MICROBIOME DATA & ANALYSIS

Research Group: Statistical Diversity Lab
PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW
@AmyDWillis  adwillis@uw.edu

Photo credit: T.D. Berry, Whitman lab, UW Madison
THANKS TO OUR SPONSOR

Museum of torture Cesky Krumlov

32 Reviews  #91 of 93 things to do in Cesky Krumlov  Museums, Specialty Museums

NamEsti svornosti 1, Cesky Krumlov, Czech Republic

“Avoidable”
Collection of old torture instruments. Dimly lit, dusty gave me an allergy. Emarrassingly insipid... read more

Reviewed October 28, 2016
Neceron, Mumbai, India via mobile

not worth it for this price

Reviewed March 16, 2018

Wouldn’t bother again

Reviewed April 5, 2018 via mobile

“You can avoid if you have other things to do”
Maybe this is not the review, I was funny at myself after out from the museum. I just have to... read more

Reviewed June 13, 2018 via mobile

Potential wasted

Reviewed June 13, 2018 via mobile
OUTLINE

- Why study microbes*?
- How do you study microbes?
- Directions for microbial ecology
- Opinions and research

* microbes = microscopic organisms; today's focus = bacteria/archaea
HUGE THANKS

- Folks from whom I pilfered material
  - Sarah Hird (UConn), Christian Mueller (Simons), Scott Handley (Wash U)
- My hardworking & brilliant research group, the Statistical Diversity Lab:
  - Bryan Martin (@BryanDMartin_), Pauline Trinh (@paulinetrinh), Kendrick Li (@KendrickLi4), David Clausen, Alex Paynter, Charlie Wolock, Jake Price (@Jake_in_the_Lab)
- Collaborators whose joint work I discuss
  - Sam Minot (Fred Hutch), Alon Shaiber & M Eren (U Chicago), Michael McLaren & Ben Callahan (NC State)
- The heroic organizers of #evomics2019 and Daniel McDonald
HUGE THANKS

YOU!

- For jumping on the 🚂
- For participating, contributing, correcting me throughout
WHY STUDY MICROBES?

- Microbial:host cells
  - Microbial:host genes
- Impact ecosystem/host health and function
  - Host associated: nutrient absorption, immune system, healing...
  - Environmental: biogeochemical cycling, origins of life...
- Highly localized communities; gene/organism transfer
FUN FACTS

- Hard to culture most microbes
- Microbes can be categorised into groups
  - Strains; taxa; x% similarity on some/all genes
- Every group has some concentration in every environment
  - possibly zero
- Every individual microbe has many genes
- *Microbes of the same strain may not have the same genes*
MICROBIAL QUESTIONS

- What strains are present?
- What genes are present?
- What microbes have what genes?
- How many microbes are there?
- How many different microbes are there?
Group exercise: (2 minutes)

Come up with a microbiome-related question that you might want to answer

Preferably one related to your area of interest
HOW DO YOU STUDY MICROBES?

It depends!
TECHNOLOGY

- The technology/technologies that you will use is driven by
  - The scientific question/questions that you have
  - Cost constraints
  - Resource constraints
  - Literature review, opinion of funding agencies, current trends...
TECHNOLOGY

- The technology/technologies that you will use is driven by
  - The scientific question/questions that you have
  - Cost constraints
  - Resource constraints
  - Literature review, opinion of funding agencies, current trends...
NEED TO KNOW TECHNOLOGIES

- Amplicon profiling
- Whole genome profiling
- Concentration profiling
- Many others...
**AMPLICON PROFILING**

- Amplify (PCR) & sequence a **HOMOLOGOUS MARKER** (amplicon) shared by all taxa

- e.g., 16S rRNA is bacterial marker gene

- e.g., 18S is marker gene for microbial eukaryotes
ACGTGCGTAG...

infer that it is from
16S is a commonly sequenced bacterial marker gene

Universal: Fancy protein reasons… ask Scott!

Balance: same in places; different in places

Not single copy
16S is a commonly sequenced bacterial marker gene

**Universal:** Fancy protein reasons… ask Scott!

**Balance:** same in places; different in places

---

**AMPLICON PROFILING**

---

Del Chierico et al. 2015

Slide modified with permission from Sarah Hird
WHY 16S?

- 16S has highly conserved sequences interspacing hypervariable regions

Slide modified with permission from Sarah Hird
WHY 16S?

