Introduction to Metagenomics

Gene Tyson
DeLong Lab (MIT)
gtyson@mit.edu

Agouron Summer Course 2008
Introduction to Metagenomics

Outline:

- Putting metagenomics in perspective
- What can metagenomics tell us about microbial communities?
- Future of metagenomics
- Tutorial - Bioinformatic tools for metagenomic data analysis
Carl Woese


most bacteria don’t grow on plates “the great plate count anomaly”
Norm Pace (University of Colorado-Boulder)
whole–cell hybridization

environmental sample

extraction

bulk DNA / RNA

PCR

cloning

community rRNA / rDNA

nucleic acid hybridization

rRNA / rDNA clones

sequencing

rRNA / rDNA sequences and database

phylogenetic trees

probe design

comparative analysis

nucleic acid probes
Known Bacterial Phylogenetic Divisions

1987
12 divisions; 12 cultured / 0 uncultured

1997
36 divisions; 24 cultured / 12 uncultured

2003
53 divisions; 26 cultured / 27 uncultured

2006
~100 divisions; 30 cultured / ~70 uncultured
very skewed representation of Bacteria

culture collection (ACM)

- Proteobacteria: 54%
- Firmicutes: 14%
- Actinobacteria: 18%
- Bacteroidetes: 6%
- Other phyla: 3%

3760 bacterial cultures

sequenced genomes

- Proteobacteria: 45%
- Firmicutes: 24%
- Actinobacteria: 7%
- Bacteroidetes: 1%
- Other phyla: 20%

71 bacterial genomes
phylogenetic trees

whole–cell hybridization

Environmental sample

Extraction

Bulk DNA / RNA

PCR

Cloning

Sequencing

Comparative analysis

Genomic sequences and database

Phylogenetic trees

Nucleic acid probes

Probe design

Nucleic acid hybridization

Community rRNA / rDNA

Cloning

Sequencing
isolate

community

Genomics

sequencing

Metagenomics
Sanger sequencing has become much cheaper
Joint Genome Institute (JGI) statistics

Courtesy of Phil Hugenholtz (JGI)
The catch… it comes in small pieces

Average Sanger read length - 750 bases (bps)

Must assemble the reads together

Giant jigsaw puzzles
Genome assembly
Genome assembly
Metagenome assembly
Metagenome assembly
Metagenome assembly
Decoding metagenomes

Environmental Sample → Extract DNA → Clone → Sheared Size selection → High throughput sequence

Library Type:
- Shotgun (small-insert) 3kb
- Fosmid (large-insert) 40 kb
- BAC (large-insert) BIG STUFF!

Assemble reads → Call genes → Bin fragments
Metagenomics projects to date:

• Metabolic profiling of environments without need for significant assembly or reference to the organism from which genes were derived (e.g. EGTs).

• Comprehensive analysis of the community that resolves gene complement of the dominant organisms and provides insight into population structure and evolution.
AMD: Coupling of microbial metabolism and mineral dissolution

FeS₂ + 3.5 O₂ + H₂O → Fe²⁺ + 2SO₄²⁻ + 2H⁺

Net increase in soluble metal

FeS₂ + 14Fe³⁺ + 8H₂O → 15Fe²⁺ + 2SO₄²⁻ + 16H⁺

Acidification

microbial up to ~10⁶ times faster

chemoautotrophy
Richmond Mine, Iron Mountain, CA
Richmond Mine
Iron Mountain CA

95% pyrite \((\text{FeS}_2)\) ore deposit
Extreme acidity \((\text{pH} < 1)\)
Warm \((30 - 50^\circ\text{C})\)
Very high ionic strength \((\text{g/l})\)
Abundant toxic metals
(molar Fe, mM Zn, Cu, As)
Iron oxidation - primary energy source
No sunlight
Limited external sources of organic C and fixed N
16S rRNA gene PCR clone library analysis

- **Sulfobacillus**
- **Ferroplasma**
- **Leptospirillum group III**
- **Leptospirillum group II**
- **“A-plasma”**
- **“G-plasma”**

10% divergence
Shotgun sequencing

 Bulk environmental DNA

 sheared

 3 – 4 kb shotgun library

 end sequence clones (f / r)

 assemble reads by alignment identity

...ACGGCTGC\text{CGTTACATCGATCAT}ACATCGATCATTTACGATACCGATTG...
Genome Scaffolding

mate pair linkage

“composite” genome scaffold
Assembling community genomic (‘metagenomic’) data

Other potential problems:
- repetitive and mobile elements
- highly conserved genes
AMD community genome sequencing

