Introduction to Metagenomics Analysis for Next Generation Sequencing Data

Noushin Ghaffari, PhD
Bioinformatics Scientist, Genomics and Bioinformatics, Texas A&M AgriLife Research
Research Scientist, Texas A&M High Performance Research Computing
Outline

• Background
• Sequencing
• Application of Next Generation Sequencing in Research
• Metagenomics
Primary NGS Applications

1. Alignment
2. Assembly (no reference/with a reference)
   • Genome
   • Transcriptome
3. RNA-Seq
4. Metagenomics
5. ChIP-Seq
6. RADSeq

Last week
Two weeks ago
Today
Next Month
Why sequencing?

Determining the sequence of nucleotides within a DNA (or RNA) fragment

How?

Using sequencing methods, such as Sanger sequencing, next generation sequencing and single-molecule techniques
Next Generation Sequencing Platforms

- Sanger
- Illumina
- PacBio
- MinION

Third Generation Sequencing Platforms

Choosing among Illumina Sequencers

- **MiniSeq**
  - MAX OUTPUT: 8 Gb
  - MAX READ NUMBER: 25 million
  - MAX READ LENGTH: 2x150 bp

- **MiSeq**
  - MAX OUTPUT: 15 Gb
  - MAX READ NUMBER: 25 million
  - MAX READ LENGTH: 2x300 bp

- **NextSeq**
  - MAX OUTPUT: 120 Gb
  - MAX READ NUMBER: 400 million
  - MAX READ LENGTH: 2x150 bp

- **HiSeq 4000**
  - MAX OUTPUT: 1500 Gb
  - MAX READ NUMBER: 5 billion
  - MAX READ LENGTH: 2x150 bp

- **HiSeq X Ten**
  - MAX OUTPUT: 1800 Gb
  - MAX READ NUMBER: 6 billion
  - MAX READ LENGTH: 2x150 bp

NGS Sequencing Workflow

DNA/RNA extraction

Library creation/amplification

Sequencing (Illumina HiSeq or Roche 454)

Data Analysis

**Pre-processing:** Base calling, Generating output sequences files (FASTQ), Quality Control (QC)

**Initial processing:** Alignment, De novo assembly

**RNA-Seq:** Normalization, Counting, Expression analysis

**Discovery:** SNP, CNV, Annotation
Illumina next-generation sequencing

Sequencing by Synthesis (SBS) Technology

- Randomly shearing DNA
- Attaching DNA fragments to the flowcell surface
- Cluster generation, “Bridge Amplification”
- Adding four labelled reversible terminators, primers, and DNA polymerase
- Determining the attached nucleotide, based on the emitted fluorescence
Sequence and Quality Scores

Quality scores measure the probability that a base is called incorrectly.

Flow-cell surface

Read

Quality Score

Quality scores

adapter sequence

sequence fragment

adapter sequence
## Comparing Sequencing Platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>Read length</th>
<th>Error rates</th>
<th>Technology</th>
<th>Portable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>&lt; 400 bp</td>
<td>Low</td>
<td>Sequencing by synthesis</td>
<td>No</td>
</tr>
<tr>
<td>PacBio</td>
<td>~ 10-15 Kb</td>
<td>High</td>
<td>SMRT – ZMW</td>
<td>No</td>
</tr>
<tr>
<td>Oxford Nanopore Technologies</td>
<td>~ 5-8 Kb</td>
<td>High</td>
<td>Nanopore protein – strand</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sequencing</td>
<td></td>
</tr>
</tbody>
</table>
Metagenomics
What is Metagenomics?

Study of communities of microbial organisms directly in their natural environments Without the need for isolation and lab cultivation of individual species

Moved from traditional BAC cloning to NGS long reads or high coverage short reads

Metagenomics Techniques

1. Whole Genome Shotgun (WGS)
2. Marker Gene
   • 16S rRNA
     • Bacteria, Archaea
   • 18S rRNA
     • Fungus, Eukaryotes

Whole Genome Shotgun (WGS) Metagenomics

- Sequencing the whole genome of the organisms present in the sample
- Facilitates discovering gene/gene function, genome structure
- Studying the evolutionary relationships for microbiomes
Whole Genome Shotgun (WGS) Metagenomics

- Sequencing the whole genome of the organisms present in the sample
- Facilitates discovering gene/gene function, genome structure
- Studying the evolutionary relationships for microbiomes
- Steps
  - Genome Assembly
  - Binning
  - Predicting and Annotating Genes

WGS Metagenomics Tools

Assembly
- Velvet, MetaVelvet, MetaVelvet-SL
- IDBA-UD
- MetAMOS pipeline: selecting assembly, scaffolding, annotating
- Genome Assemblers such as ALLPATHS, SOAP and ABySS

Binning
- LikelyBin
- PHYSCIMM
- MetaCluster
- MetaWatt
- MetaPhyler
- PhymmBL

Annotation
- MetaGeneAnnotator
- Glimmer-MG
- FragGeneScan
- MetaGeneMark
- Kraken
Marker Gene Metagenomics

