약학박사 학위논문

Establishment of Genomic/Epigenomic Analysis Methods Based on Next Generation Sequencing, and Genomic Alterations Discovery in Korean Triple Negative Breast Cancer

차세대염기서열법 기반의 유전체/후성유전체 분석 기법 확립과 한국인 삼중음성유방암의 유전 변이 발굴

2016년 8월

서울대학교 대학원
약학과 병태생리학 전공
정 해 민

Establishment of Genomic/Epigenomic Analysis Methods Based on Next Generation Sequencing, and Genomic Alterations Discovery in Korean Triple Negative Breast Cancer

차세대염기서열법 기반의 유전체/후성유전체 분석 기법 확립과 한국인 삼중음성유방암의 유전 변이 발굴

지도교수 신 영 기

이 논문을 약학박사 학위논문으로 제출함
2016년 6월

서울대학교 대학원
약학과 병태생리학 전공
정 해 민

정해민의 박사학위논문을 인준함
2016년 6월

위 원 장 _____ (인)
부위원장 _____ (인)
위 원 _____ (인)
위 원 _____ (인)
위 원 _____ (인)

ABSTRACT

Establishment of Genomic/Epigenomic Analysis Methods Based on Next Generation Sequencing, and Genomic Alterations Discovery in Korean Triple Negative Breast Cancer

Hae Min Jeong

College of Pharmacy

The Graduate School

Seoul National University

The development of Next Generation Sequencing (NGS) technology makes it possible to analyze huge amount of sequencing data with low cost, and stimulates worldwide genetic research field. Therefore, NGS will be applicable to diagnose or treat the patient with genetic disease (e.g. cancer) in the near future. NGS can be used in various genome-wide genomic/epigenomic analyzes including Targeted exome sequencing for analysing only exon regions of interesting genes, RNA

sequencing for analysing RNA expression profiles and splicing variants, Bisulfite sequencing for analysing DNA methylation and Chromatin Immunoprecipitation sequencing (ChIP-seq) for analysing histone modification. While the cost for sequencing data generation rapidly decreased with NGS technology progression, the cost for data storage and data analysis increases proportionally to the amount of sequencing data. Hence, it is important to perform NGS within interesting regions depending on the purpose of each research, increasing the efficiency of time and cost.

In the first part of this study, the efficiency of a methylated DNA immunoprecipitation bisulfite sequencing (MeDIP-BS) method, that involves a combination of methylated region enrichment using 5-mC antibodies and bisulfite conversion, was evaluated. By taking the advantage of the low cost of Methylated DNA Immunoprecipitation sequencing (MeDIP-seq) and the high resolution of Whole Genome Bisulfite sequencing (WG-BS), this method not only remarkably improves cost effectiveness, but also dramatically enhances analysis resolution, achieving base-pair resolution. In addition, by comparing this method to WG-BS, MeDIP-Seq, and Targeted Bisulfite sequencing (Targeted-BS) in analyzes using human liver and stomach samples, it is proved that MeDIP-BS is applicable for clinical diagnostics, guaranteeing cost-effective high read depth and high-resolution genome-wide DNA

methylation analysis.

In the second part of this study, genomic alterations of Korean triple negative breast cancer (TNBC) patients were profiled by targeted exome sequencing which aims at analysing target exome regions. This method had revolutionized human clinical cancer diagnosis, cancer-causing mechanism studies and processes for identifying therapeutic targets due to its cost-effectiveness compared with whole genome or whole exome NGS. The targeted exome sequencing is very advantageous at providing more reliable accuracy of mutation and copy number alteration analysis by generating sufficiently deeper coverage of sequencing reads in target exon regions at a relatively lower cost compared with whole exome NGS. In particular, HaloPlex target enrichment system had already been substantiated for its advantageous usefulness in the targeted exome NGS due to its high efficiency upon capturing the targeted regions on the exome. Through this study, for the first time to our knowledge, it is revealed that mutation and copy number variation landscapes on targeted regions for 368 cancer-associated genes from 70 Korean TNBC patients. Futhermore, some homozygous deletion genes have significant correlation with prognosis, suggesting the potential role as prognostic biomarker.

Key words: Next Generation Sequencing, DNA methylation, Methyl DNA Immunoprecipitation bisulfite, Triple Negative Breast Cancer, Targeted Exome Next Generation Sequencing, Somatic mutation, DNA repair gene

Student ID Number: 2009-21718

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

PART I

PART Ⅱ

LIST OF TABLES

INTRODUCTION

PART I

PART II

LIST OF ABBREVIATIONS

NGS	Next Generation Sequencing
MeDIP-BS	Methylated DNA Immunoprecipitation-Bisulfite Sequencing
WG-BS	Whole Genome-Bisulfite Sequencing
Targeted-BS	Targeted-Bisulfite Sequencing
MeDIP-seq	Methylated DNA Immunoprecipitation-Sequencing
TNBC	Triple Negative Breast Cancer
ER	Estrogen Receptor
PR	Progesteron Receptor
HER2	erb-b2 receptor tyrosine kinase 2

INTRODUCTION

I. Next Generation Sequencing

I-1. The principle of Next Generation Sequencing

Since DNA double helix structure was first discovered at 1950s by Francis Crick and James Watson, the genomics has been rapidly progressed with DNA sequencing technologies. Sanger sequencing was developed by Frederick Sanger in 1977, and has been the most widely used until a recent date. This technique is based on classical chain-termination method which use di-deoxynucleotide triphosphates (ddNTPs). While DNA polymerase duplicate complement sequence to DNA template, many randomly terminated DNA fragments are produced and analyzed to find out sequence of DNA sample. But, Sanger method include bad quality in the first 10~30 bases of the sequence due to primer binding and degenerated sequencing quality after 700-900 bases (1). And DNA fragment which will be sequenced, must be cloned to cloning vector before Sanger sequencing. Because of these challenges, Sanger sequencing is not suitable for huge size genomic sequencing analysis.

Next Generation Sequencing (NGS) is first developed in 2004. At first, 454FLX (Roche), Solexa genome analyzer (Illumina) and SOLiD

system (Applied Biosystems) are typical NGS platforms. Each platform is composed of complex interactions of enzymology, chemistry, high-resolution optics, hardware and software engineering (2), but the main principle of NGS is based on shotgun sequencing and massively parallel sequencing. Large DNA molecule is fragmented to appropriate size and small oligomer (Adaptor) is ligated to the end of fragmented DNA, and amplified to prepare library. Then a huge number of libraries are hybridized to the beads or flow cell, and simultaneously sequenced to generate numerous sequencing data. Finally, generated reads are assembled to the reference genome sequence (Figure 1).

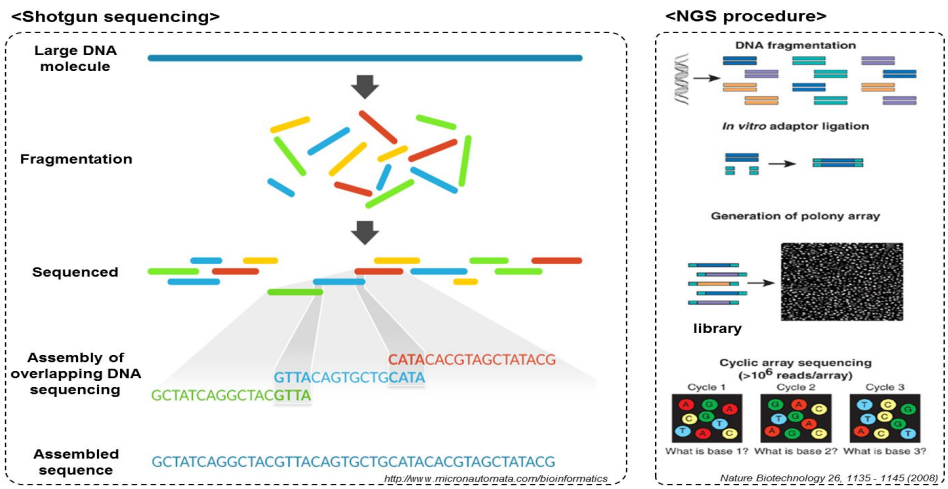


Figure 1. Shotgun sequencing and the principle of NGS. Large DNA molecule is sheared to generate DNA fragments and sequenced reads are aligned to reference sequence to confirm the sequence of original DNA molecule. Adapted from Micron, 2016 (3). Two key points of NGS are in vitro adaptor ligation and cyclic array sequencing. Adapted from Shendure J, 2008 (4).

I-2. Integrated Genomic/Epigenomic Analysis

Genomics and epigenomics are very closely cooperate to regulate various gene functions such as gene expression. Therefore, integrated genomic/epigenomic analysis is necessary to comprehensively understand a variety of biological processes and development mechanisms of genetic diseases including cancer. With development of NGS technology, the cost for DNA sequencing have been significantly reduced (Figure 2), and NGS can be applied nearly all of genomic/epigenomic research field (Figure 3).

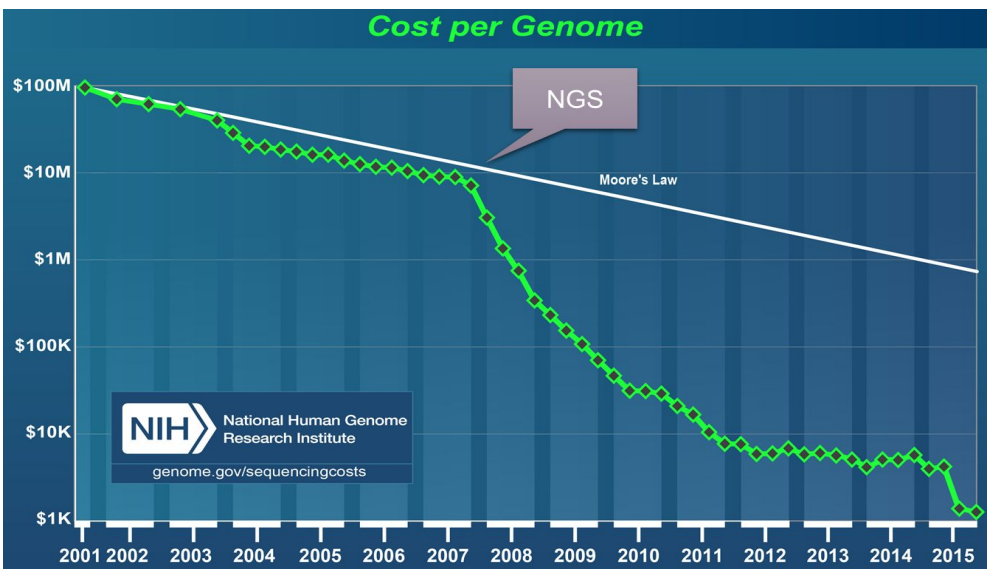


Figure 2. Sequencing cost for human genome analysis. NGS significantly reduced the cost of whole genome sequencing for a person. Adapted from NIH, 2016 (5).

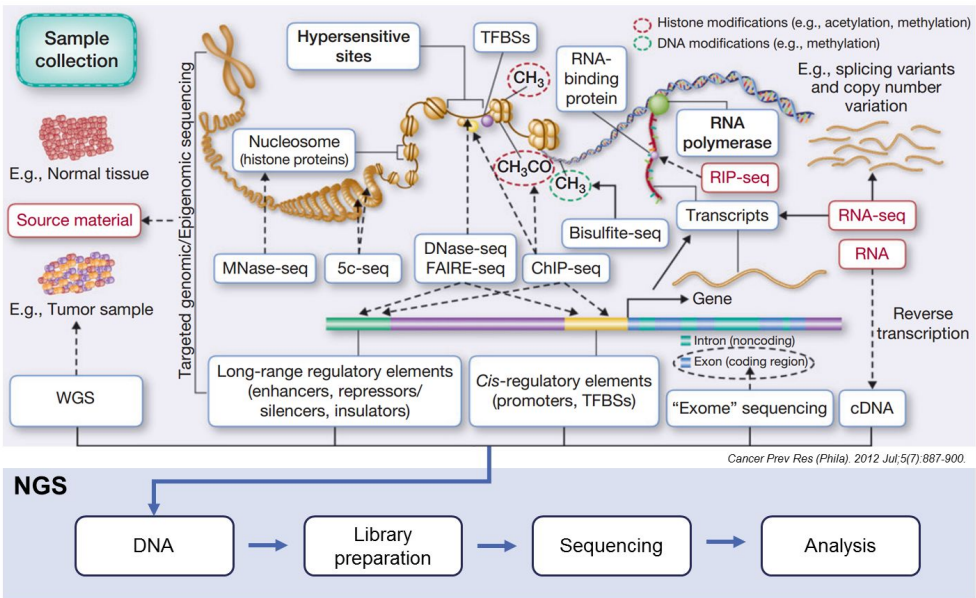


Figure 3. Various application of NGS. NGS can be applied to various genomic/epigenomic research fields by combining other experimental techniques. Adapted from Rizzo, 2012 (6).

Hence, NGS is a key technology for investigating the mechanisms underlying cancer by performing diverse genomic/epigenomic NGS and comprehensively analyzing them at the same time.

Ⅱ. The Cancer Genome

Ⅱ-1. What is Cancer?

Cancer is a collection of related diseases which caused by uncontrolled growing cells. Enlarged cancer cell mass(tumor) may disturbs the function of adjacent normal tissues and organs, and even may metastasize to distant part in the body to make secondary tumor. Benign tumors are do not spread to other tissues, and usually do not grow back after surgical tumor removal. Whereas, malignant tumors can invade not only nearby tissues, but also distant organs through blood or lymph system. When distant metastasis is detected, it is very difficult to remove all widely dispersed cancer cells by only a surgical operation. In that case, radiotherapy and chemotherapy are used to get rid of cancer cells. But, these therapies also damage normal cells causing severe side effects such as anemia, leukopenia, diarrhea, nausea, vomiting, hair loss and so on. Therefore, the importance of targeted therapy has been emphasized recently to treat cancer, which can remove only cancer cells with less side effects.

The cancer cell has many different features from normal cells that help them to grow out of control and become invasive. Figure 4 shows

typical hallmarks of cancer cells (7).

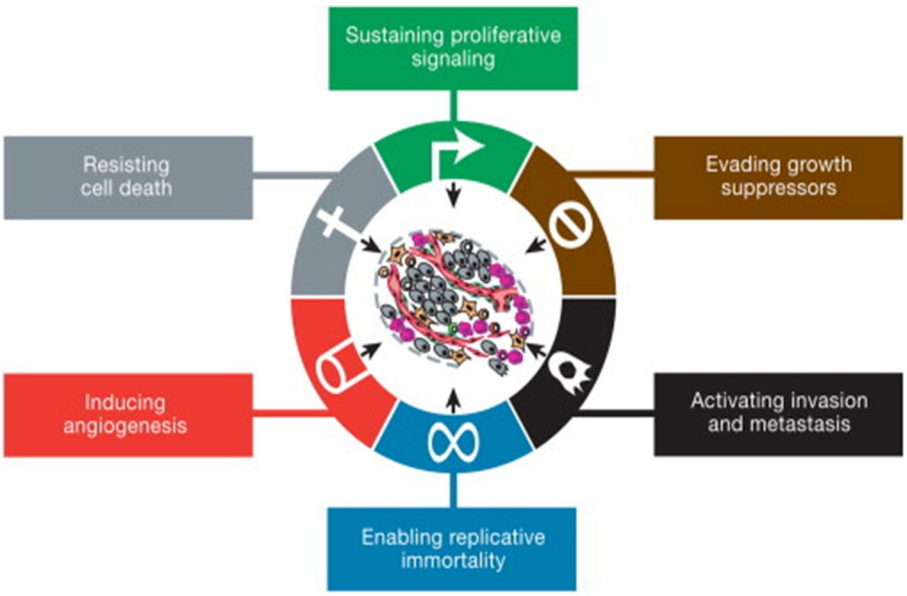


Figure 4. Hallmarks of cancer. The typical hallmarks of cancer are composed of six biological capabilities acquired during the multistep development of human tumors. Adapted from Hanahan D, 2011 (7).

II-2. How Cancer Arises?

Actually, cancer cells originate from normal cell rather than comes from outside of the body. As is well known, a gene is fundamental factor which determines diverse cell features, so it makes logical sense to assume that cancer is caused by genetic mutations. Now there is no doubt that cancer is genetic disease - it is caused by abnormal expression of genes that control cells functions, especially how they grow and divide. These genetic changes can be inherited from parents or arise during a person's lifetime as a result of errors that occur as cells devide or because of damage to DNA caused by certain environmental factors such as tobacco smoking, alcohol drinking, ultraviolet rays and so on.

There are three main types of genes that contribute to cancer development - proto-oncogenes, tumor suppressor genes, DNA repair genes. Proto-oncogenes are related to cell growth and division in normal cell. But, when these genes are mutated in diverse ways or activated than normal, they become cancer-causing genes (oncogenes) allowing cells to grow and survive when they should not. Tumor suppressor genes are also related to cell growth and division, but on the contrary, loss of function mutation of tumor suppressor genes contribute to become cancerous. DNA repair genes have roles in fixing

damaged DNA, so cells with mutations in these genes have more chances of additional mutations in other genes, resulting gradual cancer development.

II-3. Cancer Genomics and Epigenomics

Cancer is a heterogeneous disease caused by uncontrollable cell division (8). Early research recognized cancer as a genetic disorder of proliferation-related genes, but recent studies have approached cancer multidirectionally, examining resistance to cell death, angiogenesis, invasion, metastasis, and other properties (7). Many genes related to these properties can be roughly divided into 2 groups: Tumor Supressor Genes (TSGs) and oncogenes or Cancer Promoting Genes (CpGs). As the name implies, TSGs have cancer inhibitory roles, and their functions are lost in cancer cells (9). As the name implies, TSGs have cancer inhibitory roles, and their functions are lost in cancer cells. Conversely, oncogenes or CPGs have the potential to cause cancer, and they are overexpressed in many cancers. Both TSGs and oncogenes or CPGs are abnormally expressed in diverse cancers via various mechanisms, and each gene has a specific function according to the characteristics of specific cancers. Cancer develops from both genetic and epigenetic mutations, and progression is more severe when such mutations accumulate, interrupting the normal function of cancer-related genes and inducing resistance to chemotherapy that makes cancer treatment more difficult (10). The best-known epigenetic mechanism is DNA methylation which usually represses downstream

gene expression by local chromatin structure change. Human tumors were initially discovered to have global DNA hypomethylation and local DNA hypermethylation patterns, and since then, many researchers have undertaken studies of the relationship between DNA methylation and cancer. As normal cells progress to invasive cancer cells, their overall DNA methylation levels decrease, whereas CpG island hypermethylation and alteration of histone modification patterns accumulate gradually. TSG inactivation by promoter regions hypermethylation or DNA hypomethylation of highly repeated DNA regions are found in diverse cancer types In addition, DNA hypomethylation in promoter regions has recently been discovered to derepress some CPGs or proto-oncogenes that are repressed in normal cells (11, 12).

Ⅲ. Triple Negative Breast Cancer

Breast cancer commonly arise from the cells of the ducts what is called ductal carcinoma, and also begin in the cells of the lobules and other tissues of breast. Breast cancer is one of the most common cancer among women and the incidence is increasing steadily in the worldwide. But, in spite of progression in cancer diagnostic technique and therapy, breast cancer is still leading cause of cancer mortality (13).

Estrogen receptor (ER) and progesterone receptor (PR) are proteins on certain cells that can interact with each hormone to activate related signaling pathways. It is well known that these two hormones often promote the growth of breast cancer cells. So the evaluation of ER/PR expression is essential step in breast cancer diagnosis. Human epidermal growth factor receptor2 (HER2) is also important receptor protein, which promotes the growth of breast cancer cells. It is known that about 20% breast cancers show amplified HER2/neu expression.

Traditionally, pathology-based classification had been used for breast cancer categorization. But recently, there has been significant progress in knowledge of the molecular biology of breast cancer, and now molecular-based classification was well established based on some major gene expression patterns. This makes individualized therapies possible,

resulting improvements in survival rate (14). Although breast cancer is highly heterogeneous, it is usually categorized four subtypes according to the expression patterns of ER, PR and HER2 - Luminal A type (ER+ and/or PR+, HER2-), Luminal B type (ER+ and/or PR+, HER2+), HER2+ type (ER-, PR-, HER2+), Triple negative type (ER-, PR-, HER2-). Triple negative breast cancer (TNBC) which is defined as ER-, PR-, HER2- account for approximately 20% of invasive breast cancers (Table 1) (15, 16).

Subtype	Molecular features
Luminal A	ER+ and/or PR+, HER2-
Luminal B	ER+ and/or PR+, HER2+
HER2 type	ER-, PR-, HER2+
Triple negative	ER-, PR-, HER2-

Table 1. Molecular subtype of breast cancer. Breast cancer is categorized into four subtypes based on molecular features of ER, PR and HER2.

These categorization is important to predict prognosis, and it helps management of breast cancer therapy (17). In case of Luminal A, B or HER2 type breast cancer, hormone therapy or HER2 receptor therapy can be used, but TNBC has no actionable target for targeted therapy yet. That is why TNBC is obstinacy cancer showing low 5-year survival rate compare to other subtypes, and it is necessary to stratify TNBC and grasp the genetic feature of each types to discover new diagnostic marker or therapeutic target for TNBC (18).

PART I

Establishment of methylated DNA immunoprecipitation bisulfite sequencing for whole genome DNA methylation analysis

Introduction

DNA methylation is a major epigenetic mechanism that plays important roles in various biological processes, including embryonic development, X-chromosome inactivation, transposable element repression, and genomic imprinting (19, 20). DNA methylation is most frequently observed at the C5 position of cytosine, followed by guanine at CpG in vertebrates, and non-CpG sites such as CHG and CHH in plants and mammalian embryonic stem cells (21).

DNA methylation of promoter regions often down-regulates gene expression (22). In addition, during embryogenesis and tissue differentiation, epigenetic mechanisms involving DNA methylation are crucial, and likely contribute to organ-specific gene expression (22). The genome-wide DNA methylation landscape undergoes dynamic changes during cellular differentiation, and the changes in regulatory regions are particularly remarkable (23). Thus, it is important to evaluate different DNA methylation patterns in various organs to understand diverse biologic processes such as embryogenesis, developmental processes, and tissue-specific functions. Additionally, numerous studies have suggested that many diseases, including cancer, neural diseases, and autoimmune diseases, could be caused by abnormal

gene expression resulting from disrupted epigenetic regulation including epigenetic silencing and derepression (22, 24, 25).

Several next-generation sequencing (NGS) methods have been developed to profile the DNA methylation landscape on a genome-wide scale. Bisulfite conversion and methylated region enrichment using 5-methyl cytosine antibodies or methyl-CpG binding domain proteins (MBD) are commonly used to measure genome-wide DNA methylation patterns. Whole-genome bisulfite sequencing (WG-BS) can be used to directly determine the whole-genome DNA methylation landscape through fragmentation and bisulfite conversion (26). Because WG-BS can detect 5-methyl cytosine at single-base resolution, it is considered the standard method for genome-wide DNA methylation analysis. However, WG-BS is not suitable for large-scale clinical applications, because of the huge amount of data generated. In contrast, targeted bisulfite sequencing (Targeted-BS) analyzes only selected genomic regions (approximately 84Mb for the SureSelect Human Methyl-Seq kit) where, for example, cancer tissue-specific differentially methylated regions, promoters, and CpG islands and shores are presumed to be found (27). Although this method has many advantages, including a high read depth, base-pair resolution, and low cost, but never identifies any DNA methylation information from non-target regions, which account for more than 97% of the genome and, most importantly, the vast

majority of yet unknown regulatory regions. Reduced representation bisulfite sequencing (RRBS) is another widely used bisulfite sequencing method that uses MspI restriction endonuclease to enrich CpG islands and promoter regions. However, this method covers even fewer CpGs and regions than Targeted-BS (28), therefore, we exclude it from our comparison study. Methylated DNA immunoprecipitation sequencing (MeDIP-seq) uses a 5-methyl cytosine antibody to capture methylated genomic DNA fragments; then, these enriched DNA fragments are sequenced and analyzed. Although this method can detect whole-genome DNA methylation levels at relatively low cost, its resolution is too low to precisely pinpoint the methylated CpG sites in the genome (29, 30). Recently, Weyrich et al. (31) generated whole genome DNA methylation profiling data from wild guinea pig using methylated DNA-enrichment-bisulfite-sequencing (MEBS). This method, which combines MBD enrichment with bisulfite conversion, generated DNA methylation data with single base resolution at a reasonable cost.

In this study, we evaluated the cost and coverage efficiency of a methylated DNA immunoprecipitation bisulfite sequencing (MeDIP-BS) method that involves a combination of methylated region enrichment using 5-mC antibodies and bisulfite conversion. By taking the advantage of the low cost of MeDIP-seq and the high resolution of WG-BS, this method not only remarkably improves cost effectiveness, but also

dramatically enhances analysis resolution, achieving base-pair resolution, as has been shown for MEBS. . In addition, by comparing this method to WG-BS, MeDIP-Seq, and Targeted-BS in analyzes using human liver and stomach samples, we show that MeDIP-BS is applicable for clinical diagnostics, guaranteeing cost-effective high read depth and high-resolution genome-wide DNA methylation analysis. Figure 5 shows a brief outline and Figure 6 shows workflow for the four different NGS based DNA methylation sequencing methods that were used in our study.

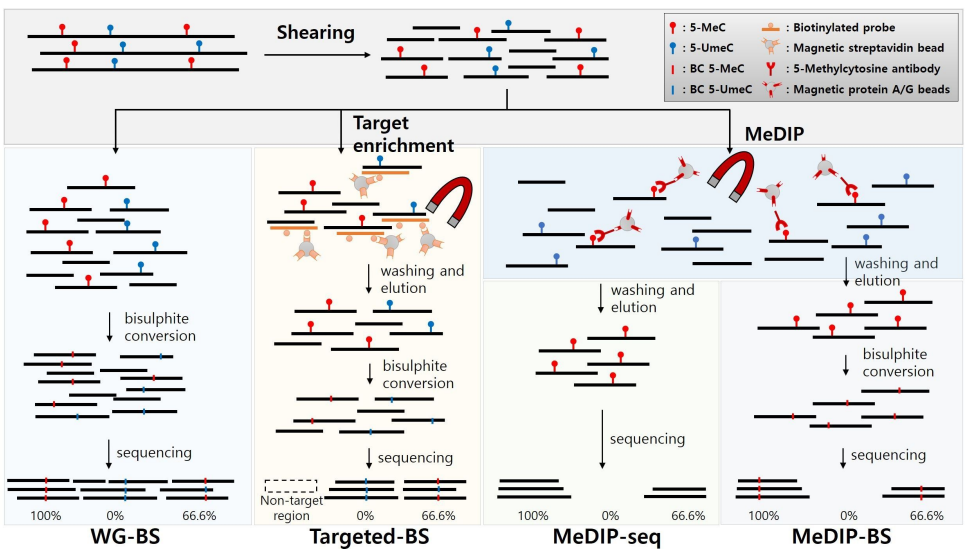


Figure 5. Schemes of the four different NGS-based DNA methylation analysis methods. Schemes of the four different NGS-based DNA methylation analysis methods included in the study. 5-MeC, 5-methylcytosine; 5-UmeC, 5-unmethylcytosine; BC 5-MeC, bisulfite-converted 5-methylcytosine; BC 5-UmeC, bisulfite-converted 5-unmethylcytosine.

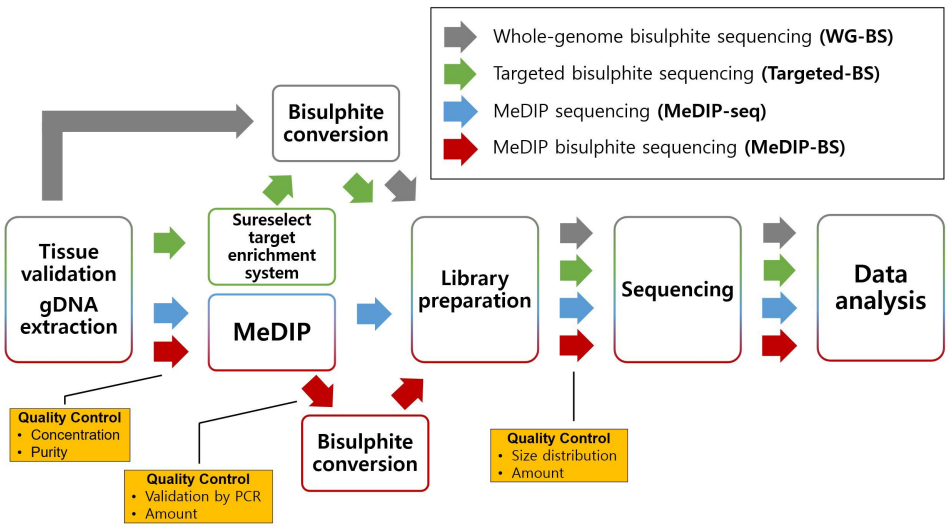


Figure 6. Workflow of four different methods. Workflow of the four different methods. The quality of the MeDIP-BS library was evaluated at each step.

Purpose of the study

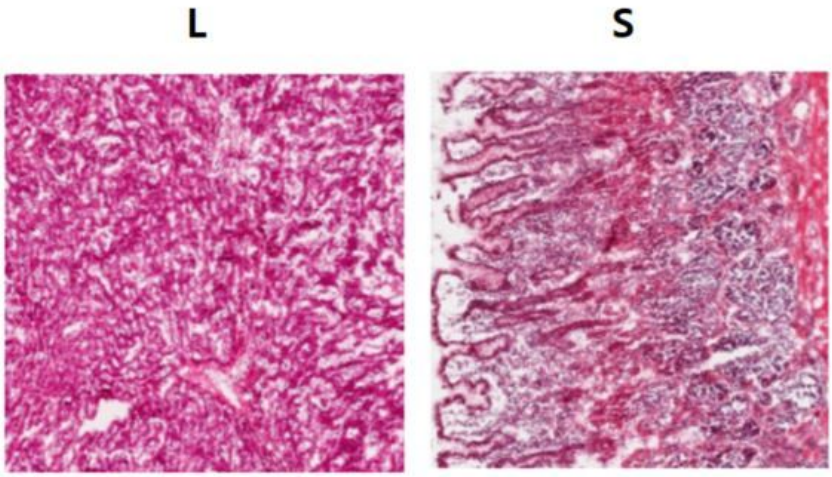
DNA methylation is well known epigenetic mechanism and plays significant roles in various biological processes including tissue specific gene expression. Several NGS based methods have been developed to profile the DNA methylation landscape on a genome-wide scale. Whole-genome bisulfite sequencing (WG-BS) can detect DNA methylation level at single-base resolution, so it is considered the standard method for genome-wide DNA methylation analysis. However, WG-BS is not suitable for large-scale clinical applications, because of the huge amount of data generated. In this study, we compared four NGS based DNA methylation analysis methods, WG-BS, targeted bisulfite sequencing (Targeted-BS), methylated DNA immunoprecipitation sequencing (MeDIP-seq) and methylated DNA immunoprecipitation bisulfite sequencing (MeDIP-BS), to find out the most efficient method which can substitute for WG-BS.