**Ferroplasma acidarmanus isolate**
- isolated in 1997
- 1.94 Mb genome of *F. acidarmanus* fer1
- annotation completed

**AMD biofilm community**
- 76 Mb shotgun library end-sequence data [133 Mb]
  - 103,462 high quality reads (~737 bp/read) [189,000 reads]
- draft assembly 1,183 scaffolds (>2kb) totaling 10.83 Mb
- draft annotation completed
Assembly of the AMD community genomic data

1183 scaffolds

biggest scaffold: 138 kb

1183 scaffolds
G+C content of community genome scaffolds

Fraction of total bases

G+C content

Bacteria 46%

Archaea 54%
High depth scaffold - *Leptospirillum* group II

Low depth scaffold - *Leptospirillum* group III
Binning assembled community genome data by GC and depth

Local read coverage
Scaffold average G+C

- Ferroplasma type II
  - 1.88 Mb

- Ferroplasma type I + G-plasma
  - 4.12 Mb

- Leptospirillum group II
  - 2.27 Mb

- Leptospirillum group III
  - 2.56 Mb
## Poisson Calculations for Shotgun Sequencing (Lander Waterman)

<table>
<thead>
<tr>
<th>Fold coverage</th>
<th>Percent of genome sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25 x</td>
<td>22%</td>
</tr>
<tr>
<td>0.50 x</td>
<td>39%</td>
</tr>
<tr>
<td>0.75 x</td>
<td>53%</td>
</tr>
<tr>
<td>1 x</td>
<td>63%</td>
</tr>
<tr>
<td>2 x</td>
<td>88%</td>
</tr>
<tr>
<td>3 x</td>
<td>95%</td>
</tr>
<tr>
<td>4 x</td>
<td>98%</td>
</tr>
<tr>
<td>5 x</td>
<td>99.4%</td>
</tr>
<tr>
<td>6 x</td>
<td>99.75%</td>
</tr>
<tr>
<td>7 x</td>
<td>99.91%</td>
</tr>
<tr>
<td>8 x</td>
<td>99.97%</td>
</tr>
<tr>
<td>9 x</td>
<td>99.99%</td>
</tr>
<tr>
<td>10 x</td>
<td>99.995%</td>
</tr>
</tbody>
</table>
Community Genome Sequencing

- Near complete recovery of two genomes
  - *Leptospirillum* group II
  - *Ferroplasma* type II
  - First sequenced member of the Nitrospira phylum

- Partial recovery of three other genomes
  - *Leptospirillum* group III
  - *Ferroplasma* type I
  - G-plasma
  - New species not recognized by 16S analysis

- Minor components sampled
  - *Sulfbacillus thermosulfooxidans*
  - A-plasma

Uncultured microorganism
Metagenomics provides:

- insight into the metabolism of organisms and overall community function
**Metabolic Network**

- EPS production (cellulose synthase)
- Motility (response to ferrous iron and oxygen gradients)
- Oxidative stress resistance
- Genes for resistance to copper, arsenite, mercury, zinc, silver, and cadmium (efflux pumps)
- Electron transport chain components and a number of novel cytochromes
- Partitioning of community essential roles
- C and N fixation
Metagenomics provides:

- insight into the metabolism of organisms and overall community function
- sequences of co-habiting / co-evolving populations for comparative analyses
Leptospirillum group II genomic variation

1 nt polymorphism / ~1,300 bases

Recent selection event (sweep)? Founder effect?
16S rRNA gene phylogeny of the *Thermoplasmatales*

- **2011:**
  - "fer1 : "fer1env" = 0.0\% 16S divergence
  - fer1 : fer2 = 0.8\% 16S divergence
### Homogeneous region of fer1env population

<table>
<thead>
<tr>
<th>ISOLATE</th>
<th>959</th>
<th>cattggccccgaatatcttcctgtcccttcgcaaggttagtaagccagaa</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYG23099.b1</td>
<td>164</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG3165.b1</td>
<td>316</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG2007.b1</td>
<td>242</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG2300.g1</td>
<td>211</td>
<td>..................................................................</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ISOLATE</th>
<th>1019</th>
<th>tagcactgggaataattttaaggctttcctggatattgaaattgaaagccctctctctg</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYG23099.b1</td>
<td>104</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG3165.b1</td>
<td>376</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG2007.b1</td>
<td>302</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG2300.g1</td>
<td>271</td>
<td>..................................................................</td>
</tr>
</tbody>
</table>

