AJMAS
https://journal.utripoli.edu.ly/index.php/Alqalam/index eISSN 2707-7179

Original Article

Germline Short Variant Discovery and Annotation Pipeline using GATK Tool

Ashraf Bourawy*, Abdalmunam Abdalla

Department of Computer Science, Faculty of Science, Omar AL-Mukhtar University, Albayda, Libya

ARTICLE INFO

Corresponding Email. *abourawy@omu.edu.ly*

Received*: 20-07-2023*
Accepted*: 05-08-2023*
Published*: 07-08-2023*

Keywords*. Germline, Variant discovery, GATK, Variant Annotation.*

ABSTRACT

Background and aims*. Identifying of variants related to genetic diseases has become affordable with the advances in whole genome sequencing (WGS) enabled by the enhancements of next-generation sequencing technology. Germline and somatic variants are discovered with the help of bioinformatics pipelines utilizing specialized tools. However, the performance and workflow of these tools are subject to evaluation. The aim of this study is to investigate the pipeline of the Genome Analysis Toolkit (GATK) tool in discovering and annotating germline short variants. In particular, this study aimed at variant calling of single nucleotide polymorphisms (SNPs) and short insertions and deletions (Indels).* ***Methods****. To accomplish our aim, several tools and packages are used in variant calling workflow. The focus of this study is on evaluating the GATK tool. There are some tools associated to work along with GATK, such as PICARD package. In addition, other tools are needed for data preprocessing before the GATK tool is applied, such as BWA and SAMTools. The human reference genome is used for mapping (alignment) purpose. Paired-end sequence reads are used as the subject of discovering germline variants. Different methods are followed in this study concerning data preprocessing, variants discovery, and variants filtering and annotation.* ***Results****. The data preprocessing steps have revealed good quality of the sequences, base quality scores, and adapter contents. These results have indicated that the sequence reads possess good quality and given the approval to proceed with downstream workflow. Using the GATK tools, variants calling has been performed where SNPs and Indels were obtained in two separate files. Filtration and annotation are applied on the discovered variants and an Excel file was obtained. This file contains the found variants which were generated by comparison with the aid of data sources from well-known databases.* ***Conclusion****. Upon completing the GATK pipeline, germline variants (SNPs and Indels) were discovered and an Excel file was produced with all information. Further analysis can be performed by specialized scientists in a convenient manner.*

Cite this article. *Bourawy A., Abdalla A. Germline Short Variant Discovery and Annotation Pipeline using GATK Tool. Alq J Med App Sci. 2023;6(2):424-432.* *https://doi.org/10.5281/zenodo.8219249*

INTRODUCTION

Germline is the set of cells that are responsible for transmitting genetic information from one generation to the next. Those cells are typically found in the gonads, which are the reproductive organs [1]. The germline is essential for preserving genetic variety in a population as they are constantly being shuffled and recombined, which is a process that shuffles the genetic material. The shuffling of genetic material helps to ensure that each offspring is different from its

parents, and that the population is not vulnerable to genetic disorders [2, 3]. The importance of germline can be related with evolution because they are the only cells that can transmit genetic mutations to future generations. Mutations can be advantageous, neutral, or harmful. Advantageous mutations can help an organism to survive and reproduce, which can lead to the evolution of new species [4].

In essence, germline variants are genetic mutations that occur in the DNA of reproductive cells. They can be passed down from parents, or they can be developed during the lifetime of an organism. There are two types of germline variants: single nucleotide polymorphisms (SNPs) and short insertions and deletions (Indels). SNPs are changes in a single DNA base pair, while Indels are changes in the number of DNA base pairs [5]. Germline variants can alter the function of genes and can be a source of genetic problems. The way to cure the diseases caused by germline variants is to modify the genetic mutations in the nuclear and mitochondrial DNA of pre-implantation embryos. This procedure is a type of gene therapy known as germline gene therapy [6].

On the other hand, a germline short variant, also known as a germline single nucleotide variant (SNV), is a type of genetic mutation that occurs in the DNA sequence of an organism's germ cells. SNVs are the most common type of genetic variation and refer to a change in a single nucleotide base (A, C, G, or T) within the DNA. They are received from one or both parents and are present in every cell of an individual's body [7]. Some germline short variants, known as benign variants, have no effect on health, while others can increase the risk of certain problems or diseases. Testing for germline short variants is often used in clinical genetics to identify individuals who may be vulnerable to certain genetic conditions [8]. Many tools are available to discover germline short variants. Some of the most popular tools are shown in Table 1:

Table 1. Popular Tools for Variant Discovery

No.	Tool (Package)	Description
1	GATK HaplotypeCaller [9]	A commonly used depth-based variant caller that is provided as part of the GATK toolkit.
2	FreeBayes [10]	Another popular standalone depth-based variant caller.
3	Beagle [11]	A popular alignment-based variant caller that is available as a standalone tool.
4	Samtools Mpileup [12]	This is a utility that can be used to generate a pileup of aligned reads, which can then be used to discover variants.
5	DeepVariant [13]	A machine learning-based variant caller.
6	VarScan2 [14]	Another machine learning tool that can be used to identify germline variants.

