QIIME2 for Microbiome Analysis: Amplicon Design and Sequencing

ROY J. CARVER BIOTECHNOLOGY CENTER
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

October 6th, 2017

Chris Wright
Assistant Director of DNA Services

Alvaro Hernandez
Director of DNA Services
Outline:

- Overview of 16S sequencing and applications
- Construction of 16S and other targeted amplicons
- Sequencing Results
- Pre-QIIME questions and concepts
Microbial Community Sequencing—3 main questions:

1) Who is there?
   - metagenomic DNA sequencing
     → 16S (or other) amplicon sequencing

2) What can they do?
   → metagenomic DNA sequencing

3) What are they doing now?
   → metatranscriptome (RNA) sequencing
Applications for Microbiome Sequencing:

- Earth Microbiome
- Human Microbiome
- Disease states (obesity, diabetes)
- Home Microbiome
- Environmental surveillance, damage assessment and habitat recovery
  - Chemical waste cleanup
  - Oil Spills
- Evolutionary Studies
- Soil/ plant function impacts
- Deep well oil drilling & recovery
- Coral reef disease
- Exercise
- Companion/farm animal diets
- Waste and wastewater treatment
- Farming
- Infections Disease surveillance

• If you have a project, it likely has a microbial component!
Targeted Amplicon Sequencing:

Requires construction of an Illumina-compatible library:
- Design and identify primers
  - 16S rRNA (bacterial/archaeal)
  - 18S (eukaryotic)
  - ITS (fungal)
- Any other gene (amoA, rpoB, nosZ, nrfA, etc.)

For each sample:
- PCR
- Pool
- Submit for sequencing
- Clean up/gel purify
- Quantitate/Qualitate
Fluidigm Multi-Targeted Amplicon Library Preparation:

- Submit genomic DNA to the U of I Biotechnology Center
- Each plate amplifies 48 samples against up to 48 primers
  - 10x multiplex capable → over 23,000 PCR reactions
- For microbial communities (16S, ITS, 18S, Archaea, protozoan, etc):
  - Typical project is 2-10 primer sets
  - Up to 24 primer sets per plate
  - 1152 PCR reactions
- 1536 barcodes available
  - Pool 32 plates per sequencing run
Fluidigm PCR Construct:
Project Design: Pre-Sequencing & Sequencing Questions

- What primer-sets to use?
- How many samples?
- How much sequencing do I need?

- DNA Services/Functional Genomics can help with:
  - Primer set choices and testing.
  - Options for sequencing.
  - Costs across different design options.

- Researcher should examine:
  - What primers are researchers in this field using?
  - Are you comparing to previously published data?
  - How much sequencing will your funding allow?
### Primer Options:

- **Ever-growing list!**
- **Includes 16S, 18S, ITS, archaeal, functional gene targets, more.**
- **Functional Genomics lab will order and test any primer-set that is not on the list.**

<table>
<thead>
<tr>
<th>Set</th>
<th>Primer Set Name</th>
<th>Forward Primer Name</th>
<th>Forward Primer Sequence</th>
<th>Reverse Primer Name</th>
<th>Reverse Primer Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>V3_V5_F357_R926</td>
<td>V3_F357</td>
<td>CCTACGGGAGGCAGCAG</td>
<td>V3_V5_R926</td>
<td>CCCTCAATTCMTTTRAGT</td>
</tr>
<tr>
<td>2</td>
<td>V1_V3_F28_R519</td>
<td>V1_F28</td>
<td>GAGTTTGATCNTGGCTCAG</td>
<td>V1_V3_R519</td>
<td>GTNTACNGCGGCKGCTG</td>
</tr>
<tr>
<td>3</td>
<td>Arch349F_806R</td>
<td>Arch349F</td>
<td>GYGASCAGKCGMGAAW</td>
<td>Arch806R</td>
<td>GGACTACVSGGGTATCTAAT</td>
</tr>
<tr>
<td>4</td>
<td>F566Euk_R1200</td>
<td>F566Euk</td>
<td>CAGCAGCCGCGTAAATTCC</td>
<td>R1200Euk</td>
<td>CCCGTGTTGAGTCAAATTAAGC</td>
</tr>
<tr>
<td>5</td>
<td>ITS1_ITS4</td>
<td>ITS1</td>
<td>TCCGTAGGTTGAACCTGCGG</td>
<td>ITS4</td>
<td>TCCTCCGCTTTATTGATATGC</td>
</tr>
<tr>
<td>6</td>
<td>ITS3_ITS4</td>
<td>ITS3</td>
<td>GCATCGATGAAGAAGCAGCAG</td>
<td>ITS4</td>
<td>TCCTCCGCTTTATTGATATGC</td>
</tr>
<tr>
<td>7</td>
<td>Euk_1391f_EukBr</td>
<td>Euk_1391f</td>
<td>GTACACACGCACGCAGC</td>
<td>EukBr</td>
<td>TGATCCTTCTGACAGGTTACCTAC</td>
</tr>
</tbody>
</table>
Sequencing Options:

