The Geneious 6.0.3 Read Mapper

Authors

Developer: Matthew Kearse
Authors: Matthew Kearse, Shane Sturrock, Peter Meintjes

Abstract

High-throughput Next-Generation Sequencing (NGS) data are increasingly ubiquitous and abundant for life science and health care research. Many applications of this technology rely on high-fidelity mappings of new sample data to a previously characterized reference sequence. There are a growing number of tools capable of performing such read mapping, including the Geneious 6.0.3 Read Mapper. This white paper describes the read-mapping algorithm included with the Geneious software package (Kearse *et al.*, 2012) and provides a comparison with other leading open-source read-mapping algorithms. Six read mapping algorithms were evaluated on Illumina HiSeq and Ion Torrent sequence data from an *Escherichia coli* - BWA (0.6.2-r126), Bowtie 1 (0.12.8), Bowtie 2 (2.0.0-beta7), SMALT (0.6.4), SOAP2 (2.20) and Geneious (6.0.3). The results demonstrate that the Geneious Read Mapper produces superior results to the other mapping algorithms on these data sets.

Introduction

The goal of a mapping algorithm is to align short DNA sequence fragments to a reference sequence. In practice, the fragments produced by sequencing machines contain a variety of systematic and random errors, and the sample data may frequently be a different strain or species from the reference sequence. This means that imperfect matches between the sample and the reference may be either error or information. In the following image variations between the sample and the reference sequence are highlighted. Visually it is easy to determine which variants are sequencing errors and which are true variations from reference.

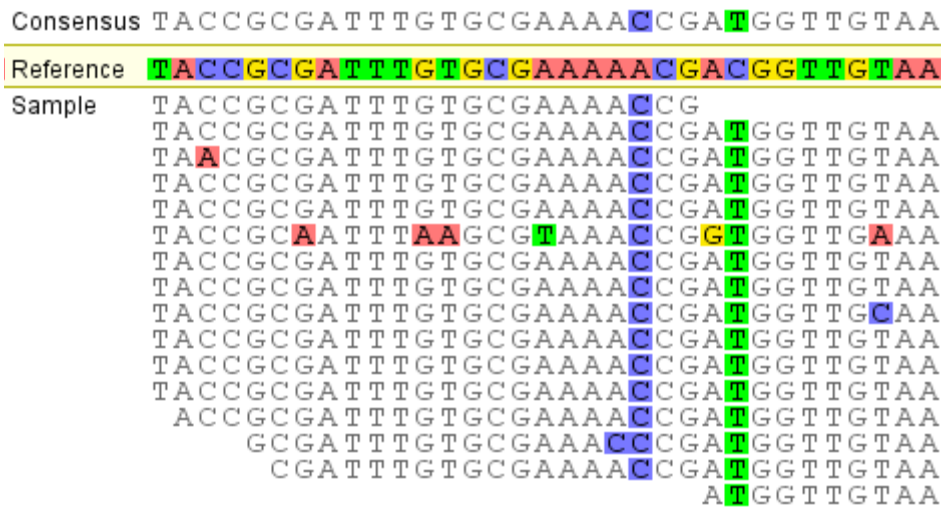


Figure 1: Highlighting sequencing errors and true variations between the sample and the reference

A reasonable approach to mapping is to find a Smith-Waterman (Smith and Waterman, 1981) alignment of each sequencing read to the reference sequence. When the sample being sequenced is highly similar to the reference sequence, this is an excellent approach. However, when the sample has diverged from the reference sequence, particularly in the presence of insertions or deletions, an independent Smith-Waterman alignment of each read to the reference is often incorrect. For example aligning each read independently would produce the alignment shown in Figure 2, whereas a correct mapping that doesn't treat each read independently should produce the alignment shown in Figure 3.

