ISSN 2007-9737

A Resource-Aware Data Processing Framework for Scalable Whole Exome Variant Analysis in an Under-Represented Asian Cohort

C Lalrinmawii*, Aparajita Dutta, Ramanujam E

National Institute of Technology Silchar, Silchar, India

{clalrinmawii21_rs, aparajita.dutta, ramanujam}@cse.nits.ac.in

Abstract. Type 2 diabetes (T2D) is a rapidly growing metabolic and genetic disorder, with Asia representing a major disease hotspot. However, Asian populations remain under-represented in genomic studies, limiting the identification of population-specific risk variants. Moreover, large-scale sequencing generates large volumes of data, posing challenges for computational efficiency, storage, and variant interpretation. This study presents a resource-efficient computational framework for whole exome sequencing (WES) analysis with integrated visual analytics. A unified pipeline was developed by integrating automated quality filtering AFastQF, Bowtie2 for alignment, and variant detection using DeepVariant. The framework is designed to minimize intermediate data generation while maintaining computational performance. The framework was evaluated on WES datasets from T2D patients, non-T2D individuals, and healthy controls from an Asian cohort. The proposed framework achieved about 24% reduction in intermediate storage requirements without increasing runtime. The analysis identified both known T2D-associated variants and those that are not in the database. Three candidate variants in WFS1, HHEX, and FTO genes were prioritized based on in silico pathogenicity predictions and absence from major population databases. The study demonstrates that efficient and scalable genomic analysis can be achieved in resource-constrained environments. The proposed framework provides a practical solution for large-scale variant discovery and interpretation in under-represented populations.

Keywords. Asia, framework, resource-aware, type 2 diabetes, variants, whole exome sequencing.

1 Introduction

Diabetes is one of the fastest growing global health concerns of the 21st century and currently ranks among the leading causes of mortality worldwide. It is estimated to become the second leading cause of disease by 2050 [12]. Type 2 diabetes (T2D), caused by genetic and environmental factors, accounts for more than 90% of all diabetes cases and represents a major public health challenge [28]. Asian populations contribute a substantial proportion of global diabetes cases, yet remain significantly under-represented in large-scale genomic studies [13].

This imbalance limits the discovery of population specific genetic determinants and reduces the effectiveness of precision medicine strategies in these regions [35]. Genome-wide association studies (GWAS) have identified numerous loci associated with T2D risk [28]. Recent advances in artificial intelligence (AI) and bioinformatics have improved the analysis and interpretation of high-dimensional genomic data, particularly for disease prediction [9, 33] and personalized healthcare applications [16].

However, large-scale sequencing analysis presents high computational challenges, particularly in low and middle income countries where resources are limited [42]. High-throughput genomic datasets require efficient quality control, alignment, variant detection, and annotation, often generating large intermediate files that increase storage and processing demands [4]. These limitations can be particularly restrictive in resource limited research environments.

A computational pipeline for efficient exome variant analysis and visualization, integrating automated genomic analysis tools, is proposed to address this challenge. The framework performs automated quality filtering using

AFastQF [20], followed by reference-based alignment using Bowtie2 [24] and variant detection through DeepVariant [2]. The pipeline reduces intermediate storage requirements without significantly increasing computational time by integrating these tools within an optimized workflow. Visual analytics components are incorporated to support exploration of variant distributions and annotation derived features, which improved the interpretability of large-scale genomic datasets. The pipeline was applied to whole exome sequencing (WES) data from diabetic and control individuals across Asian cohort, obtained from the National Center for Biotechnology Information (NCBI) [41] database. The analysis identified both previously reported and potentially novel T2D-associated coding variants, including rare deleterious missense variants in established T2D related genes. The main contributions of this study are summarized as follows:

- Design and implementation of a resource-aware data processing framework for whole exome variant analysis.
- Development of an optimized data handling strategy that minimizes intermediate storage requirements by reducing redundant data retention, achieving improved storage efficiency while preserving computational stability.
- Integration of automated quality control, read alignment, and variant detection within a unified framework, ensuring efficient data flow and reducing manual intervention in large-scale genomic analysis.
- Comprehensive evaluation of the proposed framework on WES data from an under-represented Asian cohort, demonstrating its effectiveness in identifying variants associated with T2D.

The primary contribution of this work are the systematic integration of established tools, optimization for low resource environments, and incorporating visual analytics to enhance the interpretability of large-scale genomic data. The remainder of this paper is organized as follows. Section 2 describes the materials and methods of the proposed framework, while Section 3 presents the experimental results and variant discovery outcomes. The study concludes in Section 4.

2 Materials and Methods

This section describes the materials and methods proposed for efficient processing of WES data. A reference-based variant calling approach was employed for genomic variants identification in WES data [21]. The study focused on improving the accessibility of genomic analysis techniques for low-resource environments. The overall workflow of the proposed pipeline is illustrated in Figure 1. All benchmarking experiments were conducted on a simple system with an Intel Core i7-12650H CPU, 16 GB of RAM, and Ubuntu 24.04 LTS, using 8 processing threads.

2.1 Data Acquisition and Study Cohort

WES datasets from 95 individuals of Asian origin were obtained from publicly available repositories hosted by NCBI [41]. The cohort comprised 54 individuals diagnosed with Type 2 Diabetes (T2D), 15 individuals with other forms of diabetes such as Type 1 Diabetes (T1D) [31] and Maturity-Onset Diabetes of the Young (MODY) [10], hereafter referred to as non-T2D, and 26 healthy controls.

These samples represent individuals from diverse Asian populations. Datasets were selected based on an extensive literature review to ensure representation from multiple Asian regions and the availability of high-quality publicly accessible sequencing data [36, 43, 6, 22].

2.2 Quality Filtering

Raw sequencing reads in FASTQ format were subjected to quality filtering using AFastQF [20].

The tool automatically removes low-quality reads based on average Phred scores, $Q_{avg}(r)$ less than 20 [37] and sequence length, L_s less than 36 base pairs [47]. It also trims sequencing artifacts, duplicates, and contaminants introduced during sequencing.

