

Alignment of Next Generation Paired Read Data Using NextGENe Software

Megan Manion, Kevin LeVan, Ni Shouyong and CS Jonathan Liu

Introduction

Next generation sequencing platforms like the Illumina[®] Genome Analyzer (Solexa sequencing), the Applied Biosystems SOLiD[™] System and the Genome Sequencer FLX System from Roche Applied Science (454 Sequencing) have drastically reduced sequencing costs, producing genetic data at an increased speed and quantity compared to other technologies such as Sanger sequencing and microarray methods. This has tremendously increased the quantity of sequence data that can be obtained for mutation detection. However, the short read lengths of these technologies compared to Sanger sequencing data are often not sufficient to represent a unique location within a genome which presents a challenge for accurately aligning reads to a reference and detecting variants.

The short reads also create difficulty for detecting structural variations such as large insertions and deletions (more than 1 kbp), translocations and inversions. The technique of aligning paired reads (paired end or mate-paired reads) to a reference sequence is useful for improving the accuracy of alignments and detecting these large structural variations (1, 2).

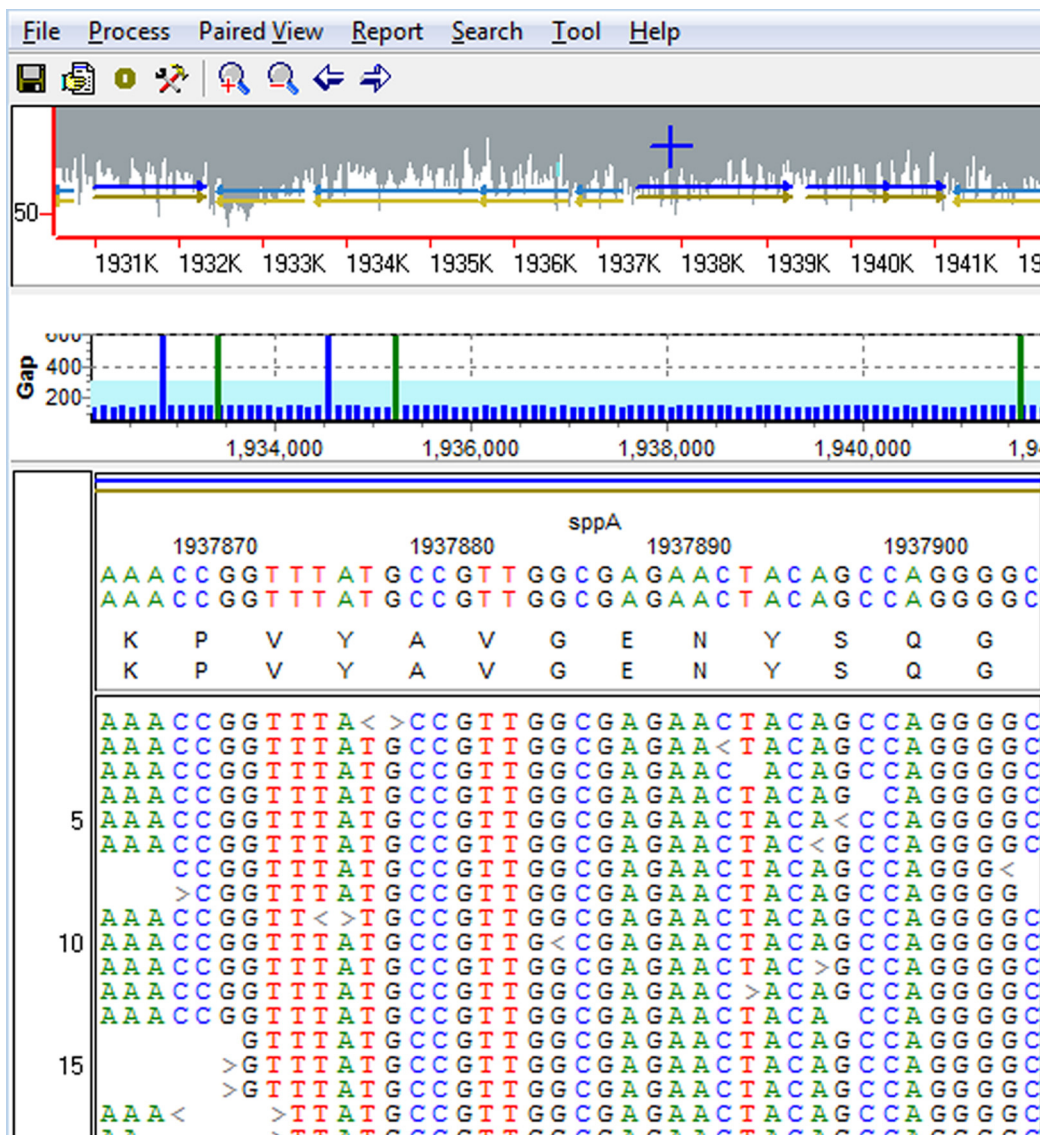


Figure 1

Figure 1: When paired reads are used for alignment to a reference sequence, the Sequence Alignment Window contains additional display features and reports. Results shown are for the alignment of paired end Illumina data with a library size of 200bp. The Paired View is displayed below the Whole Genome View and shows that most pairs are aligned with a gap distance near 200bp. Green bars indicate pairs in a localized region that are oriented in the same direction, while blue bars indicate pairs that align in opposite orientation to each other.

