Deblur
Rapidly Resolves Single-Nucleotide Community Sequence Patterns

Instructor: Prof. Elhanan Borenstein
Presented by: Judith Brener
Outline

• Clustering to OTUs
• Deblur (2017)
• Results (Comparison to Dada2)
• Summary
• Discussion Points
Outline

• Clustering to OTUs
  • Deblur (2017)
  • Results (Comparison to Dada2)
• Summary
• Discussion Points
Clustering to OTUs - *Why?*

- Absence of traditional systems of biological classification.
- Facilitates understanding of complex microbial communities.
- Overcoming sequencing errors.
Clustering to OTUs - Problems

• Sequences below the identity threshold cannot be differentiated.
• False positive outputs.
• Inflate diversity estimates.
• Merging OTUs.
Outline

• Clustering to OTUs

• Deblur (2017)

• Results (Comparison to Dada2)

• Summary

• Discussion Points
Deblur

a novel sub-OTU (sOTU) **denoising** method for **fast and accurate**
identification of sequences within a sample, with **single-nucleotide**
resolution, and can be used to **integrate large data sets**.
Deblur pipeline

Sample File → Sequence Trimming → Dereplication + Singletone Removal → Filtering

Biom Table ← de novo Chimeras Removal ← Deblur ← Multiple Sequence Alignment
Deblur Pipeline

Sample File
Deblur Pipeline

Sequence Trimming
Deblur Pipeline

Dereplication
Deblur Pipeline

Singletone Removal
Deblur Pipeline

Multiple Sequence Alignment
Deblur Algorithm - terms in use

• Greedy Algorithm - uses error profiles to obtain putative error-free sequences.

• Predicted error-derived reads assembled by:
  • $\beta(d)$ - upper error rate bound according to Hamming distance($d_H$).
  • $\alpha$ - mean probability of obtaining a misread.

• Reads $\{r_i\}$, corresponding Counts $\{c_i\}$ and $\{c'_i\}$ actual Counts.
Deblur Algorithm

\[ C_1 = 20000 \quad C_2 = 8500 \quad C_3 = 8000 \quad C_4 = 5500 \quad C_5 = 2100 \quad C_6 = 940 \quad C_7 = 430 \quad C_8 = 180 \quad C_9 = 90 \]

\[ d_H \]

\[ i = 1: \quad C'_1 = C_1 / (1 - \alpha) \]

For all \( j > i \):

\[ \beta = \beta(d_H) \]

\[ C_j = C_i - (C'_1 \times (1 - \beta)) \]
Deblur Algorithm

\[ i = 2: \quad C'_2 = \frac{C_2}{1 - \alpha} \]
Deblur Algorithm

\[ i = 2: \quad C'_2 = \frac{C_2}{1 - \alpha} \]
Deblur Algorithm

\[ C_1 = 20000 \quad C_2 = 8500 \downarrow \quad C_3 = 3000 \downarrow \quad C_4 = 5000 \downarrow \quad C_5 = 2500 \downarrow \quad C_6 = 580 \downarrow \quad C_7 = 430 \quad C_9 = 30 \downarrow \]

\[ i = 9: \quad C'_9 = C_9 / (1 - \alpha) \]
Deblur Algorithm

$C_1 = 20000$  $C_2 = 8500$  $C_3 = 3000$  $C_4 = 5000$  $C_5 = 2500$  $C_6 = 580$  $C_7 = 430$  $C_9 = 30$
Deblur Algorithm

Denoising complete

C₁ = 20000  C₂ = 8500  C₃ = 3000  C₄ = 5000  C₅ = 2500  C₆ = 580  C₇ = 430  C₉ = 30
Outline

• Clustering to OTUs
• Deblur (2017)
• Results (Comparison to Dada2)
• Summary
• Discussion Points
Results – methods compared to

• Dada2
  • Error model – unique for each sequence run
  • Fixing errors (combined in partition)
  • Open source

• UNOISE2
  • One-pass clustering
  • generate “zero-radius OTUs”
Results

- All three methods identified sOTUs with single-nucleotide differences.
Results

• All methods produced results that were close to the ground truth.

• Deblur’s output consist of some relative low abundance sOTUs that are not present in the GT.

performans on community from mock-3
Results

• Compared levels of stability of Deblur and DADA2 using technical replicates from a data set.

• Deblur showed greater stability than DADA2, indicating that a larger fraction of sOTUs from the first run were also identified in the second run.
Results

- Heat maps showing sOTUs (rows) in common with Deblur and DADA2, as well as those unique to Deblur and DADA2 (bottom, middle, and top rows, respectively)
Results – runtime Comparison
Outline

• Clustering to OTUs
• Deblur (2017)
• Results (Comparison to Dada2)
• Summary
• Discussion Points
Summary

• Rapid and sensitive means to assess ecological patterns.

• Like DADA2 and UNoise2, Deblur produces stable sOTUs which can achieve single-nucleotide resolution.

• Applicable in an automated fashion to large-scale sequencing data sets, and can integrate sequencing runs collected over time.
Deblur vs Dada2 – summary

• Deblur
  • Operates on each sample independently and removes errors
  • Amount of memory and time is significantly less than Dada2

• Dada2
  • Unique error Model which allows
  • more refined error correcting and identification of low abundance sequences.
Outline

• Clustering to OTUs
• Deblur (2017)
• Results (Comparison to Dada2)
• Summary
• Discussion Points
Discussion Points

• Which criteria should be examined when choosing which method to apply?

• Exploring the internet, I found Dada2 to be in greater use and with more discussions on it. What could be the reason?