Materials and Methods

1. Tissue sample preparation

With the approval of the Review Board Committee of Samsung Medical Center (SMC), Seoul, Korea, various tissue samples were collected from cancer patients. Among them, liver (L) and stomach (S) were collected sufficiently to perform the four different NGS based DNA methylation sequencing methods. Therefore, we selected L and S in this study because of sample availability. For sampling of adjacent normal tissue, a minimal distance from the cancer margin to the area of the tissue from which the normal adjacent tissue was taken was 5 cm. The samples were snap frozen with liquid nitrogen, and stored at -80 ℃. The frozen sections for the samples were prepared and stained with haematoxylin and eosin (H&E) for validation by a pathologist (YLC) and were confirmed to satisfy the lesion criterion (Figure 7).

(A)



(B)

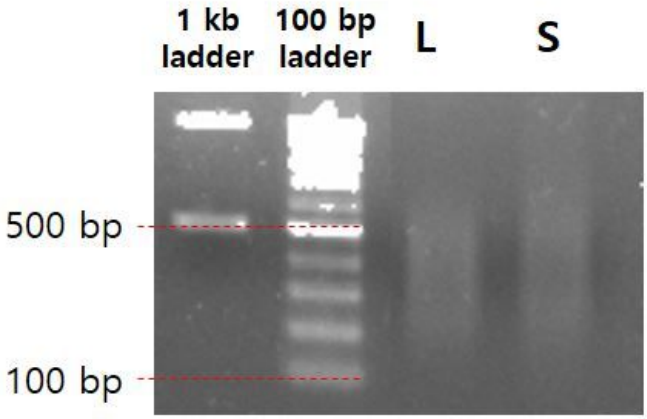
Organ	Inclusion criteria	Exclusion criteria
Liver	Parenchyma	Capsule, Inflammation
Stomach	Antrum, Mucosa	Mucosa ≤ 50%

Figure 7. Characteristics of normal tissue samples L and S. Sample L was collected from the area surrounding a metastatic liver tumour from a 60-year-old male patient, and sample S was collected from the area surrounding a gastric adenocarcinoma from a 58-year-old female patient. (A) Collected normal tissues were stained with Haematoxylin & Eosin and verified by a pathologist (YLC). (B) Normal lesion criteria.

2. Genomic DNA extraction and shearing

Genomic DNA was extracted from frozen tissue samples using the DNeasy Blood & Tissue kit (QIAGEN, Hilden, Germany) and stored at -80 ℃. The purity and concentration of the extracted DNA were

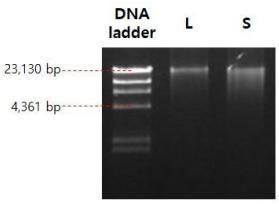
evaluated by spectrophotometry and were confirmed to meet the following criteria: concentration ≥ 100 ng/μL, 260/280 ratio ≥ 1.8, and 260/230 ratio ≥ 1.8 (Figure 8 and Figure 9). Next, the DNA was fragmented using a CovarisTM S-series S2 Focused-ultrasonicator (Woburn, MA, USA) to the appropriate size for each method.



Sample	Volume (uL)	Conc. (ng/uL)	260/280	260/230	Amount (μg)
L	150	479.00	1.91	2.28	71.85
S	150	374.00	1.84	1.89	56.10

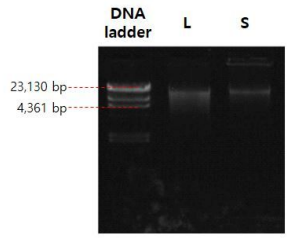
Figure 8. The quality of the gDNA extracted from the L and S (MeDIP-BS). The quality of the gDNA extracted from the L and S samples was verified and sheared to form DNA 100-500 bp fragments.

(A)



Sample	Volume (ul)	Conc. (ng/ul)	260/280	260/230	Amount (ug)
L	41	382.00	1.80	1.73	15.66
S	41	476.00	1.83	2.07	19.52

(B)



Sample	Volume (ul)	Conc. (ng/ul)	260/280	260/230	Amount (ug)
L	28	284.00	1.91	2.28	7.95
S	32	269.00	1.84	1.89	8.61

Figure 9. The quality of the gDNA extracted from the L and S (WG-BS and Targeted-BS). The quality of the genomic DNA extracted from L and S was evaluated by agarose gel electrophoresis and spectrophotometry. (A) The quality of the genomic DNA used in WG-BS and Targeted-BS. (B) The quality of the genomic DNA used in MeDIP-seq.

3. Target enrichment for Targeted-BS

Target regions were enriched using the SureSelect Human Methyl-Seq kit (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions. DNA was sheared to generate 150-200 bp fragments, hybridized to biotinylated RNA baits, which were designed to cover target regions, including CpG islands, cancer- or tissue-specific DMRs, Gencode promoters, and other regulatory feature regions. Then the hybridized DNA fragments were captured by streptavidin beads, and the unhybridized DNA fragments were washed out.

4. Methylated DNA immunoprecipitation (MeDIP)

MeDIP was performed using the MagMeDIP kit (Diagenode, Denville, NJ, USA) according to the manufacturer's instructions. First, gDNA was sheared to produce 200-500 bp fragments. Next, 1 μg of fragmented gDNA was mixed with Magbeads and 5mC antibody in IP solution. The mixture was incubated on a 4 ℃ rotator for 16 h. The following day, the gDNA-bead-antibody complexes were washed three times with MagWash buffer. The immunopreciptated DNA fragments were eluted with 100 μL of DIB buffer containing 1 μL of proteinase K at 55 ℃ for 15 min and then incubated at 100 ℃ for 15 min. The concentration was determined using a Qubit fluorometer 2.0 with the Qubit ssDNA Assay kit (Invitrogen, Carlsbad, CA, USA). Eight IPs were conducted for each sample to generate a sufficient amount of methylated DNA fragments, and the volume of the immunoprecipitated DNA was reduced to 20 μL by ethanol precipitation. The amount of captured DNA was greater than 50 ng.

5. Bisulfite conversion

Sheared genomic DNA (for WG-BS) or enriched DNA (for Targeted-BS and MeDIP-BS) was treated with sodium bisulfite to

convert unmethylated cytosine to uracil. bisulfite conversion was performed using the EZ DNA Methylation-Lightning kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions (32). Next, 20 μL of DNA obtained by each method was mixed with 130 μL of Lightning Conversion Reagent and incubated at 98 ℃ for 8 min and then at 54 ℃ for 60 min. After incubation, 600 μL of M-binding buffer was added to the mixture, and the mixture was transferred to a Zymo-Spin column. After centrifugation, the column was washed with 100 μL of M-wash Buffer and incubated with 200 μL of L-Desulphonation Buffer for 20 min. After two more wash steps, the bisulfite-converted DNA was eluted from the column using 11 μL of M-elution buffer.

6. Library preparation

Libraries were prepared using an EpiGnome Methyl-Seq Kit (Epicentre Biotechnologies, Madison, WI, USA) according to the manufacturer's instructions (33). First, the DNA synthesis primer was annealed to the bisulfite-converted DNA, and the DNA was copied using epigenome polymerase. The primer was digested with Exonuclease I, and the 3' end of the cDNA was continuously tagged with a Terminal-Tagging Oligo. After the di-tagged cDNA was purified using

the Agencourt AMPure XP system (Beckman Coulter, Brea, CA, USA), it was amplified by polymerase chain reaction (PCR). Finally, the amplified di-tagged cDNA was purified and the quality of the library was analyzed using a 2100 Bioanalyzer (Agilent).

7. Quality analysis of the MeDIP-BS library

The quality of the extracted DNA was evaluated, and the DNA was sheared into 100-500-bp fragments (Figure 8). For WG-BS, the sheared DNA was directly treated with sodium bisulfite, and for Targeted-BS, the DNA underwent a target enrichment step before bisulfite conversion. For MeDIP-seq and MeDIP-BS, we performed a methylated DNA region enrichment step, and then verified the accuracy of the MeDIP experiment. Methyl-cytosine-enriched DNA fragments were confirmed by PCR using four control primer pairs (two positive control primers, MeDNA and TSH2B, and two negative control primers, unMeDNA and GAPDH). The results indicated that immunoprecipitated DNA fragments were amplified in the positive control PCR, which targets highly methylated regions, but not in the negative control PCR, which targets unmethylated regions (Figure 10).

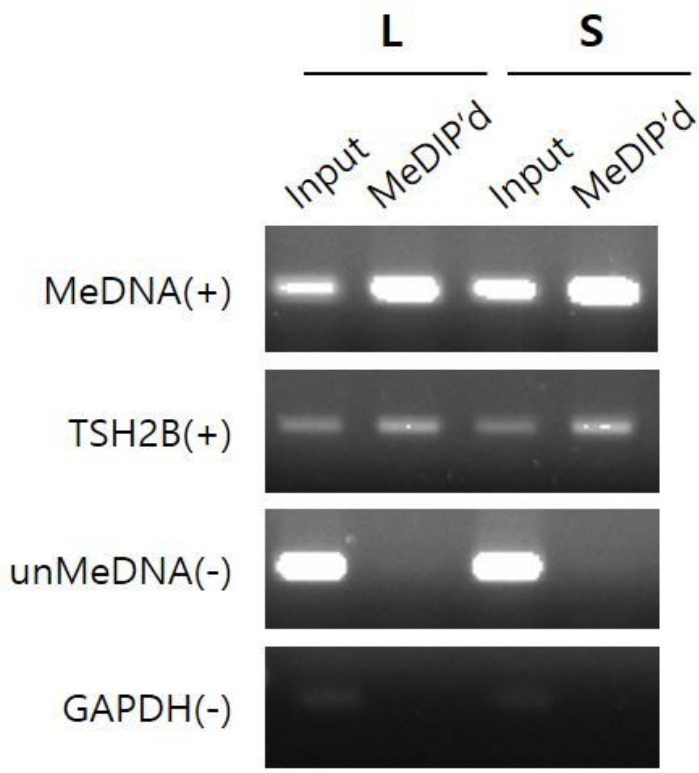
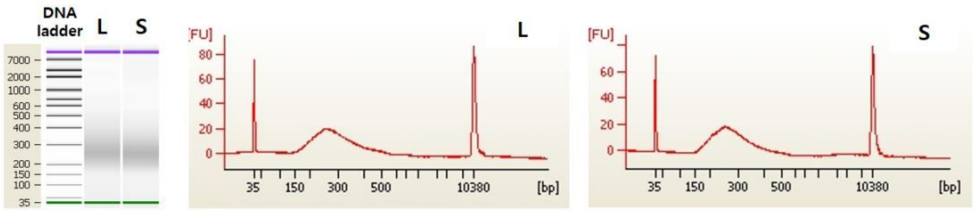


Figure 10. MeDIP validation experiment by PCR. The MeDIP experiment was validated by PCR using four control primers.

After confirmation, the MeDIP-seq library was directly prepared using the enriched DNA samples. However, for MeDIP-BS, the bisulfite conversion step was conducted before the library preparation in order to convert unmethylated cytosine to uracil. Finally, we confirmed that in the two MeDIP-BS libraries, the DNA fragment size distribution was suitable (150-500 bp) and the amount of DNA was sufficient to perform

massively parallel sequencing (Figure 11).



Sample	From[bp]	To[bp]	Corr. Area	% of Total	Average Size[bp]	Size Distribution in CV[%]	Conc. [pg/μL]	Molarity [pmol/L]
L library (1/10 dil.)	150	600	468.9	88	297	27.7	605.78	3407.7
S library (1/10 dil.)	150	600	376.3	96	276	22.9	533.7	3121.4

Figure 11. The quality of the prepared libraries. The quality of the prepared libraries was determined to be acceptable for sequencing.

8. Data analysis

All cleaned raw data were mapped to the hg19 reference genome. For bisulfite-converted data, Bismark aligner (34) was used, and for MeDIP-seq data, Bowtie2 aligner (35) was used. To estimate the coverages generated by the different methods, the mapping results from the three methods (except Targeted-BS) within each 100-bp binning window of the hg19 reference genome were compared using BEDTools (36). Reads generated by Targeted-BS were mapped to the targeted region.

To compare the methods, we calculated the methylation levels within

each 100-bp bin because of the low resolution of MeDIP-seq. In order to measure 100bp bin-wise methylation level from the MeDIP-seq data, we performed CpG-density normalization process. Since MeDIP-seq utilizes affinity-based methylated read capturing method, the possibility of capturing methylated reads is increase when the number of CpG sites increase. We counted number of CpG site within each 100bp bins and based on the measured CpG density count, each bin wise read counts is normalized to obtain CpG density normalized methylation levels. For bisulfite-treated data, each CpG site methylation level was measured using the methylation extractor embedded in Bismark and was assigned to 100-bp bins. Since read depths generated by the enrichment-based MeDIP-seq method represented methylation intensities, methylation levels were directly calculated from the read counts of the 100-bp bins. First, Spearman's rank correlation coefficient was used to compare the whole 100-bp bin methylation level of each pair of methods. Next, for the purpose of comparison among the four methods, we sorted all bins by ordering them according to methylation level in decreasing order. Because MeDIP-seq favours capturing highly methylated regions, the four methods should be compared using bins with equivalent or similar average methylation levels. Therefore, matched bins with equivalent or similar average methylation levels used for comparison between different methods may

occupy the same genomic region or different genomic regions. The concordance rate was calculated as follows: (the number of highly methylated bins in WG-BS and the other method together / the number of highly methylated bins in WG-BS) x 100 (%). A comparison of methods was performed for the top 10%, 20%, and 30% of the highly methylated bins.

Results

1. Comparison of mapping rates and read depth distributions

The sequencing reads generated by all four methods (WG-BS, Targeted-BS, MeDIP-seq, and MeDIP-BS) were mapped to the human reference sequence (hg19). As expected, WG-BS produced the highest number (1.5-1.6 billion) of raw reads, while MeDIP-BS produced the lowest number (60-80 million) of raw reads. This gap in data amount between two methods leads to a very big difference in the costs of performing NGS and analysing the data. However, the mapping rates of the WG-BS and MeDIP-BS reads were similar, ranging from ~60-70%. Due to uncertainty whether sequenced thymine is originated from existing thymine or unmethylated cytosine, the MeDIP-BS mapping rate seemed to be slightly lower than the MeDIP-seq mapping rate (Table 2 and Table 3). On the other hand, Targeted-BS showed a relatively high unique mapping rate (over 80%), which appeared to be attributable to reads being mapped only to targeted regions.

Parameter	Reads	Methods							
		WG-BS		MeDIP-seq		MeDIP-BS		Targeted-BS	
Mapping results	Raw reads	1,590,061,196		114,285,714		77,912,500		266,997,022	
	Mapped reads (unique)	974,856,134		83,647,537		51,469,294		224,672,690	
	Mapping rate (unique)	61.3%		73.2%		66.1%		84.2%	
Coverage distribution (WholeGenome)	Read depth	Bin No.	%	Bin No.	%	Bin No.	%	Bin No.	%
	No read depth	2,861,636	9.24	10,672,564	34.5	11,397,943	36.8	29,436,535	95.1
	1X ≤ read depth ≤ 5X	1,277,358	4.13	13,803,694	44.6	11,533,838	37.3	106,327	0.34
	6X ≤ read depth ≤ 10X	1,571,895	5.08	3,126,632	10.1	4,072,743	13.2	70,341	0.23
	11X ≤ read depth ≤ 25X	5,306,270	17.1	2,728,993	8.82	3,093,140	10.0	145,726	0.47
	26X ≤ read depth	19,939,626	64.4	624,902	2.02	859,121	2.78	1,197,856	3.87
	Total	30,956,785	100	30,956,785	100	30,956,785	100	30,956,785	100
Coverage distribution (AllCpGsites)	No read depth	737,022	2.61	10,857,830	38.5	8,551,983	30.3	24,320,805	86.2
	1X ≤ read depth ≤ 5X	1,858,315	6.59	13,257,839	47.0	12,577,456	44.6	479,714	1.70
	6X ≤ read depth ≤ 10X	2,406,829	8.53	2,859,784	10.1	4,004,631	14.2	300,070	1.06
	11X ≤ read depth ≤ 25X	9,271,606	32.9	1,179,980	4.18	2,582,990	9.15	615,201	2.18
	26X ≤ read depth	13,943,237	49.4	61,576	0.22	499,949	1.77	2,501,219	8.86
	Total	28,217,009	100	28,217,009	100	28,217,009	100	28,217,009	100
Coverage distribution (On-targetregion)	No read depth							-	0.00
	1X ≤ read depth ≤ 5X							10,570	1.22
	6X ≤ read depth ≤ 10X							11,243	1.30
	11X ≤ read depth ≤ 25X							36,688	4.24
	26X ≤ read depth							806,450	93.2
	Total							864,956	100

Table 2. Summary of mapping results and read depth distributions of sample L

Parameter	Reads	Methods							
		WG-BS		MeDIP-seq		MeDIP-BS		Targeted-BS	
Mapping results	Raw reads	1,519,126,682		114,285,714		63,904,099		227,218,774	
	Mapped reads (unique)	960,824,730		79,520,177		42,760,690		194,001,908	
	Mapping rate (unique)	63.3%		69.6%		66.9%		85.4%	
Coverage distribution (WholeGenome)	Read depth	Bin No.	%	Bin No.	%	Bin No.	%	Bin No.	%
	No read depth	3,011,078	9.73	11,335,942	36.6	11,899,389	38.4	29,438,943	95.1
	1X ≤ read depth ≤ 5X	1,236,469	3.99	13,869,451	44.8	11,244,101	36.3	106,205	0.34
	6X ≤ read depth ≤ 10X	1,522,329	4.92	2,771,554	8.95	3,984,620	12.9	70,351	0.23
	11X ≤ read depth ≤ 25X	5,206,358	16.8	2,361,753	7.63	2,987,700	9.65	145,598	0.47
	26X ≤ read depth	19,980,551	64.5	618,085	2.00	840,975	2.72	1,195,688	3.86
	Total	30,956,785	100	30,956,785	100	30,956,785	100	30,956,785	100
Coverage distribution (AllCpGsites)	No read depth	821,682	2.91	10,560,673	37.4	9,208,184	32.6	24,328,769	86.2
	1X ≤ read depth ≤ 5X	1,732,836	6.14	13,121,339	46.5	12,515,939	44.4	472,824	1.68
	6X ≤ read depth ≤ 10X	2,235,270	7.92	3,043,967	10.8	3,776,300	13.4	296,725	1.05
	11X ≤ read depth ≤ 25X	9,227,767	32.7	1,406,945	4.99	2,264,056	8.02	612,451	2.17
	26X ≤ read depth	14,199,454	50.3	84,085	0.30	452,530	1.60	2,506,240	8.88
	Total	28,217,009	100	28,217,009	100	28,217,009	100	28,217,009	100
Coverage distribution (On-targetregion)	No read depth							-	0.00
	1X ≤ read depth ≤ 5X							10,324	1.20
	6X ≤ read depth ≤ 10X							11,031	1.28
	11X ≤ read depth ≤ 25X							36,214	4.20
	26X ≤ read depth							803,810	93.3
	Total							861,381	100

Table 3. Summary of mapping results and read depth distributions of sample S

The coverages of the four methods were calculated within windows binned over the whole genome, all CpG sites, and target regions (in this case, only for Targeted-BS). The results showed that WG-BS reads covered more than 90% of the whole genome, and 64% of the whole genome was mapped with over 26-fold read depth. The MeDIP-seq and MeDIP-BS results showed lower coverage ranges than WG-BS, and the overall read depth distribution of MeDIP-BS was similar to that of MeDIP-seq, suggesting that the methylated DNA immunoprecipitation was properly performed and that these two methods sequenced only methylated DNA regions (Table 2-3 and Figure 12).

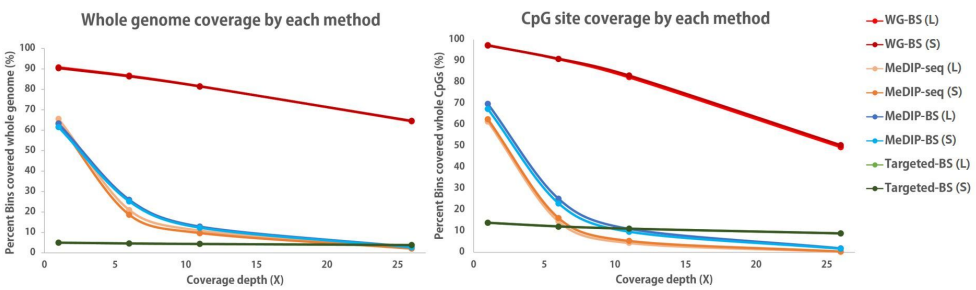


Figure 12. Read depth distribution for each method in whole-genome bins or CpG site-containing bins. WG-BS covered more than 80% of the genome and Targeted-BS covered approximately 3%. MeDIP-BS and MeDIP-seq showed similar read depth distribution.

Using the DNA methylation level distribution data from the bins in WG-BS and calculating the methylation level distribution among bins

belonging to each interval (no read depth, 1- to 5-fold read depth, 6- to 10-fold read depth, 11- to 25-fold read depth, and over 25-fold read depth, respectively) of the read depth distribution for sample L by MeDIP-BS, we found that MeDIP-BS did not capture 14.7% of the high methylation-level bins (methylation level: 70-100%), ~8% of intermediate methylation-level bins (methylation level: 30-70%) and ~6% of low methylation-level bins (methylation level: 0-30%; Figure 13). The tendency was also observed for the results of MeDIP-seq (Figure 14) and the S sample (Figure 15 and Figure 16). This deficiency may be fixed by developing new antibodies with better affinity to 5-methyl cytosine and thereby improving the efficiency of the immunoprecipitation step in this research field in the future. Despite this deficiency, the percentage of high methylation-level bins captured for each read depth interval gradually increased according to read depth. Thus, the high methylation-level bins constituted a major portion of the bins for the high read depth intervals (11- to 25-fold and over 26-fold; Figure 13-16). This indicates that the MeDIP method worked well in this study.

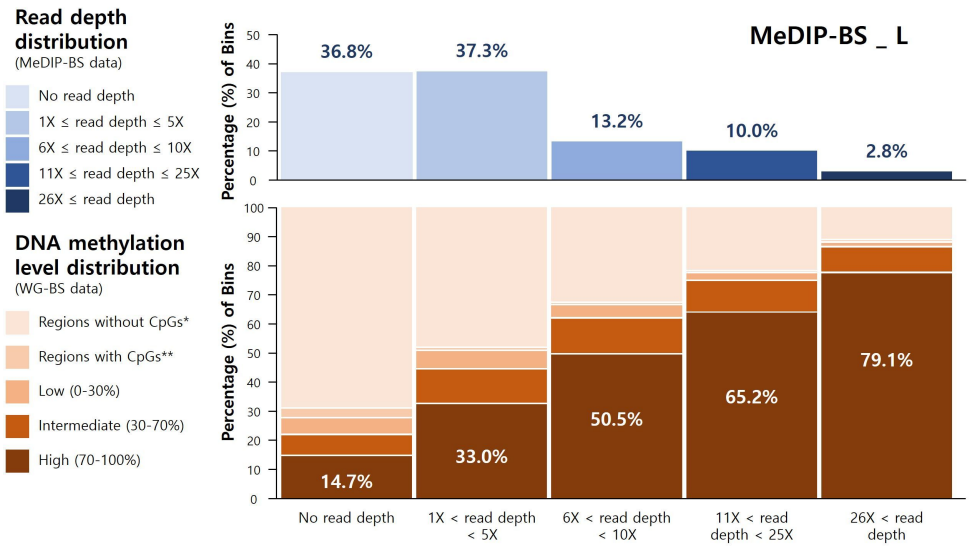


Figure 13. Correlation between DNA methylation levels and read depth of MeDIP-BS (L). DNA methylation level distribution of WG-BS according to the read depth distribution of MeDIP-BS in the L data set.

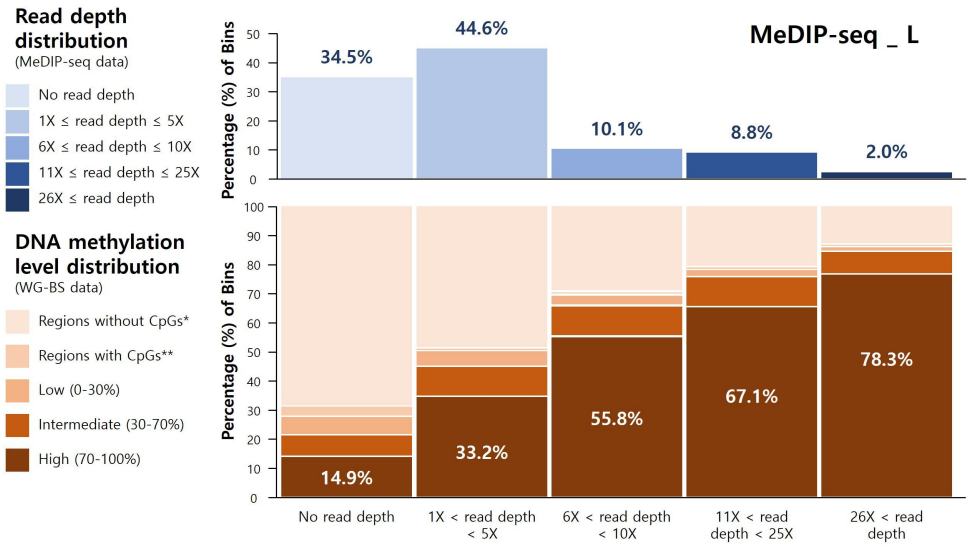


Figure 14. Correlation between DNA methylation levels and read depth of MeDIP-seq (L). DNA methylation level distribution of WG-BS according to the read depth distribution of MeDIP-seq in the L data set.

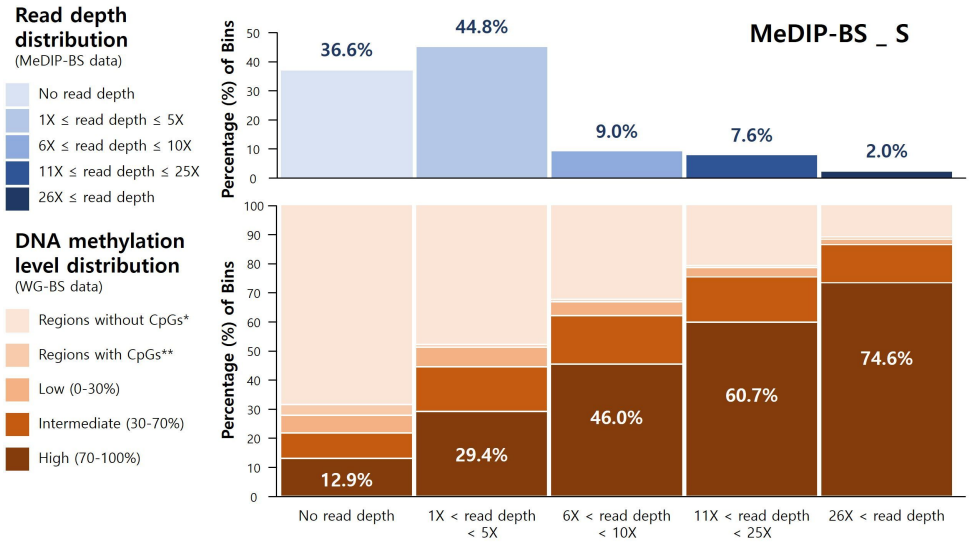


Figure 15. Correlation between DNA methylation levels and read depth of MeDIP-BS (S). DNA methylation level distribution of WG-BS according to the read depth distribution of MeDIP-BS in the S data set.

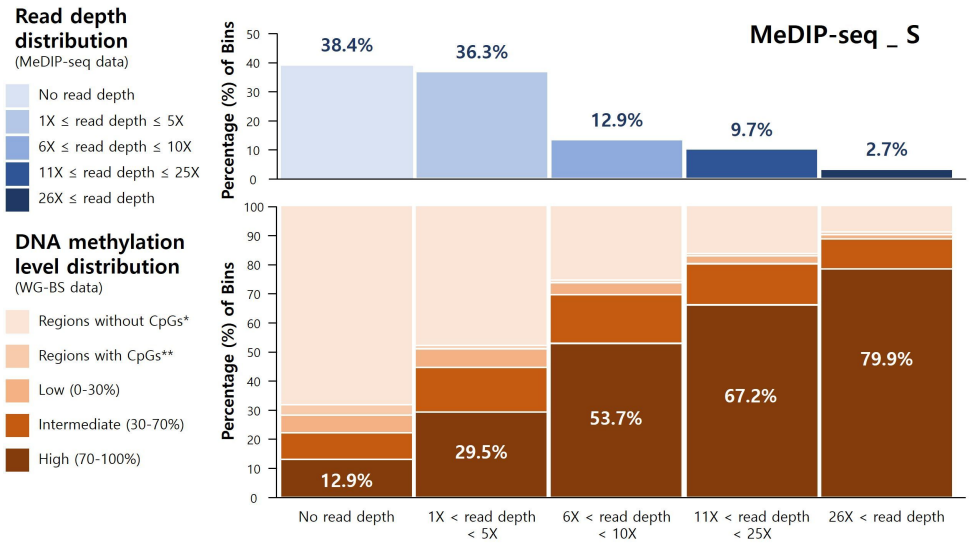


Figure 16. Correlation between DNA methylation levels and read depth of MeDIP-seq (S). DNA methylation level distribution of WG-BS according to the read depth distribution of MeDIP-seq in the S data set.

