DNASTAR's SeqMan NGen vs. Four Alternative Pipelines: Variant Detection Comparison Using Illumina Data from NA12878

To obtain reliable variant results, the accuracy of sequence alignment, consensus calling, and variant detection is of paramount importance. Throughout its history, DNASTAR has emphasized the development of exceptionally accurate software, ensuring that users will obtain the highest quality results. To assess the accuracy of DNASTAR's next-generation sequence aligner and variant caller, we compared whole exome alignment results from SeqMan NGen 12.2 to those from four alternative pipelines:

- The Burrows-Wheeler Aligner (BWA) read-mapper in combination with the Broad Institute's Genome Analysis Toolkit (GATK) Unified Genotyper variant caller
- The BWA read-mapper in combination with the GATK Haplotype Base Caller
- CLC Bio's Genomics Workbench 8.0
- Geneious 8.1

Our results demonstrate that SeqMan NGen 12.2 has fewer false negatives and has better sensitivity compared to each of the other four workflows tested. SeqMan NGen 12.2 also aligns exome data and performs variant calling an average of 3.5 times faster than the alternative pipelines.

Input Data

All software pipelines used a common set of input data derived from the HapMap/1000 CEU female, NA12878. Through the Genome in a Bottle Consortium (GIAB), the National Institute of Standards and Technology (NIST) has developed a highly accurate and well-characterized set of genome-wide reference materials for NA12878, including BED and VCF files of high-quality sequence regions and variant calls, respectively.

The GIAB call sets were built from the integration of eleven NA12878 whole human genome data sets and three exome data sets, generated across five sequencing platforms to eliminate bias from any single platform. These data can be used as a benchmark when assessing variant call accuracy.

Input data used for this comparison are described below:

- The human genome reference sequence GRCh37 (hg19).
- Illumina paired-end read exome data derived from NA12878, from the Garvan Institute of Medical Research.
- A BED file detailing the targeted regions of interest for the data set. This BED file was constructed by intersecting the GIAB (version 2.19) high confidence region BED file for NA12878 with the appropriate exome target region BED file derived from the capture manifest file. Resulting continuous sequence fragments less than 20 bases in length were not used.
- The VCF file containing the GIAB high confidence variant calls for the NA12878 genome was intersected with the BED file created above. This removes variants that are not contained in the intervals corresponding to the data set's capture region, which would otherwise be scored as "negative" calls.

We thank the researchers for their generosity in making these data publicly available.

Software Workflows

All alignment and variant detection were performed on the same 64-bit Mac Pro 6 workstation (with OS X 10.10) using each application's default settings for high-sensitivity alignment. In order to simplify timing comparisons, only one job was run at a time, with no other applications running. For each experiment, an individual data set was aligned against the entire human genome reference sequence, GRCh37 (hg19).

- DNASTAR's SeqMan NGen algorithm analyzes fully gapped alignments in-stream using a modified version of the MAQ variant caller[2] to produce variant and reference call files for each position in the intersected BED file for that experiment.

- The BWA+GATK workflows used BWA[3] for the alignment stage. Alignment results were then processed with the GATK RealignerTargetCreator and GATK IndelRealigner, followed by variant detection with either the GATK Unified Genotyper or the GATK Haplotype Base Caller. The workflow utilizing BWA and the Unified Genotyper is based on the Illumina MiSeq Reporter variant detection pipeline. Both workflows use the SAMtools[4] utility program for various procedures such as converting alignments from SAM to BAM format, removing duplicates, and sorting and indexing BAM files. (The equivalent functionality is built directly into SeqMan NGen.) Default settings were used for all components other than the GATK variant callers. For the GATK variant callers, a call quality threshold of Q=10 was used to match the default PnotRef setting for SeqMan NGen.

- CLC Bio's Genomics Workbench requires that data be imported and assembled without variant analysis being performed. After assembly, the application's "Fixed Ploidy Variant Detection" tool was used to perform variant analysis.

- Geneious was run using the recommended method for variant detection: "Medium-Low Sensitivity" with up to 5 iterations of refinement.

All workflows utilized default filters to focus the variant analysis on positions with reasonable data support, based on 1) the probability that the called base is not the homozygous reference base and 2) a specified minimum depth of coverage (Table 1). All positions meeting the coverage requirement but not the probability requirement was classified as reference calls.

