

SeqMan NGen 12.2 vs. CLC Bio Genomics Workbench 8.0: Variant Detection Comparison Using 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 for Illumina and Ion Torrent data, we compared whole exome results from DNASTAR's SeqMan NGen 12.2 with those from CLC Bio's Genomics Workbench 8.0, another commercial pipeline with a variant detection workflow.

Our results for Illumina exome data demonstrate that SeqMan NGen 12.2 has higher sensitivity, finds fewer false positives and false negatives, and has a lower false discovery rate compared to CLC Bio's Genomics Workbench 8.0. For Ion Torrent exome data, SeqMan NGen has better specificity, reports fewer false positives, and has a lower false discovery rate. For both types of data, SeqMan NGen performs variant analyses from 1.5-4.0 times faster than Genomics Workbench.

Input Data

Both 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 materials¹ 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 study are described below:

- The human genome reference sequence [GRCh37 \(hg19\)](#).
- Illumina and Ion Torrent paired-end read data. Evaluations were made for several whole exome data sets derived from NA12878 (Table 1).

Table 1. Data Sets

| Nickname | Data Source | Download Link |
|-------------|---|--|
| Garvan | Illumina exome data from the Garvan Institute of Medical Research | Download |
| ARUP | Illumina exome data from NCBI's Short Read Archive data set " ARUP exome sequencing of human HapMap individual NA12878 ," submitted by ARUP Laboratories (Salt Lake City, UT). | Download |
| Icahn | NCBI's Short Read Archive Illumina exome data set " High-coverage whole exome sequencing of CEPH/UTAH female individual (HapMap: NA12878) ," submitted by the Icahn School of Medicine at Mount Sinai (New York, NY). | Download |
| Ion Torrent | Ion Torrent "Ion Proton Target-Seq" exome data obtained from the Ion Community website . | Download (zip archive contains sequence data, intersected BED and VCF files) |

- A BED file for each data set, detailing the targeted regions of interest. Each 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.
- For each data set, the VCF file containing the GIAB high confidence variant calls for the NA12878 genome was intersected with the corresponding BED file created above. This removes variants that are not contained in the intervals for the data set's capture region, and 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 was 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 default settings included Bayesian-based removal of heterozygous indels for the Ion Torrent data set. The SeqMan NGen algorithm analyzes fully gapped alignments in-stream using a modified version of the MAQ variant caller² to produce variant and reference call files for each position in the intersected BED file for that experiment.
- CLC Bio's Genomics Workbench, by contrast, requires that data 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.

Both 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 2). All positions meeting the coverage requirement but not the probability requirement were classified as reference calls.

Table 2. Default Filters Used in Each Workflow

| Software | Default Settings | | |
|----------------------------------|---------------------------|--|---------------------------|
| | Minimum variant frequency | Probability variant is not homozygous reference base | Minimum depth of coverage |
| DNASTAR's SeqMan NGen 12.2 | 15% | Minimum PNotRef = 90% | 20 |
| CLC Bio's Genomics Workbench 8.0 | 20% | Required Variant Probability = 90% | 10 |

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 3, unshaded rows) and used to calculate a series of three statistical metrics (shaded rows).

Table 3. 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 (Tables 4 and 5).

Table 4. Accuracy Results for NA12878 Illumina Exome Data

| Data Set | Workflow | Sensitivity | Specificity | FDR | TP | FP | TN | FN | Elapsed Time* |
|----------|----------------------------------|-------------|-------------|--------|--------|-------|------------|-----|---------------|
| Garvan | DNASTAR's SeqMan NGen 12.2 | 99.56% | 99.999% | 1.29% | 15,272 | 200 | 24,882,436 | 67 | 1.3 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 |
| ARUP | DNASTAR's SeqMan NGen 12.2 | 99.51% | 99.9997% | 0.53% | 15,009 | 80 | 24,532,535 | 74 | 0.5 hr |
| | CLC Bio's Genomics Workbench 8.0 | 97.72% | 99.998% | 2.90% | 15,037 | 449 | 25,335,711 | 351 | 2.0 hr |
| Icahn | DNASTAR's SeqMan NGen 12.2 | 99.49% | 99.999% | 1.14% | 36,464 | 422 | 43,817,958 | 186 | 1.1 hr |
| | CLC Bio's Genomics Workbench 8.0 | 98.07% | 99.988% | 12.68% | 37,354 | 5,424 | 45,109,635 | 735 | 2.9 hr |

Table 5. Accuracy Results for NA12878 Ion Torrent Exome Data

| Data Set | Workflow | Sensitivity | Specificity | FDR | TP | FP | TN | FN | Elapsed Time* |
|-----------------------|----------------------------------|-------------|-------------|--------|--------|--------|------------|-------|---------------|
| Ion Proton Target-Seq | DNASTAR's SeqMan NGen 12.2 | 93.93% | 99.997% | 3.71% | 20,839 | 803 | 32,086,373 | 1,346 | 3.0 hr |
| | CLC Bio's Genomics Workbench 8.0 | 95.72% | 99.920% | 54.26% | 22,211 | 26,353 | 33,073,897 | 992 | 4.4 hr |

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

Conclusion

The results from Tables 4 and 5 demonstrate that DNASTAR's SeqMan NGen 12.2 is a fast, high accuracy read-mapper/variant caller for Illumina and Ion Torrent sequencing data.

- For each of the three sets of Illumina whole exome data, SeqMan NGen has higher sensitivity, finds fewer false positives and false negatives, and has a lower false discovery rate compared to CLC Bio's Genomics Workbench 8.0. SeqMan NGen is also an average of 3 times faster than Genomics Workbench at performing variant analysis for Illumina whole exome data.
- For Ion Torrent whole exome data, SeqMan NGen is 1.5 times faster than CLC Bio's Genomics Workbench 8.0, and also has better specificity, reports fewer false positives, and has a lower false discovery rate.

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](#).