What is an OTU?

Operational Taxonomic Units
a.k.a. phylotypes

aggregations of reads
based only on sequence similarity, independent of taxonomic name
Why OTUs?
Because we can!

We can’t:
• do whole genome sequencing of communities
• assign complete taxonomic names
• derive phylogenetic trees from millions of short tags

Need to develop new methods that are more biologically or evolutionarily meaningful, like oligotyping

Come back in 10 years…
## Pros and Cons

<table>
<thead>
<tr>
<th>OTUs</th>
<th>Taxonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel organisms</td>
<td>Universal names, not project-specific</td>
</tr>
<tr>
<td>Insufficient taxonomy</td>
<td>Meaning associated with names</td>
</tr>
<tr>
<td>Many names based on phenotype rather than genotype</td>
<td>Independent of clustering width and algorithm</td>
</tr>
<tr>
<td>Does not lump together all order or family-level classifications</td>
<td></td>
</tr>
</tbody>
</table>
Primary Aggregation Methods

Complete linkage - no two sequences are farther apart than the clustering width

Average linkage - the average distance between sequences is less than the width

Greedy clusters – sequences are less than the width to OTU representatives built de novo

Reference – sequences are less than the width to universal reference set of OTU representatives

Oligotyping – sequences are equivalent at most information-rich nucleotide positions
Cluster “Width”

**Diameter**
Sequences are never more than $D$ apart. (CL)

**Radius**
Sequences are never more than $R$ from seed or reference. (Ref, Greedy)

**Variable**
Only selected positions (Oligo)

Averaged radii (AL)
No two sequences in an OTU are >W apart
Average Linkage

Average of all pairwise distances in an OTU are $\leq W$
Greedy Clustering

Sort by abundance

First $Seq_1$ is seed of $Cluster_1$

Compare next $Seq$ to each cluster in order:

  if $\text{Distance}(Seq, Cluster) \leq W$, include it in the cluster

if $Seq$ is not a member of any cluster, create a new cluster

Greedy Clustering

Most abundant reads:
• are likely true sequences
• least likely to be errors or chimeras, and
• most likely to spawn errors and chimeras

Seed OTUs with most abundant,
Assign variation to most abundant source.
Greedy Seeds (UClust)

Each tag goes to OTU with nearest representative, not OTU with the nearest sequence
Greedy Seeds (UClust)

When equidistant, tag goes to OTU the largest number of sequences.
Closed Reference OTUs

1. Create a full-length tree of reference sequences (greengenes)
2. Cluster reference sequences
3. Select representative sequence for each cluster
4. Map each tag to nearest rep $\leq$ Width
5. Bin each tag to that reference’s OTU#, or *unclustered* if $> \text{Width}$

Open Reference OTUs

1. Perform Closed Reference OTU mapping

2. Perform *de novo* clustering on unclustered sequences

3. Subsequent datasets can be mapped to the reference and to an expanding set of de novo clusters.
Cascading Reference OTUs

1. Cluster to reference sequences at 99% (ITS dynamic)

2. Cluster to reference sequences at 97%

3. De novo cluster remaining sequences

Oligotyping

Alignment positions of greatest information

Oligotypes identical in selected positions
Ignore the rest as noise

Pros and Cons

• **Greedy, AL** - standard de novo methods, abundance dependent, project-specific

• **Reference OTU** – compare V-regions, universal IDs, dependent on reference database and reference clustering

• **Oligotyping** – increased sensitivity, no specific radius, project-specific
Primary Clustering Software

• UCLUST (USEARCH)
  Edgar (2010) *Bioinformatics*

• Oligotyping, MED

• Open/Closed Reference OTU (QIIME)

• Average Linkage with preclustering (mothur)
The Problem of Inflation

Clustering algorithms return more OTUs than predicted for mock communities.

