An Introduction to Metagenomics

Paul “Joey” McMurdie, PhD
Second Genome, Inc.

Outline for Today:

- What is metagenomics?
- What methods, theoretical basis?
- Why is it useful?
- Where is it headed?
- How can I use it?

- wet lab procedures (dry workshop)
- computational protocols, practices

Schedule for today

<table>
<thead>
<tr>
<th>Sec</th>
<th>Day</th>
<th>Start</th>
<th>End</th>
<th>Topic</th>
<th>LeadInstr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mon</td>
<td>09:00</td>
<td>10:00</td>
<td>Introduction to Metagenomics. Culture independent techniques, 16S rRNA, etc. (60-75 min)</td>
<td>Joey</td>
</tr>
<tr>
<td>2</td>
<td>Mon</td>
<td>10:00</td>
<td>11:00</td>
<td>Introduction to microbiome analysis concepts -- Exploratory data analysis, Distances, PcoA, Ordination, taxa &amp; sample-level inferences (75 min)</td>
<td>Joey</td>
</tr>
<tr>
<td>3</td>
<td>Mon</td>
<td>11:00</td>
<td>11:55</td>
<td>Introduction to microbiome analysis practices: QIIME, phylseq, reproducible research (30 min)</td>
<td>Joey</td>
</tr>
<tr>
<td>---</td>
<td>Mon</td>
<td>12:00</td>
<td>14:00</td>
<td>Lunch (120min)</td>
<td>---</td>
</tr>
<tr>
<td>4</td>
<td>Mon</td>
<td>14:00</td>
<td>17:00</td>
<td>QIIME Lab (180min)</td>
<td>Daniel</td>
</tr>
<tr>
<td>---</td>
<td>Mon</td>
<td>17:00</td>
<td>19:00</td>
<td>Dinner (120min)</td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>Mon</td>
<td>19:00</td>
<td>22:00</td>
<td>phyloseq Lab (180min)</td>
<td>Joey</td>
</tr>
</tbody>
</table>

Acknowledgements

Susan Holmes  
Former postdoc advisor, mentor, co-author

Benjamin Callahan  
DADA2 first author, slides, discussions, feedback, etc.

Holmes Group  
Helpful advice and feedback

Wolfgang Huber  
Helpful advice and feedback, creator of DESeq and DESeq2

BioC and CRAN  
Support, Feedback, Distribution of phylseq and biom

Rob Knight  
QIIME, UniFrac, etc.

Huttenhower grp  
Biobakery suite, slides, etc.

Hadley Wickham  
ggplot2, reshape2, plyr R packages, Rstudio
An Introduction to Metagenomics

Outline for Today:
- What is metagenomics?
- What methods, theoretical basis?
- Why is it useful?
- Where is it headed?
- How can I use it?
  - wet lab procedures (dry workshop)
  - computational protocols, practices

An Introduction to Metagenomics

Outline for morning lecture:
- Microbiomes and metagenomics
  - What is a microbiome?
  - Why are they important?
- Methods
  - Experimental methods
  - Analysis theory
  - Analysis tools, practices

An Introduction to Metagenomics

Outline for afternoon + evening labs:

What are microbes?

Some key differences from eukaryota (e.g. humans, plants)
- Haploid genome
- Single circular chromosome, sometimes plasmids
- Genetic malleability, metabolic diversity
- Usually no nucleus (prokaryotes)
- Relatively easy interspecies gene transfer
What are microbes?

The totality of microbes in a defined environment, especially their genomes and interactions with each other and surrounding environment.

- A population of a single species/strain is a culture, extremely rare outside of lab, some infections
- A microbiome is a mixed population of different microbial species (microbial ecosystem)

A mixed community is the norm!

Why Study Microbiomes?

