Processing Paired End Data from the Ion PGM™ Sequencer Using the Floton Paired-End Merger in NextGENe® Software

January 2012

John McGuigan, Jacie Wu, Shouyong Ni, C.S. Jonathan Liu

Introduction

The Ion PGM system now provides a paired end sequencing option in which DNA fragments can be sequenced from both directions. Paired end sequences can be merged into a single, high-quality sequence using NextGENe's "Overlap Merger" tool. This process greatly improves accuracy, especially at the 3' end, reducing the need for trimming. This approach makes it possible to improve the accuracy of the Ion Torrent platform to more than 99.8% when the "hide unmatched ends" option is used. This is especially useful for AmpliSeq™ projects. This app note demonstrates the processing of two bacterial re-sequencing projects and an AmpliSeq project.the Ion Torrent platform to more than 99.8%.

Bacterial Dataset Summary

	314 –Forward Paired and Singleton Reads	314 – After Merging	316 – Forward Paired and Singleton Reads	316 – After Merging
Average Read Length	98	118	105	126
Total Aligned Bases	35,441,511	36,186,669	417,057,466	441,903,352
Error Rate (%) for Aligned reads (including ambiguous alignment)	0.57254%	0.22043%	0.52216%	0.20656%
Equivalent Phred Score	22.42	26.57	22.82	26.85

Table 1: Read Length and Aligned Read Accuracy Improvements for two bacterial re-sequencing projects. The adapter sequence (GCTGAGGA) was trimmed for the non-merged data.

AmpliSeq Dataset Summary

	Forward Paired and Singleton Reads	Merged Reads
Number of Bases	35,062,802	25,735,553
Number of Aligned Bases (%)	33,598,317 (95.82%)	25,638,943 (99.62%)
Number of Mutations Called (> 3% frequency, > 50 counts)	119	44
Single BP Indels (Likely False Positives)	103	31

Table 2: Alignment and mutation calling comparison of an AmpliSeq dataset. For the purposes of comparison, mutations called in the amplicon regions were only filtered with an allele count and mutant percentage filter.

Procedure

1. Overlap Merger

NextGENe's Overlap Merger Tool is used to combine associated reads into a single high-quality fragment. It outputs the resulting reads in FASTQ and FASTA format. The FASTA formatted file can be used for alignment and mutation calling. NextGENe v2.20 added an option that uses the Floton method to accurately merge Ion Torrent data.