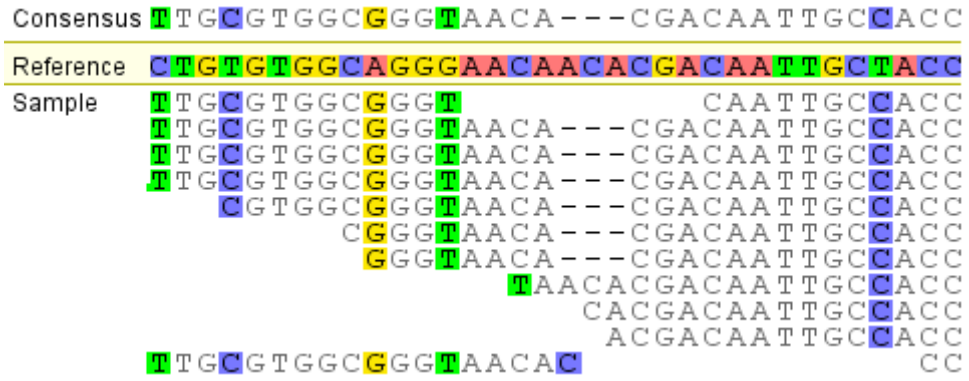


Figure 2: An independent Smith and Waterman alignment for each read

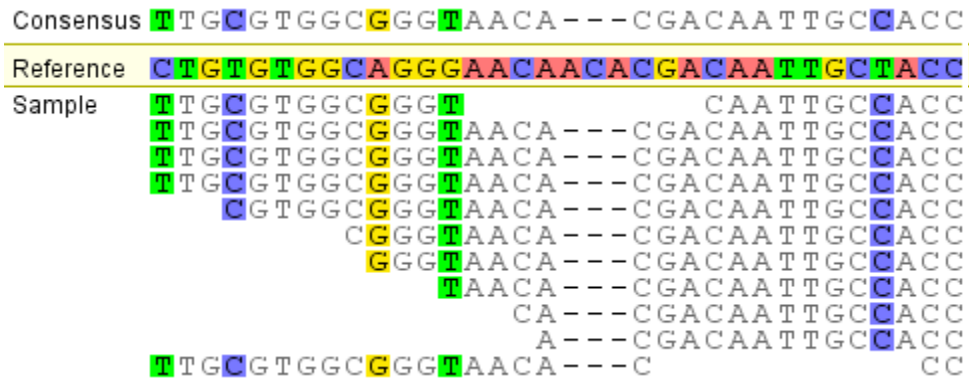


Figure 3: A correct alignment where each read is not treated independently

Importantly, Smith-Waterman is a local alignment algorithm and will tend to truncate the aligned region to improve the overall identity. While local alignment may have the beneficial effect of trimming the sequences, it is also likely that the gaps required by the correct alignment will be too costly, especially at the ends of a read, and the algorithm may truncate matching regions instead of accepting the cost of a gap (INDEL). Using a global alignment algorithm such as Needleman-Wunsch (Needleman and Wunsch, 1970) would avoid the truncation introduced by the local alignment. Realistically, determining an independent Smith-Waterman or Needleman-Wunsch alignment of each read to the reference is too computationally demanding on real data sets where there can be upwards of 1 billion sequencing reads to align to the reference sequences so mapping algorithms implement heuristics to find alignments. Geneious uses such heuristics, but also takes additional measures to ensure the mapping is globally correct rather than just an independent pairwise alignment of each read to the reference sequence.

Geneious Read Mapper Algorithm Overview

The first step implemented in the Geneious Read Mapper is the building of an index that records the location of all occurrences of all possible nucleotide sequences of a given length in the reference sequence. The exact length used for the index depends on the sensitivity chosen, but is typically in the range 10 to 15 bases, which produces a good trade off between sensitivity and performance.

For example, if our reference sequence is GATTATT and our index length is 2, then we would construct an index recording the positions that each possible subsequence starts from in the reference sequence:

AA		CA		GA	1	TA	4
AC		CC		GC		TC	

AG		CG		GG		TG	
AT	2, 5	CT		GT		TT	3, 6

For each sequencing read this table is used to identify the locations in the reference sequence of all subsequences of this length from the sequencing read. These are sorted and filtered to remove redundant adjacent matches to minimize the computation required by the algorithm later in the process.

For example, if we are searching for the location of TTAT, there are five candidate positions:
1: TT (the first two nucleotides for the query sequence) occurs at positions 3 and 6
2: TA (the 2nd and 3rd nucleotides) occurs at position 4
3: AT (the 3rd and 4th nucleotides) occurs at positions 2 and 5

These are sorted by diagonals (the difference between the position in the read and the position in the reference sequence) and nearby positions on the same diagonal are eliminated to leave three candidate positions, AT at positions 2 and 5 and TT at position 6.

For each remaining subsequence match, Geneious expands the matching region towards the ends of the sequencing read, potentially introducing gaps in regions where there is a mismatch with the reference sequence.

