



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

향상된 NGS 분석 방법을 사용한
순환 종양 DNA 분석

**Analysis of circulating tumor DNA
by using NGS-based method with
enhanced analytical performance**

2018 년 02 월

서울대학교 대학원

의학과 의과학전공

박 가 희

A thesis of the Degree of Doctor of Philosophy

**Analysis of circulating tumor DNA
by using NGS-based method with
enhanced analytical performance**

**향상된 NGS 분석 방법을 사용한
순환 종양 DNA 분석**

February 2018

**The Department of Biomedical Sciences,
Seoul National University
College of Medicine
Gahee Park**

ABSTRACT

Introduction: Interrogation of circulating tumor (ct)DNA using next-generation sequencing (NGS)-based methods have been proposed as a way to track the dynamics of tumor in real time. However, there was no standard guideline for ctDNA sequencing that I have evaluated the procedure from end-to-end to propose the optimal analysis methods for ctDNA sequencing. Chapter 1* emphasizes the importance of the recovery of unique DNA molecule from the minimal amount of starting material. After that, the systematic evaluation of each step highlights the error-prone step in the sequencing process. In Chapter 2, the utility of ctDNA sequencing has evaluated through the monitoring of tumor genomic in multi-cancer samples.

Method: To maximize the recovery rate of unique DNA molecule, I approached the ligation step during the library preparation in sequencing protocol by optimizing the temperature, time and adapter concentration. Identification of technical errors was conducted with the comparison of background error distribution from the acoustically sheared germline DNA and naturally fragmented cell-free DNA. The utility of ctDNA sequencing analysis was assessed by comparing the standard protein biomarker and imaging changes during the patients' therapeutic intervention.

Results: The modified ligation conditions for the minimal amount of starting material able to increase the recovery rate of unique DNA molecule by 20% compared to the standard conditions. A comparison of the characteristic of

acoustically sheared gDNA and naturally fragmented cfDNA revealed that gDNA constituted with 64% of C: G> A: T and 39% of C: G> G: C substitution class changes. Through testing of the series of the mild sheared conditions, the reduction of error rate was observed with an average of 40%. Furthermore, the analysis of the vicinity at the ends of the DNA fragments revealed that A> G and A> T preferentially fragmented. The enhanced analytical performance in NGS method able to establish diagnostic utility with the detection sensitivity of 100% and specificity of 97.1% as applied to cancer plasma samples. The level of ctDNA was not only highly correlated with the therapeutic response but also showed an average of two months' earlier reaction than the standard protein biomarker and imaging changes. Finally, the determination of tumor heterogeneity was observed through ctDNA analysis, which was not discovered in the matched tumor biopsies.

Conclusions: Overall, the unique characterization of cfDNA could not only emphasize the underlying cause of technical errors but also demonstrate opportunities for early detection of cancer using NGS-based technology. Ultimately, the combined approach of ctDNA and NGS sequencing analysis is believed to address unmet needs in cancer research.

*The works published in Genome Biology(1) and Scientific Reports (2).

Keywords: Cancer genomics, liquid biopsy, circulating tumor DNA, cell-free DNA, next-generation sequencing, background error

Student number: 2012-21792

CONTENT

ABSTRACT	i
CONTENT	iii
LIST OF TABLES AND FIGURES	v
LIST OF ABBREVIATION.....	vii
GENERAL INTRODUCTION	1
Cell-free DNA	3
Circulating tumor DNA	4
Current detection methods for ctDNA	4
Digital PCR.....	4
Next generation sequencing.....	5
NGS-based ctDNA analysis	7
Potential misdiagnosis from background errors	8
CHAPTER1	1 4
Practical guidelines for cell-free DNA analysis using enhanced analytical performance of NGS-based method	
INTRODUCTION.....	1 5
MATERIALS AND METHODS.....	1 8
RESULTS.....	2 2
Comparison of blood collection tubes.....	2 2
Optimization of library preparation	2 2
Optimizing statistical modeling for cfDNA analysis	2 4
Performance of optimized TDS on cfDNA and PBL DNA..	2 4
Estimation of errors derived by TDS	2 5
From sequencing reaction	2 5
Distribution of background errors	2 5
Sample preparation caused background errors	2 6
Breakpoint preferences.....	2 7

Multi-statistical adjustment for removing the background errors	2 9
DISCUSSION	3 1
CHAPTER 2	5 1
Ultrasensitive interrogation of circulating tumor DNA from cancer patients using enhanced analytical performance of the NGS-based method	
INTRODUCTION	5 2
MATERIALS AND METHODS	5 4
RESULTS	6 1
Evaluation of LOD with single mutation	6 1
KRAS mutations	6 1
Evaluation of LOD with multi-mutations	6 1
“With primary” mutation	6 1
Biopsy-free manner	6 2
Monitoring tumor burden by measuring ctDNA	6 3
Diagnostic utility	6 5
DISCUSSION	6 7
GENERAL DISCUSSION	9 4
REFERENCES	9 5

LIST OF TABLES AND FIGURES

Introduction

Figure 1 Characteristic of cell-free DNA	9
Figure 2 General ctDNA analysis schematic flow.....	10
Figure 3 General process of capture-based targeted deep sequencing	11
Figure 4 Schematic flow of ctDNA analysis using NGS-based technology	12

CHAPTER 1

Figure 1-1. Performance of cfDNA sequencing	35
Figure 1-2. Quality score of read bases in targeted deep sequencing data	36
Figure 1-3. The distribution of background errors from PBL and plasma DNA	37
Figure 1-4. Alleviation of background error by various condition of fragmentation.....	38
Figure 1-5. The fragment size distribution from PBL and plasma DNA.....	39
Figure 1-6. Evaluation of read bases from the start position.....	40
Figure 1-7. Nucleotides around the DNA breakpoint.....	41
Figure 1-8. Nucleotides around the DNA breakpoint.....	42
Figure 1-9. Frequencies of dinucleotide	43
Figure 1-10. Combination of 16 dinucleotide frequencies	44
Figure 1-11. Allele frequency of background errors from hotspot mutations	45
Table 1-1A Total amount of plasma DNA collected from Streck BCT and EDTA tube.....	46
Table 1-1B The number of genomic variants detected from Streck BCT and EDTA tube.....	47
Table 1-2 The total amount of DNA yield was compared under different ligation condition.....	48
Table 1-3 Evaluation of open-source tools and statistical analysis using	

spike-in controls.	49
Table 1-4 Performance of multi-statistical analysis for ctDNA sequencing	50

CHAPTER 2

Figure 2-1. The correlation of harbored KRAS mutations using digital PCR and enhance NGS-method from pancreatic cancer patients	70
Figure 2-2. Tumor mutations in pre-treatment cfDNA samples from 17 PDAC patients.....	71
Figure 2-3. Monitoring of ctDNA PDAC patients under therapeutic intervention.....	72
Figure 2-4. Summary of plasma mutations determined by “biopsy-free manner”	73
Figure 2-5. Allelic fraction of ctDNA and CA19-9 level depending on therapy responses	74
Figure 2-6. The number of mutations in plasma DNA.....	75
Figure 2-7. Distribution of detected genes from pleural effusion fluid and plasma DNA.....	76
Figure 2-8. The differences of allele frequencies from pleural effusion fluid and plasma DNA	77
Figure 2-9. The size distribution of pleural effusion and plasma DNA	78
Table 2-1. The limits of detection sensitivity evaluated by KRAS mutations in 14 PDAC patients.....	79
Table 2-2. Determined mutations from 17 FNA samples.....	80
Table 2-3. Evaluation of FNA mutations in baseline plasma DNA samples	81
Table 2-4. The performance of droplet digital PCR in plasma samples	82
Table 2-5. The list of somatic mutations detected in plasma samples by biopsy-free manner.....	83

LIST OF ABBREVIATION

8-oxo-G: 8-oxo-7,8-dihydroguanine

ANOVA: Analysis of variance

AP site: Apurinic-apyrimidinic site

BEAM: Beads, emulsions, amplification, and magnetics

bp: Base pair

CA: cancer antigen

CR: Complete Response

CT: Computed Tomography

CTC: Circulating tumor cell

cfDNA: Cell-free DNA

ctDNA: Circulating tumor DNA

DNA: Deoxyribonucleic acid

Dx: Diagnosis

ddPCR: droplet digital PCR

EDTA: Ethylenediaminetetraacetic acid

EGFR: Epidermal growth factor receptor

ELISA: Enzyme-linked immunosorbent assay

EUS: endoscopic ultrasound

FDA: Food and drug administration

FNA: Fine needle aspiration

gDNA: germline DNA

KRAS: KRAS proto-oncogene, GTPase

LSD: Least significance difference

NSG: Next-generation sequencing

MAF: mutant allele frequency

miRNA: micro RNA

QC: Quality control

Q score: Phred quality scores

RNA: Ribonucleic acid

SD: Stable diseases (Medical terminology)

SD: Standard deviation (Statistical terminology)

SNP: Single nucleotide polymorphisms

SNV: Single nucleotide variants

TCGA: The Cancer Genome Atlas

TP53: Tumor protein p53

PBL: Peripheral blood leukocyte

PCR: Polymerase chain reaction

PD: Progression of disease

PDAC: Pancreatic ductal adenocarcinoma

PE: Pleural effusion

PR: Partial Response

WES: Whole-exome sequencing

GENERAL INTRODUCTION

Cancer is a disease which contents uncontrollable manner of cells division and ultimately influences to nearby normal cells (3). It conquers the particular tissue and often takes a route of the blood vessel or lymph node to travel other parts of the tissue to expand the colony (4). An understanding of such a behavior revealed by comparative analysis of genomic differences in normal cells (5). The main point was cancer cells contains a fatal mistake in the DNA sequences also known as a mutation. Researchers started to target the protein which arises from the specific driver mutation to cure cancer. However, the targeted inhibitors turn out to reduce in a certain amount of time yet often the rise of the novel clones which contains the different types of mutation to evolve throughout the therapeutic intervention (6). Moreover, the characteristic of localized tumor tends to acquire similar driver mutations, but it often varies by the unique feature of individuals that the intra-heterogeneity causes the resistance of the drugs (7). To observe the unique intra-heterogeneity, the serial biopsy must be obtained to estimate the tumor growth throughout the treatment. This is near impossible due to an ethic problem and painful to patients (8).

One of the strategies to prevent the expansion of cancer cell is to detect cancer early as possible (9). The chance of success of

treatment and prevention of clonal expansion is much higher than cancer has already been metastasized and/or discovered in the late stage. Nonetheless, the procedure of tissue biopsy is done to late stage of patients, and it is often too late to eradicate the tumor mass. Therefore, there must be a start-up package with a benefit to detect fast and accurate tumor signal in the non-invasive method (10).

The computational tomography is one alternative method to detect the tumor in non-invasive manner. We now have a high resolution of computational tomography (CT) images to identify the smallest tumor. But, the cost is incredibly high, and the effect of radiation to the patient would be another side effect of increasing the chance of getting cancer. Another is a collection of blood sample from the patient and quantifies the level of according protein biomarker. Cancer antigen (CA) is a protein biomarker that related to specific types of cancer. If the level of cancer antigen is higher than the standard threshold, the assumption can be made. However, the level of protein biomarker also often miscorrelates due to the possibility of halt of molecular mechanism, some individuals have not express the certain types of protein biomarker, or the level varies on the individual's health condition (11). Recent studies suggest the alternatives of tumor biopsy or protein biomarker with the other types of resources (cancer-related exosomes, microRNA (miRNA), circulating tumor cells (CTCs), cell-

free DNA (cfDNA), and etc) can be not only collected from the plasma of blood but also from the body fluid. (10). The reason of using non-invasive biopsy collected from the blood or body fluid (hereafter, liquid biopsy) is, it allows to track the progression of the disease and figuring out the therapy response in the regular bases (10, 12). The candidates from the liquid biopsy have been evaluated with multiple types of the approach. Each of molecular biomarker candidates eliminated as their limited resources and due to the lack of the knowledge underlies the mechanisms. CTC was one of the revolutionary discovery in the cancer research. CTC claims to escape from the tumor mass but it is barely detectable in a resolution of analysis (13). Additionally, it is impossible to track in real-time. With all the dark histories, the liquid biopsy is back in business by cell-free DNA.

Cell-free DNA

The history of cfDNA began in 1948 discovered by Mendel and Metis (14). CfDNA is the naked DNA that floats in the body fluid or plasma of the blood with an average peak size of 166 base pair (bp) (Figure 1) (15). The origins of cfDNA hypothesized to be corresponded by the cell's apoptosis, necrosis, secretion, or combination of all due to its genome-wide size distribution (16). Moreover, the observation correlates to the nucleosome positioning space that it estimates to be the one wrap of chromosome (Figure 1).

As the cfDNA releases the genetic factors from the individual of cells, the analysis of cfDNA provides the vast of information not be limited to

detecting neoplastic diseases but also applicable to trauma, stroke, organ transplantation, prenatal screening for fetal aneuploidy, and etc (17, 18).

Circulating tumor DNA

Of course, cancer cells also leave out the trace to the blood stream. The cell-free DNA contains the genetic alteration is called circulating tumor DNA. In 1977, the evidence of the level of cfDNA in cancer patients represent higher than the healthy volunteers (19). Consequent results highlight the correlation of the amount of cfDNA with the existence of tumor mass(20). Hereafter, the approval of ctDNA analysis was done by applying the detection of TP53 (21) and KRAS mutations (22) in cancer patients.

However, the study of ctDNA revolutionized recently because of the lack of detection sensitivity with existed techniques. A critical fact of ctDNA is, it embedded by the massive amount of normal cfDNA. An ultrasensitive detection method is needed to detect ctDNA. Therefore, it was impossible to carry out further and walked on the spot decade ago.

Current detection methods for ctDNA

Digital PCR

Luckily, researchers realized the importance of improvement of analytical performance to implement the ultrasensitive detection methods. With a born of BEAMing (beads, emulsions, amplification, and magnetics) PCR technology (23), the capture of single DNA molecules has started the engine

about the discovering the variants with the lowest allele frequency (23). ctDNA analysis was proven as a cancer screening tool using the application of digital PCR. For instance, the food and drug administration (FDA) has tested two types of mutations (exon 19 deletion and/or L858R) that can be compensate by the tumor biopsy. The epidermal growth factor receptor (EGFR) mutations were evaluated in the non-small cell lung cancer patients. In brief, the patients who underwent the EGFR tyrosine kinase inhibitor (TKIs) tested to screen the rise of resistance mutations (24, 25) (Figure 2). Following that, the decision was made to treat with the inhibitor or otherwise.

The digital PCR application is limited to only those the patients who contain the known mutations. The chance of losing the novel signal comprises by the lack of understanding of underlying resistance mechanism and the limit of rest of patients.

Next generation sequencing

To get more information of resistance signals or tumor heterogeneity simultaneously, the implementation of genome-wide study is needed. Next-generation sequencing (NGS) technology allows analyzing the collection of genomic alterations in once that can be selected in interest target in a genome-wide region. There are two types of sequencing method: amplicon-based and capture-based sequencing. Amplicon-based sequencing amplifies the region of the target at the beginning of the sequencing and hybridizes the molecular barcode at the end. Simply, the broad range of PCR with the shorter size of

DNA fragments. On the other hand, capture-based sequencing shear the DNA at the beginning of the experiment and ligase the adapter sequences. After that, the customized RNA baits hybridize to the DNA. Therefore, it has broader and wider DNA sequences compare to amplicon-based sequences. Capture-based sequencing has known to have lower error rates than the amplicon-based sequencing. However, the methods can be exchangeable depends on the interest for cancer target.

In general, the process of targeted deep sequencing categorized in three steps (Figure 3): library preparation, target enrichment, and sequencing. The optimal experimental procedure may increase by the efficacy of the library preparation step. The step of library construction is important because the unique DNA molecule can be maximized or minimized by the optimized condition. There are three categories that impact on. Ligation, purification, and PCR amplification. In order to sustain the input DNA, the optimizing the library steps such as ligation step or adjusting the PCR cycles may help the recovery rate of DNA. Also, the higher amount of DNA has higher chance to bind the adapter that increases the recovery rate of initial input DNA molecules (2). However, there are several purification steps that the chance of losing the initial input of DNA molecules.

After generating the sequencing data, the large amount of raw DNA sequence data comes out to the world. The raw data scrutinize under the Phred quality score (Q score) by calculating the probabilities of any kinds of technical errors introduced to the base/read. Each of the base scores is then re-sorted and aligned with human genome sequences (or any interested genome

sequences, but deals with human genome in this thesis). Continuously, the aligned sequences organize with the counts of reference read counts and any alternative read counts to evaluate under the multi-statistical analysis for structural variant analysis (Figure 4).

NGS-based ctDNA analysis

Currently, the utility of NGS-based technology for ctDNA analysis has proved in numerous amounts of studies. As mentioned earlier, there must be optimized step for library preparation to have proper analysis of ctDNA. As the amount of ctDNA is limited, the recovery rates of DNA molecules are the critical points. Shortlisted to the targeted sequencing approaches, there are Tam-Seq (21), Safe-SeqS (22) for amplicon-based sequencing and CAPP-seq (23) and TEC-Seq (24) for hybrid capture sequencing. These ultrasensitive methods succeed to detect as low as 0.002% of the allelic fraction that even minimal residual diseases can be detected faster than the computed tomography (CT) images or any other cancer biomarkers by detecting ctDNA footprint.

Taking advantage of collecting the genomic alterations at once, the understanding of dynamics of tumor genomics were recognized in real-time. Roughly speaking, the main point of implementation of NGS to ctDNA analysis is, there is no need to know the prior information of tumor mutation. Moreover, the cost of NGS-based technology continuously reduces that can compete with digital PCR in a near future.

Potential misdiagnosis from background errors

Nevertheless, the inevitable problem is the background errors can be incurred either from the technology or biology. The measurement of sequencing errors was well-documented since the NGS-based technology has invented. The technical errors can be introduced by each steps of procedure. These errors are especially critical to ctDNA analysis because of the potential alleviation of false positives.

Another caveat is the biological errors. The little understanding of the biology of cfDNA makes vulnerable to apply for the screening tool. The most concern of biological errors is rise from hematopoietic cells (15). As the hematopoietic cells continuously circulate with the circulating tumor cells, the false positives can be involved. It also fluctuates the biological background that contributes the bias results. Therefore, elucidating the background noises are needed for a confident in variant calling.

In this thesis, chapter 1 focuses on how I have updated the standard operating procedure to optimize the next-generation sequencing based technology for the small amount of input DNA. Continuously, to discover the error-prone steps in NSG-based technology, the systematic evaluation was proposed by comparing the background distribution of acoustically sheared germline DNA and naturally fragmented cfDNA. In chapter 2, to prove the benefits of the ctDNA sequencing, I compared the matched tumor DNA and cfDNA and evaluated the detection sensitivity comparing to the digital PCR analysis. Lastly, I have discovered the different characteristics of cfDNA

compare by the different types of body fluid to discover the effect of cfDNA release mechanism.

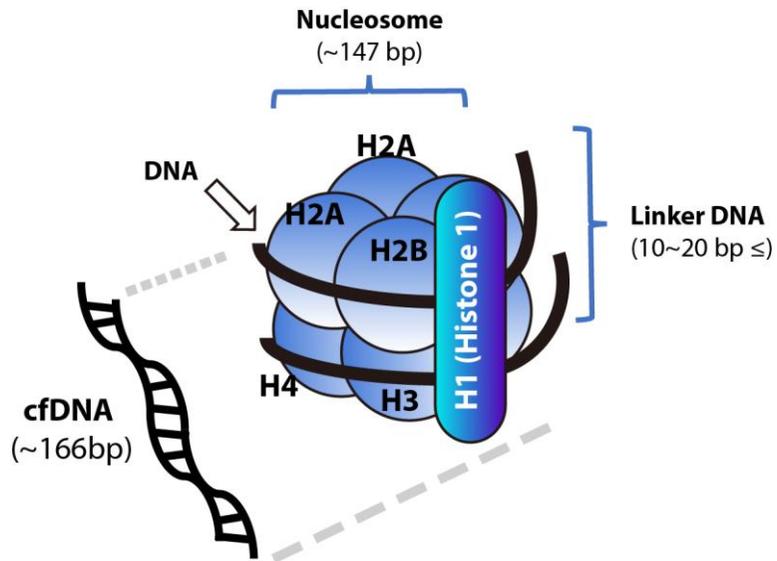


Figure 1Characteristic of cell-free DNA.

The average peak size of cfDNA is approximately 166 bp. The sequence length is similar mononucleotide that a wrap of histone core (~147 bp) and linker DNA (10-20bp).

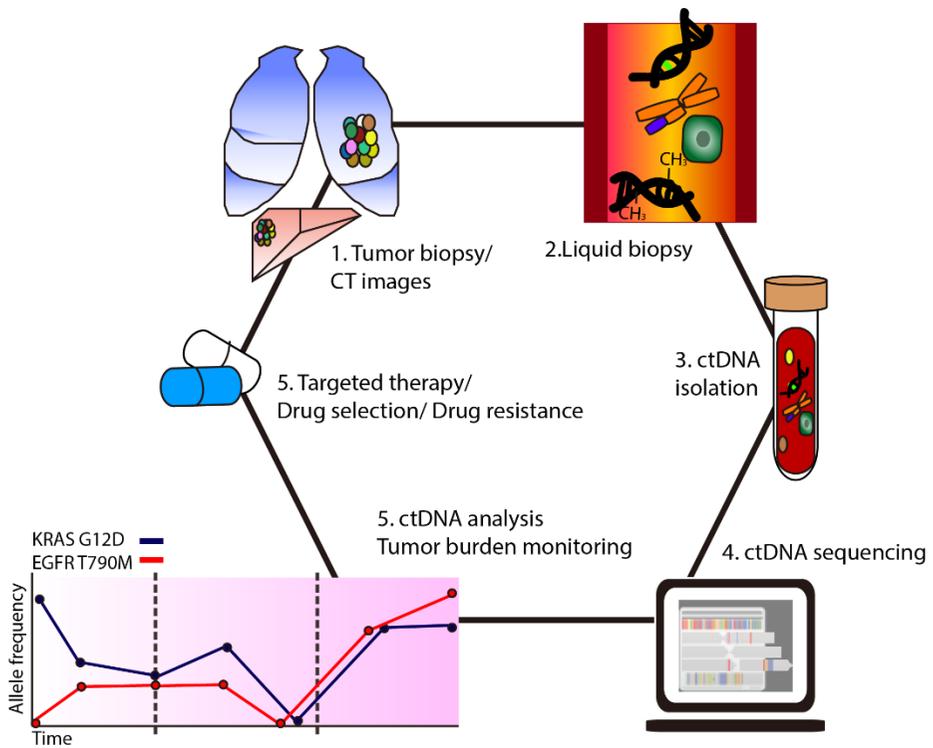


Figure 2 An overview of circulating tumor DNA analysis in clinical application

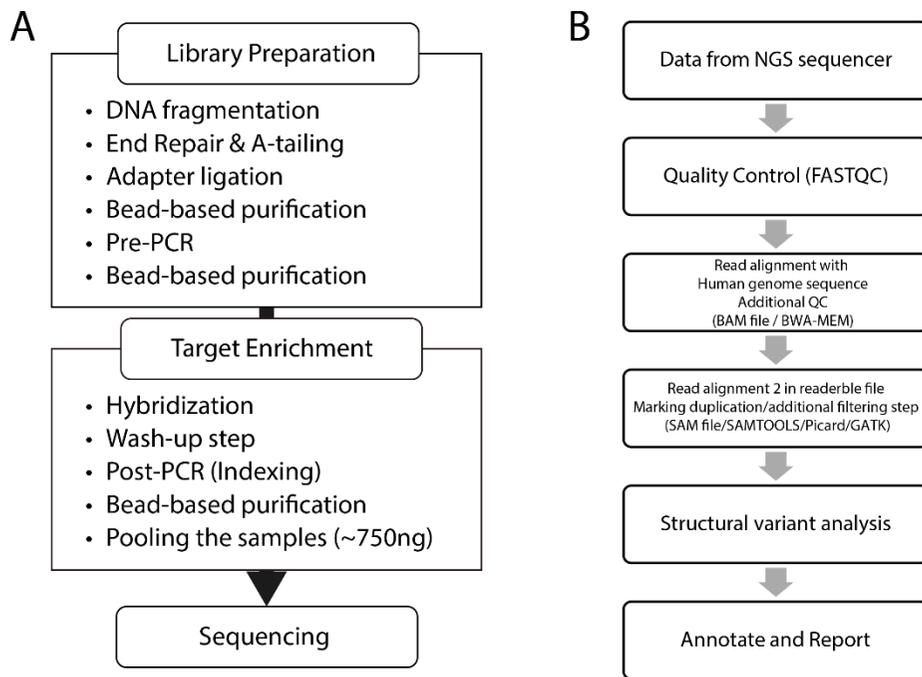


Figure 3 General process of capture-based targeted deep sequencing

(A) A general scheme of the library preparation and (B) the bioinformatics pipeline for calling genomic alterations.

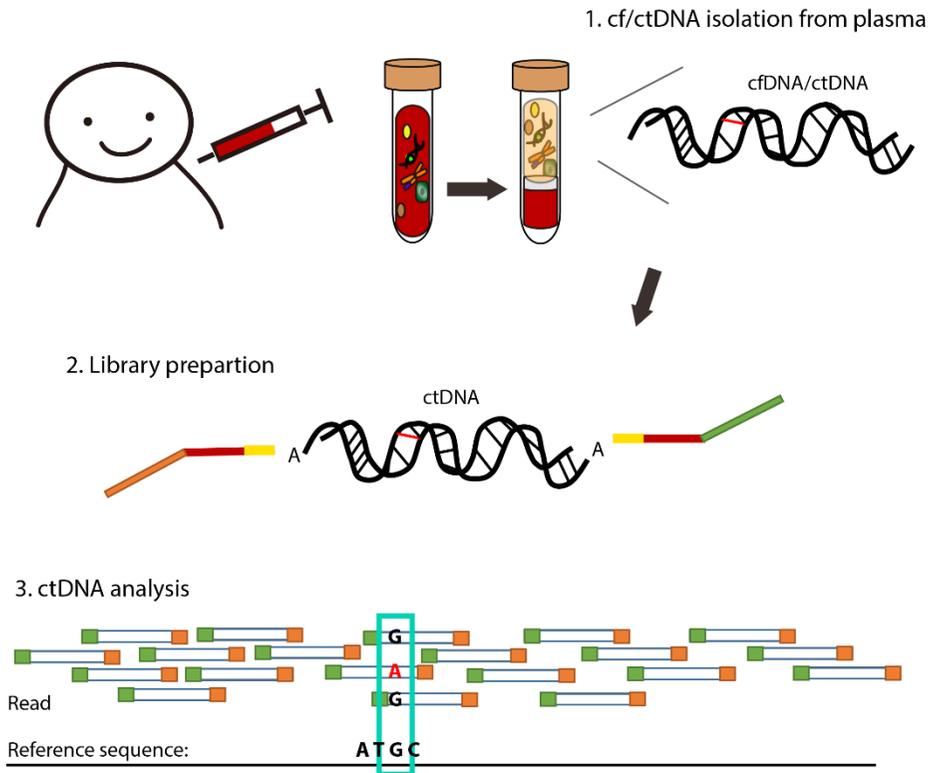


Figure 4 Schematic flow of ctDNA analysis using NGS-based technology

CHAPTER1

Practical guidelines for cell-free DNA analysis using enhanced analytical performance of NGS- based method

**This research was published in *Scientific Reports* on May 2016 and
Genome Biology on July 2017.**

INTRODUCTION

The range of the amount of cell-free DNA varies on individual samples. Theoretically, the cell-free DNA releases from every part of the somatic cells including releasing the circulating tumor DNA from cancer cells. In general, the cell-free DNA circulates in a minimal amount in the body fluid, but it relies on the condition of health that there is a possibility of having a large quantity of cell-free DNA. The concentration of cell-free DNA matters to the procedure of profiling the mutations from the cancer patients which directly associates with the detection of sensitivity and specificity. The potential of harboring the somatic mutation could be extremely low as 0.01% that the substantial recovery of unique DNA molecule is highly desirable for extensive analysis. Any loss of unique DNA molecule is critical for reducing the limit of detection, profiling the lowest cancer signal, and understanding of cancer progression. The innovative technology called digital PCR has now routinely aided to detect the lowest allelic fraction using the lowest amount of DNA. However, the argument of using digital PCR is, the prior information must be given for identifying the specific loci. It confronts the issue to many patients who do not have the particular types of known mutations as well as surveillance monitoring during the therapeutic intervention. Therefore, the implementation of genome-wide sequencing is suitable to understand the proper tumorigenesis of any kinds of cancer study.

The integration of next-generation sequencing with cfDNA has been developed recently that proved the possibility of profiling the tumor genomics

in a real-time. Nonetheless, the studies presented with the customized techniques that there is no standard guideline to implement the cell-free DNA sequencing properly. It is often hard to reproduce the experimental procedure or the bioinformatics workflow. Therefore, to set out the practical guideline for the minimal amount of starting material, I primarily reached to the ligation step in the NGS library preparation to maximize the recovery of the unique DNA molecule and the high confidence of throughput needs.

Next, the key element of high sensitivity and specificity is to discriminate the technical and biological errors from the limited amount of samples. It is well documented that the sequencing artifacts limit the analytical sensitivity (26-28). For example, errors caused by Illumina HiSeq sequencer chemistry are relatively well-understood, and therefore appropriate data filtering criteria based on this knowledge are routinely applied to generated data to remove them (29). The filtration of errors includes the removal of parts of, or entire reads containing numerous low-quality bases, to minimize downstream analysis artifacts (30). The fidelity of polymerases routinely used in the construction of sequencing libraries is well characterized (31, 32); however, it is difficult to quantify the error rate induced by DNA damage during library construction. For example, heat-induced cytosine deamination during PCR thermocycling has been suggested as a possible cause of baseline noise in Ion Torrent semiconductor sequencing data (33). Moreover, cytosine deamination occurs not only during experimental procedures such as PCR amplification (33) and formalin fixation (34, 35), but also prior to sample preparation (i.e. intrinsically or biologically) in the

original DNA templates (36). Nevertheless, it is not clear the error-prone step as well as the impact of how much of the errors relevant to cell-free DNA analysis which have been incurred during the sequencing run itself. Since technical errors are also likely to be introduced during sample preparation, library preparation, target enrichment, and/or amplification of DNA samples, a thorough characterization of such errors may facilitate the detection of method-dependent systematic errors and allow true variants to be distinguished from these errors. To determine during which step, and to what extent, a given type of error is introduced during sequencing, comparative experiments under different experimental conditions have been recommended, but are rarely performed due to practical reasons (29). Thus, no systematic analysis of the errors introduced during capture-based targeted deep sequencing has yet been conducted. To discover the systematic error-prone step in NGS-based technology, I attempted to analyze the non-reference alleles in ultra-deep coverage targeted capture sequencing data from both plasma and peripheral blood leukocyte (PBL) DNA samples. From this analysis, the rate of sequencing-artifact substitutions was estimated to be incurred during specific steps of the capture-based targeted sequencing process including DNA fragmentation, hybrid selection, and sequencing run. Based on the results, the use of mild acoustic shearing was recommended for genomic DNA (gDNA) fragmentation to minimize C:G>A:T and C:G>G:C transversion errors.

MATERIALS AND METHODS

1. Sample collection and DNA extraction

The corresponded blood samples were collected in Cell-Free DNA™ BCT tubes (Streck Inc., Omaha, NE, USA) (37) from 19 human subject. The samples were processed within 6 h of collection via three graded centrifugation steps (840 g for 10 min, 1040 g for 10 min, and 5000 g for 10 min, at 25 °C). The germline DNA were drawn from PBLs and collected from the initial centrifugation. The layer of plasma was transferred to new microcentrifuge tubes at each step. Plasma and PBL samples were stored at -80 °C until cfDNA extraction.

Germline DNAs from collected PBLs were isolated using a QIAamp DNA mini kit (Qiagen, Santa Clarita, CA, USA). Circulating DNAs were extracted from 1–5 mL of plasma using a QIAamp Circulating Nucleic Acid Kit (Qiagen). DNA concentration and purity was assessed by a PicoGreen fluorescence assay using a Qubit 2.0 Fluorometer (Life Technologies, Grand Island, NY, USA) with a Qubit dsDNA HS Assay Kit and a BR Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). The concentration of DNA and the purity quantified by using a Nanodrop 8000 UV-Vis spectrometer (Thermo Fisher Scientific) and a Picogreen fluorescence assay using a Qubit 2.0 Fluorometer (Life Technologies). The size distribution of DNA fragmentation measured using a 2200 TapeStation Instrument (Agilent Technologies, Santa Clara, CA, USA) and real-time PCR Mx3005p (Agilent Technologies) according to the manufacturer's instructions.

3. Library preparation

Genomic DNAs samples fragmented by using a standard protocol of Covaris S220 (6 min, 10% duty factor, peak incident power = 175 W, 200 cycles/burst; Covaris Inc. Woburn, MA, USA) which the average size of 150–200 bp. On the other hand, the plasma DNA was prepared without fragmentation. The construction of sequencing libraries was achieved using 200 ng (for all samples) of PBL, and 37.3 ng (on average) of plasma DNA. To conduct the effect of DNA fragmentation step of background error rate measurement, the intensity and/or duration varied using 200 ng of initial genomic DNA from HapMap samples.

Next, the libraries for PBL and plasma DNAs were constructed using a KAPA Hyper Prep Kit (Kapa Biosystems, Woburn, MA, USA) as reported in Scientific Report (2). In brief, the adjustment of the end-repair, A-tailing, adapter ligation, and PCR reactions (nine amplification cycles) prior to target enrichment were performed. A purification step was carried out using AMPure beads (Beckman Coulter, Indiana, USA) after each step. Adaptor ligation was performed using a pre-indexed PentAdapter™ (PentaBase ApS, Denmark) at 4°C overnight.

4. Sequence data processing

After acquiring the raw FASTQ file from the sequencing procedure, the BWA-mem (v0.7.5) (38) aligned the hg19 human reference to create BAM files. SAMTOOLS (v0.1.18) (39), Picard (v1.93), and GATK (v3.1.1) (40) were used for sorting SAM/BAM files, local realignments, and duplicate

markings, respectively. The duplicates, discordant pairs, and off-target reads were filtered according the instruction.

5. Background distribution analysis

The paired set of PBL and plasma DNA samples were determined a base at a position across the entire target regions to be a background allele if the following conditions were met: (1) the base was a non-reference allele; (2) the position displayed sufficient depth of coverage (i.e. $>500\times$) in the paired PBL and plasma DNA samples; and (3) the frequencies of the base in both samples did not indicate a germline variant (i.e. $<5\%$). Since the samples were collected from cancer patients, the filtration of candidate of somatic cancer variants was conducted. The genomic alteration profiled from the matched fine-needle aspiration (FNA) biopsies had been implied for the filtration. For example, *KRAS* variants were removed from the analysis if detected in the matched FNA specimens. The sequencing libraries for the primary tumors were also generated using 200 ng of input DNA according to same instruction I have mentioned above. After the removal of duplication, the depth of coverage of FNA samples was on average $987.15\times$ ($790.32 - 1476.55\times$). The position that below $250\times$ in the matched FNA biopsy and an allele if it was present at a frequency greater than 2.5% in the FNA sample was discarded to further analysis in a pair set of PBL and plasma samples.

6. Analysis of nucleotide composition and substitution rate at the near DNA break point

To obtain the frequency of mono- and dinucleotide at positions around break

point, 5'-end position of each mapped read was determined on the human reference genome and the sequence for the region of 100 bp (± 50 bp) around break point was collected. For a consistency, the collected sequences were displayed in the direction of the positive strand of the reference genome. The frequencies of nucleotides were calculated as the number of occurrence of a given mono- and dinucleotide divided total base with a quality score ≥ 30 at relative positions to break point. The frequencies were obtained for each sample and then values from 19 samples were averaged. For estimation of the frequency of mononucleotides, we displayed the position as the number of nucleotides from the first 5'-end nucleotide of the read. For dinucleotides, the number of nucleotides between the phosphodiester bond in a given dinucleotide and the break point was shown to indicate the relative position to break point. For instance, distance zero indicated that the first position was taken right before the 5'- end of the read, and the second position coincided with the beginning of the read.

Background error rates across all substitution classes were also calculated at each position relative to break point. The background alleles for each sample defined as described in the previous section were used for the analysis. For a comparison between PBL and plasma sample, substitution rate was normalized by the average rate of 1-50 bp. To remove errors occurred in Illumina sequencing platform as much as possible, we used only R1 reads whose front parts showed relatively better quality scores than those of R2 reads.

RESULTS

Comparison of blood collection tubes

The stabilized blood collection tube must be selected to avoid a loss of unique DNA molecules from the minimal amount of plasma DNA. It is critical to minimize the chance of involving the false positive variants from the lysed peripheral blood leucocytes DNA to cfDNA. A chance of accumulation of biological background errors due to the hematopoietic cells is inevitable. To test the stabilization of total amount of plasma DNA, the series of different time and temperature were measured. Streck BCT cell-free DNA blood collection tubes maintained the minimum variation of total amount of plasma DNA (154.45 ± 21.05 to 139.3 ± 18.6 , mean \pm SEM) compared to EDTA tube (138.3 ± 29.25 to 177.0 ± 20.4 , mean \pm SEM) (41) (Table 1-1A). Streck BCT cell-free blood tube also correlated higher number of detectable variants (22 ± 2 , mean \pm SEM) than EDTA tube (18.5 ± 0.5 , mean \pm SEM) (Table 1-1B). Overall, Streck BCT cell-free blood tube was selected for further analysis.

Optimization of the library preparation

The part of result is published in *Scientific Report* (2). The series of spike-in DNA was used in this test and the evaluation of commercial kits was presented by previous reports from our laboratory (2). In brief, KAPA Biosystems' Hyper Prep kit was selected for amongst the commercially available kits. To maximize the efficiency of library construction for the minimal amount of input DNA, the various ligation conditions were evaluated.

The ligation condition of (i) temperature, (ii) duration of time, and (iii) the molar ratio of adapter were considered to assess the performance of the recovery rate for unique DNA molecules. The evaluation between 16 °C to 25 °C for 15 min or 60 min did not make any differences compared to standard conditions suggested by the manufacturer recommendation (i.e., 20 °C for 15 min, data not shown). On the other hand, the ligation of temperature lowered to 4 °C and extended to overnight increased the total amount of DNA after the pre-PCR compared to 20 °C with 15 minutes (Table 1-2). After the rate of the improper mapped reads, duplication, improper pairs, and off-target reads discarded, the rate of on-target was quantified. Figure 1A shows on-target rate increased from 40% to 55% or 18% to 28% using 50 ng or 10 ng of input DNA, respectively. Moreover, the duplication rate was lowered from 35% to 19% or 60% to 50% using 50ng or 10ng of input DNA (Figure 1-1A). Next, the range of adapter concentration was tested from 136nm to 1.36Mm (Figure 1-1B). In a molar ratio of the adapter: insert, 300:1 to 30000:1 was tested using 50ng of gDNA with ligation of 4 °C overnight. Figure 1C shows the rate of duplication increased by the extension of ligation time and higher molarity of adapter. However, the purification step cleared up the potential PCR dimers and allowed to bind more DNA molecules. Therefore, the library construction was optimized with temperature of 4 °C, extension time of overnight and higher amount of adapter ratio for targeted sequencing with the minimal amount of input DNA (Figure 1-1C).

Optimizing statistical modeling for cfDNA analysis

To assess the low allelic fraction from plasma DNA, current open-source tools (42, 43) were evaluated with statistical methods suggested from previous studies in a range of spike-in controls (44, 45). As expected, the number of variants were failed to detect in both of open-source tools (Table 1-3). Despite the Fisher's exact test had detected number of variants with higher positive predicted values, the Binomial tests had shown higher sensitivity on the variants under the 0.1% allele frequency. As the sensitivity of detecting low allelic fraction is much more critical for ctDNA analysis, the Binomial statistical analysis was chosen for further analysis.

Performance of optimized TDS on cfDNA and PBL DNA

The total of 19 human subjects, including 17 pancreatic cancer patients and 2 healthy volunteers, were profiled by using optimized method described previously. On an average of 200ng of PBL DNA and 37.3 ng of plasma DNA, the average of 56.4 and 20.0 million total reads were generated in plasma and PBL DNA, respectively. The alignment rates were on average of 87.3% and 93.7% for plasma and PBL DNA. The unique coverage were determined to be $1964\times$ (1210 – 3069 \times) and $1717\times$ (1042 – 2361 \times) on average, respectively, after excluding the PCR duplication. The potential systematic bias from library or sequencing data excluded by comparing the single nucleotide polymorphisms (SNPs) from matched plasma and PBL samples. By these, a strong correlation between plasma and PBL samples were observed

($R = 0.9913$, $p\text{-value} < 0.0001$). Conclusively, the optimized methods generated sufficient amount of reads for the further analysis.

Estimation of errors derived by TDS

From sequencing reaction

The critical factor for down-stream analyses is depended on the constructing the proper background distribution from the plasma or PBL samples. Dae-soon Son and Seung-ho Shin helped to generate the proper background distribution. As mentioned in Method, the tumor-derived single nucleotide variants (SNVs) and germline SNPs excluded to avoid the potential bias. After that, Phred base quality score of non-reference background alleles was observed to distinguish any involvement of systematic bias. Most of the background alleles depicted under 30 of the base quality score, but the small bump was discovered after the base quality of 30 (Figure 1-2A). It was indistinguishable from the reference alleles (Figure 1-2B). It is critical to note the lowest fraction of errors are involved above the qualified bases because the lowest allelic fraction from plasma DNA is indistinguishable. As the background distribution was constructed after filtering out with most of the sequencing errors, the presence of the highly qualified background alleles may indicate the errors are from the other sources.

Distribution of background errors

Although both plasma and PBL DNA samples have been used as a control group for the purpose, the similarities and dissimilarities of background errors between plasma and PBL DNAs have not been elucidated. Thus, I compared the background errors from the plasma and PBL DNA. After the base quality score filtration, overall mean background rates were estimated to be 0.007% and 0.008% in plasma and PBL DNAs, respectively (Figure 1-3A). Next, with Seung ho Shin's aid, entire 12 nucleotide substitution classes of errors were examined (Figure 1-3B). The context of dependencies was revealed by incorporated information on the bases immediately 5' and 3' to each mutated. While the background frequency of each substitution class varied depending on its context, the patterns of background frequency variation associated with specific sequence contexts were strikingly similar between plasma and PBL DNAs except C:G>A:T substitution (Figure 1-3C).