- **MiSeq v3**
  - Options for 150nt, 250nt or 300nt read length
  - Output of 1M, 10M, or 20M paired-reads per run

- **HiSeq 2500**
  - Options for 150nt or 250nt read length
  - Output of 100-150M paired-reads per lane

- Cost from $785 to $3,930 per lane
Outline:

- Overview of 16S sequencing and applications
- Construction of 16S and other targeted amplicons
- Sequencing Results
- Pre-QIIME questions and concepts
Sequencing Results:

- Data (fastq) posted to password-protected FTP site
- Six-page Excel report
- Quality Score graphs for all samples & all primer sets
- Sequencing results (fastq) processed in 4 ways:
  - Unsorted
  - Demultiplexed
  - Primer Sorted
  - Primer Sorted & Demultiplexed
Sequencing Results, Run Metrics tab:

1) **Unsorted data**: This is bulk read data. It is not sorted, trimmed, demultiplexed, or processed in any way other than to remove PhiX reads.

   ➔ *Shows total number of reads generated for the project.*

2) **Demultiplexed data**: This data is sorted by index and PhiX reads are removed. No other processing is done on this data set.

   ➔ *Shows how well individual samples worked across all the primer-sets.*
Sequencing Results, Primer Sorted Metrics Tab:

3) **Primer Sorted data**: This data is sorted by PCR-specific primers and PhiX reads are removed. No other processing is done on this data set.
   - Shows number of reads for each primer-set
   - QIIME input files, maybe...

4) **Primer Sorted and Demultiplexed data**: This data is sorted by PCR-specific primers and then by index. PhiX reads are removed. No other processing is done on this data set.
   - This is the fully sorted and demultiplexed data for each sample and each primer.
Sequencing Results, Barcode File tab:

- Shows the sample ID and associated barcode (index).
- This should be added to your sample metadata mapping file.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Barcode Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CACGAAGAGC</td>
</tr>
<tr>
<td>7</td>
<td>CATACCTGAT</td>
</tr>
<tr>
<td>13</td>
<td>GACGTGCTTC</td>
</tr>
<tr>
<td>19</td>
<td>ATTTGGAGT</td>
</tr>
<tr>
<td>25</td>
<td>TCTGGTCTCA</td>
</tr>
<tr>
<td>31</td>
<td>AGGTAAGAGG</td>
</tr>
<tr>
<td>37</td>
<td>TCCTGACAGA</td>
</tr>
<tr>
<td>43</td>
<td>GCACGTGGTC</td>
</tr>
<tr>
<td>2</td>
<td>ACCATGAGTC</td>
</tr>
<tr>
<td>8</td>
<td>AATGCAGTGT</td>
</tr>
<tr>
<td>14</td>
<td>ATATGGTGGAA</td>
</tr>
<tr>
<td>20</td>
<td>ACTCAGTTAC</td>
</tr>
<tr>
<td>26</td>
<td>AAGTGCAGATG</td>
</tr>
<tr>
<td>32</td>
<td>CCACAGAGTG</td>
</tr>
<tr>
<td>38</td>
<td>AGTGTTGCAC</td>
</tr>
<tr>
<td>44</td>
<td>ACTTTCTAGC</td>
</tr>
<tr>
<td>3</td>
<td>GCCACATATA</td>
</tr>
</tbody>
</table>
Sequencing Results, Metadata file:

Metadata from a text file

Metadata is typically defined in a sample (or feature; more on that below) metadata mapping file. The QIIME 2 development team hasn’t adopted a standard set of criteria for the sample metadata mapping file, but at present the following minimum requirements are enforced:

- The file must be a tab-separated text file (TSV).
- Comment lines (i.e. lines that begin with #) may appear anywhere in the file and are ignored.
- Blank lines (i.e. empty or whitespace-only lines) may appear anywhere in the file and are ignored.
- The first non-comment, non-blank line of the file is used as the header (i.e. column labels). See note below if you’re using a QIIME 1 mapping file.
- The column labels must be unique (i.e. no duplicate values) and cannot contain certain special characters (e.g. /, \, *, ?, etc.).
- There must be at least one line of data in addition to the header.
- The first column in the table is the “identifier” column (either sample ID or feature ID, depending on the axis).
- All of the values in the first column must be unique (i.e. no duplicate values) and cannot contain certain special characters (e.g. /, \, *, ?, etc.).

Sample (and feature) metadata mapping files can be validated using Keemei, which will help identify issues while creating these files. Select Add-ons > Keemei > Validate QIIME 2 mapping file.
Sequencing Results, Sample-Primer tab:

- Shows trace files of final Fluidigm Pools

- This is like an agarose gel, with peaks equating to bands and upper and lower markers equating to a ladder.
Sequencing Results, Sample-Primer tab:

- Lists locus-specific primers used on the genomic DNA.