2. Comparison of the genome-wide DNA methylation levels measured by different methods

It is difficult to directly compare the DNA methylation level measured by MeDIP-seq with the levels measured by other bisulfite conversion-based methods which analyze the DNA methylation level at each CpG site, because the MeDIP-seq method cannot analyze the DNA methylation level at a single CpG site. To overcome this, we divided the whole human genome sequence into 100-bp bins and calculated the DNA methylation level within each bin. For MeDIP-seq, the read depth of each bin directly reflects its DNA methylation level. In this study, 5-methyl CpGs containing bins analyzed by WG-BS, Targeted-BS and MeDIP-BS covered ~50%, ~2.8%, and over 30% of the whole genome, respectively (Table 4). In addition, we ranked each bin based on the methylation level measured by each method, and compared the top 10% most highly methylated bins among the four methods. In doing so, we used the WG-BS result as a standard for assessing the other three methods.

Method	Sample	Bin methylation level	Unmapped regions		1X ≤ read depth ≤ 5X		6X ≤ read depth ≤ 10X		11X ≤ read depth ≤ 25X		26X ≤ read depth	
			Bin No.	%	Bin No.	%	Bin No.	%	Bin No.	%	Bin No.	%
MeDIP-seq	L	Regions without CpGs*	7,419,502	69.5	6,807,934	49.3	916,638	29.3	572,304	21.0	79,744	12.8
		Regions with CpGs**	278,374	2.61	74,593	0.54	10,036	0.32	7,678	0.28	1,722	0.28
		Low (0% ~ 30%)	625,003	5.86	755,441	5.47	100,441	3.21	56,388	2.07	7,585	1.21
		Intermediate (30% ~ 70%)	761,087	7.13	1,589,744	11.5	356,359	11.4	260,247	9.54	46,834	7.49
		High (70% ~ 100%)	1,588,598	14.9	4,575,982	33.2	1,743,158	55.8	1,832,376	67.1	489,017	78.3
		Total	10,672,564	100	13,803,694	100	3,126,632	100	2,728,993	100	624,902	100
	S	Regions without CpGs*	7,878,485	69.7	6,774,182	48.8	714,125	25.8	388,181	16.4	54,839	8.87
		Regions with CpGs**	338,385	2.99	74,589	0.54	8,679	0.31	5,296	0.22	1,690	0.27
		Low (0% ~ 30%)	647,318	5.72	837,080	6.04	101,985	3.68	51,306	2.17	6,373	1.03
		Intermediate (30% ~ 70%)	991,496	8.77	2,092,451	15.1	459,663	16.6	328,977	13.9	61,503	10.0
		High (70% ~ 100%)	1,453,257	12.9	4,091,199	29.5	1,487,102	53.7	1,587,993	67.2	493,680	79.9
		Total	11,308,941	100	13,869,451	100	2,771,554	100	2,361,753	100	618,085	100
MeDIP-BS	L	Regions without CpGs*	8,025,311	70.4	5,661,895	49.1	1,352,379	33.2	661,128	21.4	95,409	11.1
		Regions with CpGs**	317,293	2.78	47,349	0.41	5,320	0.13	2,185	0.07	256	0.03
		Low (0% ~ 30%)	604,260	5.30	682,886	5.92	165,919	4.07	80,274	2.60	11,519	1.34
		Intermediate (30% ~ 70%)	779,222	6.84	1,338,240	11.6	492,401	12.1	331,767	10.7	72,641	8.46
		High (70% ~ 100%)	1,671,857	14.7	3,803,468	33.0	2,056,724	50.5	2,017,786	65.2	679,296	79.1
		Total	11,397,943	100	11,533,838	100	4,072,743	100	3,093,140	100	859,121	100
	S	Regions without CpGs*	8,303,619	69.8	5,475,444	48.7	1,311,770	32.9	626,161	21.0	92,818	11.0
		Regions with CpGs**	378,566	3.18	43,093	0.38	4,808	0.12	1,936	0.06	236	0.03
		Low (0% ~ 30%)	678,724	5.70	718,247	6.39	177,277	4.45	84,235	2.82	12,529	1.49
		Intermediate (30% ~ 70%)	1,000,817	8.41	1,705,107	15.2	658,826	16.5	461,249	15.4	108,091	12.9
		High (70% ~ 100%)	1,537,663	12.9	3,302,210	29.4	1,831,939	46.0	1,814,119	60.7	627,301	74.6
		Total	11,899,389	100	11,244,101	100	3,984,620	100	2,987,700	100	840,975	100

Table 4. Bin methylation level distribution of MeDIP-seq and MeDIP-BS according to read depth distribution

Spearman correlation analysis of the methylation levels of all bins measured by different methods showed that among WG-BS, MeDIP-BS and Targeted-BS, the correlations were significantly high (r = 0.86 in L and r = 0.85 in S between WG-BS and Targeted-BS, r = 0.77 in L and r = 0.76 in S between MeDIP-BS and WG-BS, and r = 0.74 in L and r = 0.68 in S between MeDIP-BS and Targeted-BS; Figure 17 and Figure 18). In contrast, the correlations between MeDIP-seq and the other three methods were lower (r = 0.40-0.53).

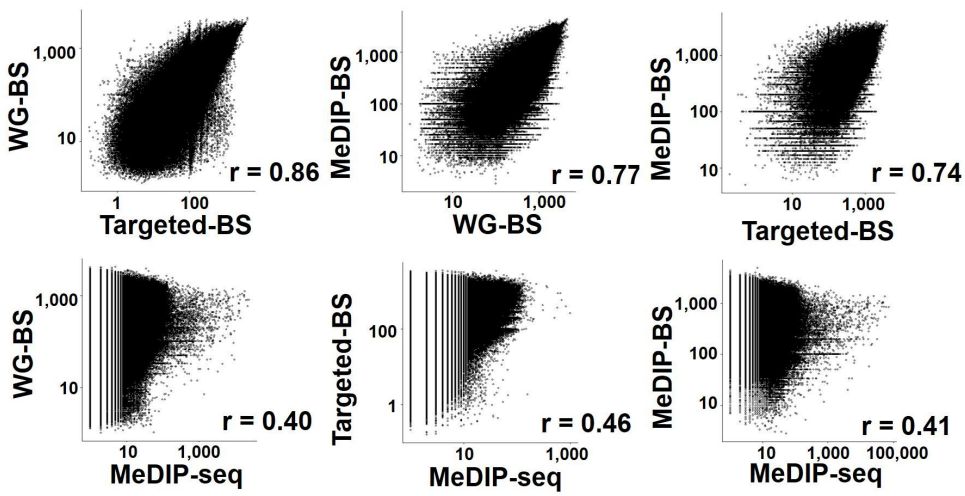


Figure 17. Comparison of DNA methylation levels among the four different methods (L). Spearman correlation analysis for L samples. The three bisulfite conversion-based methods showed relatively high correlations; however, MeDIP-seq showed very low correlation to the other methods.

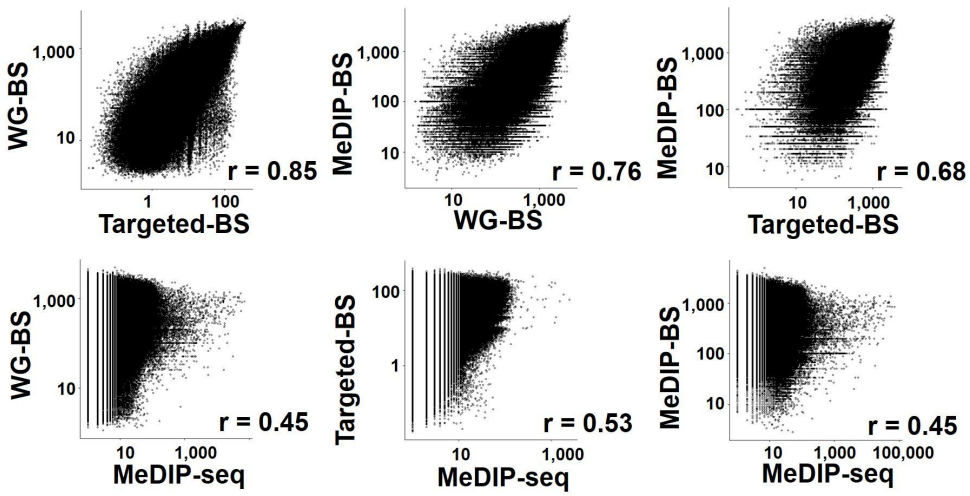


Figure 18. Comparison of DNA methylation levels among the four different methods (S). Spearman correlation analysis for S samples. The three bisulfite conversion-based methods showed relatively high correlations; however, MeDIP-seq showed very low correlation to the other methods.

In order to gain deeper insights into the low correlations between MeDIP-seq and the other three methods, we compared the bins with highest methylation levels among the four methods. In our analysis of concordance for the top 10% most highly methylated bins from each of the three methods (the L sample) compared to WG-BS, the method showing the best match was MeDIP-BS (70.2%). The concordance rate for Targeted-BS was 64.8%, and the concordance rate for MeDIP-seq was 35.3% (Figure 19).

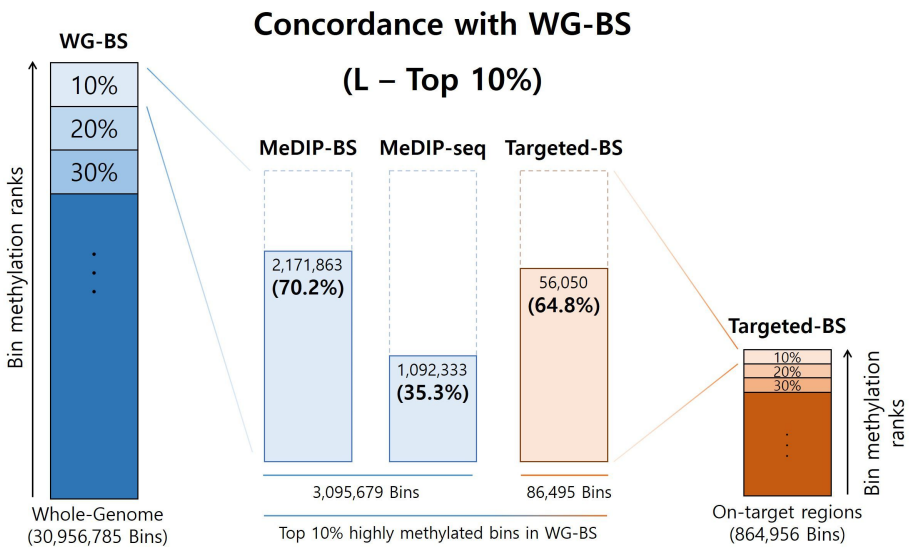
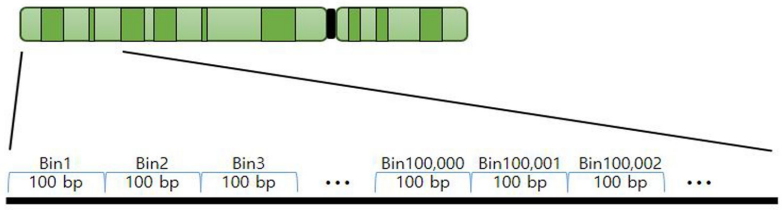


Figure 19. Concordance rate of three methods with WG-BS in the top 10% most highly methylated bins (L). Concordance with WG-BS. The top 10% most highly methylated bins in the L sample data for each method were compared. MeDIP-BS showed the highest concordance.

We also compared the top 10%, 20% and 30% most highly methylated bins from each method (L and S samples) and found that this tendency was maintained, and the concordance rates for MeDIP-BS and Targeted-BS were similar for the top 20% and 30% bins (Figure 20).

(A)



(B)

Methods	Sample	Bin No. *	% in Whole Genome
WG-BS	L	14,733,295	47.5
	S	14,804,282	47.8
Targeted-BS	L	861,381	2.78
	S	864,956	2.79
MeDIP-BS	L	10,044,341	32.4
	S	9,828,117	31.7
MeDIP-seq **	L	30,956,785	100
	S	30,956,785	100

(C)

	Top (%)	Top (Bin No.)	WG-BS-seq vs MeDIP-BS-seq		WG-BS-seq vs MeDIP-seq		Top (Bin No.)	WG-BS-seq vs Targeted-BS-seq	
			Bin No.	%	Bin No.	%		Bin No. ***	%
L	10	3,095,679	2,171,863	70.2	1,092,333	35.3	86,495	56,050	64.8
	20	6,191,357	4,444,492	71.8	2,894,483	46.8	172,991	122,885	71.0
	30	9,287,036	6,901,122	74.3	4,947,682	53.3	259,487	197,809	76.2
S	10	3,095,679	2,127,211	68.7	1,314,108	42.4	86,138	55,189	64.1
	20	6,191,357	4,318,193	69.7	3,113,074	50.3	172,276	121,039	70.3
	30	9,287,036	6,821,815	73.5	5,122,816	55.2	258,414	195,987	75.8

Figure 20. Quantification of bin DNA methylation levels for the four different methods. (A) The whole genome sequence was divided into 100-bp bins to calculate the DNA methylation level. (B) Approximately half of the whole genome was used for DNA methylation level calculation for WG-BS. The MeDIP-BS data covered more than 30% of the genome. (C) The accuracy of each method was determined by using the WG-BS results as the standard and comparing the top 10%, 20%, and 30% most methylated bins in each method to those of WG-BS. MeDIP-BS showed the highest match.

These results showed that the addition of a bisulfite conversion step to the MeDIP-seq method dramatically improved the quality of the genome-wide DNA methylation analysis and explained why the correlations between the methylation levels measured by MeDIP-seq and those measured by the other three methods are very low (Figure 17 and Figure 18).

To visualize the methylation analysis resolution measured by the four methods, we analyzed the methylation levels at CpG sites in the densely methylated promoter regions of *NIFK* (Figure 21), *PERM1* (Figure 22) and *AGRN* (Figure 23), and in the less densely methylated promoter region of *KCNIP2* (Figure 24). The methylation levels at the CpG sites in the highly and less densely methylated promoter regions were very similar in WG-BS, Targeted-BS, and MeDIP-BS, whereas the methylation levels measured by MeDIP-seq were dissimilar, and the resolution was low. In addition, Targeted-BS could not measure the methylation level in the promoter region of *AGRN*, as it is a non-targeted region, demonstrating its limitations for methylation analysis compared to the non-regionally limited analyzes of WG-BS, MeDIP-seq, and MeDIP-BS (Figure 21-24).

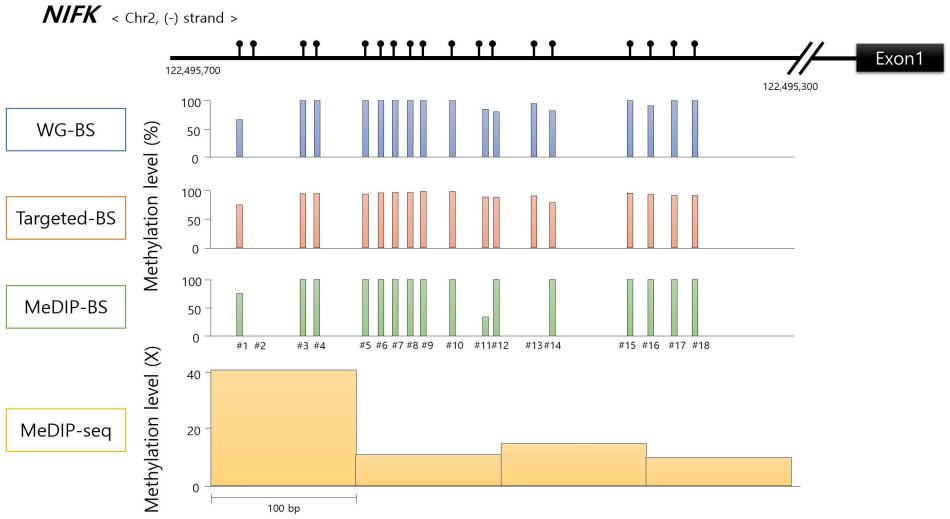


Figure 21. Methylation levels at CpG sites in promoter regions for each analysis methods (*NIFK*). The methylation level of each CpG in *NIFK* promoter region in the L data set was calculated and the read depth distribution was aligned at 100-bp intervals in MeDIP-seq.

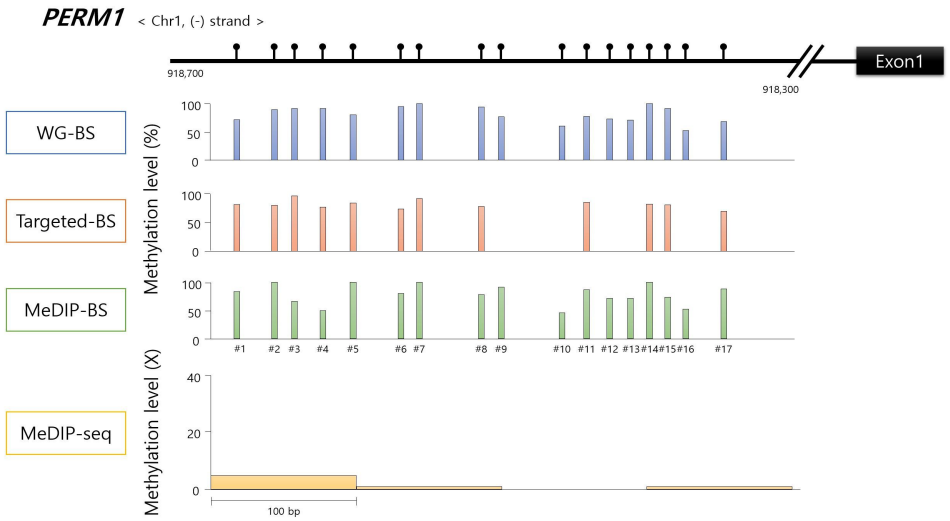


Figure 22. Methylation levels at CpG sites in promoter regions for each analysis methods (*PERM1*). The methylation level of each CpG in *PERM1* promoter region in the L data set was calculated and the read depth distribution was aligned at 100-bp intervals in MeDIP-seq.

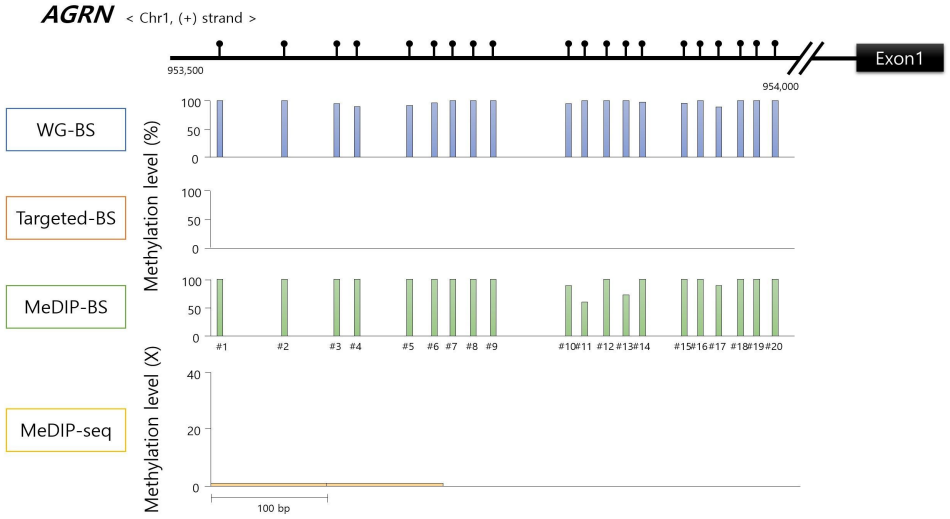


Figure 23. Methylation levels at CpG sites in promoter regions for each analysis methods (*AGRN*). The methylation level of each CpG in *AGRN* promoter region in the L data set was calculated and the read depth distribution was aligned at 100-bp intervals in MeDIP-seq.

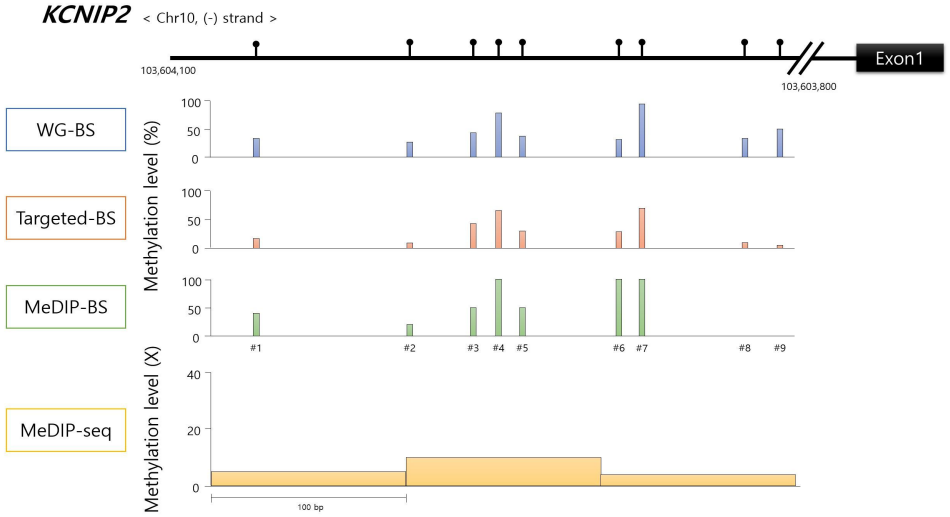


Figure 24. Methylation levels at CpG sites in promoter regions for each analysis methods (*KCNIP2*). The methylation level of each CpG in *KCNIP2* promoter region in the L data set was calculated and the read depth distribution was aligned at 100-bp intervals in MeDIP-seq.

Discussion

Epigenetic mechanisms regulate the expression of essential genes involved in diverse biological processes such as cell development, differentiation, and tissue-specific phenotyping. In particular, DNA methylation is closely associated with gene expression regulation, and it is well-known that aberrant DNA methylation may contribute to the occurrence of various diseases such as heart disease (37), Alzheimer's disease (38), and cancer (39, 40). The development of NGS technology has accelerated genetics and epigenetics research, concomitantly advancing the development of diverse DNA methylation analysis technologies. There are various DNA methylation analysis methods based on NGS, which can be generally categorized into two styles, depending on whether they contain bisulfite conversion or DNA methylation enrichment. Each method has its pros and cons, which should be carefully considered by researchers in selecting the optimal methylation analysis method for their research purposes.

Even though several studies have compared diverse NGS-based DNA methylation analysis methods (26, 29, 41-43), none have compared the MeDIP-BS with other methods. In this study, we demonstrated that MeDIP-BS reduces cost and coverage, and improves analysis resolution

by combining MeDIP with a bisulfite conversion step. Although MBD is also widely used for methylated DNA enrichment, it was reported that both MeDIP and MBD successfully enrich methylated regions showing more than 99% concordance (29, 42). However, MeDIP is more sensitive to the low dense CpG regions and exhibits a slightly more uniform read distribution than MBD (29, 44). We think that an even read distribution is more suitable for calculating CpG site methylation level, so we selected MeDIP for this study. For Targeted-BS method, two kits are generally used, the SureSelect Human Methyl-Seq kit and SeqCap Epi Enrichment kit. Although the SeqCap Enrichment kit covers more CpGs than the SureSelect Human Methyl-Seq kit (28), we selected SureSelect Human Methyl-Seq kit because of its similarity to MeDIP-BS in terms of the experimental procedure. For both MeDIP-BS and the SureSelect Human Methyl-Seq kit, bisulfite conversion is performed after enrichment; therefore, we concluded that the SureSelect Human Methyl-Seq kit is more suitable for comparison to MeDIP-BS. The main features of the four methods in this study are summarized in Table 5. MeDIP and MeDIP-BS require more DNA than the others, but this amount can be decreased by using a recently released library preparation kit, which requires only a small amount of DNA.

Feature	WG-BS	Targeted-BS	MeDIP-seq	MeDIP-BS
Amount of starting DNA	50 ng	3 μg	5 μg	8 μg
Experiment time (Library preparation)	3 hr	2 days (hybridization-16hr)	2 days (hybridization-16hr)	2 days (hybridization-16hr)
Cost/sample	$ 4,500-5,000	$ 2,000	$ 1,500-1,700	$ 1,700-2,000
Coverage	90-91 % (WholeGenome)	2.8% (Targetedregions)	63-65% (Methylatedregions)	61-63% (Methylatedregions)
Resolution	single base	single base	100-200 bp	single base
Strengths	– The most comprehensive method – Single base resolution	– Low cost – Deep analysis in target regions – Single base resolution	– Low cost	– Low cost – Single base resolution
Weaknesses	– High cost – Difficulties in data analysis	– Non-target regions are excluded	– Require a large amount of DNA – Low resolution	– Require a large amount of DNA

Table 5. The features of the four NGS based DNA methylation sequencing methods

In our sequencing effort, we generated over 960 million of WG-BS mapped reads: a sufficient quantity to guarantee the reliability of the WG-BS results, which require more than 800 million aligned reads (45). However, the number of MeDIP-BS mapped reads was approximately 50 million, which was nearly twenty times fewer than the number of WG-BS mapped reads (Table 2 and Table 3), and the cost for library preparation and sequencing were approximately three times cheaper than that for WG-BS (Table 5).

As the cost of data storage and analysis increases drastically as the amount of sequencing data increases (46), the overall cost of performing MeDIP-BS and analysing the data is significantly cheaper than the cost of WG-BS. MeDIP-seq has reduced number of required reads and is free from the weaknesses of bisulfite conversion, such as DNA damage or incomplete conversion. Moreover, the lower mapping rate of MeDIP-BS compared to the MeDIP-seq is attributable to the bisulfite conversion step. Although the lower mapping rate is an issue not only for MeDIP-BS but also for all methylation analysis methods with a bisulfite conversion step, the DNA methylation-level concordance rate between MeDIP-BS and WG-BS was significantly higher than that of MeDIP-seq and WG-BS (Figure 17-18). This suggests that bisulfite conversion after enrichment of methylated DNA fragments with 5-methylcytosine antibodies could dramatically improve the accuracy of

DNA methylation-level analysis.

Although Targeted-BS is the most suitable method in terms of cost and accuracy to meet the requirements of researchers for analysing targeted genomic regions such as promoter regions, it cannot be applied to providing insight into the methylation landscape of non-targeted genomic regions (47-49). In contrast, MeDIP-BS analyzes methylated DNA regions across the whole genome, consequently improving cost effectiveness and guaranteeing accurate genome-wide methylation analysis at base-pair resolution. Although some methylated regions are missing from the MeDIP-BS and MeDIP-seq data, showing that it is not a perfect method (Figure 13-16), the concordance rate of the DNA methylation levels of the bins from MeDIP-BS and WG-BS was significantly higher than that from MeDIP-seq and WG-BS, especially for the top 10%, 20%, and 30% most highly methylated bins (Figure 19 and Figure 20). The lower concordance rate for MeDIP-seq is mainly attributable to its low resolution for analysing DNA methylation levels.

Discovering novel biomarkers for diagnosis or therapy in refractory cancers is an urgent issue in cancer therapy, and it requires genomic/epigenomic alteration data from numerous cancer patients. Furthermore, clinical validation of candidate marker requires even more data than the discovery step. Many global projects for establishing

cancer genomic/epigenomic database are ongoing; however, much of the DNA methylation data were generated by reduced representation methods, such as RRBS or Targeted-BS. For this reason, we strongly recommend the MeDIP-BS method to clinical researchers attempting to analyze genome-wide DNA methylation status with deep read depth at base-pair resolution for numerous clinical samples in a cost-efficient way.

PART Ⅱ

Discovery of genomic alterations in 70 Korean triple negative breast cancer patients using targeted exome sequencing

Introduction

Breast cancer is one of the most prevalent cancers with over 1,300,000 cases and 450,000 deaths each year worldwide (50). Breast cancer is highly heterogeneous disease with diverse pathophysiological and clinical features, which could be caused by distinct underlying genetic, epigenetic and transcriptomic changes. Based on ER (estrogen receptor), PR (progesterone receptor) and HER2 (human epidermal growth factor receptor 2) expression, breast cancer is categorized into at three subtypes based on clinical therapy application; Hormone receptor positive type (ER+ or PR+), HER2 positive type (ER-, PR- and HER2+) and triple-negative breast cancer (TNBC) type (ER-, PR- and HER-) (50, 51). TNBC accounts for approximately 10-20% of invasive breast cancers, and the mortality rates of women with TNBC increase for 5 years after diagnosis (52, 53).

Hormonal therapy and HER2 receptor-targeting therapy can be used for treating hormone receptor positive and HER2 positive breast cancer types, respectively, whereas such therapies are not actionable upon TNBC owing to the lack of the target receptors (ER, PR and HER2) (54, 55). Moreover, there are no conventional targeted therapy for TNBC, suggesting urgent necessity of discovering novel therapeutic

targets. Even though there had been several pioneering studies on genome-wide scale for discovering diagnostic and therapeutic biomarkers in TNBC, the comprehensive effort for identifying target biomarkers for TNBCs in Korean population is still lacking (49-51).

Recently, targeted exome next generation sequencing (NGS), which aims at analysing target exome regions on the cancer genomes by using NGS, had revolutionized human clinical cancer diagnosis, cancer-causing mechanism studies and processes for identifying therapeutic targets due to its cost-effectiveness compared with whole genome or whole exome NGS (56-59). The targeted exome NGS is very advantageous at providing more reliable accuracy of mutation and copy number alteration analysis by generating sufficiently deeper coverage of sequencing reads in target exon regions at a relatively lower cost compared with whole exome NGS (57). In particular, HaloPlex target enrichment system had already been substantiated for its advantageous usefulness in the targeted exome NGS due to its high efficiency upon capturing the targeted regions on the exome (60).

In this study, for the first time to our knowledge, we present mutation and copy number variation landscapes on targeted regions for 368 cancer-associated genes in TNBC genomes from 70 Korean TNBC patients, and provide potential oncogenes and tumor suppressor genes that may be valuable biomarker targets to treat TNBC patients, for

elucidating molecular mechanisms of poor prognosis in TNBC and also for establishing NGS-based TNBC diagnosis gene panel in the future.

