Ph.D. DISSERTATION OF ENGINEERING

Next-generation sequencing error validation method for rare variant detection

매우 낮은 발생률의 변이 분석을 위한 차세대 염기서열 분석 오류 검증 방법

August 2020

ELECTRICAL AND COMPUTER ENGINEERING
GRADUATE SCHOOL
SEOUL NATIONAL UNIVERSITY
HUIRAN YEOM

Abstract

Next-generation sequencing error validation method for rare variant detection

Huiran Yeom

Electrical and Computer engineering

Graduate School

Seoul National University

The advent of next-generation sequencing (NGS) has accelerated biomedical research by enabling the high-throughput analysis of DNA sequences at a very low cost. However, NGS has limitations in detecting rare-frequency variants (< 1%) because of high sequencing errors (> 0.1~1%). NGS errors should be filtered

out for accurate analysis. Especially the ‘liquid biopsy’, which is non-invasive method to analyze cancer instead of tumor tissue biopsy, required highly accuracy of massively parallel sequencing. For the liquid biopsy analysis, the circulating tumor DNA (ctDNA) should be detected however ctDNA from tumor cell was buried in the major population of cell-free DNA which is come from normal cells. Usually the variant frequency of the ctDNA is lower than 1% for the stage 1 patient or the cancer patient after surgery. In this regard, the NGS error should be validated to distinguish true variant of the DNA sample .Accordingly, several method have been developed depending on molecular barcoding, which uses unique sequence for each DNA molecules by addition to the end of the DNA. Using the molecular barcodes, each molecules can be identified after NGS preparation including PCR amplification. Also, the NGS error can be filtered out by comparing read replicates among those with the same barcodes

However, the barcode-based methods are cost-prohibitive, especially for studying a few number (< 100) of mutation positions with rare variant frequency (< 0.1 %) such as researches for liquid biopsy, and prenatal test. This is because every barcoded DNA strands should be read 10 times although only a few position mutations is of interest.

Also, since each sequencing method (for e.g. cyclic reversible termination (Illumina) or single-nucleotide addition (Roche 454)) can introduce the same type of NGS error (such as indel or substitution), orthogonal validation of NGS error using different sequencing methods, is needed. Previous studies have used Sanger sequencing for orthogonal validation that involves high cost.

Here, I present a cost-effective NGS error validation method in a barcode-free manner. By physically extracting and individually amplifying the DNA clones of erroneous reads, I distinguish true variants of frequency > 0.003% from the systematic NGS error and

selectively validate NGS error after NGS. This method can selectively analyze erroneous reads of interest after NGS run in barcode-free manner. Therefore, I were able to reduce sequencing cost substantially (at least ten times less costly in comparison to barcode-based methods) through the selective analysis of rare variants, without the requirement for redundant barcoding reads. With this method, I achieve a PCR-induced error rate of $2.5\times10-6$ per base per doubling event, using 10 times less sequencing reads compared to those from previous studies.

Also, the previous studies have reported that trimming low-quality NGS reads based on quality score can result in the removal of a few reads of true variants, thus losing critical information from the dataset. This method offers the advantage of analyzing all variants regardless of quality control data trimming, with the possibility to preserve all information in the raw NGS result. I believe that this method can be utilized in scientific fields studying rare variants from samples of high diversity, such as metagenomics and

immune profiling.

For the application, I validated true variant of the circulating tumor DNA extracted from the patient who was diagnosed as stage 2 breast cancer. The variant was detected in PIK3CA gene after NGS error validation with this method.

In addition, this method have potential that NGS error of single-nucleotide addition sequencing can be verified orthogonally using another NGS platform of cyclic reversible termination, thus providing a high-throughput, yet cost-effective methodology.

Keywords: Next-generation sequencing, Variant frequency, Sequencing error, PCR-induced error, laser

Student Number: 2014-21681

Table of Contents

List of Tables

List of Figures

Chapter 1.

Introduction

In this chapter, a short background about liquid biopsy and current analysis method will be introduced. After introduction of background knowledge about circulating tumor DNA of liquid biopsy, state-of-art trends in analysis method such as sequencing strategy of error removal will be described. Especially, the technical innovations of error removal in next generation sequencing will be introduced. Finally, the subject of this dissertation, next generation sequencing error validation method will be presented. This method enables detection of ultra-rare variant DNA copies without any sample loss, and reduces sequencing cost substantially.

1.1. Introduction to liquid biopsy: Seeking cancer signal in the blood for early diagnosis

Liquid biopsy is a noninvasive diagnostic approach which involves the isolation of circulating tumor markers such as cell-free nucleic acids and circulating tumor cells from peripheral blood [1]. This approach is important for high accessibility to diagnose cancer compared to tissue biopsy for cancer diagnosis, which can give chance to diagnose cancer early. In 2017, 9.6 million people are estimated to have died from the various forms of cancer [2]. Every sixth death in the world is due to cancer, making it the second leading cause of death [3]. In this regard, when cancer is diagnosed early survival is more than three times higher, and liquid biopsy enabled the early diagnosis with high accessibility. Various components of tumor cells released into the blood circulation can be analyzed in liquid biopsy sampling, some of which include circulating tumor cells (CTCs) [4], circulating tumor DNA (ctDNA), cell-free RNA, tumor-educated platelets and exosomes [5]. Especially, the circulating tumor DNA (ctDNA) in the blood obtained by liquid biopsy had potential biomarker to detect cancer signal. The ctDNA derived from

tumor cell normally has somatic variant that have occurred during cancer progress and exists in the blood with a small fraction compared to cell-free DNA (cfDNA) which come from normal cell. The fraction differ according to cancer stage, in the case of stage 1 the fraction is from 0.01% to 10% and in the case of the stage 4 the fraction is from 1% to 90%. Although the presence of fragments of cell-free nucleic acids in human blood was first described in 1948, its origin and characteristics was studied actively after the advent of next-generation sequencing (NGS).

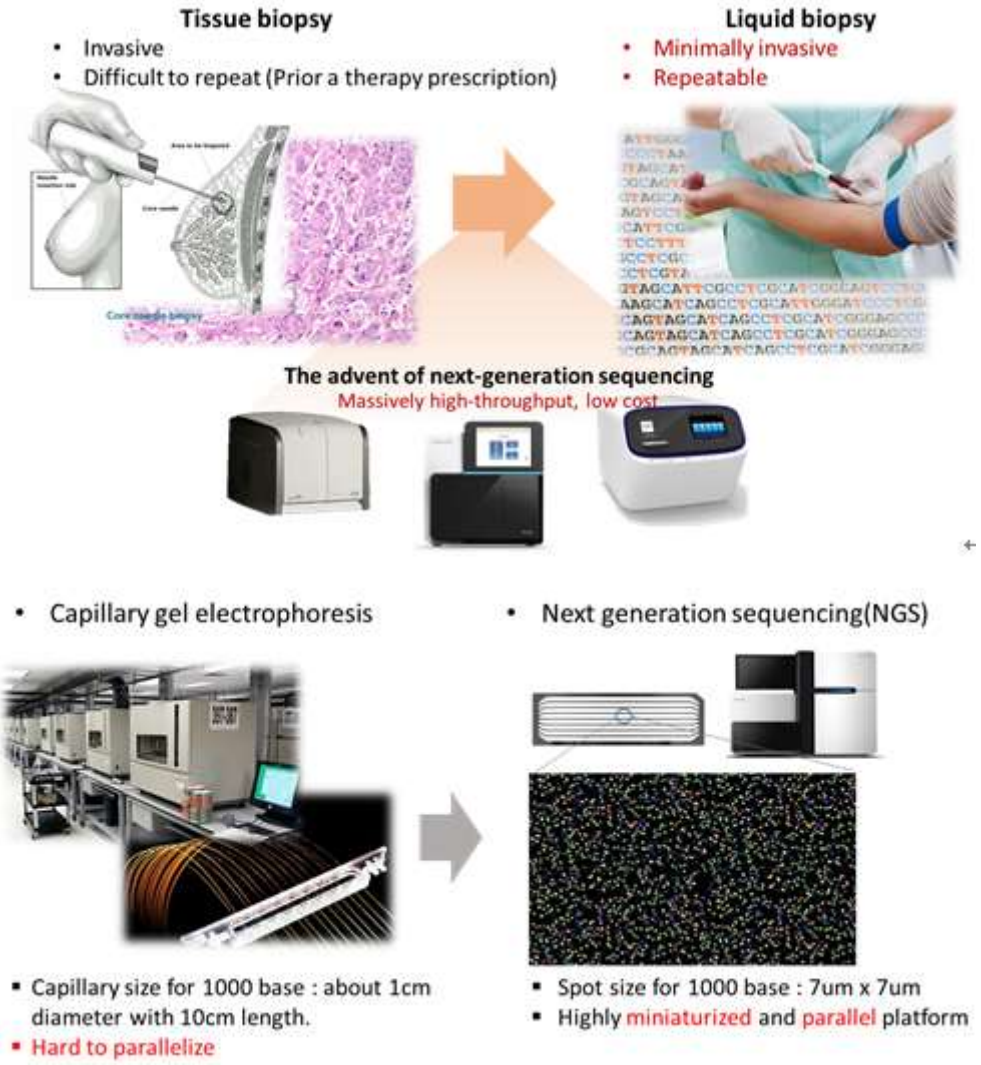


Figure 1.1 Liquid biopsy and next generation sequencing.

1.2. Introduction to next generation sequencing

High-throughput next-generation sequencing (NGS) technologies have revolutionized biological research [6] [7] and clinical fields by enabling detection of important genetic variants [8] [9] [10]. The NGS technology is highly miniaturized and enabled massively parallel sequencing compared to conventional sequencing platform, Sanger sequencing [11]. The key technology is based on two sequencing DNA method of ligation and synthesis. The ligation method add a probe sequence that is bound to a fluorophore hybridizes to a DNA fragment and is ligated to an adjacent oligonucleotide for imaging. Second one is synthesis method, which a polymerase is used and a signal, such as a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into an elongating strand (Figure 1.2) [11]. In both approach, the DNA clones were produced onto the solid surface such as glass. Because the DNA clone size is very small as from 1 micro-meter to 50 micro-meter, the throughput of DNA sequencing can be high which can read the sequences approximately 100,000 to 100,000,000 reads simultaneously (Table 1.1).

Platform	Reads
454 GS junior	~0.1M
Ion PGM 318	4~5.5M
Illumina Next seq 500/550 High output	400M, 800M
Illumina Hiseq 3000/4000	2.4B

Table 1.1 NGS platform reads capacity. NGS utilizes DNA sequencing technologies that are capable of processing multiple DNA sequences in massively parallel.

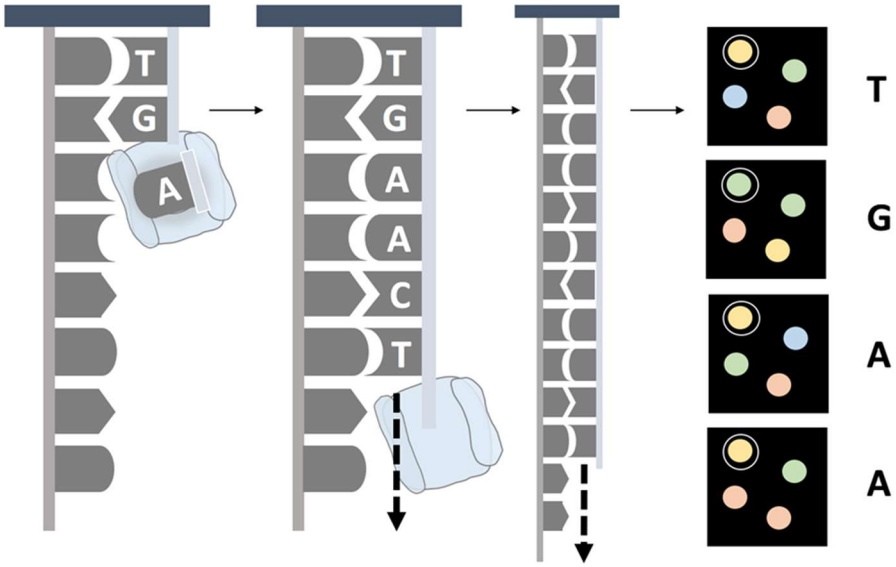


Figure 1.2 Sequencing by synthesis and fluorescence imaging. In the sequencing by synthesis method, a polymerase is used and a signal, such as a fluorophore, identifies the incorporation of a nucleotide into an elongating strand.

1.3. Next generation sequencing error

Although the next generation sequencing can give the chance to analyze genetic information in depth with low cost in high-throughput manner, there was still limitation of high error rate, the possibility to misread base in the DNA sequence. Because of this limitation, detection of the variants at a low frequency such as ctDNA in the blood was challenging with NGS analysis. However, detecting analyzing rare somatic variants is important because it provides clues towards the exact biological environment [12]. For example, detecting variants of rare frequency in cancer biology can be crucial indicators for effective treatment strategies through better understanding of the tumor heterogeneity and clonal evolution [13]. Similarly, early diagnosis of diseases by drug-resistance or organ transplant rejection requires sensitive NGS analysis with high accuracy, since the ratio of the variant is as little as below 1%. However, detection of the rare variants at a frequency below 1% remains challenging because of the high NGS error rate (0.1-1%) [14].

Platform	Error rate (error type)
454 GS junior	1%, indel
Ion PGM 318	1%, indel
Illumina Next seq 500/550 High output	<1%, substitution
Illumina Hiseq 3000/4000	0.1%, substitution

Table 1.2 The sequencing error rate of NGS platform.

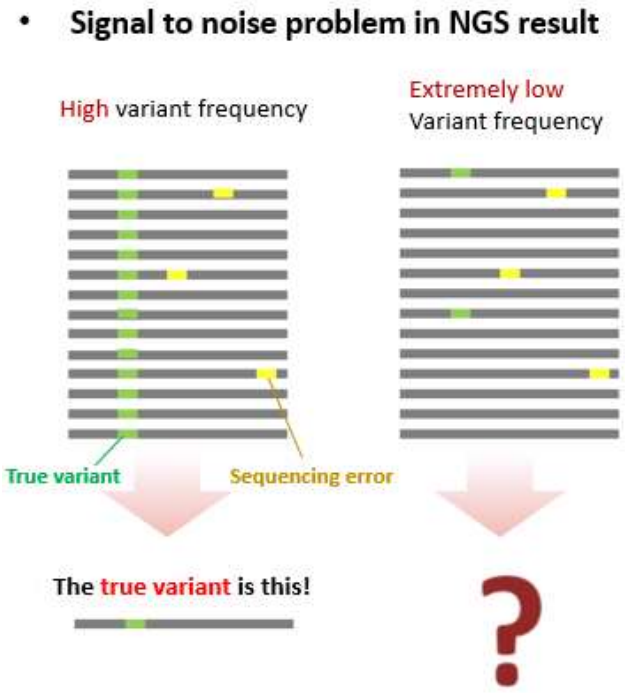


Figure 1.3 Signal to noise problem in NGS result. Detection of the rare variants at a frequency below 1% remains challenging because of the high NGS error rate (0.1–1%)

1.4. Method of consensus-based error correction

To correct the sequencing error of NGS platform, consensus-based error correction method have been developed. The approach make the true variant frequency increased more than NGS error rate by reading the DNA molecule replicates more than three. Because the NGS error occurs randomly along the DNA sequence, if replicates of the DNA molecules are read simultaneously, the true variant will be revealed at the same position of the expected DNA sequence and the randomly NGS error can be filtered out. To identify each origin of the different DNA molecules from their replicates, barcoded sequencing have been mostly used (Figure 1.3). After extracting the DNA from cell or plasma, the unique molecular identifiers (UMI) is ligated to each DNA molecule and the prepared DNA molecules are sequenced by NGS. Then, in the sequencing result, each barcoded molecules of the replicates make their read family and the true variant can be found from the aligned sequences within same barcode. Therefore, the replicate reads for filtering NGS reads I required resulting in increasing NGS sequencing cost.

1.4.1. The consensus-based error correction with barcoded sequencing

Molecular barcodes can be divided as exogenous and endogenous form [15]. Exogenous barcodes are mostly random sequences that are incorporated into either sequencing adapters or PCR primers. Endogenous barcodes describe the randomly or semi-randomly generated fragmentation points at the ends of DNA molecules in ligation-based library preparation methods [16]. The first study of the NGS error correction to detect and quantify the rare mutations with massively parallel sequencing is called as Safe-Sequencing System, 'Safe-SeqS' [15]. The Safe-SeqS involves two basic steps. First, UMI is introduced to each DNA template molecule to be sequenced. Second, each uniquely tagged template is amplified, and many daughter molecules with the identical sequence are generated. In this study, they defined "supermutant' as the identical mutation that was revealed in at least 95% of family members. With Safe-SeqS analysis of the same data, they determined that 69,505 original template molecules were assessed in this experiment (i.e., 69,505 read families of each barcode sequence, with an average of 40

members per family, were identified. All of the polymorphic variants identified by conventional analysis were also identified by Safe-SeqS. However, only eight supermutants were observed among these families, corresponding to 3.5 × 10-6 mutations/bp. Thus, Safe-SeqS decreased the presumptive sequencing errors by at least 70-fold.

