Genomic Analysis and Lineage Tracing Using Two Sequencing Platforms

Park, Sung Joon

Department of Medical Device Engineering and Management
Graduate School
Yonsei University

Genomic Analysis and Lineage Tracing Using Two Sequencing Platforms

Advisor Oh, Ji Won

**A Master's Thesis Submitted
to the Department of Medical Device
Engineering and Management
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Master of Sung Joon Park**

Park, Sung Joon

June 2025

Genomic Analysis and Lineage Tracing
Using Two Sequencing Platforms

This Certifies that the Master's Thesis
of Park, Sung Joon is Approved

Committee Chair	**Oh, Ji Won**
Committee Member	**Yang, Hun Mu**
Committee Member	**Choi, Seock Hwan**

Department of Department of Medical Device
Engineering and Management
Graduate School
Yonsei University
June 2025

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Genomic Analysis and Lineage Tracing Using Two Sequencing Platforms

Recent studies have revealed the occurrence of somatic mosaicism in human organisms, where genetically distinct cells coexist due to somatic mutations during cell division in early development or later in life. Each cell, with its unique combination of mutations, acts as a genetic barcode for lineage tracing. This discovery provides important clues for tracing the developmental origin of cells and understanding the origins and progression of cancer.

In this study, I compared and analyzed Illumina, the most widely used sequencing platform to date, and Ultima Genomics, which has recently attracted attention as a wafer-based sequencing technology, for the purpose of tracing this lineage. To identify cell lineages from zygote differentiation to adulthood, postmortem tissue samples were utilized. After collecting tissues from the anterior left and anterior right legs, I secured enough DNA through primary cell culture and single-cell clone expansion and applied a protocol that considered the characteristics of each cell type to generate high-quality whole-genome data.

Data were generated on both Illumina and Ultima Genomics platforms, followed by a comparative analysis of their quality and mutation detection performance. Statistical quality evaluation was performed based on indicators such as base-by-base error rate, reference genome coverage, and the percentage of reads remaining after removal of PCR duplicate reads. In addition, various genomic mutations such as SNVs, INDELs, CNVs, and SVs were detected and analyzed to perform a comparison between platforms. Furthermore, the annotation of genomic mutations, including BRCA mutations, and their clinical significance were evaluated to examine practical applicability.

Finally, we evaluated the fidelity with which the new technology reproduces existing lineage tracing methods by identifying major and minor cell branches based on previously reported early embryonic mutations and analyzing cell-specific somatic mutations. This study is expected to provide a basis for selecting sequencing technologies in future studies of somatic mosaicism and disease occurrence.

Key words : **Somatic mosaicism, Somatic mutations, Lineage tracing, Whole Genome Sequencing (WGS), Next-generation sequencing (NGS), Illumina, Ultima genomics, Development, Embryogenesis.**

1. INTRODUCTION

1.1 Somatic Mosaicism and Its Implications for Lineage Tracing

Somatic mosaicism refers to the phenomenon in which multiple cells with different genetic characteristics coexist within the same organism due to somatic mutations that occur during cell division in early development or later in life. [1] In the past, it was thought that all cells in multicellular organisms, including humans, shared the same genetic information. However, all cells originate from the same zygote, but over time, different genetic mutations accumulate in individual cells, forming genetic heterogeneity within the organism. However, recent advances in single-cell genome analysis and ultra-high-resolution sequencing technologies have revealed that genetic diversity can exist between cells in both normal and tumor tissues [1, 2]. These technological advances are highlighting the importance of studying somatic mosaicism to understand not only biological development but also the initiation and progression of diseases such as cancer [3].

In particular, recent studies have attempted to reconstruct the cell lineage formed during human development by analyzing somatic mutations in postmortem human tissues. [4, 5] These studies provide a powerful approach to indirectly reconstruct early developmental stages that are difficult to access in living organisms and suggest new possibilities for restoring the entire human developmental lineage based on developmental traces retained in postnatal tissues.

The core concept of these studies is that somatic mutations accumulated by cells during development serve as a genetic barcode [3, 6]. When mutations occur during the repeated divisions of the fertilized egg to create new cells, those mutations are passed on to all descendant cells. As a result, each cell has a different combination of mutations, and the pattern of mutations shared only by a specific cell population acts as a unique fingerprint that reveals the developmental path of each cell.

These genetic barcodes provide quantitative information about the timing of mutation occurrence beyond the mere presence of mutations. For example, a mutation occurring in one cell during the two-cell stage of a fertilized egg will be observed in approximately 50% of all cells, with a variant allele frequency (VAF) of approximately 50%. If the mutation occurs at the 4-cell or 8-cell stage, the VAF decreases to approximately 25% and 12.5%, respectively. Since these VAF values indirectly reflect the timing of lineage branching along the developmental time axis, they serve as key indicators for reconstructing the cell lineage tree with spatial and temporal precision [3, 7, 8].

This lineage tracing method based on somatic mutations is highly useful for understanding the origins of normal tissue development and identifying the cells from which diseases such as cancer originate. Its research value is expanding as a powerful tool for analyzing and quantifying lineage structures among cells in various physiological and pathological processes, including tumor clonal expansion, tissue regeneration, degeneration, and aging [1, 3, 8, 9].

1.2 Lineage Tracing Studies Enabled by Next-Generation Sequencing (NGS)

Mutation analysis for somatic mosaicism is mostly performed using Next-Generation Sequencing (NGS) technology. This process involves extracting DNA from each cell or group of cells, decoding the base sequence, mapping it to the reference genome, and analyzing the differences (variant calling) to detect mutations. This sequencing plays a key role in identifying the location and frequency of somatic mutations, and in inferring their timing and the cell lineage structure. There are three main sequencing strategies used in somatic mosaicism research [1, 3].

The first is the bulk sequencing method, which mixes DNA extracted from multiple cells into one sample and analyzes it. This method is relatively inexpensive and simple; however, as analysis is performed on a mixture of various clones, mutations present in a low proportion of cells may be difficult to detect due to low VAF. For example, if only 12.5% of all cells carry a specific mutation, theoretically the mutation will appear at a VAF of about 6%, and due to technical limitations, mutations in less than 5% of the cell population are likely to go undetected due to background noise [8].

The second strategy is single-cell sequencing. This method isolates individual cells and performs sequencing independently, allowing direct identification of somatic mutations that exist only in specific cells [2, 7]. However, the very limited amount of DNA obtained from a single cell necessitates whole genome amplification. Potential coverage bias or allelic dropout during this process may hinder interpretation accuracy, and there are also limitations in technical complexity and cost. Nevertheless, recent technological advancements are gradually overcoming these challenges [9].

The third strategy is based on clonal expansion. In this method, single cells are selected, and their daughter cells with identical genetic characteristics are sufficiently expanded through cell culture, after which bulk sequencing is performed on this expanded population. This strategy combines the sensitivity of single-cell analysis with the stability of bulk analysis and allows identification of mutations in a single cell at high resolution. It provides stable genome coverage and is considered a suitable method for somatic mosaicism research because it enables high-precision lineage tracing [4, 5]. However, distinguishing artifacts that may arise during cell culture from original somatic mutations remains a challenge [10, 11].

In this study, I adopted whole genome sequencing based on clonal expansion among these three strategies and conducted our analysis accordingly. I secured clones derived from single cells and analyzed their genomes to detect somatic mutations with high sensitivity. Furthermore, I aimed to infer lineage structures based on genetic similarities and differences among the clones. This approach serves as an effective analytical framework for reconstructing cell branching processes in the early stages of development and for increasing the precision of cell lineage tracing [6].

1.3 Expansion of Lineage Tracing Through the Adoption of Emerging Sequencing Platforms

Next-generation sequencing (NGS) is one of the core technologies that have driven the rapid advancement of genome research by providing significantly higher throughput than conventional Sanger sequencing [12]. NGS enables the high-speed acquisition of sequence information at the whole-genome level, facilitating precise genome interpretation such as genetic mutation analysis, gene expression studies, and cell lineage tracing [13, 14].

The most widely used NGS platform to date is the Illumina system, which adopts the sequencing-by-synthesis (SBS) method. Illumina's sequencing process begins with fragmenting sample DNA into small pieces, followed by attaching each fragment to an adapter fixed on a flow cell to form clusters. Then, fluorescently labeled nucleotides are incorporated one by one, and the fluorescent signals emitted during base incorporation are detected by a high-resolution camera to decode the sequence. This method ensures high accuracy and consistency and has been established as a standard technology in numerous genome analysis studies to date [15]. However, the requirement for repeated cycles of incorporation, signal capture, and imaging for each base limits sequencing speed and cost-efficiency. Additionally, its reliance on complex optical equipment is considered a disadvantage [16]. Meanwhile, Ultima Genomics has recently gained attention as a new sequencing technology designed to overcome these limitations of Illumina. Ultima employs a non-optical, flow-based sequencing method, which eliminates fluorescence-based image analysis and simplifies the overall process. In this method, DNA is amplified and immobilized on a wafer disk, and solutions containing each nucleotide type (dNTP) are sequentially flowed across the surface. When a nucleotide binds to DNA, a specific chemical reaction occurs, and the presence or absence of this reaction is used to detect nucleotide incorporation. By omitting fluorescence-based detection, this method offers significant advantages in sequencing speed, cost, and instrumentation simplicity.

In this study, based on the two platforms introduced above—Illumina and Ultima Genomics—I aim to compare and analyze the differences in sequencing characteristics and results under identical analytical conditions. Through this comparison, I seek to evaluate the applicability and performance of newly introduced sequencing technologies and to provide practical criteria for selecting the optimal platform in future somatic mosaicism-based lineage tracing and mutation detection studies [17].

In lineage tracing studies, the choice of sequencing method directly influences the resolution, scalability, and cost-efficiency of the analysis [9]. For instance, in conventional bulk sequencing-based lineage tracing, studies are typically conducted using a relatively small number of samples, ranging from as few as 3–5 to around 10–20 datasets. While this approach allows for an overall estimation of clonal structures and the identification of shared mutations, it has limitations in resolving fine-scale lineage relationships, particularly when rare subclones are present [18].

By contrast, studies utilizing single-cell clonal expansion methods have demonstrated significantly

higher resolution in reconstructing cell lineage trees. These studies often rely on sequencing 100 to 200 or more individual clones derived from single cells, and recent advances in cell culture and amplification protocols have enabled even larger-scale experiments [5]. Such approaches have been shown to capture early postzygotic mutations and to uncover the branching architecture of human development with greater precision [6]. However, the increase in sample throughput inevitably leads to a corresponding rise in sequencing costs and processing time.

In this context, the emergence of Ultima Genomics, a cost-effective, non-optical sequencing platform, provides an excellent opportunity for large-scale lineage tracing studies. Ultima Genomics simplifies the configuration of the sequencer instrument through flow-based sequencing chemistry, significantly improves sequencing speed, and dramatically reduces costs, and is attracting attention as a platform suitable for high-throughput analysis requiring hundreds of clone samples [17, 19]. In particular, with the recent advancement of sequencing technology, the cost of whole-genome sequencing has decreased to about $200 for Illumina's NovaSeq X platform and about $100 for Ultima Genomics, significantly increasing the feasibility of large-scale somatic mosaic studies. [20] Considering this potential, this study aimed to evaluate how applicable the Ultima Genomics platform is compared to the widely used Illumina system.

This study addresses three primary objectives. First, I compare the core technological differences and operational characteristics of Illumina and Ultima Genomics, focusing on their sequencing principles, performance, and platform requirements. Second, I investigate the respective strengths and limitations of these platforms when applied to genome analysis and somatic lineage tracing, considering factors such as variant detection sensitivity, data quality, and scalability. Finally, based on experimental data generated from both platforms, I evaluate the feasibility of adopting a new sequencing platform—specifically Ultima Genomics—for high-resolution lineage reconstruction.
By departing from the conventional single-platform paradigm and exploring the integration of distinct sequencing technologies, this study aims to propose a more accurate, scalable, and cost-effective framework for future somatic mosaicism–based lineage tracing research [4, 9].

2. MATERIALS AND METHODS

2.1 Primary Culture of Skin from Postmortem Body

Skin tissues were harvested from the anterior regions of both left and right legs of postmortem human donors within 24 hours of death (Figure 1B). Upon collection, dermal tissues were rinsed twice with phosphate-buffered saline (PBS) under sterile conditions. Enzymatic dissociation was performed using a collagenase-dispase (CD) solution (1 mg/mL in PBS) at 37°C for 4–6 hours. Following digestion, tissues were neutralized with culture medium (20% FBS DMEM), trimmed to remove residual adipose and vascular components, and cut into fragments of approximately 2–3 mm in diameter. The tissue fragments then were transferred to collagen I-coated 24-well plates (Corning BioCoat), with each well containing 200 μL of pre-warmed culture medium. Cultures were maintained in a humidified incubator at 37°C with 5% CO_2. The culture medium—Dulbecco's Modified Eagle Medium (DMEM) low glucose supplemented with 20% fetal bovine serum (FBS), 100 IU/mL penicillin, 100 μg/mL streptomycin, 2 mM L-glutamine, and 1 μg/mL amphotericin B (Fungizone; all from Gibco)—was replaced every four days to remove floating debris and non-adherent cells. Fibroblasts typically migrated out of the tissue fragments and proliferated over the course of ~2 weeks. Once confluency was reached, cells were passaged into larger culture vessels (60-mm or 100-mm dishes) for further expansion (Figure 1A).