Purpose of the study

Breast cancer is one of the most common cancer among women and leading cause of cancer mortality. Breast cancer is highly heterogeneous and categorized into at three subtypes according to clinical therapy application, based on expression pattern of ER (estrogen receptor), PR (progesterone receptor) and HER2 (human epidermal growth factor receptor 2); Hormone receptor positive type (ER+ or PR+), HER2 positive type (ER-, PR- and HER2+) and triple-negative breast cancer (TNBC) type (ER-, PR- and HER-). Unlike other subtypes, TNBC has no actionable target for targeted therapy. That's why TNBC is obstinacy cancer showing low 5-year survival rate compare to other subtypes. So, it is necessary to stratify TNBC and grasp the genetic feature of each types to discover new diagnostic marker or therapeutic target for TNBC. In this study, we analyzed somatic mutation and copy number variation landscapes of 70 Korean TNBC patients for elucidating molecular mechanisms of poor prognosis in TNBC and also for establishing NGS-based TNBC diagnosis gene panel in the future.

Materials and Methods

1. Ethics statement

This study plan for analyzing cancer genomes from Korean TNBC patients was reviewed and approved by Review Board Committee of Samsung Medical Center, Seoul, South Korea. All patients gave written informed consents for donating their tissues for performing this research. Also, this study was carried out in accordance with the Declaration of Helsinki for biomedical research with human subjects.

2. Target gene selection

TNBC associated genes were selected based on following criteria: High mutation frequency in the Catalogue of Somatic Mutations in Cancer database(COSMIC), involved in major signaling pathways, known to predict therapeutic target. Totally, 368 TNBC associated genes were selected, including 5,700 coding exon regions. Total target region size was 961,497bp (Table 6).

No.	Gene	Full name	Coverage	Source
1	ACSL3	acyl-CoA synthetase long-chain family member 3	91.50%	CCDS2455.1
2	AKAP12	Akinase(PRKA)anchorprotein12	96.10%	CCDS5229.1, CCDS5230.1
3	AKAP9	A kinase (PRKA) anchor protein (yotiao) 9	98.10%	CCDS5622.1
4	AKT1	v-akt murine thymoma viral oncogene homolog 1	98.30%	CCDS9994.1
5	AKT2	v-akt murine thymoma viral oncogene homolog 2	97.90%	CCDS12552.1
6	AKT3	v-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)	99.70%	CCDS31076.1, CCDS31077.1
7	ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	91.40%	CCDS9155.1
8	ALK	anaplastic lymphoma kinase (Ki-1)	99.40%	CCDS33172.1
9	AMFR	autocrine motility factor receptor	98.80%	CCDS10758.1
10	APC	adenomatous polyposis of the colon gene	98.60%	CCDS4107.1
11	ARAF	v-raf murine sarcoma 3611 viral oncogene homolog	98.80%	CCDS35232.1
12	ARID1A	AT rich interactive domain 1A (SWI-like)	92.70%	CCDS285.1, CCDS44091.1
13	ARID2	AT rich interactive domain 2	96.60%	CCDS31783.1
14	ASPSCR1	alveolar soft part sarcoma chromosome region, candidate 1	89.20%	CCDS11796.1
15	ATF1	activating transcription factor 1	98.50%	CCDS8803.1
16	ATM	ataxia telangiectasia mutated	98.00%	CCDS31669.1, CCDS31670.1
17	ATR	ataxiatelangiectasiaandRad3related	96.20%	CCDS3124.1
18	ATRX	alpha thalassemia/mental retardation syndrome X-linked	96.40%	CCDS14435.1, CCDS14434.1
19	AURKA	aurora kinase A	97.40%	CCDS13451.1
20	AXIN2	axin 2	96.00%	CCDS11662.1
21	BAP1	BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase)	97.70%	CCDS2853.1
22	BAX	BCL2-associated X protein	95.40%	CCDS12743.1, CCDS12744.1, CCDS12742.1
23	BGN	biglycan	90.80%	CCDS14721.1
24	BIRC7	baculoviral IAP repeat containing 7	100.0%	CCDS13513.1, CCDS13512.1
25	BLM	Bloom Syndrome	97.70%	CCDS10363.1
26	BMPR1A	bone morphogenetic protein receptor, type IA	99.70%	CCDS7378.1
27	BRAF	v-raf murine sarcoma viral oncogene homolog B1	88.70%	CCDS5863.1
28	BRCA1	familial breast/ovarian cancer gene 1	92.40%	CCDS11453.1
29	BRCA2	familial breast/ovarian cancer gene 2	97.10%	CCDS9344.1
30	BRD3	bromodomain containing 3	93.90%	CCDS6980.1
31	BRD4		99.20%	CCDS12328.1, CCDS46004.1
32	BRIP1	BRCA1 interacting protein C-terminal helicase 1	96.50%	CCDS11631.1
33	BUB1B	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)	99.90%	CCDS10053.1
34	C11orf10	chromosome 11 open reading frame 79	99.20%	CCDS8009.1
35	C15orf55	chromosome 15 open reading frame 55	98.90%	CCDS32190.1
36	CANT1	calcium activated nucleotidase 1	95.20%	CCDS11760.1
37	CCDC6	coiled-coil domain containing 6	98.90%	CCDS7257.1

38	CCNB1IP1	cyclin B1 interacting protein 1, E3 ubiquitin protein ligase	98.20%	CCDS9547.1
39	CCND1	cyclin D1	89.50%	CCDS8191.1
40	CCNE1	cyclin E1	98.00%	CCDS12419.1, CCDS46035.1
41	CD74	CD74 molecule, major histocompatibility complex, class Ⅱ invariant chain	97.40%	CCDS47309.1, CCDS47308.1, CCDS34276.1
42	CDC25C	cell division cycle 25 homolog C (S. pombe)	99.30%	CCDS4203.1, CCDS4202.1
43	CDC73	hyperparathyroidism2	97.30%	CCDS1382.1
44	CDH1	cadherin 1, type 1, E-cadherin (epithelial) (ECAD)	97.20%	CCDS10869.1
45	CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)	99.20%	CCDS10803.1
46	CDK12	cyclin-dependent kinase 12	99.10%	NM_016507, NM_015083
47	CDK2	cyclin-dependent kinase 2	95.60%	CCDS8899.1, CCDS8898.1
48	CDK4	cyclin-dependent kinase 4	98.90%	CCDS8953.1
49	CDK8	cyclin-dependent kinase 8	98.70%	CCDS9317.1
50	CDKN1A	cyclin-dependent kinase inhibitor 1A (p21, Cip1)	90.10%	CCDS4824.1
51	CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	98.30%	CCDS8653.1
52	CDKN2A	cyclin-dependent kinase inhibitor 2A (p16(INK4a)) gene	91.80%	CCDS6510.1, CCDS6511.1, CCDS34998.1
53	CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	98.00%	CCDS555.1
54	CHCHD7	coiled-coil-helix-coiled-coil-helix domain containing 7	100.0%	CCDS34895.1, CCDS6166.2, CCDS34896.1
55	CHEK1	checkpoint kinase 1	99.40%	CCDS8459.1
56	CHEK2	CHK2 checkpoint homolog (S. pombe)	90.00%	CCDS33629.1, CCDS13844.1, CCDS13843.1
57	CHN1	chimerin (chimaerin) 1	98.90%	CCDS46454.1, CCDS46455.1
58	CHUK	conserved helix-loop-helix ubiquitous kinase	98.50%	CCDS7488.1
59	CIC	capicua homolog	98.80%	CCDS12601.1
60	CNBP	CCHC-typezincfinger,nucleicacidbindingprotein	100.0%	CCDS46906.1, CCDS46908.1, CCDS46907.1, CCDS3056.1
61	COL1A1	collagen, type I, alpha 1	90.50%	CCDS11561.1
62	COX6C	cytochrome c oxidase subunit VIc	100.00%	CCDS6284.1
63	CREB1	cAMP responsive element binding protein 1	99.00%	CCDS2375.1, CCDS2374.1
64	CREB3L1	cAMP responsive element binding protein 3-like 1	97.90%	NM_052854
65	CREB3L2	cAMP responsive element binding protein 3-like 2	99.00%	CCDS34760.1
66	CRKL	v-crk sarcoma virus CT10 oncogene homolog (avian)-like	99.90%	CCDS13785.1
67	CRTC1	CREB regulated transcription coactivator 1	97.30%	CCDS42525.1, CCDS32963.1
68	CRTC3	CREB regulated transcription coactivator 3	99.50%	CCDS32331.1, CCDS45348.1
69	CSF1R	colony stimulating factor 1 receptor	96.00%	CCDS4302.1

70	CSNK1A1	casein kinase 1, alpha 1	95.80%	CCDS47304.1, CCDS47303.1
71	CSNK1D	casein kinase 1, delta	93.80%	CCDS11805.1, CCDS11806.1
72	CTNNA1	catenin (cadherin-associated protein), alpha 1, 102kDa	98.30%	CCDS34243.1
73	CTNNB1	catenin (cadherin-associated protein), beta 1	98.60%	CCDS2694.1
74	CUX1	cut-like homeobox 1	98.40%	CCDS47672.1, CCDS5720.1, CCDS5721.1
75	CXCR7	chemokine orphan receptor 1	96.60%	CCDS2516.1
76	CYLD	familial cylindromatosis gene	98.10%	CCDS42164.1, CCDS45482.1
77	DAXX	death-domain associated protein	98.20%	CCDS4776.1
78	DDB2	damage-specific DNA binding protein 2	98.60%	CCDS7927.1
79	DDIT3	DNA-damage-inducible transcript 3	98.00%	CCDS8943.1
80	DDR2	discoidin domain receptor tyrosine kinase 2	97.20%	CCDS1241.1
81	DDX5	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	99.50%	CCDS11659.1
82	DICER1	dicer1,ribonucleasetypeIII	98.80%	CCDS9931.1
83	DPYD	dihydropyrimidine dehydrogenase	97.40%	CCDS30777.1
84	DUX4	double homeobox, 4	13.00%	NM_033178
85	EBF1	early B-cell factor 1	89.90%	CCDS4343.1
86	EEF2K	eukaryotic elongation factor-2 kinase	94.90%	CCDS10604.1
87	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	92.60%	CCDS5515.1, CCDS5514.1, CCDS5516.1, CCDS47587.1
88	ELK4	ELK4, ETS-domain protein (SRF accessory protein 1)	91.80%	CCDS1456.1, CCDS1457.1
89	EML4	echinoderm microtubule associated protein like 4	99.50%	CCDS1807.1, CCDS46266.1
90	EP300	300 kd E1A-Binding protein gene	98.70%	CCDS14010.1
91	EPHA3	EPH receptor A3	99.40%	CCDS46875.1, CCDS2922.1
92	EPHA5	EPHreceptorA5	98.60%	CCDS3514.1, CCDS3513.1
93	EPHB1	EPHreceptorB1	96.80%	CCDS46921.1
94	EPHB6	EPHreceptorB6	95.80%	CCDS5873.2
95	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	97.80%	CCDS32642.1, CCDS45667.1
96	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	99.20%	CCDS44918.1, CCDS31833.1
97	ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	98.30%	CCDS42811.1, CCDS2394.1
98	ERC1	ELKS/RAB6-interacting/CAST family member 1	99.80%	CCDS8508.1
99	ERCC2	excision repair cross-complementing rodent repair deficiency, complementation group 2 (xeroderma pigmentosum D)	96.70%	CCDS33049.1, CCDS46112.1
100	ERCC3	excision repair cross-complementing rodent repair deficiency, complementation group 3 (xeroderma pigmentosum group B complementing)	99.10%	CCDS2144.1
101	ERCC4	excision repair cross-complementing rodent repair deficiency, complementation group 4	97.20%	CCDS32390.1

102	ERCC5	excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum, complementation group G (Cockayne syndrome))	96.00%	CCDS32004.1
103	ERG	v-ets erythroblastosis virus E26 oncogene like (avian)	98.20%	CCDS13658.1, CCDS46648.1, CCDS46649.1, CCDS13657.1
104	ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	97.00%	CCDS44767.1, CCDS8475.1
105	ETV1	ets variant gene 1	86.20%	NM_001163150, NM_001163149, NM_004956, NM_001163147, NM_001163148, NM_001163152, NM_001163151
106	ETV4	ets variant gene 4 (E1A enhancer binding protein, E1AF)	96.00%	CCDS11465.1
107	ETV5	ets variant gene 5	98.50%	CCDS33906.1
108	ETV6	ets variant gene 6 (TEL oncogene)	99.00%	CCDS8643.1
109	EWSR1	Ewingsarcomabreakpointregion1(EWS)	88.80%	CCDS13851.1
110	EXT1	multiple exostoses type 1 gene	96.60%	CCDS6324.1
111	EXT2	multiple exostoses type 2 gene	97.70%	CCDS7908.1
112	FAM123B	family with sequence similarity 123B (FAM123B)	98.20%	CCDS14377.2
113	FAS	tumor necrosis factor receptor superfamily, member 6 (FAS)	96.60%	CCDS7398.1, CCDS7394.1, CCDS7393.1, CCDS7395.1
114	FBXW7	F-box and WD-40 domain protein 7 (archipelago homolog, Drosophila)	96.60%	CCDS3778.1, CCDS34078.1, CCDS3777.1
115	FEV	FEV (ETS oncogene family)	85.90%	CCDS2428.1
116	FGFR2	fibroblastgrowthfactorreceptor2	93.00%	CCDS7620.2, CCDS44485.1, CCDS31298.1, CCDS44486.1, CCDS44487.1, CCDS44489.1, CCDS44488.1
117	FGFR3	fibroblast growth factor receptor 3	95.50%	CCDS3353.1, CCDS3354.1
118	FGFR4	fibroblast growth factor receptor 4	97.40%	CCDS4411.1, CCDS4410.1
119	FH	fumarate hydratase	99.50%	CCDS1617.1
120	FHIT	fragile histidine triad gene	100.0%	CCDS2894.1
121	FKBP9	FK506bindingprotein9,63kDa	75.70%	CCDS5439.1
122	FLCN	folliculin, Birt-Hogg-Dube syndrome	98.70%	CCDS32579.1, CCDS32580.1
123	FLI1	Friend leukemia virus integration 1	96.00%	CCDS44768.1
124	FLNA	filamin A, alpha	95.90%	CCDS48194.1, CCDS44021.1
125	FLT1	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	98.40%	CCDS9330.1

126	FLT4	fms-related tyrosine kinase 4	91.10%	CCDS4457.1, CCDS43412.1
127	FOS	FBJmurineosteosarcomaviraloncogenehomolog	100.0%	CCDS9841.1
128	FOXL2	forkhead box L2	99.80%	CCDS3105.1
129	FOXO1	forkhead box O1A	98.70%	CCDS9371.1
130	FRS2	fibroblast growth factor receptor substrate 2	100.0%	CCDS41809.1
131	FUBP1	far upstream element (FUSE) binding protein 1	99.40%	CCDS683.1
132	FUS	fused in sarcoma	81.80%	CCDS10707.1
133	GAB1	GRB2-associated binding protein 1	87.30%	CCDS3760.1, CCDS3759.1
134	GATA3	GATA binding protein 3	99.00%	CCDS7083.1, CCDS31143.1
135	GLI1	GLI family zinc finger 1	95.80%	CCDS8940.1
136	GLI3	GLI family zinc finger 3	96.30%	CCDS5465.1
137	GNA11	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)	99.90%	CCDS12103.1
138	GNAQ	guanine nucleotide binding protein (G protein), q polypeptide	87.70%	CCDS6658.1
139	GNAS	guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide 1	92.90%	CCDS13471.1, CCDS46622.1, CCDS42892.1, CCDS13472.1, CCDS46624.1, CCDS46623.1
140	GOLGA5	golginA5	99.50%	CCDS9905.1
141	GOPC	golgi associated PDZ and coiled-coil motif containing	97.20%	CCDS34523.1, CCDS5117.1
142	GPC3	glypican 3	99.40%	CCDS14638.1
143	GRB10	growth factor receptor-bound protein 10	96.60%	CCDS47586.1, CCDS43583.1, CCDS43582.1
144	GRIN2D	glutamate receptor, ionotropic, N-methyl D-aspartate 2D	92.80%	CCDS12719.1
145	GRM1	glutamate receptor, metabotropic 1	95.00%	CCDS5209.1, CCDS47497.1
146	GSK3B	glycogensynthasekinase3beta	96.00%	CCDS2996.1
147	GUCY1A2	guanylate cyclase 1, soluble, alpha 2	98.50%	CCDS8335.1
148	HDAC4	histone deacetylase 4	92.70%	CCDS2529.1
149	HERPUD1	homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1	99.20%	CCDS32457.1, CCDS45492.1, CCDS10771.1
150	HIST1H1B	histone cluster 1, H1b	100.0%	CCDS4635.1
151	HMGA1	high mobility group AT-hook 1	100.0%	CCDS4788.1, CCDS4789.1
152	HMGA2	high mobility group AT-hook 2	100.0%	CCDS44936.1, CCDS31854.1
153	HNF1A	transcription factor 1, hepatic (HNF1)	98.20%	CCDS9209.1
154	HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	100.0%	CCDS5397.1, CCDS43557.1
155	HOOK3	hook homolog 3	97.80%	CCDS6139.1
156	HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog	98.70%	CCDS7698.1, CCDS7699.1
157	HSP90AA1	heat shock protein 90kDa alpha (cytosolic), class A member 1	95.40%	CCDS9967.1, CCDS32160.1

158	IDH1	isocitrate dehydrogenase 1 (NADP+), soluble	98.00%	CCDS2381.1
159	IDH2	socitratedehydrogenase2(NADP+),mitochondrial	100.0%	CCDS10359.1
160	IGF1	insulin-like growth factor 1	100.0%	CCDS44960.1, CCDS9091.1, CCDS44961.1, CCDS44962.1
161	IGF1R	insulin-like growth factor 1 receptor	94.20%	CCDS10378.1
162	IGFBP3	insulin-like growth factor binding protein 3	94.70%	CCDS5505.1, CCDS34632.1
163	IKBKB	inhibitorofkappalightpolypeptidegeneenhancerinB-cells,kinasebeta	98.50%	CCDS6128.1
164	IKBKE	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase epsilon	98.10%	CCDS30996.1
165	IL6ST	interleukin 6 signal transducer (gp130, oncostatin M receptor)	98.10%	CCDS3971.1, CCDS47209.1
166	IRS4	insulin receptor substrate 4	97.20%	CCDS14544.1
167	JAZF1	juxtaposed with another zinc finger gene 1	96.40%	CCDS5416.1
168	JUN	jun proto-oncogene	98.00%	CCDS610.1
169	KDM5C	lysine(K)-specificdemethylase5C(JARID1C)	98.00%	CCDS14351.1
170	KDM6A	lysine(K)-specificdemethylase6A,UTX	98.30%	CCDS14265.1
171	KDR	vascular endothelial growth factor receptor 2	95.00%	CCDS3497.1
172	KEAP1	kelch-like ECH-associated protein 1	95.40%	CCDS12239.1
173	KIAA1549	KIAA1549	96.00%	CCDS47723.1
174	KIT	v-kitHardy-Zuckerman4felinesarcomaviraloncogenehomolog	97.40%	CCDS3496.1, CCDS47058.1
175	KLF6	core promoter element binding protein (KLF6)	97.60%	CCDS7060.1
176	KLK2	kallikrein-related peptidase 2	98.10%	CCDS12808.1, CCDS42597.1
177	KRAS	v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog	100.00%	CCDS8702.1, CCDS8703.1
178	KTN1	kinectin 1 (kinesin receptor)	97.60%	CCDS41957.1, CCDS41959.1, CCDS9725.1, CCDS41958.1
179	LHFP	lipoma HMGIC fusion partner	100.0%	CCDS9369.1
180	LIFR	leukemia inhibitory factor receptor	96.50%	CCDS3927.1
181	LPP	LIM domain containing preferred translocation partner in lipoma	96.80%	CCDS3291.1
182	MAML2	mastermind-like 2 (Drosophila)	94.00%	CCDS44714.1
183	MAP2K1	mitogen-activated protein kinase kinase 1	96.90%	CCDS10216.1
184	MAP2K2	mitogen-activated protein kinase kinase 2	97.60%	CCDS12120.1
185	MAP2K4	mitogen-activated protein kinase kinase 4	100.0%	CCDS11162.1
186	MAP2K7	mitogen-activated protein kinase kinase 7	98.50%	CCDS42491.1
187	MAP3K1	mitogen-activated protein kinase kinase kinase 1	97.90%	CCDS43318.1
188	MAP3K12	mitogen-activatedproteinkinasekinasekinase12	98.80%	CCDS8860.1
189	MAP3K14	mitogen-activatedproteinkinasekinasekinase14	97.70%	NM_003954
190	MAP3K2	mitogen-activated protein kinase kinase kinase 2	100.0%	CCDS46404.1
191	MAP3K6	mitogen-activated protein kinase kinase kinase 6	92.80%	CCDS299.1
192	MAP4K1	mitogen-activated protein kinase kinase kinase kinase 1	98.00%	CCDS42564.1
193	MAPK11	mitogen-activatedproteinkinase11	94.50%	CCDS14090.1
194	MAPK13	mitogen-activatedproteinkinase13	99.20%	CCDS4818.1

195	MAPK14	mitogen-activated protein kinase 14	94.00%	CCDS4817.1, CCDS4815.1, CCDS4816.1
196	MAPK3	mitogen-activated protein kinase 3	99.60%	CCDS10672.1, CCDS42148.1, CCDS42149.1
197	MAPK8	mitogen-activated protein kinase 8	100.0%	CCDS7226.1, CCDS7225.1, CCDS7223.1, CCDS7224.1
198	MAPK8IP3	mitogen-activated protein kinase 8 interacting protein 3	94.80%	CCDS45379.1, CCDS10442.2
199	MAPK9	mitogen-activated protein kinase 9	98.00%	CCDS4453.1, CCDS4454.1, CCDS43409.1, CCDS43410.1, CCDS47356.1
200	MCL1	myeloid cell leukemia sequence 1 (BCL2-related)	75.10%	CCDS956.1, CCDS957.1
201	MDM2	Mdm2 p53 binding protein homolog	95.60%	CCDS8986.2, CCDS44939.1, CCDS44941.1
202	MDM4	Mdm4 p53 binding protein homolog	98.90%	CCDS1447.1
203	MED12	mediator complex subunit 12	96.10%	CCDS43970.1
204	MEN1	multiple endocrine neoplasia type 1 gene	98.90%	CCDS8083.1, CCDS31600.1
205	MET	met proto-oncogene (hepatocyte growth factor receptor)	100.0%	CCDS43636.1, CCDS47689.1
206	MITF	microphthalmia-associated transcription factor	98.80%	CCDS43106.1, CCDS46864.1, CCDS46865.1, CCDS46866.1, CCDS2913.1, CCDS43107.1
207	MKRN3	makorin ring finger protein 3	97.10%	CCDS10013.1
208	MLH1	E.coli MutL homolog gene	95.90%	CCDS2663.1
209	MLL2	myeloid/lymphoid or mixed-lineage leukemia 2	95.80%	CCDS44873.1
210	MLL3	myeloid/lymphoid or mixed-lineage leukemia 3	93.40%	CCDS5931.1
211	MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	94.10%	CCDS10752.1, CCDS45487.1
212	MN1	meningioma (disrupted in balanced translocation) 1	92.20%	CCDS42998.1
213	MOS	v-mos Moloney murine sarcoma viral oncogene homolog	99.30%	CCDS6164.1
214	MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)	94.50%	CCDS8299.1, CCDS8298.1
215	MSH2	mutS homolog 2 (E. coli)	96.00%	CCDS1834.1
216	MSH6	mutS homolog 6 (E. coli)	97.60%	CCDS1836.1
217	MTOR	mechanistic target of rapamycin (serine/threonine kinase)	97.80%	CCDS127.1
218	MTUS2	microtubule associated tumor suppressor candidate 2	97.00%	NM_001033602, NM_015233
219	MUTYH	mutY homolog (E. coli)	99.90%	CCDS41322.1, CCDS44129.1, CCDS41320.1, CCDS41321.1, CCDS520.1

220	MYB	v-myb myeloblastosis viral oncogene homolog	98.40%	CCDS5174.1, CCDS47482.1, CCDS47481.1
221	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)	96.00%	CCDS6359.2
222	MYCL1	v-myc myelocytomatosis viral oncogene homolog 1, lung carcinoma derived (avian)	99.00%	CCDS30682.1, CCDS44117.1
223	MYCN	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)	86.80%	CCDS1687.1
224	NBN	Nijmegen breakage syndrome 1 (nibrin)	97.80%	CCDS6249.1
225	NCOA1	nuclear receptor coactivator 1	96.00%	CCDS1712.1, CCDS1713.1, CCDS42660.1
226	NCOA4	nuclear receptor coactivator 4 - PTC3 (ELE1)	85.30%	CCDS44394.1, CCDS44393.1, CCDS7237.1
227	NDRG1	N-myc downstream regulated 1	98.00%	CCDS34945.1
228	NF1	neurofibromatosis type 1 gene	96.50%	CCDS45645.1, CCDS42292.1, CCDS11264.1
229	NF2	neurofibromatosis type 2 gene	97.60%	CCDS13865.1, CCDS13863.1, CCDS13861.1, CCDS13862.1, CCDS13864.1
230	NFE2L2	nuclear factor (erythroid-derived 2)-like 2 (NRF2)	99.90%	CCDS42782.1, CCDS46458.1, CCDS46457.1
231	NFIB	nuclear factor I/B	98.30%	CCDS6474.1
232	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	96.70%	CCDS3657.1
233	NKX2-1	NK2 homeobox 1	90.80%	CCDS9659.1, CCDS41945.1
234	NONO	non-POU domain containing, octamer-binding	98.00%	CCDS14410.1
235	NOTCH3	notch 3	90.60%	CCDS12326.1
236	NOTCH4	notch 4	94.60%	CCDS34420.1
237	NR4A3	nuclear receptor subfamily 4, group A, member 3 (NOR1)	98.30%	CCDS6742.1, CCDS6744.1, CCDS6743.1
238	NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	96.80%	CCDS877.1
239	NTRK1	neurotrophic tyrosine kinase, receptor, type 1	96.90%	CCDS30890.1, CCDS1161.1, CCDS30891.1
240	NTRK2	neurotrophic tyrosine kinase, receptor, type 2	98.60%	CCDS35051.1, CCDS35050.1, CCDS6671.1, CCDS35053.1, CCDS35052.1
241	NTRK3	neurotrophic tyrosine kinase, receptor, type 3	97.00%	CCDS32322.1, CCDS10340.1, CCDS32323.1
242	NUTM1	nuclear protien in testis	98.90%	CCDS32190.1
243	OMD	osteomodulin	96.30%	CCDS6696.1
244	PAFAH1B2	pla+B2telet-derived growth factor beta polypeptide (simian sarcoma viral (v-sis) oncogene homolog)	82.20%	CCDS8380.1
245	PALB2	partner and localizer of BRCA2	98.80%	CCDS32406.1

246	PARP1	poly (ADP-ribose) polymerase 1	99.00%	CCDS1554.1
247	PATZ1	POZ (BTB) and AT hook containing zinc finger 1	91.70%	CCDS13895.1, CCDS13894.1, CCDS46691.1, CCDS13896.1
248	PAX3	pairedboxgene3	98.50%	CCDS2448.1, CCDS2449.1, CCDS2450.1, CCDS46522.1, CCDS42825.1, CCDS42826.1, CCDS2451.1, CCDS46523.1
249	PAX7	paired box gene 7	98.20%	CCDS44074.1, CCDS186.1, CCDS44075.1
250	PAX8	paired box gene 8	98.40%	CCDS46397.1, CCDS46398.1, CCDS42736.1, CCDS42735.1, CCDS46399.1
251	PBRM1	polybromo 1	96.60%	CCDS2860.1, CCDS2859.1, CCDS43099.1
252	PBX1	pre-B-cell leukemia homeobox 1	81.00%	CCDS1246.1
253	PCM1	pericentriolarmaterial1 (PTC4)	96.00%	CCDS47812.1
254	PDGFA	platelet-derived growth factor alpha polypeptide	93.80%	CCDS47524.1, CCDS34578.1
255	PDGFRA	platelet-derived growth factor, alpha-receptor	99.40%	CCDS3495.1
256	PDPK1	3-phosphoinositide dependent protein kinase-1	44.70%	CCDS10472.1, CCDS10473.1
257	PHOX2B	paired-like homeobox 2b	95.00%	CCDS3463.1
258	PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	93.40%	CCDS43171.1
259	PIK3CG	phosphoinositide-3-kinase, catalytic, gamma polypeptide	99.00%	CCDS5739.1
260	PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)	99.90%	CCDS3993.1, CCDS3994.1, CCDS3995.1
261	PLAG1	pleiomorphic adenoma gene 1	98.50%	CCDS6165.1, CCDS47860.1
262	PLCG2	phospholipase C, gamma 2 (phosphatidylinositol-specific)	98.50%	CCDS42204.1
263	PLD2	phospholipase D2	95.40%	CCDS11057.1
264	PLK2	polo-like kinase 2	98.60%	CCDS3974.1
265	PMS1	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	96.90%	CCDS46474.1, CCDS46473.1, CCDS2302.1
266	PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)+B464	75.70%	CCDS5343.1
267	POU5F1	POU class 5 homeobox 1	95.30%	CCDS34391.1, CCDS47398.1
268	PPARG	peroxisome proliferative activated receptor, gamma	98.10%	CCDS2609.1, CCDS2610.2
269	PPP2R1A	protein phosphatase 2, regulatory subunit A, alpha	97.70%	CCDS12849.1
270	PRCC	papillary renal cell carcinoma (translocation-associated)	99.30%	CCDS1158.1, CCDS1157.1

