

# Data Types for Heterogeneous Data

Susan Holmes,@SherlockpHolmes

STAMPS, Woods Hole, July 2019

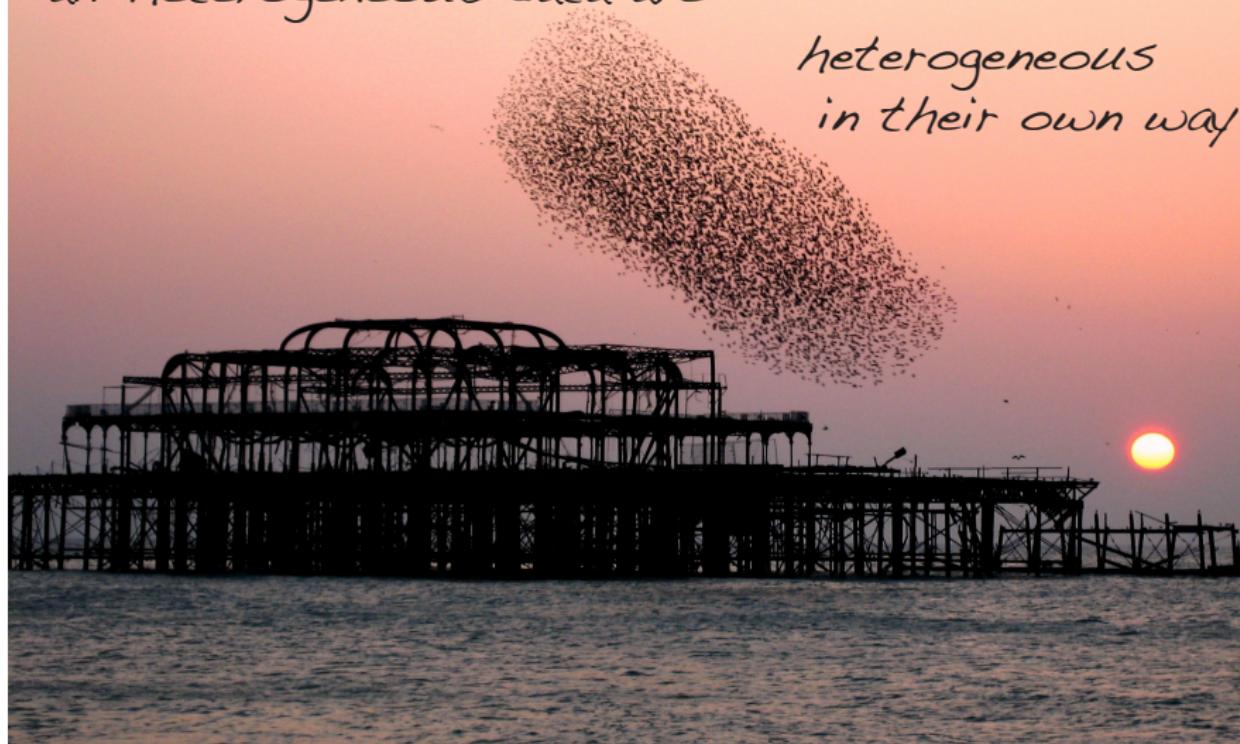
# (Big) Challenges for those of us working from the ground up

- ▶ Heterogeneity.
- ▶ Heteroscedasticity.
- ▶ Information Leaks.
- ▶ Multiplicity of Choices.
- ▶ Reproducibility.

