UNIVERSITY OF VERONA
DEPARTMENT OF BIOTECHNOLOGY

DOCTORAL PROGRAM IN BIOTECHNOLOGY
XXXVIII CYCLE

Fragment Size Profiling and Sequencer Quality Evaluation in Clinical NGS

S.S.D. BIO/18

Coordinator: Prof.ssa Flavia Guzzo

Tutor: Prof. Massimo Delledonne

Doctoral Student: Bertoli Luca

Fragment Size Profiling and Sequencer Quality Evaluation in Clinical NGS

Luca Bertoli

PhD thesis

Verona, 09 December 2025

Table of Contents

Sommario

Questa tesi indaga due tra i determinanti più critici della qualità dei dati NGS in ambito clinico: la distribuzione delle lunghezze dei frammenti di DNA (dimensione dell'inserto) e i punteggi di qualità delle basi riportati da piattaforme di sequenziamento allo stato dell'arte. Analizzando congiuntamente questi due aspetti, essa fornisce una visione integrata di come le scelte sperimentali influenzino la copertura e la genotipabilità nella genomica clinica, quantificando inoltre la relazione tra i punteggi di qualità stimati, specifici per piattaforma, e la reale accuratezza empirica dei dati.

In primo luogo, la tesi affronta un'importante lacuna nella letteratura sulla dimensione dell'inserto, laddove la lunghezza dei frammenti è stata tradizionalmente sintetizzata attraverso statistiche globali come media o mediana, senza una risoluzione dettagliata di come diverse fasce di dimensione dell'inserto influenzino copertura, mappabilità e genotipabilità. Esplorando in modo sistematico l'intero spettro delle lunghezze osservate nel sequenziamento dell'esoma e caratterizzandone le prestazioni rispetto a metriche analitiche chiave, si dimostra che inserti molto corti diminuiscono simultaneamente la mappabilità e l'efficienza di copertura, mentre inserti più lunghi mantengono una buona mappabilità, e prestazioni nella chiamata varianti, ma a costo di una efficienza di copertura ridotta. Ciò indica che un'ottimizzazione appropriata della dimensione dell'inserto può migliorare congiuntamente sia la copertura sia l'interpretabilità dei dati WES in ambito clinico. Tale analisi è implementata in un workflow completamente riproducibile basato su Snakemake, che consente una valutazione standardizzata della dimensione dell'inserto tra esperimenti e coorti differenti.

Sul lato delle piattaforme, la tesi affronta una lacuna metodologica nella valutazione dei sequenziatori, poiché la maggior parte degli studi comparativi si basa su indicatori aggregati (come i valori globali di Q30 e le metriche complessive di rilevazione delle varianti) e raramente valuta in modo sistematico quanto i punteggi di qualità specifici di ciascuna piattaforma siano ben calibrati rispetto ai reali tassi di errore, o come tale calibrazione dipenda dal ciclo di sequenziamento e dalla lunghezza dei frammenti. A tale scopo, viene introdotto un framework unificato

basato su una singola libreria PCR-free di NA12878 sequenziata su molteplici piattaforme moderne a short-read. All'interno di questo framework, i punteggi assegnati dal sequenziatore e quelli basati sull'allineamento sono confrontati lungo i cicli di sequenziamento e in intervalli di dimensione dell'inserto, fornendo una valutazione ad alta risoluzione e direttamente confrontabile della qualità dei dati tra tecnologie. Questa analisi rivela discrepanze specifiche per piattaforma tra accuratezza riportata ed empirica ed evidenzia casi in cui punteggi nominalmente simili corrispondono a profili di errore differenti, pur confermando che tutte le piattaforme valutate forniscono una proporzione elevata di basi realmente ad alta qualità, idonee per analisi cliniche.

Nel complesso, questi contributi chiariscono come la dimensione dell'inserto e la qualità del sequenziamento influenzino congiuntamente l'affidabilità dei dati NGS clinici e forniscono indicazioni pratiche per ottimizzare sia la preparazione delle librerie, sia il sequenziamento nella medicina genomica di routine.

Abstract

This thesis investigates two of the most critical determinants of clinical NGS data quality: the distribution of DNA fragment lengths (insert size) and the sequencer-reported base quality scores of state-of-the-art platforms. By jointly analyzing these two aspects, it provides an integrated view of how experimental choices shape coverage and genotypability in clinical genomics, and it quantifies the relationship between platform-specific estimated quality scores and the true empirical accuracy of the data.

Firstly, the thesis addresses a major gap in the insert-size literature, where fragment length has been traditionally summarized with global statistics such as mean or median, without resolving how different insert-size ranges influence coverage, mappability, and genotypability. By systematically exploring the full spectrum of insert sizes observed in whole-exome sequencing and characterizing their performance across key analytical metrics, it shows that very short inserts simultaneously reduce mappability and effective coverage, whereas long inserts preserve good mappability and variant-calling performance but at the cost of reduced coverage, indicating that appropriate optimization of insert size can jointly improve both coverage and interpretability of clinical WES data. This analysis is implemented in a fully reproducible workflow based on Snakemake, enabling standardized insert-size-resolved evaluation across experiments and cohorts.

On the platform side, the thesis tackles a methodological gap in sequencer evaluation, as most comparative studies rely on aggregate indicators (such as global Q30 rates and overall variant-detection metrics) and rarely assess, in a systematic way, how well platform-specific quality scores are calibrated against true error rates or how this calibration depends on cycle and fragment length. To address this, it introduces a unified framework that uses a single PCR-free NA12878 library sequenced on multiple modern short-read platforms. Within this framework, sequencer-assigned and alignment-based scores are compared across sequencing cycles and insert-size intervals, providing a high-resolution, directly comparable assessment of data quality across technologies. This analysis reveals platform-specific discrepancies between reported and empirical accuracy and

highlights cases where nominally similar quality scores correspond to different error profiles, while confirming that all evaluated platforms deliver a high proportion of genuinely high-quality bases suitable for clinical-grade analyses.

Together, these contributions clarify how insert size and sequencing quality jointly shape the reliability of clinical NGS data and provide practical guidance for optimizing both library preparation and sequencing in routine genomic medicine.

1. Introduction

1.1. Sequencing Strategies in Clinical Genomics

Modern clinical genomics relies on three primary next-generation sequencing (NGS) strategies: whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted gene panels. Each approach offers a distinct balance between genomic breadth, sequencing depth, diagnostic sensitivity, cost, and analytical complexity. Understanding their respective strengths and limitations provides essential context for the analyses presented in this thesis, which focuses on optimizing WES performance and evaluating platform-specific sequencing accuracy.

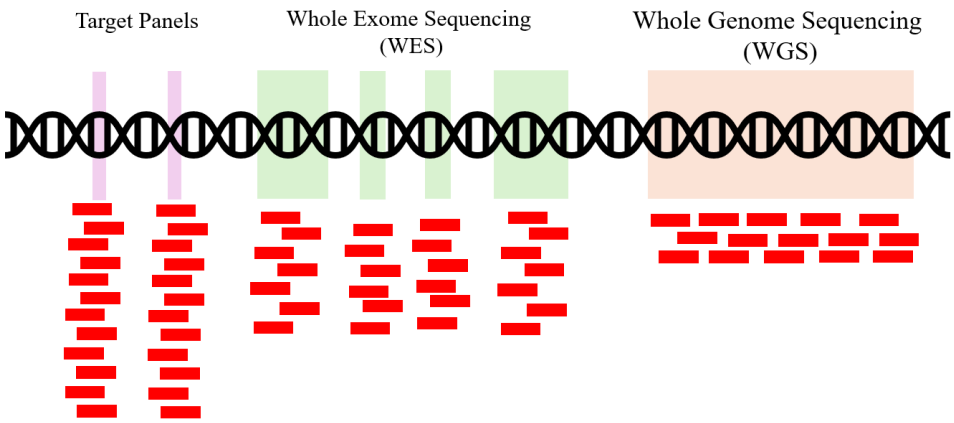


Figure 1 Sequencing paradigms. Target Panels, Whole exome Sequencing (WES), Whole Genome Sequencing (WGS).

1.1.1. Whole-Genome Sequencing (WGS)

Whole-genome sequencing focusses on the entire human genome, providing the most comprehensive view of genetic variation (Figure 1). Because WGS does not rely on hybrid-capture probes, it avoids many biases associated with GC content, repetitive regions, and homology. As a result, WGS typically achieves highly uniform coverage and allows robust detection of diverse variant types, including

- single-nucleotide variants (SNVs)
- insertions and deletions (INDELs)
- structural variants and copy-number alterations

These strengths have been widely demonstrated in population-scale projects[1] and in clinical rare-disease studies[2]. However, WGS presents practical challenges:
it is expensive, generates large data volumes, and requires substantially more computational resources and interpretation time compared with exome or panel sequencing[3].

Clinical whole-genome sequencing is typically performed at 30–40× mean depth, which provides sufficient sensitivity for germline SNV and INDEL detection while maintaining manageable sequencing cost and data volume[4].

1.1.2. Whole-Exome Sequencing (WES)

Whole-exome sequencing focuses exclusively on coding exons, which account for ~2% of the genome but contain ~85% of known pathogenic variants[5]. This balance of genomic scope and clinical relevance makes WES one of the most widely used tools in diagnostic genetics, particularly for Mendelian and rare disease diagnostics, hereditary cancer analysis and translational research (Figure 1).

Compared to whole-genome sequencing, whole-exome sequencing offers substantial advantages, including lower sequencing cost, the ability to achieve higher effective depth, reduced computational burden, and simplified interpretability[6].

However, WES is inherently limited by the constraints of hybrid-capture chemistry, which results in non-uniform exon coverage, systematic underperformance in GC-rich and repetitive regions, reduced capture efficiency in homologous gene families, and substantial platform- and protocol-dependent variability. These limitations often leave a fraction of clinically relevant exons incompletely genotyped. Moreover, because WES targets only coding regions, it does not detect variants located in intronic or regulatory regions and provides limited ability to identify structural variants, which generally require genome-wide coverage for reliable detection[7].

1.1.3. Targeted Gene Panels

Targeted gene panels sequence only a predefined subset of genes, typically those associated with the diagnostic question. Because they focus on a small genomic footprint, these assays routinely achieve very high sequencing depth, often exceeding 300x, which enables sensitive detection of low-frequency variants, rapid turnaround times, straightforward interpretation, and substantially reduced sequencing costs (Figure 1). For these reasons, targeted panels are widely used in oncology and in hereditary disease diagnostics[8].

Despite these advantages, targeted panels are inherently constrained by their limited genomic scope. They cannot detect novel or unexpected variants outside the predefined gene list and must be continually updated as clinical knowledge evolves. Consequently, although targeted sequencing excels for focused diagnostic questions, it lacks the discovery power and broader variant detection capabilities provided by WES and WGS[9].

1.1.4. Comparative Summary

These sequencing paradigms occupy complementary roles in clinical genomics:

- Whole-genome sequencing (WGS) provides the most comprehensive variant detection across coding and non-coding regions, albeit at higher cost and analytical complexity.
- Whole-exome sequencing (WES) offers an optimal balance between genomic breadth, depth of coverage, and cost, making it the standard approach for most germline diagnostic applications.
- Targeted gene panels, by focusing on predefined gene sets, maximize sensitivity, throughput, and turnaround time for highly specific clinical questions.

The work presented in this thesis uses WES to evaluate how insert size influences key analytical outcomes, such as genotypability, mapping behavior, and variant-calling accuracy, while PCR-free WGS is employed to assess the intrinsic accuracy with which different sequencing platforms generate the data.

1.2. Performance Metrics

1.2.1. On-, Near-, and Off-Target Fractions in WES

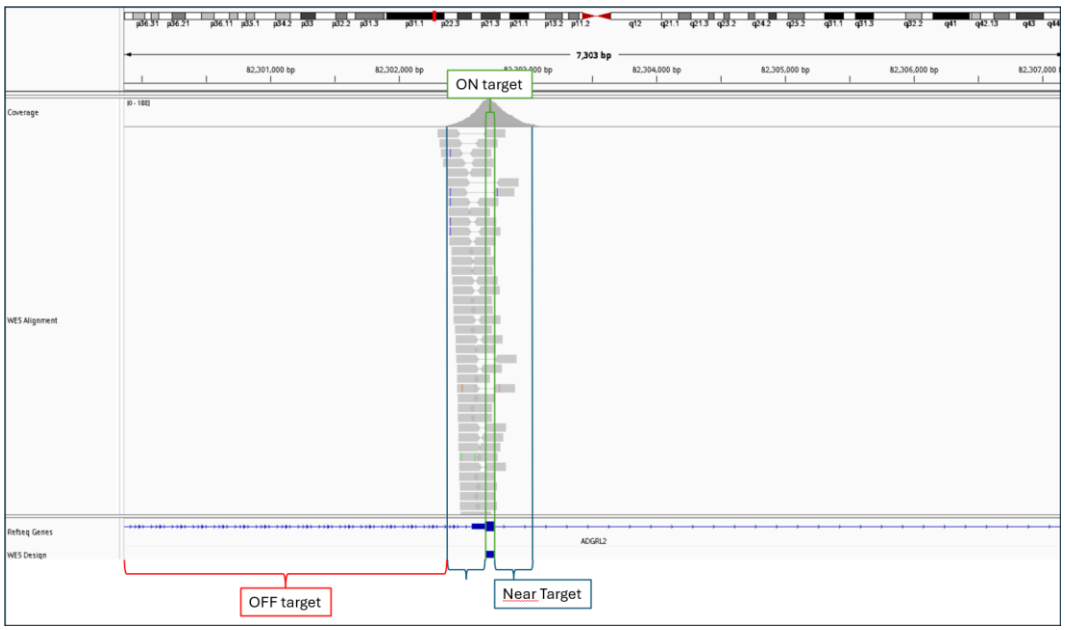


Figure 2 Integrated Genome Browser (IGV) image showing an example of design region covering a coding sequence, alignments represent WES fragments. Red delineated regions are OFF target, blue delineated regions are NRAR target and green delineated regions are ON target.

In hybrid-capture WES, the fraction of sequenced bases that map on-target, near-target, or off-target represents a key metric of assay efficiency. On-target bases fall strictly within the captured exonic regions; near-target bases map within a predefined flanking window, reflecting partially enriched fragments; off-target bases map elsewhere in the genome and represent sequencing effort not contributing to diagnostic yield (Figure 2).

High on-target rates improve both cost-efficiency and depth distribution across clinically relevant regions, whereas excessive off-target sequencing dilutes effective coverage and reduces the number of usable reads for variant calling. Near-target fractions are also informative because they capture the intrinsic variability of hybridization efficiency and the influence of insert size: longer fragments tend to extend beyond exon boundaries, increasing near-target recovery but also potentially improving mapping quality in challenging regions[10].

Together, the on/near/off-target proportions provide a sensitive readout of capture specificity, library quality, and insert-size effects, and are therefore routinely examined in clinical WES validation and quality-control pipelines.

1.2.2. Coverage Requirements in Clinical WES

Figure 3 Diagram displaying the concept of depth of coverage.

Depth of coverage refers to the number of sequencing reads that align to a given genomic position and represents one of the foundational metrics for evaluating whole-exome sequencing (WES) performance (Figure 3). Clinical diagnostics typically require ≥60X median coverage across targeted regions to ensure reliable variant calling and minimize false negatives. This threshold stems from observations that heterozygous variants often fail to meet quality filters or are missed entirely when coverage drops below ~20–30 reads, and therefore higher aggregate coverage is necessary to accommodate uneven capture efficiency and GC-related biases[11].

However, raw coverage alone does not guarantee interpretability. Uneven distribution across exons, especially in GC-rich or repetitive regions, can cause local drops that affect clinical sensitivity even when global coverage metrics appear adequate.

1.2.3. Genotypability: A Composite Metric for Clinical Reliability

Because coverage is not sufficient to ensure variant detectability, clinical sequencing increasingly relies on genotypability, defined as the percentage of bases in which a genotype can be confidently assigned. Genotypability incorporates:

- Depth of coverage (e.g., ≥4 or ≥10X)
- Base-level quality scores (Phred Q)
- Mapping quality, reflecting the confidence in read placement

A base is considered callable only when all thresholds are met simultaneously. This metric more accurately reflects clinical readiness, since regions with high coverage but low mapping quality remain uncallable and can hide clinically relevant variants[12].

1.2.4. Benchmarking with GIAB High-Confidence Variant Sets and Stratified Regions

An accurate evaluation of whole-exome sequencing (WES) performance ultimately requires a reference variant truth set that defines, with high confidence, which variants are truly present in a sample. The Genome in a Bottle (GIAB) Consortium provides such resources for a set of well-characterized human genomes, most notably NA12878, which is widely adopted as the gold standard in sequencing validation studies.

The GIAB "Gold Set" for NA12878 comprises high-confidence SNVs and indels validated across multiple sequencing technologies, library preparations, and variant callers, retaining only concordant, reproducible variants. These calls are accompanied by a high-confidence (HC) regions BED file defining genomic intervals where the reference sequence is reliable and variant calling accuracy is maximized, enabling robust, technology-independent benchmarking of performance metrics[13].

Beyond the definition of high-confidence regions, GIAB provides 13 distinct genome stratification categories via BED files, partitioning the genome into challenging contexts for standardized variant calling evaluation: Low Complexity (homopolymers, STRs, VNTRs), Functional Technically Difficult regions, Genome Specific (NA12878-specific complexities), Functional Regions (coding/non-coding), GC Content (high/low), Mappability, Segmental Duplications, Ancestry (GRCh38-only), Telomeres, XY, Other Difficult (VDJ, MHC, rDNA), and Union categories combining multiple difficulties. These stratifications reveal platform-specific weaknesses, such as reduced SNV/INDEL

recall in low-mappability (~89 Mb) and segmental duplication regions even for long-read methods.[13]

Central to these stratifications, low mappability regions identify genomic intervals where short reads align non-uniquely due to high sequence similarity. Created using GEnome Multitool (GEM)-Mapper on GRCh38/CHM13, moderately low-mappability regions allow up to two mismatches and one INDEL in 100 bp segments, while highly low-mappability regions permit zero mismatches/INDELs in 250 bp segments (matching typical short-read lengths). Final BED files union both stringencies after SAMtools/BEDtools processing, with CHM13 expanding regions (e.g., satellite/rRNA repeats) versus GRCh37/38, particularly on chromosomes 1, 9, and acrocentrics[13], [14].

1.2.5. Clinical Variant Databases: ClinVar and HGMD

In addition to benchmarking against GIAB truth sets, clinical sequencing pipelines must assess performance on clinically relevant genomic loci, positions where pathogenic or likely pathogenic variants are known to occur. Two complementary databases serve this purpose:

- ClinVar
- HGMD

ClinVar is a public, NIH-maintained archive linking human genetic variants to clinical phenotypes and expert-curated interpretations. It includes pathogenic (P), likely pathogenic (LP) variants, variants of uncertain significance (VUS) and benign annotations. It provides a broad catalogue of clinically relevant positions across the genome. For a sequencing assay to be clinically reliable, these loci must fall within callable, high-confidence regions of the sample[15].

HGMD (Human Gene Mutation Database) is a manually curated, subscription-based repository that compiles experimentally validated germline variants implicated in human inherited disease. Unlike broad variant archives, HGMD focuses specifically on mutations with established or suspected pathogenic impact, providing a comprehensive catalogue of clinically relevant genetic alterations across Mendelian disorders.

Each variant is assigned to a clinically meaningful classification, including:

- DM (Disease-causing Mutation): variants with strong experimental or clinical evidence of pathogenicity.
- DM? (Likely Disease-causing): variants supported by suggestive but not yet definitive evidence.
- DFP (Disease-associated Functional Polymorphism): variants that modulate gene or protein function and are associated with disease susceptibility.

HGMD is widely used in diagnostic and clinical genomics to interpret genetic findings, support variant classification, and contextualize newly observed mutations within the broader framework of known disease mechanisms[16].

1.3. Insert Size: Definition and Generation

1.3.1. Definition of Insert Size

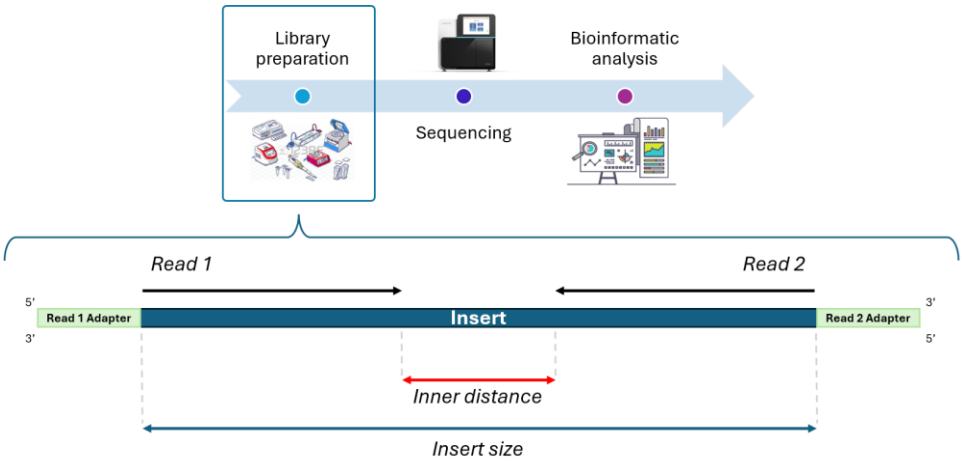


Figure 4 Diagram displaying the concept of Insert Size in paired-end sequencing.

In paired-end sequencing, the insert size refers to the length of the DNA fragment located between the two sequencing adapters ligated to the molecule during library preparation (Figure 4). In alignment files (BAM/SAM), this metric corresponds to the *template length* (TLEN), which reports the inferred distance between the forward and reverse reads.
Insert size therefore represents the physical fragment of DNA that the sequencing reads jointly interrogate, influencing mapping behavior, coverage structure, and downstream variant calling precision[17].

1.3.2. Laboratory Generation of Insert Size

Insert size is determined during library preparation and reflects a sequence of biochemical steps that shape the final fragment population. Two steps are particularly influential:

- Fragmentation, in which genomic DNA is broken into smaller fragments using mechanical (e.g., sonication) or enzymatic methods (e.g., tagmentation).
- Size selection, performed after adapter ligation using SPRI beads or gel-based approaches to enrich for fragments within a desired size range.

Manufacturers of WES kits generally recommend specific size-selection windows, typically around 200 bp, chosen to maximize hybrid-capture efficiency and mapping performance[18], [19], [20], [21]. However, despite standardized protocols, the final insert-size distribution is highly sensitive to input DNA integrity, enzymatic reaction efficiency, bead-to-DNA ratios, and cleanup conditions[22], [23]. As a result, empirical insert-size distributions rarely collapse into a single dominant value.

In practice, WES libraries frequently display broader and heterogeneous distributions. Under optimal laboratory conditions, fragmentation and size selection produce a symmetric, bell-shaped distribution centered on the intended size, reflecting uniform enzymatic activity and precise bead selection. More commonly, especially in high-throughput or clinical workflows, insert-size profiles become right-skewed, with an excess of longer fragments caused by incomplete fragmentation, partial degradation of DNA, or relaxed bead-selection thresholds. Such skewness is well documented in real exome datasets and can impact enrichment performance in long exons or regions with complex structure[24].

1.3.3. Insert-Size-dependent performance of WES

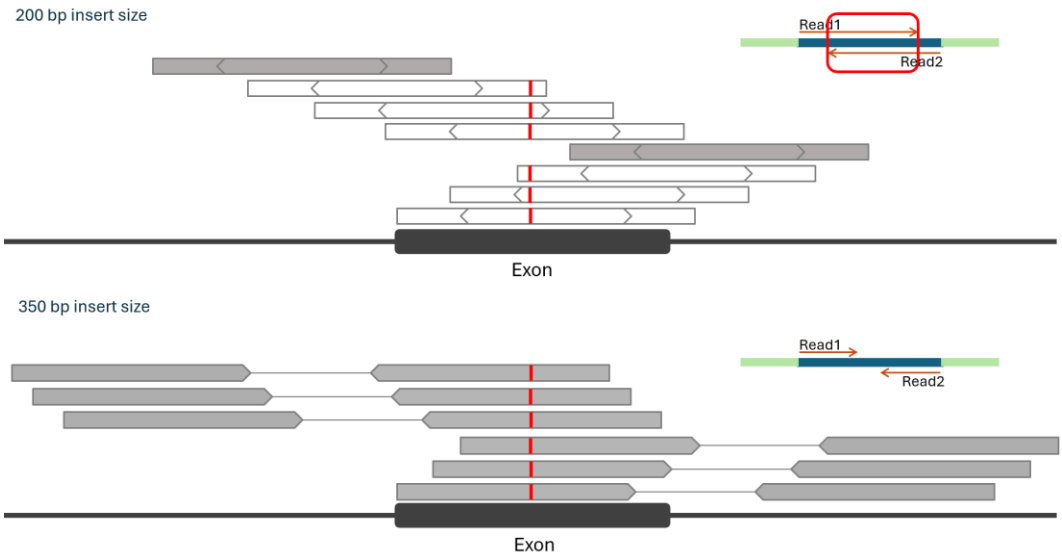


Figure 5 Example exon region where insert size affects mappability: gray fragments indicate reads mapping with high quality, whereas white fragments represent reads with mapping quality 0.

Insert size plays a central role in shaping the behavior of whole-exome sequencing data because it determines how DNA fragments interact with the underlying exon architecture. When fragments are too short, they may fail to span entire exons or provide sufficient anchoring in regions of low mappability, leading to increased mapping ambiguity, reduced coverage uniformity, and decreased genotypability in difficult loci (Figure 5). Conversely, sufficiently long fragments can improve placement accuracy, reduce mate overlap, and enhance variant detection in GC-rich or repetitive contexts[10].

Thus, the effect of insert size is not monotonic but reflects a balance between fragment length, exon length, and genomic mappability, suggesting that WES performance depends not only on the laboratory-selected mean insert size but also on the entire distribution of fragment lengths present in a library.

1.3.4. Limitations of Current Insert-Size Characterization Practices

Despite its importance, insert size is typically reported in sequencing studies using only a single summary value (the mean or median). This approach overlooks several critical aspects. Firstly, libraries may share the same mean insert size but differ dramatically in their distribution, resulting in profoundly different mapping

behaviors[23], [25]. Moreover, existing studies evaluate insert size globally, without focusing on the performance of individual insert-size intervals.

Given these gaps, there is a strong need for:

1. Insert-size-resolved analyses that characterize sequencing performance at fine-grained bin levels rather than using global summary metrics.
2. Normalization strategies allowing unbiased comparison of insert size intervals across multiple samples and cohorts.

Addressing these issues is essential for advancing WES optimization, supporting more precise library preparation guidelines, and ensuring robust clinical interpretation, especially in regions where correct genotyping is most challenging.

1.4. Evaluation of Sequencing Quality

1.4.1. Phred Quality Scores and Base-Level Accuracy

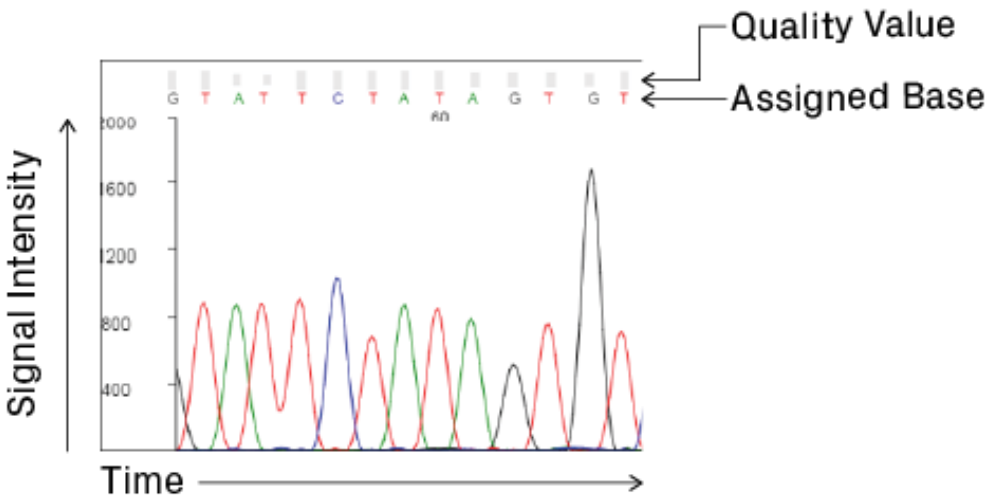


Figure 6 Sanger Sequencing chromatogram image taken from https://blog.genewiz.com/analyzing-sanger-sequencing-data

Phred quality scores were originally developed in the 1990s as part of the Phred base-calling software designed to interpret Sanger sequencing chromatograms (Figure 6). The software analyzed characteristics of the chromatogram peaks (such as peak shape, height, spacing, and resolution of overlapping peaks) to estimate the likelihood of base-calling errors. These empirical features were carefully modeled to predict error probabilities for each base call.

To translate these error probabilities into quality scores, Phred used lookup tables derived from extensive training on large sets of manually curated chromatograms with known sequence errors. These tables mapped observed chromatogram metrics to error probabilities, enabling rapid and accurate estimation of base-call confidence without requiring complex real-time computations.

The quality scores are expressed on a logarithmic scale defined as:

$$Q = -10 \times \log_{10}(P_e)$$

where $P_e$is the probability that a base call is incorrect. For example, a Q30 base corresponds to a 1 in 1000 chance of error, while a Q40 score indicates an error rate of 1 in 10,000.

This system became widely adopted due to its compact and interpretable representation of sequencing accuracy, directly linked to probabilistic models of sequencing errors. Though originally developed for Sanger sequencing chromatogram data, the Phred scoring method has been extended to next-generation sequencing (NGS), where Phred-derived quality scores remain a fundamental metric for base-level accuracy and read filtering [26], [27].

1.4.2. Modern Computation of Phred Scores in High-Throughput Sequencing

In next-generation sequencing (NGS), Phred quality scores remain the universal standard for representing base-call accuracy and are widely used for read filtering, variant calling confidence assessment, and downstream quality control pipelines. However, unlike the original Phred scores derived from analyzing Sanger chromatogram peak characteristics, NGS platforms generate Q-scores using instrument-specific, proprietary algorithms adapted to their sequencing chemistries[28][29].

These modern computations integrate diverse data sources including optical signal intensities (such as fluorescence intensities in Illumina platforms, avidity signals in other systems, or probe-anchored synthesis metrics)[30]. Additional factors encompass cluster or DNA nanoball geometry on flow cells[31] and phasing/pre-phasing statistics, (measuring synchrony loss where nucleotides lag or advance during cycles)[32]. Machine learning models or empirical lookup tables, trained on

large datasets correlating signals with known error rates, underpin these predictions[29], [32], [33].

Consequently, NGS Phred scores represent sophisticated model-based predictions of base-call error probability rather than direct empirical measurements from chromatograms. This evolution maintains the logarithmic scale $Q = -10 \times \log_{10}(P_e)$ while adapting to high-throughput optical data challenges[28], [29].

1.4.3. Modern Sequencing Platforms and Q-Score Encoding

Modern next-generation sequencing platforms rely on distinct biochemical and optical principles for DNA interrogation, each with specific advantages and implications for error profiles, Q-score calibration, and read-level performance.

Below, the four platforms evaluated in this thesis are briefly introduced, with emphasis on their sequencing chemistry, distinguishing features, and quality-score encoding schemes.

1.4.3.1. *Illumina NovaSeq X: Sequencing-by-Synthesis with Reversible Terminators*

Illumina NovaSeq X utilizes a sequencing-by-synthesis (SBS) approach, incorporating fluorescently labeled, reversible terminator nucleotides one base at a time during each cycle. After nucleotide incorporation, imaging captures the fluorescence signal, and the chemical blocking group is removed to allow the subsequent cycle. This process is supported by patterned flow cells that enable high-density cluster formation, combined with highly optimized optics and real-time signal processing to maximize accuracy[34]. The platform achieves high throughput, capable of producing terabases of data per run with ≥85% of bases higher than Q30 for 2×150 bp reads[35].

