Technologies for metagenomic data collection
Overview

• “Old school techniques”
• Next generation sequencing
• Microarrays: PhyloChip

Old school techniques

• Plating
  — Can only work with culturable strains
  — Very laborious

• Flow cytometry (unconventional)
• qPCR
  — Need primers for every species
  — Cannot identify previously unknown species
  — Laborious

“Old generation sequencing”: Sanger sequencing

• DNA is fragmented
• Cloned to a plasmid vector
• Cyclic sequencing reaction
• Separation by electrophoresis
• Readout with fluorescent tags

Current generation next generation sequencing platform comparison

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read length*</th>
<th>Gb per run*</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>GA II / Hiseq</td>
<td>2 x 100 bp</td>
<td>20+ Gb Bridge amplification</td>
</tr>
<tr>
<td>454</td>
<td>GS FLX Titanium</td>
<td>1x400-600 bp</td>
<td>~0.5 Gb Emulsion PCR amplification Pyrosequencing by extension</td>
</tr>
<tr>
<td>Applied Biosystems</td>
<td>SOLiD 3</td>
<td>2 x 50 bp</td>
<td>20+ Gb Emulsion PCR amplification Ligation-based sequencing Alignment in color space</td>
</tr>
<tr>
<td>Helicos</td>
<td>Single molecule</td>
<td>2 x 25-55 bp</td>
<td>21-28 Gb No amplification Single molecule sequencing</td>
</tr>
</tbody>
</table>

* Instrument vendors are constantly updating their technology, chemistry, and Gb in order to increase these parameters

Emulsion PCR

• Fragments, with adaptors, are PCR amplified within a water drop in oil.
• One primer is attached to the surface of a bead.
• Used by 454, Polonator and SOLiD.

Bridge PCR

• DNA fragments are flanked with adaptors.
• A flat surface coated with two types of primers, corresponding to the adaptors.
• Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
• Used by Solexa.
Roche 454 GS FLX Sequencing

• 1 million reads per run
• 400-500 base pairs per read
• Ability to multiplex (sequence many specimens in a single run) with bar codes

How does Roche 454 sequencing work?

• Nucleotides are cycled in order T, A, C, G ("flows")
• When incorporated onto a template a fluorescent signal is emitted
• The amount of fluorescence is recorded
• The reagents are washed and the flow cycle restarts

http://www.youtube.com/watch?v=vVH6GF9c58E

454 base calling

General strategy for

• Select a marker genomic region
  – Must be present in all bacteria
  – Must have regions constant for all bacteria for primer design
  – Must have regions of variability so that individual species/taxa can be distinguished
• Extract DNA from biological specimens
• Amplify the marker from extracted DNA
• Sequence multiple samples in multiplex

Targeted sequencing of 16S rRNA gene

Location of the hyper-variable regions of the 16S rRNA and a typical primer construct for amplicon library preparation

• 16S rRNA small subunit is commonly used
• Present in "all" bacteria
• Contains variable regions that (may) carry enough phylogenetic information to distinguish different bacteria on a "deep-enough" taxonomic/phylogenetic level

Limitations

• Unique sequences represent unique phylotypes, not species, due to 16S gene multiplicity
• Abundances computed with different primers are not comparable