## Heterogeneous data

Homogeneous data are all alike;  
all heterogeneous data are

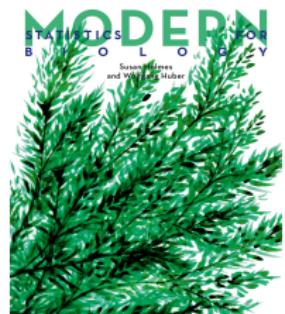
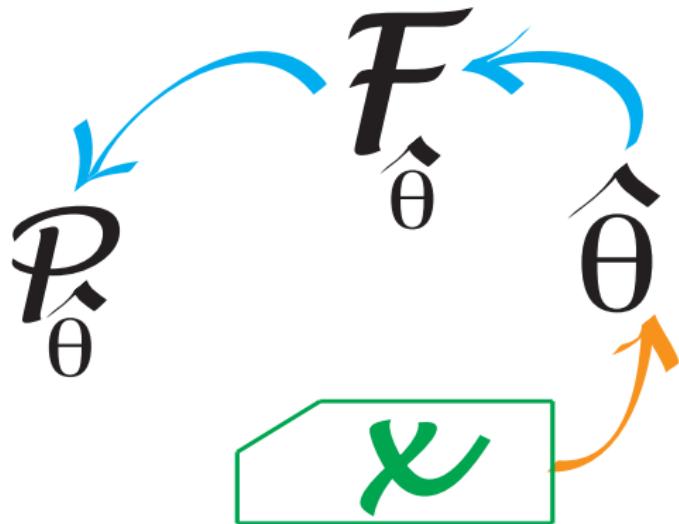
heterogeneous  
in their own way.



# 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

**vector**  $v[1] \quad v[2] \quad v[3] \quad \dots v[n]$

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

$$\begin{array}{ccc} A[1,1] & A[1,2] & A[1,3] \\ A[2,1] & A[2,2] & A[2,3] \\ A[3,1] & A[3,2] & A[3,3] \\ A[4,1] & A[4,2] & A[4,3] \end{array}$$

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

See here <http://web.stanford.edu/class/bios221/book/Generative.html>

**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  
<https://stuartlee.org/post/content/post/2019-07-09-s4-a-short-guide-for-perplexed/>.

# Human Microbiome: What are the data?

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

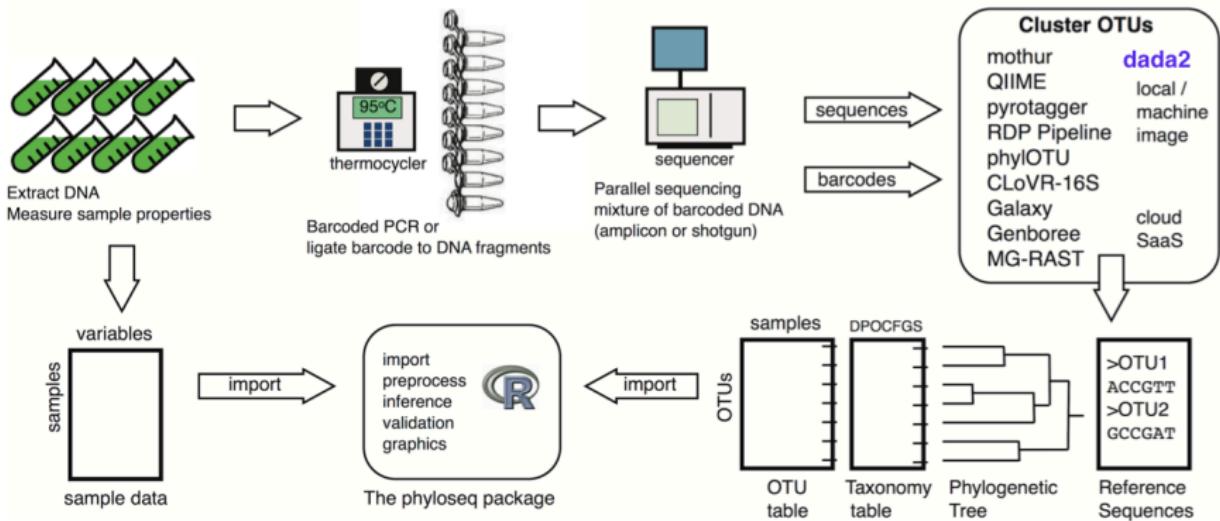
**RNA** What genes are being turned on (gene expression), transcriptomics.