Since reads in a pair are expected to align a given distance from each other with an expected relative orientation, pairs where the distance between reads or their orientation differs from the expected can indicate positions of structural rearrangements. Paired end reads, such as those from Illumina GA paired end sequencing, are generated by sequencing two ends of a molecule of DNA. Generally, the length of fragments used, known as the insert or library size is between 50-500 bp. Mate-paired reads, such as those from the SOLiD System, are generated following circularization of larger DNA fragments using an internal adaptor between ends of the fragment, then breaking of the circular DNA using restriction enzymes to produce smaller fragments with the internal adaptor between the two ends of the original fragment. Typical mate-paired library sizes are 0.6, 1, 2.5 and 4 kb. It is possible for the use of larger fragment sizes to result in some bias in the selectivity of fragments since breakage of DNA into larger fragments (0.6-4 kbp) can be less random than breaking into smaller fragment (50-500bp).


NextGENe utilizes a unique technology to align paired read data to a reference and track the distance between paired reads and the orientation of pairs. Several reports are created that provide information about the alignment of pairs and specialized display features are available for visualizing results including detected variants.

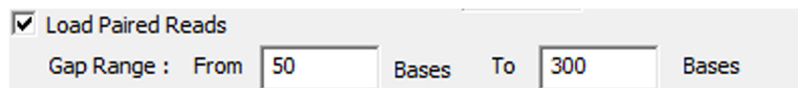
Methodology

When “Load Paired Reads” is selected in the Alignment Settings, the software attempts to align reads to locations where the gap distance between paired reads is within the expected range set by the user. If pairs cannot be aligned with a gap distance within the expected range, software will then align the reads to the best matching position and report the observed gap distance.

There are several possibilities when aligning paired read data; both reads could match and be oriented in opposite directions, both reads could match and be oriented in the same direction, one read in the pair could match while its mate is unmatched, or neither read in the pair could match. Additionally, paired read sample files occasionally include some unpaired reads that could be either matched or unmatched. NextGENe considers each of these possibilities and provides statistics for each when aligning paired read data. Different sequencing systems produce paired reads differently, so the information about the alignment of paired reads that is of the most interest will vary according to the system that was used to generate the data. For this reason, information for pairs in which both reads align in the same direction and pairs in which reads align in opposite directions are both reported in separate reports. When reads in a pair align in an orientation opposite of what is expected, this can indicate that an inversion has occurred at one of the locations.

Procedure

1. Open NextGENe’s Run Wizard by clicking on the  icon in the main toolbar.
2. Select Instrument Type.
3. Select SNP/Indel Discovery for Application Type.
 - a. Sequence Condensation and Sequence Alignment will be selected automatically.
 - b. Condensation can be deselected to align raw paired reads.
4. Click “Next” to load data.
5. Click “Load” and browse to upload both sample files generated for a lane.
 - a. If sample files are not in fasta format, or csfasta for SOLiD data, use the Format Conversion Tool to convert to fasta.
6. Click “Load” and browse to load reference file in fasta or GBK format.
7. Click “Set” and browse to specify output file location and file name.
8. Click “Next” to specify settings for Condensation.
 - a. Under Condensation Type, select Elongation Only or Error Correction Only.
9. Click “Next” to continue to Alignment Settings.
 - a. Under File Type select “Load Paired Reads” and specify expected gap range.
10. Click “Finish.”
11. Select “Run NextGENe” to begin processing project.



Load Paired Reads
Gap Range : From Bases To Bases

Figure 2

Figure 2: To align paired reads, provide an expected range for gap distance (library or insert size).

Results

The Sequence Alignment Tool is designed to view and edit the results of reads aligned to a reference sequence. When a project is aligned by NextGENe, the project is automatically opened in the Sequence Alignment Tool. Projects can also be saved and opened for review and editing at a later time. The Sequence Alignment Tool contains several views of the project including the Whole Genome Viewer and the Alignment Viewer, in addition to several other views and reports. When paired reads have been used for alignment, specialized reports and viewing options are available.

The Whole Genome Viewer is the upper pane within the display that shows a global view of the project. Coverage across the whole reference is contained within this viewer, in addition to segment breakpoints, biological information, mutation calls, indications of exon, mRNA and coding regions, and current position of aligned reads in the Alignment Viewer. The Alignment Viewer is the lower pane of the Sequence Alignment Window which contains a view of all the reads as they align to the reference sequence. Variations between the reference and sample sequences are highlighted – gray positions are variations below a set frequency, often instrumental error, and mutation calls, both SNPs and Indels, are highlighted in blue for novel variations and purple for reported variations.