2.2 Single-cell Clonal Expansion

Single-cell clonal expansion was conducted from primary cultures of skin dermis using fluorescence-activated cell sorting (FACS). Subconfluent fibroblast cultures were enzymatically dissociated using TrypLE Express (Gibco), and the resulting cell suspensions were passed through a 40-μm cell strainer (SPL) to ensure single-cellularity. Cells were sorted using a FACSAria II cell sorter (BD Biosciences) directly into 96-well culture plates (Corning), without any fluorescent staining. Plates were visually inspected to confirm single-cell occupancy under microscopy. Clonal expansion proceeded stepwise from 96-well plates to 24-well and then 6-well plates. Final expansion and harvest were performed in 100-mm culture dishes.

2.3 DNA Extraction and Quantification, Quality Control

Genomic DNA was extracted from the collected cells using the DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Briefly, tissue samples were lysed in ATL buffer and proteinase K at 56°C until complete digestion, followed by incubation with AL buffer and ethanol to facilitate DNA binding to the silica membrane. The column was washed with AW1 and AW2 buffers, and genomic DNA was eluted in AE buffer. The concentration of extracted DNA was measured using a Qubit 4 Fluorometer with the Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific), and purity was assessed by A260/A280 and A260/A230 ratios using a NanoDrop spectrophotometer. DNA integrity and fragment size distribution were evaluated using the Agilent 4200 TapeStation and D1000 ScreenTape systems. Only samples exhibiting high

molecular weight, minimal degradation, and a sufficient DNA Integrity Number (DIN) were selected for downstream processing. Following quality control, each sample was divided into two equal-volume aliquots: one for library preparation and sequencing on the Ultima Genomics platform and the other for the Illumina platform. Subsequent library quality control confirmed the expected fragment sizes—approximately 661–689 bp for Illumina libraries and 548–569 bp for Ultima Genomics libraries—with no significant presence of adapter dimers. These quality assurance steps ensured that only high-quality DNA samples proceeded with library preparation and sequencing.

2.4 Library Preparation and Whole Genome Sequencing

Illumina Library Preparation and Sequencing: Genomic DNA (100 ng) was prepared for sequencing using the TruSeq Nano DNA Library Prep Kit (Illumina, San Diego, CA) in accordance with the manufacturer's instructions. In brief, the DNA was fragmented mechanically to a size of approximately 350 bp using a Covaris ultrasonicator. The fragmented DNA was then subjected to end repair, A-tailing and ligation with Illumina-specific indexed adapters. The adapter-ligated fragments were then size-selected using sample purification beads (Beckman Coulter). The libraries were then amplified via limited-cycle PCR (8 cycles) and purified to produce sequencing-ready libraries.

The sequencing process was conducted on an Illumina NovaSeq X Plus system utilising standard SBS chemistry. Libraries were loaded onto patterned flow cells where clonally amplified clusters were generated via bridge amplification. The sequencing-by-synthesis process was conducted through the cyclic addition of reversible terminator nucleotides, with the use of fluorescence imaging facilitating the identification of bases in paired-end mode (2 × 150 bp).

Ultima Genomics Library Preparation and Sequencing: The genomic DNA was processed using the TruSeq PCR Plus Library Prep Kit (Illumina), with adaptations made to ensure compatibility with the Ultima Genomics UG100 platform. The DNA was enzymatically fragmented, followed by end-repair and A-tailing. During adapter ligation, proprietary Ultima Genomics platform-specific adapters were utilized instead of standard Illumina adapters to ensure compatibility with mnSBS sequencing chemistry. Adapter-ligated DNA was purified and subjected to PCR amplification using primers specific to Ultima Genomics, as recommended by the manufacturer. The final libraries were bead-purified prior to the process of sequencing.

The sequencing process was conducted on the Ultima Genomics UG100 instrument, which employs a flow-based sequencing-by-synthesis approach on an open flow cell design featuring a rotating silicon wafer. The process of clonal amplification of DNA fragments, followed by binding via emulsion PCR, results in the deposition of the DNA fragments onto patterned landing pads that are distributed across the wafer. Each sequencing cycle introduces a single nucleotide type across all beads, and incorporated events are detected via optical end-point imaging without the use of reversible terminators. This approach has been demonstrated to generate single-end reads of approximately 300 base pairs with high throughput and accuracy.

2.5 Mapping and Deduplication

For Illumina sequencing data, paired-end reads were aligned to the human reference genome (GRCh37 [21] and GRCh38 [22]) using BWA-MEM (v0.7.17) [23] with read group information included to facilitate downstream sample-level analyses. The resulting SAM files were converted to BAM format and sorted using samtools (v1.9) [24]. PCR duplicates were subsequently marked using Picard MarkDuplicates (v2.3.0) with default settings, retaining all reads while flagging duplicates for later filtering. This process also generated duplication metrics for quality assessment. All steps were performed within a Snakemake workflow, ensuring consistency and reproducibility through defined resource allocation and Docker-based environments.

In the case of Ultima Genomics data, raw reads generated from the UG100 sequencer were processed using UGmapper (v1.3.2), a proprietary software tool developed by Ultima Genomics. This tool performs alignment, read sorting, and duplicate marking as part of an integrated pipeline, directly outputting aligned data in the CRAM file format. Unlike conventional formats, these CRAM files contain custom tags (tp, to) that encode base-level error probabilities associated with homopolymer lengths, enhancing downstream variant calling accuracy. No additional realignment or duplicate marking was required after UGmapper processing. (Figure 1C)

2.6 Sequencing Data Quality Control

Quality metrics were evaluated using FastQC (v0.12.1) [25] and Qualimap (v2.3) [26]. to ensure the integrity of raw and aligned sequencing data. FastQC was used to assess base quality scores, per-base sequence content, GC content distribution, and levels of duplication across reads. Qualimap was applied to the aligned BAM files to evaluate overall mapping quality, coverage distribution, and alignment statistics across the genome.

All analyses were performed using default parameters. Depth of coverage and coverage uniformity across genomic regions were visually inspected using the Integrative Genomics Viewer (IGV) [27], allowing for the manual confirmation of data consistency and the detection of any potential anomalies in sequencing depth.

2.7 Variant Calling and VAF Distribution Analysis

Single nucleotide variants (SNVs) and indels were identified using the Genome Analysis Toolkit (GATK) HaplotypeCaller (v4.0.5.1) [28]. Prior to variant calling, input BAM files were preprocessed through local realignment around indels using RealignerTargetCreator and IndelRealigner from GATK v3.5, to reduce alignment artifacts and improve variant calling accuracy. Variant calling was performed in haplotype-based mode using sorted, deduplicated, and indel-realigned BAM files as input. The GRCh37 or GRCh38 human reference genome was used depending on the sample group. Resulting variant call format (VCF) files were generated per sample. Variant allele frequencies (VAFs) were calculated from the output VCFs and visualized to infer clonal structure and to identify candidate early postzygotic mutations.

2.8 CNV Analysis & SV Analysis

Copy number variants (CNVs) were detected using CNVkit (v0.9.12) [29], based on read depth information from aligned BAM files. Default copy number call parameters were used. CNV profiles were visualized as $\log_2$ copy ratio plots were exported for comparative analysis across samples and sequencing platforms.

Structural variants (SVs) were identified from whole genome sequencing data using DELLY (v1.2.6) [30], a software tool based on paired-end and split-read mapping signals. DELLY was run separately for each sample to detect deletions, duplications, inversions, and translocations. The resulting variant call format (VCF) files were filtered to retain high-confidence SVs based on default quality metrics and genotype support.

To compare SV profiles across samples and platforms, the resulting SVs were merged using SURVIVOR (v1.0.7) [31]. A maximum allowed distance of 1,000 bp between breakpoints was used to define SV equivalence across samples, and only variants supported by at least one sample were retained. The merged SV set was then used to generate a presence/absence matrix, from which pairwise correlations between samples were calculated.

This approach enabled the identification of common and platform-specific structural variants and allowed for the assessment of concordance across sequencing platforms.

2.9 Variant Annotation

Variants identified through GATK HaplotypeCaller (v4.5.0.0) were annotated using snpEff (v5.1) [32] referencing dbSNP [33], gnomAD [34], and ClinVar [35] databases. Annotation included functional classifications (e.g., missense, nonsense, frameshift) and predicted impact levels, along with population allele frequencies and clinical significance.

2.10 Mutation Signature Analysis

Somatic single nucleotide variants (SNVs) were selected by comparing four DNA samples obtained from the same individual. Variants identified in all four samples were considered germline mutations and excluded from analysis. Variants shared by two or three samples were classified as somatic mutations, while those detected in only one sample were classified as private mutations and analyzed separately. Only variants detected in both the Illumina and Ultima Genomics platforms for the same DNA sample were included in the somatic mutation analysis. In addition, platform-specific mutations, identified exclusively by either Illumina or Ultima Genomics, were analyzed separately to assess potential platform-dependent biases.

Mutational signature analysis was performed using SigProfilerExtractor (v1.1.25) [36]. For each somatic SNV, the trinucleotide context was determined to construct a 96-channel mutation profile. *de novo* mutational signatures were extracted using non-negative matrix factorization (NMF) and compared to reference signatures from COSMIC v3.4 using cosine similarity.

2.11 Sanger Sequencing Validation

Three candidate INDEL variants were validated by Sanger sequencing. Genomic DNA was amplified by PCR using variant-specific primers (amplicon size: 273–495 bp) targeting BRCA1 and BRCA2 loci. PCR was carried out using Dr. MAX DNA Polymerase (Doctor Protein Inc., Korea) with optimized thermal cycling conditions. Amplified products were purified and sequenced bidirectionally using the BigDye Terminator v3.1 Cycle Sequencing Kit on an ABI 3730xl DNA Analyzer (Applied Biosystems).

3. RESULTS

3.1 Comparison of Sequencing Data Characteristics Between the Two Sequencing Platforms

Prior to performing mutation analysis based on Whole Genome Sequencing (WGS) data generated from two sequencing platforms, a quality assessment was first conducted on each dataset. For the Illumina platform, quality assessment began with FASTQ files generated from the sequencer, whereas for the Ultima Genomics platform, evaluation was performed on CRAM files. However, to ensure a fair comparison, files in a standardized format were generated, and quality assessment was conducted using a unified analysis pipeline.

First, to evaluate the base call quality of each platform, statistics were calculated using the samtools stats function on the aligned BAM files. As evaluation metrics, the proportion of bases with a Phred score of Q30 (error rate ≤ 0.001%) or higher, and Q20 (error rate ≤ 0.01%) or higher, was measured. (Figure 2) The analysis results showed that the four samples on the left were generated using the Illumina platform, and the four samples on the right were generated using the Ultima Genomics platform. It was confirmed that the data produced by the Illumina platform exhibited higher base quality overall. Specifically, the average proportion of bases with Q20 or higher was approximately 3.03% greater for Illumina than for Ultima Genomics, and the Q30-based quality was, on average, 7.48% higher.

Next, alignment was performed for all sequenced data against the hg38 human reference genome, and the Mapped Read Rate was calculated. This analysis also utilized samtools stats, and the results showed that the data from the Ultima Genomics platform had, on average, a 0.4% higher mapping rate. However, this difference is considered to result from sample-specific characteristics rather than technical differences between the platforms. (Figure 3A)

In addition, the ratio of deduplicated reads was analyzed to determine the proportion of actual reads remaining after PCR duplicates were removed. This allowed a comparison of library complexity and duplication rate during the sequencing process. As a result, no significant difference was observed between the two platforms in terms of the deduplicated read ratio, which is interpreted as variation due to individual sample characteristics. (Figure 3B)

In general, the results of comparing sequencing data quality between the two platforms showed that the Illumina platform was superior in base quality (Phred score), while the mapping rate and deduplicated read ratio did not show significant differences between the platforms. These results suggest that although there may be differences in base-calling accuracy, both platforms provide comparable performance in terms of overall alignment and library quality. Therefore, data from both platforms can be considered a valid basis for subsequent mutation analysis.

To evaluate potential base composition bias between sequencing platforms, GC content distribution analysis was performed (Figure 4). This analysis examined the proportion of guanine (G) and

cytosine (C) bases across reads to assess whether sequencing efficiency varies according to GC content. The Illumina platform (orange line) exhibited a unimodal distribution with a clear peak around 40% GC content, aligning with the expected average for the human genome. The distribution was relatively symmetric and stable, suggesting consistent performance across a range of GC contents. In contrast, the Ultima Genomics platform (blue line) also showed a peak at a similar GC content range but demonstrated a steeper decline in read frequency as GC content increased beyond the peak. Notably, in the high-GC regions (50–70%), Ultima Genomics exhibited a sharper drop in read coverage compared to Illumina. These findings suggest that Ultima Genomics may have reduced sequencing efficiency or coverage uniformity in GC-rich regions, potentially due to technical limitations in base incorporation or amplification under high GC content. In contrast, Illumina maintained a more stable read distribution across the GC spectrum, indicating greater tolerance to GC content variation and more uniform genome coverage.

To evaluate the sequencing performance of each platform, we analyzed the relationship between coverage depth and genome fraction covered for each sample. The average of four samples per platform was plotted on a graph (Figure 5). In whole genome sequencing, a wide range of the genome must be covered even at intermediate levels of depth, so this analysis serves as a key indicator for evaluating sequencing reliability. As shown in the plot, most samples achieved more than 90% genome coverage at depths below 10X, demonstrating excellent baseline sequencing efficiency. However, no significant differences were observed between samples across platforms as the coverage depth increased. Overall, the patterns observed in both Ultima Genomics and Illumina datasets were similar. These results suggest that variation in data quality is more attributable to sample-specific factors than to differences between the two platforms, especially in studies that require high coverage or comprehensive genome-wide analysis (Table 1).

