ECOGEO Workshop 2:
Introduction to Env ‘Omics

Amplicon Analysis - mothur

Ben Tully
Bioinformatic Specialist
Center for Dark Energy Biosphere Investigations
University of Southern California
mothur Background - 1

• 1 of 2 majorly used pipelines for the analysis of 16S rRNA amplicon sequences – Qiime
  • iTagger (JGI)

• “mothur is currently the most cited bioinformatics tool for analyzing 16S rRNA gene sequences. …[T]he wiki and user forum and [contain] how you can use mothur to process data generated by Sanger, PacBio, IonTorrent, 454, and Illumina (MiSeq/HiSeq).”

In reference to reviewers: [T]he differences that they think are huge are largely cosmetic.”

Many of the same functions are available in both: cleaning sequences, clustering OTUs, determining alpha and beta diversities, generation of **BIOM** format documents
## Differences

<table>
<thead>
<tr>
<th></th>
<th><strong>Mothur</strong></th>
<th><strong>Qiime</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>All functions coded in to a single program in C</td>
<td>Written as a wrapper script in Python – individual aspect coded elsewhere</td>
<td></td>
</tr>
<tr>
<td>Fully open source</td>
<td>Open source – except for components that are not</td>
<td></td>
</tr>
<tr>
<td>Development progresses mainly through Schloss lab</td>
<td>Lots of collaboration – additional tools, etc</td>
<td></td>
</tr>
<tr>
<td>Parameter rich – lots of explanations about defaults</td>
<td>Parameter rich – limited explanation about details</td>
<td></td>
</tr>
<tr>
<td>No build in data viz</td>
<td></td>
<td>Built in data viz</td>
</tr>
<tr>
<td>OTU clustering – time intensive, heavily supported results</td>
<td>OTU clustering – faster, multiple options, results may vary</td>
<td></td>
</tr>
<tr>
<td>Classify with naïve Bayesian algorithm</td>
<td>Classify with USEARCH. Requires 2 top matches for assignment. DB order sensitive.</td>
<td></td>
</tr>
<tr>
<td>Database – default SILVA (Ref123), able to integrate other DBs</td>
<td>Database – default greengenes (2013), able to use other DBs</td>
<td></td>
</tr>
</tbody>
</table>
Remember both mothur and Qiime offer multi-day workshops that delve in to the minutiae and details of these methods – now less than 30 min

Based on the MiSeq Standard Operating Procedure - [http://mothur.org/wiki/MiSeq_SOP](http://mothur.org/wiki/MiSeq_SOP)
Reducing sequencing & PCR errors

- make.contigs
  - Schloss lab uses fully overlapping forward and reverse reads to increase fidelity of results

- Remove all sequences that have ambiguities and cannot exist in reality (too long after complete overlap join)
Align and trim

- Align sequences to the reference database
- Trim sequences outside of target hypervariable region(s)
- Important to know the position of primers on the fully expanded 16S rRNA alignment!
Essential mothur - 4

Reduce complexity of sequences

- Combine all identical sequences
- Remove chimeric sequences
- Identify sequences aligned to lineages that “cannot” be in your sample → Chloroplast, Eukaryote, etc.
Essential mothur - 5

Make OTUs

- Cluster sequences into OTUs – 97% similarity
- Produce OTU table outputs
- Assign OTUs a putative taxonomy
Analysis

- Alpha diversity - mean species diversity (no. of species present)

- Broad comparisons between samples and observed richness
  - Rarefaction – counts the number of observed OTUs as the number of sequences analyzed increases
  - Chao 1 – creates an estimate that increases the weight of the presence of rare OTUs
  - Inverse Simpson – “indication of the richness in a community with uniform evenness that would have the same level of diversity”
To perform rarefaction:

```
$ cd /home/c-debi/ecogeo/mothurdir
$ mothur

mothur > rarefaction.single(shared=stability.an.shared, calc=sobs, freq=100)
```

Open a second terminal window, navigate to /mothurdir, less output file
$ cut -f1,2 stability.an.groups.rarefaction > sample1.rarefaction

Creates file with data for rarefaction curve (nano)
### mothur (Hands-On) - 3

**Alpha diversity measures**

```
mothur > summary.single(shared=stability.an.shared, calc=nseqs-coverage-sobs-invsimpson, subsample=2441)
```

```$ less stability.an.groups.ave-std.summary```

<table>
<thead>
<tr>
<th>Sample</th>
<th>No. seqs used</th>
<th>Observed Species</th>
<th>Inv. Simpson</th>
<th>Confidence interval - L</th>
<th>Confidence interval - H</th>
</tr>
</thead>
<tbody>
<tr>
<td>F3D0</td>
<td>2441</td>
<td>132.734</td>
<td>25.716</td>
<td>24.093</td>
<td>27.574</td>
</tr>
<tr>
<td>F3D1</td>
<td>2441</td>
<td>127.36</td>
<td>34.539</td>
<td>32.539</td>
<td>37.032</td>
</tr>
</tbody>
</table>
What about bar graphs with relative abundance and taxonomy?

```
$ less stability.an.cons.taxonomy
```

```
OTU  Size  Taxonomy
0tu0001  12328

Bacteria(100);"Bacteroidetes"(100);"Bacteroidia"(100);"Bacteroidales"(100);"Porphyromonadaceae"(100);"Porphyromonadaceae"_unclassified(100);"Porphyromonadaceae"_unclassified(100);
```

```
mothur > sub.sample(shared=stability.an.shared, size=2241)
```

```
$ less stability.an.0.03.subsample.shared
```

```
<table>
<thead>
<tr>
<th>label</th>
<th>Group</th>
<th>numOtus</th>
<th>Otu0001</th>
<th>Otu0002</th>
<th>Otu0003</th>
<th>Otu0004</th>
<th>Otu0005</th>
<th>Otu0006</th>
<th>Otu0007</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>F300</td>
<td>296</td>
<td>181</td>
<td>116</td>
<td>132</td>
<td>201</td>
<td>164</td>
<td>132</td>
<td></td>
</tr>
</tbody>
</table>
Unrelated data
Analysis

- Beta diversity - measured differences between communities
- How related are communities to each other?
  - Theta YC (community structure) – considers shared OTUs and relative abundance of OTUs
  - Jaccard (community membership) – compared which OTUs are present in samples
Beta diversity measures

```
mothur > dist.shared(shared=stability.an.shared, calc=thetayc-jclass, subsample=2241)
```

Turn distance measure output in to a dendrogram

```
mothur > tree.shared(phylip=stability.an.thetayc.0.03.lt.ave.dist)
mOTHUR > quit()
```

```
$ less stability.an.thetayc.0.03.lt.ave.tre
```

Copy content

```
$ Dendroscope
```

Edit, paste
mothur (Hands-On) - 7
Statistical significance

- Involves classifying your samples into categories → good experimental design would have classify these samples before seeing results
  - E.g. surface vs deep samples or winter vs. summer

- Generate points to plot Principal Component Analysis or Non-metric Dimensional Scaling plots

- Perform AMOVA statistical tests