Assembly of Next Generation Sequence Data

Catherine Eason (Wofford College)
Amit Upadhyay (University of Tennessee)
Bhanu Rekepalli (JICS)
Outline

- DNA overview
- Background leading to problem
- Current Status in Assembly
- Methodology
- Results
- Conclusion/Future Work
DNA

- Nucleotides are the building blocks of DNA.
- Characteristics vary immensely between organisms.

What is a nucleotide?

- Nucleotide vs Base Pair

- Nucleic acid

- Adenine
- Thymine
- Guanine
- Cytosine

- Deoxyribose sugar
- Phosphate

© scienceaid.co.uk
NGS data

Cost per Genome

Moore's Law

NIH National Human Genome Research Institute

genome.gov/sequencingcosts
Paired-End Sequences

- Sequencing from both ends concurrently (by Illumina)
- Allows for detection of small frame shift mutations
- Paired information must be kept together, and in correct order

Diagram showing process of collecting paired-end reads. The genomic DNA is sequences into fragments which adaptors and primers are attached to (Green, Blue, and Purple ends). A cluster is formed and the sequences is read starting from both adaptors, producing the paired-end read.
Analysis Workflow

• **Data is collected, now what?**
  - Assembly
  - Analysis
  - Future Studies

Diagram for the complete data analysis process. Orange rectangles are the actual analysis steps while the gray rectangles represent input from outside sources.
Diagram for the complete assembly process, beginning with raw sequence data. The assembled sequences must be checked for accuracy—a difficult step. Green rectangles are the steps, gray circles a short description. And blue arrows are steps that have their own process.
Trimmomatic vs. BBTools

- Quality Control
- Assembly
- Assembly Quality Control

**Trimmomatic**
- One program designed for paired-end data that removes low-quality reads and the adaptor sequences

**BBtools**
- Set of multiple tools that provide a variety of options, including reduction of coverage and normalization
Current Assembly Algorithms

- Quality Control
- Assembly

Greedy

Overlap layout consensus (OLC)

De Bruijn

Diagram by Leland Taylor at Davidson College
SPAdes

MaSuRCA

Velvet

SOAPdenovo

ABySS
**SPAdes**

- New form of de Bruijn graph—*Multisized de Bruijn*
  - Implements new “error correction” methods
  - Allows user to backtrack over graph construction process
- Can detect “best” k-mer size (if desired)

**SOAPdenovo**

- *De novo* assembly of large, mammalian genomes
- Uses de Bruijn graph algorithm
  - Edges must be linked to existing sequence
QUAST and Statistics

- Quality Control
- Assembly
- Assembly Quality Control

N50 value

Guanine/Cytosine content

Number of Contigs

Genome Size

Quality-check programs

Some compared to published data, N50—the bigger the better, high GC % = more stable
Vibrio gazogenes

- 36 chromosomes
- Genome size?

Picture of *V. cholera* bacteria. Closely related to *V. gazogenes*
## Results

- Using Trimmomatic (quality of read)

<table>
<thead>
<tr>
<th>Kmer Size</th>
<th># of Contigs</th>
<th>Genome Size</th>
<th>N50</th>
<th>GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>514</td>
<td>4,430,394</td>
<td>17,374</td>
<td>45.27</td>
</tr>
<tr>
<td>33</td>
<td>282</td>
<td>4,467,765</td>
<td>54,782</td>
<td>45.27</td>
</tr>
<tr>
<td>55</td>
<td>215</td>
<td>4,496,327</td>
<td>68,126</td>
<td>45.27</td>
</tr>
<tr>
<td>71</td>
<td>120</td>
<td>4,555,395</td>
<td>246,573</td>
<td>45.32</td>
</tr>
<tr>
<td>Subset 51</td>
<td>201</td>
<td>4,468,133</td>
<td>61,386</td>
<td>45.30</td>
</tr>
<tr>
<td>Subset 61</td>
<td>193</td>
<td>4,485,523</td>
<td>68,483</td>
<td>45.31</td>
</tr>
<tr>
<td>Subset 71</td>
<td>180</td>
<td>4,499,332</td>
<td>79,631</td>
<td>45.32</td>
</tr>
<tr>
<td>Subset 81</td>
<td>173</td>
<td>4,510,565</td>
<td>88,093</td>
<td>45.33</td>
</tr>
<tr>
<td>Subset 91</td>
<td>88</td>
<td>4,545,153</td>
<td>262,031</td>
<td>45.36</td>
</tr>
</tbody>
</table>

