Using the data, **all** the data: the computational challenges

Susan Holmes
@SherlockpHolmes

BIRS Meeting online, June, 18th 2020

Bio-X and Statistics, Stanford University


Homogeneous data are all alike;
all heterogeneous data are
heterogeneous
in their own way.

Example 1: Human Microbiome

Joint work with David Relman and his Lab, funded by NIH TR01: Perturbations and Resilience of the Human Microbiome and March of Dimes.

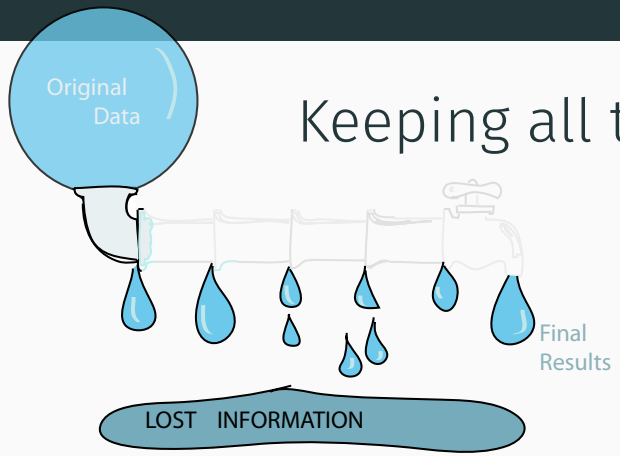
- Effect of Antibiotics.
- Colonic Cleanout.
- Diet perturbations.
- .....and March of Dimes study of pregnancy.

Example 2: Immune Cells and their role in virus response.

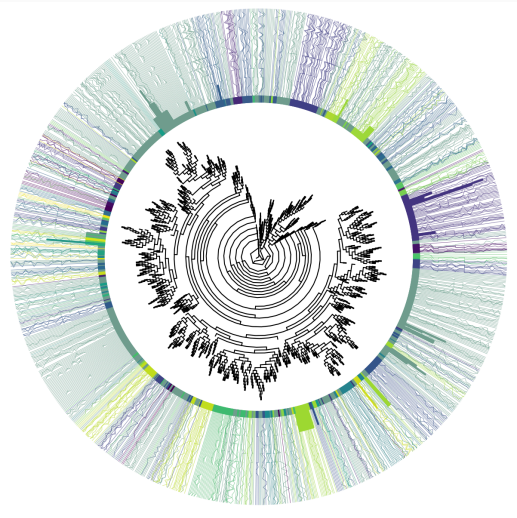
Joint work with Catherine Blish and her Lab, funded by the NIH.

- Influenza, Pregnancy and NK cells.
- NK cells and HIV response.

Challenges when working with Longitudinal Multidomain data



Keeping all the data together



- Data heterogeneity and Information leeks.
- Multidomain, multiway, multimodal, multitable data integration.
- Longitudinal, data are dependent (less information).
- Reproducibility of results across labs, experimental conditions and users.
- Confirmatory analyses: uncertainty propagation and quantification.

Paths to analysing heterogeneous systems

- You can use a list of multiple components to store the data (phyloseq in Bioconductor).
- You can use Graph or Trees to ”influence” these distances (Structured high-dimensionality).
- Mixtures are everywhere (not one parametric population).
- Latent variables or factors are an enormous resource.
- Don’t stress about choices, they are not forever (because of reproducible workflows).
- Think carefully when you throw out information of any sort.
- Be lazy: re-use and recycle methods, vocabulary and infrastructure.

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 (Sanger, 454, Illumina, MassSpec, minion, RNA-seq, Chip-seq, imaging, CyTOF, single-cell).

Heterogeneity of methods and disciplines

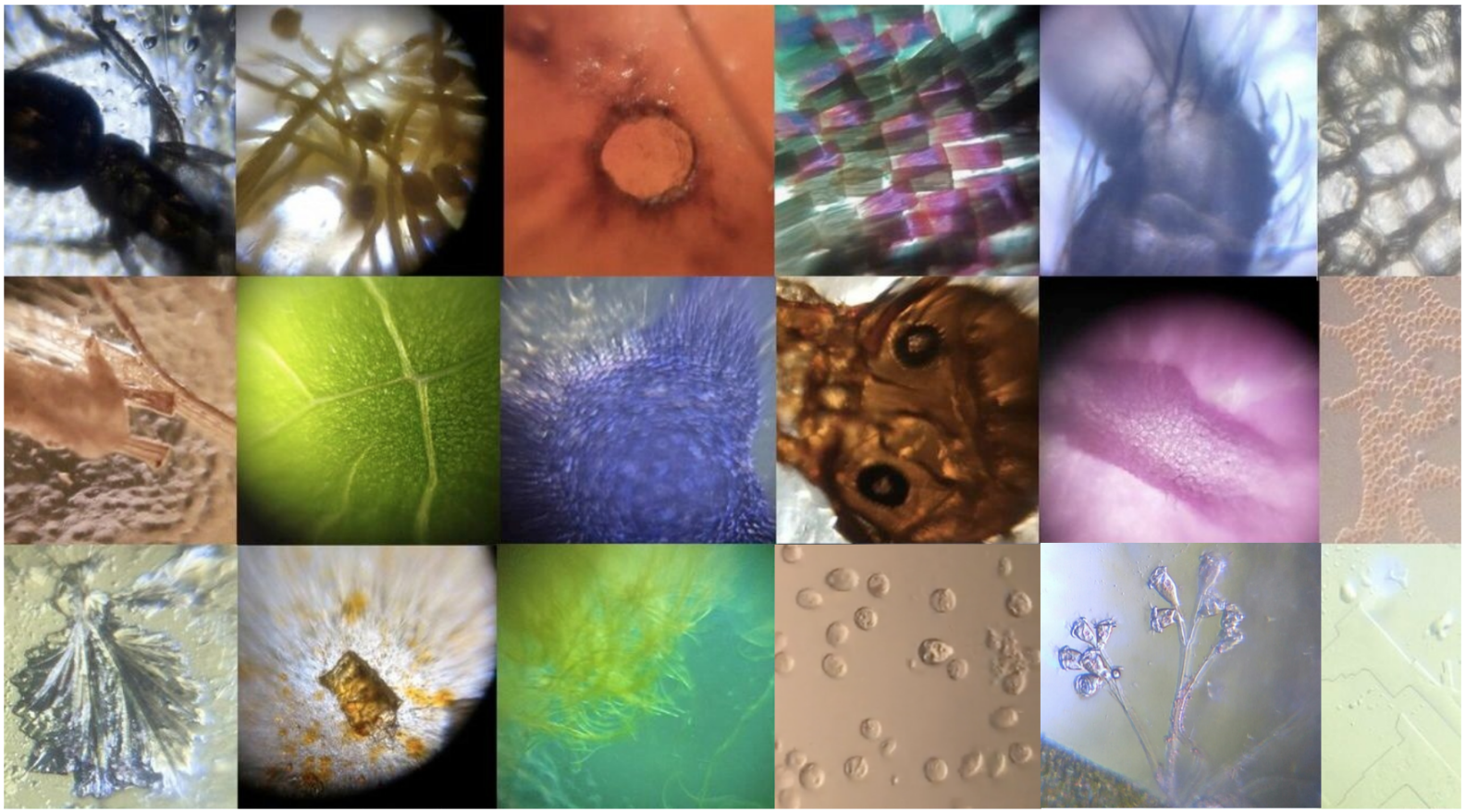
- Multiway, Multimodal, Multidomain, Multitable, Multiview, Triadic, Tensor...
- Data fusion, data integration, conjoint analyses.
- Approaches:
  - Matrix factorization, dimension reduction.
  - Latent variable estimation and Bayesian hierarchical models.
  - Algorithmic graph based methods (Nearest neighbor, Geometric Graphs).
  - Embedding methods often combine a graph and spectral decompositions.

Sankaran, Kris and Holmes, Susan P (2019) Multitable Methods for Microbiome Data Integration, Frontiers in genetics, 10, 2019.

Multiviews

S124E008613

Multiviews



Images generated using an origami- Foldscope microscope

An example: the Microbiome data

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

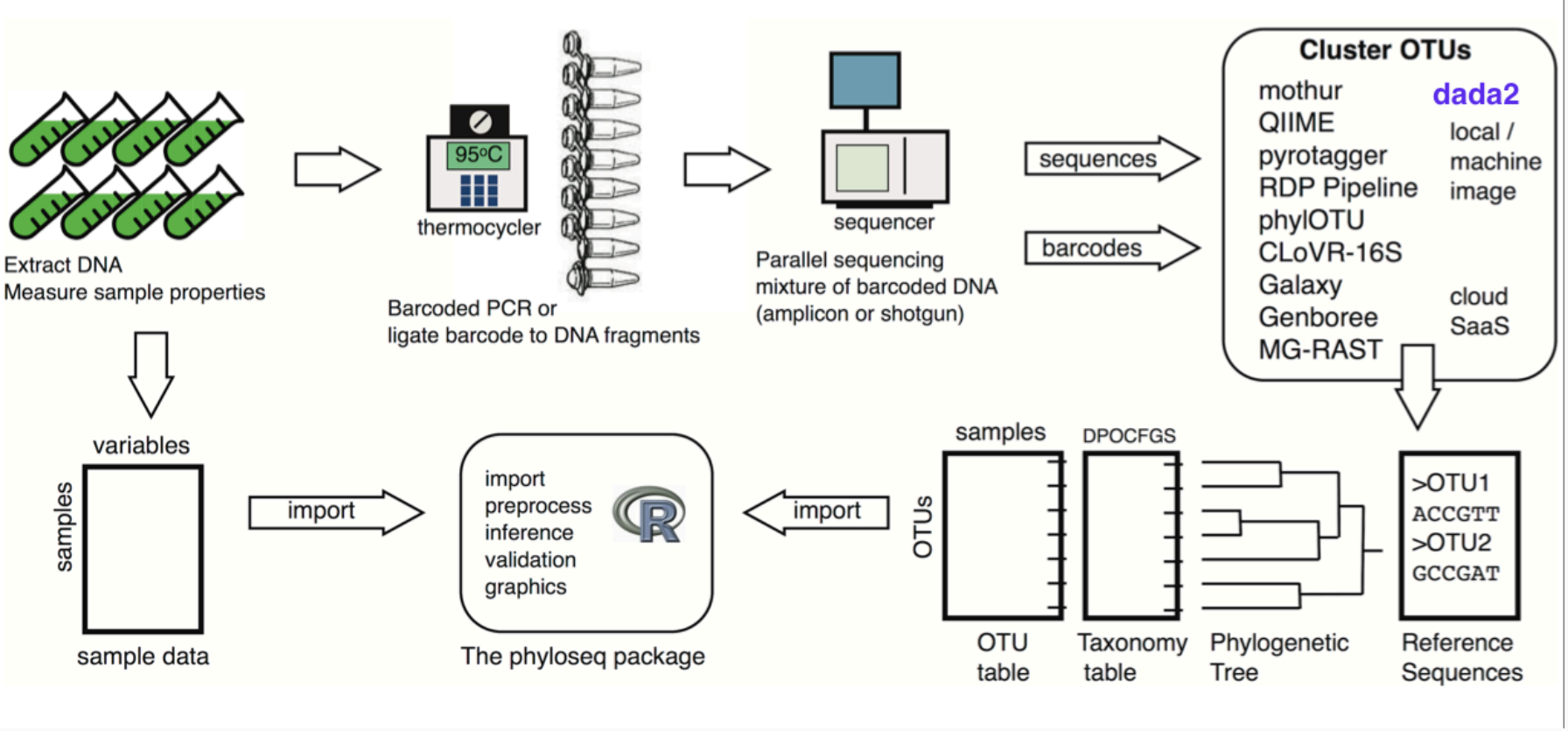
RNA What genes are being turned on (gene expression), transcriptomics.

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

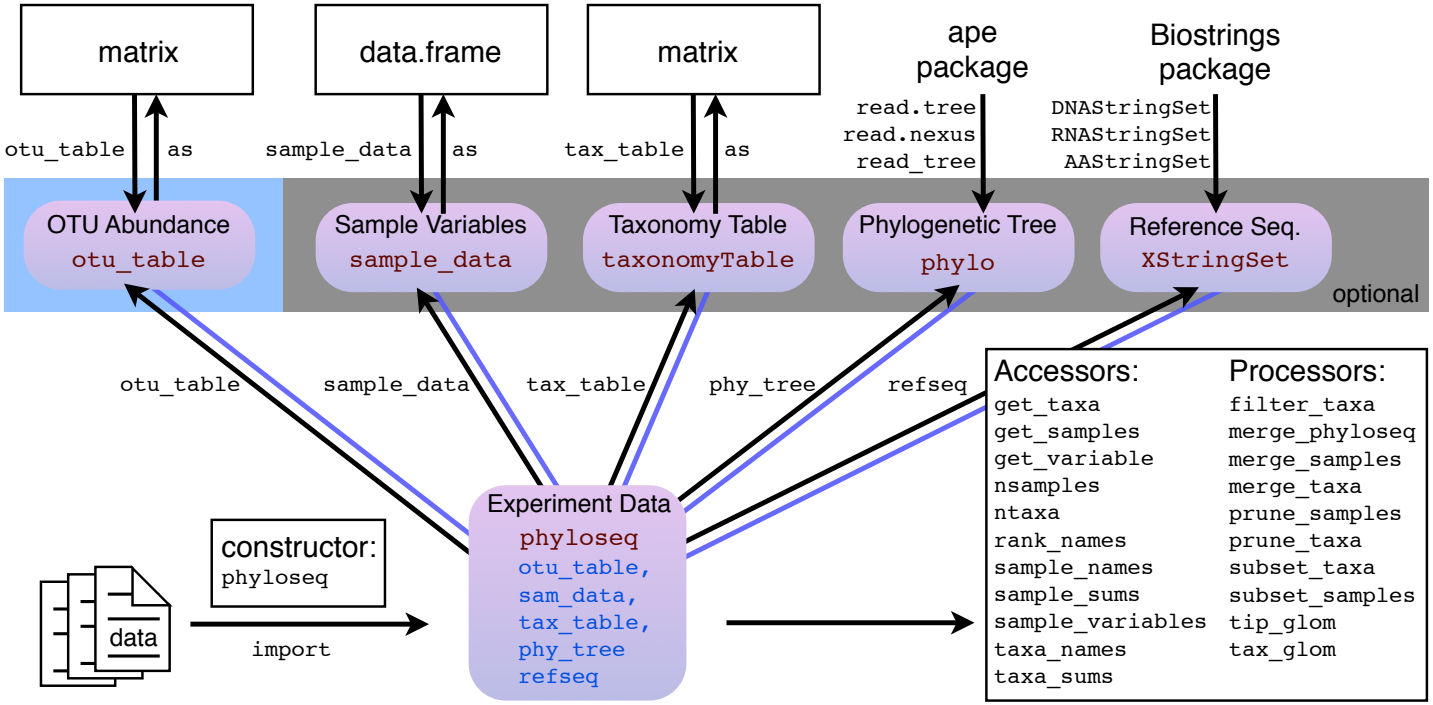
Input and data manipulation with `phyloseq`