Another representative method is duplex consensus sequencing, which ligated different UMIs to independently barcode each strand of individual DNA duplexes. This approach enable to distinguish true variant derived from one strand of the other [17]. When the strands are separately amplified, the adapted molecule contains both a UMI and mostly achieve true duplex error correction can be achieved. This method enables to detect the PCR induced error, which usually occur in a single strand, because it validates both DNA strand. In this regard, duplex sequencing method is the most accurate in removing sequencing error rate, however, it requires NGS reads more than twice compared to original barcoded sequencing.

Other methods are single-molecule molecular inversion probes (smMIPs), circular sequencing (CircSeq) and CypherSeq, etc. In the

case of smMIPs, a single oligonucleotide with overlap sequence to a DNA sample is used. The overlap sequence is hybridized to the target sample and ligated to form tagged, closed loop products that can be enriched, amplified and sequenced. The smMIP can be easily multiplexed together and solve the normalizing issue in barcode sequencing and double tagging. However, high depth of sequencing is still required.

CircSeq [18][19] improves cost-efficiency by keeping the duplicate rate more uniform, which utilize rolling circle amplification (RCA) for sample preparation. The DNA smaple are melted into very short single-stranded pieces that are then circularized and copied into concatemers via rolling-circle amplification. Since it is independent on tag-based barcoding of unlinked copies, which may have either too few or an excess of copies present, it can overcome normalizaing issue in NGS.

The Cypherseq [19] also includes rolling-circle amplification from primers targeting both strands after ligation into a circularized adapter sequence to achieve a degree of target enrichment before PCR amplification. In previous barcode based approaches, the PCR

error during sequencing process cannot be distinguished, the Cypherseq can detect only true variant without PCR error by combining RCA and target PCR amplification.

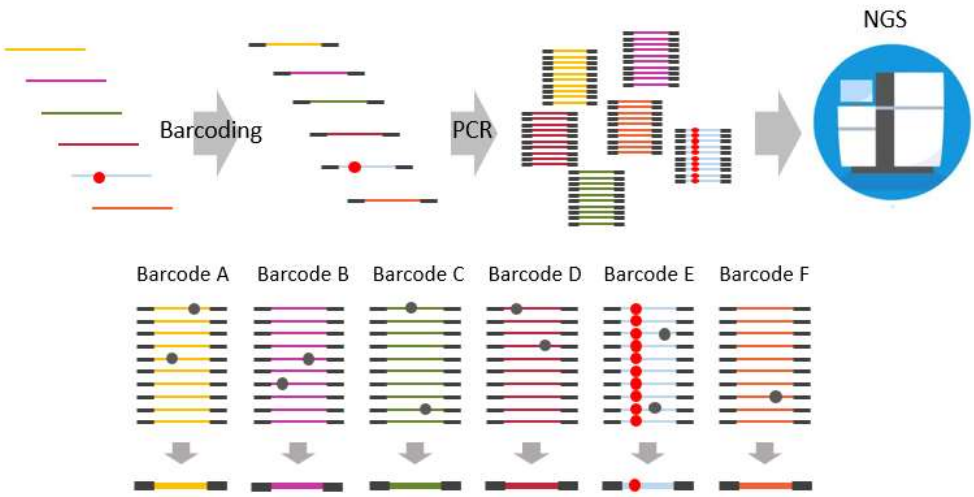


Figure 1.4 Consensus-based NGS error correction with barcoded sequencing

1.4.2. The limitation of the barcoded sequencing

When DNA template replicates can be made and sequenced, the erroneous base calls can be filtered out by establishing a consensus sequence from high-coverage sequencing reads. The consensus sequence is derived from the reads in each family of read replicates, with a typical criterion being that the read family must contain at least three read replicates before a consensus sequence is derived.

However, the number of the read replicates are mostly varied during the process of constructing the read families including barcode ligation and PCA amplification. The variation of the number of each read families can lower the fraction of the read family which contain important information such as single nucleotide variation [18]. Therefore, Additionally, the reads including the rare variants can be buried among other unnecessary reads due to non-normalized read replicates generated during sample barcoding process [13].

Considering this issue, the tag-based barcoding of unlinked copies may have either too few or an excess of copies present resulting extremely high depth (>3000x) of sequencing and it results to require more sequencing cost.

Also, the depth of coverage required for consensus building remains cost-prohibitive for low variant frequency. The lower the variant frequency is, the more redundant reads are, which translate to all sequencing reads must be replicated, regardless of whether the sequencing reads represent rare variants or not (i.e., reads with normal sequence or other non-targeted variants).

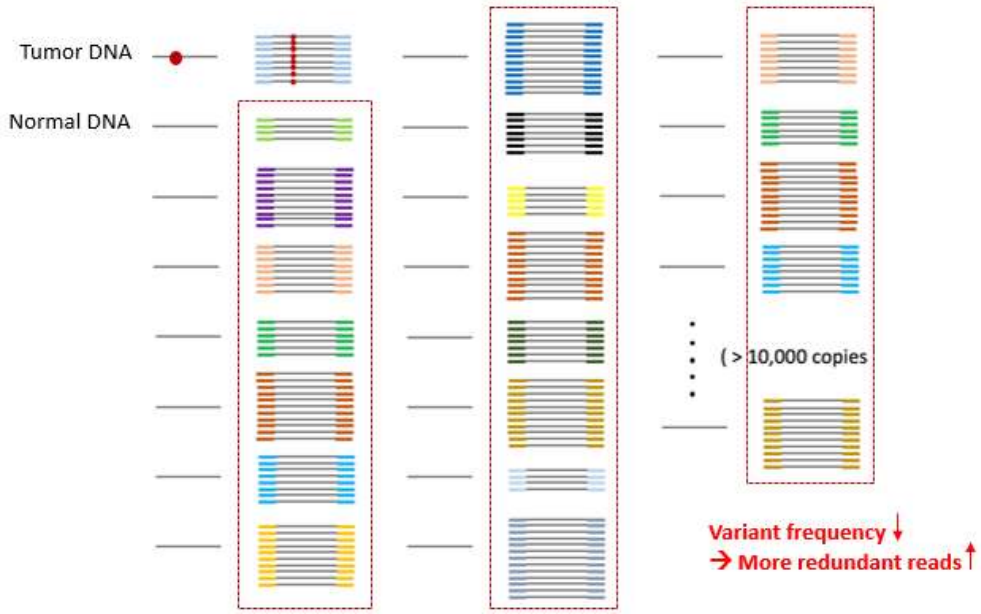


Figure 1.5 The barcoding approaches require higher NGS reads depth.

Chapter 2.

Platform development

2.1. Principle of next-generation sequencing.

2.1.1. Potential error sthisces in next-generation sequencing platform

In current NGS platforms [11], the errors occurs depending on the specific sequencing and imaging types to each platforms. For example, substitution errors can arise in platforms such as Illumina and SOLiD when incorrect bases are introduced during clonal amplification of templates. Furthermore, In the case of Illumina sequencing, the error has revealed depending on specific sequences such as repeated sequence or "GGC" with its reverse sequence that possibly arises from either single-strand DNA folding or sequence-specific alterations in enzyme preference [20]. The

single-molecule, real-time (SMRT) platform of Pacific Bioscience yields long single-molecule reads that are subject to false insertions and deletions (indels) from non-fluorescing nucleotides. Pyrosequencing (for example, Roche 454 platforms) and semiconductor sequencing (for example, Ion Torrent) have difficulty to define the length of homopolymer sequences, which results in carry-forward insertion and deletion errors.

Especially, the pyrosequencing technology in Roche 454 platforms is based on sequencing by synthesis which is performed on the bead carrying around 10 million DNA molecules amplified by emulsion PCR (emPCR) starting from one single DNA fragment. The sequencing is performed in parallel on around one million beads deposited in wells on a plate [21]. Each The sequencing is performed by cyclic flowing (T, A, C, G) of nucleotide reagents over the plate, every bead giving rise to at most one DNA sequence. Each flow produces a light signal in each of the beads, which is ‘negative flow value’ of either a very weak signal, in practice being between 0 and 0.5, indicating that no base was incorporated or ‘positive flow value’ of a stronger signal proportional to the length of a homopolymer run

[22]. In this regard, each base is called by analyzing the light signal intensity of each DNA clones on the sequencing substrate with quality information. The light signal strength from the chemical reaction in the sequencing process is the basis for correct determination of homopolymer lengths and hence responsible for data accuracy [23]. If the homopolymer sequence was too long, the light is detected with lower intensity and the sequence can be misread as shorter length. The sequencing error consists of most systematic error, such as the detected light analysis, and few molecular variants, such as polymerase synthesis error. Moreover, carry-forward errors occur when there is insufficient flushing between flows and the remaining nucleotides are present in the wells. Also, incomplete expansion of the template due to insufficient nucleotides in the flow can lead to out-of-sync readings.

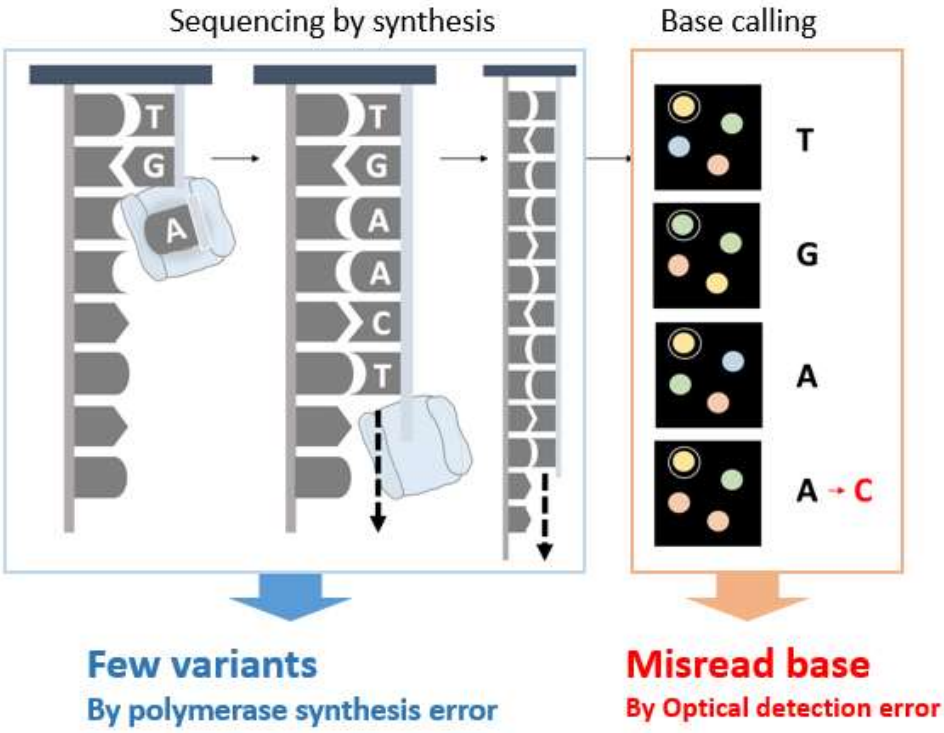


Figure 2.1 Synthesized DNA clusters will have few variants

2.2 Barcode-free of next-generation sequencing error validation

2.2.1 Synthesized DNA clones isolation

To validate NGS error in cost-efficient manner, only erroneous reads of interest can be considered selectively by excluding redundant non-interest NGS reads consumption. The erroneous reads, which are to be determined as variants or NGS errors, can be any reads of interest which need verification, or can be those harboring variations compared to a reference sequence. I approached to analyze specific DNA molecule clones corresponding to the erroneous reads after an NGS analysis. The systematic NGS errors occurred during signal detection, however, the original molecule remains unchanged [24].

To be more specific, the molecule can be changed by PCR error during sequencing by synthesis process and the probability of the PCR errors leading to false variant calls is extremely low. The reason is that each DNA clone is composed of many homogenous DNA molecules, which is same sequence. Even if the PCR errors occur during early cycles of amplification, the DNA molecules with the PCR

errors will be the small part of a DNA clone. Before calculating a probability of having false variant calls at the validation step, the extreme case can be assumed; more error-prone conditions First, assume that the length of the DNA is 400 bp, the number of homogenous DNA molecules in the DNA clone is 100, and the polymerase used for PCR has a substitution error rate of 10^{-4} / base*cycle. The 400 bp length is quite a long one for the DNA molecules in NGS platform. Also, the 100 is quite a small copy number of each clones in NGS platform and 10^{-4} / base*cycle is the order of error rate of Taq polymerase, which has a high error rate than other polymerases (Phusion, KAPA, Q5, etc). Second, in the validation step, assume that the sequence of each position is determined as a sequence that occurs more than half of the whole sequences in a DNA clone. Lastly, assume that the PCR error occurs only at the first cycle of amplification in one specific position of DNA.

Probability of occurring PCR errors in the first cycle is known as followed [25].

$$P(k) = \beta(nl, k)(cx)^k(1 - cx)^{nl-k},$$

where

k = the number of errors in the first cycle

n = the number of single-stranded copies before amplification

l = the length of DNA molecules

c = proportion of mismatches detected by a given method

x = error rate per base per cycle (error rate of polymerase)

β = coefficient of binomial distribution

Then, the probability of occurring one PCR error from one DNA read is

$$P = \beta(l,1)(x)^1(1-x)^{l-1},$$

where k = 1, n = 1, and c = 1.

Therefore, according to this assumption, the probability of leading false variant calls due to the PCR errors in the amplification is followed.

$$P(l,x,n) = \beta(l,1)(\frac{1}{3}(x)^1(1-x)^{l-1})^{0.5n},$$

where l = 400, x = 10^{-4} / base*cycle, and n = 100.

The $\beta(l,1)$ is multiplied once because all 0.5n DNA molecules must have a same PCR error in a same position. A constant 1/3 is multiplied because three bases, except a normal base are possible

candidates as the PCR error. The power of 0.5n means that 50 % of DNA molecules in a DNA cluster have a same PCR error. Applying the value of this assumption, the probability is about $2.4 * 10^{-95}$. Thus, the probability of 50 % of DNA molecules in DNA cluster having the same type of PCR error in first cycle of amplification is extremely low.

As the assumption is an extreme case, the probability will be much lower in real conditions. For example, the length of DNA (l) is usually shorter than 400 bp, polymerase error rate (x) is usually lower than 10^{-4}, and the copy number of DNA molecules in a DNA clone (n) is larger than 100 in various NGS platforms. Also, the sequence of each position is determined as a sequence that occurs more than 70~90 % of the whole sequences in a DNA clone.

As a result, the DNA clones have few number of variants and I attempted to physically isolate the DNA clones from NGS substrate to verify their sequencing errors. I performed separate PCR amplification with the isolated DNA clones. Since only true bases can be copied, rather than being mistakenly referred to as base duplication during PCR, the sequence information provided by the

amplified DNA clone did not contain errors that were incorrectly referred to as bases in previous NGS runs.

The full-process of barcode-free NGS error validation is demonstrated in Figure. 2.2. Firstly, erroneous NGS reads of interest were selected as verification targets, which have unintended variations compared to a reference sequence. Secondly, each DNA clone corresponding to the target reads was extracted from the NGS substrate using the laser retrieval system that retrieved over 40 DNA clones per one minute into 96-well plate automatically. Thirdly, the obtained DNA clones were amplified individually by PCR. As the laser retrieval system enables to isolate the DNA clones individually into each well of a 96-well PCR plate, PCR reaction can be performed right after the retrieval of the DNA clones. Also, I were able to track the corresponding NGS read information through the well location of each selected DNA clone. Finally, the amplified DNA were sequenced individually resulting in the duplicated true bases to be above 95% in the amplified molecules, the removal of NGS error of miscalled bases, and identification of true variants. I sequenced the DNA molecules by Illumina sequencing or Sanger sequencing in those cases where the

number of targets was low (<10). This method can also filter out variants, which can be damage, degradation or PCR error of DNA on the NGS substrate, occurred during the validation process [26] [27].

For NGS reads selection for verifying true variants, prior to selecting sequencing reads that needed to be validated, I constructed a hash table that mapped XY coordinates in 454 junior GS sequencing reads to pixel coordinates in the NGS chip image. The sequencing data was aligned to design sequence using basic local alignment search tool (BLAST) standalone version (BLAST-2.3.0+, NCBI). For verifying true variants of interest, I extracted the information of all sequencing reads that had variant(s) (e.g. substitution, insertion, or deletion) or a few sequencing reads that had variant(s) at the desired position from BLAST results. These extraction processes were done by the in-house python code. With the hash table, I constructed the list of pixel coordinates of each selected reads. The pixel coordinates were used as positional information.

