Introduction to metagenomic analysis

Eric A. Franzosa, Ph.D.
Galeb Abu-Ali, Ph.D.

Harvard University CFAR Workshop on Metagenomics and Transcriptomics

16 September 2014
Content

- http://huttenhower.sph.harvard.edu/content/cfar2014
Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)
What’s metagenomics?

- Total collection of **microorganisms** within a **community**
- Also **microbial community** or **microbiota**
- **Total genomic potential** of a microbial community
- Study of **uncultured microorganisms** from the environment, which can include humans or other living hosts

---

**Chemistry & Biology** October 1998, 5:R245–249

Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman¹, Michelle R Rondon¹, Sean F Brady², Jon Clardy² and Robert M Goodman¹

---

**Commensal Host-Bacterial Relationships in the Gut**

Lora V. Hooper and Jeffrey I. Gordon

---

**THE MICROFLORA AND THE PRODUCTIVITY OF LEACHED AND NON-LEACHED ALKALI SOIL**

J. E. Greaves*

Utah Agricultural Experiment Station

Received for publication July 2, 1926
Examples of metagenomic studies: Global Ocean Sampling

The Sorcerer II Expedition
Global Ocean Sampling Route

The Biodiversity of Each New Region is Different

Proteorhodopsins Vary by Region

JTC Sequencer Lab
Capacity: 240,000 sequences/day or 80 million lanes/year at 24 runs per day
The NIH Human Microbiome Project (HMP):
A comprehensive microbial survey

- **What is a “normal” human microbiome?**
- 300 healthy human subjects
- Multiple body sites
  - 15 male, 18 female
- Multiple visits
- Clinical metadata
Sequencing as a tool for microbial community analysis

Lyse cells
Extract DNA (and/or RNA)

PCR to amplify the single 16S rRNA marker gene

Classify sequence
⇒ microbe

16S amplicons

Meta’omic

Genes, Genomes, Metabolic profiling, Relative abundances, Genetic variants...

Samples

Microbes

Relative abundances

George Rice, Montana State University

Adapted from Van de Peer, 1996
What to do with your metagenome?

Basic science

- Reservoir of gene and protein functional information
- Comprehensive snapshot of microbial ecology and evolution

Translational

- Public health tool monitoring population health and epidemiology
- Diagnostic or prognostic biomarker for host disease
Composition-based analyses
Microbiome composition analyses: phylotypes and binning

**Binning**: nontrivial assignment of reads to phylotypes

**Phylotype or operational taxonomic unit (OTU)**: organisms clonal to within some tolerance (e.g. 95%); “species”

Indirect binning: BLAST etc. Relies on high similarity, reference seq.

Direct binning: analyzes seq. characteristics (GC, codons, etc.) Relies on long reads

Microbiome composition analyses: 
diversity

**Diversity**: broadly, a community’s number and distribution of organisms

Also **community composition or structure**

---

**Hamady, 2009**

### Taxonomic vs. phylogenetic

### Alpha (individual) vs. beta (shared)

### Qualitative vs. quantitative

*Schloss, 2006*
Microbiome composition analyses: alpha diversity (1-sample) scenarios

- Not diverse
- Qualitatively diverse
  - Taxonomically diverse
  - Quantitatively diverse
- Phylogenetically diverse
Microbiome composition analyses: beta diversity (2-sample) scenarios

Sample 1

Sample 2

Qualitatively diverse
Taxonomically diverse

Quantitatively diverse
Taxonomically diverse

Quantitatively diverse
Phylogenetically diverse
Which human body sites harbor the greatest microbial diversity per individual?
Which human body sites share the greatest microbial diversity among individuals?
Microbiome composition analyses: ordination

**Ordination** is the constrained projection of high-dimensional data into fewer dimensions.

Samples →

**Distance between points is Euclidean**

Microbiomes of ant castes implicate new microbial roles in the fungus-growing ant *Trachymyrmex septentrionalis*

Distance between points is a **proportional function** of their similarity

---

**PCA** or **Principal Component Analysis** guarantees the new dimensions maximize normal variation.

**NMDS** or **Nonmetric Multidimensional Scaling**, also called **PCoA** or **Principal Coordinates Analysis**, guarantees the new dimensions maximize an arbitrary similarity score (such as UniFrac beta-diversity).

Hamady, 2009
Microbiome composition analyses: ordination examples

HMP 2012

Fierer 2010
How is the gut microbiome disrupted during IBD and its treatment?

