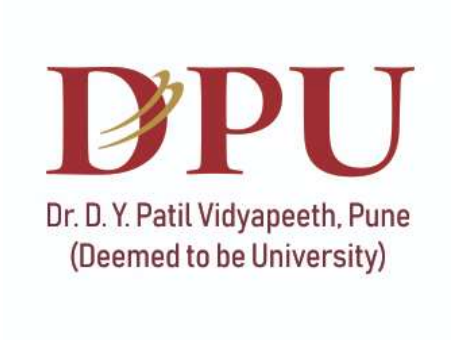


TITLE OF THE PROJECT

DEVELOPMENT OF A COMPUTATIONAL PIPELINE FOR THE IDENTIFICATION OF PATHOGENIC VARIANTS ASSOCIATED WITH IMMUNODEFICIENCY DISORDERS.

A PROJECT SUBMITTED TO DR. D. Y. PATIL VIDYAPEETH (DEEMED TO BE UNIVERSITY) IN PARTIAL FULFILLMENT OF FIVE YEARS FULL-TIME INTERGRATED PROGRAMME M.TECH. (BIOTECHNOLOGY)

SUBMITTED BY

DESHPANDE VISHWESH VIRAJ

UNDER THE GUIDANCE OF

DR PREETI PATEL

MyGenomeBox India Private Limited,
Awfis Office Solutions, 3rd floor, ABIL Imperial, Baner, Pune, Maharashtra 411045

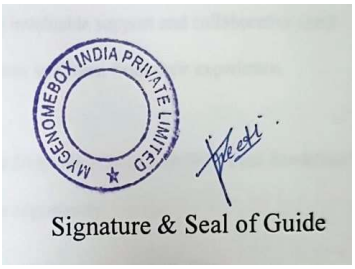
DR. D. Y. PATIL BIOTECHNOLOGY AND BIOINFORMATICS INSTITUTE

TATHAWADE, PUNE – 33

MAY 2024

CERTIFICATE

This is to certify that Mr. **Deshpande Vishwesh Viraj** has prepared this Project titled "**Development of a Computational Pipeline for the Identification of Pathogenic Variants Associated with Immunodeficiency Disorders**", under my guidance and to my satisfaction, in fulfilment of the requirement for five years M.Tech. (Integrated) Biotechnology Degree programme.

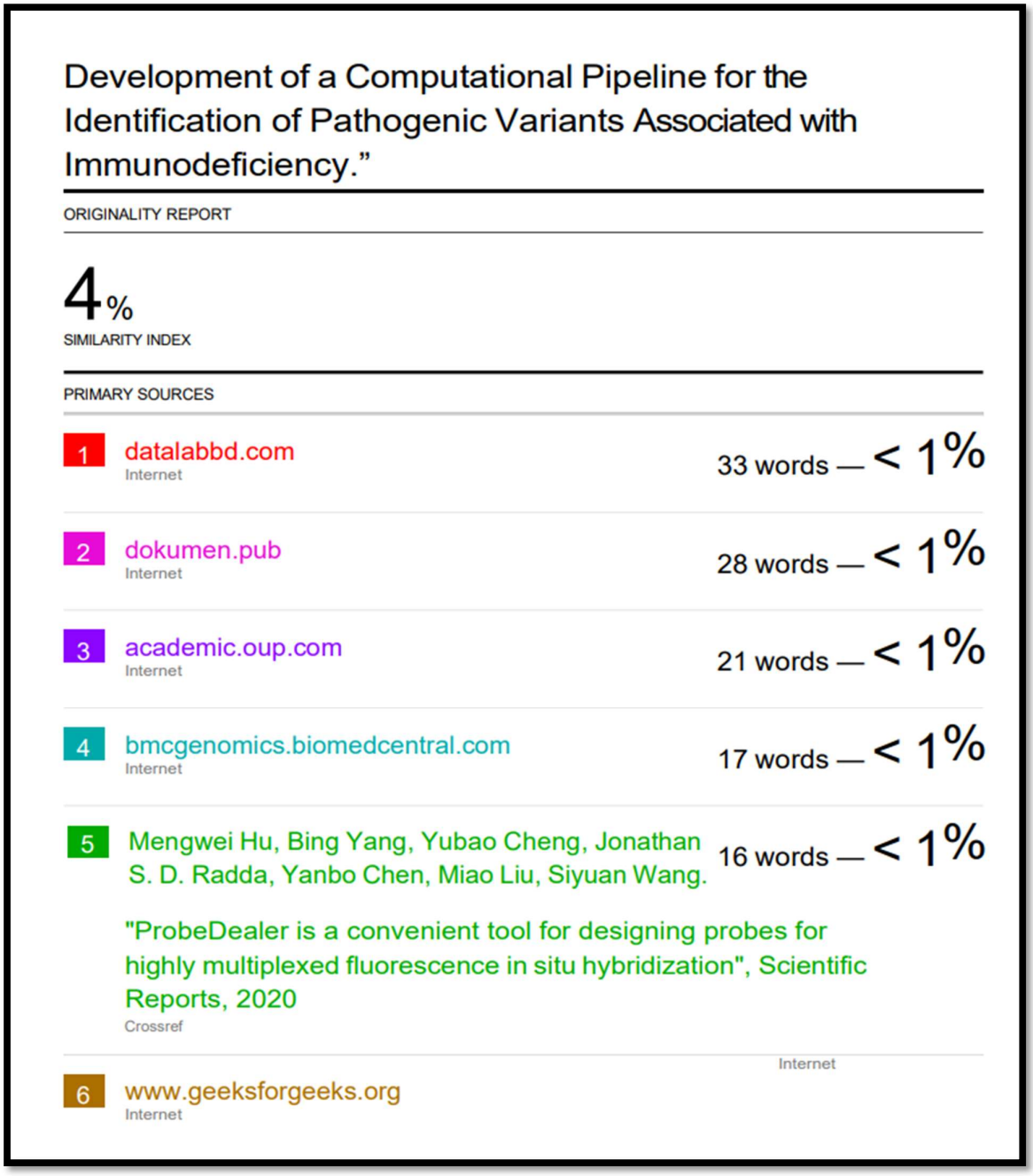


Signature & Seal of Guide

Guided By
Dr. Preeti Patel
MyGenomeBox India Private Limited,
Awfis Office Solutions, 3rd floor, ABIL Imperial, Baner, Pune, Maharashtra 411045

Director
Dr. D. Y. Patil Biotechnology and Bioinformatics Institute
Tathawade, Pune - 33

Plagiarism Report



Development of a Computational Pipeline for the Identification of Pathogenic Variants Associated with Immunodeficiency."

ORIGINALITY REPORT

4%
SIMILARITY INDEX

PRIMARY SOURCES

1	datalabbd.com Internet	33 words — < 1%
2	dokumen.pub Internet	28 words — < 1%
3	academic.oup.com Internet	21 words — < 1%
4	bmcgenomics.biomedcentral.com Internet	17 words — < 1%
5	Mengwei Hu, Bing Yang, Yubao Cheng, Jonathan S. D. Radda, Yanbo Chen, Miao Liu, Siyuan Wang. "ProbeDealer is a convenient tool for designing probes for highly multiplexed fluorescence in situ hybridization", Scientific Reports, 2020 Crossref	16 words — < 1%
6	www.geeksforgeeks.org Internet	Internet

Acknowledgment

This project is completed under the company name MyGenomeBox Private Limited. This work has been done under the guidance of Dr Preeti Patel and the co-guidance of Dr Amit Goyal.

I would like to extend my sincere gratitude to my guide and co-guide Dr. Preeti Patel and Dr. Amit Goyal for giving me the opportunity to work under them and for their invaluable support and collaborative spirit throughout my internship. I was honoured to work under them and learn from their experience.

I would like to express my deepest gratitude to our Director Dr Neelu Nawani, Dr D.Y. Patil Biotechnology & Bioinformatics Institute, Pune for providing me with this opportunity.

I would like to thank Dr Somya Basu for guiding me during my internship search phase.

Lastly, I would like to thank my family and friends who supported me in this journey.

Preface

This project, entitled **"Development of a Computational Pipeline for the Identification of Pathogenic Variants Associated with Immunodeficiency Disorders."** is submitted to the Dr. D.Y. Patil Biotechnology and Bioinformatics Institute, Pune. It contains work done by me during my internship at MyGenomeBox India Private Limited, Pune.

MyGenomeBox is a leading genomics and bioinformatics company, whose headquarters are located in South Korea and the United States of America. They are an organization dedicated to advancing personalised medicine and healthcare solutions.

During my internship at MyGenomeBox India, I had the honour of learning and executing a wide range of bioinformatics techniques and technology. I studied next-generation sequencing data analysis, and learned how to identify variants and annotate them using The Genome Analysis Toolkit (GATK). I improved my Python programming skills by writing various small scripts. Additionally, I was also involved in content writing for reports and tests and curating genetic databases, which improved my data manipulation skills using Excel.

During the period of my internship, I was involved in the development of a bioinformatics pipeline for the detection and reporting of pathogenic variations connected to primary immunodeficiency diseases. This project will help their user detect the susceptibility to primary immunodeficiency in the form of a clinical report.

Abbreviations

ALT - Alternate Allele

ACMG - American College of Medical Genetics

BAM - Binary alignment/map)

BWA - Burrows-wheeler aligner

CMA - Chromosomal Microarray Analysis

DNA - Deoxyribonucleic acid

dbSNP - Single Nucleotide Polymorphism Database

EDGC - Eone Diagnomics Genome Center

GATK - Genome Analysis toolkit

GSA - Gene set analysis

GIAB - Genome in a Bottle

GWAS - Genome Wide Association Studies

IT - Information technology

JSON - JavaScript Object Notation

Linux - Lovable Intellect Not Using XP

MGB – MyGenomeBox

NGS - Next Generation sequencing

NIPT - Non-invasive Prenatal Testing

PID - Primary immunodeficiency

QC - Quality control

REF - Reference allele

SAM - Sequence alignment/map

SID- Secondary Immunodeficiencies

SNP - Single Nucleotide Polymorphism

TB – Tuberculosis

WGS - Whole genome sequencing

CONTENTS

Overview of the Organization

1.1 Details of MyGenomeBox

MyGenomeBox stands as a leading force in genomic data, establishing the world's first shared economy platform centered around DNA. This concept is new and innovative, which allows the creation of an ecosystem that uses a highly secure cloud platform to integrate DNA data with a wide range of DNA-focused applications. MyGenomeBox's fundamental idea is to investigate, learn about, and acknowledge DNA's enormous potential as a dynamic, educational, and powerful tool for developing a variety of applications and to decipher the life secrets that are hidden in the DNA. With this method, we may help turn a person's genetic code into a blueprint for insightful understanding and useful solutions.

1.1.1 Role and Services of MyGenomeBox: -

The company provides an in-house platform that offers a secure cloud infrastructure for hosting genomic data online. Its primary function involves the transformation of genomic sequenced DNA data into a user-friendly, valuable, and application-ready format. The applications created by this platform ensure and provide a way to store personal DNA information within a highly secure cloud environment. Furthermore, this stored DNA information is readily compatible with leading genomic analysis services and companies specializing in genomic applications worldwide. The platform ensures not only security but also the ease of access to the stored DNA data, thereby establishing a robust foundation for potential collaboration with various genomic entities. As a pivotal aspect of my research, I intend to collaborate with this platform to explore its capabilities and contribute to the advancement of genomic research and analysis.

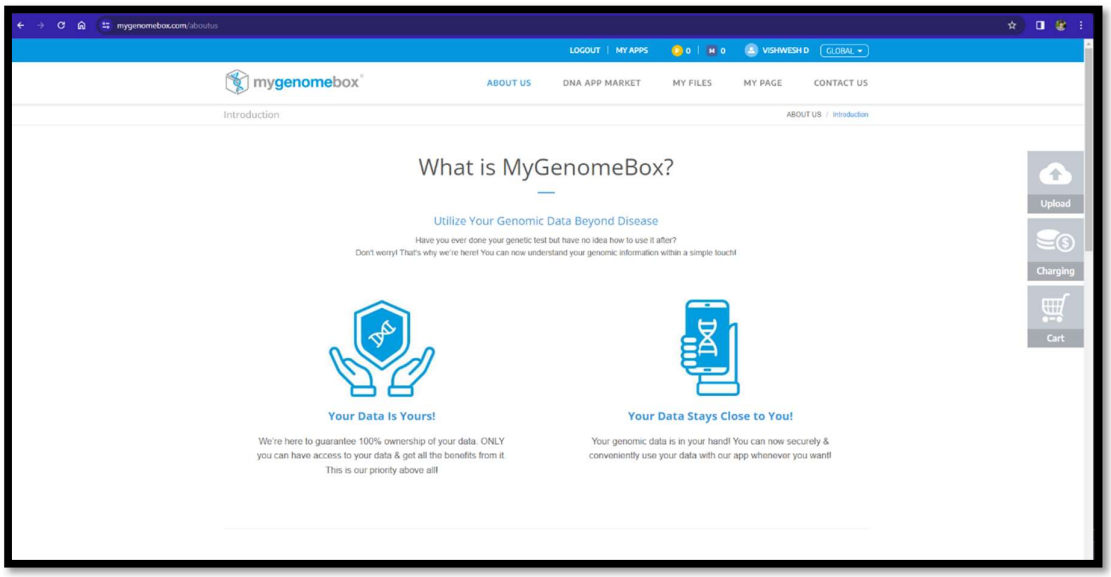


Figure 1- MyGenomeBox Website: https://www.mygenomebox.com/store

1.2 Details of Various activities of the organization

1.2.1 Sequencing Services:

MyGenomeBox, in collaboration with EDGC (Eone Diagnomics Genome Center), offers sequencing services. These services include:

- Non-invasive Prenatal Testing (NIPT),
- Chromosomal Microarray Analysis (CMA),
- Whole-genome Sequencing.

The Process:

1. Sample Collection – MyGenomeBox has partnerships with leading hospitals and research institutions across South Korea and the United States to collect samples for testing.
2. Sequencing – Samples are then sequenced in lab facilities located in South Korea; this generates a vast amount of raw genetic data.
3. Analysis and Interpretation – Once the data is generated, it is transferred to India, where the Bioinformatics team utilizes computational tools and algorithms to do analysis.

The Bioinformatics team is also actively involved in developing new genomic tests, such as the Oncology risk assessment test and more. The team is continuously exploring innovative topics to generate reports and solutions that have a positive impact on patient care.

1.2.2 Web development and Application development

MyGenomeBox has a dedicated IT team that is responsible for developing mobile applications (android and iOS) and web/platform applications. Its department is currently engaged in two exciting live projects:

- **Epi-Clock –** This project aims to determine patient's biological age based on their genetic profile.

- **Uriwell –** This project aims to find non-invasive methods for determining a person's health profile through testing their urine.