These tools employ various methods to determine germline short variants. Some tools, such as GATK HaplotypeCaller and FreeBayes, use *depth-based* methods to identify variants that are likely to be germline. Other tools, such as Beagle and Samtools Mpileup, use *alignment-based* methods to identify variants that are absent in the reference genome. *Machine learning-based* tools, such as DeepVariant and VarScan2, use machine learning models to predict germline variant. Other variant calling tools exist that vary from depth-based to machine learning-based to hybrid methods including MuTect2 [15], ANNOVAR [16], Manta [17], Strelka2 [18], Delly [19], and Lumpy [20]. Besides limitations associated with the tool itself, the choice of variant discovery tools for data analysis may significantly impact the diagnostic outcome. The use of those tools depends on the type of data under consideration, the application used, the accuracy and sensitivity needed, the computational resources available, and the cost of the tool [21].

The Genome Analysis Toolkit (GATK) is a free, simple, and widely used software package for germline variant calling [22]. GATK offers a wide variety of tools and algorithms for identifying and analyzing genetic variants, as well as for quality control and working with RNA-seq and ChIP-seq data. Although initially designed for analyzing exomes and whole genomes generated with Illumina sequencing technology, GATK can be adapted to handle a variety of other technologies and experimental designs. The tool was originally developed for human genetics, and it has since evolved to handle genome data from any organism, with any level of ploidy [22]. The developer of GATK proposed a set of guidelines, namely the GATK best practices. They offer a step-by-step guide for conducting variant discovery analysis on high-throughput sequencing (HTS) data [23].

In this study we have performed the GATK Best Practices workflow on sequenced reads obtained from [24] against the human reference genome. Data sources of well-know variants from different databases were utilized for annotating SNPs and Indels variants. Our contribution in this paper is the evaluation and investigation of the performance of the GATK tool in calling and annotating variants for further analysis.

METHODS

In order to conduct the germline variant calling pipeline, the online RStudio server on Galaxy platform has been utilized [25]. A new environment was created for our project and the required tools were downloaded and installed under the created environment. The list of the used tools is given in Table 2.

Table 2. Tools and Packages Used in the Workflow

No.	Tool (Package)	Description
1	BWA (v0.7.17-r1198)	Indexing reference genome and then mapping DNA sequences against the human reference genome.
2	SAMTools (v1.15.1)	Utilities for manipulating alignments in the SAM format.
3	GATK4 (v4.2.6.1)	Genome Analysis ToolKit for variant discovery, filtering, evaluation.
4	PICARD (v3.0.0)	A Java package needed by GATK. Comprises a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats.
5	FastQC (v0.11.9)	A quality control tool for high throughput sequence data. Written in Java. It provides a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines.
6	MultiQC (v1.13)	A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

The BWA (Burrows-Wheeler Aligner) is basically a package used for indexing and mapping sequences against a large reference genome. The BWA-MEM algorithm is used in this study for mapping reads to the reference genome because of its significance in mapping large reads. SAMTools is a suite of programs for manipulating high-throughput sequencing data in SAM format. It contains three separate packages, namely, Samtools, BCFtools, HTSlib. The reason behind including the BWA and SAMTools in this study is to perform the data preprocessing step before calling variants using GATK. The Genome Analysis ToolKit (GATK) is used for calling variants, filtering, and evaluation. The PICARD tool is a package needed by GATK for sequence manipulation. A quality control is usually needed for checking quality of sequences. For this reason, FastQC tool is used in this study to do some quality control checks on raw sequence data. On the other hand, MultiQC tool is utilized to aggregate results from other tools analyses across many samples into a single report. In addition to the above mentioned tools, the full human reference genome was also downloaded to be used in the mapping and other workflow steps. Two paired-end reads were also downloaded which would be the material for variant discovery, as shown Table 3.

The SRR062634 is a Homo sapiens (human) read for whole genome sequence with paired layout. This read was sequenced by Illumina Genome Analyzer II submitted by the Genome Center at Washington University School of Medicine in St. Louis (WUGSC). The read count is 24476109 with Base Count of 4895221800. It was a part of the study of whole genome sequencing of (GBR) British from England and Scotland HapMap population [24].

Table 3 Human Genome and Sequence Reads

No.	Genome/Reads	Description	Obtained from
1	hg38	Human genome	https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
2	SRR062634_1	Forward Read	https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/HG00096/sequence_read/SRR062634_1.filt.fastq.gz
3	SRR062634_2	Reverse Read	https://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/HG00096/sequence_read/SRR062634_2.filt.fastq.gz

The procedure of the variant calling pipeline is illustrated in Figure 1. The whole process can be divided into three main steps:

1) **Data Pre-Processing**
   In this step the human reference genome is indexed, and then the raw unmapped reads (SRR062634_1 and SRR062634_2) are mapped against the indexed reference genome. The resulted raw mapped reads in BAM format

is then processed by marking duplicates and recalibrating the base quality scores. As a result, analysis-ready reads are generated and are subject to be served as the input for the next step.

