Benchmarking whole exome sequencing pipeline for pathogenic variant detection

Ranjana Mariyam Raju[1*], Ujjwal Prathap Singh[1*] and Prashanth Suravajhala[1,2#]

1. Systems Genomics Lab, Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham, Clappana PO 690525, Kerala, India
2. Bioclues.org, Hyderabad 501511 India

*Equal contributing authors

#Correspondence: prash@bioclues.org

Abstract

Benchmarking whole exome pipelines is crucial for evaluating and comparing their performance in variant calling and clinical significance annotation. It enables researchers and clinicians to assess the accuracy, sensitivity, and specificity of different pipelines and identify the most effective and reliable ones. In this study, we evaluated and compared the performance of our in-house exome CONVEX pipeline with a widely recognized gold standard Genome Analysis Toolkit (GATK) pipeline. Four datasets were used for evaluation, three 1000 Genome Project (1KGP) datasets and one Prostate cancer (PCa) Sample. The CONVEX pipeline demonstrated a higher average transition-to-transversion (Ti/Tv) ratio, indicating enhanced precision in identifying single nucleotide variant (SNV) calls. This suggests that the CONVEX pipeline excels in effectively discerning true genetic variations from sequencing artefacts, particularly in the context of exome sequencing. Additionally, the pipeline exhibited increased sensitivity in detecting pathogenic and likely pathogenic variants in the PCa sample, further highlighting its efficacy in identifying clinically relevant somatic variants. We also conducted a trio-exome analysis with the use of two trio pipelines, namely VarScan Trio and GATK joint calling pipelines on Congenital Pouch Colon (CPC) samples from rare disease cohort which we published earlier and found that the GATK predicted a significantly higher number of somatic variants compared to VarScan. While our pipeline handles somatic variant calls, the CONVEX proved to be efficient, requiring less computational time and is comparable to the gold standard GATK pipeline.

Keywords: Exome sequencing, benchmarking, next-generation sequencing, variant calling, trio analysis

Introduction

Genomic analysis refers to the process of examining, interpreting, and understanding the information encoded in an organism's DNA sequence. This can involve various techniques and approaches, such as identifying genes, detecting genetic variations and studying gene expression patterns. High-throughput genomic analysis has been made possible by the development of next-generation sequencing (NGS) technologies, which generate massive amounts of genomic data[1]. The main feature distinguishing NGS from Sanger sequencing is its ability to perform massively parallel sequencing of DNA or RNA fragments, allowing for the simultaneous analysis of multiple samples or targets[2].

Whole exome sequencing (WES) has several advantages over whole genome sequencing (WGS) in certain contexts, for example, cost and resource efficiency as it focuses on the protein-coding regions (exons) of the genome, which constitute only about 1-2% of the entire genome. WES specifically targets the exons, where most known disease-causing mutations occur, allowing researchers to focus on the most functionally relevant parts of the genome, increasing the likelihood of identifying causative mutations[3,4]. While the targeted approach can be particularly useful for diagnosing rare genetic disorders caused by mutations in these regions, the WES or targeted approaches generate a smaller dataset compared to WGS, making data analysis and storage more feasible, and particularly beneficial for clinical research laboratories with limited computational resources. Furthermore, WES can also provide better sensitivity for detecting single nucleotide variants (SNVs) and small insertions/deletions (indels) in the coding regions compared to WGS, with limitations being the inability to detect non-coding

and structural variants[5]. As sequencing costs continue to decrease and bioinformatics tools improve, the use of WGS may become more widespread, offering a comprehensive view of the genome. In some cases, WGS may be more suitable for diagnosing rare diseases, as it can provide a more comprehensive view of the genome and identify pathogenic variants in non-coding regions. On the other hand, whole exome-trio sequencing is effective for diagnosing and screening hereditary disorders, especially for patients with undiagnosed or rare conditions. This method sequences the exomes of an affected individual and their unaffected parents to pinpoint rare, inherited variants potentially causing the condition. The resulting concise report offers in-depth, phenotype-driven interpretation, with an option to include secondary findings. Integrating DNA, RNA, and protein data can improve therapeutic stratification, potentially increasing the success of precision cancer therapy[5,6]. For conditions like epilepsy, a genetic diagnosis can aid family planning, provide answers, and prevent unnecessary investigations[7,9].

Over the years, systematic comparison of variant calls have produced comprehensive sequencing depth demonstrating higher performance by evaluating various pipelines [10,11,25,26]. Furthermore, efforts were made to establish community reference samples, data and call sets for benchmarking and detecting mutations[28-33] In this work, we performed a trio-exome analysis with the use of two trio pipelines, namely VarScan Trio and GATK joint calling pipelines on Congenital Pouch Colon (CPC) samples from our rare disease cohort([12],[13],[14] and Prostate cancer (PCa)[15] which we have published earlier. We conducted a benchmarking analysis using four datasets and a high-confidence variant call set from the Genome in a Bottle (GIAB) consortium. This truth set allowed us to validate pipeline performance and investigate the association of identified variants with disease-causing mutations in CPC and PCa. We also discuss the challenges in predicting pathogenic variants of significance based on our benchmarking results.

