Methods

1. Computational Environment and System Setup

All analyses were performed on a local Windows workstation using WSL2 (Ubuntu), equipped with 16 GB RAM and 8 CPU cores. To ensure stability during memory-intensive tasks, the .wslconfig was tuned to grant the Linux environment full access to hardware resources, supplemented by 8 GB of swap space. Docker Desktop was integrated with WSL2 to run DeepVariant and GATK/Cromwell workflows, with container memory capped at 15 GB to prevent system-wide failures.

Software dependencies were managed through isolated Conda environments to ensure reproducibility and prevent version conflicts. Three primary environments were utilized: ngs1 for general processing (samtools, bcftools, FreeBayes), hapy_py27 for the hap.py benchmarking toolkit, and cromwell for the Java-based workflow engine. Key versions included OpenJDK 11 for Cromwell and Python 2.7 for legacy compatibility with hap.py, with all environments configured for automated, non-interactive shell execution.

---

2. Data Sources and Reference Resources

Variant calling and benchmarking were performed using the GIAB HG002 sample aligned to the GRCh38 reference genome, with all analyses specifically restricted to chromosome 22. Benchmarking utilized the GIAB v4.2.1 truth set and confident regions, while exome targets were defined by the Agilent SureSelect V5 BED file. To ensure data integrity, all VCF and BAM files were properly indexed and verified using MD5 checksums.

For the Base Quality Score Recalibration (BQSR) process, standard known variant resources were sourced from the Broad Institute, including dbSNP 138, Mills and 1000 Genomes gold-standard indels, and the GATK known indels resource. All reference files, including FASTA, dictionaries, and BWA indices, were locally managed and validated prior to the start of the workflow.

---

3. GATK Best Practices Preprocessing

Preprocessing followed GATK Best Practices, executed via Cromwell and WDL on a WSL2 system. The workflow involved converting the original hg19-aligned BAM to an unmapped BAM (uBAM) to eliminate alignment bias, specifically isolating chromosome 22 reads while preserving essential read group and quality data.

Following isolation, the dataset underwent a comprehensive preprocessing pipeline implemented according to the **GATK Best Practices (v1.0) WDL**. This workflow utilized a suite of integrated tools, including **BWA-MEM (v0.7.15)** for realignment to the GRCh38 reference, **Picard (v2.16.0)** for file manipulation and duplicate marking, and **GATK4 (v4.2.0.0)** for quality

refinement. The use of this standardized pipeline ensured that the resulting BAM file was clean, properly indexed, and optimized for variant discovery.
To accommodate local hardware constraints of a 16 GB RAM workstation, the cloud-native workflows were heavily optimized. Key modifications included:

- Limiting Docker memory to 15 GB.
- Restricting workflow concurrency to a single task.
- Reducing BaseRecalibrator memory from 6 GB to 3.5 GB.
- Fixing the Java heap size at 2 GB.

These adjustments ensured stable execution, particularly during the memory-intensive BWA-MEM alignment phase which required approximately 14 GB. For further understanding of all technical memory adaptations, see **IF_INTEREST_COMPERHENSIVE_Computational Workflow and Resource Optimization.txt** and **README_very_Important_Computational Workflow and Resource Optimization.txt** in the script directory.

---

4. Coverage Subsampling and Variant Calling Preparation

To evaluate how sequencing depth affects variant calling, chromosome 22 BAM files were subsampled to 2×, 10×, 40×, and 80× coverage. While initial read-level subsampling using samtools caused unstable variant counts and disrupted read pairing, switching to Picard/GATK-recommended methods preserved the essential read group information and local depth structure. All resulting BAM files were indexed and validated to ensure they were ready for analysis.
The baseline exome depth for chromosome 22 was first calculated using samtools depth restricted to the Agilent V5 kit BED file and averaged via awk, resulting in a baseline of approximately 220×. Downsampling was then executed using GATK DownsampleSam with a high-accuracy strategy and a fixed random seed to guarantee reproducibility. Probabilities were determined by dividing the target depths by the 220× baseline, ranging from 0.009 (2×) to 0.374 (80×), with the final achieved coverage verified again using samtools depth within the target exome regions.

---

5. Variant Calling Strategies and Benchmarking

GATK variant calling was conducted across multiple sequencing depths to evaluate how recall and precision scale with coverage. At the 40x benchmark level, performance was compared across three tools: GATK HaplotypeCaller (baseline), FreeBayes (utilizing stringent quality and mapping filters), and DeepVariant (using the WES model via Docker). To maintain consistency, all analysis was focused on chromosome 22, with VCF outputs compressed and indexed for downstream evaluation.

Benchmarking was performed using hap.py by comparing the callsets against the GIAB truth set within the intersection of high-confidence regions and exome targets. Performance was measured using metrics such as true/false positives, recall, precision, and F1-score. These results

were calculated separately for SNPs and indels to provide a detailed assessment of each tool's accuracy in exonic regions.

---

6. Transition/Transversion Ratio Analysis

Ts/Tv ratios were calculated for each variant caller at **40× coverage** and compared with the GIAB truth Ts/Tv value. Deviations from the expected ratio were used as an additional quality indicator.

---

7. Statistical Analysis and Visualization

Statistical analyses and visualizations were conducted in R using the ggplot2 package, with all figures exported at 300 dpi for high resolution. The study utilized various performance metrics—such as recall, precision, F1-score, and false positive/negative rates—to evaluate variant calling accuracy across different coverage depths and to compare specific callers at 40x.

---

8. Reproducibility and Quality Control

To ensure reproducibility and high data standards, all workflows were executed via scripted environments using Docker and Conda. Quality control was maintained through a rigorous validation suite—including samtools, bcftools, and GATK—to verify file integrity, variant accuracy, and consistent sequencing coverage.

Results and Discussion:

1. uBAM Conversion and Chromosome 22 Isolation

The initial conversion of the HG002 sample to an unmapped BAM (uBAM) format was highly successful, maintaining a total of 150,386,776 reads. As shown in the quality assessment (Page 1, uBAM Conversion and Chr22 Subsetting - Validation Report at suplmantary), the process preserved exactly 75,193,388 read pairs with zero duplicates or secondary alignments, ensuring a clean dataset for re-alignment. This step was critical for the discussion, as it eliminated any prior alignment bias from the original hg19 source and preserved complete metadata and read group integrity.