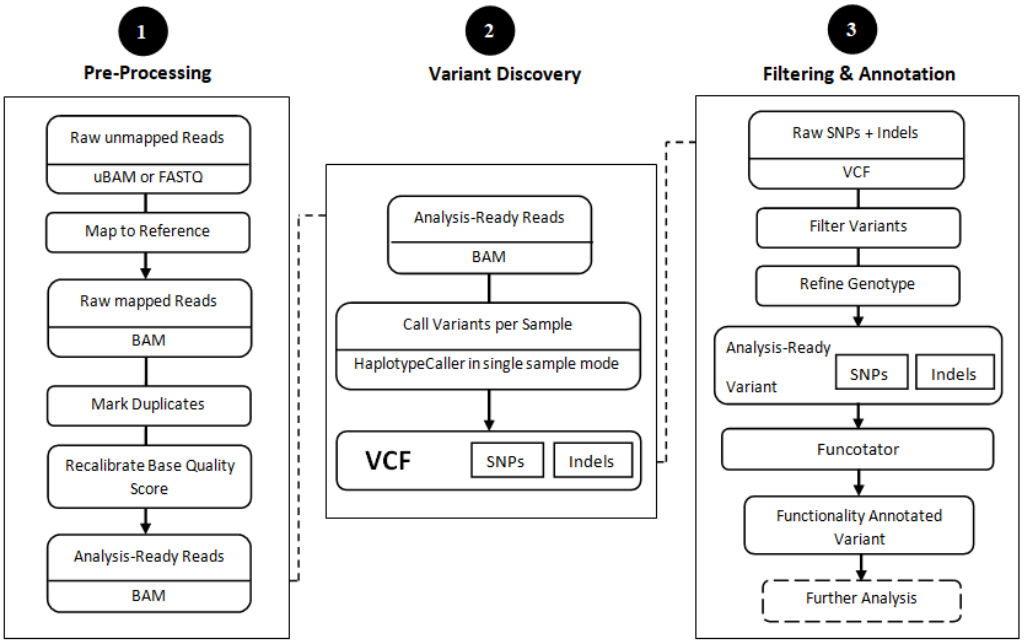

Figure 1. Overview of variant calling pipeline

2) Variant Discovery

Variant calling using GATK is applied to the output of the first step (Data Pre-processing), which is a BAM formatted file. The GATK HaplotypeCaller is used for this purpose, which consequently produces a raw variant call format (VCF) file. Several consequence steps are applied to extract single nucleotide polymorphism (SNPs) and short insertions and deletions (Indels) files.

3) Filtering and Annotation

Upon obtaining the raw SNPs and Indels files from the previous step, filtering and annotation procedure steps are applied. Starting with filtering, the GATK tool is used along with the reference genome to filter variants, refine genotypes, and annotate variants. The annotation of variants is accomplished by using the GATK Funcotator (**Func**tional Ann**otator**) tool. An analysis-ready variant file is obtained in a VCF format. The GATK VariantsToTable tool is used afterwards to produce the last results in a table which can be opened by Excel application for more analysis and reporting process.

RESULTS AND DISCUSSION

The procedure illustrated in Figure 1 has been conducted and results were obtained. In the data preprocessing step, the FastQC tool was applied on both forward and reverse reads (SRR062634_1 and SRR062634_2). The results of this quality control step were obtained in an html file format. The purpose of the FastQC tool is to provide quality control checks on raw data which gives a report on whether this data has any problems before proceeding for further downstream pipeline. Basic statistics of the SRR062634_1 and SRR062634_2 are summarized in Tables 4 and 5, respectively. The sequence length is 100 base pair (bp) and the GC% (Guanine-Cytosine) content is 40%.

Table 4. Basic Statistics of SRR062634_1

Table 5. Basic Statistics of SRR062634_2

Measure	Value
Filename	SRR062634_1.filt.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	24148993
Total Bases	2.4 Gbp
Sequences flagged as poor quality	0
Sequence length	100
%GC	40

Measure	Value
Filename	SRR062634_2.filt.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	24148993
Total Bases	2.4 Gbp
Sequences flagged as poor quality	0
Sequence length	100
%GC	40

The results of per base quality for the both reads are depicted in Figures 2 and 3, respectively. As clearly seen from both figures, all bases pass the poor base sequence quality scores. In Figure 2, base scores ranges between 20 and 38, whereas in Figure 3 base scores ranges between 28 and 38. This indicates that bases in SRR062634_2 achieve higher scores than those of SRR6062634_1. However, both are considered of getting high base scores for further analysis. The averages of scores per sequence are illustrated in Figures 4 and 5. Supporting the previous figures, the average score for SRR062634_1 is about 36, whereas for SRR062634_2 is about 37.

