Merging Paired End Reads Using NextGENe® 's Condensation Tool®

February 2010

Jacie Wu, Megan Manion, Kevin LeVan, Ni Shouyong, CS Jonathan Liu

Introduction

The short read lengths produced by Next Generation sequencers such as the Illumina Genome Analyzer can create difficulty for accurate analysis of data. Additionally, relatively high error rates (compared to other technologies such as Sanger sequencing) further complicate the analysis of next-gen sequencing data. For these reasons, lengthening the short reads prior to analysis and statistically correcting or removing errors is a valuable tool to improve alignment accuracy, indel detection and assembly. A novel, highly accurate method to elongate reads, utilizing paired end reads, has been developed by SoftGenetics for its NextGENe software.

Sequencing Paired End reads is a useful technique which produces reads in pairs such that each pair of reads are a known distance from each other in the genome. This is accomplished by preparing DNA fragments of a certain length (200 bp, for example). This fragment size, or library size, is the distance between each pair of reads. Sequencing is then done from each end of the fragment, producing two paired reads. NextGENe's paired end merging technique takes advantage of paired end information, along with the additional coverage from sequenced overlapping DNA fragments, to produce long reads spanning the entire library size with an extremely low error rate.

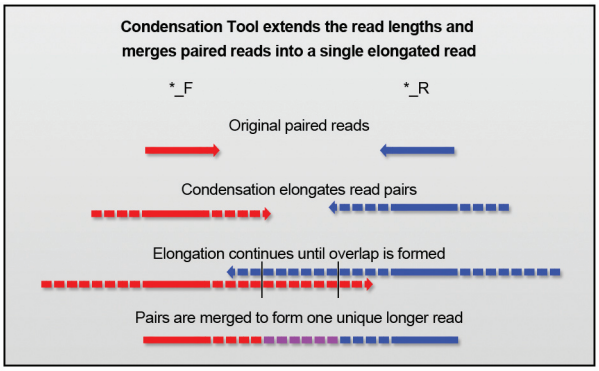

Figure 1: Multiple Cycles of Condensation can be used to elongate the paired reads, forming an overlap between the two, which allows one single read spanning the entire library to be created from the pair.

Methodology

The Paired End Merging application of NextGENe software uses overlapping elongated reads to create a link between two paired reads. The Condensation Tool clusters similar reads using a 12 bp anchor sequence as well as flanking shoulder sequences (of varying length). Using the Elongation method for Condensation, overlapping reads that are clustered together can be used to correct sequencing errors and extend read lengths. Condensation methods are also available to reduce read count by merging redundant reads (Consolidation) and to maintain original read lengths and read count while removing errors only (Error Correction). Consecutive cycles of Condensation can be used to continually lengthen reads.

Paired end reads can be merged by elongating the paired reads to the point that there is overlap between the two reads. This allows the paired reads to be joined together to form one continuous, longer read. The number of elongation cycles required depends on the read lengths and the library size. Each cycle of Condensation will generally increase the average read length to 1.6 the original length for shorter (<=36 bp) reads and to 6 bases less than twice the original length for longer (>36 bp) reads. These values may be reduced with an average depth of coverage less than 30x. For 75 bp reads from a 200 bp library, for example, a single cycle of elongation allows the paired reads to overlap and be linked together. For 35 bp reads from a 200 bp library, three cycles of elongation allows for the linking of the paired reads. Reads should be extended until a significant portion of the paired reads (roughly 15% of the elongated read length) will be expected to overlap.

Original Read Length	35 bp	50 bp	75 bp
Avg Read Length After 1 Cycle of Elongation	56 bp	88bp	138 bp
Avg Read Length After 2 Cycles of Elongation	90 bp	160 bp	
Avg Read Length After 3 Cycles of Elongation	144 bp		

Table 1: Average read lengths after elongation for varying original read lengths

The result of merging pairs is an extremely accurate elongated read with a very low error rate. The paired reads are merged only if the overlapping regions match between the reads. Errors resulting from sequencing chemistry, basecalling or the initial assembly by elongation will not match with the paired read so the pair will not be merged. Many chemistry and basecalling errors are corrected by Condensation prior to merging paired reads.

