



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

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

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



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

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

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

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

[Disclaimer](#)

의학박사 학위논문

Genomic and transcriptomic analyses
reveal tandem *DHFR* gene amplification
and mismatch repair gene mutations
in methotrexate resistant HT-29

유전체 및 전사체 분석을 활용한
항암제 (MTX) 내성 HT-29 세포주의
tandem *DHFR* 유전자 증폭
특성 및 기전 연구

2019년 2월

서울대학교 대학원

의과학과 의과학 전공

김 아 름

유전체 및 전사체 분석을 활용한
항암제(MTX) 내성 HT-29 세포주의
tandem *DHFR* 유전자 증폭

특성 및 기전 연구

지도 교수 김종일

이 논문을 의학박사 학위논문으로 제출함

2018년 10월

서울대학교 대학원

의과학과 의과학 전공

김 아 름

김아름의 의학박사 학위논문을 인준함

2019년 1월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Genomic and transcriptomic analyses
reveal tandem *DHFR* gene amplification
and mismatch repair gene mutations
in methotrexate resistant HT-29

by

Ahreum Kim

A thesis submitted to the Department of Medicine in partial
fulfillment of the requirements for the Degree of Doctor of
Philosophy in Biomedical Science at Seoul National
University College of Medicine

January 2019

Approved by Thesis Committee:

Professor _____ Chairman

Professor _____ Vice chairman

Professor _____

Professor _____

Professor _____

Abstract

Genomic and transcriptomic analyses reveal tandem *DHFR* gene amplification and mismatch repair gene mutations in methotrexate resistant HT-29

Ahreum Kim

Major in Biomedical Science

Department of Biomedical Science

Seoul National University Graduate School

The massively parallel sequencing technology known as next-generation sequencing (NGS) has been currently developed and evolved for cancer genome research to obtain the molecular microscope findings and treatment of disease. The time and cost for NGS analysis have been greatly reduced, so the mechanisms from the basic mechanism of human evolution to the complicated mechanism underlying how genetic changes have driven the resistance of cancer cells under anti-cancer drugs have been comprehensively investigated

through advancements in NGS technologies. Therefore, the combination of these NGS technologies has contributed to cancer research such as diagnosis, management, and treatment by identifying and elucidating the molecular tumor profiling and it would play an important role in the future of cancer treatment and of personalized medicine in cancer research.

DHFR gene amplification is present in methotrexate (MTX) resistant colon cancer cells and in acute lymphoblastic leukemia. The region of chromosome 5q14 contains many genes as well as *DHFR* gene, and little is known about *DHFR* gene amplification at this position since quantifying amplification size and recognizing the involved repetitive rearrangements in gene amplification position require extra time and efforts with limited technologies and bioinformatics. Also, there is no clear way to assemble the complete structure of the amplified region with short read (read length < repeat length) as they cannot resolve repetitive regions or identify junction reads. Single molecule real-time (PacBio SMRT) sequencing and BioNano optical genome mapping (read length > repeat length), which provide exceptionally long read lengths, have the potential to overcome these limitations and allow for complete assembly of the region.

Here I have proposed an integrative framework to quantify the amplified region and detect structural variations, which are large, complex DNA segments involving repeats by using a combination of

technologies, including single molecule real-times sequencing, next generation optical mapping, and high throughput chromosome conformation capture (Hi-C).

The amplification units of 11 genes from *DHFR* gene to *ATP6AP1L* gene position on chromosome 5 (~2.2Mbp) and tandem gene amplification about twentyfold longer amplified region than control have been identified by several NGS technologies such as optical mapping and single molecule real-times sequencing, and its abnormally increased expression and complicated splicing patterns were characterized by RNA sequencing data. The novel inversion (chr5:80,618,750-80,631,409) at the *DHFR* gene of amplified region was detected which might stimulate chromosomal breakage for gene amplification

Using Hi-C technology, the high adjusted interaction frequencies which indicated the inter-chromosomal contact and significant adjusted p-value were detected on the amplified unit and unsuspected position on 5q in MTX resistant HT-29 sample compared to control. It might explain that chromosomal structure from the start position of the amplified unit (80.6Mb - 82.8Mb) to end of 5q (109Mb-138Mb) could have the complex network of spatial contacts to harbor the gene amplification. Also, the increased relative copy number, the several newly identified topologically associating domains (TADs), and extrachromosomal double minutes (DMs) on this amplified region, which were not detected

by other technologies, were identified and described for finding the association with the gene amplification mechanism.

Interestingly, the novel frameshift insertions in most of *MSH* and *MLH* genes were identified, which could cause the dysregulation of mismatch repair pathway under MTX condition and play an important role on the rapid progression of gene amplification as well as being resistant to MTX. Considering the several characteristics of variable size of tandem gene amplification patterns with homogeneously staining chromosome regions (HSRs), extrachromosomal DM suggested that the gene amplification might be produced from the Breakage-fusion-bridge (BFB) cycles.

Overall, the characterized tandem gene amplified unit, more complicated interaction on intra-chromosome 5, inversion of the amplification unit as well as the mutations in *MSH* and *MLH* genes can be the critical factor for identifying the mechanism of genomic rearrangements, and these findings may give new insight into the mechanism underlying the amplification process and evolution of resistance to drugs. Therefore, the comprehensive approach of combined advanced technologies is a powerful tool for interpretation of cancer genomes, and this will provide the depth of insight to identify the most important therapeutic mechanism and new targets of the anti-cancer drug.

Keywords: Gene amplification; Structural variation; Methotrexate;
Next generation technologies; Resistance to drugs; *DHFR* gene

Student number: 2015-22027

CONTENTS

Abstract	i
Contents	vi
List of Tables	vii
List of Figures	ix
List of Abbreviations	xiii
Introduction	1
Material and Methods	6
Results	28
Discussion	87
References	96
Abstract in Korean	106

LIST OF TABLES

Table 1. The estimation of copy number and expression of <i>DHFR</i> gene in MTX resistant clone (C1-2) at each cycle.....	44
Table 2. The comparison of detected structural variants between control and MTX resistant HT-29.....	45
Table 3. The detected variants on chromosome 5 in MTX resistant HT-29.....	46
Table 4. The identified genes with high segmented coverage (segmented coverage >20) on chromosome 5.....	47
Table 5. The comparison of mutations between control and MTX resistant HT-29 over whole chromosomes.....	49
Table 6. The comparison of mutations between control and MTX resistant HT-29 on chromosome 5.....	50
Table 7. The detection of novel frameshift insertions in MTX resistant HT-29 compared to control.....	51
Table 8. The comparison of alternative splicing patterns on the amplification unit between MTX resistant and control samples.....	52
Table 9. The differentially expressed genes (DEGs) in MTX resistant HT-29.....	53
Table 10. The topologically associating domains (TADs) with high intra-	

chromosomal interactions on chromosome 5.....54

LIST OF FIGURES

Figure 1. Schematic workflow.....	5
Figure 2. The morphological change of MTX resistant colon cancer cells	55
Figure 3. The <i>DHFR</i> gene expression and copy number among MTX resistant clone (C1-2), control, and reference.....	56
Figure 4. The visualization of the amplified <i>DHFR</i> gene at 5q arm in MTX resistant C1-2.	57
Figure 5. The patterns of amplified <i>DHFR</i> gene at 5q arm in MTX resistant C1-2.	58
Figure 6. The comparison of the <i>DHFR</i> gene amplification pattern and signal type of FISH between MTX resistant clone (C1-2-4) and control	59
Figure 7. Karyotyping in MTX resistant and control samples.....	60
Figure 8. The detection and characterization of SVs in MTX resistant HT-29	61
Figure 9. The visualization of inter-chromosomal genomic rearrangements in MTX resistant HT-29 and control samples	62
Figure 10. The comparison of segmented coverage over whole	

chromosomes between MTX resistant HT-29 and control samples.	63
Figure 11. The comparison of segmented coverage and structural variants over amplified region between MTX resistant HT-29 and control samples.....	64
Figure 12. The multiple-read view of alignment over amplified region with Ribbon	65
Figure 13. The single-read view and dot plot of the read over the amplified region.....	66
Figure 14. The scaffolding of PacBio long reads over amplified region compared to hg 38	67
Figure 15. The difference of non-synonymous mutations between MTX resistant HT-29 and control over whole chromosomes.....	68
Figure 16. The comparison of non-synonymous mutations between MTX resistant HT-29 and control on chromosome 5.....	69
Figure 17. The comparison of MMR expression level.....	70
Figure 18. The comparison of expression level between MTX resistant and control samples on chromosome 5.....	71
Figure 19. The expression level in 5q 14.2 region.....	72

Figure 20. The comparison of expression level in the amplified regions between MTX resistant and control samples.....	73
Figure 21. The visualization of mapped reads on the amplified region from transcriptome data	74
Figure 22. The identification of junctions between exons over amplification unit.	75
Figure 23. The comparison of five different alternative splicing events between MTX resistant and control samples.	76
Figure 24. The identification of differentially expressed genes and enrichment with KEGG pathways	77
Figure 25. Genome mapping over the amplified region.....	78
Figure 26. The detection of structural variants over the amplified region in genome mapping.....	79
Figure 27. Genome-wide view of intra-chromosomal interactions.....	80
Figure 28. Intra-chromosomal interactions on chromosome 5.....	81
Figure 29. The topologically associating domains (TADs) on chromosome 5.....	82
Figure 30. The topologically associating domains (TADs) on chromosome 5 and its adjusted interaction frequencies.....	83
Figure 31. The comparison of adjusted intra-chromosomal interactions	

between MTX resistant HT-29 and control samples.....84

Figure 32. The comparison of relative copy number between MTX resistant and control samples.....85

Figure 33. The mechanism of tandem gene amplification under MTX.....86

LIST OF ABBREVIATIONS

A3SS: Alternative 3' splice site

A5SS: Alternative 5' splice site

BAM: Binary alignment map

BFB: Breakage-fusion-bridge

CNV: Copy number variation

DEG: Differentially expressed gene

DEL: Deletion

DHFR: Dihydrofolate reductase

DUP: Duplication

FDR: False discovery rate

FISH: Fluorescent In Situ Hybridization

FPKM: Fragment per kilo base per million

GATK: Genome analysis toolkit

GSEA: Gene set enrichment analysis

Hi-C: High throughput chromosome conformation capture

IGV: Integrative genomics viewer

INV: Inversion

INVDUP: Inverted duplication

KEGG: Kyoto encyclopedia of genes and genomes.

Lsign: Sum of the sign of the entries in the lower triangle

Lvar: the variance of the lower triangle

MTX: Methotrexate

MXE: Mutually exclusive exon

ncRNA: non-coding RNA

NGS: Next generation sequencing

PacBio: Pacific Bioscience

RI: Retained intron

RNA-seq: RNA sequencing

SE: Skipped exon

SMRT: Single molecule real-time

SNV: Single nucleotide variation

Score: Conner score

STAR: Spliced transcripts alignment to a reference

SV: Structural variation

TAD: Topologically associating domains

Usign: Sum of the sign of the entries in the upper triangle

Uvar: The variance of the upper triangle

VSD: Variance stabilizing data

WGS: Whole-genome sequencing

Introduction

The cancer genome, which is too large and complicated to understand at a molecular level, has exploded in the recent years by multiple technological advances such as high-throughput sequencing, known as “next generation sequencing (NGS)” [1-3]. Developing NGS which have been greatly reduced cost and time for sequencing helps to identify the complicated mechanism of how the individual tumor cells have a unique set of genetic alterations and how genetic changes have driven the cancer adaption under anti-cancer drugs as well as the basic mechanism of human evolution and cancer development at the molecular level [4-6]. Therefore, realizing and utilizing NGS technology is necessary for developing cancer treatment by investigating the fundamental characteristics of tumor genome, and it has the potential to change the future of cancer treatment and advance the promise of personalized medicine in cancer research [7, 8].

Gene amplification, which is abnormal copy number increase in a specific region of genome under a selective condition, is predominant in human cancer and associated with overexpression of oncogenes such as *MYC*, *MYCN*, and *ERBB*, which lead to the abnormal cell proliferation and replication [9-11]. Its chance for gene amplification is more frequent than genomic mutation event in mammalian cells since its rate of gene

amplification is greater than mutation rate [12], so amplification often suggests a poor prognosis and the tumorigenic potential [13]. However, no molecularly targeted agents have been specifically developed for its treatments preventing gene amplification because of its chromosomal complexity and the technical limitations, so it is critical to find predictive and prognostic biomarkers for gene amplification specific medicine helping to improve the outcome of patients and optimize therapy decisions [14, 15].

In addition, gene amplification is regarded as the indicator of drug resistant sample in cancer and mammalian cells [16], so it will be important to identify the genetic features or pathways that promote amplification in tumors which might be a potential therapeutic target by preventing evolution of resistance to drugs, which is designed to arrest or eradicate the tumor [17]. However, drug resistant cells have high copies of a specific gene, but its mechanism is fully unknown at molecular level since genomic rearrangements and repetitive sequence have always presented technical challenges with sequence alignment and assembly programs, which is inaccessible by short reads previously [18, 19].