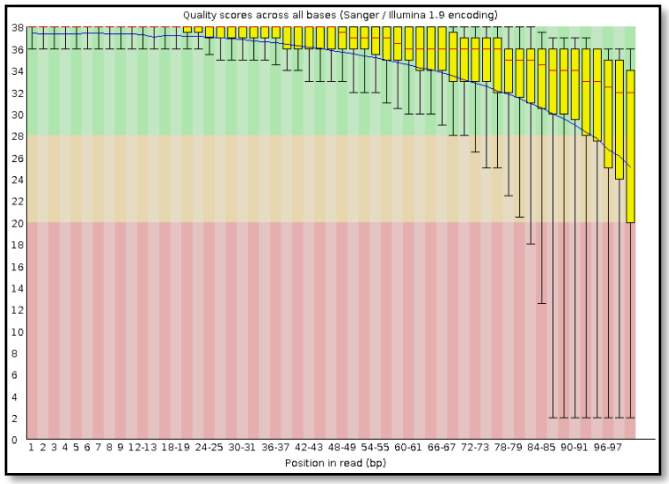

Figure 2. Per base sequence quality of SRR062634_1

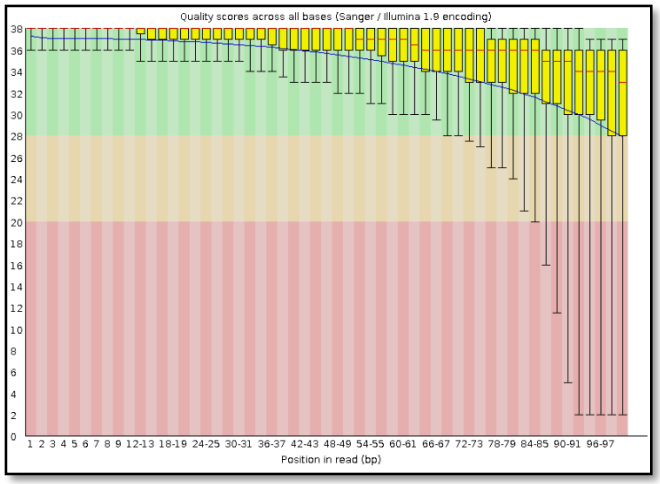

Figure 3. Per base sequence quality of SRR062634_2

The per sequence GC content for SRR062634_1 and SRR062634_2 are presented in Figures 6 and 7, respectively. The GC content distribution is slightly different than the theoretical distribution. This difference was flagged by the FastQC tool. However, no other peaks are observed and so this deviation does not affect the quality of analyzing the both reads. Similarly, the adapter contents for both reads are illustrated in Figures 8 and 9, respectively. This check is a must because if any adapter content is present, then these adapters have to be trimmed and removed before proceeding to aligning the reads. Fortunately, both reads do not have any adapter contents as shown in Figures 8 and 9.

