16S rRNA Gene Amplicon Survey: Study Design and Case Study

Considerations for a Longitudinal Case Study of Antibiotic Treatment and Virus Infection

Scott A. Handley, PhD
Assistant Professor
Department of Pathology & Immunology
Washington University School of Medicine
Rationale

• 16S amplicon surveys are extensively used to study the mouse bacterial microbiome in a large variety of contexts
  • e.g. disease, nutrition, sociology, neuroscience, etc.

• Frequently fail due to poor study design
  • Batch effects
    • Cage, paternity/breeding, facility, origin effects
  • Co-housed survival studies (specific example)

• Statistical considerations
  • Detecting signal from noise
  • Minimize variance
  • Filtering out misbehaved data

• Many of these principles apply to other data types (RNAseq)

Image credit: Davide Bonazzi/@Salmanart

“Mouse microbes may make scientific studies harder to replicate” Kelly Servick. Science Aug 16, 2016

Today’s Case Study

Case Study: Effect of Antibiotics on Viral Pathogenesis

Cage and Mouse-to-Mouse Effects

Virus

Antibiotics

Kool-Aid

Time
Cage Effects: 14 days post-treatment (pre-infection)

**Phyla**

**Beta diversity**
Individual Mouse Isolation Schema

Ampicillin (n=30) or Kool-Aid (n=15)

Day -14  Day -11
A1  A1
A2  A2
A3  A3
A4  A4
A5  A5

Co-housed 1 Week

Virus

Day -1  Day 2  Day 4  Day 6
A1  A1  A1  A1
A2  A2  A2  A2
A3  A3  A3  A3
A4  A4  A4  A4
A5  A5  A5  A5

Survived?
No  Yes  No  Yes  No

Pre-treatment  Post-treatment  Pre-infection  Post-infection
Amplicon Surveys (Highly Opinionated!) Best-practices

It’s the classic garbage in, garbage out all over again ...
16S rRNA Amplicon Survey

**Study Design**

Environmental samples
- DNA extraction
- Genomic DNA
- PCR and sequencing
- 16S rRNA sequencing
- Sequence comparison
- Phylogenetic trees

**Laboratory**

**Bioinformatics, Ecological Analysis and Statistics**

Side note: Amplicon Surveys vs. Metagenomics

Please hold your throwing tomatoes ...
16S Amplicon Surveys vs Metagenomics?


Nayfach S., Pollard KS. Cell. Aug 25;166(5):1103-16
Most of Your Decision Will Boil Down to $$$

- Our labs per sample costs:
  - 16S = $17.50 per sample
  - Metagenome = $225.00 per sample
    - Has been estimated to be as low as $100 per sample
- Study we will discuss today: 270 samples
  - $4,725 vs. $27 - $60,750
- Other considerations:
  - Understanding analytical space
  - Data storage
What are the stages of a 16S amplicon computational workflow and how can we create optimal data for analysis?
Raw Data
- QA / QC
- Clustering
- De-replication / Counting
- Chimera Removal
- Taxonomic Assignment
- Phylogenetic Tree
- Sample QA / QC
- Taxon Filtering
- Ecological Analysis

Software
- QIIME
- MOthur
- Dada2 / R
- Phyloseq

Researcher Input
* * * * ** * * ** ** ** **** ** ** **
Sequence Clustering

16S RNA Amplicons → Amplicon Clusters

- 97% Similarity
- > 97% identical to OTU
- OTU’s are 3% different
- Ambiguous

- UCLUST
- UPARSE
- SWARM
- SUMACLUST
- OTHERS
Recognized Problems with Sequence Clustering

- **False-positives:** 1,000s of OTUs when only 10s of sequences are present
  - Due to clustering artifact / noisy sequences
    - Inflates richness (# of species)
    - Sparse matrices
- **Poor taxonomic resolution** defined by arbitrary radius (e.g. 97%)
- **Increased financial cost:** poor data efficiency
- **Increased computational cost:** Clustering is quadratic
- **Unstable:** Sequence and count frequently depend on input order
There is some hope

Open-source sequence clustering methods improve the state of the art.

~7,778 citations!!!

http://benjineb.github.io/dada2/R/SotA.html
Step 1: Initial guess. All sequences + errors

Step 2: Initial error model

Step 3: Unlikely error under model. Recruit errors. Update the model

Step 3: Reject more sequences under new model & update

Convergence: All errors are plausible

What does all of this work get you?

- Raw Data
- QA / QC
- Clustering
- De-replication / Counting
- Chimera Removal
- Taxonomic Assignment
- Phylogenetic Tree
- Sample QA / QC
- Taxon Filtering
- Ecological Analysis

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASV 1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ASV 2</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>456</td>
</tr>
<tr>
<td>ASV 3</td>
<td>112</td>
<td>101</td>
<td>98</td>
<td>10</td>
</tr>
<tr>
<td>ASV 4</td>
<td>435</td>
<td>435</td>
<td>382</td>
<td>3</td>
</tr>
<tr>
<td>ASV 5</td>
<td>76</td>
<td>83</td>
<td>68</td>
<td>145</td>
</tr>
</tbody>
</table>
- More noisy than reality
- Bad for statistical inference
- Multiple hypothesis testing
- Poorly defined, difficult to separate distributions

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU 1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>OTU 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OTU 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OTU 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASV 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ASV 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ASV 3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ASV 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Making Things Normal

