Bugs in the arctic, discovering microbial diversity

James A. Foster
The Initiative for Bioinformatics and Evolutionary Studies (IBEST)
Biological Sciences, Bioinformatics and Computational Biology
University of Idaho
Bugs in the arctic, discovering microbial diversity

James A. Foster
The Initiative for Bioinformatics and Evolutionary Studies (IBEST)
Biological Sciences, Bioinformatics and Computational Biology
University of Idaho
Welcome to Spitsbergen


Receding glaciers.

Photo: Galen R Frysinger, with permission

WebCam Ny-Alesund - Kongsfjorden by NILU
Sampling emergent diversity

- Sample DNA along a age-variant transect
  - up to 10 samples per site
  - time since exposure: 5y, 19y, 40y, 63y, 100y, and 150y
  - “chronoclines” sample ecosystems by age
- Associate sequences with “species”
- Quantify variation between, within age groups
Bioinformatics problems

Biological questions:
How do soil bacterial respond to retreating glaciers? How do microbial soil communities change?
Bioinformatics problems

Biological questions:
How do soil bacterial respond to retreating glaciers? How do microbial soil communities change?

✦ Estimate $\alpha$ diversity: number of “species” in each sample and age group
✦ Estimate $\beta$ diversity: amount of variation in “species” between age groups
✦ Determine which species (no quotes) are present in each sample (not part of this talk)
Lots of data (post QC)

<table>
<thead>
<tr>
<th>Age</th>
<th>Samples</th>
<th>Sequences</th>
<th>DNA Mbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5y</td>
<td>9</td>
<td>35,092</td>
<td>8.77</td>
</tr>
<tr>
<td>19y</td>
<td>10</td>
<td>41,494</td>
<td>10.37</td>
</tr>
<tr>
<td>40y</td>
<td>8</td>
<td>33,665</td>
<td>8.42</td>
</tr>
<tr>
<td>63y</td>
<td>9</td>
<td>41,767</td>
<td>10.44</td>
</tr>
<tr>
<td>100y</td>
<td>8</td>
<td>41,178</td>
<td>10.29</td>
</tr>
<tr>
<td>150y</td>
<td>8</td>
<td>40,210</td>
<td>10.05</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>233,406</td>
<td>58.35</td>
</tr>
</tbody>
</table>
Lots of data (post QC)

<table>
<thead>
<tr>
<th>Age</th>
<th>Samples</th>
<th>Sequences</th>
<th>DNA Mbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>5y</td>
<td>9</td>
<td>35,092</td>
<td>8.77</td>
</tr>
<tr>
<td>19y</td>
<td>10</td>
<td>41,494</td>
<td>10.37</td>
</tr>
<tr>
<td>40y</td>
<td>8</td>
<td>33,665</td>
<td>8.42</td>
</tr>
<tr>
<td>63y</td>
<td>9</td>
<td>41,767</td>
<td>10.44</td>
</tr>
<tr>
<td>100y</td>
<td>8</td>
<td>41,178</td>
<td>10.29</td>
</tr>
<tr>
<td>150y</td>
<td>8</td>
<td>40,210</td>
<td>10.05</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>233,406</td>
<td>58.35</td>
</tr>
</tbody>
</table>

Note: A SMALL run, max is 37GB/8hr run max, 1.6 Bbp/day
Roche 454: a genome a day
Roche 454: a genome a day

**b**

**Emulsion PCR**

8 hours

- Anneal ssDNA to an excess of DNA capture beads
- Emulsify beads and PCR reagents in water-in-oil microreactors
- Clonal amplification occurs inside microreactors
- Break microreactors and enrich for DNA-positive beads

sstDNA library → Bead-amplified ssDNA library

**c**

**Sequencing**

7.5 hours

- Well diameter: average of 44 μm
- 400,000 reads obtained in parallel
- A single cloned amplified ssDNA bead is deposited per well

Amplified ssDNA library beads → Quality filtered bases
Roche 454: a genome a day

b  Emulsion PCR
8 hours

- Anneal ssDNA to an excess of DNA capture beads
- Emulsify beads and PCR reagents in water-in-oil microreactors
- Clonal amplification occurs inside microreactors
- Break microreactors and enrich for DNA-positive beads

sstDNA library  →  Bead-amplified ssDNA library

University of Idaho

JAF CS/UI Spitsbergen 4.30.09
Bioinformatics objectives
Bioinformatics objectives

Initiative for Bioinformatics & Evolutionary Studies
Bioinformatics objectives

determine species
Bioinformatics objectives

determine species
Bioinformatics objectives

determine species

cluster by species
Bioinformatics objectives

determine species

cluster by species

5y old

19y old

150y old
Bioinformatics objectives

- determine species
- cluster by species
- cluster by age
**Bioinformatics objectives**

- determine species
- cluster by species
- cluster by age

Explain data in terms of biological processes and age (tell a story)
Bioinformatics objectives

Too much data: 233K sequences!