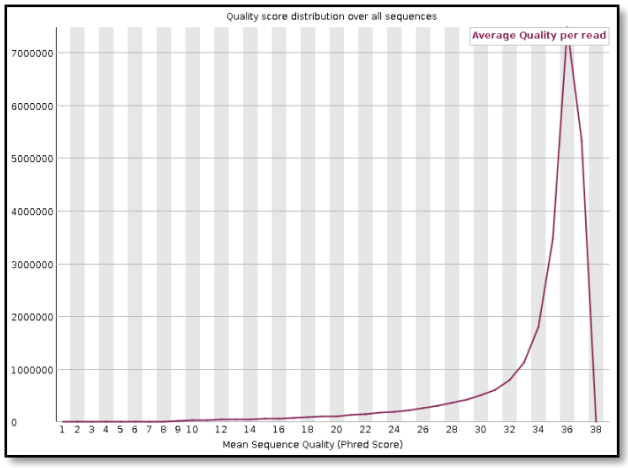

Figure 4. Per sequence quality scores of SRR062634_1

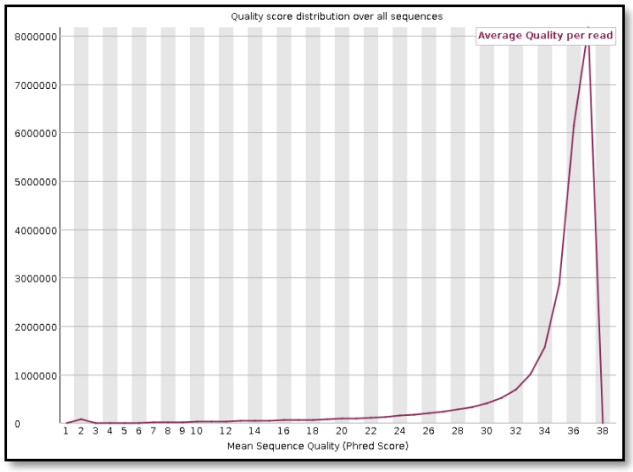

Figure 5. Per sequence quality scores of SRR062634_2

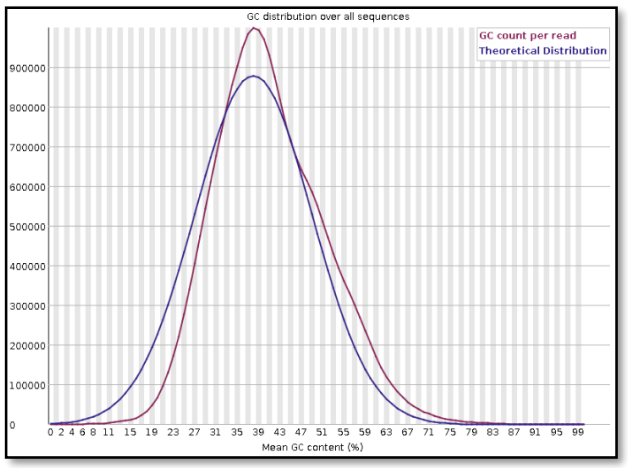

Figure 6. Per sequence GC content of SRR062634_1

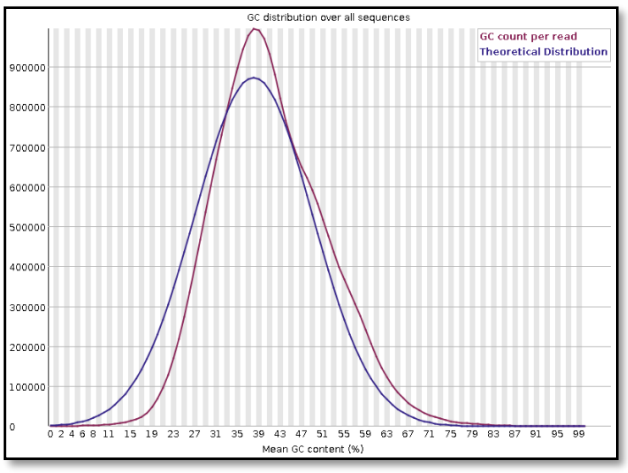

Figure 7. Per sequence GC content of SRR062634_2

Upon making sure that the sampled reads passed the quality control checks, these reads are then aligned (mapped) against the human reference genome (hg38). This produces raw mapped reads files in BAM format. Due to sequencing and amplification errors, duplicate sequences may occur in the reads which may lead to over presentation in the results. These duplicates are not removed but can be marked. Using the gatk MarkDuplicatesSpark is efficient for marking duplicates, which will be placed in the second column of the BAM file as a bitwise flag. When applying the SAMTools flagstat on the reads, results showed about 429229 flagged duplicates. These flagged duplicates will be ignored in the upcoming downstream processing.

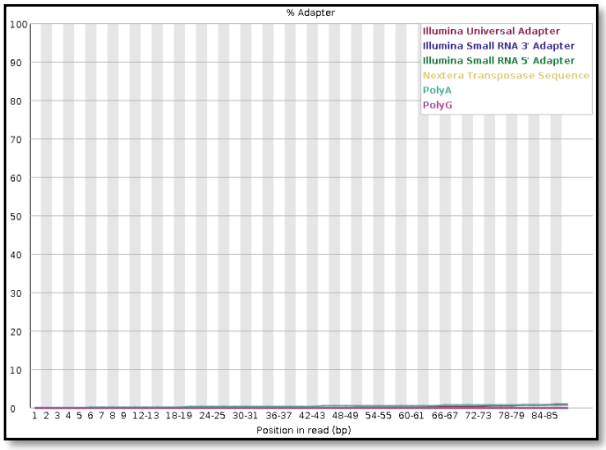

Figure 8. Adapter content of SRR062634_1

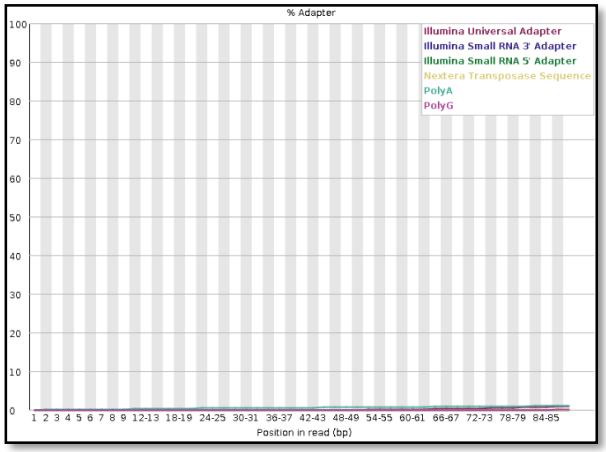

Figure 9. Adapter content of SRR062634_2

The next step in data preprocessing is the base quality recalibration. This step is further divided into two main steps: Firstly, building the model for calibration and secondly, applying the model to adjust the bases quality scores. The GATK BaseRecalibrator and ApplyBQSR tools are used for the two mentioned steps, respectively. To this end, all data preprocessing steps are performed and the analysis-ready reads files in BAM format are generated which are consequently used as an input for the next step known as variant discovery. However, before proceeding to variant discovery step, some metrics can be collected such as alignment summary metrics and insert size metrics using gatk tools. To visualize these collected metrics, MultiQC tool is used which generates an html file report containing the mentioned metrics. A histogram of insert size for all reads is also generated depicting the distribution of the insert size, as shown in Figure 10.