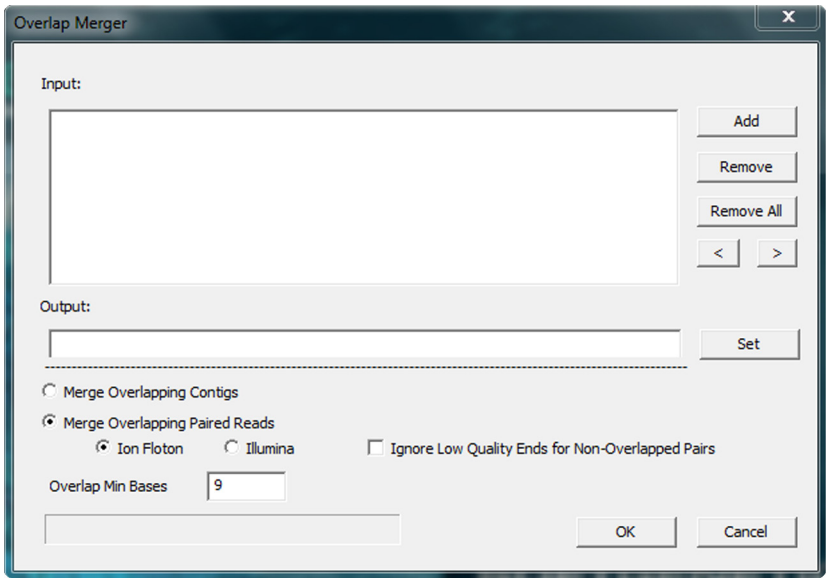

Figure 1: The Overlap Merger Tool

2. Load Data and Specify Alignment settings

The Project Wizard will guide you step-by-step through project set-up

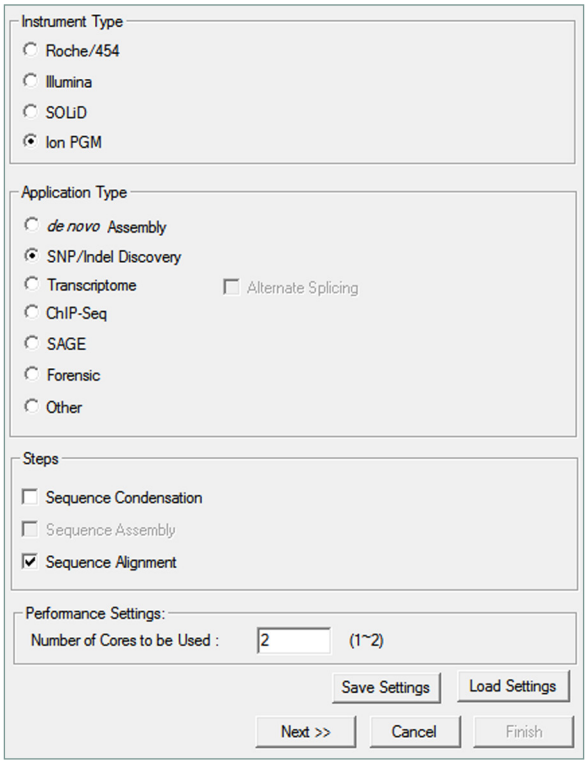

Figure 2: To begin setting up your analysis project select Ion PGM under Instrument Type, SNP/Indel Discovery under Application Type, and Sequence Alignment under Steps. Sequence Condensation can be left deselected.

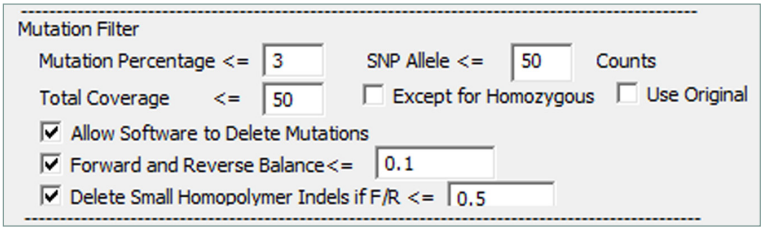

Figure 3: Mutation filtering settings for a typical AmpliSeq project. The mutation filter settings can be adjusted as needed to be more sensitive or more specific. In this case 50 mutant alleles are required, meaning positions with 500x coverage can detect mutations down to 10%, and 2000x coverage can detect mutations down to 3% (the mutation percentage cutoff). The number of reads in the forward direction must be at least 1/10th the number of reads in the reverse direction (and vice versa) for all mutant alleles, and small homopolymer indels have a higher balance requirement.

3. Visualize Results and Export Reports

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

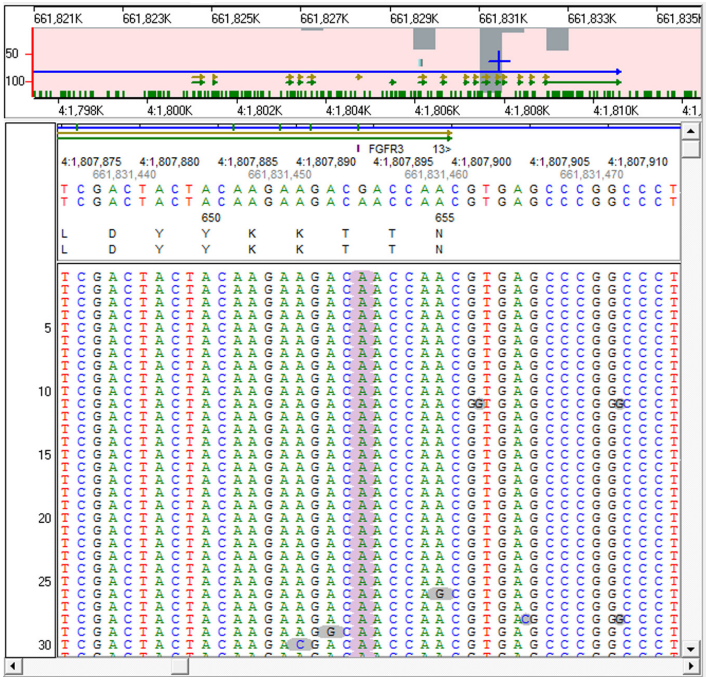

Figure 4: Alignment results for the merged reads AmpliSeq dataset displayed in the NextGENe Viewer. The highlighted position contains a known mutation in the NA12878 Hapmap sample.