Continuing with the above example the three candidate subsequences are expanded to form the following alignments:

G**AT**TATT TT**AT**	GATT**AT**T TT**AT**	GATTA**TT** **TT**AT

Each fully expanded result is assigned a score based on the number of matches, mismatches and gaps introduced and the highest scoring result is used as the final location to which the read will be mapped. Reads that map equally well to multiple locations can either be mapped to a random best location, not mapped at all, or mapped to all locations at the discretion of the user.

Paired reads have their score slightly adjusted to favor those pairs that are closest to their expected insert size. For example, if two reads with an expected insert size of 500 bases maps perfectly to locations that are 5000 bases apart, but one of the reads mapped with a single mismatch at a location approximately 500 bases from its pair, then this second location would be selected.

Running the Geneious Read Mapper algorithm (Figure 4) with default settings obtains results comparable to the best read mappers available, but at higher sensitivity settings it outperforms

other mappers as demonstrated in the results section below. The results are significantly improved by the use of an iterative system (new in Geneious 6), where the Geneious Read Mapper maps reads to the consensus sequence from the previous iteration. The reads are converted back to mappings relative to the original reference sequence and the process is repeated. This allows more reads to be mapped to variable regions, makes reads better align to each other in INDEL regions (important for downstream analyses such as variant calling), and reduces the likelihood of reads mapping to an incorrect location in near perfect repeat regions.

In addition to the primary mapping algorithm and fine-tuning iteration, there are many heuristics and minor algorithms used throughout the mapping and iterative processes to improve the quality of results. For example, allowing a single mismatch in the seed, correct handling of circular genomes, consistently choosing the same one of many equally optimal results and weighting reads differently during consensus calling based on the number of mismatches to the reference.

As well as providing excellent results, the Geneious Read Mapper is also easy to use. It is integrated into the Geneious software platform, so researchers need not be familiar with command line tools to run the algorithm. Geneious is also agnostic with respect to input data file formats and the sequencing machines that created the data, so researchers do not need to concern themselves with the details of file formats or sequencing-technology specific errors or artifacts. The major considerations for a researcher are to assign the correct reference sequence and to select the desired speed/sensitivity trade-off.

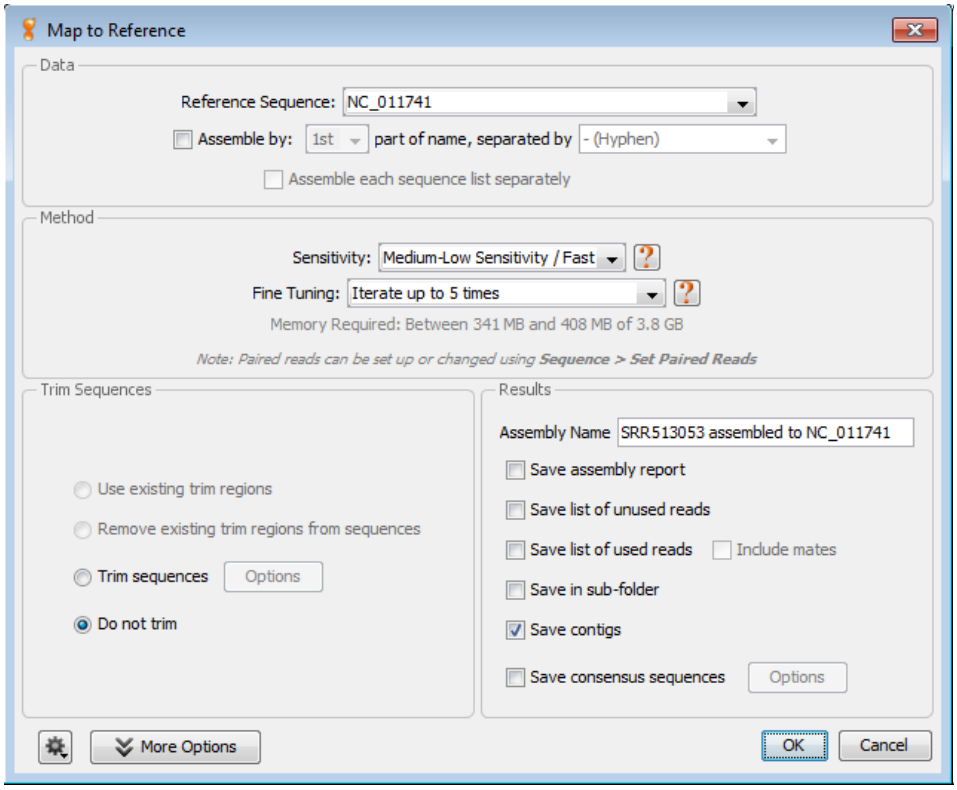


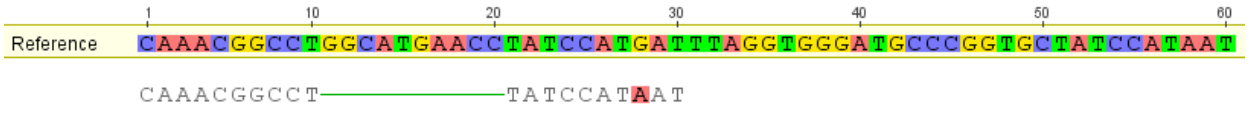
Figure 4: The Geneious Read Mapper settings

