Heterogeneous Data

Susan Holmes,@SherlockpHolmes

STAMPS, Woods Hole, July 25, 2019
(Big) Challenges for those of us working from the ground up

- Heterogeneity.
- Heteroscedasticity.
- Information Leaks.
- Multiplicity of Choices.
- Reproducibility.
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
  - Spatial Information
  - Rankings
- Amounts of dependency: independent/time series/spatial.
- Different technologies used (Illumina, MassSpec, NMR, RNA-seq).
Statistics: separate the model from the data

See a complete book:
http://bios221.stanford.edu/book/
Read data are counts, the data are not compositional.... the parameters are!

- After perturbations amounts of bacteria go up & down.
- Remove contaminants using read numbers (decontam and BARBIE).
- Estimating depth bias requires read numbers.
- Some bacteria live in symbiosis with others.
- We need the read depths for variability/standard error estimation and uncertainty quantification.
- Data transformations can be used to remove ”multiplicative error” and equalize the variance.
Glossary (statistical)

- Probabilistic.
- Statistical.
- Estimation.
- Parameter.
- Parametric.
- Nonparametric.
- Independent / dependent.
- Bias.
Glossary (computational)

**vector** \( v[1] \quad v[2] \quad v[3] \quad \ldots \quad v[n] \)

**matrix** The dimension of \( A \) here is 4 by 3.

\[
\begin{array}{ccc}
\end{array}
\]

**factor** A categorical variable with levels: "O", "AB", "B" etc.


**list** A container for different objects, usually of different type and dimensions.

**data.frame** A list whose components are of the variables, usually of different types.

**status** response/explanatory: the "formula" \( y \sim x \).
Mix and Match data types

- **Matrix**: Has to be either all numerics or all characters.
- **Mixed type**: `data.frame`, a special type of list, the name of each component is the variable name.
- **Complex list**: S4 object with special components called "slots".
Resources for dealing with different data types and structures in R

- If you’ve never used R:
  (a lecture 0 link here: Lab0-DirectoryFiles.html).
- About data types: LabDataTypes.html
- Hadley’s chapter on data structures:
  http://adv-r.had.co.nz/Data-structures.html
- Advanced resource: the special S4 data types
Human Microbiome: What are the data?

**DNA**  A biomarker DNA count (16sRNA-gene especially).
**DNA**  All the Genomic material present (shotgun metagenomics).
**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.
phyloseq

Joey McMurdie (joey711 on github).
Available in Bioconductor.
Heterogeneous Data Objects
Input and data manipulation with **phylloseq**
(McMurdie and Holmes, 2013, Plos ONE)
As always in R: object oriented data.
phyloseq

data structure & API

Experiment Data

phyloseq

constructor:

phyloseq

otu_table, sam_data, tax_table, phy_tree, refseq

Accessors:

get_taxa
get_samples
get_variable
nsamples
ntaxa
rank_names
sample_names
sample_sums
sample_variables
taxa_names
taxa_sums

Processors:

filter_taxa
merge_phyloseq
merge_samples
merge_taxa
prune_samples
prune_taxa
subset_taxa
subset_samples
tip_glom
tax_glom

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

Input
- sample data
- OTU cluster output

Import
- import_biom
- import_mothur
- import_pyrotagger
- import_qiime
- import_RDP

Preprocessing
- filter_taxa
- filterfun_sample
- genefilter_sample
- prune_taxa
- prune_samples
- subset_taxa
- subset_samples
- transform_sample_counts

Summary / Exploratory Graphics
- plot_network
- plot_heatmap
- plot_ordination

Direct Plots
- plot_richness
- plot_tree
- plot_bar

Inference, Testing
- bootstrap
- permutation tests
- regression
discriminant analysis
- multiple testing
gap statistic
- clustering
- procrustes

work flow
phyloseq

plot_ordination()

plot_network()

plot_bar()

plot_heatmap()

plot_tree()

plot_richness()
# Main title

This is an [R Markdown](my.link.com) document of my recent analysis.

## Subsection: some code

Here is some import code, etc.

```{r}
library("phyloseq")
library("ggplot2")
physeq = import_biom("datafile.biom")
plot_richness(physeq)
```

Our Goal with Collaborators:

- Reproducible analysis workflow with R-markdown
- Better Reproducibility

**microbiome data**

```
source.Rmd
```

---

Complete HTML5

knitr::knit2html()

**markdown**

(code + console) + 

**figures**

phyloseq +

`ggplot2` +

etc.

