Microbial genome pipeline

Li Liu Ph.D.
Institute of Microbiology,
Chinese Academy of Sciences
2014-9-12
Next Generation Sequencing

Sequencing and Annotating microbial genome
Introduction to sequencing and application
What is sequencing?

Finding the sequence of a DNA/ RNA molecule

What can we sequence?

http://cancergenome.nih.gov/newsevents/multimedialibrary/images/CancerBiology
Sanger sequencing

• Up to 1,000 bases molecule
• One molecule at a time

• Widely used from 1970-2000
• First human genome draft was based on Sanger sequencing
• Still in use for single molecules

http://www.genomebc.ca/education/articles/sequencing/
High Throughput Sequencing

Next Generation Sequencing (NGS) / Massively parallel sequencing

• Sequencing **millions** of molecules **in parallel**
• Do not need prior knowledge of what you’re sequencing

<table>
<thead>
<tr>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 sequencing</td>
</tr>
<tr>
<td>SOLiD/PGM</td>
</tr>
<tr>
<td>HiSeq/MiSeq</td>
</tr>
</tbody>
</table>

We will discuss Illumina’s platform only
## Summary of Sequencers

1 = *de novo* BACs, plastids, microbial genomes.
2 = transcriptome characterization.
3 = targeted re-sequencing.
4 = *de novo* plant and animal genomes.
5 = re-sequencing and transcript counting.
6 = mutation detection.
7 = metagenomics.
8 = other (ChIP-Seq, etc.)

<table>
<thead>
<tr>
<th>Platform</th>
<th>Current company</th>
<th>Former company</th>
<th>Sequencing method</th>
<th>Amplification method</th>
<th>Claim to fame</th>
<th>Primary applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>Roche</td>
<td>454</td>
<td>Synthesis</td>
<td>emPCR</td>
<td>First Next-Gen Sequencer, Long reads</td>
<td>1*, 2, 3*, 4, 7, 8*</td>
</tr>
<tr>
<td>Illumina</td>
<td>Illumina</td>
<td>Solexa</td>
<td>Synthesis</td>
<td>BridgePCR</td>
<td>First short-read sequencer; current leader in advantages†</td>
<td>1*, 2, 3*, 4, 5, 6, 7, 8</td>
</tr>
<tr>
<td>SOLiD</td>
<td>Life Technologies</td>
<td>Applied Biosystems</td>
<td>Ligation</td>
<td>emPCR</td>
<td>Second short-read sequencer; low error rates</td>
<td>3*, 5, 6, 8</td>
</tr>
<tr>
<td>HeliScope</td>
<td>Helicos</td>
<td>N/A</td>
<td>Synthesis</td>
<td>None</td>
<td>First single-molecule sequencer</td>
<td>5, 8</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>Life Technologies</td>
<td>Ion Torrent</td>
<td>Synthesis (H⁺ detection)</td>
<td>emPCR</td>
<td>First Post-light sequencer; first system &lt;$100 000</td>
<td>1, 2, 3, 4, 8</td>
</tr>
<tr>
<td>PacBio</td>
<td>Pacific Biosciences</td>
<td>N/A</td>
<td>Synthesis</td>
<td>None</td>
<td>First real-time single-molecule sequencing</td>
<td>1, 2, 3, 7, 8</td>
</tr>
<tr>
<td>Starlight‡</td>
<td>Life Technologies</td>
<td>N/A</td>
<td>Synthesis</td>
<td>None</td>
<td>Single-molecule sequencing with quantum dots</td>
<td>1, 2, 7, 8</td>
</tr>
</tbody>
</table>
# Illumina Sequencers

