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.

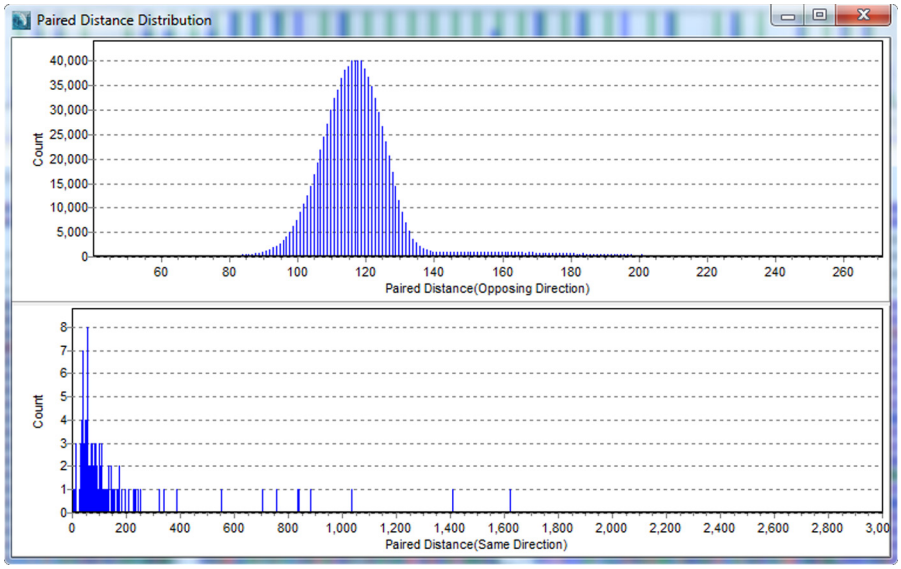


Figure 1: The Gap Distribution report provides a visualization of the distance between paired reads. A normal distribution, centered over the insert size, is expected.

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

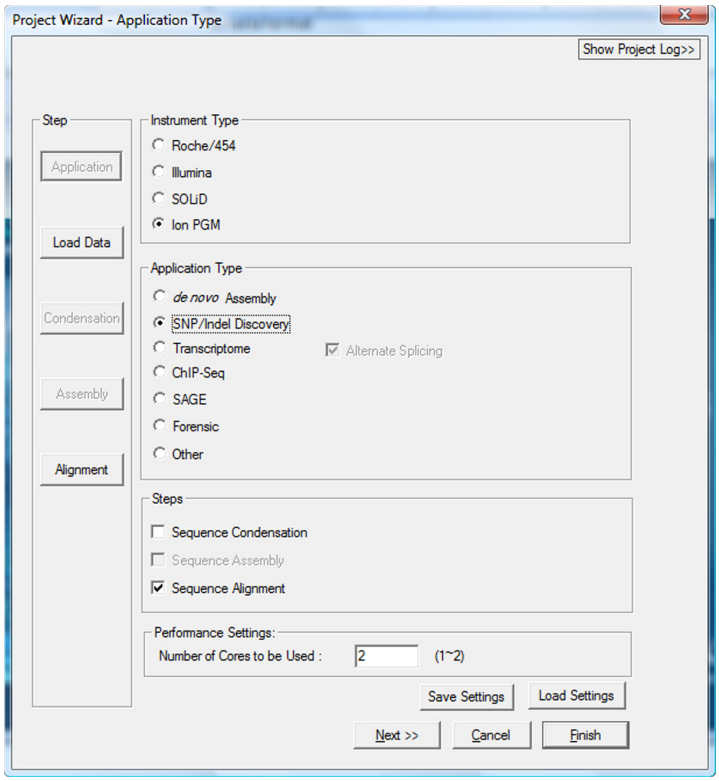


Figure 2: To begin setting up your analysis project select Ion PGM under Instrument Type, SNP/Indel Discovery under Application Type, and Sequence Alignment under Steps. Sequence Condensation can be left deselected.