Table 1. Data Yield and Mean Coverage per Sample

	Category	ALL_Fb1-3_G11	ALL_Fb13-4_G7	ARL_Fb12-2_H6	ARL_Fb5-4_C2
Illumina	Yields (Gb)	147.80	147.49	153.71	151.95
	Mean Coverage (X)	43.51	42.86	43.79	45.96
Ultima Genomics	Yields (Gb)	182.43	135.04	147.45	137.76
	Mean Coverage (X)	54.10	39.60	42.40	41.95

3.2 Comparison of Called Variants Between Sequencing Platforms

To perform lineage tracing and downstream analysis on single-cell clones using Whole Genome Sequencing (WGS) data, we first generated a VCF file containing mutation information from the BAM file of each sample to identify mutations compared to the reference genome (hg38). Mutation calling was performed using the HaplotypeCaller of GATK, with default parameters applied and no additional filtering process used. Before conducting a full-scale mutation analysis, we visualized the Variant Allele Frequency (VAF) distribution to identify the overall characteristics of the generated call set, and variants with a VAF value of 1 were excluded from the analysis. The Illumina data showed a typical bell-shaped VAF distribution, with some samples also revealing high-frequency mutations with a VAF close to 0.9. (Figure 6A) Conversely, data from the Ultima Genomics platform showed a concentration of mutations in the low allele frequency region (around 0.1), with a significantly larger total number of mutations (4.8 million to 5.1 million) compared to Illumina (approximately 3.2 million). (Figure 6B)

To more clearly examine this difference in distribution, all mutations were separated into SNVs and INDELs and analyzed. When SNVs were extracted from the Illumina VCF and analyzed, the resulting distribution maintained the same bell-shaped curve observed earlier. (Figure 7A) The SNVs from Ultima Genomics also showed a distribution similar to that of Illumina, with a reduced presence of low VAF variants compared to the overall mutation set. (Figure 7B) SNVs commonly identified between the two platforms and those identified specifically by each platform were visualized through a Venn diagram, with mutations detected across all eight samples considered germline mutations and excluded from the analysis. (Figure 8) As a result, Illumina-specific SNVs were relatively more numerous than those from Ultima Genomics, but the overall distribution patterns between the two platforms were similar.

On the other hand, a clear difference was observed in the analysis of INDELs. The VAF distribution of Illumina INDELs showed a significantly higher number of mutations at VAF 0.9 or above, and many mutations were also found in the 0.6–0.8 range, indicating relatively high VAF values compared to SNVs. (Figure 9A) In contrast, the distribution of Ultima Genomics INDELs was left-skewed, with a predominance of mutations having low VAF values. (Figure 9B) While the number of INDELs identified in Illumina was approximately 680,000, Ultima Genomics recorded approximately 2.4 million to a maximum of 2.7 million, nearly four times as many.

Specifically, mutations with VAF values of 0.2 or lower were considered likely false positives, prompting additional analysis. A Venn diagram was also created for the INDEL variants from both platforms, excluding germline mutations detected across all eight samples. This analysis revealed that the number of INDEL mutations uniquely detected by Ultima Genomics was approximately 20 times higher than those detected by Illumina, indicating a clear difference in INDEL detection characteristics between the platforms. (Figure 10)

In order to analyze the characteristics of the INDEL mutations identified above more precisely, additional analysis was performed by dividing INDELs into homopolymer INDELs and non-homopolymer INDELs. Homopolymer INDELs are defined as insertion or deletion mutations consisting of repeated identical bases. In this study, an INDEL was classified as homopolymer if its length was 3 bp or longer, and one base (A, T, G, or C) constituted 80% or more of the total length

within the INDEL region.

As a result of the analysis (Table 2), PolyA and PolyT were observed more frequently than PolyC and PolyG in homopolymer INDELs on both the Illumina and Ultima Genomics platforms, suggesting an asymmetric tendency of occurrence depending on base composition. In addition, the overall number of homopolymer INDELs was slightly higher on the Ultima Genomics platform than on the Illumina platform, indicating that this difference may result from platform-specific differences in base insertion/deletion processing. (Figure 11)

On the other hand, for non-homopolymer INDELs—excluding homopolymer INDELs—the number of mutations detected on Ultima Genomics was found to be approximately four times higher than that on Illumina. This suggests that beyond simple technical differences, there may be unique error patterns or artificial mutation artifacts specific to the Ultima Genomics platform.

Therefore, these results imply that when analyzing INDEL mutations using the Ultima Genomics platform, it is essential to employ a filtering method that accounts for the potential presence of false positives, particularly in non-homopolymer regions. Establishing an analysis strategy that incorporates platform-specific characteristics is crucial, and the development of appropriate validation and correction methodologies will be necessary in future studies.

Mutation signature analysis was performed to compare the biological signals and technical characteristics of mutations detected on different sequencing platforms. The analysis included common mutations identified only on both Illumina and Ultima Genomics platforms in the same cell, and platform-specific mutations detected exclusively on each platform. Mutations commonly observed in four DNAs were excluded as germline mutations, and mutations observed only on one of the platforms were analyzed separately. Mutations shared by two or three DNA samples, as well as unique mutations detected in only one DNA sample, were considered somatic mutations and subjected to analysis.

Mutation signature analysis results for commonly detected mutations showed very high similarity with cosine similarity of 0.99 and correlation of 0.989, indicating excellent agreement between the original data and the modeled signature. Considering that the sample was skin fibroblast, SBS1 and SBS5, which are clock-like signatures that accumulate over time, stood out along with SBS7a and SBS7b signatures induced by UV, and some SBS58 signatures were suggested to be due to sequencing errors. This confirmed that biological mutation signals were consistently well restored on both platforms. (Figure 12)

Analysis targeting mutations detected specifically on the Illumina platform showed relatively high similarity with cosine similarity of 0.949 and correlation of 0.889, and SBS5, SBS58, and SBS96 contributed as major signatures. In particular, signatures related to the action of AID enzymes related to immune response were also observed, suggesting biological significance. However, the reproducibility of some signatures was limited due to the relatively high L1/L2 error and KL divergence values, which may reflect the possibility of technical bias unique to the platform. (Figure 13A)

On the other hand, the signature analysis of mutations observed only on the Ultima Genomics platform showed the lowest similarity among the analyses so far, with cosine similarity of 0.76 and correlation of 0.46. In addition, the L1/L2 error was 50~66% and the KL divergence was very high at 0.237, suggesting that the reproducibility of the original signature was low. Although SBS96 and SBS5 were identified as major signatures, SBS96 had limitations in interpretation as its biological

function has not been clearly known to date. Overall, the signature pattern was not clear and was dispersed, raising the possibility of platform-specific error characteristics or bias in the analysis process. (Figure 13B)

In summary, while the signature based on common mutations stably restored biological signals, different signature patterns appeared when targeting only platform-specific mutations, and this tendency was observed more clearly on the Ultima Genomics platform. This suggests that platform-specific technical characteristics have an impact on mutation signature analysis, and demonstrates the need to consider platform-specific biases in future sequencing-based mutation interpretation.

Table 2. Comparison of Homopolymer and Non-Homopolymer INDEL Counts

Variant Counts		**polyA**	**polyT**	**polyC**	**polyG**	**Homo polymer**	**Non-homo polymer**
ALL_Fb1-3_G11_I	**Illumina**	14,930	13,523	375	335	29,163	607,317
ALL_Fb13-4_G7_I		14,860	13,417	381	341	28,999	603,660
ARL_Fb12-2_H6_I		14,856	13,496	368	354	29,074	605,704
ARL_Fb5-4_C2_I		15,160	13,840	393	336	29,729	615,257
ALL_Fb1-3_G11_U	**Ultima Genomics**	19,125	15,960	272	269	35,626	2,654,955
ALL_Fb13-4_G7_U		15,913	13,485	222	240	29,860	2,418,196
ARL_Fb12-2_H6_U		15,978	13,432	221	214	29,845	2,347,992
ARL_Fb5-4_C2_U		16,230	13,709	224	231	30,394	2,378,122

3.3 Identification of Platform-Specific Variants Through Variant Annotation

Variant annotation was performed for further analysis of the variant call set. The variant annotation process involves evaluating the pathological relevance of each variant, its functional impact at the protein level, and other factors by utilizing various genome and disease-related databases based on sequencing data. Since the medical records of the cadaver donor used in this study confirmed a history of breast cancer, annotation was carried out to identify mutations strongly associated with breast cancer. However, as the samples were derived from normal cells rather than tumor tissue, the analysis focused on germline mutations instead of somatic mutations. After integrating the VCF files generated from all eight samples, mutations in breast cancer-related gene regions, including BRCA1 and BRCA2, were filtered. During this process, mutations classified as 'IMPACT = HIGH'—indicating a significant effect on protein structure or function—were prioritized. Additionally, the ClinVar database was used to annotate the disease types and pathological implications associated with each mutation. (Figure 14)

As a result, mutations at three genetic loci were classified as 'Pathogenic,' and all were confirmed to be related to breast-ovarian cancer. These mutations were marked as 'ORIGIN = 1' in the ClinVar database, indicating that they are germline-derived. All three mutations were INDELs caused by the deletion of a single nucleotide, and were therefore interpreted as having a high potential to alter protein structure. (Table 3)

Notably, the first and third mutations were not detected in the Illumina platform data but were observed exclusively in the Ultima Genomics platform data. In contrast, the second mutation was identified by both platforms and may serve as an example of cross-platform detection consistency. Finally, to verify whether the variants identified through variant calling represented true mutations, visual inspection of aligned reads was performed using the BAM files for each sample. (Figure 15)

Additionally, to verify whether the variants identified at the three loci were true variants, validation was performed using Sanger sequencing as an orthogonal method. The Sanger sequencing results revealed that the INDELs detected exclusively in the Ultima Genomics platform at the first and third loci were not present in any of the four DNA samples, indicating that they were false positives. In contrast, the variant at the second locus was consistently observed in all four DNA samples, confirming it as a true variant. (Figure 16)

These findings suggest that INDELs frequently observed specifically in the Ultima Genomics platform may include a high rate of false positives. Therefore, a platform-specific filtration strategy for INDEL variants is necessary.

Table 3. Annotation of Selected BRCA1/BRCA2 Mutations

CHROM	POS	REF	ALT	GENE	IMPACT	ORIGIN	CLNDN	CLNSIG
13	32,363,371	TG	T	BRCA2	HIGH	1	Breast-ovarian_cancer,_familial,_susceptibility_to,_2|Hereditary_breast_ovarian_cancer_syndrome	Pathogenic
17	43,045,763	TC	T	BRCA1	HIGH	1	Hereditary_breast_ovarian_cancer_syndrome	Pathogenic
17	43,091,999	TA	T	BRCA1	HIGH	1	not_provided|Breast-ovarian_cancer,_familial,_susceptibility_to,_1	Pathogenic

3.4 Comparison of CNV Profiles Between Sequencing Platforms

Next, we performed an analysis of CNV (Copy Number Variation) for structural mutation analysis at the whole genome level. The CNVkit tool was used for the analysis, which is a tool that can visualize copy number variability in the entire genome based on the Read Depth of the sequencing data. The graph presented in the figure was generated using CNVkit and shows the variation in Read Depth observed in the sequencing data of each platform.

Copy number variants (CNVs) were analyzed using CNVkit, and both Illumina and Ultima Genomics platforms produced consistent and stable copy number profiles. While minor fluctuations in read depth were observed in some regions, these did not lead to significant differences in the segmented CNV calls between platforms. No platform-specific artifacts or signal distortions were detected, indicating that CNVkit performs robustly across different sequencing technologies when using the same analysis conditions. (Figure 17)

3.5 Comparison of SV Profiles Between Sequencing Platforms

Finally, we performed an analysis on structural variants (SVs). SVs were derived from data produced by each sequencing platform (=Illumina and Ultima Genomics), and the analysis included various types such as inversion, translocation, and large insertion/deletion using an SV detection tool. However, due to the nature of SVs, the precise definition of breakpoints or variant lengths can vary across sequencing platforms and analysis tools, making direct comparison and relevance evaluation, unlike single nucleotide variants (SNVs), challenging.

Accordingly, in this study, we utilized the SURVIVOR tool to comprehensively analyze the SV detection results and compare the similarity between samples. SURVIVOR provides a pairwise comparison function that can quantitatively calculate the SV similarity between two samples after merging SVs detected in different samples based on criteria. Through this, SV-based correlation analysis was performed on a total of 8 sequencing data. (Figure 18)

The analysis results showed that the SV similarity between data generated on the same platform tended to be higher than that between data generated on different platforms derived from the same sample. In particular, the correlation between data produced on the Ultima Genomics platform was observed to be slightly higher than that between Illumina platforms. This suggests that the analysis

results of Ultima Genomics may show higher reproducibility and consistency within the platform.

These results show that there is a batch effect due to the technical characteristics of each platform and suggest caution in interpretation that may occur when directly comparing platforms. In future analyses, it is judged that a follow-up strategy to correct such platform-specific bias is necessary.

3.6 Branch confirmation for 8 data based on previous Lineage Tracing studies

Based on the Early Embryonic Mutations (EEM) information confirmed in previous analyses, we inferred the lineage (branch) of the newly sequenced data by comparing it with the lineage tree presented in a previous study [10]. This was done by defining the lineage position of each sample by confirming whether the corresponding mutations exist in the newly generated sample, if the branch-defining mutations obtained in the previous study can be used as a reliable reference point.