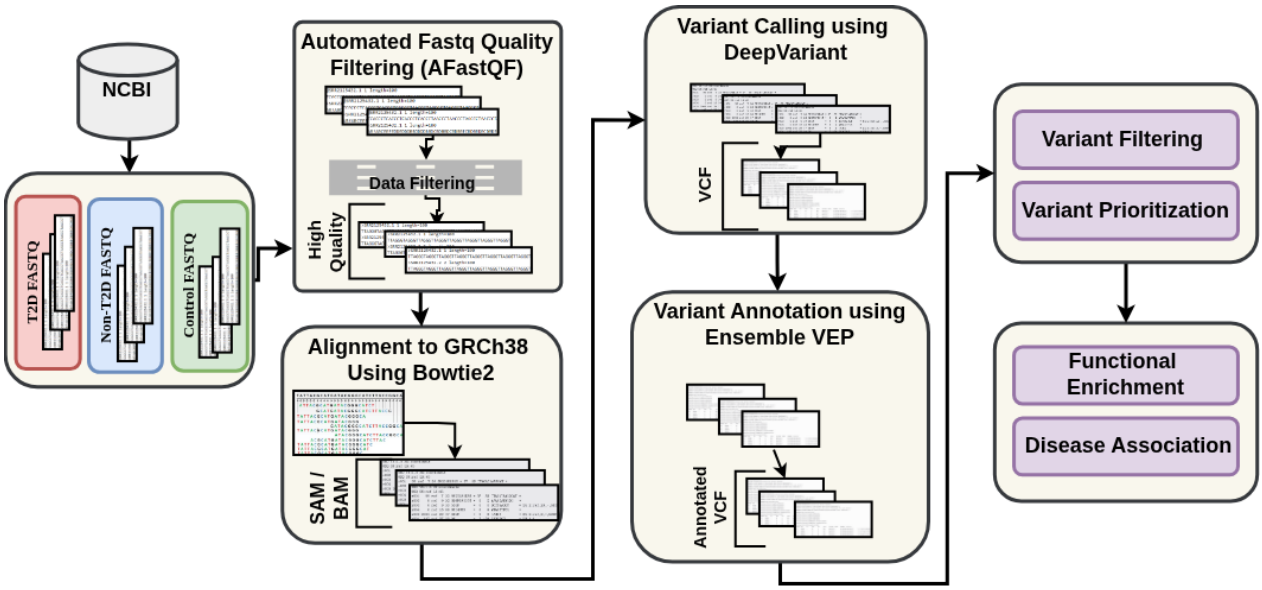

Fig. 1. A schematic workflow of the proposed integrated framework

The tool introduces duplicate detection and removal during raw data preprocessing. AFastQF was benchmarked against commonly used quality control tools, including Fastp [7], Trimmomatic [5], and Cutadapt [30], focusing on processing speed and reduction in intermediate file size.

Complete algorithm for the AFastQF tool is shown in Algorithm 1.

2.3 Read Alignment

After quality filtering, the processed reads were aligned to the GRCh38 human reference genome using Bowtie2 [24] following Algorithm 2. Bowtie2 was selected due to its high alignment speed and accuracy for short-read sequencing data [45].

The alignment performance of Bowtie2 was benchmarked against other widely used aligners, including BWA-MEM [25] and Minimap2 [26], to evaluate differences in computational time and storage requirements.

The resulting Sequence Alignment Map (SAM) files were converted to Binary Alignment Map (BAM) and sorted format using Samtools [27] for efficient storage and downstream processing.

2.4 Variant Detection

Variant detection was performed on the aligned BAM files using DeepVariant [39] following Algorithm 3. DeepVariant employs a pre-trained convolutional neural network to accurately identify single nucleotide variants and small insertions and deletions from sequencing data. The tool has demonstrated high precision in short-read WES datasets [2]. DeepVariant was benchmarked against several widely used variant callers, including GATK [44], FreeBayes [14], and Clair3 [48]. Variant calling accuracy was assessed using the Real Time Genomics evaluation toolkit (rtg-tools) [8]:

2.5 Variant Annotation

The variants identified by DeepVariant were subsequently annotated using the Ensembl Variant Effect Predictor (VEP) (release 112) [32]. VEP was configured to incorporate multiple functional prediction scores and population allele frequency datasets. The annotation process provided detailed information on variant consequences, including predicted effects on protein-coding regions, pathogenicity predictions,

Algorithm 1 Quality Filtering of FASTQ Reads using AFastQF

Require: Raw WES FASTQ data [No of samples S = 95, Class = T2D, non-T2D, and Control]
Ensure: High quality filtered reads with maximum size reduction
Initialize empty sets *CleanSamples* and *FilteredReads*
for each sample s in FASTQ dataset **do**
  Initialize empty set *TempReads*
  for each read r in sample s **do**
    if r is duplicate **then**
      Discard r
    else
      TempReads $\leftarrow r$
    end if
  end for
  CleanSamples $\leftarrow$ TempReads
end for
for each sequence s in *CleanSamples* **do**
  Load reads $\{r_1, r_2, ..., r_n\}$ from sequence s
  Compute average Phred score $Q_{avg}(r)$ of read r in sequence s
  if sequence Length $(L_s) < 36$ **then**
    Discard sample s
  else
    if $Q_{avg}(r) \geq 20$ **then**
      Discard s
    else
      FilteredReads $\leftarrow s$
    end if
  end if
end for
Remove residual contaminants from *FilteredReads*
return *FilteredReads*

Algorithm 2 Read Alignment of filtered reads using Bowtie2

Require: High-quality filtered reads *FilteredReads*, GRCh38 reference genome
Ensure: Sorted and indexed BAM files
Initialize empty set *AlignedBAM*
for each sample s in *FilteredReads* **do**
  Load reads from sample s
  $SAM_s \leftarrow$ Bowtie2(s, GRCh38) ▷ Align reads to reference
  $BAM_s \leftarrow$ Samtools view(SAM_s) ▷ Convert SAM to BAM
  $BAM_s^{sorted} \leftarrow$ Samtools sort(BAM_s)
  Samtools index(BAM_s^{sorted})
  AlignedBAM $\leftarrow BAM_s^{sorted}$
end for
return *AlignedBAM*

and known disease associations. Functional prediction scores include Sorting Intolerant from Tolerant ($SIFT$) [34], Polymorphism Phenotyping v2 ($PolyPhen$-2) [3], Combined Annotation Dependent Depletion ($CADD$) [40], and Rare Exome Variant Ensemble Learner ($REVEL$) [17]. Population level allele frequencies from the 1000 Genomes Project (AF_{1KG}) [11] and gnomAD (AF_{gnomAD}) [19] were annotated along with clinical and auxiliary annotations, including Clinical Significance ($CLIN_SIG$) and Phenotype association ($PHENO$) from ClinVar [23]. Existence on Database of Single Nucleotide Polymorphisms ($dbSNP$) [38] variant database maintained by NCBI. These annotations enabled systematic filtering and prioritization of variants relevant to T2D analysis and is expressed in Equation 1:

$$ANN_s = \mathcal{F}_{\text{VEP}}\Big(VCF_s,\ \mathcal{S},\ \mathcal{A},\ \mathcal{C}\Big), \qquad (1)$$

where ANN_s denotes the annotated variant set. $VCF_s = \{v_1, v_2, \ldots, v_n\}$ represents the set of variants for sample s, and $\mathcal{F}_{\text{VEP}}$ denotes the variant annotation function implemented using VEP. The set of functional prediction scores $\mathcal{S}$ is defined as $\mathcal{S} = \{SIFT, PolyPhen\text{-}2, CADD, REVEL\}$, the population allele frequency set $\mathcal{A}$ is defined as $\mathcal{A} = \{AF_{1KG}, AF_{gnomAD}\}$, and the clinical annotation set $\mathcal{C}$ is defined as $\mathcal{C} = \{CLIN_SIG, PHENO, dbSNP\}$.