For the DNA clone isolation, I used a laser retrieval system to which precisely separated the micro-objects by focusing the radiation pressure of the pulsed laser on the desired target. The laser

retrieval system include pulse laser (Q-Switched Nd:Yag laser, Minilite, Continuum), true-color charge-coupled device (CCD) camera (Guppy PRO F-146C, ALLIED), two motorized stages, and one inverted microscope (IX71, Olympus) with a ×10 objective lens. Also, to achieve high-throughput separation, I automated to rigorously isolate target DNA cluster without human intervention through in-house LabVIEW program. For automated laser retrieval system, the exact location of the DNA clone on the NGS plate should be calculated to isolate accurately. Therefore, I approached with two computational methods by considering shorter processing time. First, I developed an image stitching method, which recognized the features on the NGS plate and detected the corresponding center with the decimal value coordinate rather than the integer. Since the offset between different images was not approximated to an integer, the error was not accumulated even if a lot of images (i.e. hundreds) are stitched along one axis. Then, I developed an analytic 'diffusion-like mapping' to calculate the transformation matrix by applying a point pattern matching algorithms, such as invariant to translations, rotations, and scale changes. In order to calculate the location of the

desired particles immediately, the matrix is analytically derived from the least-square error estimation of multiple two-dimensional points. Therefore, the exact location of the DNA clones of interest was obtained with high accuracy and in a short time. Over 2500 DNA clusters were retrieved per one hthis into 96- well or 384-well plates. And each retrieved beads were amplified separately through PCR conditions of initial denaturation at 95 ° C for 3 min followed by 26 cycles of 95 ° C for 30 s, 64 ° C for 15 s, 72 ° C for 30 s, and final elongation at 72 ° C for 5 min with Taq polymerase 2x pre-mix (BioFact).

For validation sequencing, it was performed by Illumina Miseq (Celemics, Korea) or Sanger sequencing (Macrogen, Korea). For comparing var- iants before and after direct NGS error validation, each sequencing reads were aligned to design sequence (dapA gene of E. coli) using BLAST or Burrows-Wheeler Aligner (BWA) mem aligner (http://sthisceforge.net/projects/ bio-bwa/files/) followed by processing with SAMtools; view, sort, and mpileup (http://www.htslib.org/doc/samtools.html). For calling variants, I used VasrScan; pileup2csn (http://varscan.sthisceforge. net/using-

varscan.html). Finally, each sequencing variants (>80-95% of consensus reads) were compared excluding low reads (>2% of average depth) from Illumina sequencing results.

2.2.2 Erroneous sequence validation

The NGS error was validated with sequence-known DNA sequencing. In order to construct a library of DNA samples with known nucleotide sequences, it is known that almost no mutation occurs. By targeting the essential gene of Escherichia coli MG1655 (dapA), it is possible to minimize the mutants of the DNA molecule [28]. The target gene region (261 bp) was amplified by colony PCR, and each DNA strand of the PCR product was individually cloned by the vaccinia DNA topoisomerase I cloning method. In addition, the plasmid was extracted from the clone and the sequence was confirmed by Sanger sequencing [29]. Using this sequence-validated DNA sample, I sequenced through the 454 Junior GS sequence and selected target reads with known sequence variants. The detailed protocol is like below. Plasmids were extracted from monoclonal E. coli clones followed by PCR amplification (95 ° C for 2 min followed by six cycles of 98 ° C for 30 s, 62 ° C for 15 s, 72 °

C for 30 s, and final elongation at 72 ° C for 2 min) with KAPA HiFi HotStart ReadyMix (KAPA Biosystems). For preparing DNA templates to accumulate PCR- induced error, I extracted E. coli genomic DNA by using DNeasy blood & tissue kit (Qiagen), and performed 60 cycles of PCR with the E. coli genomic DNA (Supplementary Figure 5). The PCR protocol was according to standard PCR protocol of Phusion® High-Fidelity DNA Polymerase (M0530).

The results of NGS are 15,126 bases (0.147%) and 15,024 bases (0.148%) respectively, which are indels and substitution, which can be expected to be the wrong calling base for NGS errors. As a result of NGS, the sample size was statistically calculated to determine if the variant call is true or systematic error. DNA clusters corresponding to 1817 reads (total of 160,281 bases) including 817 indels and 1048 substitutions were selected. As a result, it was confirmed that 99.47% of variant calls occurred only in NGS results, and that there were no variants in the verification sequence results. Notably, all 817 indel variants were misreading artifacts in NGS sequences, except for one indel error. One indel error with a 'C '

inserted at the 89th position in the sequence may have resulted from a DNA synthesis error in the primer sequence (80th to 99th). In addition, 0.53% of variant calls were true variants, a true mismatch present in both 454 and validation sequence results. The cause of the mismatch may be DNA damage [30] due to sample preparation and storage, or contamination due to mixing of DNA molecules of similar sequence.

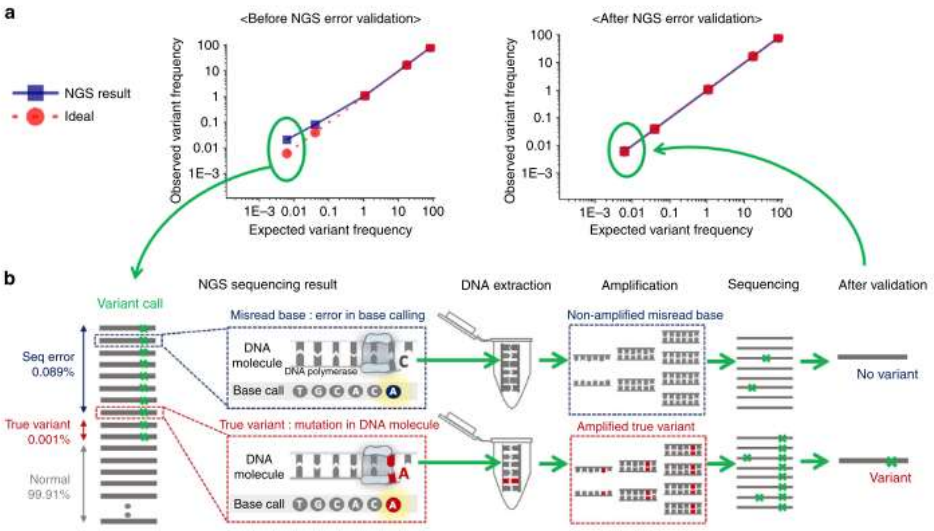


Figure 2.2 The barcode-free NGS error validation method through the DNA clone isolation.

2.3 The sensitivity of barcode-free of next-generation sequencing error validation

2.3.1 Sample preparation

To establish the sensitivity of barcode-free NGS error verification, the method determined the limit of detection using a spike-in DNA library with different variant fractions diluted from 0.01% to 90%. I assumed that the miscalled base of NGS errors called more variants than expected Variant Frequency (VF) at each position. I tried to see if I could identify the misused error of a rare VF (<1%) of a DNA sample. To distinguish spike-in DNA samples (0.01-90%) representing each VF in the NGS run, the DNA samples had different variants with mutations at different positions. Prior to performing NGS, DNA samples were quantified by real-time qPCR (Applied Biosystems, 7500 Fast) and diluted from 0.01% to 90% (0.01%, 0.1%, 1%, 10%, 90%). In addition, labeling each DNA sample with different variants allowed to accurately verify the expected frequency of the mixture after NGS runs from 0.002% to 95.6%.

In more detail, the whole process of preparing 5 spike-in DNA samples with different variant frequency (VF) is described. First,

colonies were picked and sequences of target region were verified by Sanger sequencing, confirming that each DNA samples had one real mutation at different positions. Secondly, DNA samples were quantified by real-time qPCR (Applied Biosystems, 7500 fast) and then diluted to make VF from 0.01 % to 90 % (0.01 %, 0.1 %, 1 %, 10 %, and 90 %). Final qPCR was done to confirm the VF of diluted DNA samples and the range of VF was from 0.002% to 95.6%. Lastly, DNA samples were labeled using different primers, followed by next-generation sequencing (Roche, 454 GS Junior). Three replicates of each DNA sample were prepared and quantified by qPCR. 15 reactions (5 samples x 3 replicates) were prepared, and SYBR Green I was used as the fluorescent dye. The following table shows the threshold cycle (Ct) values. The Ct value was used to calculate the relative amount of the DNA sample using the relative standard curve method.

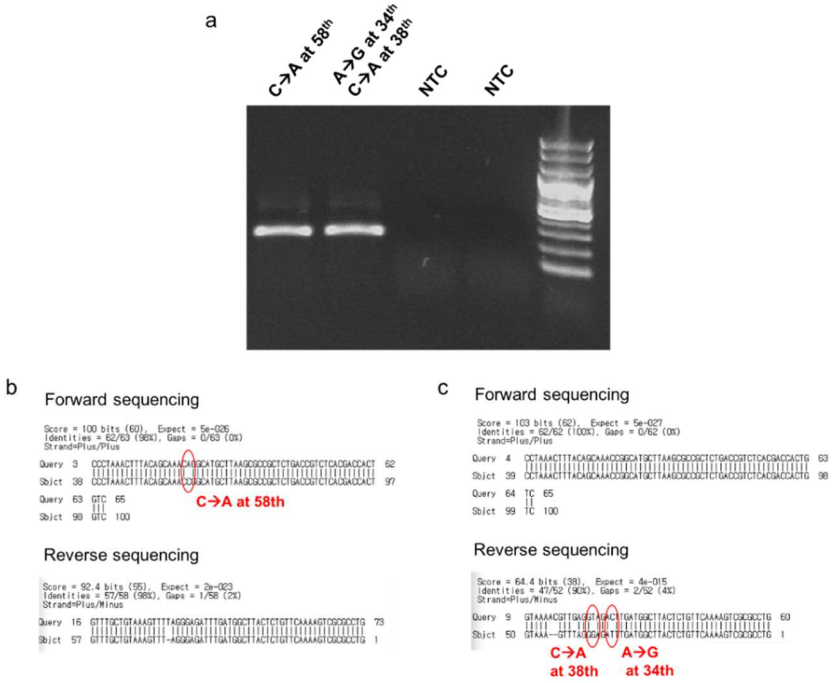


Figure 2.3 Monoclonal DNA templates. (a) Gel electrophoresis image of PCR product of *dapA* gene (261 bp) deom *E.coli.* Each PCR product was separated into the plasmid vector by Vaccinia DNA topoisomerase I and cloned (Biofact, All in One™ PCR Cloning Kit). With Sanger sequencing, I could identify each insert DNA fragments of plasmids has its own mutation at a specific position, which can be caused by polymerase error or damage such as oxidation or hydrolysis. (b) Sample #1 has a variant at 58th position. (c) Sample #2 has two variants at 34th and 38th position. NTC = No template control; In PCR reaction, water was added instead of DNA template.

BioFact™ 100 bp Plus DNA Ladder included for size reference.

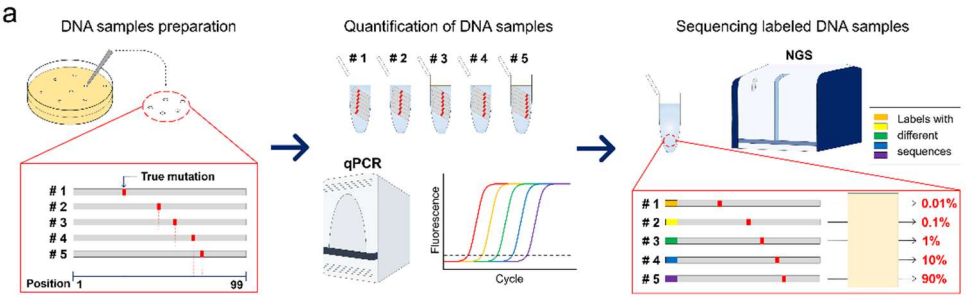
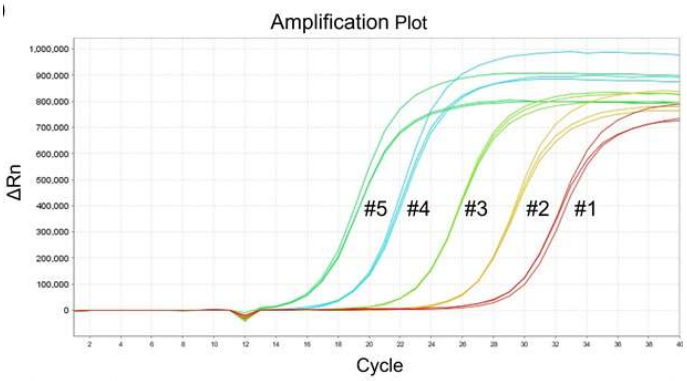


Figure 2.4 Preparation of 5 spike-in DNA samples with variant frequency (VF) from 0.01 % to 90 %.



Sample Number	Ct_replicate#1	Ct_replicate#2	Ct_replicate#3	Average Ct
#1	28.33	28.32	28.76	28.47
#2	25.58	25.67	25.69	25.65
#3	22.15	22.12	22.07	22.11
#4	18.32	18.34	18.41	18.36
#5	15.66	15.5	15.69	15.62

Figure 2.5 The result of the final qPCR. Three replicates of each DNA samples were prepared and quantified by qPCR.

Sample Number	VF (qPCR)	VF (Repeat 1)	VF (Repeat 2)	VF (Repeat 3)	VF (Repeat 4)
#1	0.014%	0.000%	0.002%	0.003%	0.003%
#2	0.093%	0.102%	0.105%	0.031%	0.040%
#3	1.032%	0.615%	0.629%	0.797%	1.074%
#4	13.278%	3.602%	4.055%	19.745%	17.107%
#5	85.568%	95.681%	95.209%	79.424%	81.776%

Figure 2.6 The variant frequency of each DNA samples from qPCR and fthis repeated experiments.

2.3.2 DNA clone isolation corresponding to erroneous sequence

The NGS results found an unexpected variant at 5 positions, giving a total of 164,332 and 806 total readings from 4 replicates. Rare variants of less than 1% of VF were filled by the incorrect base of NGS errors. Sequencing results showed 13.7-fold variants with an average of less than 1% of expected VF (R2 = 0.77, <VF 1%). Tried to validate all unexpected variants individually for all VFs.

A DNA clone with an unexpected variant in the raw NGS data was extracted and verified as an NGS error. In the second replicate, the variant (C to T at position 31) was detected five times more than expected, but two specific variants (C to T at position 31 and 38 at position 31) were found in the sequence. It was confirmed that the positions C to A) existed. It may have occurred during the emulsion PCR step due to the two DNA templates in the emulsion. The R^2 value can be calculated using the observed variant frequency (VF) data and the expected VF data. In the raw NGS results without NGS error validation, the R^2 value was 0.77 below 1% VF. However, the R^2 value after NGS error verification without barcode was 0.98. This means

that NGS errors were properly filtered. DNA spike-in samples were diluted in five-digit different variant fractions with VF 0.01% to 90% and corresponding VF was measured from 0.002% to 95.6%. And I obtained 806 suspicious readings from 4 DNA substrates (4 replication experiments). The number of reads obtained from each duplicate is 254, 240, 188, and 124 reads, respectively. The number of suspicious variants is 819 bases because some reads have multiple variant calls.

The validation eliminated the NGS error, thus reducing the observed VF. 0.05% VF 90% reduction, 1.2% VF 10% reduction, 4.5% VF 1% reduction, 65% VF 0.1% reduction, 88 VF 0.01% reduction. Variant calls with NGS results can reduce mean VF by 0.57 fold below VF 1% (R2 = 0.98, <VF 1%), sensitively distinguishing actual variants from NGS error at VF 1%. . The low throughput readings on the 454 sequencing platform (<100,000) limited sensitivity detection but allowed validation of rare variants down to VF 0.003%.

Also, of the two given scenarios, in case a), this method requires a lot of cost for validating NGS errors in many variant sites, while in case b), this method is more useful to identify the NGS errors with

lower cost than the barcoding methods. In other words, as more read numbers are verified, the cost and time increase linearly during this validation method. However, in most cases, rare mutations are buried in the more frequent NGS errors. In this case, the minimum number of reads to be verified will be limited according to the NGS error rate. Therefore, the number of variant sites to be analyzed and the number of reads containing the target sites are important factors in determining the practicality of this method. In that manner, this method will be more effective in cases similar to case b). Case a) is a situation where there are many variant sites, and it is necessary to verify whether all variant calls generated for each variant site are NGS errors. In this case, if there are fewer DNA molecules per site (Figure. 2.9), the number of reads to be analyzed absolutely is small, but not smaller than the number of reads with NGS errors. Thus, the number of reads to be verified depends on the number of NGS errors at the sites, resulting in linear increase in cost, according to the number of variant sites.

In case b), only a small number of variant sites can be analyzed, but similarly, the number of reads to be verified depends on the

number of absolute variant calls. Therefore, as shown in Figure b, all variant calls (NGS error + variant) generated for a single variant site should be verified, and since the absolute number of variants is large, many reads should be verified. However, considering that this platform is used for rare variants, the number of reads to be verified should still be small. Considering both cases, the number of variant sites is the main factor determining the cost for NGS error validation and measuring allele fraction.