- 16S has highly conserved sequences interspacing hypervariable regions
- Primers targeting the conserved regions allow us to pull out the hypervariable regions for sequencing
WHY 16S?

- 16S has highly conserved sequences interspacing hypervariable regions
- Primers targeting the conserved regions allow us to pull out the hypervariable regions for sequencing
- New(ly more common): full length 16S sequencing
ACGTGCGTAG…

infer that it is from
Cheap, easy, popular… historical reasons

Most (but not all) taxa amenable

Severe distortions (PCR, primers, index hopping)

Discussed later
WHOLE GENOME PROFILING

- Whole genome sequencing (WGS)
  - Shear all DNA and sequence fragments
  - *Functional potential*

- Commonly called “metagenomics”
  - metagenome = all the genomes
ACGTGCGTAG…

infer that it is from
WHOLE GENOME PROFILING

- Multilocus
- Gene content! Not just markers
- More expensive (getting cheaper)
- Sequence non-microbial genes
- Widely thought to be less distortion
CONCENTRATION PROFILING
Just believe me that there are more bacteria in some places than others, ok?
CONCENTRATION PROFILING

- Develop primers to target region
  - Region determines *what* concentration
- Amplify (qPCR) and count (calibrate) to see how many instances of that region there are
TECHNOLOGY

- The technology/technologies that you will use is driven by
  - The scientific question/questions that you have
  - **Cost constraints**
  - **Resource constraints**
  - Literature review, opinion of funding agencies, current trends...
$\$ \text{COMPARISON}$

- Costs *can vary wildly*… here are some recent ballparks:
  - $16S = $17/sample
  - $WGS = $100-200 per sample
  - 250 samples: $16S = $5k, $WGS = $25k-$50k
Costs can vary wildly... here are some recent ballparks:

- **16S** = $17/sample
- **WGS** = $100-200 per sample
- 250 samples: 16S = $5k, WGS = $25k-$50k

---

**Evaluating the Information Content of Shallow Shotgun Metagenomics**

Benjamin Hillmann, a Gabriel A. Al-Ghalith, b Robin R. Shields-Cutler, c Qiyun Zhu, d Daryl M. Gohl, e Kenneth B. Beckman, e Rob Knight, d,f,g and Dan Knights a,b,c

P Published online 2018 Nov 13. doi: 10.1128/mSystems.00069-18

PMCID: PMC6234283
PMID: 30443602
Costs can vary wildly… here are some recent ballparks:

- **16S** = $17/sample
- **WGS** = $100-200 per sample
- **250 samples**:
  - 16S = $5k
  - WGS = $25k-$50k

Other considerations: non-microbial contamination, storage, analysis…

---

Published online 2018 Nov 13. doi: [10.1128/mSystems.00069-18](https://doi.org/10.1128/mSystems.00069-18)

Evaluating the Information Content of Shallow Shotgun Metagenomics

**PMCID**: PMC6234283
**PMID**: 30443602
$\$ CO\text{MPARISON}$

- Costs can vary wildly…
  - $16S = $17/sample
  - $WGS = $100-200$ per sample
  - 250 samples: $16S = $5k, $WGS = $25k-$50k

- Other considerations: non-microbial contamination, storage, analysis…
MICROBIAL POPULATIONS

- Group exercise: (2 minutes)
  - Come up with a microbiome-related question that you might want to answer
  - Come up with a microbiome-related question that you could answer
    - How does sequencing technology influence what you can study?
ONCE YOU HAVE YOUR DATA...
ONCE YOU HAVE YOUR DATA...
ONCE YOU HAVE YOUR DATA...

- preprocessing
- cleaning
- (iterating)
- analysis
CLEANING AND PREPROCESSING

- Preprocessing
  - sometimes the same as cleaning
  - more often: processing the data into biological units

- Cleaning
  - Basic checks: determine whether sequencing went entirely/a little/not at all wrong
BIOLOGICAL UNITS

- The units that come off your sequence are usually not immediately useful:
  - … AAACTCTATCTATCTACTXTCGCGCGTACGCGTCAT…
  - … AAACTCTAGCTATCTACTTTTCGCTXGGTACGCCTCAT…
  - … AACCCCTCGCAGCACCAGCACAACACACTACCA…
  - … AACTCCGTAACAACACTACTACTACTACCACACAG…
- Idea: group data into units that simplify analysis and are biologically meaningful
BIOLOGICAL UNITS: TAXONOMIC PROFILING

