IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPREHENSIVE WORKFLOW FOR VARIANT CALLING PIPELINE COMPARISON AND ANALYSIS USING R PROGRAMMING

[1]Mansi Ujjainwal, [2]Preeti Chaudhary

[1]MSc Bioinformatics, [2]Mtech Bioinformatics,
[1]Amity Institute of Biotechnology

[1]Amity University, Noida, India

Abstract:
The aim of the article is to provide variant calling workflow and analysis protocol for comparing results of the two using two variant calling platforms. Variant calling pipelines used here are predominantly used for calling variants in human whole exome data and whole genome data. The result of a variant calling pipeline is a set of variants(SNPS, insertions, deletions etc) present in the sequencing data. Each pipeline is capable of calling its certain intersecting and certain unique variants. The intersecting and unique variants can further be distinguished on the basis of their reference SNP ID and grouped on the basis of its annotation. The number of variants called can be humongous depending upon the size and complexity of the data. R programming packages and ubuntu command shell can be used to differentiate and analyse the variants called by each type of pipeline.

***Index Terms* – Whole Exome Sequencing, Variant Calling, Sequencing data, R programming**

I. Introduction

The Human Genome Project started in 1990, makes up the single most significant project in the field of biomedical sciences and biology. The project was set out to change how we see biology and medicine. The project was set out to sequence complete genome of *Homo sapiens* as well as several microorganisms including *Escherichia coli*, *Saccharomyces cerevisiae*, and metazoans such as *Caenorhabtidis elegans*. The purpose behind such a mega-level project was to revolutionize the fields of medical practice, gene discovery, and biological research and to develop strategic preventive measures for genetic susceptibilities. Apart from these the HGP had many more salient goals. To achieve these, it was necessary to use technology that can give reliable, trustworthy data. The need of sequencing of any genome is to get complete, accurate catalogue of the genes, their transcriptional products, messenger RNA transcripts, and their translating proteins [1].
The HGP was completed in 2003; the result gave us the future reference sequence of the human genome. This sequence provided basis for studying the nature of variations, particularly the phenomenon of Single Nucleotide Polymorphism (SNP). For humans, studying SNP typing is a necessity, it is powerful tool to perform genetic analysis; it enables us to unleash the association of loci at specific sites in the genome with genetic diseases. Learning the association of a particular gene with its disease phenotype provides more information on associated gene products, getting to the cause of the disease and taking preventive measures. The some of the tools utilised by HGP were microarray technology, Sanger's sequencing, Polymerase Chain Reaction (PCR). Duration of HGP faced evolution of sequencing technology, High Throughput Sequencing (HTS) came into existence decades after Sanger's sequencing [2]. HTS is a fast, cheap method of sequencing genome. The need for a cheap and fast sequencing method was fulfilled by development of second-generation sequencing methods, or next generation sequencing (NGS). NGS platforms are capable of performing massive parallel sequencing. In this method, DNA fragment derived from previous steps which involve cutting of DNA hence producing millions of fragments from a single sample are sequenced together. Parallel sequencing method plays a major role behind the HTS methods. This approach allows an entire genome to be sequenced in time lesser than one day [3]. The exome is only 1% of the total genome. Exome sequencing is much more cost effective, fast and fulfils the requirements of the clinician. Exome sequencing-based research is emerging as a popular path for associating coding regions with their phenotypes. NGS is coupled with efficient DNA capture kits which enable exome sequencing. We use DNA capture kits to trap the specified exome sequences [4]. In this process, we come across both common and novel coding regions. Exome sequencing studies enable the unbiased discovery of coding variations for subsequent association testing for complex traits. Analysing the data from sequencing machines, for further understanding. For this purpose over the years pipelines have been developed. It is a common practice to compare such pipelines and their results to understand which works more efficiently. However, each pipeline has its own sets of prerequisites. Availability of tools has made it possible to come up with different workflows performing the same set of actions on a set of given data with variable efficacy. This article provides workflow for two pipelines and an R programming based algorithm on how to compare the results obtained from two pipelines.

II. Materials and Methods

The following tools are to be used in the same order on the sequencing data.