The *DHFR* gene amplification at chromosome 5 has been hallmark of methotrexate (MTX) responsiveness and resistance in colon cancer cells as well as acute lymphoblastic leukemia, which is known as an antifolate drug and inhibits dihydrofolate reductase (*DHFR*) by inhibiting

DNA synthesis and cell division [20-22]. It is previously well-known that the amplified *DHFR* gene generates two major DNA segments, extrachromosomal double minutes (DMs) and intra-chromosomal homogeneously staining chromosome regions (HSRs), but the molecular mechanism for these products from the amplified region and methods for detection and characterization are still unclear since they are detected only in advance stage tumors and accompany more repetitive and complicated sequences to assemble [23-25]. Therefore, identifying and deciphering the chromosomal abnormalities in gene amplification require many time, efforts, and a more detailed study to ascertain as well as technical support [26].

In this study, the combination of new technologies including single molecule real-time (PacBio SMRT) sequencing and optical genome mapping (BioNano Genomics) which generates the long ranged genomic data (reads size: ~10Kb), and high throughput chromosome conformation capture (Hi-C) for inter- and intra-chromosomal interactions have been used for identifying the involved repetitive rearrangements with amplified segments and interpreting gene amplification mechanism in MTX resistant colon cancer cell line (HT-29) (**Fig. 1**).

This study provide the optimized experimental methods to select the MTX resistant and homogeneously gene amplified samples from original HT-29, which have heterogeneously amplified genome. Also,

this allowed us to accurately quantify amplification size and recognize the drastic differences of chromosomal abnormalities and structural variants compared to MTX sensitive sample which are difficult to be extensively analyzed by the previous technologies such as short read sequencing and Fluorescence In Situ Hybridization (FISH) because of the technological limitations. Also, it would help to understand the principles of genome, impacts of genetic rearrangements on cancer cells and, by extension, drug resistance mechanism. Now, I shall introduce the characteristics of gene amplification and possible mechanism when the HT-29 colon cancer cells lines adapt to anti-cancer drug by using the comprehensive approach of combined advanced and new technologies.

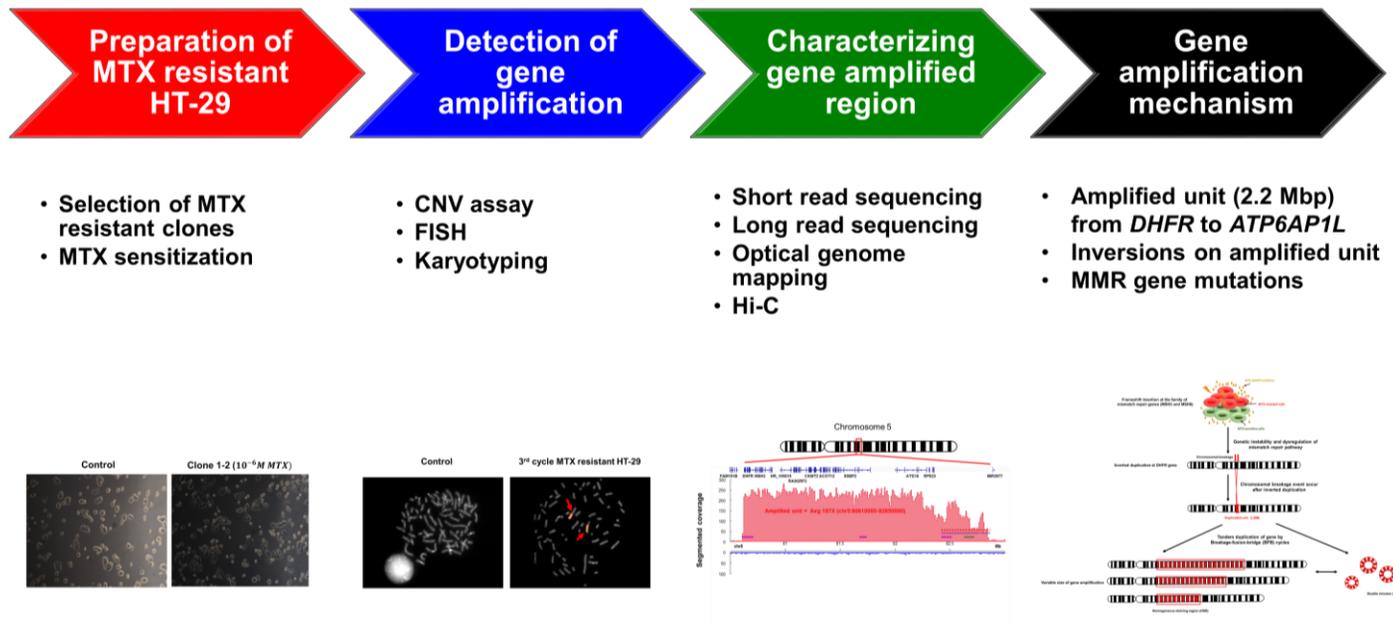


Figure 1. Schematic workflow. The experimental and computational analyses were performed on methotrexate resistant colon cancer cell line (HT-29) in order to characterize gene amplified region and understand the underlying mechanism.

MATERIALS AND METHODS

Preparation for MTX resistant HT-29

After curating a list of cancer cell lines based on the previous journal [24, 27-30], the human colon adenocarcinoma cell line HT-29 was chosen since it can be easily adapted to grow in high concentrations of MTX and concomitantly develop amplification of the *DHFR* gene. The HT-29 cell lines were targeted and maintained for developing the MTX resistant cancer cells as described previously [31]. For this study, the HT-29 cell lines were obtained from the Korean Cell Line Bank (KLCL) in Seoul National University Hospital.

Before using the MTX solution, the 100mg yellowish methotrexate hydrate powder (Tokyo chemical industrial M1664) was mixed with 1.967ml dimethyl sulfoxide (DMSO, Sigma-Aldrich) for 5 minutes by pulse-vortexing until there is no any residue at the bottom, and the solution was aliquoted to 100 microliters in 1ml Eppendorf tube and kept in -20°C for next use.

To generate MTX resistant HT-29, the IC₅₀ values was identified and the MTX concentration was optimized at several MTX concentrations from 10⁻⁸ mol/L to 10⁻² mol/L, and the increasing concentrations (from 10⁻⁸ mol/L to 10⁻⁶ mol/L) of this drug was added to a limited number of HT-29 cells (~3 × 10⁵) in five T25 cm² flasks with RPMI medium 1640

(Gibco) culture supplemented with 10% fetal bovine serum (Gibco) and 1% penicillin-streptomycin (10,000 U/mL, Gibco). The previous culture media was changed with ten times higher MTX concentration for every week.

After drug sensitization for three weeks, in order to detach the cells, the culture media was removed and the T25 cm² flasks was washed with the 1X PBS (Gibco) several times, and then 1ml of the 0.25% Trypsin – EDTA solution (Gibco) was added to each flask and incubated for 5 minutes. The 9 ml of 10⁻⁶ mol/L MTX culture media was added to detached cells, and centrifuge was used to spin down cells for 5 min at 1200 rpm. The media was aspirated off, and the 10 ml of 10⁻⁶ mol/L MTX culture media was added by pipetting up and down to achieve a single cell suspension.

The detached cells were counted by a hemocytometer, and the cells were serially diluted to get 1-2 cells/mL as the previous study described [32]. The 1 ml of diluted media was added to each well of 48-well cell culture plate [33]. The seeding density became 1-2 cells / well. The cells were grown under 10⁻⁶ mol/L MTX until the individual colonies of cells were big enough to see by naked eyes.

For picking each individual colonies of resistant cell, 3.2mm diameter cloning discs (Sigma) was used, which had been soaked in 0.25% trypsin EDTA, and 30 colonies were isolated from 48-wells, which were

well-separated with other colonies. The isolated colonies were transferred onto the T25 cm² for clonal expansion, and 3 isolated clones (C1-2,C4-3,C8-22) out of 30 clones, which were fast growing cells, were maintained in another T25 cm² flasks under the increasing concentrations (from 10⁻⁸ mol/L to 10⁻⁶ mol/L) of MTX as described above.

Sensitization studies for MTX resistant HT-29 clones

After obtaining the MTX resistant clones, the clones and control cells (HT-29) were passaged 30-35 times under MTX-free condition. For the second and third cycle of stepwise treatment of MTX concentration, 3 X 10⁵ cells of each clone as well as the parental HT-29 cells were sub-cultured in T25 cm² flasks. A stepwise treatment of MTX concentration from 10⁻⁸ mol/L to 10⁻⁶ mol/L was applied same as the first cycle did, and the MTX resistant cells, which were suffered from increased MTX concentration, were maintained under the 10⁻⁶ mol/L MTX concentration.

The morphological change of cells were assessed and captured by the light microscope detection with the several resolutions for every concentration and cycles. The 1-2 cells/mL from the diluted C1-2 colonies, which had most fast-growth rate and high *DHFR* copy numbers, were transferred to another T25 cm² flasks by using the cloning disc method and sub-cultured in order to reduce the heterogeneity of

amplification patterns.

RNA extraction

In this work, genomic RNA was extracted using TRIzol (Invitrogen) following manufacturer's protocol [34]. The $\sim 2 \times 10^5$ cells were lysed and homogenized by adding 0.3 ml of TRIzol accompanying with the incubation for 5 minutes. The 0.2 ml chloroform (Amresco) was added for lysis and incubated for 3 minutes. The samples were centrifuged for 15 minutes at 12,000 X g at 4 °C. The colorless aqueous phase was transferred into a new tube, and the 0.5ml of isopropanol (Emsure) was added to the tube. After incubation for 10 minutes, it was centrifuged for 10 minutes at 12,000 X g at 4 °C, and then the supernatant was removed.

The pellet was re-suspended with 75% ethanol (Emsure), and the supernatant was removed after centrifuging for 5 minutes at 7500 X g at 4 °C. The pellet was dried by vacuuming. The total RNA at the bottom of the tube was re-suspended with 20 μ L of RNase-free water and incubated at 55 °C for 10 minutes. The RNA yield was determined by Nanodrop spectrophotometer (ND-1000).

Reverse transcriptase PCR analysis

For reverse transcription, the 5µg total RNA was used for SuperScript First-Strand Synthesis System (Invitrogen) using oligo (dT), following manufacturer's protocol. For each reaction, 5 µl RNA, 1µl 10mM dNTP mix, 1µl primer, and 3µl DEPC-treated water were mixed in 0.5ml tube.

The mixture was incubated at 65°C for 5 minutes and placed on ice for 5 minutes. The 9 µl of prepared reaction mix (2 µl 10X RT buffer, 4 µl 25mM MgCl₂, 2 µl 0.1M DTT, 1 µl RNaseOut) was added to each tube and incubated at 42 °C for 2 minutes. 1 µl of SuperScript II RT was added to each tube and incubated at 42 °C for 50 minutes. The mixture was placed on 70 °C for 15 minutes and chilled on ice. 1 µl of RNase H was added to the collected reaction tube, and it was stored at -20 °C.

Real Time qPCR analysis

DHFR gene, *hMSH3*, *RASGRF2*, and *B2M* were investigated in MTX resistant HT-29, control (HT-29), and reference samples (NA19982) using RT-qPCR.

The TaKaRa EX HS was used for real-time qPCR using probe detection. The general reaction mixture (0.25 µl Ex Taq HS, 5 µl of 10X

Ex Taq buffer, 4 µl dNTP mixture, 200ng template, 50 pmol primer1 and primer 2 and sterile distilled water up to 50 µl) was prepared as described in the product manual. The PCR primers were as follows: DHFR-F 5'-GGGGTTTTCCATAGTCA-3' and DHFR-R 5'-GCCTCCAGTTTGCTTAC-3', hMSH3-F 5'-ATTTTAGAAGGGGTGGTG-3' and hMSH3-R 5'-TTAGGGGAAATTTAGATGCT-3', RASGRF2-F 5'-ATTTTGATTGAGAGGGAAGTA-3' and RASGRF2-R 5'-CAAGTTGATGTCGGAGTT-3', B2M-F 5'-CCAAGTCACGGTTTATTCT-3' and B2M-R 5'-TATTGCCAGGGTATTTCA-3', RASGRF2-2-F 5'-TGGGGAGGGAAATAGAC-3' and RASGRF2-2-R 5'-CTGCAGGAGGGTTACAA-3'. 1.0 µM of primers and 200 ng of total RNA was used for each 50 µl reaction. For the reactions, the standard protocol parameter [30 cycles (95°C – 10 sec, 55°C – 30 sec and 72°C – 1 min)] was used.

Agarose electrophoresis and gel visualization

The 1% agarose gel (Bio-Rad) was prepared by mixing 1 g of agarose and 100ml of 1X TBE in Erlenmeyer flask, and the mixture was heated by the microwave for 30 seconds. The cooled gel was poured into the casting chamber, and the casting combs were added into the appropriate slots. The gel was placed into the gel rig with the 1X TBE, which was sufficient to cover the gel enough.

Once the combs were removed, the 5 μ l of products with addition of 1 μ l of Dye loading STAR and DNA ladder were loaded on 1% agarose gel. The samples were electrophoresed at 80v for 75 minutes. The gels were transferred onto the UV light box, and the band image was captured. The target gene expression was estimated by Gene Analyzer program.

DNA extraction

The gDNA was extracted by using QIAamp DNA Mini Kit (Qiagen) and protocol, which was provided in Qiagen website. 1×10^6 cells, which was grown in a monolayer, were detached by trypsinization as described above.