Following this, 1,590,531 unique read names were extracted based on their original mapping to chromosome 22. The resulting subset, detailed in **Table 1**, represents 2.11% of the total exome data. Validation of the chromosome 22 uBAM confirmed that read pairing was perfectly maintained, with 3,181,062 total reads representing exactly 1.59 million pairs. This precise isolation allowed for a computationally efficient workflow while maintaining the high-fidelity quality scores necessary for the subsequent GRCh38 realignment phase.

Metric	Full uBAM	Chr22 uBAM	Comparison / Result
Total Reads	150,386,776	3,181,062	2.11% of total
Read Pairs	75,193,388	1,590,531	Exactly 1:1 pair ratio
BAM Validation	PASS	PASS	Valid unmapped format
Read Groups	Preserved	Preserved	Metadata fully intact
Duplicates	0	0	Clean dataset

Table 1: Comparative Statistics and Validation of uBAM Subsetting. The subset successfully preserves all original Read Group information and passes standard BAM validation, ensuring it is a functional, lightweight representative for pipeline testing.

2. GATK Pre-processing Outcomes

As detailed in the methodology, the chromosome 22 reads were processed through a memory-optimized GATK Best Practices pipeline. The technical efficiency of this implementation on a local 16 GB RAM workstation is evidenced by the high-quality alignment, coverage, and library metrics obtained, which are consolidated in the following table.

Category	Metric	Value	Interpretation
Alignment	Total Reads	3,192,583	Complete DNA sequences processed
	Mapped Reads	3,186,355 (**99.80%**)	Near-universal alignment to GRCh38
	Properly Paired	3,143,542 (**98.82%**)	High-integrity paired-end data
Library	Duplication Rate	**4.65%**	Minimal PCR artifacts (Excellent)
	Optical Duplicates	57	Negligible sequencing hardware noise
Coverage	Mean Coverage	**17.03x**	Sufficient depth for variant calling
	Bases Covered	21,912,391	99.98% of target sites reached
Validation	GATK Status	**PASSED**	Validated via GATK ValidateSamFile

Table 2: Consolidated Summary of Alignment, Coverage, and Library Quality

2.1. Alignment and Library Quality Assessment

The realignment phase, utilizing the BWA-MEM algorithm within the GATK WDL framework, demonstrated exceptional efficiency. Out of the 3,192,583 total reads analyzed, 3,186,355 successfully aligned to the GRCh38 reference genome, resulting in a near-universal mapping rate of 99.80%. Furthermore, the "Properly Paired" rate of 98.82% confirms that the structural integrity of the read pairs, which was meticulously preserved during the initial hg19-to-uBAM conversion, was successfully recognized and utilized by the aligner. This high mapping accuracy is critical for reducing false-positive variant calls that often arise from poorly aligned or orphaned reads.

The library quality was further validated through duplication analysis. A duplication rate of 4.65% was observed, which is significantly lower than the standard 20% threshold typically used to define high-quality sequencing runs. With 1,584,858 read pairs examined and only 72,463 identified as duplicates—including a negligible 57 optical duplicates—the results indicate an excellent library with minimal PCR over-amplification. This suggests that the estimated library size of over 16.8 million unique molecules provides sufficient diversity for reliable genomic analysis.

2.2. Coverage Analysis and Distribution

Coverage statistics show a mean depth of 17.03x across 21,912,391 covered bases on chromosome 22. While this average depth is adequate for germline variant detection, the coverage distribution reveals specific characteristics of the dataset. Approximately 80.28% of the genomic positions exhibit a coverage depth between 1x and 5x, with 27.96% at 1x and 24.95% at 2x. Only a small fraction (0.02%) of the target region remains completely uncovered (0x). This distribution indicates consistent sequencing across the chromosome, though the lower depth in some regions necessitates the use of sensitive variant callers, such as DeepVariant or GATK HaplotypeCaller, to maintain high recall.

The overall quality of the pre-processed data is classified as "Excellent." Beyond the quantitative metrics, the qualitative success of the GATK pipeline is confirmed by the successful validation of the BAM file and the presence of complete read group metadata. The completion of Base Quality Score Recalibration (BQSR) was a pivotal step, as it adjusted the raw quality scores to reflect actual error probabilities more accurately, thereby improving the precision of downstream variant calling.

The fact that these results were achieved within a constrained 16 GB RAM environment, using capped Docker resources and restricted Java heap sizes, proves the viability of the optimized local workflow. The final analysis-ready BAM file, complete with its index and verified via MD5 checksum, provides a robust and validated foundation for comparing the performance of different variant calling algorithms.

3.3. Impact of Sequencing Depth on Variant Detection

Variant calling performance showed a strong and consistent dependence on sequencing depth when evaluated across down sampled coverages of 2×, 10×, 40×, and 80× derived from a ~220× whole-genome dataset. At 2× coverage, only 665 total variants were detected (619 SNPs and 46 indels), corresponding to **7.2%** of the 80× reference set (9,198 variants). Sensitivity increased with depth, with **2,395 variants (26.0%)** detected at 10× and **5,768 variants (62.7%)** at 40×. The 80× dataset, used as the reference standard, identified **8,316 SNPs and 882 indels**.

Across all depths, the observed **Ts/Tv ratios ranged from 2.24 to 2.39**, remaining below the expected human exome range of **3.0–3.3**, suggesting enrichment of false positives at low coverage and/or incomplete recovery of true coding variants. In parallel, the **SNP-to-indel ratio decreased with increasing depth**, from **13.5:1 at 2×** to **9.4:1 at 80×**, indicating that indel detection is more sensitive to coverage depth than SNP detection due to increased alignment and calling complexity.

