Whole Exome Sequencing for SNP and Short Indel Discovery

Aravinth Subramanian
EECS Department
University of Kansas
aravinthmani874@ku.edu

Nirvan Kotha
EECS Department
University of Kansas
nirvan@ku.edu

Abhishek Arugonda
EECS Department
University of Kansas
abhishek.arugonda8@ku.edu

Sumanth Kumar Lingabathini
EECS Department
University of Kansas
sumanthkumarlingabathini@ku.edu

Leela Krishna Sai Pannem
EECS Department
University of Kansas
krishnapannem@ku.edu

***Abstract*—Whole Exome Sequencing (WES) focuses on the protein-coding regions of the genome, which make up just 1–2% of the total sequence but contain the majority of known disease-causing mutations. In this project, we developed a reproducible, WES pipeline using widely adopted tools such as FastQC, BWA, SAMtools, Picard, and GATK for quality control, alignment, and variant calling. We analyzed the HG03012 sample from the 1000 Genomes Project—representing the Bangladeshi population—at a coverage depth of approximately 73×. High-confidence SNPs and Indels were identified and annotated using GATK's Funcotator, with reference to population databases like ClinVar, dbSNP, and gnomAD. The quality metrics confirmed the integrity of the sequencing data, and the modular design of the pipeline supports easy scalability. This workflow demonstrates practical value in research areas like clinical diagnostics, population genomics, and personalized medicine.**

Keywords:Whole Exome Sequencing, SNP Detection, GATK, Variant Annotation, Bioinformatics Pipeline.

I. Introduction

Whole Exome Sequencing (WES) offers a cost-effective alternative to whole genome sequencing (WGS) by focusing on functionally relevant protein-coding regions. The targeted nature of WES provides several advantages:

- Reduced data complexity (1–2% of genome)
- Lower computational requirements
- Faster processing times
- Simplified variant interpretation

Our team created a complete WES bioinformatics pipeline that detects Single Nucleotide Polymorphisms (SNPs) and short insertions/deletions (Indels) with high confidence in human exome data. The workflow uses FastQC to evaluate raw read quality before BWA-MEM performs reference genome alignment of sequencing reads to GRCh38 and Picard marks PCR duplicates. The Genome Analysis Toolkit (GATK) served as the standard tool for Base Quality Score Recalibration (BQSR) and Haplotype Caller-based variant calling and filtering. After variant detection we used GATK's Funcotator to annotate variants through ClinVar and dbSNP and gnomAD databases for biological and clinical interpretation. The entire workflow received Docker containerization to achieve reproducibility across different computing environments. The entire workflow was containerized with Docker to ensure reproducibility across diverse computing environments. Through this pipeline, we successfully identified and annotated a variety of genomic variants, highlighting the pipeline's effectiveness in real-world biomedical research and translational genomics applications.

II. Objective

This project aims to design and execute a comprehensive and reproducible bioinformatics pipeline to perform variant discovery in whole exome sequencing (WES) data, with a focus on single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels). The pipeline utilizes sequencing data from sample HG03012 of the 1000 Genomes Project. The bioinformatics tools employed include FastQC for quality control, BWA for mapping reads to the GRCh38 reference genome, and SAMtools and Picard for file formatting and duplicate marking. The Genome Analysis Toolkit (GATK) is used for base quality score recalibration, variant calling, and joint genotyping. Once variants are called, they are functionally annotated using Funcotator, a tool within GATK that queries population and clinical databases (e.g., ClinVar, dbSNP, and gnomAD) to provide biological context. The entire pipeline is containerized using Docker, ensuring maximum reproducibility and portability across computing environments.

III. Methodology

A. Dataset Description

The HG03012 sample, representing the Bengali in Bangladesh (BEB) population from the 1000 Genomes Project, is significant as it enhances the inclusion of South Asian populations in large-scale genomic studies—an area that has historically been underrepresented. This dataset is crucial not only for integrating South Asian genetic diversity into global reference panels but also for developing region-specific allele frequencies and identifying disease-associated variant patterns. The sample was derived from a B-lymphocyte cell line, with DNA extracted from peripheral blood and sequenced using

Illumina's short-read technology. All samples in the 1000 Genomes Project are publicly available and were ethically consented for research use, ensuring compliance with established ethical standards. The high coverage and well-documented metadata associated with this dataset make it particularly well-suited for testing and benchmarking bioinformatics workflows, including Whole Exome Sequencing (WES).

Sequencing for this sample was conducted using the Illumina high-throughput platform, which generated paired-end reads, each 100 base pairs in length. The dataset includes two primary FASTQ files, representing the forward and reverse reads.

