Next-generation sequencing (NGS) for plant research

Presented by Daisuke Tsugama

Email: tsugama@res.agr.hokudai.ac.jp
Tel: 011-706-2471
Room: S268 (Lab of Crop Physiology)
Slides used for this class can be downloaded at http://www.agr.hokudai.ac.jp/botagr/sakusei/materials.html
This class ...  

• Introduces theories and applications of NGS, which is now very popular in plant research, from an experimental biologist’s viewpoint  

• Aims at letting you know  
  ✓ what is NGS  
  ✓ what is usually done in NGS data analysis  
  ✓ applications of NGS  
  ✓ NGS is not something to fear  

• Assesses you on the basis of a small test attached to the end of the handout
1. What is NGS like?
   • Sequencers for NGS
   • Basics of NGS data analysis

2. Applications of NGS
   • RNA-Seq
   • Genome sequencing
   • RAD-Seq
   • MutMap and QTL-Seq
   • Others
1. What is NGS like?
   • Sequencers for NGS
   • Basics of NGS data analysis

2. Applications of NGS
   • RNA-Seq
   • Genome sequencing
   • RAD-Seq
   • MutMap and QTL-Seq
   • Others
## Sequencers for NGS

<table>
<thead>
<tr>
<th>Sequencer</th>
<th>Company</th>
<th>Output</th>
<th>Read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS-FLX</td>
<td>454 Life Sciences (Roche)</td>
<td>~400 Mb</td>
<td>~500 b</td>
</tr>
<tr>
<td>Ion Proton</td>
<td>Life Technologies (Thermo)</td>
<td>~10 Gb</td>
<td>~200 b</td>
</tr>
<tr>
<td>HiSeq 2500</td>
<td>Illumina</td>
<td>~1 Tb</td>
<td>~200 b</td>
</tr>
<tr>
<td>PacBio RS II</td>
<td>Pacific Biosciences</td>
<td>~1 Gb</td>
<td>~40 kb</td>
</tr>
</tbody>
</table>

* Output (b / run) = read length (b/read) × # of reads
Ion Proton semiconductor

[Link to Ion semiconductor sequencing]

Illumina HiSeq 2000

## Sequencers for NGS

<table>
<thead>
<tr>
<th>Sequencer</th>
<th>Company</th>
<th>Output</th>
<th>Read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS-FLX</td>
<td>454 Life Sciences (Roche)</td>
<td>~400 Mb</td>
<td>~500 b</td>
</tr>
<tr>
<td>Ion Proton</td>
<td>Life Technologies (Thermo)</td>
<td>~10 Gb</td>
<td>~200 b</td>
</tr>
<tr>
<td>HiSeq 2500</td>
<td>Illumina</td>
<td>~1 Tb</td>
<td>~200 b</td>
</tr>
<tr>
<td>PacBio RS II</td>
<td>Pacific Biosciences</td>
<td>~1 Gb</td>
<td>~40 kb</td>
</tr>
</tbody>
</table>

HiSeq and PacBio have been gaining popularity
Illumina NGS technology

DNA to be sequenced

DNA fragmentation

Addition of adapters

Annealing sites for the bridge PCR

Index (barcode) for multiplex analysis

Annealing sites for sequencing primers
Illumina NGS technology

DNA to be sequenced

DNA fragmentation

Addition of adapters

Bridge PCR & Cluster formation

Glass flow cell covered with primers for the bridge PCR
Illumina NGS technology

1. DNA to be sequenced
2. DNA fragmentation
3. Addition of adapters
4. Bridge PCR & Cluster formation
5. Signal detection (~100 times)
Illumina NGS technology

• Single-end read: obtained by only one primer
  
  ![Single-end read diagram](image)
  
  ~100 b

• Paired-end read: obtained by two primers
  
  ![Paired-end read diagram](image)
  
  ~100 b
  
  ~100 b

• Multiplex analysis: uses more than two indexes
  
  Sample A-derived read
  
  ![Sample A diagram](image)