Environmental Science
- Critical elemental cycles (carbon, nitrogen, sulfur, iron, …)
- Pollution control, cleanup
- Ecology / Evolution (chloroplasts, mitochondria, genetic evolution, …)

Industrial Applications
- Wastewater treatment (V. cholera, algal blooms, etc.)
- Bioprospecting (novel enzymes, compounds)
- Novel biosynthesis
- Fermentations: Consortia (yogurt) / wild (kombucha, Belgian ales)

Human Health
- Protection from pathogens (e.g. Clostridium difficile)
- Absorption/Production of nutrients in the gut
- Possible Role in chronic diseases (obesity, Crohn's/IBD, other autoimmune, UTIs, periodontitis, …)
What is a microbiome?


1–10 times more microbial cells than human cells... depends on timing of your last bowel movement

Typical human microbiome < 2 kg
Some provocative oversimplifications…

Microbes can…

1. “Kill you by acute infection”
2. “Prevent same infection”
3. “Make you fat(ter)”
4. “Give you a heart attack”
5. “Give you cancer”
6. “Rescue you from cancer”

Can you guess the condition / scenario?

Microbes can make you fat(ter)…

More similar

More different


C. difficile infection

Microbes can make you fat(ter)...

- Lean (n = 10) & obese donors (n=9)
- Colonization of germ-free wild-type mice with microbiota from obese donors causes significant increase in total body fat
- Total body fat content was measured before and after a 2-week colonization
- Confirm that the ob/ob microbiome has an increased capacity for dietary energy harvest


Gut microbes promote cardiovascular disease

- Gut flora required for production of TMAO
- Supplementing diet with choline or TMAO promotes atherosclerosis (mouse)
- Gut flora suppression (Abx) inhibits dietary choline enhanced atherosclerosis


Colorectal Cancer (CRC)

- Microbes affect colonic bile pool exposure, drug metabolism, and mortality-correlated compounds
- Microbe-produced secondary bile acids are among these.
- Gut microbial metabolism may play role in beneficial or detrimental effects of certain foods


Groundwater: Chlorinated Solvents

- Chlorinated solvents can contaminate groundwater.
- Microbes can metabolize these contaminants.
- Understanding microbial interactions with chlorinated solvents is crucial for environmental remediation.

Marine picoplankton, most abundant organism on Earth?

- *Prochlorococcus* appears to be the most abundant organism on the planet
- Huge light harvesting proteins
- Its density can reach up to 100 million cells per liter
- It can be found down to a depth of 150 m in all of the intertropical belt
- Picoplankton synchronize cell division at the same time every day —> biological clock

Vertical distribution of the photosynthetic picoplankton populations determined by flow cytometry in the tropical Pacific (OLIPAC cruise, 1994).

Yellowstone National Park

- **Octopus Spring**
  - 90° to 93°C
  - Extremely low in nutrients
  - Contains abundant biomass
  - Home to “oldest” known bacteria

- **Obsidian Pool**
  - 75° – 95°C
  - High iron (II) hydrogen sulfide
  - Extensive diversity (previously unknown)


Symbiosis: sea-floor vent tube worm

- *Riftia pachyptila*
- Seafloor hydrothermal vent

End: Biological Motivation

Questions before moving on?

Exercise: How many species are present?

1  2

Confer amongst yourselves. We’ll take a poll.

The great “plate count” anomaly

- Cultivation-based cell counts are orders of magnitude lower than direct microscopic observation.
- This is because microbiologists are able to cultivate only a small minority of naturally occurring microbes.
- Our nucleic-acid derived understanding of microbial diversity has rapidly outpaced our ability to culture new microbes.

Why is microbiome research new?
Considering that…

- We have a bacterial endosymbiont in all our cells!
- Humans have always coexisted with bacteria
- We’ve known about bacteria for a few hundred years

- Historically prokaryotic biology has been focused on microbes that can be grown to large quantities/densities in the lab, especially pathogens; or can be distinguished under the microscope.
- An example of “searching where the light is”…

Why is microbiome research new?