For base quality assessment, NovaSeq X employs a binned Phred quality scoring scheme, outputting a limited number of discrete Q-values. Although this binning reduces data file size and computational demands, it also compresses the quality information, potentially masking subtle variations in base-call accuracy across sequencing cycles and individual bases[36], [37].

1.4.3.2. GeneMind SURFSeq 5000: SBS-Derived Chemistry with Optimized Optics

GeneMind SURFSeq 5000 utilizes sequencing-by-synthesis (SBS) chemistry derived from the reversible-terminator approach initially developed by Illumina. The core principle involves the incorporation of fluorescently labeled nucleotides one base at a time, followed by imaging after each cycle. GeneMind has introduced proprietary modifications to components including polymerases, fluorophores, optical detection systems and quality score scheme. Additionally, the platform employs optimized optics, achieving Q30 ≥90% in balanced mode and Q30 ≥95% in enhanced mode, targeting clinical sequencing workflows[38].

Despite being relatively recent to the market, SURFSeq 5000 has been adopted by clinical laboratories, with chemistry and quality scoring approaches that broadly resemble Illumina's SBS systems but include proprietary enhancements affecting accuracy and workflow[39].

1.4.3.3. Element AVITI: Avidity Sequencing

The Element AVITI platform employs Avidity Sequencing with polony-based amplification via rolling circle amplification (RCA). Library fragments hybridize to surface primers on the flow cell, where RCA generates long DNA strands containing hundreds of copies of the original template, forming localized clusters called polonies. These polonies provide multiple synchronized binding sites for sequencing reagents, eliminating PCR error propagation seen in traditional bridge amplification[30], [40].

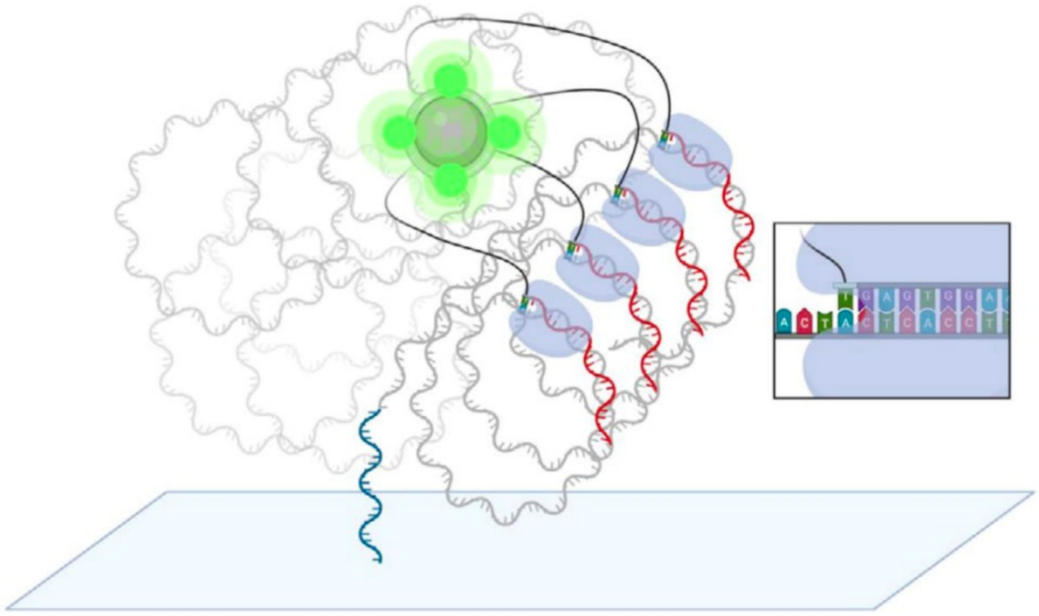


Figure 7 Image displaying the fluorescent avidity molecule bound to the respective polony[30].

Sequencing chemistry uses multivalent "avidites" (polymer-nucleotide substrates with multiple nucleotide heads) that bind simultaneously to complementary bases across the polony. This high-avidity binding forms stable complexes between the avidites and target bases, producing fluorescence signals monitored in real-time. Four dye-labeled avidites interrogate each cycle, with images captured per field of view for base identification. The engineered high-fidelity polymerase and multivalent binding mechanism reduce phasing/pre-phasing rates, supporting signal detection across high polony densities[30], [40].

Quality score calibration follows the Phred standard using four predictors trained on empirical error data:

- maximum intensity per polony across color channels;
- intensity ratio, defined as $(A + 1)/(B + 1)$where A is the highest intensity and B the second highest intensity across channels;
- phasing/pre-phasing estimates;
- median intensity ratio of the lowest 10% intensity polonies.

This process yields continuous Q-scores with >90% Q30 for 2x150bp runs and routine Q40+ achievement, strongly correlating with post-alignment mismatch rates[30], [41].

1.4.3.4. MGI T1+: DNBSEQ and cPAS Chemistry

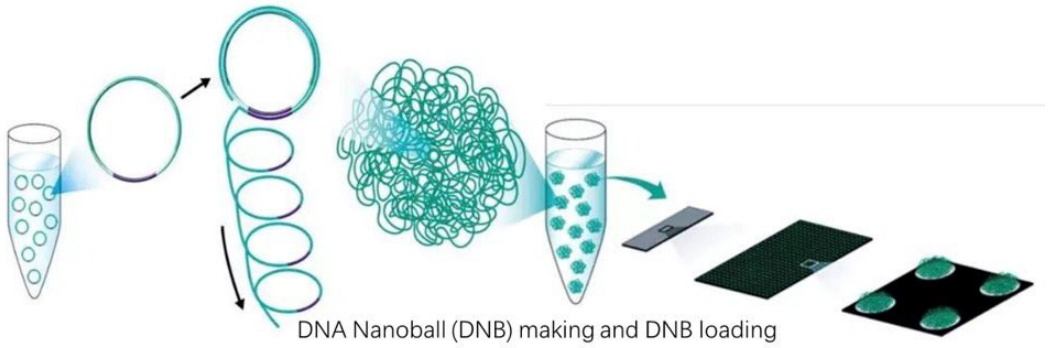


Figure 8 Image displaying the process of producing DNA Nanoballs (DMBs) and loading on the flowcell (https://decodescience.com.au/partners/mgi/).

The MGI T1+ platform employs DNBSEQ technology utilizing rolling-circle amplification (RCA) to generate DNA nanoballs (DNBs) from circularized DNA templates. Library fragments are circularized and amplified via RCA on the flow cell surface, forming densely packed DNBs that provide multiple synchronized template copies for sequencing[31], [42].

Sequencing chemistry uses combinatorial Probe-Anchor Synthesis (cPAS), where fluorescently labeled DNA probes hybridize to the exposed base in each DNB. Rather than direct nucleotide incorporation, short probes (anchor + interrogation nucleotide) bind specifically to the template base via ligation, producing detectable fluorescence signals. Four color-coded probes per cycle identify A/C/G/T, with unbound probes washed away before imaging and cleavage for the next cycle. This probe-based approach eliminates reversible terminators and reduces incorporation errors[31], [42].

Quality score calibration uses proprietary algorithms that integrate signal intensity ratios, probe binding specificity metrics, and DNB uniformity across the patterned flow cell. The platform emits continuous Q-scores with high aggregate quality (>95% Q30 for PE150 runs), reflecting stable signal profiles from uniform DNB distribution and PCR-free amplification that minimizes duplication artifacts[31][42].

1.4.4. Limitations of Modern Q-Score Schemes

Despite their ubiquity, platform-specific Q-scores suffer critical limitations:

- Q-scores estimate error probability via proprietary models (optics, phasing, ML calibration) rather than directly measuring post-alignment mismatches. Model misspecification or sequence context can cause divergence.
- NovaSeq X and SURFSeq 5000 emit discrete Q-values, obscuring subtle accuracy gradients against continuous schemes (AVITI, MGI).
- Distinct calibration procedures may yield different Q-distributions even for identical samples.

These limitations necessitate empirical validation via alignment identity alongside Q-scores for fair platform comparisons.

1.4.5. Platform Benchmarking Studies: Current Approaches and Limitations

Recent literature has benchmarked NGS platforms using reference samples like NA12878, primarily focusing on aggregate quality metrics and variant calling performance. For instance, Kim et al. (2021) compared MGI DNBSEQ and Illumina NovaSeq, finding comparable SNV sensitivity but did not provide detailed empirical validation of base-level quality scores, limiting insights into platform calibration differences[31]. Sun et al. (2024) evaluated GeneMind SURFSeq 5000, reporting high Q30 rates, but their study lacked cycle- and position-resolved accuracy assessments, reducing granularity in quality evaluation[39]. Arslan et al. (2023) demonstrated Element AVITI's high concordance between Q-scores and empirical mismatches, but the study did not standardize benchmarking across multiple platforms using the same library, limiting cross-comparability[30].

2. Aim of the Thesis

This thesis aims to elucidate how insert size and sequencing platform jointly affect the performance and reliability of clinical sequencing data. Sequencing behavior is characterized at all possible insert size levels by analyzing fixed, normalized insert-size bins (ISBs) instead of relying on global Gaussian-like distributions, enabling the identification of intervals that maximize on-target efficiency, mapping quality, genotypability and ultimately improve variant calling in whole-exome sequencing. Building on these optimized insert-size ranges, this thesis develops and tests on real-world data, a reproducible pipeline that evaluates how short-read sequencing platforms translate library properties into actual data quality, relating reported Phred scores to empirical error rates across sequencing cycles and fragment lengths, aiming to minimizing sequencing errors and improving the reliability of clinical genomic data.

3. Methods

3.1. Characterization of Target Region Lengths in WES Designs

To characterize the length distribution of the regions captured by Twist Exome 2.0 + Comp Exome Spike-In kit, we first obtained the target design BED file, which specifies the genomic intervals targeted by the capture probes. Each interval was processed using a custom python script. For all regions included in the design, the total number of regions and bases, the mean and median length were computed.

To further inspect the region length distribution, the target design regions were separated in the following categories based on their length:

a) 100 bp;
b) 101–200 bp;
c) 201–300 bp;
d) ≥ 301 bp.

For each category, we quantified both the number of regions and the number of bases falling within the corresponding length range. Percentages were computed relative to the full set of targeted regions and total targeted bases, respectively.

Finally, the number of regions for each length and the base content for each region length were graphically represented as histograms using matplotlib library.

3.2. Comparison of different insert size profiles

To evaluate the shape of the insert size distribution, two representative samples processed with the Illumina DNA Prep with Exome 2.5 Enrichment kit were selected. Library preparation was performed following Illumina's standard operating procedures for both the alpha and the final commercial version of the kit. The libraries were sequenced on the Illumina NovaSeq 6000 and NovaSeq X platforms respectively. After sequencing, adapters and low-quality bases were removed using *fastp* (v0.23.4), and the reads were aligned to the hg38 reference genome using *BWA-MEM2* (v2.2.1)[43], [44]. Insert size distributions were then computed using Picard *CollectInsertSizeMetrics* (v3.4.0), which reports fragment-length frequencies based on properly paired primary alignments. Finally, graphical

visualization of the distributions was performed using stepwise histograms generated with the *Matplotlib* Python library.

3.3. Generation of Twist WES samples for insert size bin generation

The characterization of insert-size performance presented in this work was conducted through the reanalysis of 195 clinical WES samples, organized into three sequencing batches of 65 samples each. Genomic DNA integrity was assessed using the Agilent TapeStation system, and DNA quantity was measured using the Qubit fluorometric assay. Only samples that passed these quality-control steps were processed following the Twist library preparation and hybrid capture protocol, using the Twist Exome 2.0 + Comp Exome Spike-In probe set (Twist EF 1.0 library prep and Standard Hybridization v2 capture kit). Libraries were evaluated for fragment size and concentration before hybrid capture using the TapeStation and Qubit fluorometric assay. The same quality control was applied to the final captures in addition to the qPCR assay required to precisely evaluate the captures' concentration before sequencing. Sequencing was performed on the Illumina NovaSeq 6000 platform using S4 flow cells. Following the run, raw base call (BCL) files were demultiplexed into FASTQ format using *bcl2fastq* (v2.19.0.316).

3.4. Generation of NA12878 WES replicates for variant-calling benchmarking

To evaluate whether the insert-size effects observed in clinical WES samples also translate into differences in variant-calling accuracy, we analyzed a set of 22 WES replicates of the NA12878 reference genome. All samples were prepared using the GENEQUALITY library preparation and hybrid-capture kit , starting from genomic DNA derived from the Coriell NA12878 reference cell line. The 22 WES libraries were sequenced across two independent sequencing batches, one comprising 9 samples and the other 13. These samples were analyzed as a single batch. Following sequencing, raw BCL files were converted to FASTQ format using the instrument vendor's demultiplexing pipeline.

3.5. SnakeBin Workflow for Insert Size Bin (ISB) WES Analysis

To precisely analyze sequencing performance at various insert size levels, we developed SnakeBin, an automated workflow implemented in *Snakemake*

(v7.32.4)[45]. The pipeline creates datasets stratified by insert size and with normalized depth or fragment count, allowing insert-size-dependent behavior to be evaluated without confounding technical effects, while providing a fully scripted framework that ensures analyses are both robust and reproducible. The pipeline is publicly available at: https://github.com/LucaBertoli/SnakeBin

3.5.1. Pipeline configuration

The SnakeBin workflow is fully configurable through a YAML file, which allows users to specify all required inputs, reference files, computational settings, and parameter related to the insert size intervals, without modifying the pipeline itself. The configuration file defines:

- the list of samples to be processed, provided as a CSV file containing the sample names;
- the output directory, where all intermediate and final results are generated;
- the reference genome directory, including the FASTA sequence and relative files required by the tools for the processing (*BWA-MEM2* indexes, *Picard* reference interval list and dictionary).

The configuration file also sets the number of computational threads allocated to each tool (e.g. *Fastp*, *BWA-MEM2*, *Samtools[46]*, *BGZIP*), ensuring efficient use of available resources.

Insert-size handling is controlled through three core parameters: the minimum and maximum insert size to consider in the analysis, and the fixed bin width used to generate the Insert Size Bins (ISBs).

For the analyses presented in this thesis, insert sizes from 1bp to 1000bp with a fixed width of 50bp were evaluated.

The configuration file also specifies the normalization targets used throughout the workflow, including the desired number of fragments per insert-size bin (e.g., 30 million) and the mapped-coverage levels for coverage-based downsampling. The pipeline is designed to handle any arbitrary list of downsampling targets, performing all requested normalizations in a single execution. In line with standard

requirements for clinical-grade WES data, a target coverage of 60x was used as the primary normalization level in this work.

The configuration also allows the user to specify an arbitrary list of BED files defining genomic target regions. Each of these region sets is processed independently, and all are used for computing coverage, ON/NEAR/OFF-target metrics, and other region-specific statistics. In this work, the primary target definitions correspond to the Twist 2.0 + Comp Exome Spike-In and GENEQUALITY designs, but the pipeline supports any number of custom BED/interval inputs.

Additional genome annotations, including truth-set VCFs (Mills, dbSNP, HapMap, Omni, Phase1), are provided to support base-recalibration. Finally, the configuration file includes the full executable path for each bioinformatics tool used by the pipeline.

3.5.2. Coverage-based downsampling workflow

In the first SnakeBin sub-workflow, each ISB is normalized to the same mapped coverage. This is first performed by merging the FastQ files of each sample (specified in the *samples.csv*). The resulting FastQ pair is then trimmed with *fastp* (v0.23.4), which automatically removes adapter sequences via mate overlap analysis, polyG sequences, low quality and short reads. The alignment against the reference genome with *BWA-mem2* (v2.2.1) is then performed.

Afterward, ISB are then generated, by splitting the raw aligned BAM file according to the ISB parameters defined in the *config.yaml* configuration file. The splitting procedure is performed by manually parsing the BAM file and separating each fragment based on its alignment length (TLEN field). Then, each ISB-specific BAM undergoes:

1. removal of duplicate fragments (*Picard MarkDuplicates* v3.4.0),
2. base quality score recalibration (*GATK BaseRecalibration* and *GATK ApplyBQSR* v4.6.2.0),
3. overlapping-mate clipping (*bamUtil clipOverlap* v1.0.15),

4. coverage, uniformity and genotypability quantification (*bedtools* v2.19.1, *GATK CollectHsMetrics* v4.6.2.0).

The user-defined target depths (*config.yaml*) are applied uniformly across bins via random downsampling using *sambamba* (v1.0.1). This ensures that coverage differences do not bias comparisons between insert-size intervals. Normalized ISBs are finally used to compute the final metrics such as mapped fragments (*samtools flagstat* v1.19.2), average coverage (*bedtools coverage* v2.19.1), uniformity of coverage and ON/NEAR/OFF target bases (*GATK CollectHsMetrics* v4.6.2), genotypability (*GATK CallableLoci* v3.8.0), optional variant calling and variant filtering according to *GATK Hard Filters* procedure (*GATK HaplotypeCaller/VariantFiltration* v4.6.2). Additionally, mean, median mode of the fragment's mapping quality are computed. The average insert size of each bin was used as threshold for NEAR target bases detection.

3.5.3. Fragment-based downsampling workflow

A second sub-workflow equalizes the number of input fragments contributing to each ISB. In contrast to the coverage-based downsampling workflow, where coverage is artificially normalized to the same depth, this approach intentionally preserves the effects of insert size and duplication rate, all of which influence the final mapped coverage.

In this workflow, each sample is first trimmed using *fastp* (v0.23.4) and independently aligned to the reference genome using *BWA-MEM2* (v2.2.1). After alignment, ISBs are generated for every sample by splitting the BAM file according to insert size, and each bin is converted back to FASTQ format (*samtools bam2fq*, v1.19.2). Every ISB is then downsampled by sequentially selecting a fixed number of fragments, ensuring that all samples contribute equally to the final dataset.

The downsampled FASTQs belonging to the same ISB are subsequently merged, realigned, deduplicated (*Picard MarkDuplicates*, v3.4.0), base-quality–recalibrated (*GATK BaseRecalibrator/ApplyBQSR*, v4.6.2[47]) and clipped (*bamUtil clipOverlap* v1.0.15). The resulting BAM files are used to compute the same mapping-based metrics and optional variant calling as in the coverage-based workflow.

Although this module is fully implemented in the pipeline, only the results obtained using the coverage-based downsampling workflow are presented in this thesis. This choice was made because coverage normalization enables a direct and controlled comparison of insert-size bins by removing variability introduced by differences in sequencing depth. In contrast, the fragment-based approach preserves additional factors such as duplication rate and capture efficiency, which, while biologically relevant, introduce confounding effects that complicate the interpretation of insert-size–specific performance. Since the primary objective of this thesis is to isolate and quantify the impact of insert size on sequencing performance, the coverage-based workflow was considered more appropriate.

3.5.4. Genotypability Analysis of ClinVar and HGMD Positions Across Insert Size Bins

To assess whether insert size influences the genotypability of clinically relevant positions, we compared the genotypable fraction of ClinVar (2025_03_07) and HGMD PRO (2025.1) variants across the Insert Size Bins (ISBs) generated by the SnakeBin workflow. For this analysis, we used the three batches of Clinical samples, processed through the coverage-normalized (60X) workflow.

For each ISB, genotypability was evaluated by jointly considering:

1. The Twist 2.0 + Comp Exome Spike-In target design regions.
2. GIAB genome stratifications v3.5 “Low Mappability (all)” regions which overlap with the target design.

Callable files, which contain genotypable positions of the genome as reported by *GATK CallableLoci* (v3.8.0), were first intersected with the target design regions mentioned above, the result was then intersected again with ClinVar (full set and pathogenic/likely pathogenic subset) and HGMD (full set and DM/DM?/DFP subset) variants using *bedtools intersect* (v2.31.1)[48]. To evaluate the genotypability of clinical loci in complex-to-align regions of the WES target design, a further intersection with the GIAB low mappability regions mentioned above was performed.

For each ISB, results were averaged across the three Clinical WES sequencing batches. The final output consisted of the percentage of genotypable ClinVar and HGMD sites across insert size bins for (i) the entire target design and (ii) low-mappability regions.

3.5.5. Variant Benchmarking against GIAB GoldSet

The variant calling analyses presented in this thesis were performed on NA12878 replicates, applying SnakeBin pipeline with variant calling and subsequent filtering modules enabled. This allowed us to evaluate the direct impact of insert size on variant detection while avoiding the biological confounding present in multi-individual merges.

Variant calling was performed separately on each insert-size bin (ISB) after coverage normalization to 60X. Benchmarking was carried out using *RTG-Tools vcfeval* (v3.12.1)[49], comparing the called variants against the GIAB NA12878 v4.2.1 Gold Set. Performance was evaluated across three genomic contexts:

1. GENEQUALITY target design, representing the regions targeted by the kit;
2. GIAB High Confidence Regions v4.1.2 for NA12878, enabling variant benchmarking in reliable regions of the genome;
3. GIAB Low Mappability Regions (Genome Stratifications v3.5), allowing assessment of insert-size effects in challenging genomic contexts.

As for the genotypability of clinical loci, GIAB High Confidence Regions and Low Mappability Regions BED files were restricted to those regions overlapping the WES target design. Together, these 22 NA12878 WES replicates provided a controlled experimental setting to quantify how insert size influences variant discovery and accuracy when benchmarked against high-quality truth sets.

3.6. Generation and sequencing of NA12878 PCRfree WGS library for sequencer quality evaluation

To enable a controlled comparison of sequencing performance across technologies, a single whole-genome library of the reference sample NA12878 was prepared using the KAPA HyperPrep DNA PCR-Free kit (Roche) with mechanical fragmentation using the Covaris S220. The PCR-free protocol was selected to avoid

amplification-derived artifacts and ensure that all sequencing platforms processed exactly the same biological material, thereby making any observed differences attributable solely to the sequencing platform.

Four replicates of the same library were prepared and pooled together (before sequencing) in order to obtain sufficient library yield for all sequencing runs.

The resulting WGS library was then sequenced in parallel on four different next-generation sequencing platforms, each operated according to the manufacturer's recommended conditions:

- Illumina NovaSeq X, using a 10B flowcell
- Element AVITI, using a Cloudbreak Freestyle High Output 1B flowcell
- MGI DNBSEQ-T1+ (T1 Plus), using a FCS flowcell
- GeneMind SURFSeq 5000, using a FCM flowcell

For all platforms, 150bp paired-end sequencing was performed using the standard read configuration provided by each instrument.

Although the same starting material was used across all four platforms, differences in library processing should be noted. While Illumina NovaSeq X, GeneMind SURFSeq 5000 and Element AVITI (in Cloudbreak Freestyle modality) are natively compatible with KAPA HyperPrep DNA PCR-Free kit, the workflow for MGI T1 Plus requires an additional conversion step to enable library circularization and the generation of DNA nanoballs (DNBs). This process involves the introduction of a phosphate group through a limited PCR amplification (5 cycles). Potential biases arising from this step were not specifically investigated in the present work and remain an area for future study.

3.7. Workflow for the evaluation of Sequencer accuracy.

To perform the sequencing quality evaluation presented in the second part of this thesis, we developed a dedicated computational workflow implemented in a collection of Python 3 scripts orchestrated by *Snakemake* (v7.32.4). The workflow is designed to quantify, at base-level and read-level, both the empirical alignment accuracy, expressed as identity, and its predicted sequencing accuracy, expressed

as Phred Q-score assigned by each sequencing platform. The full code is publicly available at: https://github.com/LucaBertoli/SnakeMake_Sequencing_Identity

The core processing relies on the *pysam* library, a Python wrapper to *htslib*, which enables efficient access to BAM alignment files and provides optimized utilities for extracting per-base sequence content, Phred quality scores, alignment coordinates, CIGAR operations, and tag information[17], [46], [50].

All analyses are performed on each NA12878 KAPA PCRfree WGS sequencing instance, pre-processed through adapter removal with *fastp* (v0.23.4, with quality and length filters disabled) and aligned to the hg38 reference genome with *BWA-MEM2* (v2.2.1). In order to avoid artefacts arising from poorly assembled loci, the analysis was restricted to the GIAB NA12878 high-confidence regions (v4.2.1). To reduce noise originating from poorly mapping reads, a stringent mapping quality filter was applied, excluding from the analyses all the reads with mapping quality less than 60. Additionally, as described in the next sections, to prevent true biological variants and limit systematic alignment artefacts to be considered errors, any mismatches or INDELs overlapping SNVs/INDELs in a input VCF files were not considered sequencing errors. For this study, provided that the dataset comprehends four different sequencing instances of the same library, the input variants used as error filters comprehend the common variants called by GATK HaplotyperCaller v4.6.2.0 on the raw BAM files of each sequencer. The common variants, shared by each sequencer were identified first by performing variant normalization and left-alignment and then by intersecting the results respectively with bcftools norm and isec v.1.21.

3.7.1. Base-Level Accuracy Stratification

To characterize how alignment base accuracy relates to the sequencer Q-scores, the workflow first computes the alignment identity separately for every Phred score observed in the dataset. Using *pysam*, it evaluates each aligned base and determines whether it is a match or mismatch relative to the reference, applying the VCF-based error filtering.

Insertions are recorded by considering each inserted base individually as either correct or erroneous, while deletions, by not having associated base-level Q-score,

are excluded from this analysis. Clipped bases are never included, and supplementary or secondary alignments are ignored.

Using *get_aligned_pairs* pysam function, the workflow loops every aligned base of each read, counting in a specific dictionary data-structure the correct/erroneous mismatches and correct/erroneous insertions. These counters are stored in both unfiltered and VCF-filtered manner. The latter are later used to compute the alignment identity with the following formula:

$$\text{Identity(Q)} = 1 - \frac{Mismatch\_errors(\text{Q}) + Insertion\_errors(\text{Q})}{\text{Match(Q)} + \text{Mismatch(Q)} + \text{Insertions(Q)}}$$

Where the *Identity(Q)* is the identity of the bases of a specific sequencer Q-score, *Mismatch_errors(Q)* and *Insertion_errors(Q)* are VCF-filtered counts of mismatch and inserted bases with a specific sequencer Q-score, *Match(Q), Mismatch(Q)* and *Insertions(Q)* are the total number of matching, mismatching and inserted bases with a specific sequencer Q-score.

The resulting filtered alignment identity is graphically represented as a curve across sequencer-assigned Q-score levels, which serves as a direct benchmark for evaluating the correctness and calibration of each platform's quality-scoring model. To facilitate visual comparison with the sequencer-assigned Q-scores (especially at high quality levels), the per-base identity values were also converted into alignment-based Q-scores using the standard Phred transformation formula:

$$Alignment\ Qscore = -10 * log10(1 - Identity)$$

3.7.2. Definition of a Unified Binning Scheme for sequencing Q-scores

As shown later in the Results section, the Illumina NovaSeq X and GeneMind SURFSeq 5000 platforms do not emit the full range of Phred quality values; instead, they adopt proprietary binned Q-score schemes, in which base qualities are discretized into predefined levels. To improve cross-platform comparability, a Unified Binning Scheme was applied, grouping all sequencer-assigned Q-scores into three categories:

- Low Qscore (Q0–19),
- Medium Qscore (Q20–29),

- High Qscore (Q30+).

All bases were reassigned to one of these bins and treated accordingly.

Using this harmonized quality framework, the base-quality-stratified alignment-identity analysis are repeated, aggregating bases from each platform into the Unified Binning Scheme prior to computing alignment identity.

3.7.3. Stratification of Sequencer Quality and Error Rates by Sequencing Cycle

To investigate how sequencing accuracy evolves along the read, the workflow extends the base-level analysis by stratifying both VCF-filtered Error Rates and the sequencer-assigned Q-score as a function of the sequencing cycle. For each read, only primary alignments are considered, soft- and hard-clipped bases are excluded, and all mismatches/insertions/deletions overlapping input variants are filtered according to the procedures described above.

Error Rates were calculated independently for each sequencing cycle, by aggregating all bases sequenced at a given cycle and computing the fraction of matches after VCF-filtering. In this way, Error Rates are evaluated solely as a function of cycle, without stratification by base quality.

In parallel, the sequencer accuracy is computed for each cycle by averaging the Phred Q-scores of all bases emitted at that cycle. Since Phred scores are logarithmic, base-level qualities are first converted to error probabilities; these probabilities are then averaged across the cycle, and the result is transformed back into a Phred Q-score. This produces a per-cycle estimate of the accuracy predicted by the instrument.

3.7.4. Sequencing Cycle stratification under the Unified Quality Binning Scheme

To further inspect the relationship between base quality and sequencing cycle, the workflow extends the previous analysis by grouping the bases according to the Unified Binning Scheme. For each sequencing cycle and each quality bin, two metrics are computed:

1. Alignment identity, based on VCF-filtered matches and mismatches within that cycle and quality class.
2. Fraction of bases emitted at that cycle that belong to the corresponding quality bin, normalized by the total number of bases at that quality bin.

As in the previous analysis, only aligned bases from primary reads are retained, and overlapping input variants are removed from the error set.

This joint stratification allows the workflow to distinguish whether cycle-dependent biases arise from true identity fluctuation within each Q-score class, or from a shift in the distribution of emitted qualities as sequencing progresses.

3.7.5. Computation of Read-Level Alignment Identity

To bridge from base- to read-centered quality evaluation, the workflows integrates modules for computing average read sequencing quality, error counts read alignment identity. To do so, the workflow uses pysam to scan every aligned read and quantify all discrepancies between the read and the reference, integrating now both mismatches, insertions and deletions.

The script filters the input data to ensure that unmapped, secondary, or supplementary alignments are discarded. This guarantees that identity measurements reflect only primary, uniquely aligned reads.

Like in the previous analyses, within each retained read, only the aligned portion is considered, Soft-clipped bases are not included. As in the per-base analysis, mismatch positions are reconstructed from the MD tag, while inserted and deleted bases are extracted from the CIGAR string. All observed mismatches and indels are then VCF-filtered, removing from the error counts any discrepancy that corresponds to input variants (in this study common variants called on the four platforms).

The script outputs, for every read, a set of identity measures:

- mismatch-only identity;
- mismatch-only VCF-filtered identity;
- identity including indels;

- VCF-filtered indel-inclusive identity.

These metrics are accompanied by detailed per-read statistics such as aligned read length, the number of mismatches, and the number of inserted or deleted bases (both filtered and unfiltered).

The per-read identity files are then aggregated, computing the overall distribution of errors across all reads in a dataset. This includes the average VCF-filtered indel-inclusive identity and the proportion of reads containing zero, one, two, three, or more sequencing errors.

3.7.6. High-Quality reads analyses

In parallel with whole-dataset analyses, dedicated pipeline modules processed high-quality reads (average quality score ≥30, defined as the mean of underlying per-base error rates).

Per-read identity and error distribution were first computed for these reads. They were then reprocessed to stratify VCF-filtered errors by read-level error count (1, 2, 3, or >3 errors), deriving the proportion of bases with low/medium/high-quality (Q0–19, Q20–29, Q≥30) and exporting error genomic coordinates to BED files for overlap and stratification analyses.

Platform-specific error overlaps were generated using Python's UpSetPlot library; genomic context stratification employed bedtools intersect v2.31.1 against GIAB v3.5 BED files:

- GRCh38_AllTandemRepeats.bed.gz
- GRCh38_AllHomopolymers_ge7bp_imperfectge11bp_slop5.bed.gz
- GRCh38_gclt25orgt65_slop50.bed.gz).

3.7.7. Comparison of Sequencer Q-scores and Error Rates across Insert Size Intervals

To evaluate whether the accuracy of sequenced reads depends on the length of the underlying DNA fragments, and to complement the cycle-stratified analysis, a dedicated analysis module was developed, that jointly computes the sequencer-assigned read quality distribution and error rates, stratified by insert size.