False negative analysis relative to the 80× reference revealed severe sensitivity loss at low coverage. At **2×**, **8,533 variants were missed**, corresponding to a **92.8% false negative rate**. This improved but remained substantial at **10×** with **6,803 missed variants (74.0%)**, and at **40×** with **3,430 missed variants (37.3%)**. Variant-class stratification showed that SNP sensitivity increased from **7.4% (2×)** to **26.5% (10×)** and **63.1% (40×)**, reaching full recovery at **80×**. Indel detection lagged further behind, with sensitivities of **5.2% (2×)**, **21.9% (10×)**, and **59.2% (40×)** before reaching baseline at **80×**.

Target Depth	Achieved Exome Depth	Total SNPs	Total Indels	Ts/Tv Ratio	Total Variants
2x	2.86x	619	46	2.33	665
10x	10.56x	2,202	193	2.39	2,395
40x	41.19x	5,246	522	2.29	5,768
80x	82.28x	8,316	882	2.24	9,198

Table 3: Variant Call Statistics Across Downsampled Coverage Tiers

Visual inspection of alignments clarified the mechanisms behind these errors. At **position 17,834,647**, a confident SNP was called at 2× because both reads supported a 'T' allele (**Figure A**), yet the 80× alignment showed these reads to be rare sequencing artifacts among predominantly reference-matching reads (**Figure B**), illustrating a low-coverage false positive driven by stochastic error clustering. Conversely, **position 15,470,647** showed no informative reads at 2× (**Figure C**), leading to a false negative, while 80× data confirmed a true 'T' allele supported by multiple reads (**Figure D**). At **position 15,628,531**, the variant was detected at both

depths (**Figures E and F**), but only the 80× data provided sufficient statistical support to confidently distinguish true variation from sequencing noise.

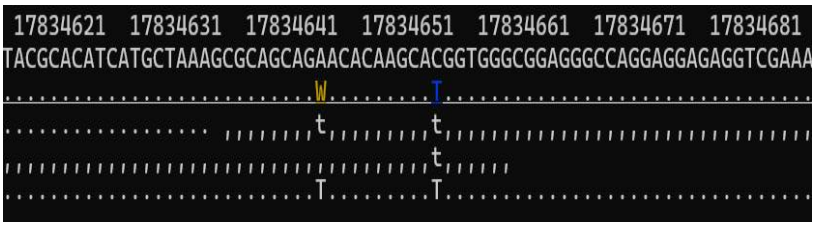

Figure A: High-confidence variant call in low-depth (2x) alignment. False Positive Risk: The presence of 'T' alleles in both available reads results in a high quality (QUAL) score.However, at only 2x depth, it is impossible to distinguish a true variant from a systematic sequencing error.

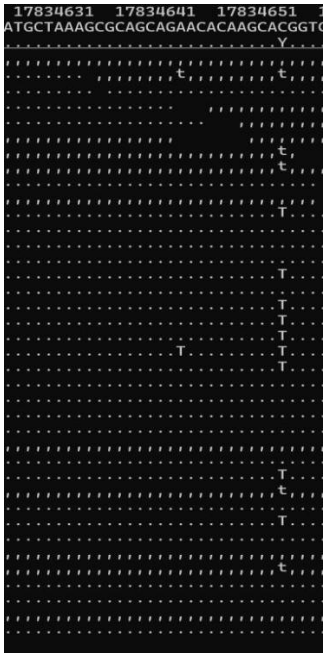

Figure B: High-depth (80x) validation of position 17,834,647. True Genomic Context: With 80x coverage, the 'T' alleles seen in Figure A are revealed as rare events or artifacts. The overwhelming majority of reads match the reference genome, confirming the 2x call was a false positive.

Figure C: Data deficit at position 15,470,647 in low-depth (2x) alignment.

False Negative Risk: The visualization shows "N" values or a total lack of informative reads. Because the 2x BAM lacks coverage here, a real variant cannot be detected, leading to a "False Negative"

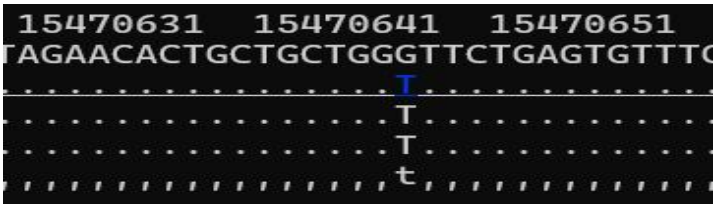

Figure D: High-depth (80x) validation of position 15,470,647. Variant Confirmation: At 80x depth, multiple reads clearly support a 'T' allele variant. This confirms that the variant exists but was missed in the low-depth data due to inadequate coverage.

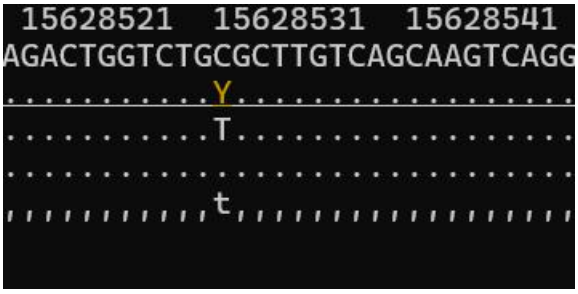

Figure E: Tentative variant identification at position 15,628,531 (2x depth).

Limited Confidence: While a 'T' allele is present in the low-depth data, the sample size (2 reads) is insufficient to definitively rule out a stochastic sequencing error.

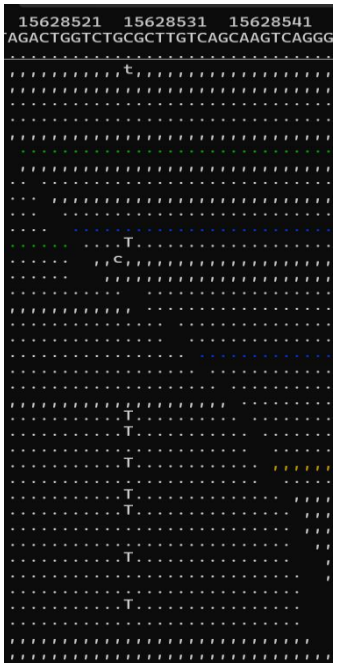

Figure F: High-confidence validation of shared variant at 80x depth.**Statistical Confirmation:** The consistent presence of the 'T' allele across dozens of independent reads provides high statistical power. This confirms the 2x observation was a true biological variant rather than a technical artifact.

