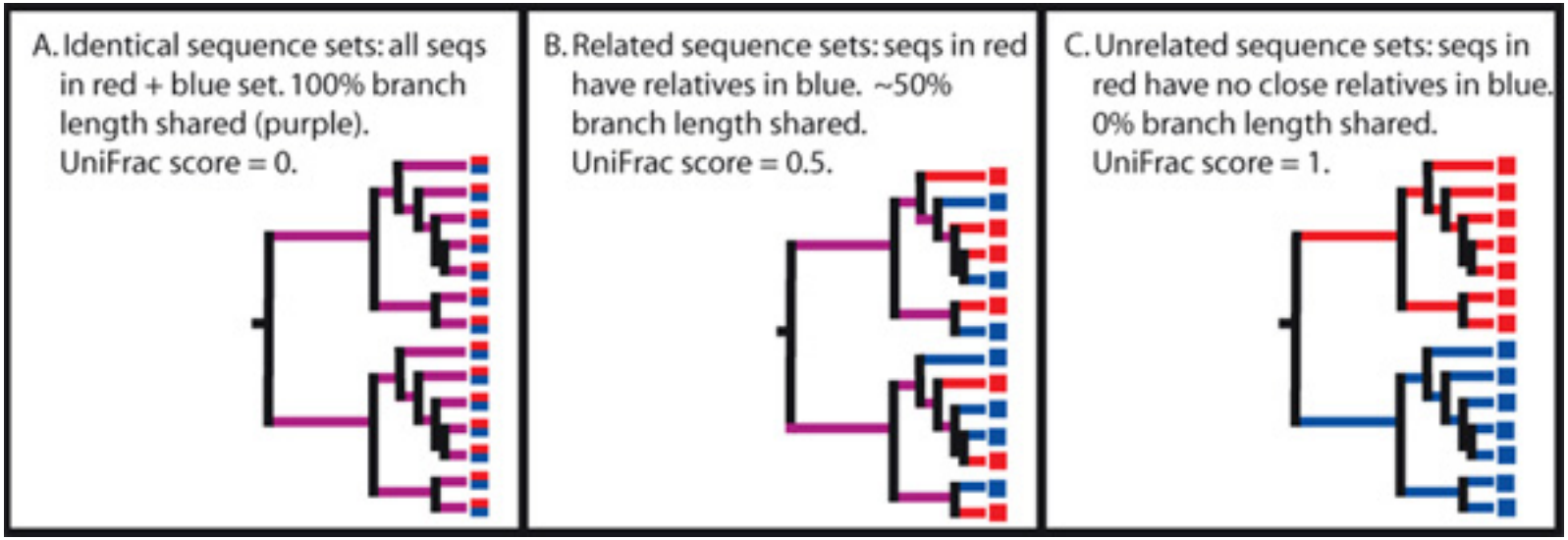
Examples using Phyloseq: Wiki Ordinate

Part IV

Combine and Compare Trees, Graphs and Contingent Data

Unifrac Distance (Lozupone and Knight, 2005)

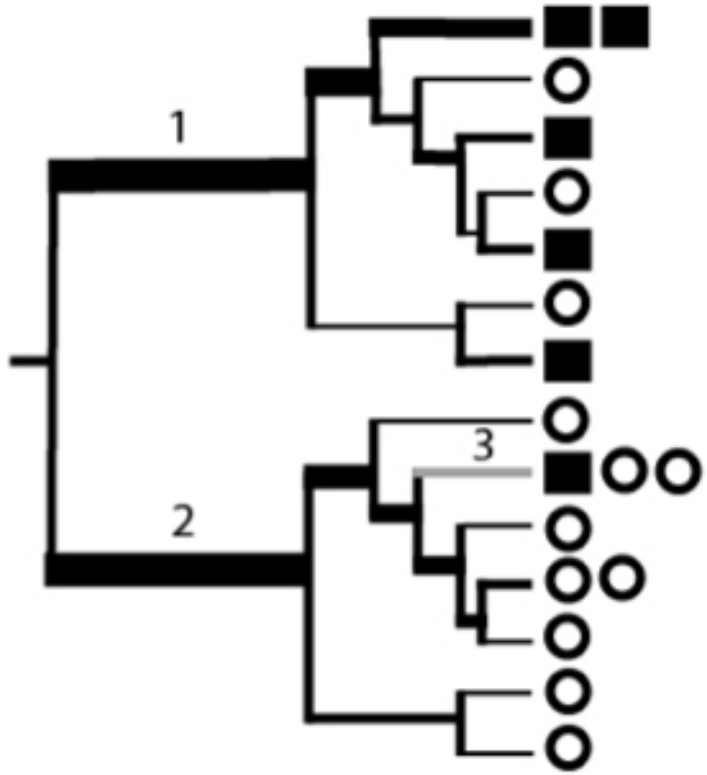
is a distance between groups of organisms that are related to each other by a tree.
Suppose we have the OTUs present in sample 1 (blue) and in sample 2(red).
Question: Do the two samples differ phylogenetically?
It is defined as the ratio of the sum of the lengths of the branches leading to members of group A or members of group B but not both to the total branch length of the tree.



Weighted Unifrac distance A modification of UniFrac, weighted UniFrac is defined in (Lozupone et al., 2007) as

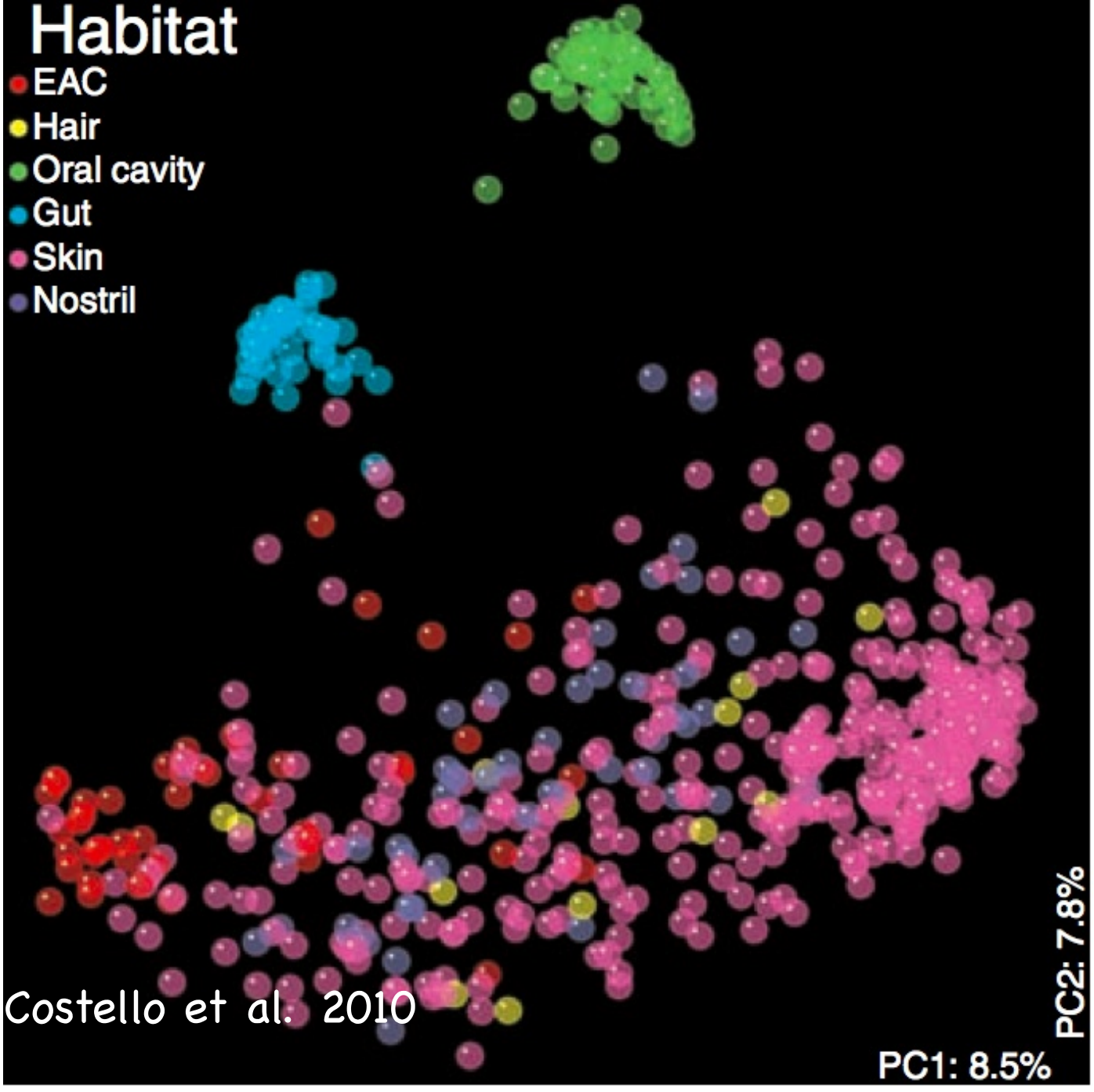
$$\sum_{i=1}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

- $n =$ number of branches in the
- b_i = length of the ith branch
- A_i = number of descendants of ith branch in group A
- A_T = total number of sequences in group A



[8]

or can be seen as by probabilists as the Wasserstein distance (earth movers) [6].