Materials and Methods

Benchmarking datasets for pipeline

Four datasets were analyzed with one dataset having a high confidence variant call set. FASTQ files of human exome HapMap/1000 Genomes project (1KGP) CEU female NA12878 (accession No.: SRR098401) were downloaded from NCBI- Sequence Read Archive (SRA- https://www.ncbi.nlm.nih.gov/sra last accessed on June 10, 2023). The WES of NA12878 was performed using the HiSeq2500 platform and Nextera Rapid Capture Exome and Expanded Exome kit. The human reference genomes GRCh38 were downloaded from the Ensembl[16]. Next, a NA12878 high confidence call set to version 4.2.2 by the GIAB consortium was used for pipeline performance validation across all datasets. The variant set along with a BED file was downloaded from NCBI. This GIAB variant set, created by integrating 14 different datasets from five sequencers, is a 'gold standard' variant dataset publicly available for systematic comparison of variant callers (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/ last accessed on March 12, 2025) [27]. For the analysis of pathogenic variant distribution, pathogenic, likely pathogenic variants and variants of uncertain/unknown significance (VUS) with non-conflicting interpretations were selected from ClinVar v.20230520. For added validation, two genomes from the 1KGP-CHS (East Asian Ancestry) and FIN (European Ancestry) populations were taken (SRA IDs: ERR032022, ERR031490) apart from the PCa samples from the cohort used in our Systems Genomics lab as discussed earlier.

Trio datasets

For the trio-exome analysis, we employed the VarScan Trio and GATK Joint Variant Calling Pipeline, one GIAB trio of NA12878 proband with father NA12891 and mother NA12892. Father and Mother FASTQ files were downloaded from 1KGP (A map of human genome variation from population-scale sequencing 2010; SRA IDs: SRR098359, ERR034529). Furthermore, two of the original CPC samples (Z12 and Z19 families) from a previous study conducted in our Systems Genomics lab were taken for conducting a retrospective study on the rare variants found using two trio variant calling pipelines with updated ClinVar records.

Benchmarking datasets			
Dataset	**Coverage**	**Illumina Sequencing Platform**	**Accession IDs**
NA12878	110x	HiSeq2500	SRR098401
HG00284	110x	NextSeq 2000	ERR031490
HG00448	110x	NextSeq 2000	ERR032022
PCa	110x	NextSeq 2000	Not published
Trio Exome Samples			
NA12878	110x	HiSeq2500	SRR098401
NA12891	110x	HiSeq2500	SRR098359
NA12892	11x	HiSeq2500	ERR034529
Z12	110x	NextSeq 500	SRR21615279
Z13	110x	NextSeq 500	SRR21615280
Z14	110x	NextSeq 500	SRR21615281
Z19	110x	NextSeq 500	SRR21615287
Z20	110x	NextSeq 500	SRR21615288
Z21	110x	NextSeq 500	SRR21615289

Table 1: Datasets used for analyses

Pipeline

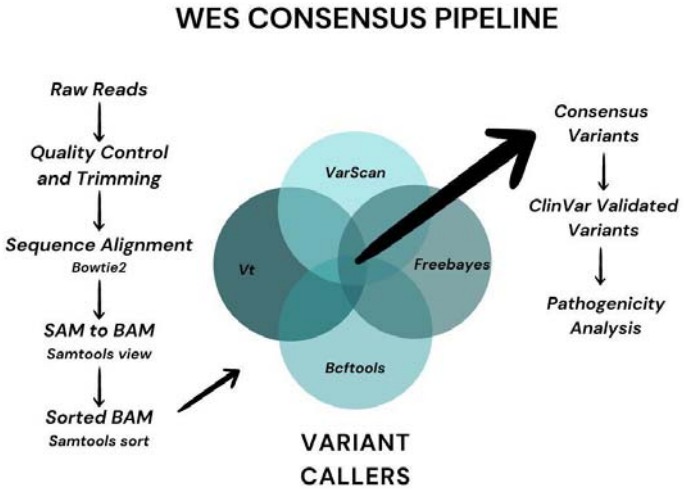


Figure 1: The Convex pipeline workflow

We have used our former WES pipeline[17] and enhanced it to call it CONsensus Variant EXome (CONVEX; https://github.com/prashbio/CONVEX) pipeline which works with four open-source variant callers in detecting variants for a comprehensive and increased sensitivity (Singh et al. 2023.,

communicated). In summary, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ last accessed on September 2, 2023) was used to evaluate the quality of all samples, as we examined key characteristics of the raw reads, such as their duration, quality scores, and base distribution, to judge the data efficiency and eliminate low-quality reads. The nucleotide distribution across all reads has been obtained thanks to Phred's optimal base quality scores, and the best quality score for GC content was set to a 50% threshold. The alignment of the reads was done using the tool Bowtie2[18] followed by the conversion of sequence alignment mapping (SAM) to binary alignment mapping (BAM) and sorting of BAM by Samtools. This analysis-ready BAM file was given to 2 of the variant callers, *i.e.* Freebayes and Vt while Bcftools and VarScan require an additional requirement of the mpileup BAM file from Samtools as input to call variants. All the variant call sets were compared with each other using awk commands on chromosome number, chromosome position and alternate allele called by the variant callers. The CONVEX variants obtained were channelled through downstream analysis of validating with the ClinVar database with further genomic variant filtration as per the American College of Medical Genetics and Genomics (ACMG;[19]) and Association of Molecular Pathology (AMP; [20])guidelines.

