Computational Performance of a Germline Variant Calling Pipeline for Next Generation Sequencing

Jie Liu¹, Xiaotian Wu¹, Kai Zhang², Bing Liu², Renyi Bao², Xiao Chen², Yiran Cai², Yiming Shen², Xinjun He², Jun Yan², and Weixing Ji¹

¹School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China
²Nanjing YiGenCloud Institute, Nanjing, China

Abstract—With the booming of next generation sequencing technology and its implementation in clinical practice and life science research, the need for faster and more efficient data analysis methods becomes pressing in the field of sequencing. Here we report on the evaluation of an optimized germline mutation calling pipeline, HummingBird, by assessing its performance against the widely accepted BW A-GATK pipeline. We found that the HummingBird pipeline can significantly reduce the running time of the primary data analysis for whole genome sequencing and whole exome sequencing while without significantly sacrificing the variant calling accuracy. Thus, we conclude that expansion of such software usage will help to improve the primary data analysis efficiency for next generation sequencing.

Index Terms—Next generation sequencing, variant calling, germline mutation, genomics

I. INTRODUCTION

Next generation sequencing has revolutionarily changed the way of basic life science research and clinical practice for cancer, infectious and genetic diseases [1], [2]. Population based large scale sequencing at national level helped to collect baseline variation information for various genetic traits from healthy and susceptible individuals [3]–[6]. With rapid increase of the output from new sequencing technologies and the gradual decrease of per base sequencing cost, whole genome sequencing and whole exome sequencing have migrated from national and multi-national research projects to daily practical clinical applications especially in the field of cancer and genetic disease diagnostics [5], [7], [8]. As it is the inherited character for next generation sequencing, the initial deposition of large amount sequencing data provided not only immediate assistances needed for its requisition, but also a rich data resource for further exploration.

BW A is a widely accepted tool for aligning sequencing reads to a reference genome [9], [10]. In particularly, BW A consists of three different algorithms: BW A-backtrack, BW A-SW and BW A-MEM. BW A-MEM is the optimized algorithm for short reads mapping. Even though BW A implements thread-level parallelism, it takes a long time for large-scale datasets. To solve this problem, many researchers used different strategies such as Hadoop, Spark, MPI to accelerate BW A [11]–[15]. It contains optimized sequencing reads preparation steps including reads sorting, duplicates marking and base quality recalibration. elPrep prepares sequencing reads and produces identical results as SAMtools [19]–[21] and Picard [22]. elPrep can be run in multithread and entirely in memory, thus avoiding repeated file I/O and significantly reducing running time for sequencing reads preparation.

The Genome Analysis Toolkit (GATK) is a set of bioinformatic tools for analyzing high-throughput sequencing (HTS) and variant call format (VCF) data [14], [15]. The toolkit is well established for germline short variant discovery from whole genome and exome sequencing data. In this study, we used GATK 4.1.2.0 as the benchmark to evaluate the performance of HummingBird pipeline.

elPrep is a high-performance tool for preparing sequencing reads for variant calling in a sequencing pipeline [16]–[18]. It contains optimized sequencing reads preparation steps including reads sorting, duplicates marking and base quality recalibration. elPrep prepares sequencing reads and produces identical results as SAMtools [19]–[21] and Picard [22]. elPrep can be run in multithread and entirely in memory, thus avoiding repeated file I/O and significantly reducing running time for sequencing reads preparation.

The power of next generation sequencing is to collect as much genetic information as possible from sequenced samples; the large amount of data generated thus require computationally intensive method by professional bioinformaticians to extract the genetic information initially looked for. However, it usually takes up to a few days to process a whole genome sequencing data and costs unacceptably long time when there are a large number of samples with standard BW A [9], [10] for alignment and the Genome Analysis Toolkit (GATK) from Broad Institute for short variant calling for primary data analysis [14], [15], [23]. Thus, more reliable, efficient and reproducible sequencing data analysis tools are needed to meet the needs of a speedy variant calling [24], [25].

In this paper, we developed HummingBird pipeline which can reduce the clock time for germline variant calling without significantly sacrificing the variant calling accuracy. Here we evaluated a germline mutation calling pipeline, HummingBird pipeline, that was developed by YiduCloud (Beijing) company, and reported that HummingBird is a fast and accurate mutation calling pipeline that can be expanded for sequencing data analysis.

