Assembly of Bacterial Genomes from Ion PGM™ Data using NextGENe® software's Floton™ Assembler with new normalization technology (v 2.3.1)

October 2012

Jacie Wu, Ning Wan, John McGuigan, Shouyong Ni, CS Jonathan Liu

Introduction

The Ion PGM system is fundamentally different from most other sequencing systems. It is a "post-light" technology because it doesn't depend on detection of light emission from nucleotide incorporation. Instead, it uses a silicon chip containing millions of individual pH meters. Its flow-based approach detects pH changes caused by release of hydrogen ion during incorporation of unmodified nucleotides in DNA replication. Because of this different approach to sequencing, the instrument and reagents are much less expensive and it has a unique error profile- most errors are indels rather than substitutions, especially in homopolymer regions.

These errors are more problematic for assembly than substitution errors because of the increased complexity of gapped comparison. However, SoftGenetics has developed the Floton™ assembler for NextGENe® Software which is able to treat these homopolymer errors as substitution errors. In doing so, it is possible to correct the errors during assembly. This method condenses the sequence into flow calls of individual bases and the number of bases in each flow (Figure 1). By converting the sequence data into this format, the indels are essentially converted into substitution errors (different base count numbers), allowing for faster computation time and correction of most homopolymer errors.

Figure 1: Conversion of base calls into flow calls.

NextGENe version 2.3.0 added a new "normalization" option- this can improve assembly speed 3-fold with equal or greater results. This algorithm works by generating "flow-mers" in the reads and calculating the frequency in order to estimate coverage of that sequence. This is similar to existing methods for discarding reads prior to assembly when the coverage is higher than necessary [arXiv:1203.4802v2]. Unlike other methods, it is designed to remove low-quality or short reads and retain long, high-quality reads that are needed for assembly. Additionally, it is specialized for Ion Torrent data because of the unique flow-mer indexing. NextGENe version 2.3.1 simplifies the analysis settings and improves poor quality read filtering, making the Floton assembler even easier to use while generating improved results.

After the optional normalization step, the remaining reads are each indexed with several flow-mers.
This information is used during the first two steps of the assembly:

1. Condensation - reads that share flow-mer indexes are compared and used to generate high-quality consensus contigs. The same read may be used in several condensation contigs.
2. Combination - an iterative process checks for condensation contigs that contain the same reads in order to discover and merge overlaps.
3. A final overlap merger step carefully combines the combination contigs into the final assembly contigs.

Procedure

1. The format conversion tool is used to convert the original FASTQ or SFF files to FASTA format while performing some quality filtering and trimming.
2. The instrument (Ion PGM) and application (de novo assembly) are selected on the first page of the project wizard.
3. The assembly settings are selected (figure 2). "Small Genome" is the default preset. Default settings were used for all assemblies in this analysis- the normalized assemblies used a 40x target level of coverage.
   - Coverage normalization can be applied by selecting this option and specifying a level of coverage. 40x is recommended for projects that have much higher coverage (>80x).
   - The results of the first two assembly stages (condensation and combination) can be saved separately. By default NextGENe only saves the final results (after overlap merger).
   - A simple cut-off option is available to filter out short contigs.
   - The index settings and low frequency removal options can be customized, or the "automatic" option can be used to allow NextGENe to use parameters estimated to be optimal.
   - Error toleration can be adjusted based on the quality of the dataset- increasing it for datasets with more errors can increase the size of assembled contigs.
4. Assembly is started, or more projects are set up for consecutive processing.

