Performance and Precision Characterization of the Bina Genomic Analysis Platform

Overview

This document summarizes the performance and fidelity of the Bina platform for human genome sequence analysis. The Bina platform features highly optimized software co-designed with custom hardware appliances to greatly accelerate genome sequence analysis while simultaneously achieving best-in-class accuracy. The goal of this document is to characterize the Bina platform on the standard aspects of genome analysis, from alignment through variation calling. The platform includes the most commonly used open-source tools for sequence alignment, small variant calling and structural variation detection. It also includes a proprietary aligner with key performance advantages, infrastructure for efficient pipelining and parallelization of standard workflows, and a user interface that facilitates the management, QC and tracking of sequence analyses.

Bina Genomic Analysis Pipelines

The Bina Genomic Analysis Platform provides integrated pipelines of best-in-class tools for analyzing whole-genome sequencing (WGS) and whole-exome sequencing (WES) data. The platform currently supports sequence data generated on the Illumina HiSeq and MiSeq sequencers. It is based on the best practices for secondary sequence analysis recommended by the Broad Institute, and incorporates the following steps and tools:

1. Alignment

Alignment against the human reference genome (for WGS or WES) may be performed either using the popular BWA aligner ([1], version 0.7.5a) or with the proprietary Bina aligner [4] – an optimized hash-based aligner featuring superior speed and high-fidelity indel detection.

2. Sorting

Sorting is performed using Bina's custom in-memory sorter. When the Bina aligner is used sorting takes place concurrently with alignment, resulting in further savings in turnaround time.

3. Duplicate marking

Duplicate reads are marked for subsequent exclusion in variation calling using Bina's custom sorter. Such reads can distort observed allele ratios and therefore impact the accuracy of variant calls.

4. Indel realignment

The realignment of reads in the vicinity of indels increases consistency across multiple alignments and improves the accuracy of variant calling. This step employs the GATK toolkit ([2,3], version 1.6.0).

5. Base quality score recalibration

This step recalibrates base quality scores using the BQSR capability contained in the GATK toolkit. The recalibration yields more accurate estimates of quality that attempt to correct for variations in quality with sequence context and machine cycle.

6. SNV and indel calling

The UnifiedGenotyper within the GATK toolkit is used for SNV and indel detection and scoring.

7. Variant quality score recalibration (VQSR)

This module, also within the GATK toolkit, is used to associate each called SNV or indel with a calibrated error probability, estimated from a Gaussian mixture model applied to a set of designated true sites.

8. Copy number variations (CNVs) and structural variations (SVs) (WGS only)

Copy Number Variation Detection is performed using the widely used tool CNVNator ([5], version 0.2.7), which proceeds based on an analysis of the read depth profile, correcting for GC content and making several other refinements to improve accuracy. Structural variations are also detected using BreakDancer ([6], version 1.1.2), which identifies anomalous paired-end alignments and jointly analyzes them. Currently the platform is limited to reporting intrachromosomal events detected by these tools. Support for interchromosomal events is coming soon.

Data for Benchmarking

- Three lanes of 2x100bp paired-end Illumina reads for the NA12878 genome, a cell line that has been studied extensively using multiple technologies and which is available from the Coriell repository. After alignment, these data correspond to an average coverage depth of 37x over the human reference genome.
- Publicly available exome data based on the Agilent (v2) selection kit. The entire data set features 228 million reads, corresponding to 17Gb of reads data.

Bina Platform Specifics for Benchmarking

The paired-end whole-genome sequence data were processed on a 4-node, 64-core Bina appliance with two pipelines:

1. Bina Aligner + GATK

   This pipeline features an improved version of the Bina aligner for high-quality, long NGS sequence reads [4] alongside a best-practices workflow based on the GATK UnifiedGenotyper. The pipeline features Bina's in-memory sorter, which is used concurrently with alignment to minimize turnaround time. In addition, the pipeline incorporates publicly available tools for structural variation detection: CNVNator for copy number variation detection and BreakDancer for structural variation detection from paired-end sequence data.

2. Accelerated BWA + GATK

   This pipeline features an optimized, parallelized version of the BWA aligner and a best-practices workflow based on the GATK UnifiedGenotyper. The pipeline also features Bina's sorter, but not concurrently with alignment.

