이학석사 학위논문

차세대 염기서열 분석을 이용한 대장암 환자의 혈액내 종양DNA 검출 및 예후 예측에 관한 연구

Use of novel computational filtering to reduce background next-generation sequencing errors in circulating tumour DNA of metastatic colorectal cancer patients

2018년 2월

서울대학교 융합과학기술대학원

분자의학 및 바이오제약학과

강 준 규

차세대 염기서열 분석을 이용한 대장암 환자의 혈액내 종양DNA 검출 및 예후 예측에 관한 연구

지도교수 김 태 유

이 논문을 이학석사 학위논문으로 제출함

2017년 12월

서울대학교 융합과학기술대학원

분자의학 및 바이오제약학과

강 준 규

강 준 규 의 이학석사 학위논문을 인준함

2017년 12월

위 원 장 ______________ (인)

부 위 원 장______________ (인)

위 원______________ (인)

Use of novel computational filtering to reduce background next-generation sequencing errors in circulating tumour DNA of metastatic colorectal cancer patients

by

Jun-Kyu Kang

(Directed by Tae-You Kim, M.D., Ph.D.)

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Molecular and Biopharmaceutical Sciences at the Seoul National University, Seoul, Korea

December 2017

Approved by thesis committee :

Professor______________Chairperson

Professor______________Vice Chairperson

Professor______________

ABSTRACT

Use of novel computational filtering to reduce background next-generation sequencing errors in circulating tumour DNA of metastatic colorectal cancer patients

Jun-Kyu Kang
Department of Molecular Medicine
and Biopharmaceutical Sciences
World Class University Graduate School of
Convergence Science and Technology
Seoul National University

Next-generation sequencing (NGS) technology is emerging as a major technique for genotyping circulating cell-free DNA (cfDNA) and for patient monitoring. However, Results of NGS is subject to numerous errors. In this study, we isolated circulating cfDNA and genomic DNA from 39 available tumours from 54 patients with advanced colorectal cancer (CRC). Deep targeted sequencing was performed for a panel of 10 genes that are recurrently mutated in CRC. To reduce sequencing error, we devised a ‘de-noising’ procedure and calculated the concordance of somatic variants between cfDNA and tumour tissue sequencing data. The sensitivity, specificity, and accuracy for somatic alterations in the 10 genes were increased from 84.5%, 74.6%, and 76.9% to 87.3%, 92.0%, and 91.1%, respectively, after de-noising. This approach improved the detection of somatic alterations in advanced CRC cfDNA. We could selectively detect clinically important somatic alterations for variant allele frequencies of 0.27%–79.42%. Patients with high cfDNA concentrations had more detectable somatic mutant fragments and larger liver metastatic lesions than patients with lower concentrations. These results demonstrate the suitability of de-noised deep targeted sequencing for cfDNA genotyping, and provide insights into strategies for monitoring metastatic lesions in patients with advanced CRC.

Keywords : colorectal cancer (CRC), cfDNA, NGS, prognosis, de-noising

Student Number： 2016−26001

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Tumour genotyping is useful for characterizing tumour lesions, but genotyping of tissue biopsies is associated with disadvantages such as biased results and difficulty in obtaining longitudinal samples, as well as difficulties in accessing the tumour in some cases. There is thus a need for a more comprehensive biomarker that can represent the overall tumour state.

Analysis of circulating cell-free DNA (cfDNA) in the blood represents a feasible method for both the early detection of cancer and for monitoring the status of cancer patients. Circulating cfDNA is present in the plasma, serum, and urine, and comprises 140–170 base pair (bp) fragments that become separated from the tumour as a result of apoptosis, necrosis, or secretion [1]. cfDNA may include a variety of cancer-derived mutated genes [2], which can be used clinically for cancer diagnosis and patient monitoring using simple liquid biopsies.

Among the many available sequencing methods, next-generation sequencing (NGS), unlike biased molecular tests, allows comprehensive genome analysis. Its high sensitivity of detection and multiplexed interpretation make NGS a suitable method for non-invasive genotyping of cancers using cfDNA [3]. Targeted exome sequencing is a low-cost NGS technique that can detect gene mutations specific to the target treatment. For example, acquired resistance to cetuximab in colorectal cancer (CRC) patients was monitored using liquid biopsy with NGS technology [4]. However, NGS has been associated with problems such as reproducibility of sequencing results, and sequencing and validation errors [5] [6]. Typical sequencing errors involve 8-oxoG, generated during sample preparation for NGS, and cytosine deamination, which are major causes of baseline noise in NGS, as well as reading errors due to the presence of homo-polymer regions in the genome sequence [5] [7] [8].

We aimed to overcome these issues and devised a clinically feasible method for monitoring circulating tumour-driven DNA in non-invasive liquid biopsies. We focused on improving the quality of the NGS results of deep targeted exome sequencing, and setting robust methods for monitoring metastatic CRC (mCRC) patients.

MATERIALS AND METHODS

Patient cohort and ethics statement

Fifty–four patients with phase lll–lV mCRC were recruited for cfDNA genotyping. Clinical information including gender, age, and pathological information was collected. All patients provided written informed consent prior to any study–specific procedures, including liquid biopsy, tissue biopsy, and genetic testing. The study protocol was reviewed and approved by the Institutional Review Board of Seoul National University Hospital and conducted in accordance with the recommendations of the Declaration of Helsinki for biomedical research involving human subjects.

Tumour tissue samples

Among 39 tumour specimens, 24 samples were formalin−fixed, paraffin−embedded (FFPE) tissues and 15 were fresh frozen tissues. Genomic DNA was isolated from each sample using a Qiagen DNA FFPE Tissue Kit (Qiagen, Hilden, Germany) for FFPE samples and a QIAamp DNA Mini Kit (Qiagen) for fresh frozen tissues. After isolation, the concentrations and purities of genomic DNA were measured using a spectrophotometer (ND1000, Nanodrop Technologies, ThermoFisher Scientific, MA, USA).

Blood samples and cell−free DNA isolation and quantification

Whole blood (4–6 ml) was collected into EDTA tubes during routine phlebotomy from patients who volunteered to donate blood samples for research purposes. Blood samples were centrifuged with Ficoll solution at 1,500 × g for 15 min. Plasma was separated by centrifugation at 16,000 × g for 10 min to remove cell debris, and 1−ml aliquots were placed in Eppendorf tubes and stored at −80C before extraction. This protocol was carried out within 20 min of collection to prevent degradation of cfDNA. cfDNA was isolated from aliquots (1 ml) of plasma using a QIAamp circulating nucleic acid kit (Qiagen) with the QIAvac 24 Plus vacuum manifold, following the manufacturer's instructions, and quantified using a 2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA). PBMC were separated following this protocol. Genomic DNA was isolated from PBMC using a QIAamp DNA Mini Kit (Qiagen).

Deep target exome sequencing

A DNA NGS library was constructed using a Celemics NGS DNA library prep kit. For cfDNA, a random barcode was introduced in P7 index sites to recover more reads, which were assumed to be PCR duplicates based on a previous analysis method. Solution-based target enrichment was performed at Celemics, Inc. using a custom target capture panel. Captured DNA libraries were sequenced using an Illumina HiSeq 2500 platform (Illumina, San Diego, CA, USA) 2 × 150 bp paired-end mode.

Illumina adaptor sequences were removed from raw fastq data using Trimmomatic (v0.33). To reduce NGS errors from low-quality bases, the read length was trimmed from 151 to 101 bp, and reads were filtered according to this condition (mean QS < 20). For cfDNA data, PCR duplicates were removed by comparing random barcodes at the P7 index sites and read contents. Filtered fastq files were aligned to the hg19 reference genome using Burrows–Wheeler Aligner (v0.7.10) “aln” and “sampe” algorithm. Aligned SAM files were converted into BAM files and sorted using SAMtools (v1.1). For PBMC and tumour tissue data, PCR duplicates were removed with Picard tools (v1.115) “MarkDuplicates” algorithm. Local realignment around known indel sites and base quality score recalibration were performed with GATK (v2.3−9). After generating pileup files with SAMtools mpileup, variants were called using Varscan2 (v2.4.0) “mpileup2snp” and “mpileup2indel”. For variant calls, QS < 30 bases were ignored and a strand bias filter was applied. Called variants were annotated with ANNOVAR (v2013−08−23) and other in−house programs.

De−noising

Variants from the cfDNA and PBMC datasets that met one of the following conditions were selected as candidate noise variants: i) variants supported with only one read, and ii) variants with < 2% VAF. Among the candidate noise variants, single nucleotide polymorphisms reported more than once or indels reported more than twice in the COSMIC databases were eliminated. A total of 27,528 variants were selected as a blacklist for the de−noising procedure.