Bias for cultivable microbes, especially pathogens

- Culture-based methods fail to detect most microbes
- Microbes are easy to miss (except pathogens)
- Most microbes are NOT pathogens (even the human-associated)

Availability of tools limited to last 3 decades

- Discovery of culture-independent techniques
- PCR, fast & cheap DNA sequencing, microarrays, etc

Discovery of Culture Independent Techniques

- 1977 rRNA as evolutionary marker - Woese & Fox PNAS
- 1985 Polymerase Chain Reaction (PCR) - K. Mullis Science
- 1985 “Universal” Primers for rRNA sequencing - N. Pace PNAS
- 1996 Large, curated rRNA database (RDP) - Maidak Nuc. Acids Res
- 2001 term “microbiome” coined by Joshua Lederberg
Discovery of *Culture Independent Techniques*

**Ribosome**

- rRNA has both catalytic and structural function.
- The small and large subunits have different lengths, 2nd-structure, 3D shape; but must work together.
- All of the catalytic activity of the ribosome is carried out by the RNA; the proteins reside on the surface and seem to stabilize the structure.

**Small subunit “16S” rRNA**

- Ubiquitous - present in all known life (viruses don’t count)
- Functionally constant translation, 2ⁿ-structure
- Evolves slowly - mutations more rare than for protein-coding genes
- Large - information for evolutionary inference
- No exchange - Limited examples of rRNA gene-sharing between organisms

---

**Metagenomics: Nucleic acid sequencing as a tool for microbial community analysis**

Single microbiome:

1. Break all cells, extract all DNA (gDNA)
2. PCR-amplify a *universal gene* from gDNA
3. DNA sequencing from pool of amplified genes
4. Cluster sequences according to species
5. Count each species and make a tree

---


**Metagenomics: Nucleic acid sequencing as a tool for microbial community analysis**

Many microbiomes in parallel:

1. Break all cells, extract all DNA (gDNA)
2. PCR-amplify a universal gene from gDNA using bar-coded primers, diff code for each sample
3. DNA sequencing from pool of amplified genes
   4a. "De-multiplex" barcode, ID source sample
4. Cluster sequences according to species
5. Count each species and make a tree

---

**Culture Independent Techniques:**

- Universal Gene census
- Shotgun Metagenome Sequencing
- Transcriptomics (shotgun mRNA)
- Proteomics (protein fragments)
- Metabolomics (excreted chemicals)

---

**Nucleic acid sequencing as a tool for microbial community analysis**

Lyse cells
Extract DNA (and/or RNA)

PCR to amplify a single marker gene, e.g. 16S rRNA

DNA sequeencer

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

Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)
Sequencing as a tool for microbial community analysis

Who’s there?
(Taxonomic profiling)

What are they doing?
(Functional profiling)

What does it all mean?
(Statistical analysis)

A Summary of Meta’omics

- Piles of short DNA/RNA reads from >1 organism
- You can...
  - Ecologically profile them
  - Taxonomically or phylogenetically profile them
  - Functionally profile them – gene/pathway catalogs
  - Assemble them
- Prior knowledge is helpful
- Caution: Correlation ≠ Causation
  - Most ‘omics results require lab confirmation

An Introduction to Metagenomics

Outline for morning lecture:

- Microbiomes and metagenomics
- What is a microbiome?
- Why are they important?
- Methods
  - Experimental methods
  - Analysis theory
  - Analysis tools, practices
Introduction to Microbiome / Metagenome Analysis Concepts

- **Sequence Processing (OTUs)**
  - Denoising
  - Chimera detection
  - Construction of sequence clusters (OTUs)

- **Comparing microbiomes**
  - Distances, Diversity
  - Exploratory Data Analysis
  - Ordination Methods
  - hierarchical dendrogram
  - extract patterns from a plot
  - clusters - gap statistic
  - gradient - regression, modeling, etc.

- **Identifying important microbes/taxa**
  - projected points, coinertia (plots)
  - inferential testing
  - modeling

- **Comparing microbiomes**
  - Distances, Diversity
  - Exploratory Data Analysis
  - Ordination Methods
  - hierarchical dendrogram
  - extract patterns from a plot
  - clusters - gap statistic
  - gradient - regression, modeling, etc.

