A Methodology Study for Metagenomics using Next Generation Sequencers

Presenter: Sushmita Singh

An ABRF 2011-12 DSRG Study
Definitions?!

Metagenomics: Metagenomics is the study of metagenomes, genetic material recovered directly from environmental samples. The broad field may also be referred to as environmental genomics, ecogenomics or community genomics.

Microbiome: A microbiome is the totality of microbes, their genetic elements (genomes), and environmental interactions in a defined environment. A defined environment could, for example, be the gut of a human being or a soil sample.

From Wikipedia (http://en.wikipedia.org/wiki/Metagenomics)
Metagenomics Methodology

How many reads will I need?

Which kit should I use for DNA extraction?

How much DNA?

Do you need the amplified DNA?

Shot-gun?

It’s got to be 16S, right?!

Which ‘V’ region/s should I amplify?

It has to be V3! or maybe V5?

Metagenomics! It has to be pyrosequencing!!

Barcoding? Will you do it or should I?

How much will it cost?

Can I do this on an Illumina platform? It’s cheaper!
Good Candidates!

DSRG (DNA Sequencing Research Group)
★ Penn State University
★ Tufts University
★ McGill University
★ University of Minnesota
★ University of Michigan
★ New York University
★ Stower’s Institute
★ Harvard Medical School
★ Ohio State University

At Core Facilities we have,
★ a ‘global’ perspective
★ cater to a ‘diverse’ client group

Given expected magnitude of study, this will be an extended, two year study
Experimental Plan

DNA extraction
- Qiagen
- MoBio
- SCODA

Sequencing approach
- WGS
- 16S rRNA

Platform
- 454: 400bp Ti
- illumina: 2x150 GAIIx

Genomic DNA
Experimental Plan

Genomic DNA

- DNA extraction
  - Qiagen
  - MoBio
  - SCODA

Sequencing approach
- WGS
- 16S rRNA

Platform
- 454: 400bp Ti
- illumina: 2x150 GAIIx
Methodology Step 1: DNA Extraction Methods

Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags

Gary D Wu, James D Lewis, Christian Hoffmann, Ying-Yu Chen, Rob Knight, Kyle Bittinger, Jennifer Hwang, Jun Chen, Ronald Berkowsky, Lisa Nessel, Hongzhe Li, Frederic D Bushman
<table>
<thead>
<tr>
<th><strong>Qiagen</strong></th>
<th>‘Traditional’ / popular kits in most labs</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Qiagen DNA blood midiprep for soil sample</td>
<td></td>
</tr>
<tr>
<td>★ QIAamp Stool Kit for stool sample</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Mo Bio</strong></th>
<th>Customized kits for genomic DNA isolation for soil, microbial, biofilm, fecal, water, plant, blood etc</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Powerlyzer Soil DNA kit for soil</td>
<td></td>
</tr>
<tr>
<td>★ PowerSoil™ kit for stool</td>
<td></td>
</tr>
</tbody>
</table>

**SCODA (Synchronous Coefficient of Drag Alteration):** “A novel, proprietary electrophoresis technology that selectively concentrates nucleic acids from large sample volumes to a small, concentrated amount of gel or buffer. This is accomplished by means of rotating electric fields that act uniquely on long, charged polymers and leave other molecules unaffected”.

Here, SCODA was used in conjunction with Mo Bio kits

For details on technology, check out [Boreal Genomics Poster #166](#)
Experimental Plan

Genomic DNA

DNA extraction

Qiagen

MoBio

SCODA

Sequencing approach

WGS

16S rRNA

Platform

454: 400bp Ti
illumina: 2x150 GAIIx
Methodology Step 2: Sequencing Approach

Shot-gun vs tag sequencing

Whole Genome Sequencing (WGS): Information obtained only as good as the assemblers and existent databases

16S rRNA gene sequencing is currently the method of choice for phylogenetic reconstruction, nucleic acid based detection and quantification of microbial diversity.
Experimental Plan

DNA extraction

- Qiagen
- MoBio
- SCODA

Sequencing approach

- WGS
- 16S rRNA

Platform

- 454: 400bp Ti
- illumina: 2x150 GAIIx
<table>
<thead>
<tr>
<th><strong>Platform?!</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrosequencing, continues to be the method of choice since the first Next Generation sequencer, Roche 454 GS 20 generating a read size of 100 bp, first appeared on the scene, now with 400+ bp read sizes</td>
</tr>
<tr>
<td>Illumina? With read sizes increasing (2x150 bp), the Illumina Sequencers are increasingly being considered as a more cost-efficient alternative to pyrosequencing. Several studies using WGS sequencing approach are already published. Studies sequencing the 16S rRNA are now coming in</td>
</tr>
</tbody>
</table>
Microbiome Profiling by Illumina Sequencing of Combinatorial Sequence-Tagged PCR Products