With Ramnik Xavier, Bruce Sands
How is the gut microbiome disrupted during IBD and its treatment?

With Ramnik Xavier, Bruce Sands
Meta’omic analyses

Cost per Raw Megabase of DNA Sequence

Moore’s Law

National Human Genome Research Institute

genome.gov/sequencingcosts
Typical shotgun metagenome and metatranscriptome analyses

Taxonomic Profiling

Functional Profiling

Assembly

Samples

Microbes

Relative abundances

Samples

Genes or Pathways

Relative abundances

eggNOG4.0

MetanRef

KEGG
Profiling microbial communities and ecology at species-level resolution

ARTICLE

Enterotypes of the human gut microbiome

Bacterial Ecosystems Divide People Into 3 Groups, Scientists Say

Gut

Vaginal
Are there discrete "types" of typical human microbiomes?

Dan Littman
Species prevalence vs. species abundance

![Graph showing species prevalence vs. median non-zero abundance.]

<table>
<thead>
<tr>
<th>site</th>
<th>bug</th>
<th>log10 med non0 abund</th>
<th>prevalence</th>
<th>tp 1-2 stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stool</td>
<td>s__Prevotella_copri</td>
<td>-0.3</td>
<td>0.11</td>
<td>1.00</td>
</tr>
<tr>
<td>Tongue_dorsum</td>
<td>s__Actinomyces_odontolyticus</td>
<td>-1.8</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Buccal_mucosa</td>
<td>s__Actinomyces_odontolyticus</td>
<td>-3.1</td>
<td>0.47</td>
<td>0.78</td>
</tr>
<tr>
<td>Supragingival_plaque</td>
<td>s__Actinomyces_odontolyticus</td>
<td>-3.1</td>
<td>0.38</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Typical shotgun metagenome and metatranscriptome analyses

Taxonomic Profiling

Functional Profiling

Assembly

Samples

Microbes

Relative abundances

Genes or Pathways

Relative abundances
Metagenomic analyses: gene calling and proxygenes

Extrinsic gene calling: BLAST etc. (proxygenes)

Intrinsic gene calling: ORF detection from seq.

Krause, 2006
Yooseph, 2008

Orphelia: Hoff, 2009
MetaGene: Noguchi, 2006

HMM models

Dalevi, 2009

BLAST

(a) Length distribution of protein-coding ORFs
(b) Length distribution of random ORFs
Metagenomic analyses: molecular functions and biological roles

**Orthology:**
Grouping genes by conserved sequence features
COG, KO, FIGfam...

**Structure:**
Grouping genes by similar protein domains
Pfam, TIGRfam, SMART, EC...

**Biological roles:**
Grouping genes by pathway and process involvement
GO, KEGG, MetaCyc, SEED...

Table 1: Glycoside hydrolases and carbohydrate-binding modules

<table>
<thead>
<tr>
<th>Family</th>
<th>Plan HMW name</th>
<th>Known activities</th>
<th>Tentative classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycoside hydrolase catalytic domains***</td>
<td><strong>GH1</strong> Glyco, hydro_1</td>
<td>β-Glucosidase, β-galactosidase, β-mannosidase, others</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td><strong>GH2</strong> Glyco, hydro_2</td>
<td>β-Galactosidase, β-mannosidase, others</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td><strong>GH3</strong> Glyco, hydro_3</td>
<td>β-1,4-Glucosidase, β-1,4-xylanase, β-1,3-glucanase, arabinofuranosidases, others</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td><strong>GH4</strong> Glyco, hydro_4</td>
<td>N-Acetylgalactosaminidase, N-acetylgalactosamidase, others</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td><strong>GH5</strong> Glyco, hydro_5</td>
<td>Cellulase, β-1,4-endoglucanase, β-1,3-glucanase, β-1,4-endoxylanase, others</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td><strong>GH8</strong> Glyco, hydro_8</td>
<td>Cellulase, β-1,3-glucanase, β-1,4-endoxylanase, β-1,4-endomannanase, others</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td><strong>GH9</strong> Glyco, hydro_9</td>
<td>Endoglucanase, cellulbiohydrolase, β-glucosidase</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td><strong>GH10</strong> Glyco, hydro_10</td>
<td>Xylanase, β-1,3-endoxylanase</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td><strong>GH11</strong> Glyco, hydro_11</td>
<td>Xylanase, β-1,3-endoxylanase</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td><strong>GH12</strong> Glyco, hydro_12</td>
<td>N-Acetylglucosaminidase and related enzymes</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td><strong>GH13</strong> Glyco, hydro_13</td>
<td>N-Acetylglucosaminidase and related enzymes</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td><strong>GH15</strong> Glyco, hydro_15</td>
<td>Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td><strong>GH16</strong> Glyco, hydro_16</td>
<td>β-Hexosaminidase, lectin-A-hemagglutinin</td>
<td>35</td>
</tr>
</tbody>
</table>

