Metagenomics
• Term first used in 1998 by Jo Handelsman
• "the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species”

Milestones in metagenomics

• 1985: Idea for sequencing the environment first proposed
  – Norman Pace and colleagues (ASM News)
• 1991: First published 16S rRNA“metagenomic” study (Schmidt et al., J. Bacteriol.)
• 2003: Sargasso sea project: > 2000 species, 148 novel bacteria (Venter et al., Science)
• 2004: Shotgun sequencing of seawater: > 5000 different viruses (Breibart et al., PNAS)
• 2004: Complete bacterial genomes assembled from environmental samples (Tyson et al., Nature)
• 2006: First published environmental sequences generated using NGS(454) technology (Poinar et al., Nature)
Metagenomics opens our eyes to the hidden world

- Viruses are the most abundant biological entities
- Culturable bacteria represent only about 1% of the total bacterial population
- Unculturable microorganisms form the vast majority of lifeforms on earth
From descriptive biology...

- Early studies focused simply on describing the microbial communities in different environments
  - How many species?
    - In soil
    - In sea water
    - In hot springs
    - In acid drainage
    - In poop

- Only addressed alpha diversity
  - Done with Sanger sequencing
...To hypothesis testing

• NGS sequencing → higher throughput, lower cost
• Allow testing of how microbial communities differ:
  – On various substrates
  – Based on climate
  – After environmental perturbation
  – When in competition

• Comparing populations
  – Examination of Beta diversity
Metagenomic Strategies

• Total genomic DNA (RNA for some viruses)
  – Gene discovery

• Target genes
  – Antibiotic resistance
  – “Detoxification” genes
  – 16S rRNA
16S rRNA gene sequencing

- Highly conserved regions
  - Identical in all bacteria
  - Single PCR primer pair can amplify 16S rRNA genes from diverse bacteria

- Highly variable regions
  - Conserved within species
  - Divergent between species

CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications

Image from Alimetrics.net
General 16S rRNA sequencing strategy

- Isolate “environmental” DNA
- Amplify 16S rRNA genes using PCR and primers recognizing conserved regions
  - Incorporate sequencing adaptors into primers
  - (add barcodes/tags to primers for multiplexing)
- Perform NGS
  - usually Roche 454 (longer read lengths)
- Use sequence data to identify types and abundances of bacterial “species”
- Measure community diversity (alpha)
- Compare diversity between communities, locations, treatments, etc. (beta)
Extracting “information” from sequences

- Types of bacteria
- Relative abundances of species identified
- Interspecific relationships

Step 1: Pick operational taxonomic units (OTUs)
Step 2: Take a single sequence from the cluster to represent the OTU
Step 3: Compare each representative OTU sequence with a 16S rRNA gene database
OTU picking

• Form clusters consisting of reads with highly similar sequences
  – *De novo* picking
    • Reads are clustered by comparing amongst themselves (slow)
  – Reference-based
    • Reads are clustered by comparing with a reference dataset (fast but less informative)
De novo OTU picking

- 97% OTU "species"
- Representative sequence
- 3%
Closed reference-based picking

Representative sequence

failed sequences
Open reference-based picking
Assign taxonomy information to OTUs

• Representative sequences are used to search database(s) of 16S rRNA genes from known bacterial species