### Heterogeneous region of fer1env population – 2 variants

<table>
<thead>
<tr>
<th>ISOLATE</th>
<th>361</th>
<th>gaatatggtgcaacacataattgcacatggtccacaggcaagggcaatagcagtccagtaaga</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYG35496.g1</td>
<td>530</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG3784.b2</td>
<td>266</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG66165.g1</td>
<td>298</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG6944.b1</td>
<td>224</td>
<td>..................................................................</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ISOLATE</th>
<th>421</th>
<th>tttgaggctcaatgaaagccttcatgaagataataaatgtaatcgcaccccgttaagggat</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYG35496.g1</td>
<td>490</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG3784.b2</td>
<td>326</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG66165.g1</td>
<td>238</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG6944.b1</td>
<td>164</td>
<td>..................................................................</td>
</tr>
</tbody>
</table>

### Only fer1env sequence type

<table>
<thead>
<tr>
<th>ISOLATE</th>
<th>661</th>
<th>acactgggttcacatggtcacttatttatattcatgttcatcttaaacagagt</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYG7614.g2</td>
<td>399</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG31767.g1</td>
<td>308</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG7766.b1</td>
<td>199</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG44653.b1</td>
<td>351</td>
<td>..................................................................</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ISOLATE</th>
<th>721</th>
<th>aagggaccttcaagatgggaatatgtcctcccatgaaatgtatcgggtcgagcact</th>
</tr>
</thead>
<tbody>
<tr>
<td>XYG7614.g2</td>
<td>339</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG31767.g1</td>
<td>248</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG7766.b1</td>
<td>139</td>
<td>..................................................................</td>
</tr>
<tr>
<td>XYG44653.b1</td>
<td>411</td>
<td>..................................................................</td>
</tr>
</tbody>
</table>
### evidence for inter-*strain* recombination

(glycosyltransferase)

<table>
<thead>
<tr>
<th>Query</th>
<th>Length</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>fer1.144c</td>
<td>121</td>
<td>aagtcaacaactgagcaatcactttcacggaattatagacaatatttattttatatgtagatcctttgca</td>
</tr>
<tr>
<td>XYG7465.g2</td>
<td>215</td>
<td></td>
</tr>
<tr>
<td>XYD1690.g1</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>XYG64953.g1</td>
<td>198</td>
<td>...t.t.a...a.c...g...aa...t...c.g</td>
</tr>
<tr>
<td>XYG17515.g3</td>
<td>374</td>
<td>...t.t.a...a.c...g...aa...t...c.g</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>Length</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>fer1.144c</td>
<td>211</td>
<td>gaagagtatataatgtaatattttatgtgaaatgaaacgctctctgtgaaagataatagaaagatcggatcgccaggagaatat</td>
</tr>
<tr>
<td>XYG7465.g2</td>
<td>365</td>
<td></td>
</tr>
<tr>
<td>XYD1690.g1</td>
<td>250</td>
<td>...c...a...a...c...a...g...c...g...t...t...a</td>
</tr>
<tr>
<td>XYG64953.g1</td>
<td>348</td>
<td>...c...t...g...c...a...a...c...a...g...c...g...t...t...a</td>
</tr>
<tr>
<td>XYG17515.g3</td>
<td>524</td>
<td>...c...t...g...a...a...a...a...a...a...a</td>
</tr>
</tbody>
</table>

- observe linkage between isolate-type and non-isolate type sequence (mate-pairs)
- observe transitions within single reads...
Ferroplasma type II inter-strain variation

Tyson et al. Nature, 2004
Are combinatorial variants a strategy for fine-tuning environmental optimization?