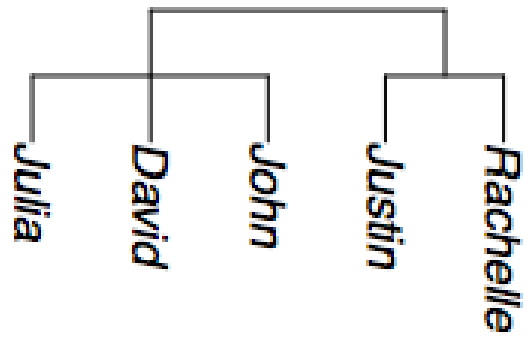


Costello et al. 2010

Rao's Distance

We start with a distance between individuals.
The heterogeneity of a population (H_i) is the average distance between members of that population.
The heterogeneity between two populations (H_{ij}) is the average distance between a member of population i and a member of population j.
The distance between two populations is

$$D_{ij} = H_{ij} - \frac{1}{2}(H_i + H_j)$$



	Julia	David	John	Justin
David	2			
John	2	2		
Justin	4	4	4	
Rachelle	4	4	4	2

Group 1 = Julia, David, John
Group 2 = Justin, Rachelle

$$H_1 = 0 \cdot 1/3 + 2 \cdot 2/3 = 4/3$$
$$H_2 = 0 \cdot 1/2 + 2 \cdot 1/2 = 1$$
$$H_{12} = 4 \cdot 1$$
$$D_{12} = H_{12} - \frac{1}{2}(H_1 + H_2)$$
$$= 4 - \frac{1}{2}(\frac{4}{3} + 1) \approx 2.8$$

Decomposition of Diversity

If we have populations $1, \ldots, k$ with frequencies $\pi_1, \ldots, \pi_k$, then the diversity of all the populations together is

$$H_0 = \sum_{i=1}^{k} \pi_i H_i + \sum_i \sum_j \pi_i \pi_j D_{ij} = H(w) + D(b)$$

Double Principal Coordinate Analysis

Pavoine et al. 2004 developed a method known as DPCoA [15] implemented in `ade4` [4].
Suppose we have n species in p locations and a (euclidean) matrix Δ giving the squares of the pairwise distances between the species. Then we can

- Use the distances between species to find an embedding in $n-1$ -dimensional space such that the euclidean distances between the species is the same as the distances between the species defined in Δ.
- Place each of the p locations at the barycenter of its species profile. The euclidean distances between the locations will be the same as the square root of the Rao dissimilarity between them.
- Use PCA to find a lower-dimensional representation of the locations.

Give the species and communities coordinates such that the inertia decomposes the same way the diversity does.

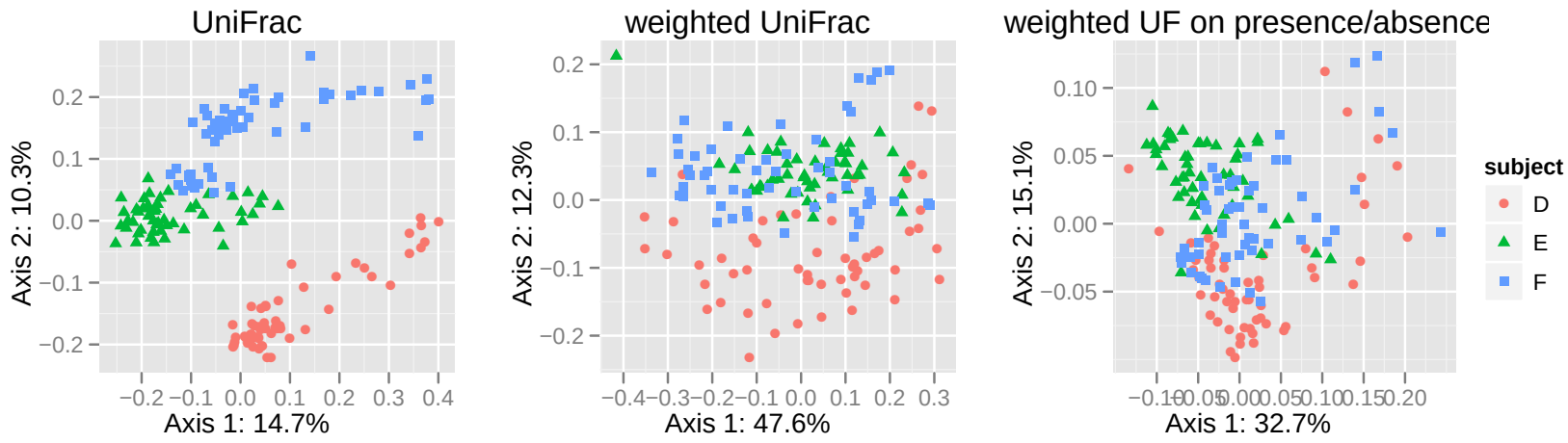
Fukuyama and Holmes, 2012.

Original description	New formula	Properties
square root of Rao's distance based on the square root of the patristic distances	$[\sum_i b_i(A_i/A_T - B_i/B_T)^2]^{1/2}$	Most sensitive to outliers, least sensitive to noise, upweights deep differences, gives OTU locations
$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	Less sensitive to outliers/more sensitive to noise than DPCoA
fraction of branches leading to exactly one group	$\sum_i b_i \mathbf{1}\{\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T} \geq 1\}$	Sensitive to noise, upweights shallow differences on the tree

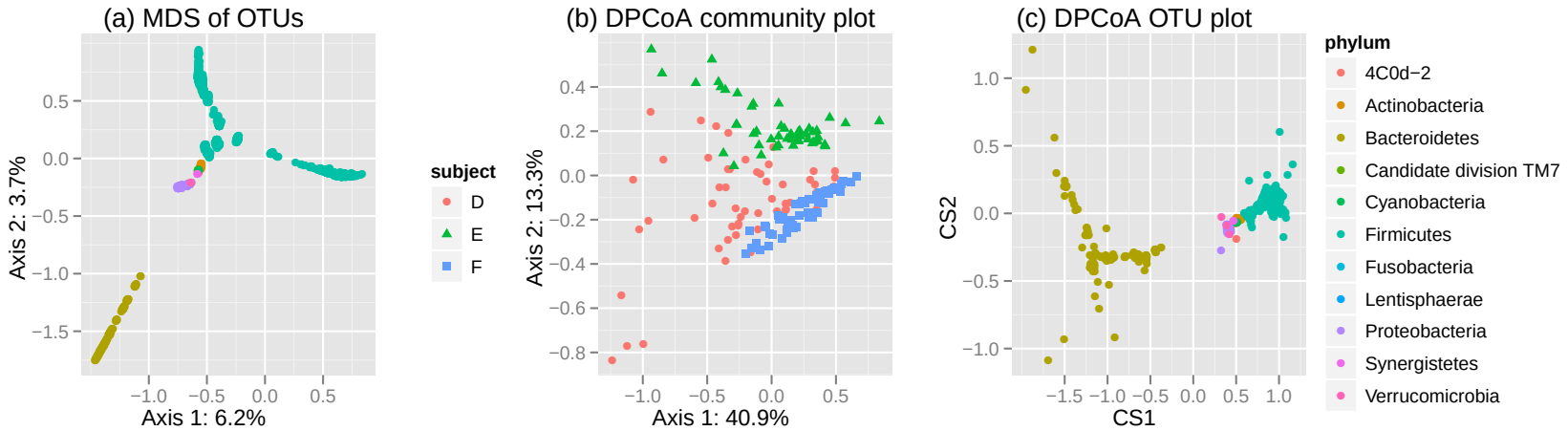
Summary of the methods under consideration. ``Outliers'' refers to highly abundant OTUs, and noise refers to noise in detecting low-abundance OTUs (see the text for more detail).

Antibiotic Time Course Data

Measurements of about 2500 different bacterial OTUs from stool samples of three patients (D, E, F)
Each patient sampled $\sim$ 50 times during the course of treatment with ciprofloxacin (an antibiotic).
Times categorized as Pre Cp, 1st Cp, 1st WPC (week post cipro), Interim, 2nd Cp, 2nd WPC, and Post Cp.



