Biostatistical Analysis and Visualization of the Microbiome

Brandie Wagner
Associate Professor
Department of Biostatistics and Informatics &
Department of Pediatrics
Gartner Hype Cycle

- Technology Trigger
- Peak of Inflated Expectations
- Trough of Disillusionment
- Slope of Enlightenment
- Plateau of Productivity

COLORADO SCHOOL OF PUBLIC HEALTH
Bioinformatic pipeline

• Agreement in basic steps but different tools are used.

• Evaluation of currently available bioinformatics tools is needed
  • it is unclear how widely results from different approaches can vary and thus affect downstream analyses

• Some of this work has already been initiated
  • Minimal reporting standards are being developed
Nuances of microbiome data

- High dimensional & Small sample sizes ($p > n$)
- Correlated (hierarchical phylogenetic tree)
- Counts
  - *Non-normal (non-continuous; non-negative)*
  - *Bounded by total number of sequences*
- Zero-inflated
  - *Artifact of compiling a dataset from different samples*
- Variable sequencing effort
  - *Variations in total # sequences obtained for each sample*
- Compositional data (relative abundance)
- Complex designs (time course & repeated measures)
Typical Characteristics of Microbiota Data

• Why do we care?
  – All methods have underlying assumptions which are often violated by this data
  – “All models are wrong, some are useful.”

• Ignoring these characteristics by using common statistical methods
  – can bias estimated parameters and standard errors,
  – result in spurious correlation and
  – produce misleading conclusions

• “Microbial data has been known to present unique statistical challenges that defy most conventional models” (Tang and Nicolae)
## Typical data – heirarchical classifications

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Sample1</th>
<th>Sample2</th>
</tr>
</thead>
<tbody>
<tr>
<td>/Root/Bacteria/&quot;Bacteroidetes&quot;/&quot;Bacteroidia&quot;/&quot;Bacteroidales&quot;/&quot;Porphyromonadaceae&quot;/Paludibacter</td>
<td>74</td>
<td>19</td>
</tr>
<tr>
<td>/Root/Bacteria/&quot;Firmicutes&quot;/&quot;Clostridia&quot;/Clostridiales/&quot;Ruminococcaceae&quot;/Oscillibacter</td>
<td>500</td>
<td>138</td>
</tr>
<tr>
<td>/Root/Bacteria/&quot;Firmicutes&quot;/&quot;Clostridia&quot;/Clostridiales/&quot;Lachnospiraceae&quot;/Robinsoniella</td>
<td>137</td>
<td>156</td>
</tr>
<tr>
<td>/Root/Bacteria/&quot;Actinobacteria&quot;/Actinobacteria/Bifidobacteriales/Bifidobacteriaceae/Bifidobacterium</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

Phyla

Genus
Relative Abundance

500 (25%)  250 (25%)  120 (30%)
The use of relative abundance indicates an increase in *Pseudomonas* between days 1 and 3 (left).

Simple example: 3 samples collected over time for a single subject. Community is simplified to Pseudomonas vs everything else.
The use of relative abundance indicates an increase in *Pseudomonas* between days 1 and 3 (left)

whereas if total bacterial load is considered, the absolute abundance decreases slightly (right).

In addition to community composition, there is evidence to suggest also looking at overall community size.
Simple design analysis strategy

- Ordination
  - PCA
  - PCoA

- Summarize communities using ecological parameters
  - Alpha Diversity
    - Richness
    - Evenness

- Univariate comparisons
  - Non-parametric
  - Negative binomial (DEseq - genome biology 2010)
  - Beta-Binomial (BBseq - bioinformatics 2011)
Pubmed search of recent microbiome papers – majority include ordination plots
Multivariate analyses

• Analytic approaches used in ecology provide several multivariate tools to investigate microbiota data.
  – Their application to microbiota data is not straightforward given the data characteristics described earlier, mainly the unit-sum constraint.
  – This feature complicates most multivariate methods because of its impact on covariance and correlation structures.
  – This constraint can make statistically independent components appear correlated.
  – Ignoring the compositional nature can lead to spurious correlations and misleading results.
The effect of the compositional nature of the data on PCA is demonstrated in the curved shape configuration of the PC scores. This is typical for data with closure problems where samples with high values for a component must necessarily have low values of the other components.
PCA was performed after using a simple transformation. This example is used to compare PCA plots of compositional data before and after transformation.
Simple design analysis strategy

- Ordination
  - PCA
  - PCoA

- Summarize communities using ecological parameters
  Alpha Diversity
  - Richness
  - Evenness

Beta Binomial (BBseq - bioinformatics 2011)

Negative binomial (DEseq - genome biology 2010)
Community biodiversity

• Diversity - "the variety and abundance of species in a defined unit of study"
  – combination of richness and evenness
  – there are numerous commonly used diversity indices, each of which can yield very different results;

“The term ‘species diversity’ has been defined in such various and disparate ways that it now conveys no information other than ‘something to do with community structure’; species diversity has become a nonconcept”.