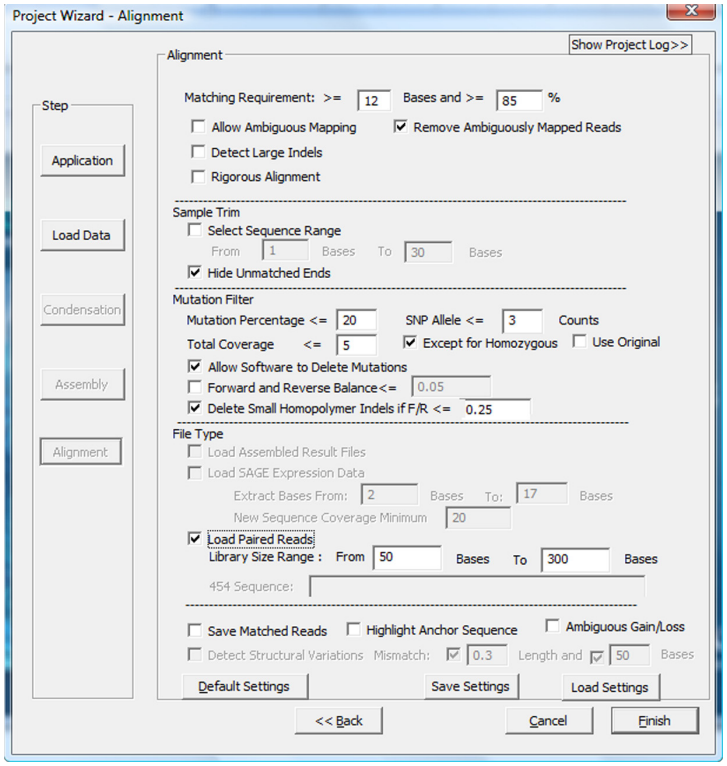


Figure 3: On the Alignment settings page, select the Load Paired Reads option and input a range for the library size

3. Visualize Results and Export Reports

The NextGENe Viewer is used to display analysis results and generate reports.