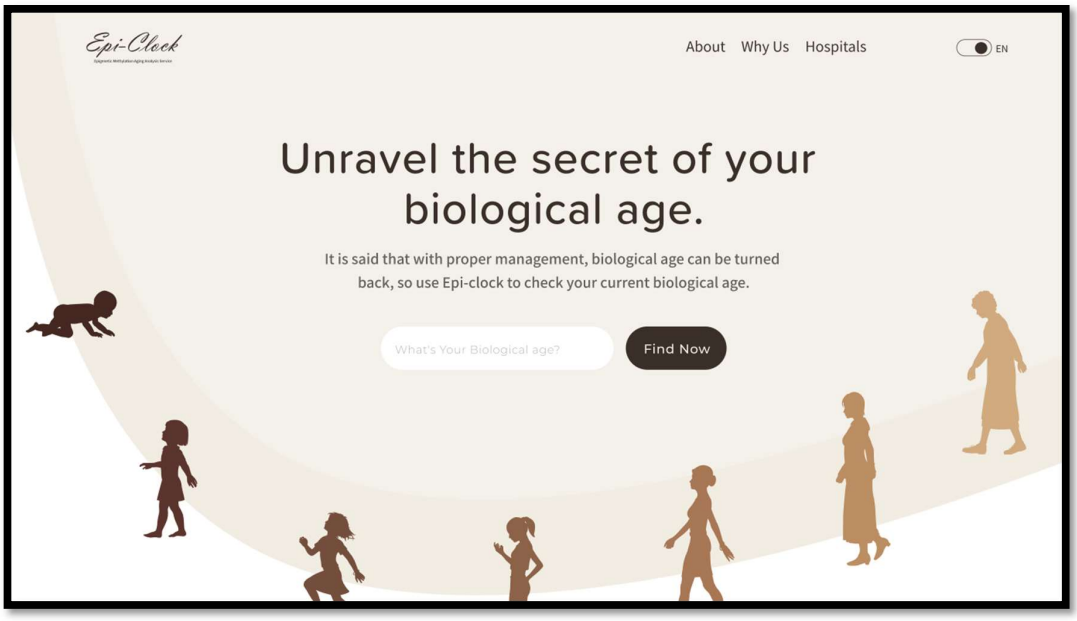


Figure 2- Epiclock Website Homepage

1.2.3 Apps developed By MyGenomeBox

There are over 100 live and working applications on MGB platform. Some of them I have provided below: -

Tuberculosis Susceptibility

Tuberculosis (TB) is a global health threat and is caused by the bacterium Mycobacterium tuberculosis. Susceptibility to TB varies from person to person due to different combinations of genetic and environmental factors. The risk of tuberculosis is increased by environmental factors, which include living in close quarters, having a compromised immune system, and coming into intimate contact with an infected person. Genetic differences in genes such as NRAMP1, IFNG, and IL12B can change the immune system's response to the bacteria, which can impact the risk of developing active tuberculosis as well as the severity of the illness. Environmental variables are typically more relevant than genetic factors, even if they do play a role, development of TB is not guaranteed by genetic factors. Determining an individual's genetic markers can assist in identifying people who are more susceptible, supporting focused preventative initiatives, and perhaps lowering the spread of tuberculosis.

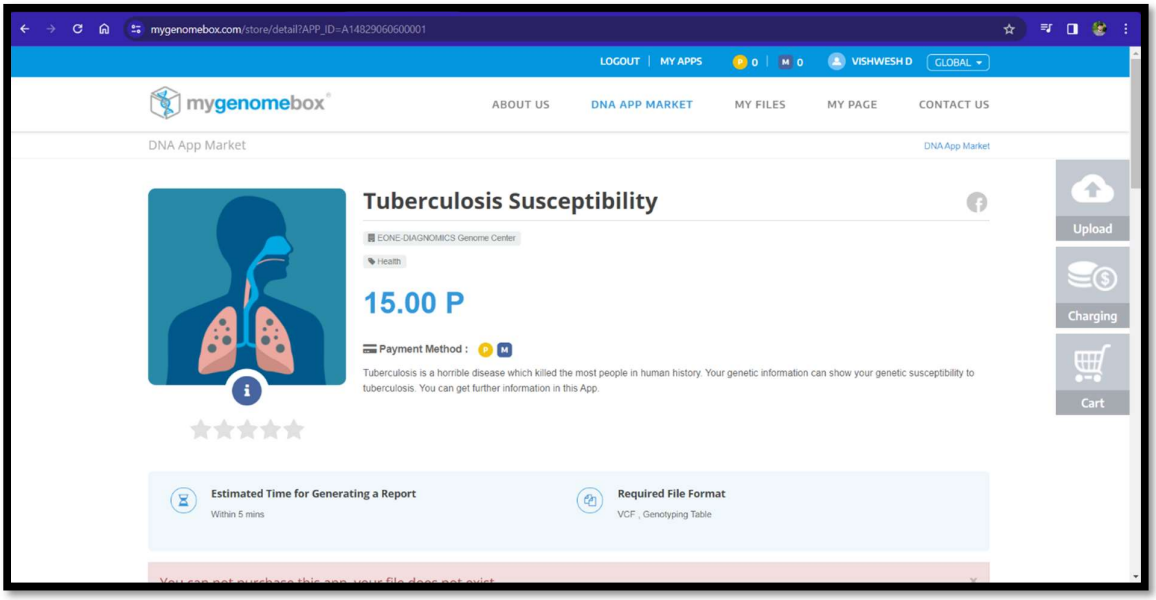


Figure 3 – Tuberculosis Susceptibility App

Coronavirus risk assessment report.

Evaluating both genetic and environmental components is crucial to accurately identifying COVID-19 risk factors. Although COVID-19 may infect everyone, each person's susceptibility and degree of infection differ depending on their genetic composition. Numerous genetic indicators, including polymorphisms in the ACE2 receptor and the HLA gene complex, have been linked by scientific studies to the results of COVID-19. The ACE2 receptor allows the virus to enter human cells, and genetic differences in this receptor can affect an individual's susceptibility and the degree of infection. Like this, changes in HLA genes affect how well the immune system fights off the infection. For developing public health policies and treatment plans, it is essential to comprehend the relationship between genetics and COVID-19 susceptibility. With the use of this genetic data, those who are more susceptible can be identified, facilitating the creation of specialized treatments and focused preventative actions.

Figure 4 – Coronavirus risk assessment App

1.3 Details of Organizational Structure and Department-wise Workflow

Hierarchical Structure of MyGenomeBox: -

CEO

The Chief Executive Officer of MyGenomeBox is Y.T. Park. CEO is the person who is responsible for all the major decisions and oversees in every department, and ensures the company is running smoothly. He is the main man; all the decisions go through him. He is the highest level of management in Cooperate World.

Director

Director and India Head of MyGenomeBox is Dr. Amit Goyal. He is responsible for all the activities at the India branch. He makes sure the organization is going in the right direction and make sures every objective is complete.

He is the Director of IT and Bioinformatics Department at MyGenomeBox India.

MyGenomeBox India is divided into two parts:

1. **Bioinformatics Department –** The Bioinformatics Department is the analytical core of the company. It consists of bioinformaticians, data scientists and interns. They are responsible for transforming raw genetic data into understandable insights through the analysis of genomic data. Utilizing cutting-edge algorithms, the team performs variant calling, annotation, and prioritization. Extensive research is conducted on various topics aimed at improving healthcare. In addition to this, they are also involved in content creation, like generative reports, and test results. Additionally, the team actively participates in brainstorming sessions for new ideas for the research project.

    - **Intern –** As an intern in the bioinformatics department I was involved in the research and development of new projects. I was part of a project involved in data analysis of genomic data to find pathogenic variants. Along with that, I was also involved in writing some python scripts. Additionally, I was also part of the content team and generated reports for clients.

2. **Information and Technology Department –** The IT department is technical core of the company. It consists of Java developers, python developers, PHP developers and UI/UX designers. They are working on various healthcare projects.

1.4 Theory Behind the project

Mutations

Mutations are changes/alterations in DNA sequences. Mutations can arise spontaneously or induced by various internal and external factors such as errors in the DNA replication process during cell division, exposure to harmful radiations, exposure to viruses, and harmful mutagens. These mutations affect structure and functions of proteins which are encoded by the genes. If there is a defect in the gene, then there is an alteration in the protein encoded by it which can lead to serious conditions. There are two types of mutations:

- Somatic Mutations – These are the changes in the DNA of somatic cells (non-reproductive body cells). They are not passed on to offspring.
- Germline Mutations – These are the changes in the DNA of reproductive cells of the body. They have an impact on next generation as they are inherited by offspring.

In the context of immunodeficiencies, if there is a mutation in a gene that is responsible for immune function, can lead to immunodeficiency disorders. For example, mutation in the gene CD40L can hamper immune function and, in this case, stop the production of B cells, which can result in immune disorders.

SNP (Single Nucleotide Polymorphism)

SNPs are the most common type of genetic variation found in human DNA sequences. These mutations cause single base pair changes in DNA. Changes in the DNA at specific locations in the genome result in different alleles at the locus among individuals in a population. For example, normal sequence in an individual is ACTG. Due to external or internal factor, SNP could occur, converting C to G, so the sequence becomes AGTG.

SNPs as Biomarkers

A biological marker, or biomarker for short, is an objective metric that records an organism's or cell's current state. They can be molecules, genes, proteins, cells, or characteristics observed during imaging. They are very important in biological research as they give information about disease diagnosis, prognosis, and treatment.

SNPs act as important biological markers. Specific SNPs, within genes associated with immune function, can act as indicators of an individual's susceptibility or presence of primary immunodeficiency.

These SNPs may:

- Directly disturb protein function
- Influence gene expression
- Serve as Genetic marker

These SNP biomarkers are important for many research areas and are studied by many researchers world-wide. There are databases of these SNP biomarkers where all researched biomarkers are present; databases like dbSNP, and GWAS are some examples. Here, all the information about that SNP is provided. A special identification number is given to each SNP in database; this number is called rsids.

Variant Classification

Not every mutation or SNP is harmful, some are benign and don't have any effect on the individual. To distinguish between benign and pathogenic variants, a standardized classification system is created.

This classification system was created by The American College of Medical Genetics and Genomics (ACMG). They have set certain guidelines based on evidence like population, frequency, and functional impact.

The main categories are:

- **Pathogenic: -** These variants have significant evidence suggesting they are directly responsible for the development of disease.

- **Likely Pathogenic: - T**hese variants have a very high probability of causing disease. Usually, this probability is more than 85%.

- **Variant of Uncertain Significance (VUS): -** These are the variants that have very little information to suggest their pathogenicity.

- **Likely Benigne: –** These are the variants that have very low probability of causing disease.

- **Benign: –** These variants are unlikely to cause any diseases.

Note: - We have only taken germline variants.

MygenomeBox examines these reported biomarkers related to various diseases, traits and characteristics. After this literature review, research is conducted by the bioinformatics team based on this research tests, reports or applications is developed.

In this project, we have concentrated on Immunodeficiency disorders specifically primary immunodeficiency disorders.

2. Details of Task Performed by candidate.

2.1. Details of activities carried out by candidate

2.1.1 Initial work-

I worked as an intern in the Bioinformatics team, and one of my primary responsibilities was exploring the in-house cloud platform of MGB. This platform hosts over 104 developed applications that are related to various fields. I did the comprehensive analysis and examined these applications, where my focus was on studying content and closely examining the reports generated, This I did using sample DNA data provided to me. This allowed me to understand and get to know the ins and outs of the platform as an end user. At the same time, I also studied previously generated reports and tried to replicate the results with sample DNA data by writing small Python scripts. The result of this was presented by me to my team every week until I successfully replicated a few reports.

2.1.2 Researching new ideas for the project: -

The next step was to look for innovative ideas for new projects. This involved an extensive research exploration; we delved into diverse aspects of genomics. Studying various genes and their effect on lifestyle, and studying various diseases and their associated genes. The main objective was to design and create cutting-edge apps or new techniques that can enhance healthcare and are entertaining, user-friendly, and educational. This involved reading various research papers and looking for information in different databases. The data collected was carefully sifted through, leading to the identification and shortlisting of promising concepts. These ideas were then presented to our collaborative team for evaluation and consideration. After doing extensive research, we pinpointed the topic Primary immunodeficiency.

Although the simultaneous examination of hundreds of genes made possible by next-generation sequencing (NGS) has transformed the diagnosis of primary immunodeficiencies (PIDs), doctors still face substantial

obstacles due to the amount and complexity of the obtained data. Currently, diagnostic workflow often relies on manual review of variant data, a process that is prone to human error.

So, we decided to build a computational pipeline which simplifies the processing of NGS data and effectively finds pathogenic variants linked to PID

Immunodeficiency

In healthy individuals, immune responses operate in two phases. The first line of defense is our innate system, which we all inherit and which offers a quick response that is not specific to a particular microbe, and it is considered an in-general response. The adaptive immune system activates when infection breaches our initial line of defense, and it takes several days for it to customize its defense against that particular invader. This targeted attack leaves behind immune memory, which allows an easier response if that specific invader invades again (a particular microbe in this case). In a person with an immunodeficiency disorder, one or more components of either the adaptive or innate immune response are impaired, resulting in the body being unable to effectively resolve infections or disease.

Immunodeficiencies comprise a range of conditions where the immune system's ability to protect against pathogens is compromised. These diseases can range in severity from moderate to severe, increasing the risk of recurring infections and potentially deadly side effects.

There are two types of immunodeficiency disorders: -

- **Primary Immunodeficiencies (PID):** – PIDs are often congenital and result from mutations or flaws in the genetic code. These disorders often are identified in early childhood or infancy. They often arise from genetic defects that regulate how the immune system grows and functions.

  Common examples of PIDs include:

  - **Severe Combined Immunodeficiencies (SCID)**: - This rare group of genetic diseases is caused by mutations in genes involved in immune functions. It affects body's ability to fight off infections, making person prone to illness. This is typically diagnosed early in life, usually in the infancy stage, and it requires early medical attention. Techniques such as bone marrow

transplantation or gene therapy are used to boost immune function. If proper treatment is not provided, then SCID can be life-threatening.

- **X-Linked Agammaglobulinemia (XLA):** - This is a genetic disorder in which the body fails to produce mature B cells, resulting in a severe immune deficiency. The condition mainly affects men and puts them at risk for infections, especially in the respiratory tract. XLA results from a mutation in the BTK gene that promotes B cell maturation, and their function is regulated by X-linked agammaglobulinemia (XLA). The treatment consists of lifelong immunoglobin replacement therapy to prevent infection.

- **Secondary Immunodeficiencies (SID)** – These are the immunodeficiencies which are result of factors other than genetic mutations. These factors include certain medications, chronic diseases like HIV/ AIDS, chemotherapy, and other conditions that compromises immune functions. They are developed later in life and are not present from the birth.

The computational pipeline developed in this project focuses on primary immunodeficiency.

2.1.3 Curation of the Genetic Markers and Building Database

A very important component of this variant analysis pipeline is making well-curated database. To ensure the accuracy and comprehensiveness of variant identification, a dedicated database was constructed. This database serves as a reference for prioritization and variant detection in patient samples. A thorough literature review was conducted to identify genes and genetic markers. This involved studying scientific research papers, clinical studies, and review articles. From this, we aim to include genes and markers which have good research backing and also recently discovered genes. This information, such as rsid and chromosome position, was verified using the company's GSA chip. The database that we created was double checked by our director. Selection of the markers was done on basis two things:

1. **Relevance to the disease**
2. **Pathogenicity according to the ACMG guidelines.**

Selected markers were stored in excel file.