One potential criticism of the Geneious Read Mapper is the higher memory requirements when compared to other algorithms. For example, Geneious requires ~14 GB (10 GB for single iteration mapping) compared to about 2.5 GB for Bowtie[1]. With modern machines, where 16 GB of memory costs around $100, the 14 GB used by Geneious is not a concern.

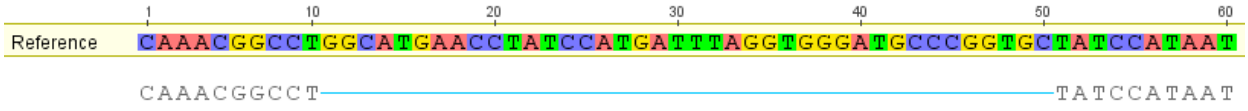
Quality Comparison

Evaluating the quality of read mapping algorithms is complex. For a more detailed discussion on challenges associated with this see Holtgrewe *et al*. (2011). One new challenge that arose during this study is that the gold standard for quality as used by Holtgrewe and colleagues is actually below the quality of results produced by the Geneious Read Mapper. For example a naive mapping algorithm may choose to map a read to a location where it matches perfectly to the reference sequence, but, in fact, the read should be mapped to a location where it doesn't match perfectly. Here we describe two scenarios in which mapping a read to a location perfectly can be incorrect.

1) Paired distances should be taken into account. A read at its expected paired distance with a single mismatch is more likely to be correct compared to mapping it at a distance of 10 times its expected distance without any mismatches. For example, imagine two 10 bp reads with an insert size of 30. If we favor mapping at the correct insert size, the result would be:



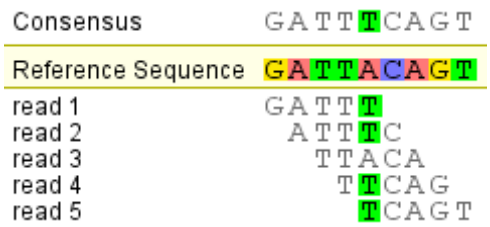
But if we favor no mismatches, the result would be



At only twice the expected distance, either result could be correct and we can't say with certainty which one is, but if in the second case the distance between the perfect matches was say 10,000 bp, then most likely the mapping with a single mismatch is correct.

2) Other reads may provide a strong indication that the sample does not match the reference sequence at a location.

1 http://bowtie-bio.sourceforge.net/index.shtml



In the above example there is evidence that the sample differs from the reference sequence and therefore 'read 3' may be better mapped elsewhere even though it perfectly matches the reference sequence at this location.

Putting aside problems such as this, evaluating quality is difficult. When using real sample data for an entire genome, even with a known reference we can't even be sure what the correct results should be for our sample. On the other hand, using simulated data where the answer is known doesn't accurately test how well the algorithm will perform on real data.

To ensure that we know both the correct alignment result and use real data for our analysis, we took Illumina HiSeq 2000 90 bp paired reads[2] from a whole genome re-sequencing sample of *E. coli K-12*, but limited it to a single well-characterised reference gene, *yghJ* where the correct alignment results are known. All 5,411,112 reads of the whole genome sample data set were mapped to NC_000913[3] (*E. coli str. K-12 MG1655*) using both the Geneious Read Mapper at high sensitivity and Bowtie. The reads where both pairs fully intersected the *yghJ* gene and where each read in the pair was at least 15% identical to the reference were extracted to form the 5,060 paired read data set and was identical for both the Geneious Read Mapper and Bowtie.