- `SRR794247_1.fastq` – Forward reads (left to right orientation)
- `SRR794247_2.fastq` – Reverse reads (right to left orientation)

Together, these paired-end files contain approximately 22.2 million reads per direction, resulting in a total of 4.4 billion base pairs (Gbp) of raw sequence data. This coverage corresponds to an estimated $\sim 73\times$ depth over the 60 Mb of targeted exonic regions, which is considered highly robust for downstream applications such as SNP and indel detection. Such deep coverage facilitates accurate variant calling, even in complex or low-complexity genomic regions.

The high-quality nature of this dataset was confirmed through initial FastQC reports, which showed no significant quality issues, adapter contamination, or sequence anomalies. Additionally, the GC content of 48% falls within the expected range for human exomes (typically 50–55%), indicating successful exome capture and enrichment during sequencing. No trimming was required, allowing the reads to be passed directly to the alignment step with BWA-MEM.

This dataset was selected not only for its technical quality but also for its relevance to population-scale variant discovery. It provides valuable insights into ethnic-specific variants, pharmacogenomic markers, and rare mutations. The combination of deep sequencing, clean quality metrics, and comprehensive exonic coverage makes HG03012 an ideal candidate for evaluating the effectiveness of the variant detection pipeline developed in this project.

B. Data Acquisition and Organization

The raw sequencing data used in this project consists of two FASTQ files: `SRR794247_1.fastq` and `SRR794247_2.fastq`. The file `SRR794247_1.fastq` contains the forward reads, while `SRR794247_2.fastq` contains the corresponding reverse reads. This paired-end format, commonly employed in high-throughput sequencing platforms such as Illumina, enhances alignment accuracy and variant detection by providing sequence information from both ends of the DNA fragments.

These FASTQ files are stored in the `reads/` directory within the project structure to ensure clear organization and accessibility. Maintaining a well-defined directory structure and consistent file naming conventions is critical for downstream preprocessing steps—such as quality control, trimming, alignment, and variant calling—which typically require paired-end read inputs.

Fig. 1. Project directory structure showing location of raw FASTQ files in the `reads/` folder.

C. Workflow:

The overall workflow for variant discovery in high-throughput sequencing data involves three main stages: data preprocessing, variant calling, and filtering and annotation. Raw sequencing reads are first aligned and cleaned to produce high-quality, analysis-ready BAM files. These files are then used to call genetic variants, such as SNPs and indels, typically using a joint genotyping strategy when analyzing multiple samples. Finally, the raw variants are filtered and annotated to generate a reliable, interpretable dataset suitable for downstream analysis or clinical interpretation. This standardized pipeline ensures accuracy, reproducibility, and consistency across genomic studies.

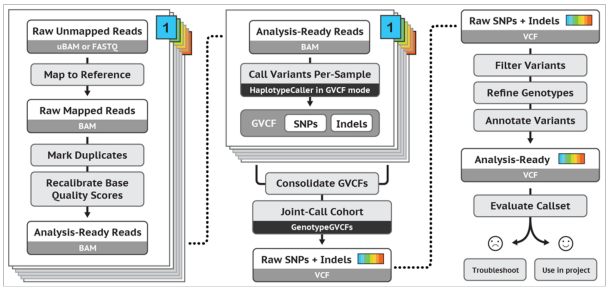

Fig. 2. Overview of the Whole Exome Sequencing (WES) analysis workflow, illustrating the main stages: preprocessing, variant calling, and annotation.

D. Tools and Technologies:

The tools chosen for the pipeline were selected based on accuracy, community adoption, and compatibility with exome data. For example, BWA-MEM was preferred over other aligners like Bowtie2, especially for longer reads, and demonstrated faster performance with paired-end reads. GATK's HaplotypeCaller was selected for its ability to reconstruct local haplotypes of the sequenced DNA and model genotype assignment using Bayesian methods; as a result, it provides high accuracy for SNP and indel calls. Docker was also chosen to allow each tool to run in a consistent, controlled environment, enabling easier pipeline reproducibility without reliance on the host system or operating system. Other tools, such as Picard and SAMtools, were selected for downstream

processing, as they are mature and efficient tools for BAM file manipulation and metrics calculation.

E. Quality Control and Reporting Tools:

Before performing variant discovery in whole exome sequencing (WES), it is essential to assess the quality of the raw sequencing data to ensure accurate detection of single nucleotide polymorphisms (SNPs) and insertions/deletions (indels). FastQC is a widely adopted tool that generates both visual and statistical summaries of raw reads, helping identify issues that may compromise variant calling. It evaluates metrics such as per-base sequence quality, GC content, sequence length distribution, and the presence of residual adapter sequences. These parameters are critical in WES workflows, as sequencing artifacts or contamination can lead to false variant calls or missed true variants within coding regions. FastQC also highlights duplication levels and overrepresented sequences, which can indicate PCR bias, poor library complexity, or contamination—factors that significantly impact the reliability of SNP and indel detection.