3.7.8. Per-read extraction of insert size, sequencer quality and alignment identity

First, the aligned BAM file is iterated using *pysam* excluding uninformative reads and bases (unmapped, secondary, or supplementary reads and clipped bases).

For every valid read, four key metrics are computed:

1. Insert size, obtained from the template length for paired-end data or from the read length for unpaired reads.
2. Average sequencer-assigned Q-scores, calculated as in the previous sections, by a logarithmic average, across the read length, of the underlying sequencer-assigned base error rates.
3. VCF-filtered mismatch and indel error counts.
4. VCF-filtered indel-inclusive alignment identity.

The script produces a compressed TSV file containing for each primary alignment, in addition to the mentioned metrics, the read name and read orientation (R1/R2).

3.7.9. Stratification of accuracy by insert size

The TSV file generated in the previous step is then processed aggregating the per-read values into insert size bins of 50 bp (150–800 bp). Leveraging the read orientation tag reported in the TSV file, the comparison was performed separately for Read 1 and Read 2.

Graphical visualization is performed by generating, for each insert size bin, distributions (violin plots) of the average read sequencer Q-score and plots representing the error rates of the bases assigned to each bin. These paired distributions allow a direct visual comparison of how predicted and observed accuracy vary as fragment length increases and between mates.

3.7.10. Cross-platform comparison of variant calling accuracy

The four platforms were then compared in terms of variant calling performance. First, the aligned BAM files were post-processed by performing duplicate removal with *Picard MarkDuplicates* (v3.4.0), base recalibration with *GATK BaseRecalibration/ApplyBQSR* (v.4.6.2.0), and *bamUtil ClipOverlap* (v.1.0.15), Afterward, in order to ensure even coverage conditions, downsampling was

performed with *sambamba* (v.1.0.1) to a target average mapped coverage of 30X. Variant calling was performed with *GATK HaplotypeCaller* (v4.6.2) and GATK HardFilters filtering procedure was applied. The resulting variants were compared against NA12878 GIAB GoldSet (v4.2.1) using *RTG-Tools vcfeval* (v3.12.1), stratifying the metrics using the following GIAB genome stratifications (v.3.5):

- High Confidence Regions
    - HG001_GRCh38_1_22_v4.2.1_benchmark.bed
- Tandem Repeats
    - GRCh38_AllTandemRepeats.bed
- Homopolymers
    - GRCh38_AllHomopolymers_ge7bp_imperfectge11bp_slop5.bed

The described evaluation was performed on PASS variants, both on the entire callset and separately for SNVs and INDELs. The latter two were separated using GATK SelectVariants (v4.6.2). The resulting metrics were graphically visualized with a custom *python matplotlib* script.

4. Results

4.1. Characterization of Insert Size in WES experiments

4.1.1. Most exons are short, but half of the exonic bases lie in long regions.

	Regions		Bases	
Size	Number	Percentage (%)	Number	Percentage (%)
All	287,879	100	37,453,133	100
<= 100 bp	149,088	52	5,333,367	14
101 – 200 bp	101,269	35	14,358,939	38
201 – 300 bp	20,998	7	4,971,696	13
>= 301 bp	16,524	6	12,789,331	34

Table 1 For different length of Twist 2.0 + Comp exome Spike In target design regions, the number of regions, number of bases and relative percentage with respect to the total.

To characterize the exon size distribution in our WES data set, we analyzed the length of the regions included in the Twist 2.0 + Comp exome Spike In target design. We first calculated the mean and median region lengths. Then, to further assess the predominance of short and long regions, we determined the percentage of regions falling into the following length categories: less than 100 bp, 101–200 bp, 201–300 bp, and greater than 300 bp (Table 1).

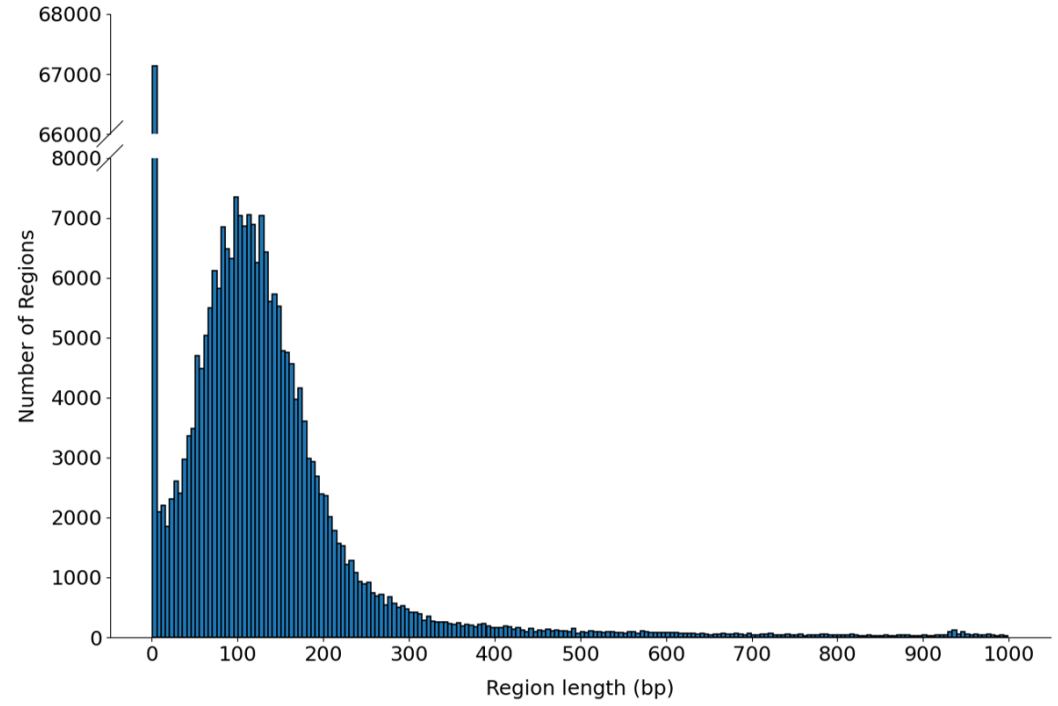


Figure 9 Distribution of the number of regions for each length, computed on Twist 2.0 + Comp exome Spike In target design.

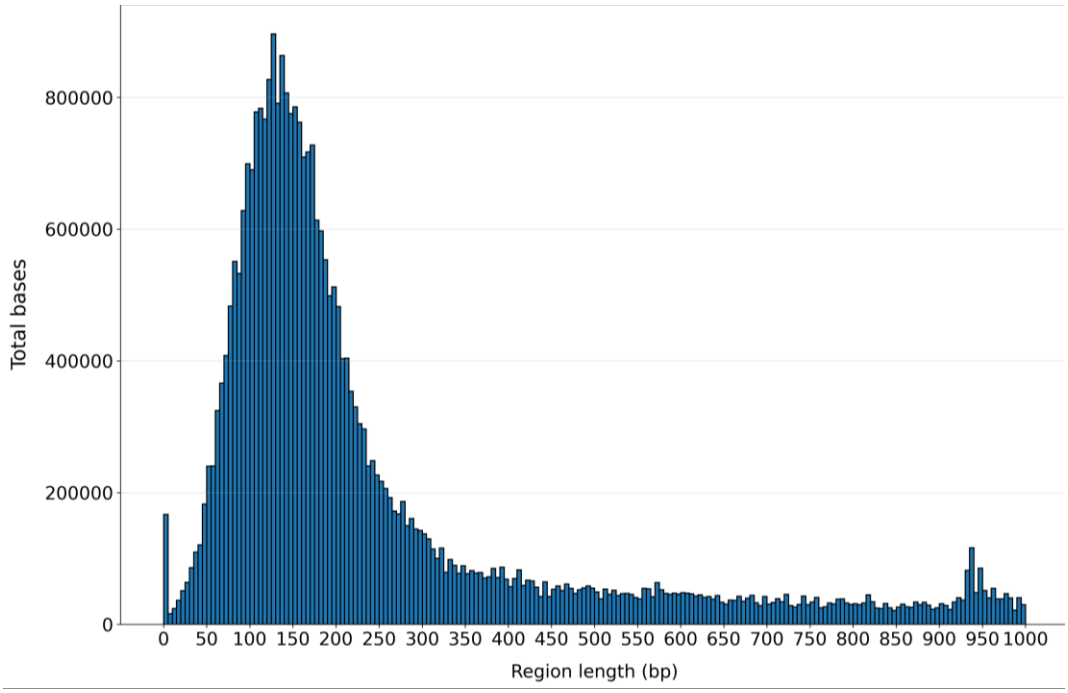


Figure 10 Distribution of the bases for each region length, computed on Twist 2.0 + Comp exome Spike In target design.

The regions of interest (ROI) exhibit relatively short design lengths, with a mean of 130 bp and a median of 97 bp. This can be explained by the high number of short regions in the target design. In fact, 52% of the design regions have lengths shorter than 100bp, and 35% between 101 and 200bp (Table 1, Figure 9). However, despite the predominance of short regions, a significant proportion of the bases is concentrated in longer regions: nearly half (47%) of the total exome bases lie in exons longer than 200 bp. In particular, 13% of the bases belong to regions 201-300bp long, and 34% in regions longer then 301bp (Table 2, Figure 10). This highlights an asymmetric distribution of exon sizes, where a small number of longer regions contribute substantially to the overall coverage of the exomic genome.

4.1.2. Not only insert size matters - but also its distribution.

Commercial WES protocols (e.g., Twist, Roche) are typically optimized to generate libraries with an average insert size of approximately 200 bp. However, previous findings have shown that longer inserts improve multiple aspects of WES performance, including alignment quality and variant calling[10]. In real datasets, insert size does not converge to a single representative value but instead follows a Gaussian-like distribution. As a consequence, sequenced samples present fragments with different lengths, which contribute unevenly to coverage and mappability, making the mean insert size alone insufficient to predict performance.

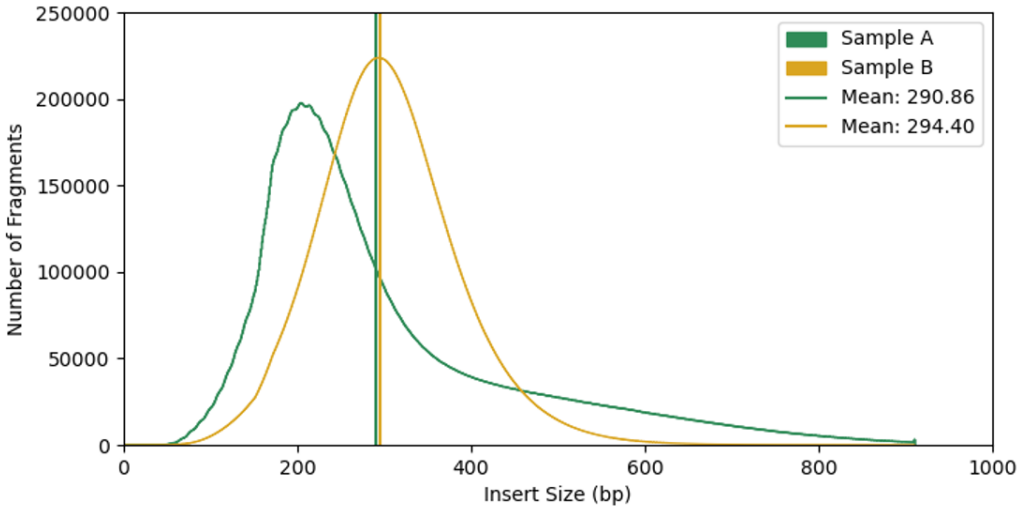


Figure 11 Distribution of insert size of two Illumina Exome 2.5 representative samples. Sample A presents a symmetrical gaussian distribution, while Sample B a right skewed distribution.

The two examples shown in Figure 11 demonstrate that insert size distributions with the same mean can nonetheless differ substantially in shape. Both samples exhibit an average insert size of ~290 bp, but their distributions are markedly different.

In Sample A, despite the expected mean fragment length defined by the protocol, the distribution still contains a proportion of both shorter and longer fragments. This pattern, represented by a symmetric Gaussian shape, reflects the natural variability introduced during library preparation and indicates that even well-optimized WES protocols generate a spread of insert sizes around the target mean.

In contrast, Sample B illustrates how insert size distributions can deviate substantially from symmetry. Here, the distribution is not centered around the mean but instead shows a marked right skew, with a large fraction of fragments shorter than the average insert size and a long tail toward larger fragments. Such skewed distributions reveal that samples with the same nominal mean insert size may in practice contain very different proportions of short and long fragments, potentially influencing both coverage efficiency and alignment performance.

4.1.3. Development of the SnakeBin Pipeline for Insert-Size-Resolved WES Analysis

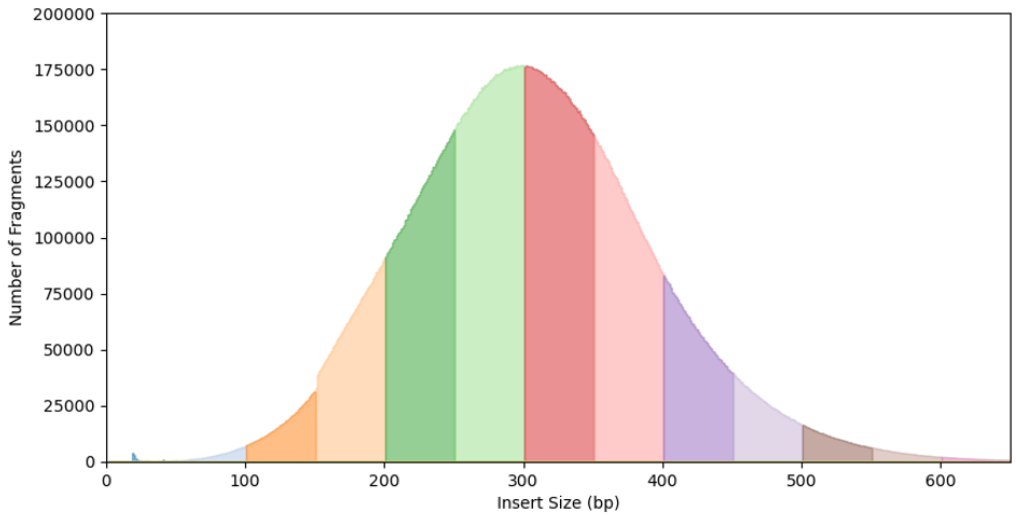


Figure 12 Histogram of fragment lengths showing the full insert-size distribution. Colors represent 50bp insert size intervals.

A systematic evaluation of insert size in whole-exome sequencing requires comparing fragments that differ not only in length but also in abundance, coverage, and mapping properties. As the insert size distribution of real WES samples typically follow a Gaussian-like shape, mid-sized fragments are highly abundant, whereas very short and very long fragments occur only rarely. As a result, the tails of the distribution contain too few reads to be reliably analyzed in a single experiment (Figure 12).

To enable a meaningful and robust characterization of all insert-size intervals, the aggregation of multiple samples into large merged datasets was needed. Although merging reads originating from different individuals prevents biologically valid variant calling, it is fully appropriate for computing mapping-based statistics (such as coverage, ON/NEAR/OFF-target rates, mapping quality, and genotypability) and crucially provides enough fragments to analyze even the sparsely populated insert-size bins.

Because no existing workflow supported this type of insert-size-aware analysis, I developed SnakeBin, a new Snakemake-based pipeline tailored for generating and evaluating insert-size normalized WES datasets. SnakeBin addresses the two central challenges of this problem:

1. Insert-size distributions are uneven, leading to strong imbalances across bins.
2. Comparisons must be made on depth- or fragment-normalized datasets rather than raw read counts.

The pipeline automatically:

- merges multiple samples to ensure adequate representation across the entire insert-size distribution,
- partitions the combined dataset into fixed-width Insert Size Bins (ISBs),
- normalizes each ISB (either by mapped coverage or by fragment count), and
- computes ISB-specific metrics including mapped coverage, average/median/modal fragment mapping quality, ON/NEAR/OFF-target bases, and optionally variant calling.

Figure 13 provides an overview of this structure: multiple WES samples are merged, the full insert-size distribution is divided into predefined 50 bp ISBs, and each interval is processed independently. Full implementation details of SnakeBin are provided in the Methods section.

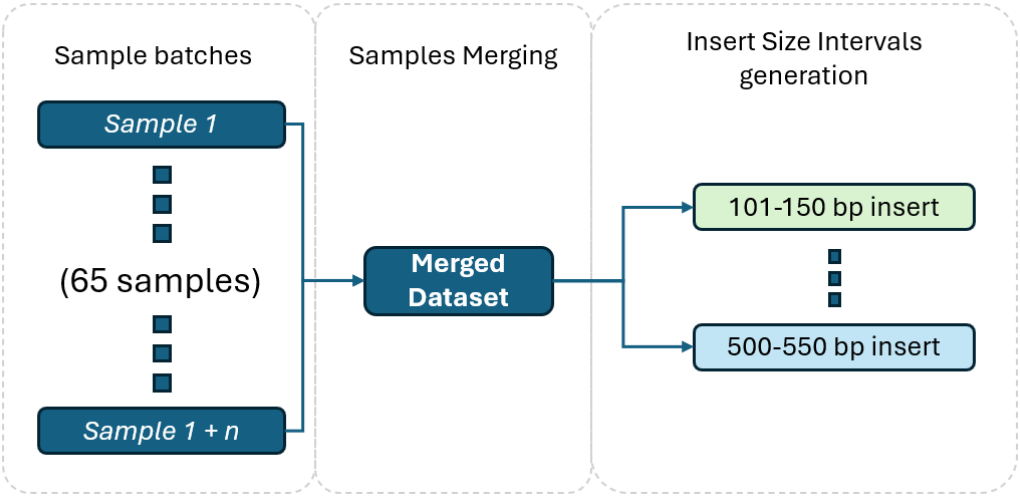


Figure 13 Schema showing how in the developed pipeline, multiple WES samples from the same batch are merged to create a high-coverage dataset, which is then split into predefined insert-size intervals (ISBs).

4.1.4. Insert Size Distribution and Coverage Across Three WES Batches

To obtain sufficient coverage for downstream analyses, we selected three batches of 65 clinical WES samples each and processed them in parallel using the SnakeBin Coverage-based downsampling workflow.

All samples were analyzed across the full insert-size range (1-1000 bp, 50 bp intervals), resulting in a total of 2,582,958,055, 2,492,952,110, and 2,514,895,020 fragments for Batch 1, Batch 2, and Batch 3, respectively (Table 2).

ISB	Batch 1		Batch 2		Batch 3	
	Nr of Fragments	Mapped Coverage	Nr of Fragments	Mapped Coverage	Nr of Fragments	Mapped Coverage
1-50	229,545	0.36	174,279	0.30	234,386	0.33
51-100	4,636,449	7.59	3,778,558	6.51	4,814,755	7.72
101-150	**43,991,165**	**80.78**	**36,734,860**	**70.13**	**44,978,827**	**82.68**
151-200	**210,932,618**	**412.54**	**182,742,342**	**371.36**	**206,875,575**	**409.00**
201-250	**494,301,497**	**995.45**	**463,328,928**	**962.58**	**483,765,223**	**989.64**
251-300	**584,184,271**	**1,258.68**	**581,719,571**	**1,282.97**	**574,068,089**	**1,261.44**
301-350	**475,966,730**	**1,043.47**	**479,640,217**	**1,076.18**	**463,759,087**	**1,041.42**
351-400	**327,971,356**	**675.15**	**326,456,216**	**689.73**	**315,787,507**	**667.32**
401-450	**203,533,101**	**395.16**	**198,354,323**	**396.53**	**194,019,291**	**386.92**
451-500	**116,445,924**	**213.35**	**110,724,785**	**209.46**	**110,494,820**	**207.72**
501-550	**62,191,006**	**107.73**	**57,596,053**	**103.25**	**59,023,840**	**104.66**
551-600	31,322,611	51.46	28,212,469	48.06	29,866,283	50.07
601-650	14,983,157	23.46	13,131,329	21.36	14,444,023	23.01
651-700	6,890,814	10.35	5,882,501	9.19	6,788,192	10.33
701-750	3,074,915	4.48	2,558,580	3.88	3,144,402	4.62
751-800	1,294.354	1.83	1,066,526	1.56	1,434,086	2.04
801-850	568,801	0.79	471,596	0.68	700,944	0.97
851-900	258,322	0.36	218,869	0.31	368,498	0.51
901-950	120,848	0.17	105,416	0.16	206,443	0.29
951-1000	60,571	0.10	54,692	0.10	120,749	0.18

Table 2 For each batch of clinical samples analyzed, the number of mapped fragments in each ISB and the resulting mapped coverage.

As expected for WES libraries, the insert size bins (ISBs) followed a Gaussian-like distribution, with fragment abundance peaking in the central portion of the distribution and decreasing toward the tails. Across the three batches, the least populated ISBs were those between 951–1000 bp, containing only 54k–120k fragments, corresponding to an average coverage of 0.10–0.18X. Conversely, the most populated ISBs were those between 251–300 bp, with 574M–584M fragments, reaching extremely high coverage levels of ~1250–1280X (Table 2).

Coverage scaled proportionally with fragment abundance across all bins. For instance, the 101–150 bp ISB reached 70–83X, the 151–200 bp bin 371–413X, and the 201–250 bp bin 963–995X, in line with their fragment counts. The central interval (251–300 bp) consistently exhibited the highest coverage across all batches (1259–1283X), while coverage dropped below 1X in bins above 751 bp (Table 2).

For downstream analyses, we retained only the insert size intervals that reached a minimum mapped coverage of at least 60X across all three batches (101bp - 550bp),

ensuring that all selected ISBs contained enough data for reliable and comparable performance assessments (Table 2).

4.1.5. Evaluation of Coverage Efficiency across ISBs

To quantify sequencing efficiency across insert size intervals, we estimated the number of mapped fragments required to reach 60X coverage for each insert size bin (ISB).

As shown in Figure 14, the number of fragments needed to achieve 60X coverage varies substantially across the insert size spectrum and follows a U-shaped convex trend.

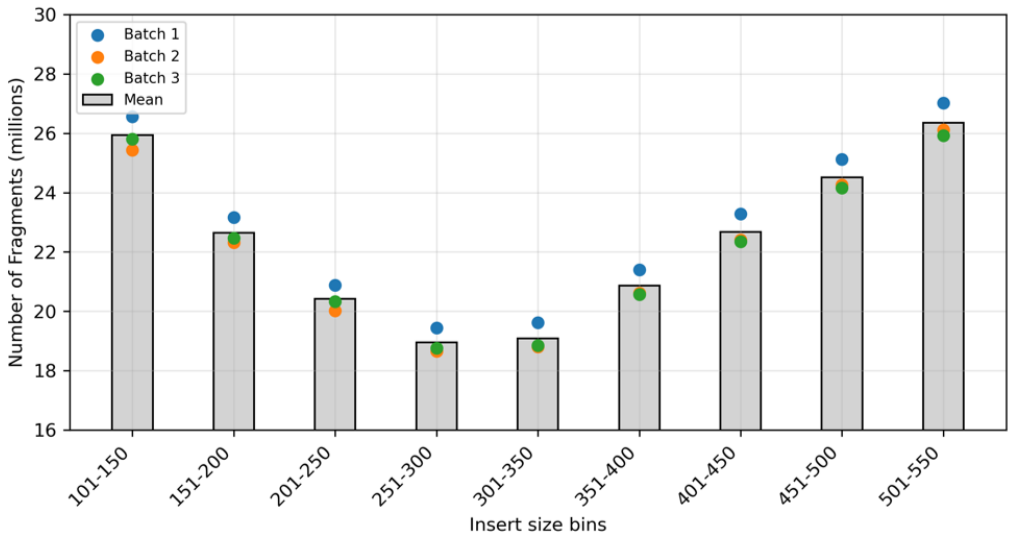


Figure 14 Dotplot representing the number of fragments needed for 60X mapped coverage and bars representing the average of the three sequencing runs.

Short insert sizes required the largest sequencing effort. The 101–150 bp interval needed the highest number of fragments (~26 million), while the 151–200 bp interval required slightly fewer fragments (~23 million). The number of required fragments decreased progressively with increasing insert size and reached a minimum in the 251–300 bp interval, which consistently required only ~19 million fragments across all three batches.

Relative to this optimal range, shorter fragments were markedly inefficient:

- The 101–150 bp bin required 1.34× more fragments than the 251–300 bp bin (~34% additional sequencing).
- The 151–200 bp bin required 1.20× more fragments than the 251–300 bp bin (~20% additional sequencing).

Efficiency decreased again for longer fragments. Insert sizes above ~350 bp showed a progressive increase in the number of fragments required to reach 60X, reaching ~26 million fragments in the 501–550 bp interval, consistent with the expected loss of effective coverage due to reads extending outside the boundaries of target regions.

Overall, these results indicate that insert sizes between 251 and 300 bp minimize the sequencing effort required to reach 60X mapped coverage, whereas both shorter (<200 bp) and longer (>400 bp) fragments require substantially more mapped fragments to achieve the same depth.

4.1.6. Effect of Insert Size on ON-, NEAR-, and OFF-Target Bases

After assessing the number of fragments required to attain 60X mapped coverage across insert size intervals, we next examined how fragment length affects the distribution of ON-, NEAR-, and OFF-target bases. This analysis provides insight into why certain insert sizes require substantially more sequencing effort, and how fragment length influences the efficiency with which bases map within the intended target regions.

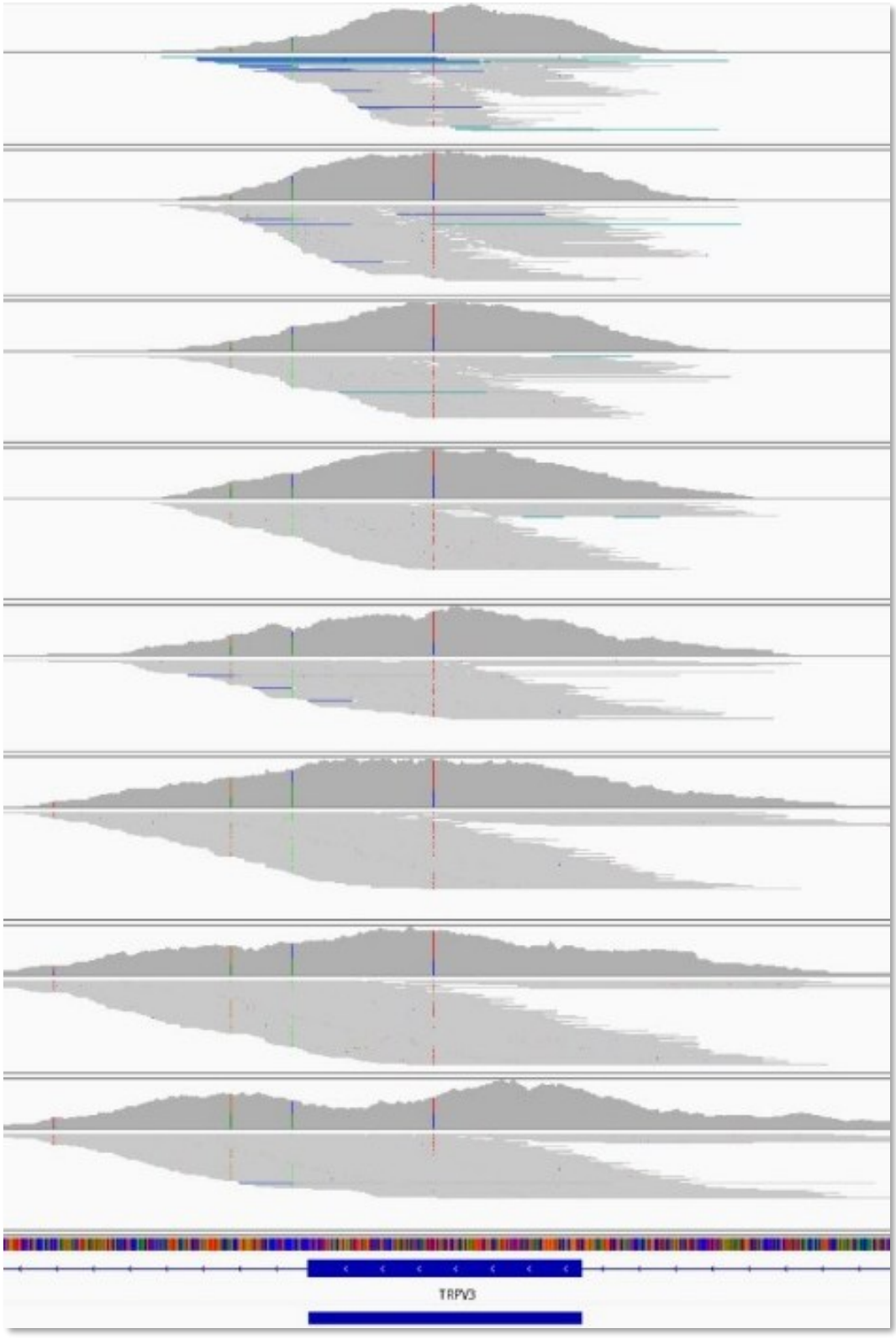


Figure 15 IGV screenshot from Batch 1, representing the 60X mapped coverage ISB alignments, from 101-150 to 450-500bp, on the lower part the blue region represents a target region of Twist 2.0 + Comp. Exome Spike In capture kit.

Figure 15 shows an IGV screenshot with representative BAM alignments extracted from one of the three sequencing batches, covering ISBs from 101–150 bp up to 451–500 bp. As the insert size increases, the portions of each fragment extending outside the targeted regions become progressively larger. Short fragments are almost entirely contained within the capture intervals, whereas long fragments span far beyond the boundaries of the design, resulting in a growing amount of read sequence mapping outside the intended target.

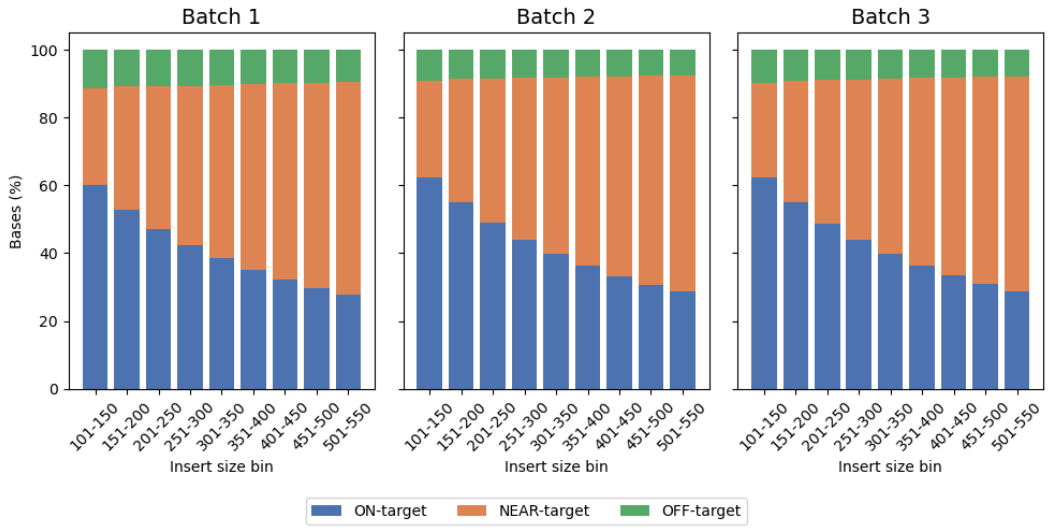


Figure 16 Barplot showing the percentage of bases that map ON target, NEAR target and OFF target. Near target distance was defined as the mean insert size observed in the sample.