Figure 5 – Few databases which I have used

2.1.4 Identification and Analysis of genetic variants using GATK

Genome sequencing and Next generation sequencing

The most thorough understanding of an organism's genetic makeup is provided via whole genome sequencing (WGS). It involves deciphering a cell's whole DNA sequence and identifying variants that impact characteristics, ancestry, health, and illness. The gold standard in the past was Sanger sequencing, which took a long time and was very expensive as it involved joining the human genome piece by piece. Genomic research was transformed by the introduction of next-generation sequencing (NGS), which greatly parallelized the procedure. NGS methods sequence millions or billions of DNA fragments at once, in contrast to Sanger's read-by-read approach. This has several benefits, such as much lower expenses, more throughput, and improved accuracy. NGS includes a variety of methods that have been tailored for objectives. For instance, whole-exome sequencing concentrates on the coding exons of genes, which are known to contain the majority of disease-causing mutations.

NGS and FASTQ files

The raw output from Next Generation Sequencing is typically stored in the form of FASTQ files. These files contain the DNA sequences of the reads along with corresponding quality scores, which indicate the confidence level of each base call. The massive volume of data generated by NGS, however, presents significant challenges for analysis and interpretation. Raw FASTQ files require specialized bioinformatics

tools and pipelines to extract meaningful information. These FASTQ files are then processed using different tools and software to obtain a file that contains all the genetic variants in that individual, this file is called as a VCF file.

To address the complexities of NGS data analysis, we have utilized the Genome Analysis Toolkit (GATK4).

Genome Analysis Toolkit (GATK 4)

The Genome Analysis Toolkit (GATK) is an industry standard software tool created by the Broad Institute to assist with analyzing high throughput sequencing data, especially in terms of its accuracy in detecting genetic variation, e.g. single nucleotide polymorphisms (SNPs) and insertions/deletions. It is specifically developed for whole genome and exome analysis. It is widely used in both research and clinical settings. The Latest version of GATK allows you to use of several application functions for processing alignment files, and VCF files.

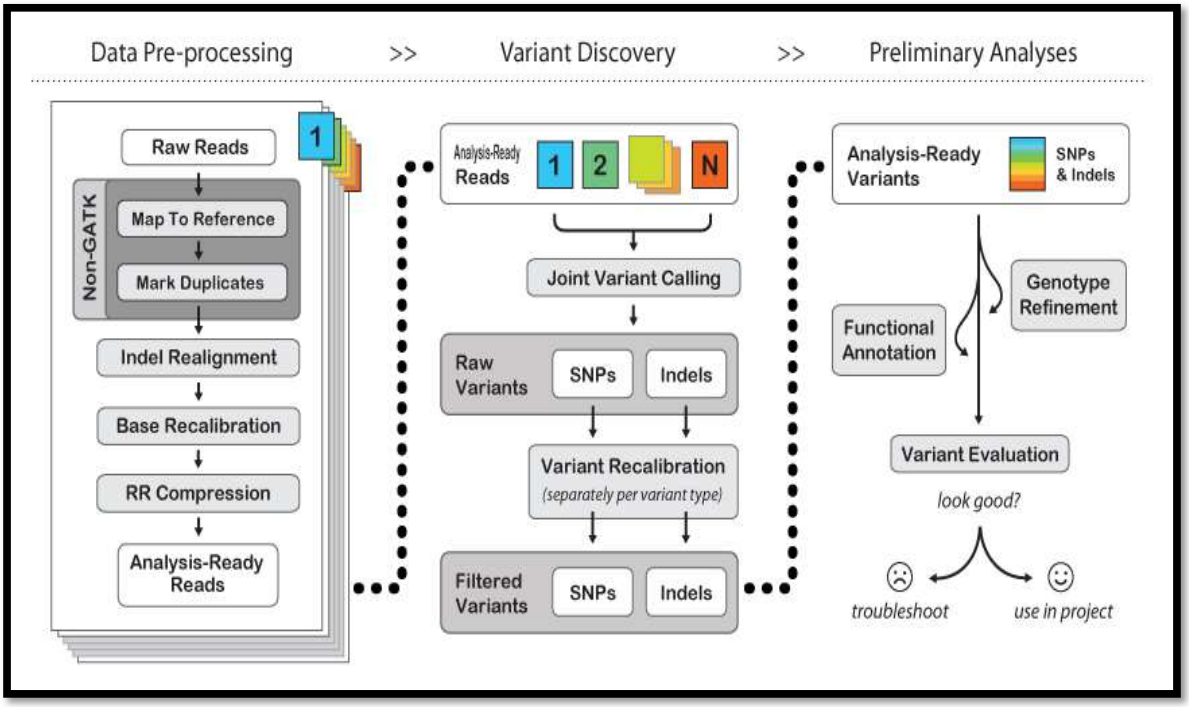


Figure 6- GATK Best Practise Workflow

2.1.5 Additional Tools which were used in variant analysis pipeline:

1. **Samtools**

   Samtools is a set of tools which are widely used in bioinformatics for processing high-throughput sequencing data in SAM(sequence alignment/map) and BAM(binary alignment/map). They are widely used for changing alignment to SAM, BAM and CRAM formats. They are also used in indexing, sorting, merging and analysing these files.

   In this pipeline we have used samtools for:

   - Conversation – From SAM file to BAM file
   - Sorting
   - Indexing.

Figure 7- Samtools Home Page

2. **BWA-MEM2 (Burrows-Wheeler Aligner)**

BWA-MEM2 is an upgraded more effective version of BWA-MEM (Burrows-Wheeler Aligner). This is the software which is used to align the sample sequence with the reference genome. It is most suited for analysing Illumina sequence data, which is the data used in my project. BWA-MEM2 has more speed, accuracy, and memory usage as compared to BWA-MEM so that is why we chose this.

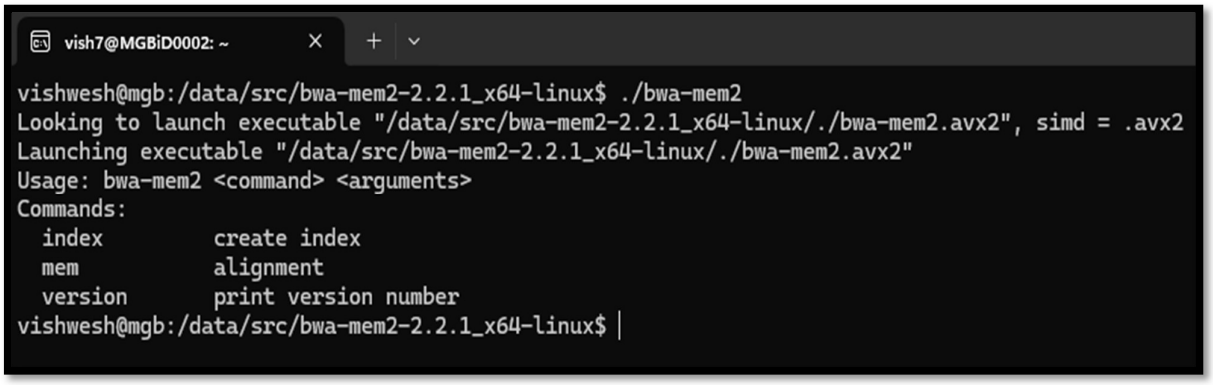


Figure 8- BWA-MEM2 running

3. **GATK(HaplotypeCaller) –**

In our pipeline, we have used HaplotypeCaller for variant calling. HaplotypeCaller takes a local assembly-based approach, which allows for the accurate identification of both single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels). It employs sophisticated algorithms to assemble individual haplotypes, or sequences of DNA variants, and then probabilistically determines the most likely set of haplotypes that best explain the observed data.

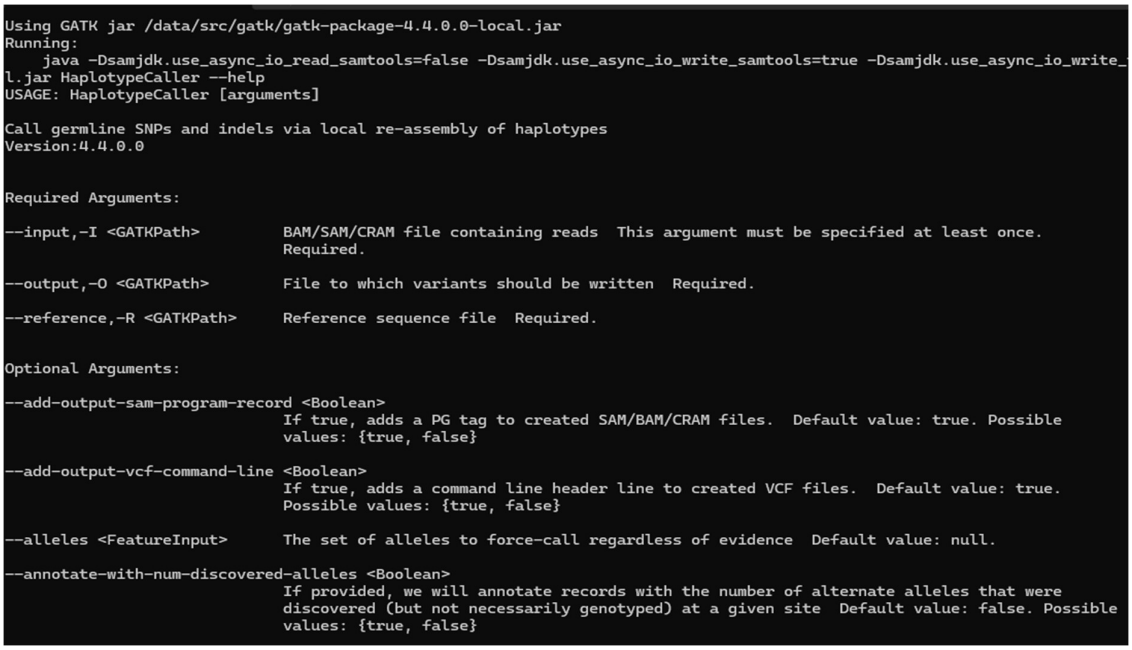


Figure 9- HaplotypeCaller Running

4. **Python Programming Language**

For our project, we used Python programming language due to its versatile nature and flexibility. Python was used to develop custom scripts which does the next processing of pipeline like variant filtration, prioritization, and report generation. Python has rich ecosystem of libraries available for data manipulation.

We have used following libraries in our code. Pandas, Pdfkit, PythonDocx, and Json.

2.2 Processes covered by candidate

2.2.1 Data Curation

We built a curated database specific to immunodeficiency. We followed the following process: -

1. **Source Identification:** – Profound data research was conducted, and relevant information was gathered from various databases like Clinvar, the Human Gene Mutation Database, SNpdeia, Genome- Wide Association Study (GWAS), Online Mendelian Inheritance in Man (OMIM). These databases have already curated markers, and we downloaded them.

2. **Literature Review:** – In addition to the public databases, scientific literature was also significantly reviewed. This was done in the hope of identifying a novel marker related to immunodeficiency. This involved manually extracting information from relevant publications and clinical reports.

3. **Data Selection and Filtering**: - The data that we got from the above two sources was stored in an Excel file and the next step was filtering and selecting the right genetic marker. This was done through two methods -
    - **Association with Immune Disorders:** – Variants that are part of the gene that are directly involved in any immune function and involved in causing immune disorders were prioritized.
    - **Pathogenicity Classification:** – Variants that are pathogenic or likely pathogenic by ACMG guidelines were only included.

4. **Data Cleaning:** – After all the variants were identified, the curated database was cleaned for better accuracy and consistency.
    - Removing duplicates – Duplicates were identified and removed, this was done through a Python script.

Database Structure: -

The resulting database was structured to include the following information:

- **rsID:** – This is a reference SNP cluster ID from dbSNP.
- **Gene:** – Gene symbol associated with variant.
- **Amino Acid Change:** – Amino acid caused by variants.
- **Reference allele (Ref):** - Reference base present at variant position.
- **Alternate allele (Alt):** - Alternate allele present at variant position.
- **Reference Position:** – Genome coordinates of that variant on reference genome.
- **Zygosity:** - The zygosity of the variant (homozygous, heterozygous, hemizygous).
- **Pathogenicity: -** The predicted or known pathogenicity of the variant.
- **PMID: -** The PubMed identifier (PMID) of the publication(s) supporting the variant's association with the disease.

	CLN_RSID	Ref	Alt	CLN_Gene	AA change	CLN_Clinical_Significance	CLN_PMID	CLN_Disease
2								
3	rs781050795	G	A	ACP5	p.Gly109Arg	Pathogenic	https://doi.org/10.1038/ng.749	SPENCDI
4	rs121908722	G	A	ADA	p.Arg156His	Pathogenic	https://doi.org/10.1172/jci116833	SCID
5	rs121908714	G	T	ADA	p.Arg101Leu	Pathogenic	https://doi.org/10.1172/jci116833	SCID
6	rs121908731	G	A	ADA	p.Val129Met	Likely pathogenic	https://doi.org/10.1002/(sici)1098-1004(1998)11:6%3C482::aid-humu15%3E3.0.co;2-e	SCID
7	rs121908733	C	T	ADA	p.Arg149Trp	Likely pathogenic	https://doi.org/10.1002/(sici)1098-1004(1998)11:6%3C482::aid-humu15%3E3.0.co;2-e	SCID
8	rs121908724	G	A	ADA	p.Gly20Arg	Pathogenic	https://doi.org/10.1006/clin.1994.1026	SCID
9	rs761242509	G	A	ADA		Pathogenic	https://doi.org/10.1002/humu.1380050309	SCID
10	rs751635016	T	A	ADA	p.Arg282Gln	Pathogenic	https://doi.org/10.1111/cge.12257	SCID
11	rs778809577	C	T	ADA	p.Arg235Trp	Likely pathogenic	https://doi.org/10.1016/j.clim.2011.04.011	SCID
12	rs121908716	G	A	ADA	p.Arg211His	pathogenic	https://doi.org/10.1182/blood.V91.1.30	SCID
13	rs1452483770	C	T	ADA	p.Pro104Leu	Pathogenic	https://doi.org/10.1093/hmg/2.8.1307	SCID
14	rs1233957241	C	A	ADA	p.Pro126Gln	Likely pathogenic	https://doi.org/10.3389%2Ffimmu.2016.00466	SCID
15	rs387906267	A	G	ADA		Likely pathogenic	https://doi.org/10.18176/jiaci.0147	SCID
16	rs121908723	G	A	ADA	p.Gly216Arg	Pathogenic	http://www.ncbi.nlm.nih.gov/pmc/articles/pmc1683191/	SCID
17	rs121908740	C	T	ADA	p.Arg211Cys	Likely pathogenic	https://www.ncbi.nlm.nih.gov/pubmed/8051429	SCID
18	rs121908714	G	A	ADA	p.Arg101Gln	Pathogenic	https://doi.org/10.1172/jci112050	SCID
19	rs77563738	C	A	BTK	p.Trp281Ter	Likely pathogenic	https://doi.org/10.1046/j.1365-3083.2001.00967.x	XLA
20	rs202134424	T	A	BTK	p.Tyr598Asn	Pathogenic	https://doi.org/10.1016/j.jaip.2018.09.004	XLA
21	rs148936893	C	A	BTK		Pathogenic	https://pubmed.ncbi.nlm.nih.gov/34241796	XLA
22	rs1423056320	G	T	BTK	p.Leu652Pro	Likely pathogenic	https://doi.org/10.15586/aei.v49i2.62	XLA
23	rs542412710	T	G	BTK	p.Ala607Asp	Likely pathogenic	https://doi.org/10.3389%2Ffimmu.2023.1252765	XLA
24	rs2115075086	C	A	BTK	p.Glu589Gly	Likely pathogenic	https://doi.org/10.1016/j.jaci.2014.12.1226	XLA
25	rs2115044716	A	T	BTK	p.Trp252Ter	Pathogenic	https://doi.org/10.1172/jci112050	XLA
26	rs1404170214	A	T	BTK	p.Lys430Glu	Pathogenic	https://doi.org/10.1172/jci112050	XLA
27	rs1554271741	G	A	BTK	p.Tyr2755Ter	Pathogenic	http://dx.doi.org/10.5507/bp.2013.011	XLA
28	rs765680532	A	G	BTK	p.Arg288Gln	Pathogenic	https://doi.org/10.1002/iid3.1049	XLA
29	rs145474800	C	T	BTK	p.Val335Asp	Likely pathogenic	https://doi.org/10.1093/hmg/3.1.79	XLA
30	rs794729673	C	T	BTK	p.Trp252Ter	Pathogenic	https://doi.org/10.1093/hmg/2.8.1307	XLA
31	rs571517554	C	T	BTK	p.Thr33Pro	Likely pathogenic	https://doi.org/10.1093/hmg/2.8.1307	XLA
32	rs397514686	A	G	BTK	p.Tyr100Ter	Pathogenic	https://doi.org/10.1016/j.jaip.2018.09.004	XLA
33	rs387907352	G	T	BTK	p.Lys12Asn	Pathogenic	https://doi.org/10.1172/jci112050	XLA
34	rs1272762412	C	A	CARD11		Pathogenic	http://dx.doi.org/10.1016/j.gendis.2019.09.015	HIGM1
35	rs2131441761	A	G	CARD11	p.Arg30Trp	Likely pathogenic	http://dx.doi.org/10.1016/j.gendis.2019.09.015	Hyper-IgM syndrome type 1
36	rs398122363			CARD11				Hyper-IgM syndrome type 1