Concordance analysis

After variant calling, mutation types detected in cfDNA, PBMC, and tumour tissue were identified as single nucleotide variants and insertions or deletions (Indels). For the 10 genes, concordance was calculated based on deep targeted exome sequencing data of tumour tissue. All positions indicating mutations were screened, except for synonymous mutations. Mutations in the same positions in PBMC as in cfDNA or tumour tissue were considered to be germline mutations, while mutations not detected in PBMC were considered to be somatic mutations. Finally, mutations detected in cfDNA were compared with mutations detected in tumour tissue with VAF > 5%. We then defined these mutations as concordant. For hotspot mutations, we screened all positions indicating clinically significant somatic mutations based on COSMIC [9] data using Integrative Genomics Viewer [10].

Analysis of metastatic lesions

Patients in this study had metastatic lesions in the liver, lung, peritoneum, and other organs (Table 1). Abdominal and chest CT scans were examined and the metastatic tumour burdens in the liver and lung were estimated by calculating the sum of the longest diameter of the tumour in the same section.

Statistical analysis

Receiver operating characteristic curve analysis was performed to compare results before and after de-noising, using IBM SPSS 23 (IBM Corporation; Armonk, NY). Survival differences between patient groups were estimated using log-rank tests on Kaplan–Meier curves, using GraphPad Prism version 7.00 for Windows (www.graphpad.com; GraphPad Software, La Jolla, CA, USA).

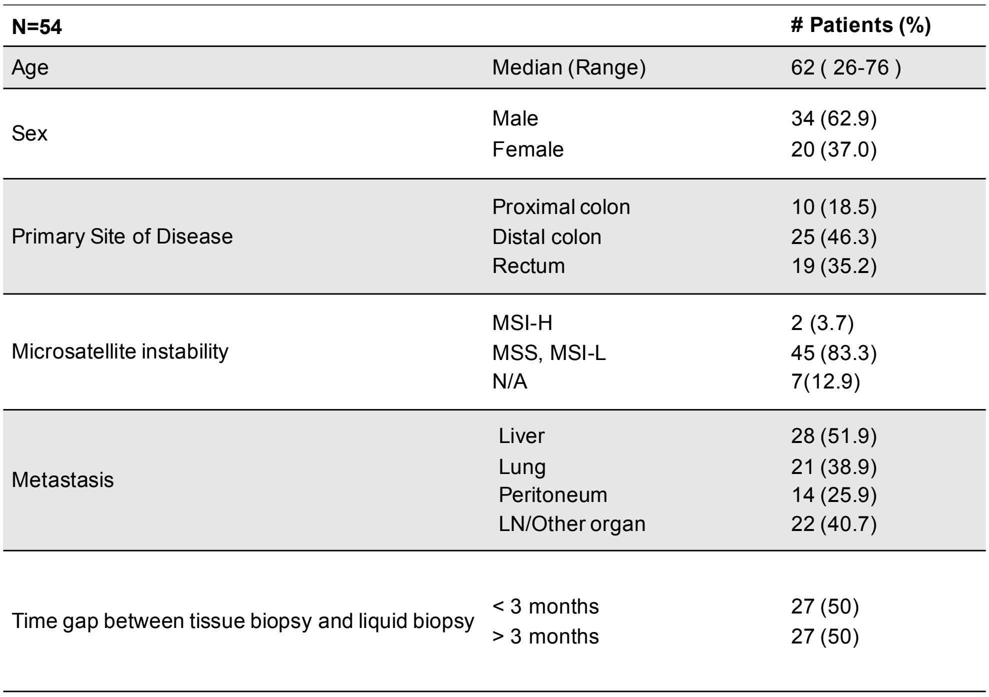
RESULT

Patient characteristics and de-noising strategy

We assessed the accuracy of tumour genotyping for cfDNA using deep targeted sequencing in a cohort of 54 patients with mCRC. The primary sites of disease were the proximal colon (n=10), distal colon (n=25), and rectal colon (n=19). Metastatic lesions were present in the liver (n=28), lung (n=21), peritoneum (n=14), and lymph nodes/other organs (n=22). Patients were divided according to the time interval between tissue biopsy and liquid biopsy within 3 months (n=27) and more than 3 months (n=27) (Table 1). We isolated DNA from samples including plasma, peripheral blood mononuclear cells (PBMC), and tumour tissues for each patient and performed deep targeted exome sequencing for a panel of 10 genes (KRAS, TP53, APC, BRAF, PIK3CA, SMAD4, ATM, ARID1A, ACVR2A, ATM), which are recurrently mutated in CRC [11] [12] [13]. Single nucleotide variants (SNVs) and insertion and deletion mutants (indels) in circulating cfDNA were detected using the pipelines as described. We reduced the error rate by 'de-noising'. We chose 10 patients with different mutations with no overlap among the 10 genes. With error-prone positions from these patients, the data was removed what we thought was an error in the filter condition (Methods and Fig. 1). The NGS data were then compared and analyzed for each sample (Fig. 2). In the screening result of the entire target region, cfDNA mutation calls were found to be decreased from 9,964 to 1,778 after de-noising (total mean of 82% in all 54 patients)

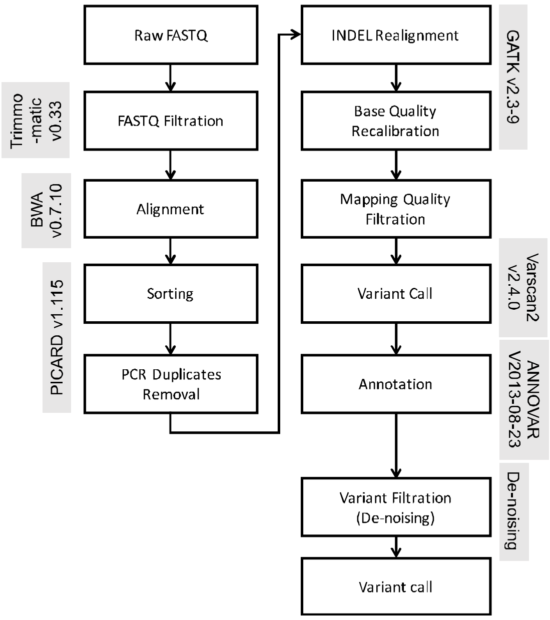
(Fig. 3A). In representative 2 cases, most error-prone alterations along targeted region were reduced after de-noising except for germline SNPs or somatic mutations which were detected in paired tumour tissue (Fig. 3B). De-noising significantly decreased the biological background rate among patients (Fig. 4). When a tumour-tissue mutation was present, the percentage of cfDNA in the plasma with at least one mutation was 97.22%. Although the variant allele frequency (VAF) was much lower when the mutant alleles were confirmed using the Integrative Genomics Viewer, the range of detected VAFs in cfDNA was 0.27%–79.42% (Table 2).

Table 1. Characteristics of phase lll–lV metastatic colorectal cancer patient cohort



N=54		**# Patients (%)**
Age	Median (Range)	62 (26-76)
Sex	Male	34 (62.9)
	Female	20 (37.0)
Primary Site of Disease	Proximal colon	10 (18.5)
	Distal colon	25 (46.3)
	Rectum	19 (35.2)
Microsatellite instability	MSI-H	2 (3.7)
	MSS, MSI-L	45 (83.3)
	N/A	7(12.9)
Metastasis	Liver	28 (51.9)
	Lung	21 (38.9)
	Peritoneum	14 (25.9)
	LN/Other organ	22 (40.7)
Time gap between tissue biopsy and liquid biopsy	< 3 months	27 (50)
	> 3 months	27 (50)

Figure 1. Pipeline for calling single nucleotide variants(SNVs) and insertion and deletion(INDEL) in circulating cfDNA with denoising.



Analysis Program	Option
Trimmomatic v0.33	trimmed from 151bp to 101bp (QS < 20)
BWA v0.7.10	alinged to hg19 reference genome (using "aln" & "sampe" algorithm)
PICARD v1.115	"MarkDuplicates" algorithm
GATK v2.3-9	Local realignment around known indel sites and base quality score recalibration(BQSR)
Varscan2 v2.4.0	bases were called (QS > 30) strand bias filter applied (> 90%) Minimum allele frequency 0.01% filter Minimum major allele depth 8 filter Minimum minor allele depth 2 filter P-value threshold 0.01
ANNOVAR v2013-08-23	Gene annotation(RefSeq) 1000 Genome allele frequency ESP6500 allele frequency COSMIC dbSNP
De-noising	Minor Allele depth ≧ 1 (alteration depth) VAF < 2% Sample overlap in same position >2 If any mutation position which have COSMIC HIT =1 for SNV or COSMIC HIT =2 for INDEL was included in 'Black list', these substitution list would be rescued for variant calling. & rescued actionable hotspots

Figure 2. Overall workflow Working flow for assessing concordance of deep targeted exome sequencing data. PBMC, peripheral blood

mononuclear cell; cfDNA, cell-free DNA; TES, Targeted exome deep sequencing.