Sample preparation caused background errors

In order to generate PBL DNA and plasma DNA sequencing data under the exact same condition, the optimized experimental protocol had to apply excluding the fragmentation step. Hence, the elevation of C: G>A: T hypothesized as due to DNA damage of the fragmentation step. The condition of milder acoustic shearing was applied to test whether the levels of C: G>A: T transversion disappeared. The intensity and/or shortened duration of acoustic shearing decreased the rate of C: G>A: T transversion in PBL DNA dramatically (Figure 1-4). Thus, the standard DNA fragmentation protocol

elevated the rate of C: G>A: T substitution owing to DNA damages, which could be alleviated under an appropriate fragmentation condition. By typical oxidative base lesion causes the formation of 8-oxo-7, 8-dihydroguanine (8-oxo-G) under C:G > A:T errors, enzyme-linked immunosorbent assay (ELISA) was performed by Yun Jeong Kim. By providing the samples with the serial attenuation of acoustic energy level, the 8-oxo-G attenuated by the assay (ANOVA p value = 6.0×10^{-7}). Therefore, it was definite that the standard protocol of DNA fragmentation step caused the elevation of C:G > A:T and C:G > G:C transversions. Taken together, the background rates were very similar between plasma and PBL DNA samples. Although PBL DNA displayed significantly higher C: G>A: T transversion rate than plasma DNA, our data suggested that an appropriate fragmentation condition abolished the elevation in the substitution rate. The results also suggested that germline DNA fragmented under the proper condition could be an alternative resource to estimate site-specific error rate distributions of plasma DNA in normal controls.

Breakpoint preferences

By considering the fragmentation step introduced the background errors, the end of fragmentation had to be associated with the mechano-chemical breakage of DNA (Figure 1-5). To characterize the preferences of breakpoint, the end bases of DNA fragments were assessed. First up was taking the both of aligned read 1 and read 2 and count the each of the nucleotides according

to the start position to 20 bp of the sequences. Figure 1-6 displays roughly estimated read counts from the randomly selected genomic regions. At a glance, I noticed the different ratio was shown at the first two bases. As sequencing platform is renounced with bad quality scores at the first four consecutive read bases (46, 47), the high quality of base scores near the end of DNA fragments were examined. As depicted in Figure 1-6B, the bad quality score were cumulated at the first five bases. Interestingly, the quality score was increased from the second bases which correlated with rough data (Figure 1-6A). Another hypothesis from Figure 1-6A was the different ratio of nucleotide bases. Noticeable differences clarified under categorization of the substitution classes (Figure 1-7A) which supported the previous data that has the dependency of fragmentation. Plasma DNA had different preferences of nucleotide changes (G>A then G>T) than PBL DNA. By taking advantage of naturally fragmented plasma DNA, the substitution rate was compared from the start point of read bases. Noticeably, the substitution rate of A with either G or T (ex, A>K) was significantly elevated at the first base in PBL DNA compared to plasma DNA (Figure 1-7B). The result indicated that the DNA damage did not induce A>K substitution. On the other hand, the substitution of neither C:G > A:T nor C:G > G:C errors was observed which are the most commonly associated with acoustic shearing (Figure 1-7B).

As I noticed the substitution of residue A might be associated with mechano-chemical breakage of DNA. The frequency of mononucleotide around DNA breakpoint was analyzed. By observing the fluctuation of frequencies of mononucleotide, A residue was predominantly presented in PBL DNA

(Figure 1-8). To get proximal examination, the total of 16 dinucleotides of frequencies around DNA breakpoint was analyzed. As expected, CA, TA and GA were susceptible to cleavage (Figure 1-8 and 9). Additionally, in order of $CG > CA > TA \sim GA$, the cleavage rate of phosphodiester bonds were reduced (Figure 10). Taken together, the acoustic shearing has fragmented DNA at the 5' A nucleotide residue preferentially.

Multi-statistical adjustment for removing the background errors

By noticing the potential background errors could be involved from qualified bases, a series of bioinformatics pipeline was framed under the binomial tests. The basic statistical model was followed by cancer personalized profiling by deep sequencing (CAPP-Seq(44)). There are two types of statistical pipelines to evaluate the significance of SNVs: “with primary” and “without primary.” With-primary pipeline tests the known mutations to plasma DNA detected from the primary tumor or tumor biopsy that called from the open-source tools such as VarScan2 or MuTect. On the other hand, without primary pipeline test is also known as “biopsy-free manner.” The answers are unknown; it opens the possibility for detecting any mutations from plasma DNA except the germline mutations. Although the framework of CAPP-seq statistical pipeline has established strictly with high sensitivity and specificity, the pipeline must be tested in order to built-in to the in-house system. There is five stepwise flow to categorize the variants (Method). Briefly, beginning with strand bias test, the total of 500 read counts must be meet or higher counts to

meet the criteria. The allele frequency of plasma evaluates with matched paired PBL DNA allele frequency and decides the allele frequency by adjustment under binomial test with position by position from the background noise distribution. After that, the filtered allele frequencies tested to entire background noise distribution that adjusting by multiple tests by Bonferroni and FDR significance level of 0.05. The variant candidates now considered as the outlier format under the adjusted read counts and Bonferroni p -value. However, I noticed CAPP-Seq has missed the concept of the batch effect. As the step of pooling was included in prior to sequencing process, the batch effect must be considered. Proximal gap dealt with differently with own pipeline process and implemented from the section of Takai et al.(48) pipeline (Method). Table 1-5 shows the number of variants has reduced the number by each step. Overall, the improvement of multi-statistical analysis was optimized for detection of ctDNA.

DISCUSSION

The significant discovery in this study is elucidating the fraction of background errors caused by acoustic shearing. Although the fraction of errors were relatively lower than previous studies, comparison between PBL DNA and plasma DNA surely clarified the acoustic shearing caused the rate of C:G>A:T and C:G >G:C transversion error mainly. The errors were constituted in “guanine” nucleotide rather than other three types of nucleotides. It can be explainable by the characteristic of guanine that has more susceptible to oxidation lesions owing to its potential of oxidation. Mechanically speaking, 8-oxo-G, G to T transversion substitution via dA:8-oxo-G pair, rouse from the process of shearing reported by Costello et al. (49) and can be reduced by the antioxidants. The fraction of errors were >20% comparatively to the present data (>1%) that their errors perhaps have exported with the shearing and the contamination. Moreover, the previous study highlighted the sequences of errors were CCG: CGG >CAG: CTG specifically. On the other hand, present study shows NCG: CGN > NAG: CTN. By taking unique feature of plasma DNA, the errors were aroused by the acoustic shearing rather than typical oxidative lesion product of 8-oxo-G; the direct C:G>A:T transversions which are the products of secondary oxidative lesion of 8-oxo-G, including imidazolone, guanidinohydantoin, and spiroiminodihydantoin which are known for causing C:G>G:C transversions(50, 51). Overall, the oxidation of guanine residues may responsible to cause both of C:G>A:T and C:G>G:C errors by acoustic

shearing.

While the present study was under review, the Chen et al.(52) reported the majority of errors were posed in the 1000 Genome Project and The Cancer Genome Atlas (TCGA) data sets. They reported the errors were the false negative variants that contained the allele frequencies of 1-5%. By analyzing those variants, they found the most prevalent substitution was C:G > A>T followed by A:T > T:A. Moreover, they presented DNA damages are caused by the purification step and alternated by the range of EDTA from the TE buffer. Continuously, the study performed the 1× TE (comprising 10 mM Tris (pH 8) and 1 mM EDTA) reduced the C: G> A: T and A: T> T: A errors. Checking up the buffer concentration immediately after the reports were found to be using exact same 1× TE buffer for DNA shearing and the error rates from the present data also had lower than Chen et al. By these, the present data showed not only the origins of errors but also supported by comparing with plasma DNA that the discovery from the study contributed for the improvement of utilizing the method of capture-based deep sequencing.

Most of previous studies attenuated their errors by adding the DNA repair enzyme. In this study, the errors were attenuated by modifying the standard condition of acoustic shearing: lowering the power and allowing the longer DNA fragments. The direct comparison presented the errors were reducible according to the recommendation of manufacturers for the fragmentation of input DNA with a median size of 150- 200bp. The advantage of reducing the

errors by modifying condition is surely by increasing the quality of bases at the end of reads and efficiency of data output. However, disregarding the fact of the library recovery rate of input DNA must be considered. It was evident from the present data that the on-target rate was reduced by 15-25% compared to the standard condition of acoustic shearing.

Although the present study mainly focused on the DNA damage due to 8-oxo-G, another common mechanism of DNA damage, apurinic-apyrimidic (AP) site, was evaluated by demonstrating the ELISA. AP site damage is involved in the DNA base excision repair that repairing the damages of mismatched DNA sequences by creating a nick at the backbone of the phosphodiester of the AP site. The damage is commonly due to depurination and/or depyrimidation (53). Through demonstrating ELISA, the level of acoustic energy was correlated by the AP sites during the steps of fragmentation (ANOVA, p value = 4.7×10^{-7}), but did not fully provided the reasons of increasing the error ratio of the A>G and/or A>T at the end of the DNA fragments. Hence, the mechano-chemical breakage of DNA is the strong candidate by causing the A>K errors at the end of fragments of DNA which was proved by comparing the plasma DNA. To have stronger evidence, similar data experimental data sets were found by the public data. Two independent studies were evaluated (54, 55). The data were generated under the whole-exome sequencing (WES) and had used the same COVARIS machine under the standard manufacturer's recommendation. The parallel comparison found the A>K errors at the end of DNA, but the errors were

lower than the present data. One general potential source could be from the differences of ligating the enzymes. While present data were facilitated with the KAPA Hyper enzymes, the two public data were made by the SureSelect enzyme that contains the T4 DNA polymerase and Klenow fragment. Due to patent, KAPA Hyper enzymes were blinded. Moreover, the time of extension might have increased the dimers that the present data had not eliminated the errors completely. As the end repair enzymes and time modified, the fidelity of end repair perhaps have influenced the fraction of errors in this present data set.

In a new regular feature of comparing plasma DNA with PBL DNA, it was noteworthy that the cleavages were preferred to cumulate at the 5' phosphodiester bonds of A residue appeared in PBL DNA. It intrigues to investigate the biological background noises. As Newman et al (44, 56) mentioned the background rate was imposed at the hotspot variants, the recurrent hotspot mutations were examined. There were no distinctive differences between the plasma and PBL samples (Figure 1-11 A). As the recurrent hotspots appeals to have predominant across the targeted regions, the background rate of tumor protein p53 (TP53) was observed (Figure 1-11 B). TP53 is also the region that frequently mutated in most of tumors compared to hematological malignancies. This point out the fact that if the background errors contain higher rate of TP53 variants in plasma DNA then the contribution can be reliably to acclaim that the pre-neoplastic cells have derived rather than the hematopoietic lineage cells (57). Since there were no

distinctive differences in either of plasma and PBL DNA, it makes sense that there is minimal impact of biological background at cancer hotspots. Taken together, the data sums up about the technical noises contributed much higher than the biological noises.

It is critical that the origin of ctDNA is unclear up to now. In fact, the data shows the higher ratio of A:T>T:A and C:G>T:A transversion errors in plasma DNA that another hypothesis must be set out to solve how the cells contributed to release and turned out to become a cell-free DNA. Another suggestion would be the healthy volunteer samples may be the alternative source for the standardization of background metrics as the circumstances of background errors are not randomly distributed entirely. In summary, the systematic analyses of technical and biological background noises helped to establish the proper statistical testing to further downstream of ctDNA analysis.

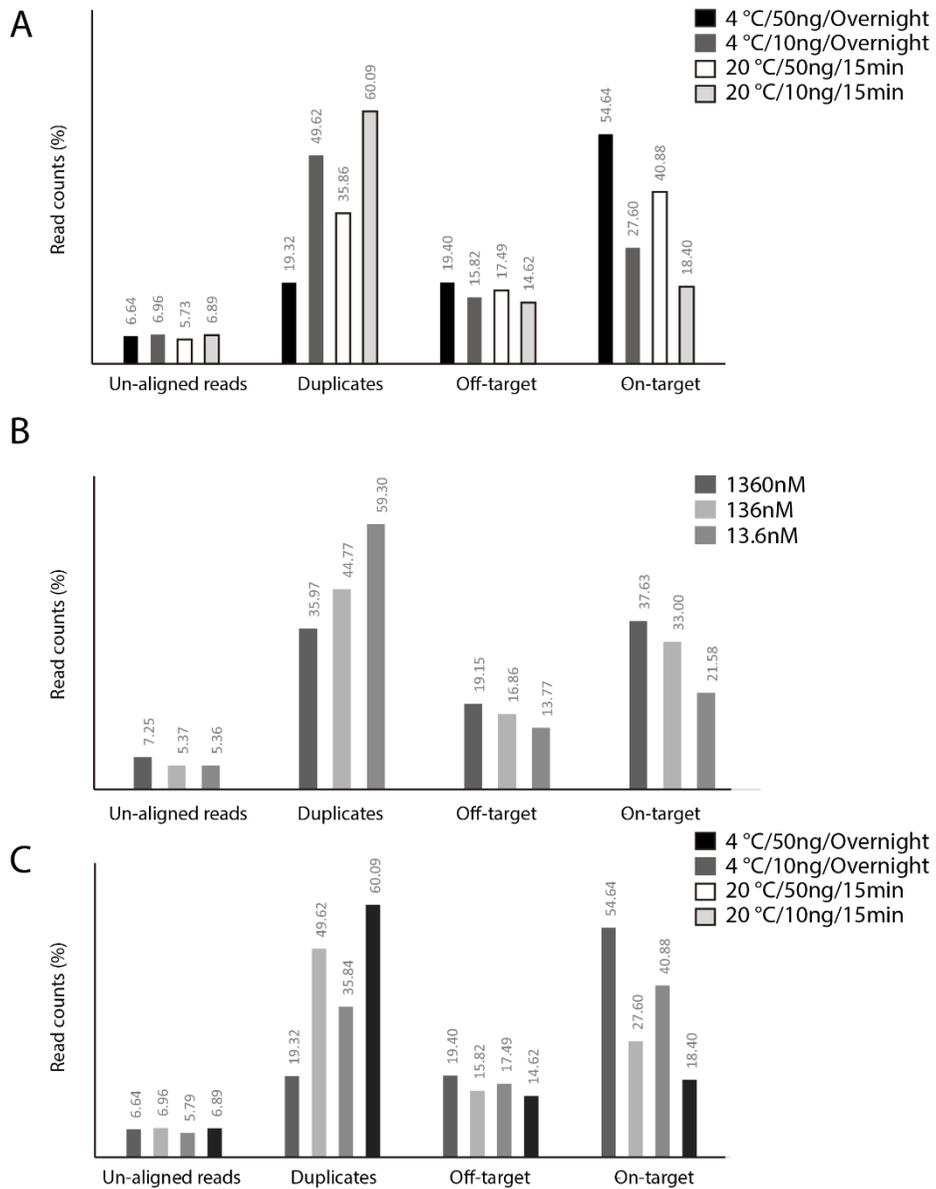


Figure 1-1. Performance of cfDNA sequencing

Performance of cfDNA sequencing by (A) adjustment of time, temperature and (B) molar ratio of adapters. (C) The comparison of optimized ligation step with standard condition.

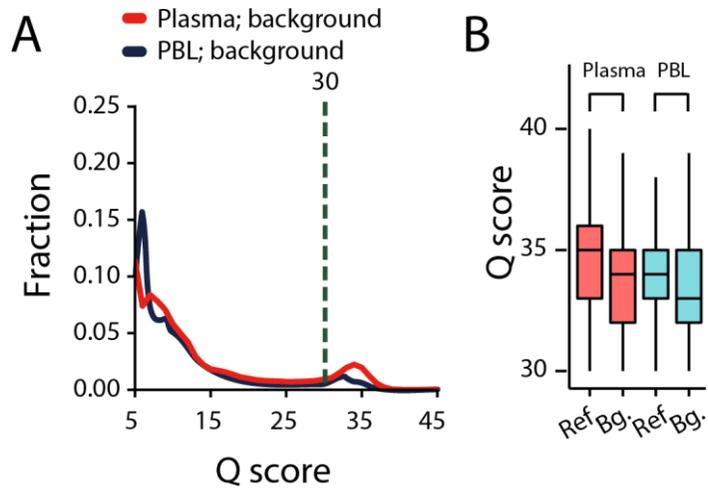


Figure 1-2. Quality score of read bases in targeted deep sequencing data

The distribution of background allele visualized with the density plot. The small fraction of background allele discovered above the quality score of 30. (B) The comparison of reference allele and background allele distribution gathered from both of plasma and PBL DNA samples.

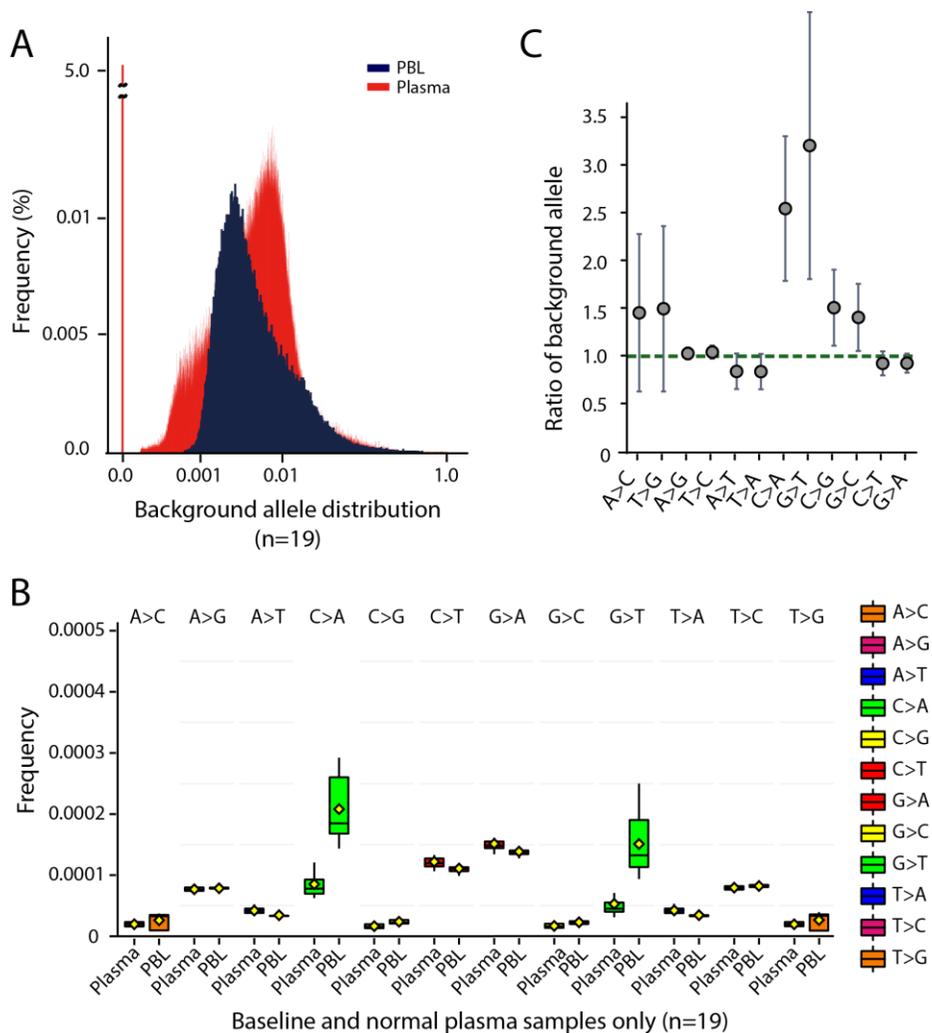


Figure 1-3. The distribution of background errors from PBL and plasma DNA

The distribution of background noise from PBL and plasma DNA were analyzed under substitution classes. (A) The distribution of background alleles from PBL and plasma DNA. (B) Substitution classes were compared between PBL and plasma DNA. (C) The ratio of substitution classes were determined by dividing plasma DNA by PBL DNA.

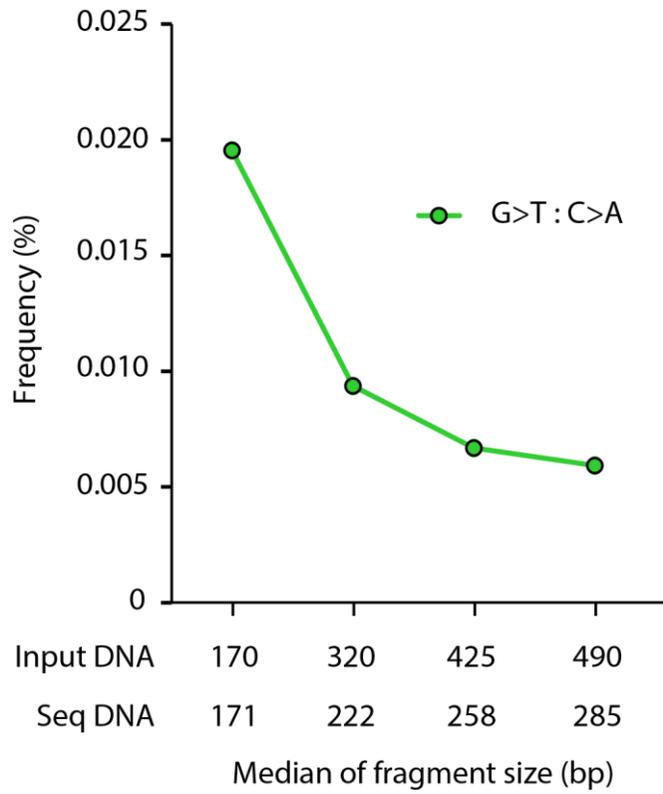


Figure 1-4. Alleviation of background error by various condition of fragmentation

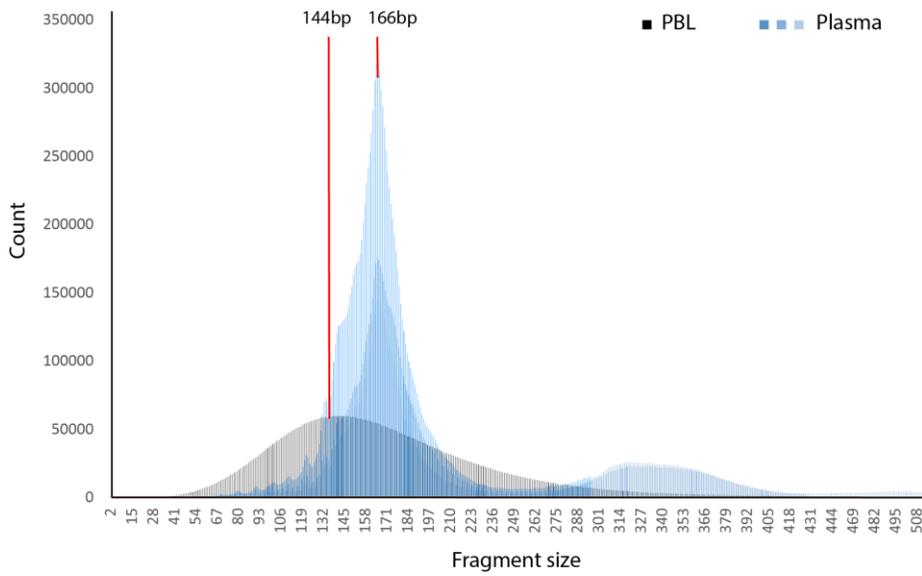


Figure 1-5. The distribution of fragment size from PBL and plasma DNA
 The distribution of fragment size in PBL and plasma DNA sample. The maximum peaks are 144 bp and 166bp for PBL and plasma DNA, respectively.

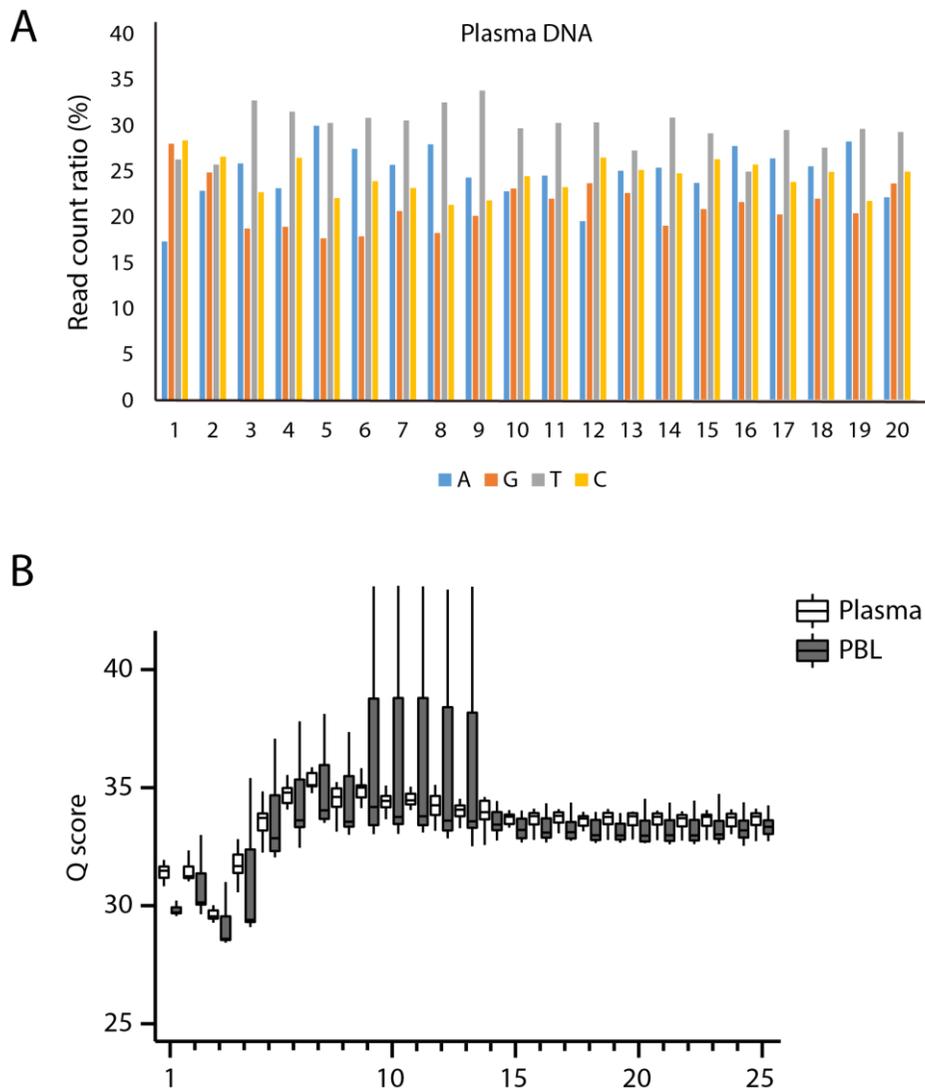


Figure 1-6. Evaluation of read bases from the start position

(A) The estimated average of read counts from the start position up to 20 bp of the sequencing reads. Random selected region, EGFR in this figure, showed for pilot test for the noticing the fluctuation of nucleotide sequences.

(B) The base quality score was examined from the starting point of read from plasma and PBL DNA.

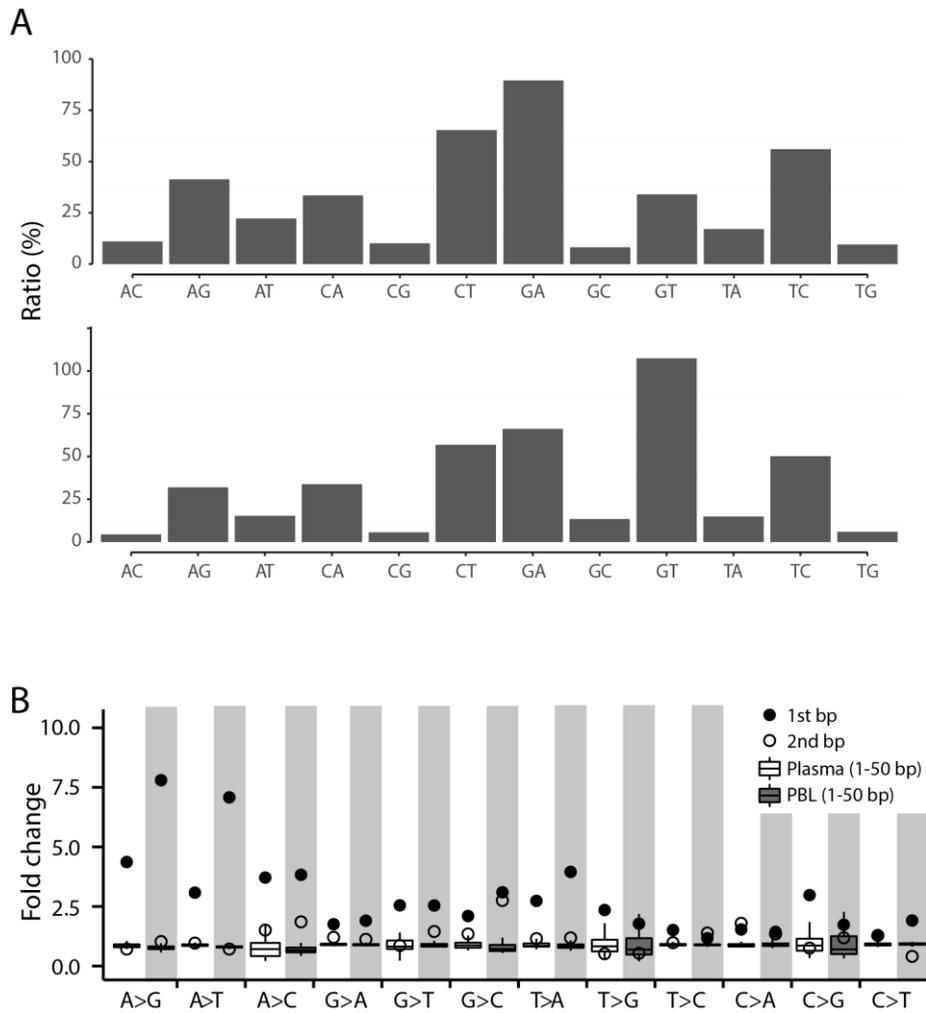


Figure 1-7. DNA breakage preference

The preferential DNA breakage was observed by substitution classes. (A) The ratio of read counts depended upon the 16 substitution classes. (B) The breakpoint of DNA across the substitution classes compared between plasma and PBL DNA.

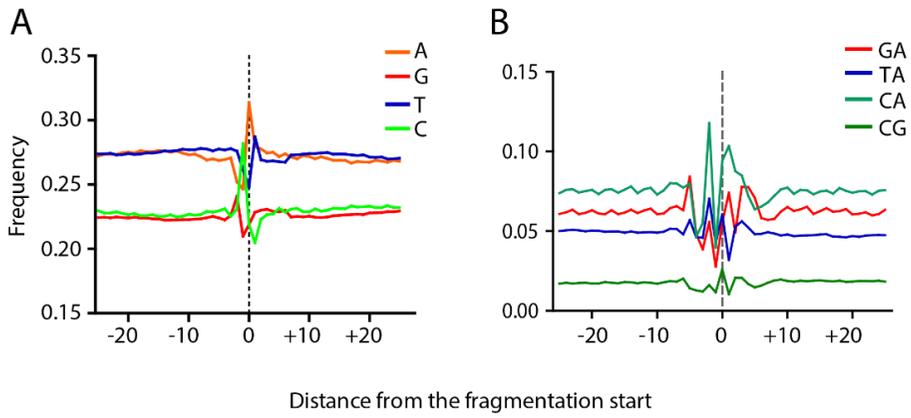


Figure 1-8. Nucleotides around the DNA breakpoint

Nucleotide around the DNA breakpoint was analyzed by (A) mononucleotide level and (B) dinucleotide level.

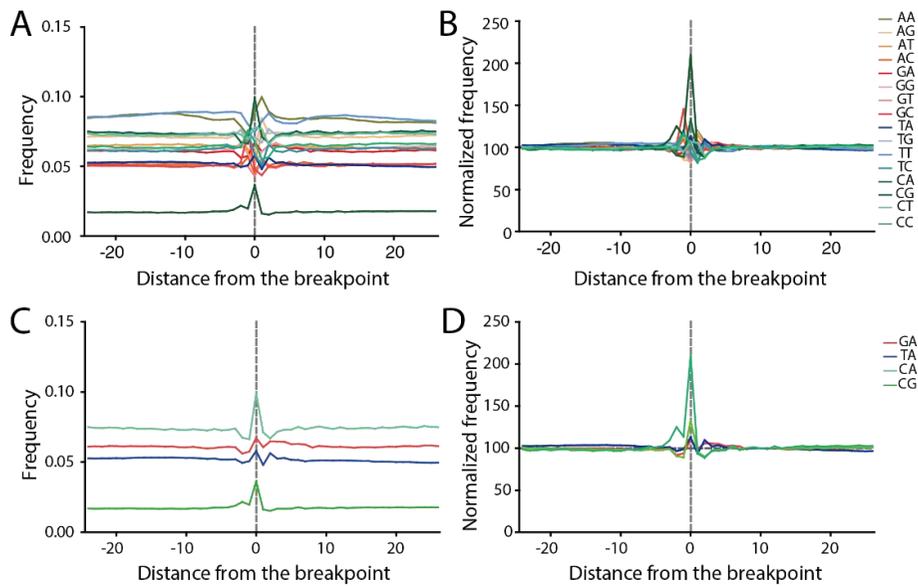


Figure 1-9. Frequencies of dinucleotide

Dinucleotide frequencies around the DNA breakpoint. (A) The frequencies and (B) normalized frequencies of dinucleotide across the 16 substitution classes were analyzed around the DNA breakpoints. The selected four substitution classes were depicted by the (C) frequency and (D) normalized frequencies.

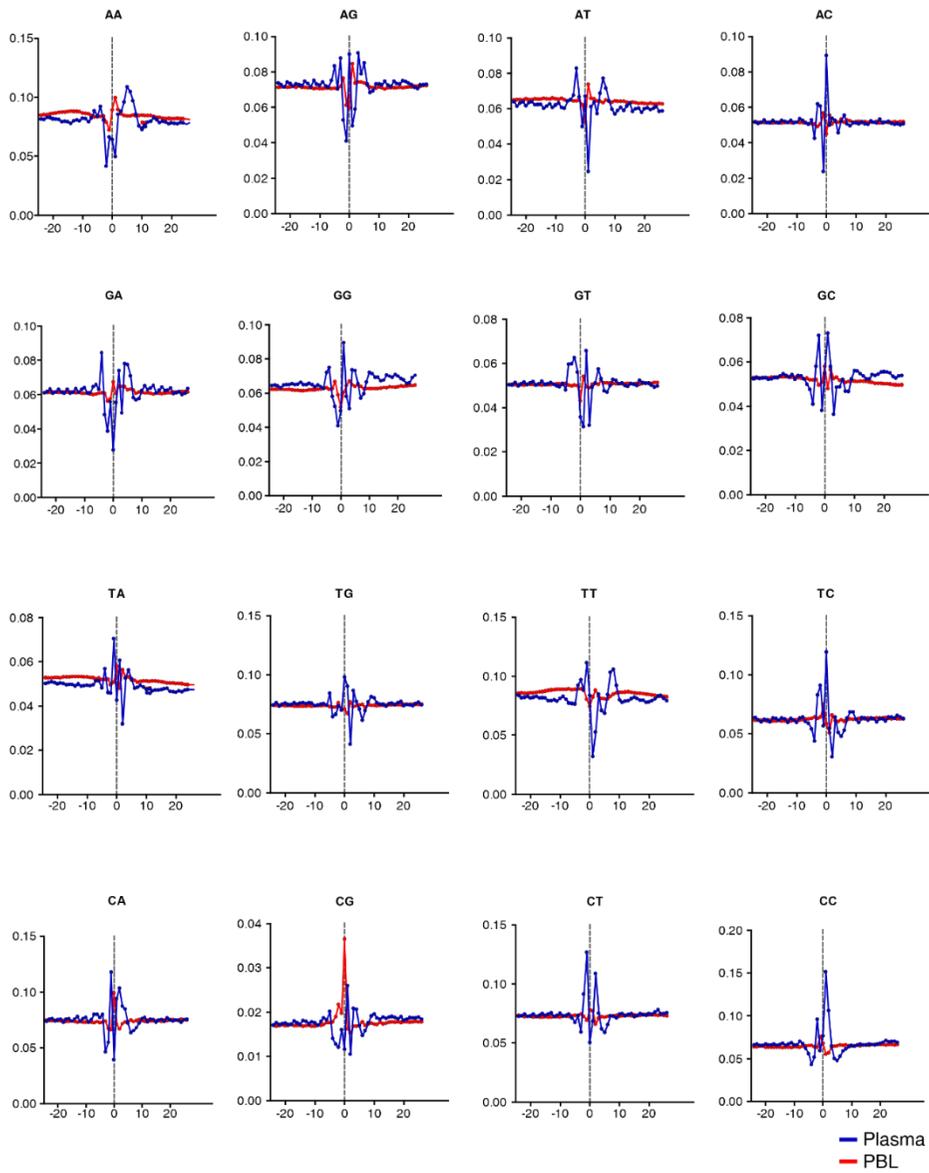


Figure 1-10. Combination of 16 dinucleotide frequencies

The combination of 16 dinucleotide frequencies were depicted around the DNA breakpoint from PBL and plasma DNA.

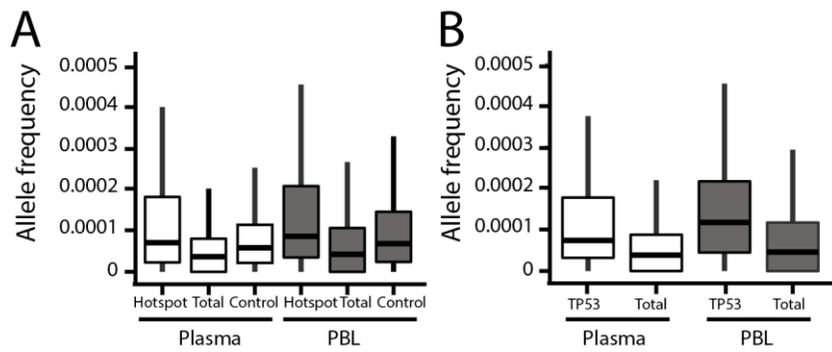


Figure 1-11. Allele frequency of background errors from hotspot mutations

(A) The average of background allele frequencies from recurrent hotspot mutations and (B) TP3 from PBL and plasma DNA.

Table 1-1A. Total amount of plasma DNA collected from Streck BCT and EDTA tube

Time	Streck BCT tube				EDTA tube			
	Day 1		Day 4		Day 1		Day 4	
Conc. (ng/ul)	133.4	175.5	138.3	140.2	109	167.5	208.3	145.7
SEM	9.66				20.76			

Table 1-1B. The number of genomic variants detected from Streck BCT and EDTA tube

Chr	Position	Ref.	GeneID	Variant	Streck Tube		EDTA tube	
					Day 1	Day 4	Day 1	Day 4
chr2	212812097	T	ERBB4	C	√	√	√	√
chr4	1807894	G	FGFR3	A	√	√	√	√
chr4	55141055	A	PDGFRA	G	√	√	√	√
chr4	55152040	C	PDGFRA	T	√	√	√	√
chr4	55972974	T	KDR	A	√	√	√	√
chr4	55962546	-	KDR	G	-	-	√	√
chr4	55980239	C	KDR	T	√	√	√	√
chr5	112175589	C	APC	T	√	√	-	-
chr5	112175770	G	APC	A	√	√	√	√
chr5	149433597	G	CSF1R	A	√	√	√	√
chr9	139399409	CAC	NOTCH1	-	-	-	-	√
chr10	43613843	G	RET	T	√	√	√	√
chr10	43615572	A	RET	T	√	-	-	-
chr11	534242	A	HRAS	G	√	√	√	√
chr13	48941623	T	RB1	C	√	-	-	-
chr13	49033902	T	RB1	C	√	-	-	-
chr17	7579471	G	TP53	-	-	-	√	√
chr17	7579472	G	TP53	C	√	√	√	√
chr19	17954157	C	JAK3	A	√	-	-	-

Table 1-2. The total amount of DNA yield was compared under different ligation condition

DNA (ng) /condition	20°C/ 15min	4°C/Overnight
50	18.6	41.6
50	19.9	35
10	3.68	9.88
10	4.74	9.1

Table 1-3. Evaluation of open-source tools and statistical analysis using spike-in controls

Expected AF (%)	Sensitivity (%)				Positive predicted value (%)			
	VarScan2	MuTect	Fisher's exact test	Binomial test	VarScan2	MuTect	Fisher's exact test	Binomial test
5	16	62.5	93.8	72.9	80	90	94.5	62.4
2.5	0	35.7	90.7	80.7				
1	0	6.25	27.1	43.8	0	25	55	43.6
0.5	0	7.86	14.3	28.6				
0.25	0	6.25	6.3	39.6	0	66.6	100	29.5
0.1	0	0.71	0	30				

Table 1-4. Performance of multi-statistical analysis for ctDNA sequencing

	Binomial	Z-test	Proximal gap	ANNOVAR filter	Final decision
Total variants	113358	1316	1278	30	27

CHAPTER 2

Ultrasensitive interrogation of circulating tumor DNA from cancer patients using enhanced analytical performance of the NGS-based method

INTRODUCTION

The countless of studies have been approached with the genome-widely to understand the molecular mechanism underlies the tumorigenesis. Nonetheless, the studies focused on the localized tumor biopsy that often contradict to the clinical outcome. It is relevantly due to the intra- and inter-tumor heterogeneity in cancer that evolves with the vast amount of mutations. To interrogate the status of tumor precisely, the real-time monitoring system must be conducted along the vast collection of genomic variants. Therefore, many studies have been analyzed the association of tumor biopsy and liquid biopsy such as protein biomarker to chase the change of tumor architecture. Nevertheless, the protein biomarkers stays as the standard biomarker (58) although lack the detection sensitivity and specificity and limiting its role complimentary on monitoring disease burden (11, 59). To compensate the protein biomarker, the cell-free DNA analysis has been evaluated in multiple cancer samples. The limits of detection varied among the location of cancer that deadly disease such as pancreatic cancer (PDAC) has not been highlighted on the benefits of ctDNA analysis.

In this chapter, I have evaluated the utility of ctDNA analysis using previously described method compare to digital PCR. As most of PDAC has constituted with the KRAS mutation over 90% (59), the detection sensitivity of KRAS mutations was benchmarked for the circulating tumor DNA analysis (48, 60-64). Next, I evaluated the benefits of interrogating in the genome-wide scale compared to selective point mutations collected from matched biopsy

samples.

As the release of circulating tumor DNA is not only obtained from the plasma DNA but also from the other types of body fluid, the characteristic of cfDNA was evaluated to analyze the confounding factor of biological interference from the release mechanism of cell-free DNA. To compare the characteristic, I approached with different types of cell-free DNA sample obtained from pleural effusion and plasma in matched lung cancer patients. The distribution of size fragmentation from collected cell-free DNA showed the significant differences. The results might support the subset of the biology of cell-free DNA as well as suppressing the biological noise while analyzing the circulating tumor DNA.