271	PRKAR1A	protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1)	100.0%	CCDS11678.1
272	PRKCA	protein kinase C, alpha	96.70%	CCDS11664.1
273	PRKCB	proteinkinaseC,beta	95.10%	CCDS10619.1, CCDS10618.1
274	PRKD1	proteinkinaseD1	100.00%	CCDS9637.1
275	PRKDC	proteinkinase,DNA-activated,catalyticpolypeptide	97.50%	NM_006904, NM_001081640
276	PRUNE2	prune homolog 2 (Drosophila)	97.10%	CCDS47982.1
277	PTCH1	Homolog of Drosophila Patched gene	92.50%	CCDS43851.1, CCDS6714.1, CCDS47996.1, CCDS47995.1
278	PTEN	phosphatase and tensin homolog gene	97.90%	NM_000314
279	PTK2	PTK2 protein tyrosine kinase 2	92.50%	CCDS6381.1
280	PTK2B	PTK2B protein tyrosine kinase 2 beta	96.20%	CCDS6058.1, CCDS6057.1
281	PTPRD	protein tyrosine phosphatase, receptor type, D	97.50%	CCDS43786.1, CCDS6472.1
282	RAD51L1	RAD51 homolog B (S. cerevisiae)	92.20%	CCDS9789.1
283	RAF1	v-raf-1 murine leukemia viral oncogene homolog 1	94.00%	CCDS2612.1
284	RB1	retinoblastoma gene	97.00%	CCDS31973.1
285	RECQL4	RecQ protein-like 4	91.40%	NM_004260
286	RELA	v-rel reticuloendotheliosis viral oncogene homolog A (avian)	95.90%	CCDS31609.1, CCDS44651.1
287	RET	retproto-oncogene	97.10%	CCDS7200.1
288	RGL1	ral guanine nucleotide dissociation stimulator-like 1	96.80%	CCDS1359.1
289	RICTOR	RPTOR independent companion of MTOR, complex 2	97.70%	CCDS34148.1
290	RIPK1	receptor (TNFRSF)-interacting serine-threonine kinase 1	98.80%	CCDS4482.1
291	RNF213	ringfingerprotein213	95.80%	CCDS32761.1
292	ROCK1	Rho-associated, coiled-coil containing protein kinase 1	89.40%	CCDS11870.2
293	ROS1	v-ros UR2 sarcoma virus oncogene homolog 1 (avian)	97.10%	CCDS5116.1
294	RPS6KA2	ribosomal protein S6 kinase, 90kDa, polypeptide 2	96.70%	CCDS34570.1, CCDS5294.1
295	RPS6KA3	ribosomal protein S6 kinase, 90kDa, polypeptide 3	99.60%	CCDS14197.1
296	RPS6KA4	ribosomal protein S6 kinase, 90kDa, polypeptide 4	88.30%	CCDS8073.1
297	RPS6KB2	ribosomal protein S6 kinase, 70kDa, polypeptide 2	99.70%	CCDS41677.1
298	RPTOR	regulatory associated protein of MTOR, complex 1	96.60%	CCDS11773.1
299	RUNX1T1	runt-related transcription factor 1; translocated to, 1 (cyclin D-related)	97.60%	CCDS47891.1, CCDS6256.1, CCDS6257.1
300	SDHB	succinate dehydrogenase complex, subunit B, iron sulfur (Ip)	100.0%	CCDS176.1
301	SDHC	succinate dehydrogenase complex, subunit C, integral membrane protein, 15kDa	88.30%	CCDS1230.1, CCDS41432.1, CCDS41431.1
302	SDHD	succinate dehydrogenase complex, subunit D, integral membrane protein	87.70%	CCDS31678.1
303	SETD2	SET domain containing 2	96.00%	CCDS2749.2
304	SFPQ	splicing factor proline/glutamine rich(polypyrimidine tract binding protein associated)	97.00%	CCDS388.1

305	SLC29A1	solute carrier family 29 (nucleoside transporters), member 1	96.00%	CCDS4908.1
306	SLC45A3	solute carrier family 45, member 3	98.90%	CCDS1458.1
307	SMAD2	SMAD family member 2	94.20%	CCDS11934.1, CCDS45863.1
308	SMAD3	SMAD family member 3	100.0%	CCDS10222.1, CCDS45288.1
309	SMAD4	Homolog of Drosophila Mothers Against Decapentaplegic 4 gene	99.20%	CCDS11950.1
310	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4	91.60%	CCDS45971.1, CCDS45973.1, CCDS12253.1, CCDS45972.1
311	SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1	99.10%	CCDS13817.1, CCDS46671.1
312	SMO	smoothened homolog (Drosophila)	96.20%	CCDS5811.1
313	SOS1	son of sevenless homolog 1 (Drosophila)	98.50%	CCDS1802.1
314	SOX2	SRY (sex determining region Y)-box 2	93.60%	CCDS3239.1
315	SP1	Sp1 transcription factor	94.40%	CCDS8857.1, CCDS44898.1
316	SRC	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)	95.70%	CCDS13294.1
317	SRGAP3	SLIT-ROBO Rho GTPase activating protein 3	99.40%	CCDS2572.1, CCDS33689.1
318	SS18	synovial sarcoma translocation, chromosome 18	96.70%	CCDS32807.1
319	SS18L1	synovial sarcoma translocation gene on chromosome 18-like 1	95.40%	CCDS13491.1
320	SSX1	synovial sarcoma, X breakpoint 1	93.00%	CCDS14287.1, CCDS14290.1
321	SSX2	synovial sarcoma, X breakpoint 2	8.80%	CCDS14344.1, CCDS14345.1, CCDS48129.1
322	SSX4	synovial sarcoma, X breakpoint 4	39.50%	CCDS35240.1, CCDS43934.1
323	STAT1	signal transducer and activator of transcription 1, 91kDa	99.50%	CCDS2309.1, CCDS42793.1
324	STK11	serine/threonine kinase 11 gene (LKB1)	98.20%	CCDS45896.1
325	STK36	serine/threonine kinase 36	99.70%	CCDS2421.1
326	SUFU	suppressor of fused homolog (Drosophila)	87.00%	CCDS7537.1
327	SUZ12	suppressor of zeste 12 homolog (Drosophila)	88.70%	CCDS11270.1
328	TAF1	TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa	96.30%	CCDS14412.1, CCDS35325.1
329	TAF15	TAF15 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 68kDa	65.40%	CCDS32623.1
330	TBX22	T-box 22	100.0%	CCDS14445.1, CCDS43975.1
331	TCEA1	transcription elongation factor A (SII), 1	97.80%	CCDS47858.1, CCDS47857.1
332	TCF12	transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)	96.90%	CCDS10159.1, CCDS10160.1, CCDS42042.1
333	TCF4	transcription factor 4	94.70%	CCDS11960.1, CCDS42438.1
334	TCF7L2	transcription factor 7-like 2	95.20%	CCDS7576.1

335	TEC	tec protein tyrosine kinase	96.70%	CCDS3481.1
336	TERT	telomerase reverse transcriptase	92.50%	CCDS3861.2, CCDS47186.1
337	TFE3	transcription factor binding to IGHM enhancer 3	98.70%	CCDS14315.3
338	TFEB	transcription factor EB	98.50%	CCDS4858.1
339	TFG	TRK-fused gene	99.60%	CCDS2939.1
340	TGFB3	transforming growth factor, beta 3	97.80%	CCDS9846.1
341	TGFBR1	transforming growth factor, beta receptor 1	99.50%	CCDS47998.1, CCDS6738.1
342	TGFBR2	transforming growth factor, beta receptor II (70/80kDa)	99.40%	CCDS33727.1, CCDS2648.1
343	THOC5	THO complex 5	93.30%	CCDS13859.1
344	THRAP3	thyroid hormone receptor associated protein 3 (TRAP150)	96.60%	CCDS405.1
345	TIAM1	T-cell lymphoma invasion and metastasis 1	96.30%	CCDS13609.1
346	TLN1	talin 1	96.10%	CCDS35009.1
347	TLR4	toll-like receptor 4	97.90%	CCDS6818.1
348	TMPRSS2	transmembrane protease, serine 2	98.20%	CCDS33564.1
349	TP53	tumor protein p53	86.90%	CCDS11118.1, CCDS45605.1, CCDS45606.1
350	TPM3	tropomyosin 3	99.70%	CCDS1060.1, CCDS41400.1, CCDS41401.1, CCDS41402.1, CCDS41403.1
351	TPR	translocatedpromoterregion	98.50%	CCDS41446.1
352	TRIM27	tripartite motif-containing 27	95.30%	CCDS4654.1
353	TRIM33	tripartite motif-containing 33 (PTC7,TIF1G)	95.90%	CCDS873.1, CCDS872.1
354	TRIM62	tripartitemotifcontaining62	98.10%	CCDS376.1
355	TSC1	tuberous sclerosis 1 gene	99.60%	CCDS6956.1
356	TSC2	tuberous sclerosis 2 gene	95.80%	CCDS45384.1, CCDS10458.1
357	TSHR	thyroid stimulating hormone receptor	96.10%	CCDS9872.1, CCDS32131.1
358	UBR5	ubiquitin protein ligase E3 component n-recognin 5	98.20%	CCDS34933.1
359	USP6	ubiquitin specific peptidase 6 (Tre-2 oncogene)	86.00%	CCDS11069.2
360	VHL	von Hippel-Lindau syndrome gene	62.60%	CCDS2598.1, CCDS2597.1
361	VTI1A	vesicle transport through interaction with t-SNAREs homolog 1A	100.0%	CCDS7575.2
362	WIF1	WNT inhibitory factor 1	92.60%	CCDS8971.1
363	WRN	Werner syndrome (RECQL2)	90.90%	CCDS6082.1
364	WT1	Wilms tumour 1 gene	93.70%	CCDS7877.2, CCDS7878.2, CCDS44562.1,CCDS44561.1
365	XPA	xeroderma pigmentosum, complementation group A	98.90%	CCDS6729.1
366	XPC	xeroderma pigmentosum, complementation group C	95.90%	CCDS46763.1, CCDS46764.1
367	ZNF331	zinc finger protein 331	90.80%	CCDS33102.1
368	ZNF668	zinc finger protein 668	97.90%	CCDS10701.1

Table 6. The list of target genes in our study

3. Haloplex target enrichment-based next generation sequencing-ready sample preparation and sequencing

A total of 70 Korean TNBC and matched-normal tissues were collected from Samsung Medical Center, Seoul, South Korea. Specimens were immediately frozen by liquid nitrogen or fixed by formalin to make Formalin Fixed-Paraffin Embedded (FFPE) block. Genomic DNA was extracted from frozen samples using DNeasy Blood&Tissue kit (QIAgen, 69506) according to manufacturer's instruction. After digesting and denaturing genomic DNA, targeted fragment DNAs were hybridized with biotinylated probes designed to guide circularization of target DNA fragments incorporating sequencing motifs. The targeted fragments bound to the biotinylated HaloPlex probes were retrieved with magnetic streptavidin beads. The circularized molecules were then closed by ligation, which ensures that only perfectly hybridized fragments are circularized. Only circular DNA targets were amplified by PCR, providing an enriched and barcoded amplification product, which then was sequenced on Illumina HiSeq 2000 instrument.

4. Immunohistochemistry

Slice section of each FFPE samples were stained with Hematoxylin&Eosin to select tumor region by pathologic validation. ER (estrogen receptor), PR (progesterone receptor), HER2 expression were confirmed by immunohistochemistry in each samples to check whether the sample is TNBC or not.

5. Bioinformatics analysis of SNVs and indels

Paired-end sequence raw reads were trimmed and filtered out to produce clean reads with good base quality (Phred Q score > 20). We used Burrows-Wheeler Alignment (BWA 0.5.9), Genome Analysis Toolkit (GATK) and Samtools in aligning our paired-end sequencing reads to the human reference genome hg19 and calling single nucleotide variants (SNVs) and short sequence insertions and deletions (indels), respectively. The assessment of SNVs and Indels were performed by using dbSNP135, dbNSFP COSMIC and 1000 Genomes variants databases, and software programs SNPEff, SIFT, PolyPhen2, LRT, PhyloP, Mutation_Taster, Mutation_Assessor and FATHMM, GERP_NR. In selecting the somatic non-synonymous SNVs and Indels, we maintained the criteria that the read-allele frequency at the position should be

over 20%, the absolute number of mapped reads at the position should be ≥ 15, and the SNV or Indel allele read count should be zero to the targeted sequence of the corresponding normal tissue. Those variants were confirmed by visualization in the Interactive Genomic Viewer (IGV) program and NextGene (Soft Genetics, Inc.).

6. Bioinformatics analysis of copy number alterations

The genomic copy number alterations between cancer and matched-normal tissues were assessed by using NextGENe (v2.3.1.) software that compares in median read coverage levels between target genomic regions of cancer and matched-normal tissues after global normalization of genome-wide read coverage levels. The copy number alteration was valued in terms of log2 ratio of read coverages between cancer and matched-normal tissues. The copy number alterations, whose log2 ratio value is > 1.5, were considered as amplification status, while copy number alterations showing log2 ratio value < -1.2 were considered as homozygous loss status.

7. Experimental validation of genomic alterations

Two SNV regions in TP53 were selected for experimental validation of somatic mutations. Target regions were amplified from genomic DNA from tumor and matched normal tissues (TNBC030 and TNBC045 patients) by genomic PCR, and the products were sequenced directly or cloned into T vector for Sanger sequencing. Five clones were selected for each sample. CNVs were validated by genomic qPCR, and frequently amplified gene NDRG1 and deleted genes ATM, BRCA1, BRCA2, WRN were selected for CNV validation. Genomic DNA from tumor and matched normal tissues (ATM - TNBC038 and TNBC048 patients; BRCA1 - TNBC026, TNBC031, TNBC038 and TNBC066 patients; BRCA2 - TNBC004, TNBC011, TNBC014 and TNBC068 patients; WRN - TNBC030 patient) were analyzed by genomic qPCR. Relative concentration of each sample was calculated by ddCt method using TERT as reference gene (61, 62). The information of mutated or altered genomic regions and primers used in the validation experiments is described in Table 7.

Somatic variants and CNVs for validation								Primer sequence					
Gene	Chr	NC_#	Chromosomal Position		Mutation	asssay	Patient	Name	Sequence (5' -> 3')	Length (nt)	Tm (℃)	%GC	Amplicon size (nt)
			Start	End									
SNVs													
TP53	chr17	NC_000017.10	7578212	7578212	c.637C>T	Sanger sequencing	TNBC030	TP53_SNV_F	GTT TCT TTG CTG CCG TCT TC	20	54.7	50.0	501
			7578271	7578271	c.578A>G		TNBC045	TP53_SNV_R	CTT AAC CCC TCC TCC CAG AG	20	56.5	60.0	
CNVs													
NDRG1	chr8	NC_000008.10	134276714	134277015	Amplification	Real-time PCR	TNBC022	NDRG1_CNV_F	GCT TCC TCA AAA CAC AGT TGG	21	59.0	48.0	75
								NDRG1_CNV_R	GCT GGT CAT GTG GGG TTC	18	59.0	61.0	
								NDRG1_Probe	FAM - CTT CAG CC - BHQ1	8	20.5	62.5	
WRN	chr8	NC_000008.10	30924477	30924571	Homozygous deletion	Real-time PCR	TNBC030	WRN_CNV_F	CCA GGT CTC TGT GCA TTT CA	20	59.8	50.0	77
								WRN_CNV_R	GGT AAT ACC TGA AAA CAG GAA CTG A	25	59.8	40.0	
								WRN_Probe	FAM - GAA ATG ATG AAA AAG CAA CAC A -BHQ1	22	57.8	31.8	
ATM	chr11	NC_000011.09	108129654	108129761	Homozygous deletion	Real-time PCR	TNBC048	ATM_CNV#1_F	GAA TAA TTG TTT TTA TTT CTT TGT TGC	27	49.4	22.2	100
								ATM_CNV#1_R	TTA ACA ATC GCA GGA AAA AGC	21	51.9	38.1	
								ATM_CNV#1 _probe	FAM - TGT CTT AAT TGC AGA AGA GTC CA - BHQ1	23	54.0	39.1	
ATM	chr11	NC_000011.09	108167887	108168120	Homozygous deletion	Real-time PCR	TNBC038	ATM_CNV#2_F	AAA CAA AAG TGT TGT CTT CAT GC	23	52.7	34.8	150
								ATM_CNV#2_R	GAA CTT CTT TTT CAC CAG TGT GG	23	54.4	43.5	
								ATM_CNV#2 _probe	FAM - TGC AGT TAT CCA AGA TGG CA - BHQ1	20	54.3	45.0	
BRCA1	chr17	NC_000017.11	41256016	41256252	Homozygous deletion	Real-time PCR	TNBC026 TNBC031 TNBC066	BRCA1_CNV#1_F	TTC TAC AGA GTG AAC CCG AAA A	22	53.8	40.9	150
								BRCA1_CNV#1_R	GGC TAA GGC AGG AGG ACT G	19	63.2	57.6	
								BRCA1_CNV#1 _probe	FAM - ATG GAG TCT TGC TCT GTG GC - BHQ1	20	57.4	55.0	
BRCA1	chr17	NC_000017.11	41244444	41246036	Homozygous deletion	Real-time PCR	TNBC038	BRCA1_CNV#2_F	CAG CGA TAC TTT CCC AGA GC	20	55.6	55.0	199
								BRCA1_CNV#2_R	TTG CAA AAC CCT TTC TCC AC	20	53.7	45.0	
								BRCA1_CNV#2 _probe	FAM - TGC TGA AGA CCC CAA AGA TC - BHQ1	20	54.8	50.0	
BRCA2	chr11	NC_000013.11	32915134	32915248	Homozygous deletion	Real-time PCR	TNBC011 TNBC068	BRCA2_CNV#1_F	TCC AAA GAT TCA GAA AAC TAC TTT GA	26	52.9	30.8	110
								BRCA2_CNV#1_R	GAA TGT GTG GCA TGA CTT GG	20	54.3	50.0	
								BRCA2_CNV#1 _probe	FAM - TGG AAG ATG ATG AAC TGA CAG A - BHQ1	22	53.4	40.9	
BRCA2	chr11	NC_000013.11	32929258	32929478	Homozygous deletion	Real-time PCR	TNBC004 TNBC014	BRCA2_CNV#2_F	CAT TGA TGG ACA TGG CTC TG	20	53.7	50.0	168
								BRCA2_CNV#2_R	TGA AAG GCA AAA ATT CAT CAC A	22	51.5	31.8	
								BRCA2_CNV#2 _probe	FAM - CAA AAA CAA CTC CAA TCA AGC A - BHQ1	22	52.3	36.4	
Reference gene													
TERT	chr5	NC_000005.9	1253628	1253725		Real-time PCR		TERT_CNV_F	GGC CTG AGT GAG TGT TTG G	19	59.0	58.0	98
								TERT_CNV_R	TGG ACA CTC AGC CCT TGG	18	60.0	61.0	
								TERT_Probe	FAM - CTT CAG CC - BHQ1	8	20.5	62.5	

Table 7. Genomic regions of somatic variants and CNVs and primer information for validation experiments

8. Protein-protein interaction network and gene expression analysis

We used STRING (Search Tool for the Retrieval of Interacting Genes/Proteins), KEGG (Kyoto Encyclopedia of Genes and Genomes) and DAVID (Database for Annotation, Visualization and Integrated Discovery) for analyzing oncogenic and tumor-suppressive pathways in the TNBC cells. Using copy number alteration, RNA-Seq expression and mutation data of human clinical TNBC samples in TCGA (The Cancer Genome Atlas) database, we also performed the comparative analysis of our Korean TNBC samples.

Results

1. Clinicopathological information of TNBC patients and statistics of targeted exome sequencing

Clinicopathological information of 70 Korean TNBC patients and influence of each features on disease free survival (DFS) or distant metastasis free survival (DMFS) are described in Figure 25. Recurrence were occurred in 15 patients and among them, 8 patients suffered a distant metastasis. Average follow-up period was 4.88 years. The high grade of primary tumor stage (pT) appeared to be associated with the increment of the risk of recurrence (pT2 - recurrence: HR=3.025, p value=0.094, pT2 - distant metastasis: HR=2.819, pvalue=0.207, pT3 - recurrence: HR=6.598, p value=0.108) (Figure 25).

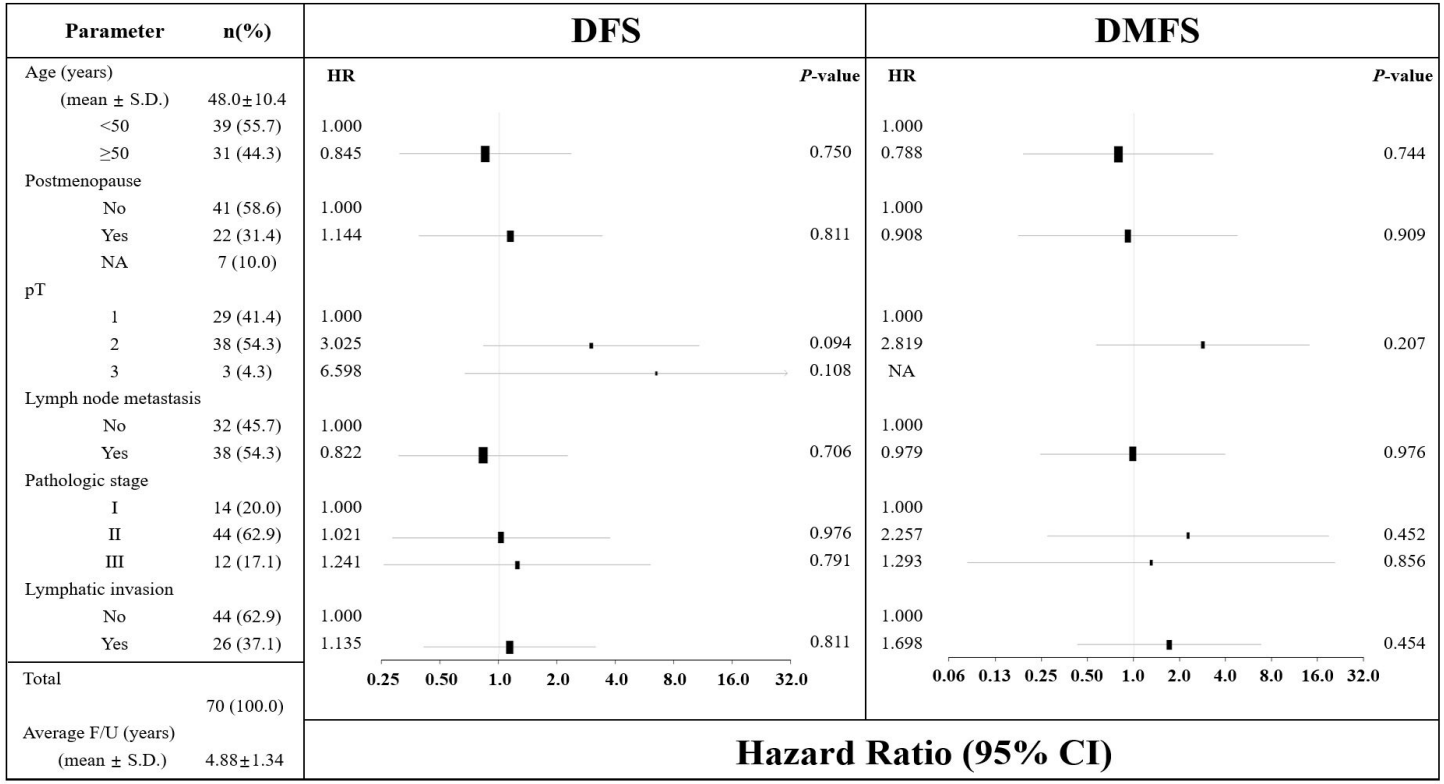


Figure 25. Results of the influence of the clinicopathological features on the DFS or DMFS in 70 Korean TNBC patients. Abbreviations: CI, Confidence interval; DFS, Disease Free Survival; DMFS, Distant Metastasis Free Survival; F/U, Follow-up period; TNBC, Triple Negative Breast Cancer; HR, Hazard ratio.

We performed targeted exome sequencing using tumor lesion samples and paired adjacent normal samples from 70 Korean TNBC patients. Average target coverage depth was 130.36X for tumor samples and 139.71X for normal samples. The target regions, whose read coverage depths are >2X and >100X, account for over 93% and over 40% of the entire target region in those samples, respectively (Table 8).

Target Sequencing Statistics	Tumor sample	Normal sample
Target Territory (bp)	2,364,198	2,364,198
Average Target Coverage (X)	130.4	139.7
% of 1x Target Bases	4.252	4.351
% of Target Bases ≥ 2x	93.35	93.12
% of Target Bases ≥ 10x	86.23	86.69
% of Target Bases ≥ 20x	78.47	79.56
% of Target Bases ≥ 30x	71.67	73.13
% of Target Bases ≥ 40x	65.65	67.36
% of Target Bases ≥ 50x	60.33	62.21
% of Target Bases ≥ 100x	40.64	42.57

Table 8. Average target coverage and percentage of target bases according to read coverage depths

2. Mutational landscape analysis in 70 Korean TNBCs

A total of 292 somatic single nucleotide variants (SNVs) were detected in 70 Korean TNBC patients. Among these, 220 (68%) were

novel SNVs which have not been reported in COMSIC or dbSNP database. Among 30 somatic INDEL mutations identified, 21 were somatic deletions including 11 (4%) novel deletions and 9 were somatic insertions including 7 (2%) novel insertions (Figure 26 and Table 9).

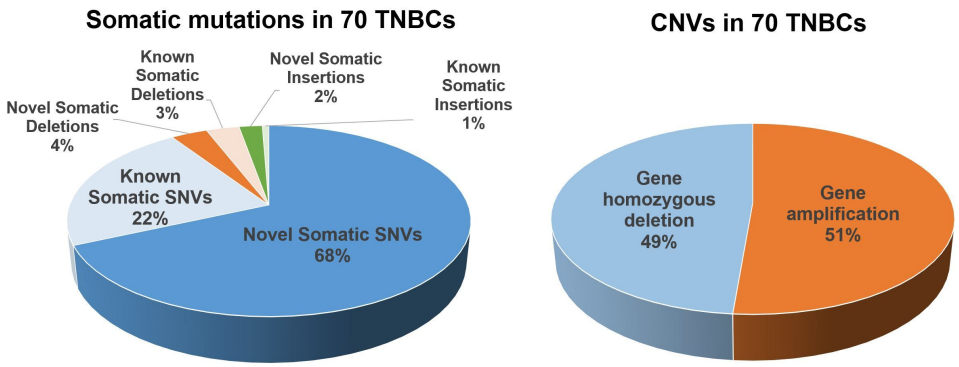


Figure 26. Somatic variants and CNVs in Korean TNBCs genomes. Somatic SNVs accounts for most somatic variants (90%), and gene amplification and gene homozygous deletion comprise almost sample portion of CNVs.

Somatic variants		Mutation Number	Total	Number of mutated Genes
Somatic SNVs	Novel	220	292	157
	COSMIC or dbSNP	72		
Somatic Deletions	Novel	11	21	
	COSMIC or dbSNP	10		
Somatic Insertions	Novel	7	9	
	COSMIC or dbSNP	2		
Copy number alterations genes				Number of altered Genes
Amplification				365
Homozygous deletion				346

Table 9. Summary of somatic variants and CNVs in Korean TNBCs genomes

Those SNVs and INDELs occurred in 157 genes, and the average number of mutated genes per TNBC is 6.44 (range: 0-15) (Figure 27).

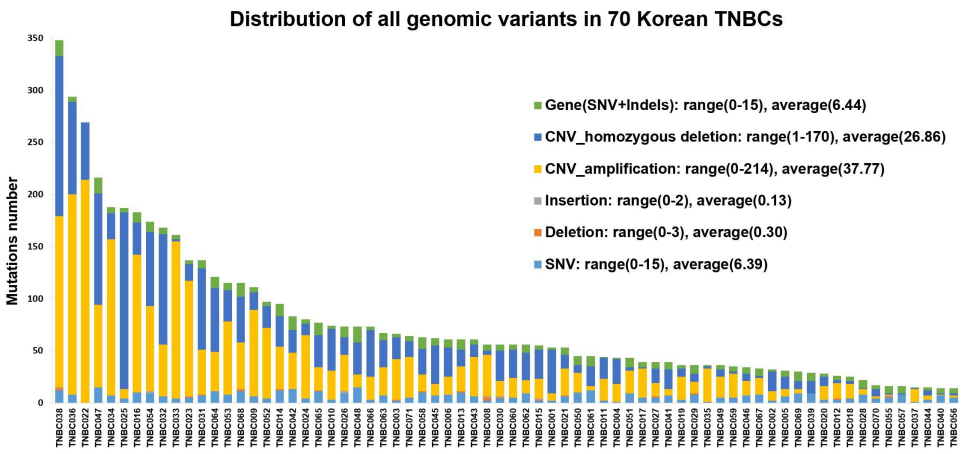


Figure 27. Distribution of genomic variants in 70 Korean TNBC patients. The number of somatic variants and CNVs differ from patient to patient. Especially, CNVs has greatly diverse range from 1 to 170 genes in homozygous deletion and 0 to 214 genes in gene amplification. TNBC038 contains the highest number of genomic alterations.

Most frequently mutated genes were TP53 (64%), NOTCH4 (27%), NOTCH3 (20%), GNAS (17%), BRD4 (14%), MN1 (14%), MLL2 (13%), PAX8 (13%), EXT1 (11%), PIK3CA (11%), ETV4 (10%), GLI3 (10%), HOOK3 (10%), MYCL1 (10%), SRGAP3 (10%) and so on (Table 10).