Table for the assembly of Trimmomatic trimmed data through SPAdes showing number of contigs, genome size, N50, and GC content statistics for k-mer sizes 21, 33, 55, 71 and a random 50% subset of data's statistics for k-mer sizes 51, 61, 71, 81, and 91.

<table>
<thead>
<tr>
<th>Kmer Size</th>
<th># of Contigs</th>
<th>Genome Size</th>
<th>N50</th>
<th>GC %</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>16</td>
<td>11,398</td>
<td>690</td>
<td>42.96</td>
</tr>
<tr>
<td>33</td>
<td>17</td>
<td>11,766</td>
<td>690</td>
<td>41.00</td>
</tr>
<tr>
<td>55</td>
<td>1,385</td>
<td>968,669</td>
<td>685</td>
<td>46.87</td>
</tr>
<tr>
<td>71</td>
<td>444</td>
<td>4,448,857</td>
<td>18,563</td>
<td>45.33</td>
</tr>
<tr>
<td>Subset 51</td>
<td>1,481</td>
<td>4,321,140</td>
<td>4,296</td>
<td>45.39</td>
</tr>
<tr>
<td>Subset 61</td>
<td>309</td>
<td>4,459,372</td>
<td>29,329</td>
<td>45.30</td>
</tr>
<tr>
<td>Subset 71</td>
<td>206</td>
<td>4,481,934</td>
<td>55,249</td>
<td>45.30</td>
</tr>
<tr>
<td>Subset 81</td>
<td>172</td>
<td>4,499,317</td>
<td>75,768</td>
<td>45.32</td>
</tr>
<tr>
<td>Subset 91</td>
<td>159</td>
<td>4,519,076</td>
<td>100,098</td>
<td>45.34</td>
</tr>
</tbody>
</table>

Table for the assembly of Trimmomatic trimmed data using SOAPdenovo2. showing number of contigs, genome size, N50, and GC content statistics for k-mer sizes 21, 33, 55, 71 and a random 50% subset of data's statistics for k-mer sizes 51, 61, 71, 81, and 91.
Results

- Using BBtools (bbnorm and bbtrim)

<table>
<thead>
<tr>
<th>SPAdes (BBtools)</th>
<th>SOAPdenovo2 (BBtools)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Kmer Size</strong></td>
<td><strong># of Contigs</strong></td>
</tr>
<tr>
<td>21</td>
<td>506</td>
</tr>
<tr>
<td>33</td>
<td>263</td>
</tr>
<tr>
<td>55</td>
<td>190</td>
</tr>
<tr>
<td>71</td>
<td>106</td>
</tr>
</tbody>
</table>

Table for the assembly of Bbtool trimmed data through SPAdes showing number of contigs, genome size, N50, and GC content statistics for k-mer sizes 21, 33, 55, and 71.

Table for the assembly of Bbtool trimmed data through SOAPdenovo2. show number of contigs, genome size, N50, and GC content statistics for k-mer sizes 21, 33, 55, and 71.
Conclusions

- Trimmomatic: no negative effect on assembly process
- Genome size ~4.5 million bp

Future Goals

- Collective scripts for all four aspects of NGS pipeline project
  - Genome assembly
  - Genome annotation
  - RNA-seq
  - Variant calling
- Collective script for all steps of assembly
- Web Interface for ease of access
References


Acknowledgments

I would like to thank the CSURE program funded by the NSF and partnered with JICS, ORNL, and the University of Tennessee for the opportunity to work on this project.