Overview of Assembly Techniques 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
  - Transcriptome assembly
  - Evaluations
  - Hands-on experiments
Primary NGS Applications

1. Alignment
2. **Assembly** (no reference, with a reference)
   - Genome
   - Transcriptome
3. RNA-Seq
4. Metagenomics
5. ChIP-Seq
6. RADSeq
**Why sequencing?**

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

- Ultimately completing the genome of non-model organisms, e.g. *Pacific whiteleg shrimp*
- Human genome project, $3.8$ Billion, 13 years to complete (1990-2003), 8-9x, coverage, 27 GBases

**How?**

Using sequencing methods, such as Sanger sequencing, next generation sequencing and single-molecule techniques
NGS Sequencing Workflow

DNA/RNA extraction

Library creation/amplification

Sequencing (Illumina HiSeq or PacBio Sequel)

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
SHORT READS

• Illumina
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 \textit{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.
Quality Score

**Illumina Quality Score**

- Phred-like algorithm: similar to scoring for Sanger sequencing
- Quality score of a given base, $Q$, is defined as:
- $e$: estimated probability of the base call being wrong

$$Q = -10 \log_{10}(e)$$

<table>
<thead>
<tr>
<th>Quality Score</th>
<th>Probability of Incorrect Base Call</th>
<th>Inferred Base Call Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (Q10)</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20 (Q20)</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30 (Q30)</td>
<td>1 in 1000</td>
<td>99.90%</td>
</tr>
</tbody>
</table>
FASTQ Format

Illumina 1.8+, Phred+33, raw reads typically (0, 41)

Read 1
@HWI-EASXXX:96:96:1:1:7939:13150 1:N:0:
TTCTCCCCCTTCTCGTTTCTTCCACCCGCG CCTATTCCTCGCTCTCTCCTTCCTT
+
BEHBHGDA(DA>C)CAEAHHHHGHHGHCDF@CDCE@EGGDDHH?HG@GDDGFGGGGE=
@HWI-EASXXX:96:96:1:1:14632:1706 1:N:0:
CACGAGACGAGAAGAGAAATGGGGAGGAGTCAGAGAGAGAGAGGGGAAGGGGAGGGAGAGGAGAGG
+
HHHHHFGD(GCGECCGHHBDGEGGGGGG>HFHDBG2D8C>C)C-@D?:A>ECECAA0A=:+B0A?+;AD<@DB>5=A@@

Read 2
@HWI-EASXXX:96:96:1:1:7939:13150 2:N:0:
CAAGGAACGAGAGAAGAGAAATGGGGAGGAGTCAGAGAGAGAGAGGGGAAGGGGAGGGAGAGGAGAGG
+
4111666664:8:8:8:8:8:8:8:8@4428447778+46665752288844444@
@HWI-EASXXX:96:96:1:1:14632:1706 2:N:0:
ACCTTCTCCTCCATCCTCCTCCTCCTCCCCTCCTCTCTCTCTCTCTCTCTCCT
+
-555598888@C@C@C@C@C@C@C@C4444444@40::6465689998@@@::4447677544::@@::@@@@@##
Choosing Illumina Sequencer!

- **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

LONG READS

• Pacific Biosciences (PacBio)
• Oxford Nanopore Technologies - MinION
Long reads - PacBio

- Single Molecule Real Time Sequencing (SMRT) Methodology
- Fluorescent dyes
- Zero Mode Waveguide (ZMW)
- DNA polymerase is immobilized at the bottom of a ZMW

http://www.nature.com/scientificamerican/journal/v294/n1/full/scientificamerican0106-46.html
http://science.sciencemag.org/content/323/5910/133.full
PacBio Sequel

~10 GB per SMRT Cell
1M ZMW/SMRT Cell
Up 16 SMRT/week
10 hour run time/SMRT
Avg. read 10-15kb

~10x jump over RSII
Long reads – Oxford NanoPore

- Oxford Nanopore Technologies
- Nanopore: a small hole (nanometer)
  - used to identify DNA sequence, passing through nanopore
- Single DNA molecule is sequenced

http://www.nature.com/scientificamerican/journal/v294/n1/full/scientificamerican0106-46.html
http://www.kurzweilai.net/single-molecule-electronic-dna-sequencing#prettyPhoto
TAMU holds patent
Dr. Higgin Bailey
NGS Read Specifications

## Comparing Sequencing Technologies

<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>
Why assembly?