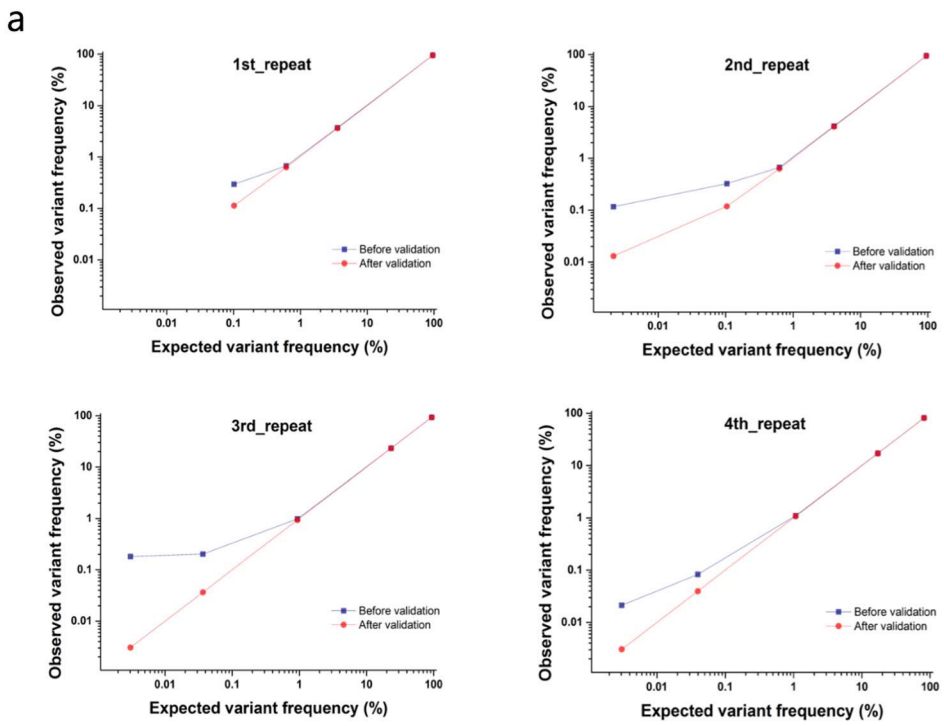


Figure 2.7 Barcode-free NGS error validation for detecting spike-in DNA sample varying amounts from 95.6 % down to 0.002 % in fthis repeats.

	1st replicate	2nd replicate	3rd replicate	4th replicate
0.01 %	66	52	61	99
0.1 %	92	102	58	13
1 %	26	18	14	9
10 %	63	66	0	0
90 %	11	3	53	3

Figure 2.8 The number of retrieved DNA reads for measuring the sensitivity of this validation method.

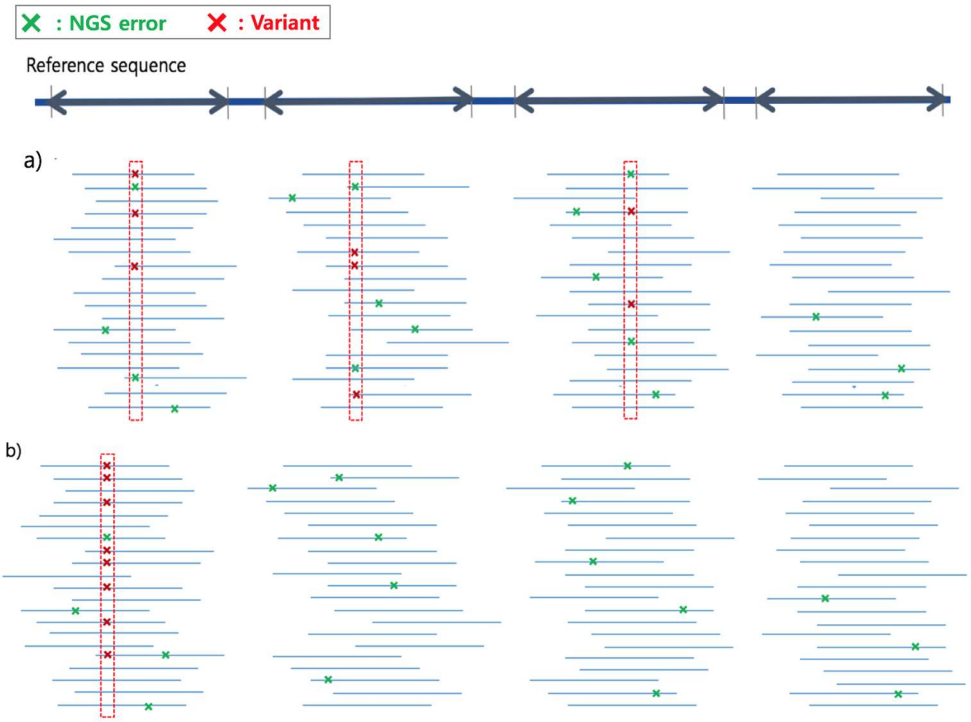


Figure 2.9 Sequence alignment on reference sequence. Case a) is a situation where there are many variant sites, and it is necessary to verify whether all variant calls generated for each variant site are NGS errors. In this case, if there are fewer DNA molecules per site, the number of reads to be analyzed absolutely is small, but not smaller than the number of reads with NGS errors. Thus, the number of reads to be verified depends on the number of NGS errors at the sites, resulting in linear increase in cost, according to the number of variant sites. In case b), only a small number of variant sites can be analyzed, but similarly, the number of reads to be verified depends

on the number of absolute variant calls. Therefore, as shown in Figure b, all variant calls (NGS error + variant) generated for a single variant site should be verified, and since the absolute number of variants is large, many reads should be verified. However, considering that this platform is used for rare variants, the number of reads to be verified should still be small.

Considering both cases, the number of variant sites is the main factor determining the cost for NGS error validation and measuring allele fraction.

2.4 Distinguishing PCR-induced error from NGS error

2.4.1 Sample preparation

I investigated whether PCR-induced errors that occur during PCR thermocycling can be distinguished from NGS errors with fewer reads (<10) than in previous studies [31] [15]. To construct the DNA template, a variation of the DNA template (261 bp) was introduced using an extended PCR protocol with 60 cycles of PCR, resulting in 43 doubling events, per base. Over 0.01% of VF resulted in accumulated variants.

Also, the template is from essential gene sequence of *E.coli.* genomic DNA. The essential gene is *dapA* encoding most of the enzymes leading to DAP production are essential to *E.coli* [28]. Therefore, mutants that lack the gene product required to maintain vigorous growth fall into the same category as mutants that have a "true lethal" mutation. Therefore, a particular gene may be classified as essential by genetic footprinting, but may result in a corresponding viable deletion mutant. Therefore I designed the primer sequence for targeting the essential gene and made amplicon through PCR

amplification.

PCR-induced error (per base per doublings) was calculated as {(True variants) / (Total sequence length)}/doublings. For true variants, I counted the bases according to variants validated through this barcode-free NGS error validation method. For total sequence length, I counted all bases sequenced in 454 sequencing result but the primer region was excluded to avoid DNA synthetic error. For measuring doublings, I quantified gDNA copies before and after PCR amplification through real-time qPCR (Applied Biosystems, 7500 fast) and divided the amplified DNA copies measured after PCR amplification by the initial DNA copies. PCR mixture for qPCR was followed as before PCR amplification: gel-purified E. coli gDNA (see in Methods—Library construc- tion) 1 μl, 10 μM, forward primer 1 μl, 10 μM, reverse primer 1 μl, KAPA SYBR FAST qPCR Master Mix (2×) 10 μl, nuclease-free water up to 20 μl. After PCR amplification: the amplified DNA sample after three steps of 60 cycles PCR 1 μl, 10 μM, forward primer 1 μl, 10 μM, reverse primer 1 μl, KAPA SYBR FAST qPCR Master Mix (2×) 10 μl, nuclease-free water up to 20 μl.

This DNA sample was used to run a 9898 read NGS run containing 2,197,356 bases. Since PCR-induced errors can occur anywhere in the DNA sequence, I extracted all DNA clones with variations at any position compared to the designed sequence. Following NGS error validation, the distribution was observed for the number of PCR-induced errors along the sequence. In addition, the primer region was excluded to avoid counting DNA synthesis errors that could occur during DNA primer synthesis. This results show that NGS errors occurred more frequently at the end of the sequence and in homopolymer sequences. However, PCR errors occurred randomly. From NGS runs, the characteristics of NGS errors (~ 1% per base) should be nearly identical within the same sequence, but for DNA samples, errors due to PCR (<0.01% per base) are even different. I will. Prepared by various enzymes. This is because the NGS error is too high to identify the true variant with a lower variant frequency and to embed the true variant to characterize only the NGS error. In this experiment, two DNA samples were prepared with Q5 polymerase (NEB) and KAPA polymerase (KAPA Biosystems) to perform NGS. I also analyzed variant correlations at all sequence

positions (n = 261) between samples before and after NGS error validation. For raw data without NGS error validation, the error rate was similar (R^2 = 0.88) following the same sequence. However, after NGS error validation, the error rates were not correlated (R^2 = 0.36).

In the NGS results, the most frequently occurring variant call was a “G” insertion error at position 173 near the homopolymer sequence of “GGG”. However, I have confirmed that the 216 insertion errors at this position are artifacts, except for a single variant of the "G" to "A" substitution. From the NGS results, 1879 substitutions (49.93% of total substitution errors) and 3571 indels (24.97% of total indels) were selected for analysis of PCR error types. Upon validation, there were 235 substitutions and a true variant of 4 indels.

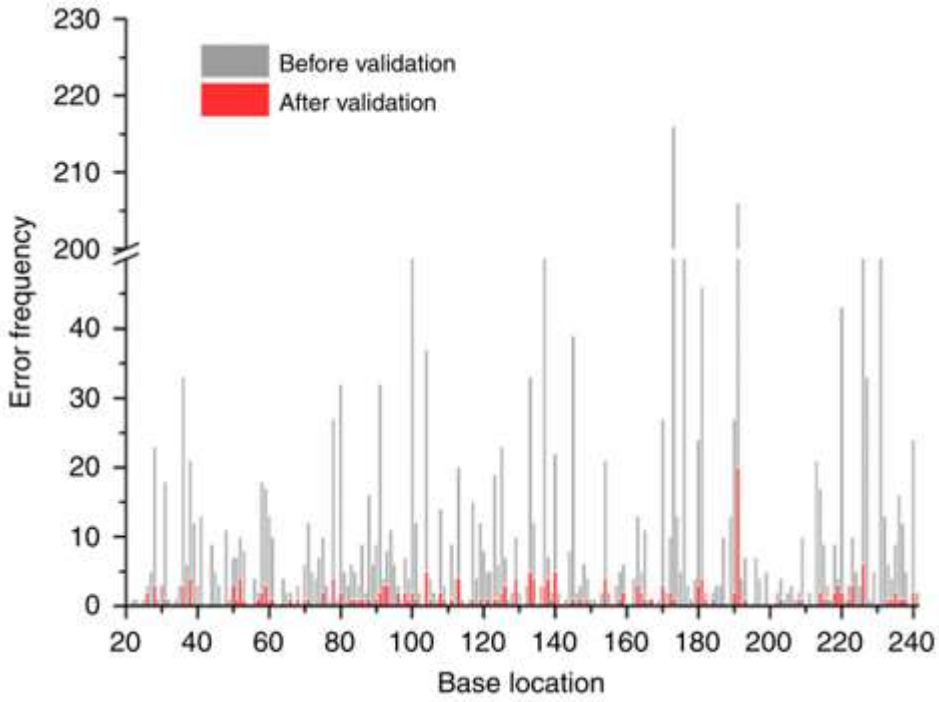


Figure 2.10 Distribution of NGS and PCR-induced errors in the emplate sequence.

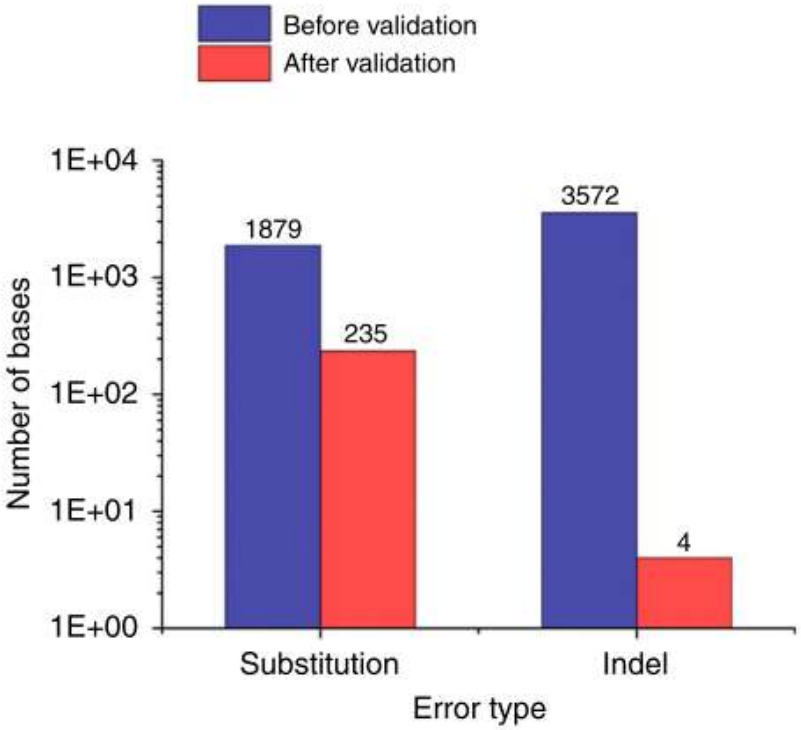


Figure 2.11 Verification of true variants according to error type.

In addition, there is a variant in the sequence result that I wanted to see if the base was read as error-free. Therefore, 700 DNA clones were randomly selected from a total of 904 error-free reads and extracted from the NGS substrate by laser retrieval system. As a result, it was confirmed that all DNA clones were error-free and that the DNA molecule had no mutation. Therefore, only true variants validated this way were used to calculate a PCR-induced error rate of 2.5×10^{-6} per base per doubling event. The calculated error rate values were correlated when compared to previous reports [29][18][32] in which the error rate introduced by the same polymerase (Phusion High Fidelity PCR Master Mix, NEB) was measured.

Other methods of measuring errors by PCR required 10 or more reads in the read family to generate consensus sequences and exclude NGS errors. However, this method is at least 10x more efficient in reducing the number of reads required, since NGS errors can be directly verified from the raw data after NGS is performed.

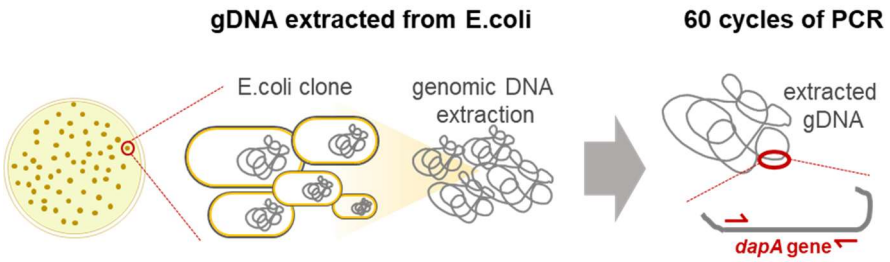


Figure 2.12 PCR induced template preparation process. After genomic DNA was extracted from *E.col,* dapA gene which is essential gene

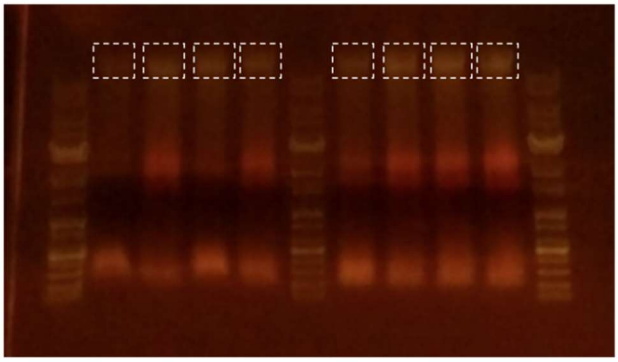


Figure 2.13 Gel electrophoresis image of the gDNA extracted from E.coli. The extracted gDNA was run on a 0.5 % agarose gel, followed by purifying gDNA from the gel.

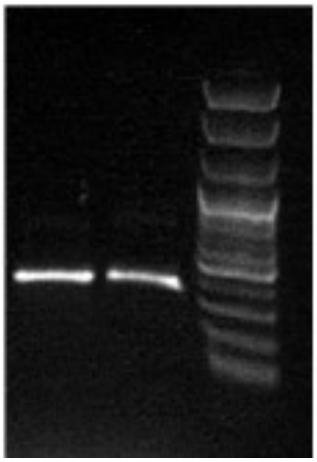


Figure 2.14 The purified gDNA was amplified through PCR with the primer for 1step.

