Analyzing Soil Microbial Community Sequencing Data with the CHTC

Thea Whitman - Department of Soil Science - UW-Madison
@theawhitman - github.com/theawhitman - whitmanlab.soils.wisc.edu
Key questions

• What drives organic matter (OM) cycling in soils?

• How can microbial ecology inform our understanding of OM cycling?

• How do forest fires impact the microbial community?
Two approaches

- Amplicon Sequencing
- Environmental Shotgun Metagenomics
Approach: Amplicon Sequencing

Sample (e.g. water, soil, sediment, faeces, biopsy, ...)

DNA extraction → DNA → PCR → DNA sequencing

PCR-amplified rRNA genes

rRNA gene sequences

BLAST-search rRNA sequence database with millions of taxonomically classified rRNA sequences (e.g. RDP, Silva)

Counts of OTUs per sample

OTU | Species   | Sample1 | Sample2 | Sample3 |
---|-----------|---------|---------|---------|
1  | E.coli    | 17      | 0       | 335     |
2  | S.aurus   | 231     | 11800   | 45      |
3  | unknown   | 30      | 0       | 0       |

Operational Taxonomic Units (OTUs)
Approach: Amplicon Sequencing

Typical study:
- ~200 samples
- 10k-100k seq/sample
- 3-5 Gb total

Counts of OTUs per sample

BLAST-search rRNA sequence database with millions of taxonomically classified rRNA sequences (e.g. RDP, Silva)

Operational Taxonomic Units (OTUs)
Approach: 
Amplicon Sequencing

Sample (e.g. water, soil, sediment, faeces, biopsy, ...)

DNA extraction → DNA → PCR → PCR-amplified rRNA genes → DNA sequencing

Typical study:
~200 samples
10k-100k seq/sample
3-5 Gb total

Approach:
Copy custom R installation with relevant packages (224Mb);
Run R script

<table>
<thead>
<tr>
<th>OTU</th>
<th>Species</th>
<th>Sample1</th>
<th>Sample2</th>
<th>Sample3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E.coli</td>
<td>17</td>
<td>0</td>
<td>335</td>
</tr>
<tr>
<td>2</td>
<td>S.aurus</td>
<td>231</td>
<td>11800</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>unknown</td>
<td>30</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Counts of OTUs per sample

BLAST-search rRNA sequence database
with millions of taxonomically classified rRNA sequences (e.g. RDP, Silva)
Approach: Amplicon Sequencing

Typical study:
~200 samples
10k-100k seq/sample
3-5 Gb total

QC and data processing:
~50 Gb memory
3-5 Gb disk
16 CPUs - 5h30

Approach:
Copy custom R installation with relevant packages (224Mb);
Run R script
Approach: Amplicon Sequencing

Typical study:
~200 samples
10k-100k seq/sample
3-5 Gb total

QC and data processing:
~50 Gb memory
3-5 Gb disk
16 CPUs - 5h30

Analysis and visualization:
Off-cluster in R or Jupyter R notebooks

Copy custom R installation with relevant packages (224Mb);
Run R script
Challenges: Amplicon Sequencing

• Almost ~tractable on desktop/laptop
• Working interactively vs. with scripts
• Rapidly evolving tools / best practices
• Bringing analysis into the classroom
Log2-fold change with fire (controlling for pH, vegetation, and %C in soil) using metagenomeSeq (Paulson et al. 2016)

Each point represents one OTU, sized by mean relative abundance. Faint OTUs were not significant ($p_{adj}<0.05$)
Results:
Amplicon Sequencing

A) Co-occurrence
- Positive
- Negative

Fire Response
- Positive
- Negative
- Neutral

One taxon
Positive co-occurrence
Negative co-occurrence
Approach: Environmental Shotgun Metagenomics

1. DNA extraction
2. Fragmentation
3. DNA sequencing
4. Assembly

- >seq1
- GCCGTAAGTCC...
- >seq2
- TATGCCGGA...
- >seq3
- ...

