Anatomy of taxonomic profiling data

- Patterns of variation in taxonomic profiling data
- Visualizing the data and statistical summaries
Statistical properties: diving into data
Overview of generic* amplicon workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

sequencing facility ➔ fastq files ➔ demultiplex (split samples by barcodes) ➔ quality filter/trim (remove adapters/primers) ➔ fasta files

might be done by sequencing facility

Some tools:
- sabre
- fastq_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

Some tools:
- trimmomatic
- fastx_filter (usearch/vsearch)
- bbdal sn (bbtools suite of tools)
- filterAndTrim (dada2)

generate OTUs ➔ count table ➔ fasta file taxonomy ➔ resolve ASVs ➔ dereplication ➔ chimera removal

Standard outputs

<table>
<thead>
<tr>
<th>Seq</th>
<th>Sample_A</th>
<th>Sample_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq_1</td>
<td>0</td>
<td>428</td>
</tr>
<tr>
<td>Seq_2</td>
<td>306</td>
<td>323</td>
</tr>
<tr>
<td>Seq_3</td>
<td>217</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Analysis

Some tools:
- most workflows
- SEPP/TIPP
- RDP classifier

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpecEase
- DESeq2

Beta diversity
e.g. dissimilarity metrics
ordination
hierarchical clustering

Alpha diversity
- richness
evenness
diversity

Taxonomic summaries

Some tools that provide whole workflows:
dada2 runs within R (ASVs)
usearch/vsearch runs at the command line (ASVs and OTUs)
mother runs at the command line (OTUs only currently)

qiime2 provides a multi-interface environment that employs processing tools like those above, infrastructure for easily documenting all processing performed, and interactive visualizations

Common study designs

Cross-sectional
population (cohort) studies

Prospective
long-term follow-ups

Longitudinal
ecosystem dynamics

Case-control & Intervention
targeted experimental testing
Organisms and samples are not independent understanding & modeling the (latent) structure(s)
From individuals to populations, follow-ups, and multimodal data

Individual: \( S \times 1 \)
Population: \( S \times N \)
Longitudinal cohort: \( S \times N \times T \)

Sequence Variants / OTUs

"Multi-modal" longitudinal cohort

- OTU
- Metagenome
- Metabolome
se <- SummarizedExperiment(
  assays,
  rowData,
  colData,
  exptData
)

rowData(se)
asays(se)
exptData(se)
rowData(se)$entrezId assays(se)$count exptData(se)$projectId

colData(se)$tissue
se$tissue
se %in% CNVs
Abundance matrix

Open data:
Fecal microbiota in 1000 western adults
(Lahti et al. Nature Comm. 2014)
Core microbiota
only few species are prevalent (shared)
in population at a high abundance
Shared core microbiota in healthy adults depends on analysis depth and prevalence

"Blanket analysis" github.com/microbiome

Estimate frequency in the core for each phylotype & bootstrap for confidence intervals

Core & prevalence
prevalence(x)
core(x)
core_members(x)
Rare Biosphere in Human Gut: A Less Explored Component of Human Gut Microbiota and Its Association with Human Health

Where less may be more: how the rare biosphere pulls ecosystems strings

Alexandre Jousset, Christina Bienhold, Antonis Chatzinotas, Laure Gallien, Angélique Gobet, Viola Kurm, Kirsten Küsel, Matthias C Rillig, Damian W Rivett, Joana F Salles, Marcel G A van der Heijden, Noha H Youssef, Xiaowei Zhang, Zhong Wei & W H Gera Hol

The ISME Journal 11, 853–862(2017)
Relative versus absolute abundance: quantitative microbiome profiling

RMP vs. QMP: drastic effect on conclusions!

Normalizing library size?

If sample A has been sampled deeper than sample B, we the counts can be expected to be higher.

Compositional data: Divide by the total number of reads per sample (compositional abundance)

Problem: Abundant taxa may distort the ratios.
Transformations

\texttt{transform(x, \texttt{“compositional”})}
\texttt{transform(x, \texttt{“clr”})}
\texttt{transform(x, \texttt{“log10p”})}
\texttt{transform(x, \texttt{“hellinger”})}
\texttt{transform(x, \texttt{“identity”})}
Normalization and microbial differential abundance strategies depend upon data characteristics