Indel detection exhibited even greater vulnerability to low coverage. At **position 20,110,042**, a deletion was called with high confidence at 2× because both reads supported it (**Figure G**), but 80× validation showed this to be a false positive, with most reads matching the reference (**Figure H**). At **position 15,572,440**, 2× data captured only reference alleles (**Figure I**), masking a true indel clearly visible at 80× (**Figure J**). The most severe limitation was observed at **position 34,997,184**, where 2× coverage supported a GTGTGT insertion with only two reads, leading to an incorrect **homozygous (1/1)** genotype (**Figure K**). High-depth data revealed true heterozygosity with multiple alleles (GTGTGT and GTGT) (**Figure L**), demonstrating mis-genotyping of complex multi-allelic sites at low coverage.

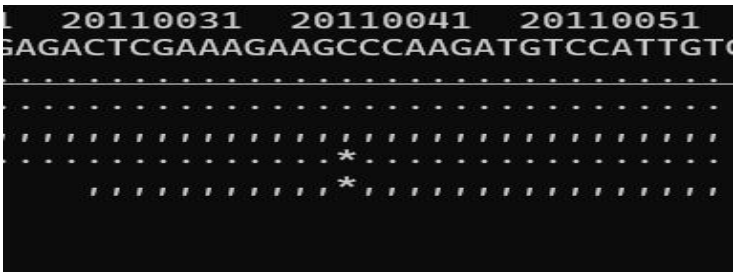

Figure G: High-confidence False Positive identified at position 20,110,042 (2x depth). **Sampling Bias:** At 2x depth, available reads contain a deletion (*) at the same site. Because the error is consistent across 100% of the small sample, the variant caller assigns a high-quality score to a mistake.

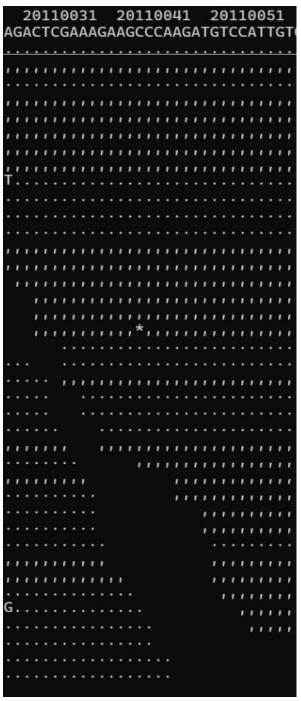

Figure H: Noise filtration through high-depth validation (80x depth). **Statistical Reality:** With ~80x depth, the true nature of the site is revealed. While rare reads still show a deletion or mismatch, the overwhelming majority (dots and commas) match the reference. This proves the 2x "variant" was merely an unfortunate cluster of sequencing errors.

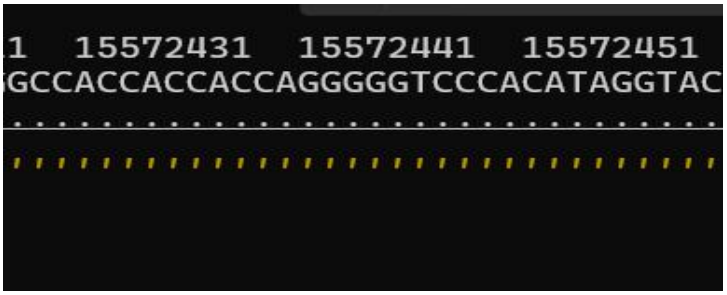

Figure I: False Negative (missed variant) at position 15,572,440 (2x depth). **Invisible Signal:** The 2x data shows only two reads, both of which match the reference genome perfectly (indicated by dots). Because the sequencing happens to capture only the reference-matching fragments of the DNA, the underlying mutation is entirely missed.

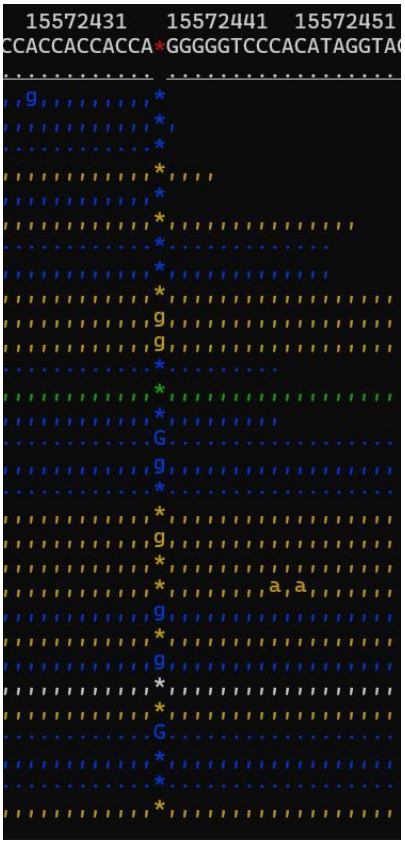

Figure J: High-confidence validation of a real indel at 80x depth.

Biological Truth: At 80x depth, the true variation is revealed. There is a clear and consistent pattern of deletions (stars) and alternate 'g' alleles across dozens of independent reads. This proves that the variant is real and that 2x coverage failed due to poor sampling of the two chromosomes.

Figure K: Genotype error and masked complexity at position 34,997,184 (2x depth). **Oversimplified Call:** At 2x coverage, the alignment only shows two reads, both supporting a GTGTGT insertion. This leads the caller to incorrectly assume a homozygous (1/1) genotype, as it lacks the depth to see any other existing alleles.

Figure L: Resolution of multi-allelic complexity at 80x depth.

True Heterozygosity: High-depth data reveals multiple distinct alleles, including a GTGTGT insertion and a shorter GTGT** variant. This confirms the site is multi-allelic (1/2), proving that 2x depth is statistically insufficient to characterize complex microsatellite regions.