The rest of the paper is organized in this way: Section 2 compares BW A-GATK and HummingBird in methodology; Section 3 analyzes the results of BW A-GATK and HummingBird under different sample sets; the conclusions are given in Section 4.

Corresponding authors: Weixing Ji(jwx@bit.edu.cn) and Xinjun He(xinjun.he@yigencloud.cn).
TABLE I
INFORMATION OF DATASETS USED IN THIS STUDY

| Sample Name     | Type    | Clean Reads Files                                      | Clean Reads #   | Clean Reads File Size |
|-----------------|---------|-------------------------------------------------------|----------------|----------------------|
| NIST7035_L001  | WES     | NIST7035_TAAGGCAGA_L001_R1_001_trimmed.clean.fq.gz    | 19612287       | 1.5G                 |
|                 |         | NIST7035_TAAGGCAGA_L001_R2_001_trimmed.clean.fq.gz    | 19612287       | 1.5G                 |
| NIST7086_L001  | WES     | NIST7086_CGTACTAG_L001_R1_001_trimmed.clean.fq.gz    | 20900829       | 1.6G                 |
|                 |         | NIST7086_CGTACTAG_L001_R2_001_trimmed.clean.fq.gz    | 20900829       | 1.6G                 |
| SRR098401      | WES     | SRR098401_1.clean.fq.gz                               | 80253885       | 5.1G                 |
|                 |         | SRR098401_2.clean.fq.gz                               | 80253885       | 5.3G                 |
| SRR742200      | WES     | SRR742200_1.clean.fq.gz                               | 24576684       | 2.0G                 |
|                 |         | SRR742200_2.clean.fq.gz                               | 24576684       | 2.0G                 |
| SRR098359      | WES     | SRR098359_1.clean.fq.gz                               | 60017306       | 3.7G                 |
|                 |         | SRR098359_2.clean.fq.gz                               | 60017306       | 4.1G                 |
| SRR768308      | WGS     | SRR768308_1.clean.fq.gz                               | 69675029        | 5.5G                 |
|                 |         | SRR768308_2.clean.fq.gz                               | 69675029        | 5.9G                 |
| ERR091571      | WGS     | ERR091571_1.clean.fq.gz                               | 196150792       | 16.0G                |
|                 |         | ERR091571_2.clean.fq.gz                               | 196150792       | 16.0G                |
| NA12878_1000G  | WGS     | NA12878_merge1_1000G.clean.R1.fq.gz                    | 741582047       | 32.0G                |
|                 |         | NA12878_merge1_1000G.clean.R2.fq.gz                    | 741582047       | 33.0G                |
| NA12878_NIST   | WGS     | NA12878_merge2.R1.clean.fq.gz                          | 236287755       | 23.0G                |
|                 |         | NA12878_merge2.R2.clean.fq.gz                          | 236287755       | 25.0G                |
| ERR262997      | WGS     | ERR262997.R1.clean.fq.gz                              | 593216693       | 49.0G                |
|                 |         | ERR262997.R2.clean.fq.gz                              | 593216693       | 50.0G                |

II. Evaluation Methodology

HummingBird pipeline is such an endeavor developed to reduce the clock time for germline variant calling by optimizing the standard BWA-GATK software interaction with maximizing the usage of hardware. As in commonly accepted BWA-GATK pipeline, the HummingBird pipeline is built with a modified BWA alignment algorithm (HB-BWA) but preserving the underlying mathematical model. The mapped sequence reads in bam files are further piped into elPrep software [16]–[18], a tool kit written in GO, that outputs sorted, duplication marked and recalibrated sequence reads, which is next processed with an optimized, C++ based GATK HaplotypeCaller (HB-HaplotypeCaller) for variant calling.