- **Identifying important microbes/taxa**
  - projected points, coinertia (plots)
  - inferential testing
  - modeling
OTUs - Operational Taxonomic Unit

PCR to amplify a single marker gene, e.g. 16S rRNA

Amplicons

Lyse cells
Extract DNA

DNA sequencing

“OTU Clustering”

OTUs - Operational Taxonomic Unit

Sample Inference from Noisy Reads

Lyse cells
Extract DNA

DNA sequencing

“OTU Clustering”

Sample sequences
amplicon reads

Errors

Make OTUs

sample sequences
amplicon reads
OTUs

Errors

Make OTUs

DADA2

Slide adapted from slide by Curtis Huttenhower; not necessarily with permission ©:-)

Slide graciously provided by Benjamin Callahan, not necessarily with permission ©:-)
Sample Inference from Noisy Reads

sample sequences → amplicon reads → OTUs

Errors → DADA2 → Make OTUs

(OTUs are not strains)

OTUs: Lump similar sequences together
DADA2: Statistically infer the sample sequences

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

The true shape of an error cloud

DADA2: Error Model

- Counts, unique sequence
- NOT AN ERROR
- Effective Hamming Distance (number of substitutions from presumed parent)

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Initial guess: one real sequence + errors

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Infer initial error model under this assumption.

Pr(i → j) =

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)
Reject unlikely error under model. Recruit errors.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.97</td>
<td>10^{-2}</td>
<td>10^{-3}</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>C</td>
<td>10^{-2}</td>
<td>0.97</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>G</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
<td>0.97</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>T</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Update the model.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.997</td>
<td>10^{-3}</td>
<td>10^{-2}</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>C</td>
<td>10^{-3}</td>
<td>0.997</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>G</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
<td>0.997</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>T</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Reject more sequences under new model.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.998</td>
<td>10^{-4}</td>
<td>2x10^{-3}</td>
<td>2x10^{-4}</td>
</tr>
<tr>
<td>C</td>
<td>2x10^{-4}</td>
<td>0.999</td>
<td>3x10^{-3}</td>
<td>3x10^{-4}</td>
</tr>
<tr>
<td>G</td>
<td>3x10^{-3}</td>
<td>3x10^{-4}</td>
<td>0.999</td>
<td>3x10^{-4}</td>
</tr>
<tr>
<td>T</td>
<td>3x10^{-4}</td>
<td>3x10^{-4}</td>
<td>3x10^{-4}</td>
<td>0.998</td>
</tr>
</tbody>
</table>
**Convergence:** all errors are plausible

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.998</td>
<td>1x10^-4</td>
<td>2x10^-4</td>
<td>2x10^-4</td>
</tr>
<tr>
<td>C</td>
<td>5x10^-3</td>
<td>0.999</td>
<td>3x10^-1</td>
<td>1x10^-1</td>
</tr>
<tr>
<td>G</td>
<td>3x10^-1</td>
<td>3x10^-6</td>
<td>0.999</td>
<td>6x10^-1</td>
</tr>
<tr>
<td>T</td>
<td>2x10^-6</td>
<td>2x10^-6</td>
<td>3x10^-1</td>
<td>9.998</td>
</tr>
</tbody>
</table>

**DADA2 Advantages: Real Data**

*Lactobacillus crispatus* sampled from vaginal microbiome 42 pregnant women

**DADA2 Advantages**

**Analytical**
- Single nucleotide resolution
  - genotypes/strains instead of 97% OTUs
- Lower false positive rate
  - Better error model, easier to ID chimeras

**Computational**
- Linear scaling of computational costs
  - Exact sequences are inherently comparable, so samples can be processed independently.