2.6 Variant Filtering and Prioritization

The identified variants were subjected to filtering and prioritization procedure within the proposed computational framework to identify variants with potential relevance to T2D. Variants were first filtered using population allele frequency information from databases integrated through VEP annotations. Rare variants with low global allele frequencies were retained for further analysis, as such variants are more likely to contribute to disease susceptibility. Functional impact predictions were then evaluated using multiple computational scoring systems.

Variants predicted to have deleterious or damaging effects on protein structure or function were prioritized, and filtering was further restricted to protein-coding missense variants that may alter gene function. Variants identified in T2D individuals were subsequently compared with those present in non-T2D diabetic and healthy control cohorts to identify variants uniquely enriched in the T2D group, thereby reducing background genetic variation and improving candidate prioritization.

The filtered T2D specific variants were cross referenced with established variant databases,

Algorithm 3 Variant Calling of aligned data using DeepVariant

Require: Sorted and indexed BAM files *AlignedBAM*, GRCh38 reference genome
Ensure: High-confidence variant calls in VCF format
1: Initialize empty set *VariantCalls*
2: **for** each sample s in *AlignedBAM* **do**
3: Load sorted BAM file BAM_s^{sorted}
4: $VCF_s \leftarrow$ DeepVariant(BAM_s^{sorted}, GRCh38) ▷ Call SNPs and indels
5: *VariantCalls* $\leftarrow VCF_s$
6: **end for**
7: **return** *VariantCalls*

including dbSNP, the 1000 Genomes Project, gnomAD, and ClinVar, to identify potentially novel variants. Variants lacking entries in these databases and without reported population allele frequency information were considered candidate novel variants. Additional functional filtering criteria were applied to retain biologically relevant mutations, including $SIFT$ predictions classified as *deleterious* ($SIFT \leq 0.05$) [34], $PolyPhen$-2 predictions classified as *probably damaging* ($PolyPhen$-2 ≥ 0.85) [3], $CADD$ scores indicating deleteriousness ($CADD \geq 20$)[40], and $REVEL$ scores suggesting possible deleterious effects (REVEL ≥ 0.5) [17].

Only variants with predicted *MODERATE* or *HIGH* functional impact and extremely low or absent allele frequency (<0.001) in population databases were retained. The resulting candidate variants were categorized into those located within previously reported T2D-associated genes identified through GWAS [29] and those occurring in genes not previously linked to T2D. Let V denote the set of annotated variants obtained from the WES pipeline. Filtering is applied to obtain the prioritized variant subset V^* as seen in Equation 2 and Equation 3:

$$V^* = \left\{ v \in V \;\middle|\; \begin{array}{l} AF(v) < 0.001 \\ \wedge\ IMPACT(v) \in \{\text{MODERATE, HIGH}\} \\ \wedge\ \text{SIFT}(v) \leq 0.05 \\ \wedge\ \text{PolyPhen}(v) \geq 0.85 \\ \wedge\ \text{CADD}(v) \geq 20 \\ \wedge\ \text{REVEL}(v) \geq 0.5 \end{array} \right\}, \quad (2)$$

$$V_{\text{novel}} = \left\{ v \in V^* \;\middle|\; \begin{array}{l} v \notin D_{\text{dbSNP}} \cup D_{\text{1000G}} \cup D_{\text{gnomAD}} \\ \qquad \cup D_{\text{ClinVar}} \end{array} \right\}, \quad (3)$$

where $AF(v)$ represents the population allele frequency of variant v, and the prediction scores correspond to functional impact estimators obtained during variant annotation. The symbols D_{dbSNP}, D_{1000G}, D_{gnomAD}, and $D_{ClinVar}$ represent the sets of variants reported in the dbSNP, gnomAD, and ClinVar databases, respectively. The subset V^* represents prioritized variants satisfying the defined biological and computational constraints, while V_{novel} denotes variants not present in existing population and clinical variant databases.

The resulting prioritized variant set enables efficient downstream visualization and exploration of disease associated genomic variation patterns within the proposed analytical framework.

2.7 Functional Enrichment and Disease Association

Genes harbouring prioritised variants were analysed using disease, pathway, and gene ontology enrichment resources to determine biological relevance. Statistical significance was assessed using multiple testing correction methods. Multi-database resources were utilized to identify potentially relevant variants, and different functional enrichment analysis were performed on the identified variants to analyze their impact. Genes carrying variants with high combined functional impact scores, were analysed across multiple annotation resources, including Gene Ontology (GO) [1], Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways 2021 Human [18] and Reactome Pathways 2024 [15] via the Enrichr platform [46].

For each analysis, enrichment scores and adjusted *p*-values were calculated, with statistical significance determined using the Benjamini–Hochberg correction to account for multiple testing. Only pathways with adjusted $p < 0.05$ were considered significant. The combination of these procedures and analyses was used to identify candidate variants in the under-represented Asian cohort. The identification of proven T2D affecting variants demonstrates the reliability of the pipeline. Employing multiple predictive features and functional enrichment

of candidate variants prioritized by in silico pathogenicity scoring further increases confidence in the identified variants.

The proposed framework incorporates a visual analytics module to facilitate intuitive exploration of high-dimensional genomic data.

The complete end-to-end workflow of the proposed framework is presented in Algorithm 4, summarizing the sequential processing steps from raw data preprocessing to variant prioritization and analysis.

3 Results

The results follow the same workflow stages described in the methodology, including data acquisition, quality filtering, read alignment, variant detection, annotation, variant prioritization, and functional interpretation of the identified variants.

3.1 Data Acquisition and Study Cohort

WES datasets were obtained from publicly available repositories, comprising 95 individuals including 54 T2D patients, 15 non-T2D individuals, and 26 healthy controls from Western, Eastern, and Southern Asian cohorts. The T2D group, comprising 45 males and 19 females, had a mean age of 51 ± 11 years. The non-T2D cohort consisted of 15 males from Eastern Asia with a mean age of 30 ± 20 years, while the control group included 24 males and 2 females from Western Asia with a mean age of 45 ± 7 years.

Overall, the dataset comprised 84 males and 21 females with a mean age of 42 ± 7 years. Raw FASTQ files were retrieved from the NCBI Sequence Read Archive (SRA), BioProjects PRJNA288292, PRJNA656961, PRJNA579845 and PRJEB5736 using the SRA Toolkit and processed through the proposed pipeline. All samples were generated on Illumina platforms. Each exome dataset contained approximately 4-6 billion base pairs with 16-24 GB of data per sample, necessitating efficient quality control and storage optimization for reliable downstream analysis.