- If two sequences are the same, should be grouped together
  - Very unlikely that two sequences are the same
- If two sequences are the same enough, should be grouped together
  - Idea: clustering!
BIOLOGICAL UNITS: 16S

= observed 16S sequence
= observed 16S sequence

= similar sequences

= less similar sequences
= observed 16S sequence

= similar sequences

= less similar sequences
OTU clustering: Make *operational taxonomic units* by clustering at x% similarity

- 97% is common for 16S, but a little arbitrary
BIOLOGICAL UNITS: OTUs

OTU clustering: Make operational taxonomic units by clustering at x% similarity

- Assign OTU the taxonomy of “most central” sequence
Why are A and B different?

Option 1: sequencing errors

Option 2: actually different
BIOLICAL UNITS: 16S DATA

Why are A and B different?

Option 1: sequencing errors

Option 2: actually different

These options should be distinguished!
We can estimate sequencing error rates

So can estimate how much observed sequences should vary around “true” sequence
BIOLOGICAL UNITS: ASVs

• We can estimate sequencing error rates

• So can estimate how much observed sequences should vary around “true” sequence
BIOLOGICAL UNITS: ASVs

- We can estimate sequencing error rates
- So can estimate how much observed sequences should vary around “true” sequence

A & B are from the same 16S sequence

C & D are similar but are from different 16S sequences (observed difference more than explainable by error rate)
BIOLOGICAL UNITS: ASVs

• Source sequences are called Amplicon Sequence Variants (ASVs)

• DADA2
  • ASV construction
  • Less spurious diversity

A & B are from the same 16S sequence

C & D are similar but are from different 16S sequences (observed difference more than explainable by error rate)
DADA2: ASV ALGORITHM

Open-source sequence clustering methods improve the state of the art.

http://benjineb.github.io/dada2/R/SotA.html

64 sequences

Slide modified with permission from Scott Handley
BIOLOGICAL UNITS: 16S

- Biological unit of 16S is 16S sequence
  - i.e. 16S amplicon sequence variants
- 16S sequences need to be clustered…
  - Old: into *operational clusters*
  - Modern: into *sequence variants*
BIOLOGICAL UNITS: WGS

- Many options
  - Genomes
  - Genes
  - Co-abundant genes
  - Others
Shotgun metagenomics, from sampling to analysis

Christopher Quince1,7, Alan W Walker2,7©, Jared T Simpson3,4, Nicholas J Loman5 & Nicola Segata6©

Diverse microbial communities of bacteria, archaea, viruses and single-celled eukaryotes have crucial roles in the environment and in human health. However, microbes are frequently difficult to culture in the laboratory, which can confound cataloging of members and understanding of how communities function. High-throughput sequencing technologies and a suite of computational pipelines have been combined into shotgun metagenomics methods that have transformed microbiology. Still, computational approaches to overcome the challenges that affect both assembly-based and mapping-based metagenomic profiling, particularly of high-complexity samples or environments containing organisms with limited similarity to sequenced genomes, are needed. Understanding the functions and characterizing specific strains of these communities offers biotechnological promise in therapeutic discovery and innovative ways to synthesize products using microbial factories and can pinpoint the contributions of microorganisms to planetary, animal and human health.
ASSEMBLY

- Every genome is a puzzle, break into pieces, put pieces back together
  - Different microbes contain same genes
  - Microbes of same strain can have very similar genomes
    - e.g., SNVs, same genome but missing gene/operon
  - Can’t assume equal coverage across genomes/samples
    - low coverage => can’t piece puzzle together
    - high coverage => expensive
ASSEMBLY

- Assemblers turn reads into \((\sim 10^4 - 10^6)\) contigs
- No single assembler “best”
  - Many use de Bruijn graphs: break reads into k-mers; find path
  - Inconsistent coverage is huge challenge
- Options: MEGAHIT, MetaSPAdes, others
  - MEGAHIT: “more genes that can be annotated in complex communities”
- Review article: “Use more than one!”
BINNING

- Contigs come from what genomes? How many genomes?
- Binning groups **contigs** into **genomes**
- Supervised & unsupervised
  - Choice dictated by reliability/availability of reference genomes
- Balance between automation and refinement
  - Anvi’o: helps with manual refinement
  - (More later)
Mini review

Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics

Karel Sedlar *, Kristyna Kupkova, Ivo Provaznik

Department of Biomedical Engineering, Brno University of Technology, Technicka 12, Brno, Czech Republic

COMPUTATIONAL AND STRUCTURAL BIO TECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj

UNSUPERVISED BINNING

COMPOSITION BASED

- SOMs ∅
  - Abe et al. 2008
- LikelyBin ∅
  - Kislyuk et al. 2009
- SCIMM
  - Kelley et al. 2010
- 2TBinning ∅
  - Saeed et al. 2011
- MetaWatt
  - Strous et al. 2012
- VizBin ∅
  - Laczny et al. 2015

ABUNDANCE BASED

- AbundanceBin
  - Wu et al. 2011
- Canopy
  - Lander et al. 2014
- MBBC
  - Wang et al. 2015

HYBRID

- CompostBin
  - Chatterji et al. 2008
- MetaCluster
  - Yi et al. 2012
  - Albersen et al. 2013
- CONCOCT ∅
  - Almeberg et al. 2013
- MaxBin
  - Wu et al. 2014
- GroopM ∅
  - Ineffert et al. 2014
- MetaBAT
  - Kang et al. 2015
- COCACOLA
  - Lu et al. 2016
- MyCC ∅
  - Lin and Liao 2016

Fig. 2

1. Introduction

2.2. Abundance based binning

2.4. Input data

References
Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software


Table 2

<table>
<thead>
<tr>
<th>Genome binner (% contamination)</th>
<th>Recovered genomes (% completeness)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt;50%</td>
</tr>
<tr>
<td>Gold standard</td>
<td>753</td>
</tr>
<tr>
<td>CONCOCT &lt;10%</td>
<td>275</td>
</tr>
<tr>
<td></td>
<td>267</td>
</tr>
<tr>
<td>MetaWatt 3.5 &lt;10%</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>476</td>
</tr>
<tr>
<td>MetaBAT &lt;10%</td>
<td>247</td>
</tr>
<tr>
<td></td>
<td>234</td>
</tr>
<tr>
<td>MyCC &lt;10%</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>220</td>
</tr>
<tr>
<td>MaxBin 2.0 &lt;10%</td>
<td>390</td>
</tr>
<tr>
<td></td>
<td>356</td>
</tr>
</tbody>
</table>

UNSUPERVISED BINNING
Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy


Nature Microbiology 3, 836–843 (2018) | Download Citation
ASSEMBLY-FREE WGS

- Can map reads to genomes (often not faster than assembly; high FP)
- Better idea: use specific genes (not all genes)
- **MetaPhlAn**
  - Core & marker genes
  - Great for human mb
BIOLOGICAL UNITS: GENES

- From genomes… genes!
- Adapt tools from single-genome world
  - Challenge: microbial genes mostly uncharacterized
BIOLOGICAL UNITS

- Challenges with WGS include
  - lots of genes
  - Choice of database has enormous impact

- Advantage: lots of redundancy = genes that occur together
  - Genes that consistently occur together arguably biological unit
  - CAGs = co-abundant genes; grouping of genes that are consistently present/absent together across samples
CAGs AS BIOLOGICAL UNITS

Work lead by Sam Minot (Fred Hutch)

- Co-abundant gene (CAG) construction algorithm
- No databases
- Reproducibly associated with disease

CAG grouping
BIOLOGICAL UNITS TO CLEANING

- Once you have your data sorted into biological units, you may need to do some cleaning

- Often cleaning = filtering

  - e.g., low yield

  - e.g., low quality score data

  - e.g., likely sequencing errors (sometimes low abundance)

  - e.g., contaminants (e.g., with decontam)
SUMMARY: FIRST HALF

- Microbes, their relevance, questions
- Technology to study microbes
- Processing data into meaningful units
- Next up: analysis; open problems
BREAK
ANALYSIS

- The type of data that you have affect how you will analyse
  - e.g., compositional/relative/absolute
- The questions that you have affect how you analyse
  - e.g., exploratory/confirmatory
# SCENARIO

## ABSOLUTE ABUNDANCE

<table>
<thead>
<tr>
<th>Environment</th>
<th>Microbe A</th>
<th>Microbe B</th>
<th>Microbe C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENVIRO 1</td>
<td>5</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>ENVIRO 2</td>
<td>10</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>