(McMurdie and Holmes, 2013, Plos ONE)

As always in R: object oriented data.

phyloseq

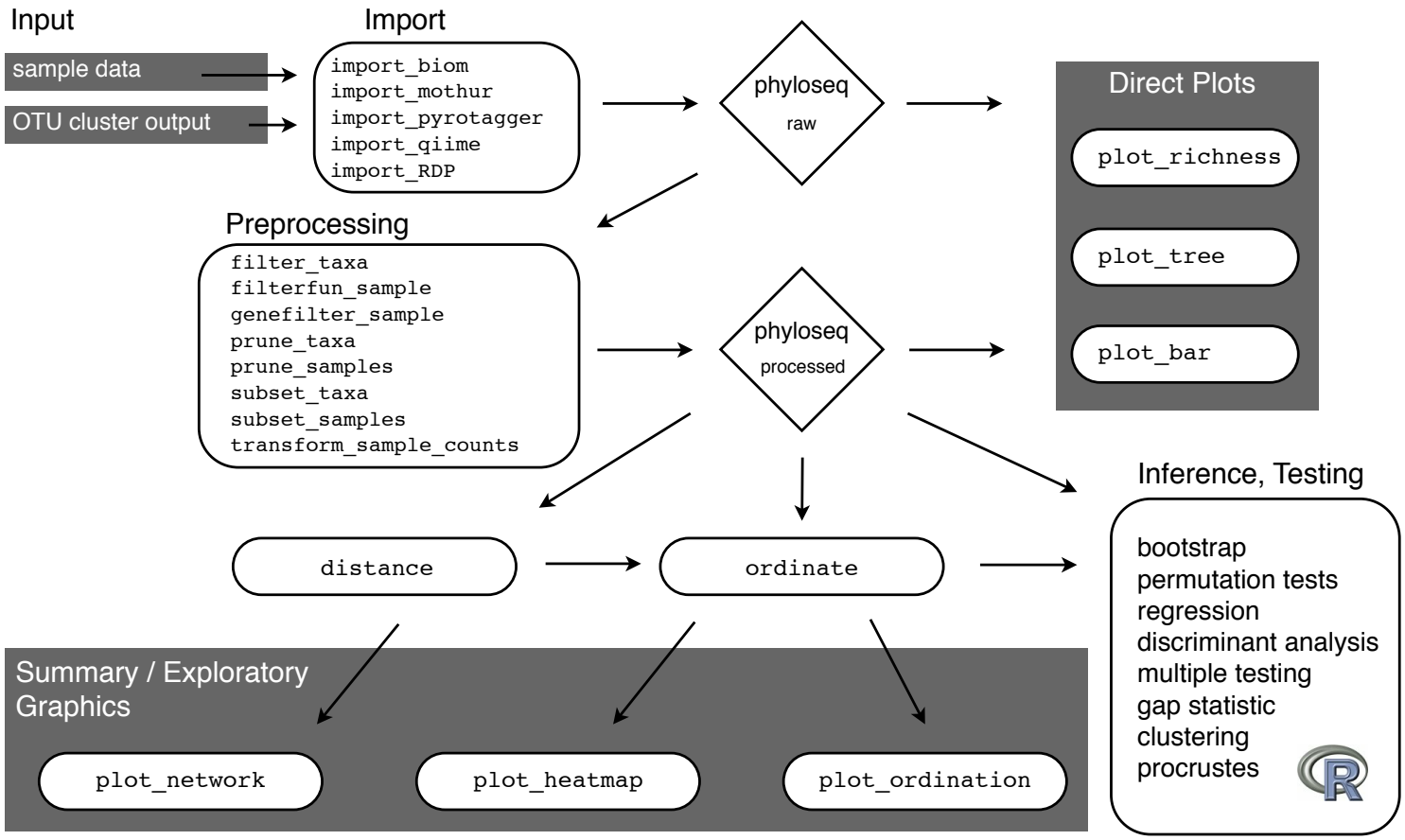
data structure & API



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

phyloseq

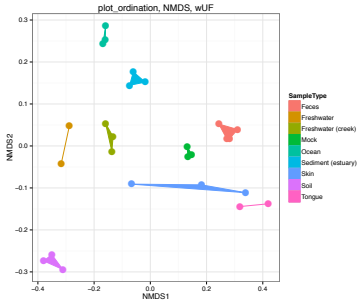
work flow

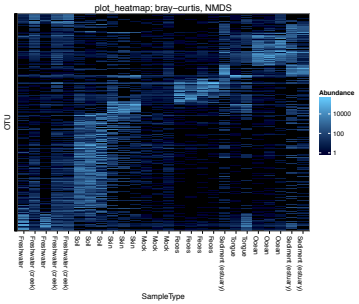


phyloseq

graphics

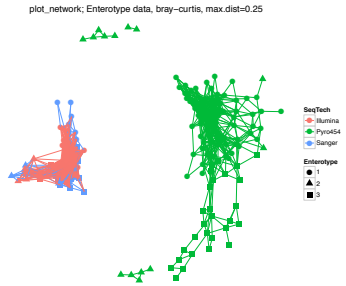
plot_ordination()

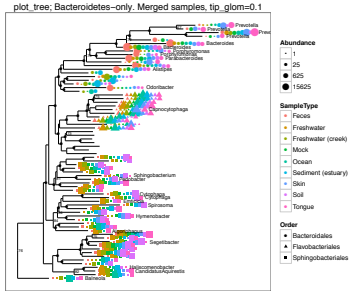




plot_heatmap()

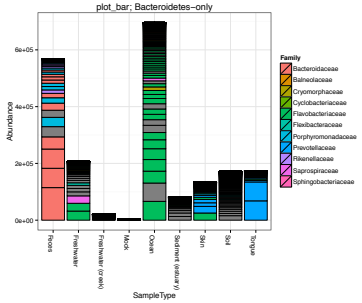
plot_network()

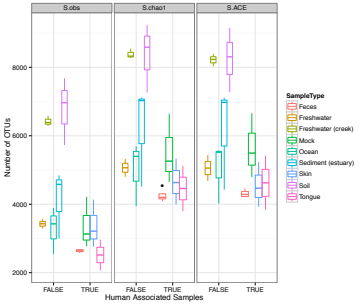




plot_tree()

plot_bar()



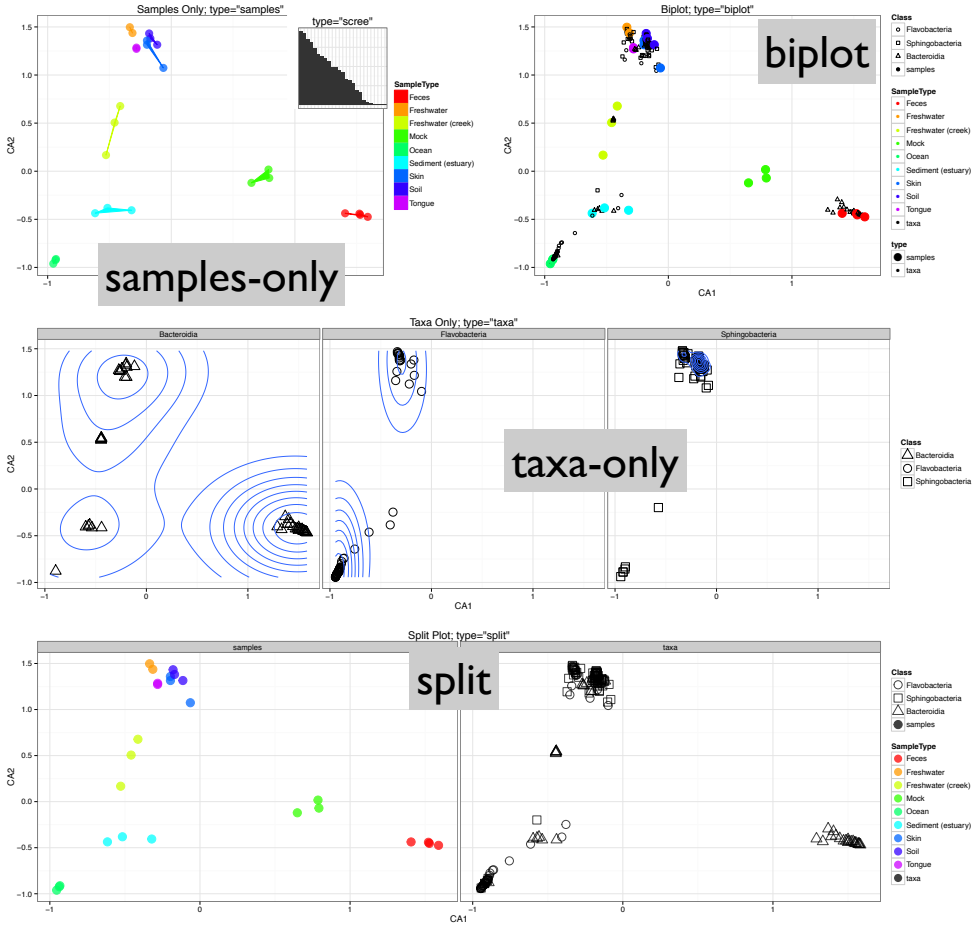


plot_richness()

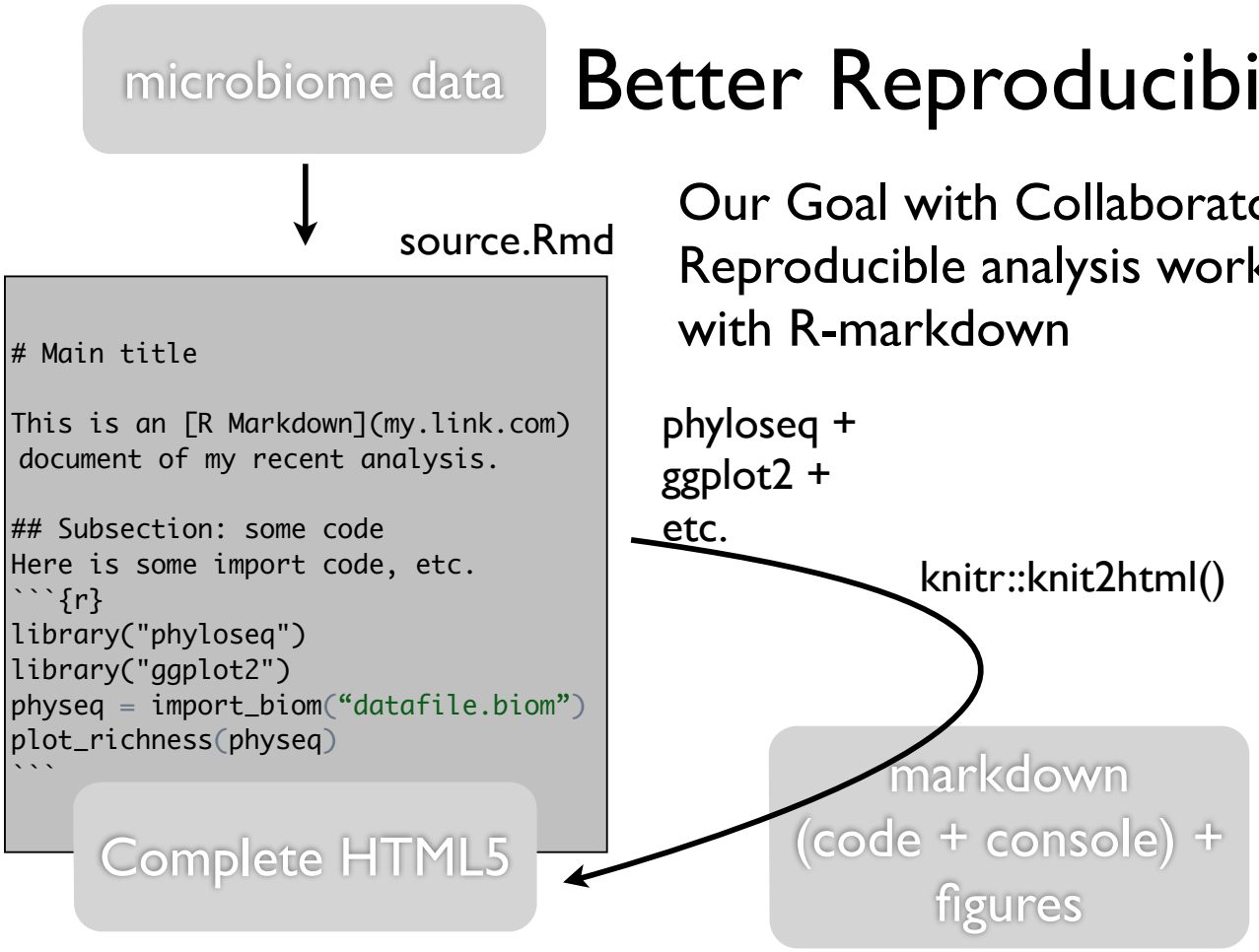
phyloseq

graphics

plot_ordination()