Turnbaugh, 2009

DeLong, 2006

Warnecke, 2007

B
Niche specialization in human microbiome function

Metabolic modules in the KEGG functional catalog enriched at one or more body habitats

- Most processes are “core”: <10% are differentially present/absent even by body site
- Contrast zero microbes meeting this threshold!
- Most processes are habitat-adapted: >66% are differentially abundant by body site

LEfSe:
LDA Effect Size
Nonparametric test for microbial and metagenomic biomarkers
http://huttenhower.sph.harvard.edu/lefse
Which *functions* of the gut microbiome are disrupted by IBD?

- Over *six times* as many microbial metabolic processes disrupted in IBD as microbes.
  - If there’s a transit strike, everyone working for the MBTA is disrupted, not everyone named Smith or Jones.
  - Phylogenetic distribution of function is *consistent* but *diffuse*

- During IBD, microbes...
  - Creating most amino acids
  - Degrading complex carbs.
  - Producing short-chain fatty acids
  - Taking up more host products
  - Dodging the immune system
  - Adhering to and invading host cells

---

Stop
- Creating most amino acids
- Degrading complex carbs.
- Producing short-chain fatty acids

Start
- Taking up more host products
- Dodging the immune system
- Adhering to and invading host cells
Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)
The two big questions...

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)
Reference genomes

• IMG alone now contains ~3,100 bacterial genomes
  – Plus ~100 archaeal, ~100 eukaryotic, and a few thousand viruses
  – About half final and half draft

• These comprise 1,222 bacterial species
  – 652 genera, 278 families, 130 orders, 66 classes, 33 phyla
  – 2,383 total clades

• And roughly 10M genes

• These genes and genomes are a tremendous resource to:
  – Identify conserved markers that can be used to infer phylogeny
  – Identify unique markers that can be used to infer taxonomy
  – Relate the microbial members of a community to their metagenomic functional potential
MetaPhlAn
Metagenomic Phylogenetic Analysis

Reference Genomes

Short Reads
MetaPhIAn
Metagenomic Phylogenetic Analysis

Reference Genomes

A X Y

X B Y

X Y C

Short Reads
MetaPhlAn overview

A is a core gene for clade Y

A is a unique marker gene for clade Y

ChocoPhlAn (offline pipeline)
- Identify all core genes for all clades
- Screen core genes for unique marker genes
- Select most representative marker genes

Available reference genomes

Unique marker genes DB

MetaPhlAn
- Blast reads against the marker genes
- Assign, count, normalize reads

Metagenome
Evaluation of MetaPhlAn accuracy

(Validation on high-complexity uniformly distributed synthetic metagenomes.)
Evaluation of MetaPhlAn performance

- >50 times faster than earlier methods
- 450 reads/sec (BLAST)
- Up to 25,000 reads/sec (bowtie2)
- Multi-threaded
- Easily parallelizable
MetaPhlAn in action

MetaPhlAn in action: strain profiling

- In practice, not all markers are present
- Individual-specific marker “barcodes”
- Often very stable over time
Plan

• Informal survey
• Metagenomics concepts & examples
• Tools for taxonomic profiling
  • MetaPhlAn
• Tools for functional profiling
  • HUMAnN
  • ShortBRED
  • PICRUSt
• Tools for testing associations
  • LEfSe
  • MaAsLin
  • CCREPE
• Resources
• Research vignette (time permitting)
The two big questions...