The detached cells were collected in 10ml culture medium and centrifuged at 300 Xg for 5 minutes. The pellet were re-suspended in 200 μ l PBS of 1.5 mL micro-centrifuge tube. The 20 μ l of QIAGEN protease K and 200 μ l buffer AL were added to sample in order. The mixed sample was incubated at 56°C for 11 minutes. The 200 μ l of 100% ethanol (Emsure) was added to the sample, and well-mixed sample was applied into the Mini spin column and centrifuged at 6000 Xg for a minute.

The 500 µl Buffer AW1 was added into the mini spin column, which was placed in a clean collection tube, and the column was placed on the centrifuge at 6000 Xg for 1 min. The 500 µl Buffer AW2 was added to the column in a clean collection tube, and it was centrifuged at full speed for 1 min. The 200 µl Buffer AE was added into the mini spin column in a clean 1.5 ml micro-centrifuge tube, and it was placed at room temperature for a minute. The gDNA was collected from the mini column by centrifuging at 6000 Xg for 1 min. The quality and quantity of extracted gDNA were measured by Nanodrop spectrophotometer (ND-1000).

Detection of *DHFR* gene amplification

The copy number of three MTX resistant clones, which was suffered from step-wise MTX concentration adjustment, was measured by TaqMan Copy Number Reference Assays (Assay ID: Hs02208275_cn, Gene: dihydrofolate reductase (DHFR), Assay Location: chr5:79929803) and Vii7 technology to select the clone with highly amplified *DHFR* genes from clones.

For performing copy number assays, TaqMan copy number assays protocols, which were provided by AppliedBiosystem, were followed (<http://docs.appliedbiosystems.com/>). The previously extracted gDNA

was diluted to make 5 ng/μL by using nuclease-free water. The final reaction volume was 20 μL by adding 10 μL of 2X genotyping Master Mix, 1 μL copy number assay, 1 μL copy number reference assay, and 4 μL nuclease-free water into the prepared 4 μL of the diluted g DNA samples. The mixture was loaded in the 96-well reaction plates.

The mixtures in the 96-well reaction plates were loaded in to the real-time PCR instrument (ViiA 7 real-time PCR system), and we performed the running process the plate using the parameter [2 stages: Hold (95°C – 10 min) and 40 cycles (95°C – 15 sec and 60°C – 60 sec)]

The copy number of the *DHFR* gene was estimated by ViiA™ 7 Real-Time PCR System (ViiA™ 7 Software v1.X; 7500 Software v2.0). The relative quantitation (RQ) was computed by the measured comparative Ct method, and it was computed by the Ct difference (Δ Ct) between MTX resistant samples and reference sample (NA19982). The Δ Ct values of MTX resistant sample to the reference sample, which have two copies of the *DHFR* gene, were calculated by multiplying two by the relative quantity. There were three equations involved in the computing the copy number from the Ct value.

$$i. Expression = 2^{-((Sample Ct - Reference Ct) - (Control Ct - Reference Ct))}$$

ii. *Rate between reference and sample*

$$= \frac{\text{Expression for Sample}}{\text{Expression for Reference}}$$

iii. *Copy number = 2 copies (reference) × Rate*

Fluorescence in situ hybridization and karyotyping

The position of *DHFR* gene from (C1-2, C4-3, C8-22) clones, which had a relative high growth rate compared to other clones, were detected by Fluorescence in situ hybridization (FISH), and the additional FISH was performed on ten sub-clones originated from C1-2, which had highest *DHFR* gene copies, to choose the clone which had most homogeneous amplified pattern. FISH and karyotyping were performed by MacroGen Ltd (Seoul, Korea).

For FISH analysis, the cultured cells of C1-2, C1-2-4, HT-29, and several clones were hybridized with a biotinylated a satellite probe specific for *DHFR* at chromosome 5. The FISH Tag DNA red detection kit (Invitrogen) was used for FISH analysis. The manual provided in the product's website was followed.

Before beginning, the binding buffer was prepared by adding 4 ml of 100% isopropanol in the 6ml of Component B. The wash buffer was ready to use by adding 11.8ml of 100% ethanol to 3.2ml of Component C Sodium. Also, bicarbonate solution was made by adding 1ml of

nuclease-free water into the sodium bicarbonate powder.

The working solution of DNase I was prepared by using the solution in kit. The nick translation reaction was done by mixing several components such as 5 μ L 10X nick translation buffer, 5 μ L 0.1M DTT, 5 μ L 10X DNA nucleotide mix, 1 μ L DNA, 1.7 μ L DNA polymerase I, 4 μ L DNase I working solution. After incubation at 15 $^{\circ}$ C for 2 hours, the 50 μ L nuclease-free water was added to the solution and vortexed for 10 seconds.

For purifying the amine-modified DNA, the synthesis reaction in addition with 400 μ L binding buffer was transferred to the spin column with a collection tube, and it was centrifuged at 10,000 X g for a minute. 650 μ L of wash buffer was added to the column and centrifuged again at 10,000 X g for a minute. The elution buffer was applied to the spin column in the new collection tube and centrifuged the column at 10,000 X g for a minute.

To elute the DNA from the collection tube, 10 μ L of 3M sodium acetate, 1 μ L of glycogen, 39 μ L of nuclease-free water added to the collection tube. In addition of 250 μ L of 100% ethanol and incubation at -20 $^{\circ}$ C for 30 minutes, the pellet was collected by the centrifugation at 10,000 X g for 10 minutes. The pallet was washed with the 400 μ L of 70% ethanol and dried out. The pellet was mixed with the 5 μ L nuclease-free water

and incubated at 37 °C for 5 minutes.

The determined and adjusted concentration of the sample was labeled with the Amine-modified DNA with fluorescent dye. The 1 µg of DNA was denatured by incubation at 96°C for 5 minutes and centrifuged for 10,000 X g for 3 minutes. The 3 µL of sodium bicarbonate was added to the sample, and the reactive dye was re-suspended in the 2 µL of DMSO and vortexed. The 2 µL of reactive dye was added to DNA sample and vortexed. The sample was incubated for an hour under the light-off, and the 90 µL water was added.

For the purifying the fluorescent dye-labeled DNA, 400 µL of binding buffer was added to the labeling reaction, and it was transferred to the spin column with a collection tube and centrifuged at 10,000 X g for a minute. The 650 µL of wash buffer was added to the column and centrifuged at 10,000 X g for a minute. The spin column was placed in a clean collection tube, and 55 µL elution buffer was added to the column and centrifuged at 10,000 X g for a minute.

The purified fluorescent dye-labeled DNA was eluted by adding 10 µL of 3M sodium acetate, 1 µL of glycogen, 39 µL of nuclease free water. The 250 µL 100% ethanol was added and incubated at -20 °C for 30 minutes, and centrifuged at 10,000 X g for 10 minutes. The pellet was collected, added with 400 µL of 70% ethanol, and finally dried out. The

10 μ L of nuclease free water was added and incubated at 37 °C for 5 minutes.

For the hybridization, the standard procedures to produce metaphase chromosome was followed as previously described [35, 36]. The *DHFR* red signal and 5p12-green signal were identified and counted at the anaphase and metaphase with the 1000x magnification and triple (RGB) filter.

Karyotyping was performed to visualize and map *DHFR* genes and to detect chromosomal abnormalities. Standard procedures were followed as previously described without modification [37]. The new media with colcemid (Sigma), which had the 0.02 mg/ml final concentration, was added to the cultured cells and incubated for an hour.

The slides were prepared and fixed by adding the methanol : acetic acid (3:1) and placed on the ice. The cells were washed with the 2XPBS and added with the DMEM plus serum. The pallet was collected after centrifugation at 1000 rpm for 3 minutes and added with 1 ml 0.56% KCL dropwise, and added with the 4 ml of 0.56 % KCL. The cells were incubated at room temperature for 6 minutes and collected with the centrifugation at 5000 rpm for 5 minutes.

The 1ml of methanol: acetic acid (3:1) was added to the pellet for fixing, and additional 4 ml of fix solution was added to cells. The fixing process was done several times and ended up with 1 ml total volume.

The water in slides was removed, and the cell suspension was dropped on the surface of the slide and dried. The slides were stained with Giemsa (Sigma GS-500) for 10 minutes and washed with the water. The image of chromosome was captured under microscope.

RNA-seq analysis and transcriptome profiling

The RNA-seq for MTX resistant HT-29 and control samples was performed to compute the gene expression and investigate the differentially expressed genes. Total RNA from $\sim 1 \times 10^6$ cells in T25 cm² flask was extracted and purified by using RNAiso Plus (Takara Bio Inc.) and RNeasy MinElute (Qiagen Inc.) respectively. The quality and quantity were identified by 6000 Nano LabChip, and the libraries for each samples were prepared as previously described [38].

The libraries were sequenced by Illumina HiSeq 2000 to obtain ideal coverage (depth 100X) followed a previous study [39]. The obtained reads were mapped to the human reference genome (GRCh38) by using the Spliced Transcripts Alignment to a Reference (STAR) tool to produce analysis-ready BAM files. The key principles of the processing and analysis steps such as sorting, mark duplicated, Split'N'CigarReads, and mapping quality was followed in GATK

website provided. The mapped reads (BAM file) was visualized by the mapped sequence analysis tool (SeqMonk version 1.42.0).

To estimate the expression of each gene, the raw reads were counted by HTSeq-0.6.1 tool and normalized to Variance Stabilizing Data (VSD) expression by R-3.3.0 package 'DESeq2'. Fragments Per Kilobase Million (FPKM) values were calculated by R package 'edge R' and converted to \log_2 values [40]. The median centered gene expression was computed by Cluster 3.0 software from FPKM expression, which subtracts the row-wise median from the expression values in each row. The median centered VSD and FPKM expression were visualized in a heatmap by Java Treeview.

Variant discovery analysis

Variant calling was performed on transcriptome datasets. The duplicated sites from analysis-ready BAM files was filtered with Picard program, and variants were called and filtered by removing spurious and known RNA-editing sites in VCF format. The variant discovery analysis was performed by observing the step-by-step recommendations which were provided by Genome Analysis Toolkit (GATK) to obtain high quality variants [41, 42].

In order to determine the exact single nucleotide polymorphisms (SNVs) from the call set, the variants were filtered out according to the several conditions. First, the cut-off for quality by depth (QD) was 3.0, which was the variant confidence score divided by the unfiltered depth of coverage, and the variants were filtered out less than 3.0.

Second, the variants were filtered out when Fisher Strand (FS) was >30.0 , which indicated the Phred-scaled p -value using Fisher's Exact test for detecting strand bias [43]. The identified and filtered variants were annotated by using RefSeq genes and the ANNOVAR tool. The identified non-synonymous variants were compared between control and MTX resistant HT-29 samples.

Differentially expressed genes (DEGs) analysis

Differentially expressed genes between MTX resistant HT-29 and control (HT-29) samples were analyzed by R packages 'DESeq2' and 'edgeR' with the specific criteria (P -value < 0.05 , $|\text{Log}_2(\text{fold change})| \geq 1$, and $\text{baseMean} \geq 100$) from the computed ht-seq raw read counts. The differentially expressed genes which were mainly expressed in each MTX resistant HT-29 and control samples were used for enrichment analysis with Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets via Gene Set Enrichment Analysis (GSEA).

Alternative splicing event analysis

The junctions between exons from the mapped reads (BAM file) were visualized by Sashimi plot from Integrative Genomics Viewer (IGV 2.3.63). The exon-inclusion levels, which was defined with junction reads from RNA sequencing results, was subsequently processed by rMATS.3.2.5 [44]. The five different types of alternative splicing event (SE:Skipped exon, MXE: Mutually exclusive exon, A5SS: Alternative 5' splice site, A3SS: Alternative 3' splice site, and RI: Retained intron) were identified and compared between MTX resistant HT-29 and control samples. The number of significant events was selected by using both junction counts and reads on target.

PacBio long read sequencing analysis

Genomic DNA was extracted from MTX resistant HT-29 (C1-2-4) and control sample (HT-29) by using the Gentra Puregene Cell kit (Qiagen). The library for PacBio was prepared under Pacific Biosciences recommended procedure. The overall procedure for PacBio data generation followed a previous study [45]. The obtained PacBio long reads were aligned to the human genome (version GRCh38) with BWA-mem aligner, and the pre-processing pipeline on BWA-mem

website was followed for both technically and biologically high quality.

The segmented coverage was estimated by depth of coverage option of GATK and pre-processed BAM files, and the coverage difference between MTX resistant HT-29 sample and control, which was bigger than 20, was selected for determining the amplified region, and the amplified region was annotated with the gene symbol and position.

Detection of genomic variants and amplification units

The structural variants (deletion, duplication, inverted duplication, translocation, and inversion) in MTX resistant HT-29 and control samples were analyzed by Sniffles from sorted PacBio output (BAM) [26]. The BAM files were converted to binned copy numbers across a genome by Copycat.

The detected genomic rearrangements were visualized by SplitThreader software (<http://splitthreader.com/>) from VCF files (Sniffles) and read coverage files (Copycat) [46]. Also, the single read view and multiple reads view of alignment results with structural variants were displayed by the online visualization tool “Ribbon”(<http://genomeribbon.com/>) [47].

Optical genome mapping

Optical mapping with the PacBio assembly data was performed by BioNano assembler software (Irys System, BioNano Genomics) for scaffolding and more accurate sequence. The DNA was extracted by IrysPrep Plug Lysis Long DNA Isolation Protocol which was provided in Bionano Genomics [48].