Figure 10- Screenshot of the Database

2.2.2 Variant Calling and Data Analysis pipeline:

In this project for variant calling, we have used the latest version of GATK, which is GATK 4.5.0, released on 14 December 2023.

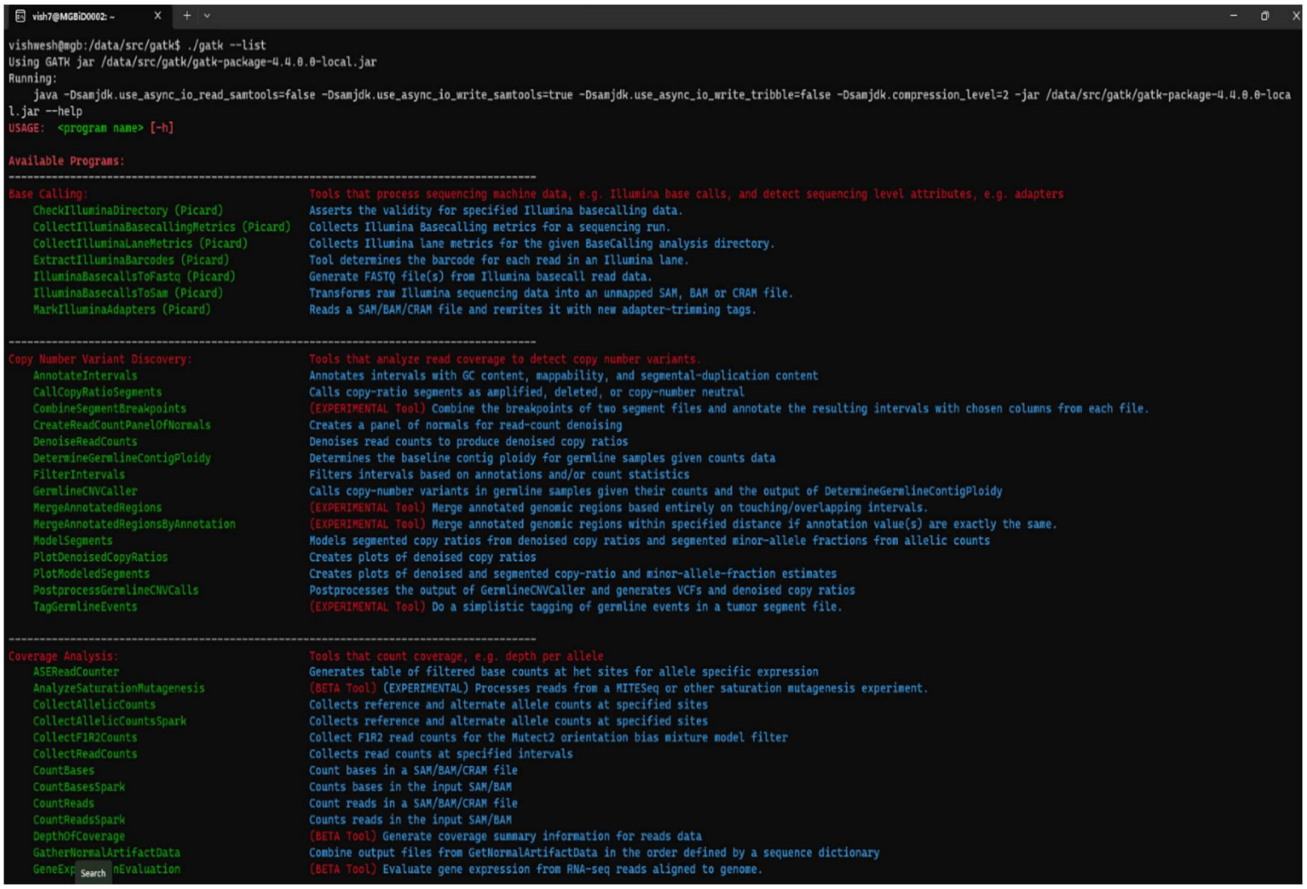


Figure 11- GATK Main Page

Prerequisites to tun GATK pipeline

- **Linux Platform (We used Ubuntu)**
- **Java installed on system**
- **Data:**

  To develop and validate our pipeline we have used data from the Genome in Bottle (GIAB) Website. The data is of Whole Exome Sequencing. The data belongs to the Garvan Institute of Medical Research, is of an unknown female from United States Of America, and was downloaded from:

  (http://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/)

There were two paired-end sequencing reads:

R1 = /home/Vishwesh/Gatk/reads/Garvan_NA12878_Whole_Exome_R1.fastq.gz

R2 = /home/Vishwesh/Gatk/reads/Garvan_NA12878_Whole_Exome_R2.fastq.gz

- **Reference:**

  The Human reference genome was downloaded from the UCSC browser.

  https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz

  hg38 = /home/Vishwesh/Gatk/reference/hg38/hg38.fa

Data Preprocessing:

Before we align our raw sequence reads to reference genome, it is important to perform preprocessing steps to ensure good quality of data.

1. **Quality Control: Firstly, we performed Fastqc on our sample files**

   ./fastqc home/Vishwesh/Gatk/ arvan_NA12878_Whole_Exome_R1.fasta

   This gave us an Html report like this: -

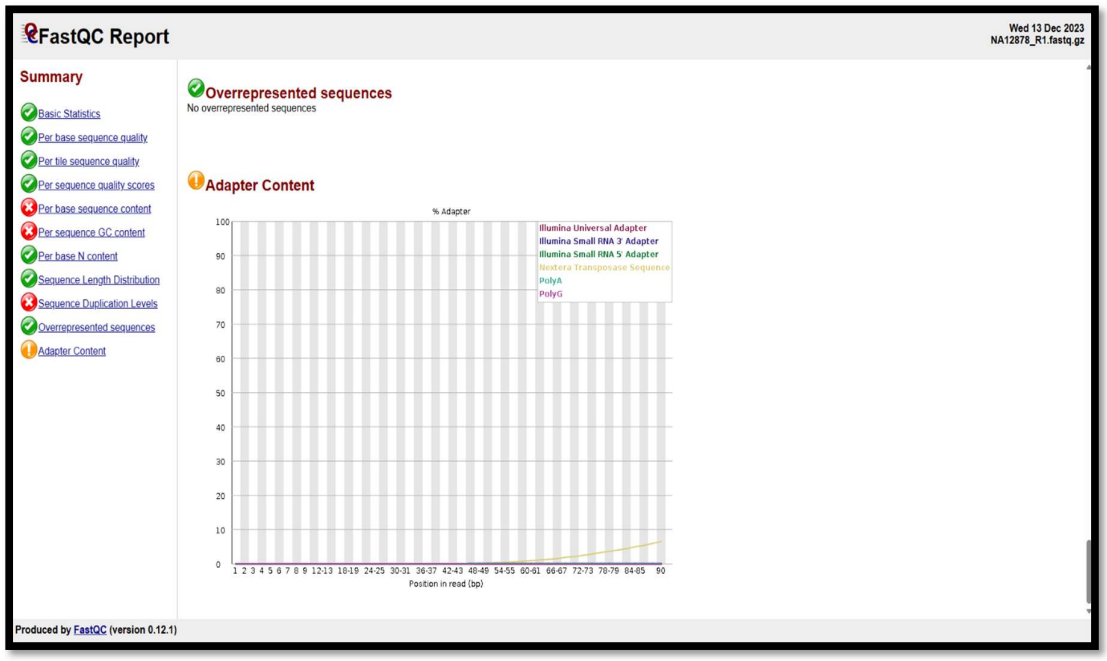


Figure 12 - This clearly shows there is an adapter sequence present in our

2. Trimming Adapter Sequence.

To remove these adapter sequences, I have used cutadapt. Cutadapt is a python-based tool so, Python should be present in your system.

Adapter sequences for the data were given in information file of sample.

cutadapt -a <adapter sequence> - output – input.

This command was run for read1.

cutadapt -a CTGTCTCTTATACACATCTCCGAGCCCACGAGAC -o output_R1.fastq.gz -p /home/vishwesh/Gatk/Reads/R1.fastq.gz

This gave us a trimmed file which we ran through Fastqc again and this was the result:

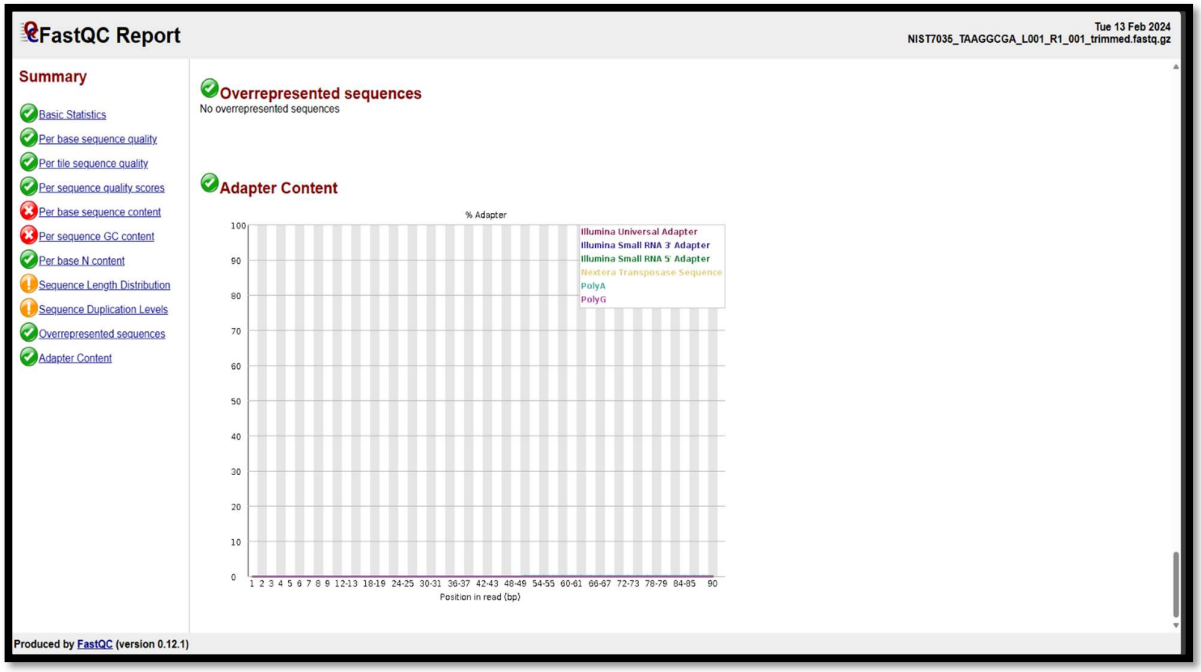


Figure 12 - This shows adapters are removed. This was done for both R1 and R2 files our data.

This step gives us trimmed and quality-controlled reads which are ready to be aligned with the reference genome.

Variant Calling:

This is the step where we identify variants (differences) between the reference genome and the sample sequenced genome.

Step 1: Indexing Reference Genome

This creates data structures and allows rapid access to specific regions of the genome. It improves the efficiency of alignment and variant calling steps.

bwa-mem2 index hg38.fa

Step 2: Alignment with BWA-MEM2

As stated earlier we have used the bwa-mem2 algorithm to align processed sequence reads (R1, R2) to the indexed hg38 genome. The output of this file is SAM file, which contains information for each read, including its genome position, and any error compared to the reference genome.

/home/vishwesh/bwa-mem2-2.2.1_x64-linux/bwa-mem2 mem -t 4 /data/reference_files/hg38.fa /home/vishwesh/Reads/R1.fastq.gz /home/vishwesh/Reads/R2.fasta.gz > aligned_reads.sam

***S*tep 3: Conversation to BAM file**

SAM file produced is in a human radiable format which is not efficient to store. So, we convert this SAM file to a BAM file with samtools sort, which also sorts BAM files based on genomic coordinates.

./samtools sort -bS /home/vishwesh/aligned_reads.sam > /home/vishwesh/aligned_reads.bam

Step 4: Marking Duplicates and Adding Read Groups

- Mark Duplicates - PCR duplicates are copies of the same original DNA that arises during library preparation. This can introduce bias in variant calling, so to improve efficiency and prevent extra coverage bias this step is performed. To tackle this ***MarkDuplicate*** (Picard tool) from GATK was used. This identifies and tags the PCR duplicates in a sorted BAM file.

./gatk MarkDuplicatesSpark -I /home/vishwesh/sorted_reads.bam -O /home/vishwesh/dedup_reads.bam -M /home/vishwesh/metrics.txt

- Adding Read Groups: This step ensures data is properly organized as read groups provide information about the sequencing experiment such as sample ID, library, platform and run date. This information is important when we are dealing with multiple samples at same time. This was done with the Picard Tool from GATK ***AddOrReplaceReadGroups.***

  ./gatk AddOrReplaceReadGroups I= /home/vishwesh/aligned_reads.bam O=/home/vishwesh/output_Reads.bam RGID=4 RGLB=lib1 RGPL=ILLUMINA RGPU=unit1 RGSM=20

Step 5: Base Quality Score Recalibration (BQSR)

Base quality scores which are assigned by the machine are not always accurate and can have errors. These errors can happen due to many reasons for example technical errors, and chemical fluctuations. To correct this BQSR step is run. This is a quality control step that assigns scores to each base.