  Sample B-derived read
  
  ![Sample B diagram](image)

  Sample C-derived read
  
  ![Sample C diagram](image)
PacBio NGS technology

DNA to be sequenced

DNA fragmentation

Addition of adapters

Addition of a primer and a DNA polymerase

Single molecule real-time (SMRT) sequencing

DNA polymerase (1 molecule / well)

~40-kb elongation

The detector detects only fluorescent signals retained longer than 1 msec on the bottom (around the DNA pol) of the well.
NGS data analysis

Run NGS to get reads

Assemble reads into contigs

Map reads to a reference

*Reference: a genome, transcripts, obtained contigs etc.

Evaluate mapping results for further analyses
NGS data analysis – read data

Read data are often handled in the fastq format

@MachineX:1:1:1:1#0/1
TNAGCTTTACGTATAGGCCCCC
+#!1508<iO{TRkol&389M|aR~y

Information for the read “MachineX:1:1:1:1#0/1”

@MachineX:1:1:1:2#0/1
ATTGCATTGTAAGTTGGGGCCT
+ |
... (usually a great number of reads follow)
An assembly requires a lot of memory (e.g., de novo assembly for an ~3 Gb genome requires ~150 GB memory)
NGS data analysis – mapping

Mapping: associating each read with a reference

Read           Reference

Reference:
• Known genome
• Known transcripts
• Contigs obtained by *de novo* assembly
NGS data analysis – mapping

Mapping: associating each read with a reference

Read counts are 22 for all of these fragments

Read counts for each region (or fragment) of the reference are often used to interpret the data
1. What is NGS like?
   • Sequencers for NGS
   • Basics of NGS data analysis

2. Applications of NGS
   • RNA-Seq
   • Genome sequencing
   • RAD-Seq
   • MutMap and QTL-Seq
   • Others
RNA-Seq

• Is a transcriptome analysis using NGS

• Flow:
  RNA extraction → mRNA purification →
  mRNA shearing → cDNA synthesis → NGS

• Each contig derived from a *de novo* assembly corresponds to each kind of transcripts

• Expression levels of the transcripts are evaluated with FPKM, RPKM or TPM

• They are usually used for further analyses such as clustering and a GO analysis
RNA-Seq

- **FPKM**: fragments per kb of exon per million mapped fragments
- **RPKM**: reads per kb of exon per million mapped fragments
  
  \[ \text{FPKM of the contig } A = \frac{R_A \times 10^9}{N \times L_A} \]

  - \( R_A \) = # of reads mapped to A
  - \( N \) = total # of mapped reads
  - \( L_A \) = size of A

*FPKM = RPKM when reads are all single-end*
RNA-Seq

• TPM: transcripts per million

$$\text{TPM of the contig } A = \left( \frac{R_A}{L_A} \right) / \Sigma \left( \frac{R_i}{L_i} \right) \times 10^6$$

\( R_A = \# \text{ of reads mapped to } A \)
\( L_A = \text{size of } A \)

TPM is like

The copy number of mRNA of interest /
The total copy number of mRNA
RNA-Seq

Sample 1 (N = 66)

<table>
<thead>
<tr>
<th>Contig (gene)</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rx</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Lx</td>
<td>321</td>
<td>230</td>
<td>428</td>
</tr>
</tbody>
</table>

Sample 2 (N = 66)

<table>
<thead>
<tr>
<th>Contig (gene)</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rx</td>
<td>22</td>
<td>10</td>
<td>34</td>
</tr>
<tr>
<td>Lx</td>
<td>321</td>
<td>230</td>
<td>428</td>
</tr>
</tbody>
</table>