• Usually based on 16S RNA
  • Conserved within species
  • Greatly different between species
  • Widely used for microbial ecology
• Needs a reference database to match the Operational Taxonomic Units (OTU)
  • Silva
  • Ribosomal Database Project
  • Unite
• Steps
  • Preprocessing to remove noise
  • OUT clustering and taxonomic assignment
  • Alpha diversity analysis – within sample diversity
  • Beta diversity analysis – between sample diversity
Marker Gene Metagenomics Tools

Microbial community analysis
- QIIME
- Mother
- SILVAngs
- MG-RAST
- MEGAN

Diversity analysis
- Chao
- UniFrac
- PCoA

Visualization
- QIIME
- MEGAN
- FigTree
Metagenomics Studies

• PathoMap
  • Research project by Weill Cornell Medical College to study the microbiome and metagenome of the built environment of NYC

• Cow rumen microbiome study
  • 220 bacterial and archaeal genomes assembled directly from 768 GB rumen sequenced data
  • Majority unsequenced strains and species of bacteria and archaea
  • Over 13,000 proteins predicted to be involved in carbohydrate metabolism, over 90% of which do not have a good match in the public databases
  • Assembly of hundreds of microbial genomes from the cow rumen reveals novel microbial species encoding enzymes with roles in carbohydrate metabolism
Metagenomics Web-Based Tools

MG-RAST
- Available tools, via PATRIC
- RAST: Rapid Annotation using Subsystem Technology
- Annotating the assembled contigs of a bacterial and archaeal genomes
- Quantitative insights for microbial populations, based on NGS data
MG-RAST Pipeline

Upload → Preprocessing → Dereplication → DRISEE → Screening → Gene Calling → AA Clustering 90% → Protein Identification → Annotation Mapping → Abundance Profiles → RNA detection → RNA Clustering 97% → RNA Identification → Done

Metagenome Analysis

Data Type
- Organism Abundance
- Representative Hit Classification

Data Selection
- Metagenomes
- Annotation Sources
- Max. e-Value Cutoff
- Min. % Identity Cutoff
- Min. Alignment Length Cutoff

Data Visualization
- Bar chart
- Line chart
- Table
- Heatmap
- PCoA
- Randomization

Subsystems
- Download chart data
- View Subsystems interactive chart

Sequence Breakdown
- Fixed GG
- Unknown
- Unclassified
- Annotated Proteins
- Raw Reads

Subsystems
- Amino Acids and Derivatives (2777)
- Carbohydrates
- Cell Division and Cell Wall and Capsules
- Clustering-based sub...
- Coenzymes, Vitamins, P...
- DNA Metabolism
- Domains and Sporul...
- Fatty Acids, Lipids, an...
- Iron acquisition and m...
- Membrane Transport
- Metabolism of Amin...
- Miscellaneous
- Motility and Chemotaxis

Download chart data
- 42,515 predicted functions
- 79.8% of predicted proteins
- 104.4% of annotated proteins

View Subsystems interactive chart

Texas A&M University     High Performance Research Computing  –  https://hprc.tamu.edu
UniFrac
Calculates distance between microbial communities, using phylogenetic trees

FastUniFrac
- Adapted for NGS data
- Incorporated with Galaxy tools
- The same idea as UniFrac

UniFrac vs FastUniFrac
- Input: tree
- Newick (PHYLIP package output) or Nexus (TAXA, CHARACTER, DATA, TREE blocks (Newick format))
- Tagging each sequence’s environment, Creating Sample ID map
- Analysis
- Measuring the overall difference between each pair of environments
- Clustering the environments
- Principal Coordinates Analysis (3D in FastUniFrac)

Practical Portion
Logging in to the system

- **SSH (secure shell)**
  - The only program allowed for remote access; encrypted communication; freely available for Linux/Unix and Mac OS X hosts;
- **For Microsoft Windows PCs, use MobaXterm**
  - https://hprc.tamu.edu/wiki/HPRC:MobaXterm
  - You are able to view images and use GUI applications with MobaXterm
  - or **Putty**
    - https://hprc.tamu.edu/wiki/HPRC:Access#Using_PuTTY
    - You can not view images or use GUI applications with PuTTY

- Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts
- Don’t write down passwords
- Don’t choose easy to guess/crack passwords
- Change passwords frequently
Using SSH - MobaXterm (on Windows)

- IMPORTANT POLICY INFORMATION -
- Unauthorized use of HPRC resources is prohibited and subject to criminal prosecution.
- Use of HPRC resources in violation of United States export control laws and regulations is prohibited. Current HPRC staff members are US citizens and legal residents.
- Sharing HPRC account and password information is in violation of State Law. Any shared accounts will be disabled.
- Authorized users must also adhere to all policies at: https://hprc.tamu.edu/wiki/index.php/HPRC_Policies

** WARNING: There are NO active backups of user data. **

Please restrict usage to 8 CORES across ALL Ada login nodes. Users found in violation of this policy will be SUSPENDED.