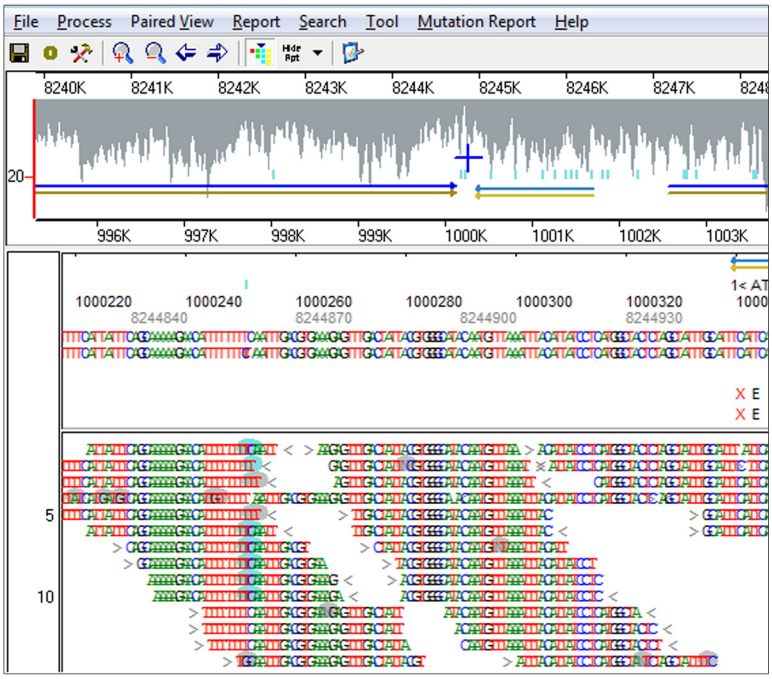

Figure 2: Raw reads aligned to the reference, prior to merging

Figure 3: Alignment of elongated, merged reads to the reference allows the detection of a 10 bp deletion

Procedure

1. Select “Illumina” under Instrument Type.
2. Select de novo Assembly under Application Type.
   Sequence Condensation and Sequence Assembly are both automatically selected under Steps.
3. Deselect Sequence Assembly. Sequence Condensation is the only step required.
4. Click Next to determine Condensation Settings.
5. Click “Inspect Input Files.”
   Known Counts and Read Lengths are automatically filled in. Supply the expected length of the assembled data in total bases, if know, to determine the expected coverage or manually select the expected coverage from the drop-down menu. This information is used to determine optimal advanced settings.
6. Select “Elongation Only” from the Condensation Type drop-down menu.
7. Click the “Paired” button next to the Condensation type menu.
8. Check the box to “Merge Overlapping Paired Reads.
9. Click “OK”
10. If more than one cycle of Condensation would be required to generate overlap between each pair, click the “Open Advanced Settings” button to edit the Number of Cycles based on the read lengths and the library size. (See Methodology Section for details.)

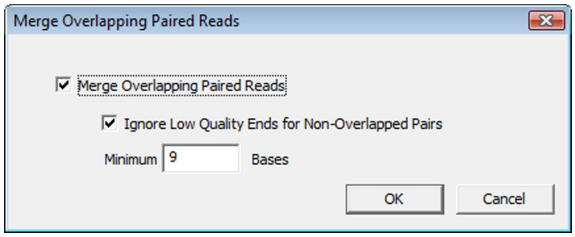

Figure 4: Settings for Paired End Merging

Discussion

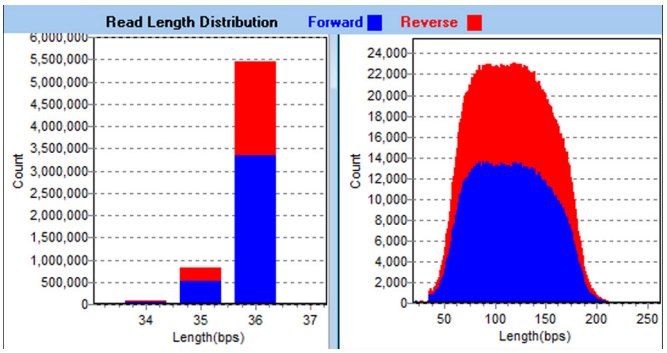

Figure 5: Read length distribution for raw reads (left) and merged reads (right)

NextGENe’s Paired End Merging function allows the generation of long, high accuracy reads from short paired end sequencing reads, such as those from the Illumina Genome Analyzer. Individual paired reads are elongated to form an overlapping region between the reads. When the overlapping regions are consistent between the paired reads the reads are merged to form one continuous read.

This process is highly accurate since any errors in sequencing or elongation will prevent the reads from being merged. These longer reads will help improve alignment accuracy and also allow for the detection of larger indels. NextGENe is able to detect deletions of up to 1/3 of the read length and insertions of up to 1/5 of the read length. Increasing the read length from 50 bp to 200 bp increases the indel detection capability from 16bp (deletion) and 10 bp (insertion) to 66 bp (deletion) and 40 bp (insertion). In addition to detecting large indels, these long, high quality reads also allow for determining the locations of other structural variants and splicing variations when used in conjunction with NextGENe software’s Detect Structural Variations option.