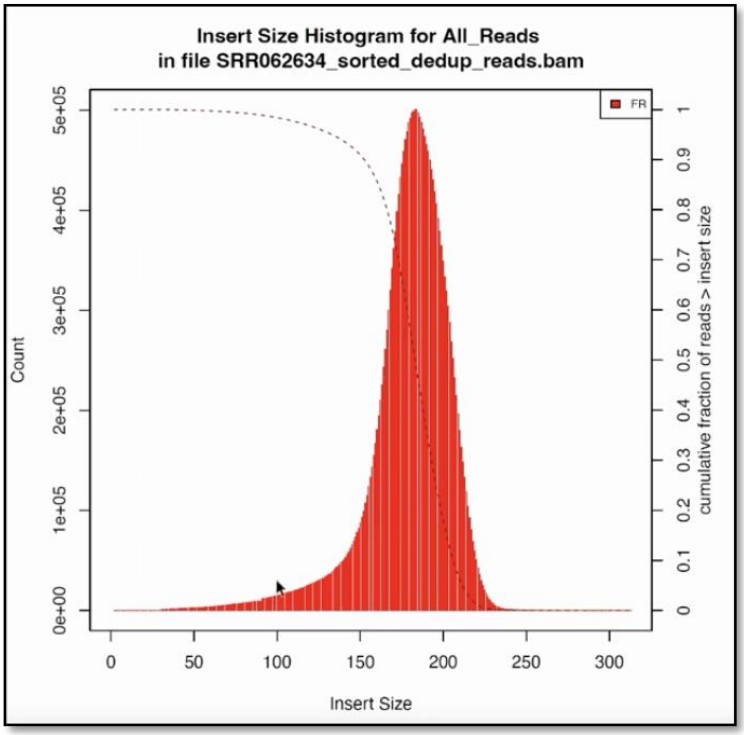

Figure 10. A Histogram of Insert Size for all reads

The variant calling (discovery) step is accomplished by applying the gatk HaplotypeCaller tool on the analysis-ready reads obtained from the previous data preprocessing step. The variants are stored in a VCF file format. To make them distinct and convenient to work on, both SNPs and Indels are extracted from the VCF file. The extracted raw SNPs and Indels are stored in separate VCF files for further processing. After finding the variants, many questions may be asked such as what region is the variant found in? exonic, intronic, regulatory, UTRs? Does this variant affect coding sequence (synonymous or non-synonymous variant)? Does it affect protein function? Is this variant disease causing or benign variant commonly found in the genome? All of these questions and information can be answered and obtained by filtering and annotating the raw SNPs and Indels.

In the filtering and annotation process, we start by filtering the raw SNPs and Indels variants in order to guarantee high confidence variant calls. Hard filtering method is utilized in this study using the gatk VariantFilteration tool. This method relies on thresholds of different filters as specified in [22] such as QualByDepth (QD) filter must be less than 2.0, FisherStrand (FS) filter must be greater than 60, and StrandOddRatio (SOR) filter must be greater than 4.0. In addition, genotype filtration (sample-level) is also performed based on thresholds such as the filtered depth (DP) filter must be less than 10 and the genotype quality (GQ) filter must be also less than 10. Upon completing the filtration process, selection of variants that pass the filters are extracted from both files of SNPs and Indels. These selected variants constitute the analysis-ready variants files which will be the input for the annotation process using GATK Funcotator tool.

The annotation of variants is accomplished by using the gatk Funcotator (**Func**tional Ann**otator**) tool along with available variants data sources from well-known databases such as Genome Aggregation Database (GNOMAD), dbSNP, ClinVar, or Catalogue of Somatic Mutations in Cancer (COSMIC). However, these data sources can be obtained directly as prepackaged data sources with Funcotator, which is the approach we followed in this study. The importance of data

sources is to help the Funcotator tool finding the variants in the analysis-ready-variants files which were obtained from previous steps. In other words, the gatk Funcotator will annotate any variants in the SNPs and Indels variant files that corresponds to variants located at the same column in the variant data source files. These data source files contain the already known variants from previous studies. After annotating the variants, we used gatk VariantsToTable tool to extract the annotated variants fields to a tab-delimited table, which can help in understanding and evaluating the variants presented in the file, as shown in Figure 11. The results obtained in this table answered the questions we set previously. In each row of the table, we can find the region in which the variant was found. In addition, it gives us the type of variant whether it is a SNP or Indels. It shows also if it affects the coding or non-coding sequence. Therefore, the whole pipeline using GATK has proven to call variants and annotate them into functionally annotated variants for any further analysis by scientists.