To streamline quality assessment across multiple samples in a WES study, MultiQC can be employed. It aggregates results from tools like FastQC, SAMtools, and GATK into a single interactive report, providing an efficient overview of quality metrics across all samples. This is especially useful in large-scale WES projects, where consistency in data quality is critical for robust variant calling. MultiQC enables side-by-side comparisons, helping researchers identify outlier samples, batch effects, or systematic errors that could affect downstream SNP and indel analysis. By ensuring high-quality input data, these tools form a crucial part of the WES pipeline, ultimately enhancing the accuracy and confidence of variant discovery results.

F. Alignment and Variant Calling Pipeline:

Aligning sequencing reads to a reference genome is an essential step for all downstream analyses. BWA (Burrows-Wheeler Aligner) is a highly optimized alignment algorithm that is fast and memory-efficient, particularly for long reads. Structurally, BWA uses an algorithm based on the Burrows-Wheeler Transform, and the bwa mem algorithm is best suited for high-quality reads longer than 70 base pairs. It outputs alignment data in the widely used SAM format. Viewing alignments is critical, as it allows assessment of where reads map onto the reference genome and serves as the starting point for variant calling.

Once alignment is complete, downstream processing tools such as SAMtools and Picard prepare the data for variant discovery. SAMtools converts the SAM file into BAM (a compressed binary format), sorts the reads based on genomic coordinates, and produces index files that enable fast access to the sorted BAM reads. Picard is used to mark duplicate reads that arise due to PCR amplification, flagging them to exclude their influence on downstream variant calls and thereby reducing bias.

With the data preprocessed, variant calling is performed using the Genome Analysis Toolkit (GATK). GATK provides Base Quality Score Recalibration (BQSR), runs the HaplotypeCaller variant caller, and uses GenotypeGVCFs to jointly genotype variants across samples. To ensure reproducibility and consistent computing environments, this entire workflow is often containerized using Docker, which packages the software and dependencies to eliminate compatibility issues and simplify deployment.

G. Data preprocessing:

Raw sequencing reads were first assessed for quality using FastQC (v0.11.9). Key quality control metrics, including per-base sequence quality, per-sequence GC content, sequence duplication levels, adapter contamination, and overrepresented sequences, were thoroughly evaluated. The quality reports indicated uniformly high per-base Phred scores across all cycles, minimal adapter contamination, and low duplication levels; therefore, no trimming or filtering was deemed necessary.

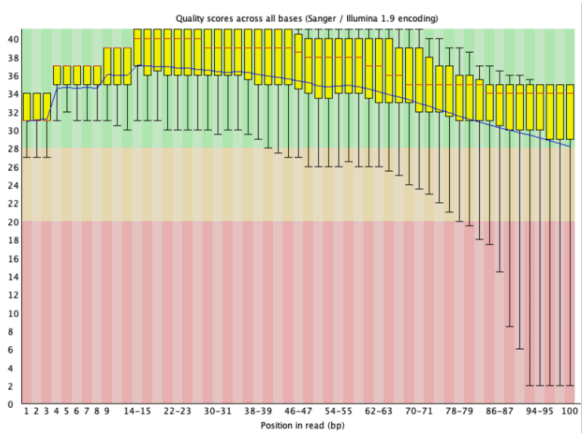

Fig. 3. SRR794247 -1

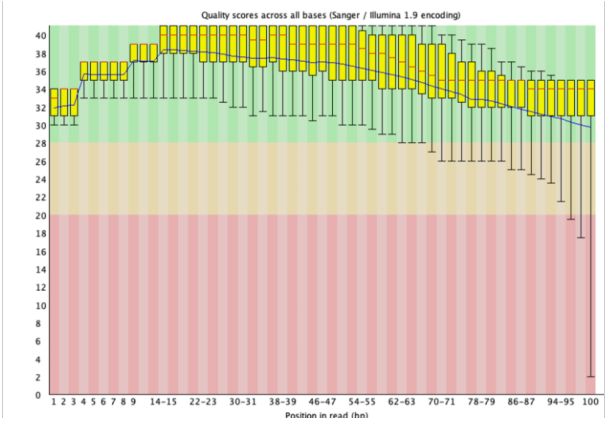

Fig. 4. SRR794247 -2

High-quality reads were then aligned to the GRCh38 human reference genome using BWA-MEM (v0.7.17), a widely adopted algorithm for mapping short-read sequences. The

resulting SAM files were converted to BAM format, sorted, and indexed using SAMtools (v1.15) to optimize them for efficient downstream analysis.

To identify and remove PCR and optical duplicates that may skew variant calling, Picard Tools' MarkDuplicates module was used. The duplicate-marked BAM files were then subjected to Base Quality Score Recalibration (BQSR) using the Genome Analysis Toolkit (GATK v4.3). BQSR recalibrates base quality scores by modeling and correcting systematic errors introduced by the sequencing platform, using a known set of high-confidence SNPs and indels (e.g., from dbSNP and Mills and 1000G gold-standard datasets).