For whole-exome analysis, the pipeline is identical to that for whole genomes for alignment and variation calling, but does not include CNV and structural variations at this time.

Hardware Configuration for Performance Tests

All computations were performed (except where indicated explicitly) on a four-node, 2U, 64-core configuration of the Bina appliance with an aggregate memory of 500 GB. However, the infrastructure is designed to scale seamlessly with larger configurations (i.e. two 4-node 2U units can function as a single cluster for operation and management).

Performance on Illumina Paired-end Reads from NA12878 (WGS and WES)

Accurate analysis of WGS data set completes in a few hours

The times taken to complete various analysis steps - alignment, sorting, indel realignment, variant calling, and variant quality score recalibration - are reported in Table 1 for the 37x whole-genome dataset of NA12878 described above. The overall turnaround time for the Bina pipeline on a single 2U-format Bina appliance was 3.9 hours starting from FASTQ files and ending in a merged, recalibrated VCF file for all the variants.

The performance of the BWA aligner has also been accelerated on the Bina platform. Using BWA and GATK, the end-to-end time of the pipeline was 5.7 hours. Note that the latency (marginal turnaround time) of the sorting stage is reduced considerably for the Bina pipeline by making it concurrent with alignment.

Table 1. Wall clock time for analysis steps on a 37x human WGS data set

Pipeline	Alignment, Sorting, Duplicate Marking	Indel Realignment	Base Quality Recalibration	Variant Calling	Variant Quality Recalibration	Total Time
BWA + GATK	3.4 h	0.26 h	0.92 h	0.55 h	0.55 h	5.9 h
Bina Aligner + GATK	1.4 h	0.40 h	0.94 h	0.59 h	0.55 h	3.9 h

Single 2U appliance can analyze more than 40 high-coverage exomes in one day

Table 2 summarizes the results for a whole-exome analysis using BWA and GATK using different run configurations. The number of nodes used can be varied to trade off turnaround time with throughput. Parallelization across multiple nodes results in a shorter turnaround time for an individual job (less than an hour using 4 nodes) but a lower overall throughput. Using single-node runs, a throughput of greater than 40 high-coverage (17 Gb) exomes per day can be achieved. The throughput for smaller exome data sets (50x – 100x) would be higher.

Table 2. Time for BWA alignment and GATK variation calling for a 17 Gb human WES data set on a Bina Appliance

# of Nodes	Alignment, Sorting, Duplicate Marking	Indel Realignment	Base Quality Recalibration	Variant Calling	Turnaround Time	Throughput for 4 nodes (exomes/day)
1	1.5 h	0.08 h	0.52 h	0.12 h	2.2 h	43.6
4	0.62 h	0.03 h	0.16 h	0.09 h	0.9 h	26.7

BWA+GATK pipeline is exactly reproducible even when parallelized

When run on the same number of nodes, two identically specified BWA+GATK runs on the same input data generate identical alignment and variation call results, regardless of the number of nodes utilized. This is demonstrated in Table 3 for a 17 Gb exome run on 8 nodes (2 x 2U Bina appliances). This reproducibility is of high value when reanalyzing data.

Table 3. Results for BWA+GATK runs are exactly reproducible when run on the same number of nodes

BWA+GATK on 8 nodes	Alignments		SNPs	SNP het/hom	SNPs in dbSNP 135	Indels	Indels het/hom	Indels in dbSNP 135
	Unique	Mulitple						
Run #1	84.99%	5.61%	33,367	1.5975	97.09%	1,874	1.2994	94.56%
Run #2	84.99%	5.61%	33,367	1.5975	97.09%	1,874	1.2994	94.56%

When two runs on the same data with BWA+GATK use different numbers of compute nodes, the results are not guaranteed to be identical. However, the differences between such runs are very small and do not result in a discernible loss of accuracy. The results for a whole-exome analysis on one, four and eight nodes are shown in Table 4. Note that the three sets of results differ by only 2 indels and fewer than 10 SNVs. Parallelization at this level for better turnaround time does not result in a detectable loss of accuracy in the results.

