Canadian Bioinformatics Workshops

www.bioinformatics.ca
Module 2
Marker Gene Based Analysis

William Hsiao
Analysis of Metagenomic Data
June 24-26, 2015

Universal phylogenetic tree based on SSU rRNA sequences
-N. Pace, Science, 1997

Learning Objectives of Module

At the end of this module, you will be able to:

• **Understand and Perform** marker based microbiome analysis
• **Analyze** 16S rRNA marker gene data
• **Use** 16S rRNA marker gene to profile and compare microbomes
• **Select** suitable parameters for marker gene analysis
• **Explain** the advantages and disadvantages of marker based microbiome analysis
rRNAs – the universal phylogenetic markers

- Ribosomal RNAs are present in all living organisms
- rRNAs play critical roles in protein translation
- rRNAs are relatively conserved and rarely acquired horizontally
- Behave like a molecular clock
  - Useful for phylogenetic analysis
  - Used to build tree-of-life (placing organisms in a single phylogenetic tree)
- 16S rRNA most commonly used
rRNA – the lens into Life

• A tool for placing organisms on a phylogenetic tree
  – Tree of life
• A tool for understanding the composition of a microbial community
  – Profiling of a community (alpha-diversity)
• A tool for relating one microbial community to another
  – Comparison of communities (beta-diversity)
• A tool for reading out the properties of a host or environment
  – Obese vs. lean; IBD vs. no IBD (classification)

-McDonald et al, RNA, 2015

Other Marker Genes Used

• Eukaryotic Organisms (protists, fungi)
  – 18S (http://www.arb-silva.de)
  – ITS (http://www.mothur.org/wiki/UNITE_ITS_database)
• Bacteria
  – CPN60 (http://www.cpndb.ca/cpnDB/home.php)
  – ITS (Martiny, Env Micro 2009)
  – RecA gene
• Viruses
  – Gp23 for T4-like bacteriophage
  – RdRp for picornaviruses

Faster evolving markers used for strain-level differentiation
DNA Extraction

• Many protocols/kits available but kit-bias exists
• HMP and other major projects standardized DNA extraction protocol -
  http://www.earthmicrobiome.org/emp-standard-protocols/dna-extraction-protocol/
• DNA Extraction can also be done after fractionation to separate out different organisms based on cell size and other characteristics -

Contaminations

Lab:
• Extraction process can introduce contamination from the lab
• Include Extraction Negative Control in your experiments!
• Especially crucial if samples have low DNA yield!

Host / Environment:
• Host DNA often ends up in the microbiome sample
• Unwanted fractions (e.g. eukaryotes) can be filtered by cell size-selection prior to DNA extraction
• Unwanted DNA can be removed by subtractive hybridization
Target Amplification

- PCR primers designed to amplify specific regions of a genome
  - In a “dirty” sample, inhibitors can interfere with PCR
  - In a complex sample, non-specific amplification can occur
  - Check your products using BioAnalyzer or run a gel
- Gel size selection may be necessary to clean up the PCR products
- Dilution may be necessary to reduce inhibition
Target Selection and Bias

- 16S rRNA contains 9 hypervariable regions (V1-V9)
- V4 was chosen because of its size (suitable for Illumina 150bp paired-end sequencing) and phylogenetic resolution
- Different V regions have different phylogenetic resolutions – giving rise to slightly different community composition results

Sample Multiplexing

- MiSEQ capacity (~20M paired-end reads) allows multiple samples to be combined into a single run
  – Number of reads needed to differentiate samples depends on the nature of the studies
- Unique DNA barcodes can be incorporated into your amplicons to differentiate samples
- Amplicon sequences are quite homogeneous
  – Necessary to combine different amplicon types or spike in PhiX control reads to diversify the sequences
One Step vs. Two Step Amplification

<table>
<thead>
<tr>
<th>One step amplification</th>
<th>Two step amplification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barcodes and sequencing adaptors included in the amplicon primers</td>
<td>Barcodes and adaptors separated from the amplicon primers</td>
</tr>
<tr>
<td>Single step PCR reaction</td>
<td>PCR followed by DNA-ligation or more PCR</td>
</tr>
<tr>
<td>Long primers; barcodes can interfere with amplification primers</td>
<td>Target primers tested independently from barcodes/adaptors</td>
</tr>
<tr>
<td>Not suitable for degenerate primer amplification</td>
<td>Compatible with random and degenerate primer amplification</td>
</tr>
<tr>
<td>Suitable if working with one amplicon type from many samples</td>
<td>Suitable if working with many types of amplicons or metagenomics in a single study</td>
</tr>
<tr>
<td>Rapid protocol and little loss of biomaterial</td>
<td>Longer protocols with more biomaterial loss</td>
</tr>
<tr>
<td>Barcoded primers for 16S available from Rob Knight’s Lab</td>
<td>Tested barcodes from Illumina, BioO, NEB, etc. available</td>
</tr>
</tbody>
</table>