<table>
<thead>
<tr>
<th>Primer target:</th>
<th>Primer name:</th>
<th>Locus-specific primer sequence:</th>
<th>Expected length (nt) of amplified region including locus-specific primer sequences:</th>
<th>Expected total fragment length (nt) with index and Illumina adaptors:</th>
</tr>
</thead>
<tbody>
<tr>
<td>16S_V3-V5</td>
<td>16s v3-v5-F357</td>
<td>5'-CCCTACGGGAGGCAGCAG</td>
<td>570</td>
<td>676</td>
</tr>
<tr>
<td></td>
<td>16s v3-v5-R926</td>
<td>5'-CCGTCAATTCTTTTATTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITS1-ITS4</td>
<td>ITS1F-10</td>
<td>5'-CCCTACGGGAGGCAGCAG</td>
<td>474+ variable</td>
<td>580+ variable</td>
</tr>
<tr>
<td></td>
<td>ITS4R-10</td>
<td>5'-CCGTCAATTCTTTTATTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMV4.5NF_AMDGR</td>
<td>AMV4.5NF</td>
<td>5'-CCCTACGGGAGGCAGCAG</td>
<td>~254 (extrapolated from final product on bioanalyzer)</td>
<td>~360 (from bioanalyzer)</td>
</tr>
<tr>
<td></td>
<td>AMDGR</td>
<td>5'-CCGTCAATTCTTTTATTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2_R6</td>
<td>F2</td>
<td>5'-CCCTACGGGAGGCAGCAG</td>
<td>~384 (extrapolated from final product on bioanalyzer)</td>
<td>~490 (from bioanalyzer)</td>
</tr>
<tr>
<td></td>
<td>R6</td>
<td>5'-CCCTACGGGAGGCAGCAG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sequencing Results, Sample-Primer tab:

- Lists CS1 and CS2 Fluidigm-specific pad sequences.
- Lists the Illumina i5 and i7 sequences.

<table>
<thead>
<tr>
<th>Sequences of Fluidigm CS1 and CS2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS1</td>
</tr>
<tr>
<td>CS2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sequences of Illumina i5 and i7 regions:</th>
</tr>
</thead>
<tbody>
<tr>
<td>i5</td>
</tr>
<tr>
<td>i7</td>
</tr>
<tr>
<td>(XXXXXXXXXX denotes index sequence placement)</td>
</tr>
</tbody>
</table>

- Lists the CS sequences to TRIM from 3’ ends.

CS sequences to trim from 3' ends of reads (and everything that appears after these sequences):
From Read 1: AGACCAAGTCTCTGC
From Read2: TGTAGAACCATGTC
Sequencing Results, Sample-Primer tab:

- Shows graphic of Fluidigm amplicon structure.
Sequencing Results, FTP-information tab:

- Details how to log into the FTP site and download data.
- TIPS:
  - Do not download to your PC or laptop—files are generally too big.
  - Do download to a server.
  - Do make at least 2 copies of your data (give one to your PI!)
  - Archive system available thru HPCBio ($200/TB/10 years)
Outline:

- Overview of 16S sequencing and applications
- Construction of 16S and other targeted amplicons
- Sequencing Results
- Pre-QIIME questions and concepts
Data Handling: Processing from the instrument

- QIIME favors bulk (non-demultiplexed) files.
- You should use the Primer Sorted fastq files.
- Each primer set must be processed separately.

- Dual-indexed samples:
  - These CAN be processed in QIIME.
  - Index read sequences are merged by DNA lab to produce one index read fastq file.
Data Handling: Trimming Data

- Posted fastqs are not trimmed for adaptors or primers
- Posted fastqs are not trimmed for quality
- You should trim adaptors and primers from reads before doing any QIIME processing.
- If sequencing read includes forward locus-specific primer, trim it after data is binned by primer.
- If sequencing read includes reverse locus-specific primer, trim it and all sequence after it.
Data Handling: Merging Overlapping Paired-Reads

- Can only pair if reads overlap—check read lengths against the length of your PCR product.
- Pair after adaptor & primer & quality trimming!
- Multiple options within and outside of QIIME (PEAR; QIIME tool)
Data Handling: Starting to QIIME:

- **Mapping File**
  - Contains all desired sample metadata.
  - Required: sample ID and barcode sequence.

- **Fastq files: Sorted or unsorted**
  - Should be primer/quality trimmed
  - Should have reads merged if needed.
Getting Started with Sequencing:

1) Chemistries, library kits, yields, protocols, turnaround times, instruments, etc. change every few weeks, months, years...

Read papers, but get in touch with us before you start. Things change....FAST!

2) Submission form for genomic DNA samples for Fluidigm amplification and sequencing are available through the Functional Genomics Lab

call or email for an appointment with us.
THANKS!

Functional Genomics Lab
(Fluidigm)
Mark Band, Director
markband@illinois.edu
356 ERML
217-244-3930

www.biotech.illinois.edu/
functionalgenomics

DNA Services Lab
/Library Prep & Sequencing/
Alvaro Hernandez, Director
aghernan@illinois.edu
329 ERML
217-244-3480

Chris Wright, Assistant Director
clwright@illinois.edu
334 ERML
217-333-4372

www.biotech.illinois.edu/htdna