Algorithm 4 WES Variant Analysis and Prioritization Framework

Require: WES FASTQ datasets S, Reference genome GRCh38
Ensure: Prioritized variant set V^* and novel variant set V_{novel}
Initialize sets:
FilteredReads $\leftarrow \emptyset$, *AlignedBAM* $\leftarrow \emptyset$, *VariantCalls* $\leftarrow \emptyset$, *AnnotatedVariants* $\leftarrow \emptyset$, $V^* \leftarrow \emptyset$, $V_{novel} \leftarrow \emptyset$
for each sample $s \in S$ **do** ▷ Quality Filtering
  Remove duplicate reads and sequencing artifacts from s
  for each read r in s **do**
    if $Q_{avg}(r) \geq 20$ and length$(r) \geq 36$ **then**
      FilteredReads $\leftarrow r$
    end if
  end for
  $SAM_s \leftarrow$ Bowtie2(*FilteredReads*, GRCh38) ▷ Read Alignment
  $BAM_s \leftarrow$ Samtools view(SAM_s)
  $BAM_s^{sorted} \leftarrow$ Samtools sort(BAM_s)
  Samtools index(BAM_s^{sorted})
  AlignedBAM $\leftarrow BAM_s^{sorted}$
  $VCF_s \leftarrow$ DeepVariant$(BAM_s^{sorted}$, GRCh38) ▷ Variant Calling
  VariantCalls $\leftarrow VCF_s$
  $ANN_s \leftarrow \mathcal{F}_{VEP}(VCF_s)$ ▷ Variant Annotation
  AnnotatedVariants $\leftarrow ANN_s$
  for each variant $v \in ANN_s$ **do** ▷ Filtering and Prioritization
    if $AF(v) < 0.001$ **and** IMPACT$(v) \in$ {MODERATE, HIGH} **then**
      if $SIFT(v) \leq 0.05$ **and** PolyPhen$(v) \geq 0.85$ **then**
        if $CADD(v) \geq 20$ **and** REVEL$(v) \geq 0.5$ **then**
          $V^* \leftarrow v$
        end if
      end if
    end if
  end for
end for
for each $v \in V^*$ **do** ▷ Identification of Novel Variants
  if $v \notin$ (dbSNP $\cup$ 1000G $\cup$ gnomAD $\cup$ ClinVar) **then**
    $V_{novel} \leftarrow v$
  end if
end for
for each gene g containing variants in V^* **do** ▷ Functional Enrichment
  Perform enrichment analysis (GO, KEGG, Reactome, OMIM)
  Compute adjusted p-values using Benjamini–Hochberg correction
end for
Retain pathways with adjusted $p < 0.05$
return V^*, V_{novel}, enriched pathways

3.2 Quality Filtering

Raw sequencing reads were subjected to quality filtering using AFastQF. Benchmarking experiments were conducted using FASTQ datasets of varying sizes with AFastQF, Fastp (v0.23.4), Trimmomatic (v0.40), and Cutadapt (v4.4). The evaluation metrics included runtime, memory usage, and dataset size reduction relative to the original FASTQ files. AFastQF preserves high quality reads while achieving at least a 50% reduction in dataset size, as shown in Table 1 and Table 2.

Table 1. Processing speed comparison of AFastQF, Fastp (v0.23.4), Trimmomatic (v0.40), and Cutadapt (v4.4) using FASTQ datasets of varying sizes. Runtime is measured in seconds. Smaller datasets (0.65–12.60 GB) were randomly selected from T2D samples, while larger datasets (22.00–87.80 GB) were generated by replicating these datasets. The output is written in seconds

Size (GB)	AFastQF	Fastp	Trimmomatic + Fastp	Cutadapt + Fastp
0.65	3	3	9.88	6.81
7.50	39.49	75	128.06	115.24
12.60	180.66	103.08	172.90	233.17
22.00	392.03	232.9	525.29	750.03
43.90	771.99	411.23	939.21	1228.77
87.80	1618.74	778.12	1690.32	2302.85

Table 2. Size reduction analysis of cleaned data against raw FASTQ data using AFastQF, Fastp 0.23.4, Trimmomatic 0.40 and Cutadapt 4.4 benchmarked on random T2D patient's FASTQ human genome data

Original size (GB)	Tools	Output Size (GB)	Size Reduction (%)
12.6	AFastQF	**4.5**	**64.29**
	Fastp	10.6	15.87
	Trimmomatic	12.2	3.17
	Cutadapt	10.8	14.29

The benchmarking analysis included adapter trimming, duplicate removal, and low average quality filtering using identical parameter settings across tools to ensure fair comparability. Among the evaluated tools, only AFastQF performs duplicate trimming directly. Fastp supports duplicate detection but not trimming, whereas Trimmomatic and Cutadapt do not support duplicate detection. Fastp's duplicate detection module was incorporated into the Trimmomatic and Cutadapt workflows during benchmarking for fair comparison. Quality filtering was performed based on average Phred quality scores. As shown in Table 1, AFastQF achieved the fastest processing times on smaller datasets and remained competitive on larger datasets despite performing the additional duplicate removal step. AFastQF demonstrated significantly greater dataset size reduction compared with other preprocessing tools as seen in Table 2, highlighting its efficiency for large-scale WES data preprocessing.

3.3 Read Alignment

Filtered reads were aligned to the GRCh38 reference genome using Bowtie2, producing high quality SAM alignment files. Bowtie2 was benchmarked against BWA-MEM and Minimap2 using identical AFastQF processed datasets. The aligners were evaluated based on alignment rate, percentage of properly paired reads, mean sequencing depth, percentage of bases covered at $\geq 20\times$, runtime, and peak memory usage. These metrics were used to assess alignment accuracy, computational efficiency, and resource utilization. The benchmarking results are summarized in Table 3.

Bowtie2 achieved the highest alignment rate of 98.4% and coverage metrics while maintaining the lowest memory footprint of 3.29 GB. In comparison, BWA-MEM achieved a 98.1% alignment rate with 6.41 GB RAM usage, while Minimap2 achieved a slightly lower alignment rate of 97.8% and coverage depth. Although Minimap2 completed alignment in less than half the runtime of Bowtie2, 173.25 s and 390.96 s respectively, its substantially higher memory consumption of 11.02 GB and reduced coverage depth of 93.9% over 95.2% of Bowtie2 limit its suitability for large-scale exome studies. Overall, Bowtie2 provided the optimal balance between alignment accuracy, computational efficiency, and resource usage. Based on these results, Bowtie2 was selected as the alignment tool for the proposed variant analysis pipeline and provide output in SAM file format.

The SAM file was converted to a BAM file, a compressed SAM format, which aids storage and file manipulation. BAM files store data in binary format, making them more storage efficient and compatible with most variant calling tools.