The Genome Analysis Tool Kit (GATK)

The GATK is a widely used tool for processing genomic data, especially for variant discovery and genotyping as it serves as the best practice for processing NGS data that are widely adopted in the genomics community[21]. After the quality control steps of FastQC and trimming of adapters with Trimmomatic[22], the read alignment was done by a high-quality aligner such as BWA-MEM to yield a SAM file[23]. The duplicates were later marked with the *Picard MarkDuplicates* to remove the duplicates that arise from PCR amplification or sequencing biases. The final step was to identify variants in the sample using GATK's HaplotypeCaller tool[21]. Over the last few years, Unified genotyper has been deprecated in GATK even as the tool uses a probabilistic approach to assemble haplotypes and call variants simultaneously, taking into account the quality scores and alignment information for each read. Further, the raw variants obtained were filtered with various tools with the SNPs and Indels were selected through the SelectVariants tool.In this process, we considered SNPs (not Indels) for further filtering with standard filters as per the GATK best practices guidelines for the refinement of variants. The filtered SNPs were then validated with the ClinVar records and genetic variation analysis was done for benchmarking with the CONVEX pipeline.

Trio-exome analysis

For the analysis of the trio-exome datasets, two variant callers having trio sample calling functionality were used (Figure 2a). The initial preprocessing of the samples was done by FastQC and Trimmomatic for all the samples in the three trio datasets. While read alignment was done using Bowtie2 for all the datasets, the individual SAM files were converted to BAM with the addition of read groups. It was further sorted and indexed with Samtools and the analysis-ready BAM files of each trio were given to the respective trio variant caller. For the VarScan pipeline, the Samtools mpileup function takes in three sorted BAM files to make one joint trio mpileup file. This file is given as input to the VarScan trio subcommand which does the variant calling with default parameters. VarScan trio utilises the family relationships and phenotypes to improve the accuracy of variant calls. It calls variants across all three samples using default parameters of --min-var-freq of 0.20 and --p-value to be 0.05. Next, it classifies any variant as Mendelian Inheritance Errors (MIE) even as it recalls such variants with adjusted settings and corrects them or the site will be marked as MIE or DENOVO. The joint variant calling workflow by GATK is a comprehensive method to detect genetic variants from sequencing data accurately. It consists of three main steps: Haplotype Caller, CombineGVCFs, and GenotypeGVCFs.The initial step, Haplotype Caller, uses local *de novo* assembly to identify variants in individual samples by reconstructing haplotypes and saving the results in the genomic Variant Call Format (gVCF). In the gVCF format, variant sites are preserved, while non-variant sites are grouped into blocks based on genotype quality during the calling process. This method allows for compression of the Variant Call Format (VCF) file without excluding any sites, thus facilitating efficient joint analysis in subsequent stages. We used the family information, such as a pedigree file and gVCF of all samples in the trio set (the father, mother, and afflicted kid), to call *de novo* variants. To identify *de novo* variants in the trio cohort, we initially generated

gVCF files for each sample within the trio using the HaplotypeCaller tool. These individual gVCF files were then merged into a multi-sample gVCF file using CombineGVCFs. Subsequently, GenotypeGVCFs were employed to call the raw variants. The result of running GenotypeGVCFs is a consolidated VCF file that contains genotype information for every variant site across all the samples. This file includes details about the genotypes, qualities of genotypes, and other pertinent annotations. GenotypeGVCFs provide flexibility in handling datasets with various ploidy levels and are capable of accommodating diploid, haploid or polyploid genomes. After this, a genotype refinement workflow (Figure 2b) was implemented which included three tools obtaining the posterior probabilities of genotypes through CalculateGenotypePosteriors, and VariantFiltration, employed to remove low-quality genotypes. Additionally, VariantAnnotator annotated *de novo* SNVs and the filtered SNVs from both pipelines were analysed for pathogenicity through concordance with the ClinVar database records and our results were compared with the previous study. Finally, the benchmarked datasets on CONVEX and GATK pipelines were analysed with a truth set. For additional verification of the ability of the developed CONVEX pipeline in comparison to the established Gold Standard GATK pipeline, both pipelines were run on a widely used benchmarking dataset, NA12878 exome sequencing data. The availability of truth set for variant calls given by the GIAB acts as a reference to which both the pipelines could be compared and the CONVEX pipeline evaluated without any bias.

