Microbiome 16S Analysis: A Quick-Start Guide

Amanda Birmingham
Center for Computational Biology & Bioinformatics
University of California at San Diego
Agenda:

- Rapid introduction to 16S microbiome studies
- Summary of analysis steps and software tools
- Minimal instruction on compute environment
- Practicum on 16S analysis with QIIME 2
  - Alternating lecture and tutorial
    - Goal: Any topic I’ve lectured about, you will get to test live (even if we don’t finish all topics)

- Notice an emphasis on speed 😊
  - Red dot on slide means I won’t be covering it in depth
  - Considerable additional material is described in the Supplemental Slides
Marker Gene Metagenomics Basics

- Approach: PCR amplicons of a conserved constitutive gene (a "marker gene") to determine identity and abundance of microbes present
  - Usually the “conserved constitutive gene” of choice is an rRNA
    - For bacteria/archaea, usually the 16S—small sub-unit (SSU) of ribosome
      - Excludes eukaryotic DNA as eukaryotes’ SSU is 18S
  - 16S rRNA widely conserved across bacteria/archaea (so shared primer sites)
    - But also has 9 hypervariable regions
      - Can be used to id different “species” and build phylogenetic trees
  - Can’t study fungi with 16S (they don’t have it) nor 18S (evolves too slowly)
    - Internal transcribed spacer (ITS) is standard fungi marker gene; 28S also used
When to Use Marker Gene Metagenomics

- When your sample is MOSTLY made up of host DNA, e.g. tumor samples
  - Shotgun reads will also be mostly host DNA, with few left over for the microbes
  - Use 16S rRNA instead, as the primers exclude eukaryotic DNA from amplification

- When you’re cheap 😊

When to Use Marker Gene Metagenomics

- When your sample is MOSTLY made up of host DNA, e.g. tumor samples
  - Shotgun reads will also be mostly host DNA, with few left over for the microbes
  - Use 16S rRNA instead, as the primers exclude eukaryotic DNA from amplification
- When you’re cheap 😊
- The good news:
  - Target gene studies are slightly cheaper to prep and sequence than shotgun ones
  - Analysis software is mature, and many studies can be analyzed on a laptop
  - Known taxa can be detected with very low (100s of reads) sequence depth
When to Use Marker Gene Metagenomics

- When your sample is MOSTLY made up of host DNA, e.g. tumor samples
  - Shotgun reads will also be mostly host DNA, with few left over for the microbes
  - Use 16S rRNA instead, as the primers exclude eukaryotic DNA from amplification
- When you’re cheap 😊
- The good news:
  - Target gene studies are slightly cheaper to prep and sequence than shotgun ones
  - Analysis software is mature, and many studies can be analyzed on a laptop
  - Known taxa can be detected with very low (100s of reads) sequence depth
- The bad news
  - No target gene distinguishes all microbes well
    - And, for a given gene, no primer pair distinguishes all microbes well
  - No other genome information (outside target gene) is captured
Common Issues in Marker Gene Studies

• Neglecting metadata
  ◦ Analysis can not test for effects of, or discard bias from, features you didn’t record!

• Picking novel 16S primers—not all created equal
  ◦ Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes

• Not taking precautions to support amplicon sequencing
  ◦ Some Illumina machines require high PhiX, low cluster density
Marker Gene Analysis Workflow

- Most critical analysis choices:
  - Whether to do OTU picking or error correction
  - What $\alpha$ and $\beta$ metrics to pick
    - Some are phylogenetically aware, some aren’t
Software Selection

• Google “16S analysis <program name>”; main contenders are

• Mothur
  ◦ Name: not an acronym (play on DOTUR, SONS)
  ◦ Philosophy: single piece of re-implemented software
  ◦ Top pro: easy to install
  ◦ Top con: re-implementation could be buggy
  ◦ Language: C++
  ◦ Model: open-source
  ◦ License: GPL
  ◦ Published: 2009
  ◦ Developed: at Umichigan
Software Selection

- Google “16S analysis <program name>”; main contenders are
  - QIIME
    - Name: Quantitative Insights Into Microbial Ecology
    - Philosophy: wrapper of best-in-class software
    - Top pro: extremely flexible
    - Top con: QIIME 2 not yet feature-complete
    - Language: python (wrapper)
    - Model: open-source
    - License: mixed
    - Published: 2010
    - Developed: At UCSD, NAU
Software Selection

- Google “16S analysis <program name>”
  - Main contenders are Mothur and QIIME
  - Both widely used
  - Both pride themselves on quality of support

- Will discuss only QIIME in this tutorial

- QIIME 1 vs QIIME 2
  - QIIME 1 won’t be supported after end of 2017
  - QIIME 2 not yet feature-complete
    - But already much easier to use!
  - This tutorial uses QIIME 2 only

- I’m not a QIIME 2 developer
  - I’m not taking credit for this tool, just demonstrating it!
Getting the Software & Data

• Not covered in this tutorial, for sake of time

• QIIME 2 is very easy to install with the Conda environment- and package-manager
  ◦ Conda is also very easy to install—either Miniconda or Anaconda versions
  ◦ Once Conda installed, QIIME 2 install is one line, e.g. for linux 64-bit, conda create -n qiime2-2017.6 --file https://data.qiime2.org/distro/core/qiime2-2017.6-conda-linux-64.txt

• Data acquisition method is project-specific
  ◦ Public data can often be pulled down from internet with wget or curl commands
  ◦ Sequencing data from a core usually available by ftp
  ◦ If all else fails, use a flash drive 😊
Get Ready To Practice!

• “Why are you making me type?!”
  ◦ QIIME 2 has a GUI—but still very under development
  ◦ QIIME 2 command-line interface is easy to install and ready to run
  ◦ Emphasize typing rather than copy/pasting commands because in your real analyses, you will need to type in the appropriate commands for your data
    ▪ Need to make realistic typing mistakes now so you know how to correct them later!
Get Ready To Practice!

• “Why are you making me type?!”
  ◦ QIIME 2 has a GUI—but still very under development
  ◦ QIIME 2 command-line interface is easy to install and ready to run
  ◦ Typing rather than copy/pasting commands because in your real analyses, you will need to type in the appropriate commands for your data
    ▪ Need to make realistic typing mistakes now so you know how to correct them later!

• Will be working in shell on the Ubuntu linux operating system
  ◦ Terminal on Mac OSX is very similar, windows ISN’T
    ▪ Most bioinformatics software doesn’t support Windows
    ▪ Use virtual machine or Cygwin

• Note that you will be training on unusually tractable data
  ◦ Beautiful, clear clustering, significant p-values, etc.
  ◦ If your own data don’t give such clear results, that doesn't mean the analysis is wrong
Tips to Help

• When typing a file name or directory path, you can use **tab completion**
  ◦ Start typing file/directory path, then hit tab—if only one file/directory matches what you already typed, shell fills that in
    ▪ Very helpful for correctly entering long file names
    ▪ If >1 matches, shell fills in as much as it can

• Press **up arrow** to get back previous commands you typed

• If you type a command, press enter, and “nothing happens”, **don’t just run it again**
  ◦ Many unix commands produce no visible output to shell—just get back command prompt
  ◦ That doesn’t mean they do nothing, so running them *again* can screw up results
  ◦ **Do not store commands in a word processing program** (or PowerPoint, etc)
    ▪ E.g., MS Word changes hyphens to “m dash”—which command line can’t understand
  ◦ Shell commands are **case-sensitive**
Pre-Acknowledgment