Better Reproducibility



Our Goal with Collaborators:
Reproducible analysis workflow
with R-markdown

Part I

An Example

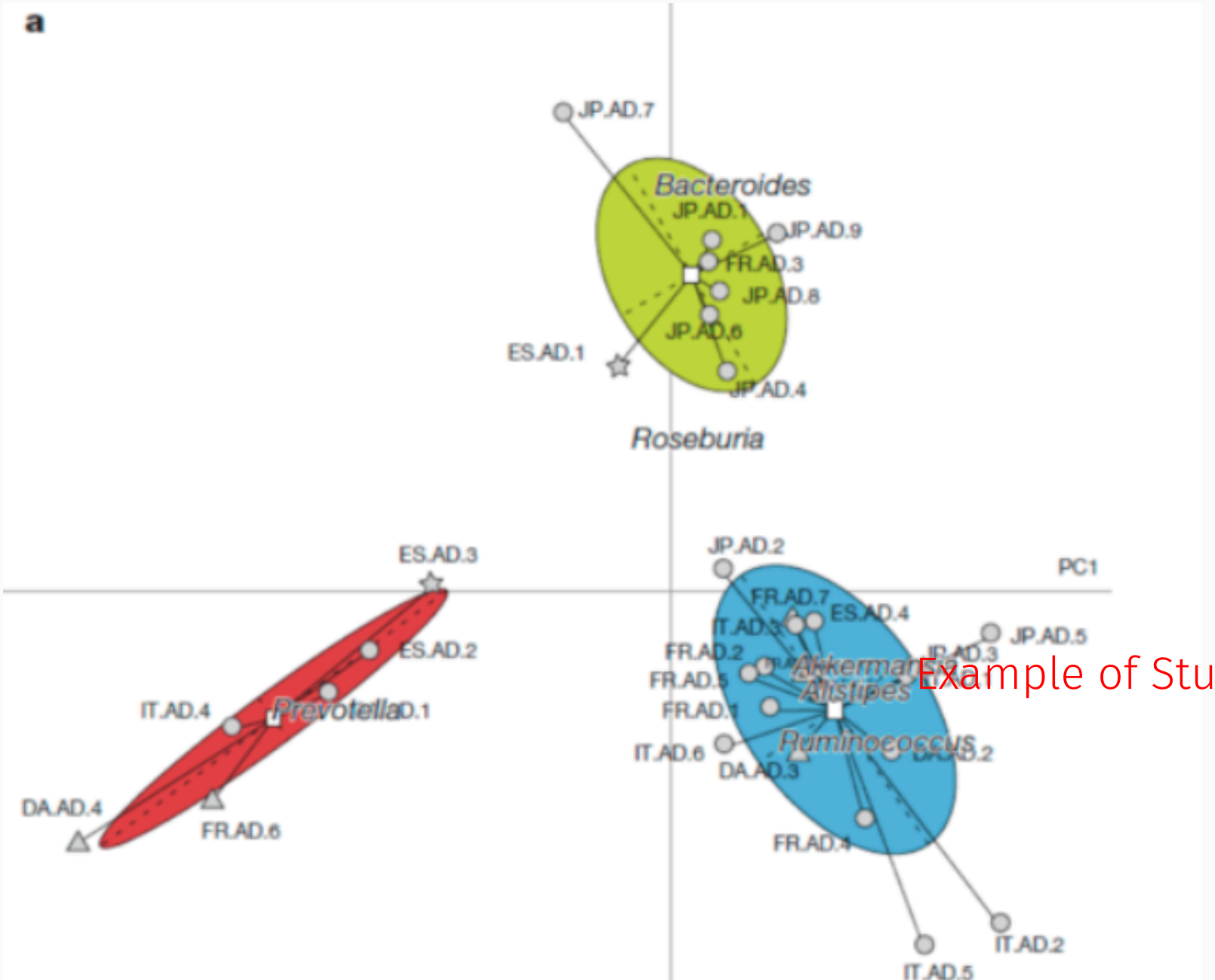
ARTICLE

doi:10.1038/nature09944

Enterotypes of the human gut microbiome

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

Our knowledge of species and functional composition of the human gut microbiome is rapidly increasing, but it is still based on very few cohorts and little is known about variation across the world. By combining 22 newly sequenced faecal metagenomes of individuals from four countries with previously published data sets, here we identify three robust clusters (referred to as enterotypes hereafter) that are not nation or continent specific. We also confirmed the enterotypes in two published, larger cohorts, indicating that intestinal microbiota variation is generally stratified, not continuous. This indicates further the existence of a limited number of well-balanced host–microbial symbiotic states that might respond differently to diet and drug intake. The enterotypes are mostly driven by species composition, but



Example of Study

Summary of the study

- Choose the data transformation (here proportions replaced the original counts).

  ... log, rlog, subsample, prop, orig.

- Take a subset of the data, some samples declared as outliers. ... leave out 0, 1, 2 ,..,9, + criteria (10).....

- Filter out certain taxa (unknown labels, rare, etc...)

  ... remove rare taxa (threshold at 0.01%, 1%, 2%,...)

- Choose a distance.

  ... 40 choices in vegan/phyloseq.

- Choose an ordination method and number of coordinates.

  ... MDS, NMDS, k=2,3,4,5..

- Choose a clustering method, choose a number of clusters.

  ... PAM, KNN, density based, hclust ...

- Choose an underlying continuous variable (gradient or group of variables: manifold).

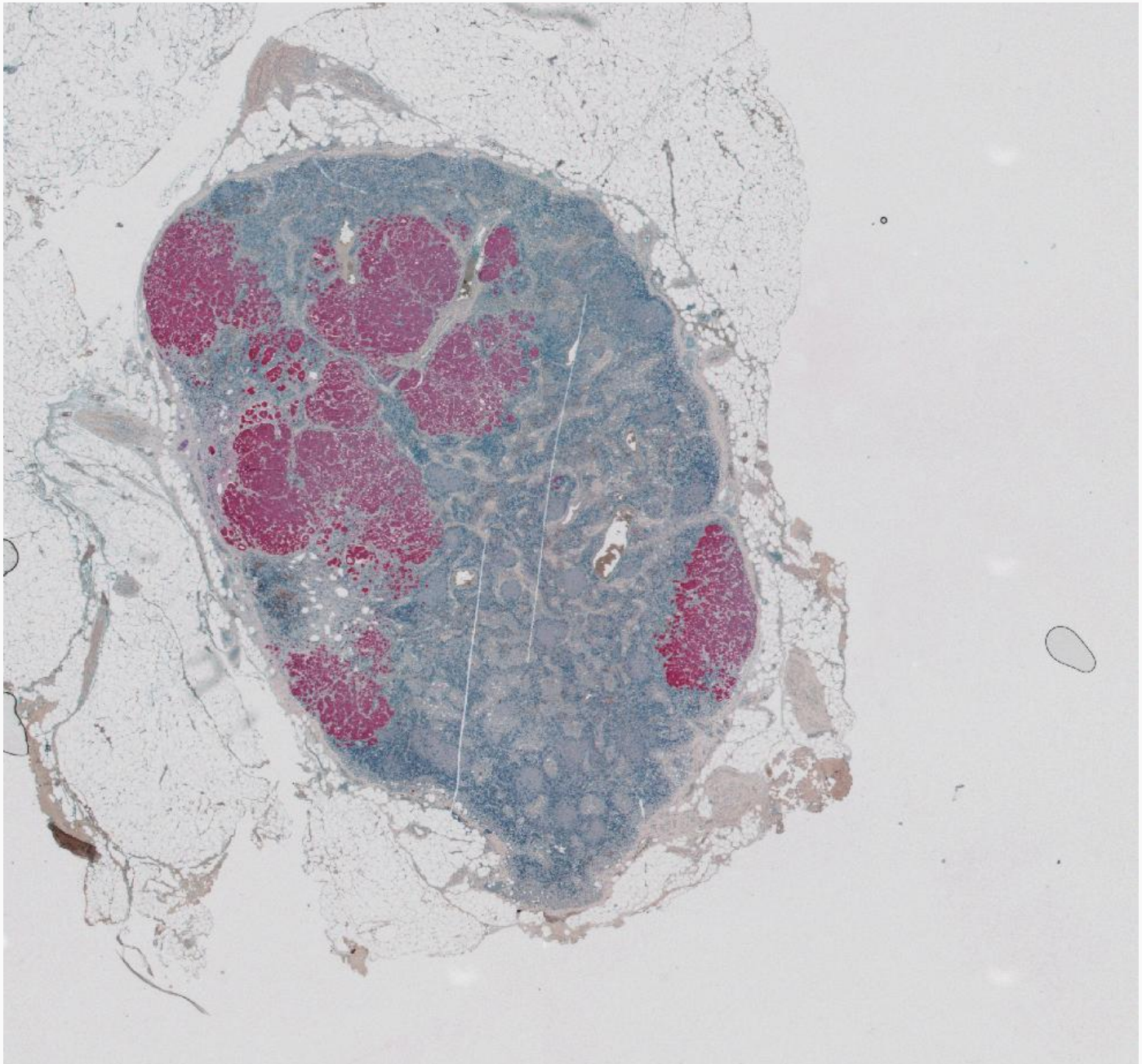
There are thus more than 200 million possible ways of analyzing this data:

$$5 \times 100 \times 10 \times 40 \times 8 \times 16 \times 2 \times 4 = 204800000$$

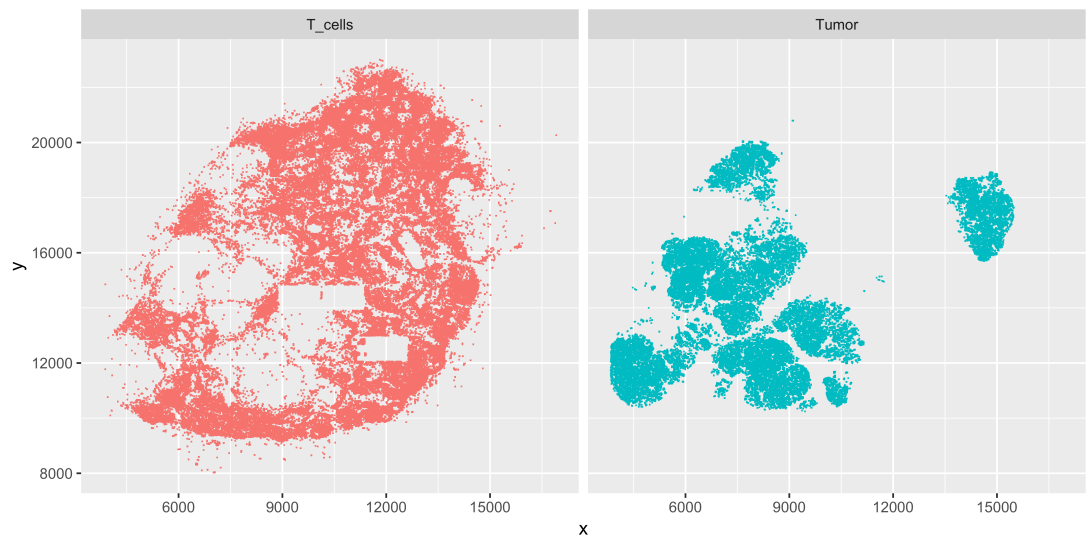
Transfer Learning for Humans: Extend existing methods

- Spatial statistics.
- Probabilistic denoising.
- Extend covariances from variables to tables.
- Nonparametric Network testing.
- Latent variables, latent clusters, latent networks, latent manifolds.
- Hierarchical, iterative methods.

Stained Lymph Node



Output Data



Setiadi AF, Ray NC, Kohrt HE, Kapelner A, Carcamo-Cavazos V, Levic EB, Yadegarynia S, Van Der Loos CM, Schwartz EJ, Holmes S, Lee PP. Quantitative, architectural analysis of immune cell subsets in tumor-draining lymph nodes from breast cancer patients and healthy lymph nodes. PLoS One. 2010;5(8).

Data Analysis

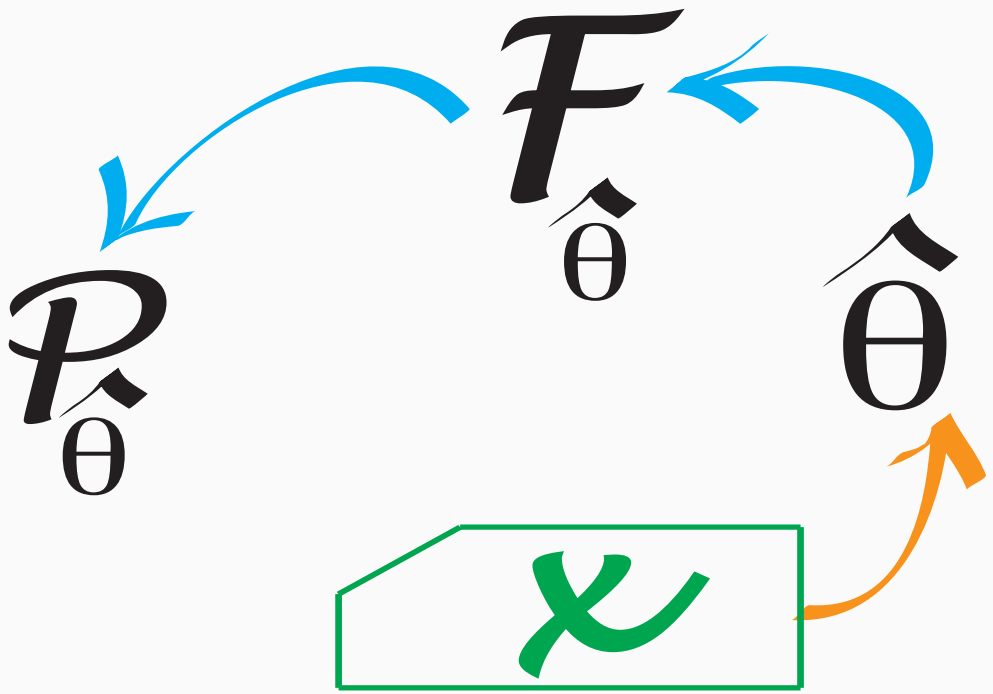
- Data are output and read into R.
  - Orders of magnitude: 100,000- 1.5 million cells per lymph node.
  - As many as 100,000 Tumor cells.
  - Only a few hundred dendritic cells.
- Spatial Analysis done with `spatstat, spdep, DCluster.`.