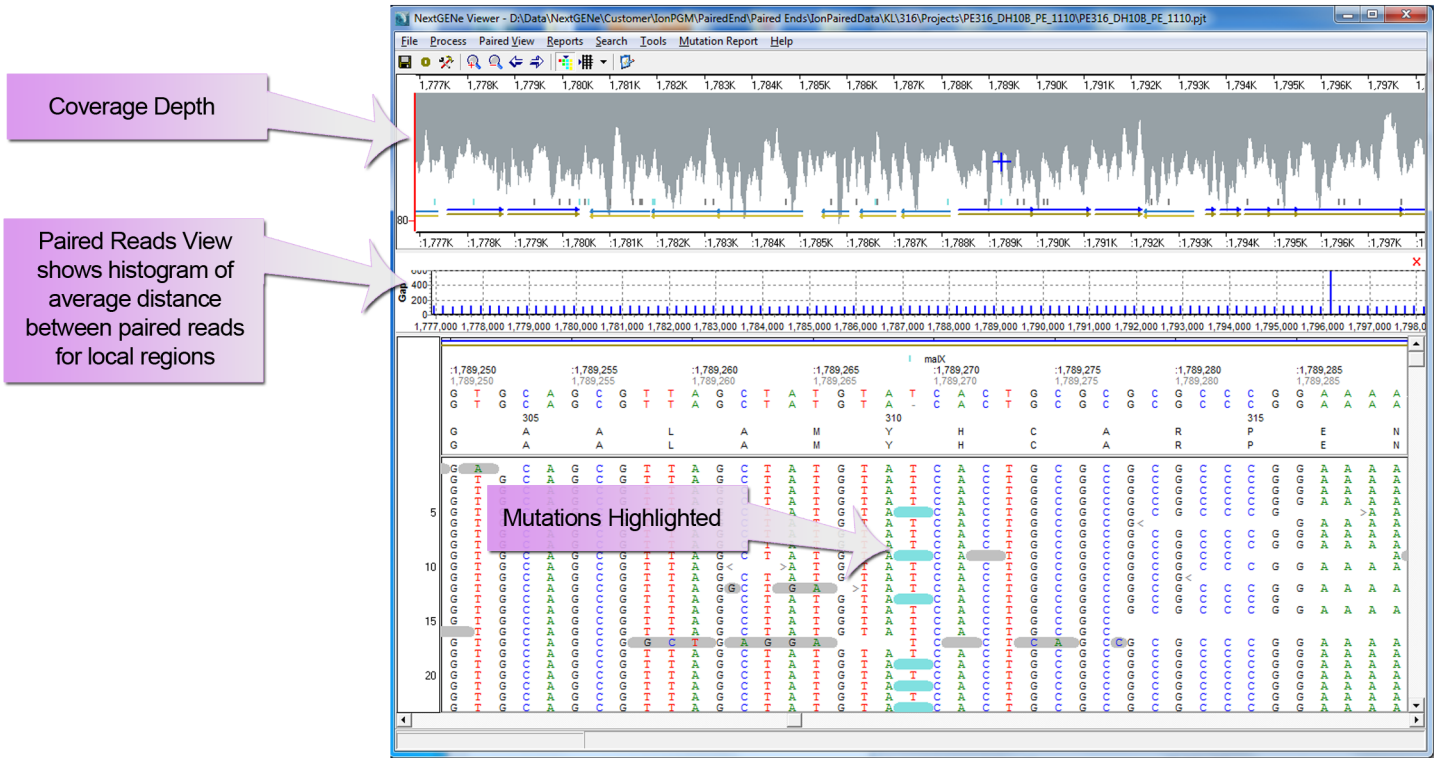


Figure 3: Paired End Alignment Results Displayed in the NextGENe Viewer

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.

Chip	314	316
Total Reads	485,293	2,294,560
Aligned Reads (% Aligned)	472,826 (97.4%)	2,224,204 (96.9%)
Average Coverage	13	49
Processing Time	2 min	6 min

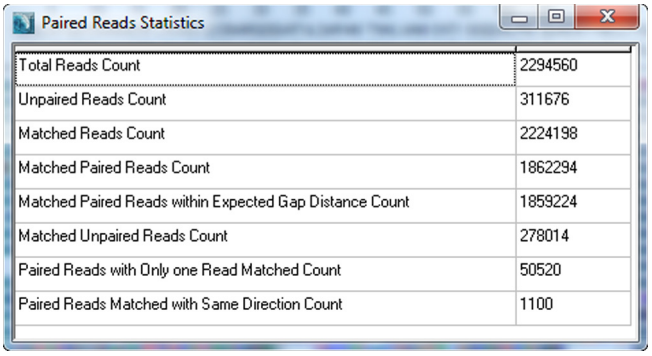
Table 1: Alignment Results

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.

		314	316
Aligned Reads	Paired End	472,826	2,224,204
	Single End	465,934	2,190,837

Table 2: Paired End vs. Single End Alignment

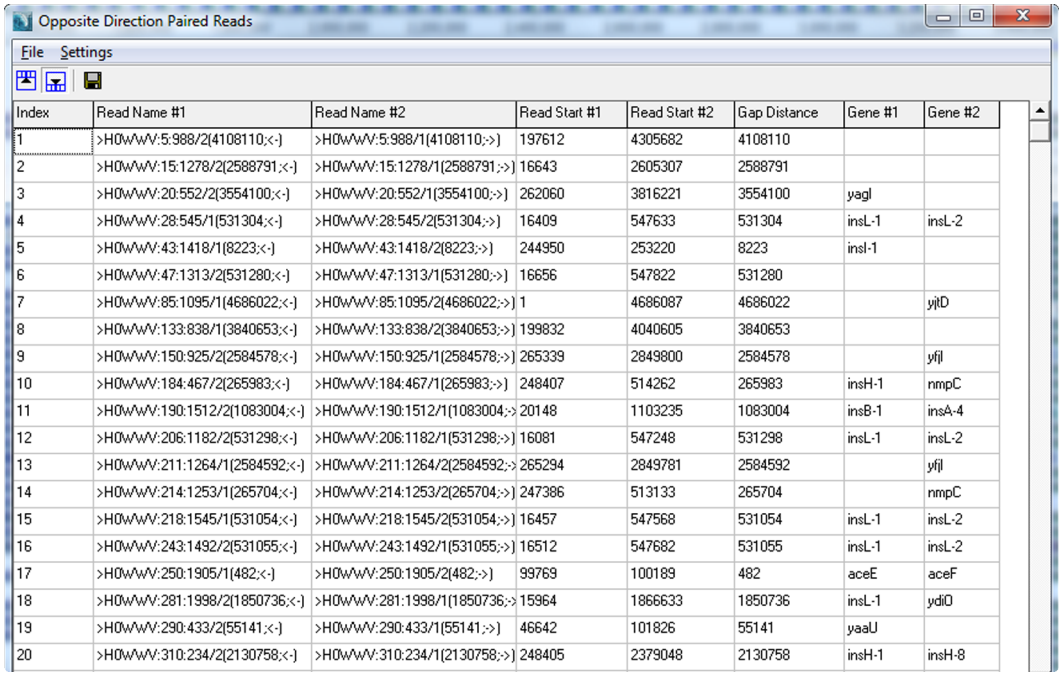
NextGENe provides several reports specifically for paired end data. Each of these reports can be accessed from the Paired View menu in the NextGENe Viewer.