**Data Transformation**

- Raw Data
- QA / QC
- Clustering
- De-replication / Counting
- Chimera Removal
- Taxonomic Assignment
- Phylogenetic Tree
- Sample QA / QC
- Taxon Filtering
- Ecological Analysis
Data Transformation

log(1 + x)  
x/sum(x)  
min(sample_sum) * x/sum(x)
Sample Outlier Detection

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASV 1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ASV 2</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>456</td>
</tr>
<tr>
<td>ASV 3</td>
<td>112</td>
<td>101</td>
<td>98</td>
<td>10</td>
</tr>
<tr>
<td>ASV 4</td>
<td>435</td>
<td>435</td>
<td>382</td>
<td>3</td>
</tr>
<tr>
<td>ASV 5</td>
<td>76</td>
<td>83</td>
<td>68</td>
<td>145</td>
</tr>
</tbody>
</table>

... n=270

... n=724
Individual Mouse Isolation Schema

**Ampicillin (n=30)**

- or **Kool-Aid (n=15)**

<table>
<thead>
<tr>
<th>Day -14</th>
<th>Day -11</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A1</td>
</tr>
<tr>
<td>A1</td>
<td>A1</td>
</tr>
<tr>
<td>A1</td>
<td>A1</td>
</tr>
</tbody>
</table>

**Virus**

<table>
<thead>
<tr>
<th>Day -1</th>
<th>Day 2</th>
<th>Day 4</th>
<th>Day 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A2</td>
<td>A1</td>
<td>A1</td>
</tr>
<tr>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
</tr>
<tr>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
</tr>
<tr>
<td>A3</td>
<td>A3</td>
<td>A3</td>
<td>A3</td>
</tr>
<tr>
<td>A3</td>
<td>A3</td>
<td>A3</td>
<td>A3</td>
</tr>
<tr>
<td>A4</td>
<td>A4</td>
<td>A4</td>
<td>A4</td>
</tr>
<tr>
<td>A4</td>
<td>A4</td>
<td>A4</td>
<td>A4</td>
</tr>
<tr>
<td>A5</td>
<td>A5</td>
<td>A5</td>
<td>A5</td>
</tr>
<tr>
<td>A5</td>
<td>A5</td>
<td>A5</td>
<td>A5</td>
</tr>
</tbody>
</table>

**Survived?**

- No
- Yes
- No
- Yes
- No

**Individual Mouse Isolation Schema**

- Pre-treatment
- Post-treatment
- Pre-infection
- Post-infection

- Co-housed 1 Week
Sample Outlier Detection – Unexpectedly Low # of Sequences
Samples that “perform” unexpectedly
Rules of Thumb for Sample Detection and Removal

• **Justify and document!!!**

• Except in extreme cases, test how sample removal alters your downstream results. Do the experiment!

• Know your data. When are you comfortable removing a sample based on your knowledge of the system

• Explore using multiple plot types

• Include enough detail to make analysis interpretable and reproducible
Understand your data better
Cleaned Data
# Feature Outlier Detection

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASV 1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>ASV 2</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>456</td>
</tr>
<tr>
<td>ASV 3</td>
<td>112</td>
<td>101</td>
<td>98</td>
<td>10</td>
</tr>
<tr>
<td>ASV 4</td>
<td>435</td>
<td>435</td>
<td>382</td>
<td>3</td>
</tr>
<tr>
<td>ASV 5</td>
<td>76</td>
<td>83</td>
<td>68</td>
<td>145</td>
</tr>
</tbody>
</table>

... n=270

n=724
Low-abundant feature removal is commonplace

• “We removed all taxa that were under 1% relative abundance and present in less than 3% of all samples.”
Sequence/Taxa Outlier Detection

*Filtering out low impact information*
Rules of Thumb for Feature Detection and Removal

• Justify and document!!!

• Except in extreme cases, test how feature removal alters your downstream results. Do the experiment!

• Know your data. When are you comfortable removing a feature based on your knowledge of the system

• Explore using multiple plot types

• Include enough detail to make analysis interpretable and reproducible
Beta Diversity Throughout the Course of the Experiment

**Colored by Cage**

- Kool-Aid
- Ampicillin
Summary

• Explore -> Document -> Test
• Does any of this really matter?
  • Sometimes?
    • Less so for community ecology measurements
    • More so for detection of differentially abundant taxa
  • Detailed exploration provides more opportunities for insights
• Don’t publish garbage data
Frequently Used 16S Analysis Techniques
Community Composition

• Broad overview
• Nothing statistical
Alpha Diversity: Richness

• Richness: Number of unique taxa (ASVs) that are observed in a sample
  • Taxonomy independent
  • Abundance independent (presence / absence)

• Loads of other Alpha diversity measures (Chao1, Shannon, Simpsons, etc.)
Richness Example
Beta Diversity

• Between sample similarity
  • Distance between one sample to all other samples
  • Multivariant
  • Can incorporate relative abundances or not
  • Most frequently displayed in an ordination plot

To learn about distance measures and ordination:
https://sites.google.com/site/mbsgustame/home
Differential Abundance Analysis

• What specific taxa are different between study groups?
  • Lots of methods
    • DeSeq2
    • Random Forest
    • LeFse
    • ANCOM
    • Gneiss
    • ...
Rest of today

• Morning: Resolve sequence variants with dada2
• Afternoon: Analyze antibiotic treated mice case study
Step 1: Initial guess. All sequences + errors

Step 2: Initial error model

Step 3: Unlikely error under model. Recruit errors. Update the model

Step 3: Reject more sequences under new model & update

Convergence: All errors are plausible

Dada2 workflow

- Select Raw Data
- QC Data
- Learn Errors
- Dereplicate
- Infer ASV