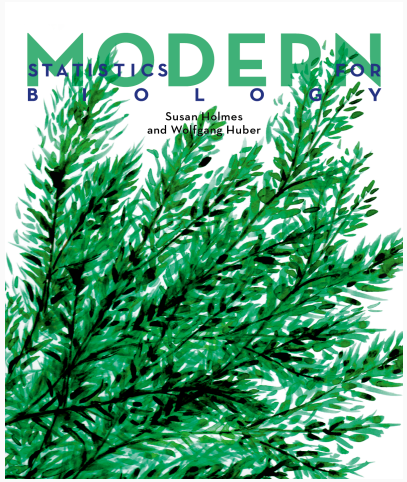
Reading Data into R

```
for ( i in (1:16)){
DCname=paste("/U/cells/P1out/",tumorsp[i],"-DCs.txt",sep
DCs=read.delim2(DCname, sep=",", header=TRUE)
list.slides[[i]]=ppp(list.all[[i]][,1],list.maxs[[i]][2]
list.all[[i]][,2],window=owin(c(0,list.maxs[[i]][1]),
c(0,list.maxs[[i]][2])),marks=as.factor(c(rep("Other",
list.counts[[i]][1]),rep("Tumor",list.counts[[i]]
[2]),rep("Tcells",list.counts[[i]][3]))))
}

list.slides[haveall[2]]
marked planar point pattern: 107347 points
multitype, with levels = DCs^^IOther^^ITcells^^ITumor
window: rectangle = [0, 11492] x [0, 12825] units
```

Using statistics:

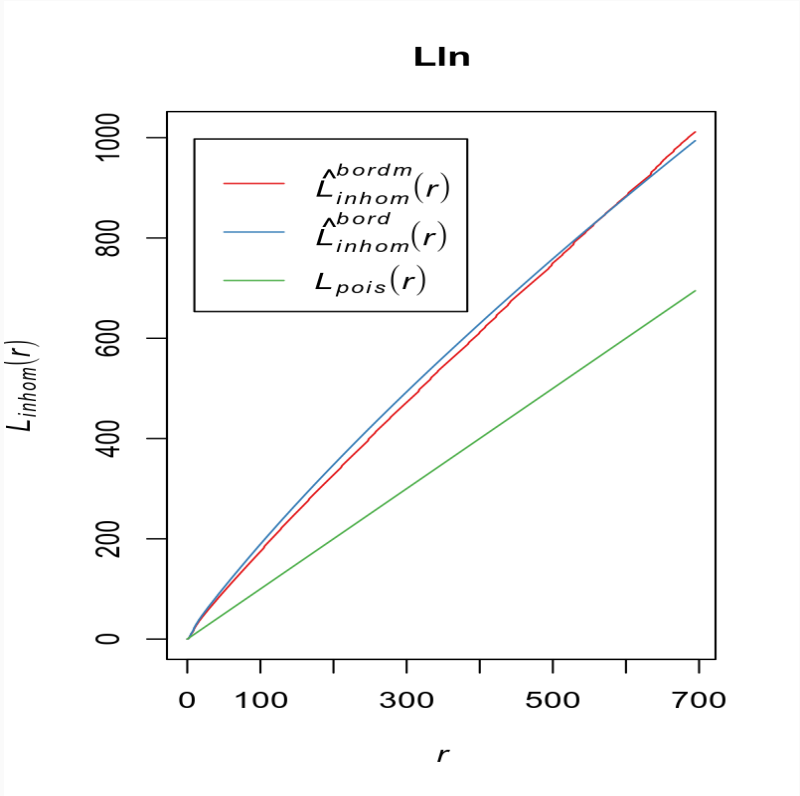




See book:
http://bios221.stanford.edu/book/

Standard spatial statistical methods

Transformation of Ripley's K statistic:



See full explanation in this chapter: `https://www.huber.embl.de/msmb/Chap-Images.html`

Part II

Features generated from raw data

(quality and frequencies)

DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan[1], Paul J McMurdie[2], Michael J Rosen[3], Andrew W Han[2], Amy Jo A Johnson[2] & Susan P Holmes[1]

We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjjneb/dada2). DADA2 infers sample

We previously introduced the Divisive Amplicon Denoising Algorithm (DADA), a model-based approach for correcting amplicon errors without constructing OTUs[5]. DADA identified fine-scale variation in 454-sequenced amplicon data while outputting few false positives[2,5].

Here we present DADA2, an open-source R package (https://github.com/benjjneb/dada2, **Supplementary Software**) that extends and improves the DADA algorithm. DADA2 implements a new quality-aware model of Illumina amplicon errors. Sample composition is inferred by dividing amplicon reads into partitions consistent with the error model (Online Methods). DADA2 is reference free and applicable to any genetic locus. The DADA2 R package implements the full amplicon workflow: filtering, dereplication, sample inference, chimera identification, and merging of paired-end reads.

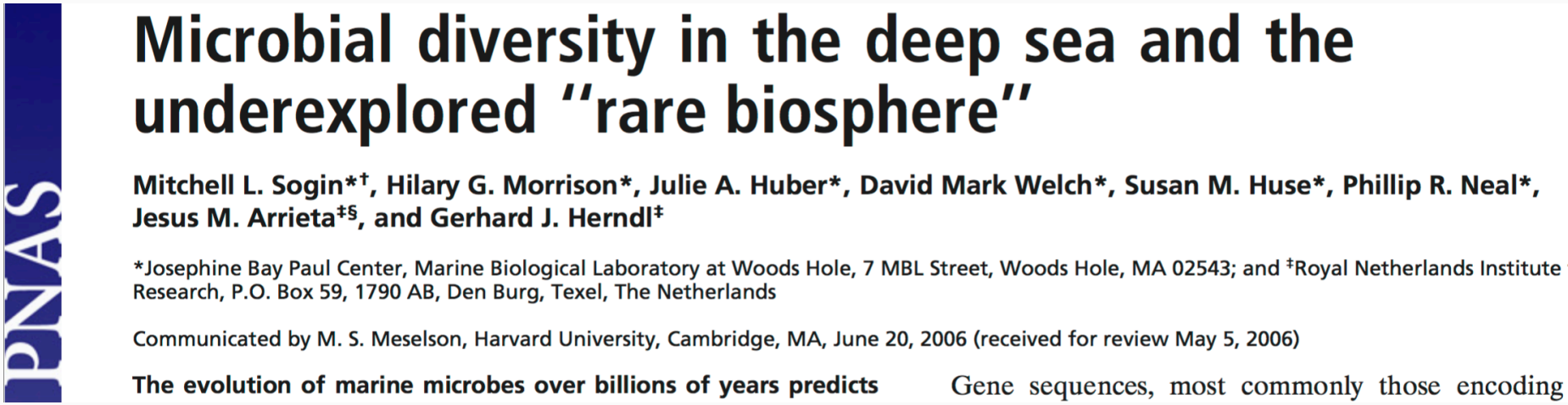
We compared DADA2 to four algorithms (Online Methods):

Diversities in the microbiome depend on the number of taxa

- α-diversity: Number of 'species'-taxa in a biological sample (from one location).
- β-diversity: Differentiation in diversity among different samples from different locations.

Extremely sensitive to noise.

Fake species:



PNAS

Microbial diversity in the deep sea and the underexplored "rare biosphere"

Mitchell L. Sogin*†, Hilary G. Morrison*, Julie A. Huber*, David Mark Welch*, Susan M. Huse*, Phillip R. Neal*, Jesus M. Arrieta‡§, and Gerhard J. Herndl‡

*Josephine Bay Paul Center, Marine Biological Laboratory at Woods Hole, 7 MBL Street, Woods Hole, MA 02543; and ‡Royal Netherlands Institute Research, P.O. Box 59, 1790 AB, Den Burg, Texel, The Netherlands

Communicated by M. S. Meselson, Harvard University, Cambridge, MA, June 20, 2006 (received for review May 5, 2006)

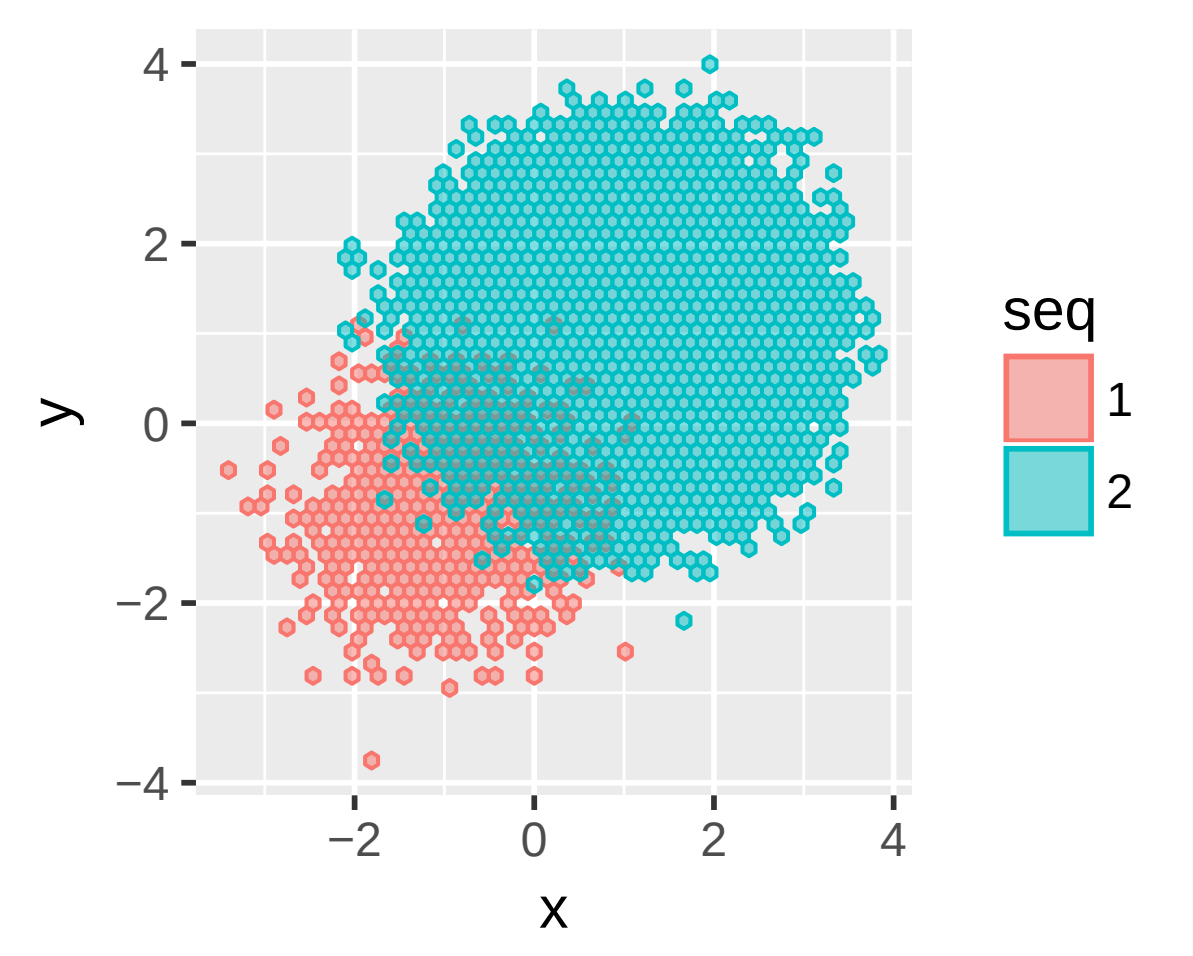
The evolution of marine microbes over billions of years predicts Gene sequences, most commonly those encoding

How many words does Professor D. know?

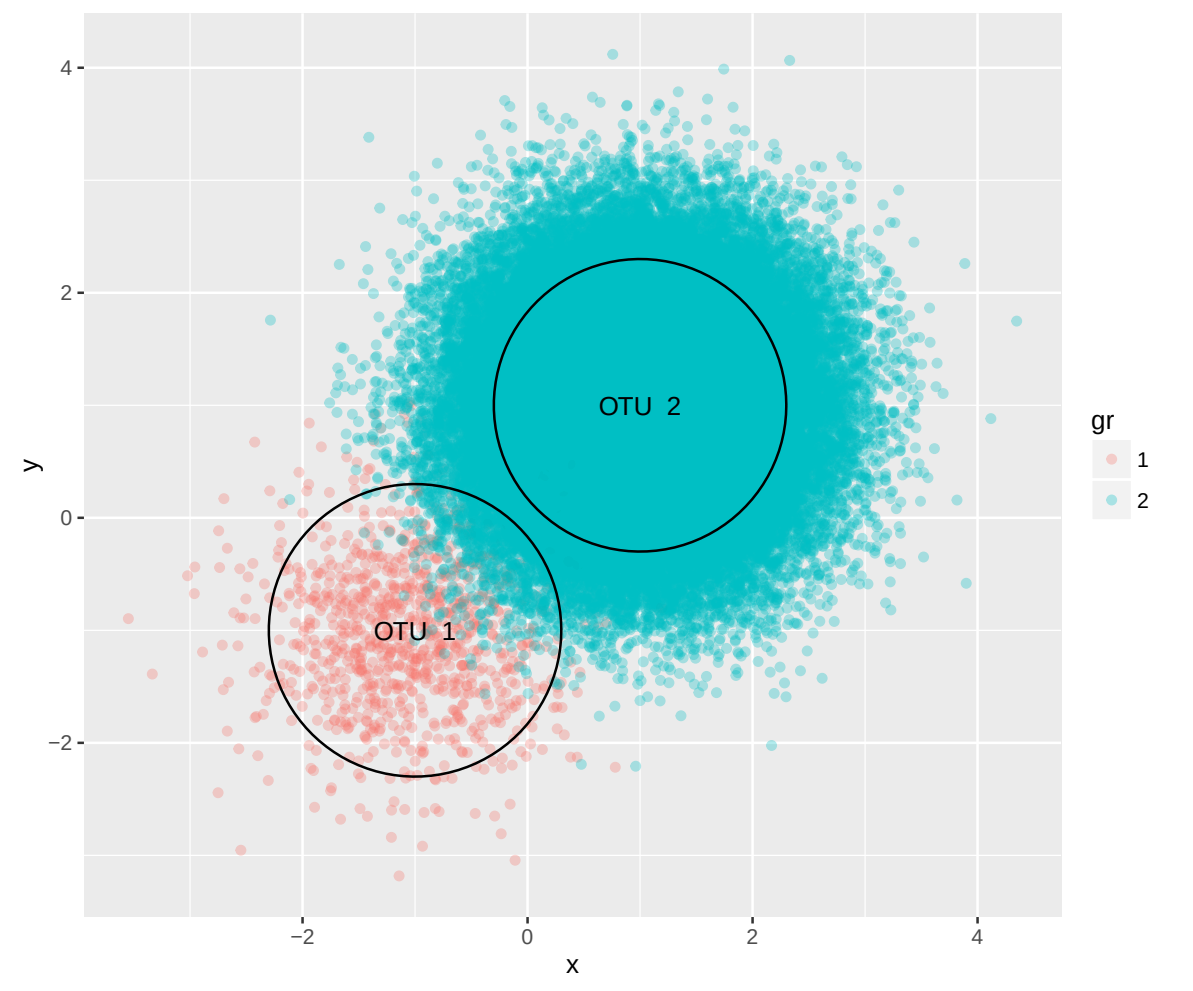
- Maybe 15,000, 20,000?
- Start sampling..... banana, bannana, bannanna, orange, orenge, muscle, musel, muscel, foreign, forene, forane,.....
- How many real words does Prof D. know?
- Use more information than the spelling....