Figure 16 reports the percentage of ON-target, NEAR-target, and OFF-target bases for each insert size bin across the three batches at 60x fold coverage. The three barplots reveal consistent patterns: although the fraction of OFF-target bases remains stable across intervals, the proportion of ON-target bases decreases steadily from approximately 60% in the shortest insert size bin (101–150 bp) to about 30% in the longest (501–550 bp). This decline is accompanied by a symmetrical increase in NEAR-target bases, indicating that longer fragments tend to extend immediately beyond the target edges rather than mapping deep into off-target regions.

Together, these observations explain the coverage inefficiencies observed for longer fragments and reinforce the importance of insert size optimization. Larger inserts disproportionately increase the amount of sequence falling just outside the designed intervals, lowering effective ON-target yield. Fine-tuning the insert size

is therefore essential to maximize ON-target mapping and minimize unnecessary overhangs beyond the target regions.

4.1.7. Insert Size Impact on Mapping Quality

After examining how insert size influences the distribution of ON-, NEAR-, and OFF-target bases, we next evaluated its effect on alignment accuracy by analyzing both median and modal Mapping Quality (MQ) across insert size bins (ISBs). MQ values were computed at normalized 60X mapped coverage for each ISB in all three batches, considering the entire target design as well as the GIAB-defined low-mappability subset (Figure 17).

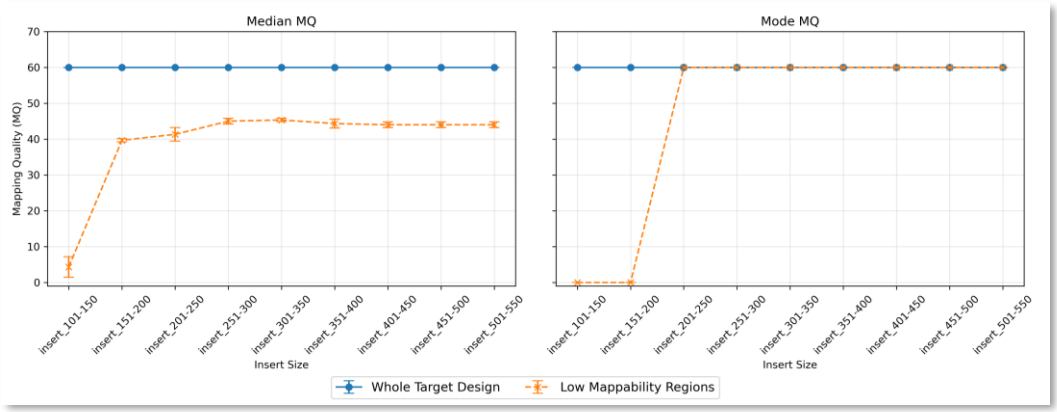


Figure 17 Plots showing, for each ISB, the median and modal mapping quality. Points represent the average of the three batches and the whiskers representing the standard deviation. In each plot, both Twist 2.0 + Comp Exome Spike In target design and GIAB genome stratification v3.5 Low Mappability Regions.

Across the whole target design, median MQ remained uniformly high for all insert size intervals, consistently close to the maximum value of 60. Similarly, the modal MQ was equal to 60 for all ISBs, showing that the majority of reads are assigned the highest mapping quality regardless of fragment length. These results confirm that alignment within targeted regions is generally stable and only minimally affected by insert size.

When restricting the analysis to low-mappability regions, however, the behavior of both metrics changed markedly. For the shortest fragments (101–150 bp and 151–200 bp), the modal MQ dropped to 0 in all three batches, indicating that most reads in these bins receive ambiguous or unreliable mapping scores due to the presence of repetitive genomic sequences, which hinder accurate and unambiguous read

placement. The median MQ was also substantially reduced for these short ISBs, reflecting the overall difficulty in confidently aligning short fragments within such challenging genomic contexts.

As insert size increased, both MQ metrics improved noticeably. Starting from ISBs ≥201 bp, the modal MQ reached 60, and the median MQ increased sharply, indicating a strong recovery in alignment confidence. For longer fragments, MQ values stabilized, showing consistently high mapping reliability across all batches.

Together, these observations show that median MQ captures the general trend of improvement, while modal MQ highlights the abrupt shift from ambiguous to confident mapping as insert size crosses the 200 bp threshold. This combined analysis makes clear that long fragments greatly enhance alignment accuracy specifically in low-mappability regions, an effect that remains hidden when evaluating only the full target design (Figure 17).

4.1.8. Insert Size effect on Genotypable bases

Following the analysis of mapping quality across insert size intervals, we next evaluated how fragment length influences the proportion of genotypable bases at a normalized depth of 60X. For each insert size bin (ISB), we computed the percentage of target bases meeting all GATK CallableLoci requirements across the three sequencing batches (minimum of 4x coverage, considering reads with MQ>=10, bases with BQ >=20 and at most 10% of the total reads with MQ<10). The results are showed in Figure 18.

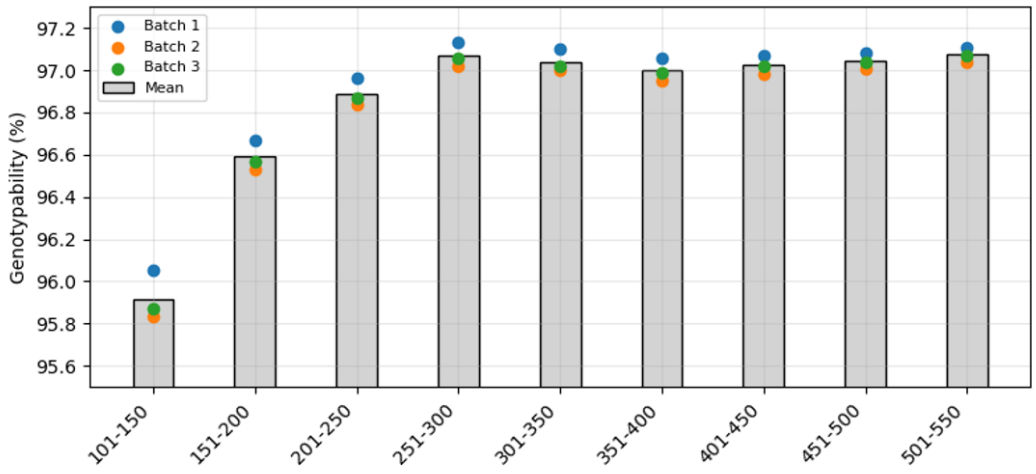


Figure 18 Dotplot representing the percentage genotypability and bars representing the average of the three batches of samples.

A clear dependence on insert size was observed. Genotypability decreased progressively for insert sizes below 251 bp, with the 101–150 bp interval showing the lowest values (approximately 95.8–95.9% across batches). Insert sizes between 151–200 bp showed partial improvement, but still did not reach the levels achieved by larger fragments.

The highest genotypability was consistently observed in the 251–300 bp interval, reaching 97.0–97.1% across all batches. Beyond this range, genotypability remained stable and did not show further meaningful improvement up to at least 550 bp. These trends closely reflect the mapping quality behavior, confirming that fragments shorter than ~250 bp exhibit reduced alignment confidence. Once insert size exceeds ~250 bp, both mapping quality and genotypability plateau at high levels, indicating that fragments in this range provide sufficient unique genomic context to resolve ambiguities (Figure 18).

Overall, these results demonstrate that genotypability declines only for insert sizes below ~250 bp, whereas insert sizes above 251 and 300 bp maximize the proportion of callable bases. Importantly, increasing insert size beyond this range does not yield additional gains once the ~250 bp threshold is surpassed.

4.1.9. Genotypable positions of clinical variant databases

Building on the improvements observed in mapping quality and overall genotypability for insert sizes above 250 bp, we next assessed whether these effects also enhance the genotypability of clinically relevant positions. Using ClinVar and HGMD as reference databases, we quantified the percentage of genotypable clinically relevant loci at 60X mapped coverage for each insert size bin, both across the whole target design and within GIAB-defined low-mappability regions (Figure 19).

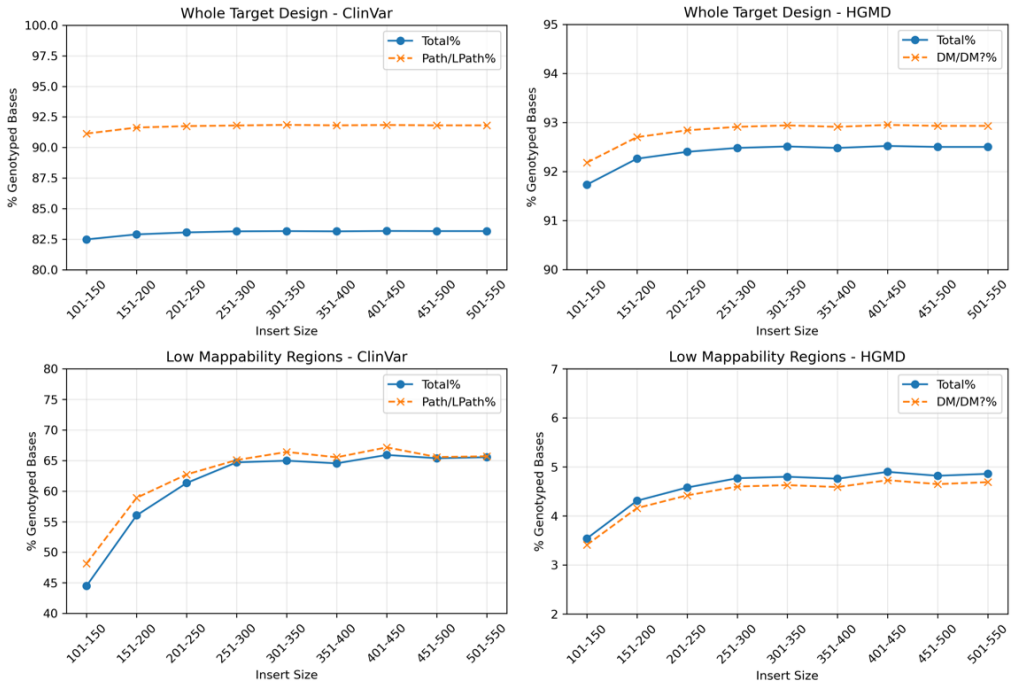


Figure 19 In both Twist 2.0 + Comp Exome Spike In target design and GIAB genome stratification v3.5 Low Mappability Regions ("all" subset) , the percentage of genotypable bases, which correspond to variants inside ClinVar and HGMD variant databases, for both the

Across the full target design, longer fragments consistently increased the genotypability of clinically relevant positions. Transitioning from short inserts (101–150 bp) to the 251–300 bp interval improved the genotypability of ClinVar and HGMD positions by approximately 1–2%. In absolute terms, this corresponds to an increase of ~23,000 ClinVar positions, including ~2,000 pathogenic or likely pathogenic sites, and ~3,700 HGMD positions belonging to the DM/DM?/DFP classes.

These effects were markedly stronger in low-mappability regions. In these challenging genomic contexts, ClinVar genotypability increased by up to 20% when moving from short to longer fragments, corresponding to ~14,000 additional ClinVar positions, including ~1,200 pathogenic or likely pathogenic sites. Similarly, HGMD genotypability increased by ~1%, representing ~2,500 additional DM/DM?/DFP positions becoming genotypable. These improvements closely parallel the mapping quality trends, where longer inserts alleviated alignment ambiguity introduced by repetitive genomic sequences (Figure 19).

Overall, these findings demonstrate that the advantages of longer insert sizes extend beyond general alignment performance and directly enhance the genotypability of clinically relevant positions, particularly within low-mappability regions. Importantly, increasing the insert size above ~250 bp enables the analysis of clinically relevant loci that are otherwise not genotypable with shorter fragments, underscoring the critical role of insert size optimization for comprehensive clinical variant assessment.

4.1.10. Analysis of NA12878 WES Replicates

To determine whether the effects of insert size on mapping quality and genotypability observed in clinical WES samples also translate into differences in variant calling, we next analyzed a dataset consisting of 22 WES replicates of the Coriell NA12878 reference cell line processed with GENEQUALITY kit. Because all samples derive from the same individual's DNA, they can be merged without introducing biological heterogeneity, allowing us to perform variant calling directly on each insert size bin (ISB). This enables a direct evaluation of whether the genotypability trends seen previously, particularly the improvements for insert sizes above ~250 bp, correspond to measurable differences in variant detection. Using NA12878 also allows benchmarking against the GIAB Gold Set, providing a high-confidence reference for assessing variant-calling accuracy.

As in the previous analyses, insert size bins were generated in 50 bp intervals, using only ISBs that reached at least 60X mapped coverage after merging all 22 samples. In the NA12878 dataset, this threshold was met only for bins between 101–150 bp and 351–400 bp; shorter and larger insert sizes did not accumulate enough fragments to reach the required depth. The SnakeBin outputs that underwent coverage normalization to 60X were used. The number of mapped fragments, the achieved coverage, and the corresponding genotypability for each selected ISB are summarized in Table 3.

Insert Size Bin	**#Mapped Fragments**	**Coverage (X)**	**%Genotypability**
101-150	24.277.742	60,23	96,28
151-200	20.070.617	59,81	96,87
201-250	17.479.881	59,88	97,21
251-300	15.640.248	59,96	97,40
301-350	15.224.009	60,03	97,42
351-400	16.474.698	60,82	97,38

Table 3 For the NA12878 ISB that reached 60X fold coverage, the number of mapped fragments, the average mapped coverage and the percentage of genotypable bases of the target design.

Consistent with the trends observed in the clinical WES samples, genotypability increased with insert size also in this dataset, confirming that shorter fragments

(<200 bp) exhibit reduced genotypability, in line with the mapping quality limitations described in the previous chapters. In the NA12878 replicates, however, the differences between the 201–250 bp and 251–300 bp intervals were less pronounced than those observed in the clinical cohort. This softer transition may reflect technical differences between the two capture designs used in the study (Twist for the clinical WES and Genequality for NA12878), which could influence how specific genomic regions respond to changes in insert size, although no formal comparison between designs was performed here.

4.1.11. Variant Calling Performance across ISBs on NA12878

Building on the genotypability results obtained for the NA12878 insert size bins, we next evaluated whether these differences translate into measurable changes in the number of detected variants and in the accuracy of variant calling. Variant calling was performed separately on each ISB that reached 60X mapped coverage, using the downsampled BAMs produced by SnakeBin pipeline.

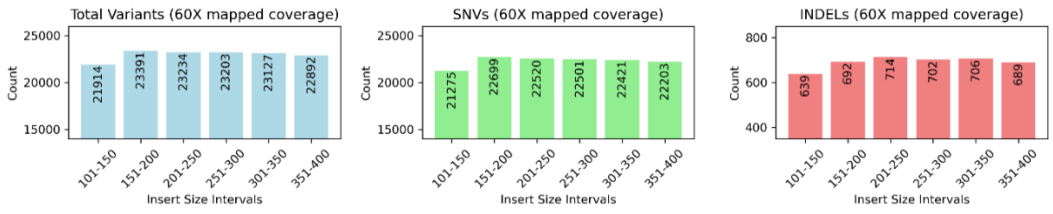


Figure 20 Number of variants called on each NA12878 ISB. The three plots show the total variants, SNVs and INDELs.

The total variant count increases from 21,914 variants in the 101–150 bp bin to a maximum of 23,391 variants in the 151–200 bp bin, after which it stabilizes across insert sizes from 201 bp upward. SNVs follow the same behavior, increasing from 21,275 SNVs (101–150 bp) to ~22,700 SNVs in the 151–200 bp interval, with minimal variation thereafter. INDELs instead exhibit a slight decrease for longer inserts, ranging from 639–714 across bins, but without major deviations once insert size exceeds ~200 bp (Figure 20).

Overall, these results show that shorter fragments (<200 bp) miss a subset of variants, consistent with their lower genotypability, while insert sizes ≥200 bp recover a larger variant set.

4.1.12. Comparison of the called variants against GIAB Gold Set

To quantify variant-calling accuracy, we compared the calls from each ISB to the GIAB NA12878 Gold Set, restricting the analysis inside the WES target design, and classified them as true positives (TP), false positives (FP), or false negatives (FN) (Figure 21).

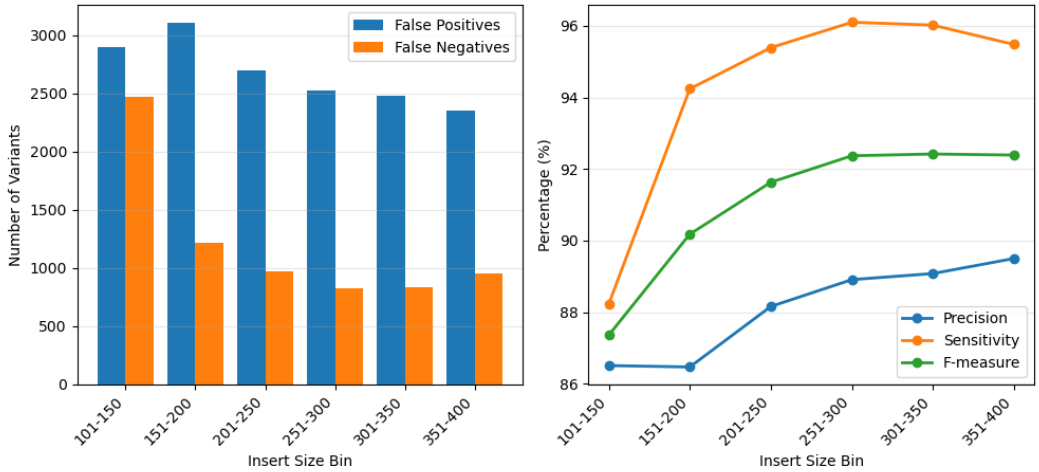


Figure 21 Comparison of the total variants called on NA12878 ISBs against GIAB NA12878 GoldSet. The analysis was restricted inside Twist 2.0 + Comp. Exome Spike In regions.

Short bins showed the highest number of errors, with FP > 2800 and FN ~2500 in the 101–150 bp bin. Both metrics improved steadily with increasing insert size: FP and FN counts dropped markedly up to the 251–300 bp interval, reflecting more reliable alignment and a higher proportion of callable positions.

These trends translated into a corresponding improvement in variant-calling metrics.

- Sensitivity increased sharply from the shortest bins to the 251–300 bp interval (peaking at 96.10%).
- Precision improved, increasing across nearly all insert size bins and reaching its maximum at 351–400 bp.
- F-measure reached its highest values (92.37-92.42) in the 251–350 bp range, with no improvements afterward.

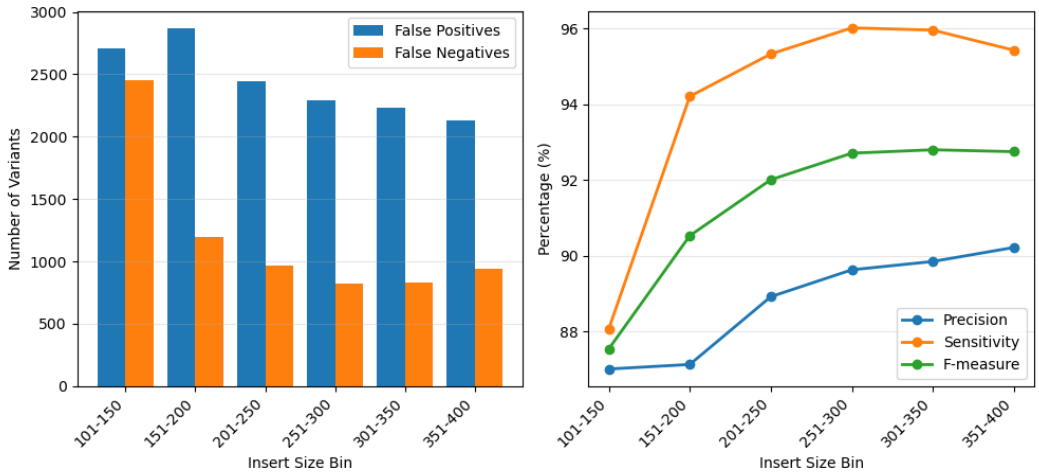


Figure 22 Comparison of the SNVs called on NA12878 ISBs against GIAB NA12878 GoldSet. The analysis was restricted inside Twist 2.0 + Comp. Exome Spike In regions.

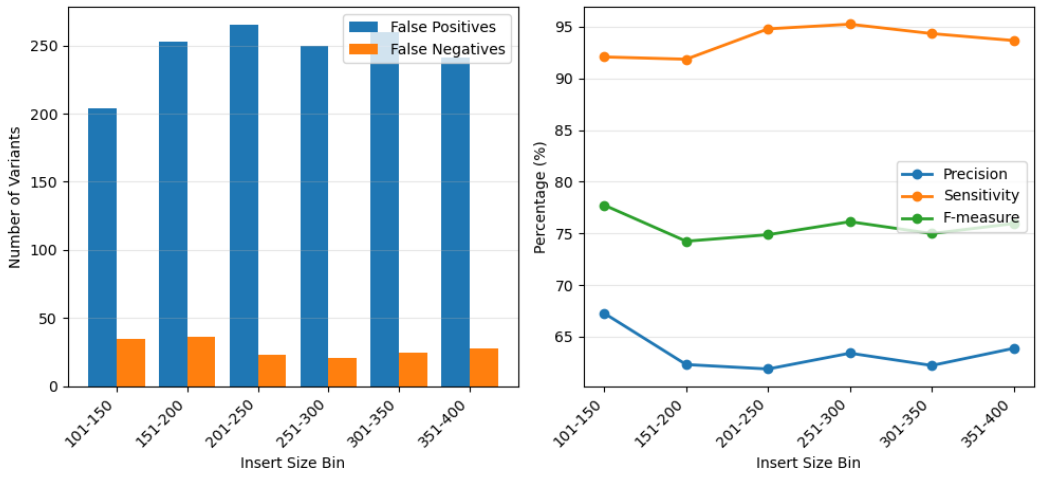


Figure 23 Comparison of INDELs called on NA12878 ISBs against GIAB NA12878 GoldSet. The analysis was restricted inside Twist 2.0 + Comp. Exome Spike In regions.

The separate benchmarking of SNVs (Figure 22) and INDELs (Figure 23) revealed distinct behaviors for the two variant types. SNVs, being the most abundant, followed trends nearly identical to the full variant set: both FP and FN counts decreased progressively as insert size increased and F-measure stabilized beyond ~250 bp (Figure 22).

INDELs, instead, showed a different pattern. Short insert size bins (101–150 bp) produced fewer false positives but slightly more false negatives, yielding higher precision but reduced sensitivity. Moreover, both FP and FN counts stabilized Once insert size exceeded ~250 bp (Figure 23).

Overall, these results align with the mapping quality and genotypability trends described in the previous chapters: insert sizes below ~200–250 bp exhibit reduced alignment accuracy and lower genotypability, which directly increases both FP and FN rates, particularly for SNVs. In contrast, insert sizes between 250 and 350 bp provide the best balance between sensitivity, precision, and F-measure, ensuring the highest variant-calling accuracy and enabling the detection of variants that remain systematically missed when using shorter fragments.

4.1.13. Comparison of the called variants against GIAB Gold Set, High Confidence Regions and Low Mappability Regions

Having established that longer insert sizes improve global variant-calling accuracy in NA12878, we next evaluated whether these effects are equally evident across distinct genomic contexts. In particular, we compared the called variants from each insert size bin (ISB) with the GIAB NA12878 Gold Set, separately analyzing the GIAB High-Confidence Regions and the GIAB Low-Mappability Regions.

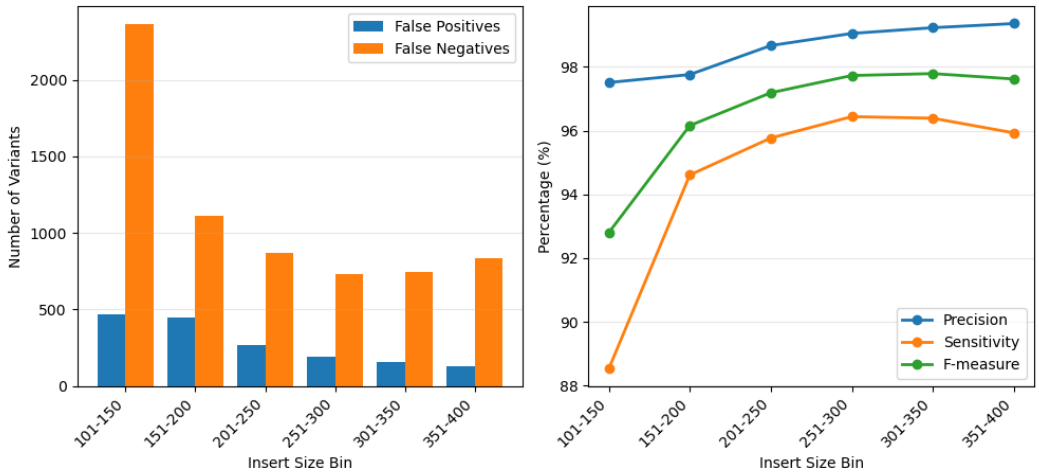


Figure 24 Comparison of the total variants called on NA12878 ISBs against GIAB NA12878 GoldSet. The analysis was restricted inside GIAB High Confidence Regions (Benchmark).

In the GIAB High-Confidence Regions, as expected, the number of false positives (FP) was low across all insert size bins. Nevertheless, the trend observed in the full dataset persisted: FP counts decreased as insert size increased. False negatives (FN) also declined with increasing ISB, although to a lesser extent. As a consequence, precision, sensitivity, and F-measure improved consistently with longer insert sizes,

reaching their highest values for ISBs above 300 bp (Figure 24). These results confirm that even in regions where alignment is generally reliable, longer fragments still provide measurable benefits for variant-calling.

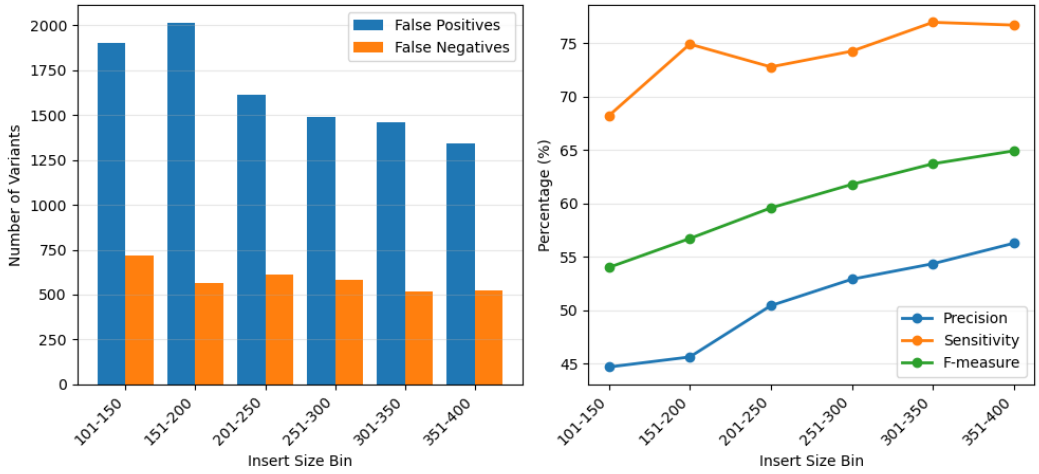


Figure 25 Comparison of the total variants called on NA12878 ISBs against GIAB NA12878 GoldSet. The analysis was restricted inside GIAB genome stratification v3.5 Low Mappability Regions ("all"subset).

The effect of insert size was also pronounced in the GIAB Low-Mappability Regions. In these challenging intervals, false positives showed a substantial reduction as ISB increased, while false negatives remain almost stable. The strong decline in FP reflects the reduced alignment ambiguity afforded by longer fragments, which provide more unique genomic context for read mapping. Consequently, precision, sensitivity, and F-measure all improved with increasing insert size (Figure 25). These results indicate that longer fragments are effective at resolving the alignment challenges in low-mappability regions, enabling more accurate and complete variant detection.

4.2. Introduction to Sequencer Quality Evaluation

Having established that insert size is a key parameter influencing mapping quality, genotypability, and variant-calling accuracy, the next question is how the quality of the sequence data itself can be further assessed and interpreted. Once the library preparation has been optimized, the sequencing platform becomes the dominant source of variability, as the type, frequency, and distribution of sequencing errors directly affect downstream analyses.

Producing data with the lowest possible error rate is therefore essential. However, the quality scores (Q-scores) reported by sequencing instruments are not direct measurements of sequencing accuracy; instead, they are estimates generated by platform-specific and proprietary models, often derived from signal-level features rather than from the true error rate observed in aligned reads. As a result, identical Q-scores may correspond to different empirical error rates across platforms.

To address this issue, developed a dedicate framework and analyzed a single PCR-free whole-genome NA12878 library sequenced in parallel on MGI T1, GeneMind SURFSeq 5000, Illumina NovaSeq X, and Element AVITI. The use of a PCR-free library minimizes amplification-related artefacts, while sequencing the same biological sample across all platforms ensures that observed differences in quality scores or mismatch rates reflect intrinsic properties of the sequencing technologies and/or individual sequencing runs, rather than biological variability.

In the experimental design adopted in this study, it was not possible to include technical replicates for individual sequencing runs. Consequently, the analyses presented here were not intended to derive platform-level performance trends, but rather to evaluate the concordance between the quality scores assigned by sequencing instruments and the empirical sequencing accuracy observed after alignment, both at the base and read levels. The use of a single library and sample enables a controlled comparison across technologies; however, the absence of replicates prevents the discrimination between run-specific artefacts and genuine technology-dependent behaviors.

Beyond the direct comparison between assigned Q-scores and alignment-based error rates, we developed a dedicated analytical method and a fully reproducible pipeline to explore how sequencing quality and empirical error rates vary as a function of sequencing cycles and insert size. Specifically, we investigated:

I. cycle-dependent fluctuations in base-quality assignment across read length,
II. differences in average read-level quality and alignment identity, and
III. interactions between insert size and base-quality profiles.

While these analyses provide valuable insights into the performance of the sequencing runs presented in this thesis, the observed trends should be interpreted as run-specific rather than as definitive, platform-wide performance signatures. Importantly, the primary goal of these analyses was to demonstrate and validate the proposed workflow, which has been designed for application to a larger cohort. A comprehensive assessment of characteristic error profiles for each sequencing technology would require replicated experiments under matched quality conditions.

4.2.1. Alignment Identity as a Function of sequencer Base Quality

The relationship between base quality and empirical alignment accuracy was evaluated by stratifying the alignment identity of the NA12878 PCR-free library according to the Phred score assigned by each sequencer (Figure 26). For each Phred score, alignment identity was computed by aggregating all bases assigned that specific quality value. This analysis provides a direct comparison between the predicted and observed probability of error, and highlights platform-specific behaviors across the full range of base quality values.