We assessed the efficiency of HummingBird pipeline by using a serial publicly available datasets. The datasets used in this paper included 12 whole genome/exome sequencing data files from NCBI/EBI database (Table 1). Table 2 shows the download links for the datasets. The raw reads of these datasets were processed with fastp [26] software to remove low quality reads. Figure 1 presents the major steps of HummingBird and BWA-GATK pipelines. The benchmarking BWA-GATK procedure first maps the sequencing reads by BWA MEM 0.7.17 to human hg19 reference genome, and then with SAMtools 1.9 for indexing, GATK 4.1.2.0 SortSam, MarkDuplicates, BaseRecalibrator and ApplyBQSR to perform sequencing reads sorting, duplicates marking and base recalibration. After the sequence reads have been prepared, GATK 4.1.2.0 HaplotypeCaller was used for variant calling. Both pipelines were run on a x86_64 standalone server with Intel(R) Xeon(R) Gold 6254 CPU/2 3.10GHz, 72 threads, 256GB memory and a 1.5TB SSD hard drive.

III. Evaluation Results

A. Running time comparison

To evaluate the performance of HummingBird pipeline, we first run the pipeline along with the standard BWA-GATK procedure on the same standalone server. We evaluated the overall running time for 12 datasets downloaded from public databases. Figure 2 compares the total time cost of HummingBird pipeline and standard BWA-GATK pipeline. We found
TABLE II
DOWNLOAD ADDRESS OF DATASETS USED IN THIS STUDY

| Sample Name | Raw Reads Download Address |
|-------------|----------------------------|
| NIST7035_L001 | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
NIST7035_TAAGGCGA_L001_R1_001_trimmed.fastq.gz
NIST7035_TAAGGCGA_L001_R2_001_trimmed.fastq.gz |
| NIST7035_L002 | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
NIST7035_TAAGGCGA_L002_R1_001_trimmed.fastq.gz
NIST7035_TAAGGCGA_L002_R2_001_trimmed.fastq.gz |
| NIST7086_L001 | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
NIST7086_CGTACTAG_L001_R1_001_trimmed.fastq.gz
NIST7086_CGTACTAG_L001_R2_001_trimmed.fastq.gz |
| NIST7086_L002 | ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome
NIST7086_CGTACTAG_L002_R1_001_trimmed.fastq.gz
NIST7086_CGTACTAG_L002_R2_001_trimmed.fastq.gz |
| SRR098401 | ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR098/SRR098401/SRR098401_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR098/SRR098401/SRR098401_2.fastq.gz |
| SRR742200 | ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR742/SRR742200/SRR742200_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR742/SRR742200/SRR742200_2.fastq.gz |
| SRR098359 | ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR098/SRR098359/SRR098359_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR098/SRR098359/SRR098359_2.fastq.gz |
| SRR768308 | ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR768/SRR768308/SRR768308_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR768/SRR768308/SRR768308_2.fastq.gz |
| ERR091571 | ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR091/ERR091571/ERR091571_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR091/ERR091571/ERR091571_2.fastq.gz |
| NA12878_1000G | ftp://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/phase3/data/NA12878/sequence_read/NA12878_NIST
ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/NA12878/NIST_NA12878_HG001_HiSeq_300x
NIST7086_CGTACTAG_L002_R1_001_trimmed.fastq.gz |
| NA12878_NIST | ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceExamples/giabdata/NA12878/NIST_NA12878_HG001_HiSeq_Exome
/131219_D00360_005_BH814YADXX/Project_RM8398/ERR091571
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR262/ERR262997/ERR262997_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR262/ERR262997/ERR262997_2.fastq.gz |
| ERR262997 | ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR262/ERR262997/ERR262997_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR262/ERR262997/ERR262997_2.fastq.gz |

that, overall, the HummingBird pipeline processed germline mutation calling much more efficient than the standard BWA-GATK pipeline, with a maximal speedup of 12.7x for the whole genome sequencing dataset SRR768308 and 4.6x for the whole exome sequencing dataset SRR098359.

To further assess the efficiency of individual steps along the pipeline, we looked into the alignment and variant calling steps respectively as elPrep has been shown to efficiently prep mapped reads for GATK HaplotypeCaller. The time cost is shown in Figure 3A and 3B respectively. We found that HB-BWA and HB-HaplotypeCaller performed relatively more efficient than the standard BWA MEM and GATK HaplotypeCaller, with the maximal speedup of 9.4x in BWA alignment step for dataset ERR262997 and 36.9x in haplotype caller step for dataset ERR091571.

Overall, we concluded that, with the help from elPrep, the HB-BWA and HB-HaplotypeCaller within the HummingBird pipeline are more efficient for processing germline variant calling than the standard BWA-GATK procedure.