Somatic Mutation Genes		Amplified Genes		Homozygous Deleted Genes	
Gene	Frequency (%)	Gene	Frequency (%)	Gene	Frequency (%)
TP53	45 (64)	*NDRG1*	36 (51)	*WRN*	30 (43)
NOTCH4	19 (27)	*UBR5*	32 (46)	*IL6ST*	22 (31)
NOTCH3	14 (20)	*PTK2*	32 (46)	*APC*	21 (30)
GNAS	12 (17)	*RECQL4*	26 (37)	*PTK2B*	20 (29)
BRD4	10 (14)	*MYC*	26 (37)	*NF1*	19 (27)
MN1	10 (14)	*IKBKE*	25 (36)	*SETD2*	18 (26)
MLL2	9 (13)	*EXT1*	25 (36)	*PTPRD*	17 (24)
PAX8	9 (13)	*CDK2*	24 (34)	*PBRM1*	17 (24)
EXT1	8 (11)	*NTRK1*	24 (34)	*MLL3*	16 (23)
PIK3CA	8 (11)	*DDR2*	22 (31)	*PCM1*	16 (23)
ETV4	7 (10)	*MCL1*	22 (31)	*PLD2*	15 (21)
GLI3	7 (10)	*TPR*	20 (29)	*PIK3R1*	15 (21)
HOOK3	7 (10)	*PARP1*	19 (27)	*CDK2*	14 (20)
MYCL1	7 (10)	*TPM3*	19 (27)	*CSF1R*	14 (20)
SRGAP3	7 (10)	*PRCC*	19 (27)	*BUB1B*	14 (20)
ARID2	6 (9)	*RNF213*	19 (27)	*CDK12*	14 (20)
COL1A1	6 (9)	*ERC1*	19 (27)	*MTOR*	13 (19)
MTOR	6 (9)	*FH*	18 (26)	*CHEK2*	13 (19)
TRIM62	6 (9)	*NBN*	18 (26)	*ATM*	13 (19)
ATM	5 (7)	*RGL1*	17 (24)	*RB1*	13 (19)
BAP1	5 (7)	*PTPRD*	16 (23)	*MAP3K1*	13 (19)
JUN	5 (7)	*TIAM1*	16 (23)	*TIAM1*	12 (17)
KDM5C	5 (7)	*NOTCH4*	16 (23)	*ERCC2*	12 (17)
PPP2R1A	5 (7)	*IGF1R*	16 (23)	*KTN1*	12 (17)
BRCA2	4 (6)	*IKBKB*	16 (23)	*BRCA1*	12 (17)
CDKN2A	4 (6)	*GATA3*	16 (23)	*TSHR*	12 (17)
FGFR3	4 (6)	*PBX1*	16 (23)	*MLL2*	11 (16)
GRIN2D	4 (6)	*MLL2*	15 (21)	*PRKDC*	11 (16)
MAP3K1	4 (6)	*FLT4*	15 (21)	*TCF4*	11 (16)
MAPK8IP3	4 (6)	*EGFR*	15 (21)	*USP6*	11 (16)
PIK3R1	4 (6)	*RPTOR*	15 (21)	*RPS6KA2*	11 (16)
PTCH1	4 (6)	*RUNX1T1*	15 (21)	*TAF1*	11 (16)
RPTOR	4 (6)	*COX6C*	15 (21)	*KIT*	11 (16)
SFPQ	4 (6)	*FLNA*	14 (20)	*MAP2K2*	11 (16)
AKAP9	3 (4)	*TSC2*	14 (20)	*EML4*	11 (16)
ATRX	3 (4)	*ATR*	14 (20)	*RPS6KA3*	11 (16)
BAX	3 (4)	*MAML2*	14 (20)	*GNAQ*	11 (16)
BRD3	3 (4)	*NTRK3*	14 (20)	*KIAA1549*	10 (14)
CD74	3 (4)	*CRTC3*	14 (20)	*PMS1*	10 (14)
CDKN1A	3 (4)	*TFEB*	14 (20)	*BRCA2*	10 (14)
CIC	3 (4)	*MLL3*	13 (19)	*CHUK*	10 (14)
EGFR	3 (4)	*ERCC2*	13 (19)	*ALDH2*	10 (14)
EPHA5	3 (4)	*SMARCA4*	13 (19)	*FGFR3*	10 (14)
FLNA	3 (4)	*EP300*	13 (19)	*TP53*	10 (14)

Table 10. The list of high frequently mutated genes

Among the somatic variants occurring in TP53 gene, 5 resulted in stop-gained mutations, and another 6 frame-shift mutations, implying that those mutations could likely give rise to destruction of p53 function in the TNBC cells. We confirmed those somatic mutations in TP53 by Sanger capillary sequencing (Figure 28). In addition, one frame-shift mutation was identified in GNAS, ARID2, JUN and MYCL1, respectively (Figure 29).

Sample	Targeted NGS	Sanger sequencing					
		PCR product	Cloned DNA				
			#1	#2	#3	#4	#5
TNBC030 (N)	637C	637C	637C	637C	637C	637C	637C
TNBC030 (T)	637C>T	637C>T	637C	637C>T	637C>T	637C>T	637C>T
TNBC045 (N)	578A	578A	578A	578A	578A	578A	578A
TNBC045 (T)	578A>G	578A>G	578A>G	578A>G	578A>G	578A>G	578A>G

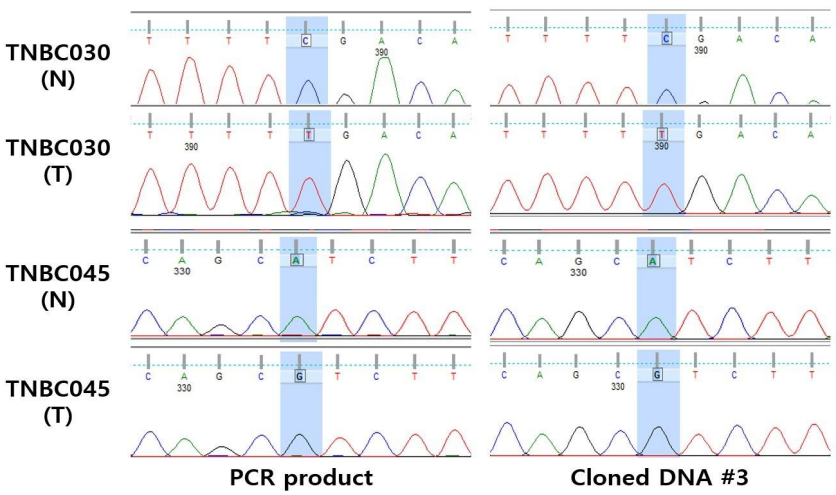


Figure 28. TP53 mutation validation by sanger sequencing. TP53 mutations 637C>T and 578A>G were validated by sanger sequencing in TNBC030 and TNBC045 samples.

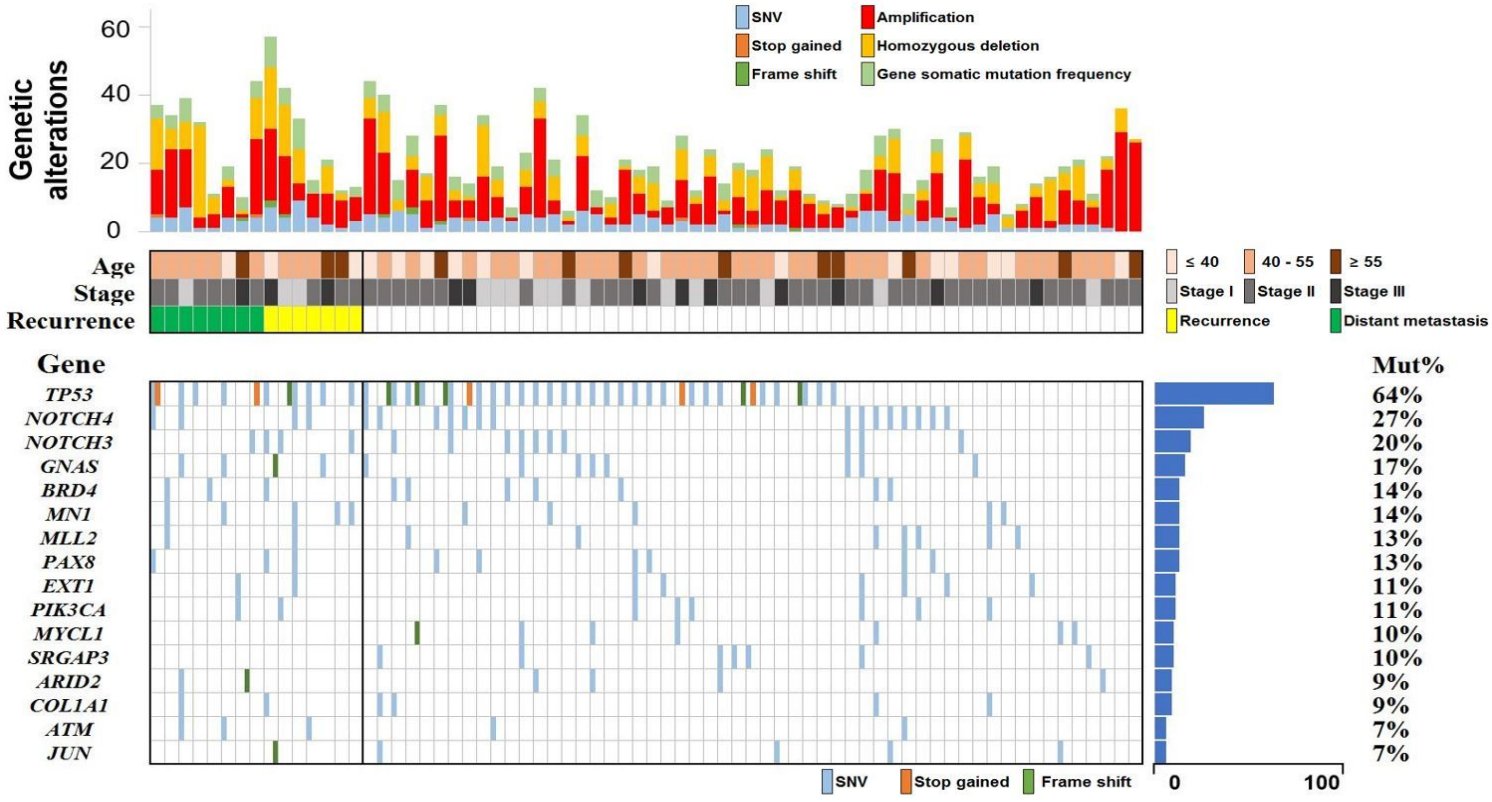


Figure 29. Landscape of high frequent somatic mutations in 70 Korean TNBC samples. High frequently occurred genetic variants in all 70 Korean TNBC samples. TP53 was the most frequently mutated gene and stop gained, frame shift mutation were mainly occurred at TP53.

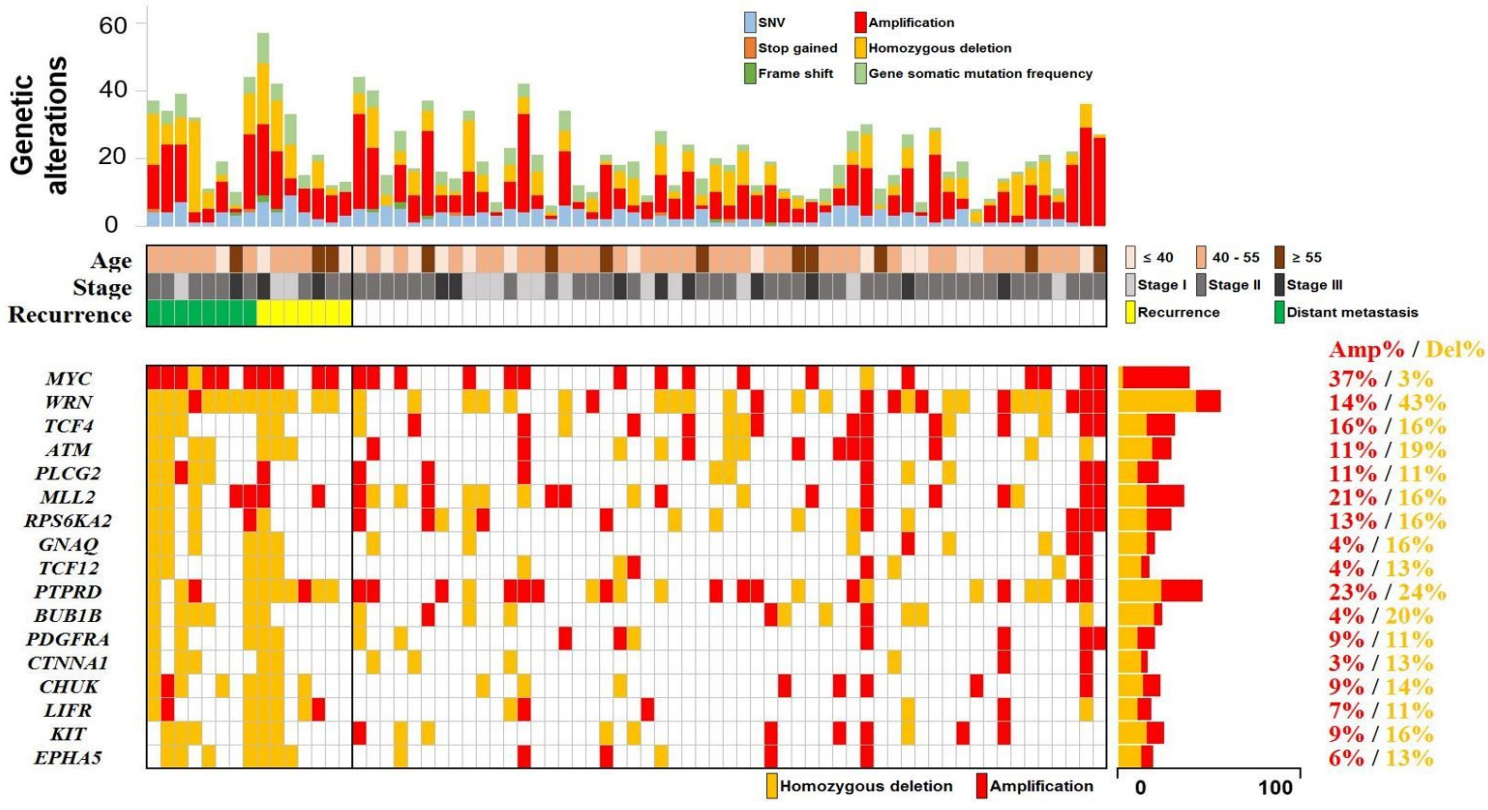


Figure 30. Landscape of high frequent copy number variations in 70 Korean TNBC samples. High frequently occurred copy number alterations in all 70 Korean TNBC samples. Some homozygous deletions were related with prognosis of TNBC patients.

The most recurrent novel mutation c.625T>G causing an amino acid change p.T209P in NOTCH4 protein was detected in nine TNBC patients (13%) TNBC patients (Table 7). In addition, another novel recurrent SNVs occurring in at least more than five TNBC patients are as follows; c.770T>G in ETV4 (p.V257G, 10%), c.148T>G in EXT1 (p.S50R, 10%), c.1264T>C in GNAS (p.S422P, 10%), c.6841C>G in NOTCH3 (p.A2281P, 10%), c.3746T>C in COL1A1 (p.E1249G, 9%), c.2482G>C in MLL2 (p.P828A, 9%), c.1103A>C in TP53 (p.H368P, 9%), c.3803A>C in ARID2 (p.N1268T, 7%) and c.118T>G in NOTCH4 (p.T40P, 7%).

Gene Name	Nucleotide Change	Amino Acid Change	Frequency (%)	Somatic mutationtype	Previously reported	Mutation Assessment							
						SIFT score	PolyPhen2				LRT score	Mutation Taster score	Mutation Assessor score
							HDIV score	HDIV pred	HVAR score	HVAR pred			
NOTCH4	c.625T>G	p.T209P	9 (13)	Heterozygous	Novel	0.01	0.9900 0.9030	D P	0.9270 0.6030	D P	0.1942	0.7871	1.3850
ETV4	c.770T>G	p.V257G	7 (10)	Heterozygous	Novel	0.11	0.8720 0.9420	P	0.8260 0.8120	P	0.0134	0.8814	1.9950
EXT1	c.148T>G	p.S50R	7 (10)	Heterozygous	Novel	0.74	0.0000	B	0.0000	B	0.0025	0.3789	0.0000
GNAS	c.1264T>C	p.S422P	7 (10)	Heterozygous	Novel	0.18	0.0030	B	0.0020	B	0.0000	0.0000	1.5250
NOTCH3	c.6841C>G	p.A2281P	7 (10)	Heterozygous	Novel	0.86	0.8380	P	0.2020	B	NA	0.5542	0.0000
COL1A1	c.3746T>C	p.E1249G	6 (9)	Heterozygous	Novel	0.00	0.7700	P	0.3270	B	0.0000	0.7868	3.4800
MLL2	c.2482G>C	p.P828A	6 (9)	Heterozygous	Novel	0.00	0.0080	B	0.0060	B	NA	NA	0.5500
TP53	c.1103A>C	p.H368P	6 (9)	Heterozygous	Novel	0.21	0.0000	B	0.0010	B	0.4522	0.0857	0.3450
ARID2	c.3803A>C	p.N1268T	5 (7)	Heterozygous	Novel	0.00	0.0540 0.0310 0.0320	B	0.0270 0.0210	B	0.0000	0.9744	0.9750
NOTCH4	c.118T>G	p.T40P	5 (7)	Heterozygous	Novel	0.03	0.0320 0.2470	B	0.0150 0.0870	B	0.1892	0.9635	2.5850
BRD4	c.2470T>G	p.T824P	4 (6)	Heterozygous	Novel	0.12	0.0000	B	0.0000	B	0.1482	0.0008	-0.6900
GLI3	c.2687T>G	p.D896A	4 (6)	Heterozygous	Novel	0.00	1.0000	D	1.0000	D	0.0000	1.0000	2.8350
HOOK3	c.62A>C	p.Q21P	4 (6)	Heterozygous	Novel	0.07	1.0000 0.9980	D	0.9980 1.0000	D	0.0000	0.8988	2.4150
MN1	c.2780G>A	p.T927R	4 (6)	Heterozygous	Novel	0.15	1.0000	D	0.9980	D	0.0000	0.9374	0.8050

MTOR	c.5480T>G	p.N1827T	4 (6)	Heterozygous	Novel	0.46	0.0000	B	0.0010	B	0.0234	0.0171	0.3450
NOTCH4	c.3064C>G	p.A1022P	4 (6)	Heterozygous	Novel	NA	0.9810	D	0.6030	P	0.0106	0.8376	0.5500
PAX8	c.695A>C	p.H232P	4 (6)	Heterozygous	Novel	0.02	0.0010 0.0000	B	0.0010 0.0020 0.0000	B	0.2301	0.0635	0.2050
PPP2R1A	c.584T>G	p.V195G	4 (6)	Heterozygous	Novel	0.07	0.9680 0.5880	D P	0.3140 0.2850	B	0.0000	1.0000	2.9600
TRIM62	c.1094T>G	p.I365S	4 (6)	Heterozygous	Novel	0.00	1.0000	D	0.9980	D	0.0000	0.9990	2.4750
ATM	c.6337A>C	p.T2113P	3 (4)	Heterozygous	Novel	0.28	0.0010	B	0.0040	B	0.6501	0.0022	0.0000
BAP1	c.626T>G	p.V209G	3 (4)	Heterozygous	Novel	0.00	1.0000	D	0.9980	D	0.0000	1.0000	3.5250
CD74	c.455T>G	p.L152R	3 (4)	Heterozygous	Novel	1.00	0.8170 0.0910 0.0150	P B	0.4990 0.0270 0.0380 0.2840	P B	0.7923	0.0338	1.1000
CDKN1A	c.93C>A	p.S31R	3 (4)	Homozygous	dbSNP	0.99	0.0000	B	0.0010	B	0.9321	0.0024	-0.1300
KDM5C	c.2254A>C	p.T752P	3 (4)	Heterozygous	Novel	0.17	0.0850 0.1990	B	0.0800 0.1730	B	0.0000	0.7922	1.9150
MAP3K14	c.2024A>C	p.H675P	3 (4)	Heterozygous	Novel	0.00	0.0000	B	0.0000	B	0.0000	0.0000	0.0000
MAPK8IP3	c.763T>C	p.S255P	3 (4)	Heterozygous	Novel	0.01	0.0100 0.7250 0.9840	B P D	0.0190 0.3270 0.6420	B P	0.0002	0.9997	1.8950
MCL1	c.116A>G	p.E39G	3 (4)	Heterozygous	Novel	NA 0.54	NA 0.0000	B	NA 0.0010	B	NA 0.0000	NA 0.0005	-0.5500
MN1	c.2773G>A	p.E925K	3 (4)	Heterozygous	Novel	0.29	0.9650	D	0.6550	P	0.0000	0.4251	0.5500

NOTCH3	c.6865G>C	p.A2289P	3 (4)	Heterozygous	dbSNP	0.37	0.0000	B	0.0000	B	NA	0.5542	0.0000
PAX8	c.665A>C	p.H222P	3 (4)	Heterozygous	Novel	0.12	0.5310 0.0030 0.3960 0.2010 0.1250	P B	0.0990 0.0020 0.0680 0.3000 0.0840	B	0.0014	0.6003	1.5450
PAX8	c.734T>G	p.Y245S	3 (4)	Heterozygous	Novel	0.03	0.6480 0.6390 0.0040 0.2720 0.7170	P B	0.4260 0.1100 0.0010 0.3280 0.2430	B	0.0168	0.2135	1.8800
PIK3CA	c.3140A>G	p.H1047R	3 (4)	Heterozygous	dbSNP COSMIC	0.16	0.6390	P	0.0850	B	0.0000	0.9999	0.0000
PIK3CA	c.821G>A	p.R274K	3 (4)	Heterozygous	dbSNP	0.03	0.9790	D	0.8920	P	0.0000	0.9997	2.1750
PIK3R1	c.367G>C	p.A123P	3 (4)	Heterozygous	Novel	0.21	0.3380	B	0.2000	B	0.0006	0.8999	1.3550
RPTOR	c.2557A>C	p.T853P	3 (4)	Heterozygous	Novel	0.29	0.2550 0.1670	B	0.1010 0.0460	B	0.0001	0.4881	1.5900
SRGAP3	c.3116T>C	p.F1039S	3 (4)	Heterozygous	Novel	0.26	0.2550 0.1650	B	0.0700 0.0320	B	0.0000	0.8194	1.7500
TP53	c.821G>T	p.R273L	3 (4)	Heterozygous	dbSNP COSMIC	0.00	0.9990 1.0000	D	0.9860 0.9900 0.9870 0.9880	D	0.0000	1.0000	3.1450
TP53	c.746G>A	p.R248Q	3 (4)	Homozygous Heterozygous	dbSNP COSMIC	0.01	1.0000 0.9940	D	0.9960 0.8820 0.9990 0.9950	D P	0.0000	1.0000	2.9700

Table 11. Top high frequent somatic mutations in TNBC genomes from 70 Korean patients

3. BRCA germline mutation analysis

The most notable phenomenon we discovered in this study was germline mutations occurring frequently in BRCA1 and BRCA2 genes, whose mutations have been well known in association with the causation of breast cancer (63, 64). Most recurrent SNVs BRCA1 c.3548A>G (p.K1183R, 56%), BRCA1 c.2613G>A (p.P871L, 61%), BRCA2 c.1114A>C (p.N372H, 46%), BRCA2 c.865A>C (p.N289H, 24%) and BRCA2 c.2971A>G (p.N991D, 13%) in our TNBC patients had also been considered as frequently occurring germline mutations in breast cancer cohorts of other studies (65, 66).

We also identified several deleterious mutations BRCA1 c.922_924delAGCinsT (p.Ser308Terfs, stop-gained, 2 patients), BRCA2 c.8363G>A (p.W2788X, stop-gained, 1 patient), BRCA1 c.2433delG (p.P811Pfs, somatic frame-shift, NM_007294.3, 1 patient) and BRCA2 c.9400delG (p.G3134Afs, somatic frame-shift, 1 patient), all of which are highly detrimental in their clinical impact assessment and had been confirmed in previous studies (67-69). For instance, the mutation c.8363G>A (p.W2788X) in BRCA2 gene caused a truncation of C-terminal part (harboring a partial portion of BRCA2-OB1 domain and the entire bodies of Tower and BRCA2-OB3 domains) in BRCA2 protein by replacing 2788th amino acid codon (tryptophan) with stop codon. In

addition, BRCA1 c.279delA (p.F93Ffs) was identified as novel germline frame-shift mutation in one TNBC patient in this study (Table 12).

Chromosomal Position	Frequency (n)	Gene Name	Chromosome	Reference Nucleotide	Genotype	Amino Acid Change
41244000	79	BRCA1	17	T	CC;TC	1183K>R
41244936	78	BRCA1	17	G	AA;GA	871P>L
41246626	4	BRCA1	17	T	AA;TA	308S>C
41246625	4	BRCA1	17	C	del	intron
41246624	4	BRCA1	17	G	delGC	FS
41219641	2	BRCA1	17	A	CA	1707H>HQ
41201205	2	BRCA1	17	A	GA	1801L>LP
41244982	2	BRCA1	17	A	AG	856Y>YH
41256909	2	BRCA1	17	A	GG;GA	93F>L
41256908	2	BRCA1	17	A	GG;GA	93F>S
41219681	2	BRCA1	17	T	del	intron
41219682	2	BRCA1	17	G	del	intron
41256907	2	BRCA1	17	A	delA	FS
41219680	2	BRCA1	17	G	delGTG	In-Frame
32906729	65	BRCA2	13	A	AC;CC	372N>NH
32906480	33	BRCA2	13	A	CA	289N>NH
32911463	18	BRCA2	13	A	GA;GG	991N>ND
32972884	4	BRCA2	13	A	AG	3412I>IV
32930598	2	BRCA2	13	T	TC	2490I>TI
32944570	2	BRCA2	13	G	AG	2788W>XW
32906558	2	BRCA2	13	T	TA	315C>SC
32906769	2	BRCA2	13	A	AG	385K>KR
32907359	2	BRCA2	13	A	AC	582T>TP

Table 12. BRCA1 and BRCA2 germline mutations

4. Copy number variation analysis

Our copy number variation analysis identified average numbers 37.77 (range: 0-214) and 26.86 (range: 1-170) of amplified and homozygous deleted genes per one TNBC patient, respectively (Figure 27). Genes showing most frequent copy number amplifications across 70 TNBC genomes were *NDRG1* (51%), *UBR5* (46%), *PTK2* (46%), *RECQL4* (37%), *MYC* (37%), *IKBKE* (36%), *EXT1* (36%), *CDK2* (34%), *NTRK1* (34%), *DDR2* (31%), *MCL1* (31%), *TPR* (29%), *PARP1* (27%), *TPM3* (27%) and *PRCC* (27%) (Table 7) (70-82). Among the 15 highly frequently amplified genes, 13 (except for *NDRG1* and *EXT1*) had been previously reported as oncogenes, significantly substantiating the reliability of our analysis result for amplified genes in this study. Consistent with a recent report of *NDRG1* amplified and overexpressed as a potential driver in non-small cell lung cancer, *NDRG1* was amplified frequently in our TNBC cohort (83). In addition, in line with a recent report of *EXT1* showing copy number gain and up-regulated expression as a prognostic signature gene in breast cancer patients, *EXT1* belonged to the list of frequently amplified genes in this study (84).

Genes showing high frequent homozygous deletion were *WRN* (43%), *IL6ST* (31%), *APC* (30%), *NF1* (27%), *SETD2* (26%), *PBRM1* (24%), *PCM1* (23%), *ATM* (19%), *RB1* (19%), *MAP3K1* (19%), *BRCA1* (17%), *TSHR*

(17%), *TP53* (14%), *CDKN2A* (13%) and *KDM6A* (13%) (Table 7). These homozygous deleted genes also had been well known as tumor suppressor genes except for IL6ST (58, 85-98). Intriguingly, a previous study reported that 60% of patients in their hepatocellular tumor cohort harbored somatic deletions in *IL6ST*, which generated in-frame deletions in gp130, subsequently facilitating ligand-independent activation of STAT3 and promoting a tumorigenesis of hepatocytes (99). *TP53*, a tumor suppressor gene showing the highest mutation frequency in this study, also exhibited its copy number homozygous losses in additional 10 TNBC patients, totaling the number of patients with abnormal *TP53* or its loss to 55. In addition to its germline deleterious frame-shift mutations in 3 TNBC patients, homozygous losses of the *BRCA1* locus were observed in TNBC genomes of another 12 patients, implying the contribution of the *BRCA1* copy number deletion to the causation of TNBC (Table 10 and Table 12).

5. Prognostic significance of homozygous deletions

We analysed proportional hazard ratio of somatic mutations or CNVs discovered in our TNBC cohort, and found that *MYC* amplification and many homozygous deletions were significantly associated with recurrence (Figure 31) or distant metastasis (Figure 32). *MYC*

amplification increased the risk of recurrence (HR=3.926, p-value=0.013) and distant metastasis (HR=6.236, p-value=0.026) together. In case of homozygous deletions, diverse homozygous deletions affected prognosis of TNBC patients including deletions of *WRN* (recurrence, HR=5.588, p-value=0.008 / distant metastasis, HR=9.479, p-value=0.036), *PTPRD* (recurrence, HR=4.415, p-value=0.005), *BUB1B* (recurrence, HR=3.638, p-value=0.018 / distant metastasis, HR=14.748, p-value=0.002) and *ATM* (recurrence, HR=5.414, p-value=0.001 / distant metastasis, HR=5.685, p-value=0.015) (Figure 31 and Figure 32).

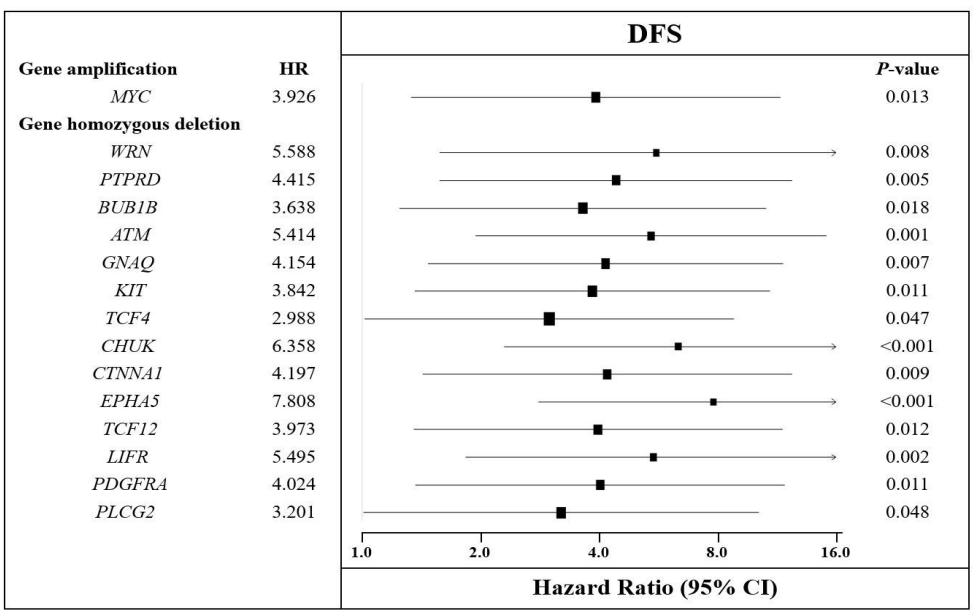


Figure 31. Proportional hazard ratio analysis of association between disease free survival and genetic alterations. *MYC* amplification and hymozygous deletion of 14 genes were related with recurrence of TNBC patients.