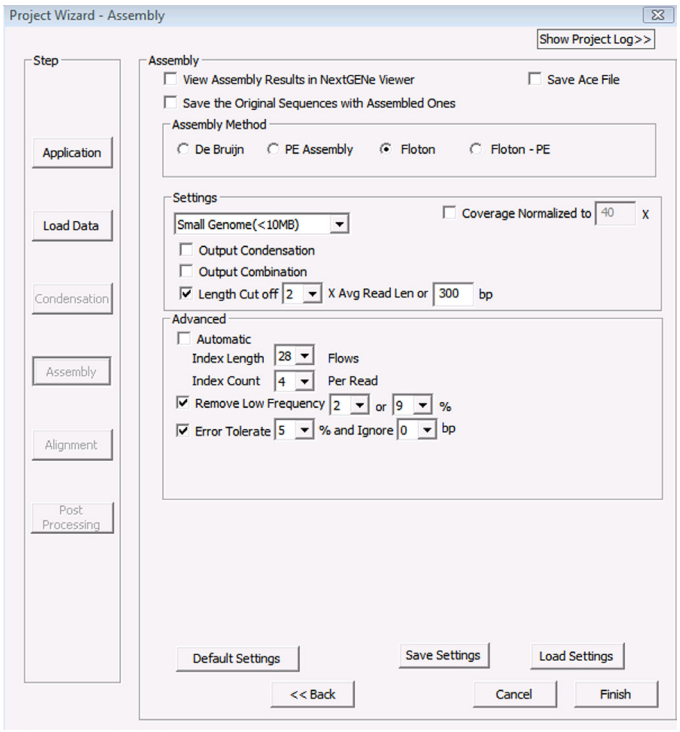

Figure 2: Default Floton Assembly Parameters (version 2.3.1)

Results

Three recent E. coli datasets from the Ion Community website were assembled. Sample information (post-format conversion) is given in table 1. Both the average and the N50 read lengths are included because the length of trimmed reads is a skewed distribution. Assembly results are summarized in table 2.

	B7-143	BEA-629	C12-245
Genome	E. coli EHEC	E. coli DH10B	E. coli DH10B
Genome Size	5.52 Mbp	4.69 Mbp	4.69 Mbp
Number of Reads	4,745,156	2,051,053	4,076,537
Read Length (Average, N50)	225 bp, 286 bp	290 bp, 373 bp	200 bp, 231 bp
Depth of Coverage	201x	130x	179x

Table 1: Pre-Assembly Read Statistics

	B7-143		BEA-629		C12-245	
	Original	40x Norm	Original	40x Norm	Original	40x Norm
Number of Contigs	*153*	107	*142*	71	*132*	96
Contig Max	*294,833*	404,715	*245,240*	249,077	*326,408*	326,348
Contig N50	*114,586*	127,112	*100,852*	107,785	*100,347*	103,441
Contig N90	*25,390*	32,988	*29,111*	38,728	*35,244*	33,984
Running Time (16-cores)	***76 min***	**21 min**	***38 min***	**13 min**	***39 min***	**15 min**

Table 2: Assembly Results and Settings Adjustments

Figure 3 shows coverage distributions before and after normalization for all three datasets. Most of the data was removed, but there was still sufficient coverage to perform assembly.

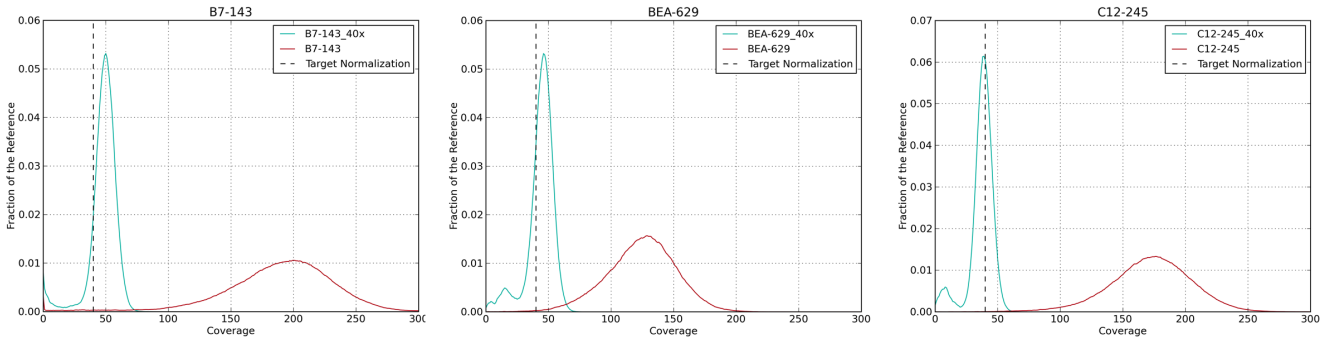

Figure 3: Coverage distribution before and after normalization for each project

Discussion

The NextGENe Floton Assembler is specially designed to handle data from flow-based sequencing technologies that tend to have indel errors rather than substitution errors. When paired with the Ion PGM system it provides a very fast and inexpensive method for de novo sequencing of bacterial and other small genomes. It is designed to run on relatively inexpensive hardware and to be easy to use.

Acknowledgements

We would like to thank Life Technologies for making these datasets available on the Ion Community website.