As a result of the analysis, the newly sequenced data showed a pattern of branching in a 3:1 ratio overall, which is consistent with the major occurrence branching pattern confirmed in the previous study. In particular, the results were clearly distinguished through comparison with the data produced by the existing Illumina platform, and it was possible to precisely classify which branch each sample belongs to base on whether it had an EEM.

This lineage structure was visualized through (Figure 19), and the lineage of the newly analyzed sample was indicated by a red line to clearly distinguish its position in the existing lineage tree. Each sample was placed in the appropriate branch based on the presence or absence of EEM, which allowed for visual verification of lineage consistency between data from the two sequencing platforms.

The phylogenetic analysis confirmed that the new sequence data can be logically connected to the existing phylogenetic branch structure, supporting that mutation-based lineage tracing is a reproducible analysis strategy regardless of the platform. Furthermore, this suggests that single-cell-derived genome information provides sufficient reliability and resolution to be utilized for developmental phylogenetic analysis, and it is expected to function as a core base data for more expanded analyses in the future.

4. DISCUSSION

This study is significant in that it performed whole-genome sequencing on single-cell clones derived from normal tissues and compared the characteristics and mutation detection performance of the Illumina and Ultima Genomics platforms. However, several limitations should be considered when interpreting the results.

First, the number of single-cell clone samples analyzed was limited to four, which made it difficult to detect private mutations unique to individual cells and ultimately reduced the resolution of the subclonal structure. To achieve high-resolution lineage tracing in single-cell genome analysis, a larger number of clones should be secured, and expanding the sample size in future studies would allow for more precise identification of genetic differences between cells.

Second, as the study was based solely on normal cells, it was not possible to identify mutations specific to tumor cells. Since tumor-derived mutations provide critical insights into clonal evolution and tumor heterogeneity, applying the same analysis to tumor-derived single-cell clones in future studies is expected to enable richer biological interpretation.

Third, for certain additional analyses such as structural variant (SV) detection and short tandem repeat (STR) profiling, the analysis tools used were not fully compatible with the Ultima Genomics platform. Ultima Genomics produces output data in CRAM format by default, which differs from the standard FASTQ-based pipelines, necessitating additional preprocessing and parameter adjustments. This introduced complexity to the analysis workflow and limited the scope of downstream analyses.

Finally, a variant filtering model specifically optimized for the Ultima Genomics platform was not applied in this study. The platform utilizes a flow-based sequencing method that detects base incorporation by measuring signal intensity as specific nucleotides are flowed over the DNA template. However, in homopolymer regions—where the same base is repeated (e.g., AAAAA or TTTT)—this method struggles to accurately quantify the number of incorporated bases due to the non-linear and saturable nature of signal intensities.

As a result, insertion/deletion (INDEL) errors are more frequent, increasing the likelihood of false positives. Recently, a machine learning–based filtering model has been developed to address this issue by learning the platform-specific error patterns and distinguishing true variants from sequencing artifacts. Although this model was not implemented in the current study, applying it in future work is expected to significantly improve INDEL detection accuracy and enhance the overall reliability of analyses performed using the Ultima Genomics platform.

5. CONCLUSION

This study conducted a comparative analysis with the current standard technology, the Illumina platform, to evaluate the applicability of the new sequencing platform, Ultima Genomics, to somatic mutation-based lineage tracing research. In terms of overall sequencing quality indicators, Ultima Genomics showed slightly lower or similar performance compared to Illumina, and SNV mutation concordance exhibited similar patterns between the two platforms.

On the other hand, the correlation between platforms was low for INDEL mutations, and it was identified that there was a platform-specific bias due to the unique technical characteristics of Ultima Genomics. In particular, there were cases where small deletion INDELs of 1 bp that were not identified in Illumina data were detected in Ultima Genomics data, and it is judged that confirmation of data including IGV for the corresponding mutations and validation through other cross-methods are necessary.

In CNV analysis, No clear differences were observed between the two platform. However, further evaluation using tumor-specific samples with copy number amplifications or deletions is needed to more clearly assess potential differences between the platforms. However, in SV analysis, the concordance between platforms was low, which is interpreted as a result of differences in analysis tools and technology.

Finally, Ultima Genomics is judged to provide sufficient resolution for capturing branch structures in single-cell-based lineage tracing. However, for precise interpretation of detailed lineage structures, a correction strategy for platform-specific variation characteristics, along with a larger number of samples, should be implemented.

FIGURES

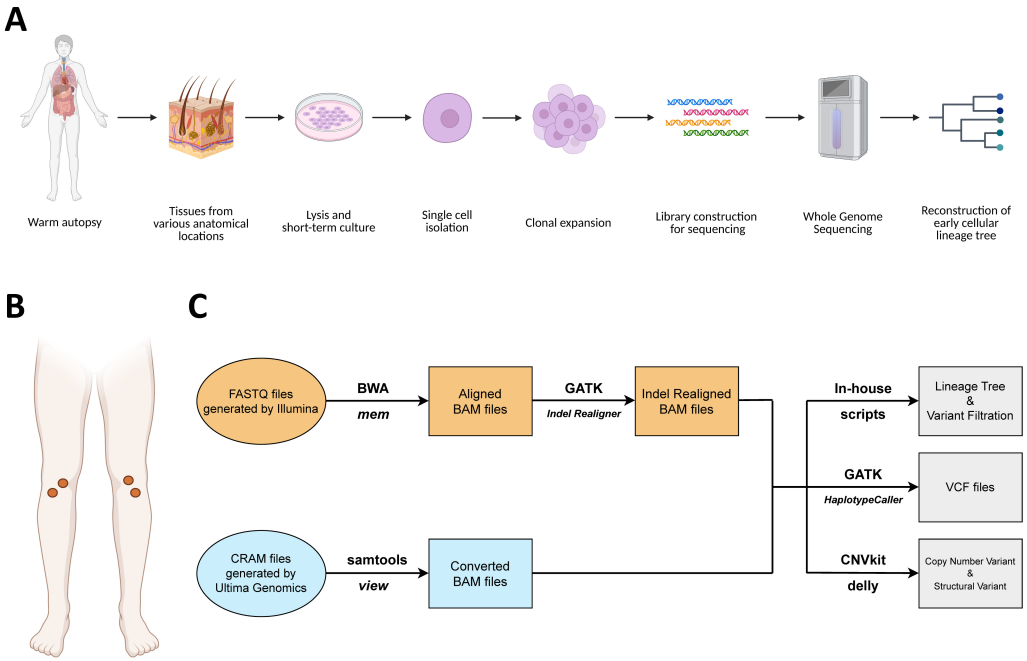

Figure 1. Overall workflow and underlying principle of the study

(A) Experimental design for lineage tracing. Skin tissues were obtained from a postmortem donor via warm autopsy and subjected to primary culture, single-cell isolation, clonal expansion, and whole-genome sequencing. (B) Sampling location: anterior regions of both lower legs. (C) Comparison of sequencing workflows between Illumina and Ultima Genomics platforms. While the initial data formats and processing steps differ, all downstream analyses were standardized using a unified Snakemake pipeline.

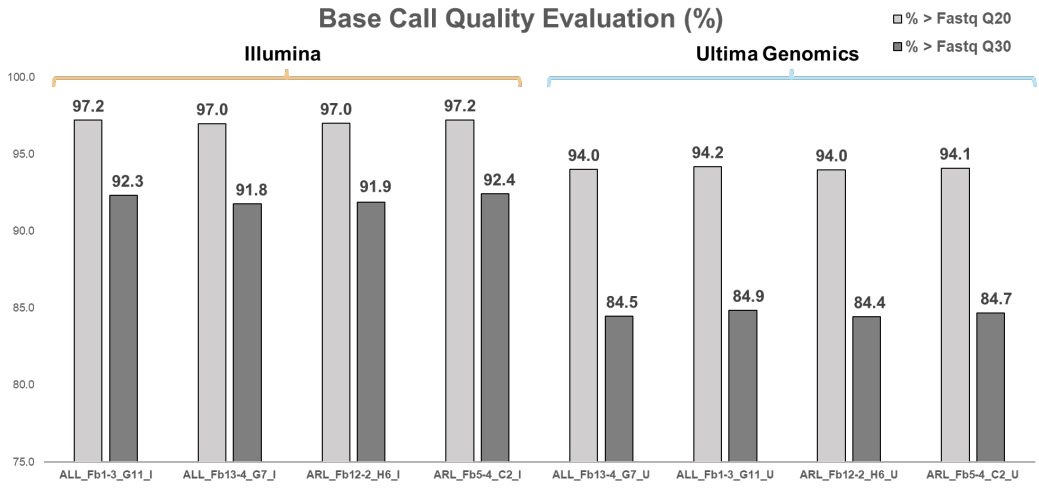

Figure 2. Comparison of Base Call Quality Between Sequencing Platforms

Prior to variant analysis, sequencing data quality was evaluated, including base call quality, mapping rate, deduplication rate, GC content, and coverage distribution. Illumina showed higher base call accuracy than Ultima Genomics, with 3.03% higher Q20 and 7.48% higher Q30 metrics on average across samples.

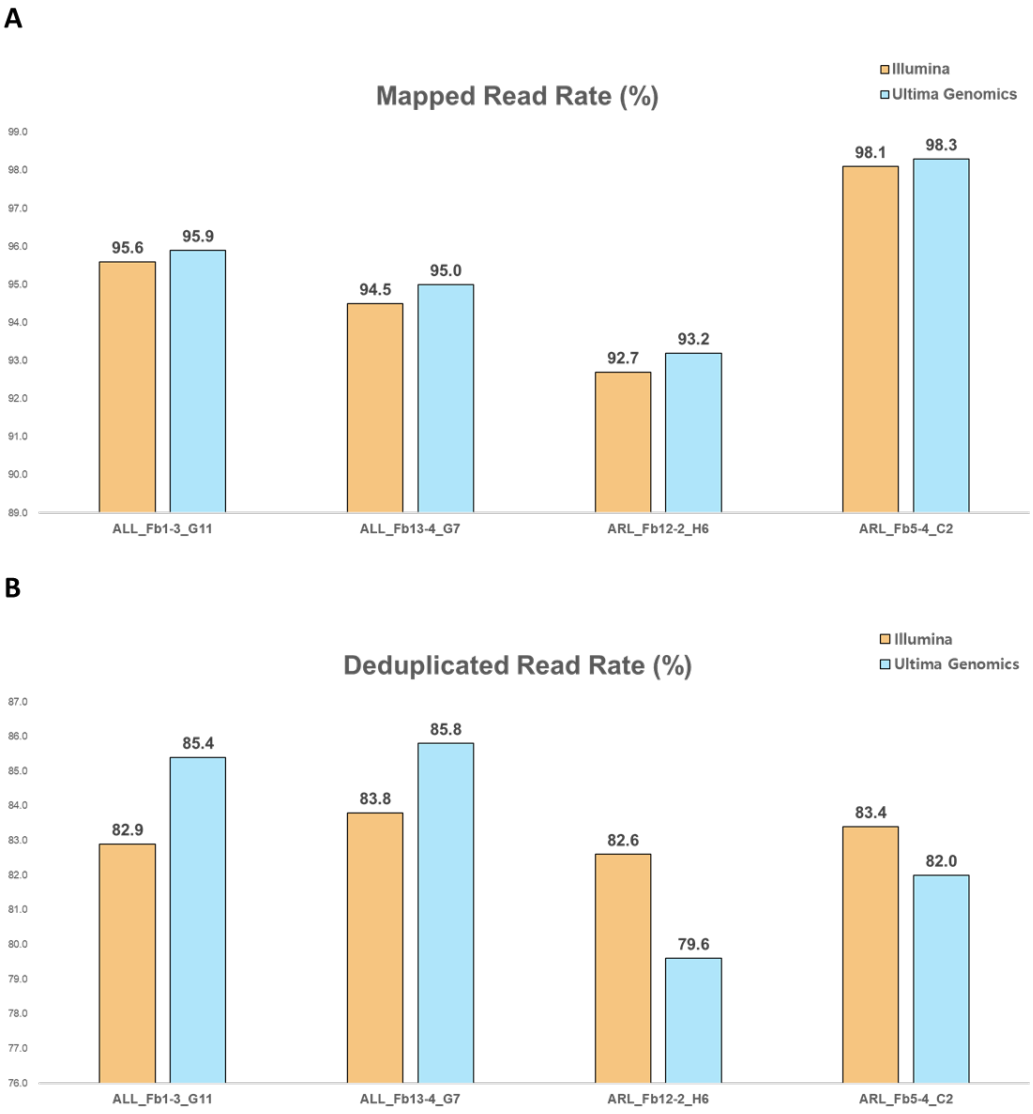

Figure 3. Comparison of Alignment Statistics Between Sequencing Platforms

(A) The Mapped Read Rate represents the proportion of reads successfully aligned to the reference genome. Ultima Genomics samples showed a slightly higher average rate (~0.4%) compared to Illumina, although this difference is likely within the margin of error. (B) The Deduplicated Read Rate, indicating the percentage of reads remaining after PCR duplicate removal, showed no notable difference between platforms, suggesting that variation is more sample-dependent than platform-dependent.