MATERIALS AND METHODS

1. Sample collection and DNA isolation

Pancreatic cancer sample

The institutional review board at Samsung Medical Center approved the study (IRB number 2014-04-048-009), and all the methods were carried out in accordance with the approved guidelines. Written informed consent was obtained from all subjects. Newly diagnosed pancreatic ductal adenocarcinoma (PDAC) patients who underwent the endoscopic ultrasound (EUS)-guided fine needle aspiration (FNA) procedure were enrolled and underwent blood draws for cell-free DNA (cfDNA) testing. A total of 120 samples, 17 FNA specimens, 34 peripheral blood leucocytes (PBLs) and 69 plasma samples was profiled from 17 patients. Among them, 17 pairs sequencing data from pretreatment plasma and PBL samples were reported in our recent study that analyzed technical sequencing errors (1).

Lung Cancer sample

The pleural effusion fluid and blood samples were collected from the 19 human subjects. The pleural effusion cell pellet and supernatants were collected separately. Other than that, the analysis and collection were followed with pancreatic cancer samples.

The pretreatment (i.e., before treatment) blood draw of the participants was collected at the time of diagnosis. Whole blood samples were collected in Cell-Free DNA™ BCT tubes (Streck Inc., Omaha, NE, USA). Plasma were prepared with three gradual steps of centrifugation (840g for 10min, 1040g for

10min, and 5000g for 10min at room temperature) while PBLs were drawn from the initial centrifugation. Collected plasma and PBL samples were stored at -80°C until cfDNA extraction. PBL germline DNA (gDNA) was isolated by QIAamp DNA mini kit (Qiagen, Santa Clarita, CA, USA). Plasma DNA was obtained from 2 to 5 mL of plasma via QIAamp Circulating Nucleic Acid Kit (Qiagen). AllPrep DNA/RNA Mini Kit (Qiagen) utilized to purify genomic DNAs from FNA tissues. The concentration and purity of DNA were examined by a Nanodrop 8000 UV-Vis spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) with Qubit dsDNA HS Assay Kit and BR Assay Kit (Thermo Fisher Scientific). The distribution of fragment size as an indicator for DNA degradation were measured using a 2200 TapeStation Instrument (Agilent Technologies, Santa Clara, CA, USA) and real-time PCR Mx3005p (Agilent Technologies) according to the manufacturer's manual.

2. Library preparation and target enrichment

Purified gDNAs were fragmented in a range of 150-200 bp using Covaris S2 (7 min, 0.5% duty, intensity = 0.1, 50 cycles/burst; Covaris Inc. Woburn, MA, USA). The libraries for FNA samples were constructed by following the manufacture's instruction of SureSelect XT reagent kit, HSQ (Agilent Technologies). The libraries for PBL and plasma DNAs were created using KAPA Hyper Prep Kit (Kapa Biosystems, Woburn, MA, USA) as described previously (26). Whereas 200 ng of PBL DNA was used to construct sequencing libraries for all samples, 37.12 ng of plasma DNA was used on average. Briefly, after end repair and A-tailing according to the

manufacturer's protocol, we performed adaptor ligation using a pre-indexed PentAdapter™ (PentaBase ApS, Denmark) at 4°C overnight. After amplification through 9 PCR cycles, the library was analyzed for its quantity and fragment size distribution and then subjected to multiplexing hybrid selection for target enrichment. Hybrid selection was performed by using customized RNA baits that targeted ~499kb of the human genome, including exons from 83 cancer-related genes (Table S1). Purified libraries were pooled up to eight and each pooled library was adjusted to a total of 750ng for a hybrid selection reaction. Target enrichment was performed following the SureSelect (Agilent Technologies) bait hybridization protocol with the modification of replacing the blocking oligonucleotide with IDT xGen blocking oligonucleotide (IDT, Santa Clara, CA, USA) for the pre-indexed adapter. After the target enrichment step, the captured DNA fragments were amplified with 13 cycles of PCR using P5 and P7 oligonucleotides.

3. Sequencing and data processing

Based on DNA concentration and the average fragment sizes, libraries were normalized to an equal concentration of 2 nM and pooled by equal volume. After denaturing libraries using 0.2 N NaOH, libraries were diluted to 20 pM using hybridization buffer (Illumina, San Diego, CA, USA) and subjected to cluster amplification according to the manufacturer's protocol (Illumina). Flow cells were sequenced in the 100-bp paired-end mode using HiSeq 2500 v3 Sequencing-by-Synthesis Kits (Illumina) and then analyzed using RTA v.1.12.4.2 or later. Using BWA-mem (v0.7.5) (27), all of the raw data were

aligned to the hg19 human reference creating BAM files. SAMTOOLS (v0.1.18) (28), Picard (v1.93), and GATK (v3.1.1) (29) were used for sorting SAM/BAM files, local realignment, and duplicate markings, respectively. Through the process, we filtered reads to remove duplicates, discordant pairs, and off-target reads.

4. SNV detection in FNA samples and statistical test for their presence in plasma

For FNA biopsy specimens, MuTect 1.1.4 and VarScan2 were employed to detect somatic single nucleotide variants (SNVs) with matched germline (i.e., PBL) samples. For VarScan2, the default parameter values were used with some modifications as previously described (44). Somatic SNVs called by at least one of the methods were retained if they were present at a frequency less than 0.5% in the matched PBL sample and higher than 4% of the tumor sample. Somatic SNVs found in the FNA samples were listed and tested for their presence in the paired plasma samples as described previously (44). After background alleles in each sample had been adjusted by position-specific error rates, it was tested if the allele frequency of a given SNV ranked in the 95th percentile of adjusted background alleles.

5. Biopsy-free SNV identification in plasma DNA

A method slightly modified from previous studies (44, 48) was established to identify somatic SNVs in the plasma sample as described in our recent study [Park et al. 2017 Oncotarget]. Firstly, positions with the strand bias under 0.9 and the total read depth over 500 were considered for the analysis. All non-reference alleles were subjected to Phred quality filtering using a threshold Q

of 30. Non-reference alleles present at a frequency below 0.5% in the matched germline DNA were subjected to the binomial test to test if a non-reference allele was significantly more abundant in plasma DNA than the matched gDNA. The multiple testing adjustments were made through the Bonferroni correction. Next, Z-tests were performed to compare frequencies of non-reference alleles with their background allele frequency distribution obtained from the other plasma DNA samples. For the comparison, a background allele frequency distribution was generated by selecting non-reference alleles in plasma DNA present at a frequency <2.5% in the paired tumor and <0.5% in the paired germline DNA with a sufficient total depth ($\geq 250\times$ in tumor tissue, $\geq 500\times$ in PBL, and $\geq 500\times$ in plasma). Additionally, the following filters were applied: (1) candidate alleles with less than eight supporting reads were discarded; (2) when there were two or more candidates within any 10 bp window, all of them with an allele frequency <20% were discarded; (3) candidates with the Bonferroni adjusted p-value higher than 10^{-18} from the z-test were discarded. We further excluded SNV candidates if found as a germline SNP in other samples processed in the same lane of a sequencing flowcell to remove false positives due to cross-contamination among multiplexed samples. Nonsynonymous, stop-gain, stop-loss, and splicing-disrupting variants were listed as the final positive calls.

6. Droplet digital PCR validation

Mutant and wild-type alleles in plasma samples were quantified by QX200 Droplet Digital PCR System (BioRad, Hercules, CA, USA). All droplet

digital PCR (ddPCR) reagents except TaqMan assays (i.e., probes/primers) were ordered from Bio-Rad. TaqMan assays for KRAS p.G12D/G12V were ordered from Bio-Rad (PrimePCR ddPCR Mutation Assay), and RB1 p.R251* and ROS1 p.I1967V assays were customized by TaqMan SNP Genotyping Assays (Thermo Fisher Scientific, Waltham, MA, USA). All assay tests were performed in parallel with no-template and wild-type gDNA controls to monitor the false positive droplets. The concentrations of wild-type and mutant DNAs (copies per ul) in each sample were calculated by manufacturer's software and their concentrations in plasma (copies per mL) were derived as describe in van Ginkel, J.H *et al.*(65)

7. Statistical test

To calculate the limits of detection, the group was labelled with “detected”, “not detected”, or “discordant.” Group “detected” was categorized if positive droplet from ddPCR and read counts was determined from cfDNA sequencing and vice versa for group “not detected”. If the results present with discordant manner either by ddPCR or cfDNA sequencing, the “discordant” was labelled and categorized into false positive (See Table 2 and Supplemental Table 4). The calculation of confidence intervals for sensitivity and specificity was evaluated under the exact Clopper-Pearson confidence intervals. In confidence intervals of positive predictive value (PPV), the standard logit confidence intervals were followed as presented in Mercaldo *et al.* (66).

The rest of all statistical significances were evaluated by two-tailed tests, and the significance level was set at 5%. One-way analysis of variance (ANOVA) with least significance difference (LSD) posthoc analysis was used to

compare means across multiple groups. A one-sample t -test was applied to compare hotspot error rates between pretreatment and peri-/post-treatment samples. For i -th patient, j -th peri-/post-treatment sample and k -th hotspot, difference (D_{ijk}) is defined as following;

$$D_{ijk} = X_{ik}^{pre-treatment} - X_{ijk}$$

Null hypothesis was that mean error rates before and after treatment were not different ($H_0 : \bar{D} = 0$ vs. $H_a : \bar{D} \neq 0$).

RESULTS

Evaluation of LOD with single mutation

KRAS mutations

The vast majority of cancer related KRAS mutations in pancreatic cancer. To evaluate the limits of detection with single point mutations, the presence of KRAS mutations in baseline samples from 17 pancreatic cancer patients had evaluated. Among the 17 patients, 13 patients (76.5%) determined to have KRAS mutation from FNA samples. KRAS mutations were detected in 10 patients (58.8%) from plasma samples (Figure 1, Table 2-1). A general clue from the data, the allele frequency of KRAS mutations in plasma DNA is relatively lower than FNA samples. The allele frequency of KRAS mutations were $21.18\% \pm 4.06$ (mean \pm standard error of the mean (SEM)) in FNA samples and $2.02\% \pm 0.67$ (mean \pm SEM) in plasma samples. There can be two reasons: the limits of detection by targeted sequencing or sampling issue. To eliminate the possibility of modest detection sensitivity, orthogonal validation was performed by digital PCR. The overall tested samples were 62 samples which included consecutive samples from 14 patients. Conclusively, the analytical detection sensitivity presented 95.65 % sensitivity (95% confidence interval (CI) 78.05 to 99.89%), 100 % specificity (95% CI 91.24 to 100 %) for detecting KRAS mutations (Figure 2-1, Table 2-1). By these, the technical issue was neglected and remained only to the biological factor.

Evaluation of LOD with multi-mutations

“With primary” mutation

To evaluate the multiple mutations compare to point mutations, the customized capture-based targeted sequencing implied to FNA samples. The total of 40 M_{FNA} (mutations found in FNA sample) was determined via 17 pancreatic cancer patients (Figure 2-1A, Table 2-2). There were failure to detect significant mutations in two patients (P11 and P28). As described in Method, the determined mutations were tested to evaluate the significance of the presence in matched plasma DNA samples ($p < 0.001$). Figure 2-1B represents the total of 28 $M_{\text{P/FNA}}$ (mutations among M_{FNA} detected in their matched plasma samples) were significantly discriminated above the background noises of plasma DNAs, resulting in 70.0% detection sensitivity with the allele frequency (MAF) of $1.60\% \pm 0.31$ (mean \pm SEM) (Table 2-3).

Biopsy-free manner

The genotyping test is important for tracking the alteration of the architecture of genomics of primary tumor, but it does not represent the entire targeted regions. To assess the plasma mutations in broader regions, biopsy-free manner was implemented. The total of 27 $M_{\text{P/TR-BF}}$ in baseline plasma samples including 15 concordantly detected in FNAs (Figure 2-1) with the mean allele frequency of $3.54\% \pm 1.38$ (mean \pm SEM). The unique 12 mutations were detected only in plasma DNA samples but not in the matched FNA sample. To eliminate the possibility of false positive reports, the digital PCR was performed for orthogonal validation. Two individual variants (ROS1 p.I1967V and RB1 p.251X) were selected from the two patients (P2 and P5) and performed digital PCR to validate their presence in 8 consecutive plasma

samples (Table 2-4). Consistent with the cfDNA sequencing results, these mutations were detected in consecutive plasma samples from the patients, indicating the variants unique to plasma were not likely to be false positives. To determine the levels of false positive due to the technical background noises from the “biopsy-free manner” method, the series of biologically replicated sequencing data using PBL DNAs from six patients were evaluated. One of replicates from each patient was paired with the other as a mock for a matched plasma sample and was processed for variant detection (Method). By testing the total of 21 replicates, there were absolutely no mutations were detected from the pipeline. By these result, the minimal false discovery rate was assured by involvement of technical background errors. Collectively, the algorithm of biopsy-free manner is feasible and useful to detect tumor mutations across the entire target regions. Limitations in genetic profiling using FNAs have been recognized as FNAs are not sufficient to represent all regional subclone events. The data suggested that somatic profiling mutations of plasma DNA in a biopsy-free manner compensate the shortcomings of FNA, revealing intra-tumor heterogeneity. Based on $M_{P/TR}$ (plasma DNA mutation across entire target regions) by combining $M_{P/FNA}$ with $M_{P/TR-BF}$, I were able to detect ctDNAs in 15 pretreatment samples suggesting the advantage of profiling broader genomic regions than KRAS hotspots.

Monitoring tumor burden by measuring ctDNA

Another merit of ctDNA analysis is the ability to monitoring the alteration of tumor genomics in real-time. Taking advantage of the serial collection of

blood draws, the responses of chemotherapy and the progression of disease (PD) examined to correlate with the level of ctDNA. Also, the level of CA 19-9 measured alongside during the clinical follow-up. Blood was collected separately for the each of the tests. Figure 2-2 displayed nine patients under the therapeutic intervention except for one patient (P27). P27 detected with no significant variants from both of the primary and plasma samples.

Consecutive samples detected spartial plasma mutations (Table 2-5, Figure 2-3), but the patient had stable diseases (SD) status along the follow-up period.

The diases progression and the therapy responses were determined under the CT images (data not shown). Among the nine patients, four patients (P2, P7, P42, and P43) presented the correlated trend in both $M_{P/TR}$ amount and CA19-9 level throughout the therapeutic intervention. On the other hand, three patient (P5, P31, and P36) had discordant level of CA 19-9 in the limited period of time but amount of ctDNA matched with the CT images. Moreover, in P11, the truncation of TP53 p.E297X was detected at the fifth plasma sampling of P11. After a month later, P11 diagnosed with liver and peritoneum metastasis. It is noteworthy that FNA sampling of P11 failed to determine any of significant mutations (Figure 2-2 and 2-3). In this context, by observing the level of $M_{P/TR}$, the alteration of $M_{P/TR}$ level was on average of 2 months ahead of the CT image changes. Figure 11 displays the overall detected mutations in plasma samples. Overall, the data suggested that tracking the level of MP/TR is better surrogate marker than CA 19-9 accounted as the monitoring of resistance of chemotherapy and/or disease progression.

Diagnostic utility

Figure 2-4 shows the comparison of ctDNA level according to the time of diagnosis (Dx) with RECIST (Response Evaluation Criteria In Solid Tumors) as a group; Complete Response/Partial Response (CR/PR), Stable Disease (SD), and Progressive Disease (PD). Figure 2-4A represent the allele frequencies of $M_{P/KRAS}$ (ANOVA, LSD, p-value = 0.084) and $M_{P/FNA}$ were not significantly different among the disease status (ANOVA, LSD, p-value=0.519, Figure 2-4B). On the other hand, the allele frequencies of $M_{P/TR}$ significantly varied among the disease statuses (ANOVA, LSD, p-value = 0.001, Figure 2-4C). The allele frequencies of $M_{P/TR}$ at the near time of PD was significantly higher (mean \pm SEM: 4.17% \pm 0.93) than those at the time of Dx (mean \pm SEM: 3.54 \pm 1.55%), SD (mean \pm SEM: 1.32 \pm 0.16%) or CR/PR (mean \pm SEM: 1.82 \pm 0.29%) (Figure 2-4C; ANOVA, LSD, p-value = 0.001). The level of CA 19-9 was also evaluated but not significantly followed the patients' disease status accordingly (ANOVA, LSD, p-value = 0.13, Figure 2-4D). These results suggested that the amount and/or allele frequency of ctDNA well indicated real-time disease status compare to the level of CA 19-9.

Next, the number of detected variants was quantified along the therapeutic intervention (Figure 2-5). As the extension of the period of treatment, it is expected to increase the number of detectable mutations as the allele frequencies of $M_{P/TR}$ varied according to the disease status. The number of detected $M_{P/TR}$ significantly different from the disease status

(Figure 2-5B; ANOVA, LSD, p-value = 5.71×10^{-8}). Moreover, the number of MP/FNA and MP/TR significantly decreased at the time of CR/PR compared to Dx, while the number of MP/TR increased at the time of PD.

Interestingly, the number of $M_{P/FNA}$ and $M_{P/TR}$ dropped after the treatment started (1.05 ± 1.11). The number of variants was the lowest (0.73 ± 1.61) at around 4 months after treatment and started to cumulate (up to 1.90 ± 1.03) as treatment period expanded (Figure 2-5 C). Collectively, the mean number of $M_{P/TR}$ per sample significantly increased depending on the duration of chemotherapy treatment (ANOVA, LSD, p-value = 0.004). The results indicated that $M_{P/TR}$ better represented real-time disease status either by allele frequency and/or a number of mutations than $M_{P/FNA}$.

DISCUSSION

In this study, the ctDNA detection methods was evaluated comparing between targeted deep sequencing with the broad-range and KRAS-oriented analysis.

The analysis highlighted the importance of considering broad-scale ctDNA analysis by allowing to characterize not only intra-heterogeneity limited by tumor biopsy but also to monitor the primary mutations which impacted on the diagnostic accuracy along the therapeutic intervention.

KRAS mutations are well-documented as initiating factor for the development of PDAC (67). However, often, the low detection sensitivity in ctDNA analysis fueled the debate about the capability of as its biomarker (60, 62).

Therefore, comparative evaluation of ctDNA detection approaches for PDAC has to done. Among the 17 pretreatment plasma samples, ctDNA detected in ten, twelve, and fifteen samples by profiling $M_{P/KRAS}$, $M_{P/FNA}$, and $M_{P/TR}$, respectively, indicating the advantage of profiling broad genomic regions on sensitivity for cancer detection. Moreover, the improved sensitivity of ctDNA detection subsequently enhanced tumor monitoring by longitudinal cfDNA analysis. For instance, in P5 and P42 patient, although a KRAS mutation was not detectable in not only pretreatment samples but also all following peri/post-treatment samples, the independent variants in other genes were coherently correlated with tumor burden (Figure 2-2). In P2 patient, the level of ROS1 p.I1967V dramatically decreased after surgical operation indicating tumor removal, although KRAS mutation was not detected before and right after the surgery (Figure 2-2A).

Despite its advantages, interrogation of broader genomic regions might result in more false positives especially when performed in a biopsy-free manner. To minimize the false positives, the stringent filtering steps was applied for calling $M_{P/TR-BF}$. Then, the filtering steps during variant calling were adequately established to minimize false discovery rate. Analyzing duplicated PBL gDNA sequencing data, the data showed that false positives due to the technical background were minimal as described in the Results section. In addition, some of $M_{P/TR-BF}$ was validated which were not detected in FNA specimens by dPCR. In present approach, taking advantage of blood sampling strengthened to neglecting the potential interruption from the biological or technological background noises. On the other hand, as the stringent filtering steps perhaps minimized the detection sensitivity. To improve the detection sensitivity, the present study merged the primary mutations and plasma mutations. However, in the future study, the limitation may overcome by adapting the molecular barcoding which will increase the uniqueness of the reads reliability of low read counts of variants.

The present study not only provides a small number of patients but also randomly selected cancer stages that limit the detection sensitivity depended upon the disease stages. Also, the study of design could not approach the “combination assay” with protein biomarkers as the recent study suggested (68). Regardless of number of patients, as if the threshold of CA 19-9 increased to 100 U/mL followed by previous study, three of patients’ data (P23, P31 and P36) cannot be included. Also, it did not affect the detection sensitivity nor compensate the diagnostic status (ANOVA, LSD, p-

value=0.59). Another obvious hurdle is cost of the targeted deep sequencing. It is evident that single mutation analysis cost cheaper and faster analysis. However, emerging evident shows the dramatic reduction of cost of sequencing may balance out in near future.

Although the present data only dealt with the plasma DNA, the cell-free DNA certainly can observe other types of body fluid. The application of the enhanced NGS-method allowed to analysis pleural effusion (PE) fluid DNA. The comparative genomic alterations were discovered in either from PE cfDNA or plasma cfDNA (Figure 2-6). Cumulative number of detected somatic mutations shared the gene feature from the PE and plasma samples. As the traditional PE test depended on the collection of PE cells from PE fluid, the allele frequency of PE cells was evaluated comparatively. Nevertheless, the allele frequency of detected somatic mutations was not correlated between the PE cfDNA and PE cell. Interestingly, the PE cfDNA and plasma cfDNA had higher correlation (Figure 2-7). Next, the detected mutations were compared with the clinical history. It turned out the correlation of clinical history and PE cfDNA had highly matched than the plasma cfDNA (data not shown). The phenomena indicate the bias of sample collection could be contributed throughout the surveillance. In addition, the allele frequencies of PE cells provided the least information compared to plasma and PE cfDNA. Another interesting factor was the PE cfDNA had the mediator role between the plasma and PE cell. The exclusive detection was observed in overall detected mutations (data not shown). The phenomenon may be contributed by the microenvironment of body fluid such as the physical or chemical effect of

the immune cells lead to the apoptosis or necrosis. The assumption could be made through the comparison of fragment size from PE cfDNA and plasma cfDNA (Figure 2-8). Therefore, it is important to show various types of body fluid can compensate the tumor biopsy as its invasive procedure and modest rate of sensitivity. In summary, the interrogation of circulating tumor DNA with targeted deep sequencing would be informative to analyze the unmet needs in cancer research.

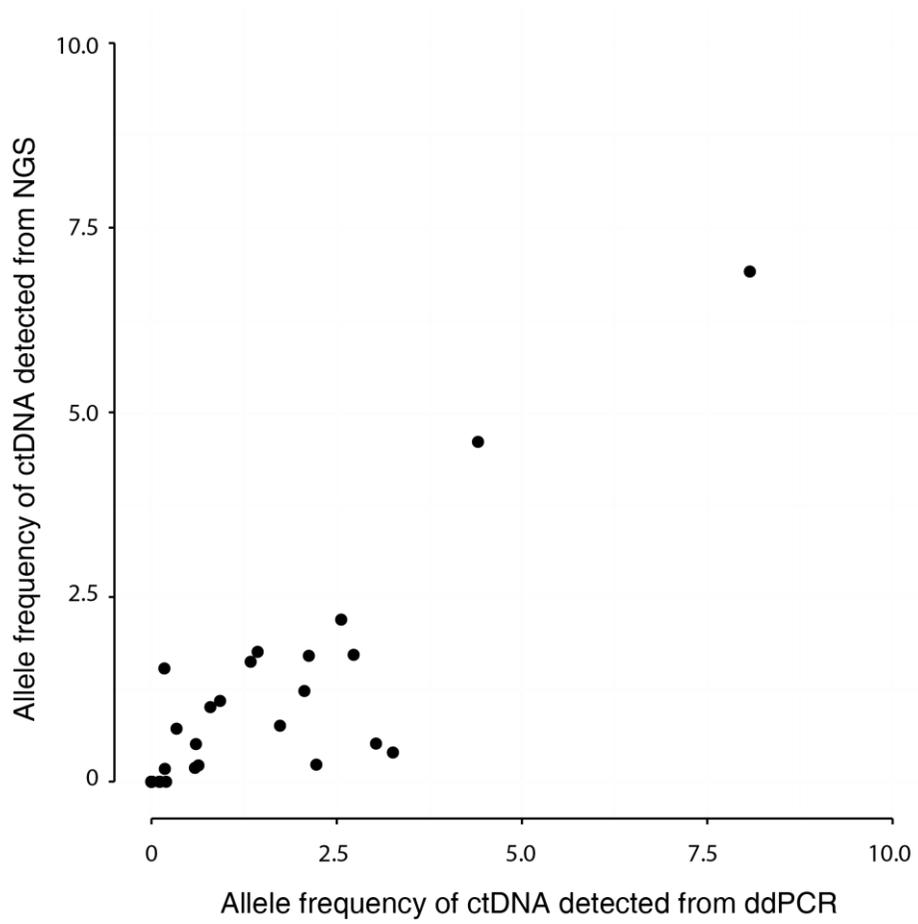


Figure 2-1. The correlation of harbored KRAS mutations using digital PCR and enhance NGS-method from pancreatic cancer patients.

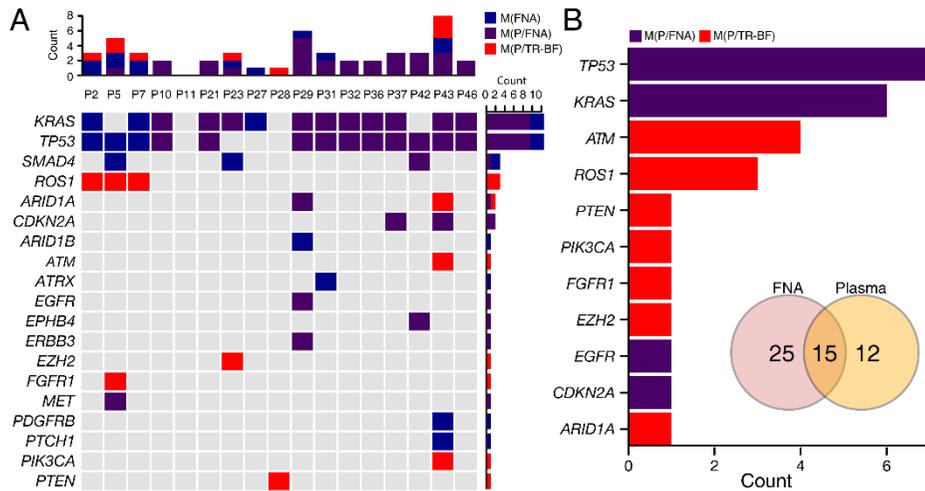
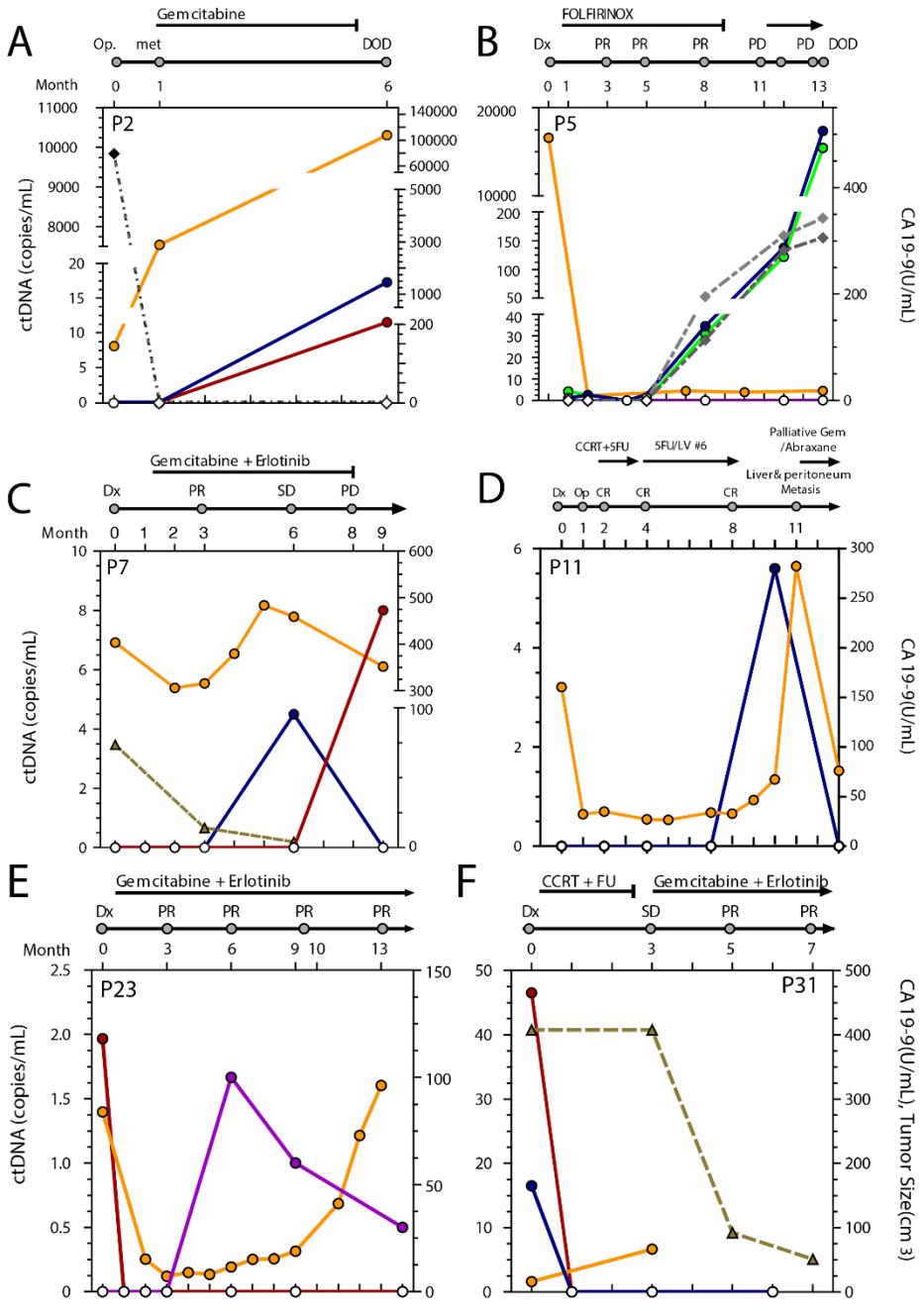


Figure 2-2. Tumor mutations in pre-treatment cfDNA samples from 17 PDAC patients

The top panel summarized the presence of detected mutation across the 17 patients depending on the detection methods (i.e., $M_{P/KRAS}$, $M_{P/FNA}$, $M_{P/TR}$). While interrogation of KRAS hotspots detected mutations ($M_{P/KRAS}$) in plasma samples from 10 patients, testing variants detected from FNA samples ($M_{P/FNA}$) and entire target regions ($M_{P/TR}$) detected tumor variants in 12 and 14 plasma samples, respectively. The oncoprint chart shows M_{FNA} and $M_{P/TR}$. If a variant is concordantly detected in both M_{FNA} and $M_{P/TR}$, the variant also corresponds to $M_{P/FNA}$. The number of affected genes for each patient is plotted the bottom of the chart. The number of samples that harbor a mutation for each gene is plotted the right side of the chart. *Four independent mutations in ATM were detected in the P43 plasma sample.



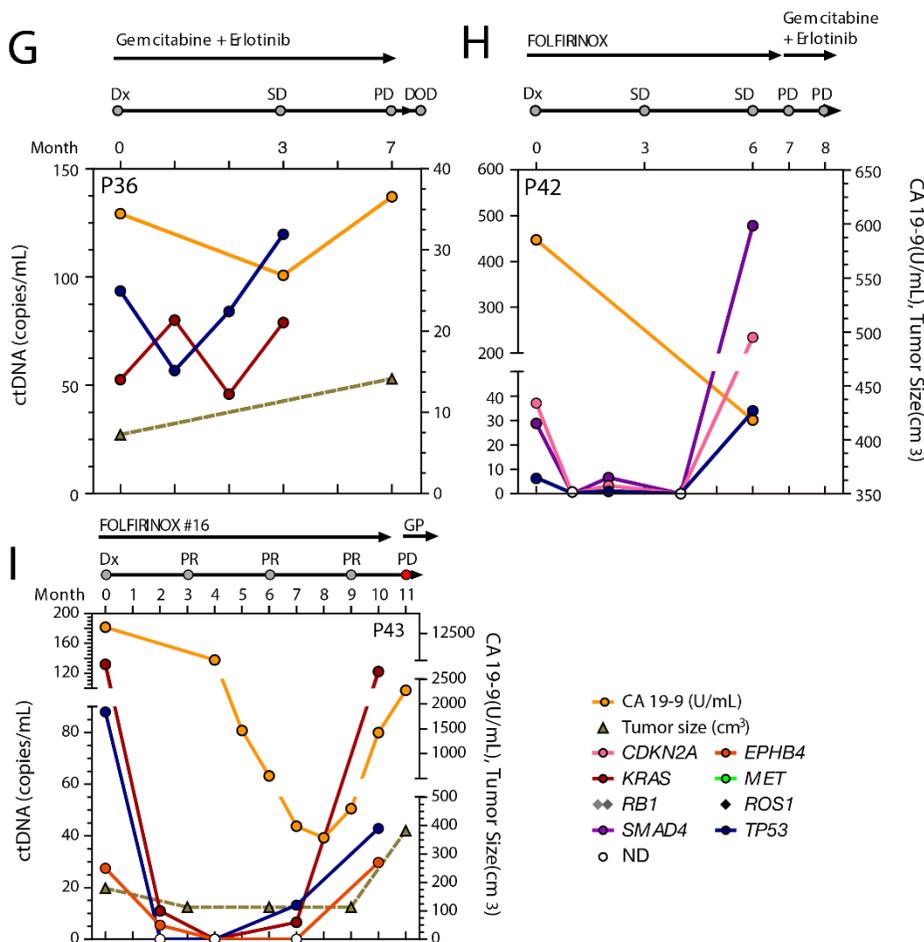


Figure 2-3. Monitoring of ctDNA PDAC patients under therapeutic intervention.

The level of ctDNA estimated by each SNV was plotted on the left y-axis for eight patients (A-D). Chemotherapeutic agents administered to each patient and therapy response evaluated based on Response Evaluation Criteria In Solid Tumors (RECIST) were displayed on top of the graph. The level of ctDNA determined either by MP/FNA (solid line) or MP/TR-BF (dotted line) was displayed in various color depending on the mutated genes. CA 19-9 level (yellow solid line) and tumor size (grey dotted line) based on CT images were plotted against the right y-axis. CCRT, concurrent chemoradiation therapy; FU, fluorouracil; CR, complete response; DOD, dead of disease; Dx, diagnosis; Met, Metastasis; ND, not detected; Op, operation; PD, progressive disease; PR, partial response; SD, stable disease.

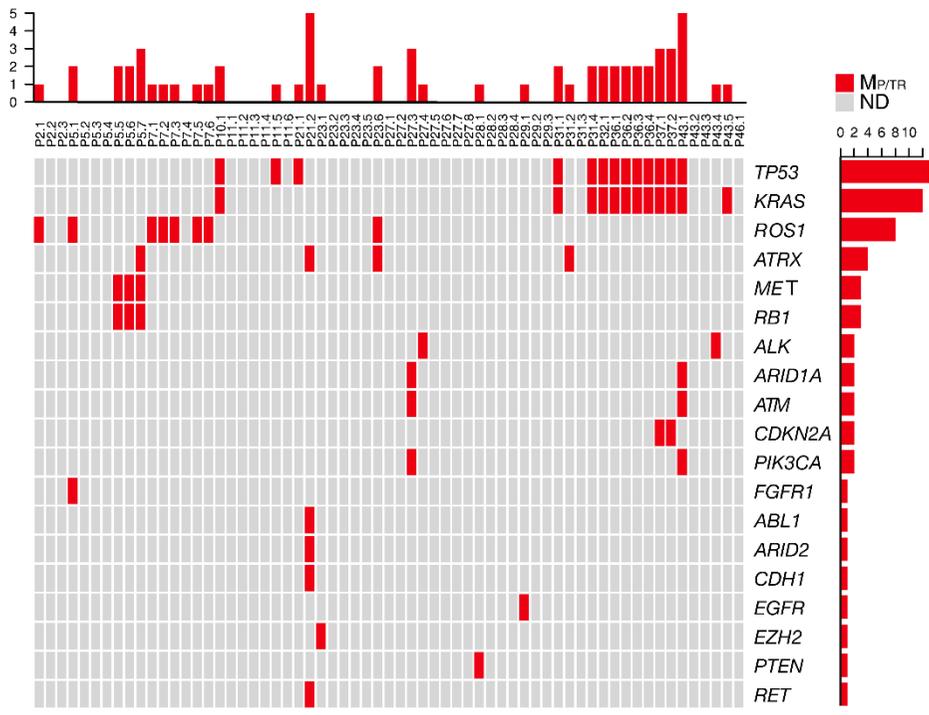


Figure 2-4. Summary of plasma mutations determined by “biopsy-free manner.”
 Total of 19 genes was determined and ordered by number of detected mutation per patient.

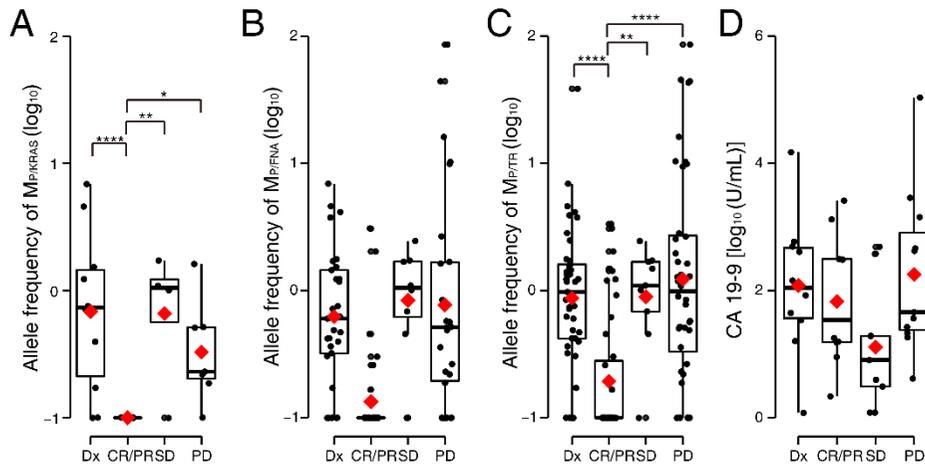


Figure 2-5. Allelic fraction of ctDNA and CA19-9 level depending on therapy responses.

The allelic frequencies of (A) $M_{P/KRAS}$ (B) $M_{P/FNA}$ and (C) $M_{P/TR}$ were box-plotted depending on their near-time therapy response evaluations. (D) CA 19-9 levels were box-plotted. All of the determined levels were displayed on a logarithmic scale. The level of statistical significance is indicated by the asterisks in the figures; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, and **** $P \leq 0.0001$. Dx, diagnosis; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

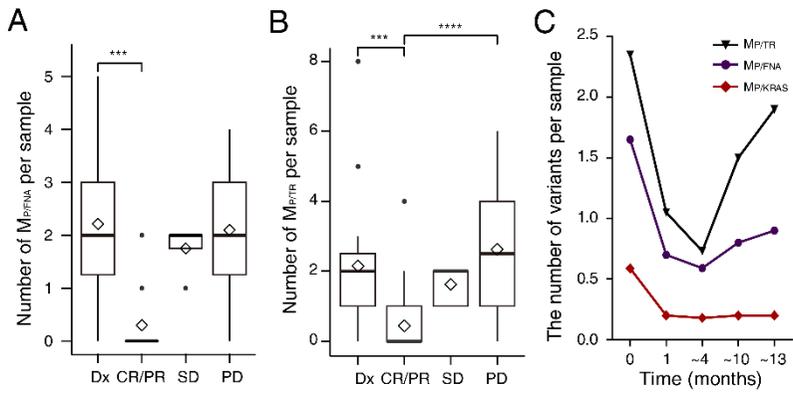


Figure 2-6. The number of mutations in plasma DNA

The number of (A) $M_{P/FNA}$ and (B) $M_{P/TR}$ presented per patient samples was categorized according to near-time disease status. (C) The number of mutations was shown depending on the period of treatment. The level of statistical significance is indicated by the asterisks in the figures; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, and **** $P \leq 0.0001$. Dx, diagnosis; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

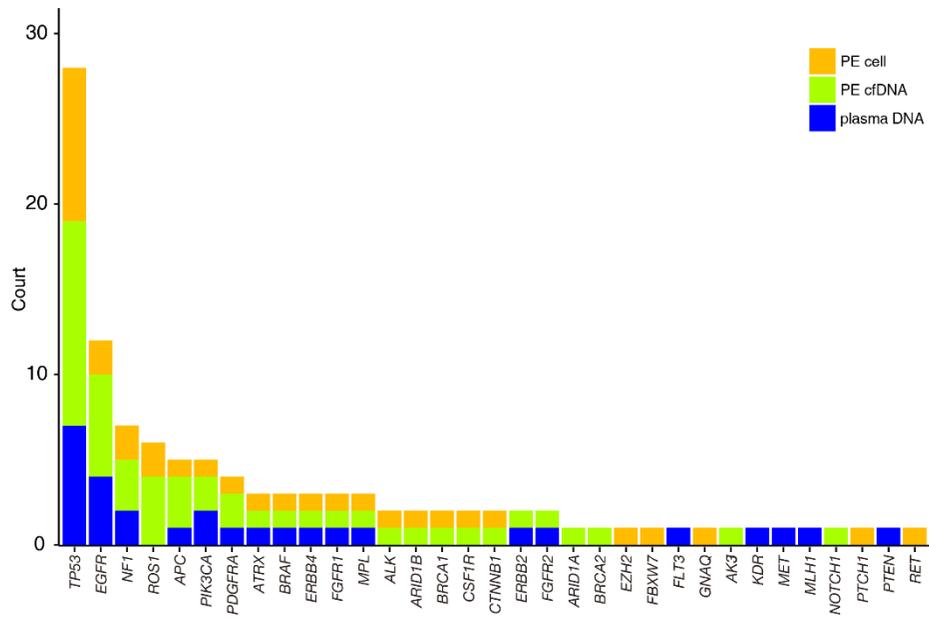


Figure 2-7. Distribution of detected genes from pleural effusion fluid and plasma DNA

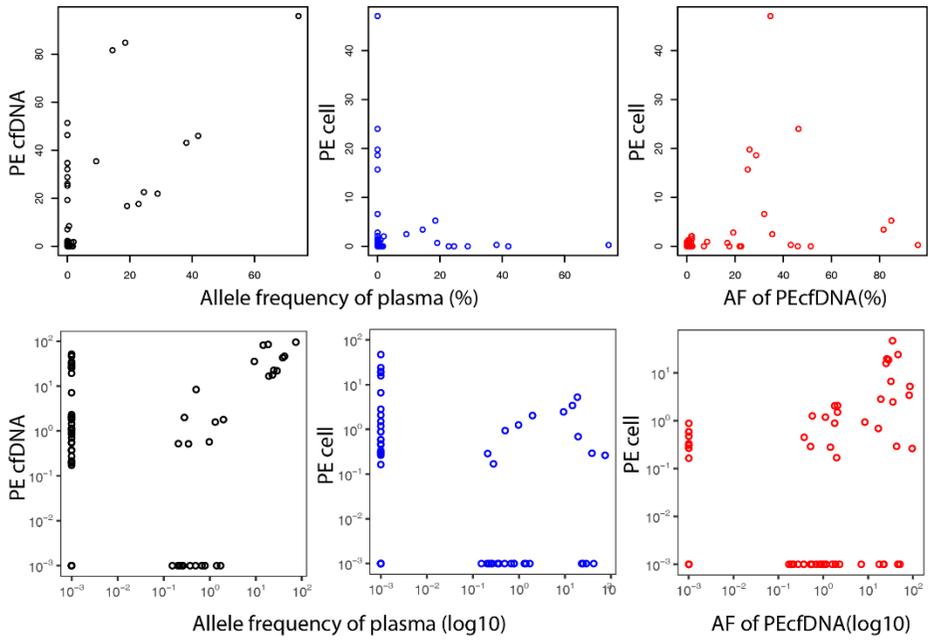


Figure 2-8. The differences of allele frequencies from pleural effusion fluid and plasma DNA

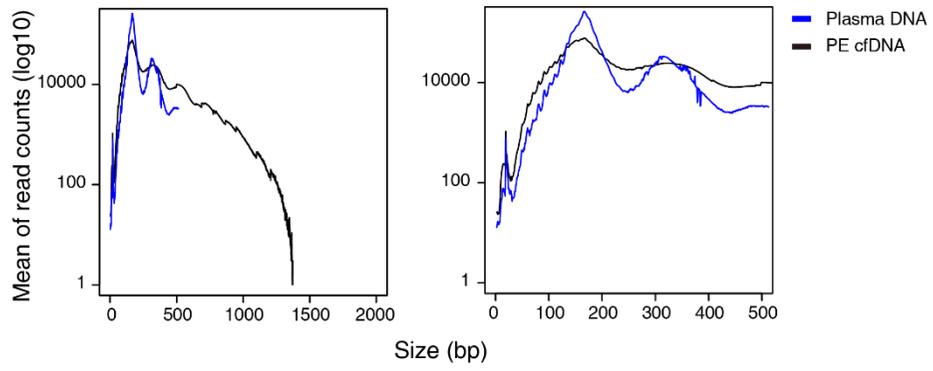


Figure 2-9. The size distribution of pleural effusion and plasma DNA

Table 2-1. The limits of detection sensitivity evaluated by KRAS mutations in 14 PDAC patients.

