de Novo Small Genome Assembly using the Floton™ Assembler in NextGENe® Software

McGuigan J.R., Manion M.R., Liu C.S.
SoftGenetics LLC, 100 Oakwood Ave Suite 350, State College, PA 16803, USA

Abstract

Ion Torrent sequencing technology has an error profile consisting mainly of indels in homopolymer regions. These errors are more problematic for assembly than substitution errors because of the increased complexity of gapped comparison. The Floton assembler for NextGENe is able to treat these homopolymer errors as substitution errors in order 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. This allows for faster computation time and correction of most homopolymer errors.

The Floton Assembler in NextGENe v2.3.0 has a new coverage normalization feature. As demonstrated in previous studies*, it is possible to remove the majority of reads prior to assembly when the coverage is higher than necessary. This allows for faster, less memory-intensive assembly. NextGENe's normalization operates on a similar principle, but uses a different approach.

**arXiv:1203.4802v2 [q-bio.GN]*

Methods

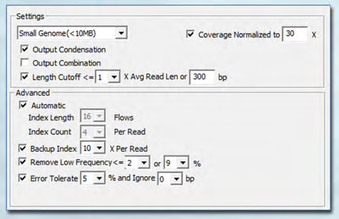
Datasets can be processed in minutes on a desktop computer running a 64-bit Windows operating system. All steps are performed with an easy-to-use point-and-click interface with no scripting required. Three datasets were processed in this analysis. Average coverage and N50 read length are listed for the data after quality trimming and filtering was performed.

Sample	Genome	Chip	Average Coverage	N50 Read Length
C12-245	DH10B	316	172	231
BEA-629	DH10B	316	124	373
B7-143	EHEC	318	191	286

All three datasets had over 120x coverage on average, but were normalized to 30x coverage prior to assembly.

NextGENe®
2nd Generation Sequence Analysis Software

Settings



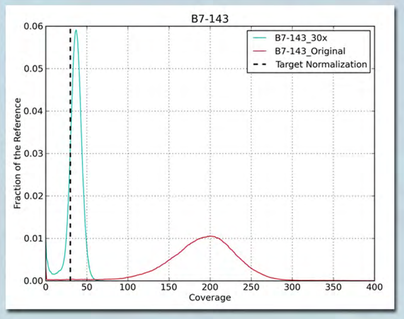
- Coverage was normalized to 30x
- The "Automatic" option will examine the estimated level of coverage and adjust the number and length of primary indexes and the low frequency removal settings.
- Any contigs less than 150 bp or 1x the average read length were removed.

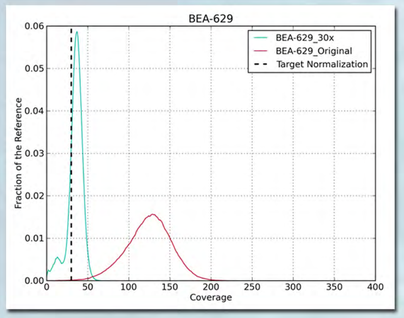
Results

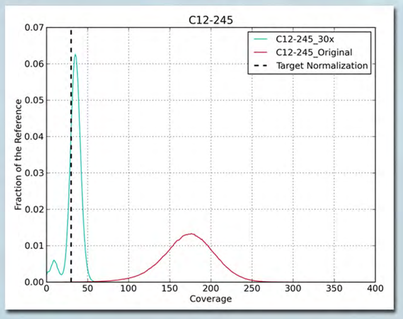
Assembly Results

Sample	Running Time (30x Norm)	Running Time (No Norm)	Number of Contigs	Max Contig Length	N50	N90	Total Length
C12-245	15 min	46 min	127	281,527	82,449	32,192	4,526,885
BEA-629	11 min	32 min	139	174,485	67,787	25,704	4,627,866
B7-143	20 min	69 min	159	292,010	111,393	25,254	5,397,238

Normalization Results







Discussion

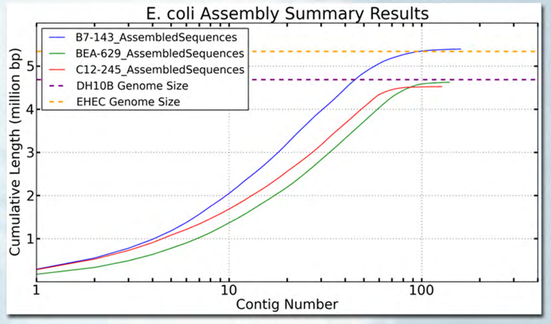
The assembly has 4 steps:

1. Normalization (optional)
2. Condensation – overlaps are found by keyword
3. Combination – overlaps are found when the same read is used in multiple condensation contigs
4. Contig Overlap Merging

NextGENe's new assembly normalization function first examines the frequency of sub-sequences in flow-space. Reads are randomly retained based on a probability that is related to three factors:

- The estimated coverage compared to the normalization level
- The quality (based on differences in frequency between different indexes in the same read
- The length of the read

The total assembly size was close to the expected size for all three samples.



Conclusion

NextGENe's new normalization method can greatly increase the speed of assembly for high-coverage datasets. The total processing time (including the time to normalize the data) was approximately 1/3 of the assembly time for the full dataset.

The Floton Assembler settings and results shown here are from version 2.3.2 in which the analysis parameters will be simplified.

Acknowledgements

All three data sets were downloaded from the Ion Torrent Community.

SOFTGENETICS®
Software PowerTools for Genetic Analysis
www.softgenetics.com