- determine species
- cluster by species
- cluster by age

5y old 19y old ... 150y old

Explain data in terms of biological processes and age (tell a story)
Bioinformatics objectives

Too much data: 233K sequences!

- Cluster by species
- Cluster by age

Explain data in terms of biological processes and age (tell a story)
Bioinformatics objectives

Too much data: 233K sequences!

- Determine species
- Cluster by species
- Cluster by age

5y old  19y old  150y old

3%

Explain data in terms of biological processes and age (tell a story)
Bioinformatics objectives
Bioinformatics objectives
Bioinformatics objectives

Cluster each of 52 samples (approx. 6k each), choose a proxy sequence
Bioinformatics objectives

Cluster each of 52 samples (approx. 6k each), choose a proxy sequence
Cluster each of 52 samples (approx. 6k each), choose a proxy sequence.
Bioinformatics objectives

Cluster each of 52 samples (approx. 6k each), choose a proxy sequence

Cluster proxies by age (approx. 40k each)
Bioinformatics objectives

Cluster each of 52 samples (approx. 6k each), choose a proxy sequence

Cluster proxies by age (approx. 40k each)

Cluster combined sequences to get species (quantify richness)

Build +/- matrix
Bioinformatics objectives

Cluster each of 52 samples (approx. 6k each), choose a proxy sequence

Cluster proxies by age (approx. 40k each)

Cluster combined sequences to get species (quantify richness)

Build +/- matrix
For each layer of data reduction
  • Estimate evolutionary distances: align with *infrnal*, generate distance matrix with *dnadist*
  • Use *dotur* for “complete clustering” (most distant neighbor) at 3% divergence
  • Select proxies (sequence with lexicographically first name)

Estimate species richness (turnover) from rarefaction curves (+/- matrices) with *dotur*, Bunge’s metric, and custom scripts
Bioinformatics details

- For each layer of data reduction
  - Estimate evolutionary distances: align with infrnal, generate distance matrix with dnadist
  - Use dotur for “complete clustering” (most distant neighbor) at 3% divergence
  - Select proxies (sequence with lexicographically first name)

- Estimate species richness (turnover) from rarefaction curves (+/- matrices) with dotur, Bunge’s metric, and custom scripts

\[ O(N^2) \] in time and space
\[ 4^4 \text{ seq} = 10^6 \text{ bp} = 10^{12} \text{ time/space} \]
IBEST Bioinformatics Core

fourtytwo: 512 AMD64 cores, 512 GB RAM, nearly 1.5 T-FLOPs capacity

servers: three at 32 GB RAM, 16 processors each; 85 TB storage

(and other stuff!)

Also used supercomputer facilities at MSU & Cornell
There are lots of species

- Abundance Based Coverage (ACE) estimate: based on number of singletons vs total seen
- Bunge parametric estimate: statistically based, uses full histogram of repeat draws

Data omitted waiting publication
Turnover rates: in progress

Compute time-weighted ratios of +/- counts (Diamond-May)
Turnover rates: in progress

Compute time-weighted ratios of +/- counts (Diamond-May)
Turnover rates: in progress

Compute time-weighted ratios of +/- counts (Diamond-May)

Estimate parameters in markov model of turnover
Conclusions

✦ Biology

- There are *thousands of species* of bacteria in arctic soil
- Number of bacterial species *increases as time* of post-glacial exposure increase
Conclusions

✦ Biology
  • There are *thousands of species* of bacteria in arctic soil
  • Number of bacterial species *increases as time* of post-glacial exposure increase

✦ Algorithmics (want a job?)
  • “Quantity has a quality all it’s own” (V.I. Lenin)
  • Need new algorithms to use new hardware
  • Database/dataset management is crucial
Future work: same tune, new lyrics

✦ Data from human microbiome
   *How do microbial communities vary between healthy and sick people?*

✦ Data from polluted soil (Yangtzee river, PRC)
   *How do microbial communities vary as pollution increases?*

✦ Data from longitudinal transects
   *How does microbial diversity change with latitude?*
Thanks!

Ursel Schüette
Zaid Abdo
Jacob Pierson
Larry Forney
Rob Lyon
The Forney-Top lab

John Bunge, Cornell

The Relational Database project, MSU

the IBEST Bioinformatics Core, partially funded by NIH COBRE and INBRE grants

JAF CS/UI Spitsbergen 4.30.09
Extra stuff

Intentionally blank
Metagenomics

- Harvest approximately first 300bp of every 16s rRNA molecule, all samples
  - Ribosome: required to translate DNA (conserved)
  - Common marker for microbial species
- Cluster by evolutionary relationships ("species")
- Analyze by chronocline