Total Reads Count	2294560
Unpaired Reads Count	311676
Matched Reads Count	2224198
Matched Paired Reads Count	1862294
Matched Paired Reads within Expected Gap Distance Count	1859224
Matched Unpaired Reads Count	278014
Paired Reads with Only one Read Matched Count	50520
Paired Reads Matched with Same Direction Count	1100

Figure 5: 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.



Index	Read Name #1	Read Name #2	Read Start #1	Read Start #2	Gap Distance	Gene #1	Gene #2
1	>H0WWV:5:988/2(4108110;<-)	>H0WWV:5:988/1(4108110;->)	197612	4305682	4108110		
2	>H0WWV:15:1278/2(2588791;<-)	>H0WWV:15:1278/1(2588791;->)	16643	2605307	2588791		
3	>H0WWV:20:552/2(3554100;<-)	>H0WWV:20:552/1(3554100;->)	262060	3816221	3554100	yagI	
4	>H0WWV:28:545/1(531304;<-)	>H0WWV:28:545/2(531304;->)	16409	547633	531304	insL-1	insL-2
5	>H0WWV:43:1418/1(8223;<-)	>H0WWV:43:1418/2(8223;->)	244950	253220	8223	insI-1	
6	>H0WWV:47:1313/2(531280;<-)	>H0WWV:47:1313/1(531280;->)	16656	547822	531280		
7	>H0WWV:85:1095/1(4686022;<-)	>H0WWV:85:1095/2(4686022;->)	1	4686087	4686022		yjtD
8	>H0WWV:133:838/1(3840653;<-)	>H0WWV:133:838/2(3840653;->)	199832	4040605	3840653		
9	>H0WWV:150:925/2(2584578;<-)	>H0WWV:150:925/1(2584578;->)	265339	2849800	2584578		yfjI
10	>H0WWV:184:467/2(265983;<-)	>H0WWV:184:467/1(265983;->)	248407	514262	265983	insH-1	nmpC
11	>H0WWV:190:1512/2(1083004;<-)	>H0WWV:190:1512/1(1083004;->	20148	1103235	1083004	insB-1	insA-4
12	>H0WWV:206:1182/2(531298;<-)	>H0WWV:206:1182/1(531298;->)	16081	547248	531298	insL-1	insL-2
13	>H0WWV:211:1264/1(2584592;<-)	>H0WWV:211:1264/2(2584592;->	265294	2849781	2584592		yfjI
14	>H0WWV:214:1253/1(265704;<-)	>H0WWV:214:1253/2(265704;->)	247386	513133	265704		nmpC
15	>H0WWV:218:1545/1(531054;<-)	>H0WWV:218:1545/2(531054;->)	16457	547568	531054	insL-1	insL-2
16	>H0WWV:243:1492/2(531055;<-)	>H0WWV:243:1492/1(531055;->)	16512	547682	531055	insL-1	insL-2
17	>H0WWV:250:1905/1(482;<-)	>H0WWV:250:1905/2(482;->)	99769	100189	482	aceE	aceF
18	>H0WWV:281:1998/2(1850736;<-)	>H0WWV:281:1998/1(1850736;->	15964	1866633	1850736	insL-1	ydiO
19	>H0WWV:290:433/2(55141;<-)	>H0WWV:290:433/1(55141;->)	46642	101826	55141	yaaU	
20	>H0WWV:310:234/2(2130758;<-)	>H0WWV:310:234/1(2130758;->)	248405	2379048	2130758	insH-1	insH-8

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.