1. Quality check for data: This was done using FastQC software developed by Babraham Institute
Next generation sequencing are capable of generating hundreds of millions of sequences in a single run. So before starting the analysis of these sequences, it becomes necessary to draw the biological conclusions. One should always start with a biological analysis plan and quality control checks to ensure that the raw data is good enough and there are no problems or biases in our data which may interfere with the final outcome. FastQC is open software that facilitates quality control of FASTQ files. This is done by carrying out a QC protocol. The end result gives quality metrics that are expressed on an interactive dashboard. This is designed to summarize individual sequencing runs in an information rich manner [5].

2. Trimmomatic (trimming tool):
The presence of poor quality sequenced data or presence of the technical sequences such as adapters in next-generation sequencing (NGS) data may result in suboptimal downstream analysis. Trimmomatic has two types of modes simple mode and palindrome mode, both of which can be used as per the user requirement. It is a java based tool; works on both paired and single end reads. The input is a fastq file. Advantage is this tool can also work on gun zipped fastq files [6].

3. BWA MEM (alignment of cleaned data with reference sequence):
BWA is an open source package that is used for mapping sequences with low divergence against an outsized reference genome, such the human genome. BWA requires an indexed reference genome before aligning the reads, which can be downloaded from igenome [7].

4. SAMTools (checking statistics of BAM files):
SAMtools is a software package that can be downloaded from online resource (github or SourceForge) .This tool is useful as it is able to perform a variety of functions such as, converting the alignment formats, sorting and merging alignments, removing PCR duplicates, generating per- nucleotide base pair position information in the pileup format, call SNPs (mpileup) and short indel variants. SAMtools is used here to check the statistics of the SAM/BAM file [8].

5. BEDtools (aligning BAM file with capture bed file): BEDTools was developed in the Quinlan laboratory of Utah and is updated by contributions from scientists all over the world. BEDtools intersect is used to find overlapping intervals in various ways. BEDTools intersect was done to get bam output files for each sample and its overlapping capture kit [9].

6. Variant Calling pipelines: Two pipelines are considered here, GATK (Broad Institute) and Sentieon

7. GATK: It is an open source collection of tools for analysing next generation sequencing data. The primary concern of the tools is on variant discovery. The tools are available for individual usage, although they are often used in a collective or a chained fashion together into complete workflows.

8. Sentieon: Sentieon DNA is a paid software and a set of tools provided by Sentieon for analysis on the Next Generation Sequencing data. Sentieon provides replacement softwares to the open source tools. Sentieon comes in a package that drives the whole pipeline, no need to install each and every tool.

7.1 Picard: is a set of tools for manipulating next generation sequencing data. The supported formats include SAM/BAM/CRAM and VCF. It is a java based tool by Broad Institute. It can be downloaded from github or SourceForge. Usage of picard includes adding or replacing readgroups using AddOrReplaceReadGroups (Picard), marking duplicates using MarkDuplicates (Picard), sorting bam file according to coordinate using SortSam (Picard).

7.2 GATK: BaseRecalibrator and GATK -ApplyBQSR generates and apply a recalibration score based on various covariates. Some of these covariates are are read group, reported quality score, machine cycle, and nucleotide context. The BaseRecalibrator generates a recal table (.txt) which is to be utilised as output for the next tool that is ApplyBQSR.

7.3 GATK -HaplotypeCaller: is the tool that will call the variants between the reference genome and the input genome sequence. The vcf file obtained from HaplotypeCaller contained SNPs and INDELs. These are separated further by hard filtering with criteria provided in GATK4 guidelines.

7.4 GATK: Select-variants and GATK Variant-filtration these tools are designed for hard filtering. The first filters the vcf file based on the type of variant is specified or selects a set of called variants (according to the requirement) from a VCF file. For instance, the vcf file generated by HaplotypeCaller contains all types of variants, in order to separate these into SNPs and INDELs we used SelectVariants. SelectVariants is to create two different vcf files, one for INDELs and one for SNPs. One can also filter other variants as per the requirement, as for this project I have taken only INDELs and SNPs in consideration. The Variant filtration is a hard filtering tool, the selected variants (previous step) are filtered on the basis of a specific criteria.

8.1 BWA-MEM (Sentieon driver): Mapping reads to the reference genome, this command is efficient to perform alignment, add readgroups and sort. Sentieon BWA-MEM allows us to skip several steps of the GATK pipeline. It can add readgroups beforehand, sort the reads, convert SAM to BAM thus skipping the SAMtools conversion step.