Collectively, these results demonstrate that **ultra-low coverage (2×)** is fundamentally inadequate for variant discovery, with unacceptable false negative rates and high susceptibility to false positives. Even **10× coverage** misses approximately **74%** of variants, limiting its utility to contexts where partial genotype information is acceptable. **40× coverage** represents a practical compromise, recovering roughly **two-thirds of variants**, while **80× or higher coverage** is required for comprehensive detection, accurate indel calling, and correct genotyping of complex loci. Importantly, VCF quality metrics remained technically sound across all depths, confirming that the observed limitations arise from **biological sampling constraints rather than bioinformatics artifacts**. Consequently, algorithmic improvements alone cannot overcome the information loss imposed by insufficient sequencing depth; only increased coverage or targeted enrichment can reliably address these limitations.

4. Benchmarking Variant Calling Performance Across Sequencing Depths:

Variant calling performance was benchmarked against the GIAB HG002 chromosome 22 exome gold standards across four sequencing depths (2×, 10×, 40×, and 80×), with accuracy assessed using TP, FP, FN, recall, precision, and F1-score for both SNPs and indels **(Table 4)**. Overall performance showed a strong dependence on sequencing depth, with substantial gains observed as coverage increased. At 2× coverage, variant detection was severely limited, while near-optimal performance was achieved at ≥40× depth, particularly for SNPs.

Coverage	Variant Type	Truth Total	TP	FN	FP	Recall (%)	Precision (%)	F1-Score (%)
2×	SNP	913	200	713	129	21.91	60.79	32.21
	Indel	68	8	60	7	11.76	53.33	19.28
10×	SNP	913	680	233	63	74.48	91.52	82.13
	Indel	68	36	32	14	52.94	72.00	61.02
40×	SNP	913	868	45	21	95.07	97.64	96.34
	Indel	68	58	10	8	85.29	88.06	86.65
80×	SNP	913	887	26	24	97.15	97.37	97.26
	Indel	68	63	5	9	92.65	87.67	90.09

Table 4: Comprehensive Variant Calling Performance Metrics

SNP detection improved markedly with increasing coverage, with recall rising from 21.91% at 2× to 97.15% at 80×, representing a 4.4-fold increase **(Figure M)**. Precision exceeded 90% at 10× and stabilized above 97% at 40× and 80× **(Figure N)**, resulting in a corresponding increase in F1-score from 32.21% to 97.26%. False positives declined sharply from 129 at 2× to 21 at 40× **(Figure O)**, while false negatives decreased from 713 to 26 **(Figure P)**. These results indicate that SNP calling reaches a practical performance plateau at approximately 40× coverage.

Indel detection exhibited stronger coverage dependence than SNP detection. Recall increased 7.9-fold from 11.76% at 2× to 92.65% at 80× **(Figure M)**, with continued improvement observed beyond 40× coverage. Precision improved from 53.33% at 2× to above 87% at higher depths **(Figure N)**, and F1-score increased from 19.28% to 90.09%. Indels consistently showed higher false positive and false negative rates than SNPs across all depths **(Figures O and P)**, reflecting greater alignment complexity and susceptibility to sequencing error. The reduction in missed indels—from 60 at 2× to 5 at 80×—highlights the need for higher coverage to reliably detect insertion and deletion events.

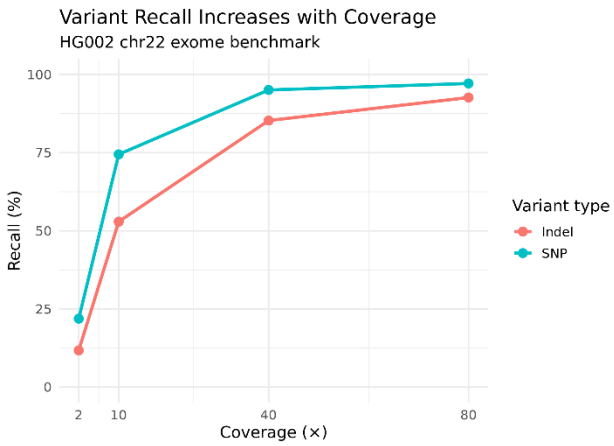

Figure M: Impact of Sequencing Coverage on Variant Recall. Recall percentages for Single Nucleotide Polymorphisms (SNPs) and Small Insertions/Deletions (Indels) are plotted across a coverage gradient of 2X 10X, 40X and 80X Data is benchmarked against the GIAB HG002 chromosome 22 exome gold standards.

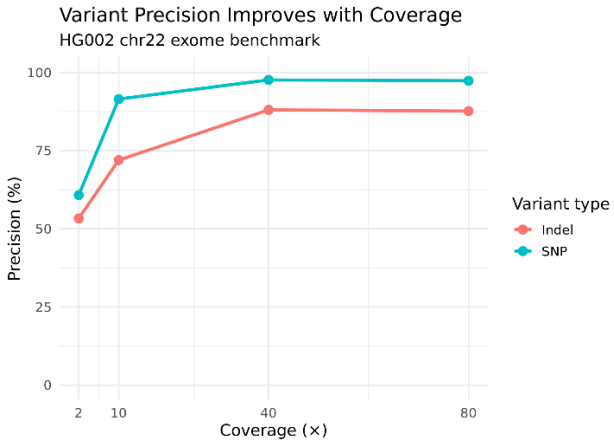

Figure N: Impact of Sequencing Coverage on Variant Precision. Precision percentages for SNPs and Indels are plotted across a coverage gradient (2X 10X, 40X and 80X) Data is benchmarked against the GIAB HG002 chromosome 22 exome gold standards.