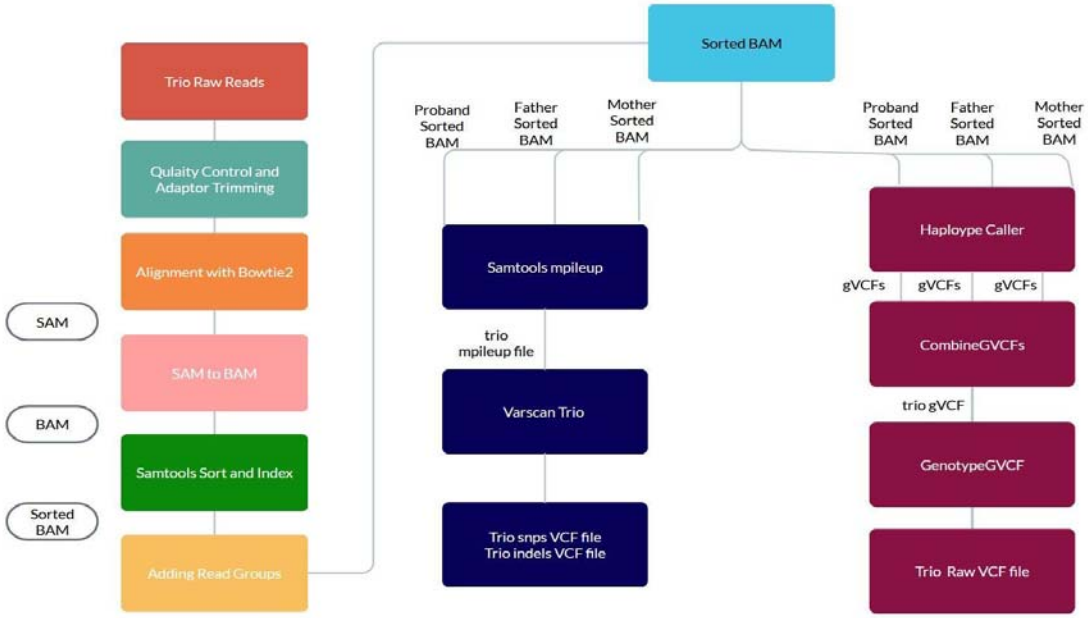
Statistics

We assessed the performance of our pipeline using three key metrics: F1 score, sensitivity, and specificity. The F1 score combines precision and recall, providing a balanced evaluation of our model's accuracy. Sensitivity measures the correct identification of positive cases, while specificity measures the correct identification of negative cases. These metrics were computed for each experimental run to evaluate our method's performance and compare it with other approaches. By utilising the F1 score, sensitivity, and specificity metrics, we aimed to evaluate our method's performance and determine its

$$
\begin{aligned}
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP} \\
specificity &= \frac{TN}{TN + FP}
\end{aligned}
$$

suitability for the task at hand.

(a)


Trio Raw Reads
Qulaity Control and Adaptor Trimming
Alignment with Bowtie2
SAM
SAM to BAM
BAM
Samtools Sort and Index
Sorted BAM
Adding Read Groups
Sorted BAM
Proband Sorted BAM
Father Sorted BAM
Mother Sorted BAM
Samtools mpileup
trio mpileup file
Varscan Trio
Trio snps VCF file
Trio indels VCF file
Proband Sorted BAM
Father Sorted BAM
Mother Sorted BAM
Haploype Caller
gVCFs
gVCFs
gVCFs
CombineGVCFs
trio gVCF
GenotypeGVCF
Trio Raw VCF file

(b)

Figure 2 (a) An overview of steps used in the trio-exome pipeline with VarScan and GATK and (b) Genotype refinement workflow

Results and Discussion

The CONVEX pipeline yielded a significant number of variants

The pipelines were executed on a server equipped with a 40-core CPU, 128 GB of RAM and L3 cache. The execution time for one exome sequenced sample was approximately 6 h 30 mins for GATK and 5 h 30 m for the CONVEX pipeline. The time required for variant calling was approximately 3 h 30 mins for GATK and 1 h 30 mins for CONVEX pipeline wherein on average 30,00,000 variants were identified in each sample. We observed that the GATK called more SNVs and indels, but the CONVEX pipeline gave a higher average transition-to-transversion (Ti/Tv) ratio (5.57 for CONVEX and 3.45 for GATK). The Ti/Tv ratio is an important criterion for assessing the quality of SNV calling, for exome sequencing, the expected Ti/Tv ratio is approximately 3.011. A higher Ti/Tv ratio usually indicates a higher accuracy of SNV calling.

	SNVs		Ti/Tv ratio	
Samples	CONVEX	GATK	CONVEX	GATK
HG00448	81156	**266818**	**5.73645**	2.56075
HG00284	73988	**267929**	**5.82906**	2.82882
PCa	68552	**1043400**	**5.46942**	5.33035

Table 2: Variant counts and Ti/Tv Ratio from the two pipelines

The genetic analysis of three samples, HG00448, HG00284, and PCa revealed significant variations in SNVs and Ti/Tv ratios (Table 2; Figure 3). The GATK pipeline consistently yielded higher SNV counts across all samples. Ti/Tv ratios indicated a higher prevalence of transitions in CONVEX SNVs, whereas the GATK SNVs showed a more balanced distribution between transitions and transversions. While we observed that the CONVEX pipeline yielded a large number of variants with conflicting pathogenicity, the variants called from both pipelines were further evaluated for pathogenicity based on the presence of variants in the ClinVar database and the concordance check of the identified variation in clinical significance.

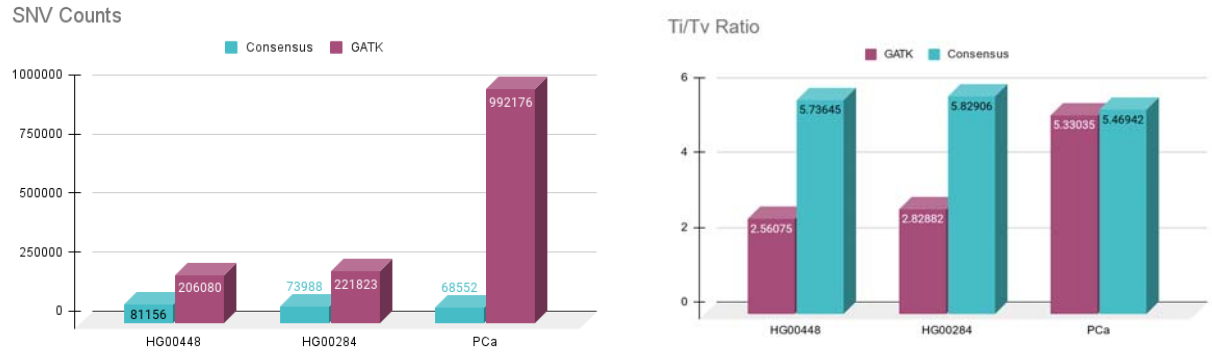


Figure 3 (a) SNV counts taken together from CONVEX and GATK pipeline and (b) Ti/Tv ratios obtained from CONVEX and GATK pipelines