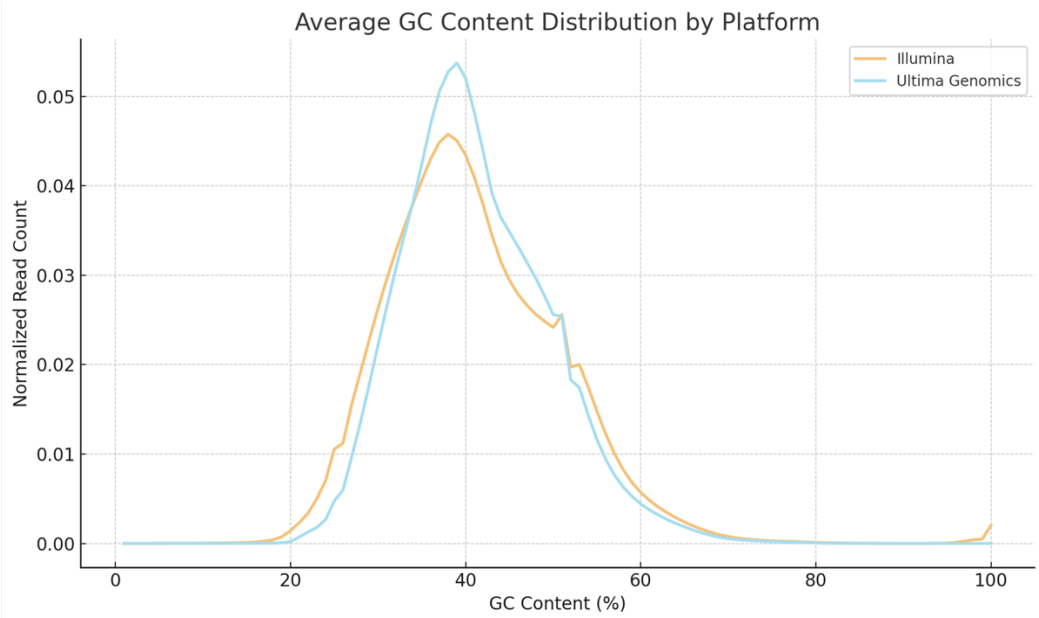

Figure 4. GC Content Distribution of Sequencing Read Count

The fourth quality metric evaluated was the average GC content distribution. Ultima Genomics showed a sharp peak near 40% GC content, indicating a narrow distribution of reads. This suggests reduced sequencing efficiency in high or low GC regions, which may affect coverage uniformity and downstream variant detection sensitivity.

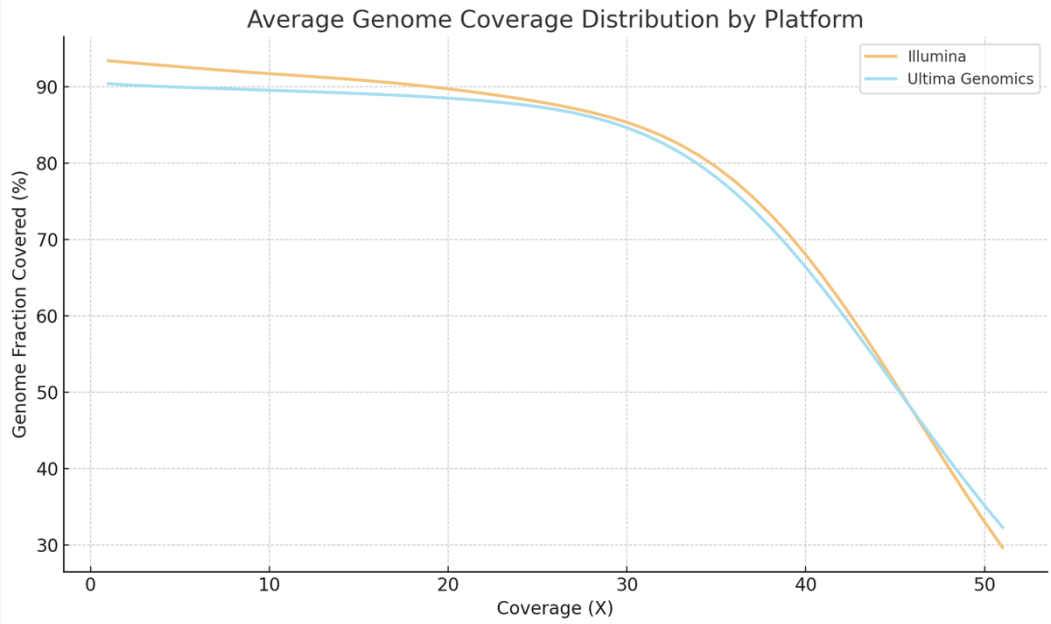

Figure 5. Genome-wide fraction coverage at varying sequencing depths

This figure illustrates the average genome coverage fraction across varying sequencing depths for eight datasets. The x-axis represents sequencing depth, and the y-axis indicates the percentage of the genome covered at or above each depth. Both Illumina and Ultima Genomics platforms showed similar patterns of coverage decay as depth increased, with no platform-specific differences observed. These results suggest that genome coverage is more influenced by sequencing yield than by platform characteristics.

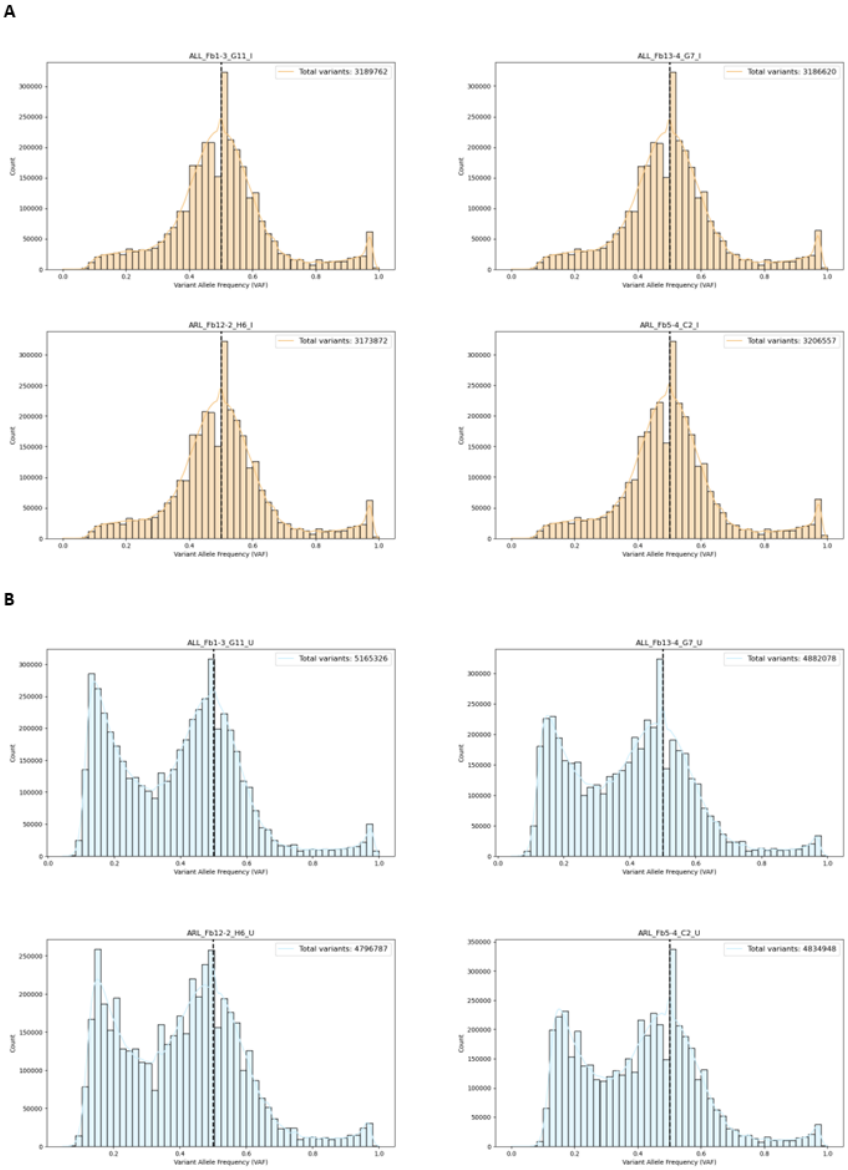

Figure 6. VAF Distribution of Total Called Variants Across Sequencing Platforms

(A) Illumina showed a typical bell-shaped VAF distribution with a peak near 0.5 and some high-frequency variants. (B) Ultima Genomics exhibited a broader spread with more low-frequency variants and a greater total variant count.

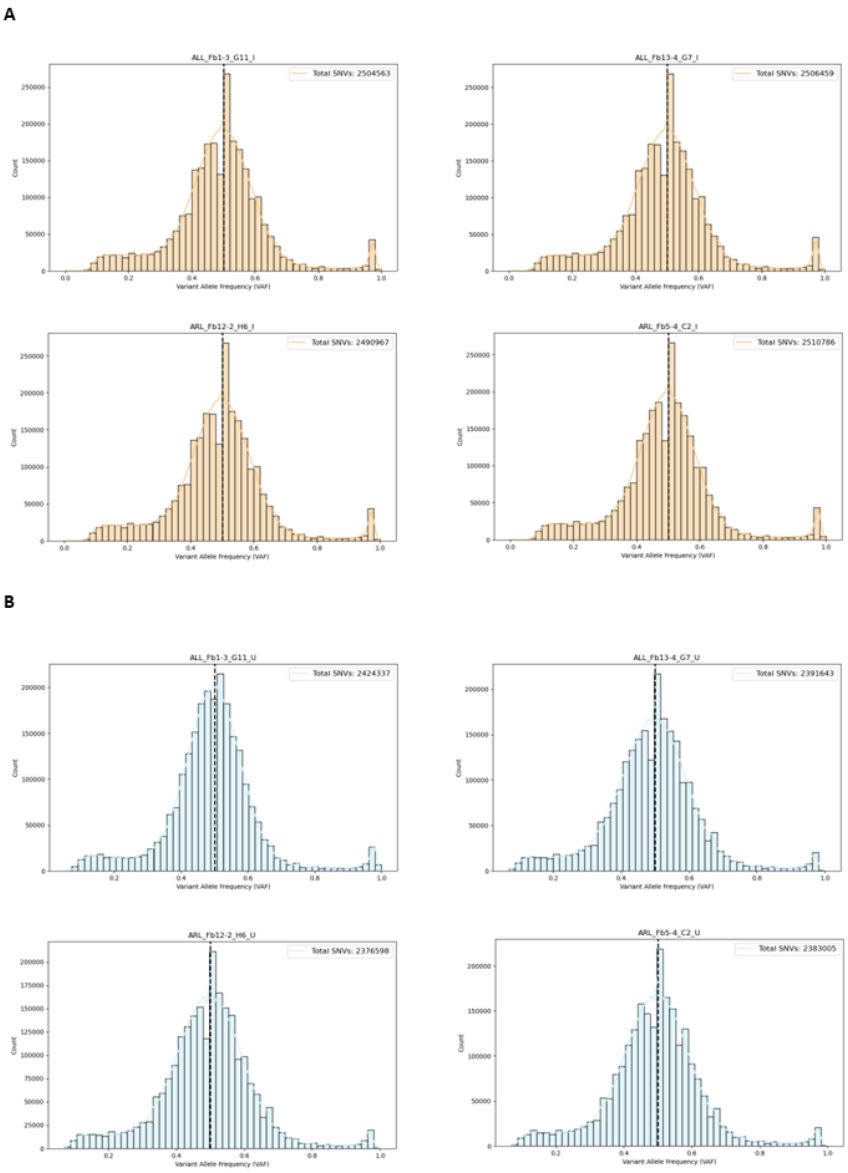

Figure 7. VAF Distribution of SNV Across Sequencing Platforms

(A) The VAF distribution of SNVs from Illumina data shows a typical bell-shaped curve with a stable variant count. (B) The Ultima Genomics platform also exhibited a similar distribution pattern and comparable total SNV counts to Illumina.

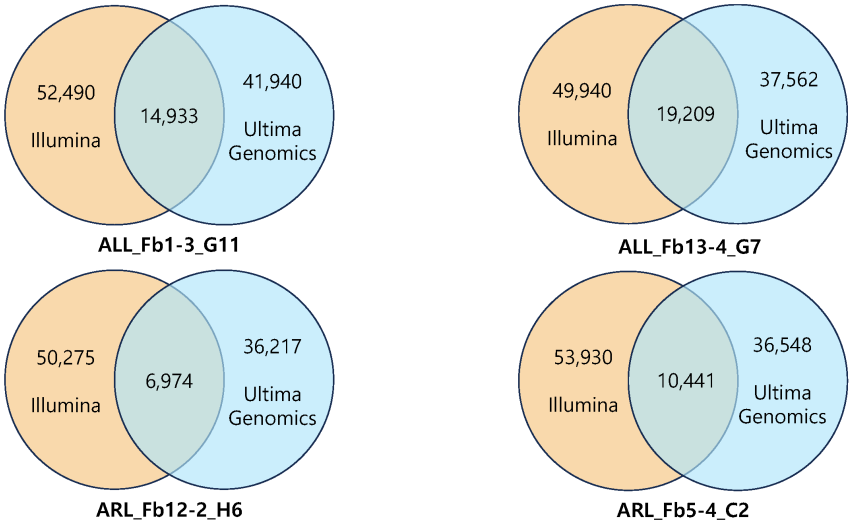

Figure 8. Venn Diagrams of SNVs Detected by Each Sequencing Platform Across Four Samples

Venn diagrams visualize the overlap and platform-specific detection of SNVs between Illumina and Ultima Genomics. Variants commonly identified in all eight samples (germline) were excluded from the comparison. While Illumina-specific SNVs appeared slightly more numerous, the overall variant patterns were similar across platforms.