<table>
<thead>
<tr>
<th>Key applications</th>
<th>MiSeq</th>
<th>NextSeq 500</th>
<th>HiSeq 2500</th>
<th>HiSeq X*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small genome, amplicon, and targeted gene panel sequencing.</td>
<td>Focused power.</td>
<td>Flexible power.</td>
<td>Production power.</td>
<td>Population power.</td>
</tr>
<tr>
<td>Everyday genome, exome, transcriptome sequencing, and more.</td>
<td>Speed and simplicity for targeted and small genome sequencing.</td>
<td>Speed and simplicity for everyday genomics.</td>
<td>Power and efficiency for large-scale genomics.</td>
<td>$1,000 human genome and extreme throughput for population-scale sequencing.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run mode</th>
<th>N/A</th>
<th>Mid-Output</th>
<th>High-Output</th>
<th>Rapid Run</th>
<th>High-Output</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow cells processed per run</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1 or 2</td>
</tr>
<tr>
<td>Output range</td>
<td>0.3-15 Gb</td>
<td>20-39 Gb</td>
<td>30-120 Gb</td>
<td>10-180 Gb</td>
<td>50-1000 Gb</td>
<td>1.6-1.8 Tb</td>
</tr>
<tr>
<td>Run time</td>
<td>5-55 hours</td>
<td>15-26 hours</td>
<td>12-30 hours</td>
<td>7-40 hours</td>
<td>&lt; 1 day - 6 days</td>
<td>&lt; 3 days</td>
</tr>
<tr>
<td>Reads per flow cell†</td>
<td>25 Million‡</td>
<td>130 Million</td>
<td>400 Million</td>
<td>300 Million</td>
<td>2 Billion</td>
<td>3 Billion</td>
</tr>
<tr>
<td>Maximum read length</td>
<td>2 × 300 bp</td>
<td>2 × 150 bp</td>
<td>2 × 150 bp</td>
<td>2 × 150 bp</td>
<td>2 × 125 bp</td>
<td>2 × 150 bp</td>
</tr>
</tbody>
</table>
Sequencing Workflow

1. Extract tissue cells
2. Extract DNA/RNA from cells
3. Sample preparation for sequencing
4. Sequencing
5. Bioinformatics analysis

Why is it important to understand the “wet lab” part?
Sample Prep

Random shearing of the DNA

Adding adaptors and barcodes

Size selection

Amplification

Sequencing
Sequencing process
Sequencing process

Leave only sequences from one direction
Sequencing process
Sequencing process
DNA sequencing type

- **De-novo sequencing** – sequencing the genome of an organism with an **unknown** genome
- **Resequencing** – sequencing the genome of an organism with a **known** genome
Basic concepts

**Insert** – the DNA fragment that is used for sequencing.

**Read** – the part of the insert that is sequenced.

**Single Read (SR)** – a sequencing procedure by which the insert is sequenced from one end only.

**Paired End (PE)** – a sequencing procedure by which the insert is sequenced from both ends.
Example of NGS’s application
--genome resequencing
Demultiplexing

Unknown inserts
Example of mapping parameters:
• Number of mismatches per read
• Scores for mismatch or gaps

Mapping parameters affect the rest of the analysis
Removing duplicates and non-unique mappings

Demultiplexing

Mapping

\[ \text{average coverage} = \frac{\text{read length} \times \text{number of uniquely mapped reads}}{\text{genome size}} \]
Demultiplexing

Mapping

Removing duplicates and non-unique mappings

Coverage profile and variant calling
coverage profile and variant calling

Removing duplicates and non-unique mappings

Coverage profile and variant calling

Variant filtering

- **Demultiplexing**
- **Mapping**
- **Removing duplicates and non-unique mappings**
- **Coverage profile and variant calling**
- **Variant filtering**

**Frequency >= 20%**

Reference Genome

**Coverage >= 5**

Reference Genome

...ACTTCGTCGAAAGG...

False SNP
Variant calling

Removing duplicates and non-unique mappings

Variant filtering

Genes and known variants

Reference Genome

... ACTTCGTGAAATG ...

... GTCCGTGATACTCCGT ...