The success of `dada2` is in it's use of the **frequencies**, often forgotten or hidden from the user if you only inventory the different sequences.

From reads to Operational Taxonomic Units

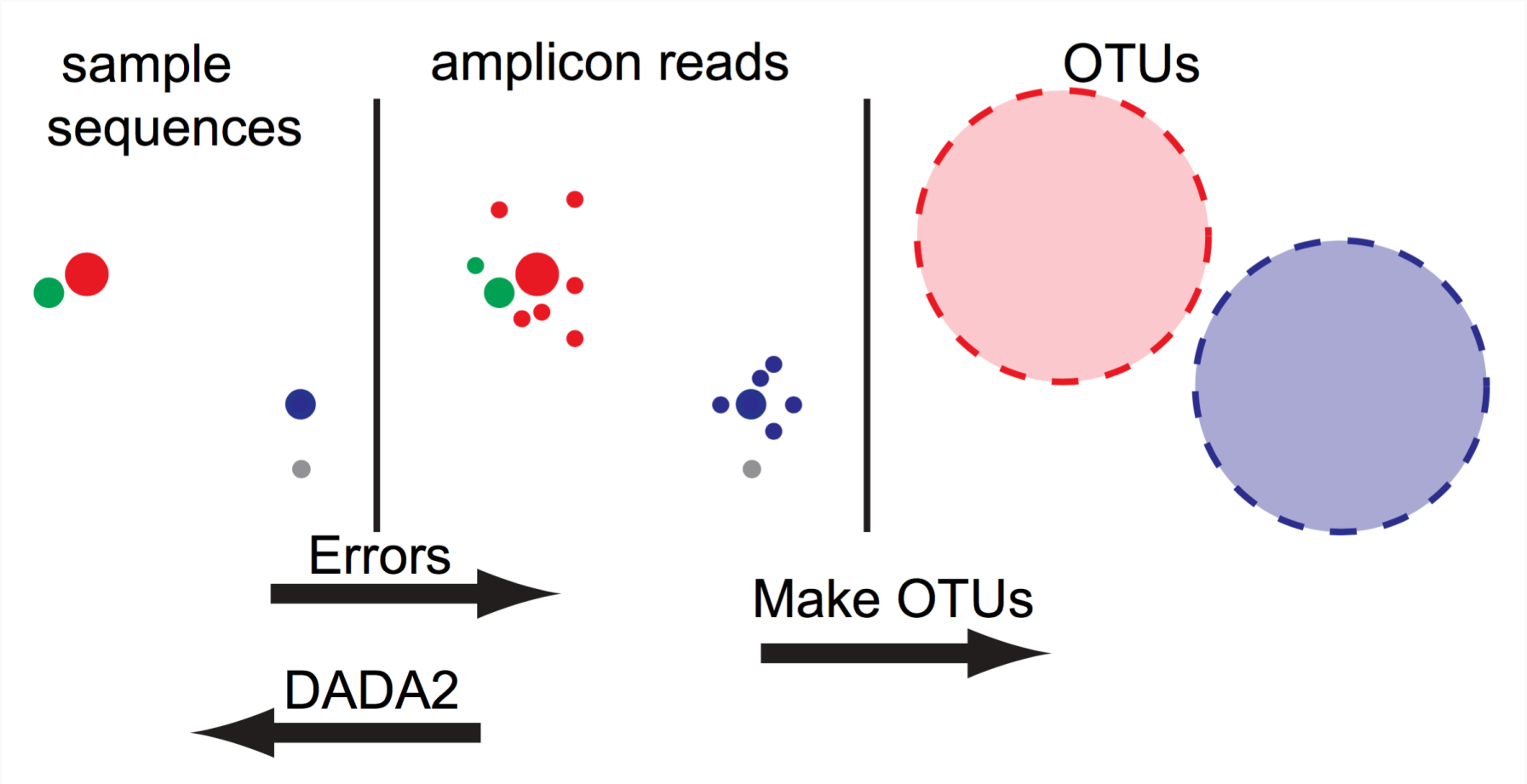


From reads to Operational Taxonomic Units



Curent practice (`qiime, mothur, rdp,...`): 97% similarity.

Probabilistic Model accounting for frequencies and distances



Error Model

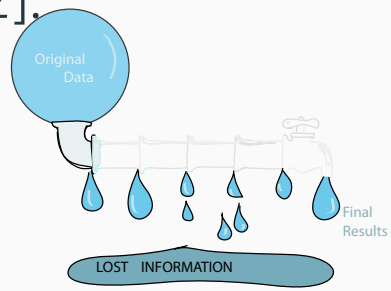
```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
      |                    |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

$$P(r|s) = \prod_{i=1}^{L} P(r(i)|s(i), q_r(i), Z)$$

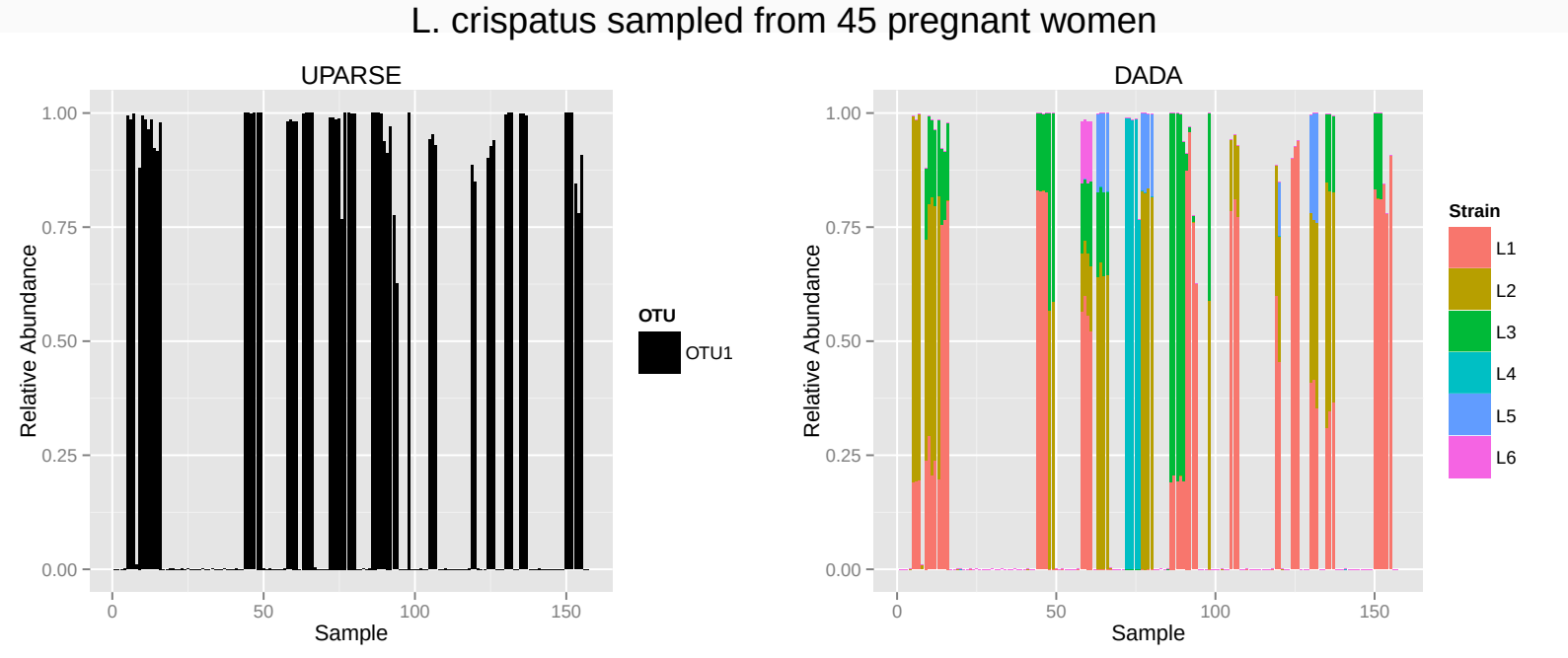
P probabilities of substitutions ($A->C$)

q Quality score (Q=30) Batch effect (run)

Use the denoised sequence instead of the OTU[2].



Higher resolution strain clustering: DADA2



R package: `http://benjjneb.github.io/dada2/R/tutorial.html`

Part III

Multitable approaches

Multi-table methods

Inertia, Co-Inertia

We generalize it in several directions through the idea of inertia.

As in physics, we define inertia as a weighted sum of distances of weighted points.

This enables us to use abundance data in a contingency table and compute its inertia which in this case will be the weighted sum of the squares of distances between observed and expected frequencies, such as is used in computing the chisquare statistic.

Another generalization of variance-inertia is the useful Phylogenetic diversity index. (computing the sum of distances between a subset of taxa through the tree).

Co-Inertia

When studying two variables measured at the same locations, for instance PH and humidity the standard quantification of covariation is the covariance.

$$sum(x1 * y1 + x2 * y2 + x3 * y3)$$

if x and y co-vary -in the same direction this will be big.

A simple generalization to this when the variability is more complicated to measure as above is done through Co-Inertia analysis (CIA):
a multivariate method that identifies trends or co-relationships in multiple datasets which contain the same samples or the same time points. That is the rows or columns of the matrix have to be weighted similarly and thus must be matchable.

RV coefficient

The global measure of similarity of two data tables as opposed to two vectors can be done by a generalization of covariance provided by an inner product between tables that gives the RV coefficient, a number between 0 and 1, like a correlation coefficient, but for tables.

$$RV(A, B) = \frac{\mathrm{Tr}(A'B)}{\sqrt{\mathrm{Tr}(A'A)}\sqrt{\mathrm{Tr}(B'B)}}$$

Example

Taxa Read counts (3 patients taking cipro: two time courses) : .

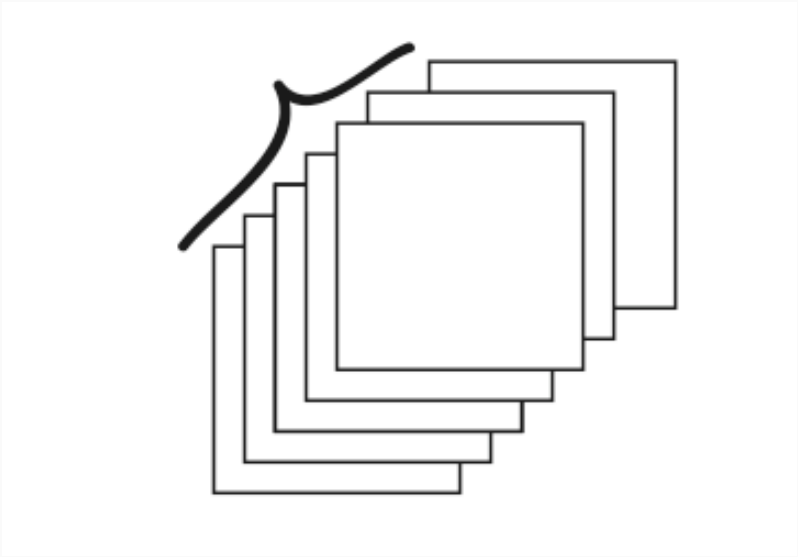
Mass-Spec Positive and Negative ion Mass Spec features and their intensities: .

RNA-seq Metagenomic data on genes :.

Here is the RV table of the three array types:

```
> fourtable$RV
           Taxa      Kegg   MassSpec+ MassSpec-
Taxa       1        0.565   0.561      0.670
Kegg       0.565     1      0.686      0.644
MassSpec+ 0.561    0.686   1          0.568
MassSpec- 0.670    0.644   0.568      1
```

Multiple table methods: STATIS and DiSTATIS



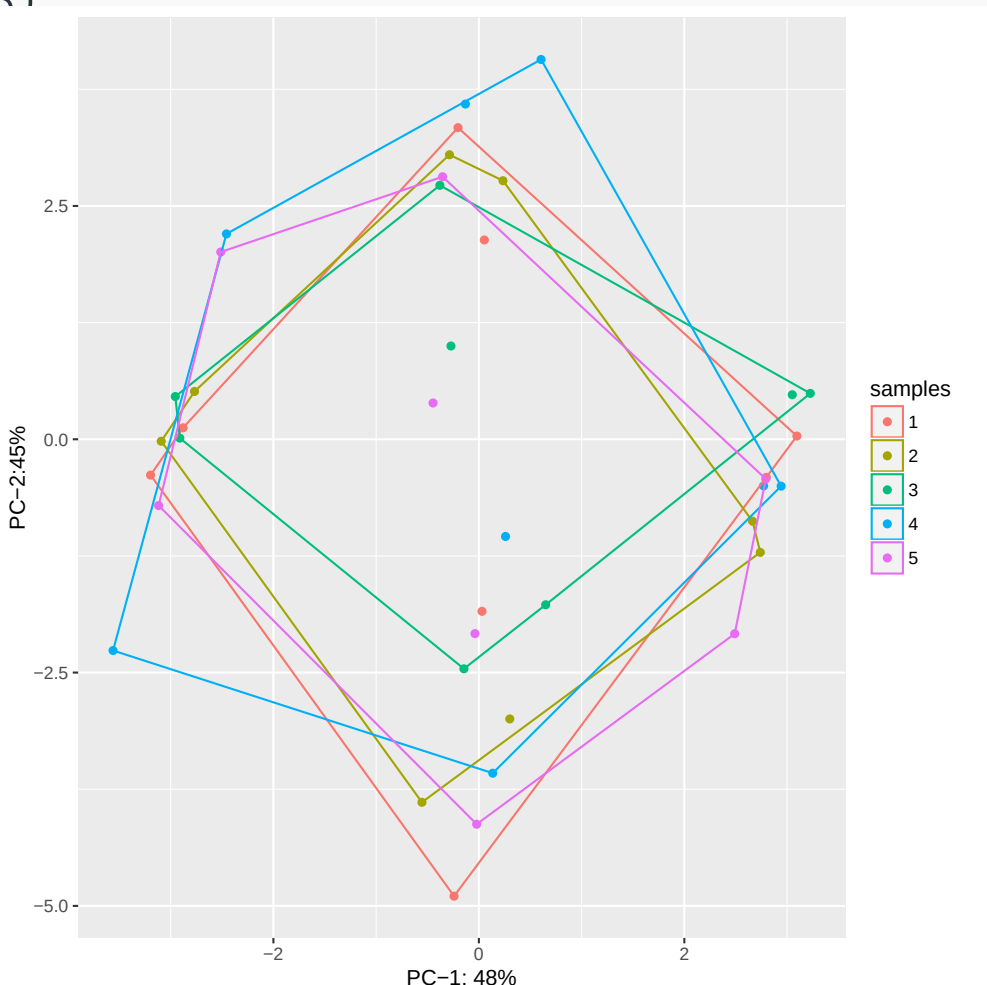
In PCA we compute the variance-covariance matrix, in multiple table methods we can take a cube of tables and compute the RV coefficient of their characterizing operators.

We then diagonalize this and find the best weighted ‘ensemble’ (PCA of PCA, Escoufier, 1975, L’Hermier des Plantes, 1976).

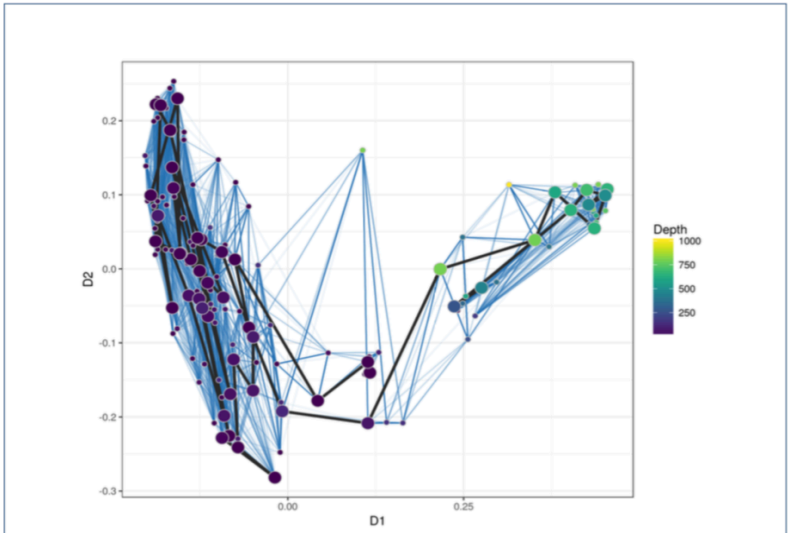
This is called the ‘compromise’ and all the individual tables can be projected onto it.

A compromise projection approach

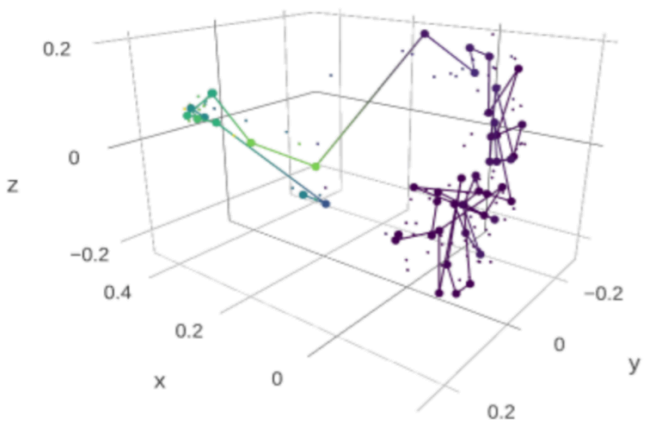
Naively overlaying projections of the principal coordinate loadings generated from different posterior samples of **S** on the same plot could show the variability of the projections (Holmes, 1985)



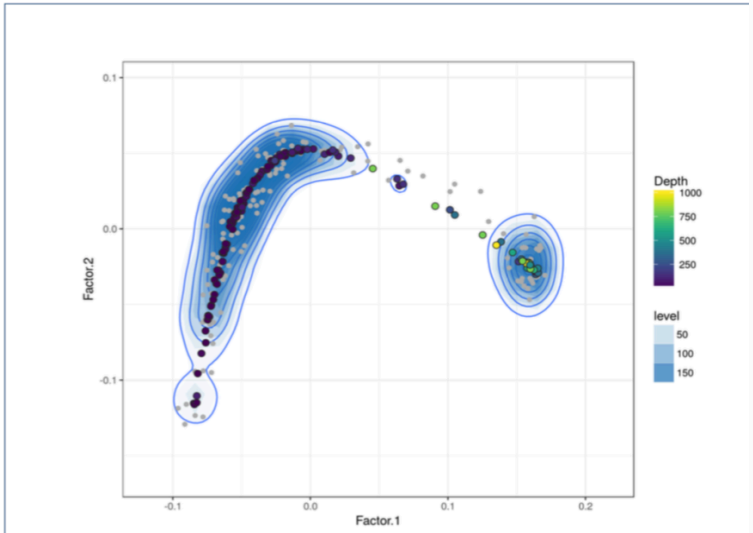
Bayesian Unidimensional Scaling (BUDS)



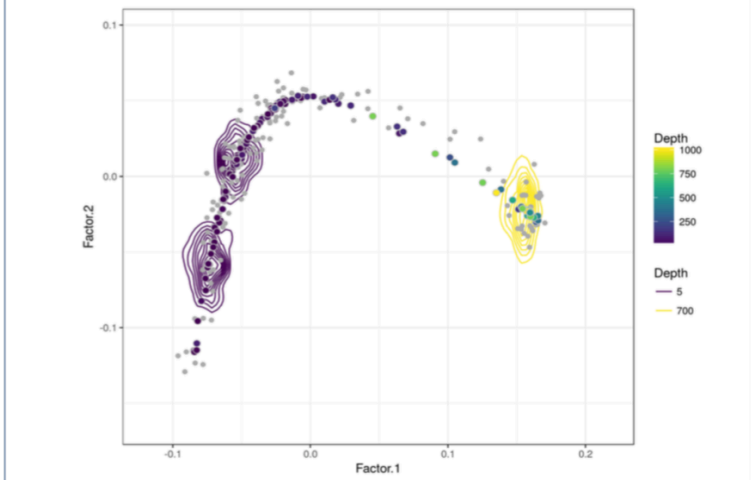
(a) Trajectory in 2D



(b) Trajectory in 3D



(a) Data density



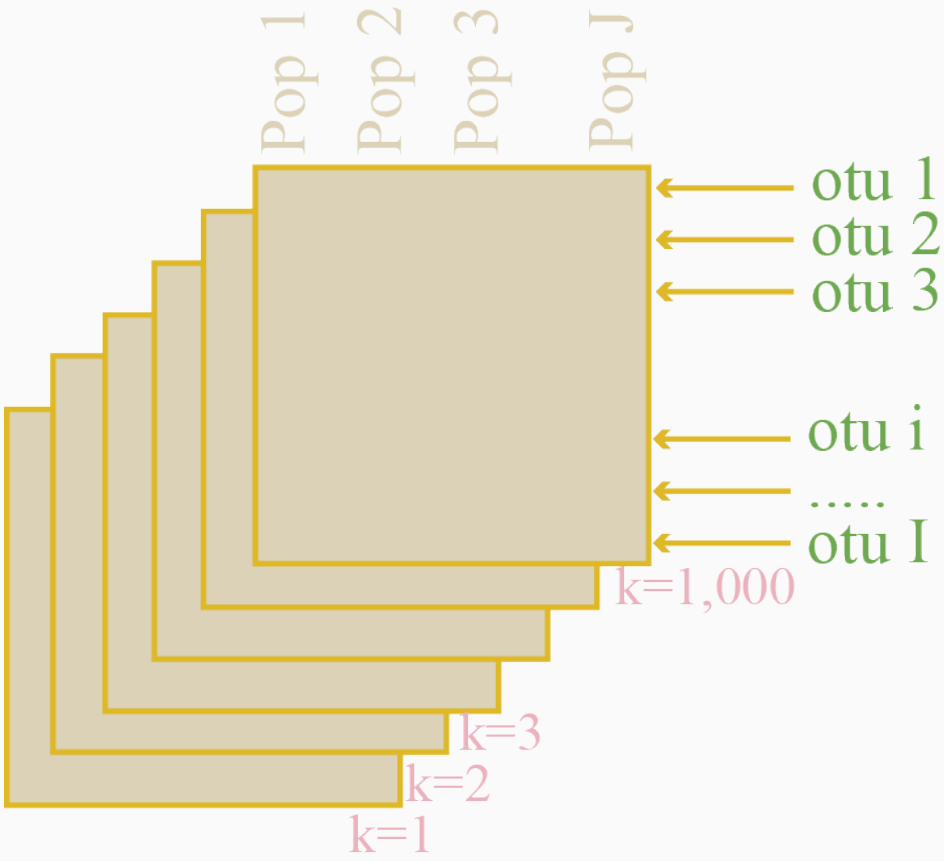
(b) Data point location confidence contours

Lan Huong Nguyen, SH, BMC Bioinformatics, 2017.

Part IV

Uncertainty quantification and propagation

Bayesian posterior uncertainty measures



Parameters for samples $\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$

Define a joint prior on these factors through the Gram matrix $(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$

The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of ASVs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j},$$

$$\epsilon_{i,j} \quad \text{iid Normal}$$

Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren, Bacallado, Favaro, Holmes, Trippa (2017, JASA).

Parameters for samples

$$\mathbf{Y}^j, j \in \mathcal{J} = \{1, \ldots, J\}$$

Define a joint prior on these factors through the Gram matrix

$$(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$$

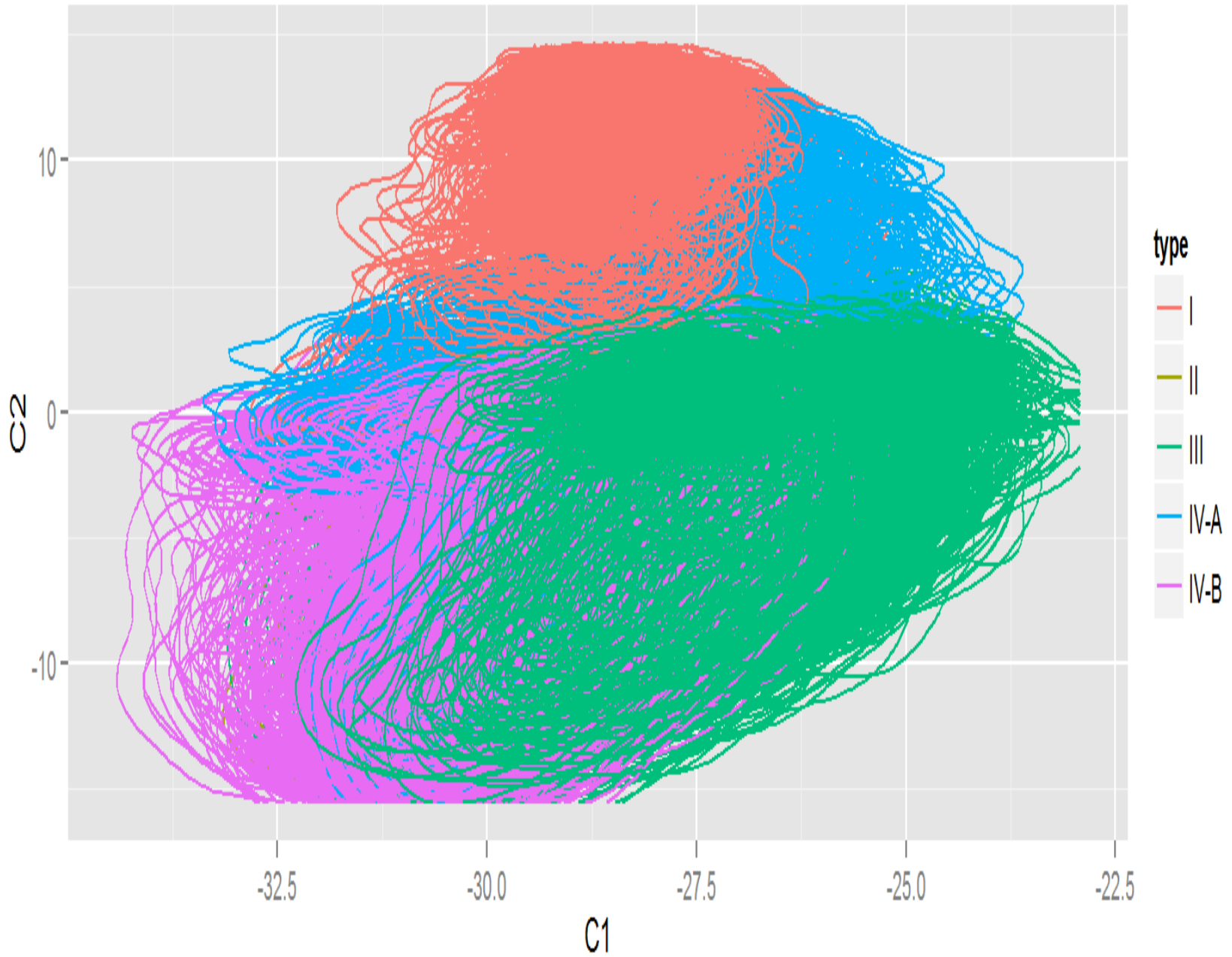
The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of OTUs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \tag{1}$$

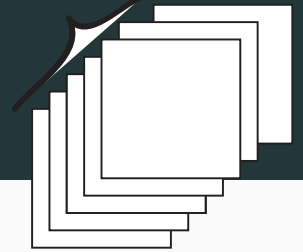
where the $\epsilon_{i,j}$ are independent Normal variables.

The methods that we consider here are all related to PCA and use the normalized Gram matrix **S** between biological samples.

S is the correlation matrix of $(Q_{i,1}, \ldots, Q_{i,J})$. Based on a single posterior instance of **S**, we can visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once.


10
0
-10
C2
-32.5
-30.0
-27.5
-25.0
-22.5
C1
type
I
II
III
IV-A
IV-B

Registration: Find $\mathbf{S}_0$



Identify a Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_K$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$.
We prefer to choose $\mathbf{S}_0$, the Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_K$.

We use the **RV** similarity metric between two symmetric square matrices $\mathbf{A}$ and $\mathbf{B}$

$$\mathrm{RV}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{AB})/\sqrt{\mathrm{Tr}(\mathbf{AA})\mathrm{Tr}(\mathbf{BB})}$$

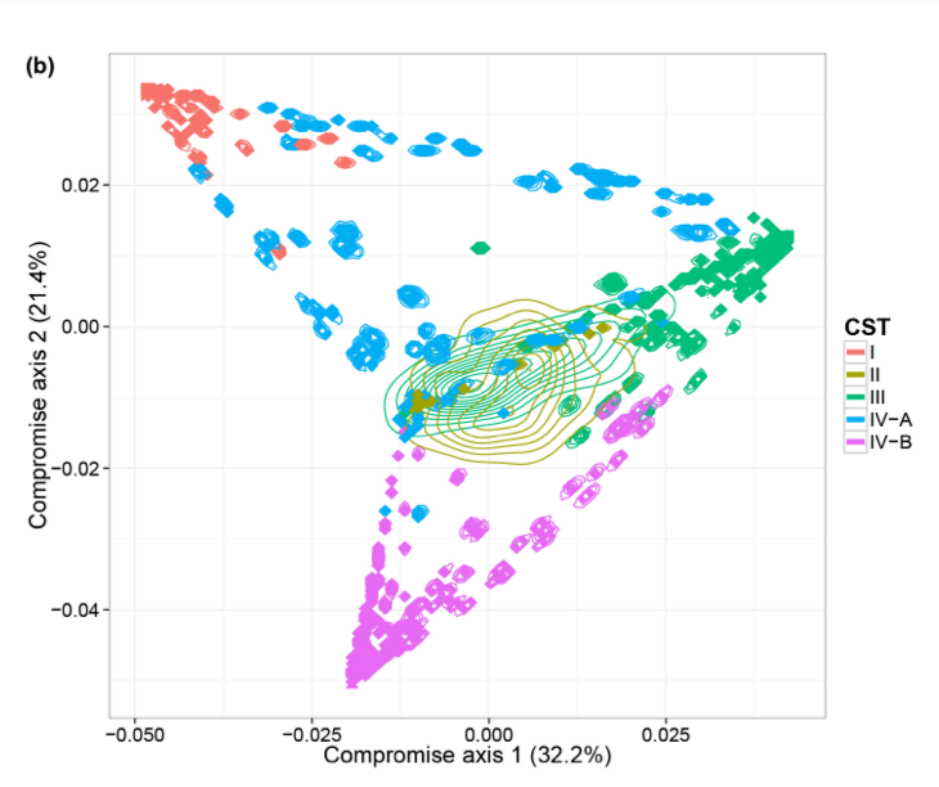
We diagonalize the **RV** matrix to obtain $\mathbf{S}_0$.

Find lower dimensional consensus space V