To begin the comparison across multiple mappers, the 5,060 paired read data set of *E. coli K-12 MG1655* for the *yghJ* gene was mapped to the *yghJ* gene from *E. coli IAI1* (NC_011741[4]) using a variety of algorithms. These two genes are 89% identical. Most of the variance comes from substitutions although there are four short INDELs. To evaluate the quality of the mapping for each read we made a record of how many mismatches the read has when aligned to the *yghJ* gene from NC_000913. Since the sample data consensus corresponds exactly to the *yghJ* gene from NC_000913 this indicates the number of errors present in each read. The consensus sequence was obtained from the mapped reads. A Needleman-Wunsch pairwise alignment of this consensus sequence with the *yghJ* gene from NC_000913 was made and the percentage of identical columns used to evaluate the Consensus Accuracy column in the following table. The number of mismatches between each mapped read and the consensus sequence were evaluated. If the number of mismatches exceeded the number of known errors for that read, then that read was considered to have been incorrectly aligned to the consensus.

2 Available from http://biomatters.com/assets/data/eColiYghjGeneData.zip

3 http://www.ncbi.nlm.nih.gov/nuccore/NC_000913

4 http://www.ncbi.nlm.nih.gov/nuccore/NC_011741

Quality Comparison Results

Algorithm	# Mapped	% Mapped	% Mapped & correctly aligned to consensus[1]	Consensus Accuracy [2]
Bowtie 1[5] (default settings)	470	9.3%	9.2%	28.8%
Bowtie 2[6] (default settings)	2,226	44.0%	43.2%	84.0%
Bowtie 2 (very-sensitive-local)	4,320	85.4%	74.7%	96.5%
SOAP2[7] (default settings)	1,316	26.0%	26.0%	48.4%
BWA[8] (default settings)	2,878	56.9%	53.1%	89.0%
SMALT[9] (default settings)	4,633	91.6%	89.6%	96.5%
Geneious[10] (single iteration, default sensitivity)	4,543	89.8%	85.6%	97.1%
Geneious (single iteration, highest sensitivity)	5,060	100.0%	96.1%	99.7%
Geneious (default settings)	5,060	100.0%	100.0%	100.0%

5 [Langmead et al, 2009]

6 [Langmead and Salzberg, 2012]

7 [Li et al., 2009a]

8 [Li and Durbin, 2009]

9 http://www.sanger.ac.uk/resources/software/smalt/

10 http://www.geneious.com/

Table 1: Quality comparison of mappers on Illumina HiSeq data

The following table provides a more graphical representation of the above results by displaying a coverage graph spanning the length of the *yghJ* gene.

Algorithm	Coverage
Bowtie 1 (default)	
SOAP2 (default)	
BWA (default)	
Bowtie 2 (default)	
Bowtie 2 (very sensitive)	
SMALT (default)	
Geneious (default)	

Figure 5: Graphical coverage plots in Geneious for each mapping algorithm

The following images provide a closer look at the alignment produced by some of the best mappers around two INDEL regions. In the first region you can see that Geneious will make reads span the three base pair gap even if it means there is only a single base pair on one side of the gap. Other algorithms don't span the gap near the end of reads or span it differently for different reads.

