CBC Data Therapy

Metagenomics Discussion
General Workflow

1. Microbial sample
2. Generate "Metagenomic" data
3. Process data (QC, etc.)
4. Analysis
Marker Genes

1. Extract DNA
2. Amplify with targeted primers
3. Filter errors, build clusters
4. Diversity analysis
Metagenomics

- Extract DNA
- Sequence random fragments
- QC, assemble, annotate
- Diversity, function analysis
Metatranscriptomics

- Extract RNA, subtract rRNA
- Sequence cDNA
- QC
- Gene expression, function
Sequencing

- Sanger
- Ion Torrent
- Roche 454
- Illumina *Seq
- Pacific Biosciences
- Nanopore
Resources (16S)

RDP II: Cole et al.
NAR (2013)

SILVA: Quast et al.
NAR (2015)

rrnDB: Stoddard et al.
NAR (2016)
Resources (Genomes)

- GenBank Genomes
- PATRIC (host and bacterial)
- GOLD (JGI metagenomes)
- Ensembl Genomes
Resources (Metagenomes)

- EBI metagenomics
- MG-RAST
- HMP DACC
Resources (Function)

- KEGG
- UniProtKB
- CARD
- Gene Ontology

Institute for Systems Genomics: Computational Biology Core
bioinformatics.uconn.edu
General Challenges/Considerations

• Sequencing errors
  • Error rates, error type (PacBio: 10% random, Illumina – 0.1% substitution)

• Chimeras
  • Amplification artifacts, cloning of restriction fragments

• 16S: different V regions give different results

• Different sequencing platforms / sampling conditions ALSO give different results

• Workflow complexity / plethora of tools
General Challenges/Considerations

• Strain-level diversity in metagenomes will often be missed by amplicon (esp. short-read) and shotgun approaches
  • This may be especially important between samples

• Taxonomy
  • Database predictions (RDP)

• Functional Annotation
  • Coverage versus accuracy
Marker Genes

- Eukaryotic Organisms (protists, fungi)
  - 18S ([http://www.arb-silva.de](http://www.arb-silva.de))

- Bacteria
  - CPN60 ([http://www.cpndb.ca/cpnDB/home.php](http://www.cpndb.ca/cpnDB/home.php))
  - ITS (Martiny, Env Micro 2009)
  - RecA gene

- Viruses
  - Gp23 for T4-like bacteriophage
  - RdRp for picornaviruses

Faster evolving markers used for strain-level differentiation
Marker Genes

- Focus on contamination reduction during preparation
- 16S rRNA contains 9 hypervariable regions (V1-V9)
- V4 was chosen because of its size (suitable for Illumina 150bp paired-end sequencing) and phylogenetic resolution
- Different V regions have different phylogenetic resolutions – giving rise to slightly different community composition results

Sequencing:
- MiSeq capacity allows multiple samples to be combined into a single run
- Number of reads needed to differentiate samples depends on the nature of the studies
  - Unique DNA barcodes can be incorporated into your amplicons to differentiate samples
Marker Genes

- QIIME (http://qiime.org)

- Mothur (http://www.mothur.org)
Bioinformatics
Overall Bioinformatics Workflow

Sequence data (fastq) → 1) Preprocessing: remove primers, demultiplex, quality filter, decontamination → 2) OTU Picking / Representative Sequences

Metadata about samples (mapping file)

1) Preprocessing:
   - remove primers
   - demultiplex
   - quality filter
   - decontamination

2) OTU Picking / Representative Sequences

3) Taxonomic Assignment

4) Build OTU Table (BIOM file)

5) Sequence Alignment

6) Phylogenetic Analysis

7) Downstream analysis and Visualization – knowledge discovery

Inputs → Processed Data → Outputs

Outputs:
- OTU Table
- Phylogenetic Tree
- Knowledge discovery
## QIIME versus MOTHUR

<table>
<thead>
<tr>
<th>QIIME</th>
<th>Mothur</th>
</tr>
</thead>
<tbody>
<tr>
<td>A python interface to glue together many programs</td>
<td>Single program with minimal external dependency</td>
</tr>
<tr>
<td>Wrappers for existing programs</td>
<td>Reimplementation of popular algorithms</td>
</tr>
<tr>
<td>Large number of dependencies / VM available</td>
<td>Easy to install and setup; work best on single multi-core server with lots of memory</td>
</tr>
<tr>
<td>More scalable</td>
<td>Less scalable</td>
</tr>
<tr>
<td>Steeper learning curve but more flexible workflow if you can write your own scripts</td>
<td>Easy to learn and works the best with built-in tools</td>
</tr>
</tbody>
</table>

http://www.mothur.org/wiki/MiSeq_SOP
Metagenomics

• Goal: Identify the relative abundance of different microbes in a sample given using metagenomics

• Problems:
  • Reads are all mixed together
  • Reads can be short (~100bp)
  • Lateral gene transfer

• Two broad approaches
  1. Binning Based
  2. Marker Based
Metagenomics

• Attempts to “bin” reads into the genome from which they originated

• Composition-based
  • Uses GC composition or k-mers (e.g. Naïve Bayes Classifier)
  • Generally not very precise and not recommended

• Sequence-based
  • Compare reads to large reference database using BLAST (or some other similarity search method)
  • Reads are assigned based on “Best-hit” or “Lowest Common Ancestor” approach
• Use all BLAST hits above a threshold and assign taxonomy at the lowest level in the tree which covers these taxa.

• Notable Examples:
  • MEGAN: [http://ab.inf.uni-tuebingen.de/software/megan5/](http://ab.inf.uni-tuebingen.de/software/megan5/)
    • One of the first metagenomic tools
    • Does functional profiling too!
  • MG-RAST: [https://metagenomics.anl.gov/](https://metagenomics.anl.gov/)
    • Web-based pipeline (might need to wait awhile for results)
  • Kraken: [https://ccb.jhu.edu/software/kraken/](https://ccb.jhu.edu/software/kraken/)
    • Fastest binning approach to date and very accurate.
    • Large computing requirements (e.g. >128GB RAM)
“MetaSPAdes showed the overall best assembly size statistics while also capturing a relatively large fraction of the expected diversity. The usage of this tool is relatively simple and convenient, being basically identical to that of SPAdes, and largely flexible regarding the format of the input data. A drawback may be the reduced sensitivity for micro diversity. However, for the majority of metagenome research questions, accurate and representative consensus genomes of species should be more than sufficient.”