**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

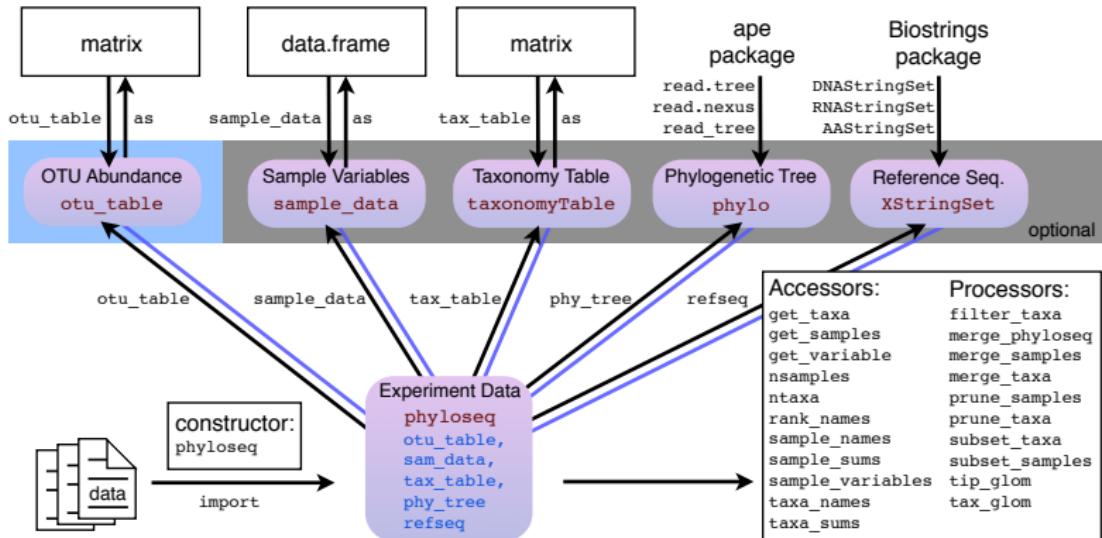
## Input and data manipulation with phyloseq

(McMurdie and Holmes, 2013, Plos ONE)

As always in R: object oriented data.