## PROPORTION DATA

<table>
<thead>
<tr>
<th># Observed</th>
<th>Microbe A</th>
<th>Microbe B</th>
<th>Microbe C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENVIRO 1</td>
<td>1.01 / 6</td>
<td>1 / 6</td>
<td>3.99 / 6</td>
<td>1</td>
</tr>
<tr>
<td>ENVIRO 2</td>
<td>0.99 / 6</td>
<td>0.99 / 6</td>
<td>4.02 / 6</td>
<td>1</td>
</tr>
</tbody>
</table>

Can compare across rows & columns.
### SCENARIO

<table>
<thead>
<tr>
<th>ABSOLUTE ABUNDANCE</th>
<th>MICROBE A</th>
<th>MICROBE B</th>
<th>MICROBE C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENVIRO 1</td>
<td>5</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>ENVIRO 2</td>
<td>10</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>

#### OBSERVED

<table>
<thead>
<tr>
<th># OBSERVED</th>
<th>MICROBE A</th>
<th>MICROBE B</th>
<th>MICROBE C</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENVIRO 1</td>
<td>4</td>
<td>5</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td>ENVIRO 2</td>
<td>9</td>
<td>9</td>
<td>37</td>
<td>55</td>
</tr>
</tbody>
</table>

Can compare across rows & columns
### SCENARIO

<table>
<thead>
<tr>
<th>ABSOLUTE ABUNDANCE</th>
<th>MICROBE A</th>
<th>MICROBE B</th>
<th>MICROBE C</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENVIRO 1</td>
<td>5</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>ENVIRO 2</td>
<td>10</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># OBSERVED</th>
<th>MICROBE A</th>
<th>MICROBE B</th>
<th>MICROBE C</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENVIRO 1</td>
<td>499</td>
<td>500</td>
<td>2001</td>
<td>3000</td>
</tr>
<tr>
<td>ENVIRO 2</td>
<td>250</td>
<td>251</td>
<td>1010</td>
<td>1511</td>
</tr>
</tbody>
</table>

Can compare across rows only.
DATA

- 16S and WGS data are compositional/relative
  - Can compare observed values within samples
  - Common (users/software): convert to proportions
    - ADW: Disagree, this loses information about precision
    - ADW: Good statistical methods model precision
- Implications for analysis
PARAMETERS

- Estimation: using information about the sample to estimate something about the population

\[
\begin{align*}
\star & = 1/3 \\
\bullet & = 1/3 \\
\square & = 1/3 \\
\star & = 4/15 \\
\bullet & = 2/15 \\
\square & = 2/5 \\
\triangle & = 1/5
\end{align*}
\]
PARAMETERS

- something about the population = “parameter”
  
  - Genus-level relative abundance of Streptococcus in your saliva right now
  
  - Proportion of Krumlovians with MRSA infections
  
  - Mean phylum-level diversity on the hands of #evomics19 faculty
PARAMETERS FOR COMPOSITIONAL DATA

- Diversity parameters: $\alpha$, $\beta$
  - sometimes called diversity "indices"
    - ADW: this terminology reflects a lack of understanding of statistical concept of parameters
- Relative abundance of taxon/gene
- Relative abundance within an environment ("enrichment")
DIVERSITY

- Low dimensional summaries of entire communities
  - α-diversity: one community
    - e.g., species richness, Shannon diversity
  - β-diversity: multiple communities
    - e.g., UniFrac, Bray-Curtis
- Diversity is relevant in lots of contexts... not just the microbiome!
DIVERSITY & PARAMETERS

- There are multiple choices to make when talking about diversity
  - Which taxonomic level? (strain/species/genus...)
  - Which diversity parameter?
  - Which estimate of the diversity parameter?
DIVERSITY & PARAMETERS

- There are multiple choices to make when talking about diversity
  - Which taxonomic level? (strain/species/genus...)
  - **Which diversity parameter?**
  - Which estimate of the diversity parameter?
Suppose we have $C$ groups in our environment in proportions $p_1, p_2, \ldots, p_C$

Any function of

- $p_1, p_2, \ldots, p_C$ OR
- $p_1, p_2, \ldots, p_C$ and some info about relationships amongst groups

is a valid $\alpha$-diversity parameter
Some examples of $\alpha$-diversity measures include:

- **Species richness:**
  \[ \sum_{i=1}^{C} p_i^2 \]

- **Simpson’s index:**
  \[ -\sum_{i=1}^{C} p_i \ln p_i \]

- **Shannon diversity:**
  \[ -\sum_{i=1}^{C} p_i \ln p_i \]

- **Shannon’s E:**
  \[ \frac{-\sum_{i=1}^{C} p_i \ln p_i}{\ln C} \]
YOU R C H O I C E

- Think: What difference do you want to highlight?