Paired View

When paired reads are used for alignment the Paired View pane is displayed below the Whole Genome Viewer and displays a histogram representing the gap distances across the whole reference. The distance between pairs within a local region that are oriented in opposite directions to each other are shown with a blue bar while distance of pairs that are oriented in the same direction are shown with a green bar.

Several Paired Reports are generated to provide a list of all pairs that align to the reference with a gap distance outside of the expected range.

The **Paired Report – Opposing Directions** includes pairs that are oriented in opposite directions with gap distances outside the expected range. A separate Paired Report is created for pairs that aligned in the same direction with a gap distance outside the expected range (**Paired Report – Same Direction**). Both reports also include information about each pair such as the position of both reads, both read names and the gap distance, as well as a histogram showing gap distance size by reference position. The expected gap distance range is highlighted across the histogram in blue.

Index	Read1 Name	Read2 Name	Read1 Start	Read2 Start	Gap Distance
1	>1378_824_691_F3	>1378_824_691_R3	34154	2723377	2689200
2	>1378_969_78_F3	>1378_969_78_R3	247948	842224	594253
3	>1378_991_574_F3	>1378_991_574_R3	437004	3946011	3508984
4	>1378_1046_612_F3	>1378_1046_612_R3	780171	858707	78513

Figure 3

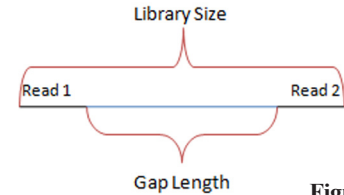


Figure 4

Figure 3: The Paired Reports provide information about the position of pairs that aligned with a gap size outside the expected range.

A Paired Report is also produced for single aligned reads (Single Reads Report). This report provides the name and position of all reads that aligned without a mate.

The **Paired Reads Gap Distribution** shows the distribution of pairs with continuous gap sizes. Two charts are shown; the top chart shows the gap sizes for pairs that are oriented in opposite directions while the bottom chart shows pairs that are oriented in the same direction. The Gap Length is not the same as the library size but is closely related. Gap Length is defined as the distance between the ends of the reads in the pair (see Figure 4). The Gap Length = Library Size – 2(read length). So, for larger library sizes, the gap length can be considered equivalent to the library or insert size.

Figure 4: Gap Length represents the distance between the reads in each pair.

Figure 5: The Paired Reads Gap Distribution chart displays the number of pairs aligned with varying gap distances. This figure shows the gap length distribution near the library size for the dataset. A normal distribution is seen here.

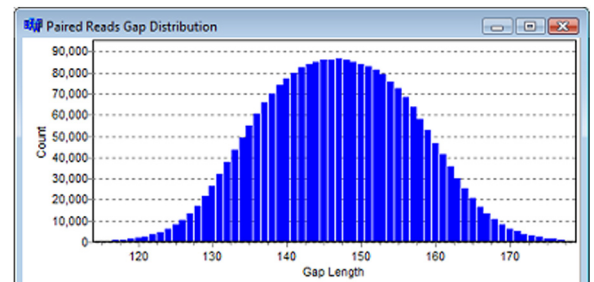


Figure 5

Figure 6: This figure shows the gap length distribution far from the library size. Reads with these gap sizes may represent structural variations.

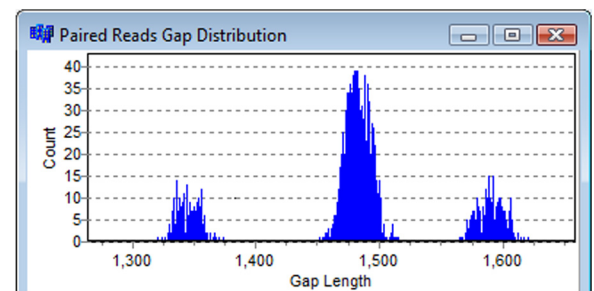


Figure 6

Paired Read Statistics are available that show various statistics related to paired reads including matched read count and matched reads with gap distance in expected range as well as reads that matched to the reference with a mate that did not match and pairs where neither read matched.

Figure 7: The Paired Reads Statistical Results Report provides information about the paired read alignment.

Discussion

NextGENE Software is able to accurately align paired reads to a reference sequence and track the gap distance and orientation between paired reads to determine large structural variations in DNA sequences. In addition, NextGENE has the capability to correct base call errors and elongate these paired-end reads prior to the alignment, enhancing the alignment accuracy.

NextGENE also includes software applications for de novo assembly (with or without use of paired reads), ChIP-Seq, Transcriptome, small RNA discovery and quantification and SAGE.

References

1. J O Korbelt et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science. 318: 420-426.
2. E Tuzun et al. 2005. Fine-scale structural variation of the human Genome. Nature Genetics. 37: 727-732.

Paired Reads Statistical Results	
Total Reads Count	5244764
Unpaired Reads Count	0
Matched Reads Count	4890795
Matched Paired Reads Count	4913090
Matched Paired Reads within Expected Gap Distance Count	4896860
Matched Unpaired Reads Count	0
Paired Reads with Only one Read Matched Count	77705
Paired Reads Matched with Same Direction Count	3094

Figure 7

Trademarks are property of their respective owners.