SeqMan NGen is a High Accuracy NGS Assembler: Assessment with NA12878 Reference Materials

With Lasergene 12, DNASTAR introduces a new SNP Validation Workflow as part of the Lasergene Genomics Suite. This workflow is designed to work with a "gold standard" set of reference materials, such as those for the HapMap/1000 Genomes CEU female NA12878 developed by the National Institute of Standards and Technology (NIST) through the Genome in a Bottle Consortium (GIAB). The purpose of the workflow is to validate the efficacy of a sample processing and sequencing procedure, whether it be for a gene panel or a whole exome/genome. The output of this workflow consists of a statistical report detailing assembly sensitivity, specificity, and accuracy. Here, we use this workflow to assess the accuracy of the SeqMan NGen® assembler with NA12878 whole exome data sets. The results demonstrate that SeqMan NGen® is a highly accurate NGS assembler with variant calling accuracies typically greater than 99.8%.

Data Sets

Genome in a Bottle (GIAB) reference materials[1]

GIAB 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. This set is publicly accessible at http://www.genomeinabottle.org/ and can be used as a benchmark when assessing variant call accuracy. The set was built from the integration of 11 NA12878 whole human genome and three exome data sets generated across five sequencing platforms to eliminate bias from any single platform. The DNASTAR SNP validation control workflow was built to utilize this benchmark, and does so by using the Genome in a Bottle BED and VCF files as a control to calculate the SNP calling accuracy.

NA12878 NGS data sets

For whole exomes, we analyzed both an Illumina paired end data set produced by the Garvan Institute and available at ftp://ftp-trace.ncbi.nih.gov/giab/ftp/technical/garvan_data/, and an Ion Torrent data set available from the Ion Community website, http://ioncommunity.lifetechnologies.com.

Software Workflow

To assess SeqMan NGen® accuracy, a series of publicly available NA12878-derived NGS data sets (Table 1) using the two leading sequencing providers, Illumina and Ion Torrent, were downloaded and run through DNASTAR's SNP validation control workflow. The SeqMan NGen® project setup wizard allows users to specify the four key data files for the workflow: 1) the reference genome to be assembled against; 2) the NGS read data set(s); 3) a BED file detailing the targeted regions of interest; and 4) the GIAB and VCF file containing the high confidence variant calls. The BED file For the validation control workflow, an intersected BED file, constructed using the Genome in a Bottle high quality region BED file and the user's specific target region BED file (or Manifest file if working with Illumina data) should be used. These directives are completed in a step-by-step process using the SeqMan NGen GUI.

Software Workflow (Continued)

For each experiment, individual data sets were each assembled against the entire human genome reference sequence, GRCh37 (hg 19), using SeqMan NGen® v12.0 running on a single standard desktop computer. 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. SeqMan NGen also uses the intersected target region file for each experiment to filter GIAB's NA12878 high confidence VCF file. Both called variant and reference positions were compared to the filtered VCF file and if present, the position was tagged with a unique ID.

For accuracy calculations, variant and reference call files for a given assembly were loaded together with the associated filtered VCF file into ArrayStar®. Only positions within the intersected high quality target regions are considered. Each position is then classified into one of four categories as follows: 1) true positives, called variants with a corresponding position in the VCF file; 2) false positives, called variants without a corresponding position in the VCF file; 3) true negatives, called reference bases without a corresponding position in the VCF file; and 4) false negatives, called reference bases with a corresponding position in the VCF file. The counts of each class were then ultimately used to calculate sensitivity, specificity and balanced accuracy (Table 1). Sensitivity measures the proportion of true positives that are correctly identified, and is calculated using the following equation: true positives / (true positives + false negatives). Specificity measures the proportion of true negatives that are correctly identified, and is calculated using the following equation: true negatives / (false positives + true negatives). Finally, the overall balanced accuracy is derived from the following: (sensitivity + specificity) / 2.

Results

Table 1. Accuracy Results*

Data	Sensitivity	Specificity	Balanced Accuracy
Illumina exome	99.655%	99.998%	99.826%
Ion Torrent exome	97.855%	99.723%	98.789%

*Given a minimum depth of coverage of 20 and PNotRef ≥ .90. PNotRef is the probability that the called base is not the homozygous reference base. This value is used as a minimum threshold for counting positives in the validation control workflow.

Conclusion

The results indicate that DNASTAR software is greater than 99.8% accurate when predicting SNPs using the DNASTAR validation control pipeline and Illumina exome data, with the understanding that differences between sequencing platforms can impact overall accuracy calculations.

The data sets described here are all publicly accessible. To replicate these calculations, please download a free trial of Lasergene 12 and utilize the cited data sets in this overview.

References

1. Zook, J. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology 32, 246-251. (2014)
2. Li, H. et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851-1858. (2008)