Figure 6: Region 1. Geneious

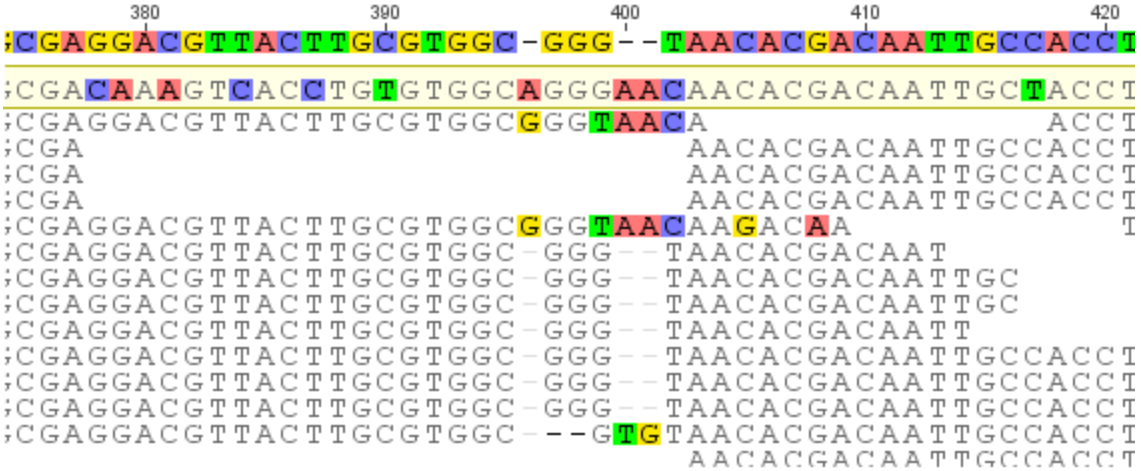


Figure 7: Region 1. Bowtie 2 - Very Sensitive

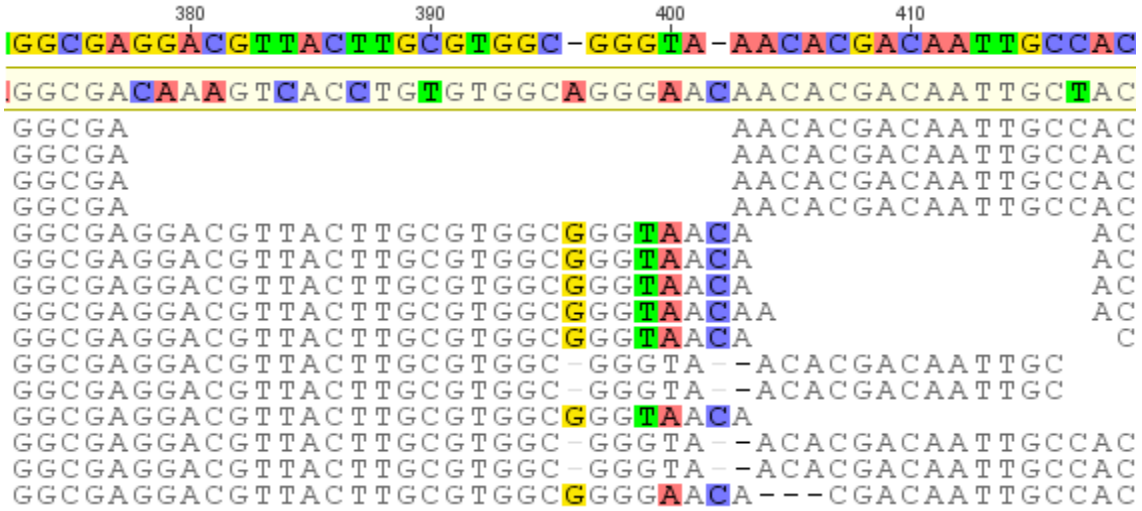


Figure 8: Region 1. SMALT

In the following region, all reads in the Geneious alignment correctly aligned the three base pair insertion including reads that terminate inside the insertion. Bowtie 2 fails to map any reads spanning this insertion even with its most sensitive settings. SMALT manages to span the region, but does so incorrectly and inconsistently.