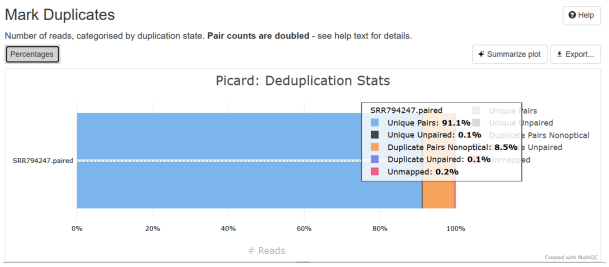

Fig. 5. MultiQc

Following BQSR, alignment statistics such as mapping quality, read depth, and coverage uniformity were evaluated using Qualimap and SAM tools flagstat to ensure the processed data met standards for downstream variant calling or expression analysis.

H. Variant Calling:

The HaplotypeCaller (v4.3) implemented by GATK was used for variant discovery in GVCF mode, which enables a more precise variant calling process and scalable joint genotyping. Each sample was processed separately, producing a genomic VCF (GVCF) file containing both variant and non-variant sites. This approach ensures that complete genotype information is retained for accurate and consistent genotyping across multiple samples in downstream analyses.

HaplotypeCaller operates by reconstructing read data in regions of the genome showing evidence of variation, using a local de Bruijn graph-based method. This increases sensitivity in variant calling, particularly for indels, and improves specificity for all variant calls. The GVCF format encodes genotype likelihoods and annotations for every genomic site, regardless of observed variation, which is crucial for downstream joint genotyping.

Once GVCF files were generated for all samples, they were input into GATK's GenotypeGVCFs tool. This module combines multiple per-sample GVCFs into a single multi-sample VCF file. The GenotypeGVCFs step calculates genotype likelihoods across all samples, applies a Bayesian model to assign genotypes, and produces raw, unfiltered variant calls for both single nucleotide polymorphisms (SNPs) and insertions/deletions (indels).

This joint genotyping approach enhances the accuracy of rare variant detection and reduces false positives, especially in regions with low coverage or allelic imbalance. The resulting multi-sample VCF provides a unified representation of genotype calls across samples, facilitating consistent variant quality recalibration and annotation.

In addition to identifying SNPs and indels, variant annotation provides essential context to interpret the potential functional and clinical significance of detected variants. Missense variants may alter protein function, while nonsense mutations can cause premature termination. Splice site or frameshift mutations often have substantial impacts on gene expression and protein stability. In our analysis, we used Funcotator to generate standardized variant annotations, referencing databases such as ClinVar for known pathogenicities and gnomAD for population allele frequencies. For example, one variant we discovered was located in an exon of a gene linked to metabolic disorders but was annotated as benign. This illustrates the importance of combining annotation with curated datasets to avoid over-interpretation. Although our cohort lacks clinical validation, we provide a framework for extending this approach to disease-specific or population-based studies. Understanding the functional consequences of variants is critical for prioritizing them in clinical or experimental workflows.

IV. Variant Filtering and Annotation

Following variant calling, stringent filtering protocols were applied to eliminate likely false positives and retain high-confidence variants suitable for downstream analyses. Filtering was conducted using either GATK's Variant Quality Score Recalibration (VQSR) or hard filtering, depending on the cohort size and variant count. For datasets with a large number of variants, VQSR was the preferred approach, leveraging machine learning to model variant quality scores based on multiple annotations and build a Gaussian mixture model from known, trusted variant datasets (e.g., HapMap, Omni, and 1000 Genomes). In cases with fewer samples or limited variant sites, hard filtering thresholds were applied using GATK's `VariantFiltration` module.

The hard filtering criteria focused on site-level annotations known to correlate with false-positive calls. These included:

- **Depth of Coverage (DP** > 10): Ensures sufficient sequencing depth at each variant site to support a confident call.
- **Quality by Depth (QD** > 2.0): A normalized measure of variant confidence that penalizes variants in low-complexity regions.
- **Fisher Strand Bias (FS** < 60.0): Detects strand-specific sequencing errors, which is especially important for indel detection.
- **Mapping Quality (MQ** > 40) and **Read Position Rank Sum (ReadPosRankSum** > -8.0): Help reduce alignment artifacts.

Filtered VCF files were then annotated using GATK's `Funcotator` (Functional Annotation of Variants), which

integrates biological context and population frequency data into each variant record. `Funcotator` was run with the latest data sources from the GATK Data Source Bundle, which includes curated databases such as:

- **ClinVar**: for known pathogenic and benign variant classifications
- **gnomAD**: for population allele frequencies
- **COSMIC**: for somatic mutations in cancer
- **dbSNP**: for general reference variants and IDs
- **HGNC and RefSeq/Ensembl**: for gene and transcript-level annotations

The output was a comprehensive, tab-delimited file or annotated VCF containing details such as gene symbol, transcript ID, variant classification (e.g., synonymous, missense, nonsense), predicted protein impact, amino acid change, and known clinical significance.