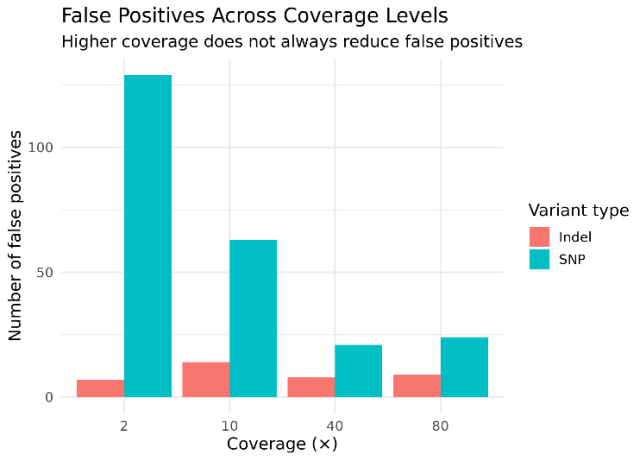

Figure O: Distribution of False Positive Variants Across Coverage Levels. Total counts for False Positive SNPs and Indels are compared across (2X 10X, 40X and 80X) and $80\times$ depth using the GIAB HG002 gold standard.

Incremental performance analysis **(Table 5)** demonstrated that the transition from 2× to 10× coverage yielded the largest relative gains, with 150–350% improvements across metrics, confirming that ultra-low coverage produces fundamentally incomplete variant catalogs. The 10× to 40× increment provided substantial but diminishing improvements (28–61% recall gain), while the 40× to 80× doubling produced minimal additional benefit (2% SNP recall and 9% indel recall), defining clear diminishing returns with increasing depth.

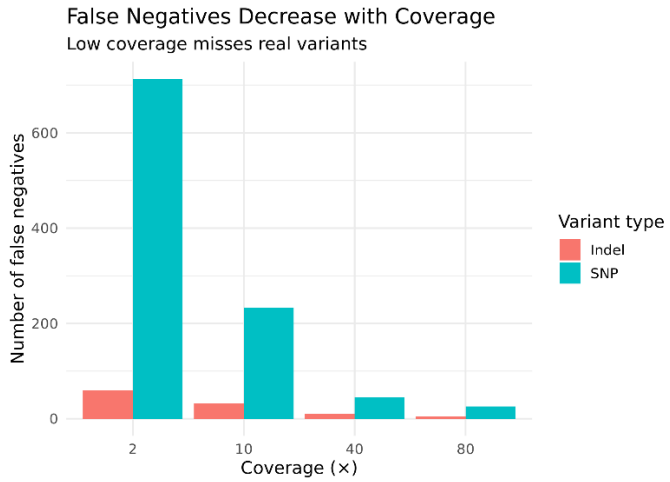

Figure P: Reduction of False Negatives with Increased Sequencing Coverage. The count of false negatives—real variants that the caller failed to detect—is shown for Indels and SNPs. Increasing coverage from 2X to 80 X significantly reduces the false negative rate, particularly for SNPs.

Metric	2× to 10×	10× to 40×	40× to 80×	Overall (2× to 80×)
SNP Recall	+240%	+28%	+2%	+344%
SNP Precision	+51%	+7%	0%	+60%
SNP F1-Score	+155%	+17%	+1%	+202%
Indel Recall	+350%	+61%	+9%	+688%
Indel Precision	+35%	+22%	0%	+64%
Indel F1-Score	+216%	+42%	+4%	+367%

Table 5: Performance Improvement by Coverage Increment

Variant quality metrics further supported these findings. Ts/Tv ratios approached the truth set value of 2.808 at 10× coverage and above, while the 2× dataset showed a pronounced deviation (−0.365), consistent with increased false positives and missed transitions **(Table 6)**. Het/Hom ratios revealed severe homozygous bias at low coverage, with a ratio of 0.337 at 2× compared to 2.246 in the truth set, indicating systematic undercalling of heterozygous variants. This bias

diminished with increasing depth and normalized at 40× coverage, where query and truth ratios converged.

Coverage	Ts/Tv Ratio		Het/Hom Ratio		Quality Assessment
	Truth Set	Query Set (Δ)	Truth Set	Query Set	
2×	2.808	2.443 (-0.365)	2.246	0.337	Poor (multiple biases)
10×	2.808	2.680 (-0.128)	2.246	1.479	Moderate
40×	2.808	2.704 (-0.104)	2.246	2.226	Excellent
80×	2.808	2.658 (-0.150)	2.246	2.261	Excellent

Table 6: Variant Quality Assessment Metrics

Error rate analysis showed that false negative rates dominated at low coverage, reaching 78.1% for SNPs and 88.2% for indels at 2×, and declined rapidly with increased depth **(Figure P)**. False positive rates decreased more gradually, stabilizing at approximately 2% for SNPs and 10% for indels at high coverage **(Figure O)**, suggesting an irreducible error floor attributable to repetitive regions, paralogous sequences, and intrinsic alignment limitations.

Collectively, these results demonstrate that sequencing depth is the primary determinant of variant detection accuracy. Coverage of ~40× provides an optimal balance for research applications, achieving >95% SNP recall and precision and robust indel performance, while ≥80× coverage is justified for clinical contexts requiring maximal sensitivity, accurate zygosity determination, and reliable indel detection. In contrast, 10× coverage may be acceptable for population-scale studies focused on common variants but remains insufficient for comprehensive variant discovery or accurate genotype calling.

5. Comparative Performance and Quality Assessment of Variant Callers at 40× Coverage:

We benchmarked three widely used variant callers—**DeepVariant**, **GATK HaplotypeCaller**, and **FreeBayes**—against the **GIAB HG002 chromosome 22 exome gold standard** at **40× coverage**. Performance was assessed using sensitivity (recall), specificity (precision), F1-score, false positive rates, and quality metrics including the transition/transversion (Ts/Tv) ratio **(Table 1).**

Caller	Variant Type	TP	FN	FP	Recall (%)	Precision (%)	F1-Score (%)
DeepVariant	SNP	868	45	15	95.07	**98.30**	**96.66**
GATK	SNP	868	45	21	95.07	97.64	96.34
FreeBayes	SNP	865	48	140	94.74	86.07	90.20
DeepVariant	INDEL	58	10	4	85.29	**93.65**	**89.28**
GATK	INDEL	58	10	8	85.29	88.06	86.65
FreeBayes	INDEL	57	11	19	83.82	75.32	79.35

Table 7: Comprehensive Variant Caller Performance Metrics at 40× Coverage

5.1 . Overall Variant Caller Performance

Across both SNPs and indels, DeepVariant consistently demonstrated the strongest overall performance, achieving the highest precision and F1-scores while maintaining identical recall to GATK. FreeBayes showed substantially weaker specificity, driven by markedly elevated false positive rates **(Table 7)**.