Patient ID	Gene	AA Change	dpcr	NGS
P2.1	KRAS	G12D	0	0
P2.1	KRAS	G12V	0	0
P2.2	KRAS	G12D	1	1
P2.3	KRAS	G12D	1	1
P5.1	KRAS	G12D	0	0
P5.1	KRAS	G12V	0	0
P7.1	KRAS	G12V	0	0
P7.2	KRAS	G12V	0	0
P7.3	KRAS	G12V	0	0
P7.4	KRAS	G12V	0	0
P7.5	KRAS	G12D	0	0
P7.5	KRAS	G12V	0	0
P7.6	KRAS	G12V	1	1
P10.1	KRAS	G12D	1	1
P10.1	KRAS	G12V	0	0
P21.1	KRAS	G12V	1	1
P21.2	KRAS	G12V	1	1
P23.1	KRAS	G12V	1	1
P23.1	KRAS	G12D	0	0
P23.2	KRAS	G12D	0	0
P23.3	KRAS	G12V	0	0
P23.4	KRAS	G12V	0	0
P23.4	KRAS	G12D	0	0
P23.5	KRAS	G12V	0	0
P23.6	KRAS	G12V	0	0
P23.7	KRAS	G12V	0	0
P27.2	KRAS	G12V	0	0
P27.2	KRAS	G12D	0	0
P27.3	KRAS	G12D	0	0
P27.4	KRAS	G12D	0	0
P27.5	KRAS	G12D	0	0
P27.6	KRAS	G12D	0	0
P27.7	KRAS	G12D	0	0

P29.1	KRAS	G12V	1	1
P29.2	KRAS	G12V	1	1
P29.3	KRAS	G12V	1	1
P31.1	KRAS	G12D	1	1
P31.2	KRAS	G12D	1	1
P31.3	KRAS	G12D	1	1
P31.4	KRAS	G12D	1	1
P32.1	KRAS	G12D	1	1
P32.1	KRAS	G12V	1	0
P36.1	KRAS	G12D	1	1
P36.2	KRAS	G12D	0	0
P36.3	KRAS	G12D	0	0
P36.4	KRAS	G12D	0	0
P37.1	KRAS	G12D	1	1
P37.1	KRAS	G12V	0	0
P37.2	KRAS	G12D	1	1
P43.1	KRAS	G12D	1	1
P43.2	KRAS	G12D	1	1
P43.2	KRAS	G12V	0	0
P43.3	KRAS	G12D	0	0
P43.3	KRAS	G12V	0	0
P43.4	KRAS	G12D	1	1
P46.1	KRAS	G12D	1	1
P46.1	KRAS	G12V	0	0

Table 2-2. Determined mutations from 17 FNA samples

Patient #	DNA	Gene	AA change	Primary VAF (%)
P2	FNA	KRAS	p.G12D	21.12
P2	FNA	TP53	p.Y236C	25.36
P5	FNA	MET	p.S907F	33.56
P5	FNA	SMAD4	p.Q256X	9.27
P5	FNA	TP53	c.96+1G>A	65.62
P7	FNA	KRAS	p.G12V	8.73
P7	FNA	TP53	p.R175H	8.45
P10	FNA	KRAS	p.G12D	13.94
P10	FNA	TP53	c.386+1G>A	14.90
P21	FNA	KRAS	p.G12V	14.69
P21	FNA	TP53	p.I30S	20.77
P23	FNA	KRAS	p.G12V	16.37
P23	FNA	SMAD4	p.R361C	23.29
P27	FNA	KRAS	p.G12D	3.55
P29	FNA	KRAS	p.G12V	4.47
P29	FNA	ARID1B	p.Q2092X	3.79
P29	FNA	EGFR	p.K189E	4.73
P29	FNA	ERBB3	p.P583S	4.15
P29	FNA	TP53	p.L194R	4.75
P31	FNA	KRAS	p.G12D	38.64
P31	FNA	ATRX	p.A3T	6.06
P31	FNA	TP53	p.F113C	54.45
P32	FNA	KRAS	p.G12D	19.73
P32	FNA	TP53	p.H154R	27.96
P36	FNA	KRAS	p.G12D	16.85
P36	FNA	TP53	p.Y220C	23.65
P37	FNA	CDKN2A	p.C100X	60.99
P37	FNA	KRAS	p.G12D	55.64
P37	FNA	TP53	p.R119L	59.97
P42	FNA	TP53	p.L226P	3.36
P42	FNA	SMAD4	p.R361C	4.50
P42	FNA	EPHB4	p.F404L	9.34
P43	FNA	KRAS	p.G12D	28.57
P43	FNA	CDKN2A	p.A97V	23.08

P43	FNA	TP53	p.D220V	29.62
P46	FNA	PDGFRB	p.P866S	17.08
P46	FNA	PTCH1	c.3606+1G>A	21.65
P46	FNA	KRAS	p.G12D	33.09
P46	FNA	PDGFRB	p.L865F	18.26
P46	FNA	TP53	p.Q167X	25.07

Table 2-3. Evaluation of FNA mutations in baseline plasma DNA samples

PlasmaID	Gene	AA Change	Allele Freq.	p-value
P5.1	MET	p.S907F	0.31	0.000325
P10.1	KRAS	p.G12D	2.19	0.000207
P10.1	TP53	c.386+1G>A	4.11	0.000206
P21.1	KRAS	p.G12V	1.23	0.000187
P21.1	TP53	p.I30S	1.70	0.000185
P23.1	KRAS	p.G12V	0.17	0.000700
P29.1	ARID1B	p.Q2092X	0.32	0.000310
P29.1	EGFR	p.K189E	1.23	0.000212
P29.1	ERBB3	p.P583S	0.47	0.000235
P29.1	TP53	p.L194R	0.42	0.000242
P29.1	KRAS	p.G12V	0.40	0.000251
P31.1	KRAS	p.G12D	6.90	0.000197
P31.1	TP53	p.F74C	3.75	0.000198
P32.1	KRAS	p.G12D	1.70	0.000209
P32.1	TP53	p.H154R	1.27	0.000211
P36.1	KRAS	p.G12D	0.76	0.000193
P36.1	TP53	p.Y88C	1.17	0.000192
P37.1	CDKN2A	p.C100X	1.77	0.000186
P37.1	KRAS	p.G12D	4.60	0.000182
P37.1	TP53	p.R119L	4.14	0.000184
P42.1	TP53	p.L226P	0.36	0.000258
P42.1	SMAD4	p.R361C	0.42	0.000221
P42.1	EPHB4	p.F404L	0.49	0.000191
P43.1	KRAS	p.G12D	1.53	0.000198
P43.1	CDKN2A	p.A97V	0.60	0.000229
P43.1	TP53	p.D220V	1.45	0.000198
P46.1	KRAS	p.G12D	0.72	0.000187
P46.1	TP53	p.Q167X	0.62	0.000195

*P31.1 ATRX discarded according to its insufficient depth coverage

Table 2-4. The performance of droplet digital PCR in plasma samples

Quantification of copies/mL of plasma DNA was calculated by following (Method derived by Ginkel et al. 2017):

$$Px = Cx * RV * \frac{EV}{TV}$$

Pmt = Mutant concentration in plasma (copies/mL); Pwt = Wild type concentration in plasma (copies/mL)

Cmt = Mutant sample concentration (copies/uL); Cwt = Wild type sample concentration (copies/uL)

RV = Total reaction volume of digital PCR (20uL)

EV= Elution volume of cfDNA (50-75uL)

TV = cfDNA volume used for dPCR reaction (8uL)

PV= Plasma volume for cfDNA extraction (1-5mL)

Patient #	Gene	AA Change	Input DNA (ng)	P V	RV (uL)	EV (uL)	TV (uL)	Pmt (copies/mL)	Pwt (copies/mL)	PCmt/PCwt (%)	Cmt (Droplet)	Cwt (Droplet)	Total droplets	Cwt (copies/ul)	Cwt (copies/20ul)	Cmt (copies/ul)	Cmt (copies/20ul)
P2.1	KRAS	G12D	5.55	3	20	70	8	0	3063	0.00	0	806	806	53	1050	0	0
P2.1	KRAS	G12V	5.55	3	20	70	8	0	2952	0.00	0	740	740	51	1012	0	0
P2.1	ROS1	I1967V	5.55	3	20	70	8	181	3710	4.87	44	868	912.00	63.6	1272	3	62
P2.2	KRAS	G12D	3.60	3	20	70	8	5	4433	0.12	1	2952	2953	76	1520	0	2
P2.3	KRAS	G12D	8.80	2	20	70	8	140	23888	0.59	19	1240	1259	273	5460	2	32
P5.1	KRAS	G12D	8.32	2	20	70	8	0	10588	0.00	0	1639	1639	121	2420	0	0
P5.1	KRAS	G12V	8.32	2	20	70	8	0	10588	0.00	0	1608	1608	121	2420	0	0
P5.1	RB1	R251*	8.32	2	20	70	8	17	23100	0.07	2	2478	2480.00	264	5280	0	4
P5.2	RB1	R251*	11.12	3	20	70	8	58	13883	0.42	12	2503	2515.00	238	4760	1	20
P5.3	RB1	R251*	21.60	2	20	70	8	7	35350	0.02	1	4559	4560.00	404	8080	0	2
P5.4	RB1	R251*	41.92	5	20	70	8	6	35000	0.02	2	8413	8415.00	1000	20000	0	3
P5.5	RB1	R251*	11.92	5	20	70	8	273	10780	2.53	104	3610	3714.00	308	6160	8	156
P5.6	RB1	R251*	13.44	5	20	70	8	952	7700	12.36	338	2513	2851.00	220	4400	27	544
P5.7	RB1	R251*	106.00	5	20	70	8	24850	33880	73.35	6751	8359	15110.00	968	19360	710	14200

P7.1	KRAS	G12V	10.32	3	20	70	8	0	8575	0.00	0	1928	1928	147	2940	0	0
P7.2	KRAS	G12V	12.80	5	20	70	8	0	6020	0.00	0	2016	2016	172	3440	0	0
P7.3	KRAS	G12V	15.68	3	20	70	8	0	15167	0.00	0	3335	3335	260	5200	0	0
P7.4	KRAS	G12V	9.52	2	20	70	8	0	12950	0.00	0	1944	1944	148	2960	0	0
P7.5	KRAS	G12V	6.27	2	20	70	8	0	6694	0.00	0	907	907	77	1530	0	0
P7.5	KRAS	G12D	6.27	2	20	70	8	0	6746	0.00	0	985	985	77	1542	0	0
P7.6	KRAS	G12V	26.72	5	20	70	8	32	5250	0.60	10	1717	1727	150	3160	1	18
P10.1	KRAS	G12D	9.20	3	20	70	8	181	7058	2.56	34	1243	1277	121	2420	3	62
P10.1	KRAS	G12V	9.20	3	20	70	8	0	8575	0.00	0	2015	2015	147	2940	0	0
P21.1	KRAS	G12V	8.48	2	20	70	8	219	10588	2.07	29	1318	1347	121	2420	3	50
P21.2	KRAS	G12V	2.82	2	20	70	8	105	4716	2.23	16	688	704	54	1078	1	24
P23.1	KRAS	G12V	2.32	2	20	70	8	41	22750	0.18	6	3004	3010	260	5200	0	9
P23.1	KRAS	G12D	2.32	2	20	70	8	0	2336	0.00	0	449	449	27	534	0	0
P23.2	KRAS	G12V	3.17	2	20	70	8	0	114	0.00	0	252	252.00	1.3	26	0	0
P23.2	KRAS	G12D	3.17	2	20	70	8	0	1733	0.00	0	265	265	20	396	0	0
P23.3	KRAS	G12V	7.76	2	20	70	8	0	12250	0.00	0	1529	1529	140	2800	0	0
P23.4	KRAS	G12V	2.59	2	20	70	8	0	3789	0.00	0	440	440	43	866	0	0
P23.4	KRAS	G12D	1.82	1	20	50	8	0	1675	0.00	0	184	184	13	268	0	0
P23.5	KRAS	G12V	28.64	5	20	70	8	0	7175	0.00	0	2351	2351	205	4100	0	0
P23.6	KRAS	G12V	11.04	5	20	70	8	0	8120	0.00	0	2776	2776	232	4640	0	0
P23.7	KRAS	G12V	102.0 0	5	20	70	8	0	9135	0.00	0	3010	3010	261	5220	0	0
P27.1	KRAS	G12D	1.25	5	20	70	8	0	676	0.00	0	236	236.00	19.3	386	0	0
P27.1	KRAS	G12V	1.25	5	20	70	8	0	546	0.00	0	183	183.00	15.6	314	0	0

P27.2	KRAS	G12D	4.22	5	20	70	8	0	1379	0.00	0	443	443	39	788	0	0
P27.2	KRAS	G12V	3.18	5	20	70	8	0	1281	0.00	0	468	468	37	732	0	0
P27.3	KRAS	G12D	9.20	5	20	70	8	0	3780	0.00	0	1100	1100	108	2160	0	0
P27.4	KRAS	G12D	8.32	5	20	70	8	0	3710	0.00	0	1096	1096	106	2120	0	0
P27.5	KRAS	G12D	4.99	5	20	70	8	0	1869	0.00	0	679	679	53	1068	0	0
P27.6	KRAS	G12D	9.76	5	20	70	8	0	1246	0.00	0	452	452	36	712	0	0
P27.6	KRAS	G12V	2.37	5	20	70	8	0	963	0.00	0	342.00	342.00	27.50	550.00	0	0
P27.7	KRAS	G12D	2.48	5	20	70	8	0	1106	0.00	0	468	468	32	632	0	0
P27.8	KRAS	G12D	1.25	5	20	70	8	0	546	0.00	0	209	209.00	15.6	312	0	0
P29.1	KRAS	G12V	4.19	2	20	70	8	114	3491	3.26	15	464	479	40	798	1	26
P29.2	KRAS	G12V	7.22	2	20	70	8	45	7018	0.64	7	1063	1070	80	1604	1	10
P29.3	KRAS	G12V	3.44	2	20	70	8	70	2310	3.03	10	309	319	26	528	1	16
P31.1	KRAS	G12D	11.12	2	20	70	8	114	1409	8.07	15	191	206	16	322	1	26
P31.2	KRAS	G12D	26.24	2	20	70	8	0	7263	0.00	0	979	979	83	1660	0	0
P31.3	KRAS	G12D	30.88	2	20	70	8	0	5023	0.00	0	637	637	57	1148	0	0
P31.4	KRAS	G12D	4.11	2	20	70	8	0	4506	0.00	0	588	588	52	1030	0	0
P32.1	KRAS	G12D	6.00	2	20	70	8	123	5766	2.12	18	808	826	66	1318	1	28
P32.1	KRAS	G12V	6.00	2	20	70	8	5	5110	0.10	1	909	910	58	1168	0	1
P36.1	KRAS	G12D	3.28	2	20	70	8	184	10588	1.74	23	1284	1307	121	2420	2	42
P36.2	KRAS	G12D	1.86	2	20	70	8	58	7263	0.80	8	978	986	83	1660	1	13
P36.3	KRAS	G12D	15.84	5	20	70	8	19	2002	0.93	6	635	641	57	1144	1	11
P36.4	KRAS	G12D	11.44	5	20	70	8	49	1796	2.73	16	585	601	51	1026	1	28
P37.1	KRAS	G12D	5.23	3	20	70	8	134	3045	4.41	31	691	722	52	1044	2	46
P37.1	KRAS	G12V	5.23	3	20	70	8	0	3127	0.00	0	842	842	54	1072	0	0

P37.2	KRAS	G12D	6.22	3	20	70	8	49	3424	1.43	11	746	757	59	1174	1	17
P43.1	KRAS	G12D	2.66	3	20	70	8	5	2678	0.17	1	533	534	46	918	0	2
P43.2	KRAS	G12D	2.45	5	20	70	8	2	1064	0.20	1	608	609	30	608	0	1
P43.2	KRAS	G12V	3.54	5	20	70	8	0	1029	0.00	0	362	362	29	588	0	0
P43.3	KRAS	G12D	2.86	5	20	70	8	0	1302	0.00	0	488	488	37	744	0	0
P43.3	KRAS	G12V	4.26	5	20	70	8	0	1061	0.00	0	347	347	30	606	0	0
P43.4	KRAS	G12D	5.50	5	20	70	8	28	2093	1.34	10	723	733	60	1196	1	16
P46.1	KRAS	G12D	2.54	2	20	70	8	6	1820	0.34	1	985	986	21	416	0	1
P46.1	KRAS	G12V	2.54	2	20	70	8	0	2774	0.00	0	496	496	32	634	0	0

Table 2-5. The list of somatic mutations detected in plasma samples by biopsy-free manner

Sample ID	Chr	Position	Ref	Alt	Function	Gene	Exonic function	AA Change	Ref#	Read Count	Total read	Allele frequency	p-value	Total average depth
P2.1	chr6	117641072	T	C	exonic	ROS1	MISSENSE	p.I1967V	NM_002944	1252	3246	38.57	0E+00	2531
P5.1	chr6	117609741	G	A	exonic	ROS1	TRUNC	p.Q2320X	NM_002944	20	2696	0.74	4E-21	2556
P5.1	chr8	38277157	G	A	exonic	FGFR1	MISSENSE	p.S385L	NM_001174064	20	1494	1.34	3E-26	2816
P7.1	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	29	2974	0.98	5E-35	2689
P10.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	30	1367	2.19	4E-46	817
P10.1	chr17	7577498	C	T	splicing	TP53	TRUNC	c.782+1G>A	NM_001126113	48	1167	4.11	2E-88	843
P21.1	chr17	7578445	A	C	exonic	TP53	MISSENSE	p.I162S	NM_001126113	41	2410	1.70	3E-60	1824
P23.1	chr7	148512600	T	C	exonic	EZH2	MISSENSE	p.K515R	NM_004456	19	2243	0.85	4E-21	1887
P28.1	chr10	89720875	G	T	exonic	PTEN	MISSENSE	p.K342N	NM_000314	39	998	3.91	1E-70	856
P29.1	chr7	55218992	A	G	exonic	EGFR	MISSENSE	p.K189E	NM_005228	28	2268	1.23	1E-36	2554
P36.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	18	2384	0.76	5E-19	1807
P36.1	chr17	7578190	T	C	exonic	TP53	MISSENSE	p.Y220C	NM_001126113	32	2724	1.17	2E-41	2018
P32.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	44	2586	1.70	1E-64	2045
P32.1	chr17	7578271	T	C	exonic	TP53	MISSENSE	p.H193R	NM_001126113	31	2441	1.27	3E-41	1906
P31.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	93	1347	6.90	4E-195	2393
P31.1	chr17	7579349	A	C	exonic	TP53	MISSENSE	p.F113C	NM_001126113	33	880	3.75	9E-59	1499
P37.1	chr9	21971101	G	T	exonic	CDKN2A	TRUNC	p.C100X	NM_058195	22	1246	1.77	1E-31	1355
P37.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	90	1956	4.60	4E-173	2045

P37.1	chr17	7578457	C	A	exonic	TP53	MISSENSE	p.R158L	NM_001126113	59	1425	4.14	3E-110	1812
P43.1	chr1	27102188	A	G	exonic	ARID1A	MISSENSE	p.N1705S	NM_006015	43	1517	2.83	4E-72	2006
P43.1	chr3	178927410	A	G	exonic	PIK3CA	MISSENSE	p.I391M	NM_006218	26	1619	1.61	3E-36	2063
P43.1	chr11	108106443	T	A	exonic	ATM	MISSENSE	p.D126E	NM_000051	29	1171	2.48	4E-46	1692
P43.1	chr11	108121733	G	A	exonic	ATM	MISSENSE	p.G514D	NM_000051	19	1620	1.17	3E-23	1930
P43.1	chr11	108143456	C	G	exonic	ATM	MISSENSE	p.P1054R	NM_000051	22	1649	1.33	2E-28	1918
P43.1	chr11	108159732	C	T	exonic	ATM	MISSENSE	p.H1380Y	NM_000051	23	1648	1.40	2E-30	2115
P43.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	24	1566	1.53	9E-33	2045
P43.1	chr17	7577505	T	A	exonic	TP53	MISSENSE	p.D259V	NM_001126113	16	1105	1.45	1E-20	1447
P5.5	chr7	116409781	C	T	exonic	MET	MISSENSE	p.S889F	NM_000245	62	3057	2.03	2E-96	2117
P5.5	chr13	48919281	C	G	exonic	RB1	TRUNC	p.S149X	NM_000321	56	3624	1.55	2E-80	2311
P5.5	chr13	48936983	C	T	exonic	RB1	TRUNC	p.R251X	NM_000321	105	3357	3.13	2E-184	2524
P5.5	chr17	7579699	C	T	splicing	TP53	TRUNC	exon4:c.96+1G>A	NM_001126113	69	2246	3.07	4E-120	1697
P5.6	chr7	116409781	C	T	exonic	MET	MISSENSE	p.S889F	NM_000245	275	2781	9.89	0E+00	2073
P5.6	chr13	48919281	C	G	exonic	RB1	TRUNC	p.S149X	NM_000321	300	3188	9.41	0E+00	2168
P5.6	chr13	48936983	C	T	exonic	RB1	TRUNC	p.R251X	NM_000321	358	3447	10.39	0E+00	2466
P5.6	chr17	7579699	C	T	splicing	TP53	TRUNC	exon4:c.96+1G>A	NM_001126113	311	1928	16.13	0E+00	1777
P5.7	chr7	116409781	C	T	exonic	MET	MISSENSE	p.S889F	NM_000245	2057	4675	44.00	0E+00	1165
P5.7	chr13	48919281	C	G	exonic	RB1	TRUNC	p.S149X	NM_000321	2191	5092	43.03	0E+00	1073
P5.7	chr13	48936983	C	T	exonic	RB1	TRUNC	p.R251X	NM_000321	2677	5907	45.32	0E+00	1100

P5.7	chr17	7579472	G	C	exonic	TP53	MISSENSE	p.P72R	NM_001126113	34	3128	1.09	6E-43	1175
P5.7	chr17	7579699	C	T	splicing	TP53	TRUNC	exon4:c.96+1G>A	NM_001126113	2313	2696	85.79	0E+00	1345
P5.7	chrX	76855029	T	C	exonic	ATRX	MISSENSE	p.K1936R	NM_000489	33	6527	0.51	7E-31	1167
P7.2	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	40	2806	1.43	1E-55	2689
P7.3	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	34	2770	1.23	1E-44	2689
P7.5	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	45	3031	1.48	2E-63	2404
P7.6	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	40	3137	1.28	1E-53	2754
P11.5	chr17	7574021	C	A	exonic	TP53	TRUNC	p.E297X	NM_001276761	28	1742	1.61	1E-39	2176
P21.2	chr9	133759935	G	T	exonic	ABL1	MISSENSE	p.G772V	NM_007313	17	1891	0.90	4E-19	1616
P21.2	chr10	43610119	G	A	exonic	RET	MISSENSE	p.G691S	NM_020975	29	1141	2.54	6E-47	1595
P21.2	chr12	46246206	G	T	exonic	ARID2	MISSENSE	p.A1434S	NM_152641	28	2165	1.29	4E-37	2150
P21.2	chr16	68857389	A	G	exonic	CDH1	MISSENSE	p.K675R	NM_004360	38	1943	1.96	8E-58	1872
P21.2	chrX	76938923	G	C	exonic	ATRX	MISSENSE	p.P609A	NM_000489	30	1065	2.82	5E-50	1530
P23.6	chr6	117642495	C	T	exonic	ROS1	MISSENSE	p.E1902K	NM_002944	19	2065	0.92	1E-21	1072
P23.6	chrX	76938208	A	G	exonic	ATRX	MISSENSE	p.F847S	NM_000489	24	3887	0.62	9E-24	741
P27.3	chr1	27102188	A	G	exonic	ARID1A	MISSENSE	p.N1705S	NM_006015	35	3306	1.06	5E-44	2077
P27.3	chr3	178927410	A	G	exonic	PIK3CA	MISSENSE	p.I391M	NM_006218	19	2986	0.64	6E-19	1831
P27.3	chr11	108106443	T	A	exonic	ATM	MISSENSE	p.D126E	NM_000051	23	2498	0.92	6E-27	1380
P27.4	chr2	29917793	C	T	exonic	ALK	MISSENSE	p.R292H	NM_004304	21	3837	0.55	9E-20	2255
P28.2	chr9	98224138	C	A	exonic	PTCH1	MISSENSE	p.Q835H	NM_001083602	383	2835	13.51	0E+00	1588

P28.2	chr10	89720875	G	T	exonic	PTEN	MISSENSE	p.K342N	NM_000314	34	4033	0.84	3E-39	856
P31.2	chrX	76938208	A	G	exonic	ATRX	MISSENSE	p.F847S	NM_000489	18	540	3.33	4E-30	2194
P36.2	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	24	2378	1.01	5E-29	1807
P36.3	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	18	1646	1.09	9E-22	1807
P36.3	chr17	7578190	T	C	exonic	TP53	MISSENSE	p.Y220C	NM_001126113	33	1962	1.68	1E-47	2018
P36.4	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	31	1804	1.72	6E-45	2365
P36.4	chr17	7578190	T	C	exonic	TP53	MISSENSE	p.Y220C	NM_001126113	47	1919	2.45	2E-76	2498
P37.2	chr9	21971101	G	T	exonic	CDKN2A	TRUNC	p.C100X	NM_058195	17	1472	1.15	5E-21	1355
P37.2	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	40	2275	1.76	3E-59	2045
P37.2	chr17	7578457	C	A	exonic	TP53	MISSENSE	p.R158L	NM_001126113	27	1620	1.67	9E-39	1812
P42.5	chr7	100417264	A	C	exonic	EPHB4	MISSENSE	p.F404L	NM_004444	71	2665	2.66	5E-119	1010
P42.5	chr11	108201015	G	A	exonic	ATM	MISSENSE	p.R2461H	NM_000051	20	1504	1.33	5E-26	1082
P42.5	chr18	48591918	C	T	exonic	SMAD4	MISSENSE	p.R361C	NM_005359	145	1413	10.26	0E+00	1160
P42.5	chr20	57415495	G	A	exonic	GNAS	MISSENSE	p.E112K	NM_016592	17	1433	1.19	6E-21	2324
P43.4	chr2	29416520	A	G	exonic	ALK	MISSENSE	p.M1478T	NM_004304	42	3499	1.20	3E-55	1126
P43.5	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	37	2279	1.62	3E-53	1263

GENERAL DISCUSSION

Various types of optimized targeted deep sequencing had been reported recently. All of the methods aim for the early cancer detection. TRACERx (69) performed a large number of patients to collect the information of the lung cancer for early cancer detection. Despite their study discovered the enormous amount of information about intra-tumor heterogeneity, the limitations of depth of sequencing, bioinformatics pipeline, and the cost of profiling had mentioned. Advanced versions of CAPP-seq (56) had clearly increased the specificity by replacing the barcoding adapter. However, it needs more stabilization by the depth of coverage. Finally, the recent study has increased the depth of coverage over 30,000x to find the early cancer detection (70). All those efforts of advancing technology turned to face another common challenging factor: the biological noises. The relationship between the tumor cells and ctDNA must be highlighted to achieve the ultimate goal of liquid biopsy with ctDNA analysis. Perhaps the investigation of the extravascular or intravascular mechanism of tumor cell may help to explain how the cells have escaped, accumulated, and released the cfDNA into the blood vessels. In summary, the characterization of the background noise of sequencing technology and biology had elucidated in this study and finalized to discriminate the ctDNA for clinical application.

REFERENCES

1. Park G, Park JK, Shin SH, Jeon HJ, Kim NKD, Kim YJ, et al. Characterization of background noise in capture-based targeted sequencing data. *Genome biology*. 2017;18(1):136.
2. Chung J, Son DS, Jeon HJ, Kim KM, Park G, Ryu GH, et al. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. *Scientific reports*. 2016;6:26732.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-74.
4. Gupta GP, Massague J. Cancer metastasis: building a framework. *Cell*. 2006;127(4):679-95.
5. Shen H, Laird PW. Interplay between the cancer genome and epigenome. *Cell*. 2013;153(1):38-55.
6. Neel DS, Bivona TG. Resistance is futile: overcoming resistance to targeted therapies in lung adenocarcinoma. *NPJ Precis Oncol*. 2017;1.
7. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;168(4):613-28.
8. Friedrich MJ. Going With the Flow: The Promise and Challenge of Liquid Biopsies. *Jama*. 2017;318(12):1095-7.
9. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*. 2017;355(6331):1330-4.
10. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature reviews Cancer*. 2017;17(4):223-38.
11. Takai E, Yachida S. Circulating tumor DNA as a liquid biopsy target for detection of pancreatic cancer. *World journal of gastroenterology*. 2016;22(38):8480-8.
12. Husain H, Velculescu VE. Cancer DNA in the Circulation: The Liquid Biopsy. *Jama*. 2017;318(13):1272-4.
13. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. *Nature reviews Clinical oncology*. 2017;14(9):531-48.
14. Mandel P, Metais P. [Not Available]. *Comptes rendus des seances de la Societe de biologie et de ses filiales*. 1948;142(3-4):241-3.
15. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*. 2016;164(1-2):57-68.
16. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer research*. 2001;61(4):1659-65.
17. De Vlaminck I, Valantine HA, Snyder TM, Strehl C, Cohen G, Luikart H, et al. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Science translational medicine*. 2014;6(241):241ra77.
18. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, et al. Presence of fetal DNA in maternal plasma and serum. *Lancet*. 1997;350(9076):485-7.
19. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer research*. 1977;37(3):646-50.
20. Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology*. 1989;46(5):318-22.
21. Sidransky D, Von Eschenbach A, Tsai YC, Jones P, Summerhayes I,

- Marshall F, et al. Identification of p53 gene mutations in bladder cancers and urine samples. *Science*. 1991;252(5006):706-9.
22. Sidransky D, Tokino T, Hamilton SR, Kinzler KW, Levin B, Frost P, et al. Identification of ras oncogene mutations in the stool of patients with curable colorectal tumors. *Science*. 1992;256(5053):102-5.
23. Li M, Diehl F, Dressman D, Vogelstein B, Kinzler KW. BEAMing up for detection and quantification of rare sequence variants. *Nature methods*. 2006;3(2):95-7.
24. Remon J, Caramella C, Jovelet C, Lacroix L, Lawson A, Smalley S, et al. Osimertinib benefit in EGFR-mutant NSCLC patients with T790M-mutation detected by circulating tumour DNA. *Annals of oncology : official journal of the European Society for Medical Oncology*. 2017;28(4):784-90.
25. Thress KS, Brant R, Carr TH, Dearden S, Jenkins S, Brown H, et al. EGFR mutation detection in ctDNA from NSCLC patient plasma: A cross-platform comparison of leading technologies to support the clinical development of AZD9291. *Lung cancer*. 2015;90(3):509-15.
26. Gundry M, Vijg J. Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. *Mutation research*. 2012;729(1-2):1-15.
27. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2012;109(36):14508-13.
28. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(23):9530-5.
29. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform*. 2016;17(1):154-79.
30. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome biology*. 2011;12(11):R112.
31. Cline J, Braman JC, Hogrefe HH. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res*. 1996;24(18):3546-51.
32. Kuchta RD, Benkovic P, Benkovic SJ. Kinetic mechanism whereby DNA polymerase I (Klenow) replicates DNA with high fidelity. *Biochemistry*. 1988;27(18):6716-25.
33. Chen G, Mosier S, Gocke CD, Lin MT, Eshleman JR. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol Diagn Ther*. 2014;18(5):587-93.
34. Wong SQ, Li J, Salemi R, Sheppard KE, Do H, Tothill RW, et al. Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours. *Scientific reports*. 2013;3:3494.
35. Do H, Wong SQ, Li J, Dobrovic A. Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. *Clin Chem*. 2013;59(9):1376-83.
36. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-21.
37. Norton SE, Lechner JM, Williams T, Fernando MR. A stabilizing reagent

- prevents cell-free DNA contamination by cellular DNA in plasma during blood sample storage and shipping as determined by digital PCR. *Clin Biochem*. 2013;46(15):1561-5.
38. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-95.
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-303.
41. Medina Diaz I, Nocon A, Mehnert DH, Fredebohm J, Diehl F, Holtrup F. Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing. *PLoS one*. 2016;11(11):e0166354.
42. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-76.
43. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013;31(3):213-9.
44. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine*. 2014;20(5):548-54.
45. Chabon JJ, Simmons AD, Lovejoy AF, Esfahani MS, Newman AM, Haringsma HJ, et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. *Nature communications*. 2016;7:11815.
46. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res*. 2015;43(6):e37.
47. Wang XV, Blades N, Ding J, Sultana R, Parmigiani G. Estimation of sequencing error rates in short reads. *BMC bioinformatics*. 2012;13:185.
48. Takai E, Totoki Y, Nakamura H, Morizane C, Nara S, Hama N, et al. Clinical utility of circulating tumor DNA for molecular assessment in pancreatic cancer. *Scientific reports*. 2015;5:18425.
49. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, 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*. 2013;41(6):e67.
50. Kino K, Sugiyama H. UVR-induced G-C to C-G transversions from oxidative DNA damage. *Mutation research*. 2005;571(1-2):33-42.
51. Kino K, Sugiyama H. Possible cause of G-C-->C-G transversion mutation by guanine oxidation product, imidazolone. *Chemistry & biology*. 2001;8(4):369-78.
52. Chen L, Liu P, Evans TC, Jr., Ettwiller LM. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*. 2017;355(6326):752-6.
53. Swenberg JA, Lu K, Moeller BC, Gao L, Upton PB, Nakamura J, et al. Endogenous versus exogenous DNA adducts: their role in carcinogenesis, epidemiology, and risk assessment. *Toxicological sciences : an official journal of the Society of Toxicology*. 2011;120 Suppl 1:S130-45.
54. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. *Nature*

biotechnology. 2011;29(10):908-14.

55. Butler TM, Johnson-Camacho K, Peto M, Wang NJ, Macey TA, Korkola JE, et al. Exome Sequencing of Cell-Free DNA from Metastatic Cancer Patients Identifies Clinically Actionable Mutations Distinct from Primary Disease. *PloS one*. 2015;10(8):e0136407.
56. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nature biotechnology*. 2016;34(5):547-55.
57. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*. 2010;2(1):a001008.
58. Lennon AM, Goggins M. Diagnostic and Therapeutic Response Markers. *Pancreatic Cancer*. New York, NY: Springer New York; 2010. p. 675-701.
59. Makohon-Moore A, Iacobuzio-Donahue CA. Pancreatic cancer biology and genetics from an evolutionary perspective. *Nature reviews Cancer*. 2016;16(9):553-65.
60. Dabritz J, Preston R, Hanfler J, Oettle H. Follow-up study of K-ras mutations in the plasma of patients with pancreatic cancer: correlation with clinical features and carbohydrate antigen 19-9. *Pancreas*. 2009;38(5):534-41.
61. Brychta N, Krahn T, von Ahsen O. Detection of KRAS Mutations in Circulating Tumor DNA by Digital PCR in Early Stages of Pancreatic Cancer. *Clin Chem*. 2016;62(11):1482-91.
62. Ako S, Nouse K, Kinugasa H, Dohi C, Matushita H, Mizukawa S, et al. Utility of serum DNA as a marker for KRAS mutations in pancreatic cancer tissue. *Pancreatology : official journal of the International Association of Pancreatology*. 2017;17(2):285-90.
63. Pietrasz D, Pecuchet N, Garlan F, Didelot A, Dubreuil O, Doat S, et al. Plasma Circulating Tumor DNA in Pancreatic Cancer Patients Is a Prognostic Marker. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2017;23(1):116-23.
64. Earl J, Garcia-Nieto S, Martinez-Avila JC, Montans J, Sanjuanbenito A, Rodriguez-Garrote M, et al. Circulating tumor cells (Ctc) and kras mutant circulating free Dna (cfdna) detection in peripheral blood as biomarkers in patients diagnosed with exocrine pancreatic cancer. *BMC cancer*. 2015;15:797.
65. van Ginkel JH, Huibers MMH, van Es RJJ, de Bree R, Willems SM. Droplet digital PCR for detection and quantification of circulating tumor DNA in plasma of head and neck cancer patients. *BMC cancer*. 2017;17(1):428.
66. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in medicine*. 2007;26(10):2170-83.
67. Yachida S, Iacobuzio-Donahue CA. Evolution and dynamics of pancreatic cancer progression. *Oncogene*. 2013;32(45):5253-60.
68. Cohen JD, Javed AA, Thoburn C, Wong F, Tie J, Gibbs P, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. *Proceedings of the National Academy of Sciences of the United States of America*. 2017;114(38):10202-7.
69. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *The New England journal of medicine*. 2017;376(22):2109-21.
70. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Science translational medicine*. 2017;9(403).

초 록

서론: 실시간으로 종양 계놈의 역학을 측정할 수 있는 방법으로 순환 종양 (ct)DNA 와 차세대 시퀀싱 (NGS) 기반 방법 구현이 제시 되었다. 그러나 혈액 속에 존재하는 정상 세포 유리 (cf)DNA의 빈도는 ctDNA의 비율보다 높아 낮은 종양 변이의 대립 유전자와 기술오차 비율 수준이 동의 선상으로 측정 될 수 있어 이의 걸 맞는 차별화 및 실용적 가이드라인과 분석 방법이 필요하다. 제 1장*에서는 고유 DNA 분자 회수율의 중요성을 강조하고 시퀀싱 과정에서 발생하는 오류들의 특성을 분석하였다. 제 2장에서는 앞선 방법 조합하여 암 환자 샘플에 적용하여 ctDNA시퀀싱의 유용성 평가 및 종양 계놈 모니터링을 실시 하였다.

방법: 적응 시료량에서 DNA분자 회수율을 극대화 하기 위해, 시퀀싱의 초기 단계인 ligation 구성 요소의 온도, 시간 및 어댑터 농도의 조정 및 최적화 하여 구현되었다. 오류의 규명은 cfDNA의 특징 중의 하나인 자연적으로 절단된 장점을 이용하여 acoustically 절단된 germline DNA와 비교 분석되었다. 암 환자 샘플들 에서 검출 된 ctDNA의 유용성은 치료 반응 및 영상 변화에 따라 평가되었다.

결과: 시퀀싱 초기 단계를 수정한 ligation 조건을 적은 시료에 적용 하였을 때 DNA 분자 회수율은 표준 조건보다 20% 높은 비율을 나타내었다. 수동으로 전단 된 gDNA와 자연적으로 단편화 된 cfDNA의 특성을 비교한 결과 gDNA에서 C : G>

A : T의 64 %와 C : G > G : C의 39 %의 substitution class 비율이 증가됨을 규명할 수 있었으며 이는 전단 과정에서 일어날 수 있는 oxo-guanine과 연관이 있다는 것을 규명할 수 있었다. 순화된 전단 조건을 통해 관련 오류률은 평균 40% 정도 제거해 낼 수 있었다. 또한, DNA 단편의 말단 부근을 분석한 결과 A > G 및 A > T 우선적으로 단편화 되는 것을 알 수 있었다. 향상된 NGS 방법은 암환자 cfDNA 샘플에 적용하여 평가하였을 때 100 % 민감도와 97.1 % 특이도를 갖춘 진단적 유용성을 확립할 수 있었다. CtDNA의 반응도는 치료 반응과 높은 상관 관계가 있었을 뿐 아니라, 표준 단백질 바이오 마커와 이미징 변화 보다 2 개월 앞선 평균 반응도를 나타내었다. 마지막으로, ctDNA 분석은 종양 생검에서 알 수 없었던 종양 내 이질성 또한 검출해 낼 수 있었다.

결론: 전반적으로 cfDNA의 독특한 특성분석을 통해 기술적인 오류의 근본 원인을 강조할 수 있었을 뿐만 아닌 NGS 기반 기술을 사용하여 암의 조기 발견 기회를 입증할 수 있는 연구였다. 궁극적으로, cfDNA와 NGS 분석의 조합 접근법은 암 연구에서 충족되지 않은 요구를 해결할 것이라 믿는다.

* 본 내용은 *Scientific Reports*와 *Genome Biology* 학술지 (참고문헌 포맷)에 출판 완료된 내용임

주요어: 암 유전체학, 액체 생검, 순환하는 종양 DNA, 무 세포 DNA, 차세대 시퀀싱, 백그라운드 오류

학 번: 2012-21792



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

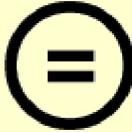
다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

향상된 NGS 분석 방법을 사용한
순환 종양 DNA 분석

**Analysis of circulating tumor DNA
by using NGS-based method with
enhanced analytical performance**

2018 년 02 월

서울대학교 대학원

의학과 의과학전공

박 가 희

A thesis of the Degree of Doctor of Philosophy

**Analysis of circulating tumor DNA
by using NGS-based method with
enhanced analytical performance**

**향상된 NGS 분석 방법을 사용한
순환 종양 DNA 분석**

February 2018

**The Department of Biomedical Sciences,
Seoul National University
College of Medicine
Gahee Park**

ABSTRACT

Introduction: Interrogation of circulating tumor (ct)DNA using next-generation sequencing (NGS)-based methods have been proposed as a way to track the dynamics of tumor in real time. However, there was no standard guideline for ctDNA sequencing that I have evaluated the procedure from end-to-end to propose the optimal analysis methods for ctDNA sequencing. Chapter 1* emphasizes the importance of the recovery of unique DNA molecule from the minimal amount of starting material. After that, the systematic evaluation of each step highlights the error-prone step in the sequencing process. In Chapter 2, the utility of ctDNA sequencing has evaluated through the monitoring of tumor genomic in multi-cancer samples.