The cultured HT-29 cells in 100mm dishes were detached by trypsin for cell counting with a hemocytometer, and the counted cells were washed by 1XPBS (Gibco). Using the Bio-Rad Plug Lysis Kit, the detached cells were put into the agarose plugs, and it was done by proteinase K digestion. After TE (Tris-EDTA) washing, the GELase enzyme was used in plugs in order to cells be melted (Epicenter).

After that, the drop dialysis was performed in the DNA, and the amount was measured by the Qubit Broad Range dsDNA Assay Kit (Thermo Fisher Scientific) after the equilibration for four days. The sing-strand nicking was applied on the 200–300 ng/μL of DNA by using Nt.BspQI nickase (New England BioLabs) from IrysPrep NLRS assay (Bionano Genomics). The nicked sites were dyed with the YOYO-1 for labeling and added into the IrysChip. The several cycles were done for obtaining the high depth of coverage. Also, the tandem repeats on amplified region in both the assembly and the raw data were identified by IrysView 2.0 software.

Preparation for high throughput chromosome conformation capture (Hi-C)

Approximately 50 million cells of MTX resistant HT-29 and control cells were used for obtaining high-throughput chromatin conformation capture (Hi-C) data sets. Two Hi-C libraries were generated by using the restriction enzyme “HindIII” as previously reported Hi-C protocol without modifications [49].

The counted cells were crosslinked with 37% formaldehyde (Sigma) and 2.5M glycine as previously described in Hi-C protocol [50], and DNA was digested with the HindIII enzyme followed by filling the 5'-overhangs. The nucleotide was labeled with biotin, and blunt end ligation was performed on the product. The biotin-labeled junctions were captured by streptavidin-coated beads, and the valid products were sequenced with paired-end by using Hi-Seq2000.

Hi-C data analysis

The raw Hi-C data files (fastq) were processed to the normalized contact matrices by HiC-Pro version 2.10.0[51]. The pipeline based on the bowtie 2 aligner and selected restriction enzyme (HindIII) was used to generate normalized contact maps as described in HiC-Pro pipeline (<https://github.com/nservant/HiC-Pro>).

Each aligned reads were assessed to determine the valid interactions and control its quality by excluding the invalid ligation products and the duplicated valid pairs. The aligned Hi-C SAM file was converted into the HiCnv format, which called copy number variations (CNVs) from Hi-C data [52]. Also, the inter-chromosomal translocations and their boundaries were detected by HiCtrans from Hi-C matrix file. The list of valid interaction output files called by HiC-pro were converted to juicebox input file and visualized by Juicebox (<https://github.com/theaidenlab/juicebox/wiki>). The topologically associating domains were identified by Arrowhead algorithm [53].

The R packages 'HiCcompare' was used to detect the differential spatial chromatic interaction on a genome-wide scale between control and MTX resistant HT-29 [54]. Using this package, the adjusted interaction frequencies were represented by adjusting joint normalization function with the adjusted p value and filtering the low average expression, which was applied by the multiple testing correction.

Statistical test

All statistical analyses were performed using R-3.3.0. The gene expression between MTX resistant HT-29 and control (HT-29) samples were compared, and the p-value was indicated by using the unpaired Student's t test or Mann-Whitney U test based on the Shapiro-Wilk

normality test. P value was less than 0.05 was considered to be a statistically significant result.

RESULTS

Determination of MTX resistant HT-29 cells

While generating the MTX resistant HT-29 cells from single cell selection and MTX sensitization as previously described [23], the dramatically morphological change was detected from rounded and circular shapes at 1st cycle to rod to irregular shape at the 2nd cycle on the selected MTX resistant clone (C1-2) (**Fig. 2**). The shape was changed back to the original shape at 3rd cycle of sensitization, which might indicate that the HT-29 cells started to be resistant and rapidly grew up under high MTX concentration.

As expected, the expression of *DHFR* gene, which was normalized by the *B2M* house keeping gene, in clones as well as C1-2 was steadily increased from 1st cycle to 3rd cycle as the HT-29 clones became resistant to MTX (**Fig. 3a**) After confirming the morphological change and increased *DHFR* expression, the copy number of *DHFR* gene in several clones was measured.

The cycle quantification values (Ct) for clones at each cycle were initially measured and used for computing rate of copy number. The detected Ct values were dropped from 26.04 to 19.99 as the number of cycle increased (**Table. 1**). Interestingly, there was a dramatic increase

of *DHFR* gene copy number between 1st cycle (0.97 copies) and 2nd cycle (54.83 copies) (**Fig. 3b**).

From the result of morphological change and quantification of *DHFR* gene expression and copy number, it was assured that the specific time period and certain condition were required to survive under MTX condition, and the amplified *DHFR* gene could be shown as the indicator of MTX resistance at the specific time point after suffering the harsh MTX condition as previously described [20]. Also, it was confirmed that the increase in *DHFR* gene expression was not proportional to the increase in *DHFR* gene copy number in each cycles [55].

Validation of gene amplification in MTX resistant HT-29

After quantifying the *DHFR* gene copy number, the amplified *DHFR* gene at 5q arm was visualized by Fluorescent In Situ Hybridization (FISH) on MTX resistant clone (C1-2) and control sample. It was found that *DHFR* gene region at 5q arm was abnormally stretched compared to control at the 2nd and 3rd cycle as expected (**Fig. 4**).

When the patterns of amplified *DHFR* gene at metaphase in C1-2 were identified, *DHFR* gene amplified region had not the spot signal but painting signal, and the FISH signal patterns were highly heterogeneous, which represented total 9 signal amplified patterns and two major patterns accounted for 44 % and 28 %, respectively (**Fig. 5**). Even if the

cells were suffered from the same condition simultaneously, the MTX effects and genetic status were different to each cell, and this would result in the several difficulties in further technical analysis from the alignment to analysis of amplified patterns [56, 57]. Therefore, optimizing and generating the homogenous amplified patterns were required before sequencing, and optimized *DHFR* gene amplified patterns were established by making the serial clone selection and sub-culturing.

The established and optimized C1-2 clone (C1-2-4), which was long-lasting under MTX condition, was visualized by FISH, and four different types of gene amplification pattern were detected; two patterns had amplified *DHFR* genes at two q arms (75 % and 12.5 %), another pattern had amplified *DHFR* genes at three q arms (8.3 %), and the other had no amplified *DHFR* genes at q arms (4.2 %) (**Fig. 6**). Overall, ~96% of cells had the *DHFR* amplified region.

From this result, it was confirmed that the majority of them was resistant to MTX and had high amplification of *DHFR* genes at 5q arm. However, this also explained that each cells in MTX resistant clone (C1-2-4) had the different *DHFR* gene amplified pattern and copy number, and they had still heterogeneous genetic status under MTX condition even if the clone which was generated by the only one cell of the MTX resistant cells experimentally [58].

Additionally, MTX resistant HT-29 and control samples were karyotyped for accurately detecting the amplified region (**Fig. 7**). Karyotype of MTX resistant clone (C1-2-4) revealed the chromosomal abnormality such as homogeneously staining region (HSR) at chromosome 5 q arm as previously described [59], but the abnormal stretch of 17q arm was also detected, not coincidentally. This results confirmed that the aneuploidy and chromosomal instability of cancer samples could cause the many changes in chromosome copy number and karyotype diversity [60]. Also, it was previously found that chromosome 17q arm amplification could be detected because of the genetic instability in colorectal cancer [61].

Analysis of the structural variants and amplification unit

After confirming the HSR on the *DHFR* gene amplified region, the five genomic structural variants (deletion, duplication, inverted duplication, translocation, and inversion) and amplified units of MTX resistant cells were analyzed to identify which genes and structural variants were involved in the gene amplification.

The total number of genomic variants in MTX resistant sample was bigger than in control, and the size of variants such as duplication and inversion were bigger in MTX resistant sample with the large number of split reads (**Fig. 8**). Also, more variants for each structural variants were

detected in MTX resistant HT-29 compared to control.

The novel structural variants on chromosome 5 were selectively compared between control and MTX resistant HT-29, and one duplication (the median size of split reads : 371,675), three Inversions (the median size of split reads : 328,697), and three inverted duplication (the median size of split reads : 1,529) were detected compared to control, which had no any variants (**Table. 2**). Also, the number of split reads was bigger than the average coverage (10X), and most CNV categories were matching, which indicated that the detected structural variants were accurately detected (**Table. 3**). While the number of translocation on chromosome 5 was decreased in MTX resistant sample compared to control, there were more detected inter-chromosomal genomic rearrangements in MTX resistant HT-29 sample (**Fig. 9**).

The \log_2 ratio of segmented coverage over whole chromosomes between control and MTX resistant sample was compared, and the high segmented coverage in MTX resistant sample was observed in chromosome 5 compared to control (**Fig. 10**). The genes which the segmented coverage was bigger than 20, was identified and annotated with its position to identify the exact amplified region which included the *DHFR* gene (**Table. 4**).

The position of amplified region was on chromosome 5 q arm around 80M to 83M, which included from *DHFR* gene as the start point to

ATP6AP1L gene, and its segmented coverage was approximately 197x between 80.6M and 82.8M (~2.2M) (**Fig. 11**). This region was regarded as the amplified unit and the MTX resistant sample had about twentyfold longer amplified region than control. This amplified size was inferred from the originally designed coverage of long read sequencing, which was 10X and the control sample exactly had, but the MTX resistant sample had abnormal high coverage (~197X) on only this region not on the other regions.

Interestingly, it was found that there were both forward and reverse strands in MTX resistant sample compared to control, which had only forward strand and small deletion on our defined amplified unit (**Fig. 12**). Originally, it has been known that the forward DNA synthesis is preferred during replication [62]. However, this result indicated that the replication was performed in the both forward and reverse direction in MTX resistant sample, and this finding was confirmed by previous article that the DNA can be replicated in the reverse direction by a backward enzyme Thg1-like proteins (TLPs) to efficiently synthesize both chains [63].

Overall, the amplification regions included the 11 genes from *DHFR* gene to *ATP6AP1L*, which were tandem gene amplification of this region. This region had inversion or inverted duplication at the end of the amplification unit, and it seemed that the tandem repeats of several genes on chr5 (2.2Mbp) were initiated and ended by inversion on the

specific sequence (**Fig. 13**). The inversion (chr5:80,618,750-80,631,409) at the start point included the *LINC01137*, *DHFR*, and *CTC325J23.2* genes (**Fig. 14**). The inverted region could lead to the genetic instability and finally stimulate chromosomal breakage for gene amplification [64]. The detected novel inversion on the amplified region in MTX resistant HT-29 sample was resulted from the palindromic sequences, which were the foundation for the several amplification mechanisms as previously described [65].

Identifying the novel mutations and its impact on gene amplification

In order to find the involved mutations on amplification mechanism, the SNVs in both samples were identified by using the transcriptome sequencing data. There were more total exonic mutations in MTX resistant HT-29 (13,982) compared to control (13,310) over all chromosomes, and 18 more exonic mutations were detected in MTX resistant HT-29 on chromosome 5 only (**Tables. 5 and 6**).

After removing the synonymous SNV, it was found that there were a few additional non-synonymous mutations (non-synonymous SNV, frameshift deletion, frameshift insertion, stop-gain, and stop-loss) in MTX resistant HT-29 than control over all chromosomes (**Fig. 15**). A few additional non-synonymous mutations in MTX resistant sample came

from chromosome 5, which the number and percentage of frameshift insertions noticeably increased in MTX resistant sample (18.1%) than in control (4.3 %) (**Fig. 16**). This might explain that the detected frameshift insertions in mRNA on chromosome 5 have concurrently occurred with the *DHFR* gene amplification to survival under MTX toxicity.

When the role of detected frameshift insertions was identified, it was recognized that the novel frameshift insertions, which the single nucleotide insertion was adenine (A) or thymine (T), were located on *MSH3* and *MSH6* genes as well as *PMS1* and *PMS2* genes in MTX resistant HT-29 sample only (**Table. 7**). The expressions of these genes except for *MSH3* were decreased in the mutated and MTX resistant HT-29 compared to control sample (**Fig. 17**).

The *MSH3* and *MSH6* genes are families of DNA mismatch repair [66] genes and known to play an important role in repairing DNA for cell division as well as cooperating intestinal tumor suppression [67]. Also, the MLH genes (*PMS1*) as well as the MSH genes (*MSH1 - 6*) are closely correlated with the abnormal status of the colon cancer, and the mutations in these genes could cause the genetic predisposition and susceptibility to Lynch syndrome in colon cancer [68-70].

Therefore, the obtained novel frameshift insertions in these genes could prevent mismatch repair function and the tumor suppression under MTX condition and stimulate the rapid progression of gene

amplification as well as being resistant to MTX. Also, the additional molecular explanation for the possible tandem gene amplification mechanism was provided, which was affected by the malfunction of MMR pathways [67, 71]. It was previously known that *MSH3* is concurrently amplified with the *DHFR* gene due to its proximity in MTX resistant cells, and this could result in the malfunction of base-base mismatch repair and affect the genetic instability as well as the degree of resistance to the cytotoxic effects under MTX condition [72].

Expression level and alternative splicing pattern on MTX resistant sample

The expression level and alternative splicing pattern were analyzed in MTX resistant sample compared to control, and there was extremely high gene expression level on the identified amplification region, which matched with the result of high coverage of amplified unit in long read sequencing data (**Fig. 18**).