• Please join me in thanking **Pedro Fernandes**
• Without his impeccably managed training computers, resources, and room, these tutorials would not be possible!
Making a Mapping File

- “Mapping file” contains metadata for study
  - Must contain info needed to process sequences and test YOUR hypotheses

- QIIME 1 required certain columns in certain order, but QIIME 2 is more flexible
  - Tab-separated text file with column labels in first line + at least one data line
    - Column label values must be unique (i.e. no duplicate values)
    - First column is the “identifier” column (sample ID)
      - All values in the first column must be unique (i.e. no duplicate values)
  - See [https://docs.qiime2.org/2017.6/tutorials/metadata/](https://docs.qiime2.org/2017.6/tutorials/metadata/)

- The easiest way to make a mapping file is with a spreadsheet
  - But Excel is not your friend!
    - Routinely corruptions gene symbols, anything interpreted as a dates, etc, & isn’t reversible
Practicum: Viewing A Mapping File

• Open Terminal
  ◦ For below, remember to try tab completion!

  source activate qiime2-2017.6
  cd tutorial-qiime2
  ls
  nano sample-metadata.tsv

• Turn on your green light when the mapping file opens for you
## Mapping File View

**GNU nano 2.0.6**

**File: sample-metadata.tsv**

<table>
<thead>
<tr>
<th>SampleID</th>
<th>BarcodeSequence</th>
<th>LinkerPrimerSequence</th>
<th>BodySite</th>
<th>Year</th>
<th>Month</th>
<th>Day</th>
<th>Subject</th>
<th>ReportedAntibioticUsage</th>
<th>DaysSinceExperimentStart</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1558</td>
<td>AGCTGACTAGTC</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>gut</td>
<td>2009</td>
<td>10</td>
<td>28</td>
<td>subject-1</td>
<td>Yes 0</td>
<td>subject-1.gut.2009-10-28</td>
<td></td>
</tr>
<tr>
<td>L1557</td>
<td>ACGACTATGCG</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>gut</td>
<td>2009</td>
<td>1</td>
<td>28</td>
<td>subject-1</td>
<td>No 84</td>
<td>subject-1.gut.2009-1-20</td>
<td></td>
</tr>
<tr>
<td>L1576</td>
<td>ACTAGTGTGTTG</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>gut</td>
<td>2009</td>
<td>2</td>
<td>17</td>
<td>subject-1</td>
<td>No 112</td>
<td>subject-1.gut.2009-2-17</td>
<td></td>
</tr>
<tr>
<td>L1585</td>
<td>AGTGCCGGTCTG</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>gut</td>
<td>2009</td>
<td>3</td>
<td>17</td>
<td>subject-1</td>
<td>No 140</td>
<td>subject-1.gut.2009-3-17</td>
<td></td>
</tr>
<tr>
<td>L25155</td>
<td>ACGATGCGGCA</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>left palm</td>
<td>2000</td>
<td>1</td>
<td>28</td>
<td>subject-1</td>
<td>No 84</td>
<td>subject-1.left-palm.2000-1-20</td>
<td></td>
</tr>
<tr>
<td>L25175</td>
<td>AGTATGCGCA</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>left palm</td>
<td>2000</td>
<td>2</td>
<td>17</td>
<td>subject-1</td>
<td>No 112</td>
<td>subject-1.left-palm.2000-2-17</td>
<td></td>
</tr>
<tr>
<td>L25284</td>
<td>ATCGACGTACGT</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>left palm</td>
<td>2000</td>
<td>3</td>
<td>17</td>
<td>subject-1</td>
<td>No 140</td>
<td>subject-1.left-palm.2000-3-17</td>
<td></td>
</tr>
<tr>
<td>L25222</td>
<td>CAGTGACGTAGT</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>left palm</td>
<td>2000</td>
<td>4</td>
<td>14</td>
<td>subject-1</td>
<td>No 168</td>
<td>subject-1.left-palm.2000-4-14</td>
<td></td>
</tr>
<tr>
<td>L35242</td>
<td>AGTACGGGCGCA</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>right palm</td>
<td>2000</td>
<td>10</td>
<td>28</td>
<td>subject-1</td>
<td>Yes 0</td>
<td>subject-1.right-palm.2000-10-28</td>
<td></td>
</tr>
<tr>
<td>L35294</td>
<td>CAGCACAGGCTA</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>right palm</td>
<td>2000</td>
<td>2</td>
<td>17</td>
<td>subject-1</td>
<td>No 112</td>
<td>subject-1.right-palm.2000-2-17</td>
<td></td>
</tr>
<tr>
<td>L35313</td>
<td>AGTACGACGTAGT</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>right palm</td>
<td>2000</td>
<td>3</td>
<td>17</td>
<td>subject-1</td>
<td>No 140</td>
<td>subject-1.right-palm.2000-3-17</td>
<td></td>
</tr>
<tr>
<td>L35341</td>
<td>CAATGACGAGAG</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>right palm</td>
<td>2000</td>
<td>4</td>
<td>14</td>
<td>subject-1</td>
<td>No 168</td>
<td>subject-1.right-palm.2000-4-14</td>
<td></td>
</tr>
<tr>
<td>L35360</td>
<td>CATTGCTATCAAC</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>right palm</td>
<td>2000</td>
<td>5</td>
<td>28</td>
<td>subject-1</td>
<td>Yes 0</td>
<td>subject-1.right-palm.2000-5-28</td>
<td></td>
</tr>
<tr>
<td>L51530</td>
<td>AATTCGATGCA</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>tongue</td>
<td>2009</td>
<td>10</td>
<td>28</td>
<td>subject-1</td>
<td>Yes 0</td>
<td>subject-1.tongue.2009-10-28</td>
<td></td>
</tr>
<tr>
<td>L51555</td>
<td>ATTTAGACGTC</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>tongue</td>
<td>2009</td>
<td>2</td>
<td>17</td>
<td>subject-1</td>
<td>No 84</td>
<td>subject-1.tongue.2009-2-17</td>
<td></td>
</tr>
<tr>
<td>L51575</td>
<td>CAGATGAGATG</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>tongue</td>
<td>2009</td>
<td>3</td>
<td>17</td>
<td>subject-1</td>
<td>No 112</td>
<td>subject-1.tongue.2009-3-17</td>
<td></td>
</tr>
<tr>
<td>L51593</td>
<td>CCTGAGTATCG</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>tongue</td>
<td>2009</td>
<td>4</td>
<td>14</td>
<td>subject-1</td>
<td>No 168</td>
<td>subject-1.tongue.2009-4-14</td>
<td></td>
</tr>
<tr>
<td>L51610</td>
<td>ATGCAGACGGTA</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>tongue</td>
<td>2008</td>
<td>10</td>
<td>28</td>
<td>subject-2</td>
<td>Yes 0</td>
<td>subject-2.gut.2008-10-28</td>
<td></td>
</tr>
<tr>
<td>L51687</td>
<td>CTAGTATCGG</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>gut</td>
<td>2009</td>
<td>3</td>
<td>17</td>
<td>subject-2</td>
<td>No 84</td>
<td>subject-2.gut.2009-3-17</td>
<td></td>
</tr>
<tr>
<td>L51670</td>
<td>CGCTGACTAGT</td>
<td>GTGCCACOMGCGCGGGTAA</td>
<td>gut</td>
<td>2009</td>
<td>3</td>
<td>17</td>
<td>subject-2</td>
<td>No 140</td>
<td>subject-2.gut.2009-3-17</td>
<td></td>
</tr>
</tbody>
</table>
Practicum: Viewing A Mapping File

• Open Terminal
  ◦ For below, remember to try tab completion!

source activate qiime2-2017.6

cd tutorial-qiime2

ls

nano sample-metadata.tsv

• Stretch the window so you can look at the contents; then, to close, type

Ctrl + x

• Mapping file errors can lead to QIIME 2 errors—or worse, garbage results!
  ◦ Keemei (pronounced ‘key may’) tool checks for errors in Google Sheets
    ▪ Chrome only, and must have Google account to use
    ▪ See http://keemei.qiime.org/
Importing Data

- After sequence data is on your machine, must be imported to a QIIME 2 “artifact”
  - Artifact = data + metadata
  - QIIME 2 artifacts have extension .qza

Note

It has been brought to our attention that the term artifact may be confusing, as it is frequently used by biologists to refer to an experimental error. We use the term artifact here to mean an object that is made by some process (similar, for example, to an archaeological artifact). Throughout our documentation and other educational materials, we try to clarify that we are talking about QIIME 2 artifacts as they are defined in this section.