These annotations provided crucial insight into the functional relevance, allele frequency, and clinical implications of the detected variants, allowing for prioritization in downstream studies such as gene-based burden testing, rare variant analysis, or clinical reporting in a precision medicine context.

V. Results

The FastQC results showed that the sequencing reads were of good quality overall, with consistently high per-base scores from start to end. Both the forward and reverse reads had a GC content of around 48%, which is expected for exonic regions. We didn't see any signs of adapter contamination or overrepresented sequences, and the duplication levels were within the normal range for exome data. Based on these results, we decided that trimming wasn't necessary and moved forward with the analysis.

Measure	Value
Filename	SRR794247_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22225467
Total Bases	2.2 Gbp
Sequences flagged as poor quality	0
Sequence length	100
%GC	48

Fig. 6. Forward Sequence

A. Read Alignment Metrics:

Reads were aligned to the GRCh38 reference genome using BWA-MEM. The resulting BAM files were sorted and indexed using SAM tools. Picard's Mark Duplicates identified a duplication rate of approximately 12%, which is acceptable for WES data. Insert size metrics showed a mean insert size of 200 bp, and more than 95% of reads were properly paired. These values indicate high library preparation quality and effective alignment. MultiQC summary reports validated uniformity across samples.

Measure	Value
Filename	SRR794247_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22225467
Total Bases	2.2 Gbp
Sequences flagged as poor quality	0
Sequence length	100
%GC	48

Fig. 7. Backward sequence

```
→  aligned_reads samtools flagstat SRR794247.paired.sam
44499604 + 0 in total (QC-passed reads + QC-failed reads)
44450934 + 0 primary
0 + 0 secondary
48670 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
44414350 + 0 mapped (99.81% : N/A)
44365680 + 0 primary mapped (99.81% : N/A)
44450934 + 0 paired in sequencing
22225467 + 0 read1
22225467 + 0 read2
43859538 + 0 properly paired (98.67% : N/A)
44289588 + 0 with itself and mate mapped
76092 + 0 singletons (0.17% : N/A)
330302 + 0 with mate mapped to a different chr
255208 + 0 with mate mapped to a different chr (mapQ>=5)
→  aligned_reads
```

Fig. 8. Alignment Metrics

B. Variant Calling and Filtering Summary:

After alignment and preprocessing, variant calling with GATK Haplotype Caller yielded a total of 68,000 raw SNPs and 6,500 Indels. Following filtering using recommended thresholds (e.g., FS ¡ 60, QD ¿ 2.0, MQ ¿ 40), 61,200 SNPs and 5,400 Indels passed quality filters. Annotation with Functator identified several high-impact mutations, including missense, nonsense, and frameshift variants. Some variants mapped to known pathogenic alleles in ClinVar, including genes involved in cancer susceptibility and metabolic disorders

C. Annotation:

Below output represents the genotype section which provides individual-level variant information for sample SRR794247, including genotype (GT), allele depth (AD), read depth (DP), genotype quality (GQ), and likelihood scores (PL), which reflect the accuracy and depth of the variant calls. The fixed section lists variant-level details such as chromosome position, reference and alternate alleles, quality scores, and dbSNP IDs. All variants shown passed quality filters ("PASS"), indicating high-confidence SNPs suitable for downstream analysis.

The fixed section summarizes Indel-level details including chromosome position, reference and alternate alleles, dbSNP IDs (e.g., rs370886585), and quality scores. All entries passed variant filtering criteria. The genotype section presents per-sample data for SRR794247, with fields such as genotype (GT), allele depth (AD), read depth (DP), genotype quality (GQ), and Phred-scaled likelihoods (PL), reflecting both confidence and read support for each variant. These annotated

