DNASTAR®
Software for life scientists

SeqMan NGen 12.2 vs. Two BWA+GATK Workflows: Variant Detection Comparison Using Illumina Data from NA12878

With the release of Lasergene 12.2, DNASTAR provides enhanced variant detection in the Lasergene Genomics Suite. To evaluate these improvements, we compared gene panel and whole exome assembly results from SeqMan NGen 12.2 to those from two industry-leading pipelines:

- The 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.

Our results demonstrate that SeqMan NGen 12.2 achieves better sensitivity and increased coverage compared to the BWA+GATK workflows. SeqMan NGen 12.2 also assembles the data and performs variant calling three times faster than the BWA+GATK pipelines.

Input Data

All three 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[1] 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 (cont.)

Input data for the software workflow comparison are described below:

1) The human genome reference sequence GRCh37 (hg19).
2) Illumina paired-end read data. Evaluations were made for both whole exome and gene panel data sets derived from NA12878.
    - Gene-panel sequence data were provided by Nephropath.
    - Whole-exome sequence data were provided by the Garvan Institute of Medical Research.
3) A BED file detailing the targeted regions of interest. BED files were constructed by intersecting the GIAB high confidence region BED file for NA12878 with the appropriate gene panel or exome target region BED file. Resulting continuous sequence fragments less than 20 bases in length were not used.
4) A GIAB VCF file containing high-confidence variant calls from the NA12878 genome.

Software Workflows

All alignment and variant detection was performed on the same Linux workstation. 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 assembled against the entire human genome reference sequence.

For SeqMan NGen 12.2, fully gapped alignments were analyzed 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. As per the default settings for this workflow, calls were filtered to a minimum PNotRef level of 90% (Q=10).

The BWA+GATK workflows used BWA[3] for the alignment stage, followed by variant detection via either the GATK Unified Genotyper or the GATK Haplotype Base Caller. Note that the workflow utilizing BWA and the Unified Genotyper is based on the Illumina MiSeq Reporter variant detection pipeline.

Software Workflows (cont.)

The BWA+GATK workflows also used 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.

For the BWA+GATK analyses, 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 PnotRef setting for SeqMan NGen.

PNotRef is the probability that the called base is not the homozygous reference base, and is used as a minimum threshold for counting positives within SeqMan Pro. The PHRED equivalents are shown below:

SeqMan Pro PNotRef	PHRED Quality Score (Q)
90%	10
99%	20
99.9%	30

Calculations

Perl scripts were used to compare variant and reference calls for a given assembly using only calls from genomic positions within the intersected high-quality target regions. This was accomplished by comparing the VCF files from GIAB with either the variant reports exported from SeqMan Pro® or the VCF files produced by GATK. Variant calls included both substitutions and small (< 10 bp) insertions and deletions. To focus the analysis on positions with reasonable data support, we applied two filters. Only positions supported by a minimum of 20 overlapping sequencing reads (i.e., coverage depth > 20) and PNotRef = 90% (Q=10) were considered as variants. All other positions meeting the depth requirement were classified as reference calls.

Calculations (cont.)

Each position was then placed into one of four categories and used to calculate a series of three statistical metrics (Table 1).

Table 1. Accuracy Metrics

Column Name	Description
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
False Discovery Rate (FDR)	The proportion of false positives among all discoveries: FP / (TP + FP)
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)

Results

For each of the three workflows, we compared the accuracy of variant detection relative to the "answer" provided by GIAB (Tables 2 and 3, next page).

The results indicate that SeqMan NGen 12.2 detects SNP and indel variants in Illumina-based sequence data with better sensitivity than do the BWA+GATK workflows. In addition, SeqMan NGen assembles more reads, resulting in more bases with sufficient coverage for analysis. This allows more true positives (TP) to be detected compared to other workflows, with only a slight increase in FDR.

In terms of assembly speed, the all-in-one SeqMan NGen workflow is three times faster than either of the multi-application BWA+GATK workflows for both gene panel and exome data.

Results (cont.)

Table 2. Accuracy Results for NA12878 Exome

Mapper	Variant Caller	TP	FP	TN	FN	FDR	Sensitivity	Specificity	Bases Covered in BED Interval	% BED Coverage	Elapsed Time*
BWA	GATK Unified Genotyper	14,831	348	24,815,483	120	2.29%	99.20%	99.999%	24,830,782	89.54%	6.7 hr
BWA	GATK Haplotype Base Caller	14,820	340	24,815,493	131	2.24%	99.12%	99.999%	24,830,782	89.54%	6.7 hr
SeqMan NGen	SeqMan NGen	14,938	375	24,905,879	45	2.45%	99.70%	99.998%	24,921,195	89.86%	2.2 hr

Table 3. Accuracy Results for NA12878 Gene Panel

Mapper	Variant Caller	TP	FP	TN	FN	FDR	Sensitivity	Specificity	Bases Covered in BED Interval	% BED Coverage	Elapsed Time*
BWA	GATK Unified Genotyper	737	17	1,027,114	7	2.25%	99.06%	99.998%	1,027,875	99.17%	1.2 hr
BWA	GATK Haplotype Base Caller	737	15	1,027,116	7	1.99%	99.06%	99.999%	1,027,875	99.17%	1.2 hr
SeqMan NGen	SeqMan NGen	744	19	1,030,056	0	2.49%	100.00%	99.998%	1,030,813	99.45%	0.4 hr

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

Conclusion

The results presented here demonstrate that SeqMan NGen is a fast, high accuracy read-mapper/variant caller for Illumina sequencing data. Compared to BWA+GATK pipelines currently in use, SeqMan NGen provides more genome coverage, identifies more true positives, and provides higher sensitivity, in one-third of the time required for the alternative pipelines.

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, assemblies 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.

All input data are publicly accessible. To replicate these calculations, please download a free trial of Lasergene 12.2 and utilize the cited input data:

- Human genome reference sequence GRCh37 (hg19)
- Garvan Institute of Medical Research whole-exome sequence data
- Intersected BED files, VCF files, and the Nephropath gene-panel sequence data can be downloaded as a single .zip archive from the DNASTAR website.

We thank Nephropath and the Garvan Institute of Medical Research for their generosity in making their data publicly available.

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.