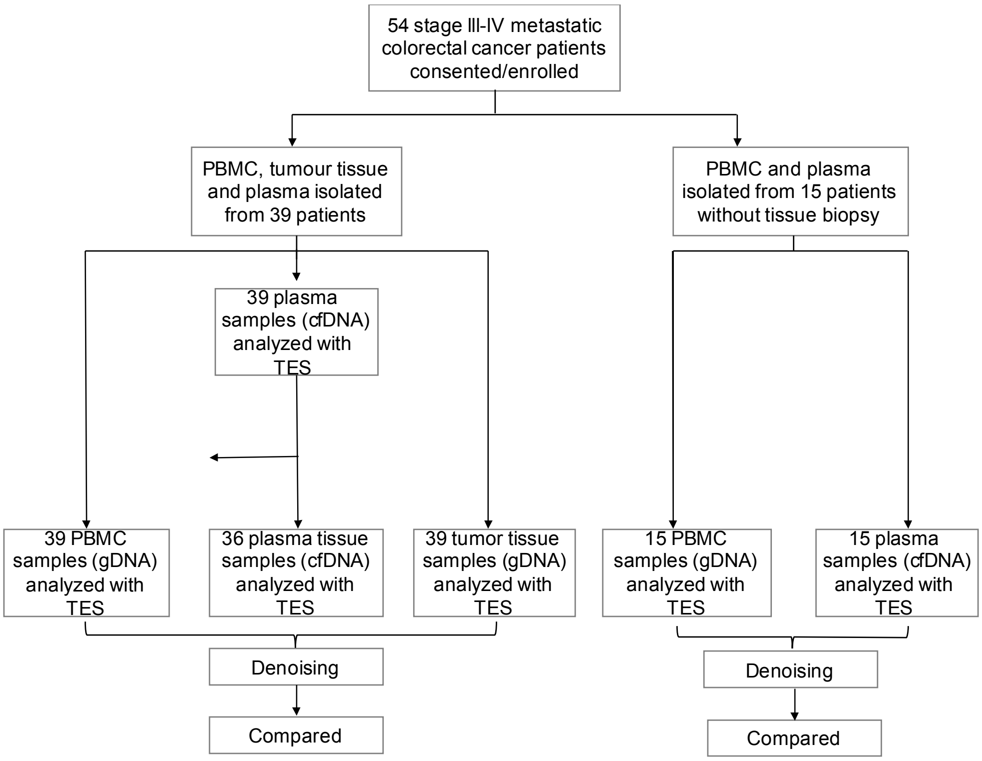
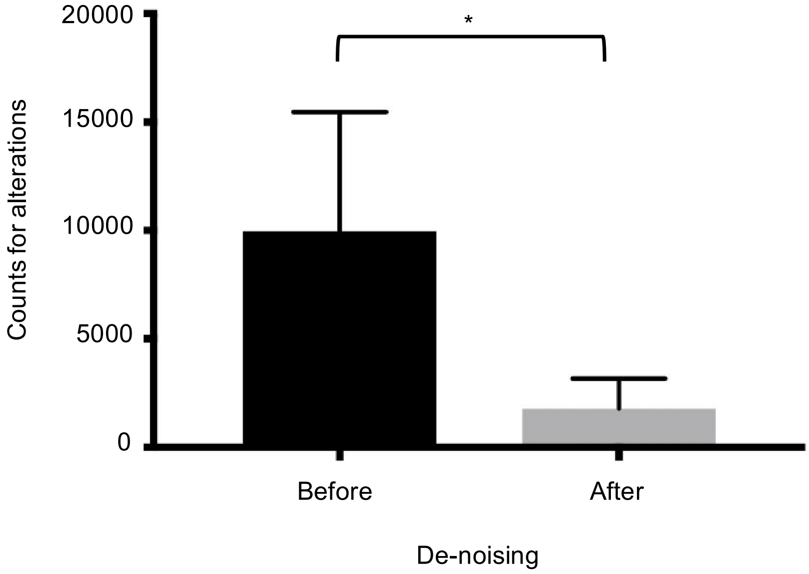

54 stage Ⅲ-Ⅳ metastatic colorectal cancer patients consented/enrolled
PBMC, tumour tissue and plasma isolated from 39 patients
PBMC and plasma isolated from 15 patients without tissue biopsy
39 plasma samples (cfDNA) analyzed with TES
39 PBMC samples (gDNA) analyzed with TES
36 plasma tissue samples (cfDNA) analyzed with TES
39 tumor tissue samples (gDNA) analyzed with TES
15 PBMC samples (gDNA) analyzed with TES
15 plasma samples (cfDNA) analyzed with TES
Denoising
Compared
Denoising
Compared

Figure 3. Development of De-noising

(A) After de−noising, biological background rate among NGS results of circulating cfDNA in mCRC patients could be reduced significantly. (*p<0.0001; unpaired t−test.)

(B) Single nucleotide variants(SNVs) were detected along targeted region in two patient's cfDNA samples (Left) All of the detected alteration was plotted. Variant allele frequency(VAF) of circulating cfDNA was from 0.019% to 100%. There were germline mutations which were also detected in paired PBMC sample. Germline mutations were marked with green plots which were detected as either 50%(heterozygous) or 100%(homozygous). SNVs were marked with red plots which were also detected in primary paired tumour tissue sample. After applying our customized de−noising strategy, most noises under 2% were corrected. Germline mutations and SNVs were remained correctly. (Right) All of mutations plotted along targeted region from patient's cfDNA sample who bear at lowest VAF mutation fragments before de−noising. After de−noising, almost germline SNPs were remained and somatic mutation detected with paired tumour tissue was remained as tumour derived cfDNA.

Fig. 3A



20000
15000
10000
5000
0
Counts for alterations
*
Before
After
De-noising

Fig. 3B

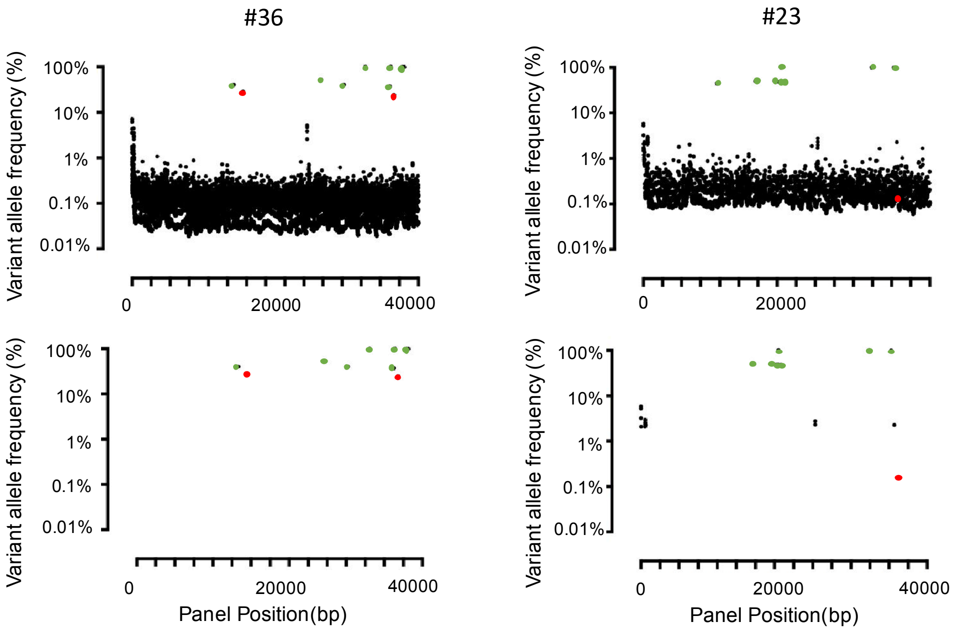

#36
#23
Variant allele frequency (%)
100%
10%
1%
0.1%
0.01%
0
20000
40000
Panel Position(bp)

Figure 4. After de−noising, biological background rate among NGS results of circulating cfDNA in mCRC patients could be reduced significantly (*p<0.0001; Mann−Whitney test)

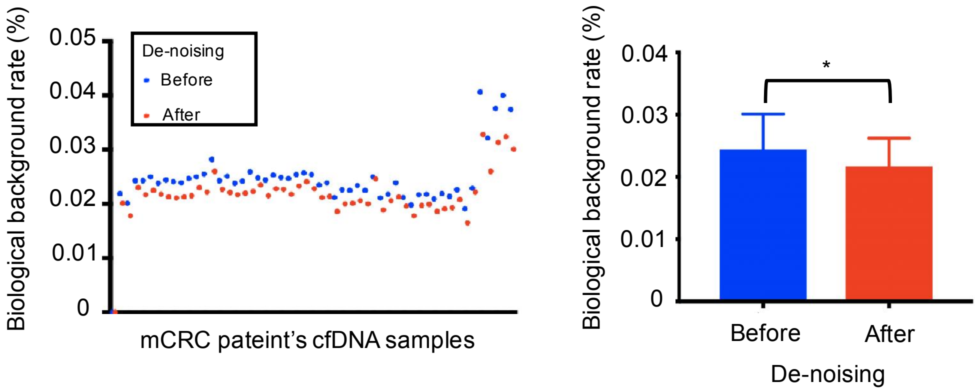


Table 2. Summary of 10 genes targeted exome sequencing

Summary of targeted deep Seqencing	
Percentage of plasma samples with at least one mutation in cell-free DNA when tissue mutation present	
	97.22%
Range of mean depth for each sample types	
cell-free DNA	**27.82~4107.52**
tumor tissue	**56.78~1460.39**
PBMC	**720.04~1536.80**
Range of detected Variant allele frequencies in cell-free DNA	
	0.27% ~ 79.42%