Table 4. Parallelization of BWA+GATK has negligible impact on results and accuracy

# of Nodes	Alignments		SNPs	SNP het/hom	SNPs in dbSNPs 135	Indels	Indel het/hom	Indels in dbSNP 135
	Unique	Multiple						
1	84.88%	5.61%	33,369	1.595	97.06%	1,874	1.299	94.6%
4	84.88%	5.61%	33,360	1.597	97.06%	1,872	1.297	94.6%
8	84.88%	5.61%	33,367	1.598	97.09%	1,874	1.299	94.6%

The Bina Aligner coupled with the GATK UnifiedGenotyper yields higher sensitivity than BWA

Table 5 shows the alignment results and SNPs called for two aligners – BWA and the Bina aligner. Both are paired with the GATK UnifiedGenotyper for variation calling, followed by VQSR. The Bina aligner has a slightly higher mapping yield. More importantly, the pipeline with the Bina aligner yields about 2% more SNP variations, with very similar rates of corroboration in dbSNP (99.8%) and transition/transversion ratio (2.12) as the BWA-based pipeline.

Table 5. Impact of Aligner Choice on SNV calls with GATK UnifiedGenotyper (WGS, post-VQSR)

Pipeline	Alignments		SNPs	SNP het/hom	SNPs in db-SNP 135	SNPs Ti/Tv	Increase in Reported Indels
	Unique	Mulitple					
BWA + GATK	92.6%	3.6%	3,151,805	1.49	99.8%	2.13	0%
Bina + GATK	93.4%	5.5%	3,221,432	1.47	99.8%	2.12	+2%

Table 6 extends the comparison to indel calls. The pipeline based on the Bina aligner results in significantly higher sensitivity: 95,770 additional indel calls, with 15% (>75,000 over the genome) increase in indels corroborated by dbSNP.

Table 6. The Bina Aligner yields more sensitive indel detection with the GATK UnifiedGenotyper than BWA (WGS, post-VQSR)

Pipeline	Indels	Indel het/hom	Indels in dbSNP 135	Increase in Reported Indels
BWA + GATK	572,548	1.23	515,328 (90.0%)	0%
Bina + GATK	668,318	1.28	592,772 (88.7%)	+15%

Summary

In this document we characterize the Bina genome analysis pipeline on Illumina data. We report the turnaround time, throughput and accuracy of the pipeline on two datasets generated from the well-studied CEU cell line NA12878 - a 37x whole-genome data set, and a 17 Gb data set corresponding to the Agilent SureSelect Human All Exon kit (v2). We describe our results with a pipeline based on tools that are commonly used for sequence analysis: the BWA aligner and GATK UnifiedGenotyper. We also report results for the Bina aligner [4], a proprietary fast and accurate hash-based aligner for long, accurate sequence reads. Here are the key findings:

- An analysis of a 37x whole-genome dataset going from reads to calibrated variation calls took less than 6 hours on a 4-node, 2U form factor Bina appliance.
- An analysis of a 17 Gb exome completed on a single node of the Bina appliance in a little more than two hours, indicating a throughput of > 40 high-coverage exome analyses in a day.
- Parallel computing to improve turnaround time does not have to come at the cost of reproducibility.
- There is still room to improve when it comes to widely-used tools in the scientific community. Acceleration of genomic analyses can be achieved without a loss in sensitivity or accuracy.
  - The Bina aligner provides faster turnaround than the BWA aligner.
  - At a similar false positive rate, the Bina aligner achieves better sensitivity for indel detection than BWA when paired with GATK UnifiedGenotyper.

References

1. Fast and accurate short read alignment with Burrows–Wheeler transform. Heng Li and Richard Durbin, Bioinformatics, 25, 1754–1760 (2009).
2. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Mark A DePristo, Eric Banks, Ryan Poplin, et al.. Nature Genetics, 43, 491–498 (2011).
3. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Aaron McKenna, Matthew Hanna, Eric Banks, et al.. Genome Res. 20, 1297-1303 (2010).
4. Fast and accurate read alignment for resequencing. John C. Mu, Hui Jiang, Amirhossein Kiani et al., Bioinformatics, 28 (18), 2366-2373 (2012).
5. CNVNator: an approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. Abyzov A, Urban AE, Snyder M, Gerstein M. Genome Res. 21(6), 974 - 984 (2011).
6. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Ken Chen, John W. Wallis et al.. Nature Methods 6, 677 - 681 (2009).