Comparing the UniFrac variants. From left to right: PCoA/MDS with unweighted UniFrac, with weighted UniFrac, and with weighted UniFrac performed on presence/absence data extracted from the abundance data used in the other two plots

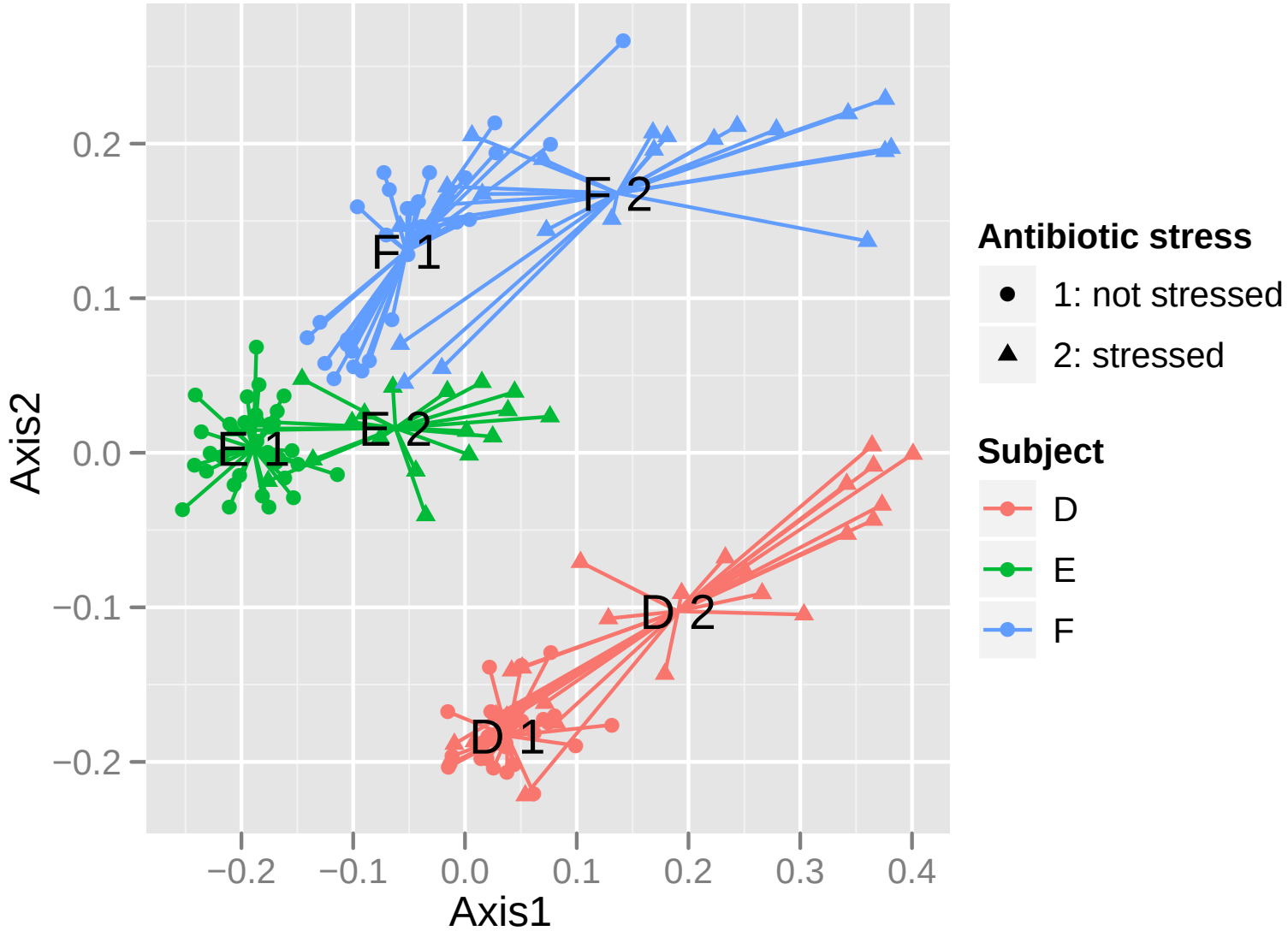


(a)

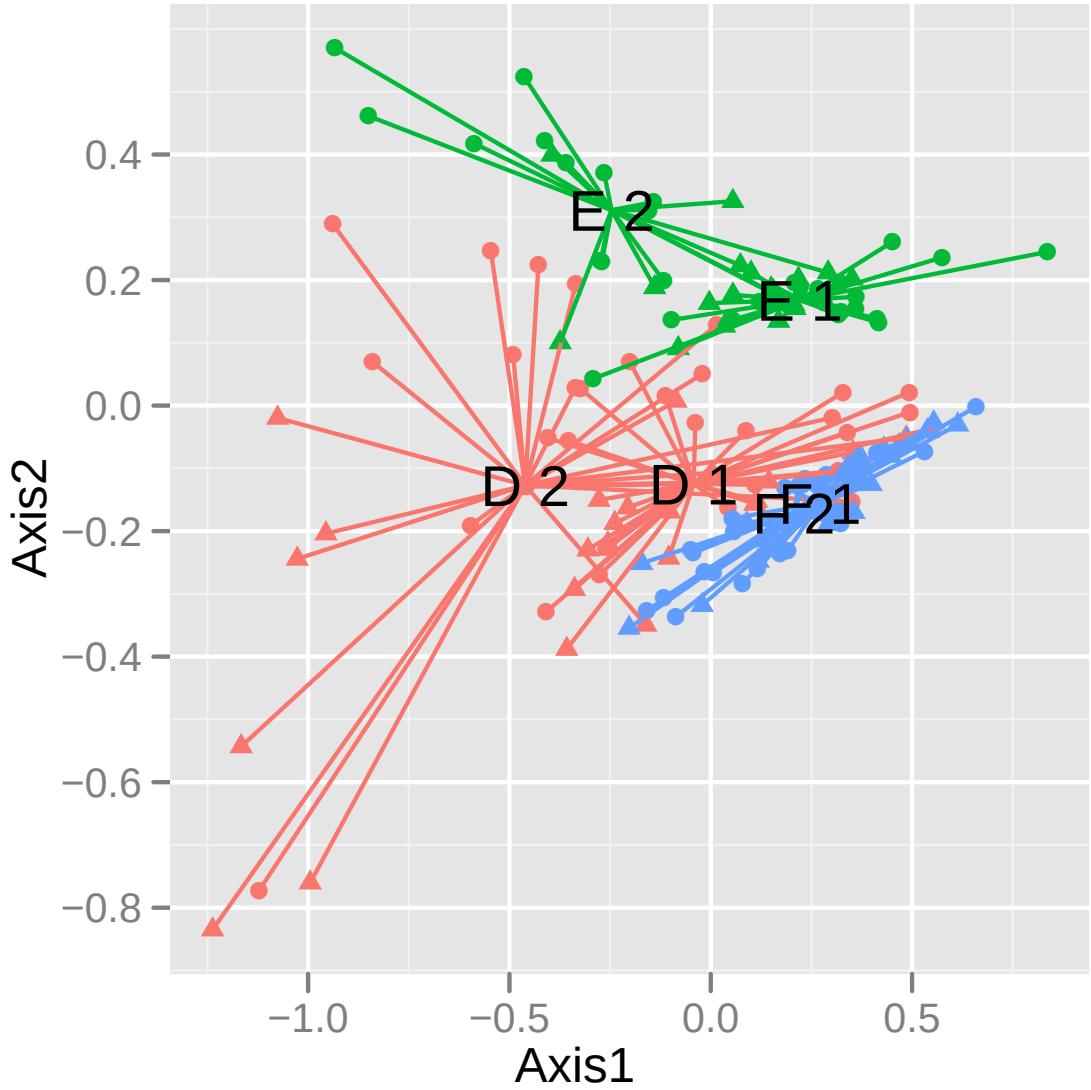
PCoA/MDS of the OTUs based on the patristic distance, (b) community and (c) species points for DPCoA after removing two outlying species.

Antibiotic Stress

We next want to visualize the effect of the antibiotic. Ordinations of the communities due to DPCoA and UniFrac with information about the whether the community was stressed or not stressed (pre cipro, interim, and post cipro were considered ``not stressed", while first cipro, first week post cipro, second cipro, and second week post cipro were considered ``stressed").
We see that for UniFrac, the first axis seems to separate the stressed communities from the not stressed communities. DPCoA also seems to separate the out the stressed communities along the first axis (in the direction associated with Bacteroidetes), although only for subjects D and E.



PCoA/MDS with unweighted UniFrac. The labels represent subject plus antibiotic condition.



Community points as represented by DPCoA. The labels represent subject plus antibiotic condition.

Conclusions for Antibiotic Stress

Since UniFrac emphasizes shallow differences on the tree and since PCoA/MDS with UniFrac seems to separate the subjects from each other better than the other two methods, we can conclude that the differences between subjects are mainly shallow ones. However, DPCoA also separates the subjects and the stressed versus non-stressed communities, and examining the community and OTU ordinations can tell us about the differences in the compositions of these communities.