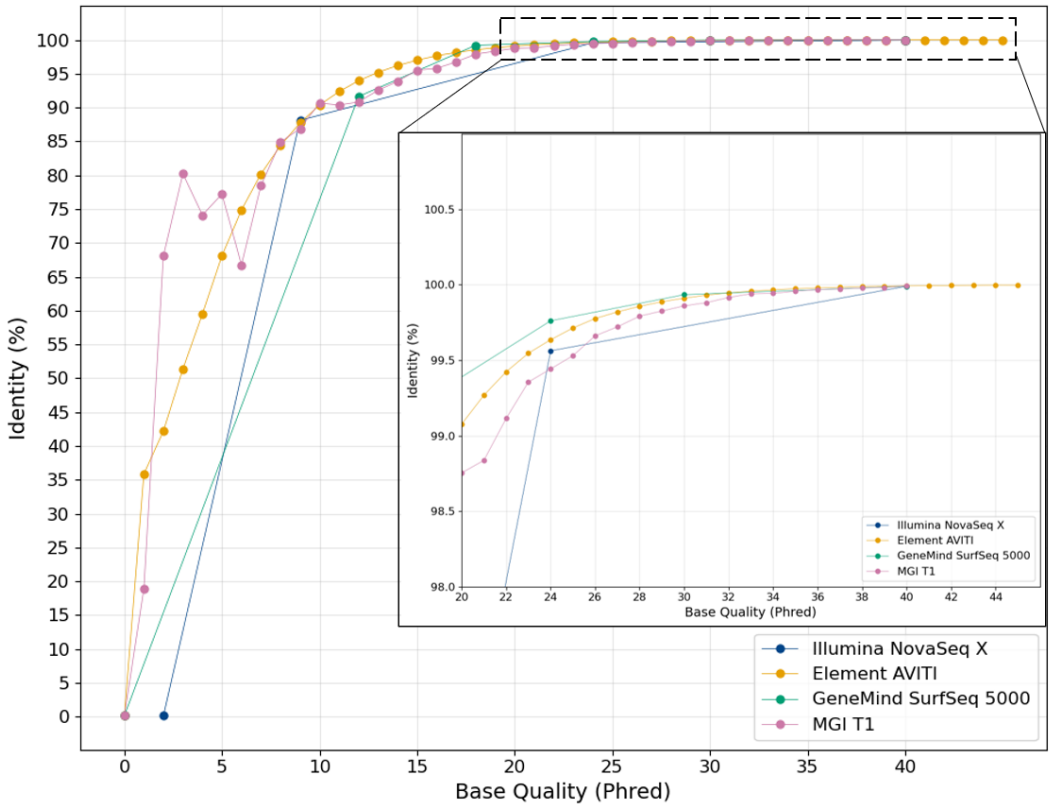


Figure 26 For each sequencer, the percentage identity computed on the aligned bases stratified by base quality assigned by the sequencing platform.

Because the Phred scale is logarithmic, differences between platforms are most pronounced at lower base qualities (Q < 20), where a shift of a few Q-points corresponds to orders of magnitude in estimated error probability. In this region, the platforms show clear separation. Element displays the highest alignment Identity at Q2, after which, MGI displays the highest alignment identity from Q3 to Q6, deviating from the expected smooth curve. Between Q6 and Q17, Element becomes generally the most accurate platform.

As base quality increases, all platforms converge towards near-perfect alignment identity. The identity-quality relationship follows a markedly asymptotic trend, with improvements in empirical accuracy becoming progressively smaller for Q-scores above Q20. Above Q24, all platforms reach alignment identities exceeding 99.5%, and differences between technologies become minimal.

Illumina and GeneMind assign base qualities in discrete, quantized bins, rather than distributing values continuously across the Phred scale. Although this design choice is motivated by the manufacturers (primarily as a strategy to reduce file size) it complicates direct quality comparison with the platforms that provide a more finely resolved quality spectrum (MGI and Element). As a result, differences observed between continuous and quantized scoring schemes may reflect algorithmic choices in quality scoring models rather than underlying biological or chemical differences in sequencing accuracy.

Overall, these results illustrate that platform-specific differences in sequencing accuracy are most visible in the low-to-intermediate quality range (Q0–Q30), whereas all four technologies achieve comparable and uniformly high alignment identities at higher Q-scores.

4.2.2. *Comparison Between Sequencer-Assigned Q Scores and Alignment-Based Q Scores*

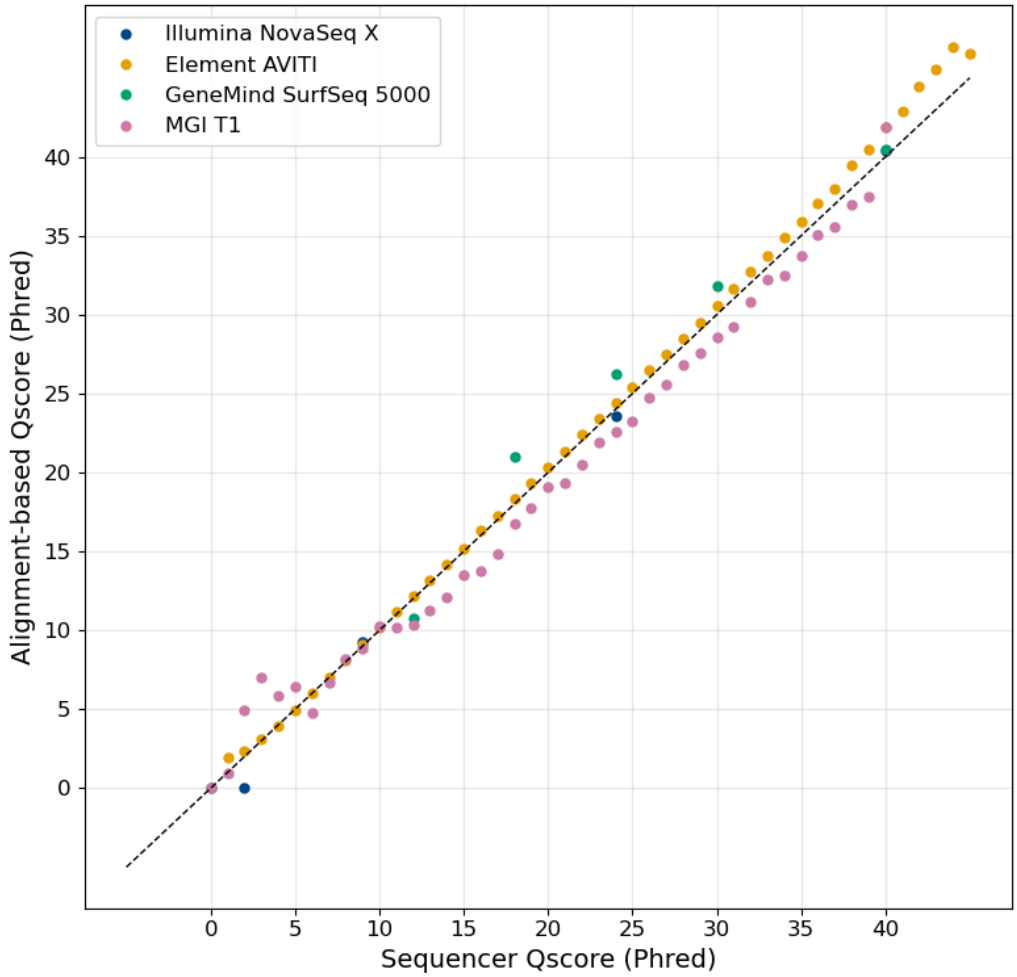


Figure 27 For each sequencer, the alignment-based Q-score computed on the aligned bases stratified by base quality assigned by the platform.

To assess how accurately each platform's quality scoring algorithm reflects the empirical accuracy of its reads, we compared the Q-scores assigned by the sequencer with identity-based Q-scores, derived directly from the identity of the aligned bases (Figure 27).

This representation provides an intuitive view of the agreement between predicted and observed (empirical) base-call accuracy:

- points on the diagonal indicate a correct quality estimate,
- points above the diagonal indicate underestimation (the sequencer assigns a lower Q than what alignment suggests),

- points below the diagonal indicate overestimation (the sequencer reports a higher Q than empirically observed).

Among the platforms evaluated, Element shows an alignment that closely follows the diagonal, indicating a strong agreement between sequencer-assigned base-quality scores and empirical alignment accuracy. At Phred scores above Q30, the reported Q-scores appear to be slightly underestimated with respect to the empirically calculated quality, suggesting a conservative quality assignment at higher confidence levels.

Illumina exhibits a generally accurate quality assignment. Aside from a slight overestimation at Q2, the platform's quality estimates remain close to the diagonal, with particularly strong agreement between Q9 and Q40, where predicted and empirical accuracies converge.

GeneMind also demonstrates good concordance between predicted and observed quality. At Q0, the assigned quality scores are in agreement with the empirical alignment accuracy. At Q12, the Q-scores appear to be slightly overestimated, whereas at intermediate quality values (Q18, Q24, and Q30) they tend to be mildly underestimated relative to the observed error rates. At high quality values (Q40), the calibration is essentially perfect, with identity-based Q-scores lying very close to the diagonal.

MGI displays a more irregular pattern. At low sequencer Q-scores (Q0-Q5), identity-based Q-scores lie well above the diagonal, indicating an underestimation of the sequencing quality of the instrument. The trend then reverses up to Q39, with identity-based Q-scores lying below the diagonal and indicating an overestimation of the sequencer-assigned Q-scores. At Q40 scores, instead, the two accuracy metrics are equivalent, indicating a perfect estimation of the base quality.

Overall, the comparison highlights differences in how accurately each platform's signal model reflects the actual error rate. While Element provides the most accurate and stable Q-score predictions across the entire quality range, Illumina and GeneMind also demonstrate strong agreement with empirical accuracy. In contrast, the MGI run mostly shows underestimation, reflecting possible suboptimal

calibration of its quality-scoring model. This analysis further emphasizes a fundamental difference among sequencing platforms, namely the binning strategy used for assigning base-quality values. While it is therefore possible to assess under- or over-calibration of Q-scores within each individual platform, direct comparisons of absolute quality scores across platforms are not meaningful, as the distribution and error rates of bases within each binned quality interval cannot be directly discerned.

4.2.3. Base-Quality Binning Schemes Across Sequencing Platforms

As shown in the previous sections, some sequencing platforms (e.g. GeneMind, Illumina) do not report base quality as a continuous Phred scale. Instead, they group quality values into a limited set of bins, a strategy primarily adopted to reduce the size of FASTQ and BAM files. This process (known as quality binning) groups different underlying signal intensities into the same reported Phred score, reducing granularity in quality representation.

Among the platforms analyzed in this study, Illumina and GeneMind implement binned quality scoring, as shown in Figure 28. Illumina groups the quality scores in 4 bins Q2-Q9-Q24-Q40, while GeneMind uses a 5 level binning schema, Q0-Q12-Q18-Q24-Q30-Q40.

In contrast, Element and MGI report continuous quality scores, producing a full-resolution Phred-scale distribution with incremental values between Q0 and Q40+. These platforms therefore preserve the native variability of the base-calling signal and provide a more detailed quality profile.

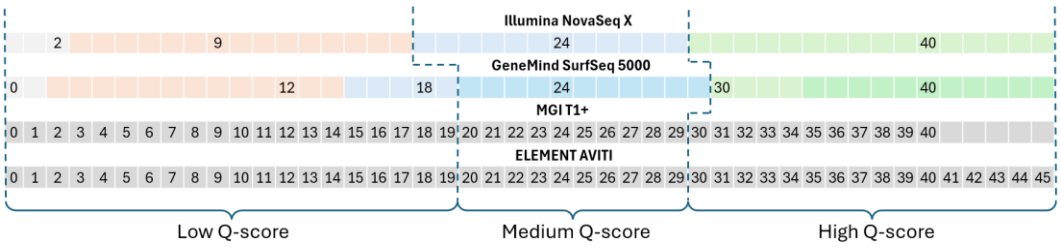


Figure 28 Quality schema used by the different platforms, Element AVITI and MGI T1 display a continuous quality score range, while Illumina NovaSeq X and GeneMind SURFSeq 5000 present a binned quality score schema.

Because these differences complicate direct cross-platform comparisons, we applied a unified binning scheme to all platforms when visualizing and comparing base-quality stratified metrics. The proposed scheme groups bases into three broad categories:

- Low Qscore (Q0-Q19),
- Medium Qscore (Q20-Q29),
- High Qscore (Q30+),

Although both GeneMind and Illumina do not align perfectly with the proposed bins, this harmonization nonetheless enhances the clarity of cross-platform comparisons and enables a more consistent interpretation of the results.

	Q bin	Low Qscore	Medium Qscore	High Qscore
Illumina NovaSeq X	**Bases (Nr)**	1,289,004,438	4,441,537,314	104,480,844,950
	Bases (%)	1.17	4.03	94.80
ELEMENT AVITI	**Bases (Nr)**	2,862,584,478	5,198,116,503	100,228,966,277
	Bases (%)	2.64	4.80	92.56
GeneMind SURFSeq 5000	**Bases (Nr)**	2,104,180,602	2,550,844,747	106,867,482,464
	Bases (%)	1.89	2.29	95.83
MGI T1	**Bases (Nr)**	1,423,101,573	2,325,868,507	143,941,417,908
	Bases (%)	0.96	1.57	97.46

Table 4 For each platform, the number and percentage of aligned bases inside each quality bin.

In line with the information provided by the manufacturers, all sequencers generated more than 90% of bases with quality scores above Q30, whereas the least populated bin is the Low Qscore, presenting only 1-2% of all bases produced by each platform (Table 4). Notably, the MGI platform reaches ~97.5% of bases with a sequencer-assigned Qscore in the highest bin (above Q30), representing the highest proportion among all tested platforms. Although these bins are unevenly populated, even the least abundant category contains far more observations than required for estimating high-quality scores. For example, computing a Q40 value (corresponding to an error rate of 0.0001) requires at least 10,000 bases across the dataset, a threshold that is comfortably exceeded for all platforms.

4.2.4. Alignment-based Q Scores stratified by Binned Base-Quality

To enable a platform-independent comparison of accuracy at matched base-quality levels, we applied the unified base-quality binning scheme (Low-Medium-High

Qscore) to all sequencers and evaluated the corresponding empirical alignment-based Qscore (Figure 29, Table 5).

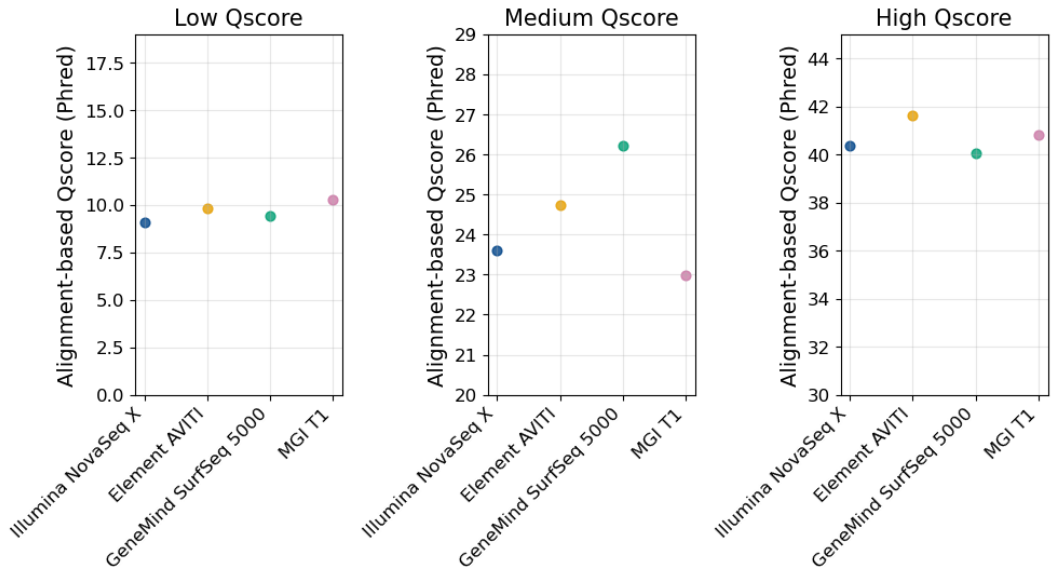


Figure 29 For the four platforms, the Alignment-based Q-scores, stratified by the binned base quality

		Binned Base Quality (Q)		
Range	Q-score	Low Qscore	Medium Qscore	High Qscore
Qscore (based on Identity)	Illumina NovaSeq X	9.07	23.59	40.36
	ELEMENT AVITI	9.84	24.74	41.61
	GeneMind SURFSeq 5000	9.40	26.23	40.04
	MGI T1	10.29	22.98	40.81

Table 5 For each platform, the identity computed on each aligned base, stratified by three the base quality bins.

Across the bins, MGI exhibits the highest empirical accuracy at the lowest quality bin (coherent with the underestimation of low Qscores observed in the previous sections), while GeneMind the highest at intermediate quality levels. In the High Qscore bin, it becomes evident that Element reaches the highest alignment-based Q-scores, whereas MGI and Illumina possess the second and third highest and GeneMind the lowest (Figure 29).

Overall, although these results do not alter the patterns observed when considering the full range of sequencer-assigned Q-scores, they confirm that for each platform (particularly for high-quality bases) the empirical Q-scores closely match the predicted values. Interestingly, the empirically measured quality of bases above Q30 is extremely high across all platforms, effectively reaching to Q40.

4.2.5. Sequencer Q-Scores and Error rate across Cycles and Platforms

To investigate how sequencing quality evolves along the reads, we examined both the sequencer-assigned Q-score (Figure 30) and the empirically measured error rate (Figure 31) across individual sequencing cycles (R1 + R2 combined). For each cycle, the Q-score of the resulting bases was averaged, and the corresponding error rate was calculated to provide an empirical assessment of sequencing accuracy. Mean base quality was computed as the average of underlying error rates, ensuring accurate estimation of per-base error probability. Alternatively, due to the logarithmic nature of Phred scores, direct arithmetic averaging of Q-scores would overestimate quality (as evident by comparing Figure 30 to Supplementary Figure S1).

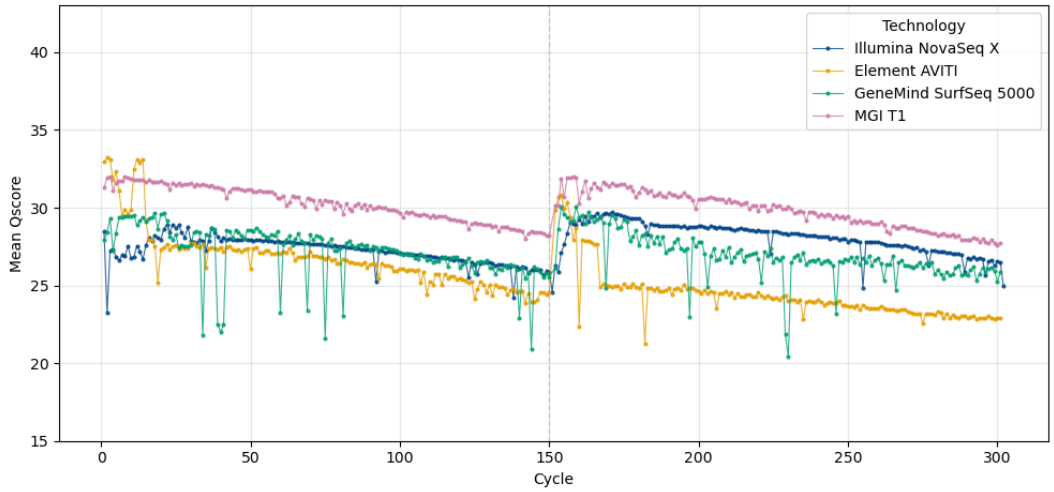


Figure 30 Average sequencer-assigned Q-score per cycle for each platform (R1 + R2 combined). For each cycle, the average Q-score was computed as the average of the underlying error probability. Cycles 1–150 correspond to Read 1, followed by Read 2 cycles.

All platforms show a gradual decrease in average quality with increasing cycle number, consistent with the typical accumulation of noise over the run. MGI, GeneMind, and Element exhibit similar values and trends between R1 and R2, whereas Illumina shows slightly higher Q-scores in R2, although the overall descending trend is maintained. GeneMind displays numerous sharp drops in quality throughout the run, with smaller, yet noticeable, negative peaks also observed in Element and Illumina. In contrast, MGI maintains a consistently smoother quality profile across all cycles, without negative peaks (Figure 30).

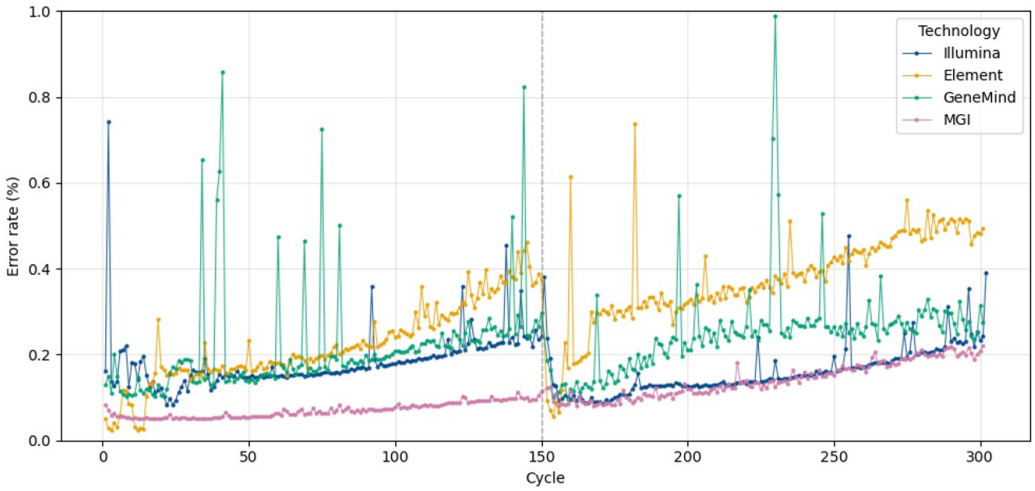


Figure 31 Error rate per sequencing cycle for each platform (R1 + R2 combined). Error rate was calculated as the ratio of mismatched, inserted, or deleted bases to the total number of bases sequenced at each cycle. Cycles 1–150 correspond to Read 1, followed by Read 2 cycles.

Consistent with the Q-score trends, all platforms demonstrate a progressive increase in error rate as sequencing cycles advance, reinforcing the pattern of quality degradation toward the terminal end of the reads (demonstrating coherence between sequencer-assigned quality and empirical error measurements, Figure 30-31). MGI exhibits a remarkably consistent error profile throughout the read length, with no prominent deviations from the overall trend. Conversely, Illumina, GeneMind, and Element are characterized by intermittent error rate spikes that align with the quality dips previously observed in the Q-score analysis (Figure 30). However, given the absence of technical replicates for the sequencing runs examined, it remains difficult to distinguish whether these sporadic peaks in error rate reflect platform-specific systematic effects or run-specific anomalies (Figure 31).

Although the absence of technical replicates limits our ability to distinguish between systematic platform-specific effects and run-dependent artifacts, the analytical approach presented here (combining cycle-by-cycle Q-score and error rate assessment) establishes a robust framework applicable to future studies with replicated runs. Such an approach will enable a more definitive characterization of sequencer performance across platforms.

4.2.6. Cycle-Dependent Shifts in Base-Quality Composition

To investigate whether the observed cycle-dependent decline in average quality results from a progressive shift toward lower-quality base emissions, we analyzed the fraction of bases assigned to each quality bin across sequencing cycles. For each quality bin, the fraction of bases emitted at each cycle was normalized by the total number of bases assigned to that bin within each platform. This normalization allows us to observe whether the production of bases in that bin follows a stable, increasing, or decreasing trend as sequencing cycles progress.

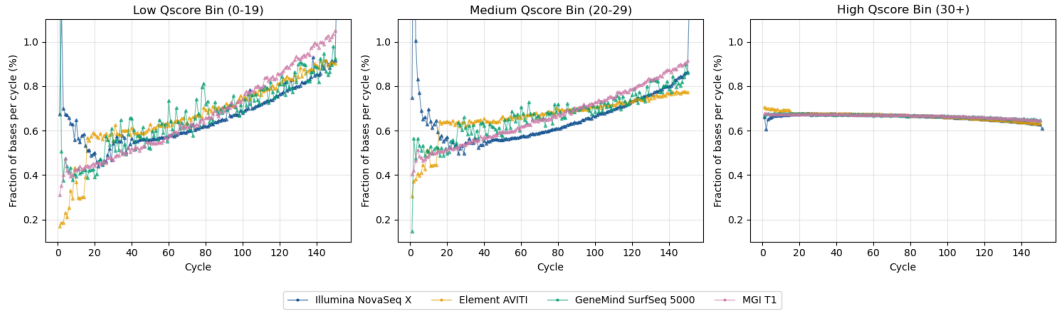


Figure 32 For each platform, the fraction of bases for each cycle, normalized by the total number of the sequencer's bases in that bin. The dataset was divided into three base quality bin according to the scheme previously described.

All platforms exhibited a consistent cycle-dependent shift in base-quality composition. The fraction of bases in the high-quality bin (Q30+) declined progressively as sequencing cycles advanced, while the proportions of medium-quality (Q20–29) and low-quality (Q0–19) bases increased correspondingly (Figure 32). This pattern was uniform across all sequencers. Supplementary Figure S2 confirms no increase in error rates within quality bins, indicating a platform-independent mechanism of quality degradation mechanisms rather than fidelity loss at constant quality levels (Figure 32).

The differences in magnitude across bins reflects the base-quality distribution described previously (see section 4.2.3), where high quality bases comprehend >90% of the total sequencer's output. Within each bin normalization amplifies fluctuations in lower-quality bins while constraining those in the dominant high quality class.

4.2.7. Cycle-Dependent Shifts in Base-Quality Composition: Read 1 and Read 2

To assess whether this compositional shift exhibits mate-specific biases, we stratified the analysis by Read 1 and Read 2 separately (Figure 33, Figure 34). Read 1 and Read 2 displayed highly consistent base-quality composition profiles across all platforms, with no appreciable differences in the cycle-dependent dynamics of low-medium-high Qscore fractions. This consistency confirms the absence of systematic mate-dependent biases in base-quality emission, in agreement with stable per-cycle alignment-based Qscore trends within quality bins observed for both mates (Supplementary Figure S3-S4) and similar quality trends between reads (section 4.2.5)

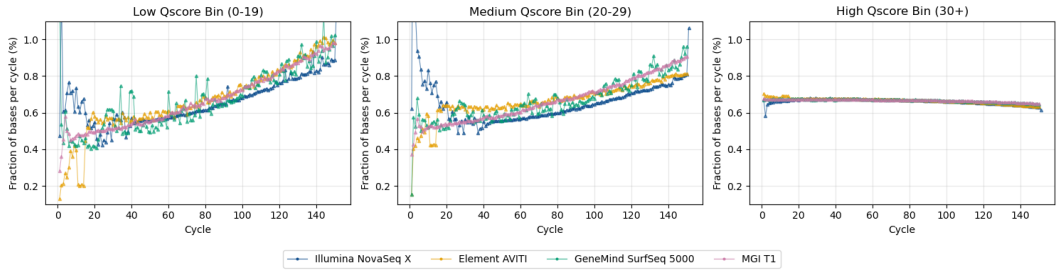


Figure 33 For the Read 1 produced by each platform, the fraction of bases for each cycle, normalized by the total number of the sequencer's bases in that bin. The dataset was divided into three base quality bin according to the scheme previously described.

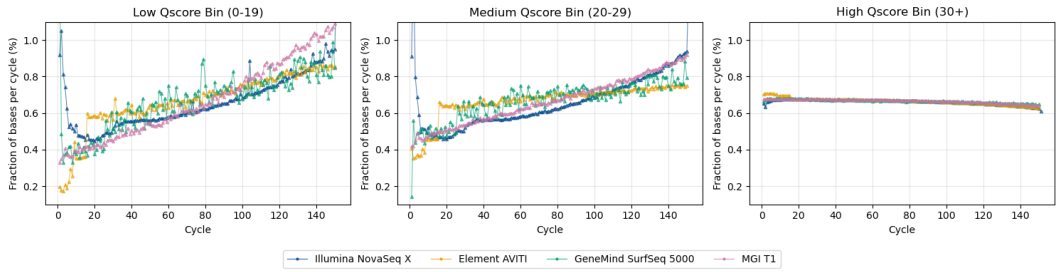


Figure 34 For the Read 2 produced by each platform, the fraction of bases for each cycle, normalized by the total number of the sequencer's bases in that bin. The dataset was divided into three base quality bin according to the scheme previously described.

4.2.8. *Insert Size Distributions Across Sequencing Platforms*

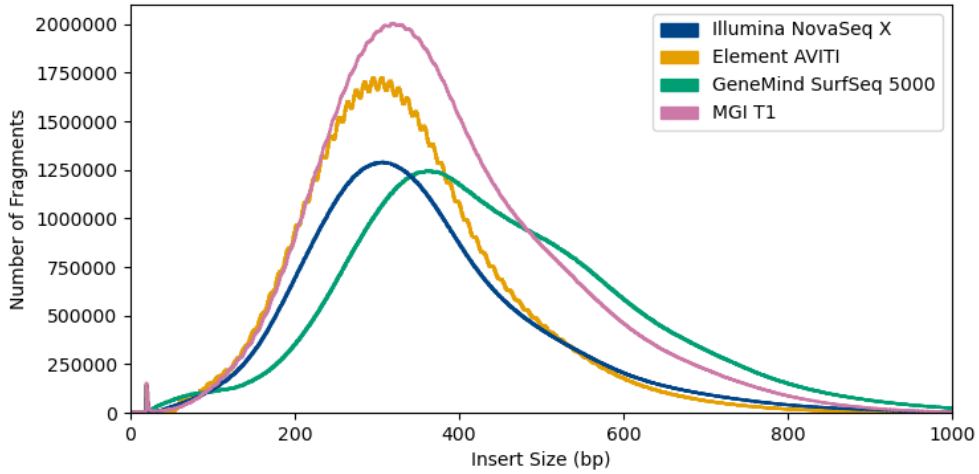


Figure 35 Insert size histogram of the library sequenced on the four platforms.

	Average	**Median**	**Mode**
Illumina NovaSeq X	355	334	307
Element AVITI	336	321	305
GeneMind SIRFSeq 5000	443	422	361
MGI T1	380	356	320

Table 6 Insert size average, median and mode of the library sequenced on the four platforms.

After identifying the optimal insert size for WES and characterizing the empirical accuracy of the four sequencing platforms at the base level, we next assessed whether sequencing quality and error rate vary as a function of insert size. This analysis is essential to determine whether the intrinsic accuracy of each platform is affected by the length of the fragment being sequenced, or whether accuracy remains stable across the insert-size spectrum.

As a preliminary step, we examined the insert size distribution generated by each platform. Although all platforms sequenced exactly the same PCR-free NA12878 library, the resulting insert-size profiles differed. Element and Illumina produced broadly similar distributions. In contrast, MGI and GeneMind showed a marked enrichment of longer fragments, particularly the latter shows a broader right tail and a clear skewed distribution toward larger inserts (Figure 35).

These differences are also reflected in the summary statistics of the insert-size distributions (Table 6). GeneMind exhibited higher mode, median, and mean insert sizes compared with the other platforms, all of which showed more compact symmetrical distributions.

4.2.8.1. Effect of Insert Size on Read-Level Sequencer Quality and Error Rates

To explore whether sequencing accuracy is affected by fragment length, we evaluated the distribution of read-level sequencer-assigned Q-scores and error rates across insert-size intervals ranging from 150 bp to 800 bp, stratified separately for Read 1 and Read 2 to detect potential mate-specific effects of insert size. The following plots show how both predicted and empirical accuracy vary as a function of insert size.