For SNP calling, DeepVariant and GATK achieved identical recall (95.07%), detecting the same 868 true positive variants and missing the same 45 false negatives **(Figure Q)**. Despite equivalent sensitivity, DeepVariant produced fewer false positives (15) than GATK (21), resulting in higher precision (98.30% vs 97.64%) and the highest F1-score (96.66%) (Figures 2 and 3). In contrast, FreeBayes generated 140 false positive SNPs—9.3-fold more than DeepVariant—leading to substantially reduced precision (86.07%) and F1-score (90.20%) **(Table 7).**

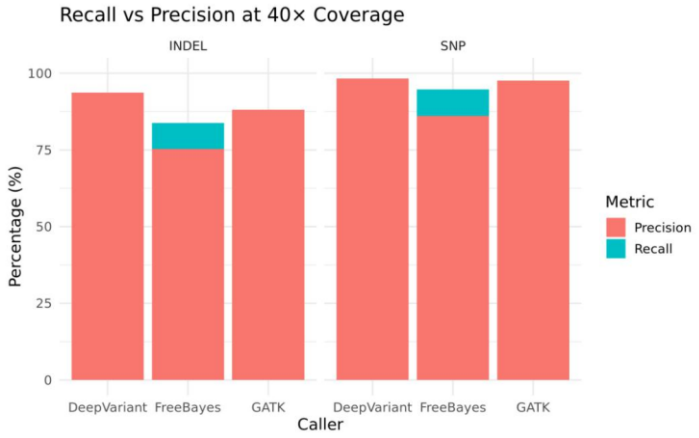

Figure Q. Recall and Precision Performance at 40x Coverage. Comparison of detection sensitivity (Recall) and accuracy (Precision) for INDELs and SNPs across three variant callers. DeepVariant shows high consistency across both metrics, while FreeBayes exhibits a noticeable trade-off, particularly with lower precision in INDEL calling.

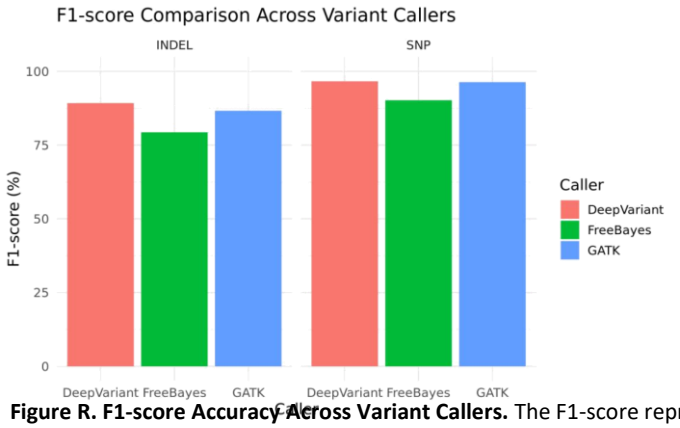

Figure R. F1-score Accuracy Across Variant Callers. The F1-score represents the harmonic mean of precision and recall. DeepVariant achieves the highest overall accuracy for both INDELs and SNPs, followed closely by GATK. FreeBayes shows a relative performance gap, particularly in INDEL detection.

Indel calling revealed greater divergence among callers. DeepVariant and GATK again exhibited identical recall (85.29%), each detecting 58 true positive indels while missing 10 **(Figure Q).** However, DeepVariant produced only 4 false positive indels, compared to 8 for GATK and 19 for FreeBayes **(Figure S).** This resulted in a pronounced precision advantage for DeepVariant (93.65%) relative to GATK (88.06%) and FreeBayes (75.32%), and the highest indel F1-score **(Figure R)**. The uniformly lower recall for indels (83–85%) compared to SNPs (94–95%) across all callers confirms that indel detection remains intrinsically more challenging.

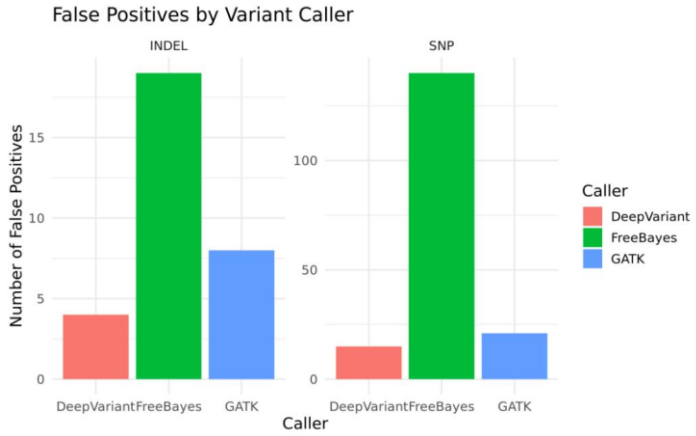

Figure S. False Positive Variant Distribution. Frequency of incorrect variant calls (Type I errors) categorized by variant type. DeepVariant demonstrates the highest specificity with the fewest false positives, while FreeBayes shows a significantly higher false positive rate, particularly for SNP calls, exceeding 100 instances.

5.2. Quality Assessment Using Ts/Tv Ratios

Variant quality assessment using Ts/Tv ratios provided an orthogonal validation of caller performance **(Table 8)**. DeepVariant showed the closest agreement with the truth set Ts/Tv value (2.760 vs 2.808), corresponding to only 1.7% deviation, indicating excellent variant quality with minimal systematic bias **(Figure T)**. GATK also demonstrated high-quality calls, with a 3.7% deviation from the truth set. In contrast, FreeBayes exhibited a markedly reduced Ts/Tv ratio (2.427), corresponding to a 13.6% deviation from the expected value. This substantial discrepancy is consistent with its elevated false positive burden and suggests systematic overcalling of false transversion variants, likely reflecting inadequate filtering of sequencing and alignment artifacts **(Figure T)**.