Table 1. Default Filters Used in Each Workflow

Software	Default Settings			Notes
	Minimum variant frequency	Probability variant is not homozygous reference base	Minimum depth of coverage	
DNASTAR's SeqMan NGen 12.2	15%	PNotRef ≥ 90%	20	A PNotRef ≥ 90% is equivalent to a Phred Q-score of 10.
BWA+GATK (both workflows)	N/A	Q=10	20	Used BWA for the alignment stage, followed by variant detection via either the GATK Unified Genotyper or the GATK Haplotype Base Caller. The workflow utilizing BWA and the Unified Genotyper is based on the Illumina MiSeq Reporter variant detection pipeline.
CLC Bio's Genomics Workbench 8.0	20%	≥ 90%	10	Data must be imported and aligned without variant analysis being performed. After alignment, the application's "Fixed Ploidy Variant Detection" tool was used to perform variant analysis.
Geneious 8.1	25%	10^{-6}	5	Used the recommended method for variant detection: "Medium-Low Sensitivity" with up to 5 iterations of refinement.

Calculations

Perl scripts were used to independently compare the accuracy of variant detection results (substitutions, and insertions and deletions of up to 10 bp) for a given alignment, relative to the "answer" provided by GIAB. Each position was then placed into one of four categories (Table 2, unshaded rows) and used to calculate a series of three statistical metrics (shaded rows).

Table 2. Accuracy Metrics

Column Name	Description
Sensitivity	The proportion of true positives that are correctly identified: TP / (TP + FN)
Specificity	The proportion of true negatives that are correctly identified: TN / (TN + FP)
False Discovery Rate (FDR)	The proportion of false positives among all discoveries: FP / (TP + FP)
True Positives (TP)	Called variants with a corresponding position in the GIAB VCF file
False Positives (FP)	Called variants without a corresponding position in the VCF file
True Negatives (TN)	Called reference bases without a corresponding position in the VCF file
False Negatives (FN)	Called reference bases with a corresponding position in the VCF file

Results

The accuracy of variant detection relative to the "answer" provided by GIAB was calculated for each workflow (Table 3).

Table 3. Accuracy Results for NA12878 using Garvan Exome Data

Workflow (Mapper/Variant Caller)	Sensitivity	Specificity	FDR	TP	FP	TN	FN	Elapsed Time*
DNASTAR's SeqMan NGen 12.2	**99.56%**	**99.999%**	**1.29%**	**15,272**	**200**	**24,882,436**	**67**	**1.3 hr**
BWA Mapper / GATK Unified Genotyper	99.09%	99.999%	1.08%	15,161	166	24,798,616	139	6.0 hr
BWA Mapper / GATK Haplotype Base Caller	99.14%	99.999%	0.97%	15,168	149	24,798,633	132	6.3 hr
CLC Bio's Genomics Workbench 8.0	98.18%	99.995%	7.41%	15,553	1,245	25,525,897	288	3.1 hr
Geneious 8.1	91.68%	99.995%	7.82%	14,827	1,257	25,872,236	1,346	2.9 hr

* Includes times to index genomes or build SeqMan NGen template mers.

Conclusion

The results shown in Table 3 demonstrate that SeqMan NGen is a fast, high accuracy read-mapper/variant caller for detecting SNP and indel variants in Illumina whole-exome sequencing data.

Compared to each of the other four pipelines tested, SeqMan NGen 12.2 has the highest sensitivity and reports fewer false negatives. SeqMan NGen also has the fastest alignment and variant-calling time of any pipeline, averaging 3.5 times faster than the other four workflows.

Other benefits to using SeqMan NGen include its Graphical User Interface (GUI), its availability on multiple platforms (Linux, Windows and Macintosh), and the ease of installing and using the software. With SeqMan NGen, jobs can be running within minutes of the initial download. In addition, SeqMan NGen works in tandem with SeqMan Pro and ArrayStar for fully integrated analysis as part of the Lasergene Genomics Suite.

To replicate these calculations, please download a free trial of Lasergene 12 and utilize the cited, publicly accessible input data.

References

1) Zook J, *et al* (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology 32, 246-251.
2) Li H, *et al* (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851-1858.
3) Li H and Durbin R (2009). Fast and Accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 15, 1754-1760.
4) Li H, *et al.* (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.