Introduction to shotgun meta’omic analysis

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

Harvard CFAR Workshop on Metagenomics
17 September 2015

Huttenhower Research Group
Harvard Chan School of Public Health
Department of Biostatistics
Plan

• Shotgun meta’omics primer
• Informal survey
• Meta’omic taxonomic profiling
  • MetaPhlAn(2)
• Meta’omic functional profiling
  • Broad functional profiling with HUMAnN(2)
  • Targeted functional profiling with ShortBRED
  • Predictive functional profiling with PICRUSt
• Downstream analyses
• Resources
• Tutorials (today and tomorrow)
Sequencing as a tool for microbial community analysis (amplicon vs. shotgun)

• **16S rRNA gene**
  • Conserved across bacteria
  • (Allows PCR amplification)

• Some regions are variable
• Permits genus-level ID

**Lyse cells**

**Extract & fragment DNA**

**Sequence short DNA reads**

**AGCTAGA**
**CCGATCG**
**TTAGCAC**
**ACTAGCA**

**Assemble into contigs**

**Map reads to reference genomes**
A note on “metagenomics” vocabulary

• Amplicon sequencing
  • PCR amplify and seq. specific marker(s)
  • Often the 16S rRNA gene (for bacteria)

• Shotgun sequencing
  • Seq. short, random DNA/RNA fragments
  • Whole metagenome shotgun (WMS)
  • Whole metatranscriptome shotgun
  • Collectively, meta’omic sequencing
Sequencing as a tool for microbial community analysis (amplicon vs. shotgun)

• Where they overlap
  • Strengths
    • Quantifying taxonomic abundance
    • Ecological statistics
    • Taxon-taxon association
    • Taxon-metadata association
  • Challenges
    • Compositional (& noisy) data
    • Difficult distributions
    • Biases from sequencing
Sequencing as a tool for microbial community analysis (amplicon vs. shotgun)

• Properties of shotgun meta’omic sequencing
  • Strengths
    • Taxonomic resolution (species, strains)
    • Functional genomics (genes, transcripts)
    • Comparative genomics
  • Challenges
    • More expensive per sample
    • Data are bigger, compute more intensive
    • Need a good reference
    • Contamination
Survey: who has/plans to work with...

- 16S data?
- shotgun data?
- metatranscriptomes?
- human vs. environmental samples?
- metagenomic assemblies?
The universal meta’omics workflow

We develop computational methods in these areas

Today we’ll be focusing on this subset
Meta’omic quality control

- Trim low-quality bases from read ends
  - [Link](http://www.usadellab.org/cms/?page=trimmomatic)
- Drop short reads
- Remove contaminant sequences
  - E.g. human genome, EST database
- Remove low-complexity sequences (?)
- Enforce end-pairing (?)
  - Not required for bioBakery tools
- Integrated workflow coming soon!
  - [Link](https://bitbucket.org/biobakery/kneaddata)
Meta’omics seeks to answer two big questions...

Who is there?  
(taxonomic profiling)

What are they doing?  
(functional profiling)
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
Profiling microbial communities and ecology at species-level resolution (HMP)

Skin (nares)
Oral (plaque)
Oral (cheek)
Oral (tongue)
Gut (stool)
Vaginal (fornix)
Profiling microbial communities and ecology at species-level resolution

Enterotypes of the human gut microbiome

Bacterial Ecosystems Divide People Into 3 Groups, Scientists Say
Are there discrete “types” of typical human microbiomes?
MetaPhlAn(2)
For meta’omic taxonomic profiling
NCBI isolate genomes
Archaea  300
Bacteria  12,926
Viruses   3,565
Eukaryota 112

protein-coding genes
49.0 million total genes

Species pangenomes
7,677 containing 18.6 million gene clusters

Core genes
Marker genes

RepoPhlAn
ChocoPhlAn (http://metaref.org)
MetaPhlAn
Metagenomic Phylogenetic Analysis

Reference Genomes

Short Reads

Nicola Segata
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

Metagenome

- Blast reads against the marker genes
- Assign, count, normalize reads
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 2.0