- Different input commands for
  - Different kinds of input data (e.g., single-end vs paired-end)
  - Different formats of input data (e.g., sequences & barcodes in same or different file)

- See “Importing data” tutorial at https://docs.qiime2.org/
Practicum: Importing Data

qiime tools import \
   --type EMPSingleEndSequences \
   --input-path emp-single-end-sequences \
   --output-path emp-single-end-sequences.qza

• A backslash \ is used to break up a command onto multiple lines
  ◦ If you prefer to type the whole command onto one run-on line, you can leave it out
Practicum: Importing Data

qiime tools import \
    --type EMPSingleEndSequences \
    --input-path emp-single-end-sequences \
    --output-path emp-single-end-sequences.qza

• What does this command actually do?
  ◦ Tells qiime to look into the folder emp-single-end-sequences ...
  ◦ For the kind of sequence files expected for EMPSingleEndSequences ...
  ◦ And load them into a new qiime artifact named emp-single-end-sequences.qza

• Note structure of arguments to qiime command
  ◦ Plugin name then method name then arguments
    ▪ Order matters
Demultiplexing

- Must assign resulting sequences to samples to analyze
- You may not need to do this!
  - If sequencing done by a core, results may be demultiplexed before returned to you

Practicum: Demultiplexing

qiime demux emp-single \\  
--i-seqs emp-single-end-sequences.qza \\  
--m-barcodes-file sample-metadata.tsv \\  
--m-barcodes-category BarcodeSequence \\  
--o-per-sample-sequences demux.qza

• Arguments have a naming convention
  ◦ Inputs (--i-<whatever>), metadata (--m-<whatever>), parameter (--p-<whatever>), output (--o-<whatever>)
  ◦ Order doesn’t matter
Practicum: Demultiplexing (cont.)

• Presumably you’d like to know how your demultiplexing worked
• QIIME 2 artifact files can’t be viewed directly (e.g., in nano)
• New concept: QIIME 2 visualization file
  ◦ Has .qzv extension
  ◦ Is intended for human (rather than computer) use
  ◦ Generally provide info via a web browser
Practicum: Demultiplexing (cont.)

qiime demux summarize \ 
   --i-data demux.qza \ 
   --o-visualization demux.qzv

• Now view the visualization, locally

qiime tools view demux.qzv
Practicum: Demultiplexing (cont.)

Demultiplexed sequence counts summary

- Minimum: 1853
- Median: 8645.0
- Mean: 7781.11784708
- Maximum: 18787
- Total: 263878
Per-sample sequence counts

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Sequence count</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4S137</td>
<td>18787</td>
</tr>
<tr>
<td>L4S63</td>
<td>17167</td>
</tr>
<tr>
<td>L4S112</td>
<td>16265</td>
</tr>
<tr>
<td>L1S8</td>
<td>12386</td>
</tr>
</tbody>
</table>
Practicum: Demultiplexing (cont.)

**qlime2**

Overview  Interactive Quality Plot

Click and drag on plot to zoom in. Double click to zoom back out to full size.
Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.
Practicum: Demultiplexing (cont.)

qiime demux summarize \
   --i-data demux.qza \
   --o-visualization demux.qzv

• Now view the visualization, locally
  qiime tools view demux.qzv

• When done examining, in Terminal, type **JUST q**
  ◦ Don’t need to hit Enter afterwards
  ◦ Beware: quitting visualization doesn’t close web page (but page becomes unreliable)
Quality Control

QIIME defaults:
- \( r = 3 \)
- \( q = 3 \)
- \( p = 0.75 \)
- \( n = 0 \)
- \( c = 0.005\% or 2 \)

Figure 1 | Quality-filtration process flow in QIIME v1.5.0.
Practicum: Quality Control

qiime quality-filter q-score \
--i-demux demux.qza \
--o-filtered-sequences demux-filtered.qza \
--o-filter-stats demux-filter-stats.qza

• This runs the command with default values for all the tuneable parameters
  ◦ To see the optional parameters, their descriptions, and their defaults, run just
    qiime quality-filter q-score
Practicum: Quality Control

```
qiime quality-filter q-score \
  --i-demux demux.qza \
  --o-filtered-sequences demux-filtered.qza \
  --o-filter-stats demux-filter-stats.qza

qiime quality-filter visualize-stats \
  --i-filter-stats demux-filter-stats.qza \
  --o-visualization demux-filter-stats.qzv
```

• Remember: for the remainder of this tutorial, any time you create a visualization file, you will need to run an additional command to view it!

```
qiime tools view yourvisualizationfilename.qzv
```
# Practicum: Quality Control

## Per-sample sequence counts

<table>
<thead>
<tr>
<th>sample-id</th>
<th>total-input-reads</th>
<th>total-retained-reads</th>
<th>fraction-retained</th>
<th>reads-truncated</th>
<th>reads-too-short-after-truncation</th>
<th>reads-exceeding-maximum-ambiguous-bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Totals</td>
<td>263878</td>
<td>186324</td>
<td>0.706099</td>
<td>245862</td>
<td>76489</td>
<td>1065</td>
</tr>
<tr>
<td>L4S137</td>
<td>18787</td>
<td>11642</td>
<td>0.619684</td>
<td>17454</td>
<td>7123</td>
<td>22</td>
</tr>
<tr>
<td>L4S63</td>
<td>17167</td>
<td>11505</td>
<td>0.670181</td>
<td>15160</td>
<td>5634</td>
<td>28</td>
</tr>
<tr>
<td>L4S112</td>
<td>16265</td>
<td>10012</td>
<td>0.615555</td>
<td>15054</td>
<td>6232</td>
<td>21</td>
</tr>
<tr>
<td>L1S8</td>
<td>12386</td>
<td>8433</td>
<td>0.680649</td>
<td>12035</td>
<td>3916</td>
<td>37</td>
</tr>
<tr>
<td>L2S240</td>
<td>11986</td>
<td>7110</td>
<td>0.593192</td>
<td>11454</td>
<td>4845</td>
<td>31</td>
</tr>
<tr>
<td>L1S87</td>
<td>11750</td>
<td>10000</td>
<td>0.851064</td>
<td>11000</td>
<td>1716</td>
<td>34</td>
</tr>
<tr>
<td>L1S105</td>
<td>11340</td>
<td>9232</td>
<td>0.814109</td>
<td>10782</td>
<td>2066</td>
<td>42</td>
</tr>
<tr>
<td>L1S208</td>
<td>11335</td>
<td>10148</td>
<td>0.895280</td>
<td>10067</td>
<td>1161</td>
<td>26</td>
</tr>
<tr>
<td>L6S93</td>
<td>11270</td>
<td>8580</td>
<td>0.761313</td>
<td>10282</td>
<td>2680</td>
<td>10</td>
</tr>
<tr>
<td>L2S175</td>
<td>10691</td>
<td>5574</td>
<td>0.521373</td>
<td>10216</td>
<td>5092</td>
<td>25</td>
</tr>
</tbody>
</table>
Feature Table Creation—The Past

