Using NextGENe® Software to assemble Illumina® MiSeq™ Paired End Reads with the Stepwise PE Assembler

November 2011

John McGuigan, Megan Manion, Kevin LeVan, Jacie Wu, CS Jonathan Liu

Introduction

Sequences that repeat throughout the genome can pose a problem for the assembly of short reads. Normally the repeat reads would assemble within the incorrect contig (figure 1). This causes two problems- the repeat regions elsewhere in the genome have reduced coverage and contigs terminate at repeat regions due to the formation of multiple ambiguous contigs.

NextGENe software solves this problem with a stepwise paired-end assembly. The end of a contig produced by assembly may indicate a repeat region. The software first calls in the reads paired to those assembled (using overlaps) at the end of the contig (up to 1 ½ times the library size from the end). These reads are then assembled and the software chooses the most complete assembly to continue the contig (figure 2). Shorter assemblies are not used because they are formed by repetitive sequences erroneously assembled together due to their similarity.

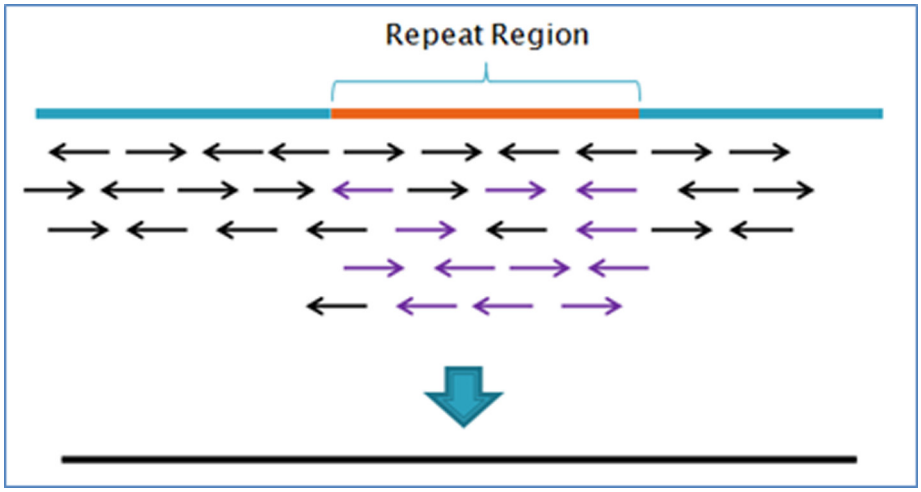


Figure 1 – Incorrect assembly of reads in a repeat region. Purple reads are from the same repeat sequence elsewhere in the genome.

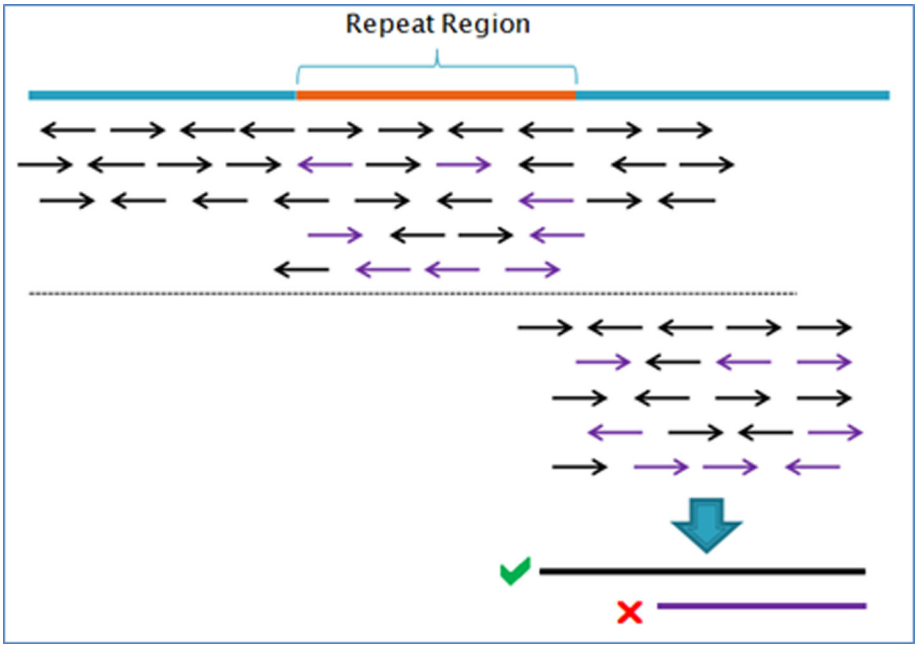


Figure 2 – NextGENe calls in the paired reads and assembles them. Reads paired to those that do not belong in the original assembly will form a separate contig (purple). Only the longest contig is used in the assembly.

Procedure

In this analysis, a 150bp PE read dataset from a MiSeq instrument was used to assemble the E coli MG1655 genome. First, the format conversion tool is used to convert the raw data (FASTQ files) to FASTA format while performing some filtering and trimming based on the quality scores (Figure 3). Next, the data is loaded using NextGENe's Project wizard. The instrument type (Illumina) and application (de novo assembly) are selected (Sequence Condensation is not used). The PE assembler is selected, and settings are adjusted (Figure 4).