Metabolic reconstruction

Phylogenetic binning

Gene finding & annotation
Approach: Environmental Shotgun Metagenomics

Current study:
- 12 samples
- ~250 M reads / sample
- 20-25 Gb / sample
- ~300 Gb total
Approach: Environmental Shotgun Metagenomics

Current study:
- 12 samples
- ~250 M reads / sample
- 20-25 Gb / sample
- ~300 Gb total

Assembly:
- ~1 Tb memory
- 1.5 Tb disk
- 32 CPUs - 12 days
- Result: 37 Gb assembly

Approach:
- Install relevant software separately and import to job; Assemble files using MEGAHIT
Approach: Environmental Shotgun Metagenomics

**Current study:**
- 12 samples
- ~250 M reads / sample
- 20-25 Gb / sample
- ~300 Gb total

**Assembly:**
- ~1 Tb memory
- 1.5 Tb disk
- 32 CPUs - 12 days
- Result: 37 Gb assembly

**Approach**
- Install relevant software separately and import to job
- Assemble files using MEGAHIT

**Mapping:**
- For each sample
- 10 Gb memory
- 66 Gb disk
- 2 CPUs - 5 hours
- Result: 12 x 20-50 Gb files

**Approach**
- Map original sequences back to assembled sequences
Approach: Environmental Shotgun Metagenomics

Current study:
- 12 samples
- ~250 M reads / sample
- 20-25 Gb / sample
- ~300 Gb total

Assembly:
- ~1 Tb memory
- 1.5 Tb disk
- 32 CPUs - 12 days
- Result: 37 Gb assembly

Database:
- ~0.5 Tb memory
- 1.5 Tb disk
- 20 CPUs
- Result: 33 Gb zipped db

Approach:
- Create database from assembly and identify potential gene locations

Database:
- ~0.5 Tb memory
- 1.5 Tb disk
- 20 CPUs
- Result: 103 Gb db

Approach:
- Create database from assembly and identify potential gene locations using Anvi’o
Approach: Environmental Shotgun Metagenomics

Current study:
- 12 samples
- ~250 M reads / sample
- 20-25 Gb / sample
- ~300 Gb total

Assembly:
- ~1 Tb memory
- 1.5 Tb disk
- 32 CPUs - 12 days
- Result: 37 Gb assembly

Database:
- ~0.5 Tb memory
- 1.5 Tb disk
- 20 CPUs
- Result: 33 Gb zipped db

Approach:
- Create database from assembly and identify potential gene locations

HMMs:
- ~1 Tb memory
- 300 Gb disk
- 40 CPUs
- Result: currently running

Approach:
- Use hidden Markov models to identify potential gene functions
Approach: Environmental Shotgun Metagenomics

Current study:
- 12 samples
- ~250 M reads / sample
- 20-25 Gb / sample
- ~300 Gb total

Assembly:
- ~1 Tb memory
- 1.5 Tb disk
- 32 CPUs - 12 days
- Result:
  - 37 Gb assembly

Database:
- ~0.5 Tb memory
- 1.5 Tb disk
- 20 CPUs
- Result:
  - 33 Gb zipped db

Approach:
- Create database from assembly and identify potential gene locations

Next up:
- Profiling - Calculate coverage statistics for each gene, across the 12 samples
  - May target only subset of genes; Skip single nucleotide variant profiling

HMMs:
- ~1 Tb memory
- 300 Gb disk
- 40 CPUs
- Result:
  - currently running
Approach: Environmental Shotgun Metagenomics

Current study:
- 12 samples
- ~250 M reads / sample
- 20-25 Gb /sample
- ~300 Gb total

Assembly:
- ~1 Tb memory
- 1.5 Tb disk
- 32 CPUs - 12 days
- Result: 37 Gb assembly

Database:
- ~0.5 Tb memory
- 1.5 Tb disk
- 20 CPUs
- Result: 33 Gb zipped db

Approach:
Create database from assembly and identify potential gene locations

Database:
- ~0.5 Tb memory
- 1.5 Tb disk
- 20 CPUs
- Result: 103 Gb db

Approach:
Use hidden Markov models to identify potential gene functions

HMMs:
- ~1 Tb memory
- 300 Gb disk
- 40 CPUs
- Result: currently running

Analysis and visualization:
Interact with database within interactive job

Profiling:
- May target only subset of genes; Skip single nucleotide variant profiling

Next up:
Profiling - Calculate coverage statistics for each gene, across the 12 samples
Soil bacterial communities are among the most diverse

(Thompson et al., 2017)
Challenges: Environmental Shotgun Metagenomics

- Copying files (sometimes not zipping is faster)
- Soil microbial communities are particularly diverse
  - Comparatively large files
  - Tools may not be optimized for soil samples
  - Weeks-long running times
- User, not developer
Results: Environmental Shotgun Metagenomics

Xenobiotic Degradation KOs

Carbohydrate Metabolism KOs

(Berry et al., in prep)
Future Results: Environmental Shotgun Metagenomics

Charcoal-associated genes

Soil-associated genes
Acknowledgements

Co-authors: Ellen Whitman, Jamie Woolet, Marc-André Parisien, Daniel Thompson, Michael Flannigan, Timothy Berry
Special thanks: Christina Koch and Lauren Michael