

Fast Whole Genome Alignment of 2nd Generation Sequencing Reads Using NextGENe® Software

Edward Bouton, Rong Ma, Yaping You, Megan Manion, Kevin LeVan, Shouyong Ni, and CS Jonathan Liu

Introduction

SoftGenetics has developed a Burrows-Wheeler transform (BWT) alignment method that includes several improvements over other methods to generate fast alignment of sequence reads to a whole large genome reference such as the human genome with high accuracy.

NextGENe's whole genome alignment method is the first to align reads from the Roche Genome Sequencer FLX System, which often contain many indels due to homopolymer errors, to a whole genome reference with high speed. The whole genome alignment algorithm is also capable of quickly aligning SOLiD™ System and Illumina® Genome Analyzer data. Additionally, NextGENe's whole genome alignment tool features complete annotation of the reference.

Alignment of high throughput short sequence reads to a large reference genome like the human genome is a difficult challenge. The Burrows-Wheeler transform is a widely accepted data compression algorithm that has been in use since 1994(1). Massively parallel sequencers such as the Illumina Genome Analyzer (Solexa sequencing technology), the SOLiD System from Applied Biosystems and the Genome Sequencer FLX System from Roche Applied Science (454 Sequencing) are capable of producing 1-200 million reads per run which has led to an interest in the usage of BWT algorithm to align this large volume of sample reads to entire genomes. The algorithm has been successfully used for this purpose by alignment programs such as Bowtie(2). In May 2009 researchers from The Wellcome Trust Sanger Institute published a paper detailing their novel BWT alignment method, BWA(3), which improved upon previous methods by allowing for the alignment of 51bp Illumina reads as well as SOLiD color-space reads.

NextGENe's whole genome alignment algorithm aligns reads to the whole genome by matching seeds smaller than the read length and then extending the alignment to find the best matching position for the whole read. This allows for the alignment of long reads and reads with indels.

	Total Reads	Read Length	Seed Size	Matched Reads	% Matched	Time (hrs)		
						8 cores	4 cores	1 core
Long Reads	1.46M	100-200 bp	17 bp	1.32M	90.4%	1.5	3	10.5
Short Reads	24.4M	51 bp	17 bp	20.7M	84.8%	1.5	2	5

Table 1

Table 1: Whole genome alignment performance for short read and long read data. Data shown is for alignment to the whole human genome. Seed size indicates the number of bases used to find perfect matching positions in the reference.

Methodology

NextGENe's whole genome alignment employs a suffix array to quickly locate the best matching location for each read. As a full suffix array utilizes an inordinate amount of disk space, NextGENe uses the Burrows-Wheeler transform (BWT) to represent the entire suffix array. An efficient rank algorithm allows the software to quickly traverse the suffix array to find the best match for each read. Along with the BWT the software maintains genome positions at every four base pairs within the genome allowing the software to monitor these locations while traversing the reference genome. This provides a quick and highly accurate alignment.

Seeds smaller than the read lengths are used to identify the best matching position within the genome. After finding the best match the alignment is expanded to align the entire read using our traditional NextGENe algorithm in order to align reads with indels and mismatches. This also allows NextGENe to work with reads longer than 120 bp including large indels. Longer reads are more likely to have a greater number of sequencing errors which preclude accurate alignment by other methods. Using 17 bp seed size with 51 bp reads, each read is compared to the reference 7 times (using move step =5). This dedicated NextGENe algorithm assures highly accurate alignments by rigorously checking matching positions while operating at high processing speed.

Burrows-Wheeler Transform		
Original sequence: gattaca\$		
Dollar sign indicates end of sequence		
Rotate sequences:	Sort Rotations	Last Column of Sorted
gattaca\$	\$gattaca	a
\$gattaca	a\$gattac	c
a\$gattac	aca\$gatt	t
ca\$gatta	attaca\$g	g
aca\$gatt	ca\$gatta	a
taca\$gat	gattaca\$	\$
ttaca\$ga	taca\$gat	t
attaca\$g	ttaca\$ga	a
BWT Transform: actga\$ta		

Figure 1

Figure 1: Burrows-Wheeler Transform algorithm

Figure 2: Settings for the Whole Genome Alignment. Seed length indicates the size of the seed to be used to determine the exact matching locations within the genome. Move step indicates the number of bases between seed start positions. Matching base percentage refers to the percent of the read required to match the reference. If a seed can be matched to a number of positions above the set threshold in the suppress setting it is likely a repeat and is disregarded.