Community Networks: Mouse experiment

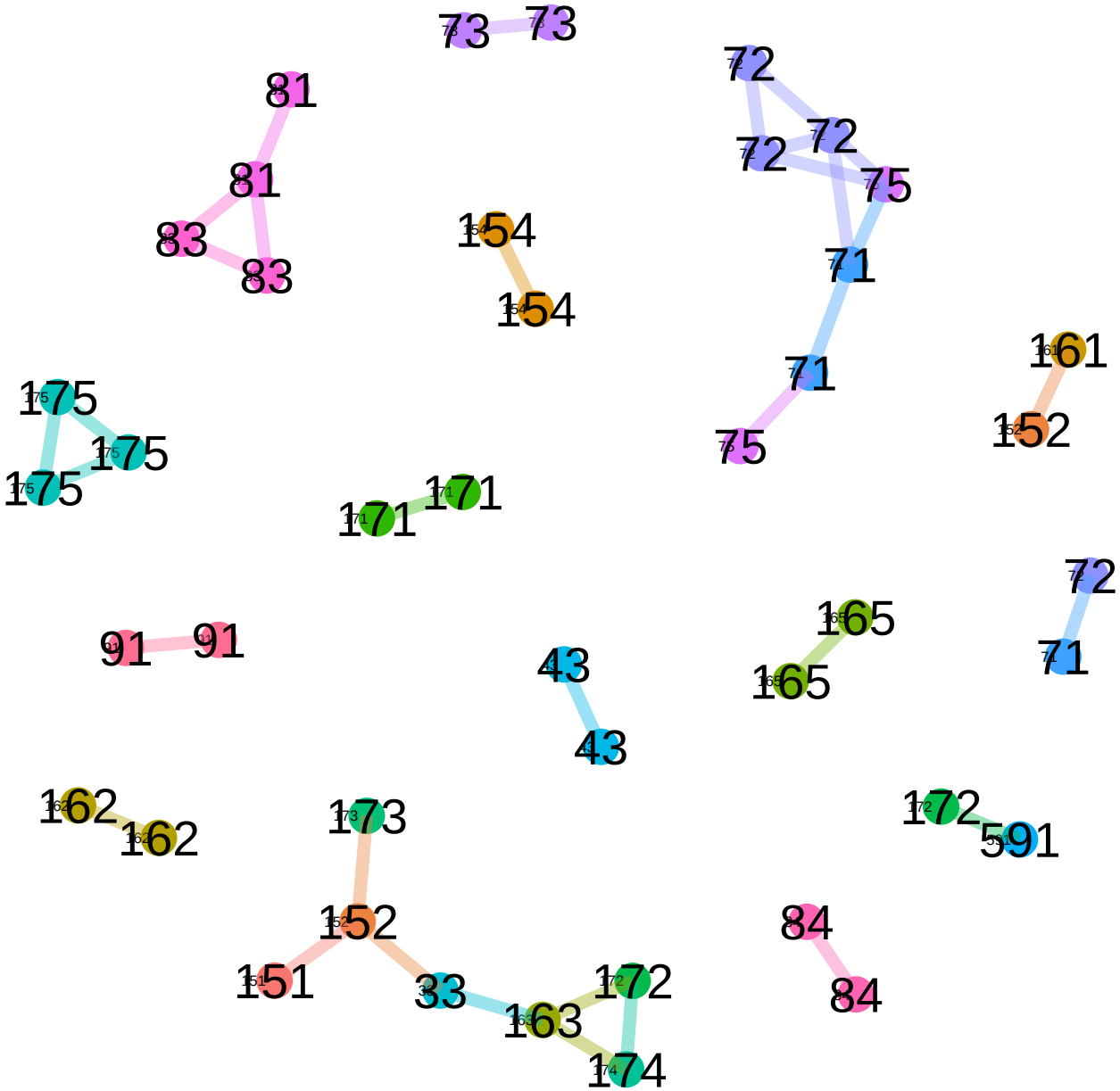
In work with David Relman and Yana Hoy, we have 6 mice measured first during a stable `reference pre-infection' time period and then over about 25 time points each and we examine the bacteria in their gut using the 16sRNA gene as an otu marker:

- Software pipeline incorporating the phylogenetic tree of relations between the species, the abundance table and the taxonomic ranks of the otus as well as the sample information.
- Measure more than just the diversity or abundances of the bacteria.
- Seek to characterize the dynamics of the bacterial communities in the gut.

Bacterial communities

- Data are normalized counts in a species $\times$ samples matrix.
- Two dual networks are possible:
  - Connect two species if the co-occur in more than 70% of the samples (using a Jaccard index).
  - Connect two samples if they share more than 70% of the bacterial species.

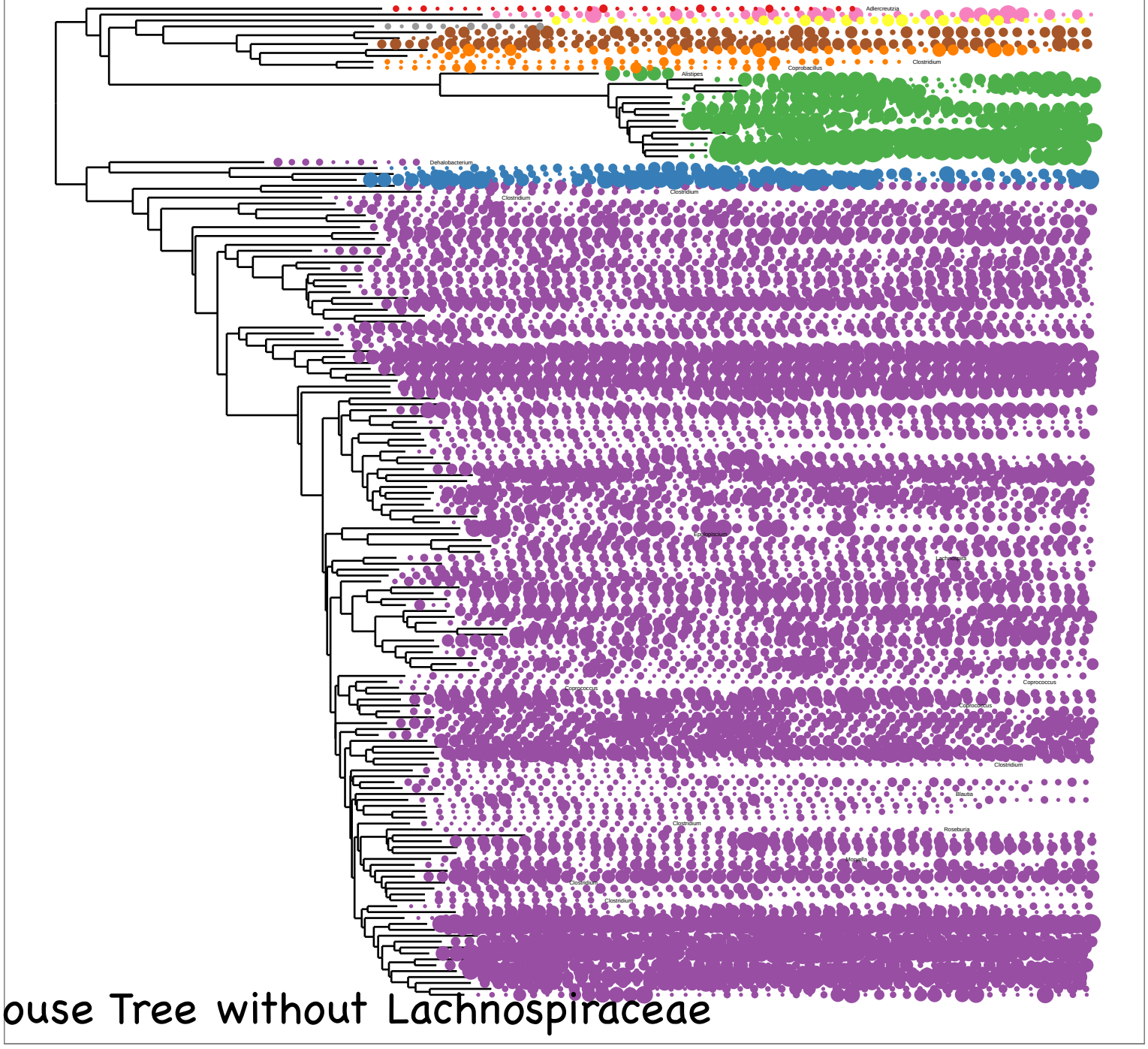
Example: Co-occurrence graph network.



Mouse Community Data

The data are filtered and combined, for the time course data we had:

```
phyloseq-class experiment-level object
otu_table()  OTU Table:      [410 taxa and 146 samples]
sample_data()Sample Data:    [146 samples by 9 sample varia
tax_table()  Taxonomy Table: [410 taxa by 7 taxonomic ranks
phy_tree()   Phylogenetic Tree[410 tips and 409 internal no
```



Mouse Tree without Lachnospiraceae

Bacteria `sharing' between mice

Using the Jaccard index that measures the co-incidence or co-occurrence of species between mice.