BQSR has two steps first one is building a machine learning model by analyzing base calls and their corresponding quality scores in an aligned BAM file, and the second step is applying this model to the original BAM file which is sorted a BAM file. This results in a BAM file which has more accurate quality scores.

- **Base Quality Score Recalibrator Tool for building model.**

  ./gatk BaseRecalibrator -I /home/vishwesh/dedup_reads.bam -R /data/reference_files/hg38.fa --known-sites /data/reference_files/gatk_resource_bundle/resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf -O /home/vishwesh/recalibration_report.table

- **Applying the Model using ApplyBQSR**

./gatk ApplyBQSR -R /data/reference_files/hg38.fa -I /home/vishwesh/dedup_reads.bam --bqsr-recal-file /home/vishwesh/recalibration_report.table -O /home/vishwesh/bqsr_output.bam

Here refernce file used is **resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf** which was downloaded as part of GATK resource bundle.

Step 6: Variant calling using GATK Haplotypecaller

Variant calling is the process in which we aim to identify genetic variants from our sequenced data. This is done through a sophisticated GATK tool called Haplotypecaller. In addition to this, we also used a bed file which contains the targeted region for Immunodeficiency which we got from the same website as sample. This bed file was initially in hg19 format which we converted to hg38 using the UCSC liftover tool.

- **Liftover hg19 to hg38**

  /path/to your file/ Your_bed_file.bed.gz hg19ToHg38.over.chain output.bed unlifted.bed

- **Variant calling**

  time ./gatk HaplotypeCaller -L hg38bed.bed -R /data/reference_files/hg38.fa -I /home/vishwesh/bqsr_output.bam -O final_Haplo.vcf

This gives us gVCF file which contains information about variants and reference sites.

Step 7: Annotation

Annotation was done using VariantAnnotator tool in GATK. This does annotation by integrating data from various databases like dbsnp, genomAD etc. It annotates variant calls based on their context from a well-established and widely used database of genetic variant. It takes variant call from our gvcf file then matches with entries in database if there is match then rsid of that certain position get copied in our new VCF file.

time ./gatk VariantAnnotator - 0 Final_VCF - V – final_haplo.vcf -D resources_broad_hg38_v0_Homo_sapiens_assembly38.dbsnp138.vcf

Here we have used dbsnp database from NCBI which was downloaded as part of GATK resource bundle.

This gives us the final vcf file which we will use in the next part of our project.

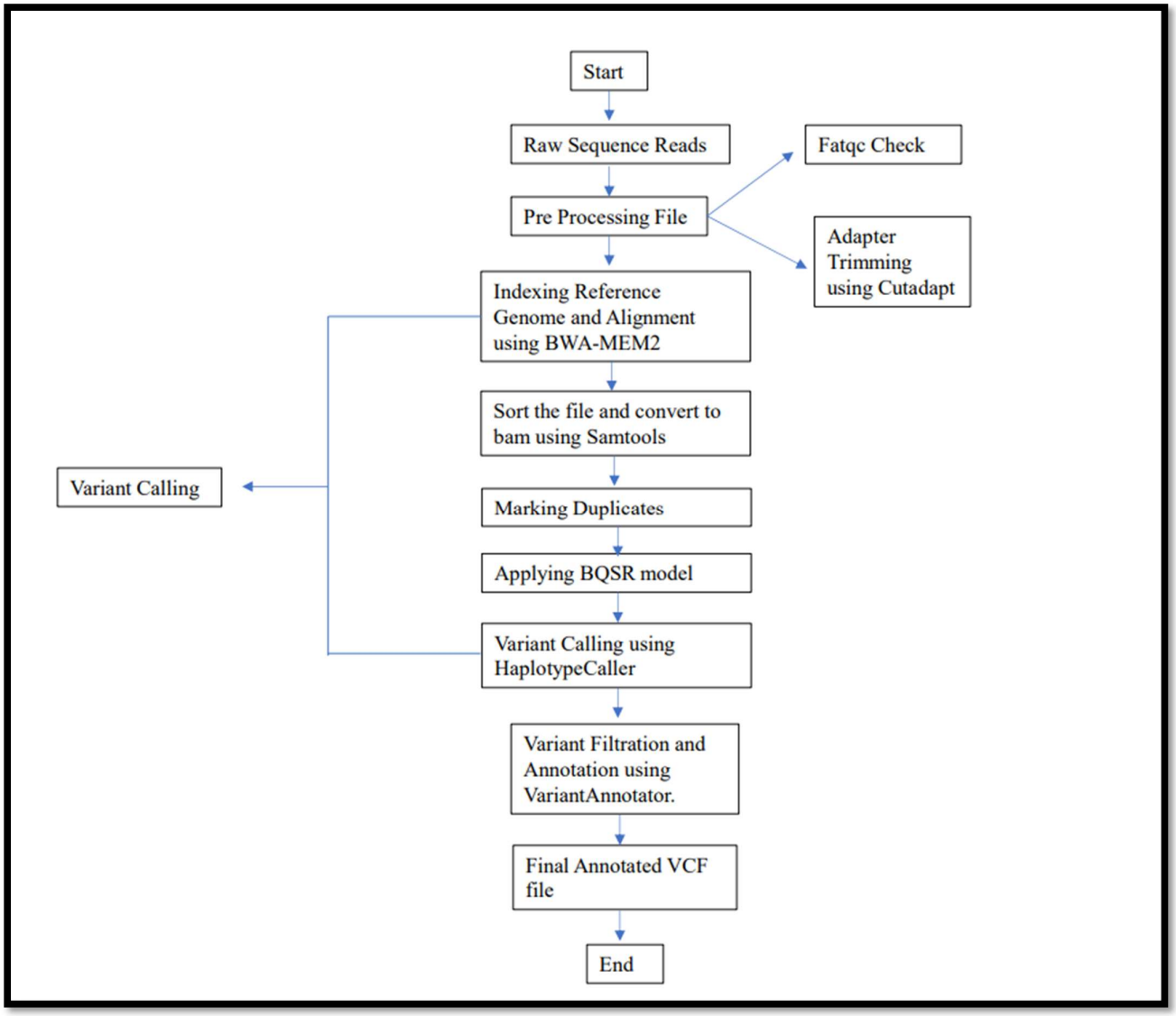


Figure 13 – GATK Workflow

2.2.3 Variant Analysis and Report Generation Pipeline

The next step after we got the annotated VCF file, is identifying the potential pathogenic variant associated with immunodeficiency in our sample file. Our team and I have developed this Python-based pipeline which takes an annotated VCF file, simplifies the pathogenic variant detection, and gives us the final clinical PDF report.

This pipeline is split into two parts:

1) The first part takes an annotated VCF file, streamlines the identification of pathogenic variants, and gives us the JSON file with identified variants.
2) The next part takes this JSON file and generates a PDF report using custom-built code.

1. Pathogenic Variant Identification Pipeline.

This is the first part of the pipeline; this code is written in Python 3 programming language.

Prerequisites for this code to run are the following files:

- Annotated VCF file (which we got from running the GATK pipeline)
- Marker file (our curated database file in text format),
- Gene description text file.
- Disease description file.

All these files should be in the same directory.

The script can be executed from a Linux terminal or Windows command prompt by providing the input VCF file name and the desired output file name on command line.

Execution of code.

> > ***Python < your code.py> -R <your output file name> -F <your vcf file.vcf>***

This is basic structure of command to run this code

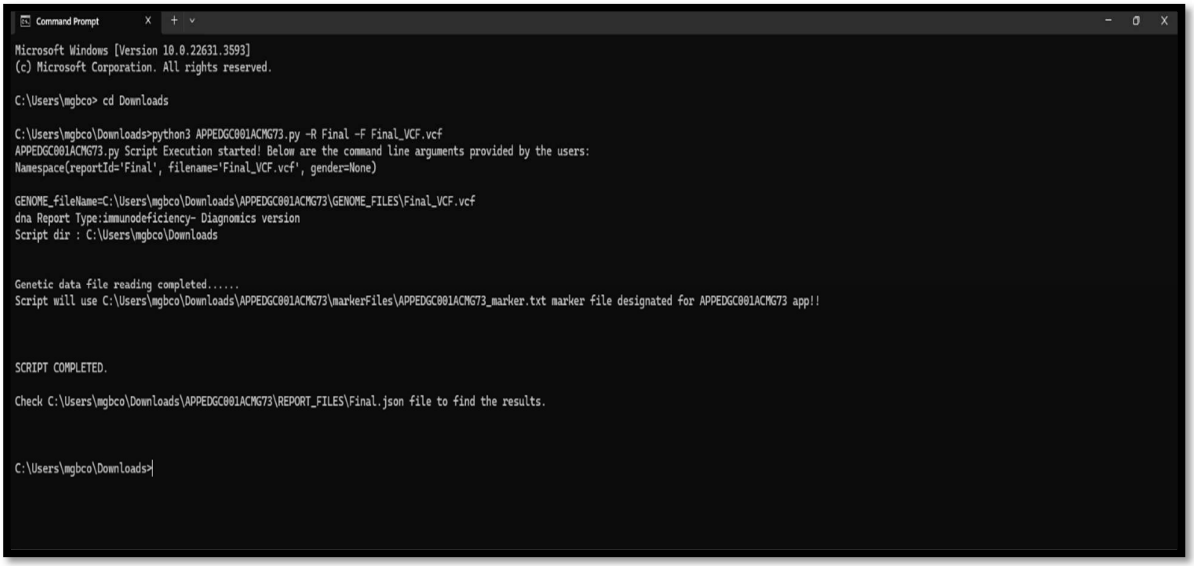
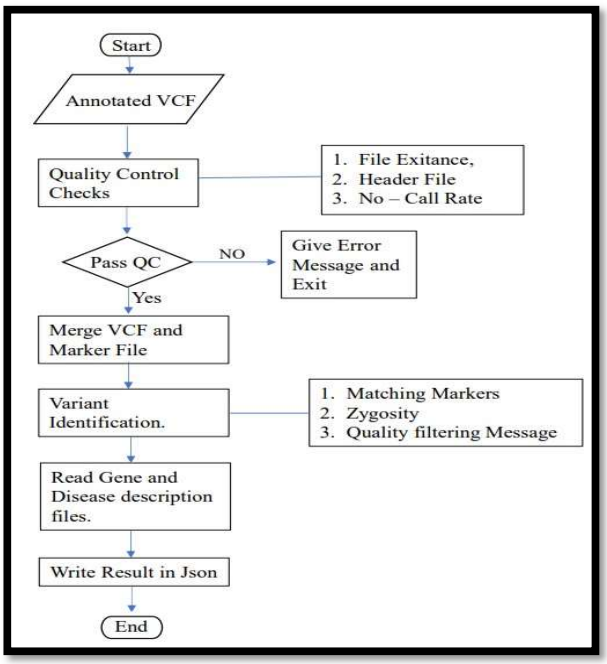


Figure 14 – Here python command is run, and our script is executed giving us Json

Explanation of Code.

The Python script utilizes the P**andas** library for data manipulation and analysis. It parses command-line arguments to get the VCF file name. Performs QC and gives the filtered variants, which are finally written in JSON format.

The flowchart below is a simpler explanation of how code is running:



The script has various quality control checks and parameters which identifies the final pathogenetic variants like:

There are three quality control checks present in code:

1. **File Existence:** – This checks if the VCF file is present and if there is data present in this file.
2. **Column Headers:** – Next script confirms the presence of essential column headers (CHROM, POS, REF, ALT) in the VCF file. These headers are required for accurate variant interpretation.
3. **No Call Rate:** – This is one of the most important parts of the code. Calculates the proportion of variants with no-call genotypes (missing data) and ensures it's below a specified threshold (2%).

Variant Identification

1. **Merging VCF and Marker File:** – This part of the code merges the information from an annotated VCF and curated database file. This step further matches the common variant in both files and removes the rest of the lines. Keping only matched variants.
2. **Zygosity determination:** - This step determines the zygosity by identifying the pattern in VCF file.

   0/1 - Homozygous

   1/1 - Heterozygous

   These columns are present in the VCF file, and code identifies with following code: -

```
#check zygosity
def Zygosity(value):
    if(value.startswith('1/1:')):
        return "Homozygous"
    elif(value.startswith('0/1:')):
        return "Heterozygous"
    elif(value.startswith('1:')):
        return "Hemizygous"
```

3. **Quality Filtering:** - Once we have assigned the Zygosity it is time to filter bad score variants. This part of the code filters out low quality score (GQ < 0). Low quality scores are deleted because they are mostly false positive result.

After these filtering steps, the script is left with a subset of variants that are more likely to be pathogenic and which are matched.

Then the matched variant along with the appropriate columns are written out in the JSON format. The JSON format is chosen for its structured, human-readable, and machine-parsable nature.

If there are no variants found in the sample. Then the code creates a JSON file which doesn't have any information but only has information about No-call rate. **This ensures that even if there is no pathogenic variant, a JSON is created which will eventually give a Negative result.**

2. Report Generation from Json File

This is the second part of the pipeline. This script is written in Python 3 Programming language.

The prerequisite to run this code are:

- Generated JSON file from previous code
- Word file (Template given by MyGenomeBox).

All this file should be present in one directory.

Libraries Included

Python-Docx: – This library is used to read and manipulate the word document. This library is used to insert the variant data from the JSON file into appropriate place-holders.

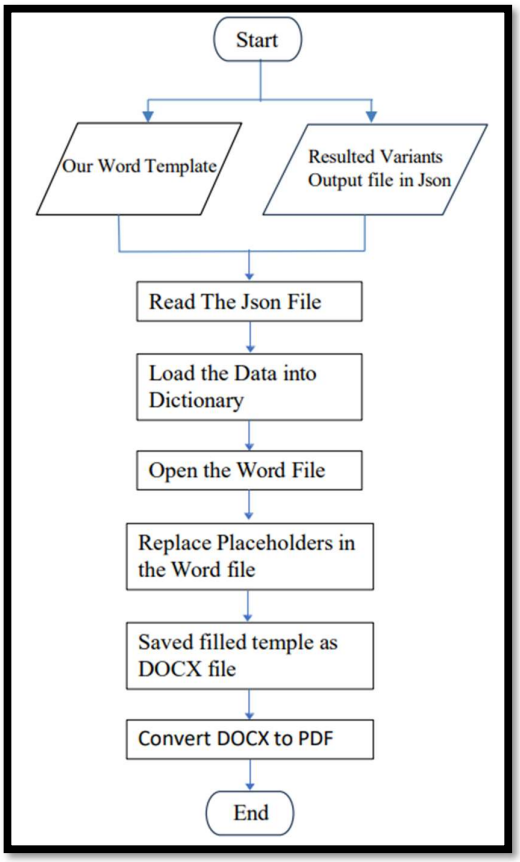
Pdfkit: – This library uses interfaces with the **wkhtmltopdf** command-line tool, is responsible for converting the populated Word document into a PDF file.