For dim 2, $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\mathbf{S}_0$ corresponding to the largest eigenvalues λ_1 and λ_2. All biological samples in V are visualized by projecting rows of $\mathbf{S}_0$ onto V:
$(\psi_1^0, \psi_2^0) = \mathbf{S}_0(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$.

Project the rows of posterior sample $\mathbf{S}_k$ onto V by $(\boldsymbol{\psi}_1^k, \boldsymbol{\psi}_2^k) = \mathbf{S}_k(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$. Overlaying all the $\boldsymbol{\psi}^k$ displays uncertainty of $\mathbf{S}$ in the same linear subspace. Posterior variability of the biological samples' projections is visualized in V by plotting each row of the matrices $(\boldsymbol{\psi}_1^k, \boldsymbol{\psi}_2^k)$, $k = 1, \ldots, K$, in the same figure.

We can see the uncertainties



Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren et al, 2017 (JASA).

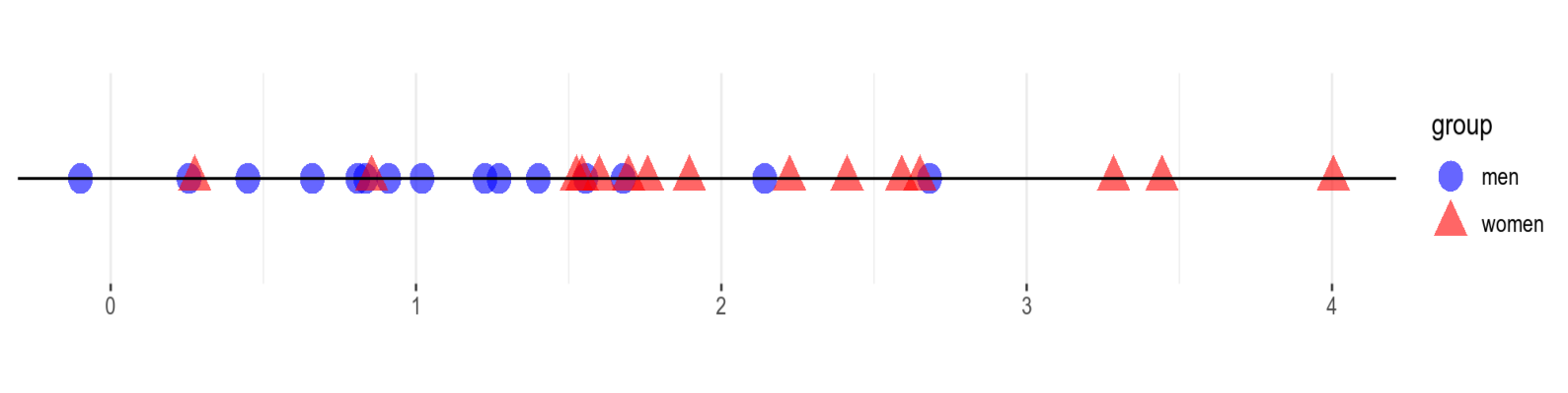
A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space $\mathcal{V}$.

Part V

Communities and networks

Graphs and Nonparametric two sample testing

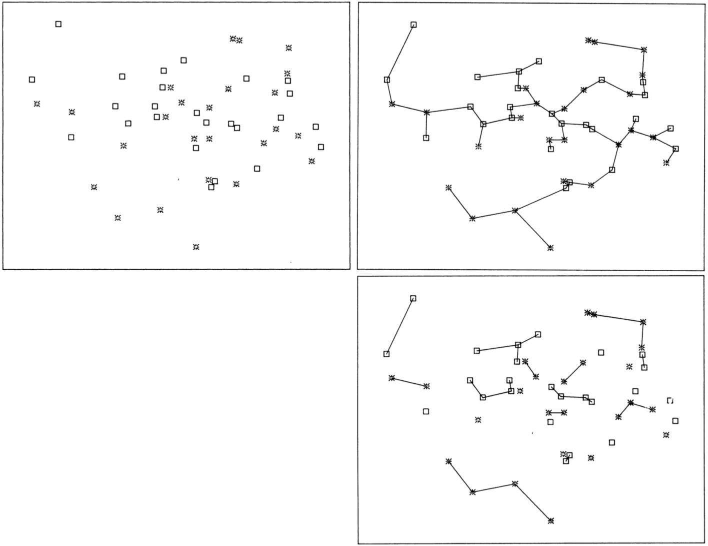
Friedman and Rafksy devised an extension of the Wald Wolfowitz test.



For univariate continuous observations:

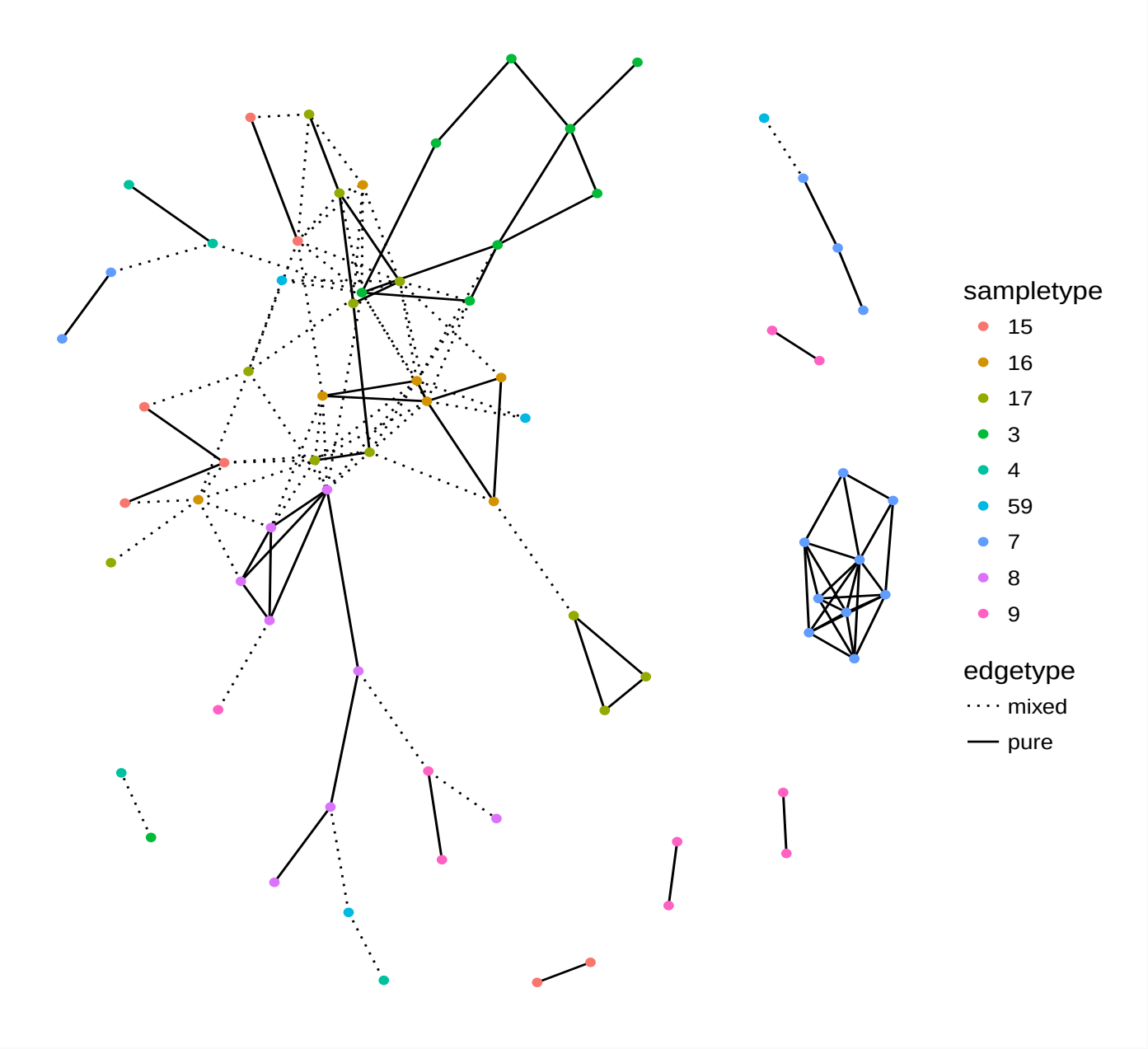
- Pool the observations.
- Rank the observations.
- Count the number of runs (sequences of observations that are from the same sample and follow each other).

The Test statistic is the total number of ”runs”.



Multivariate extension: multivariate ranking?

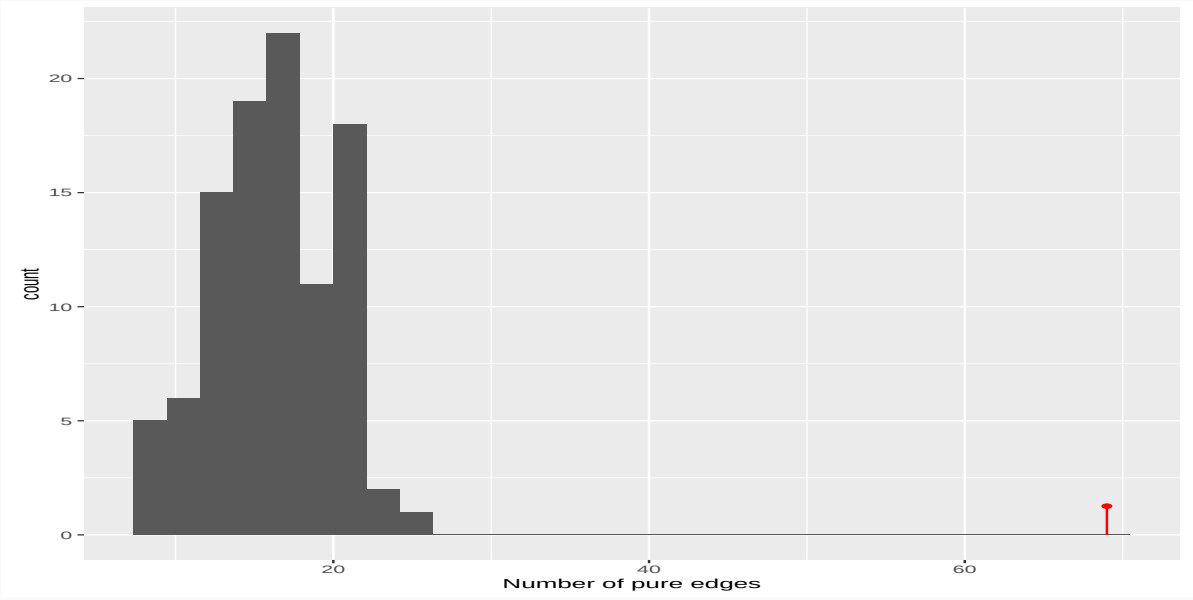
Friedman and Rafsky proposed to use minimal spanning trees as a multivariate generalization of the univariate sorted list.
Runs are ”pure” edges.
The null distribution of the test statistics can be computed using permutation tests: fix the tree and permute the labels.
Good power in finite samples for multivariate data (against general alternatives: location, spread, and shape).
Note: You can use other ”skeleton graphs” not necessarily the minimum spanning tree.


sampletype
15
16
17
3
4
59
7
8
9
edgetype
mixed
pure

Graph based Tests in practice

In our example: $F_o = 69$

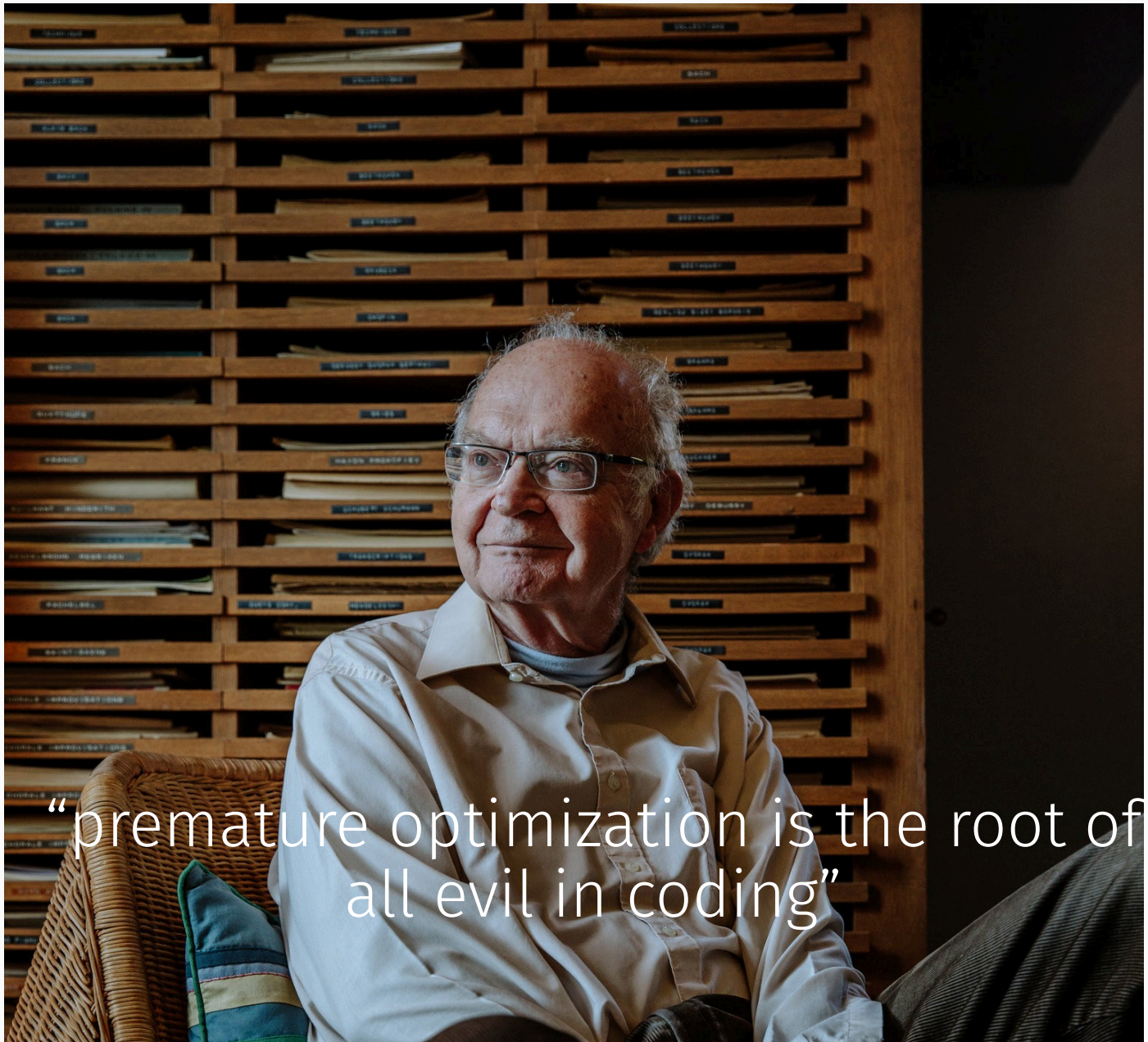
Keeping the graph fixed, permute the labels and recompute the number of pure edges. All 1000 simulated values had $F_s < 69$



so $p < 0.001$.

Different versions of this test are implemented in the R package `phyloseqGraphTest` written by Julia Fukuyama (available on CRAN).

The Yoda of Silicon Valley



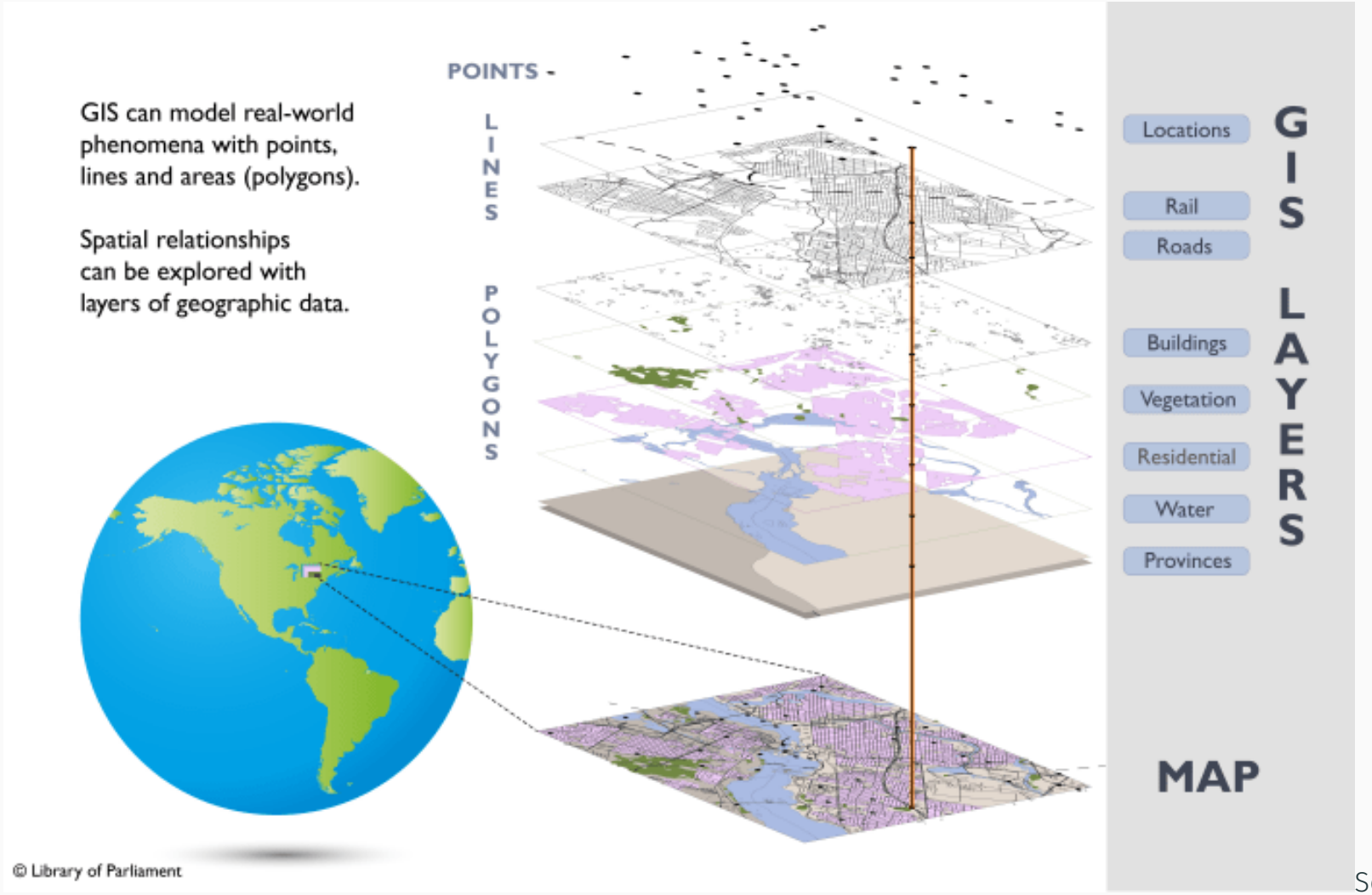
In Statistics

"premature summarization is the root of

all evil in statistics and

data science"

Keeping all the information



Source:

Ottawa, 2016, using data Natural Resources Canada, Canvec +, 2015. Software used: Esri, ArcGIS, version 10.3.1.

- Waste not, Want not supplementary material: http://joey711.github.io/waste-not-supplemental/
- Pregnancy Study:PNAS Supplement
- Prevotella and Bacteroidetes with Sue Huse and Anastassia Gorvitovskaia : https://purl.stanford.edu/fs506ff9976
- Complete Bioconductor workflow, F1000Research: http://f1000research.com/articles/5-1492/v1
- PSB Reproducible research examples (enterotypes, oral, pregnancy microbiomes): http://statweb.stanford.edu/~susan/papers/PSBRR.html
- Treelapse for antibiotics
- Short course materials open source

R packages and resources