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)
(What we mean by “function”)
HUMAnN

HMP Unified Metabolic Analysis Network

Short reads + protein families
Translated BLAST search

Weight hits by significance
Sum over families
Adjust for sequence length

Repeat for each metagenomic or metatranscriptomic sample
HUMAnN
HMP Unified Metabolic Analysis Network

Millions of hits are collapsed into thousands of gene families (KOs) 
(still a large number)

- Map genes to KEGG pathways
- Use MinPath (Ye 2009) to find simplest pathway explanation for observed genes
- Remove pathways unlikely to be present due to low organismal abundance
- Smooth/fill gaps

Collapsing KO abundance into KEGG pathway abundance (or presence/absence) yields a smaller, more tractable feature set
HUMAnN accuracy

Validated against synthetic metagenome samples (similar to MetaPhlAn validation)

Gene family abundance and pathway presence/absence calls beat naïve best-BLAST-hit strategy
HUMAnN in action
Functional potential (DNA) vs activity (RNA)

Functional metagenomics & metatranscriptomics of 8 healthy human stool samples

HUMAnN in action
Conserved potential & variable activity

What’s there: ShortBRED

- **ShortBRED** is a tool for quantifying protein families in metagenomes
  - Short Better REad Dataset

- **Inputs:**
  - FASTA file of proteins of interest
  - Large reference database of protein sequences (FASTA or blastdb)
  - Metagenomes (FASTA/FASTQ nucleotide files)

- **Outputs:**
  - Short, unique markers for protein families of interest (FASTA)
  - Relative abundances of protein families of interest in each metagenome (text file, RPKM)

- **Compared to BLAST (or HUMAnN), this is:**
  - Faster
  - More specific
What’s there: ShortBRED algorithm

• Cluster proteins of interest into families
  – Record consensus sequences

• Identify unique and common areas among proteins
  – Compared against each other
  – Compared against reference database
  – Remove all of these

• Remaining subseqs. uniquely ID a family
  – Record these as markers for that family
What’s there: ShortBRED marker identification

Prots of interest
Reference database

Cluster into families
Identify short, common regions

True Marker
Junction Marker
Quasi Marker
What’s there: ShortBRED family quantification

- Metagenome reads
- ShortBRED markers
- Translated search for high ID hits
- Normalize relative abundances
What’s there: ShortBRED is accurate

B. Antibiotic Resistance Genes Database
Correlation – 10% of Metagenome, 500 genes

Expected Abundance

Predicted Abundance

ShortBRED: 0.95
Centroids: 0.815

Six synthetic metagenomes from GemSim, spiked with known proteins of interest:
ARDB = Antibiotic Resistance
VFDB = Virulence Factors
What’s there: ShortBRED is fast

Six synthetic metagenomes from GemSim, spiked with known proteins of interest:
ARDB = Antibiotic Resistance
VFDB = Virulence Factors
Can we infer anything about function from 16S data?

16S amplicons

Lyse cells
Extract DNA (and/or RNA)

Meta’omic

PCR to amplify the single 16S rRNA marker gene

Classify sequence → microbe

Samples

Microbes

Relative abundances

Genes, Genomes, Metabolic profiling, Relative abundances, Genetic variants...

George Rice, Montana State University

Adapted from Van de Peer, 1996
**PICRUSt:** Inferring community metagenomic potential from marker gene sequencing

With Rob Knight, Rob Beiko

One can recover *general* community function with reasonable accuracy from 16S profiles.

http://picrust.github.com
Plan

- Informal survey
- Metagenomics concepts & examples
- Tools for taxonomic profiling
  - MetaPhlAn
- Tools for functional profiling
  - HUMAnN
  - ShortBRED
  - PICRUSt
- Tools for testing associations
  - LEfSe
  - MaAsLin
  - CCREPE
- Resources
- Research vignette (time permitting)
The two big questions...

Who is there?
What are they doing?

<table>
<thead>
<tr>
<th>Sample #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clade1</td>
<td>0.40</td>
<td>0.87</td>
<td>0.43</td>
<td>0.68</td>
<td>0.47</td>
<td>0.32</td>
</tr>
<tr>
<td>Clade1</td>
<td>Bug1</td>
<td>0.40</td>
<td>0.56</td>
<td>0.07</td>
<td>0.31</td>
<td>0.42</td>
</tr>
<tr>
<td>Clade1</td>
<td>Bug2</td>
<td>0.00</td>
<td>0.30</td>
<td>0.36</td>
<td>0.37</td>
<td>0.04</td>
</tr>
<tr>
<td>Clade2</td>
<td>0.60</td>
<td>0.13</td>
<td>0.57</td>
<td>0.32</td>
<td>0.53</td>
<td>0.68</td>
</tr>
<tr>
<td>Clade2</td>
<td>Bug3</td>
<td>0.11</td>
<td>0.00</td>
<td>0.10</td>
<td>0.32</td>
<td>0.15</td>
</tr>
<tr>
<td>Clade2</td>
<td>Bug4</td>
<td>0.49</td>
<td>0.13</td>
<td>0.47</td>
<td>0.00</td>
<td>0.39</td>
</tr>
</tbody>
</table>
The two three big questions...