Concordance of de-noised NGS results between liquid biopsy and tumour tissue biopsy

The concordance of deep targeted sequencing results was estimated among samples from 36 patients for whom cfDNA and genomic DNA from PBMC and tumour tissues were available. The cfDNA deep targeted sequencing results were compared with tissue deep targeted sequencing results, which produced an area under the curve of 0.86 for the detection of identical somatic variants in cfDNA and tumour tissue. This value was increased to 0.92 after application of the de-noising method (Fig. 5A). We also calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of somatic mutation detection in cfDNA based on tumour tissue. For all 10 genes, the sensitivity, specificity, PPV, NPV, and accuracy were 84.5%, 74.6%, 50.7%, 93.9%, and 76.9%, respectively, which were increased to 87.3%, 92.0%, 72.9%, 96.7%, and 91.1% after de-noising (Fig. 5B). Notably, de-noising increased the specificity and PPV by >20%. The VAF limitation for detecting cfDNA was 0.27%, indicating good sensitivity for detecting mutant fragments. Detection of TP53, APC, and the other 7 genes, except for KRAS, were generally increased after de−noising (Fig. 5C and Fig. 6). Since KRAS mutation is one of the important genetic changes in CRC patients, it was excluded from a blacklist of de−noising. Table 3. shows how cfDNA mutations were estimated according to tumour tissue mutations for the 10 genes after de−noising.

Figure 5. Performance of noninvasive tumor genotyping using denoised deep target sequencing. (n=36) (A)After denoising, performance of deep target-sequencing for diagnosis was more accurate than before; AUC, area under the curve. (B) Accuracy for detecting targeted 10 genes mutations both cfDNA and tumor tissue was increased after denoising. (C) 3 genes (KRAS, TP53, APC) mutations which were most frequently detected in CRC based on TCGA data were well concordant. PPV, positive predictive value; NPV, negative predictive value.

Fig. 5A

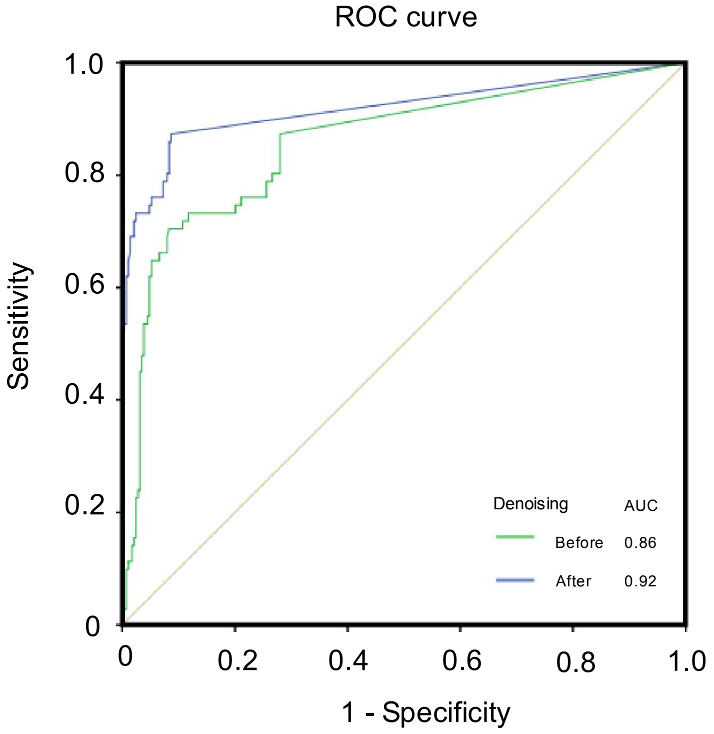
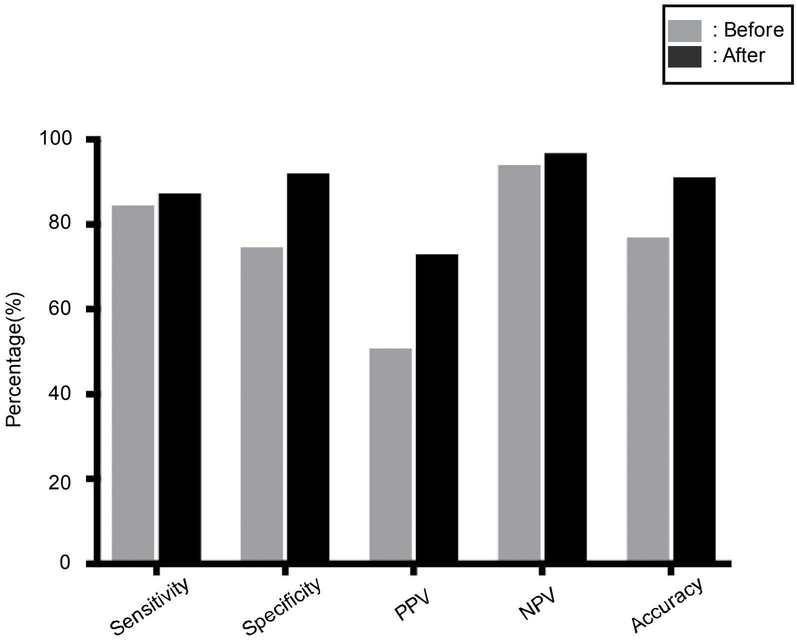

ROC curve
1.0
0.8
0.6
0.4
0.2
0
Sensitivity
0
0.2
0.4
0.6
0.8
1.0
1 - Specificity
Denoising
AUC
Before
0.86
After
0.92

Fig. 5B



: Before
: After
100
80
60
40
20
0
Percentage(%)
Sensitivity
Specificity
PPV
NPV
Accuracy

Fig. 5C

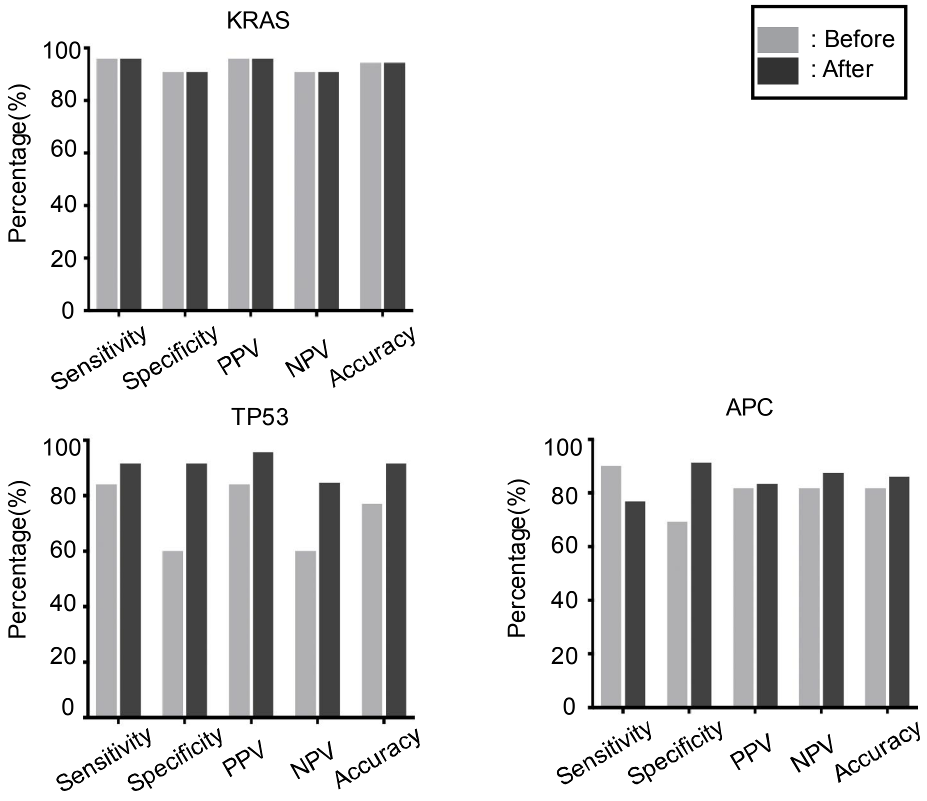

KRAS
100
80
60
40
20
0
Percentage(%)
Sensitivity
Specificity
PPV
NPV
Accuracy
: Before
: After
TP53
100
80
60
40
20
0
Percentage(%)
Sensitivity
Specificity
PPV
NPV
Accuracy
APC
100
80
60
40
20
0
Percentage(%)
Sensitivity
Specificity
PPV
NPV
Accuracy

Figure 6. Performance of noninvasive tumor genotyping using denoised deep target sequencing. The other 7 genes were calculated more accurately after denoising. PPV, positive predictive value; NPV, negative predictive value.