**** Ada Scheduled Maintenance Completed ****
The maintenance for Ada has been completed. Batch job scheduling has resumed.

Your current disk quotas are:
- Disk: 117.2M, Limit: 1419, File Usage: 37, Limit: 500000
- Type “showquota” to view these quotas again.

message of the day

your quotas
Using SSH to Access Ada

You may see something like the following the first time you connect to the remote machine from your local machine:

- Host key not found from the list of known hosts.
- Are you sure you want to continue connecting (yes/no)?

Type yes, hit enter and you will then see the following:

- Host 'ada.tamu.edu' added to the list of known hosts.
- user_NetID@ada.tamu.edu's password:
Metagenomics Practice - Tool

Mothur

- Open-source
- Serves the microbial ecology community
- DOTUR and SONS programs
- Data: Sanger, PacBio, IonTorrent, 454, and Illumina MiSeq and HiSeq
- Most cited bioinformatics tool for analyzing 16S rRNA gene sequences
- Our practical session will use Mothur to demonstrate a typical MiSeq data analysis
Metagenomics Practice - Data

• Objective: Understanding the effect of normal variation in the gut microbiome on host health
• Collected 365 (daily basis) fresh feces from mice, post weaning
• No treatment in first 150 days post weaning (dpw)

• Question: rapid change in weight observed during the first 10 dpw affected the stability microbiome when compared with microbiome in days 140 – 150 or not?
• Mock community composed of genomic DNA from 21 bacterial strains

<table>
<thead>
<tr>
<th>group</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>F3D0</td>
<td>Early</td>
</tr>
<tr>
<td>F3D1</td>
<td>Early</td>
</tr>
<tr>
<td>F3D141</td>
<td>Late</td>
</tr>
<tr>
<td>F3D142</td>
<td>Late</td>
</tr>
<tr>
<td>F3D143</td>
<td>Late</td>
</tr>
<tr>
<td>F3D144</td>
<td>Late</td>
</tr>
<tr>
<td>F3D145</td>
<td>Late</td>
</tr>
<tr>
<td>F3D146</td>
<td>Late</td>
</tr>
<tr>
<td>F3D147</td>
<td>Late</td>
</tr>
<tr>
<td>F3D148</td>
<td>Late</td>
</tr>
<tr>
<td>F3D149</td>
<td>Late</td>
</tr>
<tr>
<td>F3D150</td>
<td>Late</td>
</tr>
<tr>
<td>F3D2</td>
<td>Early</td>
</tr>
<tr>
<td>F3D3</td>
<td>Early</td>
</tr>
<tr>
<td>F3D5</td>
<td>Early</td>
</tr>
<tr>
<td>F3D6</td>
<td>Early</td>
</tr>
<tr>
<td>F3D7</td>
<td>Early</td>
</tr>
<tr>
<td>F3D8</td>
<td>Early</td>
</tr>
<tr>
<td>F3D9</td>
<td>Early</td>
</tr>
</tbody>
</table>
Looking at Data!

cd /scratch/training/NGS_metagenomics
ls -l

cd Data
ls -l

cd MiSeq_SOP
ls -l

head –n 16 F3D1_S189_L001_R1_001.fastq
Running Mothur

• Loading modules and calling the program

```
module load VSEARCH/2.3.0-intel-2016a \ Mothur/1.38.1.1-intel-2016a-Python-2.7.11
mothur
```

• To run the preprocessing script

```
cd /scratch/training/NGS_metagenomics/Data/MiSeq_SOP
mothur preprocessing.batch
```
Login and Set up

- Login to Ada using SSH or MobaXterm
- Let’s take a look at the path and create appropriate directories

```
echo $SCRATCH

cd $SCRATCH

Pwd

mkdir NGS_assembly_Oct17
mkdir NGS_assembly_Oct17/Data
mkdir NGS_assembly_Oct17/Scripts
mkdir NGS_assembly_Oct17/Outputs
```
Preprocessing of the Data

Processes done by pre-processing script:
- Making contigs for PE input data
- Mapping to reference
- Checking the alignment output
- Removing chimeras
- Assessing error rates

```
cd /scratch/training/NGS_metagenomics/Outputs
ls -l
```
Analyzing Data

Use either methods to look at the file and copy/paste the codes

```
cd /scratch/training/NGS_metagenomics/Scripts
ls -l

cp /scratch/training/NGS_metagenomics/Scripts/Analysis_Commands.txt $SCRATCH

cd $SCRATCH
cat Analysis_Commands.txt

Scp username@ada.tamu.edu:path-to-script .
```
FigTree Visualization

- Login to Ada using SSH using “-X”
- Move to the correct directory (alternatively, you can add the path to your $PATH)
  
  cd /scratch/training/NGS_metagenomics/FigTree/lib
  java -Xms64m -Xmx512m -jar figtree.jar $*

- FigTree window will appear on your screen
- Use File → Open to open a tree file (.tre)
Any question?

nghaffari@tamu.edu