FPKM

Sample 1

- A: Low
- B: High
- C: Low

Sample 2

- A: High
- B: Low
- C: Highest

Relative expression level

- A: 1
- B: 0.25
- C: 4
**RNA-Seq**

<table>
<thead>
<tr>
<th>Contig (gene)</th>
<th>Sample 1 (N = 66)</th>
<th>Sample 2 (N = 66)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rx</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Lx</td>
<td>321</td>
<td>230</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FPKM Sample 1</th>
<th>FPKM Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TPM Sample 1</th>
<th>TPM Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

- **FPKM** and **TPM** are measures of gene expression levels in RNA-Seq analysis.
Genome sequencing

- Is sequencing a genome with NGS
- >30 × coverage is usually recommended
  E.g., for the human genome (~3 Gb), getting >90 Gb reads is preferable
- $2000 / 90 Gb if HiSeq X Ten is used
- $1000 / 1 Gb if PacBio RS II is used
- Plant genomes in general have large intergenic regions with many repetitive sequences
  → PacBio RS II has advantages over HiSeq X Ten if budget is sufficient
RAD-Seq

RAD-Seq: restriction site-associated DNA sequencing

Genomic DNA →

Restriction digestion →

Addition of 1\textsuperscript{st} adapter →

Further shearing of DNA →

Addition of 2\textsuperscript{nd} adapter →

Sequencing using the 1\textsuperscript{st} adapter
RAD-Seq

Benefits

• Regions in the vicinity of the restriction sites can be deeply (again and again) sequenced (thus accuracy is good)
• SNPs (single nucleotide polymorphisms) can be detected on a genome-wide scale
  *Regions sequenced by RAD-Seq is said to be 0.1-1% of the whole genome
  If an 8 b-recognizing restriction enzyme and single-end sequencing are used, the expected coverage would be:
    $100 \times \frac{100}{4^8} = \frac{10000}{65536} = 0.152... \ (%)$
• Many samples can be handled in each run using indexes
• RAD-Seq was used for developing GWAS with sorghum etc.
GWAS: genome-wide association study

Assessment of phenotypes of various cultivars

Assessment of their SNPs

| SNP1 | CV1 | CV2 | CV3 | CV4 | CV5 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>SNP2</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>SNP3</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
</tbody>
</table>
| SNP4 | C    | C   | A   | A   | C   | ...
| SNP5 | C    | C   | C   | C   | C   | ... |

Detection of the SNPs associated with the phenotype of interest
MutMap

Was developed to accelerate gene mapping

Wild type \times \text{Mutant}

F1

Mutagenized plants

M1

F2 or M2

Genome sequencing

Detection of SNPs linked to the mutation

Frequency

SNPs
QTL-Seq

Was developed to accelerate QTL analysis

CV1 × CV2

F1

F2

Genome sequencing

Detection of SNPs linked to the phenotype

Frequency

SNPs
Others (not really for plant research)

- Exome sequencing: targets genomic regions corresponding to exons
- Amplicon-Seq: targets PCR products to find rare SNPs in genetic disease-causing genes or to analyze microbiota (communities of microorganisms)
- Whole genome bisulfite sequencing: targets genomic DNA treated with bisulfite ion, which converts unmethylated cytosine to uracil

How target DNA is prepared is important!
Summary

- Sequencers of Illumina and PacBio are often used for NGS
- Illumina sequencers output numerous short reads
- PacBio sequencers output very long reads
- It is necessary to generate contigs by de novo assembly if an appropriate reference is unavailable
- Mapping is often performed in NGS data analysis
- RNA-Seq and genome sequencing are the simplest yet the most useful applications of NGS
- It matters how to prepare or enrich target DNA
References

• Illumina sequencing technology:

• PacBio sequencing technology:

• MutMap:

• QTL-Seq:
Questions

1. It may be difficult to get a whole-genome sequence of a plant without any reference using an Illumina sequencer. Why?
2. In what situation(s), is RAD-Seq better than whole genome sequencing?
3. In RNA-Seq using model species, genome sequences are more often used as a reference for mapping than mRNA sequences. Why?
4. What would you like to do with NGS?
5. Any suggestions and/or comments?