Aligning Paired End Sequencing Data from the Ion PGM™ Sequencer Using NextGENe® Software

December 2011
Megan Manion, Kevin LeVan, John McGuigan, Shouyong Ni, CS Jonathan Liu

Introduction
Paired end sequencing can be useful for a variety of applications to provide improved alignment accuracy, greater coverage depth and to allow the detection of structural variants. The Ion PGM system provides a paired end sequencing option in which both ends of a template are sequenced. Library inserts of varying size can be used, with the library insert size determining the extent to which paired reads will overlap.

NextGENe Software can be used to quickly align paired end sequencing data from NGS systems such as the Ion PGM Sequencer. When paired end data is provided, NextGENe uses an alignment algorithm which tracks the distance between paired reads. This information can then be reviewed with several specialized reports. Standard reports, such as a mutation report and expression report are also available, along with high level visualizations in the NextGENe Viewer.

![Gap Distribution](image1.png)

**Figure 1:** The Gap Distribution report provides a visualization of the distance between paired reads. A normal distribution, centered over the insert size, is expected.

Procedure
1. Format Conversion
   NextGENe’s Format Conversion Tool is used to convert paired end data in fastq format to fasta format

2. Load Data and Specify Alignment settings
   The Project Wizard will guide you step-by-step through project set-up
Figure 2: To begin setting up your analysis project select Ion PGM under Instrument Type, SNP/Indel Discovery under Application Type, and Sequence Alignment under Steps. Sequence Condensation can be left deselected.

Figure 3: On the Alignment settings page, select the Load Paired Reads option and input a range for the library size

3. Visualize Results and Export Reports
   The NextGENe Viewer is used to display analysis results and generate reports.
Results
Two Paired End Sequencing datasets from the Ion PGM were provided by Life Technologies. Both datasets were from E. coli DH10B. The raw fastq files were converted to fasta format, with quality filtering applied to remove data with low instrument quality scores. The quality filtered fasta files were aligned to a GenBank (.gbk) reference file.

<table>
<thead>
<tr>
<th>Chip</th>
<th>314</th>
<th>316</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reads</td>
<td>485,293</td>
<td>2,294,560</td>
</tr>
<tr>
<td>Aligned Reads (% Aligned)</td>
<td>472,826 (97.4%)</td>
<td>2,224,204 (96.9%)</td>
</tr>
<tr>
<td>Average Coverage</td>
<td>13</td>
<td>49</td>
</tr>
<tr>
<td>Processing Time</td>
<td>2 min</td>
<td>6 min</td>
</tr>
</tbody>
</table>

Table 1: Alignment Results

Each sample was also processed without selecting the Load Paired Reads option, treating the data as single end fragment reads, allowing a comparison of paired end and single end data. In both cases the setting to “Remove Ambiguously Mapped Reads” was selected. Since paired end information can be used to resolve ambiguities, more reads can be aligned.

<table>
<thead>
<tr>
<th></th>
<th>314</th>
<th>316</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligned Reads</td>
<td>Paired End</td>
<td>472,826</td>
</tr>
<tr>
<td></td>
<td>Single End</td>
<td>465,934</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,190,837</td>
</tr>
</tbody>
</table>

Table 2: Paired End vs. Single End Alignment

NextGENe provides several reports specifically for paired end data. Each of these reports can be accessed from the Paired View menu in the NextGENe Viewer.
The Paired End Statistic Report shows general statistics about the paired alignment. Shown here is the report for the 316 sample.

The Paired Read Reports list read pairs that mapped at a distance from each other that is outside of the expected range. These reads can be useful for detecting possible structural variations. Two Paired Reports are available: The Opposite Direction Paired Reads Report, which includes only read pairs that map in opposite orientations (the expected case), and the Same Direction Paired Reads Report, which includes read pairs that map in the same orientation. Read pairs in the Same Direction Paired Reads Report could be indicative of an inversion. A Single Read Report is also available which lists reads that were mapped to the reference where their pair was not mapped. Each of these reports can be easily exported in tab-delimited text format.

**Figure 5:** The Paired End Statistic Report shows general statistics about the paired alignment. Shown here is the report for the 316 sample.

**Figure 6:** The Opposite Direction Paired Reads Report for the 316 dataset

**Discussion**

Paired End Sequencing can be a useful method for many sequencing projects. NextGENe Software provides a quick and accurate solution for taking advantage of paired end data from the Ion PGM sequencing system. Paired read information is utilized and tracked during alignment to the reference sequence, and it is reported in the NextGENe Viewer.

**Acknowledgement**

We’d like to thank Life Technologies for providing the Ion PGM Paired End datasets.