ACADEMIA JOURNAL OF BIOLOGY 2021, 43(2): 47–53
DOI: 10.15625/2615-9023/16006

A COMPARISON OF THREE VARIANT CALLING PIPELINES USING SIMULATED DATA

Nguyen Van Tung[1,2], Nguyen Thi Kim Lien[1], Nguyen Huy Hoang[1,2,*]

[1]Institute of Genome Research, VAST, Vietnam
[2]Graduate University of Science and Technology, VAST, Vietnam

Received 13 January 2021, accepted 8 June 2021

ABSTRACT

Advances in next generation sequencing allow us to do DNA sequencing rapidly at a relatively low cost. Multiple bioinformatics methods have been developed to identify genomic variants from whole genome or whole exome sequencing data. The development of better variant calling methodologies is limited by the difficulty of assessing the accuracy and completeness of a new method. Normally, computational methods can be benchmarked using simulated data which allows us to generate as much data as desired and under controlled scenarios. In this study, we compared three variant calling pipelines: Samtools/VarScan, Samtools/Bcftools, and Picard/GATK using two simulated datasets. The result showed a significant difference between the three pipelines in two cases. In Chromosome 6 dataset, GATK and Bcftools pipelines detected more than 90% of variants. Meanwhile, only 82.19% of mutations were detected by VarScan. In NA12878 datasets, the result showed GATK pipeline was more sensitive than Bcftools and Varscan pipeline. All pipelines showed a high Positive Predictive Value. Moreover, by a measure of run time, VarScan was the highest pipeline but GATK has an option for multithreading which is a way to make a program run faster. Therefore, GATK is more effective than Bcftools and Varscan to variant calling with a lower coverage dataset.

Keywords: Bcftools, GATK, Simulated data, Variant calling, VarScan, Dwgsim.

Citation: Nguyen Van Tung, Nguyen Thi Kim Lien, Nguyen Huy Hoang, 2021. A comparison of three variant calling pipelines using simulated data. *Academia Journal of Biology*, 43(2): 47–53. https://doi.org/10.15625/16006

**Corresponding author email*: nhhoang@igr.ac.vn

INTRODUCTION

Next-generation sequencing (NGS) known as high throughput sequencing, allow for sequencing of nucleic acid including DNA and RNA much more quickly and cheaper than previously sequencing method. Therefore, whole genome sequencing (WGS) and whole exome sequencing (WES) methods are widely applied in clinical for detecting patient's genomic variants of the genetic disease. WES is becoming a standard, more economic approach to do genome sequencing. However, the massive data was generated by NGS result in multiple challenges, including storage and bioinformatics analysis.

Currently, numerous genetic variant calling methods have been developed to identify genomic variants from whole genome or exome sequencing data. Most methods are based on the alignment of raw sequence reads against a reference genome (Li, 2014, 2012). This approach has some disadvantages including incompleteness of genomes assembly (Meyer et al., 2013), sequencing errors, and interference of single nucleotide polymorphisms on reads mapping (Iqbal et al., 2012), structural variations in the individual genomes (Sudmant et al., 2015). Therefore, the alignment-based approach has high levels of false positives of variant calling. More efficient computational methods and softwares are constantly being developed in order to provide more accurate and faster processing. When new computational methods or software tools are developed, it is essential that these software tools are more superior than existing tools with similar functionality. Normally, bioinformatics methods can be benchmarked using simulated data which allows us to generate as much data as desired and under controlled variants. Thus, computer simulation of genomic data has become increasingly popular for assessing and validating biological models.

Among many paired-end short read aligners, Bowtie2 (Langmead & Salzberg, 2012), BWA-MEM (Li, 2013) and SOAP2 (Li et al., 2009) are widely used because of fast, memory-efficient, and particularly useful for aligning repetitive reads. The most dominant genotype calling pipelines are the GATK Best Practices (DePristo et al., 2011; Van der Auwera et al., 2013). These workflows recommend read mapping by Burrows-Wheeler Aligner (BWA), post-alignment processing using Picard, and then GATK variant calling. In addition, VarScan (Koboldt et al., 2009) and Bcftools (Narasimhan et al., 2016) - two open source tools for detecting SNPs, insertions and deletions variants are widely used. In this study, we performed variant detection with two datasets including a simulated human GRCh38 chromosome 6 and a WES data set NA12878 for comparison of three variant calling pipelines using Samtools - Varscan, Samtools - Bcftools and Picard - GATK.

MATERIALS AND METHODS

Datasets

To assess variant calling pipeline, we performed with two datasets: a simulated human GRCh38 chromosome 6 and a WES data set NA12878. In the simulated dataset, the sequence of Chromosome 6 (NC_000006.12) was used as reference data to simulated Illumina paired-end reads using Dwgsim simulator (available at https://github.com/nh13/DWGSIM) with a length of reads were 150 bp. The sequence of Chromosome 6 (NC_000006.12) was used as reference data for variant calling from this dataset.