	HG00448		HG00284		PCa	
	GATK	CONVEX	GATK	CONVEX	GATK	CONVEX

Pathogenic	6	3	1	3	3	11
Likely Pathogenic	4	2	2	1	5	3
Pathogenic and likely pathogenic	0	0	1	1	7	2
VUS	30	33	30	28	134	109
Variants of conflicting pathogenicity	70	51	89	64	85	66
Total Clinvar Validated	17842	17862	18649	17398	14343	10677

Table 3: Clinical significance of variants called by the CONVEX and GATK pipeline

The counts of variants for each category of clinical significance, namely pathogenic, likely pathogenic, VUS and conflicting pathogenicity was tabulated wherein our analysis showed an almost comparable number of ClinVar validated records from both pipelines. This asserts that there was a significant difference in the initial number of SNPs called by each pipeline where the final analysis resulted in a similar performance of both pipelines. We argue that the CONVEX pipeline appeared to be more sensitive in detecting pathogenic and likely pathogenic variants in the PCa sample, while GATK showed better performance in screening VUS and ClinVar-validated variants. The CONVEX pipeline also demonstrated lower numbers of variants of conflicting pathogenicity, indicating more consistent results. On the other hand, the SNPs called by GATK and the CONVEX pipeline were comparable and fewer in the CONVEX pipeline can be attributed to the default variant filtering applied by each variant caller and the algorithms used to call the variants. So the variants obtained were deemed the most confident ones among the four variant callers. The high SNP count for the truth set of NA12878 can be due to the various samples used for making the truth set with no filtrations applied.

Benchmarking yielded variant calls from both the pipelines with ClinVar validation

Benchmarking workflows include statistical metrics such as sensitivity, precision, F1 score and specificity which provide an insight into the pipeline performance. Here both the GATK pipeline metrics and the CONVEX pipelines are compared against the truth set variants of benchmark NA12878 dataset. The variant calls obtained from both pipelines were further annotated with the ClinVar database for the clinical significance of the genetic variants providing clinical interpretations for variants (Table 4). For the current analysis, this annotation will help identify the various metrics associated with the benchmark workflow to assess the pipeline's ability to call clinically relevant variants as well as concordance with the established pipelines.

	NA12878		
	GATK	CONVEX	Truth set
Pathogenic	6	5	9
Uncertain significance variants	113	78	112
Variants of conflicting pathogenicity	110	95	116
Clinvar Validated	22062	14809	34074

Table 4: Pathogenicity analyses for GATK, CONVEX and truth set

The CONVEX pipeline had a total of 14809 ClinVar validated variants wherein 5 of them were pathogenic, 4 were identified as true positives and 1 as a false positive. While specificity was found to be 0.999971, the sensitivity of the CONVEX pipeline was 0.444, and the F1 score was around 0.571. On the other hand, the GATK pipeline, which processed a larger dataset of 22062 ClinVar validated variants, identified 6 pathogenic variants. It achieved 5 true positives and 4 false positives, yielding a higher specificity of approximately 0.999882. The GATK pipeline demonstrated a higher sensitivity of around 0.833 and an F1 score of approximately 0.667.

Comparison of trio-exome pipelines yielded distinct variants

One of the three trio-exome datasets is a 1KGP benchmark dataset that is a female and her father and mother were taken as a trio for this analysis. The other two families are CPC families(Z12 and Z19) in which only the proband is affected with the disease with the parents unaffected. The GATK trio pipeline took approximately 17 hours to finish one trio set and the VarScan trio pipeline took approximately 8 hours to complete the trio workflow. Both the pipelines included various steps of joint variant calling of trio samples and that is probably the reason for the increased time for the pipeline completion. GATK also employed additional genotype refinement as well as some joint calling specific subcommands which increases the accuracy and specificity of the variants called but also increases the total time taken to finish the pipeline.

	Total SNVs		After filtration SNVs	
Samples	GATK Trio	VarScan Trio	GATK Trio	VarScan Trio
NA12878_family	1788766	70064	677150	49051
Z12_family	1326375	145238	277160	102124
Z19_family	1408848	152132	332480	108207

Table 5: Total variants called by VarScan and GATK trio pipelines

The GATK and VarScan pipelines identified distinct sets of variants, with GATK calling significantly more due to its sensitivity-focused algorithm (Table 5; Figure 4). Both pipelines require additional stringent filtering, which was performed in subsequent steps. However, it's important to note that potential filtering artifacts may lead to a false positive rate in both pipelines. The GATK's local realignment and haplotype assembly techniques enhance its accuracy in calling variants in complex regions, potentially detecting variants missed by other callers. Filtered variants were annotated using the ClinVar database variant record VCF file (Table 6). Annotations were confirmed by checking for concordance between chromosome number, chromosome position, reference and alternate alleles in both pipeline VCF files and ClinVar records. The count of pathogenic and likely pathogenic variants was low, consistent with the rarity of the diseases in two of the samples.