3.4 Variant Detection

Variant calling was performed on the aligned reads to identify variants in a VCF file containing positional and quality information. Four widely used variant callers, including DeepVariant (v1.9.0), Clair3, GATK (v4.4.0.0), and FreeBayes (v1.3.9), were benchmarked using a common Bowtie2-aligned dataset to evaluate their computational performance and accuracy.

Table 3. Comparison of alignment performance for Bowtie2 (v2.5.4), BWA-MEM (v0.7.19), and Minimap2 (v2.28) using WES data aligned to GRCh38. Bowtie2 achieved the highest alignment and coverage metrics with the lowest memory usage, while Minimap2 showed the fastest runtime. AFastQF-processed reads were used as the common input

Aligner	Align %	Properly Paired %	Mean Depth	%≥20× Coverage	Runtime (sec)	Peak RAM (GB)
Bowtie2	98.4	97.9	93×	95.2	390.96	3.29
BWA-MEM	98.1	97.6	92×	94.7	608.45	6.41
Minimap2	97.8	97.0	91×	93.9	173.25	11.02

Table 4. Runtime, peak memory usage, and variant calling performance comparison of DeepVariant (v1.9.0), Clair3, GATK (v4.4.0.0), and FreeBayes (v1.3.9) on WES data using Bowtie2-aligned reads. Performance metrics (precision, sensitivity, and F1-score) were evaluated using rtg-tools (v3.12.1) against the GIAB NA12878/HG001 truth dataset

Tool	Runtime (sec)	Peak RAM (GB)	Precision	Sensitivity	F1-Score
DeepVariant	23,133	7.9	**99.97%**	**99.95%**	**99.96%**
Clair3	19,153	13.5	99.74%	98.22%	98.97%
FreeBayes	32,267	9.8	97.80%	94.32%	96.03%
GATK	53,867	19.8	96.12%	92.27%	94.16%

The benchmarking focused on runtime and peak memory consumption to assess time and space complexities. Clair3 achieved the fastest runtime with 19,153 s, followed by DeepVariant with 23,133 s, FreeBayes with 32,267 s, and GATK with 53,867 s. However, Clair3 required significantly higher memory of 13.5 GB, whereas DeepVariant maintained a lower 7.9 GB memory footprint while providing competitive runtime performance. FreeBayes and GATK required longer runtimes and higher memory consumption. These results indicate that DeepVariant provides the most balanced performance in terms of computational efficiency.

RealTimeGenomics Tools (rtg-tools v3.12.1) was used to compare the predicted variants against the Genome in a Bottle (GIAB) NA12878/HG001 truth dataset and to evaluate the calling accuracy. Precision, sensitivity, and F1-score were calculated for each tool, with DeepVariant achieving the highest accuracy with a precision of 99.97%, sensitivity of 99.95%, and an F1-score of 99.96%. Clair3 showed comparable performance with an F1-score of 98.97%, while FreeBayes and GATK achieved F1-scores of 96.03% and 94.16%, respectively. Given its superior accuracy and favorable balance between runtime and memory usage as seen in Table 4, DeepVariant was selected as the variant caller for the proposed pipeline.

Consequently, the combination of AFastQF, Bowtie2, and DeepVariant provided the optimal performance for the whole exome variant detection as represented in Figure 2.

3.5 Variant Annotation

Variant calling using DeepVariant produced approximately 40,000-46,000 variants per sample. These variants were annotated using the Ensembl Variant Effect Predictor (VEP) to determine genomic location, functional consequences, population allele frequencies, and clinical relevance.

Annotation incorporated multiple reference databases, including gnomAD, the 1000 Genomes Project, ClinVar, and PubMed. Functional consequences were assigned using Ensembl gene and transcript models with Sequence Ontology terms. Additional functional predictions were obtained from SIFT, PolyPhen-2, CADD, and REVEL scores to evaluate the potential pathogenicity of variants. Cross-referencing with dbSNP and ClinVar confirmed that several detected variants had been previously reported to be associated with T2D and related metabolic disorders.

The variant calling process is computationally intensive and generates substantial volumes of intermediate data at each stage of processing.

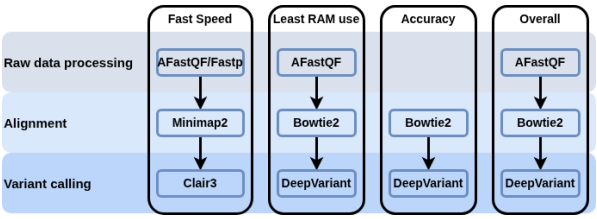

Fig. 2. Comparison of the different tool combinations based on speed, RAM requirement, accuracy, and overall

Table 5. Comparison of data output size between a conventional GATK-based workflow and the proposed pipeline

Stage	GATK Workflow	Proposed Pipeline
Raw data (FASTQ)	$\approx$ 2 TB	$\approx$ 2 TB
Cleaned reads (FASTQ)	$\approx$ 2 TB	$\approx$ 1 TB
Alignment output (SAM)	Temporary	Temporary
Processed alignment (BAM)	$\approx$ 900 GB	$\approx$ 700 GB
Variant calling (VCF)	$\approx$ 4.5 GB	$\approx$ 4.5 GB
Annotation (VCF)	$\approx$ 11 GB	$\approx$ 11 GB
Total Storage	$\approx$ **4.91 TB**	$\approx$ **3.72 TB**

A comparison of data volume produced by a conventional GATK-based workflow and the proposed pipeline is presented in Table 5. While the alignment and preprocessing stages generate data at the terabyte scale, the overall data volume decreases significantly after variant calling and annotation, where outputs are represented in compact VCF formats. The proposed pipeline reduces total storage requirements from approximately 4.91 TB to 3.72 TB, a reduction of about 24%. This reduction is achieved through improved handling of intermediate data, including partial retention of cleaned FASTQ files and more efficient management of alignment outputs, while maintaining compatibility with standard genomic analysis workflows.

3.6 Variant Filtering and Prioritization

A filtering and prioritization strategy was applied across the cohort groups to identify variants potentially associated with T2D. Across individuals, 132,513 variants were observed in the T2D group, 121,146 in the non-T2D group, and 56,993 in controls. 66,047 missense SNVs were identified as uniquely present in the T2D group and completely absent from both non-T2D and control cohorts. Similarly, 53,347 variants were unique to the non-T2D group, while 30,412 variants were specific to the control group. V^* variants were further prioritized based on predicted functional impact, allele frequency, and pathogenicity prediction scores to identify rare variants with potential biological significance. The prioritized variants were further analyzed to identify V_{novel} potentially associated with T2D. Among the unique T2D variants, 16,729 were not reported in existing literature or major population databases, including dbSNP, gnomAD, and the 1000 Genomes Project, supporting their novelty. Functional impact predictions indicated that a substantial proportion of these variants may affect protein function. SIFT classified 11,326 variants as deleterious, while PolyPhen-2 predicted 10,633 variants as probably damaging.