**DADA2: Why is this possible?**

Uses more of the information than traditional OTU clustering

<table>
<thead>
<tr>
<th></th>
<th>DADA2</th>
<th>OTUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundance</td>
<td>✓</td>
<td>Ranks only</td>
</tr>
<tr>
<td>Sequence Differences</td>
<td>✓</td>
<td>Count only</td>
</tr>
<tr>
<td>Quality</td>
<td>✓</td>
<td>No</td>
</tr>
<tr>
<td>Error Model</td>
<td>✓</td>
<td>No</td>
</tr>
</tbody>
</table>
DADA2

Divisive Amplicon Denoising Algorithm - ver.2

DADA2: High resolution sample inference from amplicon data

Benjamin J Callahan¹,², Paul J McMurdie², Michael J Rosen¹, Andrew W Han², Amy Jo Johnson² and Susan P Holmes¹
¹Department of Statistics, Stanford University
²Second Genome, South San Francisco, CA
³Department of Applied Physics, Stanford University
*Corresponding Author: benjamin.j.callahan@gmail.com

That said, we are going to use OTUs!

sample sequences
amplicon sequences
OTUs

Errors
Make OTUs

Diversity

Diversity of diversity
(diversity of greek letters used in ecology)

• \(\alpha\) – diversity within a community, \# of species
• \(\beta\) – diversity between communities (differentiation), species identity is taken into account
• \(\gamma\) – (global) diversity of the site, \(\gamma = \alpha \times \beta\), but only this simple if \(\alpha\) and \(\beta\) are independent
• Probably others, but \(\alpha\) and \(\beta\) are most common
Beta-Diversity

- Microbial ecologists typically use beta diversity as a broad umbrella term that can refer to any of several indices related to compositional differences (Differences in species content between samples)
- For some reason this is contentious, and there appears to be ongoing (and pointless?) argument over the possible definitions
- For our purposes, and microbiome research, when you hear “beta-diversity”, you can probably think: “Diversity of species composition”

http://en.wikipedia.org/wiki/Beta_diversity

Beta-Diversity

- Microbial ecologists typically use beta diversity as a broad umbrella term that can refer to any of several indices related to compositional differences (Differences in species content between samples)
- For some reason this is contentious, and there appears to be ongoing (and pointless?) argument over the possible definitions
- For our purposes, and microbiome research, when you hear “beta-diversity”, you can probably think: “Diversity of species composition”

Community Distance

Communities are a vector of abundances:

\[ \mathbf{x} = \{x_1, x_2, x_3, \ldots\} \]

- \( E.\ coli: \bullet\bullet\bullet \)
- \( P.\ fluorescens: \bullet \)
- \( B.\ subtilis: \bullet \)
- \( P.\ acnes: \)
- \( D.\ radiodurans: \)
- \( H.\ pylori: \bullet\bullet\bullet\bullet\bullet\bullet \)
- \( L.\ crispatus: \)

\[ \mathbf{x} = \{3,1,1,0,0,7,0\} \]
Community Distance Properties

- Range from 0 to 1
- Distance to self is 0
- If no shared taxa, distance is 1
- Triangle inequality (metric)
- Joint absences do not affect distance (biology)
- Independent of absolute counts (metagenomics)

The Distance Spectrum

<table>
<thead>
<tr>
<th>Categorical</th>
<th>Phylogenetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence/Absence</td>
<td>Jaccard</td>
</tr>
<tr>
<td>Quantitative Abundance</td>
<td>Bray-Curtis</td>
</tr>
</tbody>
</table>

Jaccard

\[ \text{Dist}(A, B) = 1 - \frac{(x_A > 0) \cap (x_B > 0)}{(x_A > 0) \cup (x_B > 0)} \]

Intuition: Fraction of shared types unique to one of the communities
Bray-Curtis

\[
\text{Dist}(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} = +
\]

**Intuition:** City block distance. Sum of absolute differences over total abundance.

Unifrac

\[
\text{Dist}(x, y) = + + + +
\]

**Intuition:** Fraction of shared tree unique to one of the communities.
**Weighted Unifrac**

**Intuition**: The cost of turning one distribution into the other; where the cost is the amount of “dirt” moved times the distance by which it is moved.
The Distance Spectrum