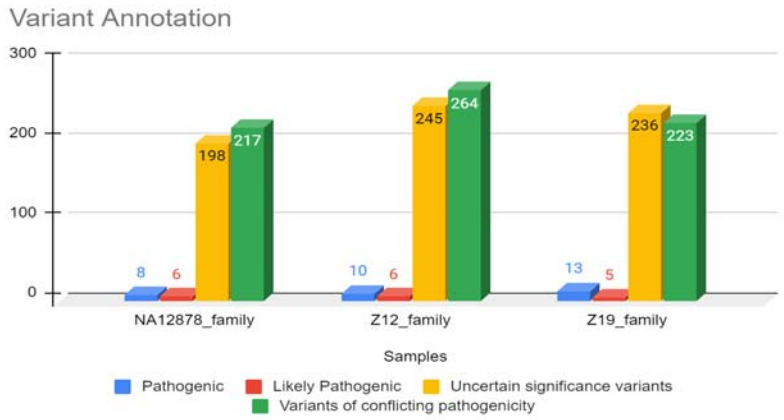


Figure 4: Count of variants after pathogenicity analysis for truth set analysis

Samples	Pathogenic	Likely Pathogenic	Uncertain significance variants	Variants of conflicting pathogenicity
NA12878_family	8	6	198	217
Z12_family	10	6	245	264
Z19_family	13	5	236	223

Table 6: Clinical significance of variants from Trio pipelines

Functional annotation of selected variants

The variants were analysed using The Genome Aggregation Database (gnomAD), gave us additional information on MAF besides Combined Annotation Dependent Depletion (CADD) scores which are key factors when analysing variants in a genome, particularly in the context of rare diseases. The accepted MAF threshold for defining rare variants in the context of rare diseases is generally considered to be less than 1% and we deemed it to identify extremely rare variants (MAF<=0.01%). Nonetheless, CADD scores are generally calculated as phred-like scores and higher scores may indicate more deleterious mutations (CADD>=10; Table 7). We couldn't consider Genomic Evolutionary Rate Profiling (GERP) scores keeping in view of the already existing stringent threshold we set.

	Chromosome Position (hg38) and Human Genome Variation Society (HGVS) Nomenclature	Ref/Alt Allele	Variant_Annotation	MAF/GnomAD	Phred_CADD	Proband	Variant type
	NC_000001.11:g.209650092G>A	G/A	nonsense	0.000006571	37	Z12	Extremely rare
T	NC_000013.11:g.31269278G>A	G/A	splice_donor_variant	0.0008022	34	Z12	Extremely rare
	NC_000004.12:g.24800212C>G	C/G	missense_variant	0.0127	23.4	Z19	Rare
A	NC_000011.10:g.94467821	G/A	nonsense	0.00008		Z19	Extremely rare

	G>A						
3	NC_000017.11:g.21300880C>T	C/T	missense_variant	NA		Z12,Z19	NA

Table 7: Functional annotation of selected trio variants

Our analysis identified novel variants in key cellular pathways related to cell proliferation, differentiation, migration, and tumorigenesis. Given the rarity of conditions like CPC (characterized by inflammation, fibrosis, and hyperplasia), we found a limited number of pathogenic and likely pathogenic variants, even though we identified several genes with a range of functional roles. For example, *SOD3* is involved in repairing oxidative damage; *MRE11A* participates in DNA double-strand break (DSB) repair; *LAMB3* plays a role in cellular organization and migration; *MAP2K3,* which regulates cell proliferation and invasion, was found to be co-expressed with other MAP kinase pathway proteins, including *MAP3K3* wherein two SNPs in the MAP3K3 gene are potential candidates for association with colon and rectal cancers. *LAMB3* is known to mediate the attachment, migration, and organization of cells into tissues during embryonic development and is also implicated in the invasive and metastatic abilities of some cancers, including colon cancer. The present study aimed to benchmark the CONVEX pipeline by comparing its performance to the widely used gold standard GATK pipeline, which comes with standard best practices guidelines. While four datasets, including three 1KGP datasets and one PCa sample, were used for evaluation, the high confidence variant call set from NA12878 was primarily used for benchmarking the CONVEX pipeline. This approach was chosen because different variant callers have varying strengths and weaknesses in detecting specific types of variants. The CONVEX pipeline integrates four open-source variant callers to improve the overall sensitivity of variant detection (Table 8).

Method/Tool	CONVEX	GATK
Type of Variants	Detects novel variants which otherwise GATK could not	Calls precise pathogenic variants
Number of variants	Calls fewer variants with higher Ti/TV	Calls more variants, relatively lower Ti/TV ratio
Sensitivity of variants	Calls high quality SNVs and differentiates true genetic variants from sequencing artefacts	Calls more SNVs with high quality

Table 8: Comparison of CONVEX and GATK on methods of calling variants.

Analysis of ClinVar validated variants showed comparable performance between the two pipelines in identifying clinically relevant variants. However, CONVEX demonstrated higher sensitivity in detecting pathogenic and likely pathogenic variants in the PCa sample, while GATK performed better in uncertain significance variants. Additionally, CONVEX showed lower numbers of variants of conflicting pathogenicity, suggesting more consistent results. To further validate the CONVEX pipeline, both pipelines were run on a widely used benchmarking dataset (NA12878 exome sequencing data) with a truth set for variant calls. The results showed comparable SNP calls between the pipelines. Benchmarking workflows, including sensitivity, precision, F1 score, and specificity, were used to assess the pipeline's performance against the truth set variants. Clinical significance annotation with the ClinVar database further aided in evaluating the pipeline's ability to call clinically relevant variants.