The read coverage of amplified region in MTX resistant sample was about 10 times higher than control in long read sequencing. Similarly, the $\log_2(\text{FPKM}+1)$ expression level from *DHFR* to *ATP6AP1L* gene was significantly over-expressed in MTX resistant HT-29 compared to control from 5-fold increase (*DHFR*) to 122-fold increase (*RASGRF2*) ($P = 0.0104$ by Mann-Whitney test; **Figs. 19 and 20**). The mapped reads

on the amplified region (chr5:80.5M-83M) were visualized from the transcriptome sequencing data (**Fig. 21**). As expected, the read depth was much higher over the amplified region in the MTX resistant HT-29.

Additionally, as the highly variable expression fold change were identified in the amplified unit, more complicated junctions between exons over the amplified region along with high read coverage were observed in MTX resistant sample whereas there was only a major pattern existed in control along with low read coverage (**Fig. 22**). Therefore, the five different types of alternative splicing patterns were estimated and compared between MTX resistant HT-29 and control sample (**Table. 8**).

The number of each alternative splicing events such as skipped exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), and retained intron (RI) in MTX resistant HT-29 was significantly higher than in control whereas the number of mutually exclusive exon (MXE) was reversely higher in control than in MTX-resistant sample (**Fig. 23**). The gene amplification might lead to a diversity of alternative splicing pattern, and this also could regulate the expression of genes which are closely associated with MTX pathways and finally result in the variable patterns of gene expression in simultaneously amplified region [73, 74].

The identification of differentially expressed genes (DEGs)

To determine which genes and gene sets involved in the MTX resistant sample compared to control, the differentially expressed genes were identified, and top 10 enriched gene sets were listed with the up-regulated and down-regulated expressed genes (**Fig. 24**). A total of 383 up-regulated and 287 down-regulated DEGs were identified in MTX resistant HT-29 sample compared to control (**Table. 9**).

Through enrichment analysis of KEGG gene sets, the up-regulated DEGs including *IL1B*, *MAPK11*, *JUN*, *MAP3K8*, *IL8*, and *CASP1* in the MTX resistant HT-29 were enriched in several signaling pathways such as MAPK signaling, toll-like receptor signaling, and NOD-like receptor signaling pathways. Interestingly, the down-regulated DEGs including *MAD2L1*, *CCNA2*, *MCM2*, *MCM4*, *FEN1*, and *CDK6* were enriched in the DNA replication, tyrosine metabolism, and cell cycle pathways, which are commonly up-regulated in colon cancer [75].

As previously reported, the down-regulated DEGs of MTX resistant osteosarcoma cell lines were enriched in the mitotic cell cycle, cell cycle, and DNA replication pathways, which were also down-regulated in our MTX resistant colon cancer cells. This result might explain the role of MTX which inhibits dihydrofolate reductase (*DHFR*) and keeps the tumor cells from proliferating in both cases [76, 77]. Still, this could not explain which mechanism is associated with *DHFR* gene amplification when

colon cancer cells are resistant to MTX, and this could not identify whether MTX targets other intracellular pathways and folate metabolism or not [78].

Optical genome mapping over the amplified region

The region of tandem amplification (80.6Mbp – 82.8Mbp) on chromosome 5 was additionally analyzed by the BioNano genome optical mapping due to the lack of the covering and mapping the whole range of the amplified region, which was too large and complicated. The contigs of genome mapping were well covered to whole chromosomes except for the amplified region, and this region had a complicated mapping with high coverage (200x) as the PacBio data results had (**Fig. 25**). Also, it was shown that the gene amplified unit had the inversions at both start and end points of amplified region as expected through the PacBio data analysis, but there was newly identified insertion at the end point of amplified region (**Fig. 26**).

These inversions at the starting and end point of amplification unit were certainly associated with amplification mechanism, and it seemed to react as the role for assisting and initiating tandem repeat amplification as previously reported [65, 79]. The identified inverted repeat could stimulate the formation of a large DNA palindrome after the breakage of an adjacent DNA double-strand occurred [80]. Therefore, this suggests

that short inversion at the start and end points in the MTX resistant HT-29 could play an important role in the initiation of gene amplification.

Chromosomal interaction and the topologically associating domains

The intra-chromosomal interactions over genome-wide view were identified and compared between MTX resistant HT-29 and control at 5kb resolution (**Fig. 27**). There was a high interaction with the clear long red line on the amplified region (5q14.1 to 5q14.2) only in the MTX resistant HT-29, and this interaction pattern was similar with the pattern when the amplification occurred in tumor sample as the previously reported [81] (**Fig. 28**).

The topologically associating domains (TADs) and several chromosomal rearrangements at chromosome 5 were identified to visualize conformation and interaction on intra or inter- chromosomes on the amplified region and to detect unsuspected chromosomal rearrangements at 500kb resolution (**Fig. 29**). The detected high intra-chromosomal interactions were involved in the amplified region (chr 5:80.6M-82.8M), and there were the several newly identified TADs at the middle and end point of this region compared to control with the high adjusted interaction frequencies (adjusted M) and adjusted p-value < 0.05 (**Fig. 30 and Table. 10**).

Interestingly, the interactions were also high on the region from 109Mbp to 138Mbp along with more TADs and high eigenvector, which delineated the compartmentalization in Hi-C [53]. This indicated that the amplified region and an adjacent region from the amplified region had both high intra-chromosomal interaction and the frequent contacts, and it seemed that the entire 5q region affected and boosted up the amplification mechanism.

In order to compare the computed intra-chromosomal interactions between MTX resistant HT-29 and control samples, the difference of adjusted interaction frequencies (adjusted M) with p-value at 500Kb resolution was analyzed between them, and the differentially interacting genomic regions at chromosome 5 were identified on amplified region, which were also statistically significant (p-value <0.05; **Fig. 31**). The adjusted M values were relatively lower in the region from 109Mbp to 138Mbp than amplified region, but this region still had the significant p-value and higher interactions compared to other positions. Therefore, chromosomal structure from the start position of the amplified unit (80Mb) to end point of 5q could have the complex network of spatial contacts to harbor the gene amplification.

Also, the relative copy number each chromosome was estimated from Hi-C data by using chromosome 2 as reference, and it was found that the relative copy number in chromosome 5 was significantly higher in MTX resistant HT-29 compared to control over whole chromosomes, but

it was not applicable to chromosome 17, which also had the stretched structure like chromosome 5 in the FISH result (**Fig. 32**).

In addition, the chromosomal rearrangement such as only DMs, which were extrachromosomal DNA and harbored the amplification of oncogenes by involving in drug resistance, were detected in the amplified region by the Hi-C data, which was not observed in the FISH data. Hi-C data could detect the unsuspected chromosomal rearrangements as well as copy number in the highly amplified region. However, this result should be confirmed by other technologies since it was not clear how to distinguish between DMs and HSRs because of the similar structure.

Overall, tandem gene amplification of several genes on chr5 (2.2Mbp) were confirmed by gene expression and gene mapping as well as HI-C data results, and more complicated interaction on intra-chromosome 5 can be the critical factor for identifying the hotspots of spatial contacts over the amplified region.

The mechanism of tandem gene amplification

All things considered, this study could propose the mechanism of gene amplification, which was occurred under the specific circumstances in the MTX resistant HT-29 cells through the Breakage-fusion-bridge (BFB) cycles as previously reported [82] (**Fig. 33**). Before gene amplification

occurred, the frameshift insertions in *MSH* genes as well as *MLH* genes over several chromosomes had been caused by MTX toxicity, and this had resulted in the genetic predisposition and dysregulation of mismatch repair pathway under MTX condition [24].

After malfunctioning the mismatch repair function under MTX condition, chromosomal breakage occurred at the *DHFR* gene position on chromosome 5 because of the inverted repeat at the start position of *DHFR* gene, and from *DHFR* gene to *ATP6AP1L* (2.2 Mb) were involved for producing the amplified unit. The end point of the amplified unit had same inverted repeat, which could indicated the stop position for gene amplification. However, this is still unknown how the specific genes were selected and involved for the gene amplification mechanism.

Finally, the variable size of gene amplification with HSRs could be produced from the BFB cycles, and the unstable HSRs could be occasionally transformed into either different size of HSR fragments or DMs accompanying the inversions at the end points. Overall, the co-amplification of *MSH3* and *DHFR* gene as well as the frame shift mutations in *MSH* and *MLH* genes continuously affected the genetic instability and enhance the resistance to methotrexate.

Table 1. The estimation of copy number and expression of *DHFR* gene in MTX resistant clone (C1-2) at each cycle.

Type of Sample	Sample Ct	Expression	Rate for Copy Number	Copy Number
Hapmap NA19982	24.99	2.52	1.00	2.00
Control (HT-29)	26.33	1.00	0.40	0.79
1st cycle MTX resistant HT-29	26.04	1.22	0.48	0.97
2nd cycle MTX resistant HT-29	20.21	69.17	27.42	54.83
3rd cycle MTX resistant HT-29	19.99	80.56	31.93	63.87

Table 2. The comparison of detected structural variants between control and MTX resistant HT-29.

No. of Events (median size)	All chromosomes		Chromosome 5	
	Control (HT-29)	MTXresistant HT-29	Control (HT-29)	MTXresistant HT-29
Deletion	13(4472)	14(2917)	2(4425.5)	3(1514)
Duplication	1(1004)	5(13534)	0	1(371675)
Inversion	0	4(331155.5)	0	3(328697)
Inverted Duplication	8(47)	15(77)	0	3(1529)
Translocation	45	50	4(chr3 to chr5) and (chr5 to chr12)	2(chr5 to chr12)

The median size of split reads is shown in parentheses.

Table 3. The detected variants on chromosome 5 in MTX resistant HT-29

chrom1	pos1	strand1	chrom2	pos2	strand2	variant_ame	variant_t ype	split	size	CNV_cate gory	category	nearby_v ariant_co unt
5	80617089	-	5	80618750	-	24	INVDUP	47	1661	matching	simple	0
5	81781592	+	5	81781648	+	50	INVDUP	10	56	neutral	simple	0
5	82245207	+	5	82573904	+	56	INV	56	328697	matching	crowded	1
5	82470789	-	5	82842464	+	58	DUP	11	371675	matching	crowded	4
5	82470887	+	5	82804501	+	59	INV	12	333614	partial	crowded	4
5	82473031	-	5	82474560	-	60	INVDUP	18	1529	matching	simple	3
5	82678370	+	5	82764579	-	62	DEL	56	86209	matching	simple	3
5	82804502	+	5	82842464	+	64	INV	47	37962	matching	simple	3
5	137519108	+	5	137520622	-	65	DEL	19	1514	neutral	simple	0
5	137686889	+	5	137688197	-	66	DEL	22	1308	neutral	simple	0
5	88112313	-	12	66057592	-	256	TRA	12	-1	neutral	reciprocal	12
5	88112381	+	12	66057683	+	284	TRA	12	-1	neutral	reciprocal	12

Table 4. The identified genes with high segmented coverage (segmented coverage >20) on chromosome 5

Chr	Segment_Start	Segment_End	Depth	Gene_Chr	Gene_Start	Gene_End	Gene_Name	Strand	Description
chr5	42810000	42820000	62.1845012	chr5	42799879	42887392	SEPP1	-	Homo sapiens selenoprotein P, plasma, 1
chr5	80620000	80630000	212.3334045	chr5	80608622	80622524	LINC01337	-	long intergenic non-protein coding RNA 1337
chr5	80630000	80640000	231.2890015	chr5	80630312	80631590	CTC-325J23.2	-	antisense RNA
chr5	80630000	80640000	231.2890015	chr5	80628227	80654983	DHFR	-	Homo sapiens dihydrofolate reductase
chr5	80640000	80650000	222.8034973	chr5	80649999	80650088	MTRNR2L2	-	Homo sapiens MT-RNR2-like 2
chr5	80740000	80750000	232.6362	chr5	80746338	80746819	RP11-241J12.3	-	antisense RNA
chr5	80810000	80820000	257.9645996	chr5	80854647	80876460	MSH3	+	Homo sapiens mutS homolog 3
chr5	80850000	80860000	236.5988023	chr5	80855506	80855843	RP11-241J12.1	-	antisense RNA
chr5	80940000	80950000	239.4228973	chr5	80947696	80960907	RASGRF2-AS1	-	antisense RNA 1
chr5	80990000	81000000	248.4263	chr5	80997182	80998047	CTD-2193P3.1	-	Known processed pseudogene
chr5	81030000	81040000	310.3222046	chr5	80980871	81230166	RASGRF2	+	Homo sapiens Ras protein-specific guanine nucleotide-releasing factor 2
chr5	81110000	81120000	226.7742004	chr5	81113384	81114852	CTD-2193P3.2	-	antisense RNA
chr5	81220000	81230000	257.3019104	chr5	81204083	81301560	CKMT2-AS1	-	Homo sapiens CKMT2 antisense RNA 1
chr5	81240000	81250000	249.2651978	chr5	81242329	81242599	CTD-2248H3.1	-	antisense RNA
chr5	81240000	81250000	249.2651978	chr5	81233284	81286397	CKMT2	+	Homo sapiens creatine kinase, mitochondrial 2 (sarcomeric)
chr5	81310000	81320000	252.9447937	chr5	81301589	81313297	ZCCHC9	+	Homo sapiens zinc finger, CCHC domain containing 9
chr5	81340000	81350000	245.8712006	chr5	81330004	81394179	ACOT12	-	Homo sapiens acyl-CoA thioesterase 12

The average segmented coverage on amplified region was approximately 197.