Method: To maximize the recovery rate of unique DNA molecule, I approached the ligation step during the library preparation in sequencing protocol by optimizing the temperature, time and adapter concentration. Identification of technical errors was conducted with the comparison of background error distribution from the acoustically sheared germline DNA and naturally fragmented cell-free DNA. The utility of ctDNA sequencing analysis was assessed by comparing the standard protein biomarker and imaging changes during the patients' therapeutic intervention.

Results: The modified ligation conditions for the minimal amount of starting material able to increase the recovery rate of unique DNA molecule by 20% compared to the standard conditions. A comparison of the characteristic of

acoustically sheared gDNA and naturally fragmented cfDNA revealed that gDNA constituted with 64% of C: G> A: T and 39% of C: G> G: C substitution class changes. Through testing of the series of the mild sheared conditions, the reduction of error rate was observed with an average of 40%. Furthermore, the analysis of the vicinity at the ends of the DNA fragments revealed that A> G and A> T preferentially fragmented. The enhanced analytical performance in NGS method able to establish diagnostic utility with the detection sensitivity of 100% and specificity of 97.1% as applied to cancer plasma samples. The level of ctDNA was not only highly correlated with the therapeutic response but also showed an average of two months' earlier reaction than the standard protein biomarker and imaging changes. Finally, the determination of tumor heterogeneity was observed through ctDNA analysis, which was not discovered in the matched tumor biopsies.

Conclusions: Overall, the unique characterization of cfDNA could not only emphasize the underlying cause of technical errors but also demonstrate opportunities for early detection of cancer using NGS-based technology. Ultimately, the combined approach of ctDNA and NGS sequencing analysis is believed to address unmet needs in cancer research.

*The works published in Genome Biology(1) and Scientific Reports (2).

Keywords: Cancer genomics, liquid biopsy, circulating tumor DNA, cell-free DNA, next-generation sequencing, background error

Student number: 2012-21792

CONTENT

ABSTRACT	i
CONTENT	iii
LIST OF TABLES AND FIGURES	v
LIST OF ABBREVIATION.....	vii
GENERAL INTRODUCTION	1
Cell-free DNA	3
Circulating tumor DNA	4
Current detection methods for ctDNA	4
Digital PCR.....	4
Next generation sequencing.....	5
NGS-based ctDNA analysis	7
Potential misdiagnosis from background errors	8
CHAPTER1	1 4
Practical guidelines for cell-free DNA analysis using enhanced analytical performance of NGS-based method	
INTRODUCTION.....	1 5
MATERIALS AND METHODS.....	1 8
RESULTS.....	2 2
Comparison of blood collection tubes.....	2 2
Optimization of library preparation	2 2
Optimizing statistical modeling for cfDNA analysis	2 4
Performance of optimized TDS on cfDNA and PBL DNA..	2 4
Estimation of errors derived by TDS	2 5
From sequencing reaction	2 5
Distribution of background errors	2 5
Sample preparation caused background errors	2 6
Breakpoint preferences.....	2 7

Multi-statistical adjustment for removing the background errors	2 9
DISCUSSION	3 1
CHAPTER 2	5 1
Ultrasensitive interrogation of circulating tumor DNA from cancer patients using enhanced analytical performance of the NGS-based method	
INTRODUCTION.....	5 2
MATERIALS AND METHODS.....	5 4
RESULTS.....	6 1
Evaluation of LOD with single mutation	6 1
KRAS mutations.....	6 1
Evaluation of LOD with multi-mutations	6 1
“With primary” mutation	6 1
Biopsy-free manner	6 2
Monitoring tumor burden by measuring ctDNA	6 3
Diagnostic utility.....	6 5
DISCUSSION	6 7
GENERAL DISCUSSION.....	9 4
REFERENCES.....	9 5

LIST OF TABLES AND FIGURES

Introduction

Figure 1 Characteristic of cell-free DNA	9
Figure 2 General ctDNA analysis schematic flow.....	10
Figure 3 General process of capture-based targeted deep sequencing	11
Figure 4 Schematic flow of ctDNA analysis using NGS-based technology	12

CHAPTER 1

Figure 1-1. Performance of cfDNA sequencing	35
Figure 1-2. Quality score of read bases in targeted deep sequencing data	36
Figure 1-3. The distribution of background errors from PBL and plasma DNA	37
Figure 1-4. Alleviation of background error by various condition of fragmentation.....	38
Figure 1-5. The fragment size distribution from PBL and plasma DNA.....	39
Figure 1-6. Evaluation of read bases from the start position.....	40
Figure 1-7. Nucleotides around the DNA breakpoint.....	41
Figure 1-8. Nucleotides around the DNA breakpoint.....	42
Figure 1-9. Frequencies of dinucleotide	43
Figure 1-10. Combination of 16 dinucleotide frequencies	44
Figure 1-11. Allele frequency of background errors from hotspot mutations	45
Table 1-1A Total amount of plasma DNA collected from Streck BCT and EDTA tube.....	46
Table 1-1B The number of genomic variants detected from Streck BCT and EDTA tube.....	47
Table 1-2 The total amount of DNA yield was compared under different ligation condition.....	48
Table 1-3 Evaluation of open-source tools and statistical analysis using	

spike-in controls.	49
Table 1-4 Performance of multi-statistical analysis for ctDNA sequencing	50

CHAPTER 2

Figure 2-1. The correlation of harbored KRAS mutations using digital PCR and enhance NGS-method from pancreatic cancer patients	70
Figure 2-2. Tumor mutations in pre-treatment cfDNA samples from 17 PDAC patients.....	71
Figure 2-3. Monitoring of ctDNA PDAC patients under therapeutic intervention.....	72
Figure 2-4. Summary of plasma mutations determined by “biopsy-free manner”	73
Figure 2-5. Allelic fraction of ctDNA and CA19-9 level depending on therapy responses	74
Figure 2-6. The number of mutations in plasma DNA.....	75
Figure 2-7. Distribution of detected genes from pleural effusion fluid and plasma DNA.....	76
Figure 2-8. The differences of allele frequencies from pleural effusion fluid and plasma DNA	77
Figure 2-9. The size distribution of pleural effusion and plasma DNA	78
Table 2-1. The limits of detection sensitivity evaluated by KRAS mutations in 14 PDAC patients.....	79
Table 2-2. Determined mutations from 17 FNA samples.....	80
Table 2-3. Evaluation of FNA mutations in baseline plasma DNA samples	81
Table 2-4. The performance of droplet digital PCR in plasma samples	82
Table 2-5. The list of somatic mutations detected in plasma samples by biopsy-free manner.....	83

LIST OF ABBREVIATION

8-oxo-G: 8-oxo-7,8-dihydroguanine

ANOVA: Analysis of variance

AP site: Apurinic-apyrimidinic site

BEAM: Beads, emulsions, amplification, and magnetics

bp: Base pair

CA: cancer antigen

CR: Complete Response

CT: Computed Tomography

CTC: Circulating tumor cell

cfDNA: Cell-free DNA

ctDNA: Circulating tumor DNA

DNA: Deoxyribonucleic acid

Dx: Diagnosis

ddPCR: droplet digital PCR

EDTA: Ethylenediaminetetraacetic acid

EGFR: Epidermal growth factor receptor

ELISA: Enzyme-linked immunosorbent assay

EUS: endoscopic ultrasound

FDA: Food and drug administration

FNA: Fine needle aspiration

gDNA: germline DNA

KRAS: KRAS proto-oncogene, GTPase

LSD: Least significance difference

NSG: Next-generation sequencing

MAF: mutant allele frequency

miRNA: micro RNA

QC: Quality control

Q score: Phred quality scores

RNA: Ribonucleic acid

SD: Stable diseases (Medical terminology)

SD: Standard deviation (Statistical terminology)

SNP: Single nucleotide polymorphisms

SNV: Single nucleotide variants

TCGA: The Cancer Genome Atlas

TP53: Tumor protein p53

PBL: Peripheral blood leukocyte

PCR: Polymerase chain reaction

PD: Progression of disease

PDAC: Pancreatic ductal adenocarcinoma

PE: Pleural effusion

PR: Partial Response

WES: Whole-exome sequencing

GENERAL INTRODUCTION

Cancer is a disease which contents uncontrollable manner of cells division and ultimately influences to nearby normal cells (3). It conquers the particular tissue and often takes a route of the blood vessel or lymph node to travel other parts of the tissue to expand the colony (4). An understanding of such a behavior revealed by comparative analysis of genomic differences in normal cells (5). The main point was cancer cells contains a fatal mistake in the DNA sequences also known as a mutation. Researchers started to target the protein which arises from the specific driver mutation to cure cancer. However, the targeted inhibitors turn out to reduce in a certain amount of time yet often the rise of the novel clones which contains the different types of mutation to evolve throughout the therapeutic intervention (6). Moreover, the characteristic of localized tumor tends to acquire similar driver mutations, but it often varies by the unique feature of individuals that the intra-heterogeneity causes the resistance of the drugs (7). To observe the unique intra-heterogeneity, the serial biopsy must be obtained to estimate the tumor growth throughout the treatment. This is near impossible due to an ethic problem and painful to patients (8).

One of the strategies to prevent the expansion of cancer cell is to detect cancer early as possible (9). The chance of success of

treatment and prevention of clonal expansion is much higher than cancer has already been metastasized and/or discovered in the late stage. Nonetheless, the procedure of tissue biopsy is done to late stage of patients, and it is often too late to eradicate the tumor mass. Therefore, there must be a start-up package with a benefit to detect fast and accurate tumor signal in the non-invasive method (10).

The computational tomography is one alternative method to detect the tumor in non-invasive manner. We now have a high resolution of computational tomography (CT) images to identify the smallest tumor. But, the cost is incredibly high, and the effect of radiation to the patient would be another side effect of increasing the chance of getting cancer. Another is a collection of blood sample from the patient and quantifies the level of according protein biomarker. Cancer antigen (CA) is a protein biomarker that related to specific types of cancer. If the level of cancer antigen is higher than the standard threshold, the assumption can be made. However, the level of protein biomarker also often miscorrelates due to the possibility of halt of molecular mechanism, some individuals have not express the certain types of protein biomarker, or the level varies on the individual's health condition (11). Recent studies suggest the alternatives of tumor biopsy or protein biomarker with the other types of resources (cancer-related exosomes, microRNA (miRNA), circulating tumor cells (CTCs), cell-

free DNA (cfDNA), and etc) can be not only collected from the plasma of blood but also from the body fluid. (10). The reason of using non-invasive biopsy collected from the blood or body fluid (hereafter, liquid biopsy) is, it allows to track the progression of the disease and figuring out the therapy response in the regular bases (10, 12). The candidates from the liquid biopsy have been evaluated with multiple types of the approach. Each of molecular biomarker candidates eliminated as their limited resources and due to the lack of the knowledge underlies the mechanisms. CTC was one of the revolutionary discovery in the cancer research. CTC claims to escape from the tumor mass but it is barely detectable in a resolution of analysis (13). Additionally, it is impossible to track in real-time. With all the dark histories, the liquid biopsy is back in business by cell-free DNA.

Cell-free DNA

The history of cfDNA began in 1948 discovered by Mendel and Metis (14). CfDNA is the naked DNA that floats in the body fluid or plasma of the blood with an average peak size of 166 base pair (bp) (Figure 1) (15). The origins of cfDNA hypothesized to be corresponded by the cell's apoptosis, necrosis, secretion, or combination of all due to its genome-wide size distribution (16). Moreover, the observation correlates to the nucleosome positioning space that it estimates to be the one wrap of chromosome (Figure 1).

As the cfDNA releases the genetic factors from the individual of cells, the analysis of cfDNA provides the vast of information not be limited to

detecting neoplastic diseases but also applicable to trauma, stroke, organ transplantation, prenatal screening for fetal aneuploidy, and etc (17, 18).

Circulating tumor DNA

Of course, cancer cells also leave out the trace to the blood stream. The cell-free DNA contains the genetic alteration is called circulating tumor DNA. In 1977, the evidence of the level of cfDNA in cancer patients represent higher than the healthy volunteers (19). Consequent results highlight the correlation of the amount of cfDNA with the existence of tumor mass(20). Hereafter, the approval of ctDNA analysis was done by applying the detection of TP53 (21) and KRAS mutations (22) in cancer patients.

However, the study of ctDNA revolutionized recently because of the lack of detection sensitivity with existed techniques. A critical fact of ctDNA is, it embedded by the massive amount of normal cfDNA. An ultrasensitive detection method is needed to detect ctDNA. Therefore, it was impossible to carry out further and walked on the spot decade ago.

Current detection methods for ctDNA

Digital PCR

Luckily, researchers realized the importance of improvement of analytical performance to implement the ultrasensitive detection methods. With a born of BEAMing (beads, emulsions, amplification, and magnetics) PCR technology (23), the capture of single DNA molecules has started the engine

about the discovering the variants with the lowest allele frequency (23). ctDNA analysis was proven as a cancer screening tool using the application of digital PCR. For instance, the food and drug administration (FDA) has tested two types of mutations (exon 19 deletion and/or L858R) that can be compensate by the tumor biopsy. The epidermal growth factor receptor (EGFR) mutations were evaluated in the non-small cell lung cancer patients. In brief, the patients who underwent the EGFR tyrosine kinase inhibitor (TKIs) tested to screen the rise of resistance mutations (24, 25) (Figure 2). Following that, the decision was made to treat with the inhibitor or otherwise.

The digital PCR application is limited to only those the patients who contain the known mutations. The chance of losing the novel signal comprises by the lack of understanding of underlying resistance mechanism and the limit of rest of patients.

Next generation sequencing

To get more information of resistance signals or tumor heterogeneity simultaneously, the implementation of genome-wide study is needed. Next-generation sequencing (NGS) technology allows analyzing the collection of genomic alterations in once that can be selected in interest target in a genome-wide region. There are two types of sequencing method: amplicon-based and capture-based sequencing. Amplicon-based sequencing amplifies the region of the target at the beginning of the sequencing and hybridizes the molecular barcode at the end. Simply, the broad range of PCR with the shorter size of

DNA fragments. On the other hand, capture-based sequencing shear the DNA at the beginning of the experiment and ligase the adapter sequences. After that, the customized RNA baits hybridize to the DNA. Therefore, it has broader and wider DNA sequences compare to amplicon-based sequences. Capture-based sequencing has known to have lower error rates than the amplicon-based sequencing. However, the methods can be exchangeable depends on the interest for cancer target.

In general, the process of targeted deep sequencing categorized in three steps (Figure 3): library preparation, target enrichment, and sequencing. The optimal experimental procedure may increase by the efficacy of the library preparation step. The step of library construction is important because the unique DNA molecule can be maximized or minimized by the optimized condition. There are three categories that impact on. Ligation, purification, and PCR amplification. In order to sustain the input DNA, the optimizing the library steps such as ligation step or adjusting the PCR cycles may help the recovery rate of DNA. Also, the higher amount of DNA has higher chance to bind the adapter that increases the recovery rate of initial input DNA molecules (2). However, there are several purification steps that the chance of losing the initial input of DNA molecules.

After generating the sequencing data, the large amount of raw DNA sequence data comes out to the world. The raw data scrutinize under the Phred quality score (Q score) by calculating the probabilities of any kinds of technical errors introduced to the base/read. Each of the base scores is then re-sorted and aligned with human genome sequences (or any interested genome

sequences, but deals with human genome in this thesis). Continuously, the aligned sequences organize with the counts of reference read counts and any alternative read counts to evaluate under the multi-statistical analysis for structural variant analysis (Figure 4).

NGS-based ctDNA analysis

Currently, the utility of NGS-based technology for ctDNA analysis has proved in numerous amounts of studies. As mentioned earlier, there must be optimized step for library preparation to have proper analysis of ctDNA. As the amount of ctDNA is limited, the recovery rates of DNA molecules are the critical points. Shortlisted to the targeted sequencing approaches, there are Tam-Seq (21), Safe-SeqS (22) for amplicon-based sequencing and CAPP-seq (23) and TEC-Seq (24) for hybrid capture sequencing. These ultrasensitive methods succeed to detect as low as 0.002% of the allelic fraction that even minimal residual diseases can be detected faster than the computed tomography (CT) images or any other cancer biomarkers by detecting ctDNA footprint.

Taking advantage of collecting the genomic alterations at once, the understanding of dynamics of tumor genomics were recognized in real-time. Roughly speaking, the main point of implementation of NGS to ctDNA analysis is, there is no need to know the prior information of tumor mutation. Moreover, the cost of NGS-based technology continuously reduces that can compete with digital PCR in a near future.

Potential misdiagnosis from background errors

Nevertheless, the inevitable problem is the background errors can be incurred either from the technology or biology. The measurement of sequencing errors was well-documented since the NGS-based technology has invented. The technical errors can be introduced by each steps of procedure. These errors are especially critical to ctDNA analysis because of the potential alleviation of false positives.

Another caveat is the biological errors. The little understanding of the biology of cfDNA makes vulnerable to apply for the screening tool. The most concern of biological errors is rise from hematopoietic cells (15). As the hematopoietic cells continuously circulate with the circulating tumor cells, the false positives can be involved. It also fluctuates the biological background that contributes the bias results. Therefore, elucidating the background noises are needed for a confident in variant calling.

In this thesis, chapter 1 focuses on how I have updated the standard operating procedure to optimize the next-generation sequencing based technology for the small amount of input DNA. Continuously, to discover the error-prone steps in NSG-based technology, the systematic evaluation was proposed by comparing the background distribution of acoustically sheared germline DNA and naturally fragmented cfDNA. In chapter 2, to prove the benefits of the ctDNA sequencing, I compared the matched tumor DNA and cfDNA and evaluated the detection sensitivity comparing to the digital PCR analysis. Lastly, I have discovered the different characteristics of cfDNA

compare by the different types of body fluid to discover the effect of cfDNA release mechanism.

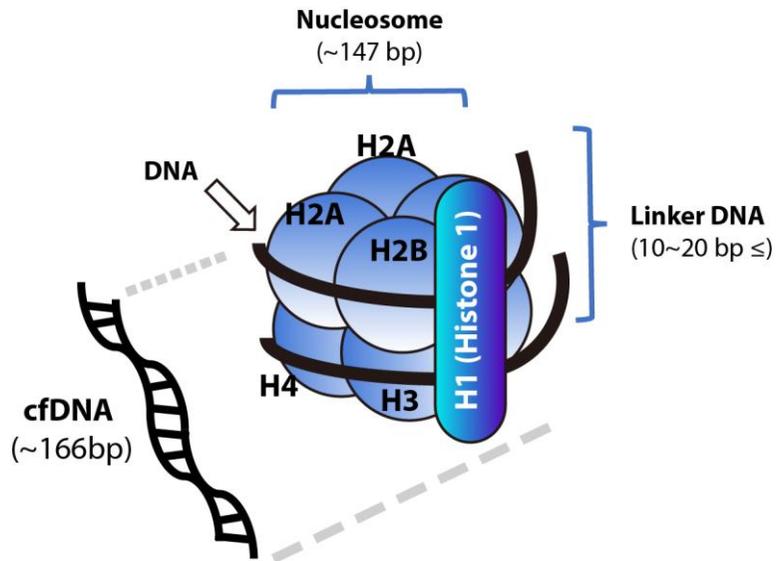


Figure 1Characteristic of cell-free DNA.

The average peak size of cfDNA is approximately 166 bp. The sequence length is similar mononucleotide that a wrap of histone core (~147 bp) and linker DNA (10-20bp).

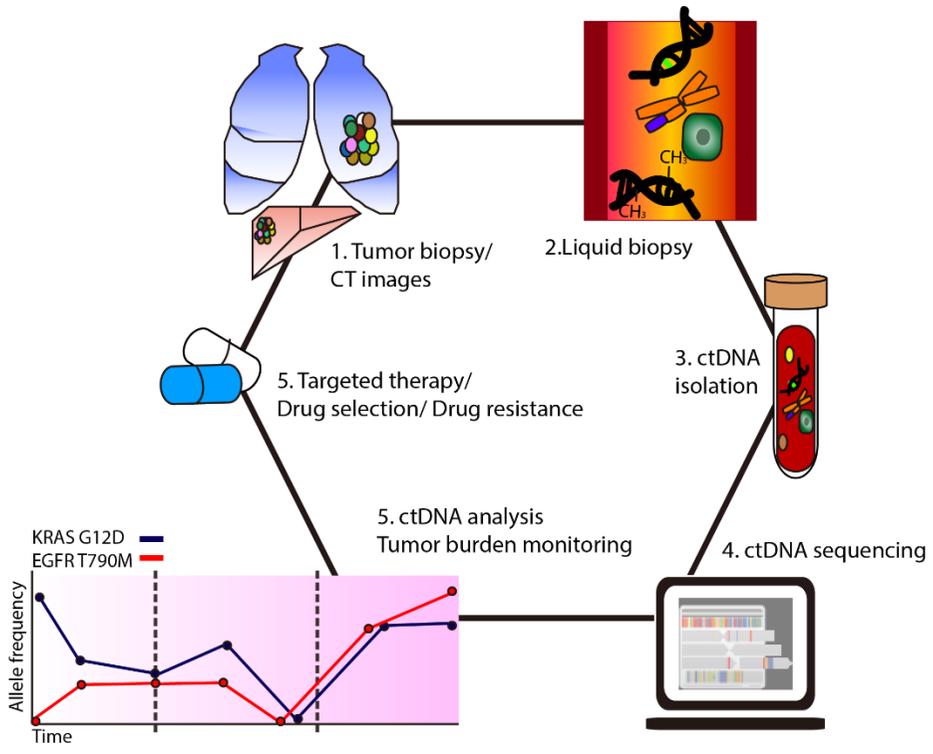


Figure 2 An overview of circulating tumor DNA analysis in clinical application

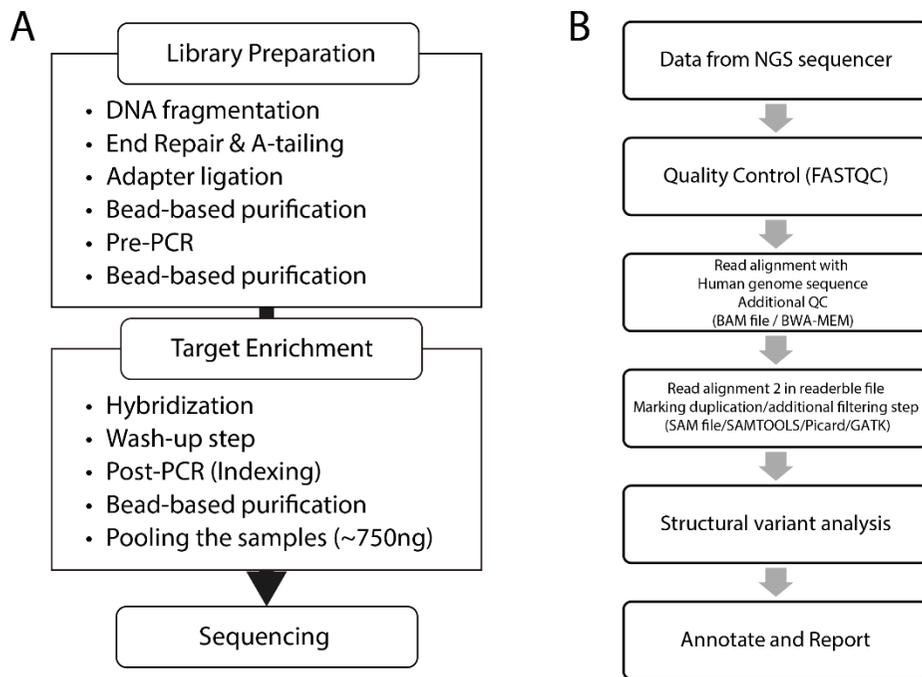


Figure 3 General process of capture-based targeted deep sequencing

(A) A general scheme of the library preparation and (B) the bioinformatics pipeline for calling genomic alterations.

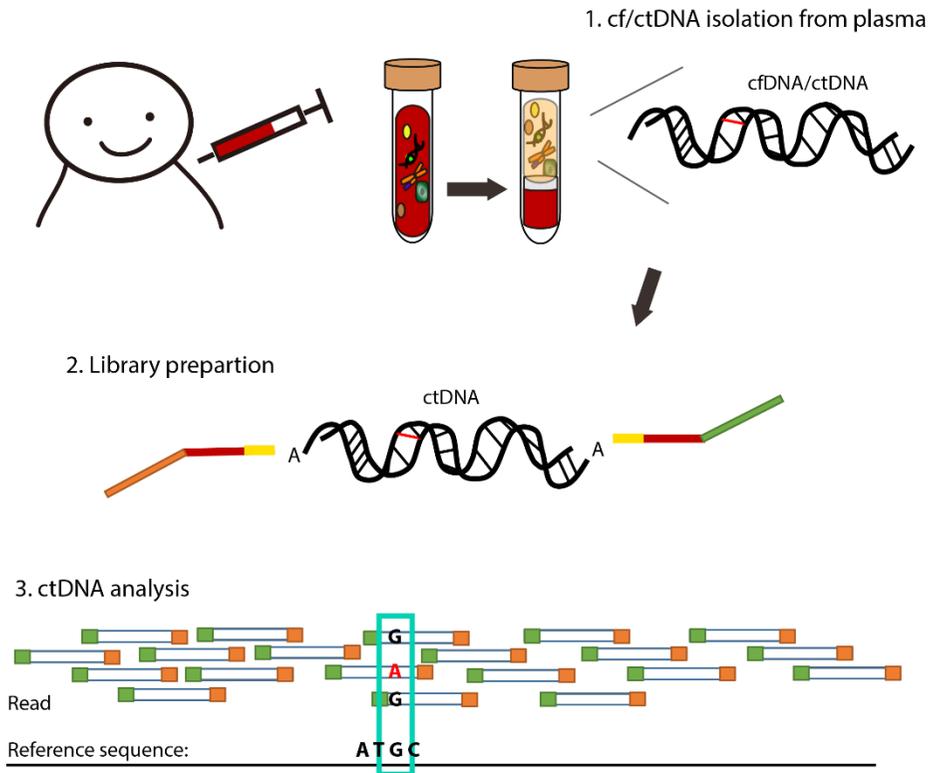


Figure 4 Schematic flow of ctDNA analysis using NGS-based technology

CHAPTER1

Practical guidelines for cell-free DNA analysis using enhanced analytical performance of NGS- based method

**This research was published in *Scientific Reports* on May 2016 and
Genome Biology on July 2017.**

INTRODUCTION

The range of the amount of cell-free DNA varies on individual samples. Theoretically, the cell-free DNA releases from every part of the somatic cells including releasing the circulating tumor DNA from cancer cells. In general, the cell-free DNA circulates in a minimal amount in the body fluid, but it relies on the condition of health that there is a possibility of having a large quantity of cell-free DNA. The concentration of cell-free DNA matters to the procedure of profiling the mutations from the cancer patients which directly associates with the detection of sensitivity and specificity. The potential of harboring the somatic mutation could be extremely low as 0.01% that the substantial recovery of unique DNA molecule is highly desirable for extensive analysis. Any loss of unique DNA molecule is critical for reducing the limit of detection, profiling the lowest cancer signal, and understanding of cancer progression. The innovative technology called digital PCR has now routinely aided to detect the lowest allelic fraction using the lowest amount of DNA. However, the argument of using digital PCR is, the prior information must be given for identifying the specific loci. It confronts the issue to many patients who do not have the particular types of known mutations as well as surveillance monitoring during the therapeutic intervention. Therefore, the implementation of genome-wide sequencing is suitable to understand the proper tumorigenesis of any kinds of cancer study.

The integration of next-generation sequencing with cfDNA has been developed recently that proved the possibility of profiling the tumor genomics

in a real-time. Nonetheless, the studies presented with the customized techniques that there is no standard guideline to implement the cell-free DNA sequencing properly. It is often hard to reproduce the experimental procedure or the bioinformatics workflow. Therefore, to set out the practical guideline for the minimal amount of starting material, I primarily reached to the ligation step in the NGS library preparation to maximize the recovery of the unique DNA molecule and the high confidence of throughput needs.

Next, the key element of high sensitivity and specificity is to discriminate the technical and biological errors from the limited amount of samples. It is well documented that the sequencing artifacts limit the analytical sensitivity (26-28). For example, errors caused by Illumina HiSeq sequencer chemistry are relatively well-understood, and therefore appropriate data filtering criteria based on this knowledge are routinely applied to generated data to remove them (29). The filtration of errors includes the removal of parts of, or entire reads containing numerous low-quality bases, to minimize downstream analysis artifacts (30). The fidelity of polymerases routinely used in the construction of sequencing libraries is well characterized (31, 32); however, it is difficult to quantify the error rate induced by DNA damage during library construction. For example, heat-induced cytosine deamination during PCR thermocycling has been suggested as a possible cause of baseline noise in Ion Torrent semiconductor sequencing data (33). Moreover, cytosine deamination occurs not only during experimental procedures such as PCR amplification (33) and formalin fixation (34, 35), but also prior to sample preparation (i.e. intrinsically or biologically) in the

original DNA templates (36). Nevertheless, it is not clear the error-prone step as well as the impact of how much of the errors relevant to cell-free DNA analysis which have been incurred during the sequencing run itself. Since technical errors are also likely to be introduced during sample preparation, library preparation, target enrichment, and/or amplification of DNA samples, a thorough characterization of such errors may facilitate the detection of method-dependent systematic errors and allow true variants to be distinguished from these errors. To determine during which step, and to what extent, a given type of error is introduced during sequencing, comparative experiments under different experimental conditions have been recommended, but are rarely performed due to practical reasons (29). Thus, no systematic analysis of the errors introduced during capture-based targeted deep sequencing has yet been conducted. To discover the systematic error-prone step in NGS-based technology, I attempted to analyze the non-reference alleles in ultra-deep coverage targeted capture sequencing data from both plasma and peripheral blood leukocyte (PBL) DNA samples. From this analysis, the rate of sequencing-artifact substitutions was estimated to be incurred during specific steps of the capture-based targeted sequencing process including DNA fragmentation, hybrid selection, and sequencing run. Based on the results, the use of mild acoustic shearing was recommended for genomic DNA (gDNA) fragmentation to minimize C:G>A:T and C:G>G:C transversion errors.

MATERIALS AND METHODS

1. Sample collection and DNA extraction

The corresponded blood samples were collected in Cell-Free DNA™ BCT tubes (Streck Inc., Omaha, NE, USA) (37) from 19 human subject. The samples were processed within 6 h of collection via three graded centrifugation steps (840 g for 10 min, 1040 g for 10 min, and 5000 g for 10 min, at 25 °C). The germline DNA were drawn from PBLs and collected from the initial centrifugation. The layer of plasma was transferred to new microcentrifuge tubes at each step. Plasma and PBL samples were stored at -80 °C until cfDNA extraction.

Germline DNAs from collected PBLs were isolated using a QIAamp DNA mini kit (Qiagen, Santa Clarita, CA, USA). Circulating DNAs were extracted from 1–5 mL of plasma using a QIAamp Circulating Nucleic Acid Kit (Qiagen). DNA concentration and purity was assessed by a PicoGreen fluorescence assay using a Qubit 2.0 Fluorometer (Life Technologies, Grand Island, NY, USA) with a Qubit dsDNA HS Assay Kit and a BR Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA). The concentration of DNA and the purity quantified by using a Nanodrop 8000 UV-Vis spectrometer (Thermo Fisher Scientific) and a Picogreen fluorescence assay using a Qubit 2.0 Fluorometer (Life Technologies). The size distribution of DNA fragmentation measured using a 2200 TapeStation Instrument (Agilent Technologies, Santa Clara, CA, USA) and real-time PCR Mx3005p (Agilent Technologies) according to the manufacturer's instructions.

3. Library preparation

Genomic DNAs samples fragmented by using a standard protocol of Covaris S220 (6 min, 10% duty factor, peak incident power = 175 W, 200 cycles/burst; Covaris Inc. Woburn, MA, USA) which the average size of 150–200 bp. On the other hand, the plasma DNA was prepared without fragmentation. The construction of sequencing libraries was achieved using 200 ng (for all samples) of PBL, and 37.3 ng (on average) of plasma DNA. To conduct the effect of DNA fragmentation step of background error rate measurement, the intensity and/or duration varied using 200 ng of initial genomic DNA from HapMap samples.

Next, the libraries for PBL and plasma DNAs were constructed using a KAPA Hyper Prep Kit (Kapa Biosystems, Woburn, MA, USA) as reported in Scientific Report (2). In brief, the adjustment of the end-repair, A-tailing, adapter ligation, and PCR reactions (nine amplification cycles) prior to target enrichment were performed. A purification step was carried out using AMPure beads (Beckman Coulter, Indiana, USA) after each step. Adaptor ligation was performed using a pre-indexed PentAdapter™ (PentaBase ApS, Denmark) at 4°C overnight.

4. Sequence data processing

After acquiring the raw FASTQ file from the sequencing procedure, the BWA-mem (v0.7.5) (38) aligned the hg19 human reference to create BAM files. SAMTOOLS (v0.1.18) (39), Picard (v1.93), and GATK (v3.1.1) (40) were used for sorting SAM/BAM files, local realignments, and duplicate

markings, respectively. The duplicates, discordant pairs, and off-target reads were filtered according the instruction.

5. Background distribution analysis

The paired set of PBL and plasma DNA samples were determined a base at a position across the entire target regions to be a background allele if the following conditions were met: (1) the base was a non-reference allele; (2) the position displayed sufficient depth of coverage (i.e. $>500\times$) in the paired PBL and plasma DNA samples; and (3) the frequencies of the base in both samples did not indicate a germline variant (i.e. $<5\%$). Since the samples were collected from cancer patients, the filtration of candidate of somatic cancer variants was conducted. The genomic alteration profiled from the matched fine-needle aspiration (FNA) biopsies had been implied for the filtration. For example, *KRAS* variants were removed from the analysis if detected in the matched FNA specimens. The sequencing libraries for the primary tumors were also generated using 200 ng of input DNA according to same instruction I have mentioned above. After the removal of duplication, the depth of coverage of FNA samples was on average $987.15\times$ ($790.32 - 1476.55\times$). The position that below $250\times$ in the matched FNA biopsy and an allele if it was present at a frequency greater than 2.5% in the FNA sample was discarded to further analysis in a pair set of PBL and plasma samples.

6. Analysis of nucleotide composition and substitution rate at the near DNA break point

To obtain the frequency of mono- and dinucleotide at positions around break

point, 5'-end position of each mapped read was determined on the human reference genome and the sequence for the region of 100 bp (± 50 bp) around break point was collected. For a consistency, the collected sequences were displayed in the direction of the positive strand of the reference genome. The frequencies of nucleotides were calculated as the number of occurrence of a given mono- and dinucleotide divided total base with a quality score ≥ 30 at relative positions to break point. The frequencies were obtained for each sample and then values from 19 samples were averaged. For estimation of the frequency of mononucleotides, we displayed the position as the number of nucleotides from the first 5'-end nucleotide of the read. For dinucleotides, the number of nucleotides between the phosphodiester bond in a given dinucleotide and the break point was shown to indicate the relative position to break point. For instance, distance zero indicated that the first position was taken right before the 5'- end of the read, and the second position coincided with the beginning of the read.

Background error rates across all substitution classes were also calculated at each position relative to break point. The background alleles for each sample defined as described in the previous section were used for the analysis. For a comparison between PBL and plasma sample, substitution rate was normalized by the average rate of 1-50 bp. To remove errors occurred in Illumina sequencing platform as much as possible, we used only R1 reads whose front parts showed relatively better quality scores than those of R2 reads.

RESULTS

Comparison of blood collection tubes

The stabilized blood collection tube must be selected to avoid a loss of unique DNA molecules from the minimal amount of plasma DNA. It is critical to minimize the chance of involving the false positive variants from the lysed peripheral blood leucocytes DNA to cfDNA. A chance of accumulation of biological background errors due to the hematopoietic cells is inevitable. To test the stabilization of total amount of plasma DNA, the series of different time and temperature were measured. Streck BCT cell-free DNA blood collection tubes maintained the minimum variation of total amount of plasma DNA (154.45 ± 21.05 to 139.3 ± 18.6 , mean \pm SEM) compared to EDTA tube (138.3 ± 29.25 to 177.0 ± 20.4 , mean \pm SEM) (41) (Table 1-1A). Streck BCT cell-free blood tube also correlated higher number of detectable variants (22 ± 2 , mean \pm SEM) than EDTA tube (18.5 ± 0.5 , mean \pm SEM) (Table 1-1B). Overall, Streck BCT cell-free blood tube was selected for further analysis.

Optimization of the library preparation

The part of result is published in *Scientific Report* (2). The series of spike-in DNA was used in this test and the evaluation of commercial kits was presented by previous reports from our laboratory (2). In brief, KAPA Biosystems' Hyper Prep kit was selected for amongst the commercially available kits. To maximize the efficiency of library construction for the minimal amount of input DNA, the various ligation conditions were evaluated.

The ligation condition of (i) temperature, (ii) duration of time, and (iii) the molar ratio of adapter were considered to assess the performance of the recovery rate for unique DNA molecules. The evaluation between 16 °C to 25 °C for 15 min or 60 min did not make any differences compared to standard conditions suggested by the manufacturer recommendation (i.e., 20 °C for 15 min, data not shown). On the other hand, the ligation of temperature lowered to 4 °C and extended to overnight increased the total amount of DNA after the pre-PCR compared to 20 °C with 15 minutes (Table 1-2). After the rate of the improper mapped reads, duplication, improper pairs, and off-target reads discarded, the rate of on-target was quantified. Figure 1A shows on-target rate increased from 40% to 55% or 18% to 28% using 50 ng or 10 ng of input DNA, respectively. Moreover, the duplication rate was lowered from 35% to 19% or 60% to 50% using 50ng or 10ng of input DNA (Figure 1-1A). Next, the range of adapter concentration was tested from 136nm to 1.36Mm (Figure 1-1B). In a molar ratio of the adapter: insert, 300:1 to 30000:1 was tested using 50ng of gDNA with ligation of 4 °C overnight. Figure 1C shows the rate of duplication increased by the extension of ligation time and higher molarity of adapter. However, the purification step cleared up the potential PCR dimers and allowed to bind more DNA molecules. Therefore, the library construction was optimized with temperature of 4 °C, extension time of overnight and higher amount of adapter ratio for targeted sequencing with the minimal amount of input DNA (Figure 1-1C).

Optimizing statistical modeling for cfDNA analysis

To assess the low allelic fraction from plasma DNA, current open-source tools (42, 43) were evaluated with statistical methods suggested from previous studies in a range of spike-in controls (44, 45). As expected, the number of variants were failed to detect in both of open-source tools (Table 1-3). Despite the Fisher's exact test had detected number of variants with higher positive predicted values, the Binomial tests had shown higher sensitivity on the variants under the 0.1% allele frequency. As the sensitivity of detecting low allelic fraction is much more critical for ctDNA analysis, the Binomial statistical analysis was chosen for further analysis.

Performance of optimized TDS on cfDNA and PBL DNA

The total of 19 human subjects, including 17 pancreatic cancer patients and 2 healthy volunteers, were profiled by using optimized method described previously. On an average of 200ng of PBL DNA and 37.3 ng of plasma DNA, the average of 56.4 and 20.0 million total reads were generated in plasma and PBL DNA, respectively. The alignment rates were on average of 87.3% and 93.7% for plasma and PBL DNA. The unique coverage were determined to be $1964\times$ (1210 – 3069 \times) and $1717\times$ (1042 – 2361 \times) on average, respectively, after excluding the PCR duplication. The potential systematic bias from library or sequencing data excluded by comparing the single nucleotide polymorphisms (SNPs) from matched plasma and PBL samples. By these, a strong correlation between plasma and PBL samples were observed

($R = 0.9913$, $p\text{-value} < 0.0001$). Conclusively, the optimized methods generated sufficient amount of reads for the further analysis.

Estimation of errors derived by TDS

From sequencing reaction

The critical factor for down-stream analyses is depended on the constructing the proper background distribution from the plasma or PBL samples. Dae-soon Son and Seung-ho Shin helped to generate the proper background distribution. As mentioned in Method, the tumor-derived single nucleotide variants (SNVs) and germline SNPs excluded to avoid the potential bias. After that, Phred base quality score of non-reference background alleles was observed to distinguish any involvement of systematic bias. Most of the background alleles depicted under 30 of the base quality score, but the small bump was discovered after the base quality of 30 (Figure 1-2A). It was indistinguishable from the reference alleles (Figure 1-2B). It is critical to note the lowest fraction of errors are involved above the qualified bases because the lowest allelic fraction from plasma DNA is indistinguishable. As the background distribution was constructed after filtering out with most of the sequencing errors, the presence of the highly qualified background alleles may indicate the errors are from the other sources.

Distribution of background errors

Although both plasma and PBL DNA samples have been used as a control group for the purpose, the similarities and dissimilarities of background errors between plasma and PBL DNAs have not been elucidated. Thus, I compared the background errors from the plasma and PBL DNA. After the base quality score filtration, overall mean background rates were estimated to be 0.007% and 0.008% in plasma and PBL DNAs, respectively (Figure 1-3A). Next, with Seung ho Shin's aid, entire 12 nucleotide substitution classes of errors were examined (Figure 1-3B). The context of dependencies was revealed by incorporated information on the bases immediately 5' and 3' to each mutated. While the background frequency of each substitution class varied depending on its context, the patterns of background frequency variation associated with specific sequence contexts were strikingly similar between plasma and PBL DNAs except C:G>A:T substitution (Figure 1-3C).

Sample preparation caused background errors

In order to generate PBL DNA and plasma DNA sequencing data under the exact same condition, the optimized experimental protocol had to apply excluding the fragmentation step. Hence, the elevation of C: G>A: T hypothesized as due to DNA damage of the fragmentation step. The condition of milder acoustic shearing was applied to test whether the levels of C: G>A: T transversion disappeared. The intensity and/or shortened duration of acoustic shearing decreased the rate of C: G>A: T transversion in PBL DNA dramatically (Figure 1-4). Thus, the standard DNA fragmentation protocol

elevated the rate of C: G>A: T substitution owing to DNA damages, which could be alleviated under an appropriate fragmentation condition. By typical oxidative base lesion causes the formation of 8-oxo-7, 8-dihydroguanine (8-oxo-G) under C:G > A:T errors, enzyme-linked immunosorbent assay (ELISA) was performed by Yun Jeong Kim. By providing the samples with the serial attenuation of acoustic energy level, the 8-oxo-G attenuated by the assay (ANOVA p value = 6.0×10^{-7}). Therefore, it was definite that the standard protocol of DNA fragmentation step caused the elevation of C:G > A:T and C:G > G:C transversions. Taken together, the background rates were very similar between plasma and PBL DNA samples. Although PBL DNA displayed significantly higher C: G>A: T transversion rate than plasma DNA, our data suggested that an appropriate fragmentation condition abolished the elevation in the substitution rate. The results also suggested that germline DNA fragmented under the proper condition could be an alternative resource to estimate site-specific error rate distributions of plasma DNA in normal controls.