Who is there?
What are they doing?
What does it all mean?

<table>
<thead>
<tr>
<th>Sample #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Profession</strong></td>
<td>Student</td>
<td>Postdoc</td>
<td>Postdoc</td>
<td>Professor</td>
<td>Student</td>
<td>Student</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>Male</td>
<td>Female</td>
<td>Female</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td><strong>Site</strong></td>
<td>Oral</td>
<td>Gut</td>
<td>Oral</td>
<td>Gut</td>
<td>Oral</td>
<td>Gut</td>
</tr>
<tr>
<td><strong>Clade1</strong></td>
<td>0.40</td>
<td>0.87</td>
<td>0.43</td>
<td>0.68</td>
<td>0.47</td>
<td>0.32</td>
</tr>
<tr>
<td>**Clade1</td>
<td>Bug1**</td>
<td>0.40</td>
<td>0.56</td>
<td>0.07</td>
<td>0.31</td>
<td>0.42</td>
</tr>
<tr>
<td>**Clade1</td>
<td>Bug2**</td>
<td>0.00</td>
<td>0.30</td>
<td>0.36</td>
<td>0.37</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Clade2</strong></td>
<td>0.60</td>
<td>0.13</td>
<td>0.57</td>
<td>0.32</td>
<td>0.53</td>
<td>0.68</td>
</tr>
<tr>
<td>**Clade2</td>
<td>Bug3**</td>
<td>0.11</td>
<td>0.00</td>
<td>0.10</td>
<td>0.32</td>
<td>0.15</td>
</tr>
<tr>
<td>**Clade2</td>
<td>Bug4**</td>
<td>0.49</td>
<td>0.13</td>
<td>0.47</td>
<td>0.00</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Properties of microbiome data

- Compositional nature ($\Sigma = 1$)
  - Abundance is relative, not absolute
- High dynamic range
- Often sparse (sample dominated by a few species)
- Noisy
- Hierarchical organization

<table>
<thead>
<tr>
<th>Site</th>
<th>Oral</th>
<th>Gut</th>
<th>Oral</th>
<th>Gut</th>
<th>Oral</th>
<th>Gut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clade1</td>
<td>0.40</td>
<td>0.87</td>
<td>0.43</td>
<td>0.68</td>
<td>0.47</td>
<td>0.32</td>
</tr>
<tr>
<td>Clade1</td>
<td>Bug1</td>
<td>0.40</td>
<td>0.56</td>
<td>0.07</td>
<td>0.31</td>
<td>0.42</td>
</tr>
<tr>
<td>Clade1</td>
<td>Bug2</td>
<td>0.00</td>
<td>0.30</td>
<td>0.36</td>
<td>0.37</td>
<td>0.04</td>
</tr>
<tr>
<td>Clade2</td>
<td>0.60</td>
<td>0.13</td>
<td>0.57</td>
<td>0.32</td>
<td>0.53</td>
<td>0.68</td>
</tr>
<tr>
<td>Clade2</td>
<td>Bug3</td>
<td>0.11</td>
<td>0.00</td>
<td>0.10</td>
<td>0.32</td>
<td>0.15</td>
</tr>
<tr>
<td>Clade2</td>
<td>Bug4</td>
<td>0.49</td>
<td>0.13</td>
<td>0.47</td>
<td>0.00</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Properties of microbiome data

- General problem: correlate microbiome features with metadata (potentially controlling for other features)
- Intuitively summarize the results