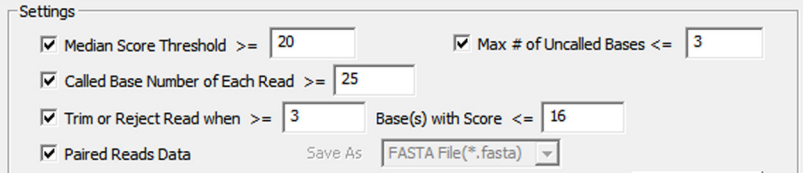


Figure 3 – Format Conversion Settings

The "word length" setting should be adjusted based on expected coverage. A word length of 23 is optimal for 20 to 30x coverage and a word length of 29 is sufficient for 50x coverage. The maximum suggested word length is 31. Longer word lengths allow for greater reduction of noise. The "high coverage limited" setting can be used to ignore coverage over a certain depth in order to improve processing speed. "Final contig merging" occurs after the PE assembly- the resulting contigs are checked for overlapping regions where they can be merged. "Reduce Memory Usage" will index only the 5' end of reads, which can result in a significant reduction in the amount of memory used for longer reads.

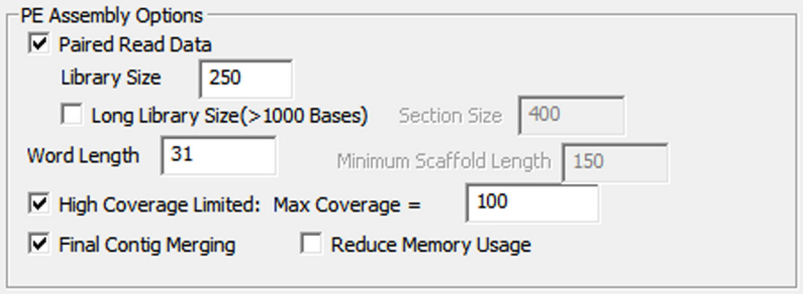


Figure 4 – Assembly Settings

Results

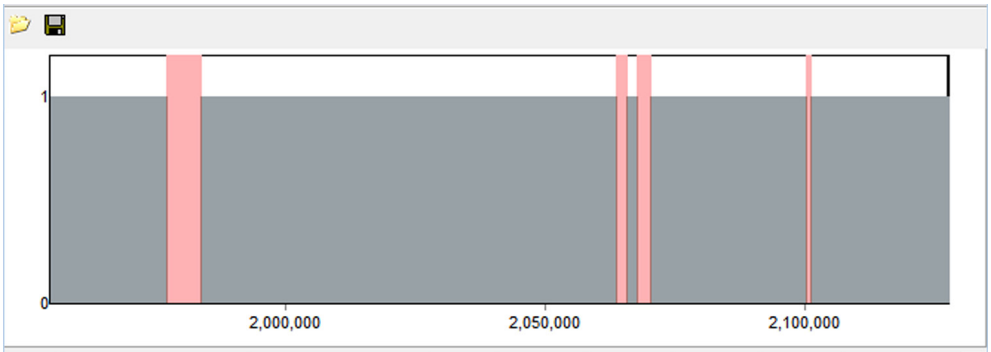
The original dataset consisted of 8,331,866 paired-end 150bp reads. After format conversion using default settings (with the paired reads option selected) there were 7,267,665 reads remaining. These reads were 137 bp on average, for an expected depth of coverage of 229x.

The assembly results are summarized in table 1.

Processing Time	1 hour, 48 min
Total Reads	7,267,665
Reads Used	7,101,508
% Reads Used	97.7%
Number of Contigs	65
Average Length	70,747 bp
Minimum Length	363
Maximum Length	494,530 bp
N50	212,730 bp

Table 1 – Assembly Results

The assembled contigs were aligned to the E. coli MG1655 reference genome. The coverage curve report was used to identify regions of the reference genome that were not covered at least once (Figure 5). Only 144,789 bp (3.12%) of the reference was not covered.



Index	Reference Position S	Reference Position E	Length	Gene Star	Gene End	CDS Star	CDS End	Descriptio
1	1	5618	5618					
2	15619	16404	786	insL	insL	1	1	
3	20120	20236	117	insB	insB	1	1	
4	226605	228685	2081	rrlH				
5	269986	274232	4247	insI	insH	1	1	
6	278711	278857	147	insB	insB	1	1	
7	290182	290338	157	insB	insB	1	1	

Figure 5 – Coverage Curve Report

Discussion

NextGENe's PE Assembler is a highly accurate and efficient tool for assembly of paired-end or mate-pair data. A small bacterial genome such as E. coli can be assembled into less than 70 contigs from a single MiSeq data run in less than 2 hours. The assembler can be run on relatively low-cost hardware- in this case the program needed less than 8 GB of RAM.

Acknowledgements

We would like to thank Illumina Inc. for supplying the data used in this analysis.