Microbiome 5, Article number: 27 (2017) | Download Citation

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilcoxon rank-sum test</td>
<td>Also called the Mann-Whitney U test. A non-parametric rank test, which is used on the un-normalized (&quot;None&quot;), proportion normalized, and rarefied matrices</td>
</tr>
<tr>
<td>DESeq</td>
<td>nbinom Test—a negative binomial model conditioned test. More conservative shrinkage estimates compared to DESeq2, resulting in stricter type I error control</td>
</tr>
<tr>
<td>DESeq2</td>
<td>nbinomWald Test—The negative binomial GLM is used to obtain maximum likelihood estimates for an OTU's log-fold change between two conditions. Then Bayesian shrinkage, using a zero-centered normal distribution as a prior, is used to shrink the log-fold change towards zero for those OTUs of lower mean count and/or with higher dispersion in their count distribution. These shrunken log fold changes are then used with the Wald test for significance</td>
</tr>
<tr>
<td>edgeR</td>
<td>exact Test—The same normalization method (in R, method = RLE) as DESeq is utilized, and for differential abundance testing also assumes the NB model. The main difference is in the estimation of the dispersion, or variance, term. DESeq estimates a higher variance than edgeR, making it more conservative in calling differentially expressed OTUs</td>
</tr>
<tr>
<td>Voom</td>
<td>Variance modeling at the observational level—library sizes are scaled using the edgeR log counts per million (cpm) normalization factors. Then LOWESS (locally weighted regression) is applied to incorporate the mean-variance trend into precision weights for each OTU</td>
</tr>
<tr>
<td>metagenomeSeq</td>
<td>fitZIG—a zero-inflated Gaussian (ZIG) where the count distribution is modeled as a mixture of two distributions: a point mass at zero and a normal distribution. Since OTUs are usually sparse, the zero counts are modeled with the former, and the rest of the log transformed counts are modeled as the latter distribution. The parameters for the mixture model are estimated with an expectation-maximization algorithm, which is coupled with a moderated t statistic</td>
</tr>
<tr>
<td>ANCOM</td>
<td>Analysis of composition of microorganisms—compares the log ratio of the abundance of each taxon to the abundance of all the remaining taxa one at a time. The Mann-Whitney U is then calculated on each log ratio</td>
</tr>
</tbody>
</table>
Data is not compositional!

Model

Observations (Data)
State diagnosis & manipulation: from specific targets to the overall ecosystem

Diet
Life style
Antibiotics
Probiotics
Prebiotics
Fecal transplants
Phylum level Classification

Proteobacteria
Firmicutes
Bacteroidetes
Genus level Classification

AS

WW
Phylogenetic trees
Abundance matrix

- Sparse
- Non-Gaussian
- Overdispersed
- Compositional
- Complex
- Stochastic
- Hierarchical

```
Sample-1  Sample-2  Sample-3
Actinomycetaceae       0       0       0
Aerococcus             0       0       0
Aeromonas              0       0       0
Akkermansia            21      36      475
Alcaligenes faecalis et rel.  1      1       1
Allistipes et rel.      72     127      34
Anaerobiospirillum      0       0       0
Anaerofustis            0       0       0
Anaerostipes caccae et rel.  176     108     27
Anaerotruncus colihominis et rel.  10     48      38
Anaerovorax odorimutans et rel.  9      10      35
Aneurinbacillus        0       0       0
Aquabacterium          0       0       0
Astroleplasma et rel.   0       0       0
Atopobium              0       0       0
Bacillus               1       1       1
Bacteroides fragilis et rel.  67     32      15
Bacteroides intestinalis et rel.  2       2       1
```
Bacterial 'abundance types'
in 1000 western adults:

~% indicates proportion among prevalent taxa

Symmetric

\[ \sim 50\% \]

Abundance (Log_{10})

Right-skewed

\[ \sim 20\% \]

Abundance (Log_{10})

Bimodal

\[ \sim 10\% \]

Abundance (Log_{10})

Rare

\[ \sim 10\% \]

Abundance (Log_{10})

Left-skewed

Fat-tailed

\[ \sim 10\% \]

Abundance (Log_{10})

Abundance histograms (one-dimensional landscapes)

Population densities for Dialister:

```r
# Load libraries
library(microbiome)
library(phylotools)
pseq <- dietswap

# Visualize population densities for specific taxa
plot_density(pseq, "Dialister") + ggtitle("Absolute abundance")

# Same with log10 compositional abundances
x <- microbiome::transform(pseq, "compositional")
tax <- "Dialister"
plot_density(x, tax, log10 = TRUE) +
  ggtitle("Relative abundance") +
  xlab("Relative abundance (%)")
```

![Absolute abundance](image1)
![Relative abundance](image2)
Problems:

- Few replicates
- Non-gaussian, discrete, positive, skewed..
- Multiple testing
Hierarchical testing (Kris Sankaran)

Tree-based methods
- StructSSI
- phylofactor
- tree-PCA
- UniFrac

Source: Susan Holmes | http://web.stanford.edu/class/bios221/Short-Phyloseq-Resources.html
EDA for finding batch effects

- negative controls
- positive controls
- batch..
Statistical aspects: summary

- Biased
- Sparse
- Non-Gaussian
- Overdispersed
- Compositional
- Complex
- Stochastic
- Hierarchical
How to choose a correct model?