Table 4. The identified genes with high segmented coverage (segmented coverage >20) on chromosome 5 (cont.)

Chr	Segment_Start	Segment_End	Depth	Gene_Chrom	Gene_Start	Gene_End	Gene_Name	Strand	Description
chr5	81590000	81800000	276.2663879	chr5	81413020	81751797	SSBP2	-	Homo sapiens single-stranded DNA binding protein 2
chr5	81850000	81860000	218.3592072	chr5	81851600	81852201	CTD-2249K22.1	+	lncRNA
chr5	81890000	81900000	226.7967072	chr5	81892489	81892721	SFMIP1	+	26S proteasome complex subunit pseudogene 1
chr5	81970000	81980000	251.2120972	chr5	81978249	81978319	AC114969.1	+	miRNA
chr5	81990000	82000000	227.0233002	chr5	81991994	81992481	ATG10-IT1	+	ATG10 intronic transcript 1
chr5	82010000	82020000	225.3244934	chr5	82009601	82010094	PPIAP11	-	peptidylprolyl isomerase A pseudogene 11
chr5	82060000	82070000	259.3533936	chr5	81972024	82276857	ATG10	+	Homo sapiens autophagy related 10
chr5	82070000	82080000	223.1358032	chr5	82078426	82078724	RN7SL378P	+	RNA, 7SL, cytoplasmic 378, pseudogene
chr5	82070000	82080000	223.1358032	chr5	82073054	82073702	ATG10-AS1	-	ATG10 antisense RNA 1
chr5	82260000	82270000	149.5509033	chr5	82265156	82265259	RPS23P5	-	ribosomal protein S23 pseudogene 5
chr5	82270000	82280000	153.1331024	chr5	82273357	82278577	RPS23	-	Homo sapiens ribosomal protein S23
chr5	82330000	82340000	174.9517975	chr5	82279461	82386977	ATP6AP1L	+	Homo sapiens ATPase, H ⁺ transporting, lysosomal accessory protein 1-like
chr5	82540000	82550000	196.9989929	chr5	82545861	82546310	CTD-2015A6.1	-	lncRNA
chr5	82580000	82590000	128.1013031	chr5	82586775	82587411	CTD-2015A6.2	-	lncRNA
chr5	82760000	82770000	92.697998	chr5	82765403	82766333	CTD-2218K11.2	+	lncRNA
chr5	82770000	82780000	115.7454987	chr5	82777796	82778586	RPL5P16	+	ribosomal protein L5 pseudogene 16
chr5	82820000	82830000	51.1856995	chr5	82824883	82825184	CTD-2218K11.3	+	pseudogene

The average segmented coverage on amplified region was approximately 197.

Table 5. The comparison of mutations between control and MTX resistant HT-29 over whole chromosomes.

Mutation types	MTX resistant HT-29	Control (HT-29)
Exonic total	13982	13310
Exonic splicing	10	9
Splicing	70	155
ncRNA exonic	6040	5876
ncRNA exonic:splicing	2	3
ncRNA splicing	21	26
3'UTR	28523	27892
5'UTR	3448	3632
Nonsynonymous SNV	5708	5730
Frameshift deletion	70	66
Frameshift insertion	858	816
Stopgain	57	55
Stoploss	9	9

Table 6. The comparison of mutations between control and MTX resistant HT-29 on chromosome 5.

Mutation types	MTX resistant HT-29	Control (HT-29)
Exonic total	577	559
Exonic splicing	1	0
Splicing	4	8
ncRNA exonic	159	206
ncRNA exonic:splicing	0	0
ncRNA splicing	1	1
3'UTR	1332	1337
5'UTR	166	164
Nonsynonymous SNV	220	226
Frameshift deletion	4	3
Frameshift insertion	51	12
Stopgain	6	4
Stoploss	0	0

Table 7. The detection of novel frameshift insertions in MTX resistant HT-29 compared to control.

MTX resistant HT-29									
Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	Gene Title	ExonicFunc.refGene	ExAC_ALL
chr2	47803552	47803552	-	T	exonic	MSH6	mutS homolog 6	frameshift insertion	.
chr2	189818083	189818083	-	A	exonic	PMS1	postmeiotic segregation increased 1	frameshift insertion	3.54E-05
chr2	189863838	189863838	-	A	exonic	PMS1	postmeiotic segregation increased 1	frameshift insertion	.
chr3	37012077	37012077	A	G	exonic	MLH1	mutL homolog 1	nonsynonymous SNV	0.2325
chr5	80654962	80654962	A	G	exonic	MSH3	mutS homolog 3	nonsynonymous SNV	0.9029
chr5	80675095	80675095	-	A	exonic	MSH3	mutS homolog 3	frameshift insertion	.
chr5	80854162	80854162	A	G	exonic	MSH3	mutS homolog 3	nonsynonymous SNV	0.8731
chr5	80873118	80873118	G	A	exonic	MSH3	mutS homolog 3	nonsynonymous SNV	0.7305
chr7	5987144	5987144	T	C	exonic	PMS2	postmeiotic segregation increased 2	nonsynonymous SNV	0.8514
chr7	5987357	5987357	G	A	exonic	PMS2	postmeiotic segregation increased 2	nonsynonymous SNV	0.3854
chr7	5987525	5987525	-	T	exonic	PMS2	postmeiotic segregation increased 2	frameshift insertion	.
chr14	75047125	75047125	G	A	exonic	MLH3	mutL homolog 3	nonsynonymous SNV	0.4126
chr14	75047180	75047180	T	C	exonic	MLH3	mutL homolog 3	nonsynonymous SNV	0.9968
Control (HT-29)									
Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	Gene Title	ExonicFunc.refGene	ExAC_ALL
chr3	37012077	37012077	A	G	exonic	MLH1	mutL homolog 1	nonsynonymous SNV	0.2325
chr5	80654905	80654905	-	CCGCAGCGC	exonic	MSH3	mutS homolog 3	nonframeshift insertion	0.0427
chr5	80654962	80654962	A	G	exonic	MSH3	mutS homolog 3	nonsynonymous SNV	0.9029
chr5	80854162	80854162	A	G	exonic	MSH3	mutS homolog 3	nonsynonymous SNV	0.8731
chr5	80873118	80873118	G	A	exonic	MSH3	mutS homolog 3	nonsynonymous SNV	0.7305
chr7	5987144	5987144	T	C	exonic	PMS2	postmeiotic segregation increased 2	nonsynonymous SNV	0.8514
chr7	5987357	5987357	G	A	exonic	PMS2	postmeiotic segregation increased 2	nonsynonymous SNV	0.3854
chr14	75047125	75047125	G	A	exonic	MLH3	mutL homolog 3	nonsynonymous SNV	0.4126
chr14	75047180	75047180	T	C	exonic	MLH3	mutL homolog 3	nonsynonymous SNV	0.9968

Table 8. The comparison of alternative splicing patterns on the amplification unit between MTX resistant and control samples.

EventType	SigEvents.JC+readsOnTarget	
	MTX resistant HT-29	Control (HT-29)
SE	3086	1732
MXE	785	3231
A5SS	526	169
A3SS	943	228
RI	2378	148

Abbreviations

Skipped exon (SE)

Mutually exclusive exon (MXE)

Alternative 5' splice site (A5SS)

Alternative 3' splice site (A3SS)

Retained intron (RI)

Table 9. The differentially expressed genes (DEGs) in MTX resistant HT-29.

Up-regulated DEGs (383 genes)	<p>PRSS22 BAIAP3 ETV7 CD22 ALOX5 TYMP BTN3A1 USP2 TRAF1 SLC2A3 CACNB1 GPR116 CLEC2D TRIB2 SPEG SIDT1 GSDBM DHRS9 SCARF1 CACNG4 RGS11 P2RX5 COL16A1 HSD17B14 ASAP3 EPB41L1 TMEM40 SLC4A11 LAG3 ARHGAP4 ANKRD24 YPEL3 PITPNM2 MYO15A NLRP1 DFNB31 CATSPERG CYP2D6 APOL4 UPK3A REC8 WFDC2 MAP1LC3A ATP11C KCND1 VGLL1 GABRE GDPD3 BMF RHOV MAP4K1 AMH IL4I1 DENND3 CCDC114 PBX4 SH3GL2 UNC5B MAP3K8 RASD1 DHX58 MYH3 MGP ADTRP ULBP1 MSH3 RASGRF2 HES1 SLC4A3 PCSK4 FN1 MLPH PADI2 GBP1 SGK1 CTGF TRPM6 LTBP2 GPR68 DUSP1 CD274 PLEKHG1 TP53AIP1 EGR1 TNFSF10 EGR2 NR4A1 C4BPB ARHGAP40 PMEPA1 ZBP1 PRICKLE4 CDKN1A RUNX2 C20orf195 IL1B C3 FOSB ID1 PRKCG ZFP36 APOL3 APOL2 APOBEC3F KRT17 LOXL1 CHRNA10 APOE TNNT2 PNPLA7 ANGPL6 C19orf66 PLXNA3 SH3BP5 PDLIM4 CKMT2 ZCCHC9 LGALS3 TRIM22 XAF1 REEP2 ALDH3B2 FBXO44 EPST11 RARRES3 RSAD2 SLC37A2 ANXA1 ELF5 EGR4 WNT10A THSD1 SCEL RTP4 CASP1 IFI44L MYOF LOXL4 DUSP5 GPR87 ADAMTS14 SEMA7A RASGEF1B ITGB7 ITGAX ZMYND15 ATHL1 SIK1 CACNG8 PADI1 SYTL1 CYR61 CTSK REN CSRNP1 KLHL3 SERPING1 LYPD6B IL18 ATG10 ACOXL NR4A2 GPR110 ANKRD29 CERS3 GOLGA7B UBE2L6 TNFRSF14 DUSP2 XDH BTG2 CCDC17 PAQR6 FOXH1 SLC25A34 SCNN1D GBP4 ATF3 IVL ELF3 IFI16 KIAA1407 ZC3H12A SHROOM1 TRPV6 OTUD1 DKFZP868J19100 HDGFRP3 ANKDD1A CERCAM PRRT2 KIFC2 RILP TRPV3 TMEM88 TRANK1 TTC21A MFSF7 THBS3 CXCL10 IL8 TMEM154 FOS CST1 STAT2 CEL KRT13 EFEMP2 SYT12 C11orf80 TDRD12 SLC26A9 SLC22A1 DDIT3 VWA3A CST6 UCP3 IDSP1 SLC03A1 C8orf31 SAMD9L JUN ODF3B ALS2CL EGR3 HCLS1 MUC16 HEPHL1 BA1 UBA7 LINC00085 HCAR2 MX2 EMILIN3 KCNQ3 TACSTD2 FAM70B DUSP8 SOCS3 IFNE MAFF HEATR7B1 SOCS1 MAPK11 STAC3 MUC1 C11orf35 TLCD2 PDIA2 CYP4F12 RPS23 CEACAM19 MIR22HG EPOR MAGI2 DNAJB13 PALM3 HES4 NOXA1 APOD C6orf222 CXCL17 CGB7 HSH2D C5AR1 BCO2 CCDC154 SLC22A20 ZNF44 PLOG2 ARC GPRASP1 SNORA31 DDO AGER C6orf25 AC096670.3 IGFL4 GPR20 FAM221B LRRC10B FAM71F2 SAP25 ATP6AP1L CPT1B KRTAP5-1 LCAT LBH UCA1 PLIN5 RP11-83B20.1 RP11-429J17.8 RP11-1036E20.9 EVPLL SMTNL1 CGCR hsa-mir-6080 ATAD3C RP11-465B22.3 AC110619.2 SNORD63 ADM5 SH3BP5-AS1 RP11-288L9.4 NTF4 AC009950.2 RP11-203J24.8 RP11-69I8.3 RAB11F1P1 SAPCD1 RP4-583P15.10 AC019349.5 GOLGA6L5 RPS23P8 RP11-34A14.3 RP11-263K19.4 RP11-67C2.2 CTD-2020K17.3 RP11-250B2.3 SAPCD1-AS1 RP11-73M7.6 NRADDP CBR3-AS1 RNF223 RP3-430N8.8 AF011889.2 BX470102.3 APOBEC3G PSMB9 UGT1A9 RP11-147I3.1 RP11-379B18.5 AF011889.5 ARHGDI3 <i>Metazoa_SRP</i> PDCL3P4 IFITM10 SCARF2 RP3-330M21.5 CTC-281B15.1 CTD-2248H3.1 RP11-510N19.5 CTC-325J23.2 CTD-2193P3.2 HMGB1P3 ALDH1L1-AS1 CTD-2193P3.1 CTC-459I6.1 RP11-459E5.1 CHKB-CPT1B RP11-496I9.1 RP11-326C3.2 MSH5-SAPCD1 RP11-167N4.2 RP11-512N21.3 RP11-631N16.2 RP11-386M24.3 MC1R RP4-647C14.3 RP11-566K11.4 RP11-521C20.3 RP11-1100L3.8 RP11-254F7.2 RP5-1085F17.3 RP11-386M24.6 AC145291.2 RP11-250B2.5 RP11-448G15.3 SPON1 LA16c-325D7.1 LINC00672 SNORD3A RP1-37N7.1 RP13-104F24.3 AC010761.10 RP5-906A24.2 MYO15B AC004510.3 SH3GL1P3 CTD-2240E14.4 CTD-3252C9.4 AC006262.5 CTD-2521M24.5</p>
Down-regulated DEGs (287 genes)	<p>CFTR SLC7A2 PLXND1 USH1C DPEP1 SLC38A5 GRAMD1B HSD17B6 MIEPE ZIC2 KITLG LTBP1 KCNQ1 GYG2 SOAT1 TMEM48 TBXAS1 SLC12A2 IPO5 PRR11 LIMS2 HMHR MCM2 PPP2R2C ENO1 PLD1 GPC4 EPB41L2 SLC4A4 SMARCA2 ABCB1 ORC1 IPO11 TPX2 TESC BIRC5 LYZ DPYSL2 TMEM38B BAMB1 PCSK5 LGALS2 C14orf105 SERPINA4 MYBL2 HNF4A FERMT1 FGF9 SLC25A15 CPPED1 KIAA1199 MCM4 CDK6 CPVL COBL MOGAT3 AGR2 SLC1A1 AMBP DKK1 MAP2K6 ANXA10 GALNTL4 SOX6 POU2AF1 CHPT1 UST PERP PRLR HGD PRKAR2A ECT2 HHLA2 PLSCR4 IGFBP2 IL1R2 LEPR AGMAT NR5A2 PROX1 PLAGL1 AKAP7 MYB SLC16A7 GDA ONECUT2 MOB3B CCDC170 LYPLA1 RPL21 FAM126A TUBA1B EREG FOXA2 TMX4 AMOT WNK4 DLGAP5 SGPP1 TSPAN8 TUBA4A VIL1 GNAI1 MGAT3 HOXD13 DLL4 PPP1R1B NR0B2 DDC RFC3 ZDHHC8P1 HMGC52 LDHA HNF1A AGT SLC19A3 SERPINE2 SMC2 HMG1A CASC5 SLC38A4 LGR5 FAM222A SLC7A1 BRCA2 DIAPH3 ZIC5 NKD1 IMPA2 GATA6 PRKACB PARP1 LBR CHAC2 CDC47 PTPRG CCNA2 FAM105A SYTL5 TACC1 LACTB2 FAM171A1 SLC43A1 ADRA2A C11orf53 SLC7A11 SLC2A13 SACS SPC25 CNKSR3 KCTD15 UBASH3B PLCL2 PPARGC1B SLC26A2 PTSS1 HKDC1 ARSE ZNF618 EDA KALRN FGFR4 SSTR5 AKR7A3 CAMK2N1 PKDC8 ARL5A KCNJ3 IHH FABP1 SAP30 MAD2L1 SPINK1 FAPB5 HNF4G CLDN2 CDX2 KBTBD6 CKB ANPEP CCDC103 RRM1 FAM83B FEN1 NPNT CXXC4 SLC38A11 GJB1 ROBO1 P2RY1 MYO7B ALCAM SERPINA6 FOXA3 TMEM37 ESCO2 LGALS4 CYCS ADH6 AGR3 CSPG4 MUC13 FBXO45 KCNE3 KBTBD11 C18orf56 GSG2 LCN15 DNAJC22 KCTD12 TMEM64 OR51E1 C3orf58 C8orf33 EPHB3 NEB ASCL2 BRI3BP SFTPA2 TARSL2 GPRIN3 NCR3LG1 MAOA CTSE RP11-57C13.3 SERPINA1 CYP2B6 SPN DPP4 S100A10 PRIM1 AKR1B10 RYR2 PAPSS2 F5 SUCNR1 GRK5 C2orf72 SP5 C10orf112 ZBTB10 LGR4 LDHAP4 MUC5AC TENM3 AC007405.2 SLC25A5-AS1 NPY6R RP1-37C10.3 RP11-103C16.2 AC007163.6 RP11-229P13.23 COL4A2-AS1 AF196970.3 AC013463.2 MYB-AS1 AC005550.3 MLK7-AS1 RPS2P5 RP11-64D22.2 TDGF1 HNF1A-AS1 RP11-67L3.5 RP11-410D17.2 RP11-382J12.1 ADH1C SLC7A11-AS1 RP11-710F7.2 RP11-115D19.1 RAD21-AS1 PRKDC CTD-2292P10.2 RP11-363E6.3 RP11-700F16.3 RP11-627G23.1 RP11-173P15.3 TMPO-AS1 RP11-186F10.2 RP11-210N13.1 RP11-386G11.10 RP11-109D20.2 LINC00261 CTD-2196E14.5 RP11-40IP9.4 CTD-2510F5.4 ZNF488 RP11-627G18.1</p>