8.2 LocusCollector and Dedup (Sentieon driver): Two different commands can run and remove and mark duplicates on the sorted and aligned BAM file. The LocusCollector collects the information on duplicates and Dedup removes the duplicates. If second algo is not used, duplicates are only marked. The remove duplicates command will shorten the file size. In GATK, the duplicates were only marked but not removed. But here, the duplicates are removed. There is no difference generated in removing or keeping the duplicates because once they are marked they will be read by the algorithm.

8.3 Indel Realignment (Sentieon driver): This command is used to perform local alignment on the indels in the input file. Indel realignment was removed from GATK 4.x version onwards that is why it couldn't be performed in the GATK pipeline.

8.4 Base quality score recalibration (Sentieon driver): Three commands are used to apply the recalibration and create a report on the BQSR. The recalibration math depends on the platform tag of the ReadGroup.

8.5 algo- Haplotyper (Sentieon driver): A single command applies BQSR calculated in the step above and generates the VCF file. The VCF file will contain the mutations reported in each sample.

8.6 CollectVCMetrics: The CollectVCMetrics algorithm collects metrics related to the variants present in the input file. Sentieon driver does not provide hard filtering tools. GATK hard filtering tools can be used to filter the variants.

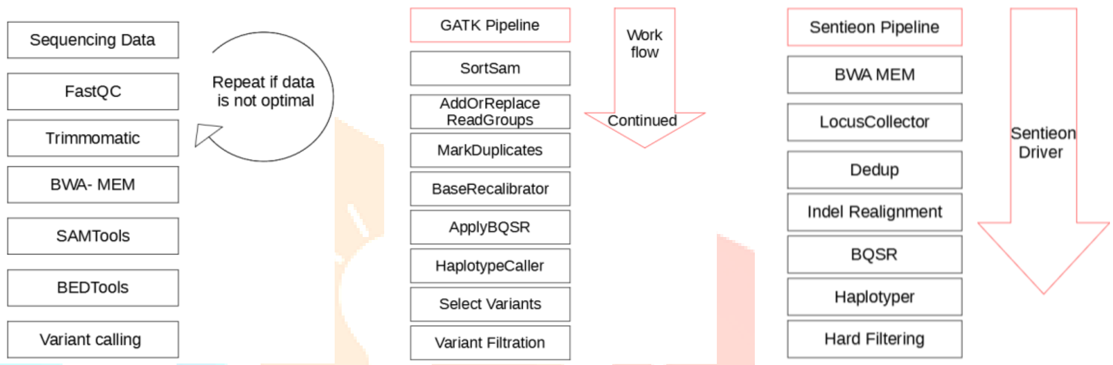


Figure 1.1 Workflow of the VariantCalling Pipelines

9. ANNOVAR: Annotation is performed to get the functional association of the genes or in this case variant recorded. The functional association is stored in different databases. The database can be gene, region or filter based. The ANNOVAR tool was developed to locate the functionally important variants in the pool of high throughput sequencing datasets. This tool is used for annotation, that is gathering the reported information regarding a variation. The variation can be an insertion, deletion, a single nucleotide variation or a copy number variation. The information is gathered from different sources such as the 1000 genomes project. There are plenty of annotation databases available that can be utilised together using ANNOVAR. Comparing the variants obtained from each of the pipeline after annotation, using their reference SNP ID (rsid). This approach will further aid in filtering out the novel variants called by each of the pipeline with their annotation [10].

10. FILTERING & PLOTTING: novel variants detected by the two pipelines. This can be done using ubuntu command shell and R programming. Use of statistical programming like R handles large datasets more efficiently without loss of quality, it also fastens the process. Using the "awk" command the required columns from the annotated VCF files can be separated. These columns can also be separated using R programming. To read the vcf files use the library vcfR. The columns can be separated using separate function. Using the drop function the unnecessary columns can be dropped. Using Ubuntu command shell to do this part is an easier and time efficient procedure. The new files are to be converted to csv or excelformat for further analysis. The package to read and write excels or csv file is "tidyverse". The functions in tidyverse allow comparing two files and generating another file that contains the result. The functions of tidyverse are anti-join and semi-join. The package can be downloaded with following command #install. packages ("tidyverse") Tidyverse has further libraries and functions that contains function that are similar to excel or LibreCalc.The ggplot package of R is to generate a graph depicting the number of rs_ids that were corresponding to the pathogenic behaviour. The type of clinical significance was discussed earlier.