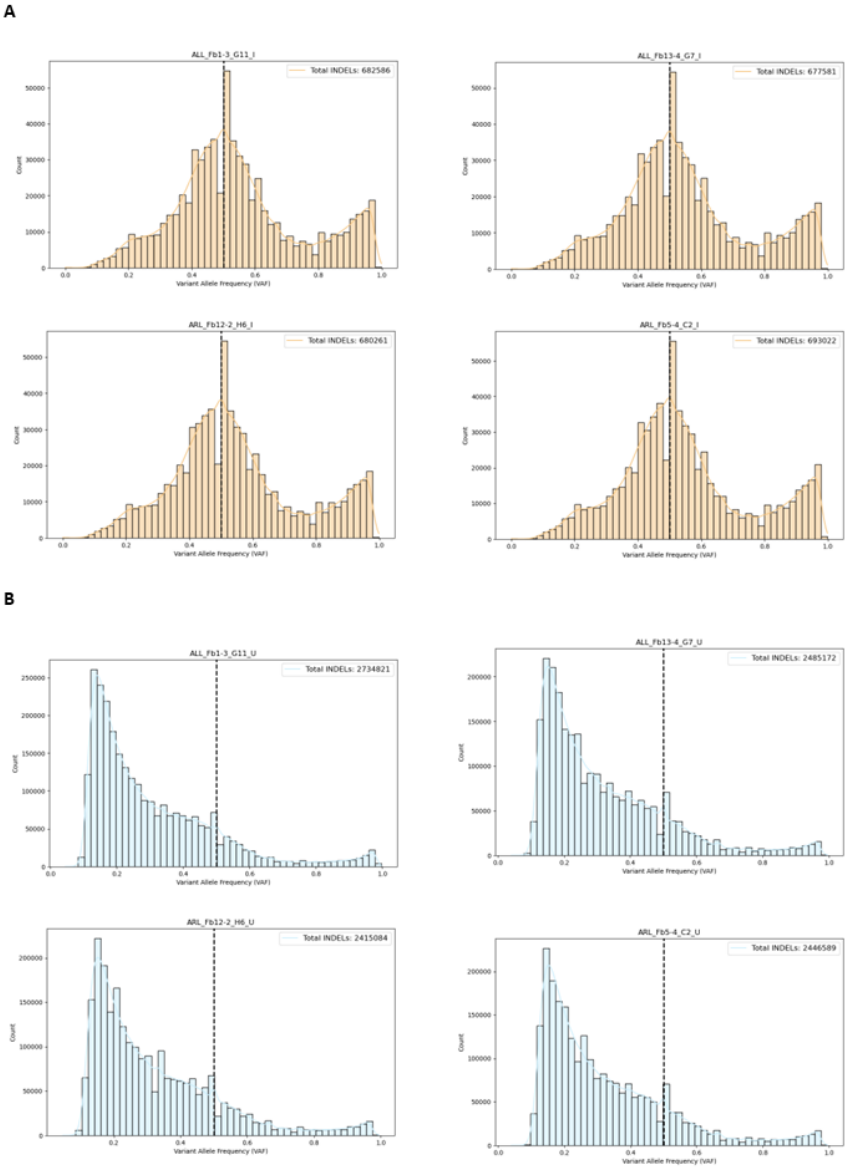

Figure 9. VAF Distribution of INDELs Across Sequencing Platforms

(A) Illumina data show a typical INDEL VAF distribution with peaks around 0.6–0.9. (B) Ultima Genomics data show left-skewed distribution, indicating a high number of low-VAF INDELs, with total INDEL counts 3–4 times higher than Illumina.

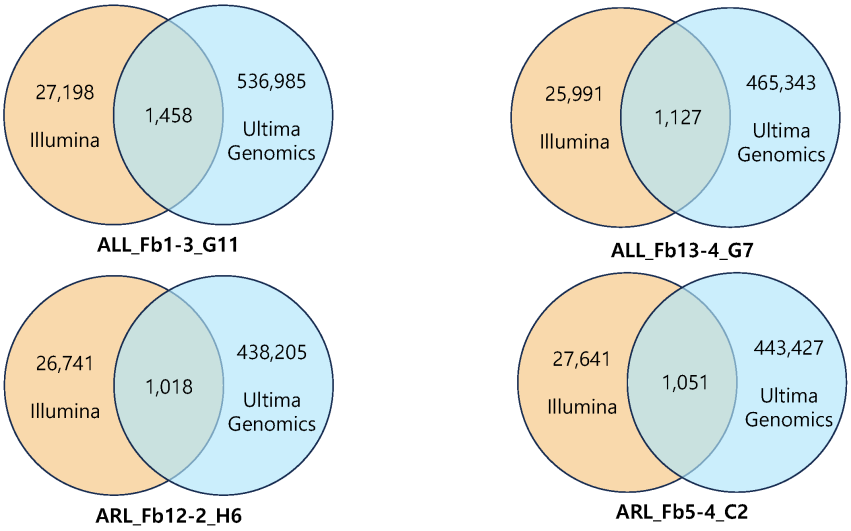

Figure 10. Venn Diagrams of INDELs Detected by Each Sequencing Platform Across Four Samples

Venn diagrams show platform-specific and shared INDELs between Illumina and Ultima Genomics for each sample. Germline variants identified across all eight datasets were excluded. The number of INDELs uniquely detected by Ultima Genomics was approximately 20 times higher than that of Illumina.

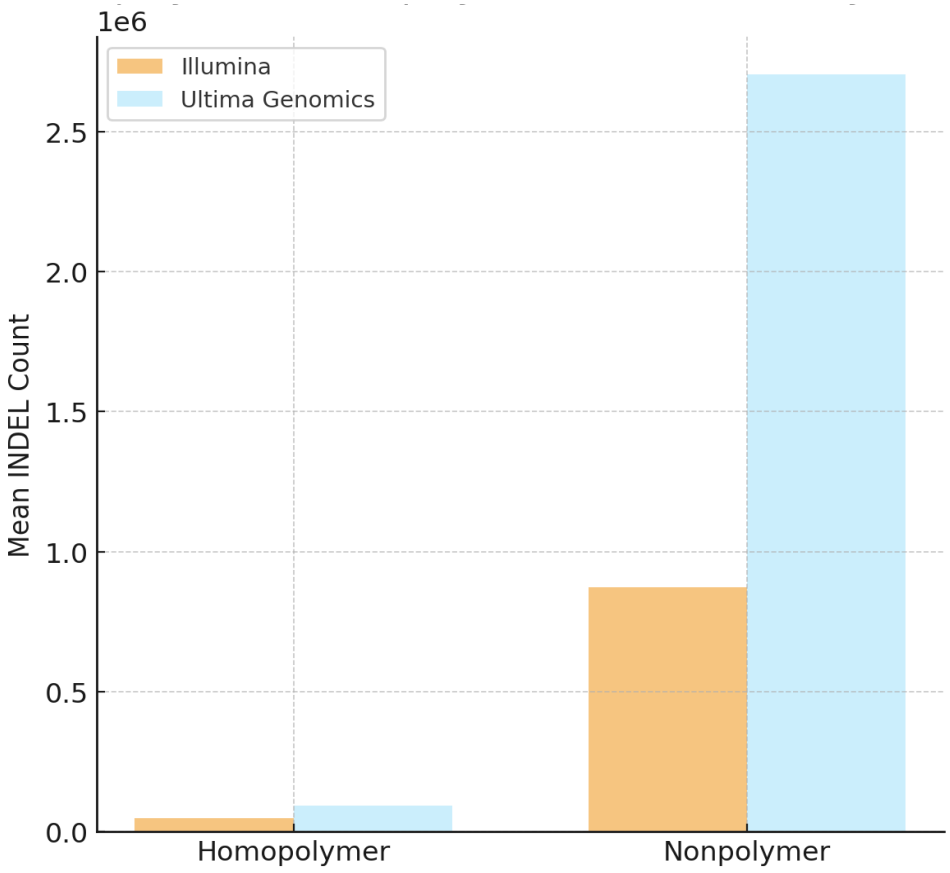

Figure 11. Comparison of Homopolymer and Non-Homopolymer INDEL Counts Across Sequencing Platforms

INDEL variants were classified into homopolymer and non-homopolymer types based on base repetition criteria (≥3 bp and ≥80% single-base content). (A) Ultima Genomics showed a slightly higher count of homopolymer INDELs than Illumina. (B) For non-homopolymer INDELs, Ultima Genomics showed ~4-fold higher counts, indicating the need for platform-specific filtering strategies.

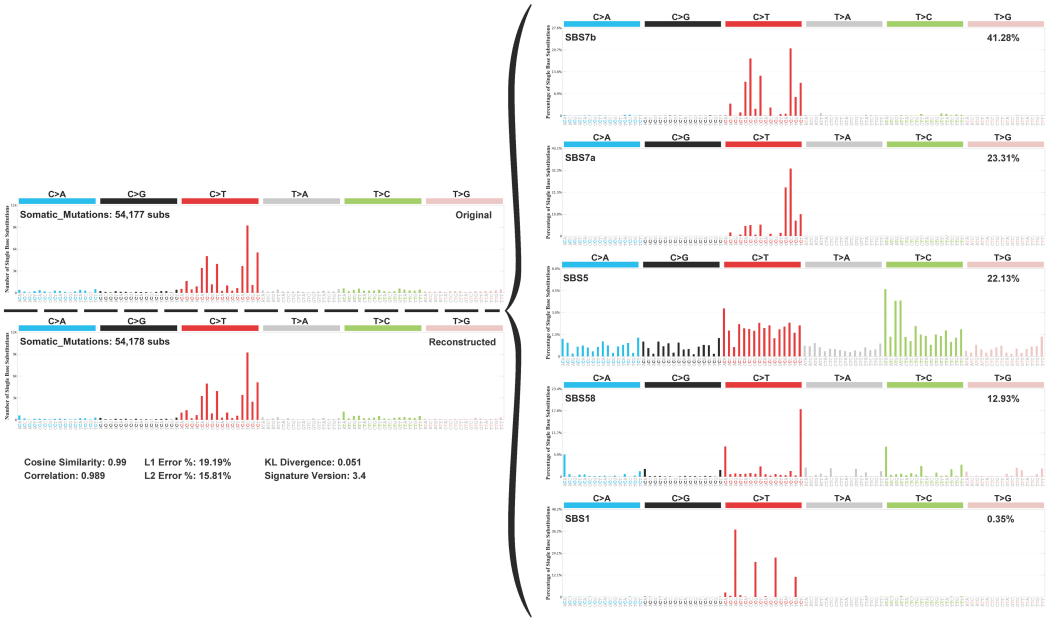

Figure 12. Shared Somatic Mutation Signatures

Mutation signatures based on somatic variants were detected on both Illumina and Ultima Genomics platforms. High concordance (cosine similarity = 0.99) supports accurate reconstruction of UV-induced (SBS7a, SBS7b) and clock-like (SBS1, SBS5) signatures.

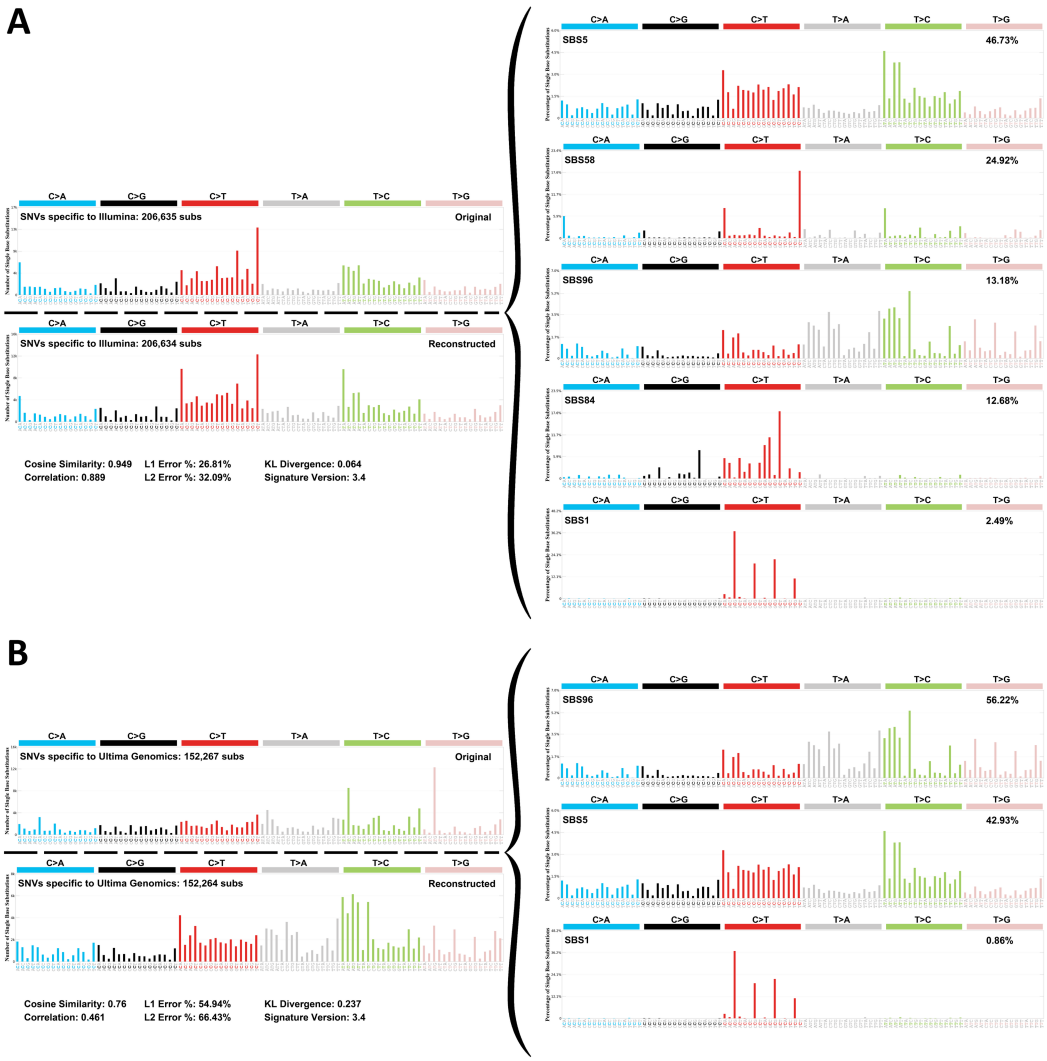

Figure 13. Platform-specific mutation signatures

(A) Mutation signature analysis of Illumina-specific SNVs showed relatively high reproducibility, with strong contributions from SBS5, SBS58, and immune-related SBS96. (B) In contrast, Ultima Genomics-specific SNVs yielded low similarity and dispersed patterns, with high L1/L2 error and KL divergence, indicating potential platform-specific technical artifacts.