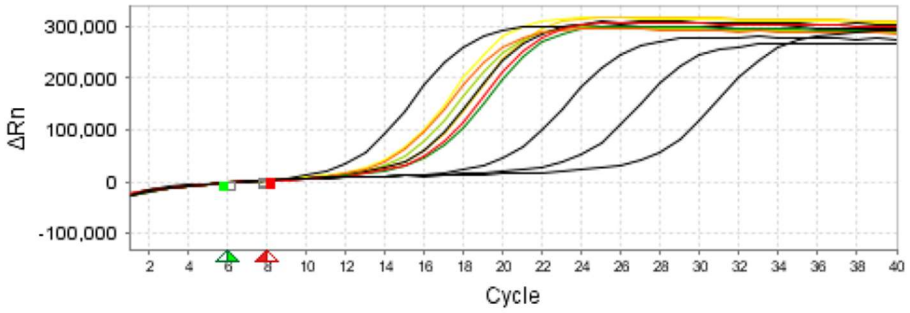


Figure 2.15 Amplification plot of real-time qPCR (Applied Biosystems, 7500 fast) with the initial gDNA template before PCR amplification. (520,549 copies of gDNA) (Black line: reference DNA template of $10^3, 10^4$, $10^5, 10^6, and\ 10^7$ copies, others: replicates of gDNA sample)

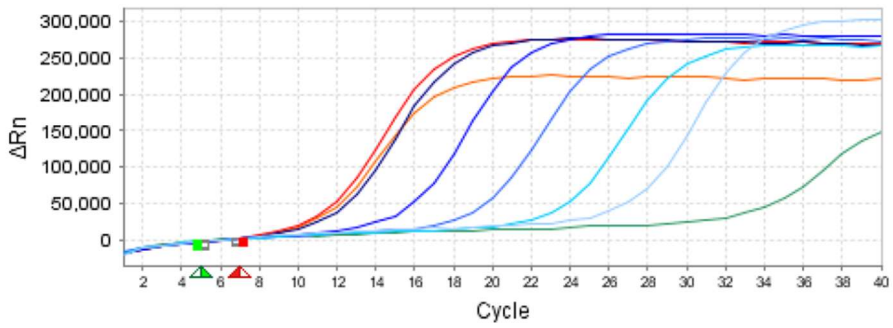


Figure 2.16 Amplification plot of real-time qPCR with the diluted (two times of 3/10000 and 1/100) gDNA copies after PCR (3,943,948 copies of gDNA), (Blue line: reference DNA template of $10^3, 10^4$, $10^5, 10^6, and\ 10^7$ copies, others: replicates of gDNA sample) Considering dilution and measured copies through qPCR, the gDNA was duplicated as 43 doublings.

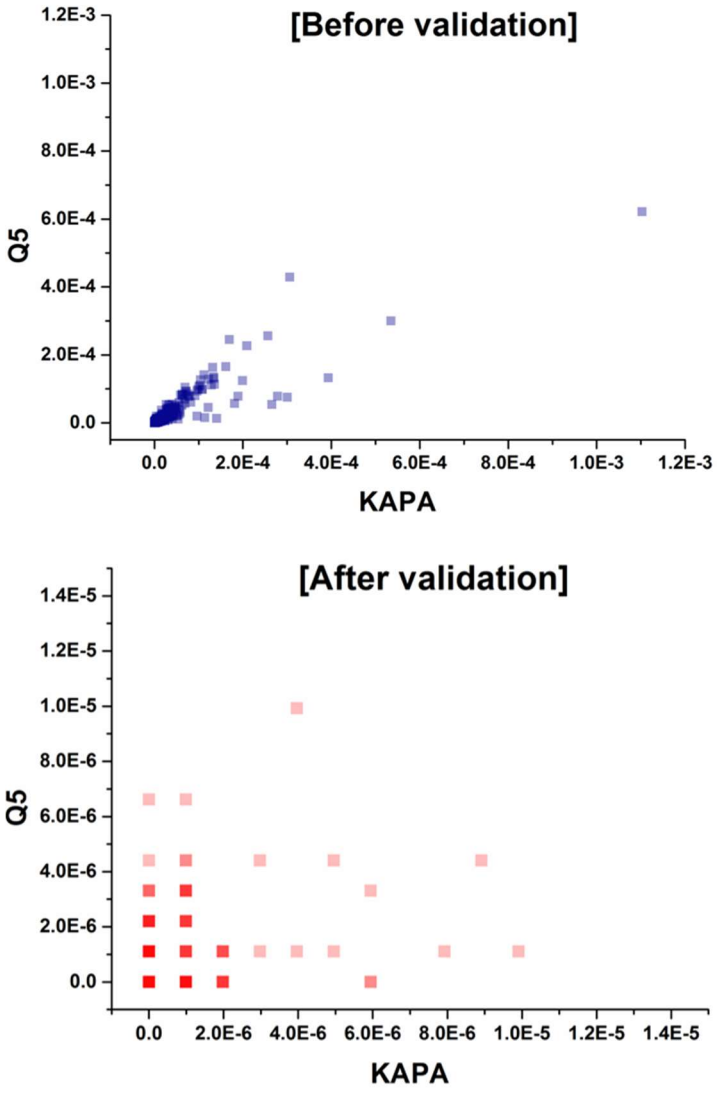


Figure 2.17 Correlation analysis for DNA sample variant rates (per base) prepared by two different DNA polymerases (KAPA and Q5 DNA polymerase).

2.5 Distinguishing PCR-induced error from NGS error

Filtered by barcodeless NGS error validation according to Q-scores above 10, 20, and 30 to see if not only NGS errors but also true variants of interest can be removed for quality control of raw data. Observed the processed variants. NGS is the result of PCR-induced errors prepared by three polymerases (Phusion, KAPA, and Q5 DNA polymerase) with more than 0.01% of the true per-base substitutions. The NGS results were filtered with the FATX toolkit quality filter to trim each NGS reading with an average Q score of less than 10, 20, and 30. Counted filtered total reads and variant calls and verified the amount of true variants that can be trimmed using the NGS verification method without barcode sequence.

As a result of Phusion polymerase, filtering using the highest quality threshold (> Q30) excluded ~ 60.2% of the true variants obtained with> Q10. That is, only 99 of the 249 true variants were identified. Furthermore, for KAPA and Q5 polymerase, the true variant was trimmed by 36.2% and 14.2%, respectively.

The number of actual variants when the quality threshold is

increased for a closer look at the quality control effect. Quality control was applied using the "p50" option. That is, if 50% of the bases have a quality score above the quality threshold, a sequence read is made. Testing confirmed that true variants started to decrease when the filtering Q-score threshold was 18, and were most reduced when the score was 24. These results show that quality control with Q-scores can lead to the loss of rare variants, especially for> Q20. In addition, the'p50 'option is usually not the choice taken to filter poor quality readings, so more data loss will occur under normal quality control situations where the'p100' option is applied.

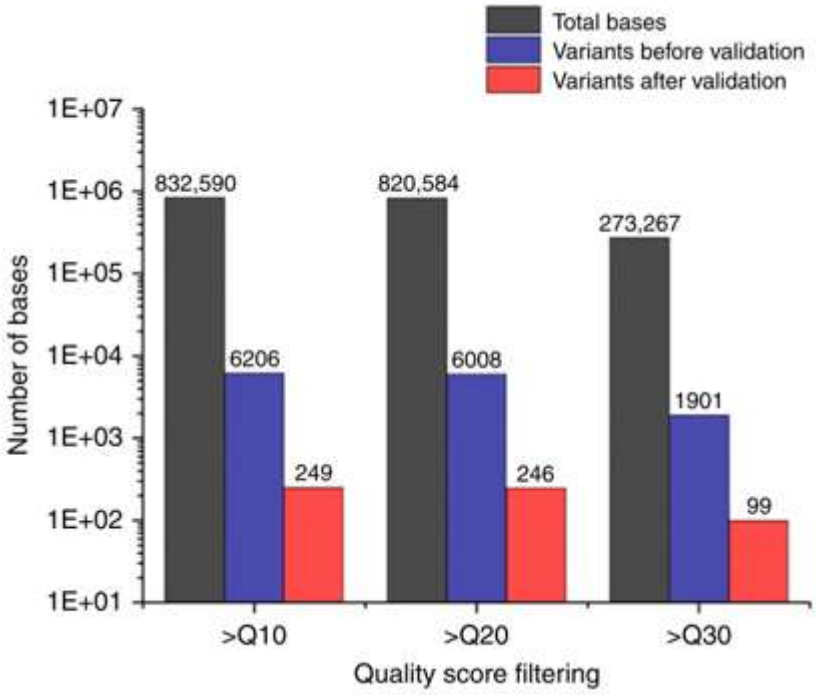


Figure 2.18 Comparison of variants after filtering raw data with the Q-score (>Q10, >Q20, and >Q30). Variants before the NGS error validation were reduced (3.26 times less) more than variants after validation (2.48 times less), which might include low Q-score of variant calls. Filtered variants according to Q-score threshold of the DNA template prepared by Phusion polymerase.

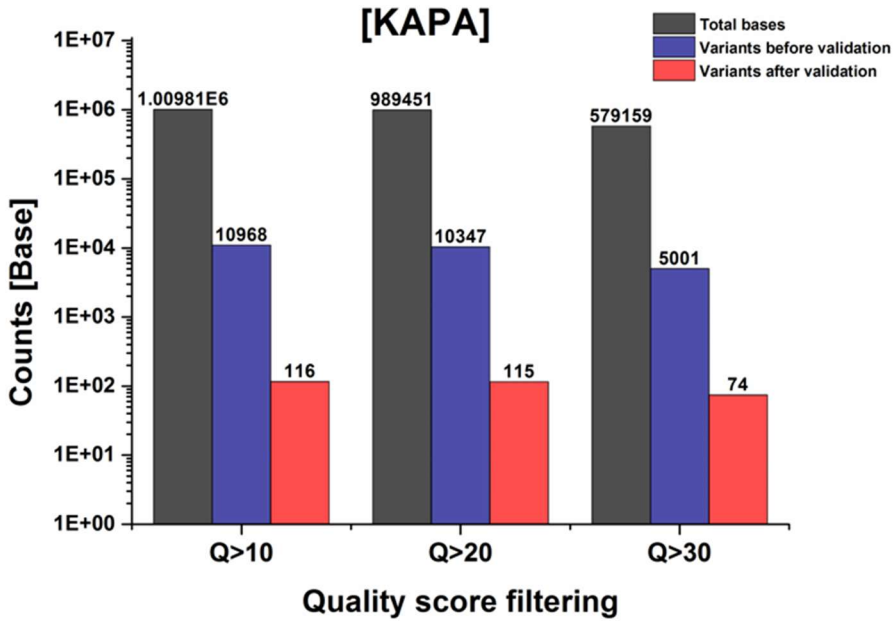


Figure 2.19 Identification of true variants after trimming raw data with the Q-score threshold. Filtered variants according to Q-score threshold of the DNA template prepared by KAPA polymerase.

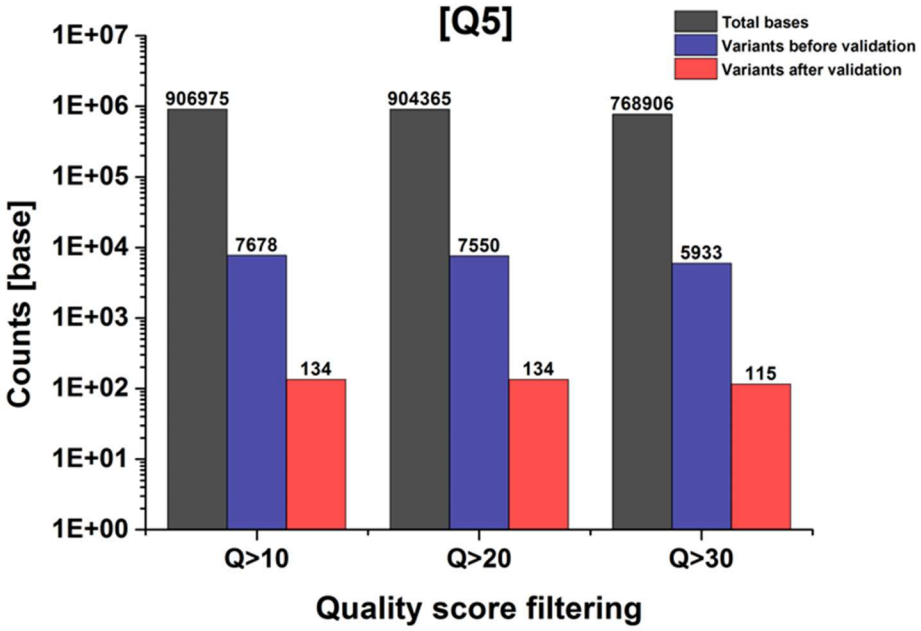


Figure 2.20 Identification of true variants after trimming raw data with the Q-score threshold. Filtered variants according to Q-score threshold of the DNA template prepared by Q5 polymerase.

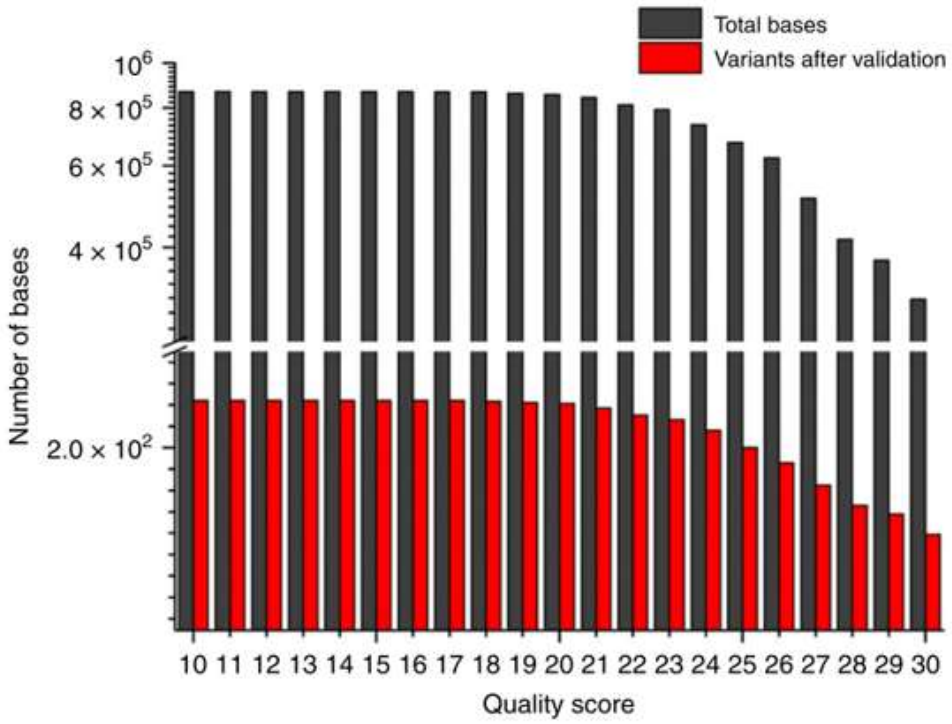


Figure 2.21 Reduction of true variants by quality control from >Q10 to >Q30. From Q- score over than 18, the true variants were confirmed to decrease by the barcode-free NGS error validation.

Chapter 3.

Circulating tumor DNA analysis

3.1. Introduction to tumor variant analysis

3.1.1. Introduction to circulating tumor DNA

Cell-free DNA (cfDNA) is released normal cells such as leukocytes and circulates freely in the bloodstream which is not necessarily from tumor origin. On the other hand, circulating tumor DNA (ctDNA) is tumor-derived fragmented DNA in the bloodstream that is not associated with cells. In cancer patients, ctDNAs represent a variable fraction of cfDNAs (ranging from 0.01% to more than 50%). Some studies have hypothesized that ctDNA is produced through the release of nucleic acids during cancer cell apoptosis or necrosis, or from tumor-derived exosomes. Certain genetic variations in cancer cells may reflect the patient's physical condition and treatment

response. Detection of DNA containing tumor-specific mutations in the peripheral blood of patients with malignant tumors may help identify dynamic changes in cancer cells. The content of ctDNA varies by tumor type and stage, and the mutation profile of an individual tumor can vary from patient to patient.

The ctDNA was extracted from plasma and usually analyzed by targeted deep sequencing. The targeted deep sequencing simultaneously uncovers new somatic mutations in genomic regions or many genes, both through a specific pure next-generation sequencing (NGS) approach and a combination of PCR and NGS. PCR-based targeted deep sequencing are tagged-amplicon deep sequencing (TAmSeq), the Safe Sequencing System (SafeSeqS) and CAncer Personalized Profiling by deep Sequencing (CAPP-Seq) [31] [33] [34].

TAm-Seq is useful for the de novo identification of rare cancer mutations and can detect cancer-specific changes with an allele frequency as low as 2%. SafeSeqS is a sequencing strategy that uses single molecule barcoding prior to PCR amplification to reduce sequencing errors and increase accuracy, with the sensitivity of

0.001%. This approach allowed us to detect single somatic mutations in ctDNA in patients with different stages of colorectal cancer using plasma samples obtained at different time points. Based on a different principle, CAPP-seq focuses on the detection and quantification of ctDNA by a probe panel composed of biotinylated DNA oligonucleotides that target repetitive mutated regions. This is an effective method for enriching and quantifying ctDNA libraries with high specificity and very low detection limits.

3.1.2. Conventional tissue biopsy and analysis

Breast cancer is one of the most common cancers, 1,300,000 cases and 450,000 deaths each year worldwide. Clinically, this heterogeneous disease falls into three basic treatment groups. The estrogen receptor (ER) positive group is by far the most abundant and diverse, and there are several genomic tests that can help predict the outcome of ER1 patients undergoing endocrine therapy. The HER2 (also called ERBB2) amplification group is of great clinical success because of HER2's effective therapeutic target, which has led to a strong effort to characterize other DNA copy number

abnormalities. Triple-negative breast cancers (TNBCs, lacking expression of ER, progesterone receptor (PR) and HER2), also known as basal-like breast cancers, are a group with only chemotherapy options, and have an increased incidence in patients with germline BRCA1 mutations or of African ancestry [5].

The breast tumor has been obtained by a biopsy using a hollow needle. The hollow needle is used by the doctor to remove a piece of breast tissue from a suspicious area that is felt or identified by imaging. The needle can be attached to a spring-loaded tool that quickly moves the needle in and out of the tissue, or to a suction device that helps pull breast tissue into the needle.