<table>
<thead>
<tr>
<th>Sample #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profession</td>
<td>Student</td>
<td>Postdoc</td>
<td>Postdoc</td>
<td>Professor</td>
<td>Student</td>
<td>Student</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>Female</td>
<td>Female</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Site</td>
<td>Oral</td>
<td>Gut</td>
<td>Oral</td>
<td>Gut</td>
<td>Oral</td>
<td>Gut</td>
</tr>
<tr>
<td>Clade1</td>
<td>0.40</td>
<td>0.87</td>
<td>0.43</td>
<td>0.68</td>
<td>0.47</td>
<td>0.32</td>
</tr>
<tr>
<td>Clade1</td>
<td>Bug1</td>
<td>0.40</td>
<td>0.56</td>
<td>0.07</td>
<td>0.31</td>
<td>0.42</td>
</tr>
<tr>
<td>Clade1</td>
<td>Bug2</td>
<td>0.00</td>
<td>0.30</td>
<td>0.36</td>
<td>0.37</td>
<td>0.04</td>
</tr>
<tr>
<td>Clade2</td>
<td>0.60</td>
<td>0.13</td>
<td>0.57</td>
<td>0.32</td>
<td>0.53</td>
<td>0.68</td>
</tr>
<tr>
<td>Clade2</td>
<td>Bug3</td>
<td>0.11</td>
<td>0.00</td>
<td>0.10</td>
<td>0.32</td>
<td>0.15</td>
</tr>
<tr>
<td>Clade2</td>
<td>Bug4</td>
<td>0.49</td>
<td>0.13</td>
<td>0.47</td>
<td>0.00</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Recall that ordination is exploratory (no $p$-values for a trend, for example)
LEfSe: LDA Effect Size
Finding metagenomic biomarkers

Data
- Class 1
- Class 2
- Features 1 through n
- Samples 1 through m
- Subclasses

Step 1
- Kruskal Wallis
- Statistical consistency

Step 2
- Wilcoxon on subclasses
- Biological consistency

Step 3
- Signed LDA log-score
- Biological effect size
Example LEfSe application:
Find $O_2$-loving bugs (controlling for body site)
Superimpose enrichments on the tree of life using GraPhlAn

LEfSe Associations

Metadata Rings
MaAsLin
Multivariate Association with Linear Models

• A more general solution for finding significant metagenomic associations in metadata-rich studies

Tim Tickle
Microbiome downstream analyses: interaction network reconstruction

Given microbial relative abundance measurements over many samples, can we detect *co-occurrence and co-exclusion relationships*?

*It’s a jungle in there – microbial interactions follow patterns from classical macro-ecology.*
Relative abundance data poses a problem for correlating metagenomic features.

**Absolute (cell) counts**

No **bug1-bug2 correlation**

**Relative abundance**

Spurious **bug1-bug2 correlation** (sequencing yields rel. ab.)
CCREPE: Compositionality Corrected by REnormalization and PErmutation

Estimating a confidence interval

Estimating the null distribution
CCREPE: Compositionality Corrected by REnormalization and PErmutation

- Synthetic evaluation
- Random sample feature/tables
- No built-in correlation structure
“Microbial co-occurrence relationships in the human microbiome.”
The **three** big questions...

**Who is there?**

**What are they doing?**

**What does it all mean?**
Plan

• Informal survey
• Metagenomics concepts & examples
• Tools for taxonomic profiling
  • MetaPhlAn
• Tools for functional profiling
  • HUMAnN
  • ShortBRED
  • PICRUSt
• Tools for testing associations
  • LEfSe
  • MaAsLin
  • CCREPE
• Resources
• Research vignette (time permitting)
Using tools through Galaxy

http://huttenhower.sph.harvard.edu/galaxy
Tutorials available online

http://huttenhower.sph.harvard.edu/biobakery
(click on your tool-of-interest)
All tools are open source

http://bitbucket.org/biobakery/biobakery
The bioBakery Virtual Machine
https://bitbucket.org/biobakery/biobakery/wiki/biobakery_wiki

Ubuntu base image preloaded and configured to run all Huttenhower lab tools; one click up-and-running via Vagrant
Thank you!

Human Microbiome Project

Owen White
Joe Petrosino
George Weinstock
Karen Nelson
Lita Proctor
Erica Sodergren
Anthony Fodor
Marty Blaser
Jacques Ravel
Pat Schloss

Bruce Birren
Mark Daly
Doyle Ward
Ashlee Earl

http://huttenhower.sph.harvard.edu
Tutorial

• Informal survey
• Metagenomics concepts & examples
• Tools for taxonomic profiling
  • MetaPhlAn
• Tools for functional profiling
  • HUMAnN
  • ShortBRED
  • PICRUSt
• Tools for testing associations
  • LEfSe
  • MaAsLin
  • CCREPE
• Resources
• Research vignette (time permitting)
Plan

• Informal survey
• Metagenomics concepts & examples
• Tools for taxonomic profiling
  • MetaPhlAn
• Tools for functional profiling
  • HUMAnN
  • ShortBRED
  • PICRUSt
• Tools for testing associations
  • LEfSe
  • MaAsLin
  • CCREPE
• Resources
• Research vignette (time permitting)