4.2.8.2. Element AVITI

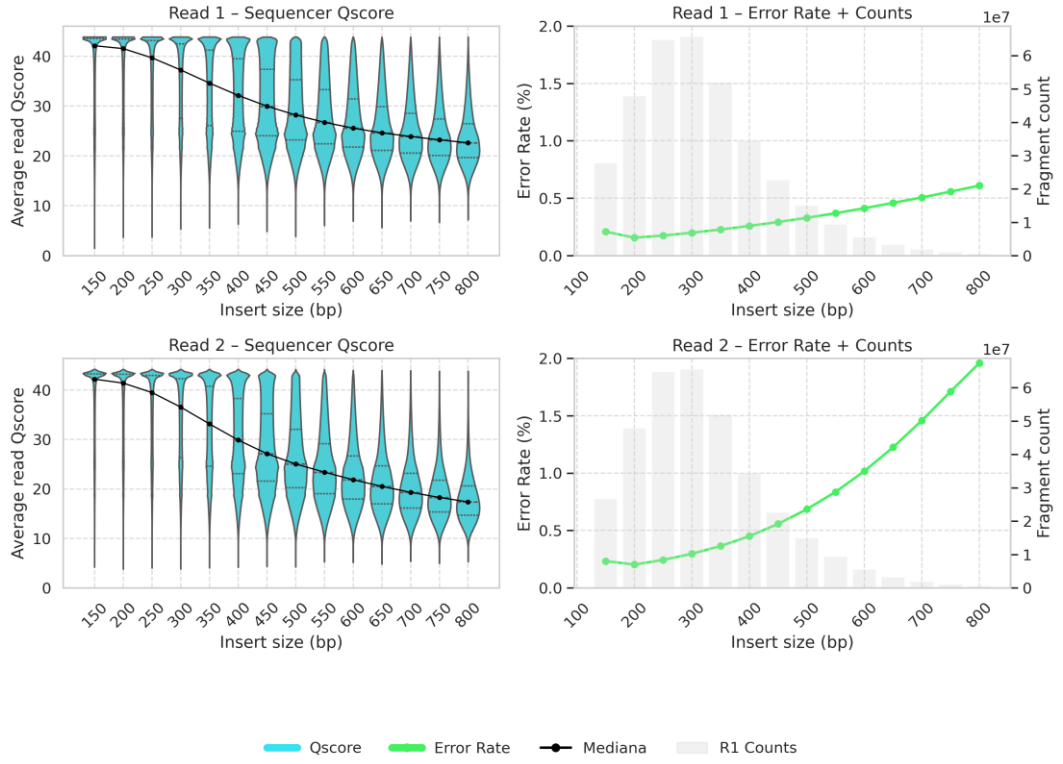


Figure 36 For Element AVITI, dividing Read1 and Read2, stratified by insert size, violin plots showing the average read Q-scores assigned by the sequencer and error rate distributions of the bases falling within each insert size bin. Background histograms show the base content in each insert size bin.

For Element (Figure 36), the distribution of average read Q-scores exhibits a clear decreasing trend. While the highest-quality reads are associated with shorter inserts, the average Q-score declines noticeably starting from 200 bp, and continues to decrease progressively up to the 800 bp interval, particularly in Read 2 where the median reaches below Q20.

Read 1 and Read 2 exhibited markedly different patterns of error rate as a function of insert size. For Read 1, the error rate remained consistently low across all insert-size intervals, reaching slightly above 0.5% at 800bp bin. In contrast, Read 2 displayed a substantially steeper increase in error rate, rising from ~0.20% at 150 bp to ~2% at 800 bp.

The divergence in error rate between Read 1 and Read 2 was substantially more pronounced than the differences observed in sequencer-assigned Q-scores. While both reads showed comparable predicted quality degradation with increasing insert size, the empirical error rate in Read 2 increased markedly more steeply, indicating a greater discrepancy between predicted and actual accuracy for longer fragments.

4.2.8.3. MGI T1+

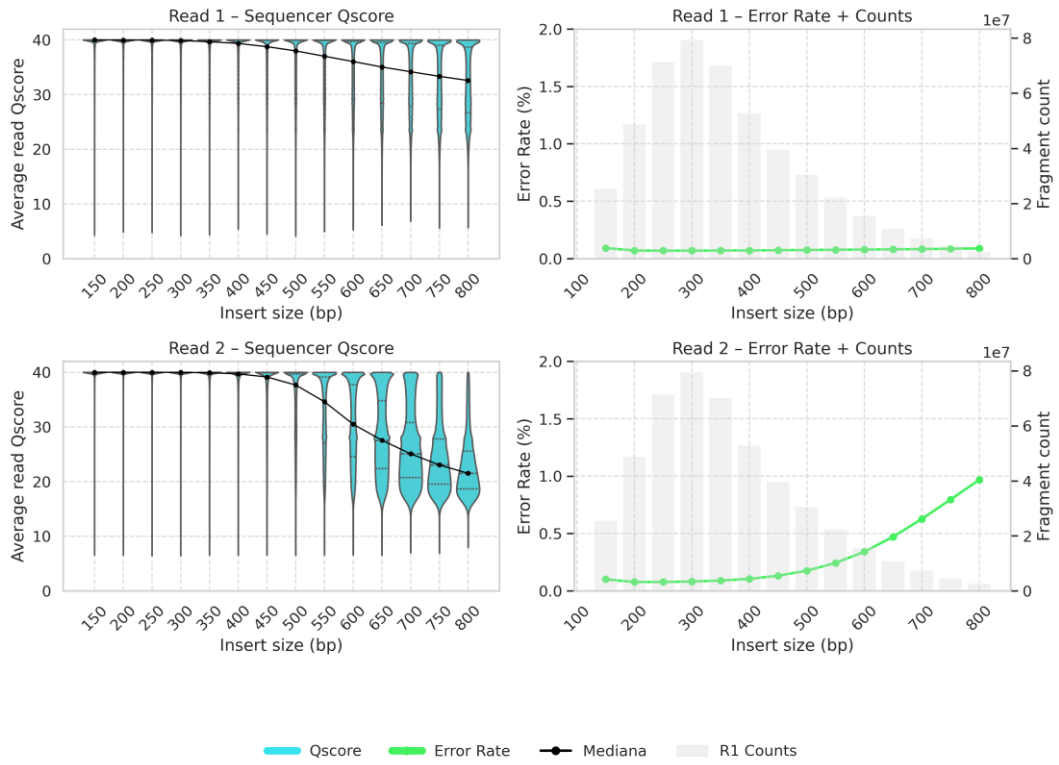


Figure 37 For MGI T1, dividing Read1 and Read2, stratified by insert size, violin plots showing the average read Q-scores assigned by the sequencer and error rate distributions of the bases falling within each insert size bin. Background histograms show the base content in each insert size bin.

For MGI (Figure 37), a pronounced mate-specific insert-size bias emerged in both sequencer-assigned Q-scores and empirical error rates. Read 1 maintained consistently high predicted quality across all insert sizes, with median Q-scores never dropping below Q30, and correspondingly low error rates that remained stable even at 800 bp. In contrast, Read 2 exhibited a marked decline in sequencer-assigned quality for inserts longer than 450 bp, reaching Q30 at 800 bp. The empirical error rate in Read 2 mirrored this trend, remaining low up to 450 bp and then rising progressively to ~1.0% at 800 bp, demonstrating alignment between predicted and actual accuracy for the second mate.

4.2.8.4. Illumina NovaSeq X

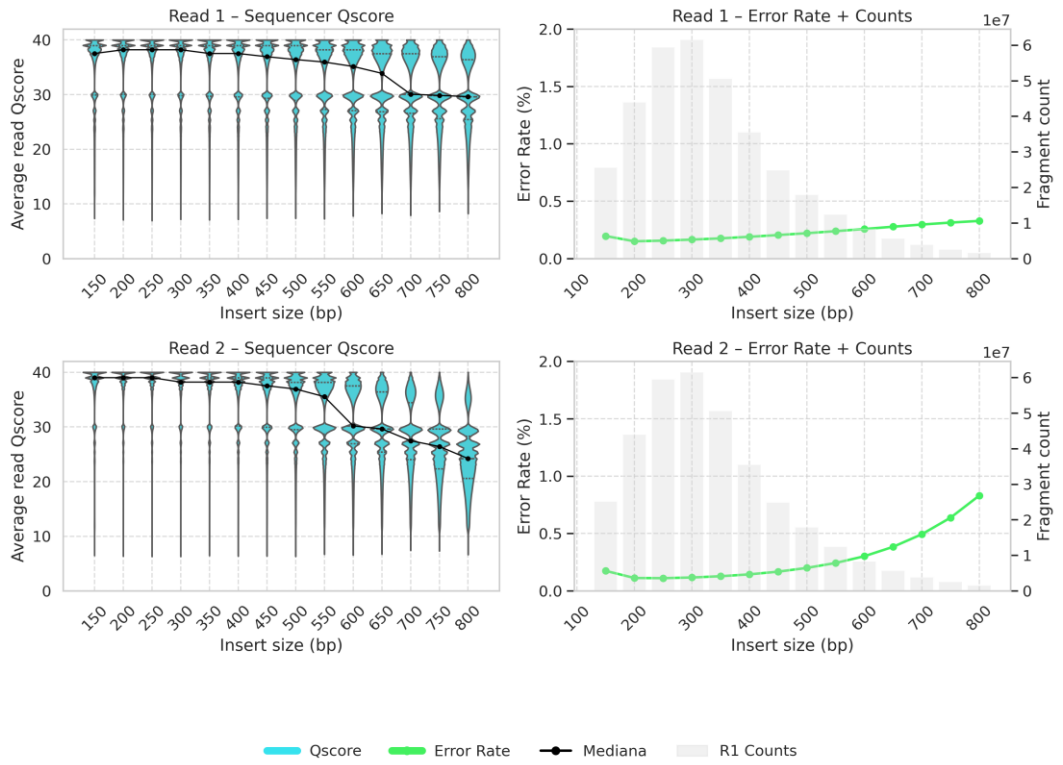


Figure 38 For Illumina NovaSeq X, dividing Read1 and Read2, stratified by insert size, violin plots showing the average read Q-scores assigned by the sequencer and error rate distributions of the bases falling within each insert size bin. Background histograms show the base content in each insert size bin.

For Illumina (Figure 38), both Read 1 and Read 2 exhibited a progressive decline in sequencer-assigned Q-scores with increasing insert size. However, Read 2 showed a steeper degradation, reaching substantially lower values, approximately Q25 at 800 bp, compared to Q30 for Read 1 at the same interval. Empirical error rates reflected this mate-dependent pattern. Read 1 showed a modest increase to ~0.3% at 800 bp, while Read 2 displayed a more pronounced rise, reaching ~0.8% at the longest insert sizes, demonstrating alignment between predicted and empirical accuracy across both mates. Nevertheless, for both reads, sequencing quality remained high up to approximately 550–600 bp, indicating adequate performance across the typical insert-size range.

4.2.8.5. GeneMind SURFSeq 5000

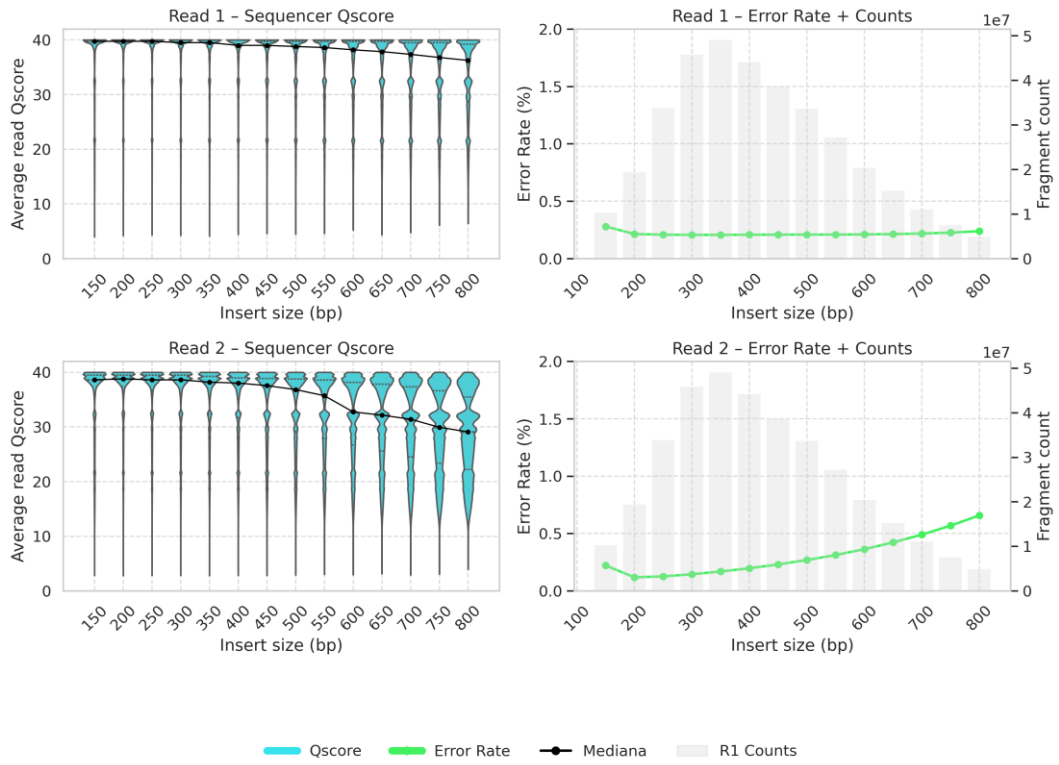


Figure 39 For GeneMind SURFSeq 5000, dividing Read1 and Read2, stratified by insert size, violin plots showing the average read Q-scores assigned by the sequencer and error rate distributions of the bases falling within each insert size bin. Background histograms show the base content in each insert size bin.

For GeneMind (Figure 39), the insert-size dependence of sequencer-assigned Qscores and error rates closely resembled the MGI pattern. Read 1 maintained consistently high sequencer-assigned quality across all insert sizes, with correspondingly low and stable error rates throughout the entire insert-size range. In contrast, Read 2 exhibited mate-specific degradation, with sequencer-assigned Qscores declining for inserts longer than 550 bp, reaching a median of approximately Q30 at 800 bp. Empirical error rates in Read 2 mirrored this trend, rising progressively to ~0.6% at 800 bp. Nevertheless, Read 1 quality remained robust across all insert sizes, ensuring reliable performance within the typical sequencing range.

4.2.8.6. *Summary of Insert-Size Effects Across Platforms*

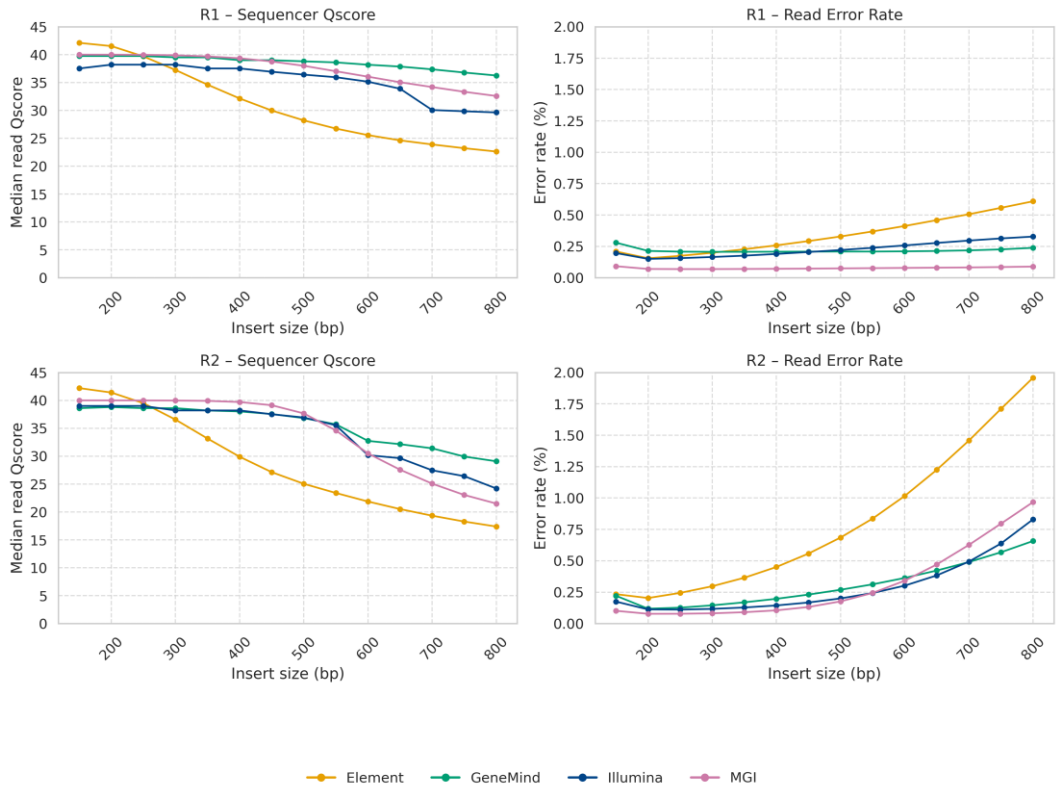


Figure 40 For the four platforms considered, dividing Read1 and Read2, the median values of average read Qscores as assigned by the sequencer and error rates for each insert size interval considered.

The comparative analysis of insert-size dependent Q-score and error rate trends revealed distinct platform-specific patterns (Figure 40). GeneMind, Illumina, and MGI displayed consistent and coherent profiles, with both sequencer-assigned quality and empirical error rates declining gradually, specifically in Read 2, as insert size increased. In contrast, Element exhibited a markedly different behavior, with quality decline emerging at shorter insert sizes and a more pronounced degradation across both reads, accompanied by a marked increase in Read 2 error rate, reaching ~2.0% at 800 bp (Figure 40).

For GeneMind, Illumina, and MGI, both predicted and empirical accuracy remained robust up to approximately 550–600 bp, indicating adequate performance within the typical insert-size range for whole-exome sequencing libraries. Element, however, showed greater insert-size sensitivity, with more pronounced quality degradation across the entire insert-size spectrum (Figure 40).

4.2.9. From base to read-level accuracy

The previous sections focused on sequencing quality at the level of individual bases, evaluating how platform-specific scoring schemes, cycle-dependent behavior and insert length influence the empirical accuracy of base calls. While these analyses provide a detailed characterization of base-level performance, they do not directly capture the accuracy of entire sequenced reads.

To bridge the analysis from base to read-level accuracy, we developed a dedicated computational workflow designed to quantify the empirical accuracy of every single read generated by each sequencing platform.

4.2.10. Read-Level Identity and Error Distribution

Sequencer	Nr reads	Mean Identity	read with 0 errors	read with 1 errors	read with 2 errors	read with 3 errors	read with >3 errors
Illumina	737,575,222	99.82%	649,410,082	54,320,532	14,303,440	6,683,848	12,857,320
Element	728,268,382	99.70%	584,660,036	80,507,174	28,007,334	12,648,913	22,444,925
GeneMind	748,278,639	99.76%	654,247,408	50,029,602	14,235,567	9,310,804	20,455,258
MGI T1	989,416,119	99.89%	913,368,804	45,662,821	12,255,525	6,362,091	11,766,878

Table 7 Average read identity and number of reads with 0-1-2-3->3 errors for each platform.

Sequencer	Nr reads	Mean Identity	read with 0 errors	read with 1 errors	read with 2 errors	read with 3 errors	read with >3 errors
Illumina	737,575,222	99.82%	88.05%	7.36%	1.94%	0.91%	1.74%
Element	728,268,382	99.70%	80.28%	11.05%	3.85%	1.74%	3.08%
GeneMind	748,278,639	99.76%	87.43%	6.69%	1.90%	1.24%	2.73%
MGI T1	989,416,119	99.89%	92.31%	4.62%	1.24%	0.64%	1.19%

Table 8 Average read identity and percentage of reads with 0-1-2-3->3 errors for each platform.

Read-level accuracy was quantified by computing the identity of every read against the reference genome and by stratifying reads according to the number of sequencing errors they contained (Table 7, Table 8).

Across the full dataset, MGI produced reads with the highest mean identity (99.89%), followed closely by Illumina, while GeneMind and Element showed the lowest average identity among the four platforms (Table 8).

A similar trend emerged when examining the distribution of error counts. MGI exhibited the largest fraction of reads with zero observed errors, exceeding 90% of all reads. Illumina and GeneMind obtained a close result, with both around 87-87%

and Element showed the lowest proportions of error-free reads (approximately 80%, consistent with their lower mean identity, Table 8).

4.2.11. Read-Level Identity and Error Distribution for High-Quality Reads

Sequencer	Nr reads	Mean Identity	read with 0 errors	read with 1 errors	read with 2 errors	read with 3 errors	read with >3 errors
Illumina	546,793,795	99.99%	536,290,747	10,220,774	201,550	31,741	48,983
Element	462,205,026	99.99%	454,934,102	7,005,927	197,170	28,534	39,293
GeneMind	577,289,400	99.99%	572,378,791	4,477,072	216,621	83,808	133,108
MGI T1	833,326,935	99.99%	817,127,737	15,159,513	717,866	124,454	197,365

Table 9 Considering only reads with an average assigned base Q-score of at least 30, the average read identity and number of reads with 0-1-2-3->3 errors for each platform.

Sequencer	Nr reads	Mean Identity	read with 0 errors	read with 1 errors	read with 2 errors	read with 3 errors	read with >3 errors
Illumina	546,793,795	99.99%	98.08%	1.87%	0.04%	0.01%	0.01%
Element	462,205,026	99.99%	98.43%	1.52%	0.04%	0.01%	0.01%
GeneMind	577,289,400	99.99%	99.15%	0.78%	0.04%	0.01%	0.02%
MGI T1	833,326,935	99.99%	98.06%	1.82%	0.09%	0.01%	0.02%

Table 10 Considering only reads with an average assigned base Q-score of at least 30, the average read identity and percentage of reads with 0-1-2-3->3 errors for each platform.

Since, as shown previously, the four sequencing platforms differ in their ability to emit high-quality bases, and this difference could be run-specific, we repeated the read-level accuracy analysis by restricting the evaluation to reads whose average quality was at least Q30 (Table 9, Table 10). This allows a fair comparison across platforms by examining reads with comparable predicted accuracy. Restricting the analysis only to high quality reads is motivated also by the fact that some platforms may perform a-priori read quality filters during sequencing or demultiplexing.

When restricting to high-quality reads, the mean read identity became identical across all sequencers, with values of 99.99% for Illumina, Element, GeneMind, and MGI. Likewise, the fraction of reads containing zero errors showed only lower differences across the platforms with respect to the previous analysis. Interestingly, the performance ranks changed, with GeneMind exhibiting the highest proportion of error-free reads (99.15%), whereas Illumina and MGI showing the lowest (98.06-98.08%, Table 10). Furthermore, consistent with the high quality of the analyzed reads, average alignment identity exceeds the theoretical error rate expected for

Q30+ scores and aligns precisely with the empirical Q40 accuracy observed for high-quality bases (section 4.2.2, Figure 29, Table 5).

4.2.12. Comparative summary: Whole Dataset vs High-Quality Reads

To synthesize the results obtained from the whole-read analysis and the high-quality (Q30+) subset, we directly compared read identity and error-free read fractions between the entire dataset and the subset of reads with average sequencer-assigned Q-score of at least 30 (Table 11).

	Whole dataset		Q30+ reads		
Sequencer	Mean Identity	%read with 0 errors	Mean Identity	%read with 0 errors	%Q30 reads
NovaSeq X	99.82%	88.05%	99.99%	98.08%	74.13%
Element AVITI	99.70%	80.28%	99.99%	98.43%	63.47%
GeneMind SURFSeq 5000	99.76%	87.43%	99.99%	99.15%	77.15%
MGI T1	99.89%	92.31%	99.99%	98.06%	84.22%

Table 11 For each platform, the average identity and percentage of reads without errors, for both the total dataset and only the reads with average assigned base Q-score of at least 30 and the relative percentage of high quality reads.

When considering all reads, the platforms differ by up to 0.2% in computed average identity and 12% in the fraction of reads without errors. However, when restricting the analysis to high-quality reads, these differences shrink dramatically. Average identity is the same, and the percentage of error-free reads varies by around 1% across platforms. This confirms that, by matching the reads by high sequencing quality, the differences in alignment-based accuracy of the four technologies becomes smaller (Table 11).

The key discriminant driving global differences in read-level accuracy is therefore the fraction of high-quality reads (Q30+) emitted by each platform. MGI produces the highest proportion of high-quality reads, approximately 84%, whereas Illumina and GeneMind both produce around 74-77% of high-quality reads. Element generate a markedly lower proportion of around 63%. These differences in high-quality read throughput fully account for the discrepancies observed in the whole-dataset metrics (Table 11).

4.2.13. Error landscape across High-Quality Reads

	Total Error Bases	Mismatch	Inserted Bases	Deleted Bases	Tandem Repeats	Homopolymers	Extreme GC%
Element	7,752,550	7,406,448	143,603	202,499	351,267	313,322	453,243
GeneMind	5,997,736	5,370,312	235,313	392,111	549,085	365,784	365,807
Illumina	11,075,329	10,489,161	201,426	384,742	506,668	552,431	752,063
MGI T1	18,200,210	14,578,102	1,424,473	2,197,635	1,546,178	2,719,749	995,360

Table 12 Absolute counts of total erroneous bases, classified by type (mismatch, insertion, deletion) and difficult genomic regions (tandem repeats, homopolymers, GC% above 65% or below 25%).

	Total Error Bases	Mismatch	Inserted Bases	Deleted Bases	Tandem Repeats	Homopolymers	Extreme GC%
Element	100%	95.54%	1.85%	2.61%	4,53%	4,04%	5,85%
GeneMind	100%	89.54%	3.92%	6.54%	9,15%	6,10%	6,10%
Illumina	100%	94.71%	1.82%	3.47%	4,57%	4,99%	6,79%
MGI T1	100%	80.10%	7.83%	12.07%	8,50%	14,94%	5,47%

Table 13 Percentage of erroneous bases out of the total errors, classified by type (mismatch, insertion, deletion) and difficult genomic regions (tandem repeats, homopolymers, GC% above 65% or below 25%).

To characterize error type distribution and identify potential platform-specific genomic biases, we stratified discrepancies in High-Quality reads by type (mismatch, insertion, deletion) and complex genomic contexts (tandem repeats, homopolymers, extreme GC% regions) across platforms (Tables 12-13).

GeneMind and MGI exhibited higher INDEL proportions relative to Illumina and Element. Complex region enrichment analysis revealed elevated error rates in tandem repeats and homopolymers for GeneMind and MGI, while extreme GC% regions showed no platform-specific bias (Tables 13).

4.2.14. Quality Distribution of Errors in High-Quality Reads

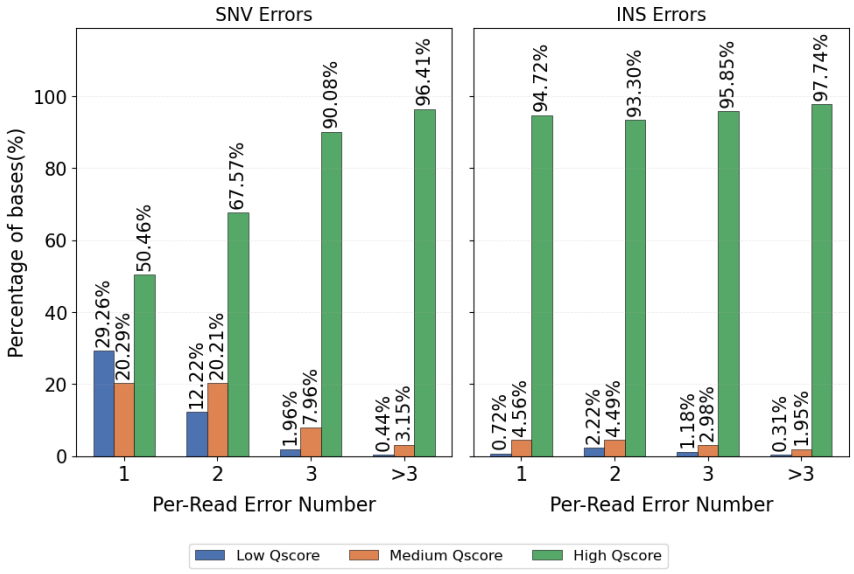


Figure 41 For Illumina NovaSeq X, the distribution of base quality scores for sequencing errors (mismatches and insertions only) in Q30+ reads, stratified by read error count (0, 1, 2, 3, >3 errors).

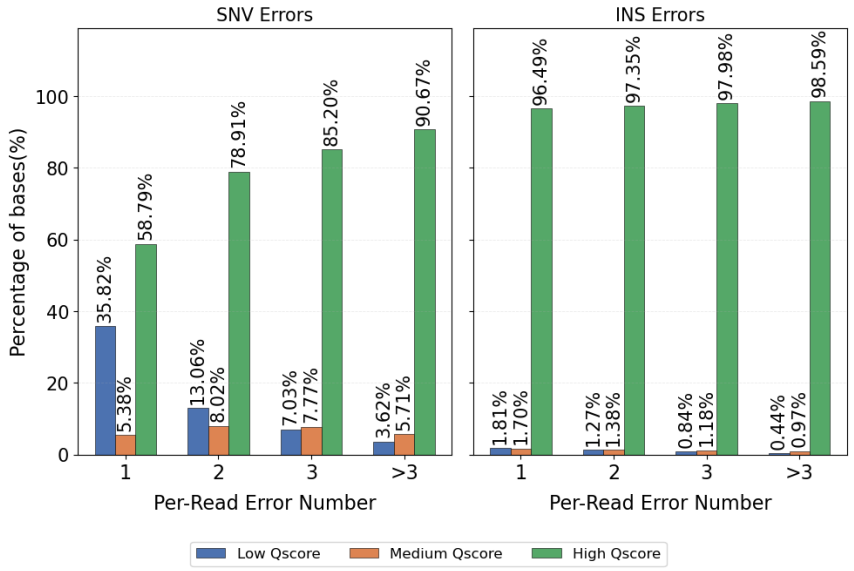


Figure 42 For GeneMind SURFSeq 5000, the distribution of base quality scores for sequencing errors (mismatches and insertions only) in Q30+ reads, stratified by read error count (0, 1, 2, 3, >3 errors).

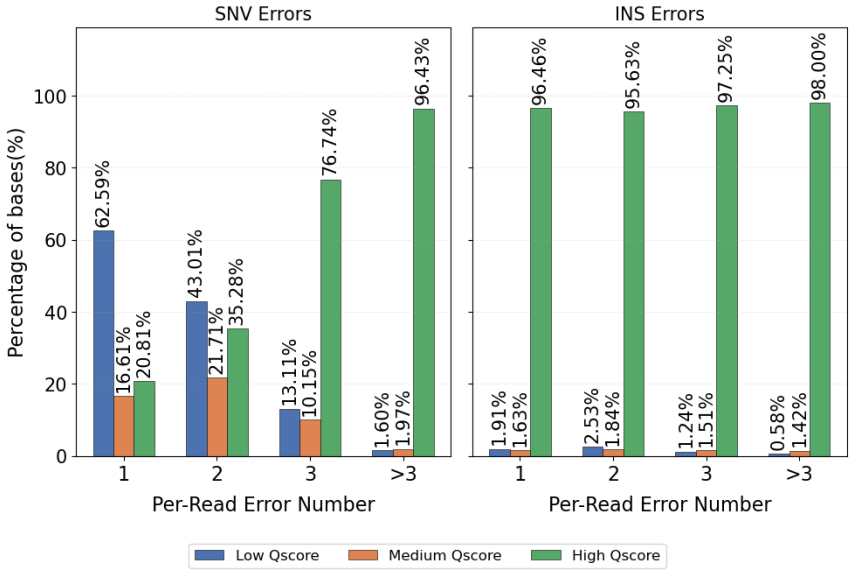


Figure 43 For Element AVITI, the distribution of base quality scores for sequencing errors (mismatches and insertions only) in Q30+ reads, stratified by read error count (0, 1, 2, 3, >3 errors).

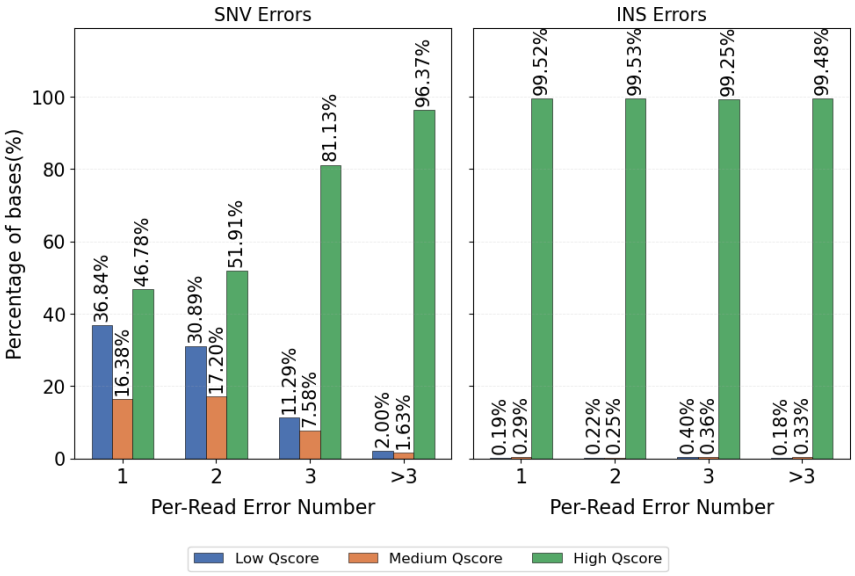


Figure 44 For MGI T1+, the distribution of base quality scores for sequencing errors (mismatches and insertions only) in Q30+ reads, stratified by read error count (1, 2, 3, >3 errors).