From the tissue, DNA is extracted and can be sequenced by targeted NGS. The target gene is can be determined by cancer subtype. The mutated genes were significantly more diverse and recurrent in luminal A and luminal B tumors than within the basal-like and HER2-rich (HER2E) subtypes. However, the overall mutation rate was lowest in luminal A subtype and highest in the basallike and HER2E subtypes. The luminal A subtype harbthised the most significantly mutated genes, with the most frequent being

PIK3CA (45%), followed by MAP3K1, GATA3, TP53, CDH1 and MAP2K4 [35]. Lumen B cancer is most frequently TP53 and PIK3CA (29% each). Luminal tumor subtypes are basal-like with TP53 mutations occurring in 80% of cases, with the exception of PIK3CA (9%), where the majority of luminally significant mutated gene repertoires are absent or rarely present. It was in sharp contrast to cancer. HER2E subtypes that frequently amplify HER2 (80%) have high frequency of TP53 (72%) and PIK3CA (39%) mutations and PIK3R1 (4%).

Conventional method for analyzing rare variant frequency of ctDNA requires replicate reads to filter out NGS errors. Assume that the variant frequency is 0.1%, the reads for sampling the ctDNA among the normal DNA is required as much as 10,000 with 10 replicates reads to filter NGS error. Therefore, the total reads to identify true variants in ctDNA is 100,000 and are linearly increased according to the number of gene panels. Given that the gene panel consists of 100 of genes, the required NGS reads is required as much as approximately 10Gb. For the case of lower variant frequency, ~0.01%, it requires NGS reads 10 times more resulting in 100Gb

(Figure 3.1). Also, the total reads can be calculated as below.

$$\text{Total reads} = \frac{10}{\text{Variant Frequency}} \times \text{replicates} \times \text{the number of sites}$$

It can be translated that NGS analysis for rare variant frequency requires tremendous reads resulting high sequencing cost. In this regards, the developed method can lower the number of NGS to analyze ctDNA of rare variant frequency. Therefore it gives chance to study large cancer patient or healthy people cohort for early diagnosis with lower sequencing cost at least 10 times. However, the limitation in this developed NGS error validation method is based on 454 sequencing platform and needed to optimize to other NGS platform. This method to verify the systematic NGS errors is, intrinsically speaking, analyzing the physically isolated DNA clones from the NGS substrate after the sequencing procedure. The key strategy stems from the observation that the systematic NGS error is caused during the signal detection process, and the enzyme-induced error (e.g. misincorporation of nucleic acids or damage during sequencing process) can be filtered out. Therefore, that the systematic NGS error causing mechanisms should the same for different NGS platforms is an important factor in considering

applicability of this method to other NGS platforms. After confirming that what I am analyzing is common for all NGS, it is important to determine technical applicability of the laser-based isolation platform to be used in other NGS platforms. For technical implementation, it is necessary to physically separate DNA molecules from the NGS substrates that are different for each and every NGS platform. Accordingly, to show that this method is applicable to other NGS platforms, I assessed the systematic NGS error causing mechanisms of other NGS platforms and the technical feasibility of the laser-based DNA molecule isolation system.

I first assessed that our platform is applicable to other NGS platforms since the systematic NGS error causing mechanisms are the same as used in this study. When the systematic errors in other NGS platforms are occurred, the major molecules remain unchanged. This is because the errors occur during signal detection, which includes phasing noise, invalid signal intensity threshold, signal decay along the increasing cycle, signal cross-talk among DNA clusters, and overlap of emission frequency spectra [36]. Although the enzyme-induced errors during the sequencing methods (i.e.

sequencing by synthesis or sequencing by ligation) changes the sequence of the physical DNAs in molecular clones on the NGS substrate, they can be filtered out using simple computational tools. This is because there is an extremely low possibility of enzyme-induced error dominating and altering the signal at the position of the DNA clusters. According to these reasons, the systematic NGS error in other NGS platforms is caused by the same mechanism, which is signal misdetection in sequencing process, as that in the NGS platform we demonstrated in the manuscript. Therefore, the same principle can be applied to other NGS platforms and the errors can be verified through our approach.

Second, in terms of technical feasibility, it is necessary to determine if the DNA clones can be separated using the optical laser system on the different NGS substrates. Since NGS platforms have diverse and different substrates, we need to optimize the laser retrieval system according to each NGS platform. For example, laser ablation is only applicable to transparent NGS substrates because the laser cannot be focused in the inner part of an opaque substrate. The previous study in our group has demonstrated

isolation of DNA clusters from Illumina sequencing plate, which is transparent. In the previous study, although the laser system could isolate two DNA clusters within a single laser spot (10um), I succeeded in verifying their sequences. In this case, the bottleneck was the large spot size (>10um) of the focused nanosecond laser, and if we use picosecond or femtosecond pulse laser we can reduce the laser spot size. When the spot size is reduced, the accuracy will increase because the accuracy of the NGS error verification depends on the ability of the system to isolate exactly one desired DNA clone from the NGS substrate. In other words, if the optical laser system is able to accurately isolate single DNA clone from the NGS substrate, this platform could be applied to other NGS platforms that use transparent substrates with high accuracy. Therefore, this method can be applied to NGS platforms using transparent substrates such as Illumina’s.

	Total reads		
Variant frequency	1%	0.1%	0.01%
Randomly shearing	**1Gb**	**10Gb**	**100Gb**
UMI ligation	1Gb	10Gb	100Gb
Duplex sequencing	3Gb	30Gb	300Gb
Our method	**0.1Gb**	**1Gb**	**10Gb**

Figure 3.1NGS reads for ctDNA valriant analysis

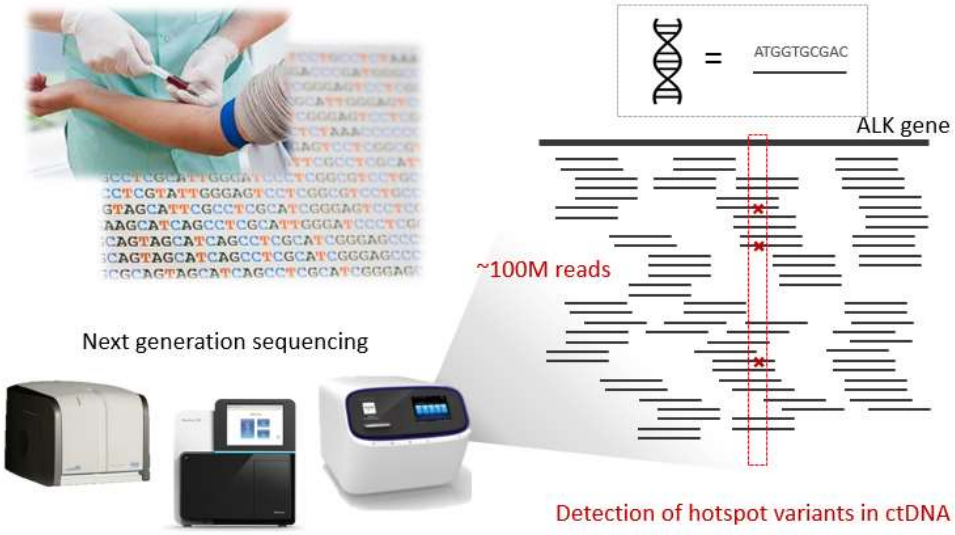


Figure 3.2 NGS enabled to detect hotspot variants in ctDNA. NGS utilizes DNA sequencing technologies that are capable of processing multiple DNA sequences in massively parallel.

3.2. Tissue biopsy and analysis for breast cancer

3.2.1. Cancer subtype information by pathological analysis

The breast tissue sample was obtained from the breast cancer patient during cancer surgery to remove the cancer tissue. The patient was diagnosed as stage 2 cancer and had luminal A subtype in the breast cancer validated from the pathological analysis of the tissue. Also, the metastasis was observed during the surgery.

3.2.2. Targeted deep sequencing

I determined 121 genes associated with breast cancer for the SNUH BCC (Seoul National University Hospital Breast Care Center Panel). The gene panel is based on the previous research [37], which the genes had a high frequency of repetitive mutations, genomic copy number amplification, deletion and altered expression in breast cancer samples. Also, the SNUH BCC panel is unique compared to other NGS-based cancer panels because it contains certain parts of the new breast cancer-related genes that are not found in other recent popular traditional cancer panels. In this regard, this SNUH

BCC panel not only targets breast cancer patients around the world, but is also ethnically directed to the Korean breasts Cancer patients for diagnostic and therapeutic prognosis.

In this research, I focused on the genetic variants on the selected 121 genes. To verify the somatic variants from tumor tissue, the reference sequence is required as baseline. The reference sequence can be obtained from the blood germ-line sample per patient. Therefore, the DNA was extracted from blood cells and analyzed by NGS and DNA extracted from cancer tissue also analyzed.

The NGS analysis pipeline is like below. NGS sequence read mapped to GRCh37 human reference genome using BWA-MEM (version 0.7.8) and default parameters. The resulting SAM file is sorted by chromosomal coordinates, followed by PCR duplicate marking using Picard (version 1.115) (Http://broadinstitute.github.io/picard/). The mapping quality scores less than 30 or mapping scores with complementary alignments were removed from the BAM file prior to further analysis. Before SNV detection, GATK (v3.5-0) IndelRealigner, I also used BaseRecalibrator to locally recalibrate the reading around the Indel

and recalibrate the base quality score of the BAM file. Then, GATK UnifiedGenotyper was used with default parameters followed by GATK VariantRecalibrator to obtain filtered variants.

As a result, the variants were called in intron region, which is not related to genetic mutation and oncogene expression.

MTOR	IKBKE	SETD2	FBXW7	IGF2R	FGFR1
EPHA2	PARP1	MST1R	MAP3K1	EGFR	PRKDC
ARID1A	AKT3	EPHA3	PIK3R1	CDK6	MYC
PIK3R3	ALK	POLQ	APC	PIK3CG	PTK2
JAK1	SF3B1	ATR	PDGFRB	MET	JAK2
NOTCH2	IDH1	**PIK3CA**	FGFR4	BRAF	CDKN2A
MCL1	ERBB4	FGFR3	FLT4	EZH2	CDKN2B
DDR2	FANCD2	PDGFRA	DDR1	KMT2C	SYK
ABL2	VHL	KIT	ROS1	ZNF703	TLR4
KDM5B	CTNNB1	INPP4B	ESR1	EIF4EBP1	ABL1

NOTCH1	C11orf30	MDM2	LTK	NCOR1	NOTCH3	AR
GATA3	PAK1	NAV3	IDH2	NF1	JAK3	
RET	MRE11A	TBX3	IGF1R	CDK12	CCNE1	
PTEN	ATM	FLT3	TSC2	ERBB2	AKT2	
FGFR2	FOXM1	BRCA2	CBFB	TOP2A	SRC	
MEN1	CDKN1B	RB1	CTCF	BRCA1	AURKA	
MALAT1	KRAS	IRS2	CDH1	RPS6KB1	GNAS	
CCND1	KMT2D	FOXA1	TP53	STK11	PTK6	
FGF4	ERBB3	MAP4K5	AURKB	INSR	RUNX1	
FGF3	CDK4	AKT1	MAP2K4	TYK2	EP300	

Figure 3.3 121 genes associated with breast cancer

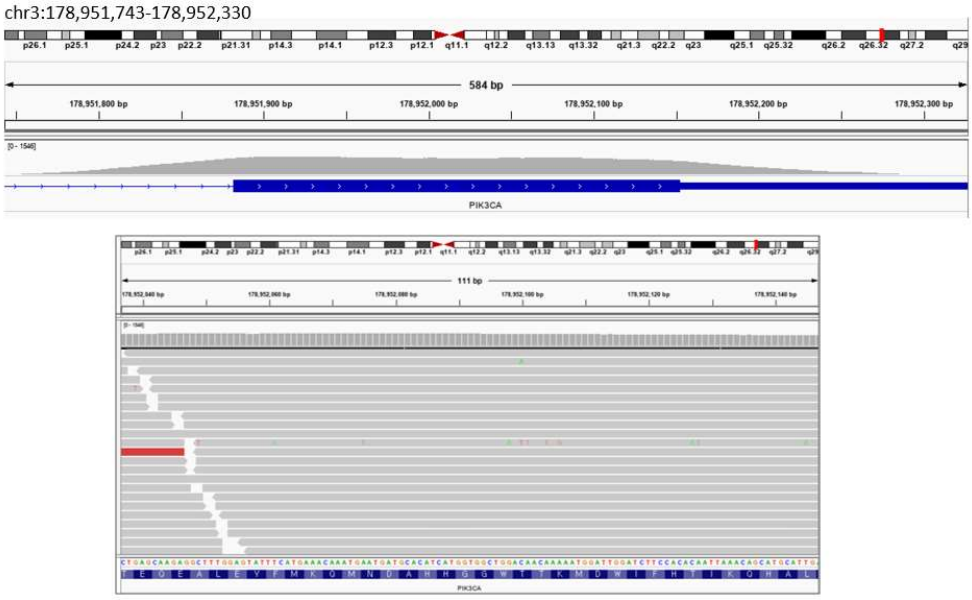


Figure 3.4 The NGS sequence reads were mapped to the GRCh37 human reference genome using BWA-MEM.

chrom	position	ref	var	normal_reads1	normal_reads2	normal_var_freq
3	47129486	T	C	36	1	2.70%
3	142215151	T	C	40	2	4.76%
3	142231027	G	A	15	0	0%

Figure 3.5 Variants in intron region.

3.3. Circulating tumor analysis by next generation sequencing error validation

3.3.1. Sample preparation from cfDNA extraction to NGS preparation

The main attraction of ctDNA analysis is that it is extracted non-invasively by blood sampling. Obtaining cfDNA or ctDNA typically requires the collection of approximately 3 mL of blood in an EDTA-coated tube. The use of EDTA is important for reducing blood clotting. The plasma and serum fractions of blood can be separated by a centrifugation step. ctDNA or cfDNA can then be extracted from these fractions. Serum tends to have high levels of cfDNA, mainly due to DNA from lymphocytes. High levels of contaminating cfDNA are not optimal as they can reduce the sensitivity of ctDNA detection. Therefore, the majority of studies use plasma for ctDNA separation. The plasma is then processed again by centrifugation to remove residual intact blood cells. The supernatant is used for DNA extraction and can be performed using commercial kit (QIAamp DNA Mini Blood kit, Qiagen).

In this experiment, PIK3CA gene was targeted to analyze somatic

variant. Therefore, the primer was designed for targeting the specific location of hot spot mutation in the PIK3CA gene. In order to verify whether the primer can hybridize and amplify only the target gene specifically, the randomly sheared genomic DNA (gDNA) extracted from HL60 cell line was used first to test it. The extracted gDNA was randomly sheared for targeting 150bp length through sonication to mimic ctDNA. Then, PCR amplification was conducted with the sheared gDNA and the designed primer set.

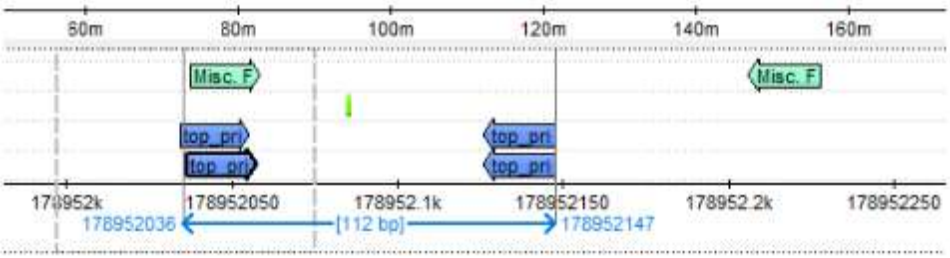


Figure 3.6 Primer design for targeting somatic variant region in PIK3CA gene.

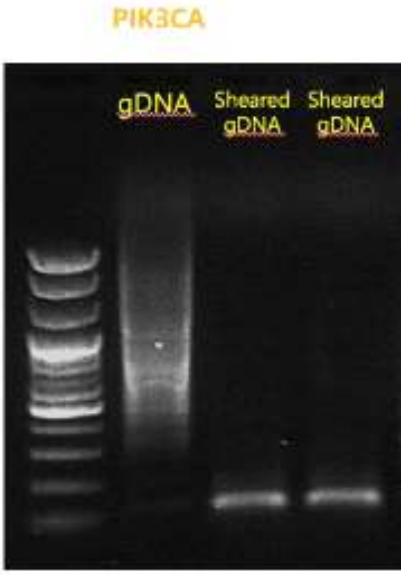


Figure 3.7 Gel electrophoresis result of 35 cycle PCR product.