The WES dataset NA12878 was sequenced using the HiSeq Illumina 2000 platform and annotated from Genome in a Bottle consortium (GiaB). The sequence of human GRCh38 was used as reference data for variant calling from this dataset.

Variant calling pipeline

The reads were mapped to reference data by order using Burrows-Wheeler Aligner (Li & Durbin, 2009). The alignment sam file then was used for all three variant calling pipelines.

In the first pipeline, we used Picard (available at http://broadinstitute.github.io/picard/) to mark and removed repeated reads, to sort and create indexes on alignment bam files. Variants were called by applying HaplotypeCaller in the GATK package version 4.1. To reduce erroneous calls, alignments were subjected to duplicate marking and local realignment by following the GATK Best Practices (DePristo et al., 2011; Van der Auwera et al., 2013).

In other pipelines, post-alignment processing was performed using Samtool (Li et al., 2009) to sort and create indexes on alignment bam files. Then variants were detected using Varscan version 2.4.3 (Koboldt et al., 2009) and Bcftools (Narasimhan et al., 2016). Schematic of the data analysis workflow was shown in Figure 1.

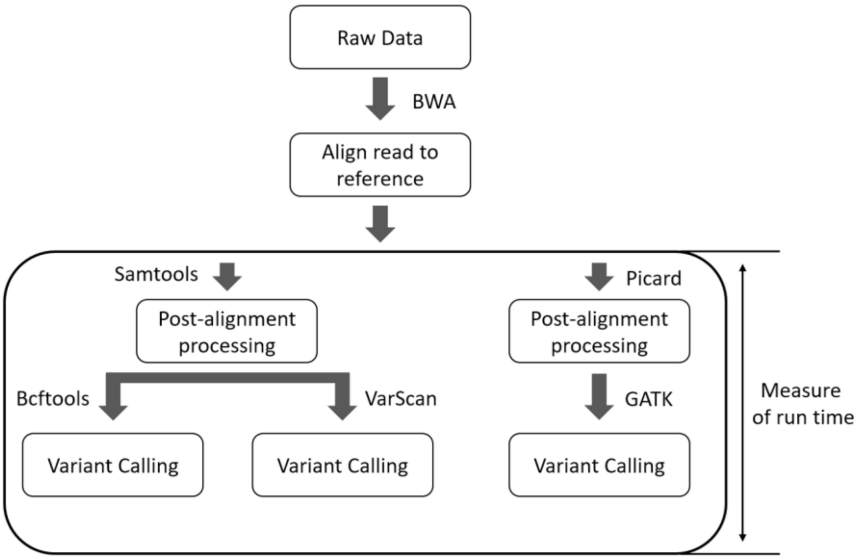

Figure 1. Schematic of the data analysis workflow

Statistical Calculations

Positive Predictive Value (PPV) and sensitivity of each tool were calculated by applying formulas below:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

True Positive (TP) is a mutation that was detected by the pipeline being tested and is one that exists in the inserted variants list; False Positive (FP) is a mutation that was detected by the pipeline being tested but is one that does not exist in the inserted variants list; False Negative (FN) is a mutation that was not detected by the pipeline being tested but is one that does exist in the inserted variants list.

Measure of run time

Three pipelines were performed by 1 core on a high-performance computer system (196 GB RAM, Intel Xeon CPU with 2.60 GHz). The run time of each pipeline was measured from post-alignment processing step to end of pipeline.

RESULTS AND DISCUSSION

Data simulation features

In a real NGS experiment, the type, number, and length of reads are determined by the specific sequencing machine and the library preparation. Computational simulators can generate a specific amount of reads with different lengths according to the sequencing technology assumed. The number of reads can be specified or estimated according to the desired coverage. Coverage is a key factor in variant calling. Tian et al., showed that 40X coverage seems sufficient for most of the callers across the full range of divergence (Tian et al., 2016). In another report, Song et al. considered that coverage of 15X was a suitable choice for obtaining a sufficient number of accurately genotyped SNPs (Song et al., 2016). In this study, we assumed the type of reads were paired-end, coverage was 30X, and length of reads was 150 bp. Computational simulation of two datasets yields 2.5 Gb sequence paired-end reads of human chromosome 6.

Mapping paired-end reads to the reference sequence

The paired-end reads were mapped against the reference sequences using the BWA aligner. The default parameter was used for mismatches in the "bwa aln" command, which means there were no more than 2 mismatches in the first 32 bp of reading. The mapping rate was defined as the ratio of mapped reads over the total number of simulated reads. In the simulated datasets, almost 100% of the read was mapped to the reference sequence and about 99.80% of the sequence reads were aligned to the reference in the NA12878 dataset.

