Course Code & Name: ENS 210 Computational Biology

Semester: Fall 2023

Date: 15.11.2023

WHOLE EXOME SEQUENCING (WES) IN HUMAN PATIENT DATA

Team Members:

Mehmet Barış Tekdemir - 29068

Türkan Doğa Gizer - 29072

Merve Karacaoğlu - 32282

Nil Özde Özgen - 30802

Samet Mert - 23511

Dila Karataş - 28852

Emir Ay -

A. INTRODUCTION

Whole Exome Sequencing is a sequencing technique that is used to identify potential disease origins in the field of bioinformatic (*Whole Exome Sequencing*, 2022). By analyzing the regions of the genome that codes proteins, WES makes it possible to identify genetic variations that are responsible for a wide range of diseases and disorders (*Whole Exome Sequencing*, 2022). This project analyzes Whole Exome Sequencing data to determine the genetic basis of specific phenotypic traits.

The reference genome used in WES indicates the healthy individual's genome and the patient genome indicates the patient's DNA sequence (Majewski et al., 2011). A comparison of the reference genome with the patient's genome is used to determine the differences (Majewski et al., 2011). Genetic variations and risk factors causing a wide range of diseases, including cancer, inherited disorders, and Freeman-Sheldon syndrome, are found using Whole Exome Sequencing (WES) (Rabbani et al., 2013). For the diagnosis of genetic abnormalities, individualized treatment planning, and identifying carriers of genetic diseases, WES is a highly effective and precise technique (Majewski et al., 2011).

The project aims to uncover possible mutations by using WES as a tool to analyze protein-coding sections of the genome, identifying differences between the patient human genome and the reference genome. This project was accomplished through a systematic framework that included data collection, quality control, data processing, sequence alignment, variant calling, and variant annotation and prioritization (**Fig.1**). In this project, the WES method was implemented using many different tools. GATK, SAMtools, Cutadapt and BWT are the main tools used in this project. At the end of the project, the 20 most critical and possible disease-causing mutations were determined and picked, after conducting an extensive literature review on the pathogenicity of observed mutations from the patient genome.

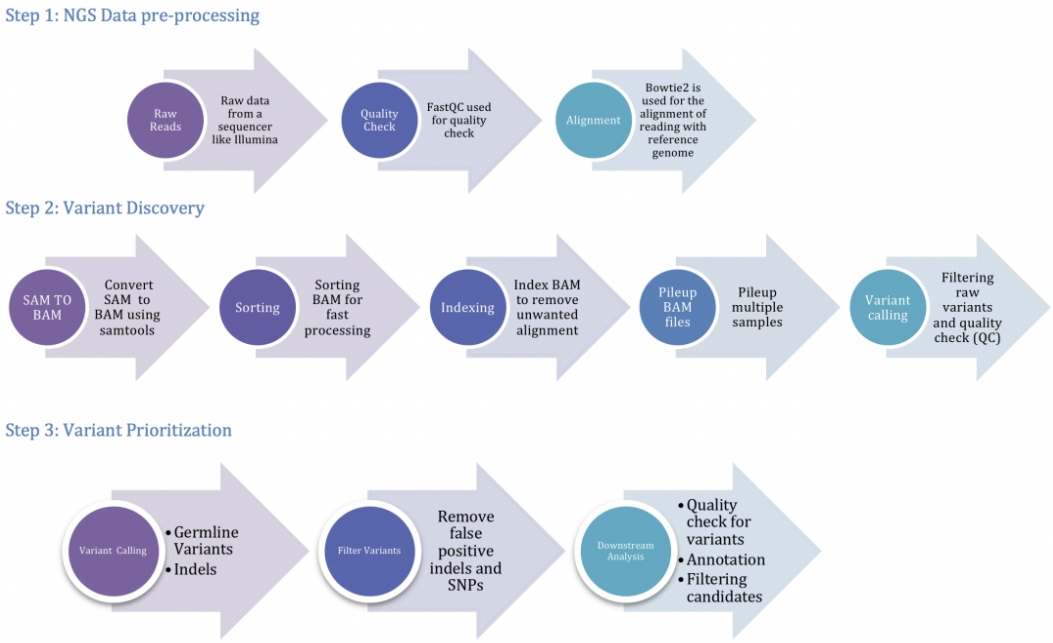

Fig.1 The pipeline of Whole Exome Sequencing (Meena et al., 2018)

B. RESULTS

In this project, we applied whole exome sequencing to identify disease-causing mutations in a patient's genome. A systematic framework that included data collection, quality control, data processing, sequence alignment, variable calling, variable annotation and variable prioritization and filtration was followed. We used various tools to analyze the data, including GATK, SAMtools, Cutadapt, and BWT. The read group could not be added due to problems that occurred while adding the read group, which should be added before the variant calling. Therefore, the variant calling step could not be started. As a result, the project was terminated before proceeding to the variant calling step.