With this prerequisite all in place, Python script can be executed from the command line, taking Json as input and giving pdf as output. This script ensures the identified variants are reported in the pdf.

Execution of the code

>> ***Python < your code.py> -R <your output file name> -F <your JSON_file.json>***

The Flowchart below is simpler explanation of the Code:



Key Factors in Code

Dynamic Content Filling –

The Script reads JSON file and dynamically populates the key-holder in the Word file which has been placed there manually. These place-holders are placed in the template Word file. This gives us the variants identified in the JSON on the Word file.

```
for key, value in data.items():
        if key in ["gene", "CLN_Disease", "ref Allele ", "Alt Allele ", "amino_acid_change", "rsid", "zygosity", "pathogenicity"]:
            for paragraph in doc.paragraphs:
                if f'{{{{ {key} }}}}' in paragraph.text:
                    paragraph.text = paragraph.text.replace(f'{{{{ {key} }}}}', str(value))
```

Figure 15 – This python code is responsible for replacing keyholders in word file.

Pdf Conversation –

Upon completing the template population, the script converts the Word document into a PDF file.

If variants are present, positive report will be generated.

If variants are not present, a negative report will be generated.

The scipt and templete are same; only JSON file generated would be different.

In the end, we get a pdf report positive or negative based on what variants are present in the Sample.

2.3 Result

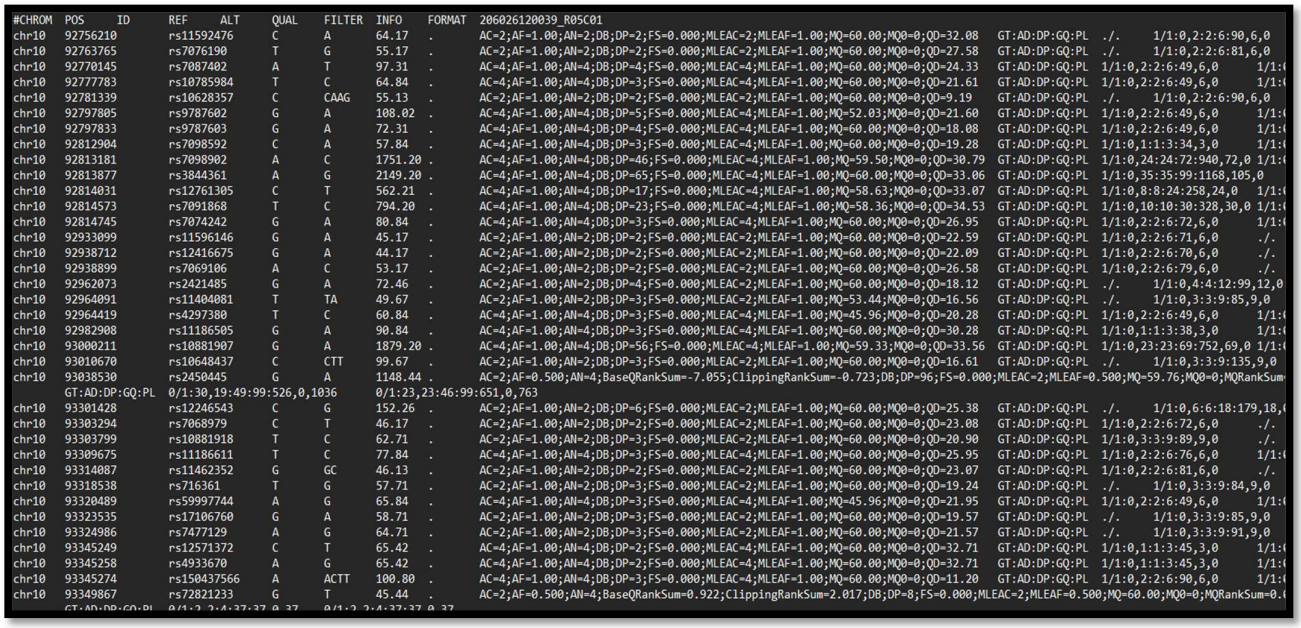
2.3.1 Immunodeficiency Database

Our deep research revealed about 3457 markers and 69 genes associated with the immunodeficiency. For each marker, relevant pathogenic and likely pathogenic criteria were established. Each variant entry included information like -

- rsid
- Gene Symbol
- Amino acid change
- Ref and Alt allele
- Reference paper

2.3.2 Variant File

GATK resulted in an annotated variant file, which contains all the rsIDs found in the sample.

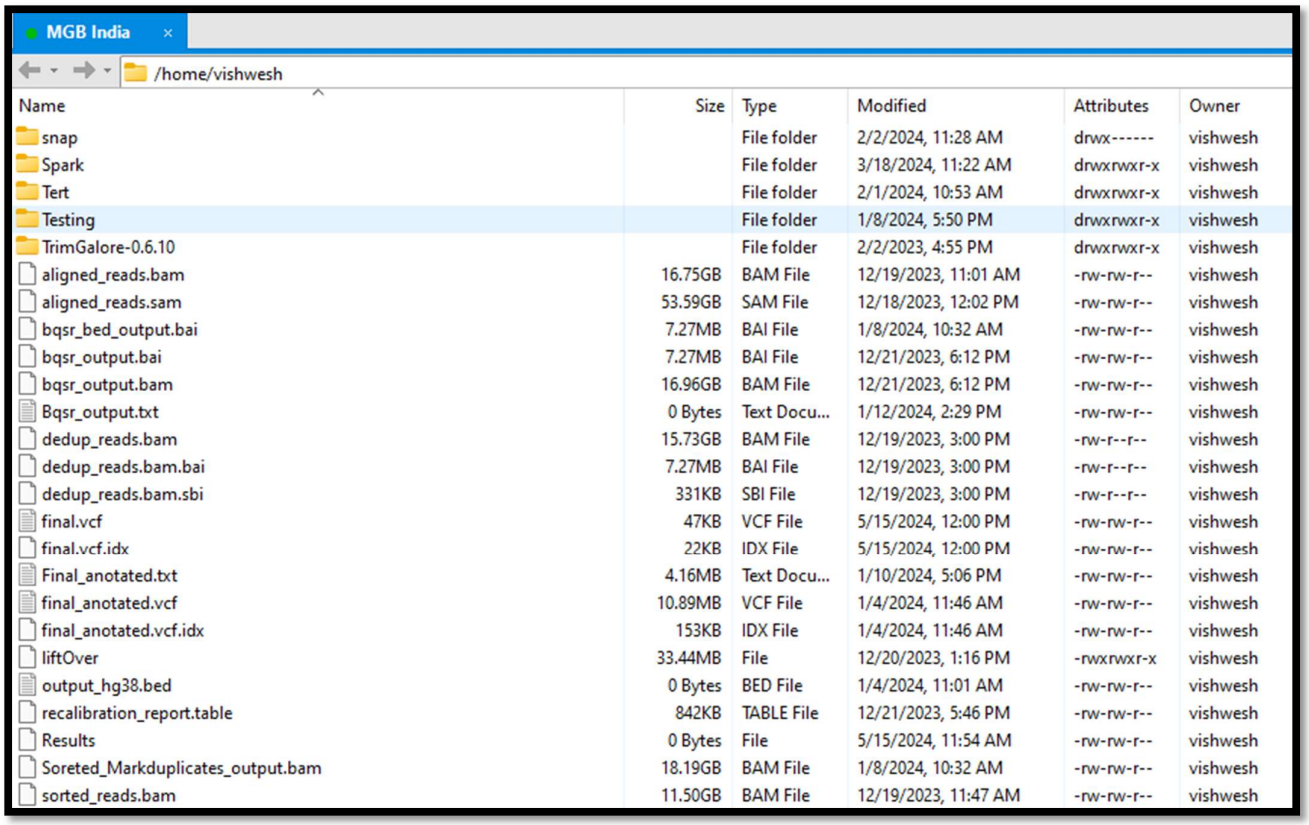


#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	206026120039_R05C01
chr10	92756210	rs11592476	C	A	64.17	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=32.08	GT:AD:DP:GQ:PL	./. 1/1:0,2:2:6:90,6,0
chr10	92763765	rs7076190	T	G	55.17	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=27.58	GT:AD:DP:GQ:PL	./. 1/1:0,2:2:6:81,6,0
chr10	92770145	rs7087402	A	T	97.31	.	AC=4;AF=1.00;AN=4;DB;DP=4;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=24.33	GT:AD:DP:GQ:PL	1/1:0,2:2:6:49,6,0 1/1:
chr10	92777783	rs10785984	T	C	64.84	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=21.61	GT:AD:DP:GQ:PL	1/1:0,2:2:6:49,6,0 1/1:
chr10	92781339	rs10628357	C	CAAG	55.13	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=9.19	GT:AD:DP:GQ:PL	./. 1/1:0,2:2:6:90,6,0
chr10	92797805	rs9787602	G	A	108.02	.	AC=4;AF=1.00;AN=4;DB;DP=5;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=52.03;MQ0=0;QD=21.60	GT:AD:DP:GQ:PL	1/1:0,2:2:6:49,6,0 1/1:
chr10	92797833	rs9787603	G	A	72.31	.	AC=4;AF=1.00;AN=4;DB;DP=4;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=18.08	GT:AD:DP:GQ:PL	1/1:0,2:2:6:49,6,0 1/1:
chr10	92812904	rs7098592	C	A	57.84	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=19.28	GT:AD:DP:GQ:PL	1/1:0,1:1:3:34,3,0 1/1:
chr10	92813181	rs7098902	A	C	1751.20	.	AC=4;AF=1.00;AN=4;DB;DP=46;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=59.50;MQ0=0;QD=30.79	GT:AD:DP:GQ:PL	1/1:0,24:24:72:940,72,0 1/1:
chr10	92813877	rs3844361	A	G	2149.20	.	AC=4;AF=1.00;AN=4;DB;DP=65;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=33.06	GT:AD:DP:GQ:PL	1/1:0,35:35:99:1168,105,0
chr10	92814031	rs12761305	C	T	562.21	.	AC=4;AF=1.00;AN=4;DB;DP=17;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=58.63;MQ0=0;QD=33.07	GT:AD:DP:GQ:PL	1/1:0,8:8:24:258,24,0 1/1:
chr10	92814573	rs7091868	T	C	794.20	.	AC=4;AF=1.00;AN=4;DB;DP=23;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=58.36;MQ0=0;QD=34.53	GT:AD:DP:GQ:PL	1/1:0,10:10:30:328,30,0 1/1:
chr10	92814745	rs7074242	G	A	80.84	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=26.95	GT:AD:DP:GQ:PL	1/1:0,2:2:6:72,6,0 1/1:
chr10	92933099	rs11596146	G	A	45.17	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=22.59	GT:AD:DP:GQ:PL	1/1:0,2:2:6:71,6,0 ./.
chr10	92938712	rs12416675	G	A	44.17	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=22.09	GT:AD:DP:GQ:PL	1/1:0,2:2:6:70,6,0 ./.
chr10	92938899	rs7069106	A	C	53.17	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=26.58	GT:AD:DP:GQ:PL	1/1:0,2:2:6:79,6,0 ./.
chr10	92962073	rs2421485	G	A	72.46	.	AC=2;AF=1.00;AN=2;DB;DP=4;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=18.12	GT:AD:DP:GQ:PL	./. 1/1:0,4:4:12:99,12,0
chr10	92964091	rs11404081	T	TA	49.67	.	AC=2;AF=1.00;AN=2;DB;DP=3;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=53.44;MQ0=0;QD=16.56	GT:AD:DP:GQ:PL	./. 1/1:0,3:3:9:85,9,0
chr10	92964419	rs4297380	T	C	60.84	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=45.96;MQ0=0;QD=20.28	GT:AD:DP:GQ:PL	1/1:0,2:2:6:49,6,0 1/1:
chr10	92982908	rs11186505	G	A	90.84	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=30.28	GT:AD:DP:GQ:PL	1/1:0,1:1:3:38,3,0 1/1:
chr10	93000211	rs10881907	G	A	1879.20	.	AC=4;AF=1.00;AN=4;DB;DP=56;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=59.33;MQ0=0;QD=33.56	GT:AD:DP:GQ:PL	1/1:0,23:23:69:752,69,0 1/1:
chr10	93010670	rs10648437	C	CTT	99.67	.	AC=2;AF=1.00;AN=2;DB;DP=3;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=16.61	GT:AD:DP:GQ:PL	./. 1/1:0,3:3:9:135,9,0
chr10	93038530	rs2450445	G	A	1148.44	.	AC=2;AF=0.500;AN=4;BaseQRankSum=-7.055;ClippingRankSum=-0.723;DB;DP=96;FS=0.000;MLEAC=2;MLEAF=0.500;MQ=59.76;MQ0=0;MQRankSum=		
	GT:AD:DP:GQ:PL	0/1:30,19:49:99:526,0,1036					0/1:23,23:46:99:651,0,763		
chr10	93301428	rs12246543	C	G	152.26	.	AC=2;AF=1.00;AN=2;DB;DP=6;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=25.38	GT:AD:DP:GQ:PL	./. 1/1:0,6:6:18:179,18,(
chr10	93303294	rs7068979	C	T	46.17	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=23.08	GT:AD:DP:GQ:PL	1/1:0,2:2:6:72,6,0 ./.
chr10	93303799	rs10881918	T	C	62.71	.	AC=2;AF=1.00;AN=2;DB;DP=3;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=20.90	GT:AD:DP:GQ:PL	1/1:0,3:3:9:89,9,0 ./.
chr10	93309675	rs11186611	T	C	77.84	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=25.95	GT:AD:DP:GQ:PL	1/1:0,2:2:6:76,6,0 1/1:
chr10	93314087	rs11462352	G	GC	46.13	.	AC=2;AF=1.00;AN=2;DB;DP=2;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=23.07	GT:AD:DP:GQ:PL	1/1:0,2:2:6:81,6,0 ./.
chr10	93318538	rs716361	T	G	57.71	.	AC=2;AF=1.00;AN=2;DB;DP=3;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=19.24	GT:AD:DP:GQ:PL	./. 1/1:0,3:3:9:84,9,0
chr10	93320489	rs59997744	A	G	65.84	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=45.96;MQ0=0;QD=21.95	GT:AD:DP:GQ:PL	1/1:0,2:2:6:49,6,0 1/1:
chr10	93323535	rs17106760	G	A	58.71	.	AC=2;AF=1.00;AN=2;DB;DP=3;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=19.57	GT:AD:DP:GQ:PL	./. 1/1:0,3:3:9:85,9,0
chr10	93324986	rs7477129	A	G	64.71	.	AC=2;AF=1.00;AN=2;DB;DP=3;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=60.00;MQ0=0;QD=21.57	GT:AD:DP:GQ:PL	./. 1/1:0,3:3:9:91,9,0
chr10	93345249	rs12571372	C	T	65.42	.	AC=4;AF=1.00;AN=4;DB;DP=2;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=32.71	GT:AD:DP:GQ:PL	1/1:0,1:1:3:45,3,0 1/1:
chr10	93345258	rs4933670	A	G	65.42	.	AC=4;AF=1.00;AN=4;DB;DP=2;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=32.71	GT:AD:DP:GQ:PL	1/1:0,1:1:3:45,3,0 1/1:
chr10	93345274	rs150437566	A	ACTT	100.80	.	AC=4;AF=1.00;AN=4;DB;DP=3;FS=0.000;MLEAC=4;MLEAF=1.00;MQ=60.00;MQ0=0;QD=11.20	GT:AD:DP:GQ:PL	1/1:0,2:2:6:90,6,0 1/1:
chr10	93349867	rs72821233	G	T	45.44	.	AC=2;AF=0.500;AN=4;BaseQRankSum=0.922;ClippingRankSum=2.017;DB;DP=8;FS=0.000;MLEAC=2;MLEAF=0.500;MQ=60.00;MQ0=0;MQRankSum=0.(		
	GT:AD:DP:GQ:PL	0/1:2,2:4:37:37,0,37					0/1:2,2:4:37:37,0,37		

Figure 16 – A portion of VCF file

Header of the file contains:– Chromosome, Position, rsIDs , Reference allele, Alternate allele. Quality score, Filter, Information

2.3.3 GATK pipeline produces a series of intermediate and final files. The below screenshot shows different files produced.