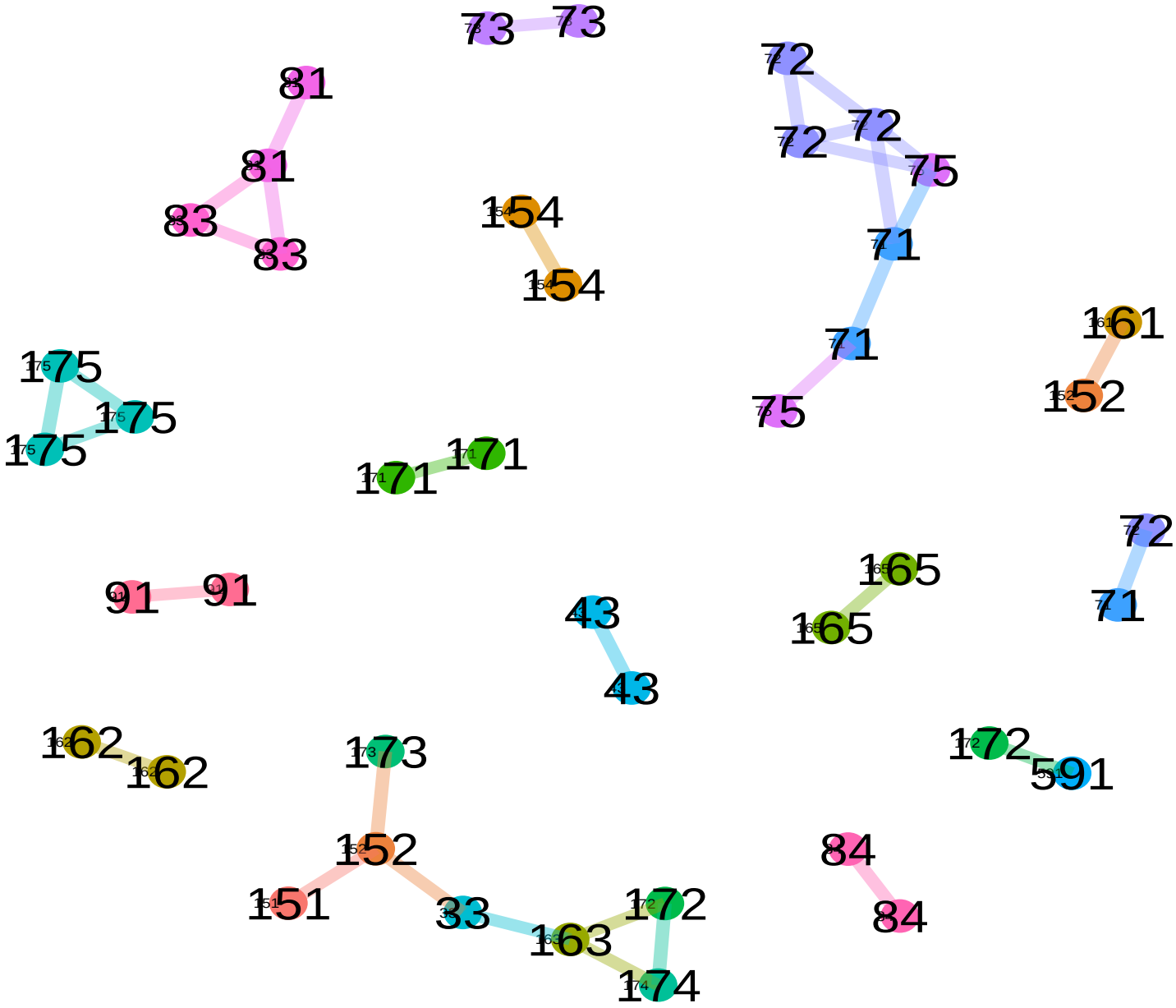
$$\text{Jaccard Similarity} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

$$\text{Jaccard Disimilarity} = \frac{f_{01} + f_{10}}{f_{01} + f_{10} + f_{11}}$$

```
mouse1
0 0 0 1 0 1 0 1 0 0 0 0 0 0 1
mouse4
1 0 0 0 0 0 0 0 0 0 0 0 0 0 1
vegdist(rbind(mouse1,mouse4),method="jaccard")
0.8
```

Bacteria `sharing' between mice as a network

```
netbaseline=make_network(phy_pifn_glom)
p=plot_network(netbaseline,phy_pifn_glom,
color="mousenames",label="mousenames",point_size=7)
+geom_text(aes(label=mousenames),size=7)
p+scale_colour_hue(guide="none")
```

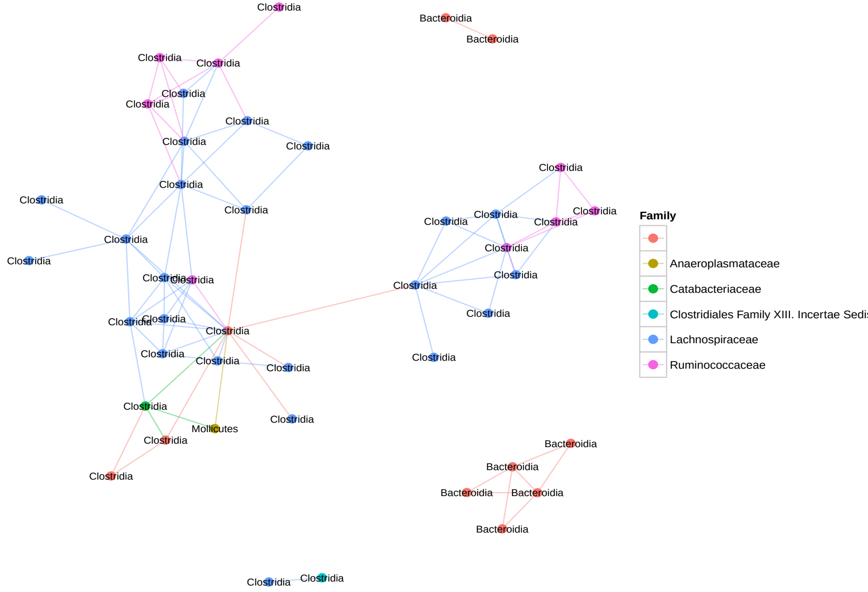

73
73
81
81
83
83
154
154
72
72
72
75
71
71
75
161
152
175
175
175
171
171
91
91
72
71
165
165
43
43
162
162
173
152
151
33
163
172
174
172
591
84
84

Does the network relate to `communities'?

Friedman and Rafsky (1979) devised a nonparametric test for multivariate data using the minimum spanning tree with any metric.
Then compute the number of `pure' edging connecting labels from the same groups compared to the mixed edges connecting labels from different groups, call F_o the observed statistic.
In our example: $F_o = 82$
Keeping the graph fixed, permute the labels and recompute the number of pure edges.
All 1000 simulated values had $F_s < 82$ so $p < 0.001$.

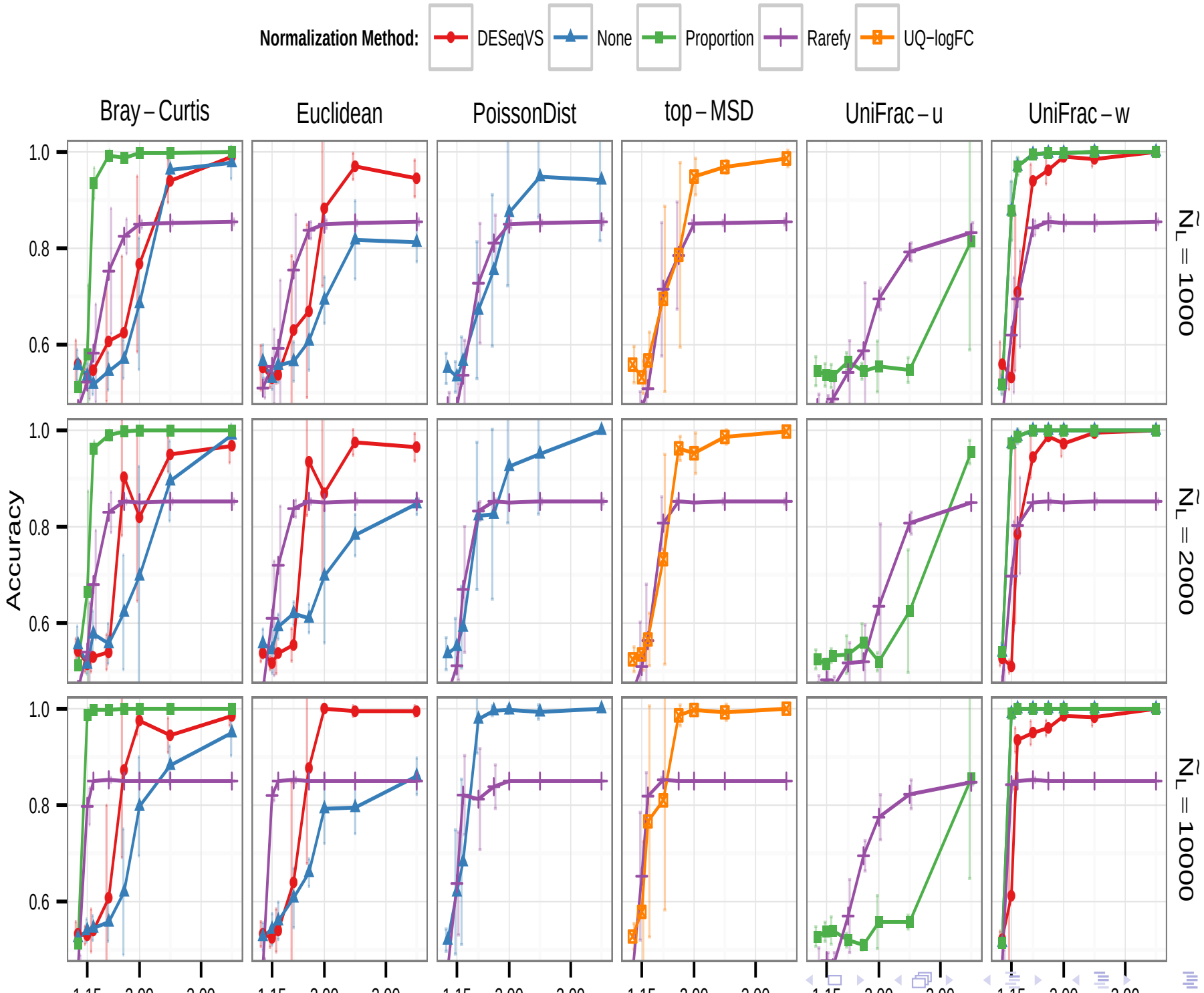
Co-occurrence networks for taxa of the baseline mice

```
p=plot_network(netbasetaxa,phy_pifn_glom,color="Family",
type="taxa",label=NULL)
p+geom_text(aes(label=Class),size=3)
```


Family
Anaeroplasmataceae
Catabacteriaceae
Clostridiales Family XIII. Incertae Sedi
Lachnospiraceae
Ruminococcaceae
Clostridia
Bacteroidia
Bacteroidia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Clostridia
Mollicutes
Clostridia
Clostridia
Bacteroidia
Bacteroidia
Bacteroidia
Bacteroidia
Bacteroidia
Clostridia
Clostridia

Improvements of Distance based clustering

Clustering accuracy in simulated two-class mixing.
Partitioning around medoids clustering accuracy (vertical axis) that results following different normalization and distance methods.
Points denote the mean values of replicates, with a vertical bar representing one standard deviation above and below.
The horizontal axis is the effect size, which in this context is the ratio of target to non-target values in the multinomials that were used to simulate microbiome counts.
Each multinomial is derived from two microbiomes that have negligible overlapping OTUs (Fecal and Ocean microbiomes in the Global Patterns dataset).
Higher values of effect size indicate an easier clustering task.


