The DADA2 Method
The amplicon inference problem

Infer the sample types and abundances \{(s, a)\} from error-ful amplicon reads \{r\}. 
sample sequences

amplicon reads

Errors

DADA2
sample sequences

amplicon reads

Errors

OTUs

Make OTUs
Motivation: Lingering problems with “OTU”

imagine sequencing reads streaming from a single true sequence...
The deeper you sequence, the more you expect to find reads outside the radius by chance.

$r = 3\%$

Motivation: Lingering problems with “OTU”
sample sequences → amplicon reads → OTUs

Errors

DADA2

Make OTUs

Error Model
An Error Model

s: ATTAACGAGATTATAACCAGAGTACGAATA...
   |       |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
An Error Model

\[
p(r | s) = \prod_{i=1}^{L} p(r(i) | s(i), q_r(i), Z)
\]

\[s: \text{ATTAACGAGATTATAACCAGAGTACGAATA...} \]
\[r: \text{AT} \boxed{\text{CA}} \text{ACGAGATTATAACA} \text{AAGAGTACGAATA...} \]
An Error Model

$s$: ATTAACGAGATTATAACCAGAGTACGAATA...

$r$: AT\textcolor{red}{C}AACGAGATTATAAC\textcolor{red}{A}AGAGTACGAATA...

\[
p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)
\]

Error process is independent across nucleotides.

Per-nucleotide transition rate depends on:
- Sample nucleotide
- Read nucleotide
- Read quality at that position
- Batch effect (eg. run)
The true shape of an error cloud reads:

Effective Hamming Distance

NOT AN ERROR

BIOLOGY

OTU
The true shape of an error cloud reads:

```
0 1 2 3 4 5 6 7

NOT AN ERROR

BIOLOGY

ERROR

DADA2
```

Effective Hamming Distance

reads

```
0 1 2 3 4 5 6 7
```

```
0 5 10 15 20 25
```
Learning Errors
Study A

But How?
Initial guess: one real sequence + errors
Infer initial error model under this assumption.

\[
\Pr(i \rightarrow j) =
\begin{array}{c|cccc}
   & A & C & G & T \\
\hline
   A & 0.97 & 10^{-2} & 10^{-2} & 10^{-2} \\
   C & 10^{-2} & 0.97 & 10^{-2} & 10^{-2} \\
   G & 10^{-2} & 10^{-2} & 0.97 & 10^{-2} \\
   T & 10^{-2} & 10^{-2} & 10^{-2} & 0.97 \\
\end{array}
\]
Reject unlikely error under model. Recruit errors.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.97</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>C</td>
<td>10^{-2}</td>
<td>0.97</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>G</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
<td>0.97</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>T</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
<td>10^{-2}</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Update the model.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.997</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>C</td>
<td>$10^{-3}$</td>
<td>0.997</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>G</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>0.997</td>
<td>$10^{-3}$</td>
</tr>
<tr>
<td>T</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>0.997</td>
</tr>
</tbody>
</table>
Reject more sequences under *new* model

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.997</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>C</td>
<td>10^{-3}</td>
<td>0.997</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>G</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
<td>0.997</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>T</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
<td>10^{-3}</td>
<td>0.997</td>
</tr>
</tbody>
</table>
Update model again

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.998</td>
<td>1x10^-4</td>
<td>2x10^-3</td>
<td>2x10^-4</td>
</tr>
<tr>
<td>C</td>
<td>6x10^-5</td>
<td>0.999</td>
<td>3x10^-6</td>
<td>1x10^-3</td>
</tr>
<tr>
<td>G</td>
<td>1x10^-3</td>
<td>3x10^-6</td>
<td>0.999</td>
<td>6x10^-5</td>
</tr>
<tr>
<td>T</td>
<td>2x10^-4</td>
<td>2x10^-3</td>
<td>1x10^-4</td>
<td>0.998</td>
</tr>
</tbody>
</table>
**Convergence**: all errors are plausible

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>0.998</td>
<td>1x10^{-4}</td>
<td>2x10^{-3}</td>
<td>2x10^{-4}</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>6x10^{-5}</td>
<td>0.999</td>
<td>3x10^{-6}</td>
<td>1x10^{-3}</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>1x10^{-3}</td>
<td>3x10^{-6}</td>
<td>0.999</td>
<td>6x10^{-5}</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>2x10^{-4}</td>
<td>2x10^{-3}</td>
<td>1x10^{-4}</td>
<td>0.998</td>
</tr>
</tbody>
</table>
Accuracy and Resolution
Accuracy: Simulated data

mothur (an)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>True abundance</td>
<td>TP:</td>
<td>978</td>
</tr>
<tr>
<td></td>
<td>FP:</td>
<td>272</td>
</tr>
<tr>
<td></td>
<td>FN:</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>cor:</td>
<td>0.935</td>
</tr>
</tbody>
</table>

Accuracy: Simulated data

mothur (an)

DADA2

TP: 978
FP: 272
FN: 77
cor: 0.935

TP: 1042
FP: 0
FN: 13
cor: 0.999

Accuracy: Mock community

![Bar chart showing the number of OTUs for different methods: DADA2, mothur, SUMAClust, Swarm, uclust, UPARSE. The y-axis represents the number of OTUs, ranging from 0 to 4000. The x-axis represents the different methods. DADA2 has the least number of OTUs, while uclust has the highest.]

Resolution: Petrel aDNA

QIIME: De novo

Petrel - Errors

DADA2

Petrel.1
Petrel.2
Petrel.3
Petrel.4

Credit: Kealoha Kinney, Michael Bunce, Andreanna Welch
Acknowledgements

Susan Holmes  Joey McMurdie  Michael Rosen

NSF  BIO-X  NIH
https://benjjneb.github.io/dada2/