Multidomain data in the Human Microbiome

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

Bio-X and Statistics, Stanford University

Stanford, 2016
Homogeneous data are all alike; all heterogeneous data are heterogeneous in their own way.
Studying the 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.
Challenges when working from the ground up

- Heterogeneity.
- Poor data quality.
- Structured high-dimensionality.
- Graph or Tree integration.
- High quality graphics.
- Reproducibility.
Part I

Heterogeneity
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 (454, Illumina, MassSpec, NMR, RNA-seq).
Human Microbiome: What are the data?

DNA  The Genetic ‘signature‘ of the bacteria (16S rRNA-gene).
RNA  What genes are being turned on (gene expression).
Proteomics 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.

Everything is data....

...... no metadata.
Source: YK Lee and SK Mazmanian Science, 2010
Heterogeneous Data Objects: S4 classes
Input and data manipulation with phyloseq (McMurdie and Holmes, 2013, Plos ONE).
DADA2: High-resolution sample inference from Illumina amplicon data

Benjamin J Callahan¹, Paul J McMurdie², Michael J Rosen³, Andrew W Han², Amy Jo A Johnson² & Susan P Holmes¹

We present the open-source software package DADA2 for modeling and correcting Illumina-sequenced amplicon errors (https://github.com/benjneb/dada2). DADA2 infers sample sequences exactly and resolves differences of as little as 1 nucleotide. In several mock communities, DADA2 identified more real variants and output fewer spurious sequences than other methods. We applied DADA2 to vaginal samples from a...
16S rRNA gene for bacterial fingerprinting

AGGTAAATCTCAATGAAATGACGGGCACCCGACAAAGCGGACGAT
H28 H29 H30

GTGGTTAAATCAGAGAACACCCTTAACCTGGCTTGACACTCCA
H30 H31 H31 H32 H33a

CGGAATTTCTGAGATAGGAATGTCGCCTCGGGAACCGTACGACAGGTG
H33b H33b ? ? H33a H34

CTGCATGCGGCTGCTCAAGCTCGGTGGTTGAGATGCTGGCCTGCC
H34 H35 H36 H36 H37 H37

AAGCGAGCGCAGGCCCTAATTCCTTGATGCGCCAGGCTCCGCCGC
H38 H38 H39 H39

AAGGAAGCTGCAGGCTGATAACGTTGAGGAAAGGGCAGCTCAGCAGT
H39 H40 H40 H38 H34

ATCAGCGCCCTTACGACCCGCGCTACACGCTGCTCAATCGGCTAC
H34 H32 H30 H41

AGGAGAAGCGACCTCAGGAAGCAGCAGCGGACCTCATAAAGTGGGCGG
H41 H41 H41

UCGGUCUGUGCCUGGCAGGACUGCCGGCUUCAUUGGAGA
H35 H36 H37 H38 H39 H40 H41
Problems involved in going from reads to ‘species’

Standard method: cluster within 97% similarity.

- Low resolution: 97% gives genus level at best
- High false positive rate: #(OTUs) $\gg$ richness.
- Big data scaling: time scales super-linearly
Probabilistic Model

Sample sequences | Amplicon reads | OTUs

Errors | DADA2 | Make OTUs
Error Model

\[ s: \text{ATTAACGAGATTATAACCAGAGTACGAATA...} \]
\[ r: \text{ATCAACGAGATTATAACAAAGAGTACGAATA...} \]

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

- \( P \) probabilities of substitutions (A \( \rightarrow \) C)
- \( q \) Quality score (Q=30)
- Batch effect (run)
Forward and Reverse quality profiles along the reads.
Frequencies of each type of nucleotide transition as a function of quality.
Accuracy: Simulated data

mothur (an)

DADA2

TP: 978
FP: 272
FN: 77
cor: 0.935

TP: 1042
FP: 0
FN: 13
cor: 0.999

Resolution: L. crispatus

Data: MacIntyre et al. Scientific Reports, 2015.

OTU O1

42 pregnant women
Resolution: L. crispatus

Data: MacIntyre et al. Scientific Reports, 2015.