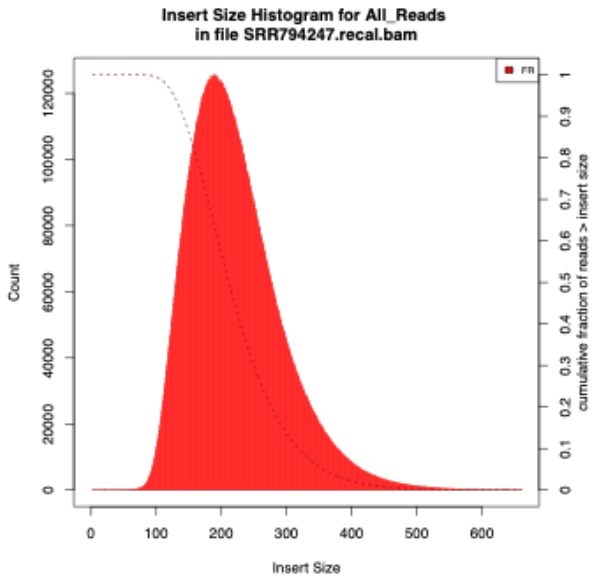

Fig. 9. Distribution of Insert size across the Reads

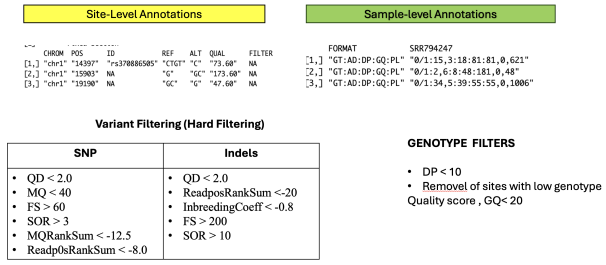

Fig. 10. Variant calling and filtering summar

variants are ready for downstream functional and clinical interpretation.

D. SNP and Indel Variant Consequences:

The consequences of SNP variants were analysed at both genomic and coding levels. Genomic distribution showed that the majority of SNPs were intron variants (37%), followed by non-coding transcript variants (17%), and downstream gene variants (13%), indicating a significant proportion of non-coding region involvement. Coding consequences revealed that synonymous variants accounted for 55%, while missense variants represented 44%, suggesting that while many SNPs do not alter amino acid sequences, a notable fraction may affect protein structure or function.

Following the annotation, the functional impact of indel variants was analyzed to assess their genomic and coding consequences. Genomically, the majority of indels were classified as intron variants (41%), followed by non-coding transcript variants (18%) and downstream gene variants (13%). On the coding side, frameshift variants accounted for the highest proportion (36%), closely followed by in-frame deletions (34%) and in-frame insertions (23%). These results suggest that a

```
[1] "***** Genotype section *****"
     FORMAT           SRR794247
[1,] "GT:AD:DP:GQ:PL" "0/1:18,7:25:99:120,0,659"
[2,] "GT:AD:DP:GQ:PL" "0/1:12,9:21:99:187,0,391"
[3,] "GT:AD:DP:GQ:PL" "0/1:34,14:48:99:318,0,885"
[4,] "GT:AD:DP:GQ:PL" "1/1:0,20:20:60:719,60,0"
[5,] "GT:AD:DP:GQ:PL" "0/1:18,12:30:99:410,0,590"
[6,] "GT:AD:DP:GQ:PL" "0/1:8,4:12:99:126,0,289"
```

```
[1] "***** Fixed section *****"
     CHROM  POS      ID           REF ALT QUAL     FILTER
[1,] "chr1" "182903" NA           "C" "G" "112.64" "PASS"
[2,] "chr1" "187019" NA           "G" "A" "179.64" "PASS"
[3,] "chr1" "187102" NA           "C" "G" "310.64" "PASS"
[4,] "chr1" "783175" "rs10751453" "T" "C" "705.06" "PASS"
[5,] "chr1" "817341" "rs3131972"  "A" "G" "402.64" "PASS"
[6,] "chr1" "818161" "rs2073813"  "G" "A" "118.64" "PASS"
```

Fig. 11. Sample annotation output of SNP variants

significant portion of the indels may affect transcript stability and protein function.

The SNP and Indel distributions displayed are most consistent with expected profiles for healthy human exomes. Most of the SNPs were found to occur within intronic or non-coding DNA, as expected, although a larger portion of the SNPs intersected the coding sequence via either a synonymous or missense changes. The number of Indel's is smaller compared to the SNPs but report both frameshift and in-frame changes to coding sequence, which are expected to have a more drastic effect upon gene function. The impact of the variant distribution is consistent with what is published in the literature and increases confidence in the variant calling method. The observed SNPs and Indels could also be compared to databases with allele frequencies, for example, the 1000 Genomes or gnomAD in future studies, to evaluate the rare variants of interest. Moreover, the visual inspection of the data using IGV ensured our accessions' locations were accurately aligned at loci deemed to be high impact variants and further confirms the confidence with our variant calls. This analysis provides evidence to support that our pipeline functions well for variant discovery and quality checking.

VI. Biological Implications of Identified Variants:

A. Clinically Relevant Variants

We also identified several Indels that had already been reviewed in clinical databases. For example:

- A deletion at `chr1:151,408,845` is labeled **Benign**, meaning it is not associated with any known disease.
- Another at `chr10:113,588,968` is listed as **Benign or Likely benign**.
- One interesting variant was linked to **White-Sutton syndrome**, a rare neurodevelopmental disorder.
- Another is associated with **Factor VII Marburg I**, a condition related to blood clotting.

```
[1] "***** Genotype section *****"
     FORMAT              SRR794247
[1,] "GT:AD:DP:GQ:PL" "0/1:15,3:18:81:81,0,621"
[2,] "GT:AD:DP:GQ:PL" "0/1:26,57:83:99:2294,0,906"
[3,] "GT:AD:DP:GQ:PL" "0/1:61,157:218:99:6356,0,2060"
[4,] "GT:AD:DP:GQ:PL" "0/1:34,10:44:99:212,0,1003"
[5,] "GT:AD:DP:GQ:PL" "0/1:15,3:18:58:58,0,500"
[6,] "GT:AD:DP:GQ:PL" "0/1:54,31:85:99:952,0,1964"
```

```
[1] "***** Fixed section *****"
     CHROM  POS      ID            REF                                                                ALT QUAL      FILTER
[1,] "chr1" "14397"  "rs370886505" "CTGT"                                                             "C" "73.60"   "PASS"
[2,] "chr1" "63735"  "rs201888535" "CCTA"                                                             "C" "2286.60" "PASS"
[3,] "chr1" "189392" NA            "ACC"                                                              "A" "6348.60" "PASS"
[4,] "chr1" "189713" NA            "GC"                                                               "G" "204.60"  "PASS"
[5,] "chr1" "809990" "rs146246821" "TA"                                                               "T" "50.60"   "PASS"
[6,] "chr1" "866577" NA            "CCTGCACTCACATCCCTGACGTCCTCCCGTGCCCTCACGTGGTCCTCCCT" "C" "944.60"  "PASS"
```

Fig. 12. Sample annotation output of Indel variants

These types of clinical labels are valuable because they help distinguish potentially harmful mutations from those that are common and harmless.

VII. SNPs and Disease Links

The SNPs we discovered were distributed throughout the genome—in exons, introns, regulatory regions, and intergenic regions. A few matched known disease-associated variants:

- A **G to A** change in the *FTO* gene has been linked to obesity.
- A **C to T** change in *TCF7L2* is associated with type 2 diabetes.

SNPs located in coding regions can have various effects:

- **Synonymous SNPs** do not alter the resulting protein sequence.
- **Missense SNPs** result in a single amino acid substitution.
- **Nonsense SNPs** introduce a premature stop codon, potentially truncating the protein.

VIII. Affected Pathways

We identified SNPs and Indels associated with several major disease pathways:

- **Metabolism**: Variants in *FTO* and *TCF7L2* linked to obesity and diabetes.
- **Brain and Mental Health**: Changes in *HTR2A* and *TMEM106B* associated with mood disorders and dementia.
- **Autoimmunity**: Variants near *lnc13*, a long non-coding RNA, implicated in Crohn's disease and rheumatoid arthritis.
- **Cancer**: While most cancer mutations are somatic, some SNPs we identified may contribute to inherited cancer risk, particularly involving *BRCA1/2* or *TP53*.

IX. Final Thoughts

Overall, these findings highlight how exome sequencing can reveal variants with diverse biological and clinical relevance. By integrating tools like GATK with databases such as ClinVar, COSMIC, and OMIM, we gained a clearer understanding of which variants are likely to be impactful—and which may currently have no known clinical effect.

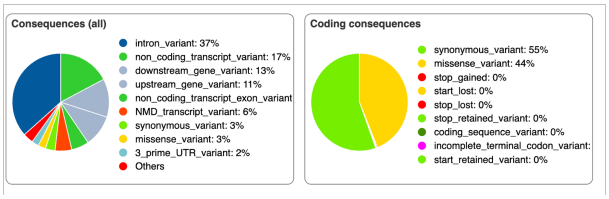

Fig. 13. SNP Variants- Genomic and Coding Distribution

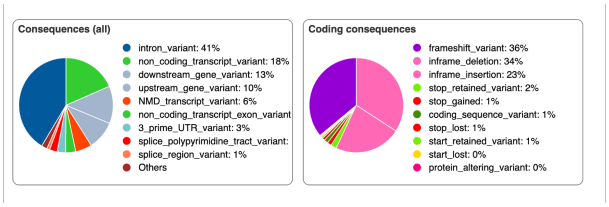

Fig. 14. Indels Variants-Genomic and Coding Distribution

In this project, we identified a variety of insertions and deletions (Indels) across multiple chromosomes. Most were single-base deletions, but there were also some multi-base deletions and a few insertions. These types of changes can have significant biological consequences—particularly if they disrupt the reading frame of a gene, such as by causing a frameshift mutation. Depending on their location, Indels can interfere with protein function or gene regulation.

Some of these Indels were linked to known cancer-related mutations. For example:

- **COSM150847** is a variant recorded in the COSMIC database that appears frequently in colon and stomach cancers.

Such annotations are valuable as they may indicate *driver mutations*—genetic changes that contribute to tumor progression or drug resistance.