Normalization Method:
DESeqVS
None
Proportion
Rarefy
UQ−logFC
Bray − Curtis
Euclidean
PoissonDist
top − MSD
UniFrac − u
UniFrac − w
Accuracy
1.0
0.8
0.6
$N_L = 1000$
$N_L = 2000$
$N_L = 10000$

Simulation Details Distinguish patterns of relationships between whole microbiome samples through normalization followed by the calculation of sample-wise distances.
Standard was to use rarefying then calculating UniFrac distances.
In some cases, formal testing of sample covariates is also done using a permutation MANOVA (e.g. `vegan::adonis` in R) with the (squared) distances and covariates as response and linear predictors.
Relative discriminating capability of each combination of normalization method and distance measure.
We will use clustering results as a quantitative proxy for the broad spectrum of approaches taken to interpret microbiome sample distances.

Normalizations in Simulation A

For each simulated experiment we used the following normalization methods prior to calculating sample-wise distances.

1. **DESeqVS**. Variance Stabilization implemented in the DESeq package.
2. **None**. Counts not transformed. Differences in total library size could affect the values of some distance metrics.
3. **Proportion**. Counts are divided by total library size.
4. **Rarefy**. Rarefying is performed as defined in the introduction, using `rarefy_even_depth` implemented in the phyloseq package. with $N_{L,min}$ set to the 15^{th}-percentile of library sizes within each simulated experiment.
5. **UQ-logFC**. The Upper-Quartile Log-Fold Change normalization implemented in the edgeR package, coupled with the top-MSD distance.

Distances in Simulation A

For each of the previous normalizations we calculated sample-wise distance/dissimilarity matrices using the following methods, if applicable.

1. **Bray-Curtis**. The Bray-Curtis dissimilarity first defined in 1957 for forest ecology.
2. **Euclidean**. The euclidean distance treating each OTU as a dimension. $\sqrt{\sum_{i=1}^{n}(K_{i1} - K_{i2})^2}$, is the distance between samples 1 and 2,n the number of distinct OTUs.
3. **PoissonDist**. Our abbreviation of `PoissonDistance`, a sample-wise distance implemented in the PoiClaClu package (Witten,2011).
4. **top-MSD**. The mean squared difference of top OTUs, as implemented in edgeR.
5. **UniFrac-u**. The Unweighted UniFrac distance (Lozupone,2005).
6. **UniFrac-w**. The Weighted UniFrac distance (Lozupone,2007).

In order to consistently evaluate performance in this regard, we generated microbiome counts by sampling from two different multinomials that were based on either the Ocean or Feces microbiomes of the Global Patterns empirical dataset. An equal number of simulated microbiome samples was generated from each multinomial. The Ocean and Feces sample classes have negligible overlapping OTUs.
Mixing them by a defined proportion allows control over the difficulty of the clustering task from trivial (no mixing) to impossible (both multinomials evenly mixed).
Clustering was performed independently for each combination of simulated experiment, normalization method, and distance measure using partitioning around medoids (PAM).
The accuracy is the fraction of simulated samples correctly clustered; worst possible accuracy is 50% if all samples are clustered. (Rarefying procedure omits samples, so its accuracy can be below 50%)

Part V

Reproducibility

Non-reproducibility: Enterotypes Study

Enterotypes Revisited

Goals already attained:

- Ways of combining heterogeneous data: distances and multivariate representation.
- Data integration `phyloseq`.
- Tree-table methods : unifrac and coinertia analysis.
- Modeling mixtures: Variance Stabilizing transformations.
- Threshold, sensitivity tests and modeling simulations.
- Reproducibility: open source standards, publication of source code and data. (`phyloseq`).

Benefitting from the tools and schools of Statisticians.....

Thanks to the R community: Chessel, Jombart, Dray, Thioulouse ade4 , Wolfgang Huber, Michael Love for tt DESeq2 and Emmanuel Paradis for ape.

Collaborators: David Relman, Alfred Spormann, Les Dethfelsen, Justin Sonnenburg, Persi Diaconis, Elisabeth Purdom.

Postdoctoral Fellows Paul (Joey) McMurdie, Alex Alekseyenko (NYU), Ben Callahan, Angela Marcobal, Serban Nacu.
Students: Elizabeth Purdom, Alden Timme, Katie Shelef, Yana Hoy, John Chakerian, Julia Fukuyama, Kris Sankaran.
Funding from CIMI, Toulouse, NIH/ NIGMS R01, NSF-VIGRE and NSF-DMS.

phyloseq

Joey McMurdie (joey711 on github).
Available in Bioconductor.
How can I (my students) learn more?
SAMSI program
http://www-stat.stanford.edu/~susan/


L. Billera, S. Holmes, and K. Vogtmann.
The geometry of tree space.
Adv. Appl. Maths, 771--801, 2001.


J. Chakerian and S. Holmes.
`distory`:Distances between trees, 2010.


Sourav Chatterjee.
Matrix estimation by universal singular value thresholding.
arXiv preprint arXiv:1212.1247, 2012.


Daniel Chessel, Anne Dufour, and Jean Thioulouse.
The ade4 package - i: One-table methods.
R News, 4(1):5--10, 2004.


P. Diaconis, S. Goel, and S. Holmes.
Horseshoes in multidimensional scaling and kernel methods.
Annals of Applied Statistics, 2007.


Steven N Evans and Frederick A Matsen.

The phylogenetic kantorovich-rubinstein metric for environmental sequence s amples.
arXiv, q-bio.PE, Jan 2010.


Jerome H Friedman and Lawrence C Rafsky.
Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests.
The Annals of Statistics, pages 697--717, 1979.


M Hamady, C Lozupone, and R Knight.
Fast unifrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data.
The ISME Journal, Jan 2009.


Susan Holmes.
Multivariate analysis: The French way.
In D. Nolan and T. P. Speed, editors, Probability and Statistics: Essays in Honor of David A. Freedman, volume 56 of IMS Lecture Notes--Monograph Series. IMS, Beachwood, OH, 2006.


S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb.
Picante: R tools for integrating phylogenies and ecology.
Bioinformatics, 26(11):1463--1464, 2010.


K. Mardia, J. Kent, and J. Bibby.
Multiariate Analysis.
Academic Press, NY., 1979.


P. J. McMurdie and S. Holmes.
Phyloseq: A bioconductor package for handling and analysis of high-throughput phylogenetic sequence data.


P. J. McMurdie and S. Holmes.
Phyloseq: Reproduible research platform for bacterial census data.
PlosONE, 2013.
April 22,.


P. J. McMurdie and S. Holmes.

Waste not, want not: Why rarefying microbiome data is inadmissible.
Plos Computational Biology, 2014.
to appear.


S Pavoine, S Ollier, and D Pontier.
Measuring diversity from dissimilarities with rao's quadratic entropy: are any dissimilarities suitable?
Theoretical population biology, 67(4):231--9, Jun 2005.


C. R. Rao.
The use and interpretation of principal component analysis in applied research.
Sankhya A, 26:329--359., 1964.

Other mixture models

If instead of modeling the read counts one uses the proportions as the random variables, with differing variances due to different library sizes, the Beta-Binomial model is the standard approach.

Laplace Distribution

Purdom and Holmes (2005)

Error Distribution for Gene Expression Data

Statistical Applications in Genetics and Molecular Biology, Vol. 4 [2005], Iss. 1, Art. 16

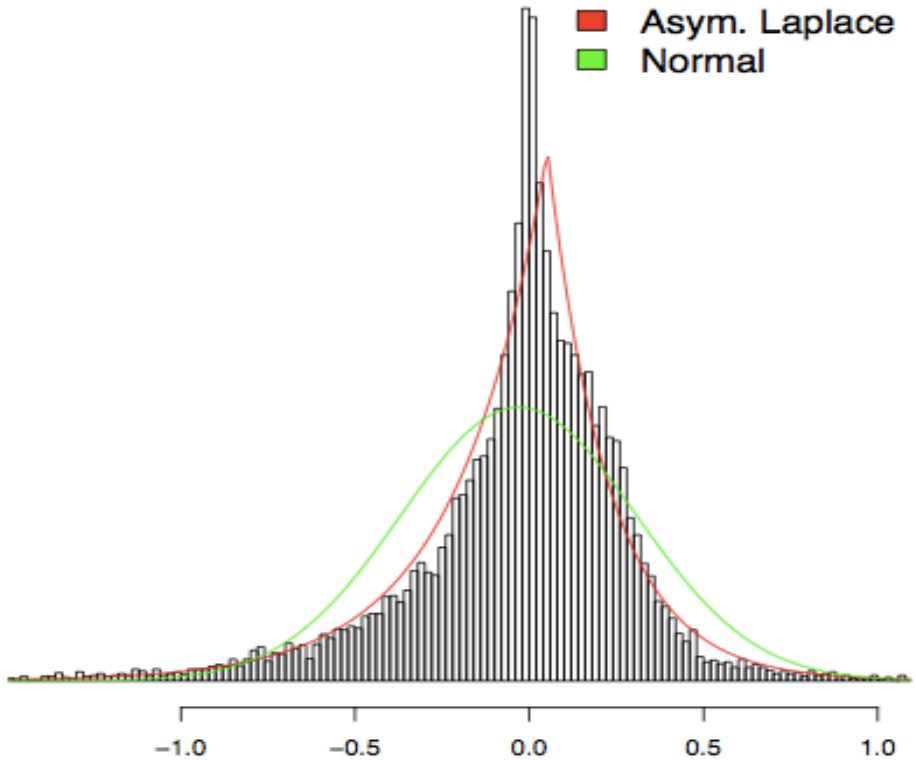


Figure 1: Histogram of gene expression of all genes on a single cDNA microarray from T-Cell Data (described below in Section 4.1). Corresponding Asymmetric Laplace and Normal distribution overlayed, with parameters estimated using maximum likelihood estimates.