- MetaPhlAn 1.0 focused on bacteria and archaea
- v2.0 adds support for eukaryotes and viruses
- ... along with many more bacteria and archaea
- v2.0 supports profiling at the strain level

https://bitbucket.org/biobakery/metaphlan2
MetaPhlAn2: synthetic evaluation

![Graph showing predicted vs expected relative abundance with data points for MetaPhlAn2, MetaPhlAn1, mOTUs, and Kraken, indicating high correlation with rho values.](https://bitbucket.org/biobakery/metaphlan2)
MetaPhlAn2: Results for HMP Skin

Visit number
Body site
Gender
Dataset
Staphylococcus caprae/capitis
Propionibacterium sp KPL1844
Merkel cell polyomavirus
Finegoldia magna
Campylobacter ureolyticus
Peptoniphilus rhinitidis
Propionibacterium granulosum
Staphylococcus epidermidis
Propionibacterium avidum
Malassezia globosa
Corynebacterium tuberculostearicum
Corynebacterium kroppenstedtii
Micrococcus luteus
Enhydrobacter aerosaccus
Polyomavirus HPyV7
Corynebacterium pseudogenitalium
Rothia dentocariosa
Haemophilus parainfluenzae
Corynebacterium matruchotii
Streptococcus mitis/oralis/pneumoniae
Corynebacterium accolens
Corynebacterium durum
Propionibacterium phage P101A
Propionibacterium phage P100D
Propionibacterium acnes

https://bitbucket.org/biobakery/metaphlan2
MetaPhlAn in action: strain profiling

- In practice, not all markers are present
- Individual-specific marker “barcodes”
- Often very stable over time
Meta’omics seeks to answer two big questions...

Who is there?
(taxonomic profiling)

What are they doing?
(functional profiling)
(What we mean by “function”)
Metagenomic analyses: gene calling and proxygenes

Extrinsic gene calling: BLAST etc. (proxygenes)

Intrinsic gene calling: ORF detection from seq.

Krause, 2006
Yooseph, 2008
Dalevi, 2009
HMM models
BLAST
Orphelia: Hoff, 2009
MetaGene: Noguchi, 2006

(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>Glycoside hydrolase catalytic domains</th>
<th>Plan WHP name*</th>
<th>Known activities</th>
<th>KEGG category</th>
<th>Core</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH1 Glyco_hyd_1</td>
<td>α-Glucosidase, β-glucosidase, β-mannosidase, others</td>
<td>22</td>
<td>Transcription</td>
<td>3</td>
<td>***</td>
</tr>
<tr>
<td>GH2 Glyco_hyd_2</td>
<td>C_G_β-Glucosidase, β-mannosidase, others</td>
<td>23</td>
<td>Translation</td>
<td>4</td>
<td>***</td>
</tr>
<tr>
<td>GH3 Glyco_hyd_3</td>
<td>C_G_β-1,4-glucosidase, β-1,3-glucosidase, α-L-arabinofuranosidase, others</td>
<td>69</td>
<td>Nucleotide metabolism</td>
<td>8</td>
<td>***</td>
</tr>
<tr>
<td>GH4 Glyco_hyd_4</td>
<td>C_G_β-1,4-glucosidase, β-1,3-glucosidase, β-1,6-endoglucanase, β-1,3-endoglucanase, others</td>
<td>56</td>
<td>Amino-acid metabolism</td>
<td>9</td>
<td>***</td>
</tr>
<tr>
<td>GH5 Glyco_hyd_5</td>
<td>C_G_β-1,3-galactosidase, β-1,4-endoglucanase, others</td>
<td>5</td>
<td>Biosynthesis of secondary metabolites</td>
<td>2</td>
<td>***</td>
</tr>
<tr>
<td>GH6 Glyco_hyd_6</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>1</td>
<td>Replication and repair</td>
<td>3</td>
<td>***</td>
</tr>
<tr>
<td>GH7 Glyco_hyd_7</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>9</td>
<td>Metabolism of other amino acids</td>
<td>4</td>
<td>***</td>
</tr>
<tr>
<td>GH8 Glyco_hyd_8</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>14</td>
<td>Glycan biosynthesis and metabolism</td>
<td>5</td>
<td>***</td>
</tr>
<tr>
<td>GH9 Glyco_hyd_9</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>46</td>
<td>Carbohydrate metabolism</td>
<td>6</td>
<td>***</td>
</tr>
<tr>
<td>GH10 Glyco_hyd_10</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>17</td>
<td>Lipid metabolism</td>
<td>7</td>
<td>***</td>
</tr>
<tr>
<td>GH11 Glyco_hyd_11</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>15</td>
<td>Biosynthesis of polyketides</td>
<td>8</td>
<td>***</td>
</tr>
<tr>
<td>GH12 Glyco_hyd_12</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>1</td>
<td>Cell growth and death</td>
<td>9</td>
<td>***</td>
</tr>
<tr>
<td>GH13 Glyco_hyd_13</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>1</td>
<td>Metabolism of cofactors and vitamins</td>
<td>10</td>
<td>***</td>
</tr>
<tr>
<td>GH14 Glyco_hyd_14</td>
<td>C_G_β-1,3-galactosidase, others</td>
<td>15</td>
<td>Energy metabolism</td>
<td>11</td>
<td>***</td>
</tr>
</tbody>
</table>

