SeqMan NGen® Assembly of Ion Torrent Exome Data With and Without Heterozygous Indel Removal: a Comparison Using NA12878 Reference Materials

With the release of Lasergene 12, DNASTAR introduced a new Validation Control workflow as part of the Lasergene Genomics Suite. Validation can be applied to a gene panel, a whole exome or a genome. As input, this workflow utilizes "gold standard" reference materials to validate the efficacy of sample processing and sequencing procedures. The workflow output consists of a statistical report detailing assembly sensitivity, specificity, and accuracy.

DNASTAR's Validation Control workflow was used to test the accuracy of Ion Torrent data assembly in SeqMan NGen 12.2. Ion Torrent data is known to have an increased error rate in homopolymeric runs[1,2], reducing the accuracy of base calling of heterozygous insertions or deletions in these regions.

In order to address this concern, the "SNP Options" dialog of the SeqMan NGen 12.2 wizard allows users to specify automatic removal of heterozygous indels. When the setting is turned "on," a Bayesian probability model is used to remove heterozygous indels while retaining homozygous variants. The accuracy of the SeqMan NGen 12.2 assembler with this setting turned on or off was tested using NA12878 whole exome Ion Torrent data.

Our findings demonstrate that automatically removing heterozygous indels in Ion Torrent data produces the most accurate assembly and variant calling. We therefore recommend that most SeqMan NGen users specify automated heterozygous indel removal when assembling Ion Torrent data. SeqMan NGen users interested in finding the largest number of True Positive results may want to consider turning off the automated heterozygous indel removal. However, these users will need to be aware that an increased number of False Positives will also be identified.

Input Data

Our analysis utilized 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.

Input Data (cont.)

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 are described below:

1) The human genome reference sequence GRCh37 (hg19).
2) A whole exome Ion Torrent data set derived from NA12878 and obtained from the Ion Community website.
3) A BED file detailing the targeted regions of interest. First, manifest files were converted to BED files. These were then intersected with the GIAB high confidence BED file for NA12878 to produce the new BED file used in this study.

   Resulting continuous sequence fragments less than 20 bases in length were not used.
4) A GIAB VCF file containing high-confidence variant calls.

Alignment and Variant Detection

To assess the accuracy of DNASTAR's variant calls, a publicly available NA12878-derived NGS data set from Ion Torrent was analyzed using the Validation Control workflow.

SeqMan NGen 12.2 was used on a desktop computer to perform alignment and variant detection. When prompted by the SeqMan NGen wizard, the user added each of the four input data files (see above). In some runs, "Bayesian-based removal of heterozygous indels" was turned on in the pre-assembly SNP Options dialog. In other runs, the same setting was turned off. For each run, the samples were then assembled against the human genome template. Both Linux and Windows operating systems yielded identical results.

Fully gapped alignments were analyzed in-stream using a modified version of the MAQ variant caller[4] to produce variant and reference call files for each position in the intersected manifest file for that experiment. SeqMan NGen also used 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. As per the default settings for this workflow, calls were filtered to a minimum PNotRef level of 90%.

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. A PNotRef of 90% is equivalent to a PHRED Quality score (Q) of 10.

Calculation of Accuracy

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 the variant reports exported from SeqMan Pro®. 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% were considered as variants. All other positions meeting the depth requirement were classified as reference calls. Positions not meeting the minimum coverage depth were ignored and did not contribute to the count of sites with coverage.

Each position was then placed into one of four categories and used to calculate a series of 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 assembly, we compared the accuracy of variant detection relative to the "answer" provided by GIAB (Table 2).

Table 2. Accuracy results for Ion Torrent exome data assembled with SeqMan NGen 12.2

Heterozygous Indels Removed?	TP	TN	FP	FN	FDR	Sensitivity	Specificity	Bases Covered in BED Intervals	% BED Coverage
No	20,614	32,838,281	7,869	890	27.63%	95.86%	99.976%	32,867,654	97.239%
Yes	20,562	32,846,208	997	946	4.62%	95.60%	99.997%	32,868,713	97.242%
% Change	-0.25%	+0.02%	-87.33%	+6.29%	-83.28%	-0.27%	+0.02%	0.00%	0.00%

As shown in Table 2, removing heterozygous indels did increase the FN rate by 6.29%. However, the increase in FN was more than offset by the desirable decrease in FP (-87.33%) and concomitant decrease in the FDR (-83.28%). Specificity did not change significantly (+0.02%), but there was a modest decline in Sensitivity (-0.26%). The removal of heterozygous indels had an insignificant impact on other accuracy metrics (≤ 0.25%).

The large decrease in FDR confirms that heterozygous calls in runs have a negative impact on the accurate assembly of Ion Torrent sequence data, and that this impact can be mitigated by specifying automatic heterozygous indel removal in SeqMan NGen 12.2.

Conclusion

In the case of Ion Torrent exome data, our results indicate that allowing automatic removal of heterozygous indels increases accuracy by greatly decreasing the False Discovery Rate (FDR).

For cases in which it is critical to identify the largest possible number of true positives, users may want to consider this alternative setting that removes heterozygous indels, although false positives will also increase with this setting.

Conclusion (cont.)

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)
- Ion Torrent data, intersected BED and VCF files can be downloaded as a single .zip archive.

References

1) Ross M, *et al* (2013). Characterizing and measuring bias in sequence data. Genome Biology 2013 14:R51.

2) Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013). Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. PloS Comput Biol 9(4): e1003031. doi:10.1371/journal.pcbi.1003031.

3) 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.

4) Li H, *et al* (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851-1858.