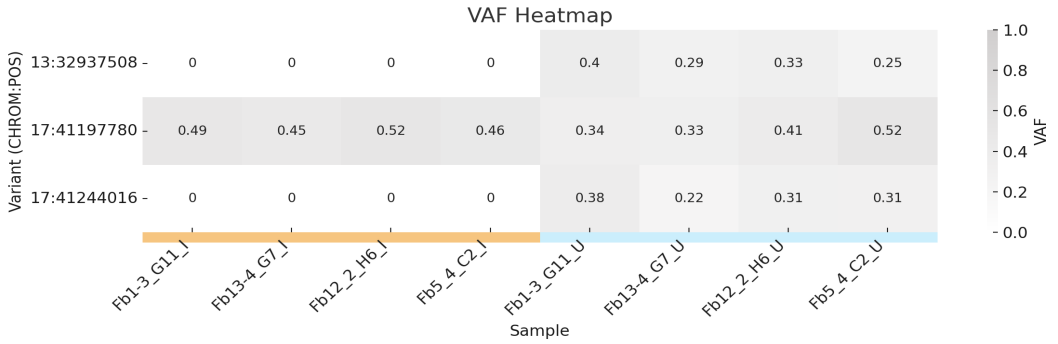

Figure 14. VAF Heatmap of Selected BRCA1/2 Variants Across Samples

Visual representation of allele frequencies for three BRCA1/2 deletion variants across all samples. Variants only observed in Ultima Genomics or in both platforms are shown, highlighting cross-platform detection differences.

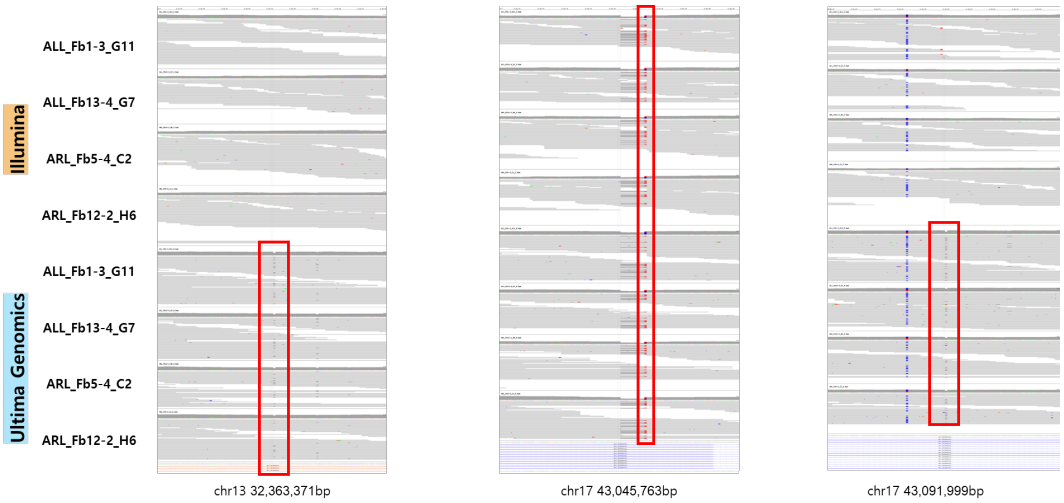

Figure 15. IGV Validation of Selected BRCA1/2 Mutations Across Sequencing Platforms

IGV visualizations confirm the presence of three BRCA-related mutations. The first and third variants are observed only in Ultima Genomics data, while the second is consistently detected in both Illumina and Ultima Genomics datasets.

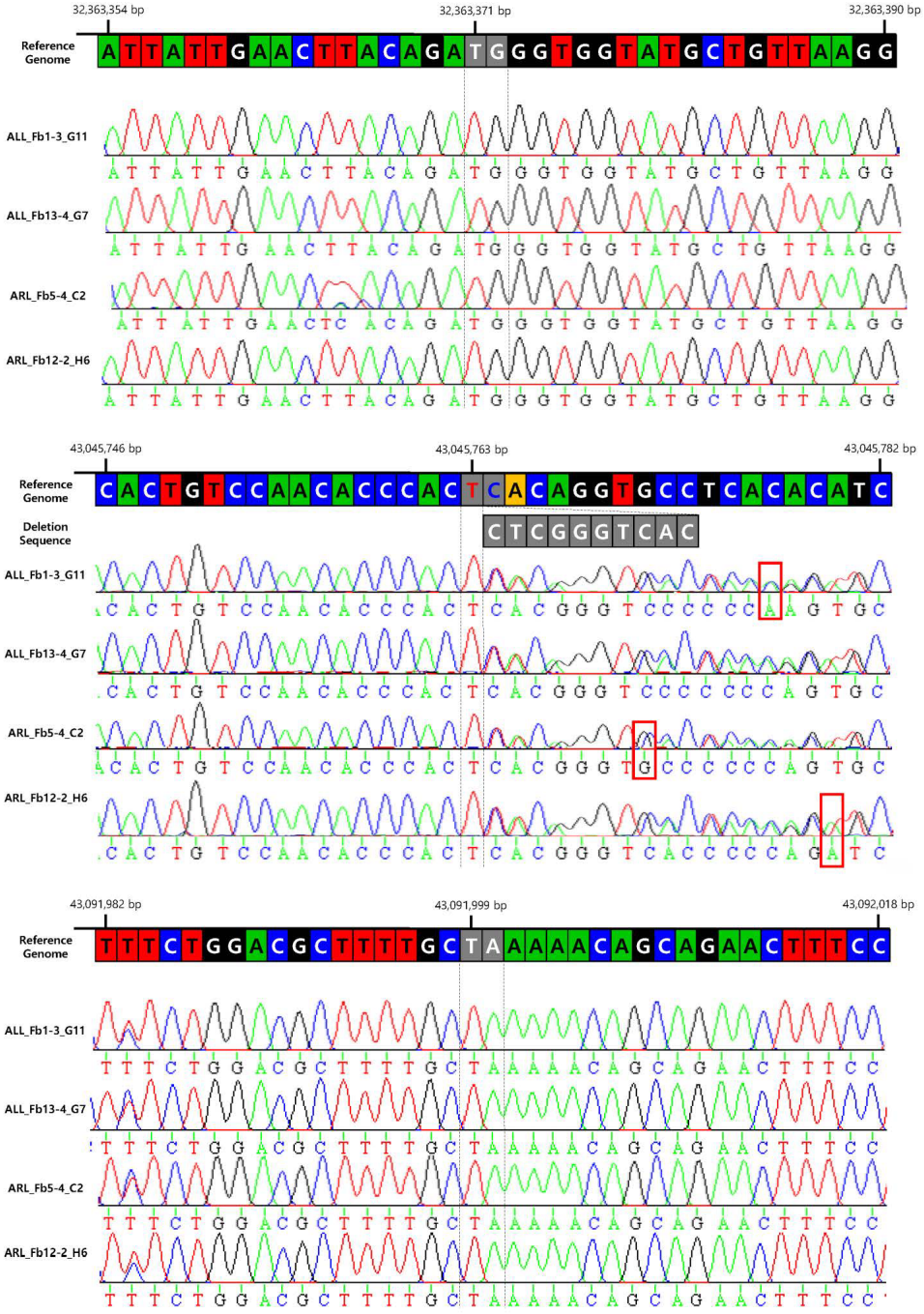

Figure 16. Sanger Sequencing Results for the Three Variant Loci

To validate the three filtered loci (chr13:32,363,371; chr17:43,045,763; chr17:43,091,999) identified through annotation, Sanger sequencing was performed as an orthogonal method. The first and third loci, where INDEL variants were detected only in the Ultima Genomics platform, were confirmed to be false positives, as no variants were observed in any of the four DNA samples. In contrast, the second locus showed a deletion consistently detected by both sequencing platforms, and this result was also confirmed by Sanger sequencing. Additionally, the reverse strand was also examined to confirm these results.

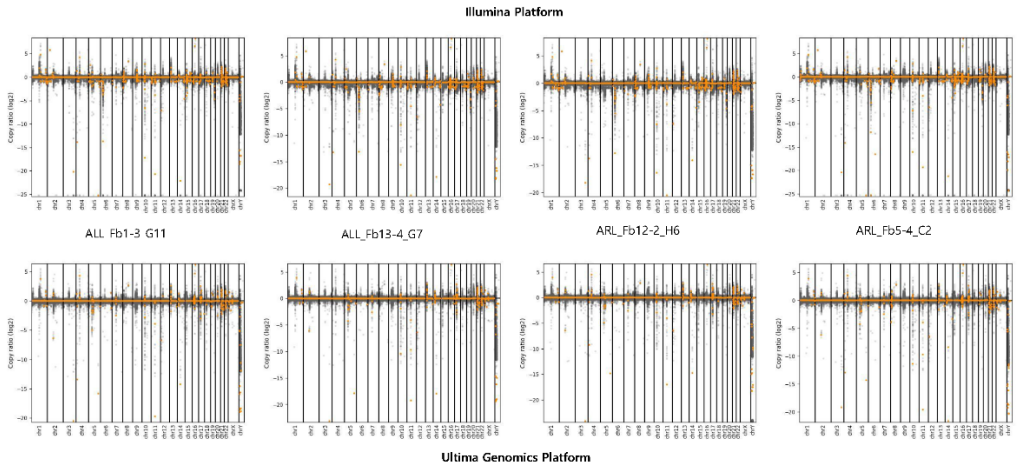

Figure 17. Genome-wide CNV Profiles Across Platforms

Copy number variation (CNV) was analyzed using CNVkit. No distinct differences were observed between the two platforms, although regions with reduced read depth due to sequencing artifacts were identified in the Illumina data.

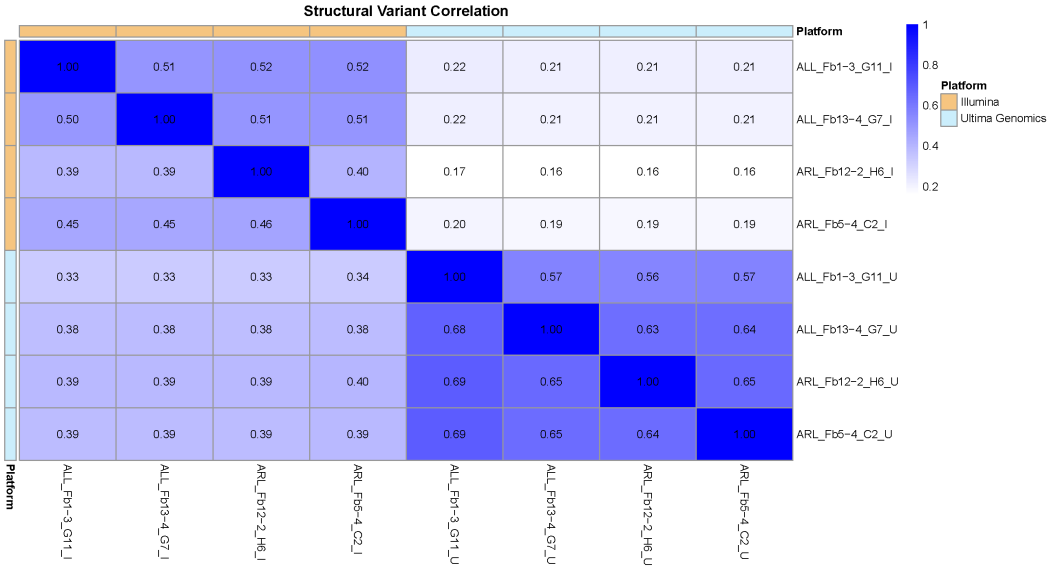

Figure 18. Pairwise Correlation of Structural Variants Across Sequencing Platforms

SVs from each sample were merged and compared using SURVIVOR to assess pairwise correlation. Higher correlations were observed within platforms, with Ultima Genomics showing slightly stronger consistency.