Name	Size	Type	Modified	Attributes	Owner
snap		File folder	2/2/2024, 11:28 AM	drwx------	vishwesh
Spark		File folder	3/18/2024, 11:22 AM	drwxrwxr-x	vishwesh
Tert		File folder	2/1/2024, 10:53 AM	drwxrwxr-x	vishwesh
Testing		File folder	1/8/2024, 5:50 PM	drwxrwxr-x	vishwesh
TrimGalore-0.6.10		File folder	2/2/2023, 4:55 PM	drwxrwxr-x	vishwesh
aligned_reads.bam	16.75GB	BAM File	12/19/2023, 11:01 AM	-rw-rw-r--	vishwesh
aligned_reads.sam	53.59GB	SAM File	12/18/2023, 12:02 PM	-rw-rw-r--	vishwesh
bqsr_bed_output.bai	7.27MB	BAI File	1/8/2024, 10:32 AM	-rw-rw-r--	vishwesh
bqsr_output.bai	7.27MB	BAI File	12/21/2023, 6:12 PM	-rw-rw-r--	vishwesh
bqsr_output.bam	16.96GB	BAM File	12/21/2023, 6:12 PM	-rw-rw-r--	vishwesh
Bqsr_output.txt	0 Bytes	Text Docu...	1/12/2024, 2:29 PM	-rw-rw-r--	vishwesh
dedup_reads.bam	15.73GB	BAM File	12/19/2023, 3:00 PM	-rw-r--r--	vishwesh
dedup_reads.bam.bai	7.27MB	BAI File	12/19/2023, 3:00 PM	-rw-r--r--	vishwesh
dedup_reads.bam.sbi	331KB	SBI File	12/19/2023, 3:00 PM	-rw-r--r--	vishwesh
final.vcf	47KB	VCF File	5/15/2024, 12:00 PM	-rw-rw-r--	vishwesh
final.vcf.idx	22KB	IDX File	5/15/2024, 12:00 PM	-rw-rw-r--	vishwesh
Final_anotated.txt	4.16MB	Text Docu...	1/10/2024, 5:06 PM	-rw-rw-r--	vishwesh
final_anotated.vcf	10.89MB	VCF File	1/4/2024, 11:46 AM	-rw-rw-r--	vishwesh
final_anotated.vcf.idx	153KB	IDX File	1/4/2024, 11:46 AM	-rw-rw-r--	vishwesh
liftOver	33.44MB	File	12/20/2023, 1:16 PM	-rwxrwxr-x	vishwesh
output_hg38.bed	0 Bytes	BED File	1/4/2024, 11:01 AM	-rw-rw-r--	vishwesh
recalibration_report.table	842KB	TABLE File	12/21/2023, 5:46 PM	-rw-rw-r--	vishwesh
Results	0 Bytes	File	5/15/2024, 11:54 AM	-rw-rw-r--	vishwesh
Soreted_Markduplicates_output.bam	18.19GB	BAM File	1/8/2024, 10:32 AM	-rw-rw-r--	vishwesh
sorted_reads.bam	11.50GB	BAM File	12/19/2023, 11:47 AM	-rw-rw-r--	vishwesh

Figure 17 – GATK files produced on our server

2.3.4 JSON output.

The initial variant identification pipeline creates a JSON output. This file served as structure repository of filtered and prioritized variants.

Below is the JSON output of the test result run on our sample.

Figure:– This shows two variants are found in our result and their appropriate columns

Figure 18 – Json file

This figure shows the identified genes and diseases according to the identified markers. The code, once identified markers, reads the disease and gene description file, finds appropriate information, and prints the gene and disease name in the JSON output.

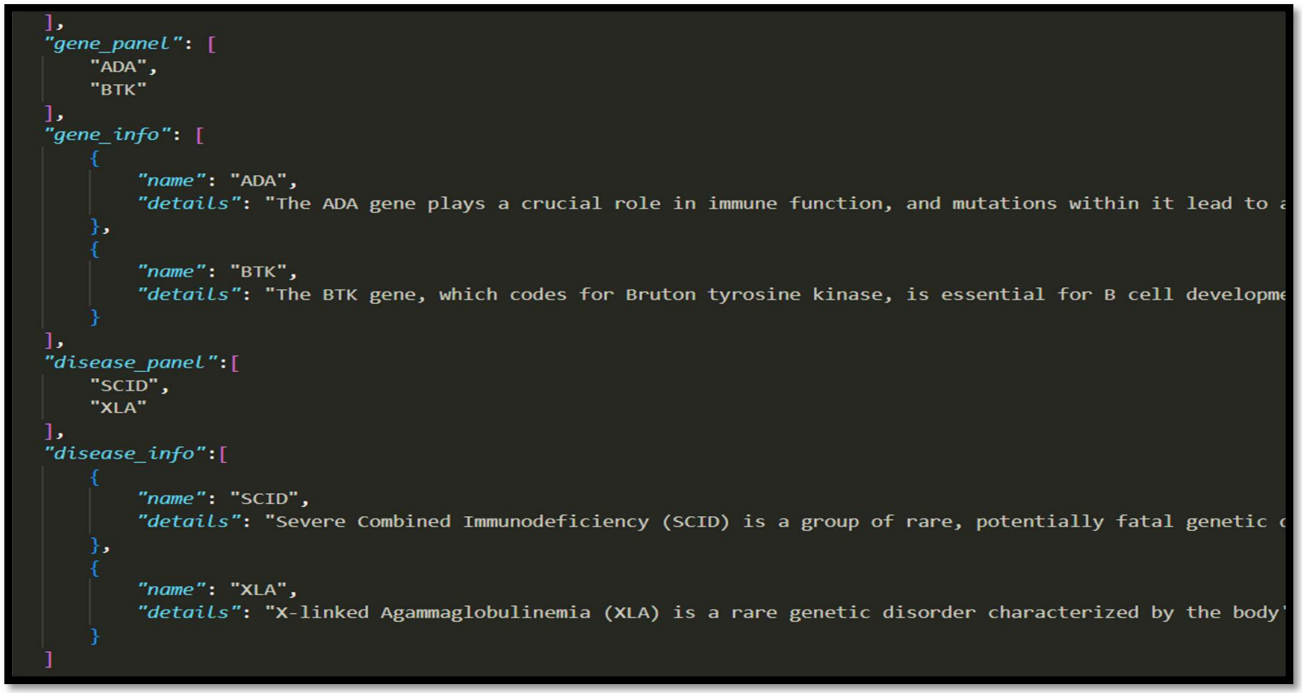


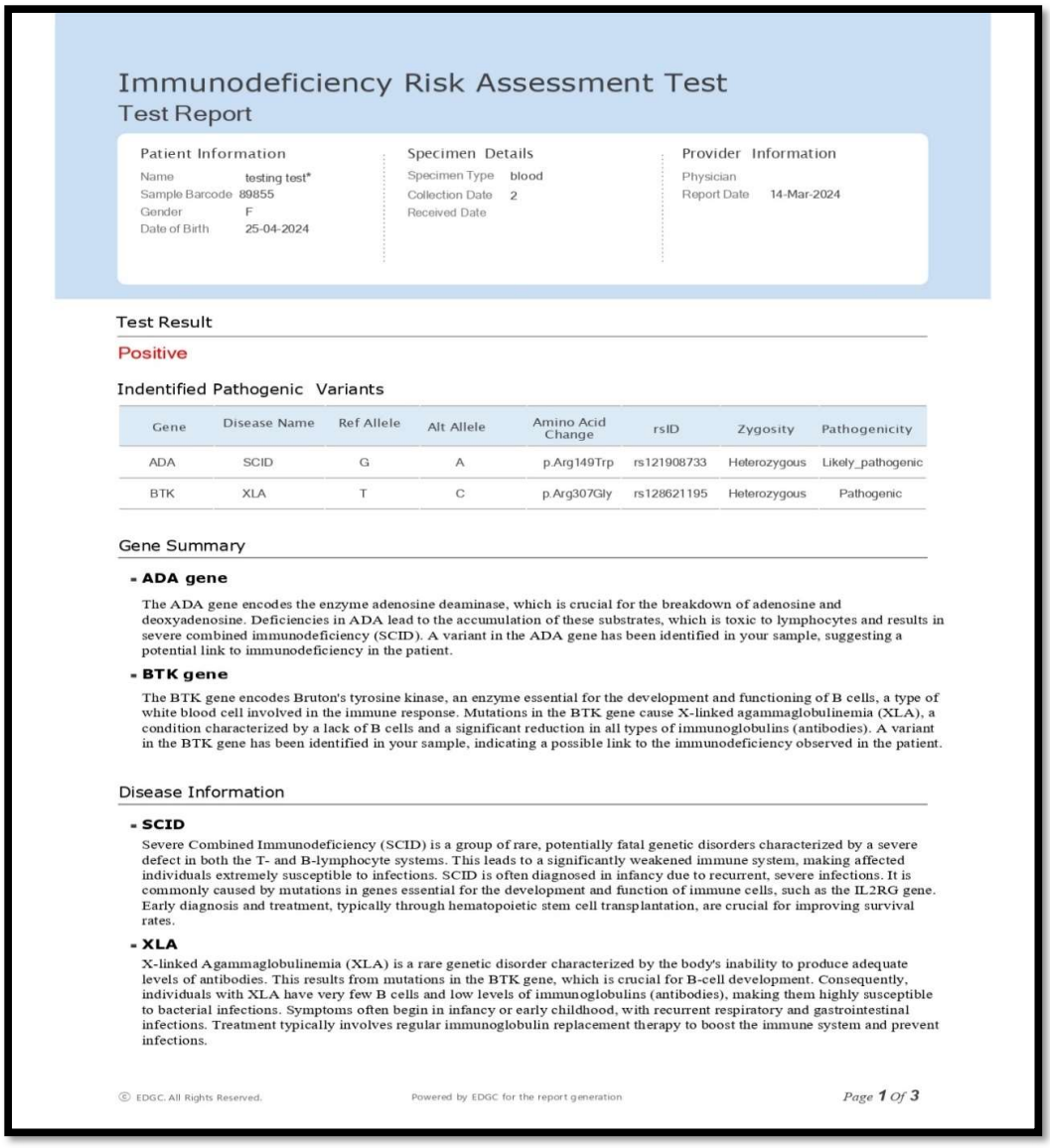
Figure 19 – JSON file

2.3.5 Final Pdf report

The final Pdf report contains 3 pages.

- The first page tells which pathogenic variant is identified.
- The second page is a disease and gene panel that has been curated by us.
- The third page has information about the test and disclaimer.

Below are the 3 pages of the pdf report that was generated when we ran our Garvan NGS samples.



Immunodeficiency Risk Assessment Test

Test Report

Patient Information		Specimen Details		Provider Information	
Name	testing test*	Specimen Type	blood	Physician	
Sample Barcode	89855	Collection Date	2	Report Date	14-Mar-2024
Gender	F	Received Date			
Date of Birth	25-04-2024				

Test Result

Positive

Indentified Pathogenic Variants

Gene	Disease Name	Ref Allele	Alt Allele	Amino Acid Change	rsID	Zygosity	Pathogenicity
ADA	SCID	G	A	p.Arg149Trp	rs121908733	Heterozygous	Likely_pathogenic
BTK	XLA	T	C	p.Arg307Gly	rs128621195	Heterozygous	Pathogenic

Gene Summary

- **ADA gene**

The ADA gene encodes the enzyme adenosine deaminase, which is crucial for the breakdown of adenosine and deoxyadenosine. Deficiencies in ADA lead to the accumulation of these substrates, which is toxic to lymphocytes and results in severe combined immunodeficiency (SCID). A variant in the ADA gene has been identified in your sample, suggesting a potential link to immunodeficiency in the patient.

- **BTK gene**

The BTK gene encodes Bruton's tyrosine kinase, an enzyme essential for the development and functioning of B cells, a type of white blood cell involved in the immune response. Mutations in the BTK gene cause X-linked agammaglobulinemia (XLA), a condition characterized by a lack of B cells and a significant reduction in all types of immunoglobulins (antibodies). A variant in the BTK gene has been identified in your sample, indicating a possible link to the immunodeficiency observed in the patient.

Disease Information

- **SCID**

Severe Combined Immunodeficiency (SCID) is a group of rare, potentially fatal genetic disorders characterized by a severe defect in both the T- and B-lymphocyte systems. This leads to a significantly weakened immune system, making affected individuals extremely susceptible to infections. SCID is often diagnosed in infancy due to recurrent, severe infections. It is commonly caused by mutations in genes essential for the development and function of immune cells, such as the IL2RG gene. Early diagnosis and treatment, typically through hematopoietic stem cell transplantation, are crucial for improving survival rates.

- **XLA**

X-linked Agammaglobulinemia (XLA) is a rare genetic disorder characterized by the body's inability to produce adequate levels of antibodies. This results from mutations in the BTK gene, which is crucial for B-cell development. Consequently, individuals with XLA have very few B cells and low levels of immunoglobulins (antibodies), making them highly susceptible to bacterial infections. Symptoms often begin in infancy or early childhood, with recurrent respiratory and gastrointestinal infections. Treatment typically involves regular immunoglobulin replacement therapy to boost the immune system and prevent infections.

Powered by EDGC for the report generation

Page 1 Of 3

This is the first page giving information about which pathogenic variants are found and their respective gene and disease information.

What are Primary Immunodeficiency disorders?

Primary immunodeficiencies (PIDs) are a group of over 400 rare, inherited disorders that disrupt the normal functioning of the immune system. When the immune system isn't working properly, individuals can become more susceptible to infections, experience autoimmune complications, and face an increased risk of developing certain types of cancer. PIDs often manifest in infancy or early childhood, but some may not become apparent until later in life.

This genetic test analyzes your DNA for specific variants (changes) in genes known to cause PIDs, providing valuable information to confirm a diagnosis, personalize treatment plans, assess risks for family members, and ultimately improve long-term health through early detection and appropriate intervention.

Below is the table containing information about all the diseases and there corresponding genes which we found while creating our database.