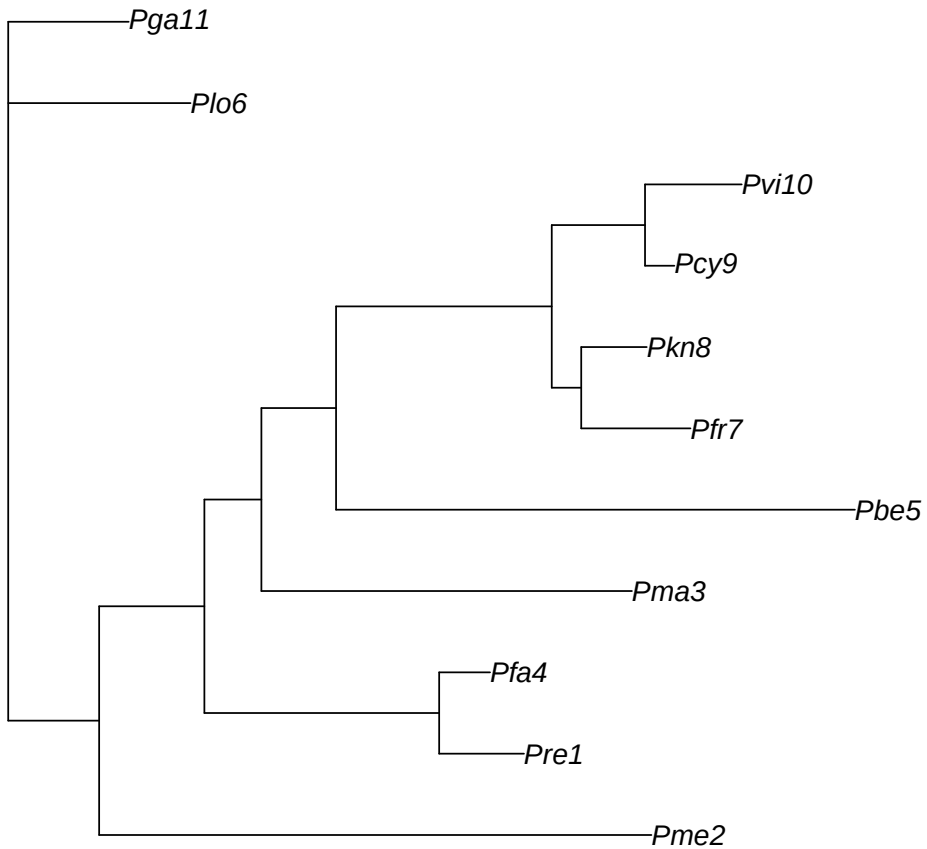
Laplace Distribution: mixture

Y can be viewed a continuous mixture of normal random variables whose scale and mean parameters are dependent and vary according to an exponential distribution:

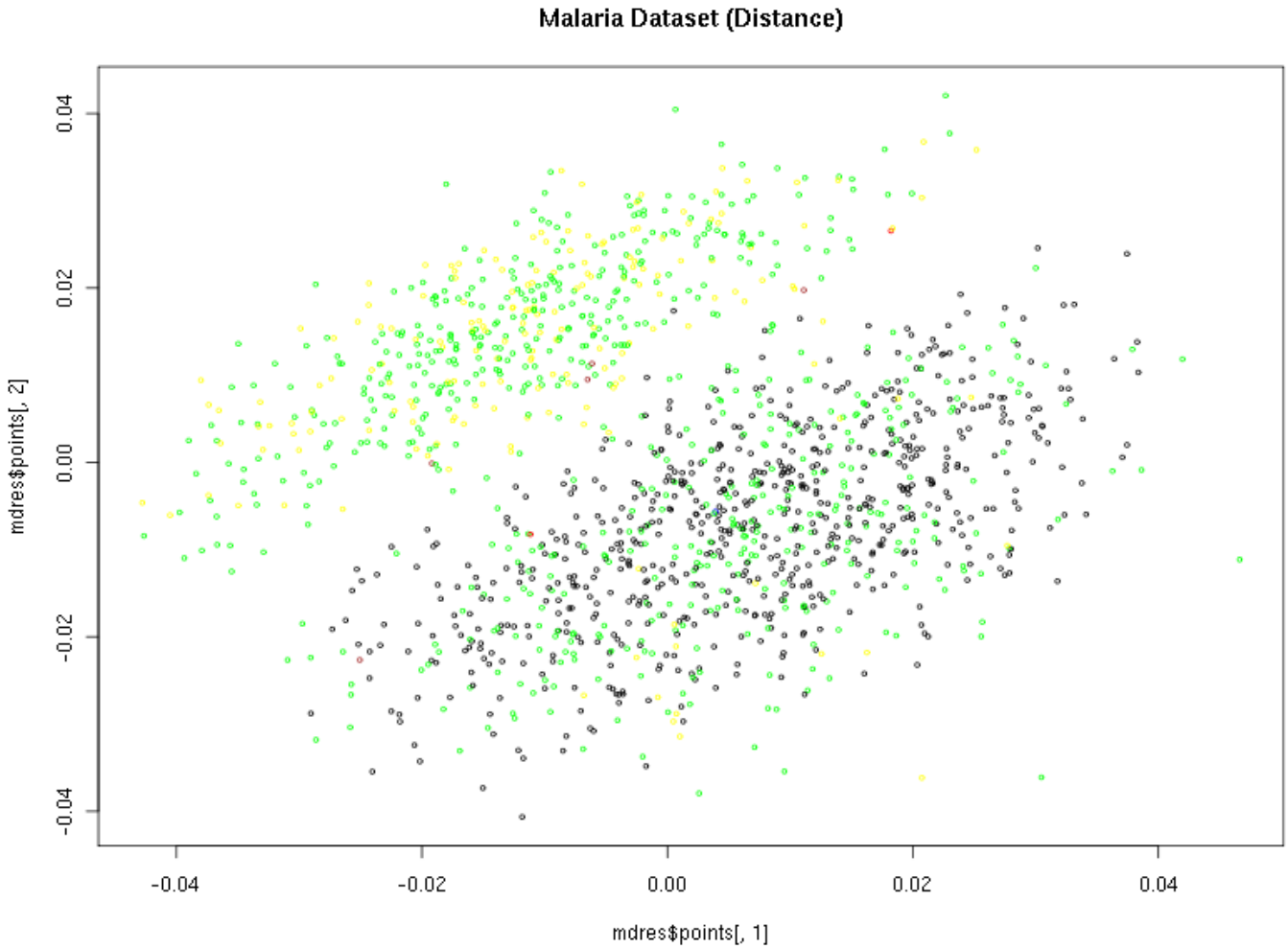
$$Y_i|W_i \sim \mathcal{N}(\theta + \mu W_i, \sigma^2 W_i)$$