42 pregnant women
Reproducible Research Workflow

Sequences and qualities
- `dada2` infer sample composition
- `phyloseq` transform, subsample, test, track
- `deseq2` differential abundance testing

Metadata and additional info
- `ggplot2` visualization
- `vegan` ecological statistics

R

Rmd
- Workflow, versions, choices

Rdata
- All data, results, one file
See complete workflow on Bioconductor channel of F1000:
http://f1000research.com/articles/5-1492/v1

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.
Heterogeneous Data Objects
Object oriented input and data manipulation with **phyloseq**

**phyloseq**

- **OTU Abundance**
  - `otu_table`
- **Sample Variables**
  - `sample_data`
- **Taxonomy Table**
  - `tax_table`
- **Phylogenetic Tree**
  - `phylo`
- **Reference Seq.**
  - `refseq`

**Accessors:**
- `get_taxa`
- `get_samples`
- `get_variable`
- `n_samples`
- `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`

**Constructor:**
- `phyloseq`
  - `otu_table`, `sam_data`, `tax_table`, `phy_tree`, `refseq`

**Data Structure & API**

- `matrix`
- `data.frame`
- `matrix`
- `ape package`
- `Biostrings package`

**Optional**

- `DNAStringSet`
- `RNAStringSet`
- `AAStringSet`
phyloseq

plot_ordination()

plot_network()

plot_bar()

plot_heatmap()

plot_tree()

plot_richness()
Part III

Combining different data types
Useful first order representation: Many Matrices

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

Holmes (2005), Duality Diagrams, matrices with metrics and measures.
Double Principal Coordinate Analysis

Pavoine, Dufour and Chessel (2004), Purdom (2010) and Fukuyama et al. (2011). Suppose we have n species in p locations and a (euclidean) matrix $\Delta$ giving the squares of the pairwise distances between the species on the tree. 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 $\Delta$.
- 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.
Antibiotic Time Course Data

Measurements of about 2500 different bacterial OTUs from stool samples of three patients (D, E, F) Each patient sampled ~ 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.
(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 with information about 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”). DPCoA separates out the stressed communities along the first axis (in the direction associated with *Bacteroidetes*), although only for subjects D and E.
Community points as represented by DPCoA. The labels represent subject plus antibiotic condition.
Conclusions for Antibiotic Stress

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.
out.dpcoa.log <- ordinate(pslog, method = "DPCoA")

evals <- out.dpcoa.log$eig
plot_ordination(pslog, out.dpcoa.log, color = "age_binned",
shape = "family_relationship") +
coord_fixed(sqrt(evals[2] / evals[1])) +
labs(col = "Binned Age", shape = "Litter")

plot_ordination(pslog, out.dpcoa.log, type = "species",
color = "Phylum") +
coord_fixed(sqrt(evals[2] / evals[1]))
Double Principal Coordinate Analysis

Samples in first plane.

Species in first plane.
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). We also have such generalizations that cover variability of points on a graph taken from standard spatial statistics.
Co-Inertia

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

\[ \text{sum}(x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3) \]

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). Co-inertia analysis (CIA) is 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.

\[
RV(A, B) = \frac{\text{Tr}(A'B)}{\sqrt{\text{Tr}(A'A)\text{Tr}(B'B)}}
\]

Example

Combining different types of data.

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

<table>
<thead>
<tr>
<th></th>
<th>Taxa</th>
<th>Kegg</th>
<th>MassSpec+</th>
<th>MassSpec-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxa</td>
<td>1</td>
<td>0.565</td>
<td>0.561</td>
<td>0.670</td>
</tr>
<tr>
<td>Kegg</td>
<td>0.565</td>
<td>1</td>
<td>0.686</td>
<td>0.644</td>
</tr>
<tr>
<td>MassSpec+</td>
<td>0.561</td>
<td>0.686</td>
<td>1</td>
<td>0.568</td>
</tr>
<tr>
<td>MassSpec-</td>
<td>0.670</td>
<td>0.644</td>
<td>0.568</td>
<td>1</td>
</tr>
</tbody>
</table>
Multiple table methods

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’. This is called the ‘compromise’ and all the individual tables can be projected onto it.
Sparse CCA, then PCA

CCA: Canonical Correlation Analysis.
PCA: Principal Components Analysis.