B. Variant calling accuracy

Recall reflects the percentage of variants obtained by the standard BWA-GATK benchmark pipeline that were called by HummingBird pipeline in each dataset. Precision is the percentage of called variants which match variants obtained by the standard BWA-GATK benchmark pipeline. F1-score is the harmonic mean of recall and precision, as a combined metric for evaluating overall accuracy. With the effort to assessment the accuracy of HummingBird pipeline, we calculated the result of precision and recall and further obtained F1-score.
on each test by the following formula:

\[ F1_{core} = \frac{precision \times recall}{precision + recall} \times 2 \]  

(1)

All these results are shown in Figure 4. It can be seen that the precision, recall and F1-score on these data samples are consistently high across different datasets, except for dataset ERR262997, a dataset generated for large insert library evaluation, with an F1-score of 0.87.

We also looked into the performance of HummingBird pipeline on the gold standard sample NA12878, one of the most studied samples with a better knowledge of its variants by using the hap.py evaluation tool at the high confidence intervals on chromosome 1-22, X and Y developed by Illumina Inc (19). Results on dataset NA12878-NIST showed that HummingBird pipeline generated the comparable accuracy score with standard BWA-GATK pipeline for both INDEL and SNP (Figure 5, additional results are shown in Table 3).

C. Variant calling consistency

To further assess the repeatability of the variant calling results, we conducted the triplicated variant calling runs of HummingBird pipeline on 10 samples, and the corresponding F1-scores are shown in Figure 6. Results showed that HummingBird has a stable variant calling ability.

IV. DISCUSSION AND FUTURE DIRECTIONS

In this study, we validated the efficiency of HummingBird germline variant calling pipeline by comparing it with the widely accepted BWA-GATK pipeline. High speedup and comparable accuracy support the superiority of HummingBird.
Fig. 6. Consistency of variant call by HummingBird pipeline. HummingBird pipeline were run in triplicates, accuracy metrics of the triplicated variant calling runs of HummingBird pipeline were evaluated on selected samples.

| TABLE III | ACCURACY METRICS OF HUMMINGBIRD TRIPLICATES RUN |
|------------|----------------------------------|
| data set   | Repeats | Recall  | Precision | F1-Score |
| NIST7035_L001 | Repeat1 | 0.9842  | 0.8926 | 0.9362 |
|            | Repeat2 | 0.9842  | 0.8927 | 0.9362 |
|            | Repeat3 | 0.9842  | 0.8927 | 0.9362 |
| NIST7035_L002 | Repeat1 | 0.9845  | 0.8943 | 0.9372 |
|            | Repeat2 | 0.9844  | 0.8943 | 0.9372 |
|            | Repeat3 | 0.9844  | 0.8943 | 0.9372 |
| NIST7086_L001 | Repeat1 | 0.9837  | 0.8951 | 0.9373 |
|            | Repeat2 | 0.9837  | 0.8952 | 0.9374 |
|            | Repeat3 | 0.9837  | 0.8951 | 0.9373 |
| NIST7086_L002 | Repeat1 | 0.9858  | 0.8928 | 0.9370 |
|            | Repeat2 | 0.9858  | 0.8928 | 0.9370 |
|            | Repeat3 | 0.9858  | 0.8928 | 0.9370 |
| SRR098359    | Repeat1 | 0.9845  | 0.9012 | 0.9410 |
|            | Repeat2 | 0.9845  | 0.9011 | 0.9410 |
|            | Repeat3 | 0.9845  | 0.9011 | 0.9410 |
| SRR742200    | Repeat1 | 0.9870  | 0.9218 | 0.9533 |
|            | Repeat2 | 0.9870  | 0.9218 | 0.9533 |
|            | Repeat3 | 0.9870  | 0.9218 | 0.9533 |
| SRR098401    | Repeat1 | 0.9827  | 0.8944 | 0.9365 |
|            | Repeat2 | 0.9828  | 0.8944 | 0.9365 |
|            | Repeat3 | 0.9828  | 0.8944 | 0.9365 |
| ERR262997    | Repeat1 | 0.8857  | 0.8733 | 0.8794 |
|            | Repeat2 | 0.8857  | 0.8732 | 0.8794 |
|            | Repeat3 | 0.8857  | 0.8733 | 0.8794 |
| NA12878_1000G | Repeat1 | 0.9624  | 0.8938 | 0.9269 |
|             | Repeat2 | 0.9624  | 0.8939 | 0.9269 |
|             | Repeat3 | 0.9624  | 0.8938 | 0.9269 |
| ERR091571    | Repeat1 | 0.9817  | 0.9233 | 0.9516 |
|             | Repeat2 | 0.9817  | 0.9234 | 0.9516 |
|             | Repeat3 | 0.9817  | 0.9233 | 0.9516 |

than BWA-GATK pipeline. Thus, it provides an alternative way of germline mutation variant calling to meet the needs of the community.