DeLong, 2006
Warnecke, 2007
Turnbaugh, 2009
Niche specialization in human microbiome function

- **LEfSe:** LDA Effect Size
  - Nonparametric test for microbial and metagenomic biomarkers
  - [http://huttenhower.sph.harvard.edu/lefse](http://huttenhower.sph.harvard.edu/lefse)

- 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
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...

  **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
Integrated functional meta’omics (Examining community DNA & RNA)

Increasing metagenomic abundance

<table>
<thead>
<tr>
<th>tetA</th>
<th>Methanogenesis</th>
<th>Spore coat protein H</th>
<th>Trp biosynthesis</th>
<th>Bacterial ribosome</th>
<th>groEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>RNA</td>
<td>DNA</td>
<td>RNA</td>
<td>DNA</td>
<td>DNA</td>
</tr>
<tr>
<td>10^{-1}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10^{-2}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10^{-3}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10^{-4}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10^{-5}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10^{-6}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10^{-7}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently over-expressed gene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently over-expressed pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently under-expressed gene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently under-expressed pathway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently over-expressed module</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently over-expressed gene</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Functional metagenomics & metatranscriptomics of 8 healthy human stool samples

HUMAnN(2)
For broad meta’omic functional profiling
HUMAnN
HMP Unified Metabolic Analysis Network

Short reads + protein families
Translated BLAST search

\[ c(g) = \frac{1}{|g|} \sum_r \frac{\sum_{a(r)} (1 - p_a) \Delta(a = g)}{\sum_{a(r)} 1 - p_a} \]

Weight hits by significance
Sum over families
Adjust for sequence length

Repeat for each metagenomic or metatranscriptomic sample

https://huttenhower.sph.harvard.edu/humann
Millions of hits are collapsed into thousands of gene families (still a large number)

- Map genes to 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 gene family abundance into pathway abundance (or presence/absence) yields a smaller, more tractable feature set

https://huttenhower.sph.harvard.edu/humann
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

https://huttenhower.sph.harvard.edu/humann
HUMAnN 2.0

- Avoid translated search where possible
- Speed up translated search with ORF-picking
- Stratify community-wide function by organism
- Focus on open gene family & pathway systems
- https://bitbucket.org/biobakery/humann2
Faster functional profiling by avoiding translated search
Faster functional profiling by avoiding translated search

Reference Genomes

A
Y
X
A
X
B
Y
X
B
Y
C
X
Y
C

Short Reads
Faster functional profiling by avoiding translated search

Reference Genomes

Short Reads

A Y
A X

X B Y
X B

Y C
Y Y C
HUMAnN 2.0 overview

Quality-controlled DNA or RNA reads → Taxonomic Profiling *(MetaPhlAn 2)* → List of abundant organisms → Nucleotide-level pan-genome mapping *(Bowie 2)* → Functionally annotated species pan-genomes *(ChocoPhlAn)*

https://huttenhower.sph.harvard.edu/humann2
HUMAnN2 accuracy
(1M read mock stool metagenome)