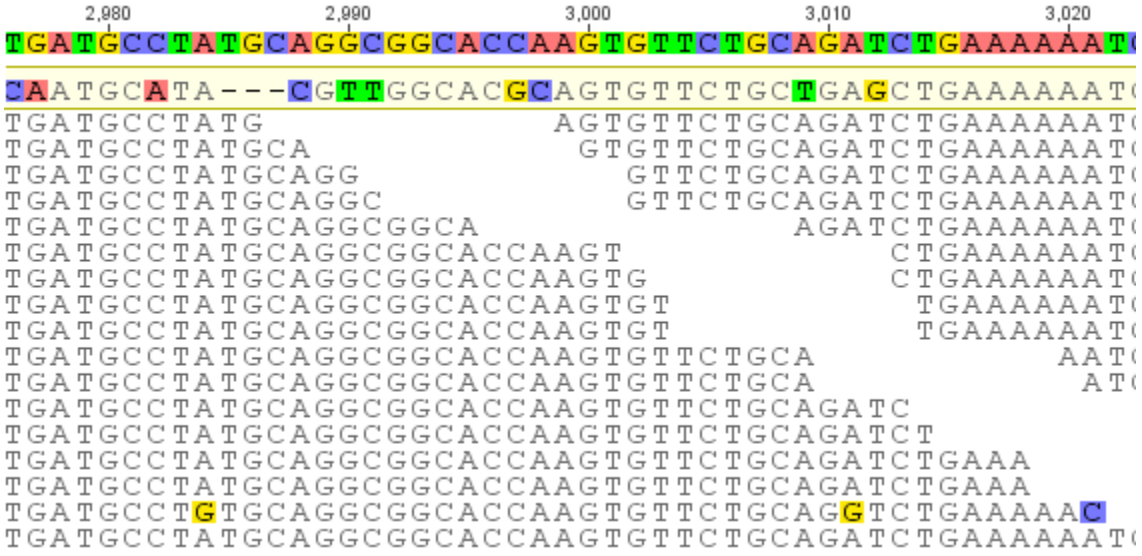


Figure 9: Region 2. Geneious

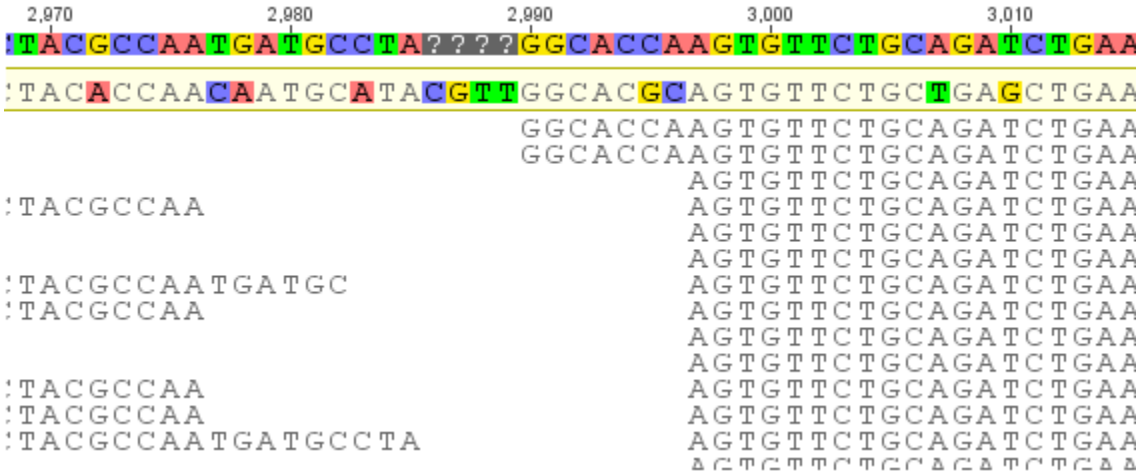


Figure 10: Region 2. Bowtie 2 - Very Sensitive

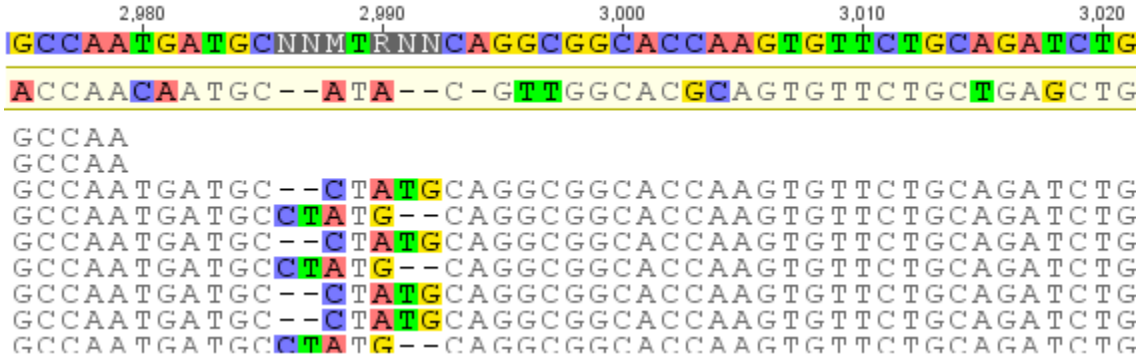


Figure 11: Region2. SMALT

The same process as outlined above for Illumina data was repeated using a whole genome Ion Torrent data set containing 2,460,975 of reads (SRR515927[11]). The reads which mapped to the *yghJ* gene created a subset of 2,536 reads, however, due to low coverage at the ends of the *yghJ* gene, the consensus sequence can't be called with accuracy and confidence. To improve the quality of the consensus calling, those reads that overlap the ends were also included and trimmed to the gene, keeping those with at least 80 bp. 80 bp was chosen since that is the minimum length read that fully intersects the *yghJ* gene. This Ion Torrent data contains a relatively high frequency of INDEL errors, so identifying the correct alignment (from a set of possible alignments) of each read to the consensus sequence isn't always possible, so the **'**% Mapped & correctly aligned to consensus' column is not included in the following table.