Results

The Ion PGM Bacterial Datasets

Two Paired End Sequencing datasets from the Ion PGM Sequencer were provided by Life Technologies. Both datasets were from E. coli DH10B. The fastq files were run in NextGENe's overlap merger tool and the resulting FASTA files were aligned to a GenBank (.gbk) reference file.

	314	316
Total Reads	621,147	7,067,580
Pairs (% of reads paired)	310,574 (90.22%)	3,533,794 (91.96%)
Merged Pairs (%)	309,283 (99.58%)	3,506,036 (99.21%)
Aligned Reads (% of Merged)	305,667 (98.83%)	3,493,594 (99.65%)
Average Coverage	8	107
Overlap Merger Processing	< 1 min	9 min, 30 seconds
Alignment and Variant Calling	< 1 min	8 min

Table 3: Merging and Alignment Results for the Bacterial Datasets

The second sequencing run and overlap merger improved the raw accuracy relative to the original sequencing run (table 1). The merge step took less than 1 minute for the 314 dataset and less than 10 minutes for the 316 dataset (table 3). In each sample over 90% of reads were paired and over 99% of paired reads were merged. Trimming based on quality scores is no longer necessary, resulting in longer read lengths (figure 5). Figure 6 shows an example of a region with corrected errors.

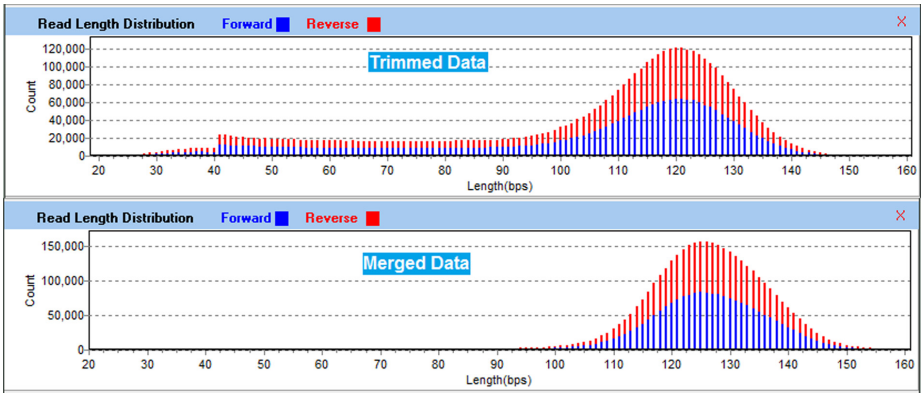

Figure 5: Read length comparison between quality-trimmed data and merged data (316 bacterial re-sequencing)

Figure 6: Several false positive calls have been corrected by performing paired-end sequencing. The left side shows the results of the original data while the right side shows the results after merging overlapping paired-end reads. One read on the left has untrimmed adapter sequence- the merged reads do not require adapter trimming, so failed trimming is not a problem.

AmpliSeq Dataset

One AmpliSeq dataset was also provided. 289,108 pairs of reads were found out of 698,817 total reads. Over 99% of these reads were merged. Low quality bases and adapter sequence (CTGAGTCGGAGACAC and GCTGAGGA) were trimmed from the un-merged data. The alignment and mutation calling results are summarized in table 2. The un-merged data had almost 3 times as many mutation calls, and most of the difference could be attributed to single base indels. Additional filters available in NextGENe could further reduce the number of false positive calls. Figure 7 shows an example of a false positive deletion that was corrected with merging.

Figure 7: A false positive deletion that was not called after using paired-end merging. The percentage of reads with the deletion decreased from 39.84% to 0.27%.

Discussion

Paired End Sequencing data from the Ion PGM allows improved accuracy by sequencing the same fragment from each direction. This provides several benefits:

- Adapter sequences are removed
- Both ends of the merged sequences have high quality basecalls
- Reads are longer on average because quality trimming is not needed
- A larger fraction of reads are able to align
- Less filtering is needed to remove false positives

NextGENe Software provides a quick and accurate method for combining these overlapping reads into single high-quality sequences using the unique Floton method.

All projects in this analysis were run on an 8-core laptop computer with 8 GB of RAM.

Acknowledgements

We would like to thank Life Technologies for supplying the data used in this analysis.