Our work, however, has inherent limitations. We could optimize the use of additional variant callers or incorporate new algorithms that specialise in detecting specific types of variants, such as indels or structural variants. Fine-tuning the variant filtering step within the CONVEX pipeline may also enhance its ability to differentiate true genetic variations from sequencing artefacts, thereby increasing the accuracy of variant calling. To further validate the CONVEX pipeline, it would be beneficial to run it on a diverse

range of benchmarked datasets and compare its performance against other established gold standard pipelines. This would provide a comprehensive evaluation of its strengths and weaknesses and establish its reliability across different datasets and variant calling scenarios. Incorporating machine learning and artificial intelligence techniques into the CONVEX pipeline could also enhance its performance. By leveraging these approaches, the pipeline could learn from large-scale datasets and improve its ability to accurately identify and annotate clinically relevant variants. Integrating external databases and resources beyond ClinVar, such as disease-specific databases or functional annotations, would further enrich the pipeline's clinical significance assessment. This would provide more comprehensive information for variant interpretation. However, we still have a persistent dearth of trio-exome samples in the public domain, which limits variant discovery.[24].

Conclusions

We conducted an evaluation of the CONVEX exome pipeline using four datasets. Our aim was to compare its performance to the widely used GATK pipeline, which is considered a gold standard in the field. While the CONVEX pipeline achieved a higher average Ti/Tv ratio compared to GATK indicating an improved accuracy in identifying SNV calls, the analysis suggests that CONVEX is particularly adept at distinguishing true genetic variations from sequencing artifacts, especially in exome sequencing data. Furthermore, the CONVEX pipeline demonstrated greater sensitivity in detecting pathogenic and likely pathogenic variants within the PCa sample, while GATK exhibited superior performance with VUS. We also found that the CONVEX pipeline offered efficiency advantages, requiring less computational time and expertise for replication. Its overall performance was comparable to the gold standard GATK pipeline, underscoring its reliability and effectiveness in variant analysis.

In summary, our evaluation indicates that the CONVEX pipeline is a valuable and efficient tool for variant calling. It shows promise for diverse sequencing data scenarios and for identifying clinically significant variants in relevant diseases. For the trio-exome analysis, it is important to note that GATK identified a significantly higher number of variants compared to VarScan. This difference is attributable to GATK's sensitivity-focused algorithm. To annotate the identified variants, we utilized the ClinVar database, which provides comprehensive information on known genetic variants and their disease associations. We believe that the CONVEX pipeline shows significant promise for variant calling and clinical annotation even as continued research and development will undoubtedly unlock its full potential in genomics and precision medicine. Further, exploring alternative variant calling pipelines and algorithms beyond GATK and VarScan may enhance both accuracy and efficiency.

Authors' contributions: RMR and UPS contributed equally to the work. RMR benchmarked the pipelines, wrote the first draft while UPS identified the datasets. UPS tweaked our previous WES pipeline which is now called CONVEX. PS ideated, supervised and curated the project, and proofread the manuscript.

Acknowledgements: The authors gratefully acknowledge the chancellor, revered Mata Amritanandamayi Devi and Bipin G Nair, Dean, Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham for their support. The authors thank Bhargavi R for tabulating Ti/Tv. PS sincerely thanks his collaborators, namely Krishna Mohan Medicherla, Praveen Mathur, Sonal Gupta and Obul R Bandapalli for the trio-exome project that they were associated with.

Funding: None

Conflicts of interests: The authors declare no competing interests whatsoever. PS is a Founder of Bioclues.org, India's bioinformatics society working for mentor-mentee relationships.

Data availability: Our CONVEX pipeline is available at https://github.com/prashbio/CONVEX; All the codes were run through the GATK/VarScan trio pipelines which were partially tweaked.