Gregory B. Gloor¹, Ruben Hummelen², Jean M. Macklam¹,², Russell J. Dickson¹, Andrew D. Fernandes¹,⁴, Roderick MacPhee²,⁵, Gregor Reig¹,²,⁵,⁶

16S V6 PCR products

- 18047360, 76nt paired-end reads
- 14960488 read pairs (0 mismatch in overlap)
- 1922084 read pairs (1 mismatch in overlap)
- 12035329 primer and tag matches (0 mismatch in overlap)
- 1501641 primer and tag matches (1 mismatch in overlap)
- 128832 ISUs (0 mismatch in overlap)
- 6334 OTUs at 95% identity (0 mismatch in overlap)

Analysis

Figure 1. Conceptual workflow of the data analysis. PCR products derived from the eubacterial V6 rrRNA region were sequenced on a single paired-end Illumina run. Reads were filtered for quality, overlapped and clustered as outlined in the text. Only reads with 0 mismatches in the overlapping region were used for further analysis.

doi:10.1371/journal.pone.0015406.g001
Experimental Plan

Sample selection

DNA extraction

- Qiagen
- MoBio
- SCODA

Sequencing approach

- Qiagen:
  - WGS
  - 16S rRNA
- MoBio:
  - WGS
  - 16S rRNA
- SCODA:
  - WGS
  - 16S rRNA

Platform

- 454: 400bp Ti
- illumina: 2x150 GAIIx

Sample selection
Two Sample types selected:

- **DNA extraction**
  - Qiagen
  - MoBio
  - SCODA

- **Sequencing approach**
  - STOOL
    - WGS
    - 16S rRNA
    - Platform: 454: 400bp Ti
      - illumina: 2x150 GAIIx
    - 1
    - 2
    - 3
    - 4
    - 5
    - 6

  - SOIL
    - WGS
    - 16S rRNA
    - Platform: 454: 400bp Ti
      - illumina: 2x150 GAIIx
    - 1
    - 2
    - 3
    - 4
    - 5
    - 6
Sample Information:

**Soil**: Forest soil was collected on Dec. 15, 2010 from the A1/A2 horizons (upper 15 cm) of a toeslope location in the Susquehanna Shale Hills Critical Zone Observatory in central Pennsylvania. Soil was spread out in the bottom of a biological safety cabinet to minimize contamination while removing large roots and rocks before air-drying for 6 hours. Soil was mixed by passing it twice through 2-mm sieves that had been flame-sterilized. Air-dried soil was kept at -20 to inhibit respiration.

**Stool 1**: The stool samples were from mantled howlers, *Alouatta palliata*, in Costa Rica. The animals ranged in age from 6-11 years but were all from the same group of animals traveling together. The samples were collected and immediately placed on dry ice, and remained on dry ice / -80C until extracted.

**Stool 2**: Human stool samples from McGill University. The samples were collected and immediately placed on dry ice, and remained on dry ice / -80C until extracted.
Study Logistics

Soil collection:
Mary Ann Bruns, PA

Soil DNA extraction:
Qiagen: PSU
MoBio: PSU
SCODA: Boreal Genomics

Stool 1 collection:
Timothy Johnson, Costa Rica

Stool DNA extraction:
Qiagen: UMN
Mo Bio: UMN
SCODA: Boreal Genomics

All samples routed to PSU

Sample QC, quantitation, aliquoting

454 sample processing
Deb Grove, PSU

454 Pyrosequencing
Deb Grove, PSU

DATA

Analysis:
Istvan Albert, PSU
Sushmita Singh, UMN

Illumina sample processing
Greg Gloor, Canada

Illumina sequencing
Bob Lyons, U Michigan

DATA

Ken Dewar, McGill

454 sample processing
Ken Dewar, McGill

454 Pyrosequencing
Ken Dewar, McGill

DATA

Analysis:
Ken Dewar, McGill
Sushmita Singh, UMN
Study Status: As of Sunday, Feb 20, 2011

Soil collection: Mary Ann Bruns, PA

Soil DNA extraction:
Qiagen: aborted
MoBio: PSU
SCODA: Boreal Genomics

Stool 1 collection: Timothy Johnson, Costa Rica

Stool DNA extraction:
Qiagen: UMN
Mo Bio: UMN
SCODA: Boreal Genomics

All samples routed to PSU

Sample QC, quantitation, aliquoting

454 sample processing
Deb Grove, PSU

454 Pyrosequencing
Deb Grove, PSU

DATA

Illumina sample processing
Greg Gloor, Canada

Illumina sequencing
Bob Lyons, U Michigan

DATA

454 sample processing
Ken Dewar, McGill

454 Pyrosequencing
Ken Dewar, McGill

DATA

Analysis:
Istvan Albert, PSU
Sushmita Singh, UMN

Analysis:
Ken Dewar, McGill
Sushmita Singh, UMN
Very, Very, VERY, Preliminary Results

Data analyzed so far is 16s rRNA sequence data for Stool 1 and Soil samples

Processing of Reads:

Reads obtained from the 454 sequencing have been processed with the mothur software.