3.7 Functional Enrichment and Disease Association

The identified variants were subsequently mapped to known T2D-associated genes curated from association studies. This analysis revealed that 2,016 variants were located within 19 established T2D susceptibility genes, while 14,713 variants were mapped to genes not previously linked to T2D. The presence of variants in well-established susceptibility genes, including *TCF7L2*, *PPARG*, *KCNJ11*, *SLC30A8*, *HHEX*, and *CDKN2A/B*, supports the robustness of the filtering strategy and highlights population specific genetic contributions in the Asian cohort. The remaining variants were distributed across 7,132 genes, representing a large candidate set for novel disease associations.

Functional enrichment analyses were performed using Gene Ontology (GO), KEGG, and Reactome via the Enrichr platform to assess biological relevance. Gene Ontology and pathway enrichment analyses consistently indicated that genes harboring high impact variants are involved in key biological processes relevant to T2D, including insulin secretion, glucose homeostasis, and β-cell signaling as seen in Figure 3. KEGG and Reactome pathway analyses shown in Table 6

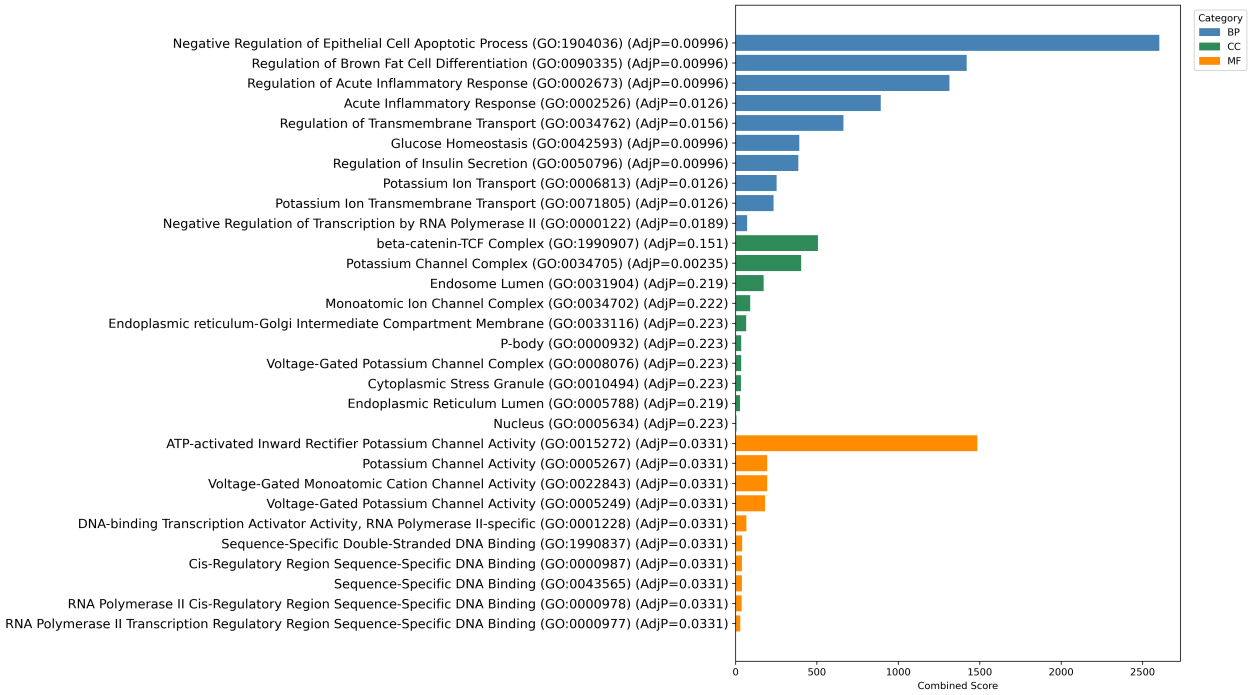

Fig. 3. Visual representation of the top 10 enriched Gene Ontology (GO) terms from each category, Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) ranked by combined score

Table 6. Top enriched pathways identified from KEGG and Reactome databases, ranked by enrichment ratio (ER) and adjusted *p*-value, highlighting key genes associated with metabolic and diabetes-related processes

Database	Pathway Name	ER[a]	Adj *p*-value[b]	Genes
KEGG	Maturity onset diabetes of the young	241.93	8.41E-07	HHEX; HNF4A; PAX4; INS
	Insulin secretion	64.71	5.76E-05	KCNJ11; ABCC8; ADCY5; INS
	Type II diabetes mellitus	86.94	3.51E-04	KCNJ11; ABCC8; INS
	Cushing syndrome	24.46	9.89E-03	TCF7L2; CDKN2B; ADCY5
	Regulation of lipolysis in adipocytes	44.24	2.01E-02	ADCY5; INS
Reactome	Regulation of Insulin Secretion	0.0513	1.23E-04	KCNJ11; ABCC8; ADCY5; INS
	Integration of Energy Metabolism	0.0370	2.27E-04	KCNJ11; ABCC8; ADCY5; INS
	Regulation of β-Cell Development	0.0714	4.21E-04	HNF4A; PAX4; INS
	Potassium Channels	0.0291	4.69E-03	KCNJ11; ABCC8; KCNQ1
	Regulation of Gene Expression in β Cells	0.0952	5.51E-03	HNF4A; INS

ER: Enrichment Ratio, calculated as Observed/Expected overlap.
Adjusted *p*-value using Benjamini–Hochberg FDR.

identified significant enrichment in pathways such as insulin secretion, maturity-onset diabetes of the young, and β-cell development.

The recurrence of established T2D genes, including *TCF7L2*, *KCNJ11*, *ABCC8*, *HHEX*, *HNF4A*, *WFS1*, and *FTO*, within these pathways supports the biological plausibility of the identified variants.

Finally, three candidate variants prioritized by in silico pathogenicity scoring missense variants were prioritized in established T2D-associated genes, including *WFS1* (chr4:6302293 T>C;

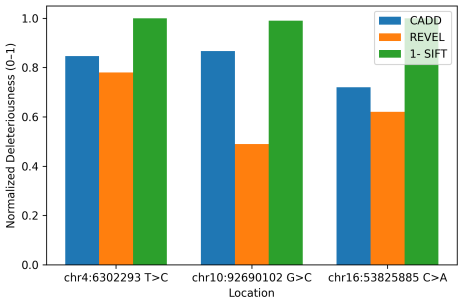

Fig. 4. Normalized functional impact scores of prioritized variants in T2D-associated genes. CADD scores were scaled to a 0–1 range, REVEL scores were used directly, and SIFT scores were inverted so that higher values consistently indicate greater deleterious potential across all metrics

p.Leu833Pro), *HHEX* (chr10:92690102 G>C; p.Gly39Ala), and *FTO* (chr16:53825885 C>A; p.Leu49Ile). All variants were absent from major population databases, indicating potential novelty. Functional predictions consistently indicate strong deleterious effects, with high CADD scores between 21.6 and 26.0, moderate-to-high REVEL scores of 0.49 to 0.78, and concordant SIFT and PolyPhen-2 classifications as deleterious and probably damaging. Sequencing metrics further support their reliability, with adequate read depth of 42 to 51, balanced variant allele fractions (0.24-0.49), and high genotype quality scores of more than 90.