OTU inflation leads to:
• alpha diversity inflation
• beta diversity inflation

Where does this inflation come from?
How can we adjust for this inflation?
Different clustering algorithms have very different effects on the size and number of OTUs created…
Inflation in Action

>200,000 V6 reads of *E. coli* clone (454)

1,042 is a few more than the expected 2

## Sequencing Noise in *E. coli*

<table>
<thead>
<tr>
<th>Distance</th>
<th>Tags</th>
<th>Seqs</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>177,697</td>
<td>2</td>
<td>82.4%</td>
</tr>
<tr>
<td>0.02</td>
<td>33,425</td>
<td>633</td>
<td>97.9%</td>
</tr>
<tr>
<td>0.03</td>
<td>3738</td>
<td>2268</td>
<td>99.6%</td>
</tr>
<tr>
<td>0.04</td>
<td>2</td>
<td>1</td>
<td>99.6%</td>
</tr>
<tr>
<td>0.05</td>
<td>635</td>
<td>306</td>
<td>99.9%</td>
</tr>
<tr>
<td>0.06</td>
<td>38</td>
<td>25</td>
<td>99.9%</td>
</tr>
<tr>
<td>0.07</td>
<td>89</td>
<td>72</td>
<td>100.0%</td>
</tr>
<tr>
<td>0.08</td>
<td>22</td>
<td>17</td>
<td>100.0%</td>
</tr>
<tr>
<td>0.09</td>
<td>0</td>
<td>0</td>
<td>100.0%</td>
</tr>
<tr>
<td>0.10</td>
<td>4</td>
<td>4</td>
<td>100.0%</td>
</tr>
<tr>
<td>&gt; 0.10</td>
<td>23</td>
<td>16</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Cluster at 3% using only tags within 3% of the “correct” sequence.
Inflation with small sequence error

Ecoli Rarefaction

Only 129 OTUs... much better!!
Regardless of clustering algorithm, an MSA cannot fully align tags whose sequences are too divergent, more tags cause more problems.

18,156 sequences and 392 positions
**MSA Limitations**

Distance Comparison

**X-axis:** 1000 distances subsampled from MSA of 40,000 sequences

**Y-axis:** 1000 distances from MSA of only the 1,000 bacterial sequences
Each V6 variant with 2nt distance, will start a new OTU.
Each V6 variant with 1nt distance, can cluster with the seed or a 2nd variant.
Average Linkage collapses errors

Clustering:
Cluster Count: 1

Clusters tend to be heavily dominated by their most abundant sequence, which strongly weights the average and smoothes the noise.
Greedy Seeds
(UClust)

Cluster Count: 1
Clustering V6 sequences within 3% of known templates (outliers removed)

E. coli (2)    S. epidermidis (1)    Clone43 (43)

MS-CL
129 / 89 / 694

PW-CL
6 / 5 / 308

MS-AL
54 / 44 / 218

PW-AL
2 / 1 / 43

MS – Multiple Sequence Alignment
PW – Pairwise Alignment
CL – Complete Linkage clustering
AL – Average Linkage clustering
No inflation of low-error reads