[Diagram showing taxonomic richness and taxonomic evenness axes]
YOUR CHOICE

taxonomic richness

species richness

taxonomic evenness
YOUR CHOICE

taxonomic richness

species richness

Shannon's E

taxonomic evenness
YOUR CHOICE

taxonomic richness

species richness

Shannon

Shannon's E

taxonomic evenness
Your choice

This is a question of parameter choice: Which parameter highlights the differences I care about?
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Richness</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Shannon</td>
<td>2.21</td>
<td>1.75</td>
</tr>
<tr>
<td>Evenness</td>
<td>0.96</td>
<td>0.90</td>
</tr>
<tr>
<td>Simpson’s</td>
<td>0.88</td>
<td>0.80</td>
</tr>
<tr>
<td>Inverse Simpson’s</td>
<td>8.17</td>
<td>4.98</td>
</tr>
</tbody>
</table>
THE PROBLEM

- In practice, we don't observe the entire community, just a sample from it

  - we don't know $C$ or $p_1, p_2, \ldots, p_C$

- We need to estimate them using the data we collected

- Research interest of ADW: how to estimate diversity
THE "CLASSICAL" APPROACH

- Substitute the observed abundances \( \hat{p}_1, \ldots, \hat{p}_c \) for the unknown, true abundances \( p_1, p_2, \ldots, p_c \) and pretend nothing happened.
  - e.g. Estimate the richness with: \( c = \# \{ i : \hat{p}_i \neq 0 \} \)
  - e.g. Estimate the Simpsons index: \( \sum_{i=1}^{c} \hat{p}_i^2 \)
ONE PROBLEM (OF MANY)

- Species richness: plug-in estimate *underestimates*
- Simpson: estimate *overestimates*
- Need new indices
- Need new estimators
HOW TO FIX

- 2 things are wrong here:
  - The bias (under/overestimation)
  - The variance (how big are the error bars — you’ll never be exactly right)
SPECIES RICHNESS

- The "species problem": how many species were missing from the sample

- Idea
  - If many rare species in sample, likely there are many missing species
  - If few rare species in sample, likely there are few missing species
  - Use data on rare species to predict # missing species

I haz C = 1
SPECIES RICHNESS

- CatchAll: mixed Poisson models
  - stable, restrictive, hard to use
- breakaway: non-mixed Poisson models
  - Higher variance, flexible models, in R
SPECIES RICHNESS ESTIMATION

- Good options
  - breakaway::breakaway(); QIIME2 breakaway plug-in
  - breakaway::chao_bunge()
  - breakaway::objective_bayes_*(
  - CatchAll

- Bad options
  - QIIME2: chao1; scikitbio…
  - R:vegan:….
- Slightly different approach:
  - Share strength across multiple samples to estimate $C$ and $p_1, p_2, \ldots, p_C$, then use network models to get variance
This idea works for estimating any diversity index (\(\alpha\) or \(\beta\)) that is a function of relative abundances.

It can also be used to estimate any diversity index that is a function of the tree.

github.com/adw96/DivNet

Coming soon...
BETA DIVERSITY

- Community 1: $p_1^{(1)}, p_2^{(1)}, \ldots, p_C^{(1)}$; Community 2: $p_1^{(2)}, p_2^{(2)}, \ldots, p_C^{(2)}$

- $\beta$-diversity parameters are usually distances between compositional vectors

- Bray-Curtis: $\beta_{BC} = 1 - \sum_{i=1}^{C} \min(p_i^{(1)}, p_i^{(2)})$

- Jaccard: $\beta_J = \%$ taxa not shared

- UniFrac: Weights phylogeny
DIVERSITY: HYPOTHESIS TESTING

- Sometimes diversity is analysed as an exploratory tool
  - e.g., ordination

- Other times you want to do inference
  - e.g., $H_0$: two communities have zero dissimilarity
  - e.g., $H_0$: communities A & B have same dissimilarity as communities A & C
HYPOTHESIS TESTING FOR DIVERSITY

- Common approach: PERMANOVA
- Critical issue: adjust for different resolution
  - Good solution = use error bars
    - `breakaway::betta(); DivNet::testDiversity`
  - (Bad solution = rarefy)
Why is estimating variance important?