• Last year: OTU (Operational Taxonomic Unit)
  ◦ “an operational definition of a species used when only DNA sequence data is available”
  ◦ Sequences at/above a given similarity threshold considered part of the same OTU
    ▪ 97% is the usual “species-level” threshold
    ─ Similarity determined using alignment (time-consuming)
    ▪ Purpose is to minimize impact of sequencing errors
    ─ But also masks fine (sub-OTU) variation in real biological sequences
  ◦ Results very difficult to compare across studies if done de novo
    ▪ “Closed reference”, “open reference” methods increase comparability require reference database

• Output is a “feature table”:
  ▪ Rows are samples
  ▪ Columns are OTUs (arbitrary identifiers if de novo, from reference database if closed reference)
  ▪ Values are frequency of reads from that OTU in that sample
Feature Table Creation—The Present

- This year: sOTU (sub-OTU) methods
  - Use error modeling to *in silico* correct sequencing mistakes
    - Sounds impossible but is actually quite accurate, with right error model
      - Error model is specific to the sequencing type (e.g., 454, Illumina Hi/MiSeq)
    - Result: only sequences likely to have been input to the sequencer
  - Options include (NOT a complete list):
    - DADA2 (2016)
    - Deblur (2017)

- Output is STILL a feature table:
  - Rows are samples
  - Columns are SEQUENCES
  - Values are frequency of reads from that SEQUENCE in that sample

Practicum: Feature Table Creation

qiime deblur denoise-16S \
  --i-demultiplexed-seqs demux-filtered.qza \
  --p-trim-length 120 \
  --o-representative-sequences rep-seqs.qza \
  --o-table table.qza \
  --o-stats deblur-stats.qza

• This command can take a few minutes to run
  ◦ So don’t worry if the command prompt doesn’t immediately return after you hit enter

• Where do you guess the number 120 came from?
Practicum: Feature Table Creation

qiime deblur denoise-16S \
    --i-demultiplexed-seqs demux.qza \
    --p-trim-length 120 \ 
    --o-representative-sequences rep-seqs.qza \ 
    --o-table table.qza \ 
    --o-stats deblur-stats.qza

• Where do you guess the number 120 came from?
  ◦ It is the length to which all sequences will be trimmed
  ◦ It was chosen by viewing the Interactive Quality Plot in demux.qzv
  ◦ You might even choose a more conservative length, like 110
Practicum: Feature Table Creation (cont.)

qiime feature-table tabulate-seqs \
  --i-data rep-seqs.qza \
  --o-visualization rep-seqs.qzv
## Feature Table Tabulation View

To BLAST a sequence against the NCBI nt database, click the sequence and then click the View report button on the resulting page.

To download a raw FASTA file of your sequences, click here.

Click on a Column header to sort the table.

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>3677e15d9d603f0da8b005f6b01af07</td>
<td>TACGTAAGGGGCAAGCTATCCCGCAGTTTACTGAGTGAATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>1b7562f8b93420d2cc279a1e811497a</td>
<td>TACAGAGGCTGCGAGGAGCGGTAAATCGGATTATTACTGAGCCGTAAGCGCTGCTGAGCCGCTGCTGATTAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>42872dc873ef667070d9a789b41840c966</td>
<td>TACGAGGCGAGCTGATCCGAGATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>51dbd685cfb177931489ebe33e69ca</td>
<td>TACGAGGCTGCGAGGAGCGGTAAATCGGATTATTACTGAGCCGTAAGCGCTGCTGAGCCGCTGCTGATTAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>8be67d8a197b54f9d04f8c98f3b1f22</td>
<td>TACGAGGCGAGCTGATCCGAGATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>54b964000a1631e547c46a28ed1a00</td>
<td>TACGAGGCTGCGAGGAGCGGTAAATCGGATTATTACTGAGCCGTAAGCGCTGCTGAGCCGCTGCTGATTAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>eccbf086fcccbe6e8c2a49d34eb146462</td>
<td>TACGAGGCGAGCTGATCCGAGATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>c18829f57effa174f5b0164cf65a38a</td>
<td>TACGAGGCTGCGAGGAGCGGTAAATCGGATTATTACTGAGCCGTAAGCGCTGCTGAGCCGCTGCTGATTAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>413256a0d28f3b577e9be9730c4a700</td>
<td>TACGAGGCTGCGAGGAGCGGTAAATCGGATTATTACTGAGCCGTAAGCGCTGCTGAGCCGCTGCTGATTAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>7595e12b779d89e8c1c26b7405a5c0d</td>
<td>TACGAGGCGAGCTGATCCGAGATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>4a5387c84bc162f31f9f3d92d6e983a5b56</td>
<td>TACGAGGCGAGCTGATCCGAGATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>79dcbe77928c7f2732b796dc2d233f</td>
<td>TACGAGGCGAGCTGATCCGAGATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
<tr>
<td>7595e12b779d89e8c1c26b7405a5c0d</td>
<td>TACGAGGCGAGCTGATCCGAGATTAGGAGCGAGTAGAAGCGTAGGATAGGCTGAAGTGAAGCCG</td>
</tr>
</tbody>
</table>
Practicum: Feature Table Creation (cont.)

```bash
qiime feature-table summarize \
  --i-table table.qza \
  --o-visualization table.qzv \
  --m-sample-metadata-file sample-metadata.tsv
```
# Feature Table Summary View

## Table summary

<table>
<thead>
<tr>
<th>Metric</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples</td>
<td>34</td>
</tr>
<tr>
<td>Number of features</td>
<td>485</td>
</tr>
<tr>
<td>Total frequency</td>
<td>102,545</td>
</tr>
</tbody>
</table>

## Frequency per sample

<table>
<thead>
<tr>
<th>Frequency</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum frequency</td>
<td>512.0</td>
</tr>
<tr>
<td>1st quartile</td>
<td>1,367.5</td>
</tr>
<tr>
<td>Median frequency</td>
<td>2,581.0</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>4,952.0</td>
</tr>
<tr>
<td>Maximum frequency</td>
<td>6,770.0</td>
</tr>
<tr>
<td>Mean frequency</td>
<td>2,016.0</td>
</tr>
</tbody>
</table>
# Feature Table Summary View