• The OTU inherits the taxonomic descriptors of the top, legitimate match

• OTUs, count, taxonomy information and metadata are stored in a BIOM table
Example BIOM table (head)

```json
{"id": "None","format": "Biological Observation Matrix 1.0.0","format_url": "http://biom-format.org","type": "OTU table","generated_by": "QIIME 1.7.0","date": "2013-07-18T19:06:06.343160","matrix_type": "sparse","matrix_element_type": "int","shape": [419, 9],"data": [[0,0,1],[1,1,1],[2,0,1],[3,2,1],[4,3,1],[5,0,1],[5,1,1],[6,4,1],[7,3,1],[8,0,1],[8,1,1],[8,2,1],[8,4,1],[9,5,1],[10,3,1],[11,1,1],[11,3,1],[12,6,1],[13,4,2],[13,6,1],[14,1,1],[14,2,1],[14,5,1],[15,7,1],[16,1,1],[16,1,1],[17,8,1],[18,8,1],[19,2,1],[20,8,1],[21,3,1],[21,4,1],[22,7,1],[23,7,1],[24,1,2],[25,3,1],[26,1,1],[27,7,2],[28,0,1],[29,8,1],[30,1,1],[31,8,2],[32,1,1],[32,3,1],[33,8,1],[34,1,1],[34,2,1],[35,1,1],[36,7,1],[37,3,3],[38,7,1],[39,7,2],[40,6,1],[41,3,1],[41,7,2],[42,3,1],[42,4,1],[43,7,1],[44,5,1],[45,4,16],[45,5,12],[46,0,6],[46,1,2],[46,7,3],[46,8,5],[47,3,1],[48,7,1],[49,5,1],[50,4,1],[51,5,1],[52,3,1],[53,1,1],[53,3,2],[53,6,2],[54,0,37],[54,1,10],[54,3,1],[54,8,4],[55,5,1],[56,0,5],[56,1,4],[56,2,1],[56,3,2],[56,4,1],[56,5,1],[56,6,3],[56,7,9],[56,8,2],[57,5,1],[58,0,1],[59,0,1],[59,1,1],[59,2,10],[59,3,2],[59,4,2],[59,5,24],[59,6,1],[60,3,1],[61,2,1],[62,7,1],[63,2,1],[64,6,1],[65,7,1],[66,1,1],[67,3,1],[68,3,1],[69,6,1],[70,7,1],[71,6,1],[72,0,2],[72,1,3],[72,8,2],[73,0,1],[73,8,2],[74,0,1],[74],...]
```
BIOM table (tail)

{
  "columns":
  [
    {"id": "PC.636", "metadata": null},
    {"id": "PC.635", "metadata": null},
    {"id": "PC.356", "metadata": null},
    {"id": "PC.481", "metadata": null},
    {"id": "PC.354", "metadata": null},
    {"id": "PC.593", "metadata": null},
    {"id": "PC.355", "metadata": null},
    {"id": "PC.607", "metadata": null},
    {"id": "PC.634", "metadata": null}
  ],
  "rows":
  [
    {
      "id": "denovo406", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__", "s__"
        ]
      }
    },
    {
      "id": "denovo407", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__[Ruminococcus]", "s__gnavus"
        ]
      }
    },
    {
      "id": "denovo408", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__", "s__"
        ]
      }
    },
    {
      "id": "denovo409", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__", "s__"
        ]
      }
    },
    {
      "id": "denovo410", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__[Eubacterium]", "s__dolichum"
        ]
      }
    },
    {
      "id": "denovo411", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Bacilli",
          "o__Turicibacterales",
          "f__Turicibacteraceae",
          "g__Turicibacter", "s__"
        ]
      }
    },
    {
      "id": "denovo412", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae"
        ]
      }
    },
    {
      "id": "denovo413", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__", "s__"
        ]
      }
    },
    {
      "id": "denovo414", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae"
        ]
      }
    },
    {
      "id": "denovo415", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__", "s__"
        ]
      }
    },
    {
      "id": "denovo416", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__[Ruminococcus]", "s__gnavus"
        ]
      }
    },
    {
      "id": "denovo417", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae",
          "g__", "s__"
        ]
      }
    },
    {
      "id": "denovo418", 
      "metadata": {
        "taxonomy": [
          "k__Bacteria",
          "p__Firmicutes",
          "c__Clostridia",
          "o__Clostridiales",
          "f__Lachnospiraceae"
        ]
      }
    }
  ]
}
Alpha diversity

• Measure of species richness at local scale
  – Locations
  – Conditions
  – Samples

• Measured in different ways:
  – Classical diversity measures
    • Shannon index, Simpson Index, etc.
  – Phylogenetic measures
    • Phylogenetic distances over whole trees
Assigning sequences to OTUs → loss of information

OTU1

?  

OTU2
Phylogenetic metrics

OTU1

Genetic distance

OTU2

Store in a distance matrix
Beta diversity

• Comparison of microbial communities based on their composition

• Metrics assess differences between these communities in:
  – Overall composition
    • Comparison of distance matrices → principle coordinates analysis
  – Composition at varying taxonomic levels
    • Species, genus, family, etc. → G-tests, ANOVA, etc.
  – Correlation studies → Pearson test
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6.0</td>
<td>5.0</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10.0</td>
<td>9.0</td>
<td>4.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9.0</td>
<td>8.0</td>
<td>5.0</td>
<td>3.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>