Hypothesis testing

Most hypothesis tests take the form

$$\frac{\text{estimate}}{\text{standard error}} \sim N(0, 1)$$
VARIANCE AND HYPOTHESIS TESTS

- If your estimate was 1, and the (true) standard deviation is 1…

<table>
<thead>
<tr>
<th>STANDARD ERROR</th>
<th>1</th>
<th>0.5</th>
<th>0.33</th>
<th>0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-VALUE</td>
<td>0.318</td>
<td>0.046</td>
<td>0.002</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
BIAS AND DIVERSITY

- Alternative approach that I loathe: rarefaction

- Idea:
  - Discover more diversity with more sequencing
  - Can’t directly compare samples with different depths
  - Randomly throw away reads until all samples have same depth

- Better idea:
  - Statistical estimation accounts for different sequencing depths!
Alternative approach that I loathe: rarefaction

Better idea: Statistical estimation accounts for different sequencing depths!
BIAS AND DIVERSITY

- Alternative approach

I loathe: rarefaction

Idea:
- Discover more diversity with more sequencing
- Can't directly compare samples with different depths
  - Randomly throw away reads until all samples have same depth

Better idea:
- Statistical estimation accounts for different sequencing depths!
BIAS AND DIVERSITY

- Alternative approach

I loathe

Idea:
Discover more diversity with more sequencing
Can't directly compare samples with different depths
Randomly throw away reads until all samples have same depth

Better idea:
Statistical estimation accounts for different sequencing depths!
DIVERSITY

- Very useful summary of (high-dimensional) compositional data… in many settings!

- Diversity is a useful *first question*

- Drawback: Changes in diversity don’t indicate what composition(s) are changing…. 
ABUNDANCE

- How do we talk about changes in the amount of something?
  - Fraction of environments with a characteristic
  - Relative abundance: proportions only
  - Relative abundance: count data
  - Absolute abundance (same tools for DE analysis e.g., DESeq2, edgeR)
ABUNDANCE: ENRICHMENT

_Enrichment_ of genes/functions/pathways: higher presence in one group vs another group

- Need to know: anvi’o
  - Amazingly powerful tools for lots of things, including WGS
  - Fantastic workflows and tutorials for all things WGS

merenlab.org/

Image credit: anvi’o development team
ENRICHMENT

- If samples the genomes came from were observed independently, the **enrichment analysis** in anvi’o gives a hypothesis test for enrichment.

- Key points: adjusts for different numbers of genomes in each group; hypothesis testing & false discovery control.
RELATIVE ABUNDANCE: COUNTS

- Observe $W_i$ counts out of $M_i$ total counts for samples $i=1\ldots n$

- For each sample have $X_i$, information about treatment/disease/source environment

- Goal: Hypothesis test for changes in mean relative abundance with $X_i$

  - Options: CORNCOB, LEFse, ANCOM, MaAsLin, gneiss, DESeq2, ALDEx2, many others
CORNCOB
COnpositional Regression for Correlated Observations with the Beta-binomial

- Latent variable model & hypothesis testing for relative abundance
- Adjusts for different depths
- Flexible model: individual microbes correlated
- Bonus: Mean and variance ("dysbiosis") testing
CORNCOB AND DESEQ2

**corncob**
- Designed for marker gene (compositional) data
- Models relative abundance, overdispersion, and correlation parameters
- Different structure for different taxa
- Uses within-taxon correlation to model zeros

**DESeq2**
- Designed for RNAseq (different data structure)
- Tests changes in abundance
- Constrained dispersion
- Individual microbes are assumed independent
OTHER ANALYSIS APPROACHES

- Networks
  - Can be very interesting… if your data is very good

- Source tracking
  - Can be very interesting… if your data is very good

- Many, many others
"I didn't collect the data that I really wanted, so I will use what I have to try to reconstruct the data that I really wanted"

- e.g., microbial concentration (16S qPCR x 16S rel abundance)
- e.g., functional information (PiCRUST)