Gene X

rs230985

G

A

Gene X
## Resequencing results

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>Reference</th>
<th>Allele/s</th>
<th>Indel/SNP</th>
<th>Phred Q-score for allele call</th>
<th>Coverage in Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>44050195</td>
<td>C</td>
<td>T</td>
<td>SNP</td>
<td>195</td>
<td>16</td>
</tr>
<tr>
<td>19</td>
<td>44111582</td>
<td>CACACAC</td>
<td>CACAC</td>
<td>INDEL</td>
<td>132</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Genotype Quality</th>
<th>Genes</th>
<th>dbSNP_132</th>
<th>1000Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,1</td>
<td>90</td>
<td></td>
<td>rs25479: C -&gt; T</td>
<td>gene_name = &quot;ZNF428&quot;; gene_biotype = &quot;protein_coding&quot;;</td>
</tr>
<tr>
<td>0,1</td>
<td>68</td>
<td></td>
<td>:: C -&gt; T</td>
<td></td>
</tr>
</tbody>
</table>
Example for further analysis

**Recessive disease:**
1. Variant not in known databases
2. Homozygous variant shared by all affected individuals
3. Same variant appears in healthy parents at heterozygous state
4. Healthy brothers can be heterozygous to the same variant

**Dominant disease:**
1. Variant not in known databases
2. Heterozygous variant shared by all affected individuals
3. The variant doesn’t appear in healthy individuals
Quality control steps in the pipeline

1. Demultiplexing
2. Mapping
3. Removing duplicates and non-unique mappings
4. Coverage profile and variant calling
5. Variant filtering
6. Genes and known variants
7. Finding suspicious variants

QC steps at each stage: QC, QC, QC, QC, QC, QC.
Introduction to microbial genome sequencing and annotation
<table>
<thead>
<tr>
<th>Tools for Genome sequencing and annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genome sequencing</strong></td>
</tr>
<tr>
<td><strong>Transcriptome sequencing</strong></td>
</tr>
<tr>
<td><strong>Genome assembly</strong></td>
</tr>
<tr>
<td><strong>Gene Prediction</strong></td>
</tr>
<tr>
<td><strong>Annotation</strong></td>
</tr>
<tr>
<td><strong>Repeatmasking</strong></td>
</tr>
</tbody>
</table>
Workflow for microbial genome

- Libraries
- Sequencing
- Assembly
- Closure
- Annotation
- Release
Workflow for microbial genome

1. Libraries
2. Sequencing
3. Assembly
4. Closure
5. Annotation
6. Release
Genome

Physical methods (Sonication)

Genomic Fragments (200 nt or 400 nt or 1kb)
Workflow for microbial genome

1. Libraries
2. Sequencing
3. Assembly
4. Closure
5. Annotation
6. Release
Short Reads

Genomic Fragments (200 nt or 400 nt or 1kb)

Low cost & Less time

Single end sequencing

454 FLX

Ion torrent

Paired end sequencing

ABI’s Solid

Illumina

Short Reads
Workflow for microbial genome

- Libraries
- Sequencing
  - Assembly
    - Closure
  - Annotation
- Release
De novo assembly strategies

- **SSAKE**
  - Warren et al., 2007
  - Uses DNA prefix tree to find k-mer matches

- **Edena**
  - Hernandez et al., 2008
  - overlap-layout algorithm adapted for short reads

- **Velvet**
  - Zerbino and Birney, 2008
  - Uses DeBruijn graph algorithm plus error correction
Workflow for microbial genome
Genome annotation

- A process of attaching biological information to sequences (contigs or chromosomes).

- Consists of two main steps:
  
  A. Identifying elements on genome a process called gene prediction (*Structural annotation*).
  
  B. Attaching biological information to these elements (*Functional annotation*).
Genome annotation

• **Structural annotation**
  - ORFs and their localization
  - Gene structure
  - Coding regions
  - Location of regulatory motifs

• **Functional annotation**
  - Biochemical function
  - Biological function
  - Involved regulation and interactions
  - Expression
Genome annotation

• Can be done manually (require human expertise) or with automated pipelines.