# phyloseq

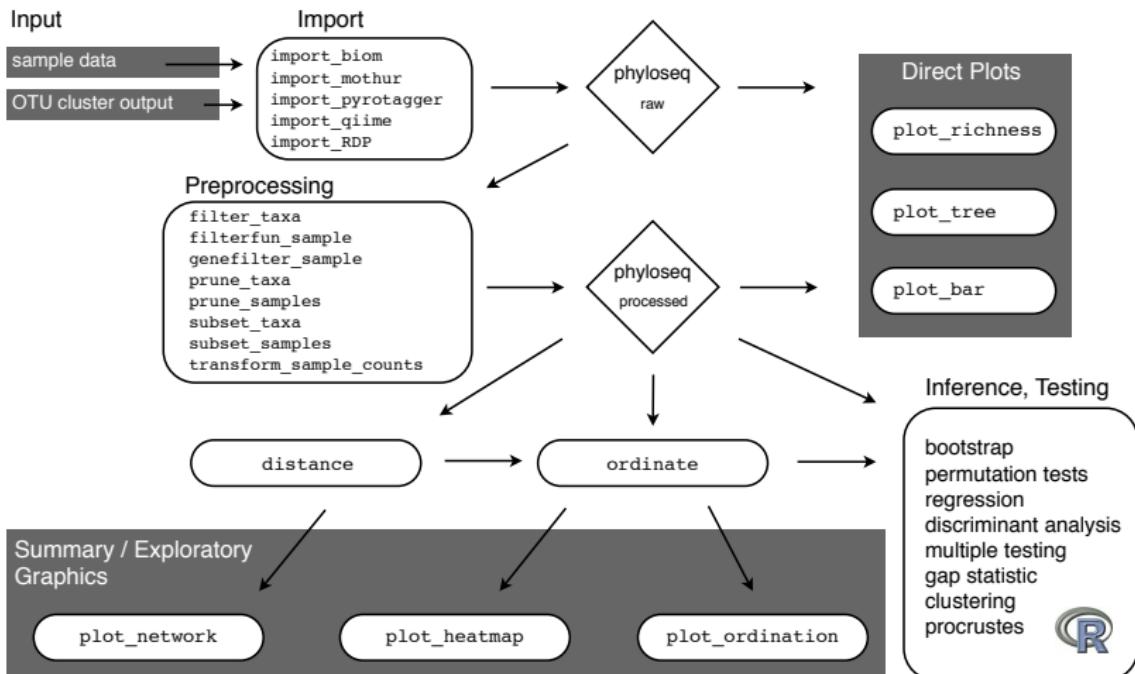
data structure & API



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

# phyloseq

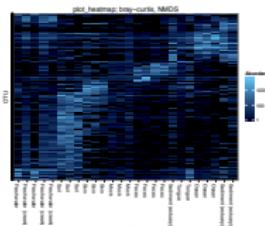
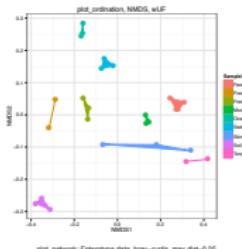
work flow



# phyloseq

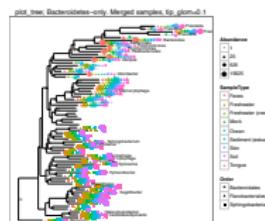
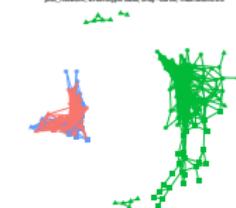
graphics

plot\_ordination()



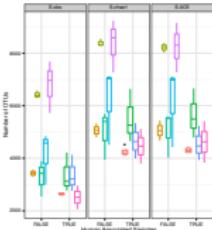
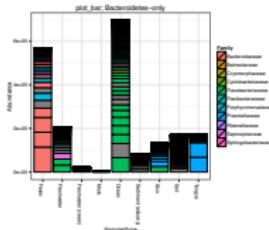
plot\_heatmap()

plot\_network()



plot\_tree()

plot\_bar()

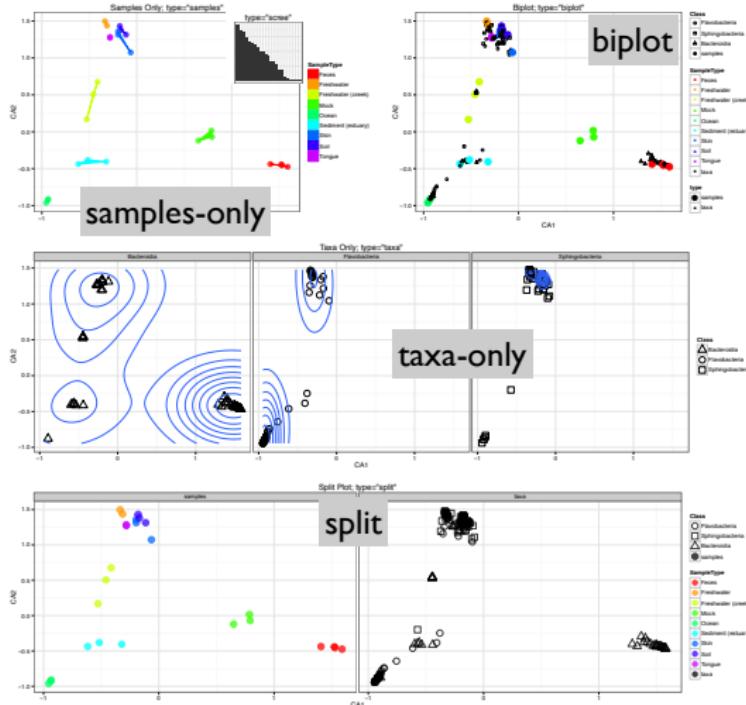


plot\_richness()