Variant calling from human chromosome 6 simulated data

To compare three pipelines, we performed variant calling by applying all pipelines from the same alignment sam file. The result showed that the pipeline using Picard/GATK Haplotypecaller and Bcftools were more sensitive than Samtool - Varscan pipeline. In total 6,474,862 variants were generated by Dwgsim, 6,419,178 variants (99.14%) were detected by GATK Haplotypecaller, 5,839,168 variants (90.18%) were detected by Bcftools. Only 5,321,774 variants (82.19%) were detected using the Samtool-Varscan pipeline. Variant calling performance, Positive Predictive Value, and sensitivity of three pipelines were showed in Table 1.

Table 1. Variant calling performance of three pipelines in Chromosome 6 dataset

Variant Caller	True Positive	False Positive	False Negative	PPV (%)	Sensitivity (%)
GATK	6,419,178	0	110,954	100	99.14
Bcftools	5,839,168	20	635,694	99.99	90.18
VarScan	5,321,774	0	1,153,088	100	82.19

Variant calling from NA12878 dataset

GATK Haplotypecaller and Bcftools pipelines were more sensitive than Samtool - Varscan pipeline. In total 3,775,119 variants were reported GiaB in NA12878 dataset, 3,744,918 variants (99.20%) were detected by GATK Haplotypecaller, 3,507,085 variants (92.89%) were detected by Bcftools. Only 3,390,056 variants (89.80%) were detected using Samtool - Varscan pipeline and 14 mutations were reported as failed by the strand filter. Variant calling performance, Positive Predictive Value, and sensitivity of three pipelines were showed in Table 2.

Table 2. Variant calling performance of three pipelines in NA12878 dataset

Variant Caller	True Positive	False Positive	False Negative	PPV (%)	Sensitivity (%)
GATK	3,744,918	0	30,201	100	99.20
Bcftools	3,507,085	18	268,034	99.99	92.89
VarScan	3,390,056	11	385,063	99.99	89.80

Measure of run time

The run time of each pipeline was measured from post - alignment processing step to end of pipeline (Table 3).

VarScan was the fastest tool in the two cases. There was no significant difference between Bcftools and GATK in the run time but GATK has an option for multithreading which is a way to make a program finish faster. All three pipelines contain pre-processing steps which maybe increase the run time. However, as these steps are either necessary or highly recommended by the authors of the tools, they are regarded as being an integral part of the variant calling process.

Table 3. Run time of the variant calling process for the three variant calling pipelines

Pipeline	Runtime Chromosome 6 dataset	Runtime NA12878	Option for multithreading
GATK	3,782 seconds	5,857 seconds	Yes
VarScan	3,015 seconds	5,123 seconds	No
Bcftools	3,856 seconds	5,967 seconds	No

The development of NGS technologies has remarkably decreased the cost of genome sequencing. This affordability of NGS allows the clinical application of WES or WGS to identify variants of personal genomes for practicing genomic medicine. This affordability of NGS allows the clinical application of WES or WGS to identify variants of personal genomes for practicing genomic medicine. Accurate variant discovery is crucial for pinpointing the causal mutations underlying human diseases. Current computational methods are generally effective in detecting ordinary variants but less so for variants located in difficult regions (Weisenfeld et al., 2014). Furthermore, each kind of data in accordance with a different method. Although the alignment-based approach has many limitations, genetic variant calling is based on the alignment of raw sequence reads against a reference genome has been widely applied on large and complex genomes (Wu et al., 2017). In this study, GATK and Bcftools pipelines were more sensitive than the Varscan pipeline. This result may be since Varscan requested a higher coverage than GATK. Koboldt et al. (2013) found that VarScan2 performed best overall with sequencing depths of 100x, 250x, 500x and 1000x required to accurately identify variants present at 10%, 5%, 2.5% and 1% respectively.

CONCLUSION

In summary, to understand the overall performance of variant callers for next generation sequencing data, we compared Samtools - Varscan, Samtools - Bcftools Picard - GATK variant calling pipelines using simulated paired-end read of chromosome 6 and NA12878 dataset from GB. GATK pipeline showed the highest sensitivity and positive predictive value. Although it wasn’t the fastest pipeline GATK has an option for multithreading which will make it run faster. Therefore, GATK is more effective than Bcftools and Varscan to variant calling with a lower coverage dataset.

Acknowledgement: This study was supported by the National Foundation for Science and Technology Development, Vietnam (NAFOSTED, grant no. 108.02-2018.305).