Ecoli Rarefaction

Count of OTUs

Count of Reads

MS-CL  3% MS-CL  3% PW-AL

2 OTUs created!!
Low-quality sequences with errors $>W$ will create errant OTUs
Error Sequences
Create Errant OTUs

with errors back up to 277 OTUs
## Relative Inflation

<table>
<thead>
<tr>
<th>Templates</th>
<th>Expected OTUs</th>
<th>MS-CL</th>
<th>PW-AL</th>
<th>SLP/ PW-AL</th>
<th>% drop</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> (n=215,618)</td>
<td>2</td>
<td>1042</td>
<td>277</td>
<td>88</td>
<td>92%</td>
</tr>
<tr>
<td><em>S. epidermidis</em> (n=197,876)</td>
<td>2</td>
<td>1267</td>
<td>323</td>
<td>123</td>
<td>90%</td>
</tr>
<tr>
<td>Clone-43 (v6) (n=202,340)</td>
<td>43</td>
<td>2473</td>
<td>458</td>
<td>229</td>
<td>91%</td>
</tr>
<tr>
<td>Clone-43 v4-5 (232nt, n=16,673)</td>
<td>43</td>
<td>126</td>
<td>51</td>
<td>49</td>
<td>61%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Natural Samples</th>
<th>Expected OTUs</th>
<th>MS-CL</th>
<th>PW-AL</th>
<th>SLP/ PW-AL</th>
<th>% drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep-sea vent Archaea (n=63,133)</td>
<td>N/A</td>
<td>709</td>
<td>483</td>
<td>432</td>
<td>39%</td>
</tr>
<tr>
<td>English Channel (n=12,851)</td>
<td>N/A</td>
<td>1154</td>
<td>880</td>
<td>788</td>
<td>32%</td>
</tr>
<tr>
<td>Human Gut (n=15,239)</td>
<td>N/A</td>
<td>803</td>
<td>625</td>
<td>511</td>
<td>36%</td>
</tr>
<tr>
<td>Sewage (n=33,082)</td>
<td>N/A</td>
<td>2383</td>
<td>1881</td>
<td>1704</td>
<td>28%</td>
</tr>
<tr>
<td>North Atlantic Deep Water (n=15,497)</td>
<td>N/A</td>
<td>1713</td>
<td>1363</td>
<td>1254</td>
<td>27%</td>
</tr>
</tbody>
</table>

**Absolute** number of errant OTUs will increase with sample size.

**Relative** number of errant OTUs will decrease with sample complexity.
Errant Read Rate = \( f(\text{error rate, length}) \)
Errant Reads - Short

Short V6 Reads (60 nt) – original 454 and HiSeq

If error rate = 0.001

Then *theoretically*:

- perfect reads = 94.2%
- exactly 1 error = 5.66%
- 0 or 1 errors = 99.8%
- 2 or more errors = 0.17%

Errant read = 2 errors  \[3\% \text{ (OTU width)} \times 60 \text{ nt} = 1.8\]

0.17\% \times 500,000 \text{ reads} = \textbf{852 errant reads}
Errant Reads – Long

Longer Reads (300 nt) - MiSeq
If error rate = 0.001
Then theoretically:

- perfect reads = 74.1%
- exactly 1 error = 22.2%
- 0 or 1 errors = 96.3%
- 2 or more errors = 0.37%
- 9 or more errors = 3.7E-9%

Errant read = 9 errors  [3% (OTU width) * 300 nt = 9]
3.7E-9 % * 1,000,000 reads = \textbf{3.7E-3% errant reads}
The Magical 3%

3% SSU OTUs = Species

and

6% SSU OTUs = Genera
History of The 3% OTUs

Uses genome-level phylogeny to define species, using 70% DNA-DNA reassociation.

70% DNA-DNA ~ 97% similarity of full-length 16S

Therefore…
97% similarity of any hypervariable region in any bacteria = species
Assessment of clustering quality.

Fig. 1. Cumulative fraction of taxa that had a specified maximum intrataxon distance (A) and total branch length (B) for each taxonomic level.

To overcome the challenge of assessing when full-length 16S rRNA gene sequences were analyzed. At each taxonomic level, sequences that did not affiliate with a known lineage (i.e., incertae sedis) were excluded. The numbers in parentheses next to the name of each taxonomic level indicate the number of taxa within that level.

What are appropriate distance thresholds for an OTU-based analysis?
What is the right cluster width?

Distances vary by hypervariable region
- Full-length does not equal V1-V3
- which does not equal V3-V5
- which does not equal V6, etc.

Distance vary by bacterial lineage

What specificity are you trying to gain?

No single clustering width is the right answer, use several (and wave your hands a lot) or use other methods like Oligotyping
Clustering Questions

• How meaningful are clusters functionally?
• When is an *errare* rare and when is it an error?
• Should it be included in an existing cluster or start its own?
• How to assign sequences if OTUs overlap?
• What is the effect of residual low quality data or chimeras?
• How sensitive are alpha and beta diversity estimates to clustering results?