Available Marker Gene Analysis Platforms

- QIIME ([http://qiime.org](http://qiime.org))
- Mothur ([http://www.mothur.org](http://www.mothur.org))
## Comparison Between QIIME and Mothur

<table>
<thead>
<tr>
<th>QIIME</th>
<th>Mothur</th>
</tr>
</thead>
<tbody>
<tr>
<td>A python interface to glue together many programs</td>
<td>Single program with minimal external dependency</td>
</tr>
<tr>
<td>Wrappers for existing programs</td>
<td>Reimplementation of popular algorithms</td>
</tr>
<tr>
<td>Large number of dependencies / VM available</td>
<td>Easy to install and setup; work best on single multi-core server with lots of memory</td>
</tr>
<tr>
<td>More scalable</td>
<td>Less scalable</td>
</tr>
<tr>
<td>Steeper learning curve but more flexible workflow if you can write your own scripts</td>
<td>Easy to learn but workflow works the best with built-in tools</td>
</tr>
</tbody>
</table>

## Overall Bioinformatics Workflow

1. **Preprocessing:** remove primers, demultiplex, quality filter, decontamination
2. **OTU Picking / Representative Sequences**
3. **Taxonomic Assignment**
4. **Build OTU Table (BIOM file)**
5. **Sequence Alignment**
6. **Phylogenetic Analysis**
7. **Downstream analysis and Visualization – knowledge discovery**

**Inputs:**
- Sequence data (fastq)
- Metadata about samples (mapping file)

**Outputs:**
- OTU Table
- Phylogenetic Tree
- Processed Data
1) Preprocessing

**Decontamination (in-silico)**

- Decontamination is typically done by mapping reads to databases containing suspected contaminants (host sequences, known contaminating sequences, etc.)
  - BLAST could be used for this purpose, but slow
  - Short-read aligners (SRA) most commonly used
- No decontamination tools can achieve 100% sensitivity or specificity
- In our own analysis, most SRAs performed similarly
- More important is the database used for the search
  - Include any host variants in your database to improve matching

1) Preprocessing

**Sample de-multiplexing**

- MiSEQ capacity (~20M paired-end reads) allows multiple samples to be combined into a single run
- Reads need to be linked back to the samples they came from using the unique barcodes
  - Error-correcting barcodes available (~2000 unique barcodes from *Caporaso et al ISME 2012*)
- De-multiplexing removes the barcodes and the primer sequences
Quality Filtering

- QIIME filtering by:
  - Minimal length of consecutive high-quality bases (as % of total read length)
  - Maximal number of consecutive low-quality bases
  - Maximal number of ambiguous bases (N’s)
  - Minimum Phred quality score

- Other quality filtering tools available
  - Cutadapt (https://github.com/marcelm/cutadapt)
  - Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic)
  - Sickle (https://github.com/najoshi/sickle)

- Sequence quality summary using FASTQC
2) OTU Picking

**OTU Picking**

- OTUs: formed arbitrarily based on sequence identity
  - 97% of sequence similarity ≈ species
- QIIME supports 3 approaches (each with several algorithms)
  - De novo clustering
  - Closed-reference
  - Open-reference
- More details about OTU picking

**De novo Clustering**

- Groups sequences based on sequence identity
- Pair-wise comparisons needed
- Hierarchical clustering commonly used
- Requires a lot of memory (hundreds Gb) and time
- Greedy algorithm using single centroids to save time (uclust)
- Suitable if no reference database is available

Navas-Molina et al. 2013, Meth. Enzym
Closed-Reference

- Matches sequences to an existing database of reference sequences
- Unmatched sequences discarded
- Fast and can be parallelized
- Suitable if accurate and comprehensive reference is available
- Allow taxonomic comparison across different markers
- Novel organisms missed/discarded