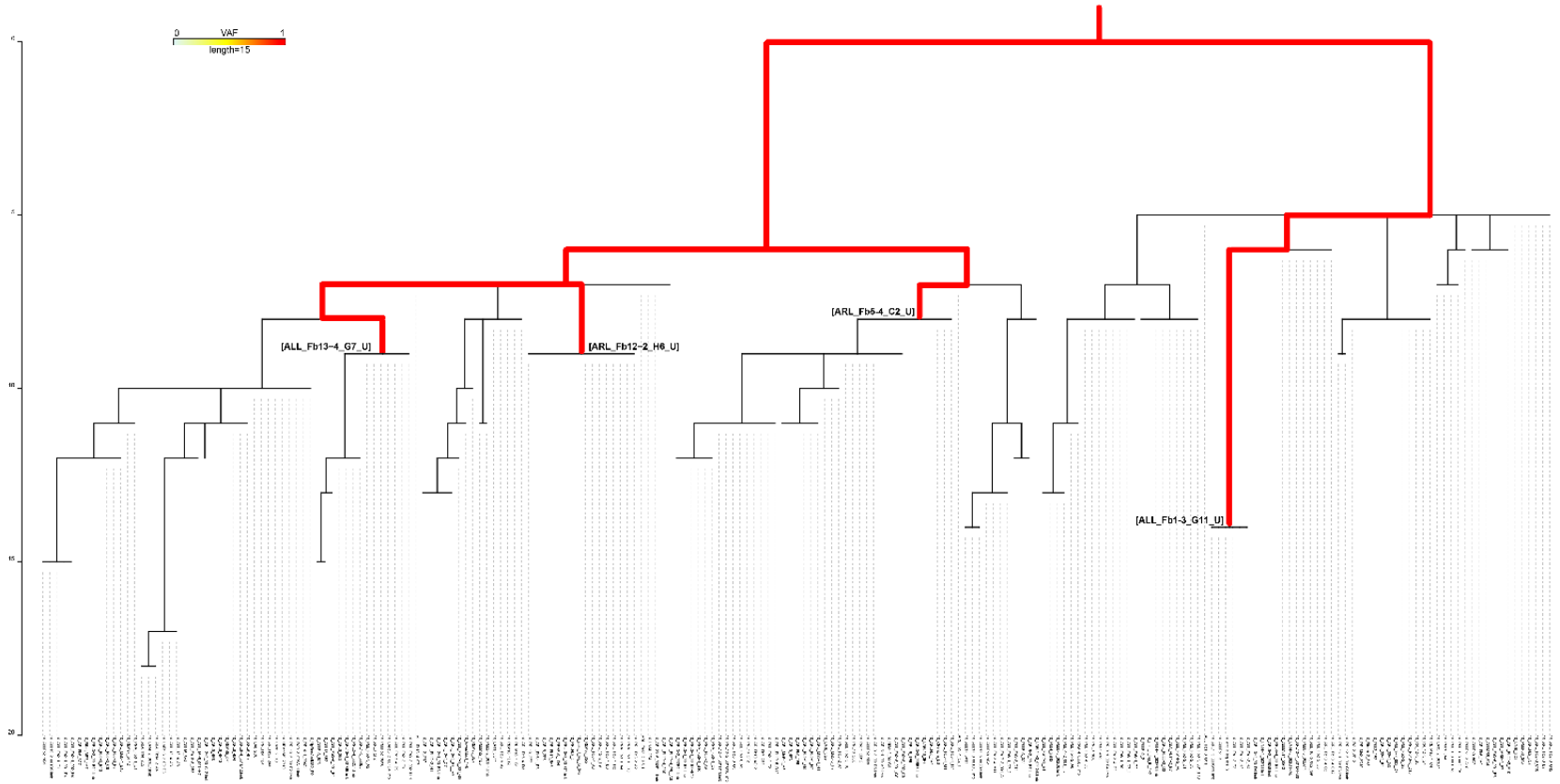

Figure 19. Lineage mapping of the four sequenced samples onto the previously established lineage tree

Samples sequenced on the Ultima Genomics platform were aligned to hg19 and mapped onto a pre-established lineage tree. This confirmed that branch assignment is feasible using variants detected by Ultima Genomics data.

REFERENCES

1. Bizzotto, S. and C.A. Walsh, *Genetic mosaicism in the human brain: from lineage tracing to neuropsychiatric disorders.* Nature Reviews Neuroscience, 2022. **23**(5): p. 275-286.
2. Evrony, Gilad D., et al., *Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain.* Cell, 2012. **151**(3): p. 483-496.
3. Lodato, M.A., et al., *Somatic mutation in single human neurons tracks developmental and transcriptional history.* Science, 2015. **350**(6256): p. 94-98.
4. Choi, S.H., et al., *Grave-to-cradle: human embryonic lineage tracing from the postmortem body.* Experimental & Molecular Medicine, 2023. **55**(1): p. 13-21.
5. Chapman, M.S., et al., *Lineage tracing of human development through somatic mutations.* Nature, 2021. **595**(7865): p. 85-90.
6. Bizzotto, S., et al., *Landmarks of human embryonic development inscribed in somatic mutations.* Science, 2021. **371**(6535): p. 1249-1253.
7. Bae, T., et al., *Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis.* Science, 2018. **359**(6375): p. 550-555.
8. Ju, Y.S., et al., *Somatic mutations reveal asymmetric cellular dynamics in the early human embryo.* Nature, 2017. **543**(7647): p. 714-718.
9. Wagner, D.E. and A.M. Klein, *Lineage tracing meets single-cell omics: opportunities and challenges.* Nature Reviews Genetics, 2020. **21**(7): p. 410-427.
10. Park, S., et al., *Clonal dynamics in early human embryogenesis inferred from somatic mutation.* Nature, 2021. **597**(7876): p. 393-397.
11. Lee-Six, H., et al., *The landscape of somatic mutation in normal colorectal epithelial cells.* Nature, 2019. **574**(7779): p. 532-537.
12. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nature biotechnology, 2008. **26**(10): p. 1135-1145.
13. Metzker, M.L., *Sequencing technologies—the next generation.* Nature reviews genetics, 2010. **11**(1): p. 31-46.
14. Behjati, S., et al., *Genome sequencing of normal cells reveals developmental*

lineages and mutational processes. Nature, 2014. **513**(7518): p. 422-425.

15. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* nature, 2008. **456**(7218): p. 53-59.
16. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nature reviews genetics, 2016. **17**(6): p. 333-351.
17. Almogy, G., et al., *Cost-efficient whole genome-sequencing using novel mostly natural sequencing-by-synthesis chemistry and open fluidics platform.* bioRxiv, 2022: p. 2022.05.29.493900.
18. Martincorena, I., et al., *Universal patterns of selection in cancer and somatic tissues.* Cell, 2017. **171**(5): p. 1029-1041. e21.
19. Grada, A. and K. Weinbrecht, *Next-generation sequencing: methodology and application.* Journal of Investigative Dermatology, 2013. **133**(8): p. 1-4.
20. Eisenstein, M., *Innovative technologies crowd the short-read sequencing market*, in *Nature*. 2023, Nature Publishing Group. p. 798-800.
21. Church, D.M., et al., *Modernizing Reference Genome Assemblies.* PLOS Biology, 2011. **9**(7): p. e1001091.
22. Schneider, V.A., et al., *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.* Genome Research, 2017. **27**(5): p. 849-864.
23. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv, 2013.
24. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
25. Andrews, S., *FastQC: a quality control tool for high throughput sequence data. 2010*. 2017.
26. García-Alcalde, F., et al., *Qualimap: evaluating next-generation sequencing alignment data.* Bioinformatics, 2012. **28**(20): p. 2678-2679.
27. Thorvaldsdóttir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Briefings in bioinformatics, 2013. **14**(2): p. 178-192.

28. McKenna, A., et al., *The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Research, 2010. **20**(9): p. 1297-1303.
29. Talevich, E., et al., *CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing.* PLOS Computational Biology, 2016. **12**(4): p. e1004873.
30. Rausch, T., et al., *DELLY: structural variant discovery by integrated paired-end and split-read analysis.* Bioinformatics, 2012. **28**(18): p. i333-i339.
31. Jeffares, D.C., et al., *Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast.* Nature Communications, 2017. **8**(1): p. 14061.
32. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* fly, 2012. **6**(2): p. 80-92.
33. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic acids research, 2001. **29**(1): p. 308-311.
34. Gudmundsson, S., et al., *Variant interpretation using population databases: Lessons from gnomAD.* Human mutation, 2022. **43**(8): p. 1012-1030.
35. Landrum, M.J., et al., *ClinVar: public archive of interpretations of clinically relevant variants.* Nucleic acids research, 2016. **44**(D1): p. D862-D868.
36. Cirulli, E.T. and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nature Reviews Genetics, 2010. **11**(6): p. 415-425.

ACKNOWLEDGEMENT

This research presented in this thesis would not have been possible without the help and support of many individuals.

First and foremost, I would like to express my deepest gratitude to my academic advisor, Professor Ji Won Oh, for his invaluable guidance and unwavering support. I am also sincerely grateful to the members of my thesis committee, Professor Seock Hwan Choi and Professor Hun Mu Yang, for their insightful comments and encouragement.

My heartfelt thanks go to Dr. Mee Sook Jun, who has guided me since the beginning of my graduate studies and provided thoughtful, meticulous advice throughout the development of this thesis. I would also like to express my sincere appreciation to Dr. Soung Hoon Lee and Dr. Nanda Mali, Dr. Seong Gyu Kwon for their continuous support during my time in the lab.

I am especially thankful to June Hyug Choi, Jae Eun Shin, Geon Hue Bae, Areum Cho, Nam Seop Lim, Hyung Bin Chun, Seok Won Jeoung, Su Rim Kim, and Joong Gil Bae, Seung Yeon Woo, Ji-in Choi, with whom I shared countless conversations, discussions, and meaningful experiences throughout the course of my master's program. Their companionship and encouragement made this journey truly rewarding. I would also like to extend my appreciation to all other lab members and the administrative staff, whose contributions ensured a smooth and productive research environment.

I am deeply thankful to the many colleagues at Theragen Bio, where I worked for about five years before beginning my graduate studies. I owe special thanks to Dr. Seong-eui Hong, who mentored me in bioinformatics during my early years, and to my lifelong mentor, Dr. Seung-il Yoo, who consistently supported me through challenges and growth. I would also like to thank my fellow bioinformatics analysis team members—Da Bin Jeon, Dong-min Shin, Ji Hwan Moon, and Eui Hyun Bae—as well as colleagues from the experimental, IT, and sales teams for their collaboration and support.

Most of all, I am profoundly grateful to my beloved family for their unconditional love and unwavering support. In particular, I would like to thank my parents and my sibling, whose endless encouragement made everything possible.

Special thanks also go to Sol Seo for insightful discussions, kind support, and advice regarding my future research endeavors. I am likewise grateful to all the colleagues whose names are not mentioned here but who have contributed meaningfully to my journey.

Lastly, I thank every reader who has opened this thesis and taken the time to engage with my work. I sincerely hope this research will offer insights and contribute to the ongoing pursuit of scientific understanding.

ABSTRACT IN KOREAN

두 가지 시퀀싱 플랫폼을 활용한 유전체 분석 및 계통 추적 연구

최근 연구에 따르면, 인간의 개체 내에서도 유전적으로 서로 다른 세포들이 공존할 수 있는 체성 모자이크 현상이 발생한다는 사실이 밝혀졌다. 이는 발생 초기 또는 생애 동안 세포 분열 중 발생한 체세포 돌연변이로 인해 나타나며, 각 세포는 고유한 돌연변이 조합을 통해 계통 추적이 가능한 유전체 바코드 역할을 한다. 이러한 발견은 세포의 발생적 기원을 추적하고, 암 발생 및 진행 경로를 이해하는 데 중요한 단서를 제공한다.

본 연구에서는 이러한 계통 추적을 목적으로, 현재 가장 널리 사용되는 시퀀싱 플랫폼인 Illumina와, 최근 웨이퍼 기반 시퀀싱 기술로 주목받고 있는 Ultima Genomics를 비교 분석하였다. 특히, 수정란이 성체로 분화하는 과정에서의 세포 계통을 확인하기 위해, 사후 조직 샘플을 활용하였다. 좌측 및 우측 앞다리에서 조직을 채취한 후 1차 배양 및 단일세포 클론 확장을 통해 DNA를 충분히 확보한 뒤, 세포 유형별 특성을 고려한 프로토콜을 적용하여 고품질의 전장 유전체 데이터를 생성하였다.

데이터는 Illumina와 Ultima Genomics 플랫폼에서 각각 생산되었으며, 두 플랫폼의 품질 및 변이 검출 성능을 비교 분석하였다. 통계적 품질 평가는 염기별 오류율, 참조 유전체 커버리지(coverage), PCR 중복 리드를 제거한 후 남은 리드 비율 등의 지표를 기준으로 수행되었다. 또한 다양한 유전체 변이를 탐지하고 분석하여 플랫폼 간 비교를 수행하였다. 아울러 BRCA 변이를 포함한 유전체 변이에 대해 주석(annotation) 및 임상적 의의를 평가하였으며, 이를 통해 실제 응용 가능성을 검토하였다.

마지막으로, 기존 문헌에서 보고된 초기 배아 돌연변이를 바탕으로 세포의 주된 계통과 하위 계통을 식별하고, 각 세포에서 특이적으로 발견되는 체세포 돌연변이를 분석함으로써, 새로운 기술이 기존 기술을 계통 추적 연구에서 얼마나 충실히 재현할 수 있는지를 평가하였다. 본 연구는 향후 체성 모자이크 현상 및 질병 발생 연구에 있어 시퀀싱 기술 선택의 기준을 제시할 수 있을 것으로 기대된다.

핵심어 : 체세포 모자이크 현상, 체세포 돌연변이, 세포 계통 추적, 전장 유전체 분석, 차세대 염기서열 분석, Illumina, Ultima Genomics, 발생 과정, 배아 발생