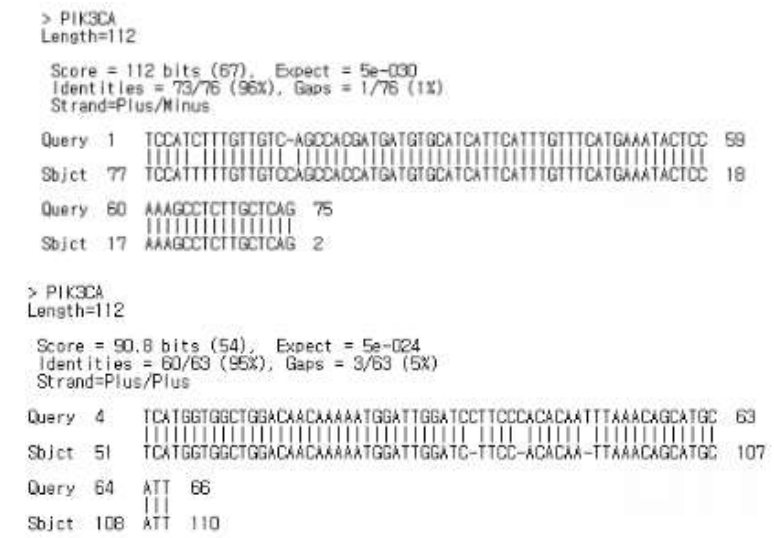


Figure 3.8 Sanger sequencing result of PCR product. (left : sequencing with forward primer, right : sequencing with reverse primer)

The verified primer set for PIK3CA gene was applied to the ctDNA extracted plasma. However, since ctDNA was captured through PCR, it is needed to consider how many sequences can be lost during the PCR. The designed primer was tested to be hybridized with ctDNA efficiently in early cycle. I prepared target sequence amplicon(112bp) and random sheared gDNA(~150bp) extracted from HL60 cells, and compare their calculated initial template copies measured in qPCR. Although the coverage of which ctDNA can be captured was not perfect, its average deviation of concentration might be under 15%.

sample	Ct	copy number	expected copy number	coverage
ref5	19.28	1.29E+05	1.29E+05	
ref4	23.61	1.29E+04	1.29E+04	
ref3	26.12	1.29E+03	1.29E+03	
ref2	29.92	1.29E+02	1.29E+02	
sheared gDNA) 6*10^2 #1	27.72	521.6145172	6.00E+02	86.94%
sheared gDNA) 6*10^2 #2	27.82	488.152981	6.00E+02	81.36%
sheared gDNA) 6*10^2 new_#1	27.35	666.6324485	6.00E+02	111%
sheared gDNA) 6*10^2 new_#2	27.35	666.6324485	6.00E+02	111%
sheared gDNA) 6*10^3 #1	23.16	10723.67124	6.00E+03	179%
sheared gDNA) 6*10^3 #2	22.56	15962.68105	6.00E+03	266%
sheared gDNA) 1*10^3 #1	25.29	2612.384912	1.20E+03	218%
sheared gDNA) 1*10^3 #2	25.56	2184.208594	1.20E+03	182%

Figure 3.9 Calculation of DNA copies from qPCR result.

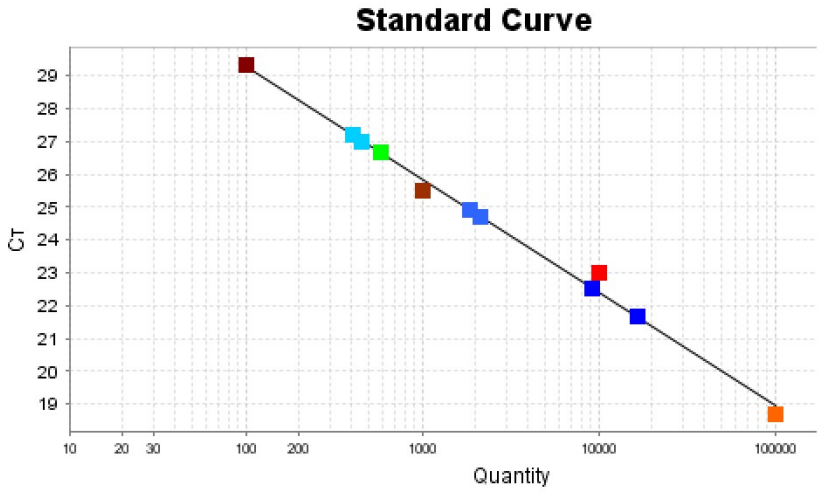


Figure 3.10 Amplification plot and standard curve from qPCR result.

With this optimized primer set for targeting PIK3CA gene, the ctDNA was captured and prepared for NGS sequencing. The preparation was consisted of three step of PCR to add sequencing adapter. The first step is PCR amplification for targeting the specific region in PIK3CA gene including hot spot mutation location. The second step is for attachment of NGS adapter, which is partial sequence to minimize hairpin structure to be hindered in PCR amplification. Then the third step is for final PCR amplification to construct completed NGS adapter in the both end sequence.

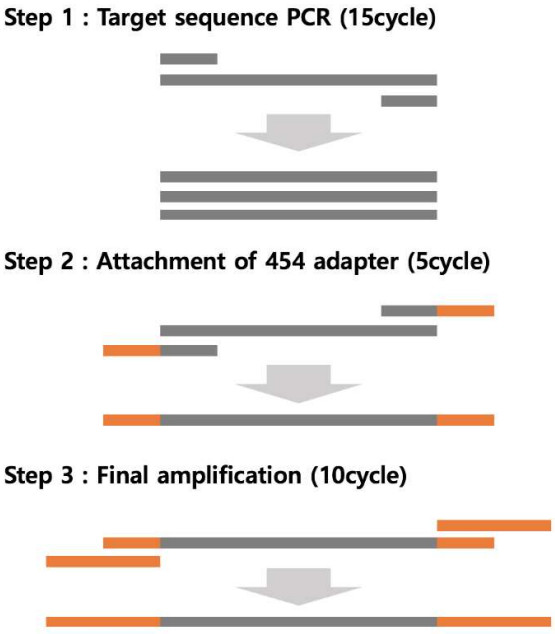


Figure 3.11 Three PCR step of sample preparation for NGS.

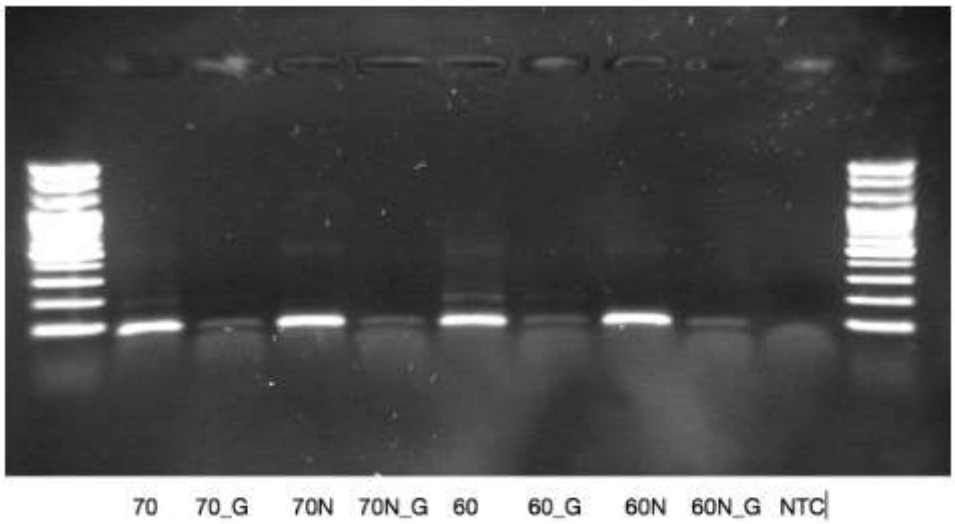


Figure 3.12 Gel electrophoresis result of PCR product after the 3rd step. (# : annealing temperature is #'C in 2nd step, #_G : annealing temperature is #'C in 2nd step and then gel-purified, N : NTC in 2nd step.)

3.3.2. Amplicon sequencing and sequencing error validation

The prepared sample was sequenced through 454 junior GS sequencing (100 cycles) according to the protocols of GS Junior from Roche 454 Life Sciences, 'emPCR Amplification Method Manual—Lib-L'. Hot spot mutation is detected in Chr3 : exon 20 (c.3140A →G and c.3140A→T), and, in this case, the corresponding position is 50th base.

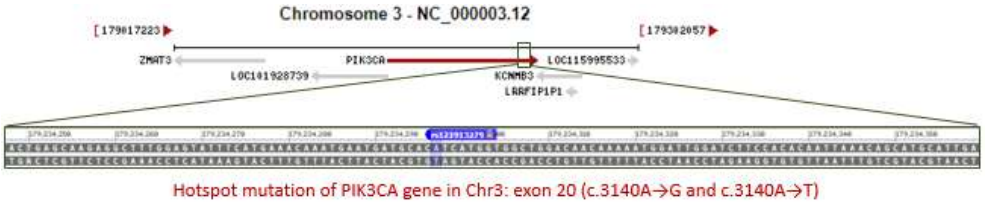


Figure 3.13 Hot spot mutation located in PIK3CA.

In 454 sequencing result, the total number of reads was 32,338. With this fastq file, for verifying true variants of interest, I extracted the information of all sequencing reads that had variant(s) (e.g. substitution, insertion, or deletion) or a few sequencing reads that had variant(s) at the desired position from BLAST results. With the

BLAST result, I observed the distribution of variants along the amplicon sequence. The result showed that the variants occurred more frequently in homopolymer sequences such as 'AAAA' or 'AAA', which is located at approximately 30^{th} base and 70^{th} base, respectively. However, the 50^{th} position was supposed to be hot spot mutation location for the breast cancer, and in this NGS result the number of variants counted at the position was two. Therefore, I isolated the DNA clones from the NGS substrate and validated whether the variant is true or artificial systematic error in sequencing process. The isolated DNA clones were amplified by PCR with the universal primer sequence and sequenced by Sanger sequencing.

As a result, the variants was validated as true which is not a systematic sequencing error. The variants were 'G' mutated from 'A'. This is related to the frequently occurred mutation in PIK3CA of the breast cancer oncogenes as published in previous researches.

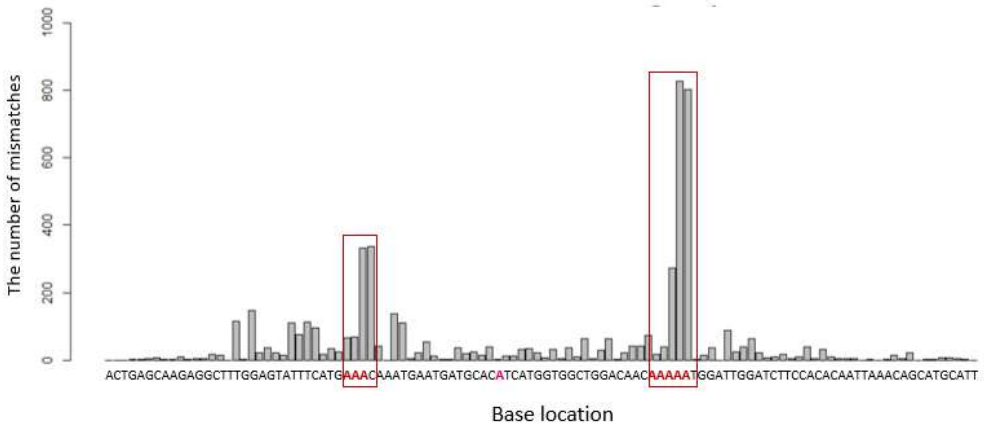


Figure 3.14 Amplicon of PIK3CA region sequencing.

917.0	1046.0> PIK3CA	49	T	C	KX7RI2S01CIS9K
925.0	3219.0> PIK3CA	49	T	C	KX7RI2S01CJJ75
928.0	1785.0> PIK3CA	49	T	C	KX7RI2S01CJSLN
939.0	725.0> PIK3CA	49	T	C	KX7RI2S01CKQJR
944.0	1724.0> PIK3CA	49	T	C	KX7RI2S01CK64E
962.0	3913.0> PIK3CA	49	T	C	KX7RI2S01CMTO7
966.0	2622.0> PIK3CA	49	T	C	KX7RI2S01CM5CG
995.0	2017.0> PIK3CA	49	T	C	KX7RI2S01CPOI7
996.0	3372.0> PIK3CA	49	T	C	KX7RI2S01CPSQM
1293.0	3953.0> PIK3CA	50	G	A	KX7RI2S01DFVUR
726.0	3906.0> PIK3CA	50	G	A	KX7RI2S01B13TG
1003.0	139.0> PIK3CA	51	A	T	KX7RI2S01CQCC9
1162.0	1517.0> PIK3CA	51	A	T	KX7RI2S01C4BX7
1209.0	1663.0> PIK3CA	51	A	T	KX7RI2S01C8GLT
1297.0	1414.0> PIK3CA	51	A	T	KX7RI2S01DF6JC
1331.0	197.0> PIK3CA	51	A	T	KX7RI2S01DI41Z
1435.0	1469.0> PIK3CA	51	A	T	KX7RI2S01DSAP7
1465.0	3271.0> PIK3CA	51	A	T	KX7RI2S01DUYXL
465.0	24.0> PIK3CA	51	A	T	KX7RI2S01BE3XM
484.0	333.0> PIK3CA	51	A	T	KX7RI2S01BGR7Z
554.0	185.0> PIK3CA	51	C	T	KX7RI2S01BMXCB

Figure 3.15 The variants in the 50th position in the amplicon sequence before NGS error correction.

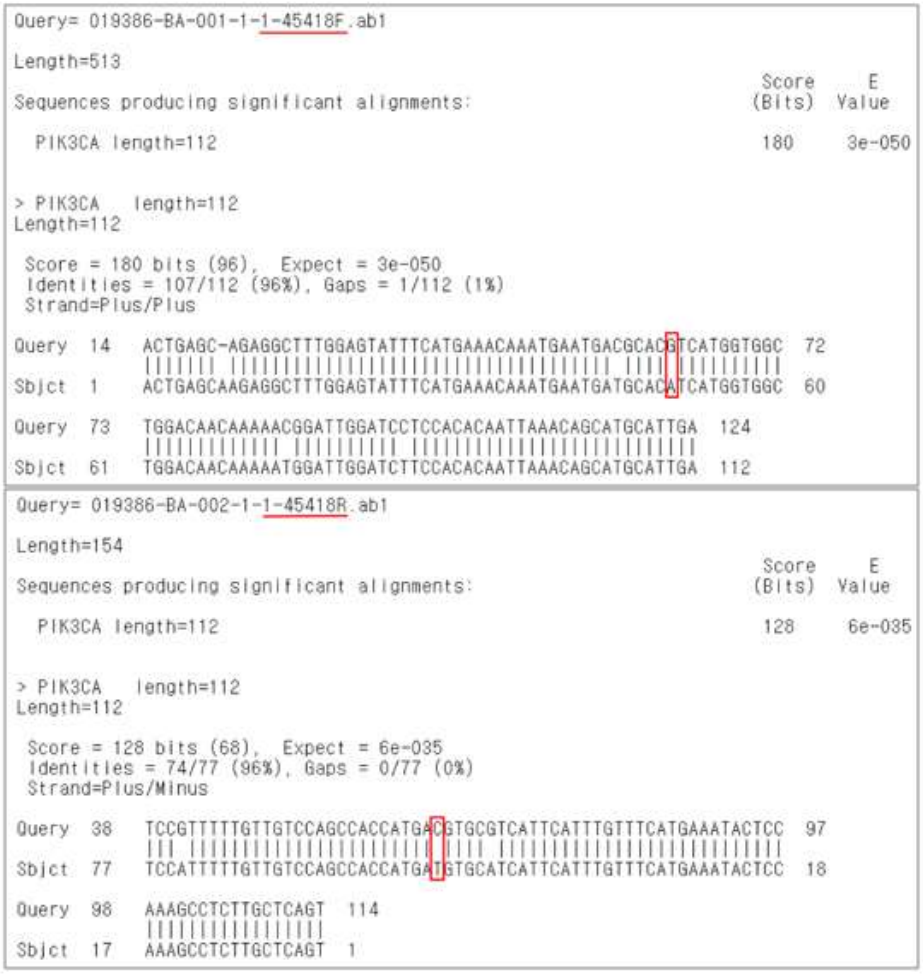


Figure 3.16 True variant was ensured by barcode-free NGS error validation.