Parametric assumptions:
1. Independent samples
2. Data normally distributed
3. Equal variances

Type of question?
- Relationships
  - Do you have dependent & independent variables?
    - Yes: Regression analysis
    - No: Correlation analysis
      - Nonparametric: Spearman's rank correlation
      - Parametric: Pearson's r

- Differences
  - Differences between what?
    - Means: Fmax test or Bartlett's test
    - Variances: One-sample t-test

Type of data?
- Continuous
- Discrete, categorical
  - Any counts < 5?
    - Yes: Fisher's exact test
    - No: Chi-square tests, one and two sample

How many groups?
- More than two: One-way ANOVA
  - Transform data?
    - Yes: -
    - No: OK

- Two: Mann-Whitney U or Wilcoxon test

Transform data?
- No: OK
- Yes: -

Parametric assumptions satisfied?
- Yes: OK
- No: Mann-Whitney U or Wilcoxon test
Generative models
Model

Observations (Data)
Generative models

Construct a model
- Incorporate prior knowledge
- Learn the model with some data

Criticize the model
- Generate artificial data
- Compare to real data
- Revise the model
- Regularize overfitting!

Validate the model
Biased cell lysis

Biased sequencing
The Poisson distribution

- This bag contains very many small balls, 10% of which are red.

- Several experimenters are tasked with determining the percentage of red balls.

- Each of them is permitted to draw 20 balls out of the bag, without looking.
\[
\frac{3}{20} = 15\% \\
\frac{1}{20} = 5\% \\
\frac{2}{20} = 10\% \\
\frac{0}{20} = 0\%
\]
7 / 100 = 7%
10 / 100 = 10%
8 / 100 = 8%
11 / 100 = 11%
### Poisson distribution: Counting uncertainty

<table>
<thead>
<tr>
<th>expected number of red balls</th>
<th>standard deviation of number of red balls</th>
<th>relative error in estimate for the fraction of red balls</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$\sqrt{10} = 3$</td>
<td>$1 / \sqrt{10} = 31.6%$</td>
</tr>
<tr>
<td>100</td>
<td>$\sqrt{100} = 10$</td>
<td>$1 / \sqrt{100} = 10.0%$</td>
</tr>
<tr>
<td>1,000</td>
<td>$\sqrt{1,000} = 32$</td>
<td>$1 / \sqrt{1000} = 3.2%$</td>
</tr>
<tr>
<td>10,000</td>
<td>$\sqrt{10,000} = 100$</td>
<td>$1 / \sqrt{10000} = 1.0%$</td>
</tr>
</tbody>
</table>
$M = \log_2 \frac{N_1}{N_2}$

$A = \log_2 \sqrt{N_1 N_2}$

two biological replicates

treatment vs control
Two component noise model

\[ \text{var} = \mu + c \mu^2 \]

shot noise (Poisson) biological noise

Small counts
Sampling noise dominant
Improve power: deeper coverage

Large counts
Biological noise dominant
Improve power: more biol. replicates
Taylor’s law (in HITChip Atlas)

Heteroscedasticity:
Variance increases with the mean

Overdispersion:
Variance increases faster than proposed by the model

Data: HITChip Atlas
Effect of shrinkage of log fold-change estimates

Key assumption:

**Taxa with similar abundances have similar sample variances**

→ Variance can be estimated with a higher precision
Dispersion and overdispersion

- Minimum variance of count data:
  \[ \nu = \mu \quad \text{(Poisson)} \]

- Actual variance:
  \[ \nu = \mu + \alpha \mu^2 \]

- \( \alpha \) : “dispersion”
  \[ \alpha = (\mu - \nu) / \mu^2 \]
  (squared coefficient of variation of extra-Poisson variability)
The NB from a hierarchical model

- Biological sample with mean $\mu$ and variance $\nu$
- Poisson distribution with mean $q$ and variance $q$.
- Negative binomial with mean $\mu$ and variance $q+\nu$
The negative binomial distribution

A commonly used generalization of the Poisson distribution with two parameters

\[ \Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \] for \( k = 0, 1, 2, \ldots \)