References

1. Goodwin, S., McPherson, J. & McCombie, W. (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351.
2. Hu T, Chitnis N, Monos D, Dinh A.(2021) Next-generation sequencing technologies: An overview. Hum Immunol. 82(11):801-811.
3. Teer JK, Mullikin JC. (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet. 2010 Oct 15;19(R2):R145-51.
4. Hansen, Mette Halskov, Haferlach, Torsten, and Nyvold, Charlotte Guldborg. (2019) A Decade with Whole Exome Sequencing in Haematology, British Journal of Haematology, 188/3,, 367–82.
5. Battke, F., Schulte, B., Schulze, M., Saskia B,.(2021) The question of WGS's clinical utility remains unanswered. Eur J Hum Genet 29, 722–723.
6. Wahjudi LW, Bernhardt S, Abnaof K, Horak P, Kreutzfeldt S, Heining C, Borgoni S, Becki C, Berg D, Richter D, Hutter B, Uhrig S, Pfütze K, Leichsenring J, Glimm H, Brors B, von Kalle C, Stenzinger A, Korf U, Fröhling S, Wiemann S. (2021) Integrating proteomics into precision oncology. Int J Cancer. 148(6):1438-1451.
7. Symonds JD, Zuberi SM, Stewart K, McLellan A, O'Regan M, MacLeod S, Jollands A, Joss S, Kirkpatrick M, Brunklaus A, Pilz DT, Shetty J, Dorris L, Abu-Arafeh I, Andrew J, Brink P, Callaghan M, Cruden J, Diver LA, Findlay C, Gardiner S, Grattan R, Lang B, MacDonnell J, McKnight J, Morrison CA, Nairn L, Slean MM, Stephen E, Webb A, Vincent A, Wilson M. (2019) Incidence and phenotypes of childhood-onset genetic epilepsies: a prospective population-based national cohort. Brain.142(8):2303-2318.
8. Acuna-Hidalgo R, Veltman JA, Hoischen A. (2016) New insights into the generation and role of de novo mutations in health and disease. Genome Biol.17(1):241.
9. Eldomery MK, Coban-Akdemir Z, Harel T, Rosenfeld JA, Gambin T, Stray-Pedersen A, Küry S, Mercier S, Lessel D, Denecke J, Wiszniewski W, Penney S, Liu P, Bi W, Lalani SR, Schaaf CP, Wangler MF, Bacino CA, Lewis RA, Potocki L, Graham BH, Belmont JW, Scaglia F, Orange JS, Jhangiani SN, Chiang T, Doddapaneni H, Hu J, Muzny DM, Xia F, Beaudet AL, Boerwinkle E, Eng CM, Plon SE, Sutton VR, Gibbs RA, Posey JE, Yang Y, Lupski JR. (2017) Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. 9(1):26.
10. Alioto, T., Buchhalter, I., Derdak, S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun 6, 10001 (2015). https://doi.org/10.1038/ncomms10001
11. Tosco-Herrera E, Muñoz-Barrera A, Jáspez D, Rubio-Rodríguez LA, Mendoza-Alvarez A, Rodriguez-Perez H, Jou J, Iñigo-Campos A, Corrales A, Ciuffreda L, Martinez-Bugallo F, Prieto-Morin C, García-Olivares V, González-Montelongo R, Lorenzo-Salazar JM, Marcelino-Rodriguez I, Flores C. (2022) Evaluation of a whole-exome sequencing pipeline and benchmarking of causal germline variant prioritizers. Hum Mutat.43(12)
12. Mathur P, Medicherla KM, Chaudhary S, Patel M, Bagali P, Suravajhala P. (2018) Whole exome sequencing reveals rare variants linked to congenital pouch colon. Sci Rep.8(1):6646.
13. Gupta S, Gupta N, Tiwari P, Menon S, Mathur P, Kothari SL, Nallapeta S, Medicherla KM, Suravajhala P .(2018) Lnc-EPB41-Protein Interactions Associated with Congenital Pouch Colon. Biomolecules.8(3):95.
14. Gupta S, Mathur P, Mishra AK, Medicherla KM, Bandapalli OR, Suravajhala P. (2023) Whole Exome-Trio Analysis Reveals Rare Variants Associated with Congenital Pouch Colon. Children (Basel).10(5):902.
15. Gupta A, Shukla N, Nehra M, Gupta S, Malik B, Mishra AK, Vijay M, Batra J, Lohiya NK, Sharma D, Suravajhala P. (2020) A Pilot Study on the Whole Exome Sequencing of Prostate Cancer in the Indian Phenotype Reveals Distinct Polymorphisms. Front Genet. 11:874.
16. Cunningham, Fiona, Amode, M Ridwan, Barrell, Daniel, Beal, Kathryn, Billis, Konstantinos, Brent, Simon, et al., Ensembl 2015, Nucleic Acids Research, 43/D1 (2014), D662–69

17. Meena, N., Mathur, P., Medicherla, K. M. and Suravajhala, P. (2018). A Bioinformatics Pipeline for Whole Exome Sequencing: Overview of the Processing and Steps from Raw Data to Downstream Analysis. *Bio-101*: e2805.
18. Langmead, B., Salzberg, S. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359
19. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST; (2011) Working Group of the American College of Medical Genetics Laboratory Quality Assurance Committee. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med. 13(7):680-5.
20. 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; (2015) 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. 17(5):405-24.
21. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 43(1110):11.10.1-11.10.33.
22. Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics.30(15):2114-20.
23. Li H, Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics.;26(5):589-95.
24. Barbitoff YA, Abasov R, Tvorogova VE, Glotov AS, Predeus AV. (2022) Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery. BMC Genomics.;23(1):155.
25. Chen, Z., Yuan, Y., Chen, X. et al. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. Sci Rep 10, 3501 (2020). https://doi.org/10.1038/s41598-020-60559-5
26. Jin J, Chen Z, Liu J, Du H, Zhang G. Towards an accurate and robust analysis pipeline for somatic mutation calling. Front Genet. 2022 Nov 15;13:979928. doi: 10.3389/fgene.2022.979928.
27. Dwarshuis N, Kalra D, McDaniel J, Sanio P, Alvarez Jerez P, Jadhav B, Huang WE, Mondal R, Busby B, Olson ND, Sedlazeck FJ, Wagner J, Majidian S, Zook JM. The GIAB genomic stratifications resource for human reference genomes. Nat Commun. 2024 Oct 19;15(1):9029. doi: 10.1038/s41467-024-53260-y.
28. Menzel, M., Martis-Thiele, M., Goldschmid, H., et al. 2024, Benchmarking whole exome sequencing in the German network for personalized medicine. Eur J Cancer, 211, 114306.
29. Fang, L. T., Zhu, B., Zhao, Y., et al. 2021, Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat Biotechnol, 39, 1151-1160.
30. Xiao, W., Ren, L., Chen, Z., et al. 2021, Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nature Biotechnology, 39, 1141-1150.
31. Kumaran, M., Subramanian, U. and Devarajan, B. 2019, Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics, 20, 342.

32. Chen, J., Li, X., Zhong, H., Meng, Y. and Du, H. 2019, Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep, 9, 9345.
33. Ahmed, Z., Renart, E. G. and Zeeshan, S. 2021, Genomics pipelines to investigate susceptibility in whole genome and exome sequenced data for variant discovery, annotation, prediction and genotyping. PeerJ, 9, e11724.