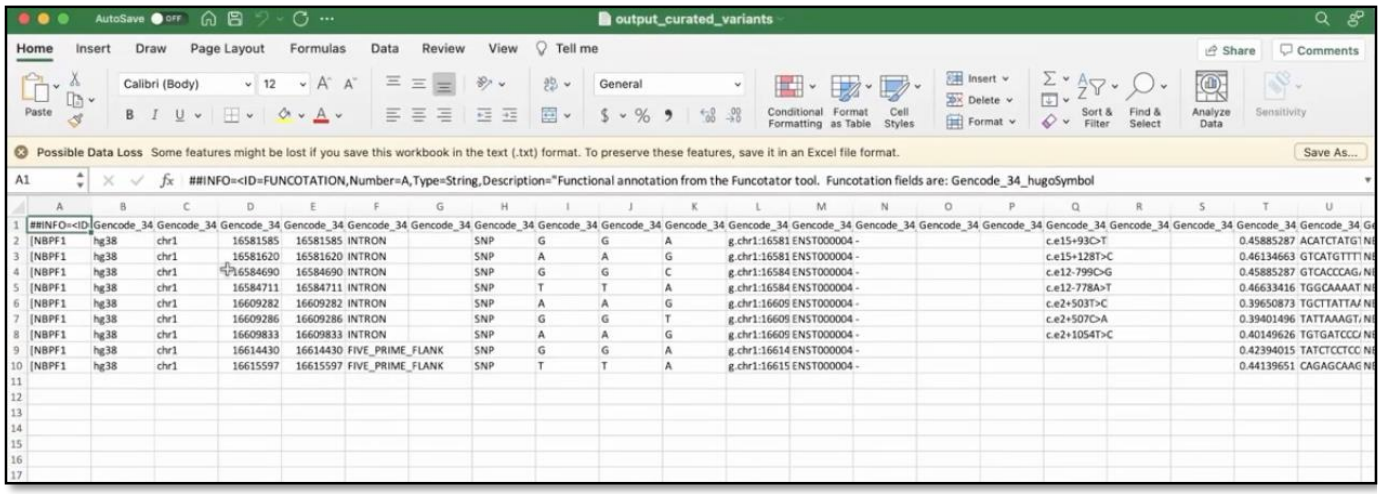

Figure 11. Resulted and Found SNPs Variants

CONCLUSION

Most of the diagnostics performed by researchers in bioinformatics are basically focusing on two variant types: single nucleotide polymorphisms (SNPs) and short insertions and deletions (Indels). In this article, the GATK tool was used for discovering and annotating variants. The results showed that the GATK pipeline was able to discover SNPs and Indels in the studied paired-end reads (SRR062634_1 and SRR062634_2). These variants were annotated using GATK Funcotator tool which generated an Excel file containing all information and specifics regarding these variants. Thus, the GATK is a critical and crucial depth-based tool which can be used for the purpose of discovering and annotating SNPs and Indels variants.

Conflict of Interest

There are no financial, personal, or professional conflicts of interest to declare.

REFERENCES

1. Conine C.C. and Rando O.J. Soma-to-germline RNA communication. Nature reviews Genetics. 2022; 23(2):73-88.
2. Ellegren H. and Galtier N. Determinants of genetic diversity. Nature Reviews Genetics. 2016; 17(7):422-433.
3. Stange M., Barrett R.D., Hendry A.P. The importance of genomic variation for biodiversity, ecosystems and people. Nature Reviews Genetics. 2021; 22(2):89-105.
4. Bergeron L.A., Besenbacher S., Zheng J., Li P., Bertelsen M.F., Quintard B. et al. Evolution of the germline mutation rate across vertebrates. Nature biotechnology. 2023; 615(7951):285-291.
5. Zhao S., Agafonov O., Azab A., Stokowy T., Hovig E. Accuracy and efficiency of germline variant calling pipelines for human genome data. Scientific reports. 2020; 10(1):20222.
6. Hasija Y. Translational biotechnology: A journey from laboratory to clinics.Academic Press. 2021.
7. Stout L.A., Kassem N., Hunter C., Philips S., Radovich M., Schneider B.P. Identification of germline cancer predisposition variants during clinical ctDNA testing. Scientific Reports. 2021; 11(1):13624.
8. Carlson J., Locke A.E., Flickinger M., Zawistowski M., Levy S., Myers R.M. et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. Nature comm. 2018; 9(1):3753.
9. McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010; 20(9):1297-1303.