```
 phyloseq: http://bioconductor.org/packages/stats/
           bioc/phyloseq/
    dada2: http://bioconductor.org/packages/stats/
           bioc/dada2/
treelapse: https:
           //krisrs1128.github.io/treelapse/
treelapse antibiotics http://statweb.stanford.
           edu/~kriss1/antibiotic.html
microbiome_pvlm: https://github.com/
           krisrs1128/microbiome_plvm
decontam:
           https://github.com/benjjneb/decontam/
adaptiveGPCA: https://cran.r-project.org/web/
           packages/adaptiveGPCA/index.html
bootLong:   https:
```

Solutions : respect the data.

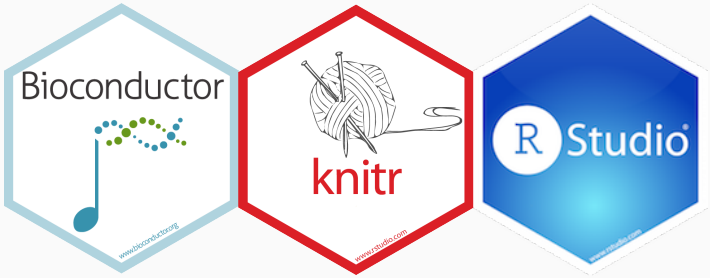
- Poor data quality, information $\longrightarrow$ quality scores & probability.
- Maintain all information $\longrightarrow$ sequences are names, retain uncertainties.
- Interpretation $\longrightarrow$ latent variables (gradients, clusters, networks).
- Reproducibility $\longrightarrow$ complete code source.
- Heterogeneity $\longrightarrow$ multicomponent objects: BioC.
- Training and collaboration $\longrightarrow$ Rmd and html.

More Solutions.

- Nonlinearity $\longrightarrow$ Transformations help.
- Dependencies other than linear $\longrightarrow$ new correlation coefficients (XiCOR).
- Local information $\longrightarrow$ collate patches and align(UMAP).
- Tree and graph integration $\longrightarrow$ multi-table analyses with kernels.
- Testing $\longrightarrow$ nested permutations and dependent (block bootstrap).
- Robustness $\longrightarrow$ sparse methods using regularization.

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

Thanks to the `R` and `Bioconductor` community and to co-authors.



Wolfgang Huber, Martin Morgan, Joey McMurdie, Ben Callahan, JJ Allaire and Rob Gentleman.

Thank you to the organizers for inviting me.

Lab Group and David Relman

Postdoctoral Fellows Paul (Joey) McMurdie, Ben Callahan, Christof Seiler, Pratheepa Jeganathan. Students: John Cherian, Diana Proctor, Lan Huong Nguyen, Julia Fukuyama, Kris Sankaran, Claire Donnat. Funding from NIH TR01 and NSF-DMS.

References

[1] Ben J Callahan, Kris Sankaran, Julia A Fukuyama, Paul J McMurdie, and Susan P Holmes. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. F1000Research, 5, 2016.

[2] Benjamin J. Callahan, Paul J. McMurdie, and Susan P. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal, (10.1038/ismej.2017.119):1–5, 2017.

[3] BJ Callahan, PJ McMurdie, MJ Rosen, AW Han, AJ Johnson, and SP Holmes. Dada2: High resolution sample inference from amplicon data. Nature Methods, 13(7):581, 2016.

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

[5] Julia Fukuyama, Laurie Rumker, Kris Sankaran, Pratheepa

Jeganathan, Les Dethlefsen, David A. Relman, and Susan P. Holmes. Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. PLOS Computational Biology, (10.1371/journal.pcbi.1005706), 2017. August 16.

[6] 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.

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

[8] Lan Huong Nguyen and Susan Holmes. Bayesian unidimensional scaling for visualizing uncertainty in high

dimensional datasets with latent ordering of observations. BMC bioinformatics, 18(10):394, 2017.

[9] Lan Huong Nguyen and Susan Holmes. Ten quick tips for effective dimensionality reduction. PLoS computational biology, 15(6), 2019.

[10] Boyu Ren, Sergio Bacallado, Stefano Favaro, Susan Holmes, and Lorenzo Trippa. Bayesian nonparametric ordination for the analysis of microbial communities. Journal of the American Statistical Association, (February), 2017.

[11] Kris Sankaran and Susan Holmes. Latent variable modeling for the microbiome. Biostatistics, pages 31–pages, 2018.

[12] Kris Sankaran and Susan P Holmes. Multitable methods for microbiome data integration. Frontiers in genetics, 10, 2019.