<table>
<thead>
<tr>
<th>Feature</th>
<th>Frequency</th>
<th># of Samples Observed In</th>
</tr>
</thead>
<tbody>
<tr>
<td>4b5eeb300368260019c1fc7a3c718fc</td>
<td>8,223</td>
<td>16</td>
</tr>
<tr>
<td>fe30ff71a38a39cf1717ec2be3a2fc</td>
<td>6,935</td>
<td>19</td>
</tr>
<tr>
<td>d29fed3ce7054f0c0699f2c03e0d1e5561</td>
<td>6,428</td>
<td>27</td>
</tr>
<tr>
<td>1d2ef5444ca750c85302cceed2473331</td>
<td>5,809</td>
<td>27</td>
</tr>
<tr>
<td>868528ca947bc57b69f9f83e6b73bae</td>
<td>5,347</td>
<td>12</td>
</tr>
<tr>
<td>154709e1600e8cada68f2b1115acc80f5</td>
<td>5,117</td>
<td>14</td>
</tr>
<tr>
<td>0305e64893ecf2d8e4149f4fc7592603</td>
<td>3,671</td>
<td>13</td>
</tr>
<tr>
<td>997056ba80681b9bddd5d9aa591eadc0</td>
<td>3,051</td>
<td>18</td>
</tr>
<tr>
<td>cb2fe01e6a2fcb1c101050edeb996ab9ea2</td>
<td>3,021</td>
<td>17</td>
</tr>
<tr>
<td>3c9c437f27aca05f9db167cd0800f1ec</td>
<td>2,358</td>
<td>18</td>
</tr>
<tr>
<td>9075bfebcce0f2f48b7c7580e67b1208c31</td>
<td>2,093</td>
<td>15</td>
</tr>
<tr>
<td>b7b63eb63e6b9f0c4627424163d20118</td>
<td>1,622</td>
<td>17</td>
</tr>
<tr>
<td>d86ef5d394f5dc6eb894f5f9a25e7426</td>
<td>1,405</td>
<td>12</td>
</tr>
<tr>
<td>a4d87f6e65c9777146c343341a9b93b4</td>
<td>1,118</td>
<td>16</td>
</tr>
</tbody>
</table>
Feature Table Summary View

Overview Interactive Sample Detail Feature Detail

Sampling depth:

1000

(Zero implies no even sampling.)

Subject

Retained 27,000 (26.33%) sequences in 27 (79.41%) samples at the specified sampling depth.
# Feature Table Summary View

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Sequence Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4S137</td>
<td>6,770</td>
</tr>
<tr>
<td>L4S63</td>
<td>5,912</td>
</tr>
<tr>
<td>L1S57</td>
<td>5,525</td>
</tr>
<tr>
<td>L4S112</td>
<td>5,523</td>
</tr>
<tr>
<td>L6S93</td>
<td>5,261</td>
</tr>
<tr>
<td>L6S20</td>
<td>5,170</td>
</tr>
<tr>
<td>L1S20B</td>
<td>5,136</td>
</tr>
<tr>
<td>L1S76</td>
<td>5,037</td>
</tr>
<tr>
<td>L1S155</td>
<td>1,347</td>
</tr>
<tr>
<td>L5S240</td>
<td>1,329</td>
</tr>
<tr>
<td>L2S309</td>
<td>895</td>
</tr>
<tr>
<td>L3S378</td>
<td>849</td>
</tr>
<tr>
<td>L3S294</td>
<td>800</td>
</tr>
<tr>
<td>L3S313</td>
<td>741</td>
</tr>
<tr>
<td>L3S242</td>
<td>660</td>
</tr>
<tr>
<td>L3S341</td>
<td>653</td>
</tr>
<tr>
<td>L3S360</td>
<td>512</td>
</tr>
</tbody>
</table>
Phylogenetic Tree Creation

• Evolution is the core concept of biology
  ◦ There’s only so much you can learn from microbes while ignoring evolution!

• Evolution-aware analyses of a dataset need a phylogenetic tree of its sequences
  ◦ *De novo*: infer tree using only sequences from dataset
  ◦ Reference-based: insert sequences from dataset into an existing phylogenetic tree
    ▪ Not all existing phylogenies are created equal—have strengths and weaknesses based on intended purpose when developed

• Phylogenetically based analyses in QIIME 2 need a **rooted** tree

Unrooted: Rooted:

---

Practicum: Phylogenetic Tree Creation

• Note: here we are doing *de novo* phylogenetic tree creation
  ◦ Not necessarily the BEST approach, but an easy to show you 😊

• No visualizations will be produced
Practicum: Phylogenetic Tree Creation

qiime alignment mafft \
 --i-sequences rep-seqs.qza \
 --o-alignment aligned-rep-seqs.qza

qiime alignment mask \
 --i-alignment aligned-rep-seqs.qza \
 --o-masked-alignment masked-aligned-rep-seqs.qza

qiime phylogeny fasttree \
 --i-alignment masked-aligned-rep-seqs.qza \
 --o-tree unrooted-tree.qza

qiime phylogeny midpoint-root \
 --i-tree unrooted-tree.qza \
 --o-rooted-tree rooted-tree.qza
So how do you actually compare microbial communities?
- Can’t just eyeball the (gigantic, sparse) feature tables and look for differences
- Instead, calculate metrics that compress a lot of info into a single number
- Then do statistical tests on metrics to look for significant differences
  - **BE CAREFUL**—microbiome data is sparse, compositional, etc, so requires unusual tests
  - QIIME 2 uses appropriate tests; if doing your own, **MUST** check the literature first

These metrics are lossy!
- No metric exposes all the information in the full feature table
  - If it did, it would **BE** the feature table
- Different metrics capture different aspects of the communities

Thus ...
- Don’t ask, “Which metric should I use?” **UNTIL** you know what you’re looking for!
Core Metrics (cont.)

- QIIME 2 calculates a smorgasbord of metrics for you with one command

- Alpha diversity
  - Shannon’s diversity index (a quantitative measure of community richness)
  - Observed OTUs (a qualitative measure of community richness)
  - Faith’s Phylogenetic Diversity (a qualitative measure of community richness that incorporates phylogenetic relationships between the features)
  - Evenness (or Pielou’s Evenness; a measure of community evenness)

- Beta diversity
  - Jaccard distance (a qualitative measure of community dissimilarity)
  - Bray-Curtis distance (a quantitative measure of community dissimilarity)
  - unweighted UniFrac distance (a qualitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)
  - weighted UniFrac distance (a quantitative measure of community dissimilarity that incorporates phylogenetic relationships between the features)
Normalization for Core Metrics

- Calculated metric values depend on sampling depth
  - Ex: circled column has more non-zero counts than others
    - Is its community really more diverse—or do we just SEE more?
    - Samples with more sequences (greater sampling depth) show more diversity

- Normalization is necessary for valid comparisons of abundance/diversity
  - “But how?!”
    - Longstanding approach: rarefaction (reduce all samples to uniform sampling depth)
    - Recent publication caused concern
    - Further work demonstrated concern is excessive
      - Normalization and microbial differential abundance strategies depend upon data characteristics. Weiss S, et al. Microbiome. 2017 Mar 3;5(1):27. (Note: I’m an author, so not objective)
Rarefaction

• What is rarefaction?
  ◦ randomly subsampling the same number of sequences from each sample
  ◦ NB: samples without that number of sequences are discarded

• Concerns:
  ◦ Too low: ignore a lot of samples’ information
  ◦ Too high: ignore a lot of samples
  ◦ Still a good choice for normalization (Weiss S, et al. Microbiome. 2017):
    ▪ “Rarefying more clearly clusters samples according to biological origin than other normalization techniques do for ordination metrics based on presence or absence”
    ▪ “Alternate normalization measures are potentially vulnerable to artifacts due to library size”

• Researcher must choose sampling depth—but how?
Sampling Depth Selection

• Don’t sweat it too much
  ◦ “Low” depths (10-1000 sequences per sample) capture all but very subtle variations

![Graphs showing variation within and between samples](image)

Fig. 2, Kuczynski, J. et al., "Direct sequencing of the human microbiome readily reveals community differences", Genome Biology, 2010