, where $W_i \sim \exp(1)$

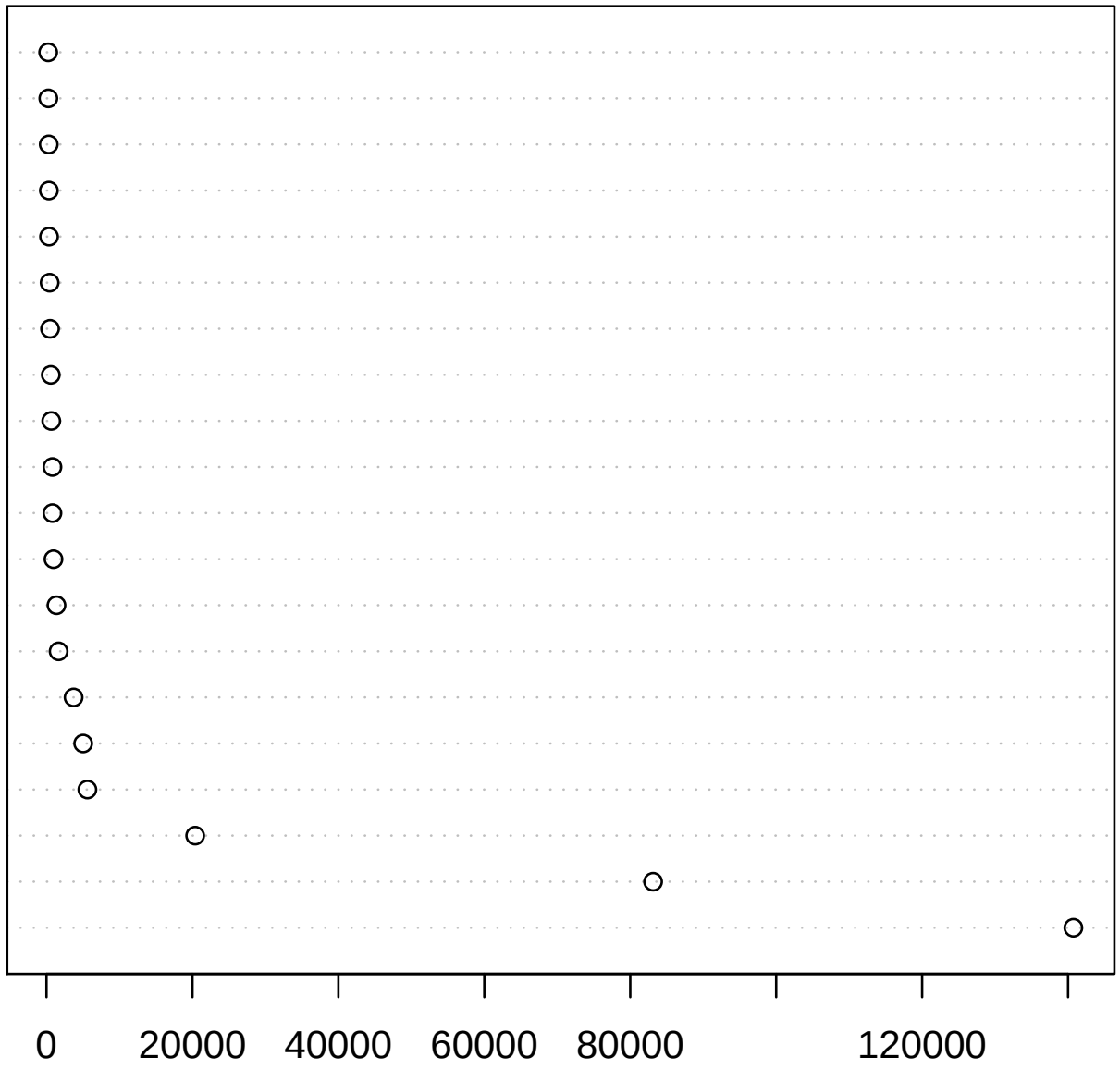
Malaria Data as seen using ape

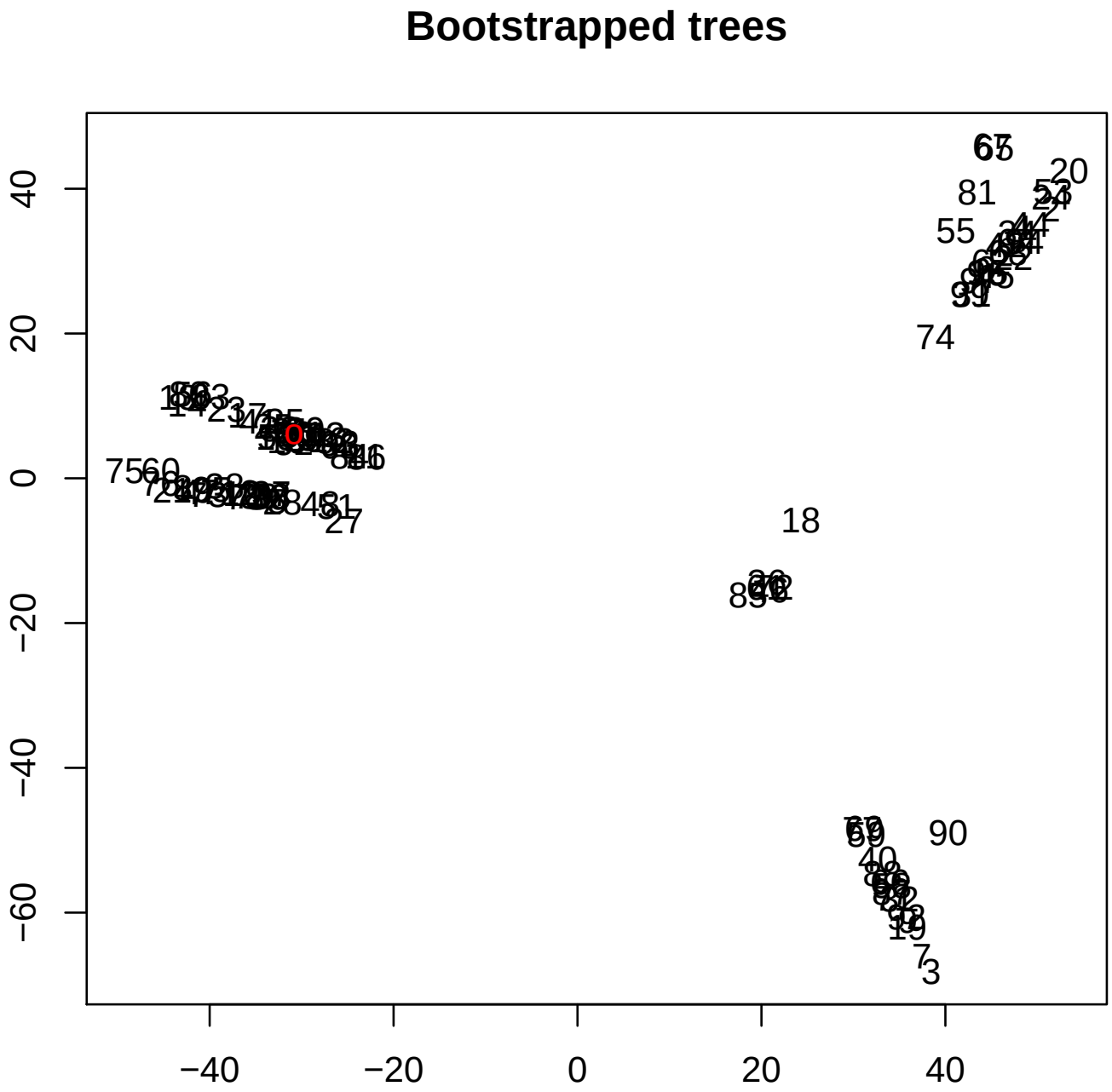


Bootstrap of Malaria Data



Eigenvalues of MDS for bootstrapped trees




Bootstrapped trees
40
20
0
−20
−40
−60
−40
−20
0
20
40
20
81
55
74
18
90
3

Data Analysis: Geometrical Approach

i. The data are p variables measured on n observations.

ii. X with n rows (the observations) and p columns (the variables).

iii. D_n is an $n \times n$ matrix of weights on the ``observations", which is most often diagonal.

iv Symmetric definite positive matrix Q,often

$$Q = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & \dots \\ 0 & \frac{1}{\sigma_2^2} & 0 & 0 & \dots \\ 0 & 0 & \frac{1}{\sigma_3^2} & 0 & \dots \\ \dots & \dots & \dots & 0 & \frac{1}{\sigma_p^2} \end{pmatrix}.$$

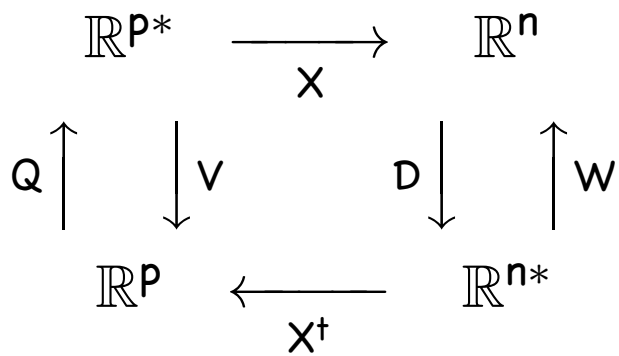
Euclidean Spaces

These three matrices form the essential ``triplet'' $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ defining a multivariate data analysis.
Q and D define geometries or inner products in $\mathbb{R}^p$ and $\mathbb{R}^n$, respectively, through

$$x^t Q y = < x, y >_Q \qquad x, y \in \mathbb{R}^p$$
$$x^t D y = < x, y >_D \qquad x, y \in \mathbb{R}^n.$$

An Operator Approach (Holmes, 2005)

- Q can be seen as a linear function from $\mathbb{R}^p$ to $\mathbb{R}^{p*} = \mathcal{L}(\mathbb{R}^p)$, the space of scalar linear functions on $\mathbb{R}^p$.
- D can be seen as a linear function from $\mathbb{R}^n$ to $\mathbb{R}^{n*} = \mathcal{L}(\mathbb{R}^n)$.
-



Properties of the Diagram

Rank of the diagram: X, X^t, VQ and WD all have the same rank.
For Q and D symmetric matrices, VQ and WD are diagonalisable and have the same eigenvalues.

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_r \geq 0 \geq \cdots \geq 0.$$

Eigendecomposition of the diagram: VQ is Q symmetric, thus we can find Z such that

$$VQZ = Z\Lambda, Z^tQZ = \mathcal{I}_p, \text{ where } \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p). \quad (2)$$