C. MATERIALS & METHODS

In this project, the workflow for Whole Exome Sequencing followed by the control of the raw data's quality, preprocessing of the data with removing the adaptors, aligning the data with reference genome data, removing read duplicates as post alignment processing, performing variant calling and variant annotation, then filtering and prioritizing the variants according to the literature review carried out in the study.

As the first step of the workflow, quality control of the data made with the FastQC tool. The main output type used to describe the quality assessment's findings is an HTML file. In the HTML file, Per base sequence quality, Per sequence quality scores, Per base sequence content, Per sequence GC content, Per base N content, Sequence Length Distribution, Sequence Duplication Levels, Overrepresented Sequences, and Adapter Content analyzed as Passed, Warning and Failed. In the second step of the workflow, adaptors which were used in the sequencing were removed from the data with the Cutadapt tool. Adaptor sequences used in the trimming were Sanger / Illumina 1.9 adaptors. The trimmed data obtained in the FASTQ file format.

In the third step of the workflow, reference genome data which is accessed from the European Bioinformatics Instıtue, and the patient data aligned with the BWA tool. The reference genome is chosen as hg38 Homo Sapiens genome. As a result of the alignment, one file was obtained in the SAM file format. In the fourth step of the workflow, post alignment processing was performed. In the first part of the post alignment processing, duplicates in the data are removed with the help of SAMtools tool. The data obtained in the BAM file format.

In the base quality score recalibration, hg38 Homo Sapiens reference genome data, lastly duplicate removed data and The Single Nucleotide Polymorphism Database (dbSNP) data which is accessed from the European Bioinformatics Instıtue used while performing it. After finishing the post alignment processing, as the fifth step of the workflow, variant calling was performed with the GATK4/SAMtools tool. The

commands were run, but due to the errors that occurred, the commands did not work and the steps after the post alignment process, which are variant calling, variant annotation and variant filtration and prioritization could not be completed.

D. DISCUSSION

In this experiment WES technique was used to determine the possible mutations in a given patient genome sequence. The project was carried out in multiple steps. Each step requires a specific tool and command.

Data Quality Control

The raw data quality in Whole exome sequencing technique is essential in order to provide more accurate data analysis. In this experiment, in the raw data quality control step, FastQC tool was used. After performing the quality control of the raw data with the FastQC tool, two HTML files were obtained. The HTML file consists of the pass, fail, and warning analyzes. In the Patient 2 FastQC Report, “Per base sequence content” got a warning, and “Per sequence GC content” failed. In the Patient 1 FastQC Report, “Adapter content” got a warning. Therefore, it can be concluded that the quality of the data is observed below a certain threshold.

Data Preprocessing

Data Processing is the second step of WES which involves several key processes to transform the initial raw sequencing data into a format suitable for downstream analysis. For data processing, Cutadapt is used to trim the adapter sequences. Cutadapt’s flexibility and customization options make it an effective tool in handling diverse datasets.

Sequence Alignment

The primary objective of sequence alignment is to map the short reads generated from the exome sequencing to a reference genome. This process identifies the genomic location of each sequenced fragment, allowing to understand where the reads originate in the reference genome. For sequence alignment the BWT tool was used while aligning reference genome sequence and the patient genome sequence.

Post-alignment Processing

The aim of the post alignment process is to ensure the accuracy of aligned sequences. Indexes, file conversions and duplication were applied in this process. This step is crucial to improve the accuracy and the quality of the data. However, In Del Realignment of the data was not performed due to the GATK4 tool which does not support this process, and GATK3 tool can perform this process.

Variant Calling

Variant calling is another crucial step in WES. It basically determines where the reference genome and the patient genome differ from each other. In this project, due to a problem with the BAM file, the read groups could not be added into data. That's why variant calling was not applied. Variant annotation and variant filtration and prioritizatiın steps also could not be applied.

E. REFERENCES

Meena, N. K., Mathur, P., Medicherla, K. M., & Suravajhala, P. (2018). A Bioinformatics Pipeline for Whole Exome Sequencing: Overview of the Processing and Steps from Raw Data to Downstream Analysis. *Bio-protocol*, *8*(8). https://doi.org/10.21769/bioprotoc.2805

Bioinformatics Workflow of Whole Exome Sequencing - CD Genomics. (n.d.). https://www.cd-genomics.com/bioinformatics-workflow-of-whole-exome-sequencing.html

Whole exome sequencing. (2022, October 30). Yale Medicine. https://www.yalemedicine.org/conditions/exome-sequencing#:~:text=Whole%20exome%20sequencing%20is%20a,cancer%20diagnosis%20and%20prenatal%20screening

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., & Jabado, N. (2011). What can exome sequencing do for you? *Journal of Medical Genetics*, *48*(9), 580–589. https://doi.org/10.1136/jmedgenet-2011-100223

Rabbani, B., Tekin, M., & Mahdieh, N. (2013). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, *59*(1), 5–15. https://doi.org/10.1038/jhg.2013.114

Tian, S., Yan, H., Kalmbach, M., & Slager, S. L. (2016). Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016-1279-z