- There are two tables in the study presented here, one for microbes and another with metabolites. 12 samples were obtained, each with measurements at 637 m/z values and 20,609 OTUs; however, about 96% of the entries of the microbial abundance table are exactly zero.
- CCA chooses a subset of available features that capture the most co-Inertia.
- We then apply PCA to this selected subset of features. In this sense, we use sparse CCA as a screening procedure, rather than as an ordination method.
## Call: CCA(x = t(X), z = t(metab), penaltyx = 0.15, penaltyz = 0.15)

## Num non-zeros u's: 5
## Num non-zeros v's: 15
## Type of x: standard
## Type of z: standard
## Penalty for x: L1 bound is 0.15
## Penalty for z: L1 bound is 0.15
## Cor(Xu,Zv): 0.974

With these parameters, 5 microbes and 15 metabolites have been selected, based on their ability to explain covariation between tables. Further, these 20 features result in a correlation of 0.974 between the two tables. The microbial and metabolomic data reflect similar underlying signals. To relate the recovered metabolites and OTUs to characteristics of the samples on which they were measured, we use them as input to an ordinary PCA.
A PCA triplot produced from the CCA selected features in from multiple data types (metabolites and OTUs). Triangles for Knockout and circles for wild type. The main variation in the data is across PD and ST samples (different diets).

Part IV

Combining graphs and covariates.
A network created by thresholding the Jaccard dissimilarity matrix. The colors represent which mouse the sample came from and the shape represents which litter the mouse was in.
Graph-based two-sample tests were introduced by Friedman and Rafsky, 1979 as a generalization of the Wald-Wolfowitz runs test. They proposed the use of a minimum spanning tree (MST) based on the distances between the samples, and then counting the number of edges on the tree that were between samples in different groups. It is not necessary to use an MST, graphs made by linking nearest neighbors or distance thresholding can also be input. No matter what graph we build between the samples, we can approximate a null distribution by permuting the labels of the nodes of the graph.
We perform a test using an MST with Jaccard dissimilarity. We want to know whether the two litters come from the same distribution.
See Shiny-Phyloseq
# 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)
Goals already attained:

- Data quality through more NGS denoising (DADA implementation)[1].
- Data integration phyloseq.
- Data normalization **Gamma-Poisson** noise model.
- High quality graphics, easy to make and change.
- Conjoint analyses of trees, networks and count data.
- Threshold, sensitivity tests and modeling simulations.
- Interactive graph visualizations: Shiny-phyloseq.
- Reproducibility: open source standards, publication of source code and data. (R).
Current work in progress

- Longitudinal analyses: antibiotic dynamics.
- NMR, Mass spec, proteomic multi-table integration within phyloseq.
- Uncertainty propagation.
In fact we know uncertainties

Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren et al, 2016 (arXiv).

A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space $V$. 
Benefitting from the tools and schools of Statisticians.......
Collaborators:

David Relman
Alfred Spormann
Elisabeth Purdom

Josh Elias
Justin Sonnenburg
Sergio Bacallado
Lab Group

Postdoctoral Fellows: Paul (Joey) McMurdie, Ben Callahan, Christof Seiler.
Students: John Cherian, Diana Proctor, Daniel Sprocket, Lan Huong Nguyen, Julia Fukuyama, Kris Sankaran.
Funding from NIH TR01 and NSF-DMS.
dada2, phyloseq

Available in Bioconductor.
How can I learn more?
http://www-stat.stanford.edu/~susan/

- **STAMPS**: MBL, Woods Hole, August 2016.
- **SISMID 2016**: Summer Institute in Statistics and Modeling for Infectious Diseases, July, Seattle.
- Postdoctoral positions.


Sandrine Pavoine, Anne-Béatrice Dufour, and Daniel Chessel.
From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis.  

Elizabeth Purdom.  
Analysis of a data matrix and a graph: Metagenomic data and the phylogenetic tree.  

C. R. Rao.  
The use and interpretation of principal component analysis in applied research.  
How to compress the data?

...without losing too much information?

Statisticians know the answer: sufficiency.

Models for noise: hierarchical Gamma-Poisson: we know how to transform the data to stabilize the variance (Delta-method).

Keeping Track of uncertainties

Joint work with Sergio Bacallado, Lorenzo Trippa, Stefano Favaro, Boyu Ren.
Bayesian Nonparametric Ordination for the Analysis of Microbial Communities
Higher resolution strain clustering: DADA2

L. crispatus sampled from 45 pregnant women