**Categorical**
- Jaccard
- Unifrac

**Phylogenetic**
- Bray-Curtis
- Weighted Unifrac

**phyloseq distances**
- manhattan
- euclidean
- canberra
- bray
- kulczynski
- jaccard
- gower
- altGower
- morisita-horn
- mountford
- raup
- binomial
- chao
- cao
- jensen-shannon
- unifrac
- weighted-unifrac

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-(}
Ordination Methods

Project high-dimensional data onto lower dimensions

\[ \begin{array}{ccccccccc}
\text{P taxa} & 0,1,5,1,0,1,2,1,0,0,9, & 7,2,0,0,0,0,0,1,0,0, & 0,0,0,0,0,0,0,1,0,0, & 0,0,0,1,0,1,2,0,0,0,5, & 0,1,0,2,0,0,1,0,0,4, & 0,0,0,1,9,1,2,5,2,0,1, & 0,0,0,0,0,1,2,1,8,0,0, & 0,0,0,0,9,4,0,0,0,0,1, \\
\text{N samples} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\text{P-dimensions} & \text{2-dimensions} & \text{Axial 2} [14.4\%] & \text{Axis 1} [39.5\%]
\end{array} \]

\text{Intuition:}
Each PC axis is projection that maximizes the area of the shadow
Equivalently - max\{sum of square of distances between points\}
Goal: “See” as much variation as possible

Multi-dimensional Scaling

Why MDS? It works with any distance!

\text{Input distance matrix can by Bray-Curtis, Unifrac, …}

MDS Details

Given distances between each observation (sample), MDS finds the closest approximation of that in lower dimensional Euclidean space.

- Algorithm starts from \( D \) inter-point distances:
  - Center the rows and columns of the distance matrix:
    \[ S = -1/2 \ H \ D^{(2)} \ H \]
  - Compute SVD by diagonalizing \( S \): \( S = U \Lambda U^T \)
  - Extract Euclidean representations: \( X = U \Lambda^{1/2} \)
- The relative values of diagonal elements of \( \Lambda \) gives the proportion of variability explained by each of the axes.
- The valued of \( \Lambda \) should always be looked at in deciding how many dimensions to retain

NMDS is similar, but minimizes a different function
(difference in distance ranks)
Exploratory Data Analysis
"Unsupervised Learning"
"Ordination Methods"

What we “learn” depends on the data.

• How many axes are probably useful?
• Are their clusters? How many?
• Are their gradients?
• Are the patterns consistent with covariates
• (e.g. sample observations)
• How might we test this?

Best Practices
• Looking for patterns (the “I-test”)
• Always look at scree plot
• Biplot (if legible)
• Use multiple distances
  • For which D is pattern strongest?
  • phyloseq (and R/Rmd) make this easy!

Technique:
Gap Statistic
• Are their gradients?
• Are they explained by one or more sample covariates?

Technique:
PC regression (statistics’ PCR)

• Are the patterns consistent with covariates?

Technique:
Permutational Multivariate ANOVA
vegan::adonis()

End:
Introduction to Microbiome / Metagenome Analysis Concepts

Questions?
Introduction to Microbiome / Metagenome Analysis Tools and Practices

1. Probably-not-comprehensive summary of metagenomic tools
2. Short sermon on the virtues of reproducible analysis
3. Introduction to phyloseq & send-off this afternoon’s lab

**I6S rRNA Databases**

- GreenGenes - [http://greengenes.secondgenome.com](http://greengenes.secondgenome.com)
- Silva - [www.arb-silva.de](http://www.arb-silva.de)
- Ribosomal Database Project (RDP) - [https://rdp.cme.msu.edu](https://rdp.cme.msu.edu)

- ~100Ks - millions of unique 16S rRNA genes
- Curated taxonomy
- Classification tools (e.g. RDP classifier, ARB, etc.)
(16S rRNA) Amplicon Sequence Processing Tools:

- QIIME (Soon ‘Qiita’) - http://qiime.org/
- mothur - www.mothur.org/
- usearch - www.drive5.com/usearch
- DADA2 - https://github.com/benjjneb/dada2

Afternoon will be spent using QIIME
Daniel has much more to say about it…

How to find biology in your meta’ome

- Looking for ecology?
  - Diversity metrics, k-mer analysis, curve fitting

- Looking for specific bugs?
  - Assembly: +novelty, -difficulty
  - Mapping: +speed/ease, -novelty

- Looking for specific processes?
  - Intrinsic annotation: +novelty, -difficulty
  - Extrinsic annotation: +sensitivity, -novelty

- Looking for variants?
  - Clustering: +specificity, -difficulty
  - Mapping: +sensitivity, -novelty

- What else?

---

Goal: expose strain level features

The picture using shotgun metagenomics and MetaPhlAn

Next step: strain-level profiling

(i) Identify
(ii) Track (e.g. across samples)
(iii) Characterize (genomically)

MetaPhlAn: Taxonomic profiling using unique marker genes

- Gene X
  - Is a core gene for clade Y
  - Is a unique marker gene for clade Y

---

Slide graciously provided by Curtis Huttenhower, not necessarily with permission O:-)
Typical shotgun metagenome and metatranscriptome analyses

Taxonomic Profiling

Functional Profiling

Assembly

Samples

Genes or Pathways

Relative abundances

Microbiome meta’omic analyses: assembly

MicroPhlAn2: Trans-kingdom profiling

http://huttenhower.sph.harvard.edu/metaphlan
Reproducible analysis of microbiome / metagenome data

- Why make the effort?
- What if I don't want someone else reproducing my analysis?
- What if I don't know how?
- Isn't it enough to provide a cursory description in the methods section with a light sprinkling of literature citations?

Illustrative example favoring reproducible analysis: “Enterotypes of the human genome”

MDS on supported distance metrics: enterotype data

http://joey711.github.io/phyloseq/gap-statistic.html

# Main title
This is an [R Markdown](my.link.com) document of my recent analysis.

## Subsection: some code
Here is some import code, etc.
```{r}
library("phyloseq")
library("ggplot2")
physeq = import_biom("datafile.biom")
plot_richness(physeq)
```

Complete HTML5

knitr::knit2html()

markdown (code + console) + figures

microbiome data

Reproducible analysis workflow with R-markdown
Reproducible analysis workflow with R-markdown

Key Packages:
- vegan
- ape
- distory
- phangorn
- picante
- metagenomeSeq

phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

Introduction

Import

Preprocessing

Filtering taxa
- filter_taxa
- filterfun_sample
- genefilter_sample

Pruning samples
- prune_taxa
- prune_samples

Subsetting
- subset_taxa
- subset_samples

Transforming sample counts
- transform_sample_counts

Import biom
- import_biom
- import_markhor

Import pyrotagger
- import_pyrotagger

Import qiime
- import_qiime

Import RDP
- import_RDP

Input

Sample data
- OTU cluster output

Output

Summary / Exploratory Graphics
- plot_network
- plot_heatmap
- plot_ordination

Direct Plots
- plot_bar
- plot_richness
- plot_tree

Inference, Testing
- bootstrap
- permutation tests
- regression
- discriminant analysis
- multiple testing
- gap statistic
- clustering
- procrustes

Summary

OTU Abundance
- otu_table

Sample Variables
- sample_data

Taxonomy Table
- tax_table

Phylogenetic Tree
- phylo

Reference Sequences
- refseq

Accessors:
- get_taxa
- get_variable
- get_samples

Processors:
- filter_taxa
- merge_phyloseq
- merge_samples
- merge_taxa
- prune_taxa
- prune_samples
- subset_taxa
- subset_samples
- transform_sample_counts

Data structure & API

phyloseq constructor:
- phyloseq

Input

sample_data as sample_data
otu_table as otu_table
sample_data as sample_data
sample_data as sample_data