Table 10. The topologically associating domains (TADs) with high intra-chromosomal interactions on chromosome 5.

chr1	x1	x2	chr2	y1	y2	score	uVarScore	lVarScore	upSign	loSign
chr5	80970000	81100000	chr5	80970000	81100000	0.634	0.047	0.058	0.599	0.522
chr5	81205000	81295000	chr5	81205000	81295000	0.944	0.017	0.035	0.633	0.722
chr5	81315000	81405000	chr5	81315000	81405000	0.853	0.038	0.006	0.700	0.611
chr5	81430000	81735000	chr5	81430000	81735000	0.701	0.094	0.103	0.593	0.503
chr5	81510000	81720000	chr5	81510000	81720000	0.982	0.045	0.080	0.803	0.639
chr5	81765000	81840000	chr5	81765000	81840000	0.776	0.051	0.144	0.609	0.516
chr5	82645000	82835000	chr5	82645000	82835000	1.389	0.084	0.113	0.647	0.884
chr5	82765000	82835000	chr5	82765000	82835000	0.888	0.016	0.169	0.518	0.607
chr5	82905000	83605000	chr5	82905000	83605000	0.219	0.435	0.341	0.411	0.419
chr5	87025000	87390000	chr5	87025000	87390000	1.347	0.314	0.319	0.581	0.655
chr5	90840000	91440000	chr5	90840000	91440000	0.737	0.354	0.392	0.414	0.439
chr5	94970000	95085000	chr5	94970000	95085000	1.354	0.075	0.119	0.688	0.521
chr5	95895000	96340000	chr5	95895000	96340000	1.020	0.343	0.290	0.498	0.535
chr5	96740000	97155000	chr5	96740000	97155000	1.037	0.327	0.308	0.402	0.537
chr5	100565000	101370000	chr5	100565000	101370000	0.851	0.378	0.316	0.457	0.406
chr5	102545000	103145000	chr5	102545000	103145000	1.206	0.332	0.337	0.537	0.484
chr5	102615000	103030000	chr5	102615000	103030000	0.924	0.331	0.312	0.423	0.451
chr5	111505000	112140000	chr5	111505000	112140000	1.311	0.323	0.323	0.466	0.660
chr5	112160000	112475000	chr5	112160000	112475000	1.194	0.294	0.331	0.651	0.417
chr5	116555000	116960000	chr5	116555000	116960000	0.665	0.271	0.290	0.410	0.459
chr5	118980000	119445000	chr5	118980000	119445000	0.730	0.350	0.408	0.474	0.483
chr5	120770000	121930000	chr5	120770000	121930000	0.675	0.366	0.394	0.401	0.468
chr5	128205000	128880000	chr5	128205000	128880000	0.979	0.367	0.318	0.439	0.536
chr5	130210000	130740000	chr5	130210000	130740000	0.673	0.357	0.339	0.408	0.413
chr5	136250000	136775000	chr5	136250000	136775000	0.507	0.321	0.421	0.403	0.460
chr5	136880000	137585000	chr5	136880000	137585000	0.615	0.321	0.307	0.476	0.432
chr5	137650000	137740000	chr5	137650000	137740000	1.380	0.106	0.090	0.567	0.656
chr5	146455000	146515000	chr5	146455000	146515000	0.985	0.045	0.131	0.500	0.500
chr5	154005000	154410000	chr5	154005000	154410000	0.873	0.370	0.355	0.484	0.478
chr5	174395000	174835000	chr5	174395000	174835000	0.967	0.369	0.339	0.411	0.425

Abbreviations

Score = corner score

Uvar = the variance of the upper triangle

Lvar = the variance of the lower triangle

Usign = -1*(sum of the sign of the entries in the upper triangle)

Lsign = sum of the sign of the entries in the lower triangle

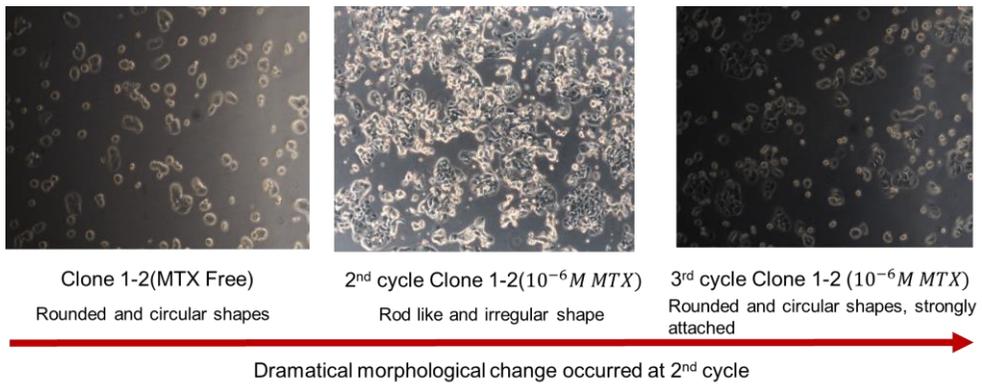


Figure 2. The morphological change of MTX resistant colon cancer cells. The MTX resistant clone (C1-2) was detected and captured by light microscope at 200x magnification. The morphological changes at each cycle were indicated under 10^{-6} mol/L MTX.

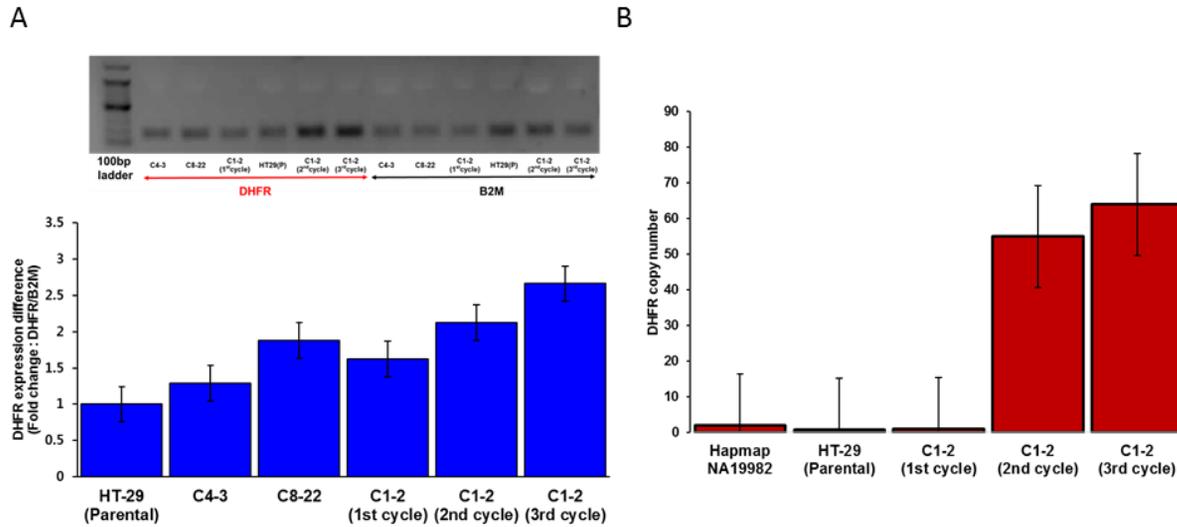


Figure 3. The *DHFR* gene expression and copy number among MTX resistant clone (C1-2), control, and reference. The expression (a) and copy number (b) of *DHFR* gene in MTX resistant clones and C1-2 clone at each cycle were estimated by qPCR. The error bars indicated standard error (SE) of the mean.

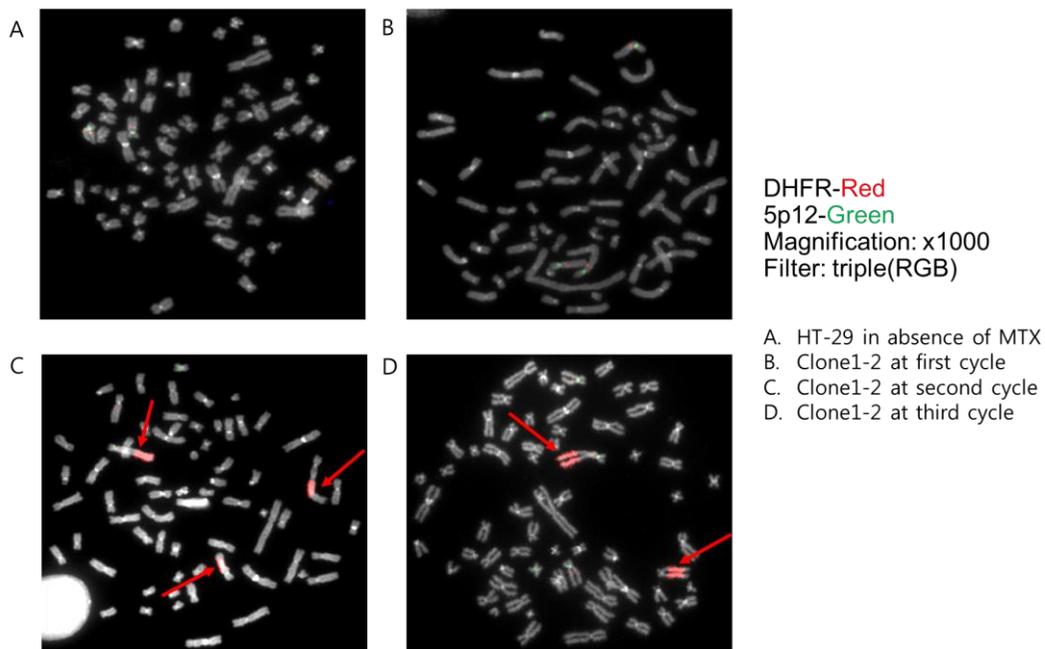


Figure 4. The visualization of the amplified *DHFR* gene at 5q arm in MTX resistant C1-2. The amplified *DHFR* gene was visualized by Fluorescent In Situ Hybridization (FISH) on MTX resistant clone (C1-2) at each cycle, and control sample with the 1000x magnification. The amplified *DHFR* gene region was indicated by the red arrow.