Can this (and other?) microbial species be defined like sexually reproducing organisms?
Detecting intra–species recombination

Ferroplasma type I
5.4 x 10^{-5} events/bp

Ferroplasma type II
5.1 x 10^{-5} events/bp
Detecting inter-species recombination

Ferroplasma type I

Ferroplasma type II

Recombination frequency: Mutation rate 1:30
Recombination frequency versus sequence divergence
Metagenomics provides:

- insight into the metabolism of organisms and overall community function
- sequences of co-habiting / co-evolving populations for comparative analyses
- possibility to detect organisms missed in 16S rRNA surveys
Detection of Novel Diversity

Mismatches with commonly used 16S rRNA gene primers

<table>
<thead>
<tr>
<th>Primer</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>21F</td>
<td>TTCCGGTTGATCCYGCCGGA</td>
</tr>
<tr>
<td>23F</td>
<td>TTCCGGTTGATCCYGGCC</td>
</tr>
<tr>
<td>1492R</td>
<td>GGTTCACCTTGTTAUUCGAUUCTT</td>
</tr>
<tr>
<td>1391R</td>
<td>GUACGGCGGTGWGTRCA</td>
</tr>
<tr>
<td>1525R</td>
<td>AGGAAGGTGATCCAGUCC</td>
</tr>
</tbody>
</table>
16S rRNA Analysis

2 new groups:

8% and 17% divergent from WTF-1
FISH (fluorescent in situ hybridization) analyses of ARMAN groups

- They are present in low abundance, but seen in ALL samples in the mine
- MUCH smaller than other cells in the mine

Since the are smaller than other cells we should be able to concentrate them by filtration…
Concentration of ARMAN groups

The ARMAN groups are highly enriched in the filtrate.
TEM characterization of the 450 nm filtrate

- The cells range in size from 170nm to 240nm (averaging 200 nm)
- The cell walls contain an S-layer
- The cytoplasm is densely packed
- 1 to 2 protrusions on each cell

Baker, B.J, Tyson, G.W. et al. Science
Metagenomics provides:

- insight into the metabolism of organisms and overall community function
- sequences of co-habiting / co-evolving populations for comparative analyses
- possibility to detect organisms missed in 16S surveys
- clues for cultivating uncultured organisms
Environmental nifH libraries confirmed the presence of only one nitrogen fixer.
Genome-directed isolation

Screened samples using FISH

Serial diluted

45 days at 37ºC

Subcultured

EUB338
All Bacteria

LF655
Leptospirillum groups I, II and III

LF1252
Leptospirillum group III only

Combined image showing all three probes

“Genome-directed isolation of *Leptospirillum ferridiazotrophum*, the key nitrogen fixer in acid mine drainage communities”

Gene W. Tyson, Ian Lo, Brett J. Baker, Eric E. Allen, Philip Hugenholtz & Jillian F. Banfield (AEM)
16S rRNA analysis of *Leptospirillum* group III culture

“Genome-directed isolation of *Leptospirillum ferridiazotrophum*, the key nitrogen fixer in acid mine drainage communities”
G.W. Tyson *et al.*, *(AEM)*
Community genomics provides:

- insight into the metabolism of organisms and overall community function
- sequences of co-habiting / co-evolving populations for comparative analyses
- possibility to detect organisms missed in 16S surveys
- clues for cultivating uncultured organisms
- basis for proteomic analyses of ‘whole’ communities
### Proteome Sample (FISH)

<table>
<thead>
<tr>
<th>Genome</th>
<th>Proteome</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>75%</td>
<td>83%</td>
<td><em>Leptospirillum</em> group II</td>
</tr>
<tr>
<td>10%</td>
<td>9%</td>
<td><em>Leptospirillum</em> group III</td>
</tr>
<tr>
<td>10%</td>
<td>8%</td>
<td>Archaea</td>
</tr>
<tr>
<td>1%</td>
<td>1%</td>
<td>non-<em>Leptospirillum</em> bacteria</td>
</tr>
</tbody>
</table>

Protein Extraction for Mass Spectrometry

Biofilm Sample
1. Wash in H$_2$SO$_4$, pH 1.1
2. Low speed centrifugation

Cellular Fraction
1. Sonicate to lyse
2. Low speed spin (remove sediment/unbroken cells)
3. High speed centrifugation

Extracellular fraction

Membrane

Soluble
Community Proteomics: 2D LC-MS/MS

Trypsin proteolysis of proteins in fractions

**COMMUNITY GENOME DATABASE**

Leptoll_scaffold_14_GENE_20
MNKWAGAVLGTKVTGLLSATAYSAELDILKPN
RVPADQIAAKAMKPPFPVTAAVIAKGEVFNFNGAGTCYTCHVGKGKGDGPAGAGMDSPRFTNH
QFDQVRTAGEMVWVVSNGSPLQPAMGFGVSAG
ITDKQAWEAVMYERSLGCDDMDCVTSADWVGKQPVHEEAASSLKPEYIGVASAH
Proteins detected in the natural community