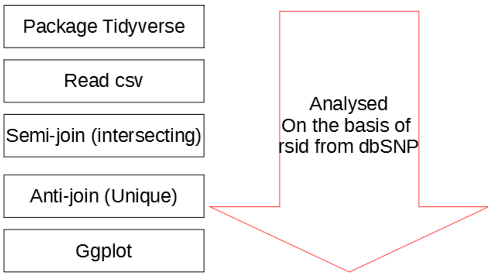
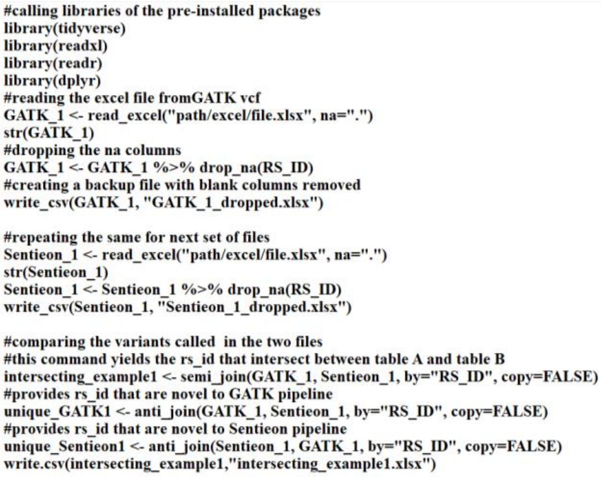


Figure 1.2 Workflow of the R programming algorithm



```
#calling libraries of the pre-installed packages
library(tidyverse)
library(readxl)
library(readr)
library(dplyr)
#reading the excel file fromGATK vcf
GATK_1 <- read_excel("path/excel/file.xlsx", na=".")
str(GATK_1)
#dropping the na columns
GATK_1 <- GATK_1 %>% drop_na(RS_ID)
#creating a backup file with blank columns removed
write_csv(GATK_1, "GATK_1_dropped.xlsx")

#repeating the same for next set of files
Sentieon_1 <- read_excel("path/excel/file.xlsx", na=".")
str(Sentieon_1)
Sentieon_1 <- Sentieon_1 %>% drop_na(RS_ID)
write_csv(Sentieon_1, "Sentieon_1_dropped.xlsx")

#comparing the variants called  in the two files
#this command yields the rs_id that intersect between table A and table B
intersecting_example1 <- semi_join(GATK_1, Sentieon_1, by="RS_ID", copy=FALSE)
#provides rs_id that are novel to GATK pipeline
unique_GATK1 <- anti_join(GATK_1, Sentieon_1, by="RS_ID", copy=FALSE)
#provides rs_id that are novel to Sentieon pipeline
unique_Sentieon1 <- anti_join(Sentieon_1, GATK_1, by="RS_ID", copy=FALSE)
write.csv(intersecting_example1,"intersecting_example1.xlsx")
```

Figure 1.3 Sample programming workflow

III. Discussion

The interpretation of results from sequencing data files is a laborious process and the usual tools such as ubuntu command shell and Microsoft office excel functions often are not able to handle such large amount of data efficiently. R programming package tidyverse provides an alternative. The above article discusses the tools that can be used in variant calling pipeline to the analysis of the variants obtained. The results can be interpreted as required.

References

1. Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology.*Science*,*300*(5617), 286-290.

2.Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., ... & Wu, L. F. (2001). Experimental annotation of the human genome using microarray technology.*Nature*,*409*(6822), 922.

3.Watson, J. D. (1990). The human genome project: past, present, and future.*Science*,*248*(4951), 44-49.

4. Bentley, D. R. (2000). The human genome project—an overview.*Medicinal research reviews*,*20*(3), 189-196.

5. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2015), "FastQC," https://qubeshub.org/resources/fastqc

6. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data

7. "Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

8. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, 15 August 2009, Pages 2078–2079,https://doi.org/10.1093/bioinformatics/btp352

9. Aaron R. Quinlan, Ira M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, Volume 26, Issue 6, 15 March 2010, Pages 841–842,https://doi.org/10.1093/bioinformatics/btq033

10. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data *Nucleic Acids Research*, 38:e164, 2010