## phyloseq

## graphics

## plot\_ordination()



microbiome data

# Better Reproducibility



source.Rmd

```
# 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)  
```
```

Complete HTML5

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

phyloseq +  
ggplot2 +  
etc.

knitr::knit2html()

markdown  
(code + console) +  
figures

# Part I

## An Example

### ARTICLE

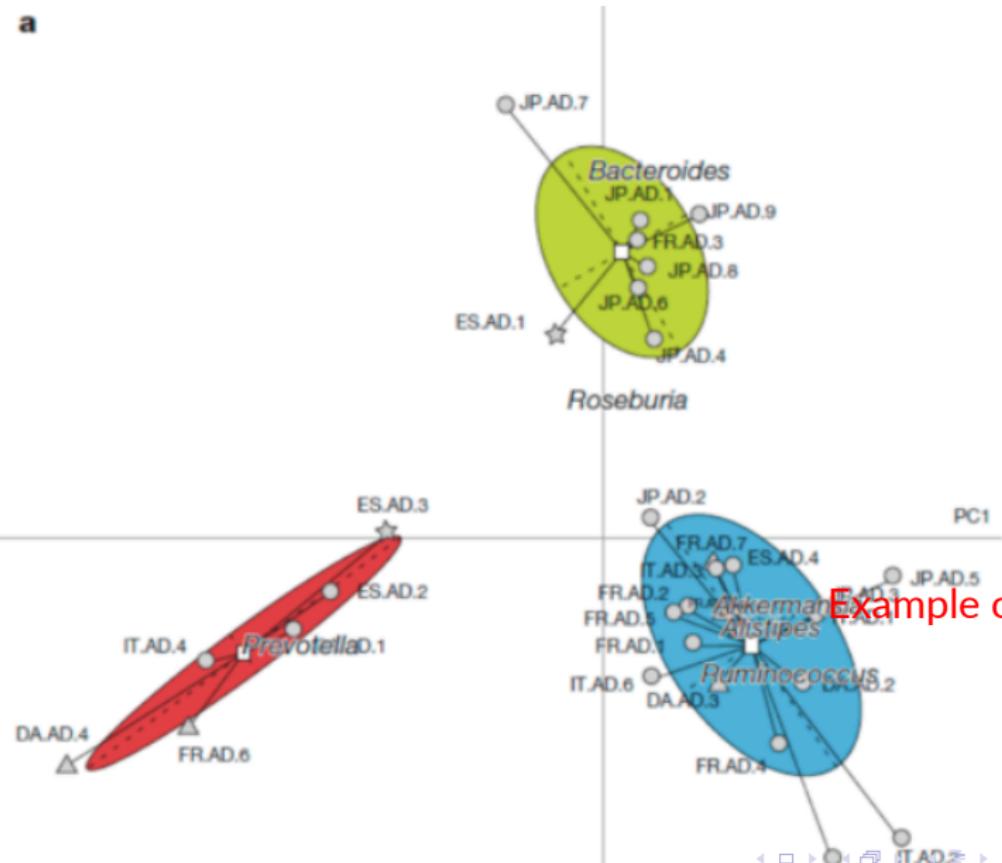
doi:10.1038/nature09944

# Enterotypes of the human gut microbiome

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

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

a



## 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:

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

# phyloseq



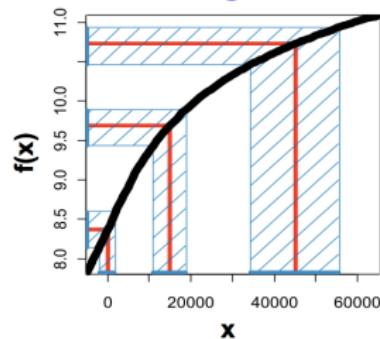
Joey McMurdie (joey711 on github).  
Available in Bioconductor.



# Heteroscedasticity

Different variances across runs and samples and bacteria.

- ▶ variance stabilizing transformation



- 
- ▶ Transformations (variance stabilizing): log, asinh, etc.  
For more information: Waste Not, want not video lecture  
Tutorial on DEseq and phyloseq
  - ▶ Rank-transformations (see LabRobustness.html).

# 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_plvm`: [https://github.com/krisrs1128/microbiome\\_plvm](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://github.com/PratheepaJ/bootLong/blob/master/vignettes/Workflow.Rmd>

**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 → latent variables (gradients or clusters).