The optimization to germline mutation variant calling on standalone machine is important but not adequate. In this study, we did not test its performance in a cluster computation environment, although the pipeline can be readily implemented. With the rapid increase of human genomics data, we need to pay more effort to implement and optimize HummingBird pipeline on distributed clusters and thus bear larger scale computing.

ACKNOWLEDGEMENTS

We thank the individuals who have made their data publicly available. We also would like to thank the public sequencing projects made the genomic data available for different researchers to explore. We thank Wei Yue, Lei Ye, Jinpeng Jiang, Yueyan Zhao, Jianhua Gao, Shiyu Han and Zhaonian Tan for their inputs during this manuscript preparation.

REFERENCES

[1] K. M. Boycott, M. R. Vanstone, D. E. Bulman, A. E. MacKenzie, Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14, 681-691 (2013).
[2] R. B. Altman et al., A research roadmap for next-generation sequencing informatics. Sci Transl Med 8, 335ps310 (2016).
[3] UK Biobank data on 500,000 people paves way to precision medicine. Nature 562, 163-164 (2018).
[4] N. Cox, UK Biobank shares the promise of big data. Nature 562, 194-195 (2018).
[5] M. J. Khoury et al., From public health genomics to precision public health: a 20-year journey. Genet Med 20, 574-582 (2018).
[6] A. Auton et al., A global reference for human genetic variation. Nature 526, 68-74 (2015).
[7] G. R. Abecasis et al., A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073 (2010).
[8] G. R. Abecasis et al., An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65 (2012).
[9] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760 (2009).
[10] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-595 (2010).
[11] L. Pirrucchini, S. Leo, G. Zanetti, SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics 27, 2159-2160 (2011).
[12] J. M. Añón, J. C. Pichel, T. F. Pena, J. Amigo, SparkBWA: Speeding Up the Alignment of High-Throughput DNA Sequencing Data. PLoS One 11, e0155461 (2016).
[13] D. Peters, X. Luo, K. Qiu, P. Liang, Speeding Up Large-Scale Next Generation Sequencing Data Analysis with pBWA. J Appl Bioinform Comput Biol 1, (2012).

[14] A. McKenna et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303 (2010).

[15] G. A. Van der Auwera et al., From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43, 11.10.11-11.10.33 (2013).

[16] C. Herzeel, P. Costanza, D. Decap, J. Fostier, J. Reumers, elPrep: High-Performance Preparation of Sequence Alignment/Map Files for Variant Calling. PLoS One 10, e0132868 (2015).

[17] P. Costanza, C. Herzeel, W. Verachtert, Comparing Ease of Programming in C++, Go, and Java for Implementing a Next-Generation Sequencing Tool. Evol Bioinform Online 15, 1176934319869015 (2019).

[18] C. Herzeel, P. Costanza, D. Decap, J. Fostier, W. Verachtert, elPrep 4: A multithreaded framework for sequence analysis. PLoS One 14, e0209523 (2019).

[19] M. A. Eberle et al., A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res 27, 157-164 (2017).

[20] H. Li et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).

[21] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. Bioinformatics 27, 2987-2993 (2011).

[22] B. Institute. (2019), pp. http://broadinstitute.github.io/picard/.

[23] M. Yi et al., Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. Nucleic Acids Res 42, e101 (2014).

[24] N. Kathiresan et al., Accelerating next generation sequencing data analysis with system level optimizations. Sci Rep 7, 9058 (2017).

[25] M. Plüss et al., Need for speed in accurate whole-genome data analysis: GENALICE MAP challenges BWA/GATK more than PEmapper/PECaller and Isaac. Proc Natl Acad Sci U S A 114, E8320-E8322 (2017).

[26] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884-i890 (2018).