10. Garrison E., Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012.
11. Browning B.L., Zhou Y., Browning S. A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics. 2018; 103(3):338-348.
12. Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O. et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021; 10(2): giab008.
13. Yun T., Li H., Chang P.-C., Lin M.F., Carroll A., McLean C.Y. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics. 2020; 36(24):5582-5589.
14. Koboldt D.C., Larson D.E., Wilson R.K. Using VarScan 2 for germline variant calling and somatic mutation detection. Current protocols in bioinformatics. 2013; 44(1):15.4. 1-15.4. 17.
15. Cibulskis K., Lawrence M.S., Carter S.L., Sivachenko A., Jaffe D., Sougnez C. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013; 31(3):213-219.
16. Wang K., Li M., Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010; 38(16):e164-e164.
17. Chen X., Schulz-Trieglaff O., Shaw R., Barnes B., Schlesinger F., Källberg M. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016; 32(8):1220-1222.
18. Kim S., Scheffler K., Halpern A.L., Bekritsky M.A., Noh E., Källberg M. et al. Strelka2: fast and accurate calling of germline and somatic variants. Nature methods. 2018; 15(8):591-594.
19. Rausch T., Zichner T., Schlattl A., Stutz A.M., Benes V., Korbel J.O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28(18):i333-i339.
20. Layer R.M., Chiang C., Quinlan A.R., Hall I.M. LUMPY: a probabilistic framework for structural variant discovery. Genome biology. 2014; 15(6):1-19.
21. Barbitoff Y.A., Predeus A.V. Negligible effects of read trimming on the accuracy of germline short variant calling in the human genome. bioRxiv. 2023; 2023.04. 28.538608.
22. Broad Institute. A genomic analysis toolkit focused on variant discovery. 2023. [Accessed: March 24, 2023]; Available from: https://gatk.broadinstitute.org/hc/en-us.
23. Caetano-Anolles D. About the GATK Best Practices. 2023. [Accessed: March 17, 2023]; Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360035894711-About-the-GATK-Best-Practices.
24. ENA. *European Nucleotide Archive*. 2015. [Accessed: March 15, 2023]; Available from: https://www.ebi.ac.uk/ena/browser/view/SRR062634.
25. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research. 2022; 50(W1):W345–W351.

AJMAS
https://journal.utripoli.edu.ly/index.php/Alqalam/index eISSN 2707-7179

إكتشاف وتعريف الطفرات القصيرة في خلايا الخط الإنتاشي عن طريق تطبيق سير العمل للأداة GATK

أشرف بوراوي*، عبدالمنعم عبد الله

قسم الحاسوب، كلية العلوم، جامعة عمر المختار، البيضاء، ليبيا

المستخلص

الخلفية والأهداف. أصبح تحديد الطفرات المتعلقة بالأمراض الوراثية ممكناً بتكلفة معقولة بسبب التقدم في عملية تسلسل الجينوم الكامل (WGSالذي تم تمكينه من خلال تحسينات الجيل التالي من تقنيات التسلسل. يتم اكتشاف الطفرات القصيرة في خلايا الخط الانتاشي الجنسي وكذلك في الخلايا الجسدية الاخرى عن طريق استخدام أدوات التقنية الحيوية. الهدف من هذه الدراسة هو التحقق ودراسة سير عمل اكتساف الطفرات وتعريفها باستخدام الاداة GATK. على وجه الخصوص، تهدف هذه الدراسة إلى اكتشاف الطفرات على مستوى النوكليوتيدات المفردة (SNPsوكذلك عمليات اضافة أوحذف عدد قليل من النوكليوتيدات (Indels) **طرق الدراسة.** لتحقيق هدف الدراسة، تم استخدام العديد من الادوات والحزم في عملية سير عمل اكتشاف الطفرات. حيث تركز هذه الدراسة على تقييم الأداة GATK بالتحديد والادوات المرتبطة معها مثل حزمة PICARD. لتمكين المعالجة المسبقة للبيانات تم استخدام بعض الادوات الاخرى مثل BWA و SAMTools. كذلك تم استخدام الجينوم البشري المرجعي لغرض المحاذاة للتسلسلات قيد الدراسة، والتي من نوع مزدوجة النهاية. **النتائج.** كشفت النتائج في خطوات المعالجة المسبقة للبيانات عن جودة جيدة للتسلسلات قيد الدراسة، وقيم جيدة لمستوى جودة النوكيلوتيدات، وعدم وجود أي مهيئات إضافية تستلزم الحذف. وكذلك بينت النتائج أن قراءات التسلسل تتمتع بنوعية جيدة ويمكن متابعة العمل في عملية اكتشاف الطفرات بطريقة صحيحة. باستخدام الاداة GATK تم اكتشاف الطفرات في خلايا الخط الإنتشاي ووضعها في ملفين منفصلين للطفرات المفردة وطفرات الاضافة والحذف القصيرة. تم تعريف وتهميش هذه الطفرات ووضعها في ملف Excel ليسهل التعامل معها. **الخاتمة.** عند تطبيق سير عمل الاداة GATK تم اكتشاف طفرات الخط الانتاشي المفردة والقصيرة المضافة او المحذوفة وتم انتاج ملف Excel يحتوي على تعريف هذه الطفرات وتحديد أماكنها لتمكين تحليلها أكثر بواسطة البحاث المتخصصين في هذا المجال.

الكلمات الدالة. الخط الإنتاشي، اكتشاف الطفرات، GATK، تعريف وتهميش الطفرات.