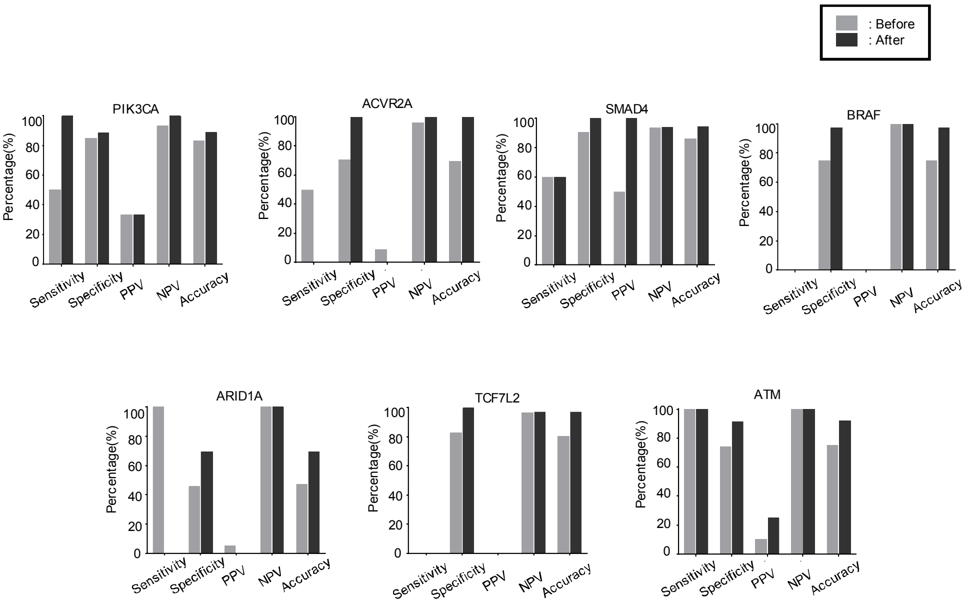


Table 3 Sensitivity, Specificity, and Diagnostic accuracy across 10 genes

N=36		Tumor-tissue genetic alteration		Sensitivity(%)	Specificity(%)	PPV(%)	NPV(%)	Accuracy(%)
cfDNA genetic alteration		MT	WT					
KRAS	MT	24	1					
	ND	1	10	96.0	90.9	96.0	90.9	94.4
APC	MT	10	2					
	ND	3	21	76.9	91.3	83.3	87.5	86.1
TP53	MT	22	1					
	ND	2	11	91.7	91.7	95.7	84.6	91.7
BRAF	MT	0	1					
	ND	0	35	-	97.2	0.0	100.0	97.2
PIK3CA	MT	2	4					
	ND	0	30	100.0	88.2	33.3	100.0	88.9
ATM	MT	1	3					
	ND	0	32	100.0	91.4	25.0	100.0	91.7
ACVR2A	MT	0	0					
	ND	0	36	-	100.0	-	100.0	100.0
ARID1A	MT	0	11					
	ND	0	25	-	69.4	0.0	100.0	69.4
SMAD4	MT	3	0					
	ND	2	31	60.0	100.0	100.0	93.9	94.4
TCF7L2	MT	0	0					
	ND	1	35	0.0	100.0	-	97.2	97.2
Total MT		62	23					
Total ND		9	266					
Total (MT + ND)		71	289	87.3	92.0	72.9	96.7	91.1

(MT : Mutation, WT : Wild type, ND : Not detected)

cfDNA mutational genotyping among patients with mCRC

Among the 10 genes, detected somatic variants in KRAS, TP53, and APC were detected in plasma in 25/36 (69.4%), 23/36 (63.9%), and 12/36 (33.3%) patients, respectively (Fig. 7). This was correlated with data from TCGA for the top three of 10 genes mutated in CRC patients. We then estimated the concordance of cfDNA somatic variants in individual patients. Before de-noising, numerous cfDNA-somatic variants detected in plasma were only counted among the patients. After de-noising, patients had 0–4 mutations in identical positions in cfDNA and tumour tissue. Thirty-two patients had at least one somatic mutation in both cfDNA and tumour tissue, three had only a cfDNA mutation or tumour tissue mutation (#5, #17, #28), and one patient had no mutation in either cfDNA or tumour tissue (#30). Somatic mutations in other genes (ARID1A, PIK3CA, ATM, and BRAF) were only detected in circulating cfDNA. There was no correlation between these results and the time interval between tissue biopsy and liquid biopsy. Twenty-eight of the 36 patients had metastatic lesions in the liver, and there was a tendency for more cfDNA somatic mutations to be detected in this group (Fig. 8). The profiles of the clinically significant somatic variants of the 10 genes detected in cfDNA suggest the existence of heterogeneity among patients with advanced CRC.

Figure 7. Comparing frequency of somatic alteration among 10 genes in plasma cfDNA and tumor tissue.

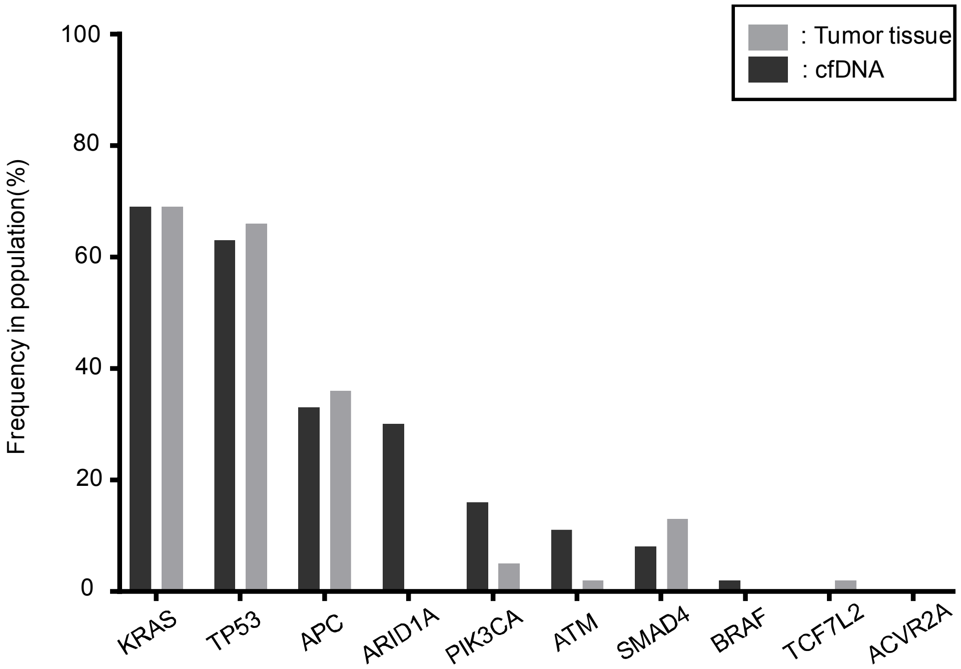
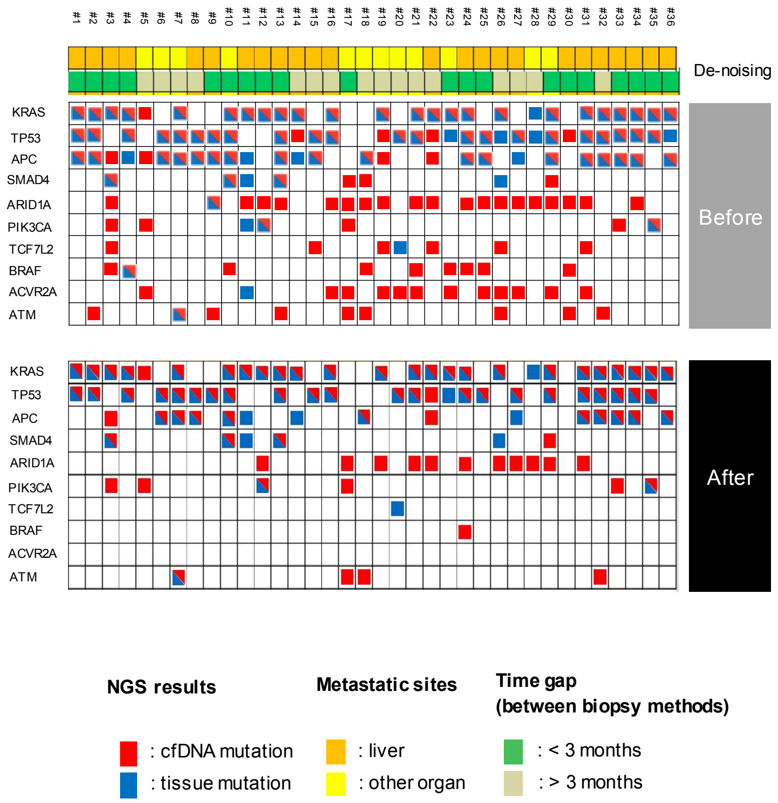


Figure 8. Compared between before and after denoising, actionable cfDNA mutations detected by deep target-sequencing Among patients, there were diversity for genotype of metastatic CRC. Somatic alteration detected in plasma cfDNA and tumor tissue marked in red and blue. If there were same position of somatic alteration in patient, marked in red and blue together. Patients who have metastatic lesions in Liver and other organ marked in orange and yellow. The patient's samples from patients which had time gap between tissue biopsy and liquid biopsy marked in green (less than 3 months) and olive (more than 3 months).