Navas-Molina et al. 2013, Meth. Enzym

Open-Reference

- First matches sequences to reference database (like closed-reference)
- Unmatched sequences then go through de-novo clustering
- Best of both worlds
- Suitable if a mixture of novel sequences and known sequences is expected
- Recommended approach, in general

Navas-Molina et al. 2013, Meth. Enzym
Representative Sequence

- Once sequences have been clustered into OTUs, a representative sequence is picked for each OTU.
- This means downstream analyses treats all organisms in one OTU as the same!
- Representative sequence can be picked based on
  - Abundance (default)
  - Centroid used (open-reference OTU or de novo)
  - Length
  - Existing reference sequence (for closed-reference OTU picking)
  - Random
  - First sequence

Chimera Sequence Removal

- PCR amplification process can generate chimeric sequences (an artificially joined sequence from >1 templates)
- Chimeras represent ~1% of the reads
- Detection based on the identification of a 3-way alignment of non-overlapping sub-sequences to 2 sequences in the search database

Overall Bioinformatics Workflow

1) Preprocessing: remove primers, demultiplex, quality filter, decontamination
2) OTU Picking / Representative Sequences
3) Taxonomic Assignment
4) Build OTU Table (BIOM file)
5) Sequence Alignment
6) Phylogenetic Analysis
7) Downstream analysis and Visualization – knowledge discovery

Module 2

3) Taxonomic Assignment

Taxonomic Assignment

- OTUs don’t have names but humans have the desire to refer to things by names.
  - we want to refer to a group of organism as Bacteroides rather than OTU1.
- Closed-reference:
  - Transfer taxonomy of the reference sequence to the query read
- Open-reference + de-novo clustering:
  - Match OTUs to a reference data using “similarity” searching algorithms
  - It is important to report how the matching is done in your publication and which taxonomy database is used
3) Taxonomic Assignment

Taxonomic Assignment

OTUs → Specify a matching algorithm → Taxonomy DBs

RDP Classifier
NCBI BLAST
rtax
tax2tree

RDP
GreenGenes
Silva

Specify a target taxonomy database

Ranked taxonomy names

Taxonomy Databases

- RDP (Cole et al 2009)
  - Most similar to NCBI Taxonomy
  - Has a rapid classification tool (RDP-Classifier)
- GreenGenes (DeSantis et al 2006)
  - Preferred by QIIME
- Silva (Quast et al. 2013)
  - Preferred by Mothur
- Choose DB based on existing datasets you want to compare to and community preference
- Caveat: only 11% of the ~150 human associated bacterial genera have species that fall in the 95%-98.7% 16S rRNA identity (Tamisier et al, IJSEM, 2015)
3) Taxonomic Assignment

**Taxonomy Summary**

Different body sites have different taxonomy composition shown as stacked bar graphs.

---

**Overall Bioinformatics Workflow**

1) Preprocessing: remove primers, demultiplex, quality filter, decontamination
2) OTU Picking / Representative Sequences
3) Taxonomic Assignment
4) Build OTU Table (BIOM file)
5) Sequence Alignment
6) Phylogenetic Analysis
7) Downstream analysis and Visualization – knowledge discovery

Inputs:
- Sequence data (fastq)
- Metadata about samples (mapping file)

Outputs:
- OTU Table
- Phylogenetic Tree
- Processed Data

Source: Navas-Molina et al 2015
4) Build OTU Table

**OTU Table**

- OTU table is a sample-by-observation matrix

<table>
<thead>
<tr>
<th></th>
<th>OTU1</th>
<th>OTU2</th>
<th>OTU3</th>
<th>OTU4</th>
<th>OTU5…</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample1</td>
<td>10</td>
<td>14</td>
<td>0</td>
<td>33</td>
<td>23</td>
</tr>
<tr>
<td>Sample2</td>
<td>5</td>
<td>0</td>
<td>54</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

- OTUs can have corresponding taxonomy information
- Extremely rare OTUs (singletons or frequencies <0.005% of total reads) can be filtered out to improve downstream analysis
  - These OTUs may be due to sequencing errors and chimeras
- Encoded in Biological observation Matrix (BIOM) format (http://biom-format.org/)

5) Sequence Alignment

**Sequence Alignment**

- Sequences must be aligned to infer a phylogenetic tree
- Phylogenetic trees can be used for diversity analyses
- Traditional alignment programs (**clustalw, muscle** etc) are slow (Larkin et al 2007, Edgar 2004)
- Template-based aligners such as **PyNAST** (Caporaso et al 2010) and **Infernal** (Nawrocki, 2009) are more suitable for large number of sequences
Phylogenetic Tree Reconstruction