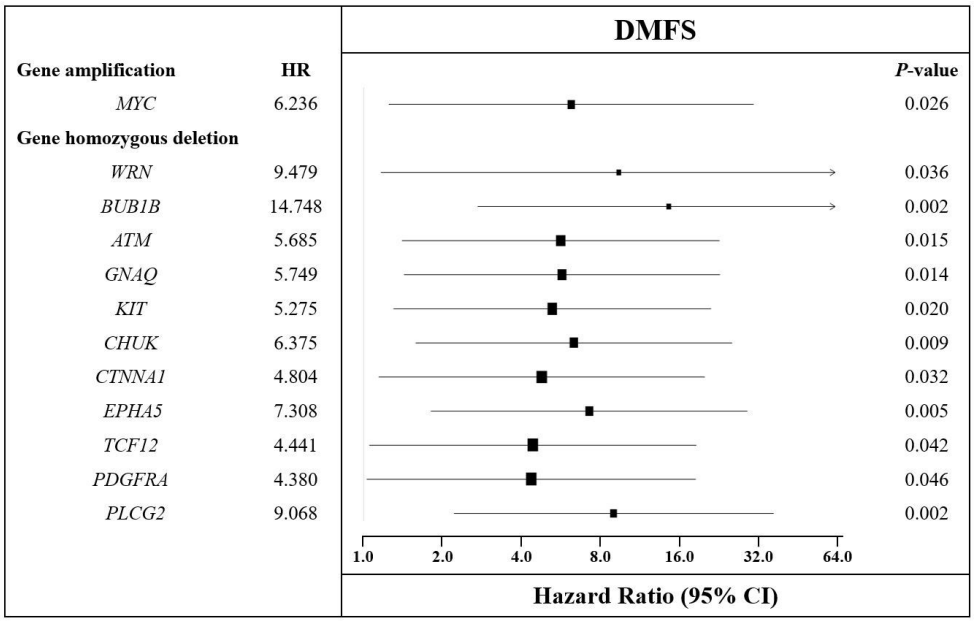


Figure 32. Proportional hazard ratio analysis of association between distant metastasis free survival and genetic alterations. *MYC* amplification and hymozygous deletion of 11 genes were related with distant metastasis of TNBC patients.

To further confirm that these alterations are related with bad prognosis of TNBC, we performed survival analyses using Kaplan-Meier method. It revealed that patients with *MYC* amplification had worse DFS (P-value=0.0099) and DMFS (P-value=0.0144) than those without *MYC* amplification. Homozygous deletions of *ATM* (DFS, P-value=0.0012 / DMFS, P-value=0.0209) and *WRN* (DFS, P-value=0.0025 / DMFS, P-value=0.0360) were also significantly related with bad prognosis of Korean TNBC patients (Figure 33).

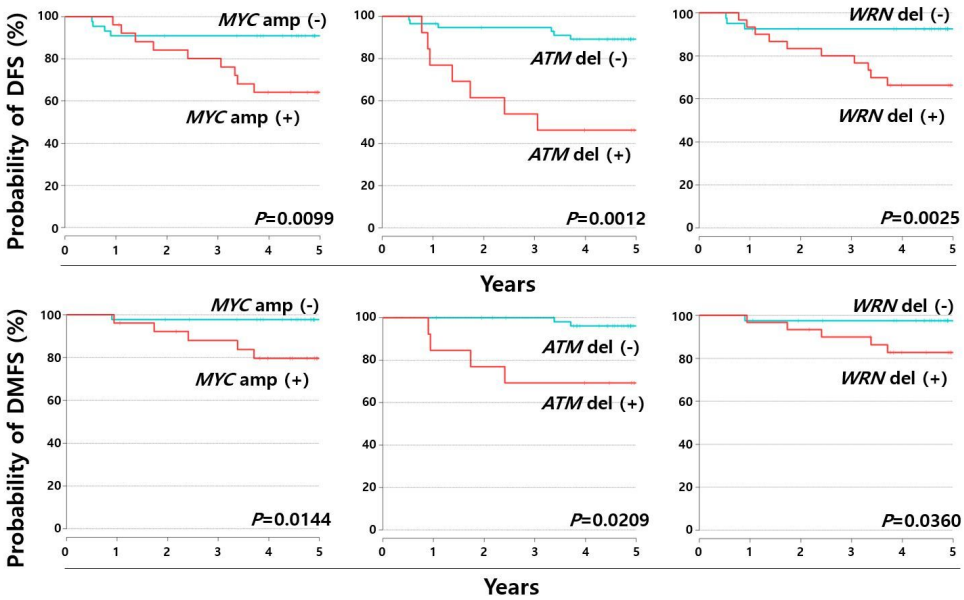


Figure 33. Survival analysis in 70 Korean TNBC patients based on *MYC* amplification and *ATM*, *WRN* homozygous deletion. *MYC* amplification and *ATM*, *WRN* homozygous deletions were related with recurrence and distant metastasis of TNBC patients.

6. New insight of TNBC novel oncogenes based on copy number alteration, expression and survival analysis

Analysis of copy number variation and mRNA expression data of breast cancer samples from TCGA (the cancer genome atlas) database revealed that *NDRG1*, *UBR5*, *EXT1* and *COX6C*, all of which showed frequent amplification in our 70 TNBC samples and had not been previously known in association with TNBC, may be novel oncogenes implicated in causing TNBC. As shown in Figure 34, very consistent

with genomic copy number gain and amplification status of *NDRG1*, *UBR5*, *EXT1* and *COX6C* loci in the clinical breast cancer samples, expression levels of their mRNAs increased significantly, suggesting that their genomic copy number amplification might contribute to the causation of TNBC. We also confirmed that expression of previously known oncogenes *MYC* and *NBN* were upregulated with increment of their copy number in those breast cancer samples (Figure 34).

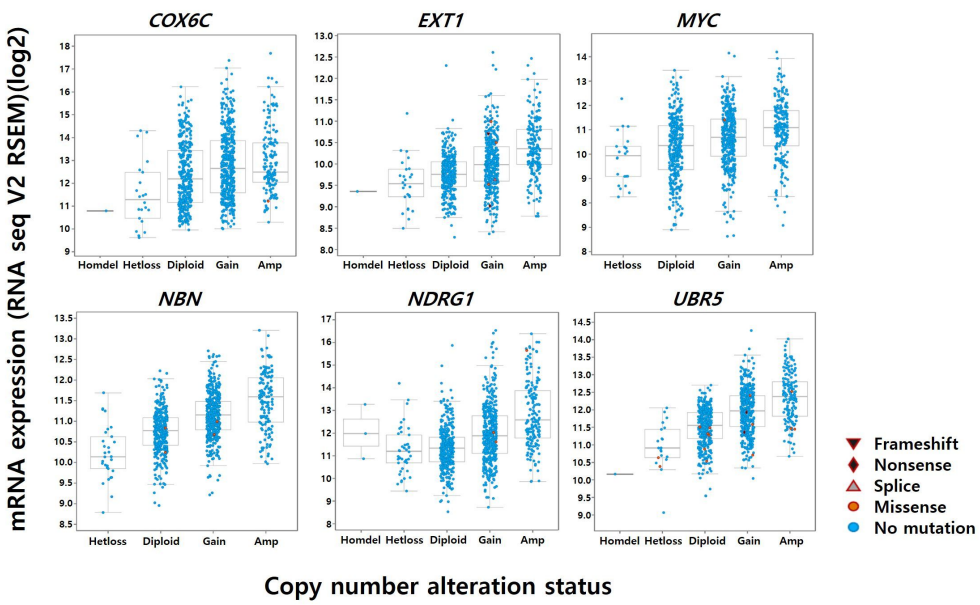


Figure 34. Increased mRNA expression level of six amplified genes in TCGA database. The expression level of six genes gradually increase with copy number amplifications.

Survival analysis revealed that survival rates of breast cancer

patients with copy number alteration of *NDRG1*, *UBR5*, *EXT1*, *COX6C*, *MYC* and *NBN* decreased significantly (Log rank test P-values: *COX6C*, 0.0073; *EXT1*, 0.0103; *MYC*, 0.0094; *NBN*, 0.0030; *NDRG1*, 0.0554; *UBR5*, 0.0122) (Figure 35).

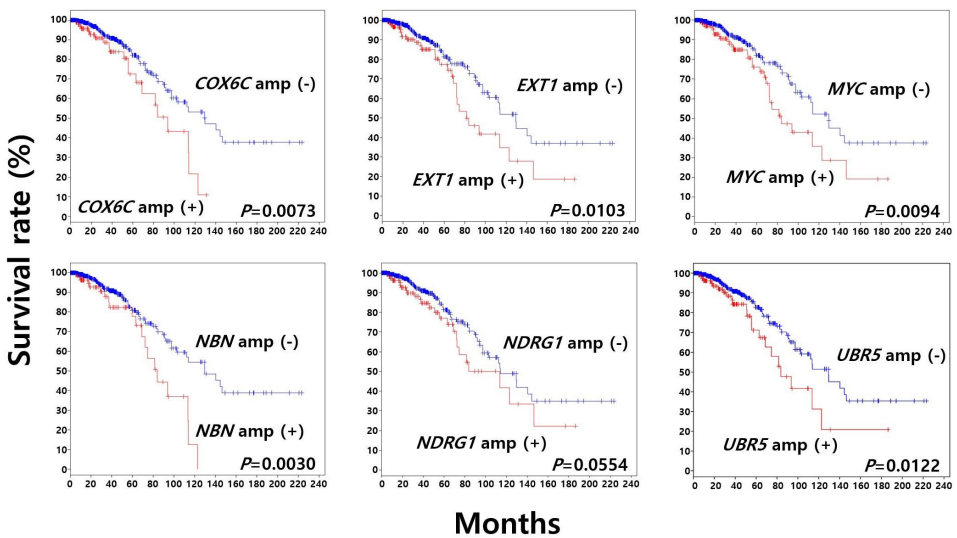


Figure 35. Amplified genes and its influence on breast cancer survival rate based on TCGA database. Amplification of six genes were significantly related with prognosis of breast cancer patients.

Furthermore, the influence of *MYC* amplification on prognosis was also confirmed in our cohort (Figure 31 and Figure 32). These results suggest strongly that NDRG1, UBR5, EXT1 and COX6C are novel TNBC oncogenes that might be very valuable therapeutic targets for the treatment of TNBC patients in the future.

7. Interaction network analysis among proteins encoded by genes with recurrent genetic alterations

We performed interaction network analysis among proteins encoded by genes displaying most frequent genetic alterations (somatic non-synonymous mutations and copy number variations) in our 70 TNBC patients using STRING (version 10) (100). As shown in Figure 36, we identified that the genetic alterations in genes (such as *TP53* and *WRN*) involved in DNA damage response are the major cause of TNBC in our cohort and also that overexpression of *NDRG1* might play oncogenic role by inhibiting p53 expression (101), through *NDRG1* expression is triggered by p53 (102). Very interestingly, our mutual exclusivity analysis result of 500 clinical breast cancer samples from TCGA database corroborates strongly a high likelihood (p value < 0.05, Fisher's exact test) of the co-occurrence of the genetic alterations in those genes (*TP53*, *MYC*, *WRN*, *NDRG1*, *NOTCH3*, *UBR5* and *BRD4*) involved in the above-mentioned interaction network, significantly substantiating the reliability of our results in this study (Figure 37).

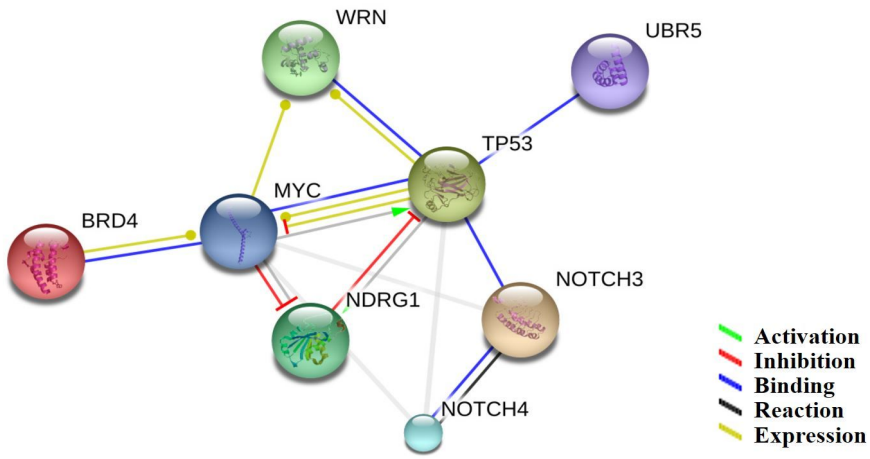


Figure 36. Interaction network analysis of genes having frequent genetic alteration in Korean TNBCs. Various DNA damage response genes are mutated in Korean TNBC patients, suggesting that the genetic alterations of these genes are the major cause of TNBC.

Gene	*TP53*	*UBR5*	*MYC*	*EXT1*	*NDRG1*	*BRD4*	*WRN*	*NOTCH3*	*NOTCH4*
TP53	---	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001	<0.000001	0.000133	0.49133
UBR5		---	<0.000001	<0.000001	<0.000001	0.009468	0.000013	0.039225	0.082807
MYC			---	<0.000001	<0.000001	0.000004	0.016019	0.164945	0.343477
EXT1				---	<0.000001	0.000022	0.051449	0.168537	0.174497
NDRG1					---	0.000016	0.082249	0.199102	0.167127
BRD4						---	0.055443	<0.000001	0.469896
WRN							---	0.432464	0.077401
NOTCH3								---	0.000001
NOTCH4									---

p-values <0.05, as derived via Fisher's Exact test are outlined in red.
p-values are not adjusted for FDR.

Legend
Strong tendency towards mutual exclusivity (0 < Odds Ratio < 0.1)
Some tendency towards mutual exclusivity (0.1 < Odds Ratio <0.5)
No association (0.5 < Odds Ratio < 2)
Tendency toward co-occurrence (2 < Odds Ratio < 10)
Stron tendency towards co-occurrence (Odds Ratio > 10)
No events recorded for one or both genes

Figure 37. Mutual exclusivity analysis of breast cancer samples from TCGA database. Various DNA damage response genes showed co-occurrence of the genetic mutations, supporting our results in this study.

Discussion

Although there had been recent studies for elucidating the clonal evolution and comprehensive mutational spectrum of TNBC that is the most malignant subtype of breast cancer, efforts for identifying mutational spectrum and therapeutic targets for TNBC patients in diverse ethnic populations are still lacking (53).

In this study, for the first time, we presented comprehensive mutational spectrum, copy number variation landscape and novel potential therapeutic targets in 70 Korean TNBC patients by performing targeted exome next generation sequencing (NGS) and copy number analysis. Compared with the cohort of Western European-North American (WENA) TNBC patients from TCGA database, our cohort of Korean TNCB patients exhibited unique feature, as well as common high-recurrent mutated genes such as *TP53* and *PIK3CA* in their mutational spectra. While *NOTCH4* was the second most frequently mutated genes in Korean TNBC patients, *TNS3* was the second most frequently mutated gene in WENA TNBC cohort (Table 13).

70 Korean TNBCs in this study		65 WENA TNBCs in TCGA database	
Gene	Frequency (%)	Gene	Frequency (%)
TP53	45 (64%)	*TP53*	35 (54%)
NOTCH4	19	*TNS3*	14
NOTCH3	14	*PIK3CA*	7 (11%)
GNAS	12	*ARHGAP5*	7
BRD4	10	*USH2A*	7
MN1	10	*MYO3A*	6
MLL2	9	*PTEN*	6
PAX8	9	*PCDHA10*	5
EXT1	8	*PPFIBP2*	5
PIK3CA	8 (11%)	*ATR*	5
ETV4	7	*RB1*	4
GLI3	7	*ZNF142*	4
HOOK3	7	*UBAP2L*	4
MYCL1	7	*GPR112*	4
SRGAP3	7	*HECW1*	4
ARID2	6	*UBR5*	4
COL1A1	6	*LRP2*	4
MTOR	6	*MDN1*	4
TRIM62	6	*COL6A3*	4
ATM	5	*SYNE1*	4
BAP1	5	*TTN*	4

Table 13. Comparison of frequently mutated genes between Korean TNBCs and WENA TNBCs

Recent studies had reported that *NOTCH4* and *NOTCH3* could be involved in causation of TNBC and breast cancer (103-105). In this regard, a novel recurrent single nucleotide variation *NOTCH4* c.625T>G

(p.T209P) occurring in 9 TNBC patients in our cohort, which replaced Threonine with Proline in fifth EGF-like domain (InterProScan 4; IPR000742) of NOTCH4 protein, might play a critical, yet unknown, role in inducing an oncogenic activity of this protein by likely affecting a binding affinity between NOTCH4 and its ligand. In addition, another recurrent novel mutations *NOTCH4* c.118T>G (p.T40P, 5 patients) and c.3064C>G (p.A1022P, 4 patients) occurred also in the EGF-like domain repeats. Unlike the three above-mentioned novel *NOTCH4* mutations, recurrent novel mutation *NOTCH3* c.6841C>G (p.A2281P, 7 patients) occurred in a region encoding N-terminal cytoplasmic domain of NOTCH3 protein, which could be implicated in upregulating downstream target oncogenes by activating transcription factor RBPJ for inducing carcinogenesis (Table 11).

Another difference in mutational spectrum between the two TNBC cohorts is that *GNAS*, *BRD4*, *MN1*, *MLL2*, *PAX8*, *EXT1*, *ETV4*, *GLI3*, *HOOK3*, *MYCL1*, *SRGAP3*, *ARID2*, *COL1A1* and *MTOR* were more frequently mutated in Korean 70 TNBC patients (Table 13). A novel recurrent variation BRD4 p.T824P (c.2470T>G, 4 patients) was embedded in a region closely adjacent to NET domain, which plays a critical role in dynamic regulation of breast cancer metastasis by conferring transcriptional activation through mediating protein-protein interactions (106, 107). Novel recurrent variation GNAS p.S422P (c.1264T>C, 7

patients) resided in a region adjacent to P-loop NTPase domain (IPR027417) in GNAS protein involved in carcinogenesis of breast cancer (108). Novel recurrent variation MN1 p.T927R (c.2780G>A, 4 patients) was embedded in Glycine-rich domain of MN1 protein, whose overexpression was implicated in accelerating leukemia onset by suppressing p53 and Bim induction (109). Interestingly, the Glycine-rich domain, which was known to play a role in nucleocytoplasmic shuttling and nuclear localization of various proteins, was a hot-spot undergoing mutations in TARDBP (TAR DNA binding protein) implicated in causing amyotrophic lateral sclerosis (110, 111). Novel recurrent variation EXT1 p.S50R (c.148T>G, 7 patients) was embedded in a non-cytoplasmic region closely adjacent to transmembrane domain in EXT1, a ER-resident type II transmembrane glycoprotein known as a tumor suppressor, whose mutation or suppression caused malignant osteosarcomas and multiple myeloma in patients (112, 113). Novel recurrent variation MTOR p.N1827T (c.5480T>G, 4 patients) resided in PIK-related kinase domain (FAT domain, IPR003151) of MTOR protein implicated in progression of TNBC (114, 115). Even though this mutation occurred in the PIK-related kinase domain, its position was located closely adjacent to both Rapamycin-binding domain and Tetratricopeptide-like helical domain. Novel recurrent variation PAX8 p.Y245S (c.734T>G, 3 patients) was located between Winged

helix-turn-helix DNA-binding domain (IPR011991) and Paired-box protein 2 C-terminal domain (IPR022130) in PAX8 involved in promoting tumor cell growth (116). Novel recurrent variation COL1A1 p.E1249G (c.3746T>C, 6 patients) was embedded in C-terminal Fibrillar collagen domain (IPR000885) of COL1A1, whose down-regulation was associated with progression of diverse cancers (117, 118). Very intriguingly, the novel point mutation occurred at the same position in TNBC genomes of all (6) of patients with COL1A1 mutation in our cohort, implying a high likelihood of the amino acid residue position (E1249) to affect a critical functional activity of COL1A1. Novel recurrent variation GLI3 p.D896A (c.2687T>G, 4 patients) resided adjacent to Zinc finger C2H2-type/integrase DNA-binding domain (IPR013087) in GLI3, whose high expression was associated with maintenance of metastatic breast cancer stem cells (119). Novel recurrent variation MLL2 p.P828A (c.2482G>C, 6 patients) was flanked by two PHD-finger domains (IPR019787) in MLL2 (also called KMT2D) involved in causing many types of cancer (120). In contrast to the novelty of the above-mentioned recurrent mutations in those cancer-associated genes in our cohort, the critical molecular role of the recurrent mutation PIK3CA p.H1047R (c.3140A>G, 3 patients) embedded in Phosphatidylinositol 3-/4-kinase catalytic domain (IPR000403) of PIK3CA has been elucidated in recent two studies (121, 122). The two

companion papers reported that PIK3CA p.H1047R (c.3140A>G) was crucial in inducing multi-potency and heterogeneity of breast tumor. In this regard, the novel recurrent mutations identified in our cohort (Table 10) are worthy of pursuit as strong candidates for studying molecular pathogenesis of triple-negative breast cancer in the upcoming future.

The most notable feature in copy number variation landscape on the cancer genomes of Korean TNBC cohort is that NDRG1 is the most frequently amplified gene and WRN is the most frequently lost gene, while in WENA cohort, EGFR and PARK2 were the top frequent amplified and deleted genes, respectively (Table 14 and Table 15).

70 Korean TNBCs inthisstudy		65 WENA TNBCs inTCGAdatabase	
Gene	Frequency	Gene	Frequency
NDRG1	36	*EGFR*	5
UBR5	32	*SEC61G*	4
PTK2	32	*NOTCH2*	3
RECQL4	26	*HMGCS2*	2
MYC	26	*PSPH*	2
IKBKE	25	*RGMA*	2
EXT1	25	*CCT6A*	2
CDK2	24	*PHKG1*	2
NTRK1	24	*HUWE1*	1
DDR2	22	*HEXA*	1
MCL1	22	*ADCY9*	1
TPR	20	*LRP6*	1
PARP1	19	*PORCN*	1
TPM3	19	*IQSEC2*	1
PRCC	19	*GRIN2A*	1
RNF213	19	*MYO9A*	1
ERC1	19	*ARHGDIB*	1
FH	18	*MYH11*	1
NBN	18	*GUCY2C*	1
RGL1	17	*SLC2A1*	1
PTPRD	16	*AKAP8*	1
TIAM1	16	*GAB2*	1
NOTCH4	16	*MAGOHB*	1
IGF1R	16	*SEC23B*	1
IKBKB	16	*PEX5*	1
GATA3	16	*HDAC6*	1
PBX1	16	*SUV39H1*	1
MLL2	15	*OTUD5*	1
FLT4	15	*ADRB3*	1
EGFR	15	*RBMX*	1
RPTOR	15	*EIF4EBP1*	1

Table 14. Comparison of frequently amplified genes between Korean TNBCs and WENA TNBCs

70 Korean TNBCs inthisstudy		65 WENA TNBCs inTCGAdatabase	
Gene	Frequency (%)	Gene	Frequency (%)
WRN	30	*PARK2*	6
IL6ST	22	*RB1*	5 (8%)
APC	21	*OR4N4*	5
PTK2B	20	*PTEN*	3
NF1	19	*TLR7*	3
SETD2	18	*PRPS2*	3
PTPRD	17	*PAPSS2*	3
PBRM1	17	*MAP3K1*	3 (5%)
MLL3	16	*ARHGAP6*	3
PCM1	16	*OR4K2*	3
PLD2	15	*ATXN3L*	3
PIK3R1	15	*OFD1*	3
CDK2	14	*TLR3*	3
CSF1R	14	*TMSB4X*	3
BUB1B	14	*PORCN*	2
CDK12	14	*HDAC6*	2
MTOR	13	*SUV39H1*	2
CHEK2	13	*OTUD5*	2
ATM	13	*HIST1H4A*	2
RB1	13 (19%)	*PRKG1*	2
MAP3K1	13 (19%)	*FAF1*	2
TIAM1	12	*ROBO2*	2
ERCC2	12	*LIPC*	2
KTN1	12	*ITGA2*	2
BRCA1	12	*MBTPS1*	2
TSHR	12	*PRKX*	2
MLL2	11	*HRH1*	2
PRKDC	11	*NDUFS4*	2
TCF4	11	*ACSL1*	2
USP6	11	*CASP3*	2
RPS6KA2	11	*PIGA*	2

Table 15. Comparison of frequently deleted genes between Korean TNBCs and WENA TNBCs

Unlike expression characteristics of a majority of oncogenes and tumor suppressor genes, NDRG1 upregulation promotes breast cancer progression and differentiation, but suppresses tumor metastasis (123, 124). In addition, NDRG1 expression is induced by DNA damage owing to binding of p53 to NDRG1 promoter and NDRG1 inactivates p53 directly (102, 125). Moreover, tumor suppressor WRN contributes to p53 induction through interaction with tumor suppressor ATM, whose coding gene locus was lost in 13 patients in our TNBC Korean cohort (126). Furthermore, the evidence that corroborates the fact that a majority of 70 Korean TNBC patients had undergone carcinogenesis by demolition of DNA damage response is reinforced by MYC amplification status in 26 patients. Between two opposing MYC-induced DNA damage responses (tumor promoting response and tumor suppressive response) acting as double-edged swords, the tumor suppressive DNA damage response pathway had been destroyed due to the copy number loss and mutation of TP53 and ATM in those patients TNBC cells, both of which encode essential proteins for the pathway (127). Instead of the destroyed tumor suppressive pathway, we hypothesize that MYC-induced tumor promoting DNA damage response pathway might exert its role for tumor progression in those Korean TNBC patients.

Another important thing we should not neglect, but rather prioritize carefully in considering the causation of TNBC in Korean patients is

how the inherited and somatic mutations and CNVs of BRCA1/BRCA2 could contribute to carcinogenesis. All of 70 Korean TNBC patients had inherited or somatic mutations of BRCA1/BRCA2 (Table 12). Those inherited and somatic mutations are deleterious frame-shift mutations or non-synonymous mutations with unknown consequences. We hypothesize that in case of Korean TNBC patients (23%) with homozygous copy number loss (12 patients) or deleterious frame-shift mutations of BRCA1 (one somatic mutation and three germline mutations), a collapse in the BRCA1-mediated p53-independent DNA damage response pathway might trigger carcinogenesis (128). However, in case of patients harboring genetic alterations (inherited or somatic mutations and homozygous copy number loss) in BRCA1, TP53 and ATM, which are observed in a large portion of our Korean cohort, a collapse in DNA damage response pathway requiring a cooperativity of BRCA1, TP53 and ATM might trigger and promote carcinogenesis (128). We also elucidated the deleterious functional consequence of genetic alterations in DNA damage response genes by analyzing protein interaction network and mutual exclusivity (Figure 36 and Figure 37). Supporting the above result, homozygous deletion of WRN and ATM, both of which are involved in DNA repair system, significantly decreased the probability of DFS and DMFS in our cohort (Figure 35). Recently, FDA approved PARP inhibitor to treat ovarian cancer patients who have deleterious

germline mutations in BRCA1 and BRCA2, both of which are involved in DNA repair system (129). Therefore, our finding strongly suggests that other DNA repair genes with deleterious mutations that we found in our cohort are potential candidate genes for usage as biomarkers for prognosis prediction or companion diagnosis of TNBC patients and for application of PARP inhibitors against their corresponding protein structures as therapeutic targets.

Another novel finding in this study is that EXT1 and COX6C might be novel diagnostic and therapeutic targets for TNBC. Previous studies reported that EXT1 could exert two opposing effects in association with carcinogenesis, that is, on the one hand, tumor suppressor in multiple cartilaginous tumors (130), and on the other hand, oncogene-like role in multiple myeloma (113). This fact generated a dilemma regarding a relatively ubiquitous role of EXT1 in association with cancer. However, we showed that EXT1 exhibited frequent amplification in our Korean TNBC cohort, with validation of EXT1 upregulation in breast cancer samples with its amplification and remarkable decrease in overall survival rate of breast cancer patients with EXT1 genetic alteration (Figure 35), suggesting EXT1 as novel oncogene in TNBC and breast cancer. In case of COX6C, little is known regarding its functional role in association with cancer, so far. We revealed for the first time that COX6C was frequently amplified in Korean TNBC cohort and also

validated that COX6C amplification caused its upregulation and decreased survival rate in breast cancer patients, implying COX6C as novel oncogene, similar with EXT1 in TNBC and breast cancer (Figure 35).

REFERENCES

1. Ledergerber C, Dessimoz C. Base-calling for next-generation sequencing platforms. Briefings in bioinformatics. 2011;12(5):489-97.

2. Mardis ER. Next-generation DNA sequencing methods. Annual review of genomics and human genetics. 2008;9:387-402.

3. Micron. Bioinformatics. [cited 20th June 2016]. Available from: http://www.micronautomata.com/bioinformatics/.

4. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135-45.

5. National Institute of Health. The cost of sequencing a human genome. [cited 20th June 2016]. Available from: https://www.genome.gov/sequencingcosts/.

6. Rizzo JM, Buck MJ. Key principles and clinical applications of "next-generation" DNA sequencing. Cancer prevention research (Philadelphia, Pa). 2012;5(7):887-900.

7. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.

8. Evan GI, Vousden KH. Proliferation, cell cycle and apoptosis in cancer. Nature. 2001;411(6835):342-8.

9. Lee EY, Muller WJ. Oncogenes and tumor suppressor genes. Cold Spring Harbor perspectives in biology. 2010;2(10):a003236.