To investigate the quality profile of sequencing errors in high-confidence reads, we examined the base quality distribution of mismatches and insertions (excluding deletions, which lack assigned quality scores) within High-quality reads, stratified by total read-level error count (Figure 41-44).

Reads containing a single error displayed balanced quality distributions across low-/medium- and high-quality bins, with 20-58% of errors occurring in high-quality (Q30+) bases. Notably, MGI and Element exhibited lower proportions of high-Q errors (Figures 41-42), while GeneMind and Illumina showed higher fractions (Figures 43-44). This platform-specific pattern persisted in reads with two errors, where MGI/Element reached 35-51% high-Q errors compared to 68-78% for GeneMind/Illumina.

However, reads with >2 errors revealed >85% of errors occurring exclusively in high-quality bases regardless of platform, with no significant differences between technologies (Figure 41-44). This high-Q error predominance in multi-error reads is confirmed by Supplementary Figure S5, where boxplots reveal stable distributions of total bases across high/medium/low quality bins when comparing reads with varying error rates. Insertions, in contrast, showed no platform or error-count differences, with almost all errors occurring exclusively in high-quality bases across platforms (Figure 41-44).

The overwhelming predominance of high-Q base errors in multi-error reads (>2 errors) raises fundamental questions about their origin. These discrepancies may represent true sequencing errors, alignment artifacts, or mismatches due to reference genome complexity. This analysis highlights the inherent difficulty in distinguishing genuine sequencing errors from mapping ambiguities in high-confidence reads (particularly when multiple discrepancies co-occur) even after applying stringent mapping quality, regional, and variant filters.

4.2.15. Error landscape across Single-Error High-Quality Reads

The unexpectedly high base qualities in reads with multiple errors raised concerns about confounding effects on error type and context stratification. We therefore repeated the comprehensive error profiling on single-error High-Quality reads.

	Total Error Bases	Mismatch	Inserted Bases	Deleted Bases	Tandem Repeats	Homopolymers	Extreme GC%
Element	7,005,927	6,925,725	20,843	59,359	165,792	212,132	386,981
GeneMind	4,477,072	4,319,910	53,055	104,107	173,370	268,261	262,854
Illumina	10,220,774	9,997,971	60,602	162,201	251,928	450,264	649,336
MGI T1	15,159,513	13,723,474	661,900	774,139	331,821	1,715,006	782,432

Table 14 On reads with 1 error, the absolute counts of total erroneous bases, classified by type (mismatch, insertion, deletion) and difficult genomic regions (tandem repeats, homopolymers, GC% above 65% or below 25%).

	Total Error Bases	Mismatch	Inserted Bases	Deleted Bases	Tandem Repeats	Homopolymers	Extreme GC%
Element	100%	98,86%	0,30%	0,85%	2,37%	3,03%	5,52%
GeneMind	100%	96,49%	1,19%	2,33%	3,87%	5,99%	5,87%
Illumina	100%	97,82%	0,59%	1,59%	2,46%	4,41%	6,35%
MGI T1	100%	90,53%	4,37%	5,11%	2,19%	11,31%	5,16%

Table 15 On reads with 1 error, the percentage of erroneous bases out of the total errors, classified by type (mismatch, insertion, deletion) and difficult genomic regions (tandem repeats, homopolymers, GC% above 65% or below 25%).

Consistent with the previous error characterization (Section 4.2.13, Tables 12-13), single-error reads confirm MGI indel enrichment relative to other platforms, while GeneMind indel bias attenuates compared to Element and Illumina (Table 15).

Homopolymers error enrichment remains elevated (especially in MGI), mirroring full dataset patterns, while the proportion of errors in Tandem Repeats becomes almost equivalent between the platforms (Table 15).

4.2.16. Genomic Overlap of Errors Across Platforms

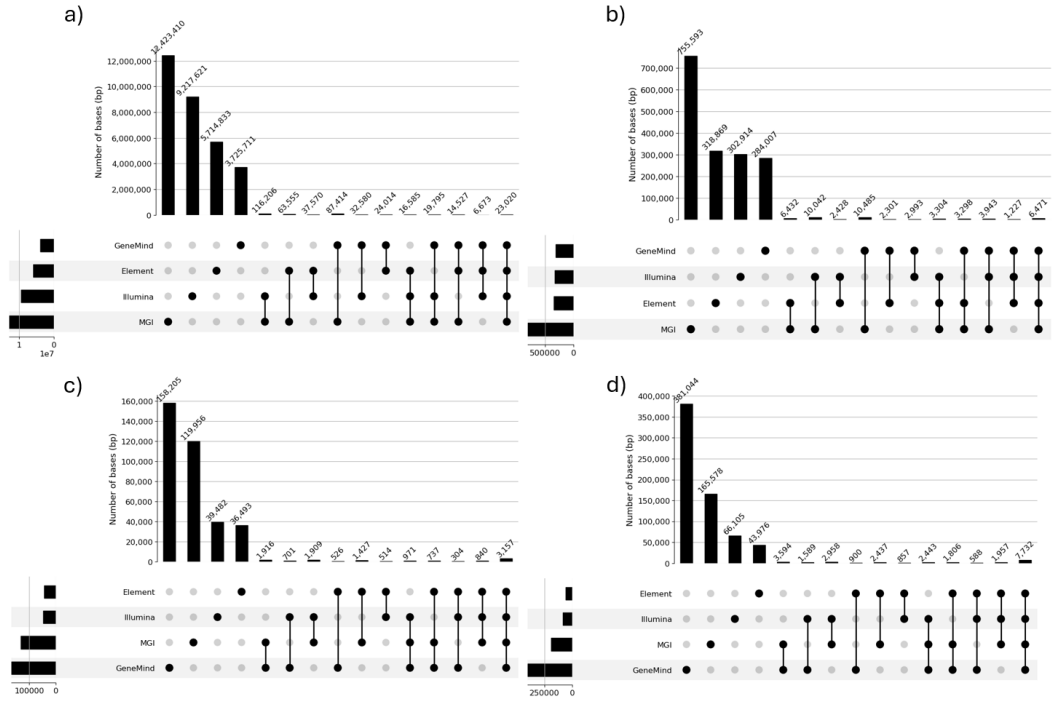


Figure 45 UpSet plots showing the overlap of error-containing genomic positions across platforms, stratified by read error count. a) UpSet plot of reads with 1 error. b) UpSet plot of reads with 2 errors. c) UpSet plot of reads with 3 errors. d) UpSet plot of reads with >3 errors.

Lastly, to determine whether observed errors represent shared systematic biases or sporadic events, we examined the overlap of error-containing genomic positions across the four platforms, stratified by read-level error count (Figure 45).

The analysis revealed that the vast majority of error-containing genomic regions were unique to individual platforms, with shared error positions representing a small fraction. This platform-private error predominance was consistent regardless of read error counts (Figure 45).

4.2.17. Variant Calling Evaluation

The Proof of any technology lies in the outcome. Even if the aim of this part of the thesis is to evaluate the consistency of sequencing quality of different instruments, the evaluation of variant calls is essential to underline the impact of differences in the sequencing technology. To this end, the variants called on each sequencing run were compared against GIAB Gold Set.

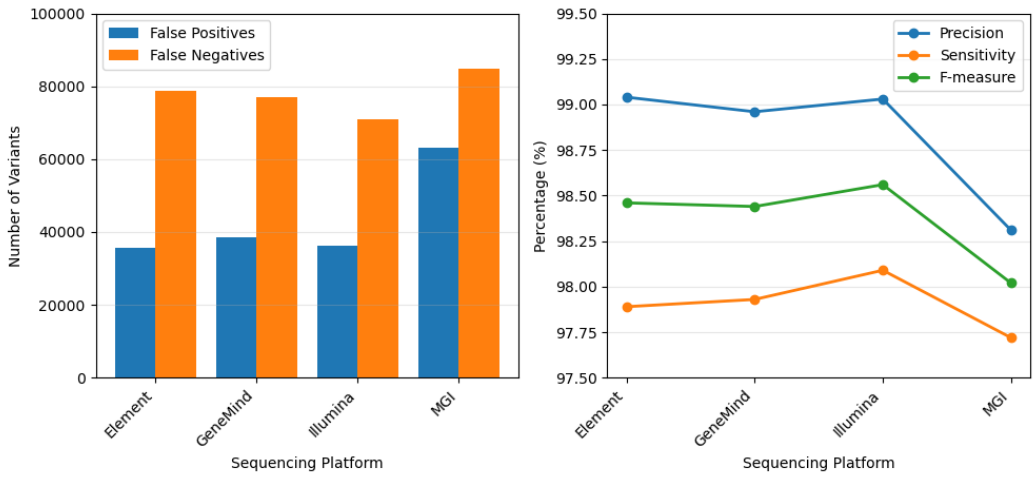


Figure 46 Variant calling benchmarking against GIAB Gold Set, evaluating variants inside GIAB NA12878 High Confidence Regions. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.

In high-confidence regions, overall variant calling performances are highly comparable across all four platforms, with Precision ranging from 98.31% to 99.04%, Sensitivity from 97.72% to 98.09%, and F-measure from 98.02% to 98.56% (Figure 46). Element, GeneMind, and Illumina show similar false positive (35,690–38,588) and false negative counts (70,919–78,720), while MGI T1 displays higher false positives (63,011) and false negatives (84,727), resulting in slightly lower Precision, Sensitivity, and F-measure values.

By stratifying the variant calls in tandem repeats and homopolymers, the patterns observed in the error stratification section translate in differences in variant calling metrics. In tandem repeats, MGI T1 shows slightly lower performances, with Precision and F-measure (45.16% and 61.01%) lower than the other platforms (46.20-47.69% and 61.97-63.25%, respectively), while Sensitivity (93.99%) is comparable to Element (93.88%) and lower than GeneMind and Illumina (95.39-

95.88%). This difference becomes more pronounced in homopolymeric regions, where MGI T1 shows higher false positive and false negative counts, translating into lower Precision, Sensitivity, and F-measure (43.49%, 93.84%, and 59.44%) compared to the other platforms (Precision 46.47-50.01%, Sensitivity 98.24-98.90%, F-measure 63.23-66.28%, Supplementary Figure 6-7).

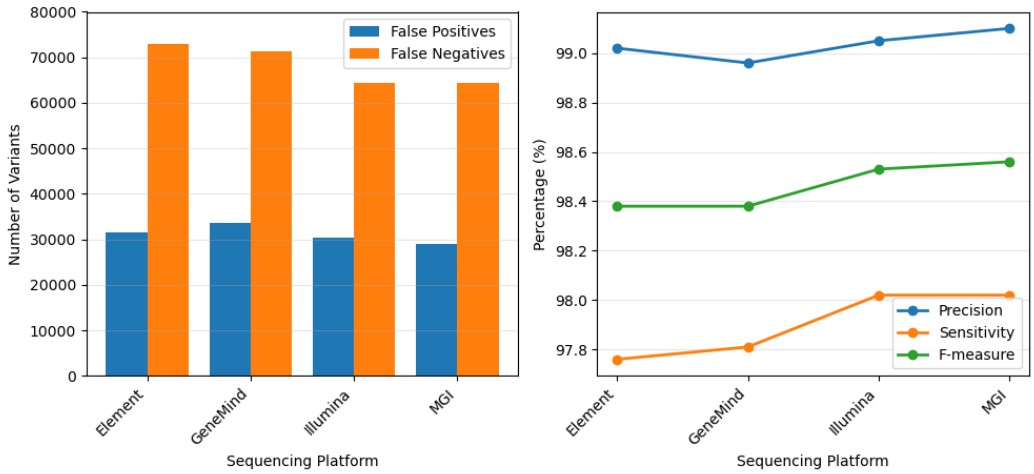


Figure 47 Variant calling benchmarking against GIAB Gold Set, evaluating SNVs inside GIAB NA12878 High Confidence Regions. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.

For SNV calling in high-confidence regions, all four sequencing platforms show highly comparable performances, with Precision ranging from 98.96% to 99.10%, Sensitivity from 97.76% to 98.02%, and F-measure from 98.38% to 98.56% (Figure 47). The main difference is a slightly lower number of false negatives in MGI T1 and Illumina (64,305 and 64,487, respectively) compared to GeneMind and Element (71,304 and 72,954), resulting in higher Sensitivity and F-measure values (Figure 47).

Similarly, SNV calling performances in tandem repeats and homopolymeric regions remain highly comparable across all platforms (Supplementary Figure 8-9). In tandem repeats, Precision ranges from 36.23% to 37.24%, Sensitivity from 93.95% to 95.89%, and F-measure from 52.59% to 53.34%, while in homopolymers Precision ranges from 57.15% to 58.22%, Sensitivity from 98.21% to 98.97%, and F-measure from 72.33% to 73.10%. Overall, these results indicate consistent SNV

detection performances across all sequencing technologies, including repetitive genomic regions (Supplementary Figure 8-9).

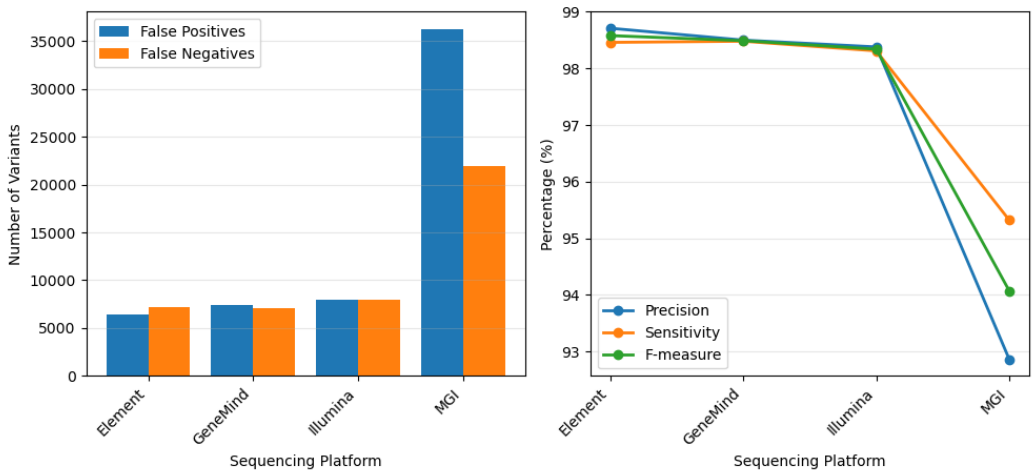


Figure 48 Variant calling benchmarking against GIAB Gold Set, evaluating INDELs inside GIAB NA12878 High Confidence Regions. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.

For INDELs in high-confidence regions, Element, GeneMind, and Illumina show highly consistent performances, with false positive counts between 6,369 and 7,989 and false negatives between 7,091 and 7,915 (Figure 48). These datasets achieve Precision, Sensitivity, and F-measure values ranging from 98.31% to 98.71%. In contrast, MGI T1 shows a marked increase in both false positives (36,265) and false negatives (21,894), resulting in substantially lower Precision (92.86%), Sensitivity (95.32%), and F-measure (94.07%).

In homopolymeric regions, Element, GeneMind, and Illumina again show comparable INDEL performances, with false positive counts ranging from 260,561 to 310,619 and false negatives from 2,865 to 4,642. These correspond to Precision values between 43.06% and 47.12%, Sensitivity between 97.90% and 98.71%, and F-measure values between 59.96% and 63.62%. MGI T1 instead shows the highest false positive and false negative counts (336,151 and 18,745), leading to a marked reduction in Sensitivity (91.54%) and F-measure (54.99%, Supplementary Figure S11).

In tandem repeat regions, Element, GeneMind, and Illumina still maintain similar results, with false positive counts between 103,018 and 118,204 and false negatives between 6,264 and 9,507, corresponding to Precision values from 54.52% to 57.19%, Sensitivity from 92.50% to 95.06%, and F-measure values from 68.90% to 70.68%. MGI T1 again shows slightly lower overall metrics, with the highest false positive and false negative counts (122,683 and 10,857), resulting in the lowest Precision (52.65%), Sensitivity (91.44%), and F-measure (66.82%).

5. Discussion

This thesis focused on two critical aspects of clinical NGS that shape the robustness of downstream analyses: the length of sequenced fragments and the actual accuracy of the generated data. Firstly, we were able to assess the performance of each insert size interval, characterizing its effect on one of the most widely used sequencing methods, whole-exome sequencing. It is commonly accepted that exons and coding regions are short and that, because the majority of clinically relevant variants are located within them, using short insert sizes (~200 bp) improves target enrichment while limiting the amount of data falling outside coding regions[24], [51], [52], [53]. While this idea of short genomic regions is partly correct, we showed that almost half of exonic bases reside in relatively long regions, which would greatly benefit from longer insert-size fragments.

At the same time, the wet-lab processes that determine insert size (for example fragmentation and size selection) are not perfect, and the resulting libraries follow an approximately Gaussian distribution with tails of both short and long fragments. These tails, particularly the short-fragment tail, strongly affect overall performance. Commonly used parameters such as mean or median insert size, which are appropriate descriptors of symmetrical normal distributions, do not always represent the underlying data. In this thesis we showed examples of samples produced with two different versions of the same kit, which share the same average insert size but display very different distribution shapes: one with an optimal, symmetric Gaussian curve and one with a heavily right-skewed distribution. In practice, although the nominal optimal insert-size range is the same, the right-skewed distributions contain a higher proportion of short fragments.

Through the development of a Snakemake-based framework, SnakeBin, and by adopting an Insert Size Bin analysis, we were able to quantify the relative contribution of each segment of the insert-size curve and independently evaluate the performance of each length interval. The analysis showed that short, medium, and long insert sizes (up to 351 bp) have a profound impact on the amount of data required to achieve clinical-grade coverage. In particular, up to one third more sequencing data are required for the shortest and longest tested inserts (101–150 bp,

501-550bp) compared with the optimal range (251–350 bp). This implies both a substantial increase in sequencing costs and a reduction in the number of samples that can be multiplexed per run. In addition to improving coverage efficiency, this work demonstrated that longer insert sizes also affect alignment quality. Previous studies had suggested that longer inserts improve mappability in difficult regions[10], [24], but the magnitude of the benefit and the effect of fine-grained insert-size optimization on alignment quality and on the ability to resolve clinically relevant loci had not been systematically quantified.

In this thesis we showed that, although the effect across the entire WES design is modest (largely because most regions are not particularly complex), long fragments (>200–250 bp) consistently improve mapping quality and genotypability by resolving alignment ambiguities, whereas shorter inserts progressively reduce mapping efficiency and the number of genotypable positions. This is especially evident when focusing on clinically relevant loci from ClinVar and HGMD. In line with the genotypability results, long fragments (>250 bp) achieve optimal coverage of clinical variants, while the reduced mappability of short fragments (<250 bp) prevents reliable genotyping at many clinical sites. In a clinical context this is particularly relevant, because it is impossible to identify a causative variant at a position that cannot be unambiguously read.

Current state-of-the-art pipelines do not rely solely on alignment confidence for variant calling; instead they use complex haplotype reassembly and statistical models to infer the most likely genotype at each genomic position[54]. This abstraction allows correct variant calls even in regions where traditional Smith–Waterman-derived aligners such as BWA-MEM2 assign low mapping quality. For this reason, it was essential to validate the trends observed in mappability and genotypability using a variant-calling benchmark against the GIAB NA12878 Gold Set. In this benchmark we confirmed that short insert sizes impair both specificity and sensitivity: reduced mappability of short fragments increases the number of both false-negative and false-positive variants. When insert size is increased above 250 bp, variant-calling performance reaches an optimal plateau. These results hold in both high-confidence and low-mappability regions of the genome; interestingly, in low-mappability regions the main improvement is a reduction in false positives.

This indicates that the enhanced mappability of longer fragments not only improves coverage efficiency but also reduces alignment artifacts, directly contributing to more reliable clinical sequencing data. Though in most cases, clinical interpretation of WES experiments may not benefit in fine-grade optimization of these aspects (e.g. clearly inherited clinically annotated variants in easy to interpret genomic regions) the optimization of sequencing cost, mappability and signal-to-noise ratio, may be a key to the resolution of complex edge cases.

Modern short-read sequencing platforms estimate base error rates through predictive algorithms trained on instrument-specific signal models, rather than by directly observing mismatches after alignment[28], [29]. These models are proprietary and tightly coupled to the underlying chemistry, which may differ substantially between sequencing-by-synthesis systems such as Illumina NovaSeq X and GeneMind SURFSeq 5000 and polony or DNA-nanoball chemistries such as Element AVITI and MGI T1+[30], [32], [33], [40], [42]. Because model design and training are not standardized across vendors, it is essential to evaluate each platform with a uniform, data-driven framework that measures how well reported quality scores reflect the true accuracy of the sequenced bases.

In this study, sequencing of a single PCR-free NA12878 library across four different platforms under vendor-recommended conditions enabled the development of an automated and fully reproducible workflow. By excluding complex genomic regions, ambiguous alignments, and errors detected as systematic variants, the workflow attempts to focus on genuine sequencing errors, providing a clear view of platform-specific behaviors. The developed and validated pipeline proved effective in capturing these patterns and offers a robust framework for evaluating sequencing quality and empirical error rates. Importantly, although the current analyses are limited to a single run per platform, this workflow lays the groundwork for systematic comparisons across technologies and can be scaled to larger cohorts, enabling more comprehensive and robust assessments of sequencing performance in future studies.

The first step was a per-base analysis to test whether sequencer-assigned Q-scores are consistent with empirical accuracy. Here, the platforms showed markedly

different behaviors, with some coming close to a perfect calibration and others exhibiting clear biases at specific quality levels. For example, Element's quality scores were almost perfectly aligned with the observed error probabilities, whereas MGI tended to underestimate error rates at very low Q, overestimate them at intermediate Q, and then match the empirical accuracy very well at Q40. The comparison was further complicated by the fact that several instruments emit heavily binned Q-scores instead of continuous values, making direct cross-platform comparison of nominal quality more difficult. To address this, an Unified Binning Scheme was introduced, collapsing all scores into three standard classes of low, medium, and high quality; this allowed a direct comparison of accuracy across platforms within the same quality bands. Under this common scheme, differences between platforms remained evident, especially at medium and high Qscores. However, it is important to note that almost all bases produced by every platform fall into the high-quality bin, so the discrepancies observed at low and medium quality affect only a small fraction of the data and differ in their practical impact across instruments. Nevertheless, all platforms produced the vast majority of bases in the high-quality bin (above 90%), and, despite some over- or underestimation in specific intervals, the Q-scores were generally consistent with alignment-derived accuracies, both on raw data and after applying the unified bins; notably, while all platforms attained empirical Q40 base quality in the high quality bin, Element was the highest, while MGI produced almost exclusively very high-quality bases.

Base-by-cycle analyses showed a clear general trend. Reported quality decreased as sequencing cycles progressed while empirical error rate increased, a phenomenon already described in the literature as a consequence of cumulative phasing, pre-phasing, and signal decay effects in high-throughput sequencers[33]. This downward trend was present on all platforms. Platform-specific fluctuations highlighter how these results need to be considered run-specific and further replicated analyses need to be performed to draw platform-specific conclusions.

To better understand the origin of this quality decline, base composition were examined across cycles within the low, medium, and high-quality bins. The results indicated that the main driver of the quality drop is not a dramatic increase in the error rate of high-quality bases, but a progressive reduction in the proportion of

bases assigned to the high-quality bin and a corresponding increase in medium- and low-quality bases as cycle number increases. This shift is consistent with a possible gradual degradation of the optical signal and/or cluster synchronization over time in all chemistries, leading to more conservative or uncertain quality assignments in late cycles.

Given the central role of insert size demonstrated in the first part of the thesis, the relationship between quality, error rates, and fragment length was also investigated. Overall, the average Q-scores assigned by the instruments decreased as insert size increased, while errors remained low and stable for inserts up to approximately 550-600 bp in most platforms, indicating that the underlying mapping remained robust even when the predicted error probability slightly increased. Some platforms proved more tolerant to long inserts than others: GeneMind maintained the highest average Q-scores across the insert-size spectrum, whereas Element showed the steepest decline in reported quality with increasing fragment length. Nonetheless, Element offers different run recipes optimized for specific fragment length ranges, which may partly account for its behavior with long inserts[55]. While no strong mate-dependent bias emerged in the per-cycle quality analysis, it was present for the insert-based analysis. In general, the drop in sequencing quality and rise in error rate were more pronounced in Read 2 than in Read 1 across all platforms. Together, these observations are encouraging, especially when considered alongside the optimal insert-size window identified for WES (approximately 251–300 bp), which lies in a range where all platforms maintain high empirical and predicted quality.

Finally, read-level analyses on the full datasets highlighted substantial differences between platforms in both average alignment identity and the fraction of error-free reads. When all reads were considered, Element showed the lowest overall identity and the smallest proportion of reads without errors, whereas MGI achieved the highest values for both metrics. However, these differences could be run-specific or influenced by platform-specific internal filtering performed before or during base-calling and demultiplexing. To mitigate this effect, the comparison was repeated on filtered datasets retaining only High-Quality reads (average Q30+). Under these conditions, performance differences became much smaller. All platforms reached similar mean alignment identity, with only minor discrepancies

in the fraction of error-free reads. Importantly, in these filtered datasets the mean Q30 values reported by each sequencer were consistently matched by the empirical read-level alignment identities, confirming that Q30 is a reliable target once low-quality data are removed. Overall, the main quality differences between platforms were driven by the proportion of high-quality data produced, rather than by large disparities in the accuracy of the retained high-quality reads. This underscores the central role of rigorous post-sequencing quality control and filtering.

To further characterize the nature of the identified sequencing errors, we investigated their base-quality profiles within High-Quality reads. This analysis provides additional insight into how reliably sequencer-assigned quality scores reflect true error probabilities. Reads containing a single error largely behaved as expected, with a substantial fraction of discrepancies occurring at low- or medium-quality bases. In contrast, reads harboring more than two errors, as well as insertion events, were almost exclusively associated with high-quality (Q30+) bases across all platforms. The predominance of high-quality scores among these errors introduces an important source of uncertainty. Although such events represent only a small fraction of the total data, their quality profiles are inconsistent with a simple sequencing-error model. This suggests that at least a subset of high-Q discrepancies may not correspond to genuine sequencing errors, but instead suggest alignment artefacts, local reference ambiguities, or unresolved genomic complexity. These observations underscore the difficulty of unambiguously distinguishing true sequencing errors from mapping-related artefacts when multiple discrepancies co-occur within the same read. On the other hand, the overlap of error-containing genomic positions across the four platforms was limited, with the vast majority of positions being platform-specific rather than shared, regardless the read error count. Despite the residual uncertainty in the origin of high-quality discrepancies, the small number of common error positions provides confidence in the analytical approach and supports the effectiveness of the filtering strategy applied.

The analysis of error type and genomic context, performed both on the full dataset and on reads containing a single error, revealed platform-specific patterns that further contextualize the observed quality differences. In particular, MGI (and to a

lesser extent GeneMind) showed a higher enrichment of indel errors, especially within complex genomic contexts such as homopolymeric regions and tandem repeats. Although this observation may be influenced by the lack of technical replicates and should therefore be interpreted cautiously, it is consistent with previously reported characteristics of similar sequencing chemistries in the literature[56], [57], [58]. These findings suggest that error composition and genomic context, in addition to overall error rates, are important dimensions to consider when comparing sequencing technologies.

Furthermore, variant calling performance across platforms is consistent with the sequencing error patterns. SNV detection is highly comparable across all technologies, reflecting the overall robustness of substitution calling and the limited impact of platform-specific differences in base quality calibration. In contrast, INDEL calling shows greater divergence, particularly for MGI T1, which exhibits reduced performance relative to the other platforms. This effect is most pronounced in repetitive genomic contexts, such as homopolymers and tandem repeats, where higher false positive and false negative rates lead to decreased precision and F-measure. These observations are in agreement with the increased indel-related error burden and context-specific error enrichment previously described.

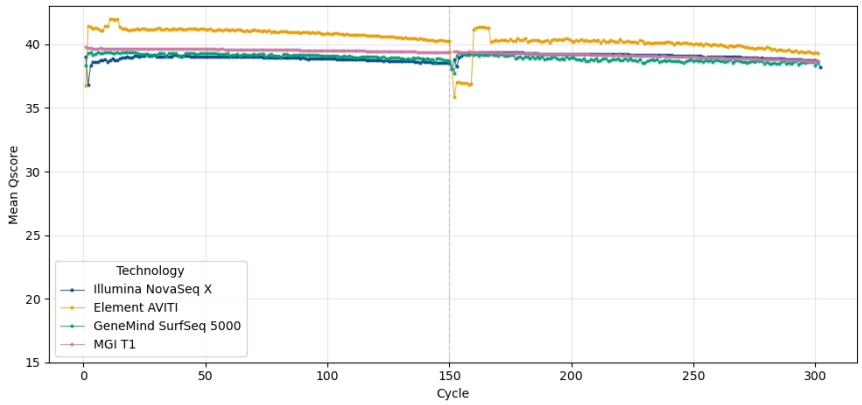
Taken together, these results highlight that optimizing clinical NGS requires treating library design and sequencing platform as a single, integrated system rather than as independent components. Insert size determines how effectively exons are captured, mapped, and genotyped, while platform-specific quality models control how sequencing errors are encoded and subsequently filtered; only when an empirically optimal insert-size window is combined with well-calibrated, high-quality data do WES experiments achieve the best balance between coverage efficiency, sensitivity, and specificity. Within this framework, SnakeBin and the sequencing quality assessment workflow provide complementary tools: the former refines library preparation by identifying the most performant insert-size intervals, and the latter validates whether different sequencers can exploit those intervals to produce reliable high-confidence reads, showing that careful post-sequencing quality control could narrow the gap between platforms and make the most of libraries tuned to the 250-350 bp range.

6. Limitations and Future Directions

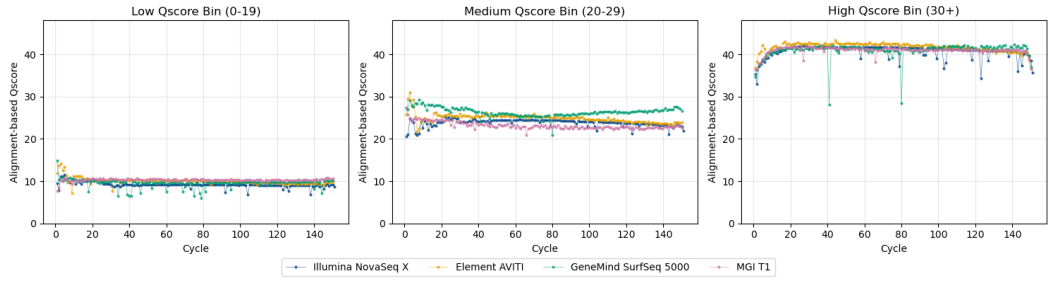
A few limitations of this work should be acknowledged, together with possible directions for future development. First, two different exome-capture kits were used for the clinical WES cohorts and for the NA12878 WES replicates, a choice driven by the availability of already sequenced datasets; employing a single, shared capture design for both parts would further consolidate the observed insert-size effects on coverage, mappability, and variant calling. Second, the cross platform comparison of sequencers did not include technical replicates such as multiple runs or lanes; while this does not affect how each instrument assigns quality scores to individual bases, it may influence both the proportion of high quality bases produced and error rate patterns. Future studies with replicated runs would better characterize this variability and allow more precise conclusions to be drawn regarding intrinsic differences between sequencing platforms and underlying chemistries, distinguishing them from variability arising from run-to-run or lane-specific effects. Lastly, comparisons across conditions are primarily based on descriptive statistics rather than formal hypothesis testing. This choice reflects the nature of high-throughput sequencing data, where the extremely large number of observations can lead to inflated statistical significance even for minimal differences. In addition, observations are not fully independent, as bases originate from the same biological samples, sequencing reads, and genomic regions, violating key assumptions of standard statistical tests. For these reasons, emphasis was placed on consistent trends and reproducibility across datasets rather than on formal statistical significance.