Breakpoint preferences

By considering the fragmentation step introduced the background errors, the end of fragmentation had to be associated with the mechano-chemical breakage of DNA (Figure 1-5). To characterize the preferences of breakpoint, the end bases of DNA fragments were assessed. First up was taking the both of aligned read 1 and read 2 and count the each of the nucleotides according

to the start position to 20 bp of the sequences. Figure 1-6 displays roughly estimated read counts from the randomly selected genomic regions. At a glance, I noticed the different ratio was shown at the first two bases. As sequencing platform is renounced with bad quality scores at the first four consecutive read bases (46, 47), the high quality of base scores near the end of DNA fragments were examined. As depicted in Figure 1-6B, the bad quality score were cumulated at the first five bases. Interestingly, the quality score was increased from the second bases which correlated with rough data (Figure 1-6A). Another hypothesis from Figure 1-6A was the different ratio of nucleotide bases. Noticeable differences clarified under categorization of the substitution classes (Figure 1-7A) which supported the previous data that has the dependency of fragmentation. Plasma DNA had different preferences of nucleotide changes (G>A then G>T) than PBL DNA. By taking advantage of naturally fragmented plasma DNA, the substitution rate was compared from the start point of read bases. Noticeably, the substitution rate of A with either G or T (ex, A>K) was significantly elevated at the first base in PBL DNA compared to plasma DNA (Figure 1-7B). The result indicated that the DNA damage did not induce A>K substitution. On the other hand, the substitution of neither C:G > A:T nor C:G > G:C errors was observed which are the most commonly associated with acoustic shearing (Figure 1-7B).

As I noticed the substitution of residue A might be associated with mechano-chemical breakage of DNA. The frequency of mononucleotide around DNA breakpoint was analyzed. By observing the fluctuation of frequencies of mononucleotide, A residue was predominantly presented in PBL DNA

(Figure 1-8). To get proximal examination, the total of 16 dinucleotides of frequencies around DNA breakpoint was analyzed. As expected, CA, TA and GA were susceptible to cleavage (Figure 1-8 and 9). Additionally, in order of $CG > CA > TA \sim GA$, the cleavage rate of phosphodiester bonds were reduced (Figure 10). Taken together, the acoustic shearing has fragmented DNA at the 5' A nucleotide residue preferentially.

Multi-statistical adjustment for removing the background errors

By noticing the potential background errors could be involved from qualified bases, a series of bioinformatics pipeline was framed under the binomial tests. The basic statistical model was followed by cancer personalized profiling by deep sequencing (CAPP-Seq(44)). There are two types of statistical pipelines to evaluate the significance of SNVs: “with primary” and “without primary.” With-primary pipeline tests the known mutations to plasma DNA detected from the primary tumor or tumor biopsy that called from the open-source tools such as VarScan2 or MuTect. On the other hand, without primary pipeline test is also known as “biopsy-free manner.” The answers are unknown; it opens the possibility for detecting any mutations from plasma DNA except the germline mutations. Although the framework of CAPP-seq statistical pipeline has established strictly with high sensitivity and specificity, the pipeline must be tested in order to built-in to the in-house system. There is five stepwise flow to categorize the variants (Method). Briefly, beginning with strand bias test, the total of 500 read counts must be meet or higher counts to

meet the criteria. The allele frequency of plasma evaluates with matched paired PBL DNA allele frequency and decides the allele frequency by adjustment under binomial test with position by position from the background noise distribution. After that, the filtered allele frequencies tested to entire background noise distribution that adjusting by multiple tests by Bonferroni and FDR significance level of 0.05. The variant candidates now considered as the outlier format under the adjusted read counts and Bonferroni p -value. However, I noticed CAPP-Seq has missed the concept of the batch effect. As the step of pooling was included in prior to sequencing process, the batch effect must be considered. Proximal gap dealt with differently with own pipeline process and implemented from the section of Takai et al.(48) pipeline (Method). Table 1-5 shows the number of variants has reduced the number by each step. Overall, the improvement of multi-statistical analysis was optimized for detection of ctDNA.

DISCUSSION

The significant discovery in this study is elucidating the fraction of background errors caused by acoustic shearing. Although the fraction of errors were relatively lower than previous studies, comparison between PBL DNA and plasma DNA surely clarified the acoustic shearing caused the rate of C:G>A:T and C:G >G:C transversion error mainly. The errors were constituted in “guanine” nucleotide rather than other three types of nucleotides. It can be explainable by the characteristic of guanine that has more susceptible to oxidation lesions owing to its potential of oxidation. Mechanically speaking, 8-oxo-G, G to T transversion substitution via dA:8-oxo-G pair, rouse from the process of shearing reported by Costello et al. (49) and can be reduced by the antioxidants. The fraction of errors were >20% comparatively to the present data (>1%) that their errors perhaps have exported with the shearing and the contamination. Moreover, the previous study highlighted the sequences of errors were CCG: CGG >CAG: CTG specifically. On the other hand, present study shows NCG: CGN > NAG: CTN. By taking unique feature of plasma DNA, the errors were aroused by the acoustic shearing rather than typical oxidative lesion product of 8-oxo-G; the direct C:G>A:T transversions which are the products of secondary oxidative lesion of 8-oxo-G, including imidazolone, guanidinohydantoin, and spiroiminodihydantoin which are known for causing C:G>G:C transversions(50, 51). Overall, the oxidation of guanine residues may responsible to cause both of C:G>A:T and C:G>G:C errors by acoustic

shearing.

While the present study was under review, the Chen et al.(52) reported the majority of errors were posed in the 1000 Genome Project and The Cancer Genome Atlas (TCGA) data sets. They reported the errors were the false negative variants that contained the allele frequencies of 1-5%. By analyzing those variants, they found the most prevalent substitution was C:G > A>T followed by A:T > T:A. Moreover, they presented DNA damages are caused by the purification step and alternated by the range of EDTA from the TE buffer. Continuously, the study performed the 1× TE (comprising 10 mM Tris (pH 8) and 1 mM EDTA) reduced the C: G> A: T and A: T> T: A errors. Checking up the buffer concentration immediately after the reports were found to be using exact same 1× TE buffer for DNA shearing and the error rates from the present data also had lower than Chen et al. By these, the present data showed not only the origins of errors but also supported by comparing with plasma DNA that the discovery from the study contributed for the improvement of utilizing the method of capture-based deep sequencing.

Most of previous studies attenuated their errors by adding the DNA repair enzyme. In this study, the errors were attenuated by modifying the standard condition of acoustic shearing: lowering the power and allowing the longer DNA fragments. The direct comparison presented the errors were reducible according to the recommendation of manufacturers for the fragmentation of input DNA with a median size of 150- 200bp. The advantage of reducing the

errors by modifying condition is surely by increasing the quality of bases at the end of reads and efficiency of data output. However, disregarding the fact of the library recovery rate of input DNA must be considered. It was evident from the present data that the on-target rate was reduced by 15-25% compared to the standard condition of acoustic shearing.

Although the present study mainly focused on the DNA damage due to 8-oxo-G, another common mechanism of DNA damage, apurinic-apyrimidic (AP) site, was evaluated by demonstrating the ELISA. AP site damage is involved in the DNA base excision repair that repairing the damages of mismatched DNA sequences by creating a nick at the backbone of the phosphodiester of the AP site. The damage is commonly due to depurination and/or depyrimidation (53). Through demonstrating ELISA, the level of acoustic energy was correlated by the AP sites during the steps of fragmentation (ANOVA, p value = 4.7×10^{-7}), but did not fully provided the reasons of increasing the error ratio of the A>G and/or A>T at the end of the DNA fragments. Hence, the mechano-chemical breakage of DNA is the strong candidate by causing the A>K errors at the end of fragments of DNA which was proved by comparing the plasma DNA. To have stronger evidence, similar data experimental data sets were found by the public data. Two independent studies were evaluated (54, 55). The data were generated under the whole-exome sequencing (WES) and had used the same COVARIS machine under the standard manufacturer's recommendation. The parallel comparison found the A>K errors at the end of DNA, but the errors were

lower than the present data. One general potential source could be from the differences of ligating the enzymes. While present data were facilitated with the KAPA Hyper enzymes, the two public data were made by the SureSelect enzyme that contains the T4 DNA polymerase and Klenow fragment. Due to patent, KAPA Hyper enzymes were blinded. Moreover, the time of extension might have increased the dimers that the present data had not eliminated the errors completely. As the end repair enzymes and time modified, the fidelity of end repair perhaps have influenced the fraction of errors in this present data set.

In a new regular feature of comparing plasma DNA with PBL DNA, it was noteworthy that the cleavages were preferred to cumulate at the 5' phosphodiester bonds of A residue appeared in PBL DNA. It intrigues to investigate the biological background noises. As Newman et al (44, 56) mentioned the background rate was imposed at the hotspot variants, the recurrent hotspot mutations were examined. There were no distinctive differences between the plasma and PBL samples (Figure 1-11 A). As the recurrent hotspots appeals to have predominant across the targeted regions, the background rate of tumor protein p53 (TP53) was observed (Figure 1-11 B). TP53 is also the region that frequently mutated in most of tumors compared to hematological malignancies. This point out the fact that if the background errors contain higher rate of TP53 variants in plasma DNA then the contribution can be reliably to acclaim that the pre-neoplastic cells have derived rather than the hematopoietic lineage cells (57). Since there were no

distinctive differences in either of plasma and PBL DNA, it makes sense that there is minimal impact of biological background at cancer hotspots. Taken together, the data sums up about the technical noises contributed much higher than the biological noises.

It is critical that the origin of ctDNA is unclear up to now. In fact, the data shows the higher ratio of A:T>T:A and C:G>T:A transversion errors in plasma DNA that another hypothesis must be set out to solve how the cells contributed to release and turned out to become a cell-free DNA. Another suggestion would be the healthy volunteer samples may be the alternative source for the standardization of background metrics as the circumstances of background errors are not randomly distributed entirely. In summary, the systematic analyses of technical and biological background noises helped to establish the proper statistical testing to further downstream of ctDNA analysis.

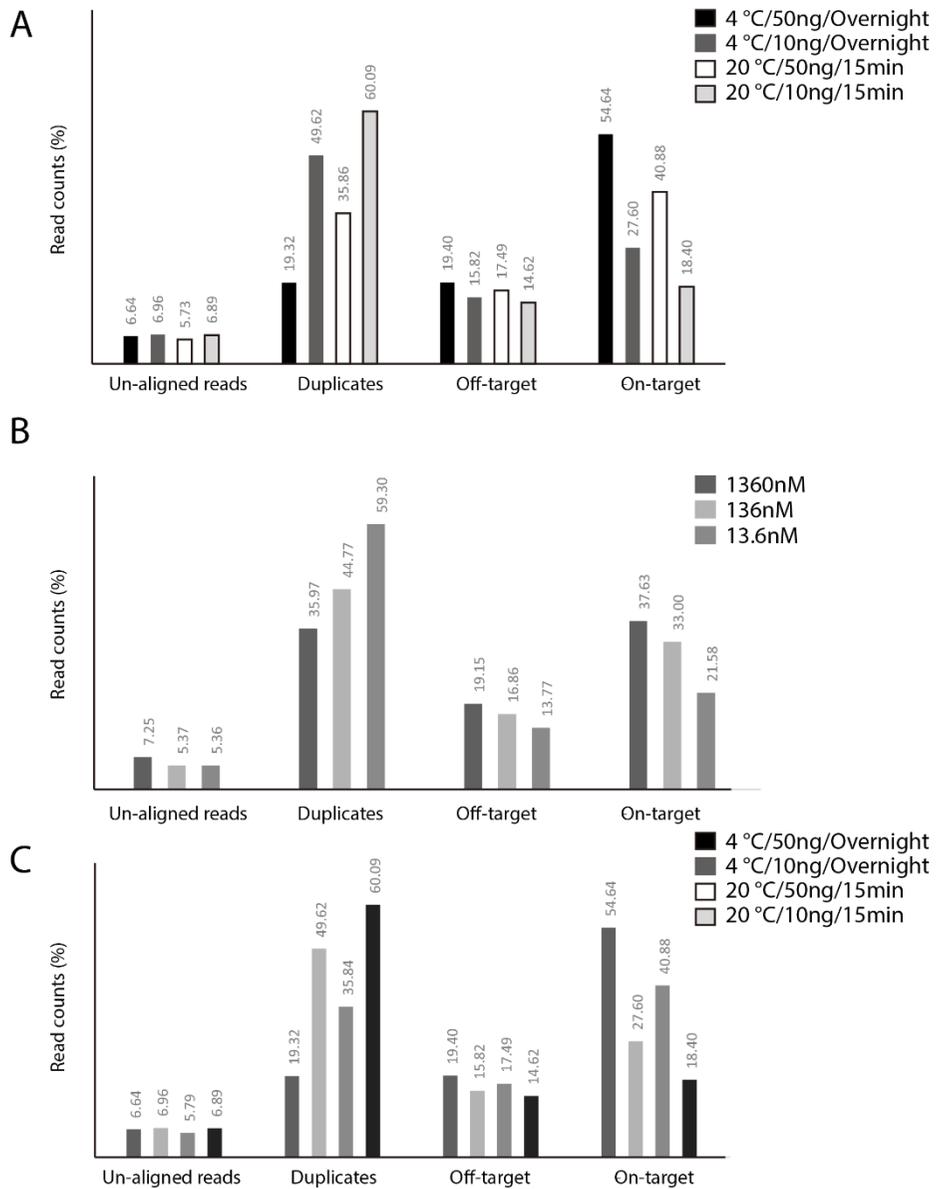


Figure 1-1. Performance of cfDNA sequencing

Performance of cfDNA sequencing by (A) adjustment of time, temperature and (B) molar ratio of adapters. (C) The comparison of optimized ligation step with standard condition.

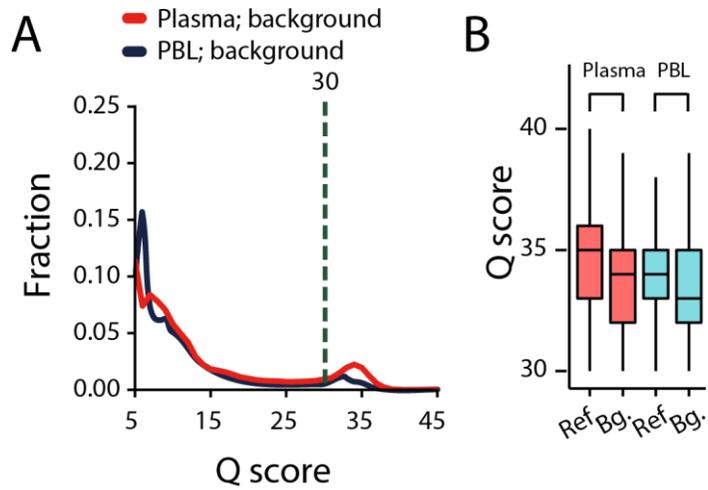


Figure 1-2. Quality score of read bases in targeted deep sequencing data

The distribution of background allele visualized with the density plot. The small fraction of background allele discovered above the quality score of 30. (B) The comparison of reference allele and background allele distribution gathered from both of plasma and PBL DNA samples.

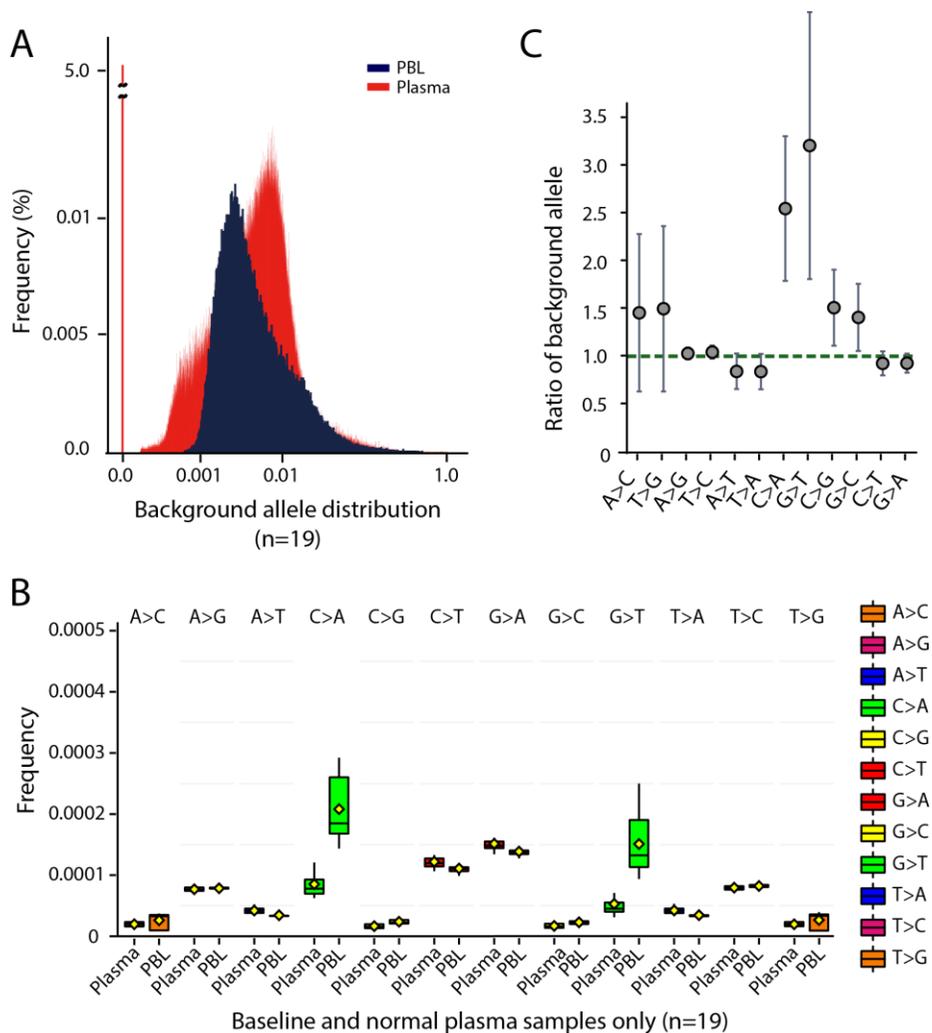


Figure 1-3. The distribution of background errors from PBL and plasma DNA

The distribution of background noise from PBL and plasma DNA were analyzed under substitution classes. (A) The distribution of background alleles from PBL and plasma DNA. (B) Substitution classes were compared between PBL and plasma DNA. (C) The ratio of substitution classes were determined by dividing plasma DNA by PBL DNA.

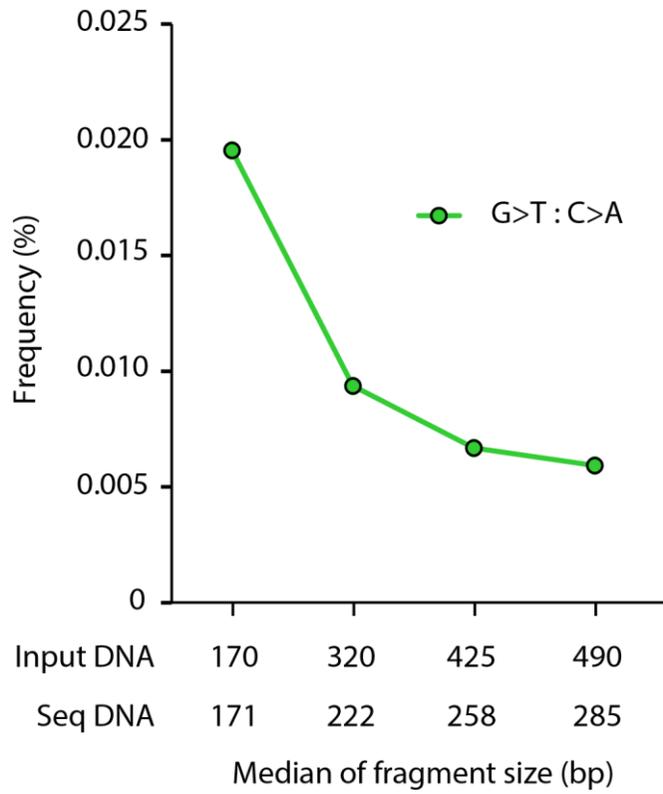


Figure 1-4. Alleviation of background error by various condition of fragmentation

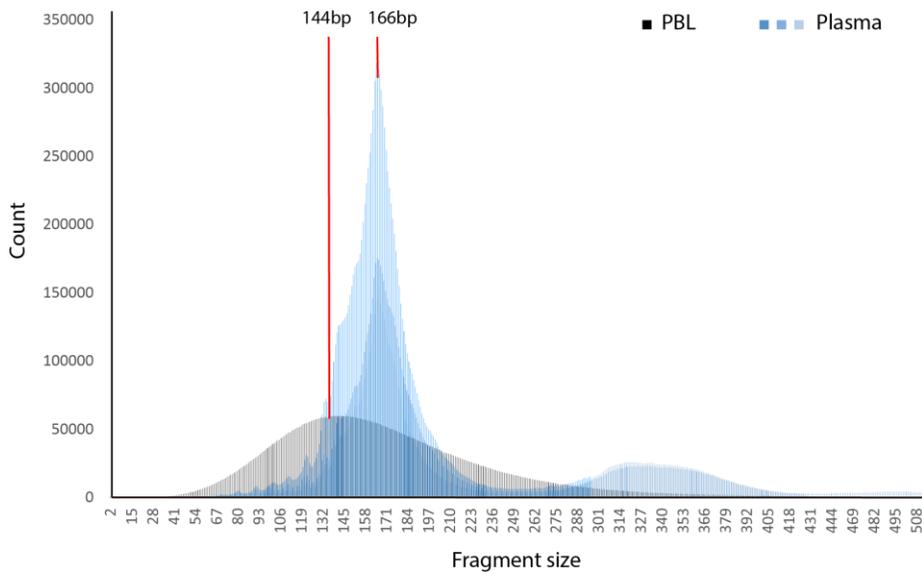


Figure 1-5. The distribution of fragment size from PBL and plasma DNA
 The distribution of fragment size in PBL and plasma DNA sample. The maximum peaks are 144 bp and 166bp for PBL and plasma DNA, respectively.

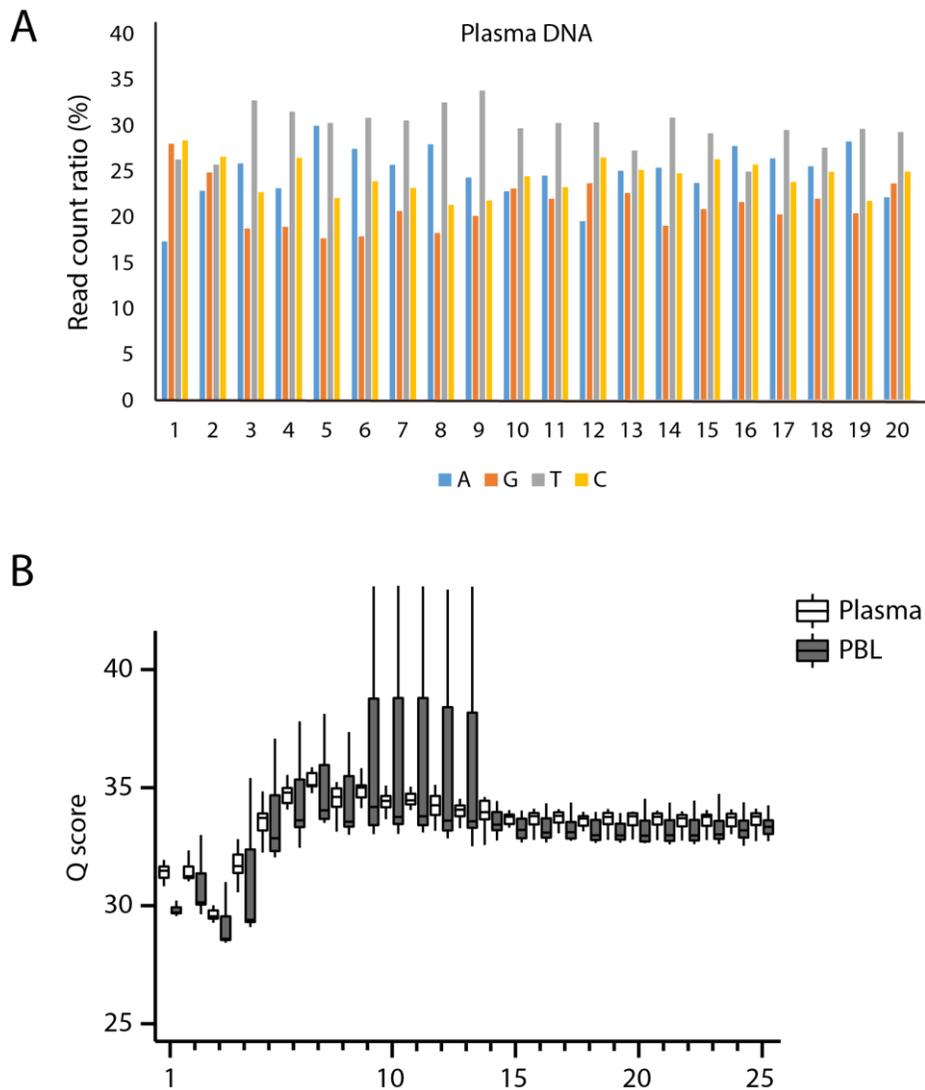


Figure 1-6. Evaluation of read bases from the start position

(A) The estimated average of read counts from the start position up to 20 bp of the sequencing reads. Random selected region, EGFR in this figure, showed for pilot test for the noticing the fluctuation of nucleotide sequences.

(B) The base quality score was examined from the starting point of read from plasma and PBL DNA.

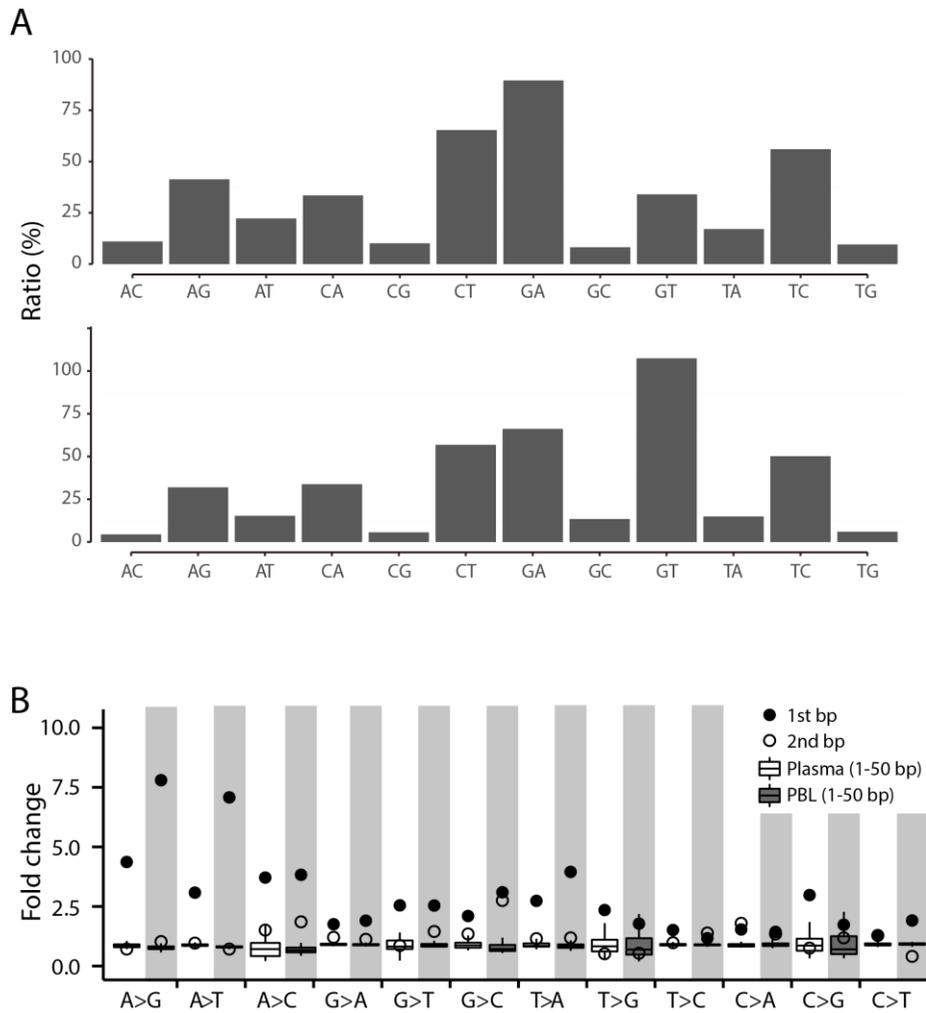


Figure 1-7. DNA breakage preference

The preferential DNA breakage was observed by substitution classes. (A) The ratio of read counts depended upon the 16 substitution classes. (B) The breakpoint of DNA across the substitution classes compared between plasma and PBL DNA.

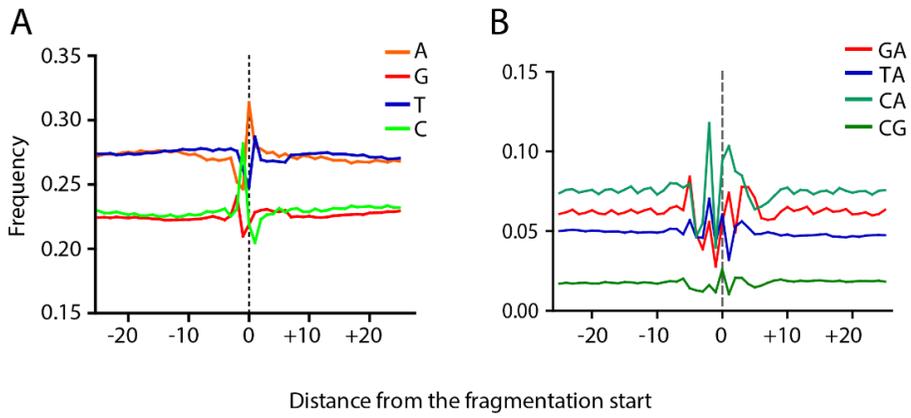


Figure 1-8. Nucleotides around the DNA breakpoint

Nucleotide around the DNA breakpoint was analyzed by (A) mononucleotide level and (B) dinucleotide level.

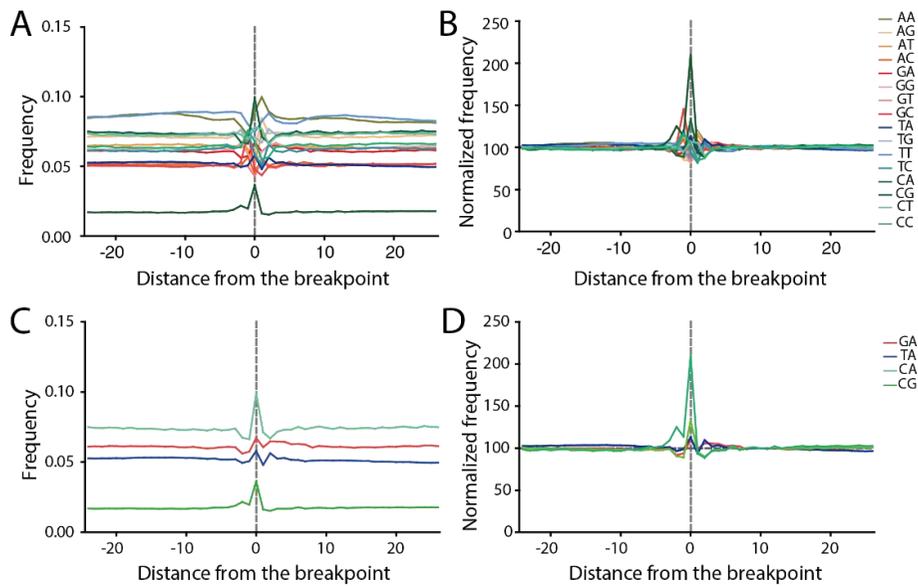


Figure 1-9. Frequencies of dinucleotide

Dinucleotide frequencies around the DNA breakpoint. (A) The frequencies and (B) normalized frequencies of dinucleotide across the 16 substitution classes were analyzed around the DNA breakpoints. The selected four substitution classes were depicted by the (C) frequency and (D) normalized frequencies.

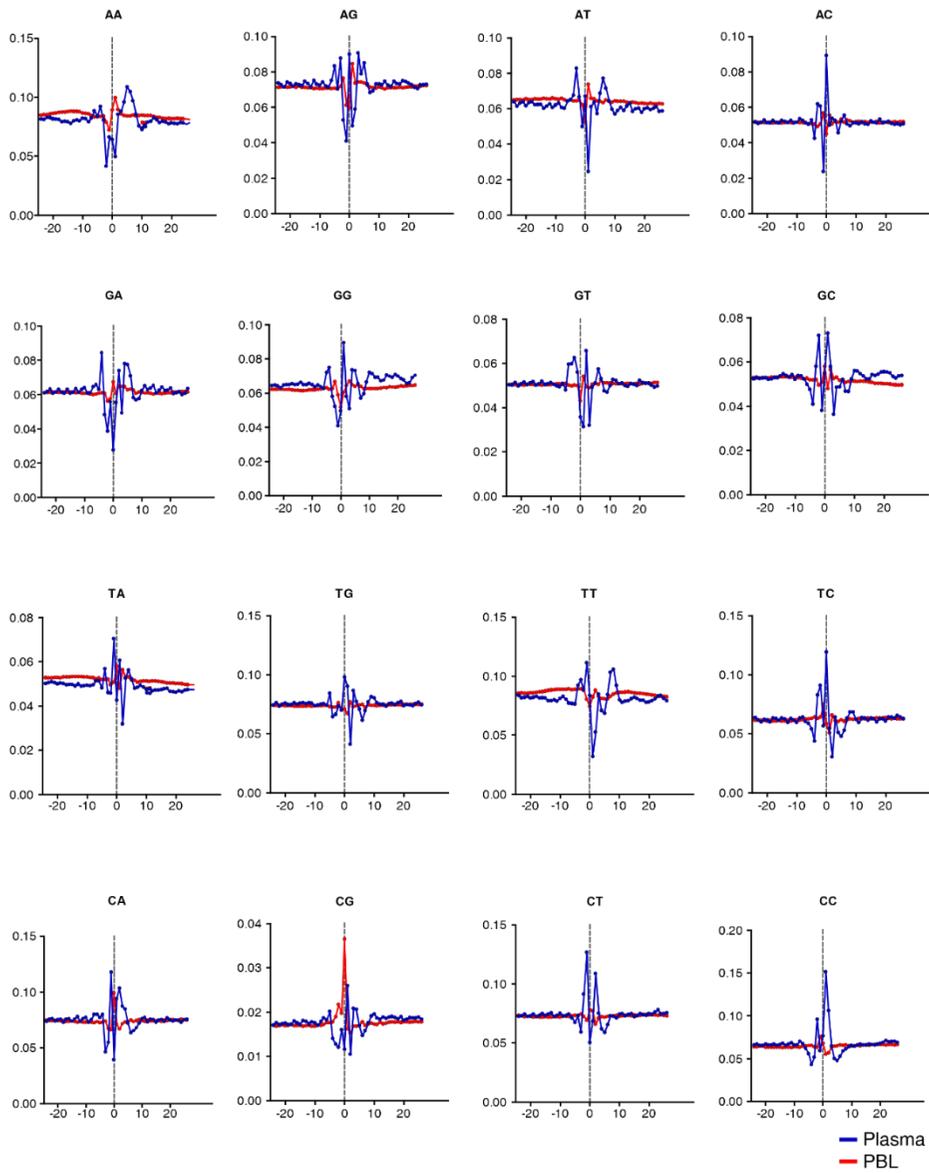


Figure 1-10. Combination of 16 dinucleotide frequencies

The combination of 16 dinucleotide frequencies were depicted around the DNA breakpoint from PBL and plasma DNA.

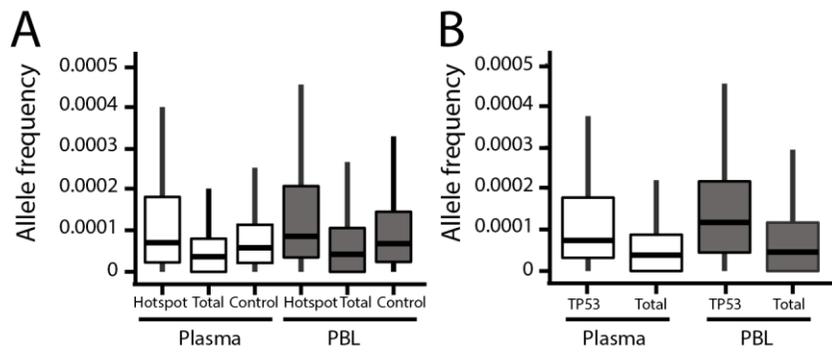


Figure 1-11. Allele frequency of background errors from hotspot mutations

(A) The average of background allele frequencies from recurrent hotspot mutations and (B) TP3 from PBL and plasma DNA.

Table 1-1A. Total amount of plasma DNA collected from Streck BCT and EDTA tube

Time	Streck BCT tube				EDTA tube			
	Day 1		Day 4		Day 1		Day 4	
Conc. (ng/ul)	133.4	175.5	138.3	140.2	109	167.5	208.3	145.7
SEM	9.66				20.76			

Table 1-1B. The number of genomic variants detected from Streck BCT and EDTA tube

Chr	Position	Ref.	GeneID	Variant	Streck Tube		EDTA tube	
					Day 1	Day 4	Day 1	Day 4
chr2	212812097	T	ERBB4	C	√	√	√	√
chr4	1807894	G	FGFR3	A	√	√	√	√
chr4	55141055	A	PDGFRA	G	√	√	√	√
chr4	55152040	C	PDGFRA	T	√	√	√	√
chr4	55972974	T	KDR	A	√	√	√	√
chr4	55962546	-	KDR	G	-	-	√	√
chr4	55980239	C	KDR	T	√	√	√	√
chr5	112175589	C	APC	T	√	√	-	-
chr5	112175770	G	APC	A	√	√	√	√
chr5	149433597	G	CSF1R	A	√	√	√	√
chr9	139399409	CAC	NOTCH1	-	-	-	-	√
chr10	43613843	G	RET	T	√	√	√	√
chr10	43615572	A	RET	T	√	-	-	-
chr11	534242	A	HRAS	G	√	√	√	√
chr13	48941623	T	RB1	C	√	-	-	-
chr13	49033902	T	RB1	C	√	-	-	-
chr17	7579471	G	TP53	-	-	-	√	√
chr17	7579472	G	TP53	C	√	√	√	√
chr19	17954157	C	JAK3	A	√	-	-	-

Table 1-2. The total amount of DNA yield was compared under different ligation condition

DNA (ng) /condition	20°C/ 15min	4°C/Overnight
50	18.6	41.6
50	19.9	35
10	3.68	9.88
10	4.74	9.1

Table 1-3. Evaluation of open-source tools and statistical analysis using spike-in controls

Expected AF (%)	Sensitivity (%)				Positive predicted value (%)			
	VarScan2	MuTect	Fisher's exact test	Binomial test	VarScan2	MuTect	Fisher's exact test	Binomial test
5	16	62.5	93.8	72.9	80	90	94.5	62.4
2.5	0	35.7	90.7	80.7				
1	0	6.25	27.1	43.8	0	25	55	43.6
0.5	0	7.86	14.3	28.6				
0.25	0	6.25	6.3	39.6	0	66.6	100	29.5
0.1	0	0.71	0	30				

Table 1-4. Performance of multi-statistical analysis for ctDNA sequencing

	Binomial	Z-test	Proximal gap	ANNOVAR filter	Final decision
Total variants	113358	1316	1278	30	27

CHAPTER 2

Ultrasensitive interrogation of circulating tumor DNA from cancer patients using enhanced analytical performance of the NGS-based method

INTRODUCTION

The countless of studies have been approached with the genome-widely to understand the molecular mechanism underlies the tumorigenesis. Nonetheless, the studies focused on the localized tumor biopsy that often contradict to the clinical outcome. It is relevantly due to the intra- and inter-tumor heterogeneity in cancer that evolves with the vast amount of mutations. To interrogate the status of tumor precisely, the real-time monitoring system must be conducted along the vast collection of genomic variants. Therefore, many studies have been analyzed the association of tumor biopsy and liquid biopsy such as protein biomarker to chase the change of tumor architecture. Nevertheless, the protein biomarkers stays as the standard biomarker (58) although lack the detection sensitivity and specificity and limiting its role complimentary on monitoring disease burden (11, 59). To compensate the protein biomarker, the cell-free DNA analysis has been evaluated in multiple cancer samples. The limits of detection varied among the location of cancer that deadly disease such as pancreatic cancer (PDAC) has not been highlighted on the benefits of ctDNA analysis.

In this chapter, I have evaluated the utility of ctDNA analysis using previously described method compare to digital PCR. As most of PDAC has constituted with the KRAS mutation over 90% (59), the detection sensitivity of KRAS mutations was benchmarked for the circulating tumor DNA analysis (48, 60-64). Next, I evaluated the benefits of interrogating in the genome-wide scale compared to selective point mutations collected from matched biopsy

samples.

As the release of circulating tumor DNA is not only obtained from the plasma DNA but also from the other types of body fluid, the characteristic of cfDNA was evaluated to analyze the confounding factor of biological interference from the release mechanism of cell-free DNA. To compare the characteristic, I approached with different types of cell-free DNA sample obtained from pleural effusion and plasma in matched lung cancer patients. The distribution of size fragmentation from collected cell-free DNA showed the significant differences. The results might support the subset of the biology of cell-free DNA as well as suppressing the biological noise while analyzing the circulating tumor DNA.

MATERIALS AND METHODS

1. Sample collection and DNA isolation

Pancreatic cancer sample

The institutional review board at Samsung Medical Center approved the study (IRB number 2014-04-048-009), and all the methods were carried out in accordance with the approved guidelines. Written informed consent was obtained from all subjects. Newly diagnosed pancreatic ductal adenocarcinoma (PDAC) patients who underwent the endoscopic ultrasound (EUS)-guided fine needle aspiration (FNA) procedure were enrolled and underwent blood draws for cell-free DNA (cfDNA) testing. A total of 120 samples, 17 FNA specimens, 34 peripheral blood leucocytes (PBLs) and 69 plasma samples was profiled from 17 patients. Among them, 17 pairs sequencing data from pretreatment plasma and PBL samples were reported in our recent study that analyzed technical sequencing errors (1).