10. Sadikovic B, Al-Romaih K, Squire JA, Zielenska M. Cause and consequences of genetic and epigenetic alterations in human cancer. Current genomics. 2008;9(6):394-408.

11. Esteller M. Epigenetics in cancer. The New England journal of medicine. 2008;358(11):1148-59.

12. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009;1(2):239-59.

13. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.

14. Polyak K. Heterogeneity in breast cancer. The Journal of clinical investigation. 2011;121(10):3786-8.

15. Boyle P. Triple-negative breast cancer: epidemiological considerations and recommendations. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2012;23 Suppl 6:vi7-12.

16. Cetin I, Topcul M. Triple negative breast cancer. Asian Pacific journal of cancer prevention : APJCP. 2014;15(6):2427-31.

17. Zubeda S, Kaipa PR, Shaik NA, Mohiuddin MK, Vaidya S, Pavani B, et al. Her-2/neu status: a neglected marker of prognostication and management of breast cancer patients in India. Asian Pacific journal of

cancer prevention : APJCP. 2013;14(4):2231-5.

18. Foulkes WD, Smith IE, Reis-Filho JS. Triple-negative breast cancer. The New England journal of medicine. 2010;363(20):1938-48.

19. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics. 2012;13(7):484-92.

20. Day JJ, Sweatt JD. Epigenetic mechanisms in cognition. Neuron. 2011;70(5):813-29.

21. Pinney SE. Mammalian Non-CpG Methylation: Stem Cells and Beyond. Biology. 2014;3(4):739-51.

22. Jeong HM, Kwon MJ, Shin YK. Overexpression of Cancer-Associated Genes via Epigenetic Derepression Mechanisms in Gynecologic Cancer. Frontiers in oncology. 2014;4:12.

23. Slieker RC, Roost MS, van Iperen L, Suchiman HE, Tobi EW, Carlotti F, et al. DNA Methylation Landscapes of Human Fetal Development. PLoS genetics. 2015;11(10):e1005583.

24. Jakovcevski M, Akbarian S. Epigenetic mechanisms in neurological disease. Nature medicine. 2012;18(8):1194-204.

25. Zhang Z, Zhang R. Epigenetics in autoimmune diseases: Pathogenesis and prospects for therapy. Autoimmunity reviews. 2015;14(10):854-63.

26. Li N, Ye M, Li Y, Yan Z, Butcher LM, Sun J, et al. Whole genome DNA methylation analysis based on high throughput sequencing

technology. Methods (San Diego, Calif). 2010;52(3):203-12.

27. Lee EJ, Luo J, Wilson JM, Shi H. Analyzing the cancer methylome through targeted bisulfite sequencing. Cancer letters. 2013;340(2):171-8.

28. Sun Z, Cunningham J, Slager S, Kocher JP. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics. 2015;7(5):813-28.

29. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nature biotechnology. 2010;28(10):1097-105.

30. Walker DL, Bhagwate AV, Baheti S, Smalley RL, Hilker CA, Sun Z, et al. DNA methylation profiling: comparison of genome-wide sequencing methods and the Infinium Human Methylation 450 Bead Chip. Epigenomics. 2015:1-16.

31. Weyrich A, Schullermann T, Heeger F, Jeschek M, Mazzoni CJ, Chen W, et al. Whole genome sequencing and methylome analysis of the wild guinea pig. BMC genomics. 2014;15:1036.

32. Holmes EE, Jung M, Meller S, Leisse A, Sailer V, Zech J, et al. Performance Evaluation of Kits for Bisulfite-Conversion of DNA from Tissues, Cell Lines, FFPE Tissues, Aspirates, Lavages, Effusions, Plasma, Serum, and Urine. PLoS ONE. 2014;9(4):e93933.

33. Khanna A, Czyz A, Syed F. EpiGnome[trade] Methyl-Seq Kit: a novel post-bisulfite conversion library prep method for methylation analysis. Nat Meth. 2013;10(10).

34. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics (Oxford, England). 2011;27(11):1571-2.

35. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9.

36. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010;26(6):841-2.

37. Serra-Juhe C, Cusco I, Homs A, Flores R, Toran N, Perez-Jurado LA. DNA methylation abnormalities in congenital heart disease. Epigenetics. 2015;10(2):167-77.

38. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nature neuroscience. 2014;17(9):1156-63.

39. Paska AV, Hudler P. Aberrant methylation patterns in cancer: a clinical view. Biochemia medica. 2015;25(2):161-76.

40. Gyorffy B, Bottai G, Fleischer T, Munkacsy G, Budczies J, Paladini L, et al. Aberrant DNA methylation impacts gene expression

and prognosis in breast cancer subtypes. International journal of cancer Journal international du cancer. 2016;138(1):87-97.

41. Walker DL, Bhagwate AV, Baheti S, Smalley RL, Hilker CA, Sun Z, et al. DNA methylation profiling: comparison of genome-wide sequencing methods and the Infinium Human Methylation 450 Bead Chip. Epigenomics. 2015:1-16.

42. Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nature biotechnology. 2010;28(10):1106-14.

43. Robinson MD, Statham AL, Speed TP, Clark SJ. Protocol matters: which methylome are you actually studying? Epigenomics. 2010;2(4):587-98.

44. Robinson MD, Stirzaker C, Statham AL, Coolen MW, Song JZ, Nair SS, et al. Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. Genome research. 2010;20(12):1719-29.

45. Ziller MJ, Hansen KD, Meissner A, Aryee MJ. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. Nature methods. 2015;12(3):230-2, 231 p following 232.

46. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome biology. 2016;17(1):53.

47. Rauscher GH, Kresovich JK, Poulin M, Yan L, Macias V, Mahmoud AM, et al. Exploring DNA methylation changes in promoter, intragenic, and intergenic regions as early and late events in breast cancer formation. BMC cancer. 2015;15:816.

48. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends in genetics : TIG. 2015;31(5):274-80.

49. Liyanage VR, Jarmasz JS, Murugeshan N, Del Bigio MR, Rastegar M, Davie JR. DNA modifications: function and applications in normal and disease States. Biology. 2014;3(4):670-723.

50. Sotiriou C, Pusztai L. Gene-Expression Signatures in Breast Cancer. New England Journal of Medicine. 2009;360(8):790-800.

51. Badve S, Dabbs DJ, Schnitt SJ, Baehner FL, Decker T, Eusebi V, et al. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc. 2011;24(2):157-67.

52. Reddy KB. Triple-negative breast cancers: an updated review on treatment options. Current oncology (Toronto, Ont). 2011;18(4):e173-9.

53. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative

breast cancers. Nature. 2012;486(7403):395-9.

54. de Rinaldis E, Gazinska P, Mera A, Modrusan Z, Fedorowicz GM, Burford B, et al. Integrated genomic analysis of triple-negative breast cancers reveals novel microRNAs associated with clinical and molecular phenotypes and sheds light on the pathways they control. BMC genomics. 2013;14:643.

55. Craig DW, O'Shaughnessy JA, Kiefer JA, Aldrich J, Sinari S, Moses TM, et al. Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. Molecular cancer therapeutics. 2013;12(1):104-16.

56. Drilon A, Wang L, Arcila ME, Balasubramanian S, Greenbowe JR, Ross JS, et al. Broad, Hybrid Capture-Based Next-Generation Sequencing Identifies Actionable Genomic Alterations in Lung Adenocarcinomas Otherwise Negative for Such Alterations by Other Genomic Testing Approaches. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015;21(16):3631-9.

57. Han SW, Kim HP, Shin JY, Jeong EG, Lee WC, Lee KH, et al. Targeted sequencing of cancer-related genes in colorectal cancer using next-generation sequencing. PloS one. 2013;8(5):e64271.

58. Singh RR, Patel KP, Routbort MJ, Aldape K, Lu X, Manekia J, et al. Clinical massively parallel next-generation sequencing analysis of

409 cancer-related genes for mutations and copy number variations in solid tumours. British journal of cancer. 2014;111(10):2014-23.

59. Shitara M, Okuda K, Suzuki A, Tatematsu T, Hikosaka Y, Moriyama S, et al. Genetic profiling of thymic carcinoma using targeted next-generation sequencing. Lung cancer (Amsterdam, Netherlands). 2014;86(2):174-9.

60. Miya F, Kato M, Shiohama T, Okamoto N, Saitoh S, Yamasaki M, et al. A combination of targeted enrichment methodologies for whole-exome sequencing reveals novel pathogenic mutations. Scientific reports. 2015;5:9331.

61. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative C(T) method. Nature protocols. 2008;3(6):1101-8.

62. Hoh BP, Sam SS, Umi SH, Mahiran M, Nik Khairudin NY, Rafidah Hanim S, et al. A novel rare copy number variant of the ABCF1 gene identified among dengue fever patients from Peninsular Malaysia. Genetics and molecular research : GMR. 2014;13(1):980-5.

63. Roa BB, Boyd AA, Volcik K, Richards CS. Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. Nature genetics. 1996;14(2):185-7.

64. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage

Consortium. American journal of human genetics. 1998;62(3):676-89.

65. Horvath A, Pakala SB, Mudvari P, Reddy SD, Ohshiro K, Casimiro S, et al. Novel insights into breast cancer genetic variance through RNA sequencing. Scientific reports. 2013;3:2256.

66. Jalkh N, Nassar-Slaba J, Chouery E, Salem N, Uhrchammer N, Golmard L, et al. Prevalance of BRCA1 and BRCA2 mutations in familial breast cancer patients in Lebanon. Hereditary cancer in clinical practice. 2012;10(1):7.

67. Machackova E, Foretova L, Lukesova M, Vasickova P, Navratilova M, Coene I, et al. Spectrum and characterisation of BRCA1 and BRCA2 deleterious mutations in high-risk Czech patients with breast and/or ovarian cancer. BMC cancer. 2008;8:140.

68. Seo JH, Cho DY, Ahn SH, Yoon KS, Kang CS, Cho HM, et al. BRCA1 and BRCA2 germline mutations in Korean patients with sporadic breast cancer. Human mutation. 2004;24(4):350.

69. Ahn SH, Son BH, Yoon KS, Noh DY, Han W, Kim SW, et al. BRCA1 and BRCA2 germline mutations in Korean breast cancer patients at high risk of carrying mutations. Cancer letters. 2007;245(1-2):90-5.

70. Subbaiah VK, Zhang Y, Rajagopalan D, Abdullah LN, Yeo-Teh NS, Tomaic V, et al. E3 ligase EDD1/UBR5 is utilized by the HPV E6 oncogene to destabilize tumor suppressor TIP60. Oncogene. 2016;35(16):2062-74.

71. Golubovskaya VM, Conway-Dorsey K, Edmiston SN, Tse CK, Lark AA, Livasy CA, et al. FAK overexpression and p53 mutations are highly correlated in human breast cancer. International journal of cancer. 2009;125(7):1735-8.

72. Peng C, Zeng W, Su J, Kuang Y, He Y, Zhao S, et al. Cyclin-dependent kinase 2 (CDK2) is a key mediator for EGF-induced cell transformation mediated through the ELK4/c-Fos signaling pathway. Oncogene. 2016;35(9):1170-9.

73. Fang H, Nie L, Chi Z, Liu J, Guo D, Lu X, et al. RecQL4 helicase amplification is involved in human breast tumorigenesis. PloS one. 2013;8(7):e69600.

74. Walz S, Lorenzin F, Morton J, Wiese KE, von Eyss B, Herold S, et al. Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. Nature. 2014;511(7510):483-7.

75. Boehm JS, Zhao JJ, Yao J, Kim SY, Firestein R, Dunn IF, et al. Integrative genomic approaches identify IKBKE as a breast cancer oncogene. Cell. 2007;129(6):1065-79.

76. Vaishnavi A, Capelletti M, Le AT, Kako S, Butaney M, Ercan D, et al. Oncogenic and drug-sensitive NTRK1 rearrangements in lung cancer. Nature medicine. 2013;19(11):1469-72.

77. Peschard P, Park M. From Tpr-Met to Met, tumorigenesis and tubes. Oncogene. 2007;26(9):1276-85.

78. Zhang H, Guttikonda S, Roberts L, Uziel T, Semizarov D, Elmore SW, et al. Mcl-1 is critical for survival in a subgroup of non-small-cell lung cancer cell lines. Oncogene. 2011;30(16):1963-8.

79. Payne LS, Huang PH. Discoidin domain receptor 2 signaling networks and therapy in lung cancer. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer. 2014;9(6):900-4.

80. Amano Y, Ishikawa R, Sakatani T, Ichinose J, Sunohara M, Watanabe K, et al. Oncogenic TPM3-ALK activation requires dimerization through the coiled-coil structure of TPM3. Biochemical and biophysical research communications. 2015;457(3):457-60.

81. Wray J, Williamson EA, Singh SB, Wu Y, Cogle CR, Weinstock DM, et al. PARP1 is required for chromosomal translocations. Blood. 2013;121(21):4359-65.

82. Weterman MJ, van Groningen JJ, Jansen A, van Kessel AG. Nuclear localization and transactivating capacities of the papillary renal cell carcinoma-associated TFE3 and PRCC (fusion) proteins. Oncogene. 2000;19(1):69-74.

83. Lazar V, Suo C, Orear C, van den Oord J, Balogh Z, Guegan J, et al. Integrated molecular portrait of non-small cell lung cancers. BMC medical genomics. 2013;6:53.

84. Horlings HM, Lai C, Nuyten DS, Halfwerk H, Kristel P, van

Beers E, et al. Integration of DNA copy number alterations and prognostic gene expression signatures in breast cancer patients. Clinical cancer research : an official journal of the American Association for Cancer Research. 2010;16(2):651-63.

85. Hickson ID. RecQ helicases: caretakers of the genome. Nature reviews Cancer. 2003;3(3):169-78.

86. Ratner N, Miller SJ. A RASopathy gene commonly mutated in cancer: the neurofibromatosis type 1 tumour suppressor. Nature reviews Cancer. 2015;15(5):290-301.

87. Fodde R, Kuipers J, Rosenberg C, Smits R, Kielman M, Gaspar C, et al. Mutations in the APC tumour suppressor gene cause chromosomal instability. Nature cell biology. 2001;3(4):433-8.

88. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature. 2011;469(7331):539-42.

89. Zhu X, He F, Zeng H, Ling S, Chen A, Wang Y, et al. Identification of functional cooperative mutations of SETD2 in human acute leukemia. Nature genetics. 2014;46(3):287-93.

90. Pils D, Horak P, Gleiss A, Sax C, Fabjani G, Moebus VJ, et al. Five genes from chromosomal band 8p22 are significantly down-regulated in ovarian carcinoma: N33 and EFA6R have a potential impact on overall survival. Cancer. 2005;104(11):2417-29.

91. Rotman G, Shiloh Y. ATM: from gene to function. Human molecular genetics. 1998;7(10):1555-63.

92. Silver DP, Livingston DM. Mechanisms of BRCA1 tumor suppression. Cancer discovery. 2012;2(8):679-84.

93. Goodrich DW. The retinoblastoma tumor-suppressor gene, the exception that proves the rule. Oncogene. 2006;25(38):5233-43.

94. Levine AJ, Momand J, Finlay CA. The p53 tumour suppressor gene. Nature. 1991;351(6326):453-6.

95. Xing M, Usadel H, Cohen Y, Tokumaru Y, Guo Z, Westra WB, et al. Methylation of the thyroid-stimulating hormone receptor gene in epithelial thyroid tumors: a marker of malignancy and a cause of gene silencing. Cancer research. 2003;63(9):2316-21.

96. Roy DM, Walsh LA, Chan TA. Driver mutations of cancer epigenomes. Protein & cell. 2014;5(4):265-96.

97. Brown VL, Harwood CA, Crook T, Cronin JG, Kelsell DP, Proby CM. p16INK4a and p14ARF tumor suppressor genes are commonly inactivated in cutaneous squamous cell carcinoma. The Journal of investigative dermatology. 2004;122(5):1284-92.

98. Pham TT, Angus SP, Johnson GL. MAP3K1: Genomic Alterations in Cancer and Function in Promoting Cell Survival or Apoptosis. Genes & cancer. 2013;4(11-12):419-26.

99. Rebouissou S, Amessou M, Couchy G, Poussin K, Imbeaud S,

Pilati C, et al. Frequent in-frame somatic deletions activate gp130 in inflammatory hepatocellular tumours. Nature. 2009;457(7226):200-4.

100. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research. 2015;43(Database issue):D447-52.

101. Chen B, Nelson DM, Sadovsky Y. N-myc down-regulated gene 1 modulates the response of term human trophoblasts to hypoxic injury. The Journal of biological chemistry. 2006;281(5):2764-72.

102. Stein S, Thomas EK, Herzog B, Westfall MD, Rocheleau JV, Jackson RS, 2nd, et al. NDRG1 is necessary for p53-dependent apoptosis. The Journal of biological chemistry. 2004;279(47):48930-40.

103. Speiser J, Foreman K, Drinka E, Godellas C, Perez C, Salhadar A, et al. Notch-1 and Notch-4 biomarker expression in triple-negative breast cancer. International journal of surgical pathology. 2012;20(2):139-45.

104. Nagamatsu I, Onishi H, Matsushita S, Kubo M, Kai M, Imaizumi A, et al. NOTCH4 is a potential therapeutic target for triple-negative breast cancer. Anticancer research. 2014;34(1):69-80.

105. Yamaguchi N, Oyama T, Ito E, Satoh H, Azuma S, Hayashi M, et al. NOTCH3 signaling pathway plays crucial roles in the proliferation of ErbB2-negative human breast cancer cells. Cancer research.

2008;68(6):1881-8.

106. Rahman S, Sowa ME, Ottinger M, Smith JA, Shi Y, Harper JW, et al. The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3. Molecular and cellular biology. 2011;31(13):2641-52.

107. Alsarraj J, Hunter KW. Bromodomain-Containing Protein 4: A Dynamic Regulator of Breast Cancer Metastasis through Modulation of the Extracellular Matrix. International journal of breast cancer. 2012;2012:670632.

108. Garcia-Murillas I, Sharpe R, Pearson A, Campbell J, Natrajan R, Ashworth A, et al. An siRNA screen identifies the GNAS locus as a driver in 20q amplified breast cancer. Oncogene. 2014;33(19):2478-86.

109. Pardee TS. Overexpression of MN1 confers resistance to chemotherapy, accelerates leukemia onset, and suppresses p53 and Bim induction. PloS one. 2012;7(8):e43185.

110. Van Dusen CM, Yee L, McNally LM, McNally MT. A glycine-rich domain of hnRNP H/F promotes nucleocytoplasmic shuttling and nuclear import through an interaction with transportin 1. Molecular and cellular biology. 2010;30(10):2552-62.

111. Pesiridis GS, Lee VM, Trojanowski JQ. Mutations in TDP-43 link glycine-rich domain functions to amyotrophic lateral sclerosis. Human molecular genetics. 2009;18(R2):R156-62.

112. McCormick C, Leduc Y, Martindale D, Mattison K, Esford LE, Dyer AP, et al. The putative tumour suppressor EXT1 alters the expression of cell-surface heparan sulfate. Nature genetics. 1998;19(2):158-61.

113. Reijmers RM, Groen RW, Rozemuller H, Kuil A, de Haan-Kramer A, Csikos T, et al. Targeting EXT1 reveals a crucial role for heparan sulfate in the growth of multiple myeloma. Blood. 2010;115(3):601-4.

114. Vicier C, Dieci MV, Arnedos M, Delaloge S, Viens P, Andre F. Clinical development of mTOR inhibitors in breast cancer. Breast cancer research : BCR. 2014;16(1):203.

115. Walsh S, Flanagan L, Quinn C, Evoy D, McDermott EW, Pierce A, et al. mTOR in breast cancer: differential expression in triple-negative and non-triple-negative tumors. Breast (Edinburgh, Scotland). 2012;21(2):178-82.

116. Li CG, Nyman JE, Braithwaite AW, Eccles MR. PAX8 promotes tumor cell growth by transcriptionally regulating E2F1 and stabilizing RB protein. Oncogene. 2011;30(48):4824-34.

117. Hayashi M, Nomoto S, Hishida M, Inokawa Y, Kanda M, Okamura Y, et al. Identification of the collagen type 1 alpha 1 gene (COL1A1) as a candidate survival-related factor associated with hepatocellular carcinoma. BMC cancer. 2014;14:108.

118. Ossandon FJ, Villarroel C, Aguayo F, Santibanez E, Oue N, Yasui W, et al. In silico analysis of gastric carcinoma Serial Analysis of Gene Expression libraries reveals different profiles associated with ethnicity. Molecular cancer. 2008;7:22.

119. Colavito SA, Zou MR, Yan Q, Nguyen DX, Stern DF. Significance of glioma-associated oncogene homolog 1 (GLI1) expression in claudin-low breast cancer and crosstalk with the nuclear factor kappa-light-chain-enhancer of activated B cells (NFkappaB) pathway. Breast cancer research : BCR. 2014;16(5):444.

120. Rao RC, Dou Y. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. Nature reviews Cancer. 2015;15(6):334-46.

121. Koren S, Reavie L, Couto JP, De Silva D, Stadler MB, Roloff T, et al. PIK3CA(H1047R) induces multipotency and multi-lineage mammary tumours. Nature. 2015;525(7567):114-8.

122. Van Keymeulen A, Lee MY, Ousset M, Brohee S, Rorive S, Giraddi RR, et al. Reactivation of multipotency by oncogenic PIK3CA induces breast tumour heterogeneity. Nature. 2015;525(7567):119-23.

123. Nagai MA, Gerhard R, Fregnani JH, Nonogaki S, Rierger RB, Netto MM, et al. Prognostic value of NDRG1 and SPARC protein expression in breast cancer patients. Breast cancer research and treatment. 2011;126(1):1-14.

124. Ellen TP, Ke Q, Zhang P, Costa M. NDRG1, a growth and

cancer related gene: regulation of gene expression and function in normal and disease states. Carcinogenesis. 2008;29(1):2-8.

125. Rahman S, Islam R. Mammalian Sirt1: insights on its biological functions. Cell communication and signaling : CCS. 2011;9:11.

126. Blander G, Zalle N, Leal JF, Bar-Or RL, Yu CE, Oren M. The Werner syndrome protein contributes to induction of p53 by DNA damage. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2000;14(14):2138-40.

127. Campaner S, Amati B. Two sides of the Myc-induced DNA damage response: from tumor suppression to tumor maintenance. Cell division. 2012;7(1):6.

128. Deng CX. BRCA1: cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. Nucleic acids research. 2006;34(5):1416-26.

129. Kim G, Ison G, McKee AE, Zhang H, Tang S, Gwise T, et al. FDA Approval Summary: Olaparib Monotherapy in Patients with Deleterious Germline BRCA-Mutated Advanced Ovarian Cancer Treated with Three or More Lines of Chemotherapy. Clinical cancer research : an official journal of the American Association for Cancer Research. 2015;21(19):4257-61.

130. McCormick C, Duncan G, Goutsos KT, Tufaro F. The putative tumor suppressors EXT1 and EXT2 form a stable complex that

accumulates in the Golgi apparatus and catalyzes the synthesis of heparan sulfate. Proceedings of the National Academy of Sciences of the United States of America. 2000;97(2):668-73.

국문초록

차세대 염기 서열법 기반의 유전체/후성유전체 분석 기법 확립과 한국인 삼중음성 유방암의 유전 변이 발굴

최근 차세대 염기 서열법 (Next Generation Sequencing, NGS) 기술이 개발되면서 저렴한 비용으로 방대한 양의 유전체 염기서열 분석이 가능하게 되었고, 이로 인해 전 세계적으로 유전체 연구가 매우 활발하게 이루어지고 있다. 따라서 곧 임상에서도 암을 포함한 여러 유전병 환자의 진단 및 치료에 NGS를 적용할 수 있을 것으로 보인다. 이러한 NGS는 단순한 염기서열 분석을 넘어서 다양한 유전체/후성유전체 분석에 적용될 수 있는데, 분석하고자 하는 유전자의 엑솜 지역만을 선별하여 분석하는 표적 엑솜 시퀀싱 (Targeted Exome Sequencing), mRNA의 발현 양상 및 접합 변형을 분석하기 위한 RNA 시퀀싱 (RNA-seq), DNA 메틸화와 히스톤 단백질 변이 등의 후성유전체 분석을 위한 중아황산염 시퀀싱 (Bisulfite Sequencing) 및 ChIP 시퀀싱 (Chromatin Immunoprecipitation sequencing, ChIP-seq) 등 NGS를 응용한 여러 방법을 통하여 유전체/후성유전체 연구를 전 유전체 수준 (genome-wide) 에서 분석할 수 있게 되었다. 기술이 발전하면서 NGS로 염기서열을 결과를 얻는 실험 자체에 대한 비용은 급격히 감소하고 있는 반면, 이로부터 얻어지는 방대한 양

의 데이터를 보관하기 위한 저장 장치와 데이터 분석에 필요한 고성능 컴퓨터 및 분석을 수행할 생물정보학 지식을 갖춘 인력에 대한 비용은 데이터의 양에 비례하여 증가하게 된다. 따라서 NGS를 이용해 유전체/후성유전체 데이터를 생산함에 있어서 무조건적으로 전장 유전체 분석(whole genome sequencing)을 수행하기 보다는, 연구의 목적에 맞추어 특정 지역만을 선별하여 NGS를 수행하고 결과를 분석하는 것이 시간과 비용을 줄이고 분석 결과의 품질을 높일 수 있는 효율적인 방법이라고 할 수 있다.

본 연구에서는 대표적인 후성유전체 기전이라고 할 수 있는 DNA 메틸화 분석 시, 기존의 메틸화 DNA 선별법 (Methylated DNA enrichment)과 중아황산염 시퀀싱을 결합한 MeDIP-BS (Methylated DNA Immunoprecipitation-Bisulfite Sequencing) 방법의 효율성을 검증하였다. 인간의 정상 간 조직과 위 조직에서 현재 일반적으로 사용되고 있는 전장 유전체 중아황산염 시퀀싱 (Whole Genome Bisulfite Sequencing, WG-BS), 표적 중아황산염 시퀀싱 (Targeted Bisulfite Sequencing, Targeted-BS), 메틸화 DNA 면역침강 시퀀싱 (Methylated DNA Immunoprecipitation Sequencing, MeDIP-seq) 방법과 MeDIP-BS 방법을 포함하여 총 네 가지 DNA 메틸화 분석 방법을 수행하여 각각 결과를 분석하였다. WG-BS 방법은 전장 유전체의 DNA 메틸화 경향을 단일 염기 수준에서 분석할 수 있는 방법으로, 여러 NGS 기반의 DNA 메틸화 분석법 중 현재 가장 표준적인 방법으로 생각되고 있다. 하지만 NGS 수행에 많은 비용이 들어가며, 또한 생산되는 데이터의 양이 너무 크기 때

문에 많은 수의 샘플을 사용한 연구에는 적합하지 않은 방법이라고 할 수 있다. MeDIP-BS는 5-methyl cytosine 항체를 이용하여 전장 유전체 중 DNA 메틸화가 되어있는 부분만을 우선적으로 선별한 뒤, 중아황산염 변환을 수행하고 NGS로 염기서열을 분석함으로써 DNA 메틸화 되어있는 지역에서 단일염기 수준으로 CpG site의 DNA 메틸화를 확인할 수 있는 방법이다. 본 연구 결과 MeDIP-BS의 데이터 생산량이 WG-BS의 5% 수준임에도 불구하고, 분석 결과의 연관성은 다른 방법들에 비해 가장 높은 것을 확인할 수 있었다 (r = 0.77). 즉, MeDIP-BS 방법이 WG-BS를 대체하여 전장유전체의 DNA 메틸화를 효율적으로 분석할 수 있는 방법임을 보여주었다.

유전체 분석에 있어서는 기존의 전장 엑솜 시퀀싱 (Whole Exome Sequencing)에서 더 나아가 분석하고자 하는 유전자만을 선별하여 분석할 수 있는 표적 엑솜 시퀀싱을 사용함으로써 더욱 효율적인 유전체 분석이 가능하게 되었다. 본 연구에서는 한국인 삼중음성유방암 70명 환자로부터 얻은 정상 조직과 암 조직을 사용하여, 고형암과 연관된 것으로 알려진 유전자 234개와 세포성장 및 키나아제 (kinase) 관련 전사 유전자 134개를 포함한 총 368개 유전자에 대해 표적 엑솜 시퀀싱을 수행하였다. 분석 결과 157개의 유전자에서 292개의 비유사 돌연변이 (non-synonymous SNVs)와 30 개의 작은 삽입/결실 (small Indel)이 발견되었다. 높은 빈도로 체세포 변이를 보이는 유전자로는 *TP53*, *NOTCH4*, *NOTCH3*, *GNAS*, *BRD4*, *MN1* 등이 있었다. 또한 유전자 복제수 변이 (Copy Number Variations, CNVs) 분석 결과 *NDRG1*, *UBT5*, *PTK2*,

RECQL4, *MYC* 등의 유전자 증폭 (gene amplification)이 발견되었으며, *WRN*, *IL6ST*, *APC*, *NF1*, *SETD2*, *PBRM1*, *PCM1*, *ATM* 등의 유전자 동형결실 (homozygous deletion)이 발견되었다. 특히 *WRN*, *ATM*과 같은 DNA 복구 유전자 (DNA repair gene) 동형결실을 보이는 환자에서 유의미하게 예후가 나빠지는 것을 확인함으로써, 해당 유전 변이가 삼중음성유방암 환자의 예후 예측에 사용될 수 있음을 보여주었다. 또한 유전성 유방암의 주요 유전자인 *BRCA1*와 *BRCA2* 유전자의 생식세포 돌연변이가 다수 발견되었으며, 위 결과들을 토대로 삼중음성유방암의 표적치료제로서 PARP 억제제가 사용될 수 있음을 확인하였다.

결론적으로 본 연구를 통해 차세대 염기 서열법을 기반으로 한 효율적인 유전체/후성유전체 분석 기법을 확립하였고, 이를 이용하여 삼중음성유방암의 진단 및 치료에 사용될 가능성을 가진 후보 바이오마커를 발굴할 수 있었다.

핵심어: 차세대염기서열법, DNA 메틸화, 메틸 DNA 면역침강 중아황산염 시퀀싱, 삼중음성유방암, 표적엑솜시퀀싱, 체세포 변이, DNA 복구 유전자

학번: 2009-21718