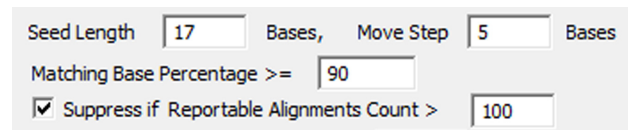


Figure 2

NextGENE's algorithm allows for rapid and accurate alignment of millions of next generation sequencing reads to whole large genomes such as human, mouse and rat.

Once reads are accurately aligned, mutation positions are identified and numerous report outputs are available.

Results

NextGENE's whole genome alignment algorithm was used to align human sequence data from the Roche Genome Sequencer FLX and the Illumina Genome Analyzer. The Roche dataset included 1.46M reads of 100-200 bp per read. These reads were aligned to the whole human reference in 10.5 hours using a single 2.8 GHz processing core on a Windows® operating system using 6.5 GB. The Illumina data is from a human male, NA12750 from the 1000 Genomes Project. The dataset includes 24.4 million 51 bp reads. These reads were aligned to the genome in 5 hrs. using a single processing core with the same specifications described above. Processing time can be significantly reduced by utilizing additional cores. (Table 1)

```
[Tuesday, June 09, 2009, 12:08:22] Whole_Gene_Alignment Begins ...
[Tuesday, June 09, 2009, 12:10:31] Load Index ...
[Tuesday, June 09, 2009, 12:11:57] Load Reference ...
[Tuesday, June 09, 2009, 12:11:57] Load&Process Sample Files ...
[Tuesday, June 09, 2009, 13:00:16] Do Statistics & Detect Mutations ...
[Tuesday, June 09, 2009, 13:32:14] Whole_Gene_Alignment Complete.
Perfectly Matched Reads Number: 16489122

[Alignment Statistics Information]
Matched Reads Count: 20690621
Unmatched Reads Count: 3588951
```

Figure 3

Figure 3: Whole genome alignment performance as displayed in the Run Log and the statinfo.txt file

Upon completion of a whole genome alignment project the results are automatically displayed in the Sequence Alignment window. The Sequence Alignment window displays coverage information, aligned reads, complete annotations and provides access to reports. NextGENE's whole genome alignment function includes annotation information which is shown in this window.

Figure 4: Whole genome alignment results are displayed in the sequence alignment window. The top pane uses gray regions to indicate coverage across the genome. As shown in this figure, the view can be zoomed in to show a detailed view of a small region. The blue, gold and green lines are used to indicate gene, CDS and mRNA locations, respectively. Tick marks are used to indicate SNP locations with blue marks indicating novel SNPs, purple indicating known and green indicating negative SNPs. The bottom pane shows aligned reads with SNP positions highlighted. The middle pane provides the reference and consensus nucleotide sequences as well as the amino acid sequences. The gene name is provided.

A mutation report lists all the detected mutations, provides annotation information for each position as well as coverage and allele frequencies. The report can be easily edited, saved and exported.