Preprocessing:
★ The reads were split by barcode and filtered for quality
★ Reads with an average quality of less than 20 over the entire read were removed.
★ No ambiguous bases and only one mismatch was allowed in the barcode sequence.
★ The reads were aligned to the silva reference database that contains 14,956 prealigned 16sRNA genes.
★ The alignment is then screened to include ones that overlap over the same region.
★ The remainder of the reads were passed through a chimera detection tool (chimera slayer).
★ Finally a pre-clustering method is applied to remove reads that are likely due to pyrosequencing errors.
★ The resulting alignment file is used to compute a pairwise distance file that is used to cluster the reads.

Taxonomical classification:
★ mothur software used to perform a bayesian classification against the silva reference template and taxonomy.
★ custom programs used to post-process and visualize the output from mothur.
# Stool 1 Data Analysis

## Read Information:

### Stool 1 Sample:

<table>
<thead>
<tr>
<th>Quality/Barcode</th>
<th>Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobio</td>
<td>51,161</td>
</tr>
<tr>
<td>Qiagen</td>
<td>63,223</td>
</tr>
<tr>
<td>Scoda</td>
<td>97,680</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>212,064</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trimmed/Unique</th>
<th>Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobio</td>
<td>6,570</td>
</tr>
<tr>
<td>Qiagen</td>
<td>7,069</td>
</tr>
<tr>
<td>Scoda</td>
<td>11,353</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>24,992</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unique/Aligned</th>
<th>Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobio</td>
<td>2,473</td>
</tr>
<tr>
<td>Qiagen</td>
<td>2,646</td>
</tr>
<tr>
<td>Scoda</td>
<td>3,759</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td>8,878</td>
</tr>
</tbody>
</table>
Classification against the SILVA gold aligned genes:

Phylum (2) level classification, mincount=10, normalized=FALSE

Counts

TaxonID

0.1.11
0.1.15
0.1.2
0.1.24
0.1.28
0.1.5

label
- 0.1.11 Cyanobacteria
- 0.1.15 Firmicutes
- 0.1.2 Actinobacteria
- 0.1.24 Proteobacteria
- 0.1.28 TM7
- 0.1.5 Bacteroidetes

variable
- Mobile
- Qiagen
- Sooda
Classification against the SILVA gold aligned genes:

The cluster heatmaps display relative abundances (percent of reads) with color scaled either by column (pink) or row (green). Higher percentage will be show as stronger color.
# Soil Data Analysis

**Read Information:**

<table>
<thead>
<tr>
<th>Quality info</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Reads</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>181,908</td>
<td></td>
</tr>
<tr>
<td>Scrapped</td>
<td>27,886</td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>154,022</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality/Barcode</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Reads</td>
<td></td>
</tr>
<tr>
<td>Mobio</td>
<td>103,831</td>
<td></td>
</tr>
<tr>
<td>Scoda</td>
<td>50,191</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>154,022</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trimmed/Unique</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Reads</td>
<td></td>
</tr>
<tr>
<td>Mobio</td>
<td>87,284</td>
<td></td>
</tr>
<tr>
<td>Scoda</td>
<td>43,721</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>131,005</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unique/Aligned</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>Reads</td>
<td></td>
</tr>
<tr>
<td>Mobio</td>
<td>53,679</td>
<td></td>
</tr>
<tr>
<td>Scoda</td>
<td>27,593</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>81,272</td>
<td></td>
</tr>
</tbody>
</table>
Classification against the SILVA gold aligned genes:

Phylum (2) level classification, mincount=25

Variable:
- Stoda
- Mobio

Label:
- 0.1.1 Acidobacteria
- 0.1.11 Candidate_division_OP10
- 0.1.2 Actinobacteria
- 0.1.20 Candidate_division_TM7
- 0.1.22 Candidate_division_WS3
- 0.1.26 Chloroflexi
- 0.1.28 Cyanobacteria
- 0.1.34 Firmicutes
- 0.1.37 Gammaproteobacteria
- 0.1.43 Planctomycetes
- 0.1.44 Proteobacteria
- 0.1.50 Verrucomicrobia
- 0.1.51 WCHB1-60
- 0.1.6 Bacteroidetes
Classification against the SILVA gold aligned genes:

The cluster heatmaps display relative abundances (percent of reads) with color scaled either by column (pink) or row (green). Higher percentage will be shown as a stronger color.
The joint effort continues …..

### DSRG Committee:

### Our Collaborators:
Bruns, MA, Gloor, G, Johnson, T, Albert, I

### Our Supporters:
Roche, Illumina, Qiagen, Mo Bio, Boreal Genomics

Thank You!