Generating the consensus of transcriptome or genome of non-model species

Reconstructing the genome and transcriptome of non-model species are essential steps in expanding our understanding of the organism and developing therapeutic approaches to fight disease.
De novo Assembly

- Pool of reads
- No Reference genome!
- Creating consensus from the reads

Contig 1: ...CTAATAATAC TATATCTTAGGTTATATATTATCTATAAGTAGCACTTAAGTAAACTATT TT TT TT TT TT TAAGTATAGTT...
Contig 2: ...AAAGTAAAACTATCTATCTAGACCATAAAATTATTTTACTTACCTGACTGAGGAAAAAAGTCTATATTAATAACT...
Contig n: ...GATCTACCTATTTTAATCTATCTAGACCATAAAAAAAGTAAAAATTAGTAATTTCTTTAAGTAATATTAAGTATTCGTTG...
De novo Assembly - 2

• Connection reads by finding common sections of kmers
  • Kmers are made from reads!
• Resolving conflicts
• Complicated process!
• Highly computational resource demanding!

a. Generate all substrings of length K from the reads
b. Generate the De Brujin graph
c. Collapse the De Brujin graph
d. Traverse the graph
e. Assembled isoforms

De novo Assembly - 3

Reference Genome Generation

• Goal: generating the reference genome for a new species, using the genomic DNA data, generated by NGS
• Main tool: de novo assembly algorithm
• Output: annotated reference genome

Major steps
• Step 1: Assembly
  • ALLPATHS-LG (large genomes, recently DISCOVAR de novo): Broad institute
• Step 2: Annotation
  • PASA: Broad institute
## De novo Assembly - 4

<table>
<thead>
<tr>
<th>Sanger</th>
<th>Next Generation Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low coverage depth</td>
<td>High coverage depth</td>
</tr>
<tr>
<td>High cost for large genomes</td>
<td>Relatively low cost, even for large genomes</td>
</tr>
<tr>
<td>Slow</td>
<td>Fast</td>
</tr>
<tr>
<td>Handles repeats well</td>
<td>Need long reads for repeated regions (e.g. PacBio, Illumina Mate-Pair)</td>
</tr>
</tbody>
</table>

**Diagram:**
- **Region 1**
  - Short reads
  - Long reads
- **Region 1, repeated**
  - Short reads
  - Long reads
Genome Assembly Tools:

- ALLPATHS
- ALLPATHS-LG (Special recipe: fragments + jumping libraries)
- DISCOVAR *de novo*
- ABySS
- EULER-SR
- SOAPDenovo
- VCAKE
- Velvet
- **Canu**
- CLC Bio Genomics Workbench
Transcriptome Assembly Tools:

• SOAPdenovo-Trans
• Trans-ABySS
• Velvet + Oases
• **Trinity**
• Rnnotator
• CLC Bio Genomics Workbench
High Quality Assembly

- Hybrid Approach
- High Coverage
- Merging
  - Metassembler
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](https://hprc.tamu.edu/wiki/HPRC:MobaXterm)
    - You are able to view images and use GUI applications with MobaXterm
  - or **Putty**
    - You can not view images or use GUI applications with PuTTY
Your Login Password

- 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
Contact the HPRC Helpdesk

Website: hprc.tamu.edu
Email: help@hprc.tamu.edu
Telephone: (979) 845-0219

Help us, help you -- we need more info
• Which Cluster
• UserID/NetID
• Job id(s) if any
• Location of your jobfile, input/output files
• Application used if any
• Module(s) loaded if any
• Error messages
• Steps you have taken, so we can reproduce the problem
Using SSH - MobaXterm (on Windows)

- message of the day
- your quotas
Using SSH to Access Ada

```
ssh user_NetID@ada.tamu.edu
```

https://hprc.tamu.edu/wiki/Ada:Access

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:
Any question?
nghaffari@tamu.edu