- 20 common gut bugs (even)
- 1M 100-nt reads
- Computed expected UniRef50 abundances from genome annotations
- Ran reads through HUMAnN2
- Compared expected and observed profiles

Strong agreement, even for closely related species (e.g. Bacteroides)

https://huttenhower.sph.harvard.edu/humann2
HUMAnN2 performance (1M read mock stool metagenome)

<table>
<thead>
<tr>
<th></th>
<th>MetaPhlAn2 Prescreen</th>
<th>Pangenome Search</th>
<th>Translated Search</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Flow</td>
<td>0.5 cpu-hours</td>
<td>0.7 cpu-hours</td>
<td>1.7 cpu-hours</td>
<td>2.9 cpu-hours</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(86% reads)</td>
<td>(14% reads)</td>
<td></td>
</tr>
<tr>
<td>Translated Search Only</td>
<td>NA</td>
<td>NA</td>
<td>12.1 cpu-hours</td>
<td>12.1 cpu-hours</td>
</tr>
</tbody>
</table>

https://huttenhower.sph.harvard.edu/humann2
HUMAnN2: Example combining DNA- & RNA-seq from 8 healthy gut microbiomes

- Dot = functional contribution of one species
- Ribosomal & peptidoglycan transcription correlate
- Ribosome biosyn. generally “over-transcribed”
- Peptidoglycan biosyn. generally “under-transcribed”
- Not a paradox, it’s consistent with the biology

https://huttenhower.sph.harvard.edu/humann2
HUMAnN2: Glycolytic processes performed by different species in Finns and Russians

https://huttenhower.sph.harvard.edu/humann2
ShortBRED
For targeted meta’omic functional profiling
The problem with short reads and regions of local homology among proteins

- Protein of interest
- Belongs to a family
- Local homology to unrelated families
- Short reads from unrelated families may map to protein of interest (spurious hits)

https://huttenhower.sph.harvard.edu/shortbred
ShortBRED Identify
Find unique markers for interesting prots

Prots of interest
Reference database
Cluster into families
Identify short, common regions

True Marker
Junction Marker
Quasi Marker

https://huttenhower.sph.harvard.edu/shortbred
ShortBRED Quantify

Use markers for highly specific profiling

Metagenome reads → ShortBRED markers → Translated search for high ID hits → Normalize relative abundances

Jim Kaminski

https://huttenhower.sph.harvard.edu/shortbred
ShortBRED Synthetic Evaluation (ABR genes)

Relative to mapping reads against full-length centroids, we are:
> Substantially more accurate (fewer false positives)
> Faster (reduced search space)

https://huttenhower.sph.harvard.edu/shortbred
ShortBRED: ABR in human gut metagenomes

https://huttenhower.sph.harvard.edu/shortbred
PICRUST
For predictive functional profiling
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
Meta’omics seeks to answer 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>
Meta’omics seeks to answer **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>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>0.40</td>
<td>0.56</td>
<td>0.07</td>
<td>0.31</td>
<td>0.42</td>
<td>0.27</td>
</tr>
<tr>
<td>Clade1</td>
<td>0.00</td>
<td>0.30</td>
<td>0.36</td>
<td>0.37</td>
<td>0.04</td>
<td>0.05</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>0.11</td>
<td>0.00</td>
<td>0.10</td>
<td>0.32</td>
<td>0.15</td>
<td>0.23</td>
</tr>
<tr>
<td>Clade2</td>
<td>0.49</td>
<td>0.13</td>
<td>0.47</td>
<td>0.00</td>
<td>0.39</td>
<td>0.45</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>
MaAsLin
Multivariate Association with Linear Models

- A more general solution for finding significant metagenomic associations in metadata-rich studies
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.)
Estimating a confidence interval

Estimating the null distribution

CCREPE: Compositionality Corrected by REnormalization and PErmutation

Emma Schwager
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?
Resources
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
Sahar Abubucker
Brandi Cantarel
Alyx Schubert
Mathangi Thiagarajan
Beltran Rodriguez-Mueller
Makedonka Mitreva
Yuzhen Ye
Mihai Pop
Larry Forney
Barbara Methe

Bruce Birren
Mark Daly
Doyle Ward
Ashlee Earl

http://huttenhower.sph.harvard.edu