Correlation between cfDNA and metastasis

We validated the correlation between cfDNA somatic variants and specific organ metastasis using cfDNA samples from 54 patients. Twenty-eight patients had metastatic liver lesions. Mutated fragments of KRAS, TP53, and APC, which were the most frequently detected genes in plasma, were more detected in patients with liver metastasis (Fig. 9). In addition, higher levels of cfDNA were detected in samples from patients with liver metastasis, compared to patients with metastasis in other organs (Fig. 10).

Kaplan–Meier plot of overall survival rate suggests worse prognosis in high level of cfDNA concentration (>10ng/mL; n=17) than in low level of cfDNA concentration (<10 ng/mL; n=35; HR = 2.784, 95% CI: 1.031 – 7.518, P=0.0054). Median survival period of patients who had either KRAS or TP53 mutant allele in plasma was 33 months and 32 months, respectively. This period was shorter than median survival period of patients without KRAS and TP53 mutant allele in plasma (63 months for KRAS negative and 77 months for TP53 negative, KRAS; HR = 1.95, 95% CI: 0.9038 – 4.206, P=0.0368, TP53; HR = 2.152, 95% CI: 0.9893 – 4.682, P=0.05, Fig. 11A, B). Moreover, we examined the abdominal CT scans for patients with liver metastatic lesions (n=28). The size of the metastatic lesion in the liver was positively correlated with the level of cfDNA in the plasma and the detection of mutant fragments (#13, #24 vs. #22, #31) (R = 0.356, Fig. 12).

Figure 9. For specific organ metastasis, concentration and mutant fragments of plasma cfDNA were higher. (n=54) High mutant fragments of KRAS were detected in liver metastatic patients. (**p=0.002; Welch's t test) High mutant fragments of TP53 were detected in liver metastatic patients. (**p=0.007; Welch's t test) High mutant fragments of APC were detected in liver metastatic

patients (*p=0.03; Welch's t test).

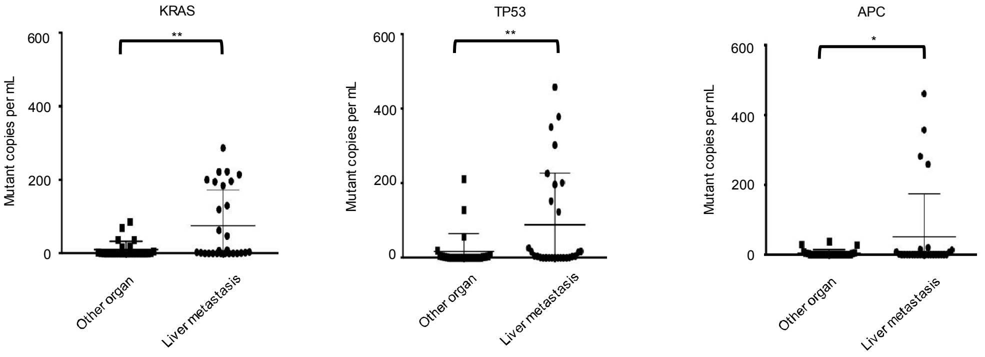


Figure 10. Medians of cfDNA concentration with metastatic lesion in liver and without were 5.43 and 7.80 each. (n=54, *p=0.05; Mann-Whitney test.)

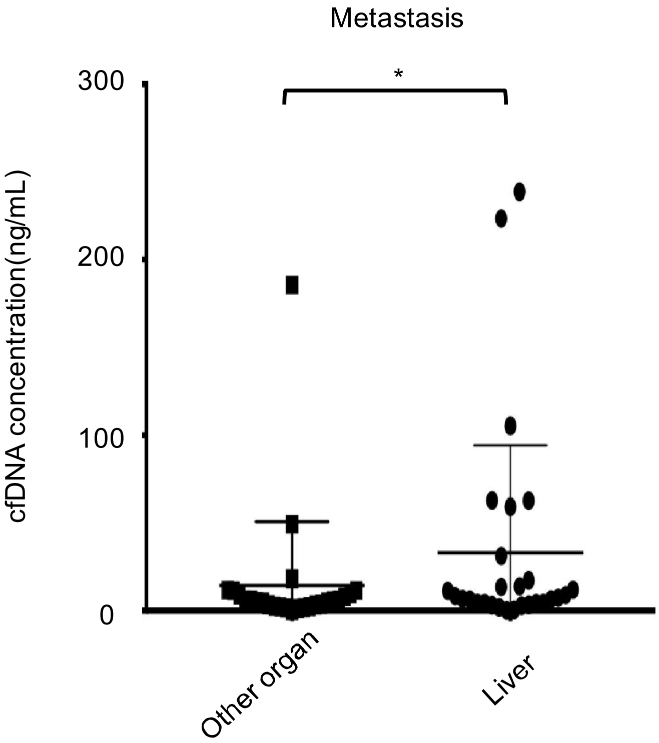
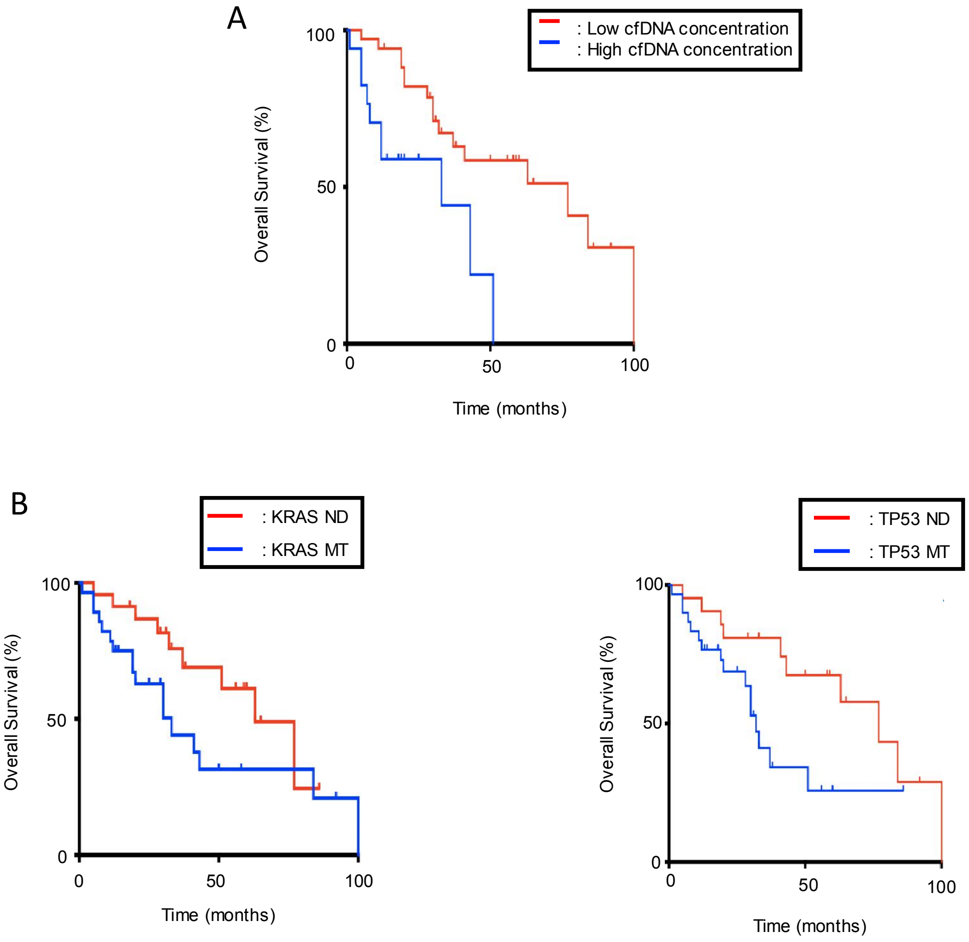


Figure 11. Correlation patients overall survival with mutant fragments in plasma cfDNA. (n=52) (A)Statistical analysis was

performed by Log-rank test.(HR = 2.784, 95% CI: 1.031 – 7.518, P=0.0054), (B)Statistical analysis was performed by Gehan-Breslow-Wilcoxon test for KRAS and Log-rank test for TP53 (KRAS; HR = 1.95, 95% CI: 0.9038 – 4.206, P=0.0368, TP53; HR = 2.152, 95% CI: 0.9893 – 4.682, P=0.05).