These variants are located within functionally relevant protein domains, including the luminal domain of *WFS1*, transcription factor domain of *HHEX*, and structural regions of *FTO*, suggesting potential biological impact. Detailed annotations are provided in Figure 4. These collective findings highlight rare, potentially pathogenic variants that may contribute to T2D susceptibility in under-represented Asian populations. Given the established roles of *WFS1*, *HHEX*, and *FTO* in glucose metabolism and pancreatic β-cell function, these variants may have potential clinical relevance.

However, further validation in larger cohorts and integration with detailed phenotype data are required.

4 Conclusion

This study presents a computational framework for efficient WES variant analysis that addresses challenges related to computational efficiency and storage management in genomic data processing. The proposed pipeline combines automated quality filtering, alignment, and variant detection, enabling reduced intermediate data storage without increasing computational requirements and reducing manual intervention. The framework enabled the identification of both known and potentially novel T2D-associated variants when applied to WES data from T2D patients, non-T2D diabetes individuals, and healthy controls from diverse Asian cohort.

Rare deleterious missense variants were identified, including candidate variants in WFS1, HHEX, and FTO. Functional relevance of these variants is further supported by their association with key biological processes related to glucose regulation and pancreatic β-cell function. Although the identified variants demonstrate strong in silico evidence of pathogenicity, clinical validation and genotype–phenotype correlation were beyond the scope of the present study. Future investigations integrating detailed clinical parameters are necessary to establish their translational and clinical relevance in T2D, particularly in resource-constrained settings.

In addition to variant discovery, the proposed pipeline demonstrates improved efficiency in data handling. A comparison with a conventional GATK-based workflow shows about 24% reduction in total storage requirements. The reduction was achieved by reducing the retention of intermediate files and optimizing the management of alignment outputs. This highlights the practical advantage of the framework for large-scale genomic analysis.

A lightweight visual analytics facilitates exploration and interpretation of high-dimensional genomic features which serves as a complementary extension to the primary computational framework. Overall, the results demonstrated that efficient genomic analysis and variant discovery can be achieved in resource constrained environments using an optimized and integrated pipeline. While further validation in

larger cohorts is required, the proposed framework provides a practical foundation for variant discovery low resource environments.

Declarations

Funding: This work was supported by the Council of Scientific and Industrial Research (CSIR), Government of India, under grant number 22WS(0023)/2023-24/EMR-II/ASPIRE.

Availability of data and materials: Data used for the analysis were collected from the NCBI public database. BioProjects PRJNA288292, PRJNA656961, PRJNA579845 and PRJEB5736[1].

References

1. **(2021).** The gene ontology resource: enriching a gold mine. Nucleic acids research, Vol. 49, No. D1, pp. D325–D334.
2. **Abdelwahab, O., Torkamaneh, D. (2025).** Artificial intelligence in variant calling: a review. Frontiers in Bioinformatics, Vol. 5, pp. 1574359.
3. **Adzhubei, I., Jordan, D. M., Sunyaev, S. R. (2013).** Predicting functional effect of human missense mutations using polyphen-2. Current protocols in human genetics, Vol. 76, No. 1, pp. 7–20.
4. **Ahmad, T., Ahmed, N., Al-Ars, Z., Hofstee, H. P. (2020).** Optimizing performance of gatk workflows using apache arrow in-memory data framework. BMC genomics, Vol. 21, No. Suppl 10, pp. 683.
5. **Bolger, A. M., Lohse, M., Usadel, B. (2014).** Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics, Vol. 30, No. 15, pp. 2114–2120.
6. **Chambers, J. C., Abbott, J., Zhang, W., Turro, E., Scott, W. R., Tan, S.-T., Afzal, U., Afaq, S., Loh, M., Lehne, B., et al. (2014).** The south asian genome. PloS one, Vol. 9, No. 8, pp. e102645.
7. **Chen, S., Zhou, Y., Chen, Y., Gu, J. (2018).** fastp: an ultra-fast all-in-one fastq preprocessor. Bioinformatics, Vol. 34, No. 17, pp. i884–i890.
8. **Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Nohzadeh-Malakshah, S., Rathod, M., et al. (2014).** Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. Journal of Computational Biology, Vol. 21, No. 6, pp. 405–419.
9. **Deepali, Goel, N., Khandnor, P. (2025).** Advances in ai-based genomic data analysis for cancer survival prediction. Multimedia Tools and Applications, Vol. 84, No. 14, pp. 14139–14166.
10. **Delvecchio, M., Pastore, C., Giordano, P. (2020).** Treatment options for mody patients: a systematic review of literature. Diabetes Therapy, Vol. 11, No. 8, pp. 1667–1685.
11. **Devuyst, O. (2015).** The 1000 genomes project: welcome to a new world. Peritoneal dialysis international: journal of the International Society for Peritoneal Dialysis, Vol. 35, No. 7, pp. 676.
12. **Duncan, B. B., Magliano, D. J., Boyko, E. J. (2025).** Idf diabetes atlas 11th edition 2025: global prevalence and projections for 2050.
13. **Fatumo, S. (2024).** Including diverse populations enhances the discovery of type 2 diabetes loci. Nature Reviews Genetics, Vol. 25, No. 2, pp. 82–82.
14. **Garrison, E., Marth, G. (2012).** Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.
15. **Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., et al. (2022).** The reactome pathway knowledgebase 2022. Nucleic acids research, Vol. 50, No. D1, pp. D687–D692.

[1] https://www.ncbi.nlm.nih.gov

16. **Göğebakan, K., Ulu, R., Abiyev, R., Şah, M. (2024).** A drug prescription recommendation system based on novel diakid ontology and extensive semantic rules. Health Information Science and Systems, Vol. 12, No. 1, pp. 27.

17. **Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016).** Revel: an ensemble method for predicting the pathogenicity of rare missense variants. The American Journal of Human Genetics, Vol. 99, No. 4, pp. 877–885.

18. **Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M. (2021).** Kegg: integrating viruses and cellular organisms. Nucleic acids research, Vol. 49, No. D1, pp. D545–D551.