• After sequences are aligned, phylogenetic trees can be constructed from the multiple alignments

• Examples:
  – FastTree (Price et al 2009) – recommended by QIIME
  – Clearcut (Evans et al 2006)
  – Clustalw
  – RAXML (Stamatakis et al 2005)
  – Muscle

Rarefaction

• As we multiplex samples, the sequencing depths can vary from sample to sample

• Many richness and diversity measures and downstream analyses are sensitive to sampling depth
  – “more reads sequenced, more species/OTUs will be found in a given environment”

• So we need to rarefy the samples to the same level of sampling depth for more fair comparison

• Alternative approach: variance stabilizing transformations
  – McMurdie and Holmes, PLOS CB, 2013
  – Liu, Hsiao et al, Bioinformatics, 2011
Alpha Diversity Analysis

- **Alpha-diversity**: diversity of organisms in one sample / environment
  - **Richness**: # of species/taxa observed or estimated
  - **Evenness**: relative abundance of each taxon
  - **α-Diversity**: taking both evenness and richness into account
    - Phylogenetic distance (PD)
    - Shannon entropy
    - dozens of different measures in QIIME

---

**Alpha Diversity (PD)**

- **Closed-reference**: Rarefaction curves different based on the OTU-picking method used. De-novo approach and Open-reference generated higher diversity than Closed-reference approach

Source: Navas-Molina et al 2015
Beta Diversity Analysis

- **Beta-diversity**: differences in diversities across samples or environments
  - UniFrac (Lozupone et al, AEM, 2005) (phylogenetic)
  - Bray-Curtis dissimilarity measure (OTU abundance)
  - Jaccard similarity coefficient (OTU presence/absence)
  - Large number of other measures available in QIIME

7) Downstream Analysis

**Beta Diversity Analysis**

- First, paired-wise distances/similarity of the samples calculated

<table>
<thead>
<tr>
<th></th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>1</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 2</td>
<td></td>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>Sample 3</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

- Then the matrix (high-dimensional data) is transformed into a lower-dimensional display for visualization
  - Shows you which samples cluster together in 2D or 3D space
  - PCoA
  - Hierarchical clustering
7) Downstream Analysis

**Beta Diversity (UniFrac)**

- all OTUs are mapped to a phylogenetic tree
- summation of branch lengths unique to a sample constitutes the UniFrac score between the samples
- shared nodes either ignored (unweighted) or discounted (weighted)
- Weighted UniFrac sensitive to experimental bias
- Unweighted UniFrac often proven more robust

**Principal Coordinate Analysis (PCoA)**

- When there are multiple samples, the UniFrac scores are calculated for each pair of samples. In order to view the multidimensional data, the distances are projected onto 2-dimensions using PCoA
7) Downstream Analysis

**PCoA**

- Each dot represents a sample
- Dots are colored based on the attributes provided in the mapping (metadata) file

Source: Navas-Molina et al 2015

---

7) Downstream Analysis

**Hierarchical Clustering**

- Each leave is a sample
- Samples “forced” into a bifurcating tree

Source: Navas-Molina et al 2015
Marker Genes vs. Shotgun Metagenomics

<table>
<thead>
<tr>
<th>Marker Gene Profiling</th>
<th>Shotgun Metagenomics Profiling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less expensive (~$100 per sample)</td>
<td>Still very expensive (~$1000 per sample)</td>
</tr>
<tr>
<td>Computational needs can be met by desktop / small server</td>
<td>Usually requires huge computational resources (cluster of computers)</td>
</tr>
<tr>
<td>Provides mainly taxonomic profiling</td>
<td>Provides both taxonomic and functional profiling</td>
</tr>
<tr>
<td>For 16S, majority of genes can be assigned at least to</td>
<td>Many more unassigned gene fragments (&quot;wasted&quot; data)</td>
</tr>
<tr>
<td>phylum level</td>
<td></td>
</tr>
<tr>
<td>Relatively free of host DNA contamination</td>
<td>Prone to host DNA contamination</td>
</tr>
</tbody>
</table>

Functional Stability vs. OTU Stability

PICRUSt

- Predicts gene content of microbiome based on marker gene survey data
- Analysis showed that in poorly-characterized community types, 16S profile works better than (predicted) functional profile to classify samples

Source: Xu et al ISME Journal 2014