◦ Retaining samples is usually more important than retaining sequences
  ▪ May care not just how many samples are left out but WHICH samples are left out
Exercise: Core Metrics Sampling Depth

• Do NOT start typing yet!

qiime diversity core-metrics \
  --i-phylogeny rooted-tree.qza \
  --i-table table.qza \
  --p-sampling-depth ??? \
  --output-dir metrics

• Note that the core metrics command requires a sampling depth
Exercise: Core Metrics Sampling Depth

• Which sampling depth should we use?
  ◦ How can we decide?
    - Work with your partner to choose a sampling depth, then answer:
      - Why did you choose this value?
      - How many samples will be excluded from your analysis based on this choice?
      - How many total sequences will you be analyzing in the core-metrics command?

qiime tools view table.qzv
**Answers: Core Metrics**

```
qiime diversity core-metrics \\
   --i-phylogeny rooted-tree.qza \\
   --i-table table.qza \\
   --p-sampling-depth 800 \\
   --output-dir metrics
```

- **My answers:**
  - Why did you choose this value?
    - Anything higher excludes \( \geq \) half of right palm samples
  - How many samples will be excluded from your analysis based on this choice?
    - 4, all from right palm of subject 1
  - How many total sequences will you be analyzing in the core-metrics command?
    - 24,000 (23.40%)
Practicum: Core Metrics

qiime diversity core-metrics \
--i-phylogeny rooted-tree.qza \
--i-table table.qza \
--p-sampling-depth 800 \
--output-dir metrics

• Note: there is no single visualization for core metrics
  ◦ We will examine a few different visualizations later
Beta Diversity

• “Between-sample” diversity
  ◦ Has similar categories, caveats as $\alpha$ diversity

• A popular phylogenetic option is 'UniFrac':
  ◦ Measures how different two samples' component sequences are

  ◦ Weighted UniFrac: takes abundance each sequence into account

Illustration courtesy of Dr. Rob Knight
Beta Diversity Ordination

• **Ordination**: multivariate techniques that arrange samples along axes on the basis of composition

• **Principal Coordinates Analysis**: a way to map non-Euclidean distances into a Euclidean space to enable further investigation
  ◦ Abbreviated as PCoA, not to be confused with PCA (Principal Component Analysis)
  ◦ Starting point is distance matrix
    ▪ NOT the full set of independent variables for each sample
  ◦ n pairwise distances are projected into n-1 dimensions
  ◦ PCA performed to reduce the dimensionality back down

• PCoA axes can’t be decomposed into independent variable contributions
  ◦ But results can be compared to metadata to identify patterns
Practicum: Beta Diversity Ordination

qiime emperor plot \
   --i-pcoa metrics/unweighted_unifrac_pcoa_results.qza \
   --m-metadata-file sample-metadata.tsv \
   --o-visualization metrics/unweighted-unifrac-emperor.qzv

• This is only showing the PCoA visualization of ONE beta diversity metric
  ◦ Not necessarily “the correct one”!
  ◦ Remember that 3 others are calculated by core-metrics alone

• To check the group significance of a different metric, just input a different file
  ◦ To find them:
    cd metrics/
    ls *_pcoa_results.qza
Beta Diversity Ordination View
Exercise: Beta Diversity Ordination

• Initial PCoA view (see previous slide) is completely independent of metadata
  ◦ Clusters/gradients/etc seen in PCoA are produced by unsupervised learning, based on
    the feature table information without any awareness of metadata

• It’s great to see clear, distinct clusters as in this dataset–but even greater if they
  can be explained by a known metadata category

• Work with your partner to answer the following question:
  ◦ Can you find a metadata category that appears associated with the observed clusters?
    ▪ Hint: Experiment with coloring points by different metadata
Answers: Beta Diversity Ordination

• My answer:
  ◦ Can you find a metadata category that appears associated with the observed clusters?
    ▪ Yep: BodySite. Note left and right palm aren’t distinct from each other, unsurprisingly
Practicum: Beta Diversity Group Significance

qiime diversity beta-group-significance \
--i-distance-matrix metrics/unweighted_unifrac_distance_matrix.qza \
--m-metadata-file sample-metadata.tsv \
--m-metadata-category BodySite \ 
--p-pairwise \ 
--o-visualization \ 
metrics/unweighted-unifrac-bodysite-significance.qzv

• Standard caveats apply—not the “one true metric”, etc
### Beta Diversity Group Significance View

**qime2**

<table>
<thead>
<tr>
<th>Method/Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method name</td>
<td>PERMANOVA</td>
</tr>
<tr>
<td>Test statistic name</td>
<td>pseudo-F</td>
</tr>
<tr>
<td>Sample size</td>
<td>30</td>
</tr>
<tr>
<td>Number of groups</td>
<td>4</td>
</tr>
<tr>
<td>Test statistic</td>
<td>10.556</td>
</tr>
<tr>
<td>P-value</td>
<td>0.001</td>
</tr>
<tr>
<td>Number of permutations</td>
<td>999</td>
</tr>
</tbody>
</table>

#### PERMANOVA Results

- Method name: PERMANOVA
- Test statistic name: pseudo-F
- Sample size: 30
- Number of groups: 4
- Test statistic: 10.556
- P-value: 0.001
- Number of permutations: 999

#### Distances to Gut

- Box plots showing distances to gut with various ranges.
Exercise: Beta Diversity Group Significance

• Work with your partner to answer these questions:
  ◦ Does the group significance analysis bear out your intuition from the ordination?
    ▪ If so, are the differences statistically significant?
    ▪ Are there specific pairs of BodySite values that are significantly different from each other?
  ◦ How about Subject?
    ▪ Hint: you will need to run a new command!
Answers: Beta Diversity Group Significance

• My answers:
  ◦ Does the group significance analysis bear out your intuition from the ordination?
    ▪ Yes
    ▪ If so, are the differences statistically significant?
      — Yes, with p <= 0.001 (bonus: why do I say “less than or equal to”?)
    ▪ Are there specific pairs of BodySite values that are significantly different from each other?
      — Yes, all of the pairs except left palm/right palm
  ◦ How about Subject?
    • qiime diversity beta-group-significance \\n    • --i-distance-matrix metrics/unweighted_unifrac_distance_matrix.qza \\n    • --m-metadata-file sample-metadata.tsv \\n    • --m-metadata-category Subject\ 
    • --p-pairwise \ 
    • --o-visualization metrics/unweighted-unifrac-subject-significance.qzv
    ▪ Nope: significance of difference of distributions of unweighted unifrac metric grouped by subject has p = 0.442
Acknowledgements

• Center for Computational Biology & Bioinformatics, University of California at San Diego
• Caporaso lab, Northern Arizona University
• Knight lab, UCSD

• QIIME 2 development team!
  ◦ Especially for the excellent “Moving Pictures” tutorial on which this one is based
Practicum: Beta Diversity Ordination

• But wait, this is time-series data! Maybe we’d like to view it on a time axis:

  qiime emperor plot \ 
  --i-pcoa metrics/unweighted_unifrac_pcoa_results.qza \ 
  --m-metadata-file sample-metadata.tsv \ 
  --p-custom-axis DaysSinceExperimentStart \ 
  --o-visualization metrics/unweighted-unifrac-emperor-bydayssince.qzv

• Standard caveats apply
Supplemental Slides
Beta Diversity Ordination View (cont.)
Practicum: Beta Diversity Correlation

qiime diversity beta-correlation \
   --i-distance-matrix metrics/unweighted_unifrac_distance_matrix.qza \
   --m-metadata-file sample-metadata.tsv \
   --m-metadata-category DaysSinceExperimentStart \
   --o-visualization metrics/unweighted-unifrac-dayssinceexperimentstart-beta-correlation.qzv

• Standard caveats apply
Beta Diversity Correlation View
Alpha Diversity

• “Within-sample” diversity
  ◦ Many different metrics exist
    ▪ Taxonomy-based (e.g., number of observed OTUs)
      – Assume everything is equally dissimilar
      – More likely to see differences based on close relatives
    ▪ Phylogeny-based (e.g., phylogenetic diversity over whole tree)
      – Treat less related items as more dissimilar
      – Better at scaling the observed differences
  ◦ The “correct” metric(s) are those relevant to your hypothesis
    ▪ Please do HAVE a hypothesis!