Lung Cancer sample

The pleural effusion fluid and blood samples were collected from the 19 human subjects. The pleural effusion cell pellet and supernatants were collected separately. Other than that, the analysis and collection were followed with pancreatic cancer samples.

The pretreatment (i.e., before treatment) blood draw of the participants was collected at the time of diagnosis. Whole blood samples were collected in Cell-Free DNA™ BCT tubes (Streck Inc., Omaha, NE, USA). Plasma were prepared with three gradual steps of centrifugation (840g for 10min, 1040g for

10min, and 5000g for 10min at room temperature) while PBLs were drawn from the initial centrifugation. Collected plasma and PBL samples were stored at -80°C until cfDNA extraction. PBL germline DNA (gDNA) was isolated by QIAamp DNA mini kit (Qiagen, Santa Clarita, CA, USA). Plasma DNA was obtained from 2 to 5 mL of plasma via QIAamp Circulating Nucleic Acid Kit (Qiagen). AllPrep DNA/RNA Mini Kit (Qiagen) utilized to purify genomic DNAs from FNA tissues. The concentration and purity of DNA were examined by a Nanodrop 8000 UV-Vis spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) and a Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) with Qubit dsDNA HS Assay Kit and BR Assay Kit (Thermo Fisher Scientific). The distribution of fragment size as an indicator for DNA degradation were measured using a 2200 TapeStation Instrument (Agilent Technologies, Santa Clara, CA, USA) and real-time PCR Mx3005p (Agilent Technologies) according to the manufacturer's manual.

2. Library preparation and target enrichment

Purified gDNAs were fragmented in a range of 150-200 bp using Covaris S2 (7 min, 0.5% duty, intensity = 0.1, 50 cycles/burst; Covaris Inc. Woburn, MA, USA). The libraries for FNA samples were constructed by following the manufacture's instruction of SureSelect XT reagent kit, HSQ (Agilent Technologies). The libraries for PBL and plasma DNAs were created using KAPA Hyper Prep Kit (Kapa Biosystems, Woburn, MA, USA) as described previously (26). Whereas 200 ng of PBL DNA was used to construct sequencing libraries for all samples, 37.12 ng of plasma DNA was used on average. Briefly, after end repair and A-tailing according to the

manufacturer's protocol, we performed adaptor ligation using a pre-indexed PentAdapter™ (PentaBase ApS, Denmark) at 4 °C overnight. After amplification through 9 PCR cycles, the library was analyzed for its quantity and fragment size distribution and then subjected to multiplexing hybrid selection for target enrichment. Hybrid selection was performed by using customized RNA baits that targeted ~499kb of the human genome, including exons from 83 cancer-related genes (Table S1). Purified libraries were pooled up to eight and each pooled library was adjusted to a total of 750ng for a hybrid selection reaction. Target enrichment was performed following the SureSelect (Agilent Technologies) bait hybridization protocol with the modification of replacing the blocking oligonucleotide with IDT xGen blocking oligonucleotide (IDT, Santa Clara, CA, USA) for the pre-indexed adapter. After the target enrichment step, the captured DNA fragments were amplified with 13 cycles of PCR using P5 and P7 oligonucleotides.

3. Sequencing and data processing

Based on DNA concentration and the average fragment sizes, libraries were normalized to an equal concentration of 2 nM and pooled by equal volume. After denaturing libraries using 0.2 N NaOH, libraries were diluted to 20 pM using hybridization buffer (Illumina, San Diego, CA, USA) and subjected to cluster amplification according to the manufacturer's protocol (Illumina). Flow cells were sequenced in the 100-bp paired-end mode using HiSeq 2500 v3 Sequencing-by-Synthesis Kits (Illumina) and then analyzed using RTA v.1.12.4.2 or later. Using BWA-mem (v0.7.5) (27), all of the raw data were

aligned to the hg19 human reference creating BAM files. SAMTOOLS (v0.1.18) (28), Picard (v1.93), and GATK (v3.1.1) (29) were used for sorting SAM/BAM files, local realignment, and duplicate markings, respectively. Through the process, we filtered reads to remove duplicates, discordant pairs, and off-target reads.

4. SNV detection in FNA samples and statistical test for their presence in plasma

For FNA biopsy specimens, MuTect 1.1.4 and VarScan2 were employed to detect somatic single nucleotide variants (SNVs) with matched germline (i.e., PBL) samples. For VarScan2, the default parameter values were used with some modifications as previously described (44). Somatic SNVs called by at least one of the methods were retained if they were present at a frequency less than 0.5% in the matched PBL sample and higher than 4% of the tumor sample. Somatic SNVs found in the FNA samples were listed and tested for their presence in the paired plasma samples as described previously (44). After background alleles in each sample had been adjusted by position-specific error rates, it was tested if the allele frequency of a given SNV ranked in the 95th percentile of adjusted background alleles.

5. Biopsy-free SNV identification in plasma DNA

A method slightly modified from previous studies (44, 48) was established to identify somatic SNVs in the plasma sample as described in our recent study [Park et al. 2017 Oncotarget]. Firstly, positions with the strand bias under 0.9 and the total read depth over 500 were considered for the analysis. All non-reference alleles were subjected to Phred quality filtering using a threshold Q

of 30. Non-reference alleles present at a frequency below 0.5% in the matched germline DNA were subjected to the binomial test to test if a non-reference allele was significantly more abundant in plasma DNA than the matched gDNA. The multiple testing adjustments were made through the Bonferroni correction. Next, Z-tests were performed to compare frequencies of non-reference alleles with their background allele frequency distribution obtained from the other plasma DNA samples. For the comparison, a background allele frequency distribution was generated by selecting non-reference alleles in plasma DNA present at a frequency <2.5% in the paired tumor and <0.5% in the paired germline DNA with a sufficient total depth ($\geq 250\times$ in tumor tissue, $\geq 500\times$ in PBL, and $\geq 500\times$ in plasma). Additionally, the following filters were applied: (1) candidate alleles with less than eight supporting reads were discarded; (2) when there were two or more candidates within any 10 bp window, all of them with an allele frequency <20% were discarded; (3) candidates with the Bonferroni adjusted p-value higher than 10^{-18} from the z-test were discarded. We further excluded SNV candidates if found as a germline SNP in other samples processed in the same lane of a sequencing flowcell to remove false positives due to cross-contamination among multiplexed samples. Nonsynonymous, stop-gain, stop-loss, and splicing-disrupting variants were listed as the final positive calls.

6. Droplet digital PCR validation

Mutant and wild-type alleles in plasma samples were quantified by QX200 Droplet Digital PCR System (BioRad, Hercules, CA, USA). All droplet

digital PCR (ddPCR) reagents except TaqMan assays (i.e., probes/primers) were ordered from Bio-Rad. TaqMan assays for KRAS p.G12D/G12V were ordered from Bio-Rad (PrimePCR ddPCR Mutation Assay), and RB1 p.R251* and ROS1 p.I1967V assays were customized by TaqMan SNP Genotyping Assays (Thermo Fisher Scientific, Waltham, MA, USA). All assay tests were performed in parallel with no-template and wild-type gDNA controls to monitor the false positive droplets. The concentrations of wild-type and mutant DNAs (copies per ul) in each sample were calculated by manufacturer's software and their concentrations in plasma (copies per mL) were derived as describe in van Ginkel, J.H *et al.*(65)

7. Statistical test

To calculate the limits of detection, the group was labelled with “detected”, “not detected”, or “discordant.” Group “detected” was categorized if positive droplet from ddPCR and read counts was determined from cfDNA sequencing and vice versa for group “not detected”. If the results present with discordant manner either by ddPCR or cfDNA sequencing, the “discordant” was labelled and categorized into false positive (See Table 2 and Supplemental Table 4). The calculation of confidence intervals for sensitivity and specificity was evaluated under the exact Clopper-Pearson confidence intervals. In confidence intervals of positive predictive value (PPV), the standard logit confidence intervals were followed as presented in Mercaldo *et al.* (66).

The rest of all statistical significances were evaluated by two-tailed tests, and the significance level was set at 5%. One-way analysis of variance (ANOVA) with least significance difference (LSD) posthoc analysis was used to

compare means across multiple groups. A one-sample t -test was applied to compare hotspot error rates between pretreatment and peri-/post-treatment samples. For i -th patient, j -th peri-/post-treatment sample and k -th hotspot, difference (D_{ijk}) is defined as following;

$$D_{ijk} = X_{ik}^{pre-treatment} - X_{ijk}$$

Null hypothesis was that mean error rates before and after treatment were not different ($H_0 : \bar{D} = 0$ vs. $H_a : \bar{D} \neq 0$).

RESULTS

Evaluation of LOD with single mutation

KRAS mutations

The vast majority of cancer related KRAS mutations in pancreatic cancer. To evaluate the limits of detection with single point mutations, the presence of KRAS mutations in baseline samples from 17 pancreatic cancer patients had evaluated. Among the 17 patients, 13 patients (76.5%) determined to have KRAS mutation from FNA samples. KRAS mutations were detected in 10 patients (58.8%) from plasma samples (Figure 1, Table 2-1). A general clue from the data, the allele frequency of KRAS mutations in plasma DNA is relatively lower than FNA samples. The allele frequency of KRAS mutations were $21.18\% \pm 4.06$ (mean \pm standard error of the mean (SEM)) in FNA samples and $2.02\% \pm 0.67$ (mean \pm SEM) in plasma samples. There can be two reasons: the limits of detection by targeted sequencing or sampling issue. To eliminate the possibility of modest detection sensitivity, orthogonal validation was performed by digital PCR. The overall tested samples were 62 samples which included consecutive samples from 14 patients. Conclusively, the analytical detection sensitivity presented 95.65 % sensitivity (95% confidence interval (CI) 78.05 to 99.89%), 100 % specificity (95% CI 91.24 to 100 %) for detecting KRAS mutations (Figure 2-1, Table 2-1). By these, the technical issue was neglected and remained only to the biological factor.

Evaluation of LOD with multi-mutations

“With primary” mutation

To evaluate the multiple mutations compare to point mutations, the customized capture-based targeted sequencing implied to FNA samples. The total of 40 M_{FNA} (mutations found in FNA sample) was determined via 17 pancreatic cancer patients (Figure 2-1A, Table 2-2). There were failure to detect significant mutations in two patients (P11 and P28). As described in Method, the determined mutations were tested to evaluate the significance of the presence in matched plasma DNA samples ($p < 0.001$). Figure 2-1B represents the total of 28 $M_{\text{P/FNA}}$ (mutations among M_{FNA} detected in their matched plasma samples) were significantly discriminated above the background noises of plasma DNAs, resulting in 70.0% detection sensitivity with the allele frequency (MAF) of $1.60\% \pm 0.31$ (mean \pm SEM) (Table 2-3).

Biopsy-free manner

The genotyping test is important for tracking the alteration of the architecture of genomics of primary tumor, but it does not represent the entire targeted regions. To assess the plasma mutations in broader regions, biopsy-free manner was implemented. The total of 27 $M_{\text{P/TR-BF}}$ in baseline plasma samples including 15 concordantly detected in FNAs (Figure 2-1) with the mean allele frequency of $3.54\% \pm 1.38$ (mean \pm SEM). The unique 12 mutations were detected only in plasma DNA samples but not in the matched FNA sample. To eliminate the possibility of false positive reports, the digital PCR was performed for orthogonal validation. Two individual variants (ROS1 p.I1967V and RB1 p.251X) were selected from the two patients (P2 and P5) and performed digital PCR to validate their presence in 8 consecutive plasma

samples (Table 2-4). Consistent with the cfDNA sequencing results, these mutations were detected in consecutive plasma samples from the patients, indicating the variants unique to plasma were not likely to be false positives. To determine the levels of false positive due to the technical background noises from the “biopsy-free manner” method, the series of biologically replicated sequencing data using PBL DNAs from six patients were evaluated. One of replicates from each patient was paired with the other as a mock for a matched plasma sample and was processed for variant detection (Method). By testing the total of 21 replicates, there were absolutely no mutations were detected from the pipeline. By these result, the minimal false discovery rate was assured by involvement of technical background errors. Collectively, the algorithm of biopsy-free manner is feasible and useful to detect tumor mutations across the entire target regions. Limitations in genetic profiling using FNAs have been recognized as FNAs are not sufficient to represent all regional subclone events. The data suggested that somatic profiling mutations of plasma DNA in a biopsy-free manner compensate the shortcomings of FNA, revealing intra-tumor heterogeneity. Based on $M_{P/TR}$ (plasma DNA mutation across entire target regions) by combining $M_{P/FNA}$ with $M_{P/TR-BF}$, I were able to detect ctDNAs in 15 pretreatment samples suggesting the advantage of profiling broader genomic regions than KRAS hotspots.

Monitoring tumor burden by measuring ctDNA

Another merit of ctDNA analysis is the ability to monitoring the alteration of tumor genomics in real-time. Taking advantage of the serial collection of

blood draws, the responses of chemotherapy and the progression of disease (PD) examined to correlate with the level of ctDNA. Also, the level of CA 19-9 measured alongside during the clinical follow-up. Blood was collected separately for the each of the tests. Figure 2-2 displayed nine patients under the therapeutic intervention except for one patient (P27). P27 detected with no significant variants from both of the primary and plasma samples. Consecutive samples detected spartial plasma mutations (Table 2-5, Figure 2-3), but the patient had stable diseases (SD) status along the follow-up period. The diases progression and the therapy responses were determined under the CT images (data not shown). Among the nine patients, four patients (P2, P7, P42, and P43) presented the correlated trend in both $M_{P/TR}$ amount and CA19-9 level throughout the therapeutic intervention. On the other hand, three patient (P5, P31, and P36) had discordant level of CA 19-9 in the limited period of time but amount of ctDNA matched with the CT images. Moreover, in P11, the truncation of TP53 p.E297X was detected at the fifth plasma sampling of P11. After a month later, P11 diagnosed with liver and peritoneum metastasis. It is noteworthy that FNA sampling of P11 failed to determine any of significant mutations (Figure 2-2 and 2-3). In this context, by observing the level of $M_{P/TR}$, the alteration of $M_{P/TR}$ level was on average of 2 months ahead of the CT image changes. Figure 11 displays the overall detected mutations in plasma samples. Overall, the data suggested that tracking the level of MP/TR is better surrogate marker than CA 19-9 accounted as the monitoring of resistance of chemotherapy and/or disease progression.

Diagnostic utility

Figure 2-4 shows the comparison of ctDNA level according to the time of diagnosis (Dx) with RECIST (Response Evaluation Criteria In Solid Tumors) as a group; Complete Response/Partial Response (CR/PR), Stable Disease (SD), and Progressive Disease (PD). Figure 2-4A represent the allele frequencies of $M_{P/KRAS}$ (ANOVA, LSD, p-value = 0.084) and $M_{P/FNA}$ were not significantly different among the disease status (ANOVA, LSD, p-value=0.519, Figure 2-4B). On the other hand, the allele frequencies of $M_{P/TR}$ significantly varied among the disease statuses (ANOVA, LSD, p-value = 0.001, Figure 2-4C). The allele frequencies of $M_{P/TR}$ at the near time of PD was significantly higher (mean \pm SEM: 4.17% \pm 0.93) than those at the time of Dx (mean \pm SEM: 3.54 \pm 1.55%), SD (mean \pm SEM: 1.32 \pm 0.16%) or CR/PR (mean \pm SEM: 1.82 \pm 0.29%) (Figure 2-4C; ANOVA, LSD, p-value = 0.001). The level of CA 19-9 was also evaluated but not significantly followed the patients' disease status accordingly (ANOVA, LSD, p-value = 0.13, Figure 2-4D). These results suggested that the amount and/or allele frequency of ctDNA well indicated real-time disease status compare to the level of CA 19-9.

Next, the number of detected variants was quantified along the therapeutic intervention (Figure 2-5). As the extension of the period of treatment, it is expected to increase the number of detectable mutations as the allele frequencies of $M_{P/TR}$ varied according to the disease status. The number of detected $M_{P/TR}$ significantly different from the disease status

(Figure 2-5B; ANOVA, LSD, p-value = 5.71×10^{-8}). Moreover, the number of MP/FNA and MP/TR significantly decreased at the time of CR/PR compared to Dx, while the number of MP/TR increased at the time of PD.

Interestingly, the number of $M_{P/FNA}$ and $M_{P/TR}$ dropped after the treatment started (1.05 ± 1.11). The number of variants was the lowest (0.73 ± 1.61) at around 4 months after treatment and started to cumulate (up to 1.90 ± 1.03) as treatment period expanded (Figure 2-5 C). Collectively, the mean number of $M_{P/TR}$ per sample significantly increased depending on the duration of chemotherapy treatment (ANOVA, LSD, p-value = 0.004). The results indicated that $M_{P/TR}$ better represented real-time disease status either by allele frequency and/or a number of mutations than $M_{P/FNA}$.

DISCUSSION

In this study, the ctDNA detection methods was evaluated comparing between targeted deep sequencing with the broad-range and KRAS-oriented analysis.

The analysis highlighted the importance of considering broad-scale ctDNA analysis by allowing to characterize not only intra-heterogeneity limited by tumor biopsy but also to monitor the primary mutations which impacted on the diagnostic accuracy along the therapeutic intervention.

KRAS mutations are well-documented as initiating factor for the development of PDAC (67). However, often, the low detection sensitivity in ctDNA analysis fueled the debate about the capability of as its biomarker (60, 62).

Therefore, comparative evaluation of ctDNA detection approaches for PDAC has to done. Among the 17 pretreatment plasma samples, ctDNA detected in ten, twelve, and fifteen samples by profiling $M_{P/KRAS}$, $M_{P/FNA}$, and $M_{P/TR}$, respectively, indicating the advantage of profiling broad genomic regions on sensitivity for cancer detection. Moreover, the improved sensitivity of ctDNA detection subsequently enhanced tumor monitoring by longitudinal cfDNA analysis. For instance, in P5 and P42 patient, although a KRAS mutation was not detectable in not only pretreatment samples but also all following peri/post-treatment samples, the independent variants in other genes were coherently correlated with tumor burden (Figure 2-2). In P2 patient, the level of ROS1 p.I1967V dramatically decreased after surgical operation indicating tumor removal, although KRAS mutation was not detected before and right after the surgery (Figure 2-2A).

Despite its advantages, interrogation of broader genomic regions might result in more false positives especially when performed in a biopsy-free manner. To minimize the false positives, the stringent filtering steps was applied for calling $M_{P/TR-BF}$. Then, the filtering steps during variant calling were adequately established to minimize false discovery rate. Analyzing duplicated PBL gDNA sequencing data, the data showed that false positives due to the technical background were minimal as described in the Results section. In addition, some of $M_{P/TR-BF}$ was validated which were not detected in FNA specimens by dPCR. In present approach, taking advantage of blood sampling strengthened to neglecting the potential interruption from the biological or technological background noises. On the other hand, as the stringent filtering steps perhaps minimized the detection sensitivity. To improve the detection sensitivity, the present study merged the primary mutations and plasma mutations. However, in the future study, the limitation may overcome by adapting the molecular barcoding which will increase the uniqueness of the reads reliability of low read counts of variants.

The present study not only provides a small number of patients but also randomly selected cancer stages that limit the detection sensitivity depended upon the disease stages. Also, the study of design could not approach the “combination assay” with protein biomarkers as the recent study suggested (68). Regardless of number of patients, as if the threshold of CA 19-9 increased to 100 U/mL followed by previous study, three of patients’ data (P23, P31 and P36) cannot be included. Also, it did not affect the detection sensitivity nor compensate the diagnostic status (ANOVA, LSD, p-

value=0.59). Another obvious hurdle is cost of the targeted deep sequencing. It is evident that single mutation analysis cost cheaper and faster analysis. However, emerging evident shows the dramatic reduction of cost of sequencing may balance out in near future.

Although the present data only dealt with the plasma DNA, the cell-free DNA certainly can observe other types of body fluid. The application of the enhanced NGS-method allowed to analysis pleural effusion (PE) fluid DNA. The comparative genomic alterations were discovered in either from PE cfDNA or plasma cfDNA (Figure 2-6). Cumulative number of detected somatic mutations shared the gene feature from the PE and plasma samples. As the traditional PE test depended on the collection of PE cells from PE fluid, the allele frequency of PE cells was evaluated comparatively. Nevertheless, the allele frequency of detected somatic mutations was not correlated between the PE cfDNA and PE cell. Interestingly, the PE cfDNA and plasma cfDNA had higher correlation (Figure 2-7). Next, the detected mutations were compared with the clinical history. It turned out the correlation of clinical history and PE cfDNA had highly matched than the plasma cfDNA (data not shown). The phenomena indicate the bias of sample collection could be contributed throughout the surveillance. In addition, the allele frequencies of PE cells provided the least information compared to plasma and PE cfDNA. Another interesting factor was the PE cfDNA had the mediator role between the plasma and PE cell. The exclusive detection was observed in overall detected mutations (data not shown). The phenomenon may be contributed by the microenvironment of body fluid such as the physical or chemical effect of

the immune cells lead to the apoptosis or necrosis. The assumption could be made through the comparison of fragment size from PE cfDNA and plasma cfDNA (Figure 2-8). Therefore, it is important to show various types of body fluid can compensate the tumor biopsy as its invasive procedure and modest rate of sensitivity. In summary, the interrogation of circulating tumor DNA with targeted deep sequencing would be informative to analyze the unmet needs in cancer research.

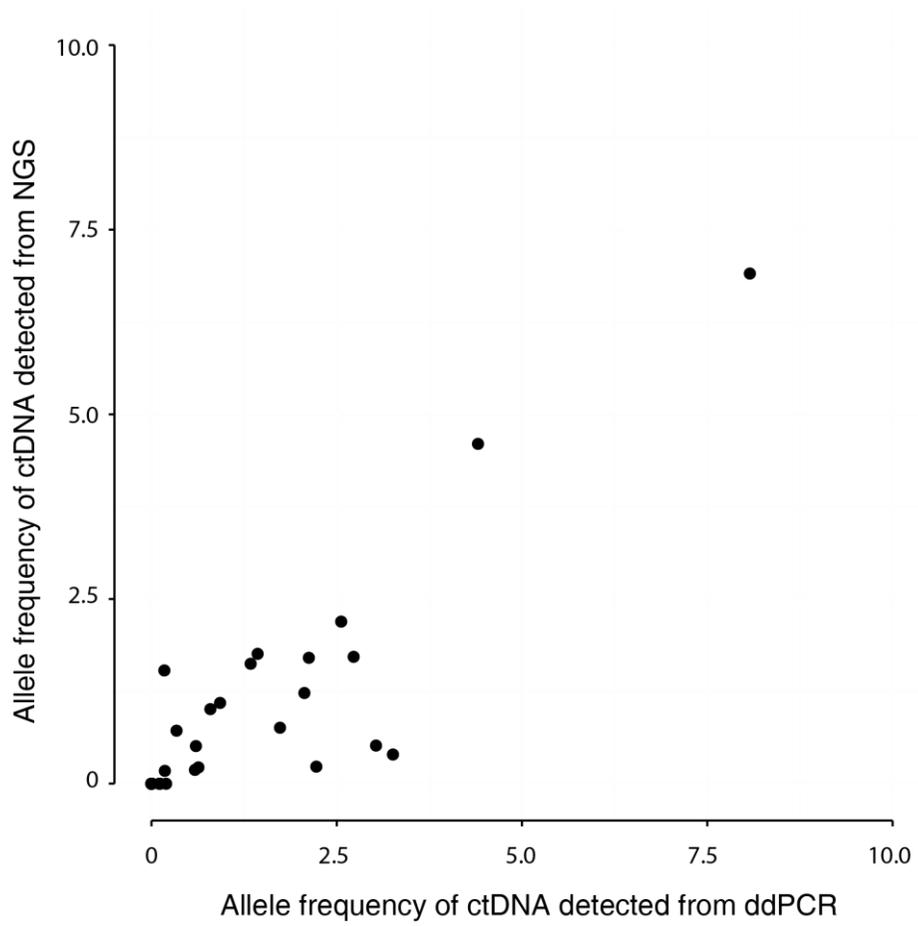


Figure 2-1. The correlation of harbored KRAS mutations using digital PCR and enhance NGS-method from pancreatic cancer patients.

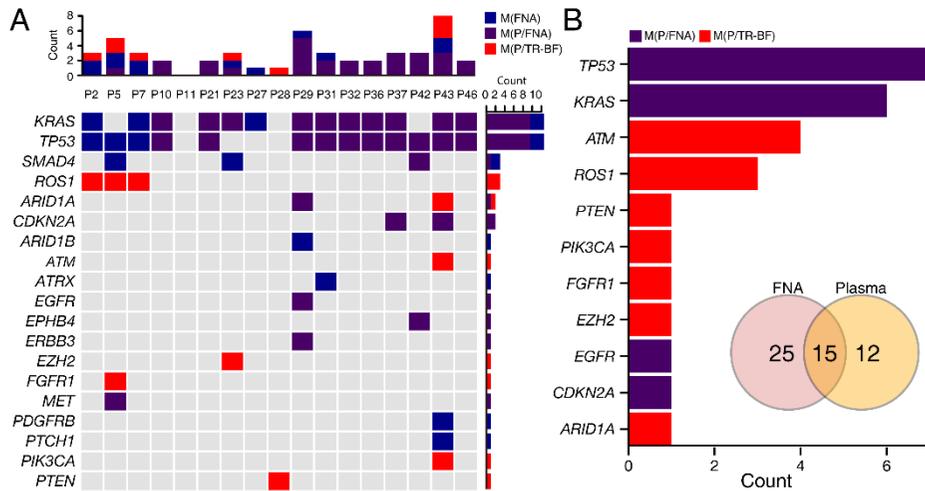
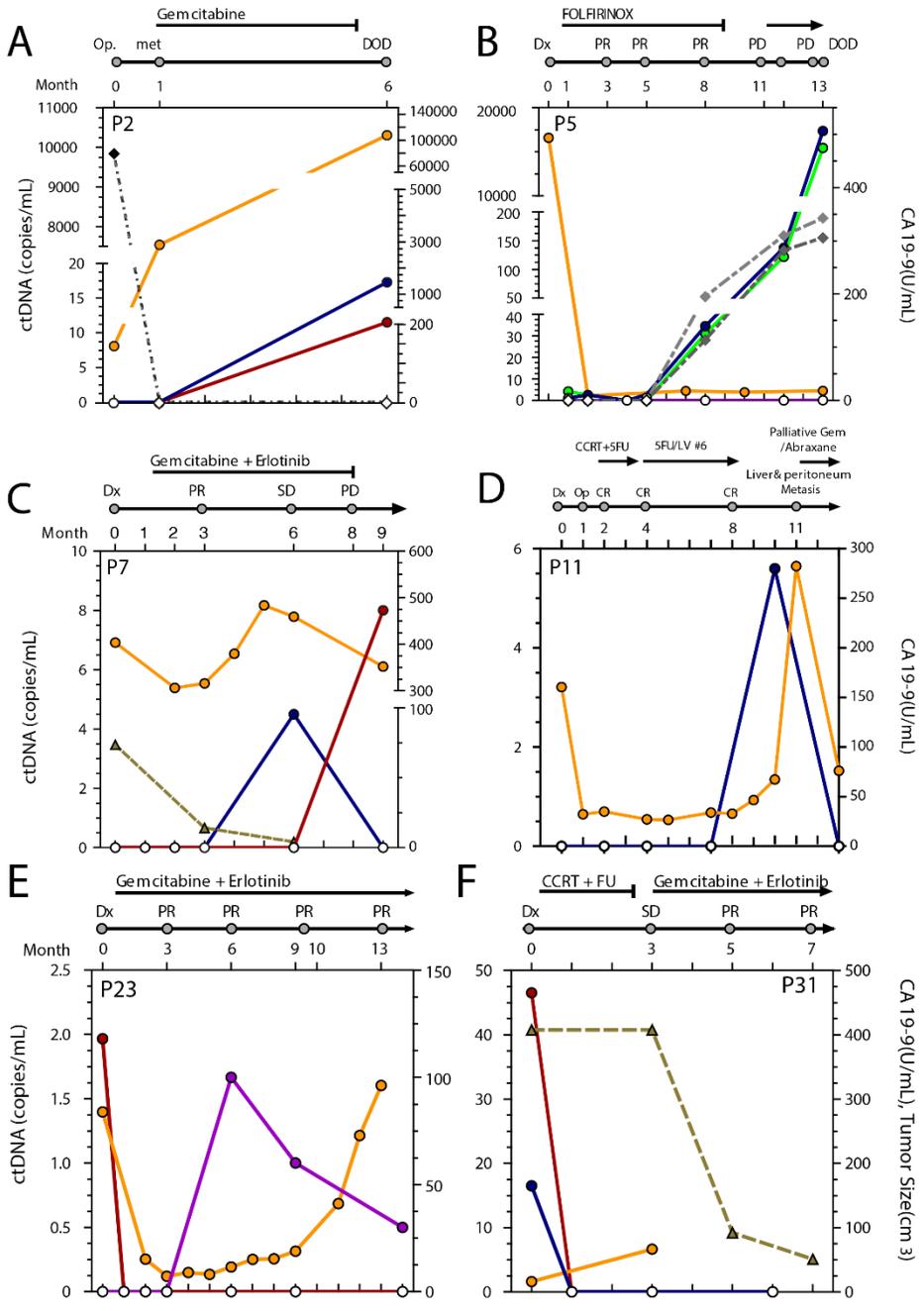


Figure 2-2. Tumor mutations in pre-treatment cfDNA samples from 17 PDAC patients

The top panel summarized the presence of detected mutation across the 17 patients depending on the detection methods (i.e., $M_{P/KRAS}$, $M_{P/FNA}$, $M_{P/TR}$). While interrogation of KRAS hotspots detected mutations ($M_{P/KRAS}$) in plasma samples from 10 patients, testing variants detected from FNA samples ($M_{P/FNA}$) and entire target regions ($M_{P/TR}$) detected tumor variants in 12 and 14 plasma samples, respectively. The oncoprint chart shows M_{FNA} and $M_{P/TR}$. If a variant is concordantly detected in both M_{FNA} and $M_{P/TR}$, the variant also corresponds to $M_{P/FNA}$. The number of affected genes for each patient is plotted the bottom of the chart. The number of samples that harbor a mutation for each gene is plotted the right side of the chart. *Four independent mutations in ATM were detected in the P43 plasma sample.



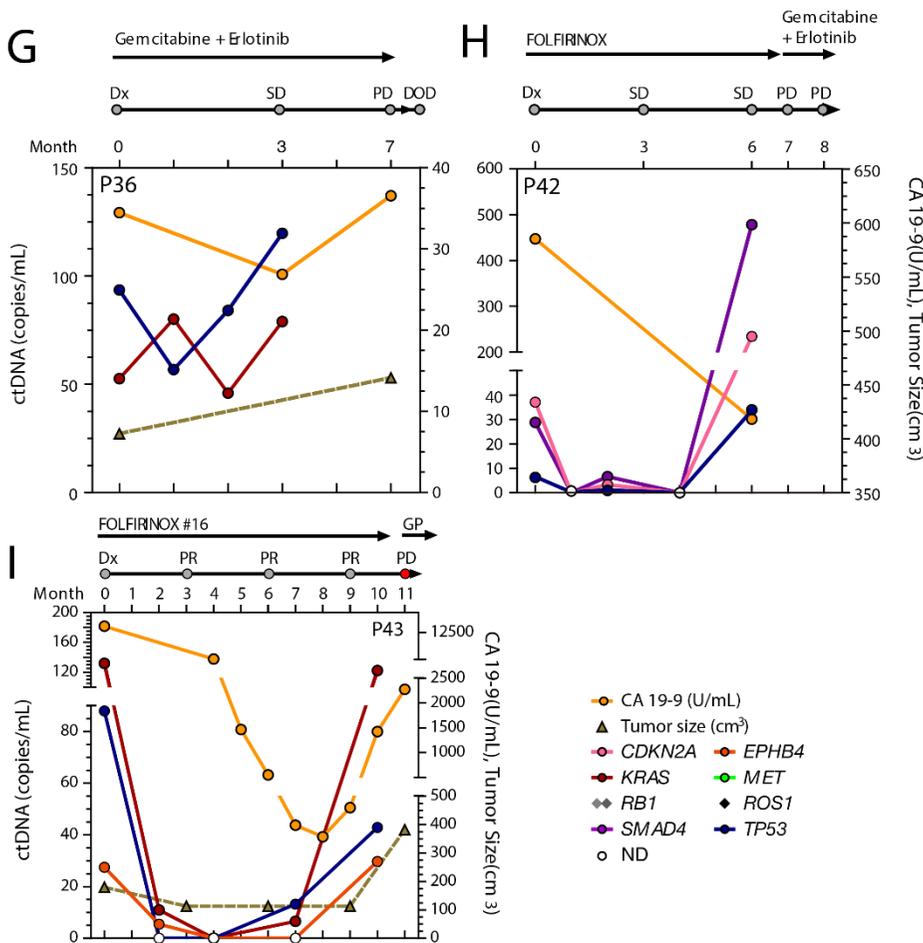


Figure 2-3. Monitoring of ctDNA PDAC patients under therapeutic intervention.

The level of ctDNA estimated by each SNV was plotted on the left y-axis for eight patients (A-D). Chemotherapeutic agents administered to each patient and therapy response evaluated based on Response Evaluation Criteria In Solid Tumors (RECIST) were displayed on top of the graph. The level of ctDNA determined either by MP/FNA (solid line) or MP/TR-BF (dotted line) was displayed in various color depending on the mutated genes. CA 19-9 level (yellow solid line) and tumor size (grey dotted line) based on CT images were plotted against the right y-axis. CCRT, concurrent chemoradiation therapy; FU, fluorouracil; CR, complete response; DOD, dead of disease; Dx, diagnosis; Met, Metastasis; ND, not detected; Op, operation; PD, progressive disease; PR, partial response; SD, stable disease.

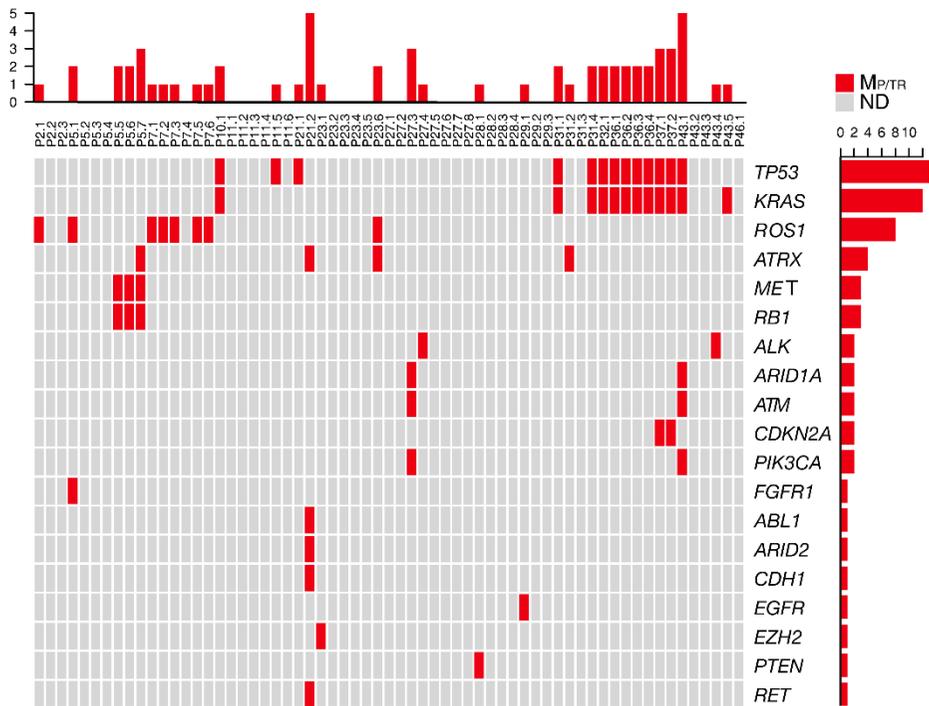


Figure 2-4. Summary of plasma mutations determined by “biopsy-free manner.”
 Total of 19 genes was determined and ordered by number of detected mutation per patient.

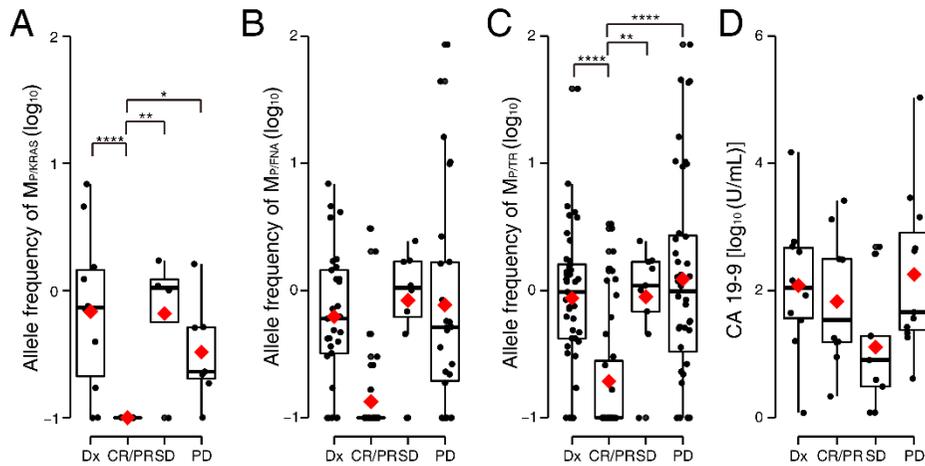


Figure 2-5. Allelic fraction of ctDNA and CA19-9 level depending on therapy responses.

The allelic frequencies of (A) $M_{P/KRAS}$ (B) $M_{P/FNA}$ and (C) $M_{P/TR}$ were box-plotted depending on their near-time therapy response evaluations. (D) CA 19-9 levels were box-plotted. All of the determined levels were displayed on a logarithmic scale. The level of statistical significance is indicated by the asterisks in the figures; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, and **** $P \leq 0.0001$. Dx, diagnosis; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

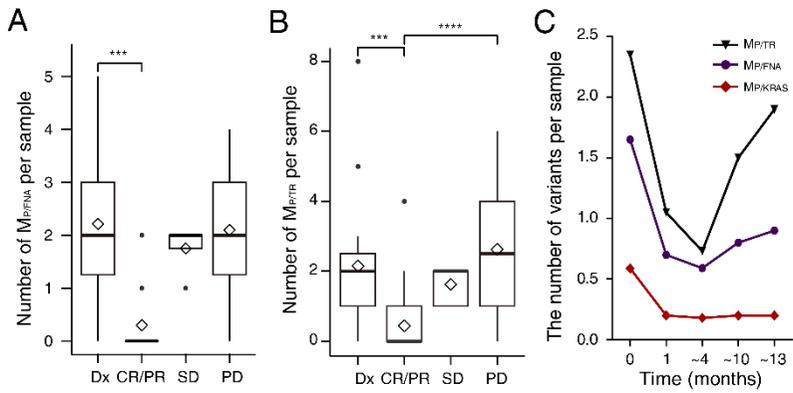


Figure 2-6. The number of mutations in plasma DNA

The number of (A) $M_{P/FNA}$ and (B) $M_{P/TR}$ presented per patient samples was categorized according to near-time disease status. (C) The number of mutations was shown depending on the period of treatment. The level of statistical significance is indicated by the asterisks in the figures; * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, and **** $P \leq 0.0001$. Dx, diagnosis; CR, complete response; PR, partial response; SD, stable disease; PD, progressive disease.

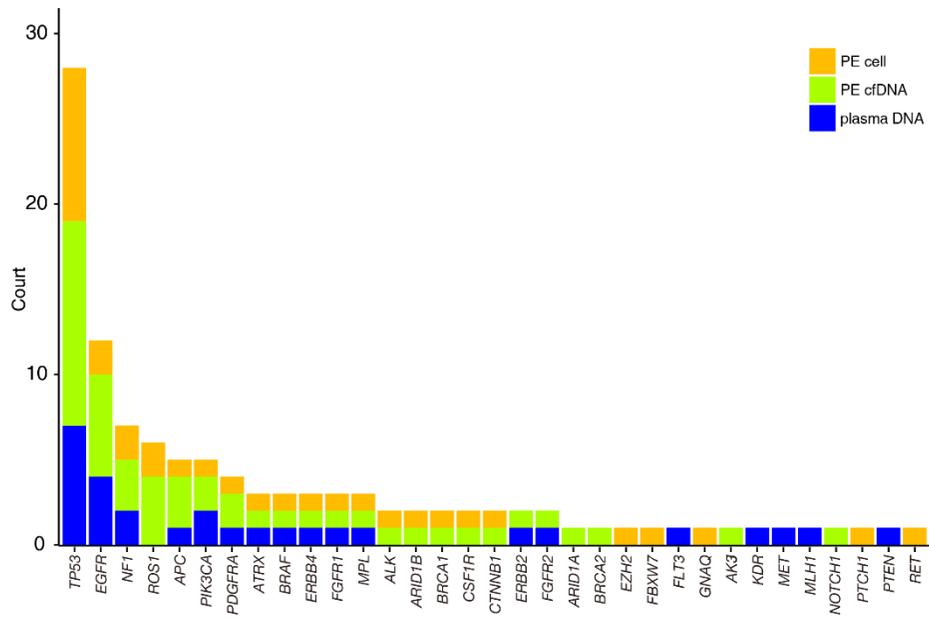


Figure 2-7. Distribution of detected genes from pleural effusion fluid and plasma DNA

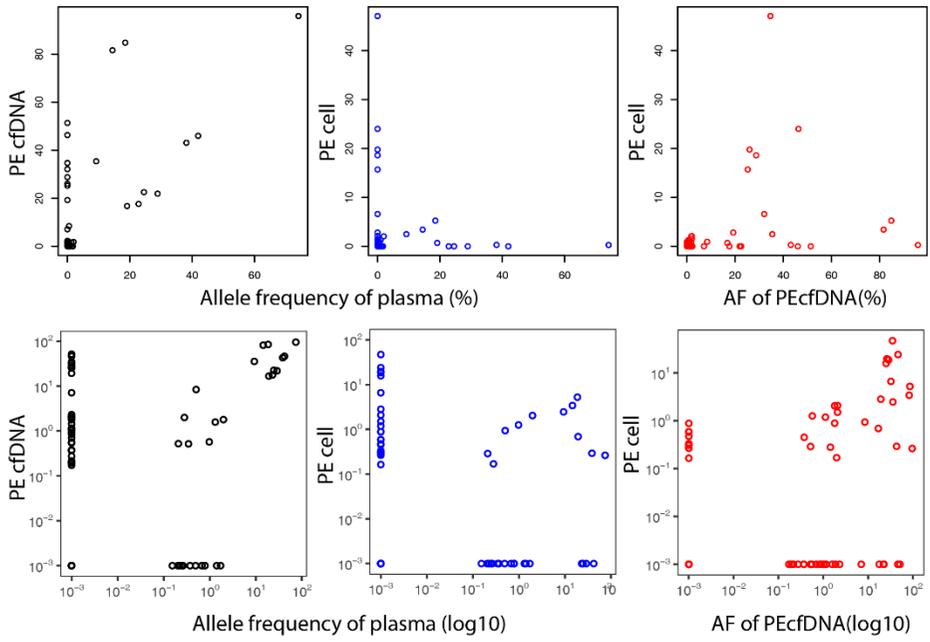


Figure 2-8. The differences of allele frequencies from pleural effusion fluid and plasma DNA

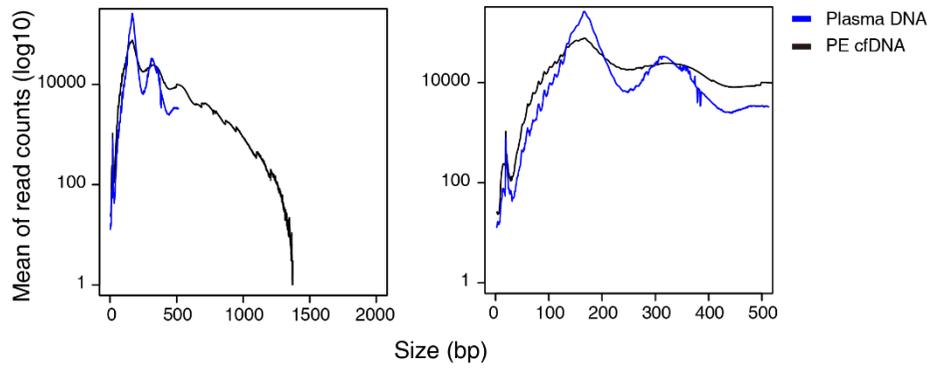


Figure 2-9. The size distribution of pleural effusion and plasma DNA

Table 2-1. The limits of detection sensitivity evaluated by KRAS mutations in 14 PDAC patients.