**Very very serious caveats! Use with extreme caution!**
CONSIDERATIONS FOR MICROBIOME SCIENCE

- Too many microbiome papers list significant associations
  - Taxon A, B, C; genes X, Y, Z are significantly higher/lower abundance in [folks with disease D]
- Observations are interesting, often unhelpful
  - Does the microbiome cause the disease, or the other way around?
- Studies involving (intelligent) interventions can help
  - e.g., paired data/longitudinal sampling
EXPERIMENTAL DESIGN

The population that you want to study may not be the population that you get to study

- Before undertaking a microbiome study, think carefully about
  - the question you want to answer,
  - the data you have access to, and
  - the questions you can answer with the data that you have access to
WHAT CAN WE DO?

- Replicate, replicate, replicate
  - Independently repeating the experiment is the gold standard for confirming a result is “real”
- Think critically
- Use plots, not p-values
WHAT ELSE CAN WE DO?

- Be honest
  - Keep all analyses that you ran, not just the final one
- Write down all of the hypotheses that you care about
  - Before doing the experiment
  - Before doing the analysis
- Your university might house a statistician; try to involve them...
  - ...in the entire process, not just calculating p-values
SUMMARY

- Technology: Taxonomic, functional, concentration profiling

- Data cleaning & preprocessing: organising data into biological units (16S = ASVs; WGS = genomes/genes/CAGs)

- Statistical estimation & hypothesis testing
  - Diversity analysis: $\alpha$, $\beta$
  - Abundance analysis: enrichment, proportions (count/proportion), abundance

- and many other things that couldn't be fit into this lecture
RESOURCES: HOW DO YOU STUDY MICROBES?

- Your university probably has a microbiology department
- Your university probably has a statistical consulting service
- STAMPS: Strategies and Techniques for Analyzing Microbial Population Structures at the MBL (Marine Biological Laboratory)
  - Apply by April 19
- The Statistical Diversity Lab @ UW

statisticaldiversitylab.com
DIRECTIONS FOR MICROBIAL ECOLOGY

- Research
  - Reproducibility
  - Calibrating sequencing results with reality
- Lab goals & wrap up
REPRODUCIBILITY

- Microbiome Quality Control Project
  - Sent same set of samples to 10+ sequencing labs, 8 bioinformatics labs
  - Compared results

- Question for you: what is the best case scenario?
REPRODUCIBILITY

- Reproducibility evaluation
  - Ideal: every lab gets identical results
  - Good enough: Not identical, but consistent ability to discriminate
  - Our qtn: Are technical replicates of Sample A more similar to each other than technical replicates from Sample B?
    - Within lab? Across lab?
  - How likely are results obtained from one lab to be reproduced in another lab?
What percentage of the time can we determine sample type based on within-lab replicates?

~95%+
What percentage of the time can we determine sample identifier based on within-lab replicates?

~90% (good labs)
What percentage of the time can we determine sample type based on another lab’s results?

~80-90%
What percentage of the time can we determine sample identifier based on another lab’s results?

~60% (good labs)
HOW MUCH WORSE IS REPRODUCIBILITY ACROSS VS WITHIN LABS?
MODELING EFFICIENCY

- Big picture goal: correct cross-lab differences
- Current step: understand how taxon abundances are distorted by sequencing process
- Approach: mock communities!
MODELING EFFICIENCY
MODELING EFFICIENCY

![Graph showing observed vs expected values]

**Taxon**
- A. vaginae
- G. vaginalis
- L. crispatus
- L. iners
- P. bivia
- S. amnii
- S. agalactia

Michael McLaren (NCSU)

Ben Callahan (NCSU)

David Clausen

146
MODELING EFFICIENCY

Michael McLaren
(NCSU)

Ben Callahan
(NCSU)

David Clausen

![Graph showing observed vs. expected efficiency for different bacterial species mixtures.](image)
MODELING EFFICIENCY
STATISTICAL DIVERSITY
LAB GOALS

- Develop statistical and computational tools for reproducible microbiome science
  - Address model misspecification
  - Make use of existing data (yours and others')
  - Model sequencing process and errors
  - Outreach: why statistical estimation and good statistical practice matters
MICROBIOME DATA & ANALYSIS

Research Group: Statistical Diversity Lab
Pl: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis    adwillis@uw.edu

Photo credit: T.D. Berry, Whitman lab, UW Madison