Gene disease information

Disease Information	Genes Associated With it
Adrenoleukodystrophy	ABCD1
Duffy Blood group system	ACKR1
Spondyloenchondrodysplasia with immune dysregulation	ACP5
Severe combined immunodeficiency	ADA,AK2,ATP6AP1,BACH2, CARD11, CARMIL2, CD3D, CD3E, CORO1A, DCLRE1C, GINS1, IKBKB, IL7R, LCK, MTHFD1, PIK3CD, RAG1, ZAP70
Vasculitis	ADA2
Hyper-IgM syndrome type 2	AICDA, CD40LG, CD40, UNG
Inherited Immunodeficiency	ADA2,BTK, CD19, IKZF1, IL17RA, NFKB2, TERT
Hermansky-Pudlak syndrome 2	AP3B1, BLOC1S3, DTNBP1, HPS1, HPS5
Platelet abnormalities with eosinophilia and immune-mediated inflammatory disease	ARPC1B
Ataxia-telangiectasia syndrome	ATM, MRE11
RASopathy	BRAF, CBL
X-linked agammaglobulinemia	BTK
MHC class II deficiency	CIITA, NR2C2AP, RFX5
Autoimmune lymphoproliferative syndrome	CTLA4, FAS, PRKCD
Granulomatous disease	CYBA, CYBB, NCF2
Autosomal recessive hyper-IgE syndrome	DOCK8
Vici Syndrome	EPG5
T-cell immunodeficiency	FOXN1
Deafness-lymphedema-leukemia syndrome	GATA2
X-linked severe combined immunodeficiency	IL2RG, MAGT1, SH2D1A
Leukocyte adhesion deficiency 1	ITGB2
T-B+ severe combined immunodeficiency	JAK3
Chédiak-Higashi syndrome	LYST
Cernunnos-XLF deficiency	NHEJ1
Griscelli syndrome type 2	RAB27A
Nijmegen breakage syndrome-like disorder	RAD50
Congenital sideroblastic anemia-B-cell immunodeficiency-periodic fever-developmental delay syndrome	TRNT1
Wiskott-Aldrich syndrome	WAS
Lazy leukocyte syndrome	WDR1

This is the second page gives little information about the what primary immunodeficiency are and features a gene disease panel that we curated from our database. We have identified 29 diseases associated with PIDs.

Test Information

Test Background

This test is designed to assess the genetic risk of developing various primary immunodeficiency disorders (PIDs). Utilizing next-generation sequencing (NGS) technology, our analysis encompasses a comprehensive evaluation of 3457 variants across 69 genes known to be associated with PIDs. These variants have been carefully selected based on their established clinical relevance and association with different immune system components and pathways.

About the Report

This report offers insights on pathogenic/likely pathogenic variants found within the genes underlying primary immunodeficiency. These gene variants were elucidated from extracted genomic material analyzed on an Illumina Global Screening Array-based assay designed to detect clinically relevant single-nucleotide polymorphisms (SNPs). Detected variants were evaluated using genotype quality, minor allele frequency, and curated clinical database for pathogenicity. The findings of this report should be considered as educational to the patient and the clinician. The findings should not be interpreted as a medical diagnosis, advice, or consultation from a professional healthcare provider. In an event of a Positive Result (Pathogenic/Likely Pathogenic variant detected), patients are advised to refer to physicians for proper follow-up examination.

Classification of Sequence Variants

- Pathogenic mutation: There is significant evidence to suggest that this variant is a dominant high-risk pathogenic variant.
- Likely pathogenic mutation: There is evidence that this variant is a dominant high-risk pathogenic variant.

Test Methods

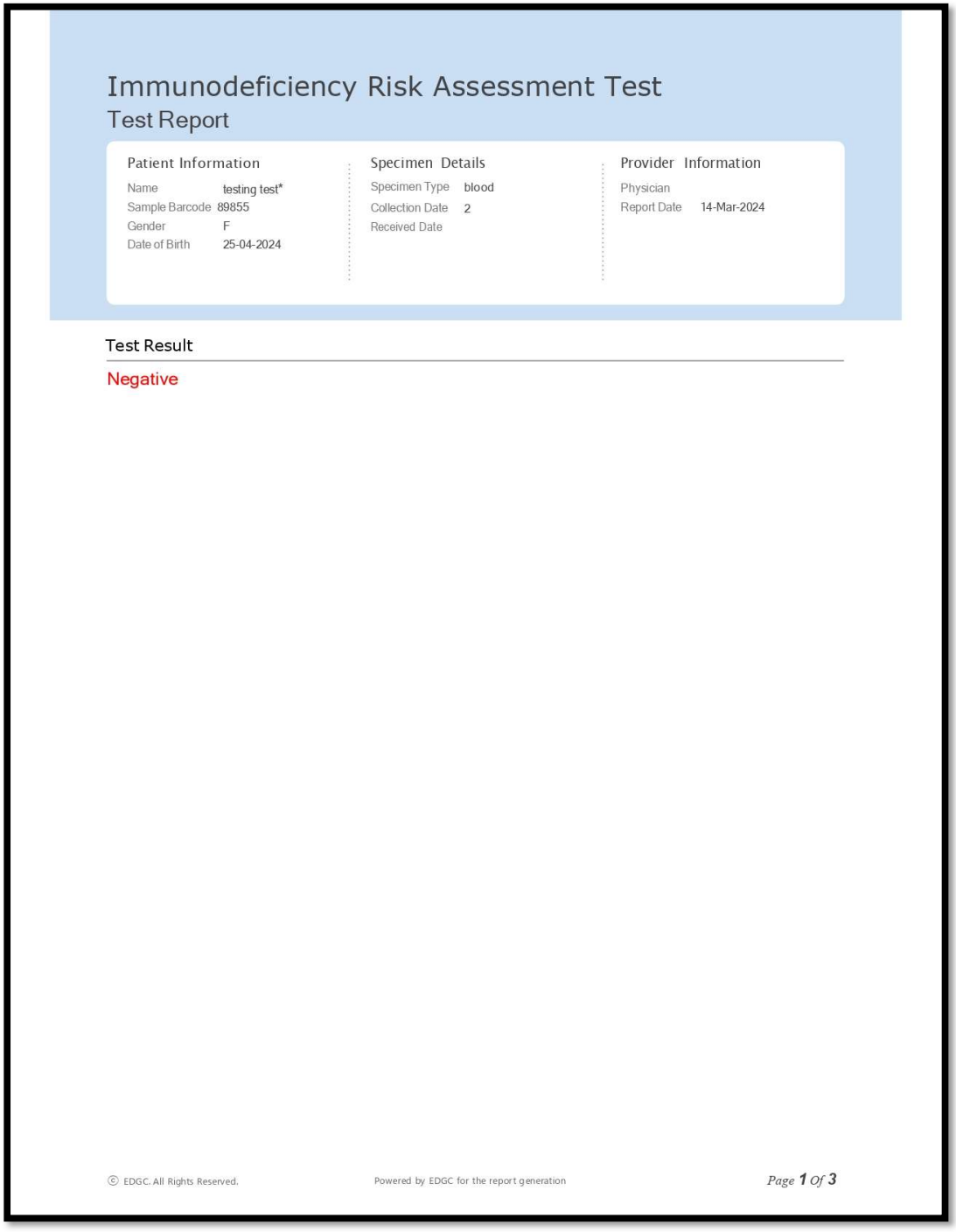
This report analyzes your genetic predisposition to primary immunodeficiencies (PIDs) using advanced next-generation sequencing (NGS) technology. Your DNA was examined for specific variants in a panel of genes known to be associated with PIDs. Variants were identified using the Genome Analysis Toolkit (GATK), annotated with relevant information, and then filtered and prioritized using a custom Python-based pipeline. Variants were classified according to established guidelines (ACMG/AMP), and their clinical significance was interpreted based on current scientific knowledge. This report does not provide a medical diagnosis but serves as an educational resource for you and your healthcare provider. Consultation with a qualified healthcare professional is strongly recommended for further interpretation and guidance.

Disclaimer

This report assesses your genetic risk for primary immunodeficiencies (PIDs) by analyzing specific variants in a curated panel of genes. While comprehensive, this test may not detect all possible disease-causing variants due to technological limitations and the vastness of the human genome. This report does not account for individual variations or non-genetic factors influencing PID risk. A negative result does not rule out the possibility of having a PID, and a positive result does not guarantee a diagnosis. The findings should be interpreted in conjunction with clinical evaluations and family history. Please consult a qualified healthcare professional for further assessment, diagnosis, and personalized guidance.

Diagnomics
5050 Murphy Canyon Road Ste 150, San Diego, CA 92123 USA
info@diagnomics.com
EDGC
EONE-DIAGNOMICS Genome Center

This is the last page it contains information about the test and has disclaimer.



Immunodeficiency Risk Assessment Test

Test Report

Patient Information

Name testing test*
Sample Barcode 89855
Gender F
Date of Birth 25-04-2024

Specimen Details

Specimen Type blood
Collection Date 2
Received Date

Provider Information

Physician
Report Date 14-Mar-2024

Test Result

Negative

Powered by EDGC for the report generation

For the samples where no pathogenic variants are found a negative report will be generated. This report will not have table instead it will just show the negative message and rest of the report will remain same.

3. Review of the task performed

3.1 Critical Analysis of work carried out

Creating this pipeline for identifying pathogenic variants in patients with primary immunodeficiencies (PIDs) is a big step in the right direction for expediting diagnosis and enhancing patient outcomes. The pipeline's strength lies in its approach, which includes a well-curated database of genetic markers linked to PID, stringent quality control methods, and an easy-to-understand report format.

A combination of Genome Analysis Toolkit (GATK) and custom Python scripts allowed analysis of the data and identifying variants, prioritizing pathogenic candidates, and generating a comprehensive PDF report for clinical interpretation.

After a review from the Korea team, this test will be live on the MygenomeBox website.

The only limitation of this pipeline is it focuses on a specific set of known primary immunodeficiency genes, which may not encompass all potential genetic causes of PIDs. Further research and expansion of the database is required to enhance pipeline accuracy.

Despite these limitations, the clear and concise report format facilitates effective communication between the laboratory and the clinician, enabling timely decision-making regarding further testing, treatment, and genetic counseling.

3.2 Challenges faced by student

Some minor challenges were faced by me during the course of my internship. Being a biotechnology student, learning about command line and Unix system, and Python scripting was initially difficult for me to get familiar with, but with practice and a proactive approach, I successfully passed these challenges. Many other challenges were faced on a day-to-day basis, which my team helped me solve.

4. References

1. Luo S. (2019). Enabling and Performance Benchmarking of a Next-generation Sequencing Data Analysis Pipeline.

2. K. R. Srinath (2017) Python – The Fastest Growing Programming Language

3. Heldenbrand, J.R., Baheti, S., Bockol, M.A. et al. Recommendations for performance optimizations when using GATK3.8 and GATK4. BMC Bioinformatics 20, 557 (2019).

4. Bellamy R. Genetic susceptibility to tuberculosis in human populationsThorax 1998;53:588-593.

5. Möller M, Kinnear CJ, Orlova M, Kroon EE, van Helden PD, Schurr E, Hoal EG. Genetic Resistance to Mycobacterium tuberculosis Infection and Disease. Front Immunol. 2018 Sep 27;9:2219.

6. Handysides S. Stalin lives, and is running the NHS. Br J Gen Pract. 2014 Sep;64(626):466.

7. Anastassopoulou, C., Gkizarioti, Z., Patrinos, G.P. et al. Human genetic factors associated with susceptibility to SARS-CoV-2 infection and COVID-19 disease severity. Hum Genomics 14, 40 (2020).

8. Godri Pollitt, K.J., Peccia, J., Ko, A.I. et al. COVID-19 vulnerability: the potential impact of genetic susceptibility and airborne transmission. Hum Genomics 14, 17 (2020).

9. Griffiths, A. J.F. (2024, May 14). mutation. Encyclopedia Britannica.

10. Justiz Vaillant AA, Qurie A. Immunodeficiency. [Updated 2023 Jun 26]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024

11. Raje N, Dinakar C. Overview of Immunodeficiency Disorders. Immunol Allergy Clin North Am. 2015 Nov;35(4):599-623.

12. Sitinjak BDP, Murdaya N, Rachman TA, Zakiyah N, Barliana MI. The Potential of Single Nucleotide Polymorphisms (SNPs) as Biomarkers and Their Association with the Increased Risk of Coronary Heart Disease: A Systematic Review. Vasc Health Risk Manag. 2023 May 5;19:289-301.

13. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015 May;17(5):405-24.

14. Fischer A. Severe combined immunodeficiencies (SCID). Clin Exp Immunol. 2000 Nov;122(2):143-9.

15. El-Sayed ZA, Abramova I, Aldave JC, Al-Herz W, Bezrodnik L, Boukari R, Bousfiha AA, Cancrini C, Condino-Neto A, Dbaibo G, Derfalvi B, Dogu F, Edgar JDM, Eley B, El-Owaidy RH, Espinosa-Padilla SE, Galal N, Haerynck F, Hanna-Wakim R, Hossny E, Ikinciogullari A, Kamal E, Kanegane H, Kechout N, Lau YL, Morio T, Moschese V, Neves JF, Ouederni M, Paganelli R, Paris K, Pignata C, Plebani A, Qamar FN, Qureshi S, Radhakrishnan N, Rezaei N, Rosario N, Routes J, Sanchez B, Sediva A, Seppanen MR, Serrano EG, Shcherbina A, Singh S, Siniah S, Spadaro G, Tang M, Vinet AM, Volokha A, Sullivan KE. X-linked agammaglobulinemia (XLA):Phenotype, diagnosis, and therapeutic challenges around the world. World Allergy Organ J. 2019 Mar 22;12(3)

16. Kim K, Seong MW, Chung WH, Park SS, Leem S, Park W, Kim J, Lee K, Park RW, Kim N. Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants. Genomics Inform. 2015 Jun;13(2)

17. Qin D. Next-generation sequencing and its clinical application. Cancer Biol Med. 2019 Feb;16(1)

18. Zhou, Y., Kathiresan, N., Yu, Z. et al. A high-performance computational workflow to accelerate GATK SNP detection across a 25-genome dataset. BMC Biol 22, 13 (2024).

19. Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li, Twelve years of SAMtools and BCFtools, GigaScience, Volume 10, Issue 2, February 2021

20. M. Vasimuddin, S. Misra, H. Li and S. Aluru, "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems," 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Rio de Janeiro, Brazil, 2019, pp. 314-324

21. MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011.

22. Šimec, Alen & Magličić,. (2014). Comparison of JSON and XML Data Formats.

23. X. Liu, S. Han, Z. Wang, J. Gelernter, and B.-Z. Yang, “Variant callers for next-generation sequencing data: a comparison study,” PLoS ONE, vol. 8, no. 9, Article ID e75619, 2013

24. G. A. Van der Auwera, M. O. Carneiro, C. Hartl et al., “From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline,” Current Protocols in Bioinformatics, vol. 43, 2013.

25. E. Garrison and G. Marth, “Haplotype-based variant detection from short-read sequencing,” http://arxiv.org/abs/1207.3907.

26. Cornish A, Guda C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Res Int. 2015;2015:456479