Patient ID	Gene	AA Change	dpcr	NGS
P2.1	KRAS	G12D	0	0
P2.1	KRAS	G12V	0	0
P2.2	KRAS	G12D	1	1
P2.3	KRAS	G12D	1	1
P5.1	KRAS	G12D	0	0
P5.1	KRAS	G12V	0	0
P7.1	KRAS	G12V	0	0
P7.2	KRAS	G12V	0	0
P7.3	KRAS	G12V	0	0
P7.4	KRAS	G12V	0	0
P7.5	KRAS	G12D	0	0
P7.5	KRAS	G12V	0	0
P7.6	KRAS	G12V	1	1
P10.1	KRAS	G12D	1	1
P10.1	KRAS	G12V	0	0
P21.1	KRAS	G12V	1	1
P21.2	KRAS	G12V	1	1
P23.1	KRAS	G12V	1	1
P23.1	KRAS	G12D	0	0
P23.2	KRAS	G12D	0	0
P23.3	KRAS	G12V	0	0
P23.4	KRAS	G12V	0	0
P23.4	KRAS	G12D	0	0
P23.5	KRAS	G12V	0	0
P23.6	KRAS	G12V	0	0
P23.7	KRAS	G12V	0	0
P27.2	KRAS	G12V	0	0
P27.2	KRAS	G12D	0	0
P27.3	KRAS	G12D	0	0
P27.4	KRAS	G12D	0	0
P27.5	KRAS	G12D	0	0
P27.6	KRAS	G12D	0	0
P27.7	KRAS	G12D	0	0

P29.1	KRAS	G12V	1	1
P29.2	KRAS	G12V	1	1
P29.3	KRAS	G12V	1	1
P31.1	KRAS	G12D	1	1
P31.2	KRAS	G12D	1	1
P31.3	KRAS	G12D	1	1
P31.4	KRAS	G12D	1	1
P32.1	KRAS	G12D	1	1
P32.1	KRAS	G12V	1	0
P36.1	KRAS	G12D	1	1
P36.2	KRAS	G12D	0	0
P36.3	KRAS	G12D	0	0
P36.4	KRAS	G12D	0	0
P37.1	KRAS	G12D	1	1
P37.1	KRAS	G12V	0	0
P37.2	KRAS	G12D	1	1
P43.1	KRAS	G12D	1	1
P43.2	KRAS	G12D	1	1
P43.2	KRAS	G12V	0	0
P43.3	KRAS	G12D	0	0
P43.3	KRAS	G12V	0	0
P43.4	KRAS	G12D	1	1
P46.1	KRAS	G12D	1	1
P46.1	KRAS	G12V	0	0

Table 2-2. Determined mutations from 17 FNA samples

Patient #	DNA	Gene	AA change	Primary VAF (%)
P2	FNA	KRAS	p.G12D	21.12
P2	FNA	TP53	p.Y236C	25.36
P5	FNA	MET	p.S907F	33.56
P5	FNA	SMAD4	p.Q256X	9.27
P5	FNA	TP53	c.96+1G>A	65.62
P7	FNA	KRAS	p.G12V	8.73
P7	FNA	TP53	p.R175H	8.45
P10	FNA	KRAS	p.G12D	13.94
P10	FNA	TP53	c.386+1G>A	14.90
P21	FNA	KRAS	p.G12V	14.69
P21	FNA	TP53	p.I30S	20.77
P23	FNA	KRAS	p.G12V	16.37
P23	FNA	SMAD4	p.R361C	23.29
P27	FNA	KRAS	p.G12D	3.55
P29	FNA	KRAS	p.G12V	4.47
P29	FNA	ARID1B	p.Q2092X	3.79
P29	FNA	EGFR	p.K189E	4.73
P29	FNA	ERBB3	p.P583S	4.15
P29	FNA	TP53	p.L194R	4.75
P31	FNA	KRAS	p.G12D	38.64
P31	FNA	ATRX	p.A3T	6.06
P31	FNA	TP53	p.F113C	54.45
P32	FNA	KRAS	p.G12D	19.73
P32	FNA	TP53	p.H154R	27.96
P36	FNA	KRAS	p.G12D	16.85
P36	FNA	TP53	p.Y220C	23.65
P37	FNA	CDKN2A	p.C100X	60.99
P37	FNA	KRAS	p.G12D	55.64
P37	FNA	TP53	p.R119L	59.97
P42	FNA	TP53	p.L226P	3.36
P42	FNA	SMAD4	p.R361C	4.50
P42	FNA	EPHB4	p.F404L	9.34
P43	FNA	KRAS	p.G12D	28.57
P43	FNA	CDKN2A	p.A97V	23.08

P43	FNA	TP53	p.D220V	29.62
P46	FNA	PDGFRB	p.P866S	17.08
P46	FNA	PTCH1	c.3606+1G>A	21.65
P46	FNA	KRAS	p.G12D	33.09
P46	FNA	PDGFRB	p.L865F	18.26
P46	FNA	TP53	p.Q167X	25.07

Table 2-3. Evaluation of FNA mutations in baseline plasma DNA samples

PlasmaID	Gene	AA Change	Allele Freq.	p-value
P5.1	MET	p.S907F	0.31	0.000325
P10.1	KRAS	p.G12D	2.19	0.000207
P10.1	TP53	c.386+1G>A	4.11	0.000206
P21.1	KRAS	p.G12V	1.23	0.000187
P21.1	TP53	p.I30S	1.70	0.000185
P23.1	KRAS	p.G12V	0.17	0.000700
P29.1	ARID1B	p.Q2092X	0.32	0.000310
P29.1	EGFR	p.K189E	1.23	0.000212
P29.1	ERBB3	p.P583S	0.47	0.000235
P29.1	TP53	p.L194R	0.42	0.000242
P29.1	KRAS	p.G12V	0.40	0.000251
P31.1	KRAS	p.G12D	6.90	0.000197
P31.1	TP53	p.F74C	3.75	0.000198
P32.1	KRAS	p.G12D	1.70	0.000209
P32.1	TP53	p.H154R	1.27	0.000211
P36.1	KRAS	p.G12D	0.76	0.000193
P36.1	TP53	p.Y88C	1.17	0.000192
P37.1	CDKN2A	p.C100X	1.77	0.000186
P37.1	KRAS	p.G12D	4.60	0.000182
P37.1	TP53	p.R119L	4.14	0.000184
P42.1	TP53	p.L226P	0.36	0.000258
P42.1	SMAD4	p.R361C	0.42	0.000221
P42.1	EPHB4	p.F404L	0.49	0.000191
P43.1	KRAS	p.G12D	1.53	0.000198
P43.1	CDKN2A	p.A97V	0.60	0.000229
P43.1	TP53	p.D220V	1.45	0.000198
P46.1	KRAS	p.G12D	0.72	0.000187
P46.1	TP53	p.Q167X	0.62	0.000195

*P31.1 ATRX discarded according to its insufficient depth coverage

Table 2-4. The performance of droplet digital PCR in plasma samples

Quantification of copies/mL of plasma DNA was calculated by following (Method derived by Ginkel et al. 2017):

$$Px = Cx * RV * \frac{EV}{TV}$$

Pmt = Mutant concentration in plasma (copies/mL); Pwt = Wild type concentration in plasma (copies/mL)

Cmt = Mutant sample concentration (copies/uL); Cwt = Wild type sample concentration (copies/uL)

RV = Total reaction volume of digital PCR (20uL)

EV= Elution volume of cfDNA (50-75uL)

TV = cfDNA volume used for dPCR reaction (8uL)

PV= Plasma volume for cfDNA extraction (1-5mL)

Patient #	Gene	AA Change	Input DNA (ng)	P V	RV (uL)	EV (uL)	TV (uL)	Pmt (copies/mL)	Pwt (copies/mL)	PCmt/PCwt (%)	Cmt (Droplet)	Cwt (Droplet)	Total droplets	Cwt (copies/ul)	Cwt (copies/20ul)	Cmt (copies/ul)	Cmt (copies/20ul)
P2.1	KRAS	G12D	5.55	3	20	70	8	0	3063	0.00	0	806	806	53	1050	0	0
P2.1	KRAS	G12V	5.55	3	20	70	8	0	2952	0.00	0	740	740	51	1012	0	0
P2.1	ROS1	H1967V	5.55	3	20	70	8	181	3710	4.87	44	868	912.00	63.6	1272	3	62
P2.2	KRAS	G12D	3.60	3	20	70	8	5	4433	0.12	1	2952	2953	76	1520	0	2
P2.3	KRAS	G12D	8.80	2	20	70	8	140	23888	0.59	19	1240	1259	273	5460	2	32
P5.1	KRAS	G12D	8.32	2	20	70	8	0	10588	0.00	0	1639	1639	121	2420	0	0
P5.1	KRAS	G12V	8.32	2	20	70	8	0	10588	0.00	0	1608	1608	121	2420	0	0
P5.1	RB1	R251*	8.32	2	20	70	8	17	23100	0.07	2	2478	2480.00	264	5280	0	4
P5.2	RB1	R251*	11.12	3	20	70	8	58	13883	0.42	12	2503	2515.00	238	4760	1	20
P5.3	RB1	R251*	21.60	2	20	70	8	7	35350	0.02	1	4559	4560.00	404	8080	0	2
P5.4	RB1	R251*	41.92	5	20	70	8	6	35000	0.02	2	8413	8415.00	1000	20000	0	3
P5.5	RB1	R251*	11.92	5	20	70	8	273	10780	2.53	104	3610	3714.00	308	6160	8	156
P5.6	RB1	R251*	13.44	5	20	70	8	952	7700	12.36	338	2513	2851.00	220	4400	27	544
P5.7	RB1	R251*	106.00	5	20	70	8	24850	33880	73.35	6751	8359	15110.00	968	19360	710	14200

P7.1	KRAS	G12V	10.32	3	20	70	8	0	8575	0.00	0	1928	1928	147	2940	0	0
P7.2	KRAS	G12V	12.80	5	20	70	8	0	6020	0.00	0	2016	2016	172	3440	0	0
P7.3	KRAS	G12V	15.68	3	20	70	8	0	15167	0.00	0	3335	3335	260	5200	0	0
P7.4	KRAS	G12V	9.52	2	20	70	8	0	12950	0.00	0	1944	1944	148	2960	0	0
P7.5	KRAS	G12V	6.27	2	20	70	8	0	6694	0.00	0	907	907	77	1530	0	0
P7.5	KRAS	G12D	6.27	2	20	70	8	0	6746	0.00	0	985	985	77	1542	0	0
P7.6	KRAS	G12V	26.72	5	20	70	8	32	5250	0.60	10	1717	1727	150	3160	1	18
P10.1	KRAS	G12D	9.20	3	20	70	8	181	7058	2.56	34	1243	1277	121	2420	3	62
P10.1	KRAS	G12V	9.20	3	20	70	8	0	8575	0.00	0	2015	2015	147	2940	0	0
P21.1	KRAS	G12V	8.48	2	20	70	8	219	10588	2.07	29	1318	1347	121	2420	3	50
P21.2	KRAS	G12V	2.82	2	20	70	8	105	4716	2.23	16	688	704	54	1078	1	24
P23.1	KRAS	G12V	2.32	2	20	70	8	41	22750	0.18	6	3004	3010	260	5200	0	9
P23.1	KRAS	G12D	2.32	2	20	70	8	0	2336	0.00	0	449	449	27	534	0	0
P23.2	KRAS	G12V	3.17	2	20	70	8	0	114	0.00	0	252	252.00	1.3	26	0	0
P23.2	KRAS	G12D	3.17	2	20	70	8	0	1733	0.00	0	265	265	20	396	0	0
P23.3	KRAS	G12V	7.76	2	20	70	8	0	12250	0.00	0	1529	1529	140	2800	0	0
P23.4	KRAS	G12V	2.59	2	20	70	8	0	3789	0.00	0	440	440	43	866	0	0
P23.4	KRAS	G12D	1.82	1	20	50	8	0	1675	0.00	0	184	184	13	268	0	0
P23.5	KRAS	G12V	28.64	5	20	70	8	0	7175	0.00	0	2351	2351	205	4100	0	0
P23.6	KRAS	G12V	11.04	5	20	70	8	0	8120	0.00	0	2776	2776	232	4640	0	0
P23.7	KRAS	G12V	102.0 0	5	20	70	8	0	9135	0.00	0	3010	3010	261	5220	0	0
P27.1	KRAS	G12D	1.25	5	20	70	8	0	676	0.00	0	236	236.00	19.3	386	0	0
P27.1	KRAS	G12V	1.25	5	20	70	8	0	546	0.00	0	183	183.00	15.6	314	0	0

P27.2	KRAS	G12D	4.22	5	20	70	8	0	1379	0.00	0	443	443	39	788	0	0
P27.2	KRAS	G12V	3.18	5	20	70	8	0	1281	0.00	0	468	468	37	732	0	0
P27.3	KRAS	G12D	9.20	5	20	70	8	0	3780	0.00	0	1100	1100	108	2160	0	0
P27.4	KRAS	G12D	8.32	5	20	70	8	0	3710	0.00	0	1096	1096	106	2120	0	0
P27.5	KRAS	G12D	4.99	5	20	70	8	0	1869	0.00	0	679	679	53	1068	0	0
P27.6	KRAS	G12D	9.76	5	20	70	8	0	1246	0.00	0	452	452	36	712	0	0
P27.6	KRAS	G12V	2.37	5	20	70	8	0	963	0.00	0	342.00	342.00	27.50	550.00	0	0
P27.7	KRAS	G12D	2.48	5	20	70	8	0	1106	0.00	0	468	468	32	632	0	0
P27.8	KRAS	G12D	1.25	5	20	70	8	0	546	0.00	0	209	209.00	15.6	312	0	0
P29.1	KRAS	G12V	4.19	2	20	70	8	114	3491	3.26	15	464	479	40	798	1	26
P29.2	KRAS	G12V	7.22	2	20	70	8	45	7018	0.64	7	1063	1070	80	1604	1	10
P29.3	KRAS	G12V	3.44	2	20	70	8	70	2310	3.03	10	309	319	26	528	1	16
P31.1	KRAS	G12D	11.12	2	20	70	8	114	1409	8.07	15	191	206	16	322	1	26
P31.2	KRAS	G12D	26.24	2	20	70	8	0	7263	0.00	0	979	979	83	1660	0	0
P31.3	KRAS	G12D	30.88	2	20	70	8	0	5023	0.00	0	637	637	57	1148	0	0
P31.4	KRAS	G12D	4.11	2	20	70	8	0	4506	0.00	0	588	588	52	1030	0	0
P32.1	KRAS	G12D	6.00	2	20	70	8	123	5766	2.12	18	808	826	66	1318	1	28
P32.1	KRAS	G12V	6.00	2	20	70	8	5	5110	0.10	1	909	910	58	1168	0	1
P36.1	KRAS	G12D	3.28	2	20	70	8	184	10588	1.74	23	1284	1307	121	2420	2	42
P36.2	KRAS	G12D	1.86	2	20	70	8	58	7263	0.80	8	978	986	83	1660	1	13
P36.3	KRAS	G12D	15.84	5	20	70	8	19	2002	0.93	6	635	641	57	1144	1	11
P36.4	KRAS	G12D	11.44	5	20	70	8	49	1796	2.73	16	585	601	51	1026	1	28
P37.1	KRAS	G12D	5.23	3	20	70	8	134	3045	4.41	31	691	722	52	1044	2	46
P37.1	KRAS	G12V	5.23	3	20	70	8	0	3127	0.00	0	842	842	54	1072	0	0

P37.2	KRAS	G12D	6.22	3	20	70	8	49	3424	1.43	11	746	757	59	1174	1	17
P43.1	KRAS	G12D	2.66	3	20	70	8	5	2678	0.17	1	533	534	46	918	0	2
P43.2	KRAS	G12D	2.45	5	20	70	8	2	1064	0.20	1	608	609	30	608	0	1
P43.2	KRAS	G12V	3.54	5	20	70	8	0	1029	0.00	0	362	362	29	588	0	0
P43.3	KRAS	G12D	2.86	5	20	70	8	0	1302	0.00	0	488	488	37	744	0	0
P43.3	KRAS	G12V	4.26	5	20	70	8	0	1061	0.00	0	347	347	30	606	0	0
P43.4	KRAS	G12D	5.50	5	20	70	8	28	2093	1.34	10	723	733	60	1196	1	16
P46.1	KRAS	G12D	2.54	2	20	70	8	6	1820	0.34	1	985	986	21	416	0	1
P46.1	KRAS	G12V	2.54	2	20	70	8	0	2774	0.00	0	496	496	32	634	0	0

Table 2-5. The list of somatic mutations detected in plasma samples by biopsy-free manner

Sample ID	Chr	Position	Ref	Alt	Function	Gene	Exonic function	AA Change	Ref#	Read Count	Total read	Allele frequency	p-value	Total average depth
P2.1	chr6	117641072	T	C	exonic	ROS1	MISSENSE	p.I1967V	NM_002944	1252	3246	38.57	0E+00	2531
P5.1	chr6	117609741	G	A	exonic	ROS1	TRUNC	p.Q2320X	NM_002944	20	2696	0.74	4E-21	2556
P5.1	chr8	38277157	G	A	exonic	FGFR1	MISSENSE	p.S385L	NM_001174064	20	1494	1.34	3E-26	2816
P7.1	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	29	2974	0.98	5E-35	2689
P10.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	30	1367	2.19	4E-46	817
P10.1	chr17	7577498	C	T	splicing	TP53	TRUNC	c.782+1G>A	NM_001126113	48	1167	4.11	2E-88	843
P21.1	chr17	7578445	A	C	exonic	TP53	MISSENSE	p.I162S	NM_001126113	41	2410	1.70	3E-60	1824
P23.1	chr7	148512600	T	C	exonic	EZH2	MISSENSE	p.K515R	NM_004456	19	2243	0.85	4E-21	1887
P28.1	chr10	89720875	G	T	exonic	PTEN	MISSENSE	p.K342N	NM_000314	39	998	3.91	1E-70	856
P29.1	chr7	55218992	A	G	exonic	EGFR	MISSENSE	p.K189E	NM_005228	28	2268	1.23	1E-36	2554
P36.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	18	2384	0.76	5E-19	1807
P36.1	chr17	7578190	T	C	exonic	TP53	MISSENSE	p.Y220C	NM_001126113	32	2724	1.17	2E-41	2018
P32.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	44	2586	1.70	1E-64	2045
P32.1	chr17	7578271	T	C	exonic	TP53	MISSENSE	p.H193R	NM_001126113	31	2441	1.27	3E-41	1906
P31.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	93	1347	6.90	4E-195	2393
P31.1	chr17	7579349	A	C	exonic	TP53	MISSENSE	p.F113C	NM_001126113	33	880	3.75	9E-59	1499
P37.1	chr9	21971101	G	T	exonic	CDKN2A	TRUNC	p.C100X	NM_058195	22	1246	1.77	1E-31	1355
P37.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	90	1956	4.60	4E-173	2045

P37.1	chr17	7578457	C	A	exonic	TP53	MISSENSE	p.R158L	NM_001126113	59	1425	4.14	3E-110	1812
P43.1	chr1	27102188	A	G	exonic	ARID1A	MISSENSE	p.N1705S	NM_006015	43	1517	2.83	4E-72	2006
P43.1	chr3	178927410	A	G	exonic	PIK3CA	MISSENSE	p.I391M	NM_006218	26	1619	1.61	3E-36	2063
P43.1	chr11	108106443	T	A	exonic	ATM	MISSENSE	p.D126E	NM_000051	29	1171	2.48	4E-46	1692
P43.1	chr11	108121733	G	A	exonic	ATM	MISSENSE	p.G514D	NM_000051	19	1620	1.17	3E-23	1930
P43.1	chr11	108143456	C	G	exonic	ATM	MISSENSE	p.P1054R	NM_000051	22	1649	1.33	2E-28	1918
P43.1	chr11	108159732	C	T	exonic	ATM	MISSENSE	p.H1380Y	NM_000051	23	1648	1.40	2E-30	2115
P43.1	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	24	1566	1.53	9E-33	2045
P43.1	chr17	7577505	T	A	exonic	TP53	MISSENSE	p.D259V	NM_001126113	16	1105	1.45	1E-20	1447
P5.5	chr7	116409781	C	T	exonic	MET	MISSENSE	p.S889F	NM_000245	62	3057	2.03	2E-96	2117
P5.5	chr13	48919281	C	G	exonic	RB1	TRUNC	p.S149X	NM_000321	56	3624	1.55	2E-80	2311
P5.5	chr13	48936983	C	T	exonic	RB1	TRUNC	p.R251X	NM_000321	105	3357	3.13	2E-184	2524
P5.5	chr17	7579699	C	T	splicing	TP53	TRUNC	exon4:c.96+1G>A	NM_001126113	69	2246	3.07	4E-120	1697
P5.6	chr7	116409781	C	T	exonic	MET	MISSENSE	p.S889F	NM_000245	275	2781	9.89	0E+00	2073
P5.6	chr13	48919281	C	G	exonic	RB1	TRUNC	p.S149X	NM_000321	300	3188	9.41	0E+00	2168
P5.6	chr13	48936983	C	T	exonic	RB1	TRUNC	p.R251X	NM_000321	358	3447	10.39	0E+00	2466
P5.6	chr17	7579699	C	T	splicing	TP53	TRUNC	exon4:c.96+1G>A	NM_001126113	311	1928	16.13	0E+00	1777
P5.7	chr7	116409781	C	T	exonic	MET	MISSENSE	p.S889F	NM_000245	2057	4675	44.00	0E+00	1165
P5.7	chr13	48919281	C	G	exonic	RB1	TRUNC	p.S149X	NM_000321	2191	5092	43.03	0E+00	1073
P5.7	chr13	48936983	C	T	exonic	RB1	TRUNC	p.R251X	NM_000321	2677	5907	45.32	0E+00	1100

P5.7	chr17	7579472	G	C	exonic	TP53	MISSENSE	p.P72R	NM_001126113	34	3128	1.09	6E-43	1175
P5.7	chr17	7579699	C	T	splicing	TP53	TRUNC	exon4:c.96+1G>A	NM_001126113	2313	2696	85.79	0E+00	1345
P5.7	chrX	76855029	T	C	exonic	ATRX	MISSENSE	p.K1936R	NM_000489	33	6527	0.51	7E-31	1167
P7.2	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	40	2806	1.43	1E-55	2689
P7.3	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	34	2770	1.23	1E-44	2689
P7.5	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	45	3031	1.48	2E-63	2404
P7.6	chr6	117710794	C	T	exonic	ROS1	MISSENSE	p.R493H	NM_002944	40	3137	1.28	1E-53	2754
P11.5	chr17	7574021	C	A	exonic	TP53	TRUNC	p.E297X	NM_001276761	28	1742	1.61	1E-39	2176
P21.2	chr9	133759935	G	T	exonic	ABL1	MISSENSE	p.G772V	NM_007313	17	1891	0.90	4E-19	1616
P21.2	chr10	43610119	G	A	exonic	RET	MISSENSE	p.G691S	NM_020975	29	1141	2.54	6E-47	1595
P21.2	chr12	46246206	G	T	exonic	ARID2	MISSENSE	p.A1434S	NM_152641	28	2165	1.29	4E-37	2150
P21.2	chr16	68857389	A	G	exonic	CDH1	MISSENSE	p.K675R	NM_004360	38	1943	1.96	8E-58	1872
P21.2	chrX	76938923	G	C	exonic	ATRX	MISSENSE	p.P609A	NM_000489	30	1065	2.82	5E-50	1530
P23.6	chr6	117642495	C	T	exonic	ROS1	MISSENSE	p.E1902K	NM_002944	19	2065	0.92	1E-21	1072
P23.6	chrX	76938208	A	G	exonic	ATRX	MISSENSE	p.F847S	NM_000489	24	3887	0.62	9E-24	741
P27.3	chr1	27102188	A	G	exonic	ARID1A	MISSENSE	p.N1705S	NM_006015	35	3306	1.06	5E-44	2077
P27.3	chr3	178927410	A	G	exonic	PIK3CA	MISSENSE	p.I391M	NM_006218	19	2986	0.64	6E-19	1831
P27.3	chr11	108106443	T	A	exonic	ATM	MISSENSE	p.D126E	NM_000051	23	2498	0.92	6E-27	1380
P27.4	chr2	29917793	C	T	exonic	ALK	MISSENSE	p.R292H	NM_004304	21	3837	0.55	9E-20	2255
P28.2	chr9	98224138	C	A	exonic	PTCH1	MISSENSE	p.Q835H	NM_001083602	383	2835	13.51	0E+00	1588

P28.2	chr10	89720875	G	T	exonic	PTEN	MISSENSE	p.K342N	NM_000314	34	4033	0.84	3E-39	856
P31.2	chrX	76938208	A	G	exonic	ATRX	MISSENSE	p.F847S	NM_000489	18	540	3.33	4E-30	2194
P36.2	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	24	2378	1.01	5E-29	1807
P36.3	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	18	1646	1.09	9E-22	1807
P36.3	chr17	7578190	T	C	exonic	TP53	MISSENSE	p.Y220C	NM_001126113	33	1962	1.68	1E-47	2018
P36.4	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	31	1804	1.72	6E-45	2365
P36.4	chr17	7578190	T	C	exonic	TP53	MISSENSE	p.Y220C	NM_001126113	47	1919	2.45	2E-76	2498
P37.2	chr9	21971101	G	T	exonic	CDKN2A	TRUNC	p.C100X	NM_058195	17	1472	1.15	5E-21	1355
P37.2	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	40	2275	1.76	3E-59	2045
P37.2	chr17	7578457	C	A	exonic	TP53	MISSENSE	p.R158L	NM_001126113	27	1620	1.67	9E-39	1812
P42.5	chr7	100417264	A	C	exonic	EPHB4	MISSENSE	p.F404L	NM_004444	71	2665	2.66	5E-119	1010
P42.5	chr11	108201015	G	A	exonic	ATM	MISSENSE	p.R2461H	NM_000051	20	1504	1.33	5E-26	1082
P42.5	chr18	48591918	C	T	exonic	SMAD4	MISSENSE	p.R361C	NM_005359	145	1413	10.26	0E+00	1160
P42.5	chr20	57415495	G	A	exonic	GNAS	MISSENSE	p.E112K	NM_016592	17	1433	1.19	6E-21	2324
P43.4	chr2	29416520	A	G	exonic	ALK	MISSENSE	p.M1478T	NM_004304	42	3499	1.20	3E-55	1126
P43.5	chr12	25398284	C	T	exonic	KRAS	MISSENSE	p.G12D	NM_033360	37	2279	1.62	3E-53	1263

GENERAL DISCUSSION

Various types of optimized targeted deep sequencing had been reported recently. All of the methods aim for the early cancer detection. TRACERx (69) performed a large number of patients to collect the information of the lung cancer for early cancer detection. Despite their study discovered the enormous amount of information about intra-tumor heterogeneity, the limitations of depth of sequencing, bioinformatics pipeline, and the cost of profiling had mentioned. Advanced versions of CAPP-seq (56) had clearly increased the specificity by replacing the barcoding adapter. However, it needs more stabilization by the depth of coverage. Finally, the recent study has increased the depth of coverage over 30,000x to find the early cancer detection (70). All those efforts of advancing technology turned to face another common challenging factor: the biological noises. The relationship between the tumor cells and ctDNA must be highlighted to achieve the ultimate goal of liquid biopsy with ctDNA analysis. Perhaps the investigation of the extravascular or intravascular mechanism of tumor cell may help to explain how the cells have escaped, accumulated, and released the cfDNA into the blood vessels. In summary, the characterization of the background noise of sequencing technology and biology had elucidated in this study and finalized to discriminate the ctDNA for clinical application.

REFERENCES

1. Park G, Park JK, Shin SH, Jeon HJ, Kim NKD, Kim YJ, et al. Characterization of background noise in capture-based targeted sequencing data. *Genome biology*. 2017;18(1):136.
2. Chung J, Son DS, Jeon HJ, Kim KM, Park G, Ryu GH, et al. The minimal amount of starting DNA for Agilent's hybrid capture-based targeted massively parallel sequencing. *Scientific reports*. 2016;6:26732.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646-74.
4. Gupta GP, Massague J. Cancer metastasis: building a framework. *Cell*. 2006;127(4):679-95.
5. Shen H, Laird PW. Interplay between the cancer genome and epigenome. *Cell*. 2013;153(1):38-55.
6. Neel DS, Bivona TG. Resistance is futile: overcoming resistance to targeted therapies in lung adenocarcinoma. *NPJ Precis Oncol*. 2017;1.
7. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;168(4):613-28.
8. Friedrich MJ. Going With the Flow: The Promise and Challenge of Liquid Biopsies. *Jama*. 2017;318(12):1095-7.
9. Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*. 2017;355(6331):1330-4.
10. Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature reviews Cancer*. 2017;17(4):223-38.
11. Takai E, Yachida S. Circulating tumor DNA as a liquid biopsy target for detection of pancreatic cancer. *World journal of gastroenterology*. 2016;22(38):8480-8.
12. Husain H, Velculescu VE. Cancer DNA in the Circulation: The Liquid Biopsy. *Jama*. 2017;318(13):1272-4.
13. Siravegna G, Marsoni S, Siena S, Bardelli A. Integrating liquid biopsies into the management of cancer. *Nature reviews Clinical oncology*. 2017;14(9):531-48.
14. Mandel P, Metais P. [Not Available]. *Comptes rendus des seances de la Societe de biologie et de ses filiales*. 1948;142(3-4):241-3.
15. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*. 2016;164(1-2):57-68.
16. Jahr S, Hentze H, Englisch S, Hardt D, Fackelmayer FO, Hesch RD, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer research*. 2001;61(4):1659-65.
17. De Vlaminck I, Valantine HA, Snyder TM, Strehl C, Cohen G, Luikart H, et al. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Science translational medicine*. 2014;6(241):241ra77.
18. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, et al. Presence of fetal DNA in maternal plasma and serum. *Lancet*. 1997;350(9076):485-7.
19. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer research*. 1977;37(3):646-50.
20. Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology*. 1989;46(5):318-22.
21. Sidransky D, Von Eschenbach A, Tsai YC, Jones P, Summerhayes I,

- Marshall F, et al. Identification of p53 gene mutations in bladder cancers and urine samples. *Science*. 1991;252(5006):706-9.
22. Sidransky D, Tokino T, Hamilton SR, Kinzler KW, Levin B, Frost P, et al. Identification of ras oncogene mutations in the stool of patients with curable colorectal tumors. *Science*. 1992;256(5053):102-5.
23. Li M, Diehl F, Dressman D, Vogelstein B, Kinzler KW. BEAMing up for detection and quantification of rare sequence variants. *Nature methods*. 2006;3(2):95-7.
24. Remon J, Caramella C, Jovelet C, Lacroix L, Lawson A, Smalley S, et al. Osimertinib benefit in EGFR-mutant NSCLC patients with T790M-mutation detected by circulating tumour DNA. *Annals of oncology : official journal of the European Society for Medical Oncology*. 2017;28(4):784-90.
25. Thress KS, Brant R, Carr TH, Dearden S, Jenkins S, Brown H, et al. EGFR mutation detection in ctDNA from NSCLC patient plasma: A cross-platform comparison of leading technologies to support the clinical development of AZD9291. *Lung cancer*. 2015;90(3):509-15.
26. Gundry M, Vijg J. Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. *Mutation research*. 2012;729(1-2):1-15.
27. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2012;109(36):14508-13.
28. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(23):9530-5.
29. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform*. 2016;17(1):154-79.
30. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome biology*. 2011;12(11):R112.
31. Cline J, Braman JC, Hogrefe HH. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res*. 1996;24(18):3546-51.
32. Kuchta RD, Benkovic P, Benkovic SJ. Kinetic mechanism whereby DNA polymerase I (Klenow) replicates DNA with high fidelity. *Biochemistry*. 1988;27(18):6716-25.
33. Chen G, Mosier S, Gocke CD, Lin MT, Eshleman JR. Cytosine deamination is a major cause of baseline noise in next-generation sequencing. *Mol Diagn Ther*. 2014;18(5):587-93.
34. Wong SQ, Li J, Salemi R, Sheppard KE, Do H, Tothill RW, et al. Targeted-capture massively-parallel sequencing enables robust detection of clinically informative mutations from formalin-fixed tumours. *Scientific reports*. 2013;3:3494.
35. Do H, Wong SQ, Li J, Dobrovic A. Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. *Clin Chem*. 2013;59(9):1376-83.
36. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415-21.
37. Norton SE, Lechner JM, Williams T, Fernando MR. A stabilizing reagent

- prevents cell-free DNA contamination by cellular DNA in plasma during blood sample storage and shipping as determined by digital PCR. *Clin Biochem*. 2013;46(15):1561-5.
38. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-95.
39. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-303.
41. Medina Diaz I, Nocon A, Mehnert DH, Fredebohm J, Diehl F, Holtrup F. Performance of Streck cfDNA Blood Collection Tubes for Liquid Biopsy Testing. *PLoS one*. 2016;11(11):e0166354.
42. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-76.
43. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013;31(3):213-9.
44. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nature medicine*. 2014;20(5):548-54.
45. Chabon JJ, Simmons AD, Lovejoy AF, Esfahani MS, Newman AM, Haringsma HJ, et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. *Nature communications*. 2016;7:11815.
46. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res*. 2015;43(6):e37.
47. Wang XV, Blades N, Ding J, Sultana R, Parmigiani G. Estimation of sequencing error rates in short reads. *BMC bioinformatics*. 2012;13:185.
48. Takai E, Totoki Y, Nakamura H, Morizane C, Nara S, Hama N, et al. Clinical utility of circulating tumor DNA for molecular assessment in pancreatic cancer. *Scientific reports*. 2015;5:18425.
49. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, 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*. 2013;41(6):e67.
50. Kino K, Sugiyama H. UVR-induced G-C to C-G transversions from oxidative DNA damage. *Mutation research*. 2005;571(1-2):33-42.
51. Kino K, Sugiyama H. Possible cause of G-C-->C-G transversion mutation by guanine oxidation product, imidazolone. *Chemistry & biology*. 2001;8(4):369-78.
52. Chen L, Liu P, Evans TC, Jr., Ettwiller LM. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*. 2017;355(6326):752-6.
53. Swenberg JA, Lu K, Moeller BC, Gao L, Upton PB, Nakamura J, et al. Endogenous versus exogenous DNA adducts: their role in carcinogenesis, epidemiology, and risk assessment. *Toxicological sciences : an official journal of the Society of Toxicology*. 2011;120 Suppl 1:S130-45.
54. Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. *Nature*

biotechnology. 2011;29(10):908-14.

55. Butler TM, Johnson-Camacho K, Peto M, Wang NJ, Macey TA, Korkola JE, et al. Exome Sequencing of Cell-Free DNA from Metastatic Cancer Patients Identifies Clinically Actionable Mutations Distinct from Primary Disease. *PloS one*. 2015;10(8):e0136407.
56. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nature biotechnology*. 2016;34(5):547-55.
57. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*. 2010;2(1):a001008.
58. Lennon AM, Goggins M. Diagnostic and Therapeutic Response Markers. *Pancreatic Cancer*. New York, NY: Springer New York; 2010. p. 675-701.
59. Makohon-Moore A, Iacobuzio-Donahue CA. Pancreatic cancer biology and genetics from an evolutionary perspective. *Nature reviews Cancer*. 2016;16(9):553-65.
60. Dabritz J, Preston R, Hanfler J, Oettle H. Follow-up study of K-ras mutations in the plasma of patients with pancreatic cancer: correlation with clinical features and carbohydrate antigen 19-9. *Pancreas*. 2009;38(5):534-41.
61. Brychta N, Krahn T, von Ahsen O. Detection of KRAS Mutations in Circulating Tumor DNA by Digital PCR in Early Stages of Pancreatic Cancer. *Clin Chem*. 2016;62(11):1482-91.
62. Ako S, Nouse K, Kinugasa H, Dohi C, Matushita H, Mizukawa S, et al. Utility of serum DNA as a marker for KRAS mutations in pancreatic cancer tissue. *Pancreatology : official journal of the International Association of Pancreatology*. 2017;17(2):285-90.
63. Pietrasz D, Pecuchet N, Garlan F, Didelot A, Dubreuil O, Doat S, et al. Plasma Circulating Tumor DNA in Pancreatic Cancer Patients Is a Prognostic Marker. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2017;23(1):116-23.
64. Earl J, Garcia-Nieto S, Martinez-Avila JC, Montans J, Sanjuanbenito A, Rodriguez-Garrote M, et al. Circulating tumor cells (Ctc) and kras mutant circulating free Dna (cfdna) detection in peripheral blood as biomarkers in patients diagnosed with exocrine pancreatic cancer. *BMC cancer*. 2015;15:797.
65. van Ginkel JH, Huibers MMH, van Es RJJ, de Bree R, Willems SM. Droplet digital PCR for detection and quantification of circulating tumor DNA in plasma of head and neck cancer patients. *BMC cancer*. 2017;17(1):428.
66. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. *Statistics in medicine*. 2007;26(10):2170-83.
67. Yachida S, Iacobuzio-Donahue CA. Evolution and dynamics of pancreatic cancer progression. *Oncogene*. 2013;32(45):5253-60.
68. Cohen JD, Javed AA, Thoburn C, Wong F, Tie J, Gibbs P, et al. Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. *Proceedings of the National Academy of Sciences of the United States of America*. 2017;114(38):10202-7.
69. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *The New England journal of medicine*. 2017;376(22):2109-21.
70. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Science translational medicine*. 2017;9(403).

초 록

서론: 실시간으로 종양 게놈의 역학을 측정할 수 있는 방법으로 순환 종양 (ct)DNA 와 차세대 시퀀싱 (NGS) 기반 방법 구현이 제시 되었다. 그러나 혈액 속에 존재하는 정상 세포 유리 (cf)DNA의 빈도는 ctDNA의 비율보다 높아 낮은 종양 변이의 대립 유전자와 기술오차 비율 수준이 동의 선상으로 측정 될 수 있어 이의 걸 맞는 차별화 및 실용적 가이드라인과 분석 방법이 필요하다. 제 1장*에서는 고유 DNA 분자 회수율의 중요성을 강조하고 시퀀싱 과정에서 발생하는 오류들의 특성을 분석하였다. 제 2장에서는 앞선 방법 조합하여 암 환자 샘플에 적용하여 ctDNA시퀀싱의 유용성 평가 및 종양 게놈 모니터링을 실시 하였다.

방법: 적응 시료량에서 DNA분자 회수율을 극대화 하기 위해, 시퀀싱의 초기 단계인 ligation 구성 요소의 온도, 시간 및 어댑터 농도의 조정 및 최적화 하여 구현되었다. 오류의 규명은 cfDNA의 특징 중의 하나인 자연적으로 절단된 장점을 이용하여 acoustically 절단된 germline DNA와 비교 분석되었다. 암 환자 샘플들 에서 검출 된 ctDNA의 유용성은 치료 반응 및 영상 변화에 따라 평가되었다.

결과: 시퀀싱 초기 단계를 수정한 ligation 조건을 적은 시료에 적용 하였을 때 DNA 분자 회수율은 표준 조건보다 20% 높은 비율을 나타내었다. 수동으로 전단 된 gDNA와 자연적으로 단편화 된 cfDNA의 특성을 비교한 결과 gDNA에서 C : G>

A : T의 64 %와 C : G > G : C의 39 %의 substitution class 비율이 증가됨을 규명할 수 있었으며 이는 전단 과정에서 일어날 수 있는 oxo-guanine과 연관이 있다는 것을 규명할 수 있었다. 순화된 전단 조건을 통해 관련 오류률은 평균 40% 정도 제거해 낼 수 있었다. 또한, DNA 단편의 말단 부근을 분석한 결과 A > G 및 A > T 우선적으로 단편화 되는 것을 알 수 있었다. 향상된 NGS 방법은 암환자 cfDNA 샘플에 적용하여 평가하였을 때 100 % 민감도와 97.1 % 특이도를 갖춘 진단적 유용성을 확립할 수 있었다. CtDNA의 반응도는 치료 반응과 높은 상관 관계가 있었을 뿐 아니라, 표준 단백질 바이오 마커와 이미징 변화 보다 2 개월 앞선 평균 반응도를 나타내었다. 마지막으로, ctDNA 분석은 종양 생검에서 알 수 없었던 종양 내 이질성 또한 검출해 낼 수 있었다.

결론: 전반적으로 cfDNA의 독특한 특성분석을 통해 기술적인 오류의 근본 원인을 강조할 수 있었을 뿐만 아닌 NGS 기반 기술을 사용하여 암의 조기 발견 기회를 입증할 수 있는 연구였다. 궁극적으로, cfDNA와 NGS 분석의 조합 접근법은 암 연구에서 충족되지 않은 요구를 해결할 것이라 믿는다.

* 본 내용은 *Scientific Reports*와 *Genome Biology* 학술지 (참고문헌 포맷)에 출판 완료된 내용임

주요어: 암 유전체학, 액체 생검, 순환하는 종양 DNA, 무 세포 DNA, 차세대 시퀀싱, 백그라운드 오류

학 번: 2012-21792