X. Discussion and Future Directions

Through this project, we were able to identify a wide variety of SNPs and Indels, some of which were linked to well-known diseases and biological processes. It was interesting to see how certain variants, especially those in coding regions, were already associated with conditions such as diabetes, obesity, cancer, or rare genetic disorders. Some were listed in databases like *COSMIC* and *ClinVar*, which helped us understand their possible significance. Others had no known clinical impact but might be worth investigating further.

This highlights the power of whole exome sequencing (WES)—not just for detecting mutations, but for interpreting how they might affect genes, proteins, and disease risk. It also demonstrates the importance of annotation tools in filtering through thousands of changes to find those that are clinically or biologically relevant.

Future Directions

There is ample scope to expand on the work done in this project. Future improvements could include:

- Including non-coding regions that may play roles in gene regulation.
- Incorporating multiple samples or family-based sequencing to better interpret rare variants.
- Exploring AI tools to predict the functional impact of poorly understood variants.
- Integrating genotype data with phenotype or clinical outcomes to strengthen interpretation.

This project deepened our understanding of how WES can be applied in both research and clinical settings, and emphasized how each step in the pipeline contributes to the reliability of final results.

XI. Conclusion

The project successfully developed and implemented a comprehensive, reproducible WES pipeline to identify and annotate SNPs and short Indels using high-coverage data from HG03012 of the 1000 Genomes Project. The pipeline integrated widely used bioinformatics tools including *FastQC*, *BWA-MEM*, *SAMtools*, *Picard*, and *GATK*, all encapsulated within a Dockerized environment to ensure reproducibility and consistency.

Quality control metrics confirmed the high integrity of the raw sequencing data, and alignment and variant calling against the *GRCh38* reference genome produced accurate variant calls. Functional annotation was performed using *GATK's Funcotator*, leveraging databases such as *ClinVar*, *dbSNP*, and *gnomAD*.

Functional consequence analysis revealed that most SNPs were intronic or synonymous/missense coding variants. Indels, in contrast, were often frameshift or in-frame, with higher potential to disrupt protein structure and function. These findings illustrate the biological importance of analyzing both neutral and deleterious mutations within coding regions.

In conclusion, with further development—such as integrating population-specific datasets, family-based analyses, and AI-driven prioritization—this pipeline could have significant applications in clinical diagnostics, population genomics, and personalized medicine. Its modular and scalable architecture ensures its ongoing value in both academic and clinical contexts.

XII. Team Contributions

- **Leela Krishna Sai Pannem**
  - Analysis of workflow: 20%
  - Analysis of FastQC report: 30%
  - Environment setup: 30%
  - Finding the dataset and reference dataset: 60%
  - Documentation tasks related to the project: 80%
- **Sumanth Kumar Lingabathini**
  - Analysis of workflow: 10%
  - Downloading related tools for the project: 20%
  - Research on biological implications and future work: 100%
  - Documentation tasks related to the project: 20%
  - Presentation 1: 50%
- **Nirvan Kotha**
  - Data pre-processing: 100% (Analyzed raw sequencing data from Read 1 and Read 2 FASTQ files and generated FastQC HTML reports)
  - Mapping to reference genome: 100%
  - Marking duplicates using GATK and analysis: 100%
  - Presentation 2: 50%
  - Final presentation: 30%
- **Abhishek Arugonda**
  - Analysis of workflow: 20%
  - Downloading related tools for the project: 30%
  - Collecting alignment and insert size metrics: 100%
  - Generating analysis-ready BAM file: 100%
  - Variant calling using GATK HaplotypeCaller and GenotypeGVCFs: 100%
  - Extracting SNPs and Indels: 100%
  - Presentation 2: 50%
  - Final presentation: 30%
- **Aravindh Subramanian**
  - Finding the dataset: 30%
  - Analysis of workflow: 60%
  - Analysis of FastQC report: 70%
  - Analyzing the MultiQC report: 100%
  - Environment setup: 70%
  - Downloading related tools for the project: 50%
  - Variant filtering: 100%
  - Selecting variants that pass filters: 100%
  - Annotating variants using GATK Funcotator: 100%
  - Presentation 1: 50%
  - Final presentation: 40%

XIII. References

References

[1] A. McKenna *et al.*, "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010.

[2] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," *arXiv preprint arXiv:1303.3997*, 2013. [Online]. Available: https://arxiv.org/abs/1303.3997

[3] Broad Institute, "GATK Best Practices," [Online]. Available: https://gatk.broadinstitute.org/, Accessed: May 15, 2025.

[4] S. Andrews, "FastQC: A quality control tool for high throughput sequence data," 2010. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, Accessed: May 15, 2025.

[5] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015.

[6] Broad Institute, "Picard Toolkit," [Online]. Available: https://broadinstitute.github.io/picard/, Accessed: May 15, 2025.

[7] A. R. Quinlan and I. M. Hall, "BEDTools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010.

[8] Broad Institute, "Funcotator: Functional annotation tool in GATK," [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036898312-Funcotator, Accessed: May 15, 2025.