Colorado School of Public Health
<table>
<thead>
<tr>
<th>Example 1: relatively even communities</th>
<th>( \alpha ) diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Taxon</strong></td>
<td>1</td>
</tr>
<tr>
<td>Community A</td>
<td>60%</td>
</tr>
<tr>
<td>Community B</td>
<td>30%</td>
</tr>
<tr>
<td><strong>Ratio B/A</strong></td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Example 2: uneven communities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Taxon</strong></td>
</tr>
<tr>
<td>Community A</td>
</tr>
<tr>
<td>Community B</td>
</tr>
<tr>
<td><strong>Ratio B/A</strong></td>
</tr>
</tbody>
</table>

Table 1. Ecological indices for two example communities illustrating varying results across different diversity measures. In both examples, community B is twice as diverse as community A (by doubling richness), whereas the communities across the two examples represent a near doubling of diversity due to evenness. The inference based on the values of the diversity indices differ dramatically depending on which measure is used.
Diversity

• The fundamental unity of diversity measures related all diversity indices under a unifying theory.

• By transforming a diversity index to a profile curve, the diversity indices can be related using a single equation, differing only in a single parameter $q$ (order of diversity):

$$\text{true diversity of order } q = q D = \left( \sum_{i=1}^{n} p_i^q \right)^{1/(1-q)},$$

where $q$ is non-negative and $p$ is some function of frequency.
Figure 2. The fundamental unity of diversity measures is a unifying theory under which diversity indices are related. This figure shows how the diversity profile curve measures change with different q values. Example measures are also provided inside the box with their corresponding q value.

- All diversity indices can be thought of as weighted components of evenness and richness.
- In the true diversity equation, the order, q, determines how much weight the measure gives to abundant versus rare species (evenness).
Diversity profile curve

Example 1: relatively even communities

<table>
<thead>
<tr>
<th>Taxon</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>60%</td>
<td>0</td>
<td>40%</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>30%</td>
<td>30%</td>
<td>20%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Example 2: uneven communities

<table>
<thead>
<tr>
<th>Taxon</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community A</td>
<td>98%</td>
<td>2%</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Community B</td>
<td>0</td>
<td>1%</td>
<td>49%</td>
<td>49%</td>
<td>1%</td>
</tr>
</tbody>
</table>
Diversity profile curve
Different communities with similar alpha

Alpha = 0.04 0.06
Hill Beta = 1.97 1.88
Beta diversity

- Partition diversity into independent components
  - evenness and richness
  - collections of libraries (beta diversity).
- Similarly, differing results are obtained across beta diversity indices
  - due to differences in weighting of the components.
  - a unifying theory called new beta.
  - The most popular similarity indices (Jaccard, Sorenson, Horn and Morisita-Horn) have been shown to be monotonic transformations of new beta.
- Although this area is still under development there is general consensus on the ideal properties of a beta diversity measure
  1) *alpha and beta components are mathematically independent, that is, the value of alpha should not affect the value of beta and vice versa*;
  2) *gamma is completely determined by alpha and beta; and*
  3) *alpha is never greater than gamma.*
Simple design analysis strategy

- Ordination
  - PCA
  - PCoA
- Summarize communities using ecological parameters
  - Alpha Diversity
    - Richness
    - Evenness
- Univariate comparisons
  - Non-parametric
  - Negative binomial (DEseq - genome biology 2010)
  - Beta-Binomial (BBseq - bioinformatics 2011)
  - interactions and multivariate structure of the bacterial communities.
    - These microbes are not present in isolation and understanding the entire community is important.
- Machine learning, networks, multinomial dirichlet
General approach to modeling

**Multivariate**

multinomial

Multinomial - dirichlet

**Univariate**

![Diagram showing relationships between distributions]

\[ f(Y = y|\pi) = \begin{cases} 
(1 - \pi) \cdot pdf & y > 0 \\
\pi + (1 - \pi) \cdot pdf & y = 0 
\end{cases} \]
Data visualization tools

Radial/ Spider plot

Pie chart

Stacked bars

Area plot

Colorado School of Public Health
Data visualization tools

Cladogram

Networks
Conclusions

- Statistical analysis of microbiota data is an active area for statistical methods research.
- A better understanding of which methods can be applied and when statistical approaches are robust to violation of assumptions is needed.
- The ultimate goal is to have a set of standard analytic tools available, supported by mathematical theory and sufficient evaluative comparisons, to facilitate microbiota research.
Newest Trends

• Define healthy community
• Modify microbiota
  – Antibiotics
  – Probiotics
• Personalized medicine
• Longitudinal data
• Multi-omics
Thanks

MiRC Collaborators:
J. Kirk Harris
Charles Robertson
Dan Frank

Clinical collaborators:
Frank Accurso
Edith Zemanick
Scott Sagel
Jordana Hoppe
Pete Mourani
Steve Abman
Sophie Fillon
Kristen Demourelle