MT : Mutant in plasma / ND : Not detected in plasma

Figure 12. Correlation metastatic tumor burden in liver with mutant

fragments in plasma cfDNA. There was trend about size of metastatic lesion in liver (n=28, 3 excluded too disseminated to measure metastatic lesion in liver, R=0.356)

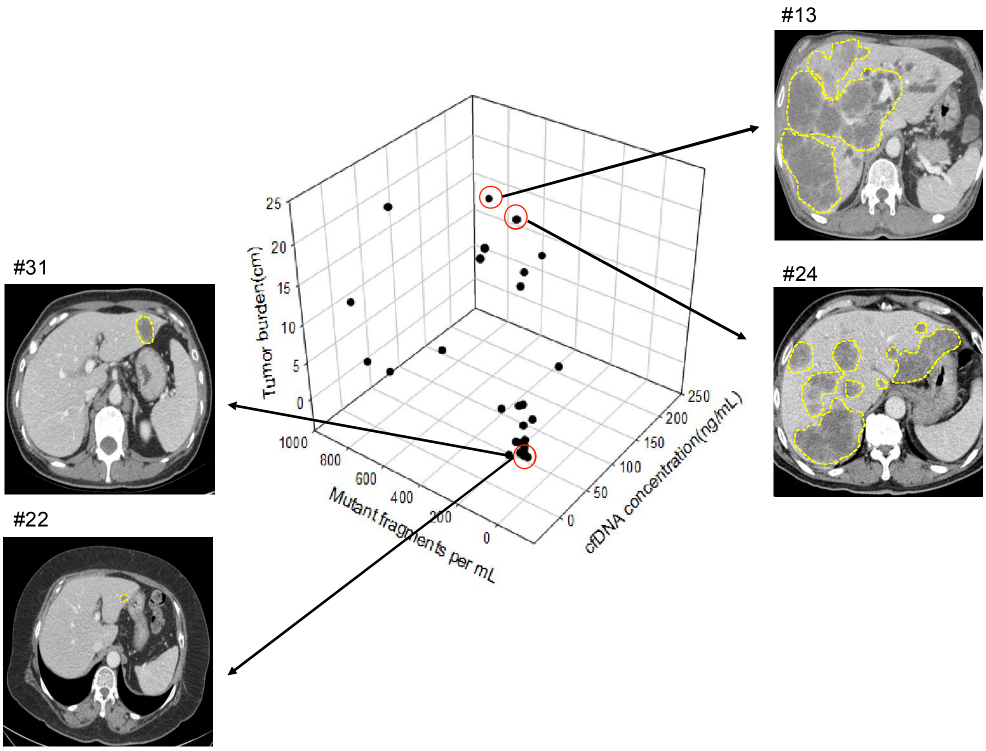
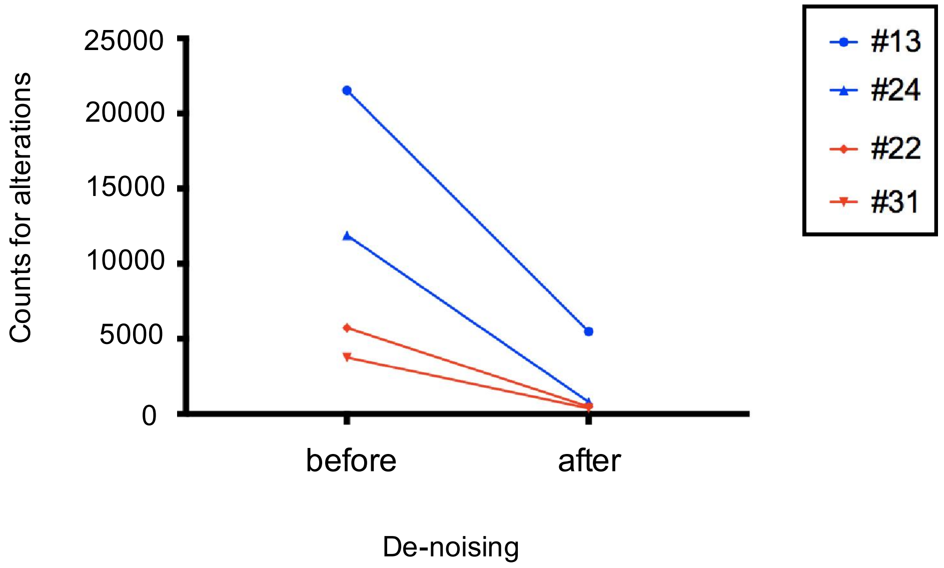


Figure 13. Difference of mutational alterations depending on

metastatic burden size in liver. (heavy burden; blue, light burden; red)



DISCUSSION

In this study, we devised a method for reducing the errors associated with deep targeted sequencing to provide a clinically feasible method for monitoring mCRC. Our results indicate that de-noising reduced NGS errors, allowing cfDNA analysis to be used to determine the status of patients with mCRC. Error-prone positions were notably reduced by de-noising. Although mutant fragments in the plasma of patients with light or heavy size of metastases in the liver have a positive trend as mentioned, the reduced rate between before and after the de-nosing was similar to the overall mean values (Fig. 13).

Numerous studies have used NGS for genotyping cfDNA alterations in plasma for diagnosing and evaluating patient status. For example, circle sequencing uses a circular library for highly accurate sequencing [14], while other methods using barcodes and additional bioinformatics pipelines include the safe-sequencing system (Safe-SeqS), duplex sequencing, tagged-amplicon deep sequencing (Tam-Seq), and personalized cancer profiling by deep sequencing (CAPP-Seq) [15] [16] [17] [18] [19]. These studies all detected somatic alterations in circulating cfDNA. However, a few studies have only reported on the concordance of hotspot alterations in circulating cfDNA, while our study targeted the whole exome. Moreover, bioinformatics has been recognized as an important clinical tool for analyzing valuable NGS data from raw sequencing data [20]. Similar to our approach, previous studies have used advanced methods to reduce NGS error, including integrated digital error suppression for improved detection, and methods analyzing the base-position error rate [21] [22]. Previous studies used healthy controls, but we were able to filter out error-prone sites using the biological background error of the cancer patients, without the need for healthy controls, thus reducing cfDNA background effects caused by genetic factors in the healthy controls themselves. However, despite the high quality of the results, these studies only analyzed hotspot somatic alterations [23]. In contrast, our study demonstrated concordance between deep targeted sequencing results for cfDNA and tumour tissues at positions representing not only hotspot somatic mutations, but also other positions. Although healthy controls were used to suppress NGS errors in recent studies, the biological background was also detected in cancer patients [19]. So, our group used the biological background in mCRC patients and reduced its rate (Supplementary Fig S3). The somatic mutation in plasma of patient who have KRAS mutation with lowest VAF (0.27%) was remained after de-noising (Fig 1B).

However, de-noising reduced the sensitivity of APC and other targeted genes (ACVR2A and ARID1A), possibly because of the presence of error-prone positions in both cfDNA and tumour tissue that were below our cutoff value before de-noising (Supplementary Fig. S4). Because de-noising is based on removing error-prone positions, such as homo-polymer regions, the procedure could reduce the values required for accuracy. Nevertheless, there was some dis-concordance between cfDNA and tumour tissue (Fig. 3). Firstly, some somatic alterations were detected in circulating cfDNA but not in tumour tissue. As tumour tissue biopsies (invasive) are selective, unlike liquid biopsies, somatic alterations present in only a small part of the tumour may not be detected in tissue biopsies. Regarding the metastatic patient cohort, somatic alterations that were not detected in circulating cfDNA might have been derived from the metastatic lesions. Secondly, some somatic alterations were detected in tumour tissue but not in circulating cfDNA; although circulating cfDNA is derived from the tumour as a result of apoptosis, necrosis, secretion, and circulates in the peripheral blood1, levels of specific mutations may be too low to be detected by NGS. This may be because of the presence of newly-occurring minor sub-clones in the tumour, with low levels of somatic mutations. These differences may be explained by intra-tumoural heterogeneity [24].