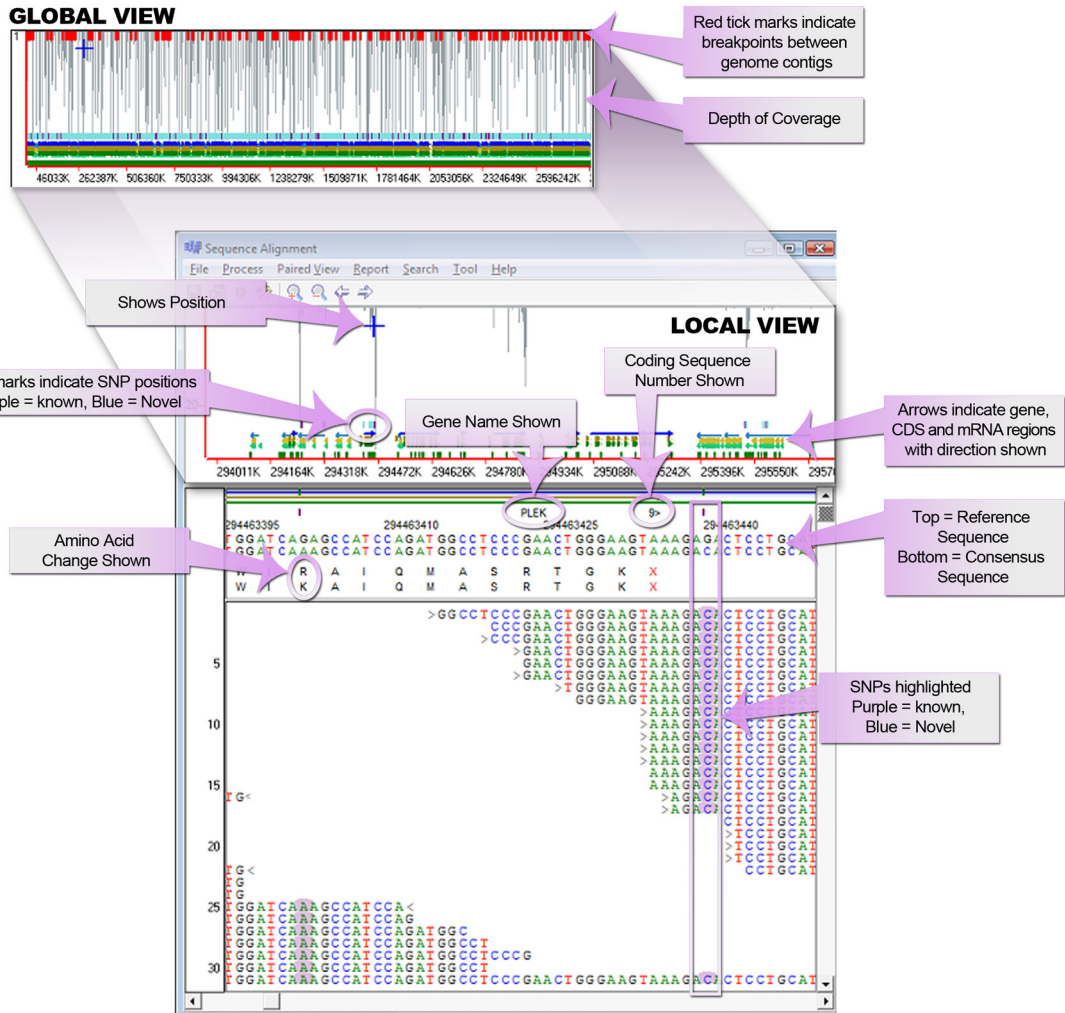


Figure 4

Figure 5: The mutation report lists all identified mutations and provides information for each including the gene name, chromosome position, reference nucleotide, coverage, allele frequencies and dbSNP identification. Clicking on the dbSNP ID provides a direct hyperlink to the NCBI website.

Index	Reference Position	Gene	Chr	Reference Nucleotide	Coverage	A (%)	C (%)	G (%)	T (%)	Ins (%)	Del (%)	SNP db_xref
9594	294461806	PLEK	2	A	4	0.00	0.00	75.00	0.00	0.00	25.00	
9595	294461808	PLEK	2	T	4	0.00	0.00	75.00	0.00	0.00	25.00	
9596	294463402	PLEK	2	G	7	100.00	0.00	0.00	0.00	0.00	0.00	rs1063479
9597	294463440	PLEK	2	G	18	0.00	100.00	0.00	0.00	0.00	0.00	rs1050181
9598	294463758	PLEK	2	C	33	0.00	0.00	0.00	100.00	0.00	0.00	rs6713721
9599	294463758	PLEK	2	C	33	0.00	0.00	0.00	100.00	0.00	0.00	rs6713721

Figure 5

Discussion

NextGENe's alignment algorithm is a valuable tool for aligning massively parallel sequencing reads to large references such as the entire human genome. The algorithm is powerful in its speed and its ability to align reads with indels and long reads. The software also provides detailed annotations not available with other programs. Currently available references including human, mouse and rat can be downloaded from SoftGenetics website using a customer password or sent directly upon request. Custom indexes can be created for any reference genome by contacting SoftGenetics.

The whole genome alignment tool is included in the NextGENe software package which also contains applications for SNP and Indel detection for small genomes or genomic regions, ChIP-Seq, small RNA detection/quantification, de novo assembly, transcriptome analysis and SAGE studies. NextGENe also provides the unique Condensation Tool which can be used for instrumental error reduction.

References

1. Burrows, M. and Wheeler, D. J. 1994. A block-sorting lossless data compression algorithm. Technical Report 124.
2. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. 2009. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology* 10.
3. Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* (Oxford, England).

Trademarks are property of their respective owners.