11 http://www.ncbi.nlm.nih.gov/sra/SRR515927

Algorithm	# Mapped	% Mapped	Consensus Accuracy
Bowtie1	4	0.2%	3.6%
SOAP2	3	0.1%	3.6%
Bowtie 2 (default settings)	1,090	43.0%	84.3%
Bowtie 2 (very sensitive)	2,028	80.0%	99.2%
BWA	42	1.7%	21.3%
SMALT	2,158	85.1%	99.3%
Geneious (single iteration, default sensitivity)	2,306	90.9%	99.6%
Geneious (single iteration, highest sensitivity)	2,532	99.8%	99.7%
Geneious (default settings)	2,535	99.96%	100.0%

Table 2: Quality comparison of mappers on Ion Torrent data

Performance Comparison

Mapping times are fairly unimportant relative to the importance of obtaining correct results, but they are still of interest. The above results (which focused only on quality) were for a single gene and were generated almost instantly for all algorithms. So in order to be evaluate

performance, all 5,411,112 reads from the original Illumina dataset were mapped to *E. coli IAI1* with default settings. For downstream visualization and analysis, read mappings usually need to be sorted. The Geneious Read Mapper automatically sorts and indexes results as part of its mapping process so for those mappers that don't perform their own sorting, the results were converted to BAM format and sorted using SAMtools 0.1.18[12].

Algorithm	Ubuntu 12 Duration (minutes:seconds) (Mapping + Sorting[13])	Windows 7 Duration (minutes:seconds)
Bowtie 1	3:39 (1:00 + 2:39)	5:07 (2:38 + 2:29)
Bowtie 2	6:42 (3:50 + 2:52)	
BWA	5:50 (3:02 + 2:48)	
SOAP2	7:11 (1:41 + 3:22[14] + 2:08)	
SMALT	3:19 (0:44 + 2:25)	
Geneious (single iteration)	1:31	1:22
Geneious (default settings - 5 iterations)	5:34	4:33

Table 3: Performance comparison of mappers on Illumina HiSeq data

All comparisons were run on an Intel Core i7-2600 3.4GHz (four physical cores, eight logical cores) with 16 GB RAM. All durations are mean durations from at least two runs, and are provided only to give an approximate indication of performance of each algorithm. Actual run times may vary from run to run, even on the same hardware and data.

Discussion

High fidelity mapping such as that demonstrated by the Geneious Read Mapper in version 6.0 is critical to ensure downstream analyses can be performed with confidence in terms of minimizing

12 [Li et al., 2009b]

13 Some mappers have a shorter sorting duration because they map fewer reads thus have fewer reads to sort.

14 SOAP2 doesn't produce a standard file format, so this is the duration to convert the results to SAM format using soap2sam.pl

error rates and maximizing repeatability of results across multiple input samples. One such analysis is variant calling between a sample dataset and a known reference, which can be performed scalably on high throughput data for multi-exon amplicon panels or whole genomes and on multiple data sets simultaneously. The output of the variant analyses is a variant report for all differences between the sample and the reference, specifically identifying various types of SNPs and INDELs.

Variant reports can be used in disease diagnostics where a researcher is looking for any variants from among a set of known disease causing variants to assign a disease condition and prognosis for a patient based on known metadata. Alternatively, variant reports can be used to identify candidate causal variants from among unknown variants by additional filtering on metadata fields, and in tandem with tools such as SIFT and PolyPhen the effects (severity, loss-of-function, gain-of-function etc.) of those variants can also be determined. These variant reports are an ideal output for clinical analysts in CLIA certified laboratories that need to orthogonally evaluate variants and fundamental researchers attempting to identify novel causative relationships between genotype and phenotype.

These results demonstrate that the Geneious Read Mapper in version 6.0 produces reliable and accurate alignments through regions of relatively low identity (89%) where two major types of polymorphisms, SNPs and INDELs, are present using two of the major sequencing technologies, Illumina and Ion Torrent. In comparison to other mappers, the default settings of the Geneious Read Mapper give far superior results compared with existing read mappers and in terms of speed is not dissimilar from the fastest mappers available. Geneious also integrates variant calling making it an ideal integrated suite of tools for diagnostic applications that require confirmation by clinical analysts in CLIA certified laboratories or exploratory research of novel genotype-phenotype associations.

References

Kearse, M., Moir R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P.L., and Drummond, A.J. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* (2012) **28** (12): 1647-1649

Holtgrewe, M., Emde, A.-K., Weese, D., and Reinert, K. (2011). A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics*, **12**:210.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**:R25

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**(14):1754-1760.

Li, R., Yu, C., Li, Y., Lam, T-W., Yiu, S-M., Kristiansen, K. and Wang J. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**(15): 1966–1967.

Li H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map (SAM) Format and SAMtools. **25**(16): 2078–2079.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**(1):195-197.