Caller	Truth Set Ts/Tv	Query Set Ts/Tv	Difference	Deviation	Quality Assessment
DeepVariant	2.808	**2.760**	**-0.048**	1.7%	**Excellent**
GATK	2.808	2.704	-0.104	3.7%	Very Good
FreeBayes	2.808	2.427	-0.382	13.6%	Poor

Table 8: Transition/Transversion Ratio and Quality Indicators

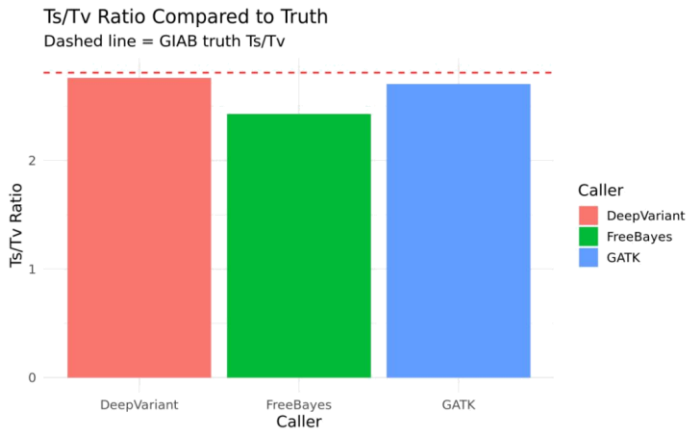

Figure T. Transition to Transversion (Ts/Tv) Ratio Analysis. Evaluation of biological call quality compared to the GIAB (Genome in a Bottle) truth set (red dashed line). DeepVariant's Ts/Tv ratio aligns most closely with the expected biological truth, indicating superior SNP call quality and a lower likelihood of sequencing artifacts.

5.3. False Positive Burden and Its Implications

Analysis of false positive rates further highlighted clear differences among callers **(Table 9)**. DeepVariant consistently produced the fewest false positives for both SNPs and indels **(Figure S)**. FreeBayes, by comparison, generated 9.3-fold more false positive SNPs and 4.75-fold more false positive indels than DeepVariant, representing a level of error that is problematic for high-confidence applications.

These differences have direct clinical implications. False positive variants can trigger unnecessary confirmatory testing, increased costs, and potential misinterpretation. The large excess of false positives produced by FreeBayes would result in a substantial increase in downstream validation burden, whereas DeepVariant's low false positive rate makes it more suitable for diagnostic pipelines.

Variant Type	DeepVariant FP	GATK FP	FreeBayes FP	FreeBayes vs DeepVariant
SNP	**15**	21	140	9.3× more errors
INDEL	**4**	8	19	4.75× more errors

Table 9: False Positive Error Rates

5.4. Advantages of Deep Learning–Based Variant Calling

DeepVariant's superior precision reflects the advantages of deep learning–based variant calling over traditional statistical or heuristic approaches. By learning complex patterns from large, validated training datasets, DeepVariant effectively discriminates true variants from sequencing artifacts and alignment errors. This is evident in the 9.3-fold reduction in SNP false positives

compared to FreeBayes and the 40% reduction relative to GATK (15 vs 21 false positives; **Figure S**).

Notably, the identical recall observed for DeepVariant and GATK for both SNPs and indels indicates equivalent sensitivity and shared false negative profiles. This suggests that the variants missed by both callers represent inherently difficult cases, such as low-coverage regions or complex genomic contexts. When combined with superior precision, this establishes DeepVariant as strictly superior to GATK for most applications, even though the absolute F1-score improvement is modest due to GATK's already high baseline performance.

5.5. Limitations of FreeBayes

FreeBayes showed consistently degraded performance across all evaluated metrics. The high number of false positive SNPs and indels, together with the severely depressed Ts/Tv ratio, indicates fundamental limitations in its Bayesian haplotype-based approach under these conditions. The 13.6% deviation in Ts/Tv ratio represents unacceptable systematic bias that could confound downstream analyses, including mutation spectrum studies, evolutionary inference, and clinical interpretation **(Figure T)**.

Although FreeBayes offers advantages in computational efficiency and direct haplotype reconstruction, these benefits are outweighed by its poor specificity in contexts where accurate variant calling is critical.

5.6. Persistent Challenges in Indel Detection

The consistently lower performance for indel detection across all callers **(FigureQ)** confirms that indels remain technically challenging regardless of algorithmic strategy. Indels often occur in repetitive or homopolymer regions, generate ambiguous alignments, and are associated with higher sequencing error rates. DeepVariant's substantially higher indel precision relative to FreeBayes (**Figure Q**) demonstrates that deep learning approaches provide a meaningful advantage in resolving these ambiguities.

The 4.75-fold reduction in indel false positives achieved by DeepVariant has particular clinical importance, as frameshift and splice-site indels often have severe functional consequences. At the same time, the ~15% of indels missed by all callers underscores the need for orthogonal validation strategies, especially for clinically relevant variants.

5.7. Precision–Recall Trade-offs and Clinical Relevance

The benchmarking results reveal distinct precision–recall trade-offs among callers. DeepVariant and GATK maintain high precision while achieving maximal recall, representing optimal operating points for most research and clinical workflows (**Figure Q**). FreeBayes, in contrast, sacrifices precision for negligible recall gains: its marginal difference in SNP recall relative to DeepVariant comes at the cost of a dramatic increase in false positives (**FigureS**).

In clinical diagnostics, false positives are particularly costly, both financially and ethically. Each false positive variant may prompt confirmatory testing, additional analyses, or genetic counseling. The large disparity in false positive rates between FreeBayes and DeepVariant translates directly into increased costs and heightened risk of misdiagnosis.

5.8. Quality Metrics as Orthogonal Performance Indicators

The Ts/Tv ratio analysis demonstrates the value of biological quality metrics as complements to conventional accuracy statistics (**FigureT**). DeepVariant's near-perfect alignment with the expected Ts/Tv ratio confirms the absence of systematic bias, whereas FreeBayes's large deviation flags fundamental issues not fully captured by TP/FP/FN counts alone. This highlights the importance of incorporating orthogonal biological indicators in variant caller benchmarking.