• Pipelines available:
  - PGAAP (NCBI)
  - RAST server
  - IMG-ER,
  - ISGA
  - MAKER (for eukaryotes).
Workflow for microbial genome

- Libraries
- Sequencing
- Assembly
- Closure
- Annotation
- Release
Genome submission to NCBI (GenBank)

• NCBI (GenBank) accepts both complete and incomplete genomes (contigs produced after genome assembly).
Publications

• Whole genome assembly and annotation of microbes with preliminary analysis can be published in reputed journals like Journal of Bacteriology (http://jb.asm.org/) and Eukaryotic cell (http://ec.asm.org/).

• Other journals are Genome Biology, Genome Reaserch and Nature Biotechnology(according to the analysis done).
Burkholderia sp. SJ98

- Degrade a number of aromatic compounds, e.g., p-nitrophenol, o-nitrobenzoate, p-nitrobenzoate, and 4-nitrocatechol (Pandey G, et. al. 2002), 2-chloro-4-nitrophenol (Pandey J, et al. 2011), and 3-methyl-4-nitrophenol (Bhushan B, et. al. 2000).

**Burkholderia sp. SJ98 genome sequence**

- Roche’s 454 FLX
- Short Reads
- Nebwler 2.5.3
- Contigs
- RAST, tRNA-scan v1.21 and RnAmmer v1.2

**Genome size**
- 7.89-Mb

**Large contigs**
- 79

**Protein coding genes**
- 7,364

**rRNAs**
- 3

**tRNAs**
- 51
Genome Sequence of the Nitroaromatic Compound-Degrading Bacterium Burkholderia sp. Strain SJ98

Shailesh Kumar, Surendra Vikram and Gajendra Pal Singh Raghava

Strain RKJ300 is capable of utilizing 4 nitrophenol, 2-chloro-4-nitrophenol, and 2, 4-dinitrophenol as sole sources of carbon and energy (Ghosh A, et al. 2010).

### Rhodococcus imtechensis sp. RKJ300

- **Illumina GAIIx**
- **Short Reads**
- **NGS QC toolkit v2.2.1**
- **Filtered Short Reads**
- **SOAPdenovo v1.05**
- **Contigs**
- **RAST, tRNA-scan v1.21 and RNAmmer v1.2**

<table>
<thead>
<tr>
<th>Genome size</th>
<th>8.231-Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs produced</td>
<td>178</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>8,059</td>
</tr>
<tr>
<td>rRNAs</td>
<td>5</td>
</tr>
<tr>
<td>tRNAs</td>
<td>49</td>
</tr>
</tbody>
</table>
Strain K1T is known to possess various enzymatic activities, such as lipase, γ-glutamyl transferase, glycine arylamidase, and Glu-Gly-Arg-arylamidase (Vikram S et al. 2012).

**Imtechella halotolerans K1^T**

- **Illumina HiSeq**
- **Short Reads**
- **NGS QC toolkit v2.2.1**
- **Filtered Short Reads**
- **Velvet v1.2.03**
- **Contigs**
- **RAST, tRNA-scan v1.21 and RNAmmer v1.2**

<table>
<thead>
<tr>
<th>Genome size</th>
<th>3.087-Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs produced</td>
<td>44</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>2,738</td>
</tr>
<tr>
<td>rRNAs</td>
<td>3</td>
</tr>
<tr>
<td>tRNAs</td>
<td>38</td>
</tr>
</tbody>
</table>
The strain is capable of gelatin liquefaction. All the strains of the genus Marinilabilia were reported to decompose various biomacromolecules.
CHALLENGES

Removal of artifacts in short reads ??

Several assemblers available, which is best ??

Genome assembly of short reads ??

Annotation and validation of assembled genome ??
Challenges of genome sequencing

• Data produce in form of short reads, which have to be assembled correctly in large contigs and chromosomes.
• Short reads produced have low quality bases and vector/adaptor contaminations.
• Several genome assemblers are available but we have to check the performance of them to search for best one.
Thanks for your attention