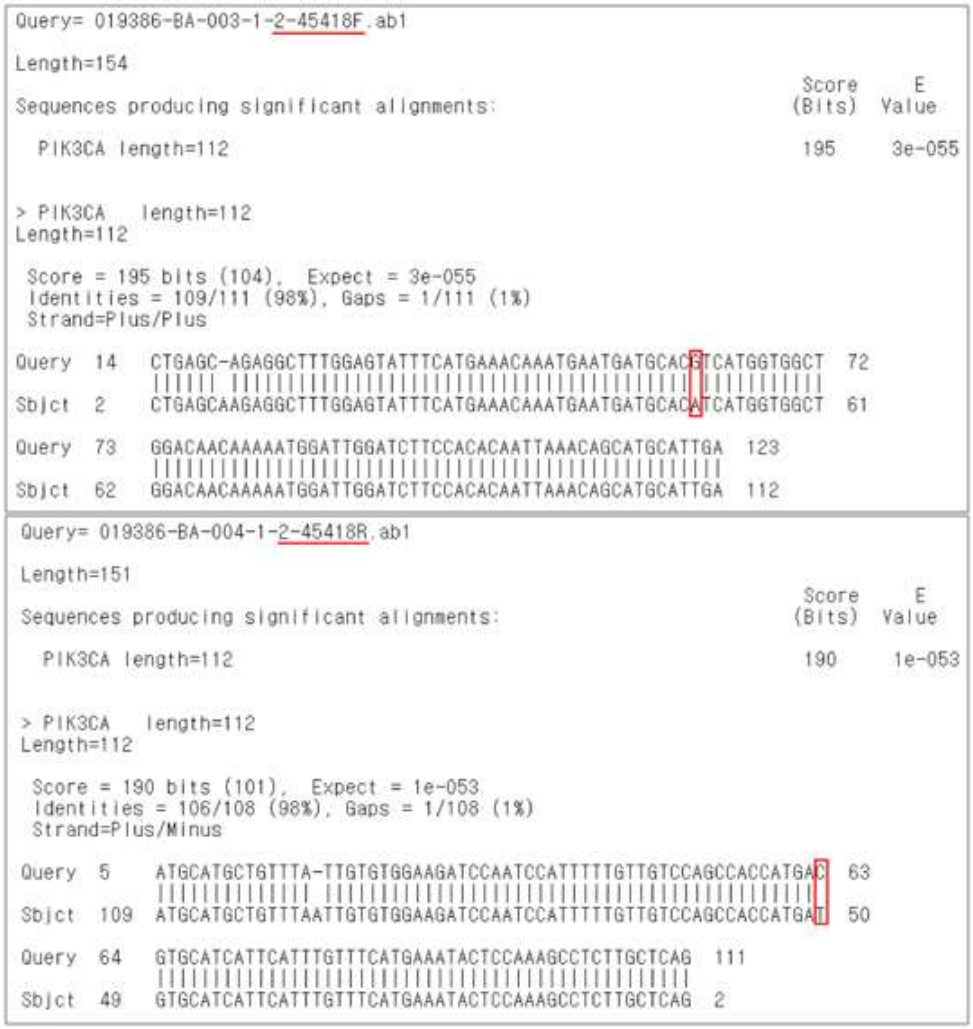


Figure 3.17 True variant was ensured by barcode-free NGS error validation.

Chapter 4.

Conclusion

In summary, I have developed a platform for directly inspecting falsely recalled base NGS errors from raw NGS data without the need for barcode sequence or quality control data processing. This method confirmed that true variants (more than 0.003% of VF) can be distinguished from NGS errors. In addition, previous studies characterized PCR-induced errors (per 2.5×10^{-6} bases per doubling) filled in by NGS errors (approximately 1% per base), at least 10x less than the number of base sequences used. This method avoids extra NGS sample preparation to distinguish NGS errors from actual variants. This can result in the loss of DNA samples during additional steps such as adding barcodes or DNA purification. In addition, this method allows detection of ultra-rare variants by preserving information on rare variant DNA copies from the original sample through quality control filtering of the entire raw NGS data.

This method can optionally be performed after performing NGS with selective read validation, which allows for selective validation of some NGS errors, reducing costs.

However, the number of variant sites analyzed and the number of leads containing the target site are important factors in determining the practicality of this method. This is because the cost of the verification sequence is proportional to the number of rare variant sites subject to verification and inversely proportional to the NGS error rate. Thus, this method is more effective when there are fewer variant sites with rare frequencies than those with a large number of variant sites. For example, this platform is effective in applications that quantify the allelic fraction of several mutation sites at infrequent frequencies. In particular, compared with the bar code system, this method is cost effective when the number of target variant sites is less than approximately 10,000 sites in a single analysis, given that state-of-the-art technology with an NGS error rate is 0.1% and a barcoding sequence depth is 10 (typically done at depth> 10). Also, if the NGS error rate decreases in the future, this method will be even more advantageous for validating more variants.

Therefore, this method can be used to study low-frequency ultra-rare mutants such as circulating tumor DNA or hotspot mutations in highly diverse samples.

Although this method has been demonstrated using certain types of NGS platforms, the basic principle of validating sequence errors by separating physical DNA from NGS sequencing substrates can be applied to other types of NGS platforms. This is because the root cause of NGS errors in both types of sequencing methods (ie, synthetic sequencing and ligation sequencing) occurs during signal detection itself and is not enzymatic (e.g. misincorporation of nucleic acids or damage during signal detection of sequencing process). Proper optimization of the separation technique, such as laser spot size optimization, is necessary for accurate separation of DNA clusters on the Illumina platform.

For the circulating tumor DNA analysis, the breast tissue sample was obtained from the breast cancer patient of stage 2. With the circulating tumor DNA extracted from the blood, the PIK3CA gene was targeted and the corresponding primer was designed to make ~100bp amplicon. The variants were counted nearby homopolymer

sequence before NGS error validation while the variants were removed after NGS error vadliation. In this experiments, the hotspot mutation of PIK3CA gene was detected as A to G at 50^{th} position in the amplicon sequence.

Bibliography

[1] O. Kim *et al.*, "Whole Genome Sequencing of Single Circulating Tumor Cells Isolated by Applying a Pulsed Laser to Cell-Capturing Microstructures," *Small*, vol. 15, no. 37, pp. 1-8, 2019.

[2] S. W. Song *et al.*, "One-Step Generation of a Drug-Releasing Hydrogel Microarray-On-A-Chip for Large-Scale Sequential Drug Combination Screening," *Adv. Sci.*, vol. 6, no. 3, 2019.

[3] M. Naghavi *et al.*, "Global, regional, and national age-sex specifc mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016," *Lancet*, vol. 390, no. 10100, pp. 1151-1210, 2017.

[4] A. C. Lee *et al.*, "OPENchip: An on-chip: In situ molecular profiling platform for gene expression analysis and oncogenic mutation detection in single circulating tumour cells," *Lab Chip*, vol. 20, no. 5, pp. 912-922, 2020.

[5] S. Alimirzaie, M. Bagherzadeh, and M. R. Akbari, "Liquid biopsy in breast cancer: A comprehensive review," *Clin. Genet.*, vol. 95, no. 6, pp. 643-660, 2019.

[6] S. Chang *et al.*, "A high-throughput single-clone phage fluorescence microwell immunoassay and laser-driven clonal retrieval system," *Biomolecules*, vol. 10, no. 4, pp. 1-12, 2020.

[7] H. Yeom *et al.*, "Cell-free bacteriophage genome synthesis using low cost sequence-verified array-synthesized oligonucleotides," *ACS Synth. Biol.*, 2020.

[8] G. M. Frampton *et al.*, "Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing," *Nat. Biotechnol.*, vol. 31, no. 11, pp. 1023-1031, 2013.

[9] Y. Choi *et al.*, "High information capacity DNA-based data storage with augmented encoding characters using degenerate bases," *Sci. Rep.*, vol. 9, no. 1, pp. 1-7, 2019.

[10] J. Noh *et al.*, "High-throughput retrieval of physical DNA for NGS-identifiable clones in phage display library," *MAbs*, vol. 11, no. 3, pp. 532-545, 2019.

[11] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing

technologies," *Nat Rev Genet*, vol. 17, no. 6, pp. 333–351, 2016.

[12] Y. Choi, H. Choi, A. C. Lee, and S. Kwon, "Design and synthesis of a reconfigurable DNA accordion rack," *J. Vis. Exp.*, vol. 2018, no. 138, pp. 1–12, 2018.

[13] H. Yeom *et al.*, "Barcode-free next-generation sequencing error validation for ultra-rare variant detection," *Nat. Commun.*, vol. 10, no. 1, pp. 1–8, 2019.

[14] O. Zagordi, R. Klein, M. Däumer, and N. Beerenwinkel, "Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies," *Nucleic Acids Res.*, vol. 38, no. 21, pp. 7400–7409, 2010.

[15] M. W. Schmitt, S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, and L. A. Loeb, "Detection of ultra-rare mutations by next-generation sequencing," vol. 2012, pp. 1–6, 2012.

[16] J. J. Salk, M. W. Schmitt, and L. A. Loeb, "Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations," *Nat. Rev. Genet.*, vol. 19, no. 5, pp. 269–285, 2018.

[17] S. R. Kennedy *et al.*, "Detecting ultralow-frequency mutations by Duplex Sequencing," 2014.

[18] D. I. Lou, J. A. Hussmann, R. M. Mcbee, A. Acevedo, R. Andino, and W. H. Press, "High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing," *Proc. Natl. Acad. Sci.*, vol. 110, no. 49, pp. 19872–19877, 2013.

[19] M. T. Gregory *et al.*, "Targeted single molecule mutation detection with massively parallel sequencing," *Nucleic Acids Res.*, vol. 44, no. 3, pp. 1–11, 2015.

[20] K. Nakamura *et al.*, "Sequence-specific error profile of Illumina sequencers," *Nucleic Acids Res.*, vol. 39, no. 13, 2011.

[21] A. Gilles, E. Meglécz, N. Pech, S. Ferreira, T. Malausa, and J. Martin, "Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing," 2011.

[22] S. Balzer, K. Malde, and I. Jonassen, "Systematic exploration of error sources in pyrosequencing flowgram data," *Bioinformatics*, vol. 27, no. 13, pp. 304–309, 2011.

[23] C. Ledergerber and C. Dessimoz, "Base-calling for next-generation sequencing platforms," vol. 12, no. 5, 2011.

[24] C. Endrullat, J. Glï ¿ ½kler, P. Franke, and M. Frohme, "Standardization and quality management in next-generation sequencing," *Appl. Transl. Genomics*, vol. 10, pp. 2–9, 2016.

[25] J. Reiss, M. Krawczak, M. Schloesser, M. Wagner, and D.

N. Cooper, "The effect of replication errors on the mismatch analysis of PCR-amplified DNA," *Nucleic Acids Res.*, vol. 18, no. 4, pp. 973-978, 1990.

[26] T. Kivioja *et al.*, "Counting absolute numbers of molecules using unique molecular identifiers," *Nat. Methods*, vol. 9, no. 1, pp. 72-74, 2012.

[27] P. Liao, G. A. Satten, and Y. J. Hu, "PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies," *Genet. Epidemiol.*, vol. 41, no. 5, pp. 375-387, 2017.

[28] S. Gerdes *et al.*, "Experimental determination and system level analysis of essential genes in Escherichia coli MG1655," *J. Bacteriol.*, vol. 185, no. 19, pp. 5673-5684, 2003.

[29] M. S. Hestand, J. Van Houdt, F. Cristofoli, and J. R. Vermeesch, "Polymerase specific error rates and profiles identified by single molecule sequencing," *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, vol. 784-785, pp. 39-45, 2016.

[30] M. Costello *et al.*, "Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation," *Nucleic Acids Res.*, vol. 41, no. 6, pp. 1-12, 2013.

[31] I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein, "Detection and quanti fi cation of rare mutations with massively parallel sequencing," vol. 108, no. 23, 2011.

[32] J. Cline, J. Braman, and H. Hogrefe, "PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases.," *Nucl. acids res.*, vol. 24, no. 18, pp. 3546-51, 1996.

[33] A. M. Newman *et al.*, "An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage."

[34] T. Forshew *et al.*, "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA," vol. 4, no. 136, 2012.

[35] D. C. Koboldt *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61-70, 2012.

[36] D. Laehnemann, A. Borkhardt, and A. C. McHardy, "Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction," *Brief. Bioinform.*, vol. 17, no. 1, pp. 154-179, 2016.

[37] S. Kim *et al.*, "PHLI-seq: constructing and visualizing cancer genomic maps in 3D by phenotype-based high-throughput laser-aided isolation and sequencing," *Genome Biol.*, vol. 19, no. 1,

pp. 1-17, 2018.

국문 초록

약 15년 전인 2003년에 인간 유전체 분석이 4조원이라는 천문학적인 비용을 들여 완성되었다. 그러나 2006년에 등장한 차세대 염기서열 분석 (Next Generation Sequencing, NGS) 기술을 시작으로 한 염기서열 분석 기기의 비약적인 발전은 현재 우리가 현재 우리가 100만원에 인간 유전체를 전부 분석할 수 있는 시대에 살 수 있게 해 주었다. 이는 기존 엄청난 분석 비용 때문에 제한적이었던 생명공학, 의학, 약학 등의 생물학적 연구에 혁명적인 발전을 가져오게 하였고, 오늘날에는 임상에서 질병의 진단과 처방을 위해서 사용되기 위한 단계에 있다.

그러나 NGS 분석이 임상에서 쓰이기에 대두되고 있는 문제는 높은 오류율로, 염기서열을 잘못 읽는 경우가 많다는 것이다. 이 문제는 특히 암 조기진단을 위한 DNA 변이 분석에 있어서 치명적이다. 암 발생 초기에는 일반세포에 비해서 매우 낮은 비율 (< 0.1%) 로 암세포가 존재하는데 비록 적은 비율로 존재할 지라도 시간이 지날수록 매우 왕성하게 분열하여 생체 내 조직을 파괴하기 때문에 조기에 이 암세포들을 발견하는 것이 중요하다. 그러나 현재는 NGS의 높은 오류율 (> 0.1%) 로 인하여 낮은 변이율의 DNA 변이 분석을 해야하는 경우,

NGS 분석결과에서 NGS 분석 오류와 DNA 변이의 구별이 불가능한 실정이다. 따라서, 낮은 변이율의 DNA 변이를 감지하기 위해서는 NGS 오류를 검증할 수 있는 기술이 필요하다. 이러한 NGS분석의 높은 오류율은 암 조기진단 이외에도 산모 내 태아 유전자 검사 (비침습성 산전 검사), 장기 이식 거부반응 검사 등과 같이 낮은 비율의 DNA 변이를 검사해야하는 경우에 걸림돌이 되고 있다.

이를 해결하기 위하여 NGS의 오류를 검증할 수 있는 기술을 개발하였다. 기술의 핵심은 NGS 오류가 염기서열 분석과정에서 DNA에 변이가 있는 것이 아니라 광학적 감지에 오류가 있는 것임을 착안한 것이다. 차세대 염기서열분석은 각 염기 (A, T, G, C) 에 빛을 내는 물질을 달아 광학적으로 감지해 내는 원리인데, 이 때에 광학적 감지 오류로 인하여 마치 돌연변이가 있는 것처럼 분석하게 되는 것이다. 이 원리를 바탕으로 NGS 분석에서 오류로 읽힌 DNA 분자들만을 레이저로 추출하여 복제 후에 NGS 분석 결과와 독립적으로 재분석하고자 하였다. 그 결과 NGS 분석결과에서는 DNA 변이로 분석되었으나 실질적으로는 NGS의 분석과정에서 생긴 광학적 감지 오류임을 밝힐 수 있었다. 본 방법을 통해 NGS 광학적 감지 오류를 정확하게 구별 해 냄으로써 최종적으로는 0.003%의 변이율을 가지는 DNA 변이까지 NGS 분석이 가능함을 보였다.

또한, 본 방법은 기존 NGS 오류를 검증하고자 하는 접근에서 벗어난 새로운 방법으로, NGS 기기 자체에서 정해지는 품질 점수 (Q-score)에 의존하는 기존 검증방법의 한계점을 극복하였다. 이 품질 점수는 NGS 기기 자체의 알고리즘에 의해 결정되는 것으로 NGS의 근본적인 오류를 검증하기에는 한계를 가진다. 하지만 본 방법은 레이저로 추출해 낸 DNA 분자를 다른 염기서열 분석 기기로 재분석할 수 있게 함으로써 염기서열 분석 품질 점수에 의존하지 않고 NGS 오류를 검증할 수 있다.

본 오류 검증방법을 통하여 실제 암환자의 혈액 내의 종양 유래 DNA를 분석함으로써 임상에 적용가능한지에 대한 실험을 검증하였다. 해당 환자는 유방암 2기의 환자로서 luminal A type의 subtype으로 진단된 환자였다. 따라서 환자의 암 특이적 변이를 확인하기 위하여 조직과 혈액에서 각각 NGS 분석을 실시하였다. 그 결과 조직 분석에서는 인트론 영역에서만 변이가 발견되었으며 따라서 유전자와 관련된 종양 특이적 변이는 발견되지 않았다. 혈액 분석을 위해서는, 혈액 10ml을 추출하여 플라즈마 분리 후 DNA만을 추출하였으며, 환자의 유방암 subtype인 PIK3CA 유전자에 대하여 변이를 분석하고자 하였다. 이를 위해 해당 유전자 특이적인 프라이머를 디자인 하여 PCR 증폭을 통해 샘플을 준비하였다. 해당 PCR 증폭물을 NGS

분석하였으며 그 결과 오류 검증 전에는 PIK3CA 유전자 염기서열 중 'A'가 반복되는 부분에 variant calling이 많이 발생하는 것을 확인하였으며, 암 특이적인 변이에 해당하는 염기서열 위치에서는 상대적으로 적은 개수의 variant calling이 나타난 것을 관찰하였다. 따라서 본 NGS 오류 검증방법으로 관심있는 영역인, PIK3CA의 암 특이적 변이 위치에 발생한 variant calling에 대하여 NGS 오류를 검증하고자 했다. 암 특이적 변이 위치에서는 총 2개의 variant가 calling 되었으며, 이에 해당하는 DNA 클론을 NGS 기판으로부터 분리하여 PCR 증폭 후 재분석 해보았다. 그 결과 해당 위치에서는 NGS 오류 없이 모두 실제 변이였음을 검증할 수 있었다.

주요어 : 차세대 염기서열 분석, 분석 오류 검증, 액체 생검,
학번 : 2014-21681