- 17% of predicted genes in the community
- 48% of proteins predicted from the *Leptospirillum* group II (1362 proteins)
- validation of 572 hypothetical proteins

Functional Category Analysis of Expressed Proteins

Prevalence of hypothetical proteins

Environmental context clues to function
Importance of hypothetical proteins

31% of detected proteins are hypothetical

212 operons encoding proteins of ascribable function and expressed hypothetical proteins

\textit{e.g.:} operon encoding 8 flagellar proteins also encodes 4 detected hypothetical proteins

Detected complete operons of hypothetical proteins

\textit{e.g.:} 3 gene operon that is \textit{Leptospirillum group II}
Most abundant proteins overall
(compiled top 50 from each fraction)

17% hypothetical

13% ribosomal

11% chaperones

9% thioredoxins

8% radical defense

Protein folding and radical defense
Using MS/MS data to infer localization
% Sequence of Reads

Environmental Sample (Complexity)

Acid Mine Drainage Biofilm (Low)
Sargasso Sea (Moderate)
Soil (High)
Environmental Gene Tags (EGT)

Identify genes in sequence data (contigs to unassembled reads) from *multiple* environmental samples

Assign genes to their gene family, or higher level groupings

Compare relative abundance of different gene families according to habitat
Primary challenges for metagenomics in more complex environments

- Cost
- Sequencing efficiency

Improved assembly algorithms/heuristics

Resolution of strain heterogeneity
- Post-genomic databases
- Molecular evolution

learn from simple communities
454 pyrosequencer

- Sanger
  - 0.7 Mbp
  - 700 bp
  - 0.1 $\$$

- GS20
  - 35 Mbp
  - 100 bp
  - 0.03 $\$$

- FLX
  - 100 Mbp
  - 200 bp
  - 0.01 $\$$

- XLR
  - 400 Mbp per run
  - 400 bp reads
  - 0.003 $\$$ per base
Metagenomics projects

23 completed and 130 ongoing metagenomic projects
JGI metagenomic projects

2005
- Gutless worm (MPI)
- Planktonic archaea (MIT)
- EBPR sludge (UW/UQ)
- Groundwater (ORNL)

2006
- AMD nanoarchaea (UCB)
- Alaskan soil (UW)
- Termite hindgut (CalTech)
- TA-degrading bioreactor (NUS)
- Antarctic bacterioplankton (DRI)
- Hypersaline mats (UCol)
- Soil archaea (UW)
- Korarchaeota enrichment (Diversa)
Metagenomics is just the first level
Bioinformatic tools for metagenomic data analysis

MEGAN
- blast-based tool for exploring taxonomic content

MG-RAST (SEED, FIG)
- rapid annotation of metagenomic data, phylogenetic classification and metabolic reconstruction

CAMERA (JCVI, Calit2, UCSD)
- metagenomic data repository and blast server
Getting started

http://urkaryote.mit.edu/files/

Download MEGAN and metagenomic datasets BLAST files:

Index of /files

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td>04-Aug-2006 10:57</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>11009 PR.fna</td>
<td>25-Jun-2008 14:09</td>
<td>1k</td>
<td></td>
</tr>
<tr>
<td>MEGAN macco 2bosal.ome</td>
<td>13-Jun-2008 14:21</td>
<td>15.5M</td>
<td></td>
</tr>
<tr>
<td>MEGANmanual.pdf</td>
<td>13-Jun-2008 14:21</td>
<td>815k</td>
<td></td>
</tr>
<tr>
<td>OWZ 20m allDNA.fna.5..&gt;</td>
<td>15-Jun-2008 23:29</td>
<td>30.1M</td>
<td></td>
</tr>
<tr>
<td>OWZ 20m allDNA.fna.5..&gt;</td>
<td>24-Jun-2008 17:57</td>
<td>31.9M</td>
<td></td>
</tr>
<tr>
<td>OWZ 45m allDNA.fna.5..&gt;</td>
<td>15-Jun-2008 23:27</td>
<td>30.0M</td>
<td></td>
</tr>
<tr>
<td>OWZ 60m allDNA.fna.5..&gt;</td>
<td>15-Jun-2008 23:58</td>
<td>34.7M</td>
<td></td>
</tr>
</tbody>
</table>