REFERENCES

DePristo M. A., Banks E., Poplin R., Garimella K. V., Maguire J. R., Hartl C., Philippakis A. A., del Angel G., Rivas M. A., Hanna M., McKenna A., Fennell T. J., Kernytsky A. M., Sivachenko A. Y., Cibulskis K., Gabriel S. B., Altshuler D., Daly M. J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43: 491–498.

Ewing B., Green P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*., 8: 186–194.

Ewing B., Hillier L., Wendl M. C., Green P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.,* 8: 175–185.

Iqbal Z., Caccamo M., Turner I., Flicek P., McVean G., 2012. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.*, 44: 226–232.

Koboldt D. C., Chen K., Wylie T., Larson D. E., McLellan M. D., Mardis E. R., Weinstock G. M., Wilson R. K., Ding L., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinforma. Oxf. Engl.*, 25: 2283–2285.

Koboldt D. C., Larson D. E., Wilson R. K., 2013. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al 44: 15.4.1-15.4.17.

Langmead B., Salzberg S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9: 357–359.

Li H., 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinforma. Oxf. Engl.*, 30: 2843–2851.

Li H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio.

Li H., 2012. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinforma. Oxf. Engl.*, 28: 1838–1844.

Li H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25: 1754–1760.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.*, 25: 2078–2079.

Li R., Yu C., Li Y., Lam T. W., Yiu S.

M., Kristiansen K., Wang J., 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinforma. Oxf. Engl.*, 25: 1966–1967.

Meyer L. R., Zweig A. S., Hinrichs A. S., Karolchik D., Kuhn R. M., Wong M., Sloan C. A., Rosenbloom K. R., Roe G., Rhead B., Raney B. J., Pohl A., Malladi V. S., Li C. H., Lee B. T., Learned K., Kirkup V., Hsu F., Heitner S., Harte R. A., Haeussler M., Guruvadoo L., Goldman M., Giardine B. M., Fujita P. A., Dreszer T. R., Diekhans M., Cline M. S., Clawson H., Barber G. P., Haussler D., Kent W. J., 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, 41: D64–D69.

Narasimhan V., Danecek P., Scally A., Xue Y., Tyler-Smith C., Durbin R., 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinforma. Oxf. Engl.*, 32: 1749–1751.

Song K., Li L., Zhang G., 2016. Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Sci. Rep.*, 6: 35736.

Sudmant P. H., Rausch T., Gardner E. J., Handsaker R. E., Abyzov A., Huddleston J., Zhang Y., Ye K., Jun G., Hsi-Yang Fritz M., Konkel M. K., Malhotra A., Stütz A. M., Shi X., Paolo Casale F., Chen J., Hormozdiari F., Dayama G., Chen K., Malig M., Chaisson M. J. P., Walter K., Meiers S., Kashin S., Garrison E., Auton A., Lam H. Y. K., Jasmine Mu X., Alkan C., Antaki D., Bae T., Cerveira E., Chines P., Chong Z., Clarke L., Dal E., Ding L., Emery S., Fan X., Gujral M., Kahveci F., Kidd J. M., Kong Y., Lameijer E. W., McCarthy S., Flicek P., Gibbs R. A., Marth G., Mason C. E., Menelaou A., Muzny D. M., Nelson B. J., Noor A., Parrish N. F., Pendleton M., Quitadamo A., Raeder B., Schadt E. E., Romanovitch M., Schlattl A., Sebra R., Shabalin A. A., Untergasser A., Walker J. A., Wang M., Yu F., Zhang C., Zhang J., Zheng-Bradley X., Zhou W., Zichner T., Sebat J., Batzer M. A., McCarroll S. A., The 1000 Genomes Project Consortium, Mills R. E., Gerstein M. B., Bashir A., Stegle O., Devine S. E., Lee C., Eichler E. E., Korbel J. O., 2015. An integrated map of structural variation in 2,504 human genomes. *Nature,* 526: 75–81.

Tian S., Yan H., Neuhauser C., Slager S. L., 2016. An analytical workflow for accurate variant discovery in highly divergent regions. *BMC Genomics*, 17(1): 703.

Van der Auwera G. A., Carneiro M. O., Hartl C., Poplin R., Del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K. V., Altshuler D., Gabriel S., DePristo M. A., 2013. From FastQ data to high confidence variant calls: the genome analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma*., 43: 11.10.1–11.10.33.

Weisenfeld N. I., Yin S., Sharpe T., Lau B., Hegarty R., Holmes L., Sogoloff B., Tabbaa D., Williams L., Russ C., Nusbaum C., Lander E. S., MacCallum I., Jaffe D. B., 2014. Comprehensive variation discovery in single human genomes. *Nat. Genet*., 46: 1350–1355.

Wu L., Yavas G., Hong H., Tong W., Xiao W., 2017. Direct comparison of performance of single nucleotide variant calling in human genome with alignment-based and assembly-based approaches. *Sci. Rep*., 7: 10963.