7. Supplementary material

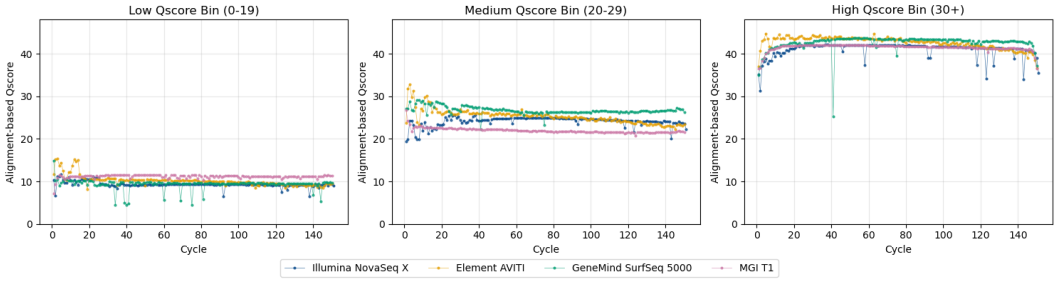
Supplementary Figure S1 Average sequencer-assigned Q-score per cycle for each platform (R1 + R2 combined). For each cycle, the average Q-score was computed as the arithmetic average of the bases Qscores. Cycles 1–150 correspond to Read 1, followed by Read 2 cycles.



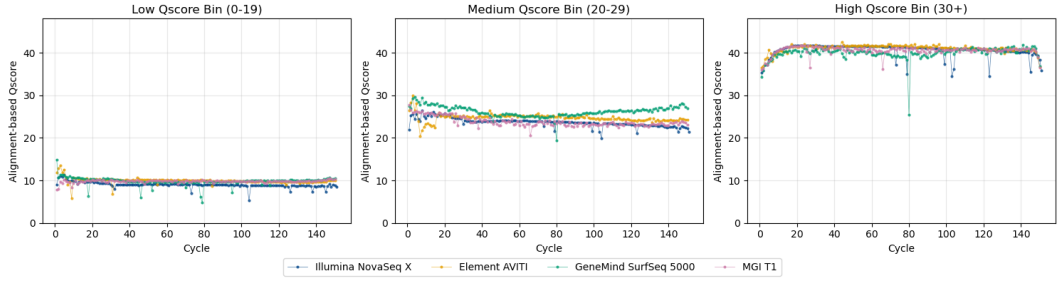
Supplementary Figure S2 For each platform, empirical alignment identity across sequencing cycles for bases classified into Q-score bins: low (Q0-20), medium (Q20-30), high (Q30+).



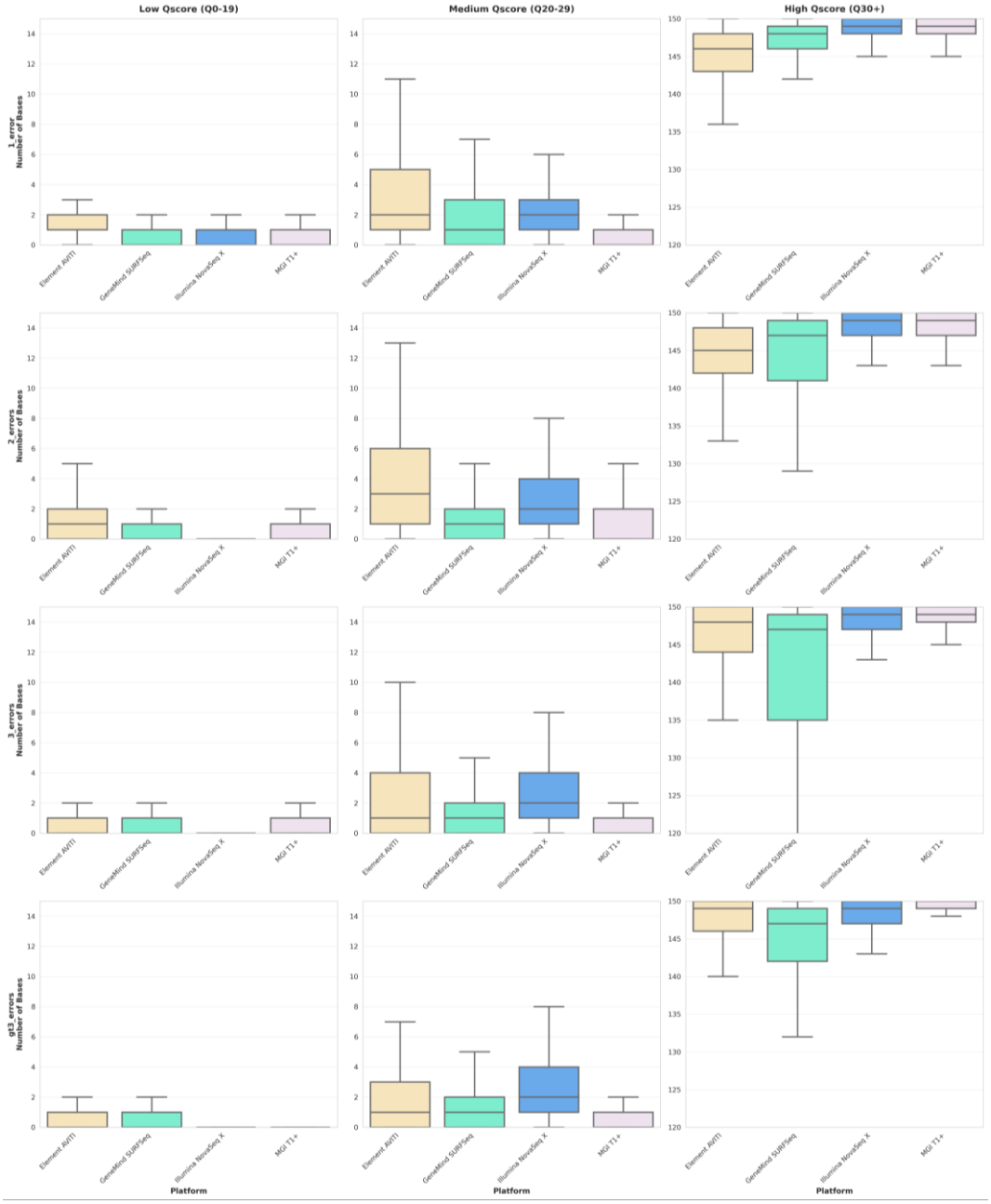
Supplementary Figure S3 For each platform, Read 1 empirical alignment identity across sequencing cycles for bases classified into Q-score bins: low (Q0-20), medium (Q20-30), high (Q30+).



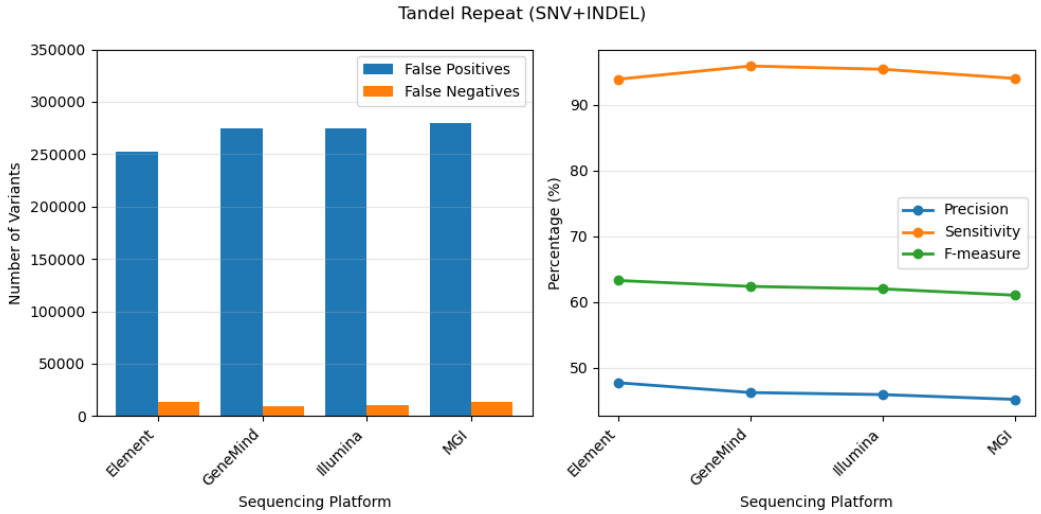
Supplementary Figure S4 For each platform, Read 2 empirical alignment identity across sequencing cycles for bases classified into Q-score bins: low (Q0-20), medium (Q20-30), high (Q30+).



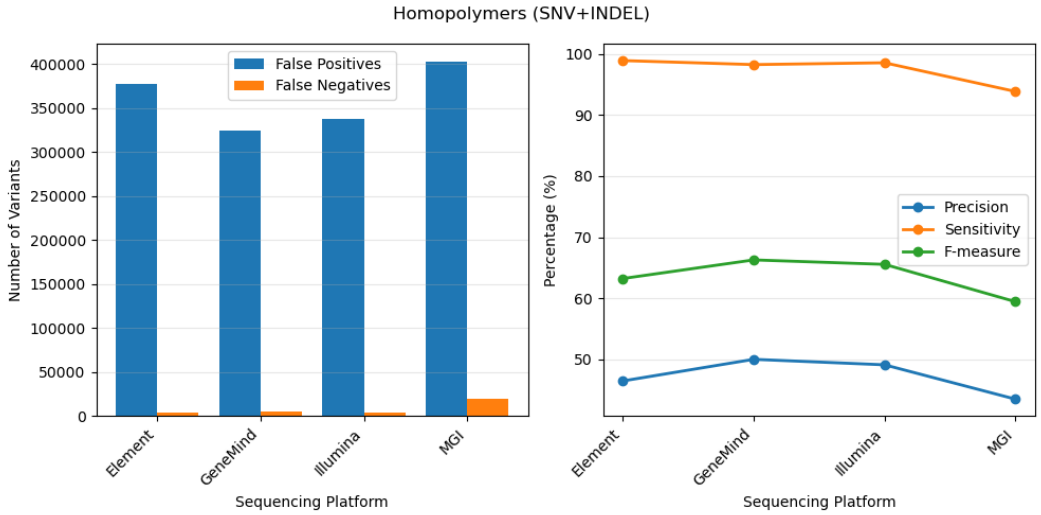
Supplementary Figure S5 For each platform, boxplots showing Phred score distributions across all bases (correct + erroneous) in reads containing 1, 2, 3, or >3 errors. Panels from left to right correspond to low (Q0-20), medium (Q20-30), high (Q30+) base quality bins.



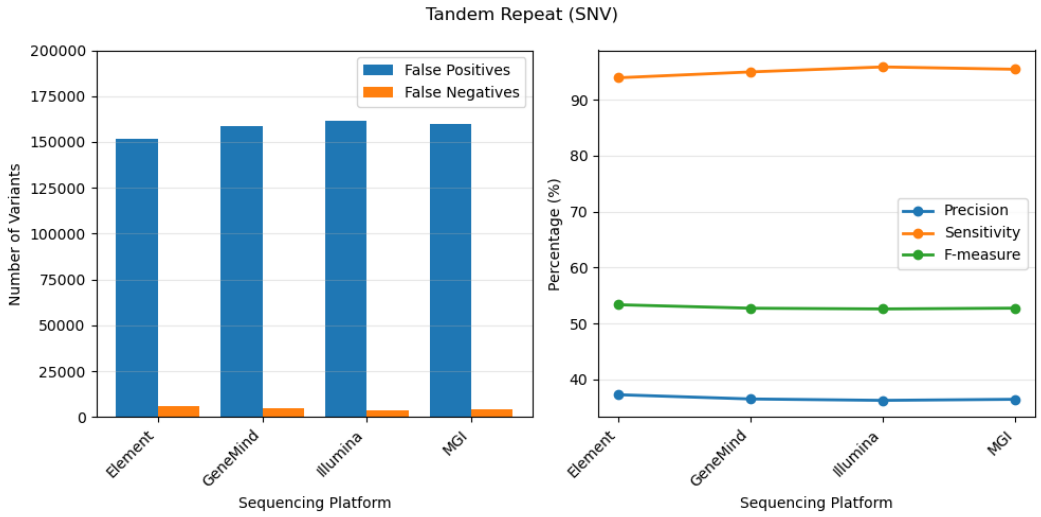
Supplementary Figure S6 *Variant calling benchmarking against GIAB Gold Set, evaluating variants inside Tandem Repeats, as defined in GIAB Genome Stratifications v3.5. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.*



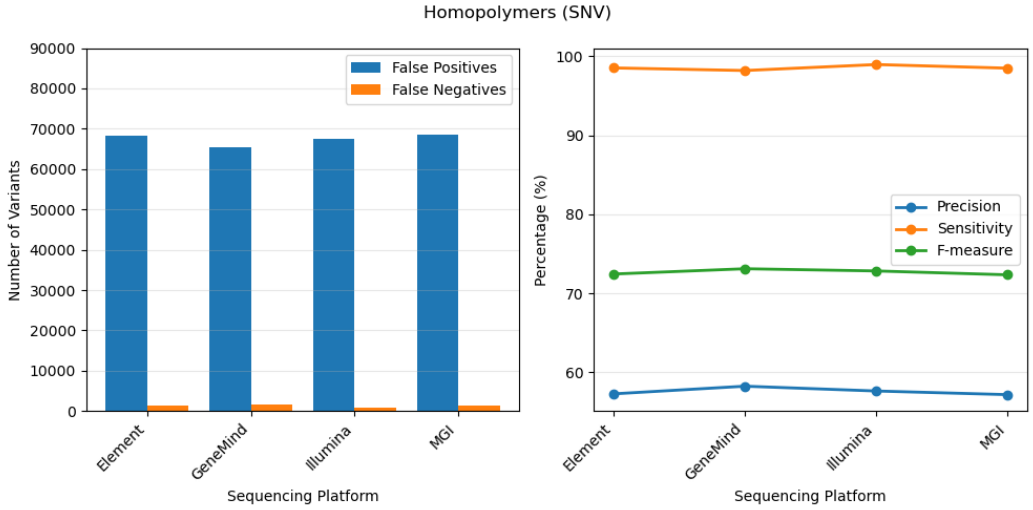
Supplementary Figure S7 *Variant calling benchmarking against GIAB Gold Set, evaluating variants inside Homopolymers, as defined in GIAB Genome Stratifications v3.5. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.*



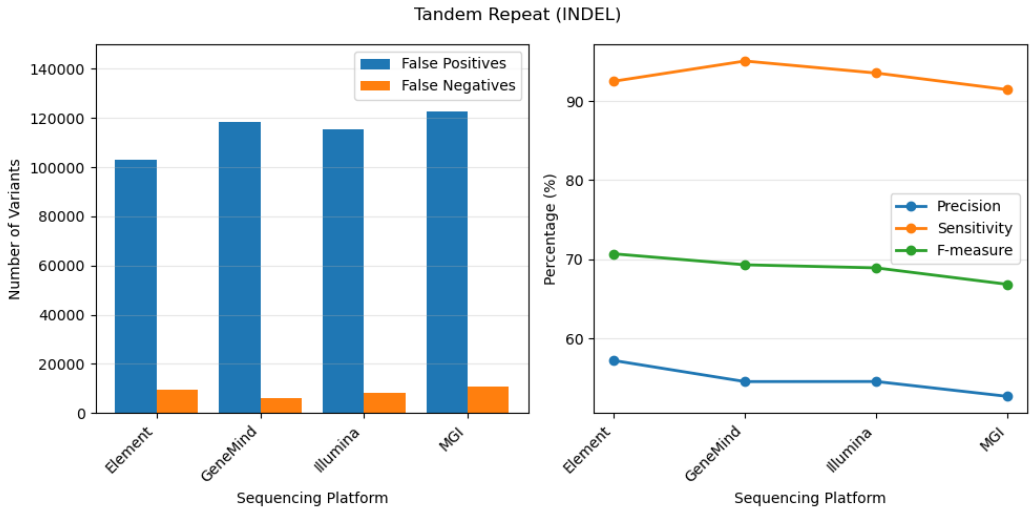
Supplementary Figure S8 Variant calling benchmarking against GIAB Gold Set, evaluating SNVs inside Tandem Repeats, as defined in GIAB Genome Stratifications v3.5. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.



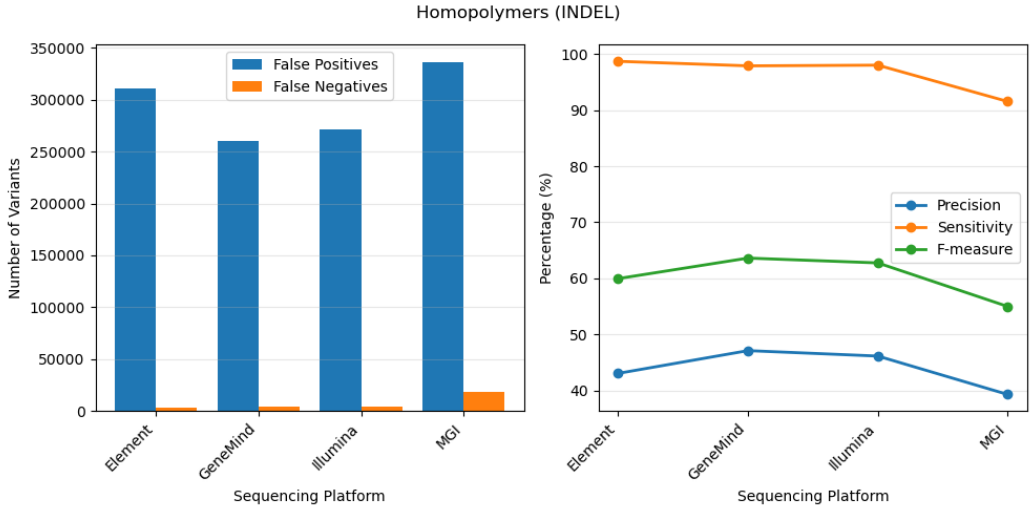
Supplementary Figure S9 Variant calling benchmarking against GIAB Gold Set, evaluating SNVs inside Homopolymers, as defined in GIAB Genome Stratifications v3.5. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.



Supplementary Figure S10 Variant calling benchmarking against GIAB Gold Set, evaluating INDELs inside Tandem Repeats, as defined in GIAB Genome Stratifications v3.5. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.



Supplementary Figure S11 Variant calling benchmarking against GIAB Gold Set, evaluating INDELs inside Homopolymers, as defined in GIAB Genome Stratifications v3.5. For each platform the barplot shows false positive (FP) and false negative (FN) calls, while the lineplot shows the corresponding precision, sensitivity and F-measure.



8. Bibliography

[1] A. Telenti *et al.*, “Deep sequencing of 10,000 human genomes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 42, pp. 11901–11906, Oct. 2016, doi: 10.1073/PNAS.1613365113.

[2] A. C. Lionel *et al.*, “Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test,” *Genet. Med.*, vol. 20, no. 4, pp. 435–443, Apr. 2018, doi: 10.1038/GIM.2017.119.

[3] Z. Stark *et al.*, “Integrating Genomics into Healthcare: A Global Responsibility,” *Am. J. Hum. Genet.*, vol. 104, no. 1, pp. 13–20, Jan. 2019, doi: 10.1016/J.AJHG.2018.11.014.

[4] D. J. Stavropoulos *et al.*, “Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine,” *NPJ Genom. Med.*, vol. 1, Jan. 2016, doi: 10.1038/NPJGENMED.2015.12.

[5] S. B. Ng *et al.*, “Targeted capture and massively parallel sequencing of 12 human exomes,” *Nature 2009 461:7261*, vol. 461, no. 7261, pp. 272–276, Aug. 2009, doi: 10.1038/nature08250.

[6] M. J. Bamshad *et al.*, “Exome sequencing as a tool for Mendelian disease gene discovery,” *Nature Reviews Genetics 2011 12:11*, vol. 12, no. 11, pp. 745–755, Sep. 2011, doi: 10.1038/nrg3031.

[7] F. J. Sedlazeck *et al.*, “Accurate detection of complex structural variations using single-molecule sequencing,” *Nature Methods 2018 15:6*, vol. 15, no. 6, pp. 461–468, Apr. 2018, doi: 10.1038/s41592-018-0001-7.

[8] L. J. Jennings *et al.*, “Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists,” *J. Mol. Diagn.*, vol. 19, no. 3, pp. 341–365, May 2017, doi: 10.1016/J.JMOLDX.2017.01.011.

[9] T. Hu, N. Chitnis, D. Monos, and A. Dinh, “Next-generation sequencing technologies: An overview,” *Hum. Immunol.*, vol. 82, no. 11, pp. 801–811, Nov. 2021, doi: 10.1016/J.HUMIMM.2021.02.012.

[10] B. Iadarola *et al.*, “Shedding light on dark genes: enhanced targeted resequencing by optimizing the combination of enrichment technology and DNA fragment length,” *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 9424-, Jun. 2020, doi: 10.1038/s41598-020-66331-z.

[11] O. J. Dillon *et al.*, “Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders,” *Eur. J. Hum. Genet.*, vol. 26, no. 5, pp. 644–651, May 2018, doi: 10.1038/S41431-018-0099-1.

[12] C. R. Marshall *et al.*, “Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease,” *npj Genomic Medicine 2020 5:1*, vol. 5, no. 1, pp. 47-, Oct. 2020, doi: 10.1038/s41525-020-00154-9.

[13] N. Dwarshuis *et al.*, “The GIAB genomic stratifications resource for human reference genomes,” *Nature Communications 2024 15:1*, vol. 15, no. 1, pp. 9029-, Oct. 2024, doi: 10.1038/s41467-024-53260-y.

[14] S. Marco-Sola, M. Sammeth, R. Guigó, and P. Ribeca, “The GEM mapper: fast, accurate and versatile alignment by filtration,” *Nature Methods 2012 9:12*, vol. 9, no. 12, pp. 1185–1188, Oct. 2012, doi: 10.1038/nmeth.2221.

[15] M. J. Landrum *et al.*, “ClinVar: improvements to accessing data,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D835–D844, Jan. 2020, doi: 10.1093/NAR/GKZ972.

[16] P. D. Stenson *et al.*, “The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies,” *Hum. Genet.*, vol. 136, no. 6, pp. 665–677, Jun. 2017, doi: 10.1007/S00439-017-1779-6.

[17] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/BIOINFORMATICS/BTP352.

[18] “IDT xGen DNA Library Prep EZ Kit and xGen DNA Library Prep EZ UNI Kit Protocol (RUO21-0413_002 06/24) | Enhanced Reader.”

[19] “KAPA HyperCap Workflow v3.4 Instructions For Use with: KAPA HyperExome Probes, KAPA HyperExome V2 Probes, KAPA HyperCap Heredity Panel, KAPA HyperChoice, and KAPA HyperExplore Probes,” 2015, Accessed: Dec. 04, 2025. [Online]. Available: www.hyperdesign.com.

[20] Illumina, “Illumina DNA Prep with Exome v2 Enrichment Reference Guide (1000000157112),” 2023, Accessed: Dec. 04, 2025. [Online]. Available: www.illumina.com/company/legal.html.

[21] “Agilent Technologies SureSelect XT HS Target Enrichment System For Illumina Multiplexed Sequencing Platforms Protocol SureSelect platform manufactured with Agilent SurePrint Technology For Research Use Only.

Not for use in diagnostic procedures," 2006, Accessed: Dec. 04, 2025. [Online]. Available: www.agilent.com/en/contact-us/page.

[22] M. A. Quail *et al.*, "A large genome center's improvements to the Illumina sequencing system," *Nat. Methods*, vol. 5, no. 12, pp. 1005–1010, 2008, doi: 10.1038/NMETH.1270.

[23] A. Krasnenko *et al.*, "Effect of DNA insert length on whole-exome sequencing enrichment efficiency: an observational study," *Adv. Genomics Genet.*, vol. 8, pp. 13–15, Jun. 2018, doi: 10.2147/AGG.S162531.

[24] C. Pommerenke *et al.*, "Enhanced whole exome sequencing by higher DNA insert lengths," *BMC Genomics*, vol. 17, no. 1, p. 399, May 2016, doi: 10.1186/S12864-016-2698-Y.

[25] S. Das, N. K. Biswas, and A. Basu, "Mapinsights: deep exploration of quality issues and error profiles in high-throughput sequence data," *Nucleic Acids Res.*, vol. 51, no. 14, pp. e75–e75, Aug. 2023, doi: 10.1093/NAR/GKAD539.

[26] B. Ewing and P. Green, "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities," *Genome Res.*, vol. 8, no. 3, pp. 186–194, Mar. 1998, doi: 10.1101/GR.8.3.186.

[27] B. Ewing, L. D. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using phred. I. Accuracy assessment," *Genome Res.*, vol. 8, no. 3, pp. 175–185, 1998, doi: 10.1101/GR.8.3.175.

[28] R. Pereira, J. Oliveira, and M. Sousa, "Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics," *J. Clin. Med.*, vol. 9, no. 1, p. 132, Jan. 2020, doi: 10.3390/JCM9010132.

[29] Illumina, "Quality Scores for Next-Generation Sequencing," 2011, Accessed: Dec. 04, 2025. [Online]. Available: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/

[30] S. Arslan *et al.*, "Sequencing by avidity enables high accuracy with low reagent consumption," *Nature Biotechnology 2023 42:1*, vol. 42, no. 1, pp. 132–138, May 2023, doi: 10.1038/s41587-023-01750-7.

[31] H. M. Kim *et al.*, "Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing," *Gigascience*, vol. 10, no. 3, p. giab014, Mar. 2021, doi: 10.1093/GIGASCIENCE/GIAB014.

[32] G. J. Kastanis, L. V. Santana-Quintero, M. Sanchez-Leon, S. Lomonaco, E. W. Brown, and M. W. Allard, "In-depth comparative analysis of Illumina®

MiSeq run metrics: Development of a wet-lab quality assessment tool," *Mol. Ecol. Resour.*, vol. 19, no. 2, p. 377, Mar. 2019, doi: 10.1111/1755-0998.12973.

[33] "Why does the per base sequence quality decrease over the read in Illumina?" Accessed: Dec. 04, 2025. [Online]. Available: https://www.ecseq.com/support/ngs/why-does-the-sequence-quality-decrease-over-the-read-in-illumina

[34] "Chemistry and Imaging on the NovaSeq X Series Instruments | Illumina Knowledge." Accessed: Dec. 04, 2025. [Online]. Available: https://knowledge.illumina.com/instrumentation/novaseq-x-x-plus/instrumentation-novaseq-x-x-plus-reference_material-list/000007970

[35] "NovaSeq X Specifications | Capacity for high-intensity genomics." Accessed: Dec. 04, 2025. [Online]. Available: https://www.illumina.com/systems/sequencing-platforms/novaseq-x-plus/specifications.html

[36] D. R. Bentley *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature 2008 456:7218*, vol. 456, no. 7218, pp. 53–59, Nov. 2008, doi: 10.1038/nature07517.

[37] "NovaSeq X v1.2 software enables sequencing with 80% of bases >= Q40." Accessed: Dec. 04, 2025. [Online]. Available: https://www.illumina.com/science/genomics-research/articles/data-quality-q-scores.html

[38] "SURFSeq 5000* High-throughput Sequencing Platform Fast Cost-effective Efficient Scalable".

[39] Y. Sun *et al.*, "Assessing the impact of sequencing platforms and analytical pipelines on whole-exome sequencing," *Front. Genet.*, vol. 15, p. 1334075, May 2024, doi: 10.3389/FGENE.2024.1334075/FULL.

[40] E. Biosciences, "Element AVITI System Specification Sheet".

[41] "Q50+: Setting a New Standard for Sequencing Accuracy with Cloudbreak UltraQ | Element Biosciences." Accessed: Dec. 04, 2025. [Online]. Available: https://www.elementbiosciences.com/blog/enhancing-sequencing-accuracy-with-cloudbreak-ultraq

[42] "DNBSEQ Technology | MGI-tech | Leading Life Science Innovation." Accessed: Dec. 04, 2025. [Online]. Available: https://mgi-tech.eu/technology

[43] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM,” Mar. 2013, Accessed: Dec. 09, 2025. [Online]. Available: https://arxiv.org/abs/1303.3997v2

[44] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “fastp: an ultra-fast all-in-one FASTQ preprocessor,” *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/BIOINFORMATICS/BTY560.

[45] F. Mölder *et al.*, “Sustainable data analysis with Snakemake,” *F1000Research 2021 10:33*, vol. 10, p. 33, Jan. 2021, doi: 10.12688/f1000research.29032.1.

[46] P. Danecek *et al.*, “Twelve years of SAMtools and BCFtools,” *Gigascience*, vol. 10, no. 2, pp. 1–4, Jan. 2021, doi: 10.1093/GIGASCIENCE/GIAB008.

[47] A. McKenna *et al.*, “The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, Sep. 2010, doi: 10.1101/GR.107524.110.

[48] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010, doi: 10.1093/BIOINFORMATICS/BTQ033.

[49] J. G. Cleary *et al.*, “Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines,” *bioRxiv*, p. 023754, Aug. 2015, doi: 10.1101/023754.

[50] J. K. Bonfield *et al.*, “HTSlib: C library for reading/writing high-throughput sequencing data,” *Gigascience*, vol. 10, no. 2, Feb. 2021, doi: 10.1093/GIGASCIENCE/GIAB007.

[51] S. Richards *et al.*, “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.*, vol. 17, no. 5, p. 405, May 2015, doi: 10.1038/GIM.2015.30.

[52] M. Q. Zhang, “Statistical Features of Human Exons and Their Flanking Regions,” *Hum. Mol. Genet.*, vol. 7, no. 5, pp. 919–932, May 1998, doi: 10.1093/HMG/7.5.919.

[53] E. Samorodnitsky *et al.*, “Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing,” *Hum. Mutat.*, vol. 36, no. 9, pp. 903–914, Sep. 2015, doi: 10.1002/HUMU.22825.

[54] R. Poplin *et al.*, “Scaling accurate genetic variant discovery to tens of thousands of samples,” *bioRxiv*, p. 201178, Jul. 2018, doi: 10.1101/201178.

[55] “Run Planning for Sequencing | Element Biosciences Software Documentation.” Accessed: Dec. 09, 2025. [Online]. Available: https://docs.elembio.io/docs/elembio-cloud/runs/seq-run-planning/

[56] O. K. Tørresen *et al.*, “Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases,” *Nucleic Acids Res.*, vol. 47, no. 21, pp. 10994–11006, Dec. 2019, doi: 10.1093/NAR/GKZ841.

[57] N. Stoler and A. Nekrutenko, “Sequencing error profiles of Illumina sequencing instruments,” *NAR Genom. Bioinform.*, vol. 3, no. 1, Jan. 2021, doi: 10.1093/NARGAB/LQAB019.

[58] S. I. Jeanjean *et al.*, “A detailed analysis of second and third-generation sequencing approaches for accurate length determination of short tandem repeats and homopolymers,” *Nucleic Acids Res.*, vol. 53, no. 5, p. 131, Feb. 2025, doi: 10.1093/NAR/GKAF131.