• Testing approach:
  ◦ Examine alpha diversity metric by metadata values
  ◦ Test whether differences in metric distribution is different between groups (if metadata is categorical) or correlated with metadata (if metadata is continuous)
Alpha Diversity

High within-sample diversity—why?
Practicum: Alpha Diversity Group Significance

qiime diversity alpha-group-significance \ 
  --i-alpha-diversity-metrics/faith_pd_vector.qza \ 
  --m-metadata-file sample-metadata.tsv \ 
  --o-visualization metrics/faith-pd-group-significance.qzv

• Note: only showing you the group significance visualization of ONE alpha diversity metric
  ◦ Remember that 3 others are calculated by core-metrics alone
  ◦ The one I am showing is not “the correct one”—pick the one that fits your hypothesis

• To check the group significance of a different metric, just input a different vector file
  ◦ To find them:
    cd metrics/
    ls *_vector.qza
Alpha Diversity Group Significance View

The following metadata categories have been omitted because the metadata was numerous, not categorical: DaysSinceExperimentStart, Year, Day, Month.

The following metadata categories have been omitted because the number of groups was equal to the number of samples, or there was only a single category: BarcodeSequence, LinkerPrimerSequence, Description.

Alpha Diversity Boxplots

Download SVG

BodySite
Alpha Diversity Group Significance View

Kruskal-Wallis (all groups)

Result

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>16.267096747419356</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.000000000000000</td>
<td></td>
</tr>
</tbody>
</table>

Kruskal-Wallis (pairwise)

Download CSV

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>H</th>
<th>p-value</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gui (n=6)</td>
<td>left palm (n=6)</td>
<td>11.29418</td>
<td>0.000778</td>
<td>0.002333</td>
</tr>
<tr>
<td></td>
<td>right palm (n=6)</td>
<td>3.02429</td>
<td>0.057040</td>
<td>0.107791</td>
</tr>
<tr>
<td></td>
<td>tongue (n=6)</td>
<td>0.750000</td>
<td>0.386476</td>
<td>0.463771</td>
</tr>
<tr>
<td>left palm (n=6)</td>
<td>right palm (n=6)</td>
<td>0.000000</td>
<td>1.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td></td>
<td>tongue (n=6)</td>
<td>12.000000</td>
<td>0.000032</td>
<td>0.002333</td>
</tr>
<tr>
<td>right palm (n=6)</td>
<td>tongue (n=6)</td>
<td>3.240000</td>
<td>0.071891</td>
<td>0.107791</td>
</tr>
</tbody>
</table>
Exercise: Alpha Diversity Group Significance

qiime diversity alpha-group-significance \
    --i-alpha-diversity metrics/faith_pd_vector.qza \
    --m-metadata-file sample-metadata.tsv \
    --o-visualization metrics/faith-pd-group-significance.qzv

• Work with your partner to answer these questions:
  ◦ Is BodySite value associated with significant differences in phylogenetic diversity?
  ◦ Which two sites have the most significant difference in phylogenetic diversity distributions?
    ▪ Note different between p-value and q-value
  ◦ Is Subject value associated with significant differences in phylogenetic diversity?
Answers: Alpha Diversity Group Significance

- My answers:
  - Is BodySite value associated with significant differences in phylogenetic diversity?
    - Yes, with $p < 1 \times 10^{-3}$
  - Which two sites have the most significantly difference in phylogenetic diversity distributions?
    - Left palm is (equally) most significantly different from gut and tongue
      - Consider: any idea why perhaps left palm but not right?
  - Is Subject value associated with significant differences in phylogenetic diversity?
    - No, $p$ value = 0.24
Practicum: Alpha Diversity Correlation

qiime diversity alpha-correlation \
--i-alpha-diversity metrics/evenness_vector.qza \
--m-metadata-file sample-metadata.tsv \
--o-visualization metrics/evenness-alpha-correlation.qzv

• Same caveat as before:
  ◦ Only showing the correlation visualization of ONE alpha diversity metric
    ▪ Not necessarily “the correct one”!
Alpha Diversity Correlation View

The following metadata categories have been omitted because the metadata was categorical, not numerical: BodySite, Subject, BarcodeSequence, ReportedAntibioticUsage, Description, LinkerPrimerSequence

Alpha Correlation

<table>
<thead>
<tr>
<th>Test</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test statistic</td>
<td>0.3328</td>
</tr>
<tr>
<td>P-value</td>
<td>0.0723</td>
</tr>
<tr>
<td>Sample size</td>
<td>30</td>
</tr>
</tbody>
</table>

Download SVG: DaysSinceExperimentStart vs.
Taxonomic Assignment

• Sequence features or OTUs have limited utility
  ◦ At some point, you’ll want to link your findings to published work
  ◦ That requires identifying the taxonomy of each sequence feature

• Steps:
  ◦ Pick reference database
    ▪ I hear you cry, “Which one should I use?”
  ◦ Train a classifier algorithm to assign taxonomies to sequences
    ▪ Use the reference database as the training set
  ◦ Run the classifier algorithm on your sequence features
Taxonomic Assignment

• Sequence features or OTUs have limited utility
  ◦ At some point, you’ll want to link your findings to published work
  ◦ That requires identifying the taxonomy of each sequence feature

• Steps:
  ◦ Pick reference database
    ▪ I hear you cry, “Which one should I use?”
  ◦ Train a classifier algorithm to assign taxonomies to sequences
    ▪ Use the reference database as the training set
  ◦ Run the classifier algorithm on your sequence features
Common Issues in Marker Gene Studies

- Neglecting metadata
  - Analysis can not test for effects of, or discard bias from, features you didn’t record!

- Picking novel 16S primers—not all created equal
  - Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes

- Not taking precautions to support amplicon sequencing
  - Some Illumina machines require high PhiX, low cluster density

- Selecting an inappropriate reference database
  - E.g., Greengenes (16S) reference database when sequencing ITS
Marker Gene Reference Databases

- NOT a complete list:
  - Greengenes: 16S
  - Silva: 16S/18S
  - RDP: 16S/18S/28S
  - UNITE: ITS

- Another not complete list at eukref.org/databases (not just eukaryotic)

- At the very least, choose a database that includes your marker gene!
  - Beyond that, formal guidance is hard to find
  - But off the record you might get some informal guidance 😊
Taxonomic Assignment

• Sequence features or OTUs have limited utility
  ◦ At some point, you’ll want to link your findings to published work
  ◦ That requires identifying the taxonomy of each sequence feature

• Steps:
  ◦ Pick reference database
    ▪ I hear you cry, “Which one should I use?”
  ◦ Train a classifier algorithm to assign taxonomies to sequences
    ▪ Use the reference database as the training set
  ◦ Run the classifier algorithm on your sequence features
Common Issues in Marker Gene Studies

• Neglecting metadata
  ◦ Analysis can not test for effects of, or discard bias from, features you didn’t record!