---
Part I

An Example

ARTICLE

do:10.1038/nature09944

Enterotypes of the human gut microbiome


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 abundant molecular functions are not necessarily provided by abundant species, highlighting the importance of a functional analysis to understand microbial communities. Although individual host properties such as body mass index, age, or gender cannot explain the observed enterotypes, data-driven marker genes or functional modules can...
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).
- Choose a graphical representation.
There are thus more than 200 million possible ways of analyzing this data:

\[ 5 \times 100 \times 10 \times 40 \times 8 \times 16 \times 2 \times 4 = 204800000 \]
Heteroscedasticity

Different variances across runs and samples and bacteria.

- Variance stabilizing transformation

Transformations (variance stabilizing): $\log$, $\text{asinh}$, etc.
For more information: Waste Not, want not video lecture
Tutorial on DESeq2 and phyloseq

- Rank-transformations (see LabRobustness.html).
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.

hidden variables.
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.

hidden variables.
Paths in thinking about these heterogeneous systems

- Think in terms of mixtures (not one parametric population).
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.
The Yoda of Silicon Valley

“premature optimization is the root of all evil in coding”
“premature summarization is the root of all evil in statistics”
Different Levels of Dependencies

- Spatial (2 or 3 dimensional), oral microbiome study with Diana Proctor (Data and code here: ).
- Longitudinal studies:
  Equally spaced: (pregnancy).

Stability
Perturbation study (multidomain data).
Not equally distant time points.

Between point **variation** should be as equal as possible.
See Peter Diggle’s text: Analysis of Longitudinal Data, 2002.
Chapter on design of high-throughput experiments
Part II

Statistics is Hard, we are all Outsiders?
The Jargon: we need the right words

We can often find helpful information about biological phenomena by using Google or Wikipedia, because the words are quite specific: Lymphocytes, CD45, epigenetics,.....
Statistics often uses common words with hundreds of other meanings: Normal, Geometric, Independent, Expectation, Process, Mean, Bias, Conditional, Cluster, Random, Variance, generative, parameter ...
Useful Concepts with (Hard) Buzzwords

- Dependence.
- Multivariate.
- Effective sample size.
- Latent Variables.
- Noise.
- Robust.
- Spatio-Temporel Data Analysis.
- Degrees of Freedom.
- Heteroscedasticity.
False Friends
# Main title

This is an [R Markdown](my.link.com) document of my recent analysis.

## Subsection: some code

Here is some import code, etc.

```{r}
library("phyloseq")
library("ggplot2")
physeq = import_biom("datafile.biom")
plot_richness(physeq)
```

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

Better Reproducibility

knitr::knit2html()

Complete HTML5

microbiome data

source.Rmd

phyloseq +
ggplot2 +
etc.

(markdown (code + console) + figures)
Reproducible Research Workflow

Sequences and qualities

- dada2: infer sample composition

Metadata and additional info

- phyloseq: transform, subsample, test, track
- ggplot2: visualization
- deseq2: differential abundance testing
- vegan: ecological statistics

R

- Rmd: Workflow, versions, choices
- Rdata: All data, results, one file
Bioconductor workflow for microbiome data analysis: from raw reads to community analyses [version 1; referees: awaiting peer review]

Ben J. Callahan¹, Kris Sankaran¹, Julia A. Fukuyama¹, Paul J. McMurdie², ✉ Susan P. Holmes¹

Author affiliations
Grant information

This article is included in the Bioconductor channel.
Reproducible research R markdown code and data

- Waste not, want not paper, Plos Comp Bio.
- Oral Microbiome
- Web page with all code and images as they appear from the code. Complete Analysis of Colonic Cleanout Data
- Enterotypes, oral microbiome PSB 2016.
- Treelapse for antibiotics
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_pvlm
decontam:  https://github.com/benjjneb/decontam/
adaptiveGPCA:  https://cran.r-project.org/web/packages/adaptiveGPCA/index.html

Modern Statistics for Modern Biology
http://bios221.stanford.edu/book/
Solutions for microbiome analyses: respect the data.

- Poor data quality, information → quality scores & probability.
- Maintain all information → sequences are names.
- Reproducibility → complete code source.
- Heterogeneity → multicomponent objects: phyloseq.
- Training and collaboration → Rmd and html.
More problems: todo next time and further

- Multivariate statistics = analysing matrices.
- Multivariate data combined with trees and networks.
- Multidomain problems (metagenomics, RNA-seq, images, ...).
- Interpretation $\rightarrow$ latent variables (gradients or clusters).
- Model choices (transformation choices).