19. **Kore, P., Wilson, M. W., Tiao, G., Chao, K., Darnowsky, P. W., Watts, N. A., Mauer, J. H., Baxter, S. M., Consortium, G. A. D., Rehm, H. L., et al. (2025).** Improved allele frequencies in gnomad through local ancestry inference. Nature Communications, Vol. 16, No. 1, pp. 8734.

20. **Lalrinmawii, C., Dutta, A., Elangovan, R. (2025).** Afastqf: An automated fastq quality filtering of high-throughput genomic data. Proceedings of the 2024 8th International Conference on Computational Biology and Bioinformatics, Association for Computing Machinery, New York, NY, USA, pp. 59–65. DOI: `10.1145/3715020.3715047`.

21. **Lalrinmawii, C., Muppalaneni, N. B. (2023).** An overview of the workflow of next-generation sequencing data analysis. 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1655–1659. DOI: `10.1109/ICESC57686.2023.10193007`.

22. **Lalrohlui, F., Zohmingthanga, J., Vanlallawma, A., Kumar, N. S., et al. (2021).** Whole exome sequencing identifies the novel putative gene variants related with type 2 diabetes in mizo population, northeast india. Gene, Vol. 769, pp. 145229.

23. **Landrum, M. J., Chitipiralla, S., Kaur, K., Brown, G., Chen, C., Hart, J., Hoffman, D., Jang, W., Liu, C., Maddipatla, Z., et al. (2025).** Clinvar: updates to support classifications of both germline and somatic variants. Nucleic Acids Research, Vol. 53, No. D1, pp. D1313–D1321.

24. **Langmead, B., Wilks, C., Antonescu, V., Charles, R. (2019).** Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics, Vol. 35, No. 3, pp. 421–432.

25. **Li, H. (2013).** Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprint arXiv:1303.3997.

26. **Li, H. (2018).** Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, Vol. 34, No. 18, pp. 3094–3100.

27. **Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, . G. P. D. P. (2009).** The sequence alignment/map format and samtools. bioinformatics, Vol. 25, No. 16, pp. 2078–2079.

28. **Li, Z.-q., Yang, D.-j. (2025).** Advances in genome-wide association studies of type 2 diabetes across populations: A systematic review. International Journal of Diabetes in Developing Countries, pp. 1–18.

29. **Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., Payne, A. J., Steinthorsdottir, V., Scott, R. A., Grarup, N., et al. (2018).** Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature genetics, Vol. 50, No. 11, pp. 1505–1513.

30. **Martin, M. (2011).** Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, Vol. 17, No. 1, pp. 10–12.

31. **Mauvais, F.-X., van Endert, P. M. (2025).** Type 1 diabetes: A guide to autoimmune

mechanisms for clinicians. Diabetes, Obesity and Metabolism, Vol. 27, pp. 40–56.

32. **McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., Cunningham, F. (2016).** The ensembl variant effect predictor. Genome biology, Vol. 17, pp. 1–14.

33. **Nath, A., Kumar, C. J., Bharalee, R., Baruah, S., Nath, K., Hassan, S. M. (2026).** Neomirx an interpretable deep learning framework for mirna-based cancer classification and regulatory network analysis. Discover Artificial Intelligence.

34. **Ng, P. C., Henikoff, S. (2003).** Sift: Predicting amino acid changes that affect protein function. Nucleic acids research, Vol. 31, No. 13, pp. 3812–3814.

35. **Omeranovic, A., Van Long, F. N., Boubaker, A., Turgeon, A., Nabi, H. (2025).** Underrepresented populations in genomic research: a qualitative study of researchers' perspectives. BMC medical genomics, Vol. 18, No. 1, pp. 72.

36. **O'Beirne, S. L., Salit, J., Rodriguez-Flores, J. L., Staudt, M. R., Abi Khalil, C., Fakhro, K. A., Robay, A., Ramstetter, M. D., Malek, J. A., Zirie, M., et al. (2018).** Exome sequencing-based identification of novel type 2 diabetes risk allele loci in the qatari population. PloS one, Vol. 13, No. 9, pp. e0199837.

37. **O'Rawe, J. A., Ferson, S., Lyon, G. J. (2015).** Accounting for uncertainty in dna sequencing data. Trends in Genetics, Vol. 31, No. 2, pp. 61–66.

38. **Phan, L., Zhang, H., Wang, Q., Villamarin, R., Hefferon, T., Ramanathan, A., Kattman, B. (2025).** The evolution of dbsnp: 25 years of impact in genomic research. Nucleic acids research, Vol. 53, No. D1, pp. D925–D931.

39. **Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., et al. (2018).** A universal snp and small-indel variant caller using deep neural networks. Nature biotechnology, Vol. 36, No. 10, pp. 983–987.

40. **Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., Kircher, M. (2019).** Cadd: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research, Vol. 47, No. D1, pp. D886–D894.

41. **Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., et al. (2020).** Ncbi taxonomy: a comprehensive update on curation, resources and tools. Database, Vol. 2020, pp. baaa062.

42. **Shi, L., Wang, Z. (2019).** Computational strategies for scalable genomics analysis. Genes, Vol. 10, No. 12, pp. 1017.

43. **Sun, X., Sui, W., Wang, X., Hou, X., Ou, M., Dai, Y., Xiang, Y. (2016).** Whole-genome re-sequencing for the identification of high contribution susceptibility gene variants in patients with type 2 diabetes. Molecular Medicine Reports, Vol. 13, No. 5, pp. 3735–3746.

44. **Van der Auwera, G. A., O'Connor, B. D. (2020).** Genomics in the cloud: using Docker, GATK, and WDL in Terra. O'Reilly Media.

45. **Wilton, R., Szalay, A. S. (2022).** Performance optimization in dna short-read alignment. Bioinformatics, Vol. 38, No. 8, pp. 2081–2087.

46. **Xie, Z., Bailey, A., Kuleshov, M. V., Clarke, D. J., Evangelista, J. E., Jenkins, S. L., Lachmann, A., Wojciechowicz, M. L., Kropiwnicki, E., Jagodnik, K. M., et al. (2021).** Gene set knowledge discovery with enrichr. Current protocols, Vol. 1, No. 3, pp. e90.

47. **Zhang, D., Zhang, J., Du, J., Zhou, Y., Wu, P., Liu, Z., Sun, Z., Wang, J., Ding, W., Chen, J., et al. (2022).** Optimized sequencing adaptors enable rapid and real-time metagenomic identification of pathogens during runtime of sequencing. Clinical Chemistry, Vol. 68, No. 6, pp. 826–836.

48. **Zheng, Z., Li, S., Su, J., Leung, A. W.-S., Lam, T.-W., Luo, R. (2022).** Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nature Computational Science, Vol. 2, No. 12, pp. 797–803.

Article received on 24/03/2026; accepted on 20/06/2026.
**Corresponding author is C Lalrinmawii.*