In our study, cfDNA levels and mutational fragments in circulating cfDNA were correlated with the size of the liver metastatic lesion. Originally, liver is the most frequent site of metastasis for colorectal cancer, accounting for about 60% to 80% of the cases of metastatic colorectal cancer [25]. Genotyping of cfDNA may represent liver metastatic lesions better than metastasis in other organs because of the anatomic properties of the liver [26]. Our novel findings suggest that circulating mutant fragments of tumour DNA in the plasma may represent progression of liver metastasis in advanced CRC, thus providing us a useful, non-invasive biomarker (Supplementary Fig. S8). Various methods for diagnosing colorectal cancer have been used in clinic [27]. For genotyping cfDNA using

NGS platform in clinic, turnaround time(TAT) represents an important factor in terms of the clinical application. Some studies have reported TAT for detecting EGFR and KRAS mutations in plasma of patients with advanced lung cancer, using digital droplet PCR [28]. Likewise, in platforms using NGS with liquid biopsy, the tissue of origin can be identified by detecting epigenetic markers in circulating cfDNA [29] [30] [31]. It is therefore necessary to set a clinically useful TAT for the proposed cfDNA genotyping approach in CRC patients, after which we aim to focus on the use of circulating cfDNA for early diagnosis, and for monitoring drug response using longitudinal samples.

In conclusion, the results of the current study demonstrate that de-noising can correct error-prone positions to allow the detection of clinically meaningful somatic alterations in circulating cfDNA. Moreover, our findings suggest that cfDNA may be used to determine the status of mCRC patients using de-noised, deep targeted sequencing.

REFERENCE

1. Heitzer E, Ulz P, Geigl JB: **Circulating tumor DNA as a liquid biopsy for cancer.** *Clin Chem* 2015, **61**(1):112-123.
2. Kim ST, Chang WJ, Jin L, Sung JS, Choi YJ, Kim YH: **Can Serum be Used for Analyzing the KRAS Mutation Status in Patients with Advanced Colorectal Cancer?** *Cancer Res Treat* 2015, **47**(4):796-803.
3. Diaz LA, Jr., Bardelli A: **Liquid biopsies: genotyping circulating tumor DNA.** *J Clin Oncol* 2014, **32**(6):579-586.
4. Russo M, Siravegna G, Blaszkowsky LS, Corti G, Crisafulli G, Ahronian LG, Mussolin B, Kwak EL, Buscarino M, Lazzari L *et al*: **Tumor Heterogeneity and Lesion-Specific Response to Targeted Therapy in Colorectal Cancer.** *Cancer Discov* 2016, **6**(2):147-153.

5. Laehnemann D, Borkhardt A, McHardy AC: **Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction.** *Brief Bioinform* 2016, **17**(1):154-179.

6. Macarthur D: **Methods: Face up to false positives.** *Nature* 2012, **487**(7408):427-428.

7. 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):587-593.

8. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, Fostel JL, Friedrich DC, Perrin D, Dionne D *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.

9. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A *et al*: **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic Acids Res* 2011, **39**(Database issue):D945-950.

10. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotechnol* 2011, **29**(1):24-26.

11. Lee DW, Han SW, Cha Y, Bae JM, Kim HP, Lyu J, Han H, Kim H, Jang H, Bang D *et al*: **Association between mutations of critical pathway genes and survival outcomes according to the tumor location in colorectal cancer.** *Cancer* 2017.

12. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA *et al*: **Mutational landscape and significance across 12 major cancer types.** *Nature* 2013, **502**(7471):333-339.

13. Han SW, Kim HP, Shin JY, Jeong EG, Lee WC, Lee KH, Won JK, Kim TY, Oh DY, Im SA *et al*: **Targeted sequencing of cancer-related genes in colorectal cancer using next-generation sequencing.** *PLoS*

One 2013, **8**(5):e64271.

14. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL: **High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing.** *Proc Natl Acad Sci U S A* 2013, **110**(49):19872-19877.
15. Kukita Y, Matoba R, Uchida J, Hamakawa T, Doki Y, Imamura F, Kato K: **High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients.** *DNA Res* 2015, **22**(4):269-277.
16. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B: **Detection and quantification of rare mutations with massively parallel sequencing.** *Proc Natl Acad Sci U S A* 2011, **108**(23):9530-9535.
17. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA: **Detection of ultra-rare mutations by next-generation sequencing.** *Proc Natl Acad Sci U S A* 2012, **109**(36):14508-14513.
18. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, Dawson SJ, Piskorz AM, Jimenez-Linan M, Bentley D *et al*: **Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA.** *Sci Transl Med* 2012, **4**(136):136ra168.
19. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, Liu CL, Neal JW, Wakelee HA, Merritt RE *et al*: **An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage.** *Nat Med* 2014, **20**(5):548-554.
20. Oliver GR, Hart SN, Klee EW: **Bioinformatics for clinical next generation sequencing.** *Clin Chem* 2015, **61**(1):124-135.
21. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, Stehr H, Liu CL, Bratman SV, Say C *et al*: **Integrated digital error suppression for improved detection of circulating tumor DNA.** *Nat Biotechnol* 2016, **34**(5):547-555.

22. Pecuchet N, Rozenholc Y, Zonta E, Pietraz D, Didelot A, Combe P, Gibault L, Bachet JB, Taly V, Fabre E *et al*: **Analysis of Base-Position Error Rate of Next-Generation Sequencing to Detect Tumor Mutations in Circulating DNA.** *Clin Chem* 2016, **62**(11):1492-1503.

23. Bennett CW, Berchem G, Kim YJ, El-Khoury V: **Cell-free DNA and next-generation sequencing in the service of personalized medicine for lung cancer.** *Oncotarget* 2016.

24. Burrell RA, McGranahan N, Bartek J, Swanton C: **The causes and consequences of genetic heterogeneity in cancer evolution.** *Nature* 2013, **501**(7467):338-345.

25. Sunami E, Tsuno N, Osada T, Saito S, Kitayama J, Tomozawa S, Tsuruo T, Shibata Y, Muto T, Nagawa H: **MMP-1 is a prognostic marker for hematogenous metastasis of colorectal cancer.** *Oncologist* 2000, **5**(2):108-114.

26. Lalor PF, Lai WK, Curbishley SM, Shetty S, Adams DH: **Human hepatic sinusoidal endothelial cells can be distinguished by expression of phenotypic markers related to their specialised functions in vivo.** *World J Gastroenterol* 2006, **12**(34):5429-5439.

27. Kuipers EJ, Rosch T, Bretthauer M: **Colorectal cancer screening--optimizing current strategies and new directions.** *Nat Rev Clin Oncol* 2013, **10**(3):130-142.

28. Sacher AG, Paweletz C, Dahlberg SE, Alden RS, O'Connell A, Feeney N, Mach SL, Janne PA, Oxnard GR: **Prospective Validation of Rapid Plasma Genotyping for the Detection of EGFR and KRAS Mutations in Advanced Lung Cancer.** *JAMA Oncol* 2016, **2**(8):1014-1022.

29. 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.

30. Choi W, Lee J, Lee JY, Lee SM, Kim DW, Kim YJ: **Classification of Colon Cancer Patients Based on the Methylation Patterns of Promoters.** *Genomics Inform* 2016, **14**(2):46-52.

31. Laird PW: **Principles and challenges of genomewide DNA methylation analysis.** *Nat Rev Genet* 2010, **11**(3):191-203.

국문 초록

차세대 유전체분석 (NGS) 기술은 혈액을 순환하는 무세포 DNA (cfDNA)의 유전형을 분석하고 환자 모니터링을 위한 주요 기술로 부상하고 있다. 그러나 NGS의 결과에는 많은 오류가 발생할 수 있다. 이 연구에서, 전이성 대장 암 환자 54 명의 39 개의 이용 가능한 순환하는 종양 무세포DNA (cfDNA)와 게놈 DNA를 분리했다. 대장암환자에서 빈번하게 발견되는 돌연변이 10 개를 표적하는 유전자 패널을 이용하여 표적염기서열분석을 진행하였다. 염기서열분석에서 발생하는 오류를 줄

이기 위해 '오류 제거'절차를 고안하여 무세포 DNA (cfDNA)와 종양 조직의 염기서열분석 데이터에서 검출된 돌연변이 사이의 일치도를 계산하였다. 10 개 유전자의 체세포 돌연변이에 대한 민감도, 특이도 및 정확도는 오류 제거 전과 후 각각 84.5 %, 74.6 % 및 76.9 %에서 87.3 %로, 92.0 %에서 91.1 %로 증가했다. 이 접근법은 전이성 대장암 무세포 DNA (cfDNA)의 체세포 돌연변이의 검출 능력을 향상시켰다. 본 방법은 0.27 % ~ 79.42 % 범위의 유전자 돌연변이에 대해 임상 적으로 중요한 변이를 선택적으로 검출 할 수 있었다. 이외에, 높은 무세포 DNA (cfDNA) 농도로 검출된 환자는 낮은 농도로 검출된 환자보다 더 많은 체세포 돌연변이 단편과 더 큰 간 전이 병변을 검출 되었다. 이러한 결과는 무세포 DNA (cfDNA) 유전자 분석에 대한 표적염기서열분석의 오류재거의 적합성을 입증하고, 병기가 진행된 대장암 환자의 전이 병변을 모니터하는 전략에 적합한 방법으로 생각된다.

주요어 : 대장암, 무세포유전자, 혈액종양DNA, 차세대유전체분석, 예후 예측, 디노이징

학번 : 2016-26001