• Picking novel 16S primers—not all created equal
  ◦ Earth Microbiome Project recommends 515f-806r primers, error-correcting barcodes

• Not taking precautions to support amplicon sequencing
  ◦ Some Illumina machines require high PhiX, low cluster density

• Selecting an inappropriate reference database
  ◦ E.g., Greengenes (16S) reference database when sequencing ITS

• Expecting species-level taxonomy calls
  ◦ Most OTUs/features only specified to family or genus level
<table>
<thead>
<tr>
<th>Taxonomy: Expectation Vs Reality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kingdom</strong></td>
</tr>
<tr>
<td>Bacteria</td>
</tr>
<tr>
<td><strong>Phylum</strong></td>
</tr>
<tr>
<td><strong>Class</strong></td>
</tr>
<tr>
<td><strong>Order</strong></td>
</tr>
<tr>
<td><strong>Family</strong></td>
</tr>
<tr>
<td><strong>Genus</strong></td>
</tr>
<tr>
<td><strong>Species</strong></td>
</tr>
<tr>
<td><strong>Strain</strong></td>
</tr>
</tbody>
</table>
Practicum: Taxonomic Assignment

qiime taxa tabulate
   --i-data taxonomy.qza
   --o-visualization taxonomy.qzv
### Taxonomic Assignment Tabulation View

<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3677e15d86603d0a8bbf5b101af7</td>
<td>k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_Lachnospiraceae; g__; s__</td>
</tr>
<tr>
<td>1b75626f834820d2c2729a1a81f197a</td>
<td>k_Bacteria; p_Proteobacteria; c_Gammaproteobacteria; o_Pseudomonadales; f_Moraxellaceae; g_Acinetobacter; s__</td>
</tr>
<tr>
<td>42872dc578ef8070af87884184c06</td>
<td>k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_[Mogibacteriae]; g__; s__</td>
</tr>
<tr>
<td>512db85e1f77593148eb3a3e96ca</td>
<td>k_Bacteria; p_Bacteroidetes; c_Bacteroidia; o_Bacteroidales; f_Porphyromonadaceae; g_Paludibacter; s__</td>
</tr>
<tr>
<td>6bb68e1a197b54f9a4f60394840b9f22</td>
<td>k_Bacteria; p_Proteobacteria; c_Alphaproteobacteria; o_Sphingomonadales; f_Sphingomonadaceae</td>
</tr>
<tr>
<td>54b964002ad631e547c482861ed1a0</td>
<td>k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales</td>
</tr>
<tr>
<td>ecbf08d9ccbe5e82ca6920af5b144662</td>
<td>k_Bacteria; p_Acinetobacteria; c_Acinetobacteriaceae; o_Acinetomyxota; f_Micrococcales; g__; s__</td>
</tr>
<tr>
<td>c18829df5a5da174f50164c305ca38a</td>
<td>k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Streptococcaceae; g_Streptococcus; s_anginosus</td>
</tr>
<tr>
<td>4132561a3b25767e4beef73ac4a70a</td>
<td>k_Bacteria; p_Proteobacteria; c_Betaproteobacteria; o_Neisseriaceae; f_Neisseriaceae</td>
</tr>
<tr>
<td>759f5e123b71bdaea8a1c1c2cb7405ac50c</td>
<td>k_Bacteria; p_Firmicutes; c_Bacilli; o_Lactobacillales; f_Streptococcaceae</td>
</tr>
<tr>
<td>4a536704bc812fd839d9de838e5s556</td>
<td>k_Bacteria; p_Bacteroidetes; c_Flavobacteria; o_Flavobacteriales; f_[Weeksellaeae]; g__Chryseobacterium; s__</td>
</tr>
<tr>
<td>79acbc27925d4f2723789dcd2f239f</td>
<td>k_Bacteria; p_Acinetobacteria; c_Acinetobacteriaceae; o_Acinetomyxota; f_Corynebacteriaceae; g_Corynebacterium; s__</td>
</tr>
<tr>
<td>6edc59464812off7d197299f01663</td>
<td>k_Bacteria; p_Bacteroidetes; c_Bacteroidiota; o_Bacteroidales; f_Prevotellaceae; g_Prevotella; s_melaninogonica</td>
</tr>
<tr>
<td>fcd45f5c05b868000121f70906b2f21</td>
<td>k_Bacteria; p_Firmicutes; c_Clostridia; o_Clostridiales; f_[Tissierellaeae]; g_Firmicutes; s__</td>
</tr>
<tr>
<td>9540e9f5da194029f54e20920b39999</td>
<td>k_Bacteria; p_Rutaceae; c_Rutaceae; o_Rutaceae; f_[Paretomycetinae]; g_Paretomycetinae; s__</td>
</tr>
</tbody>
</table>
Practicum: Taxonomic Assignment

qiime feature-classifier classify-sklearn \
   --i-classifier gg-13-8-99-515-806-nb-classifier.qza \
   --i-reads rep-seqs.qza \
   --o-classification taxonomy.qza

qiime taxa barplot \
   --i-table table.qza \
   --i-taxonomy taxonomy.qza \
   --m-metadata-file sample-metadata.tsv \
   --o-visualization taxa-bar-plots.qzv
Taxonomic Assignment Bar Plot View
Exercise: Taxonomic Assignment

• “Level 1” = kingdom, “Level 2” = phylum, etc

• Work with your partner to:
  ◦ Visualize the taxa at level 2
  ◦ Sort the samples by BodySite
  ◦ Do you see anything suggestive?
• Gut sure seems to have a lot more Bacteroidetes than the other sites
Differential Abundance Analysis

• Why go to the trouble of assigning taxonomies?
  ◦ Probably you want to know whether any particular taxa are differentially abundant
    ▪ In different individuals, environments, time points, etc

• How to test for differential abundance?
  ◦ Remember: microbiome datasets are “compositional” (fixed sum)
  ◦ Watch out: “traditional” statistical methods perform badly for this sort of data!
    ▪ E.g., 95% false positives when you expect an FDR of 5%
  ◦ What to use instead?
    ▪ Balance trees (borrowed from geology) are currently the best known option (as of 2017)
      ─ But they aren’t implemented in QIIME 2 yet
      ─ So until they are, use previous best known option (as of 2015)
    ▪ ANCOM (ANalysis of Composition Of Microbiomes)
Practicum: Differential Abundance Analysis

```
qiime taxa collapse \
  --i-table table.qza \
  --i-taxonomy taxonomy.qza \
  --p-level 2 \
  --o-collapsed-table table-level2.qza

qiime composition add-pseudocount \
  --i-table table-12.qza \
  --o-composition-table comp-table-level2.qza

qiime composition ancom \
  --i-table comp-table-level2.qza \
  --m-metadata-file sample-metadata.tsv \
  --m-metadata-category BodySite \
  --o-visualization ancom-bodysite-level2.qzv
```
Differential Abundance Analysis View

ANCOM statistical results

Percentile abundances of features by group

[Table]

Differential Abundance View
Details: Differential Abundance Analysis

• W-statistic
  ◦ # of other items from which a single item is found to be significantly different
    ▪ With alpha=0.05 by default (can be changed)

• Percentile abundance table:
  ◦ A table of items and their percentile abundances in each group
  ◦ Rows are items
  ◦ Columns are percentile within a group
  ◦ Values are abundance of reads for given percentile for that group