Direct Plots

plot_network
plot_heatmap
plot_ordination

Summary / Exploratory Graphics

plot_bar
plot_richness
plot_tree

Inference, Testing

bootstrap
permutation tests
regression
discriminant analysis
multiple testing
gap statistic
clustering
procrustes

Workflow

Import biom
import_biom
import_markhor
import_pyrotagger
import_qiime
import_RDP

Preprocessing

Filtering taxa
- filter_taxa
- filterfun_sample
- genefilter_sample

Pruning samples
- prune_taxa
- prune_samples

Subsetting
- subset_taxa
- subset_samples

Transforming sample counts
- transform_sample_counts

distance
ordinate

Summary / Exploratory Graphics
- plot_network
- plot_heatmap
- plot_ordination

Inference, Testing
- bootstrap
- permutation tests
- regression
discriminant analysis
multiple testing
gap statistic
clustering
procrustes
phyloseq

plot_ordination()
plot_network()
plot_bar()

plot_heatmap()
plot_tree()
plot_richness()

MDS/PCoA on weighted-UniFrac distance, GlobalPatterns

ordu = ordinate(GP1, "PCoA", "unifrac", weighted = TRUE)
plot_ordination(GP1, ordu, color = "SampleType", shape = "human")

phyloseq

plot_ordination()

plot_heatmap()

plot_network; Enterotype data, bray color, unifrac

plot_bar; Bacteroidetes

plot_richness()

Ordination on bray-curits dist: Global Patterns data

plot_ordination() samples-only

joey711.github.io/phyloseq/plot_ordination-examples.html

joey711.github.io/phyloseq/distance

supported ordination methods

samples-only

taxa-only

split

graphics

graphics

graphics
plot_network

plot_ordination()

plot_bar()

plot_network()

plot_ordination()

plot_tree()

plot_heatmap()

plot_richness()

plot_network(enterotype, dist.fun = "bray", max.dist = 0.3)

plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
line_weight = 0.4, label = NULL)

g <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)

plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
line_weight = 0.4, label = NULL)

g <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)

plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
line_weight = 0.4, label = NULL)

g <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)

plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
line_weight = 0.4, label = NULL)

g <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)

plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
line_weight = 0.4, label = NULL)

g <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)

plot_tree()

plot_tree()
Schedule for today

<table>
<thead>
<tr>
<th>Sec</th>
<th>Day</th>
<th>Start</th>
<th>End</th>
<th>Topic</th>
<th>Lead Instr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mon</td>
<td>09:00</td>
<td>10:00</td>
<td>Introduction to Metagenomics. Culture independent techniques, 16S rRNA, etc. (60-75 min)</td>
<td>Joey</td>
</tr>
<tr>
<td>2</td>
<td>Mon</td>
<td>10:00</td>
<td>11:00</td>
<td>Introduction to microbiome analysis concepts -- Exploratory data analysis, Distances, PCoA, Ordination, taxa &amp; sample-level inferences (75 min)</td>
<td>Joey</td>
</tr>
<tr>
<td>3</td>
<td>Mon</td>
<td>11:00</td>
<td>11:59</td>
<td>Introduction to microbiome analysis practices: QIIME, phyloseq, reproducible research (30 min)</td>
<td>Joey</td>
</tr>
<tr>
<td></td>
<td>Mon</td>
<td>12:00</td>
<td>14:00</td>
<td>Lunch (120min)</td>
<td>---</td>
</tr>
<tr>
<td>4</td>
<td>Mon</td>
<td>14:00</td>
<td>17:00</td>
<td>QIIME Lab (180min)</td>
<td>Daniel</td>
</tr>
<tr>
<td></td>
<td>Mon</td>
<td>17:00</td>
<td>19:00</td>
<td>Dinner (120min)</td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>Mon</td>
<td>19:00</td>
<td>22:00</td>
<td>phyloseq Lab (180min)</td>
<td>Joey</td>
</tr>
</tbody>
</table>