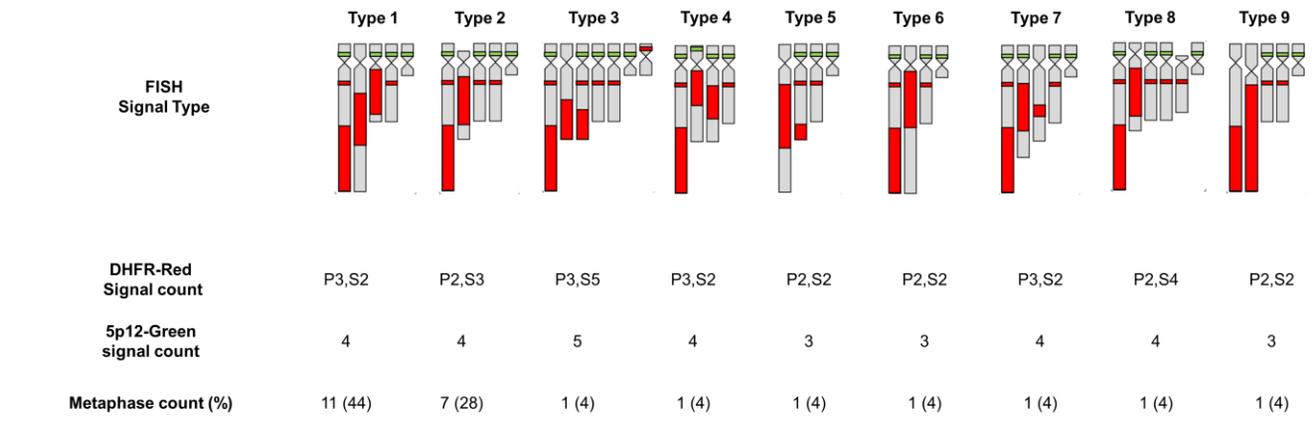


Figure 5. The patterns of amplified *DHFR* gene at 5q arm in MTX resistant C1-2. The FISH signal type of amplified *DHFR* gene in MTX resistant clone (C1-2) was displayed with the pairing signal (P) and spot signal (S). 5q12 was indicated by green signal, and *DHFR* gene was indicated by red signal. Each number of red and green signal at metaphase were counted.

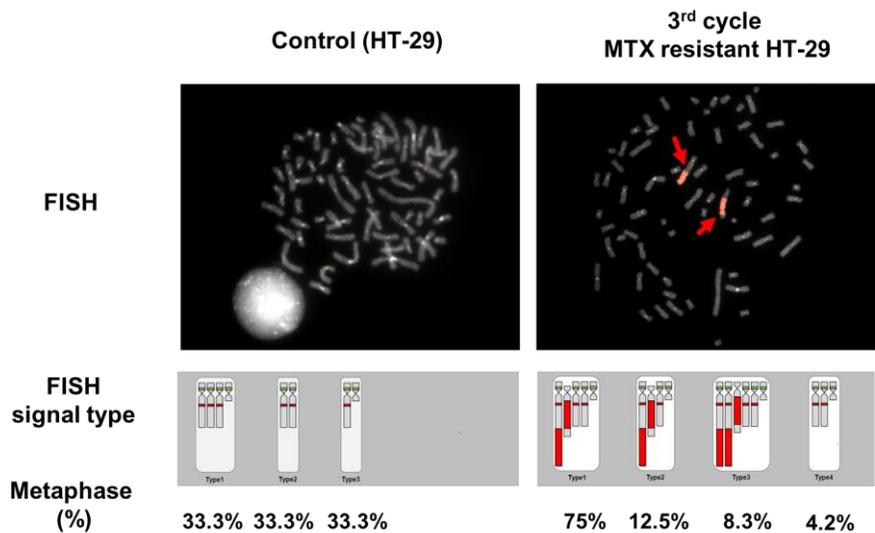


Figure 6. The comparison of the *DHFR* gene amplification pattern and signal type of FISH between MTX resistant clone (C1-2-4) and control. The subculture (C1-2-4) of C1-2 clone was visualized by Fluorescent In Situ Hybridization (FISH) with 1000x magnification, and the FISH signal type and the percentage at metaphase were compared between control and MTX resistant clone (C-1-2-4).

Control (HT-29)



3rd cycle
MTX resistant HT-29

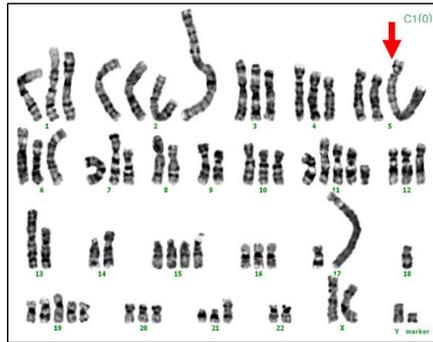


Figure 7. Karyotyping in MTX resistant and control samples. All chromosomes were karyotyped and the abnormal chromosomal shapes were detected at 5q compared to control, and this was indicated by the red arrow.

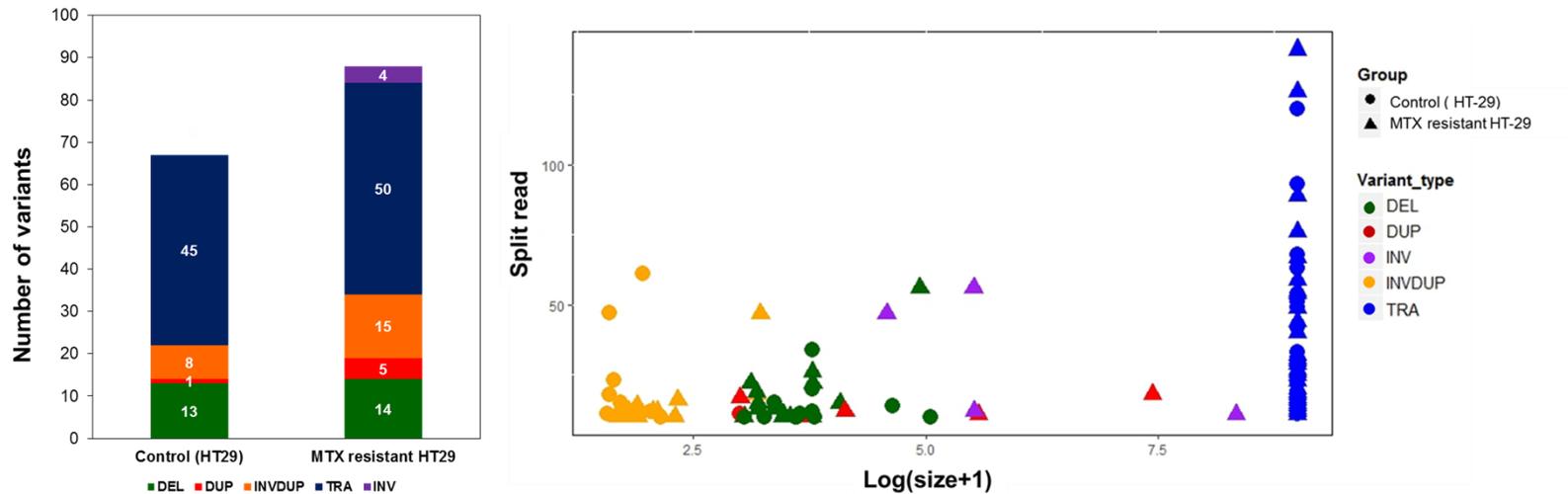


Figure 8. The detection and characterization of SVs in MTX resistant HT-29. The five genomic variants (deletion, duplication, inverted duplication, translocation, and inversion) in MTX resistant sample and control were analyzed and visualized with the bar graph. In the bar graph, the number of variants were counted and compared between two samples. In the scatter plot, the size of variants ($\log(\text{size}+1)$) and depth of split read were plotted and compared between control and MTX resistant HT 29. The group and variant type were indicated by the different shape and color.

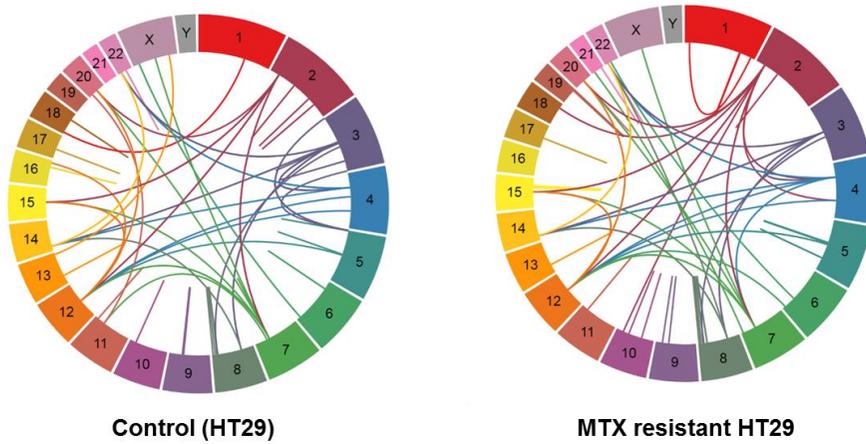


Figure 9. The visualization of inter-chromosomal genomic rearrangements in MTX resistant HT-29 and control samples. The inter-chromosomal genomic rearrangements were compared between MTX resistant HT-29 and control samples, and it was visualized by Splitthreader over all chromosomes.

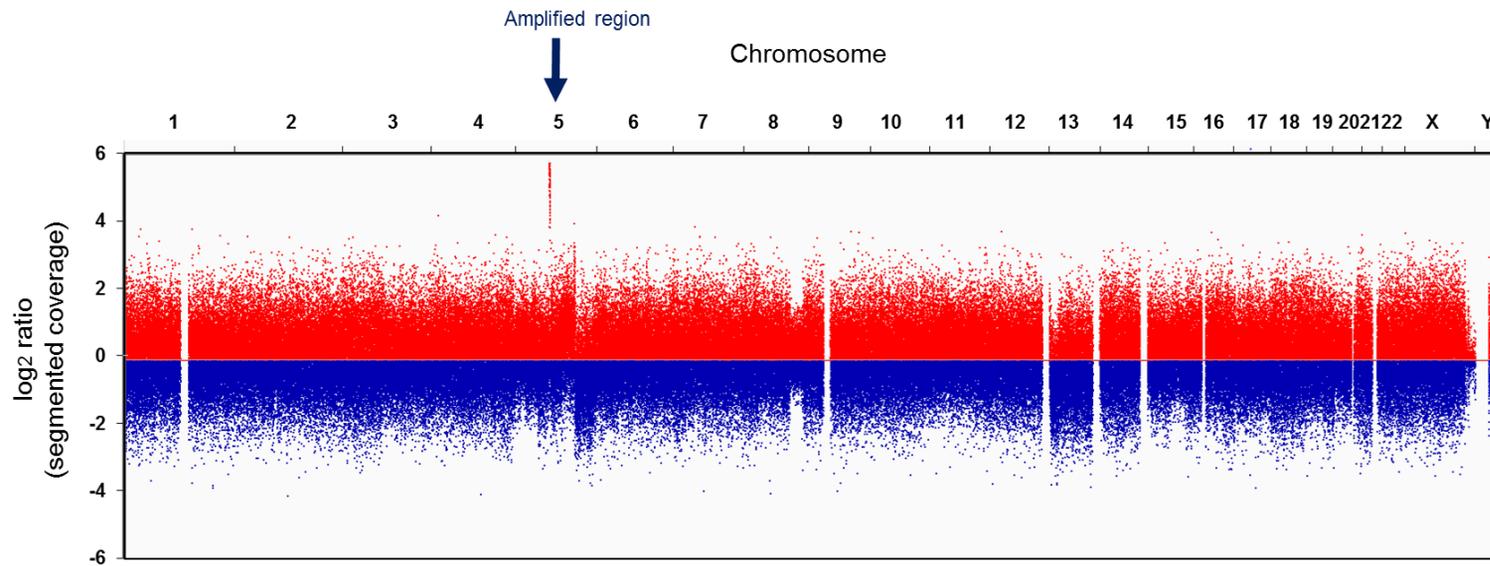


Figure 10. The comparison of segmented coverage over whole chromosomes between MTX resistant HT-29 and control samples. The log₂ ratio of segmented coverage was compared between control and MTX resistant HT-29 over all chromosome. The amplified region on 5q in MTX resistant HT-29 was indicated by the blue arrow.

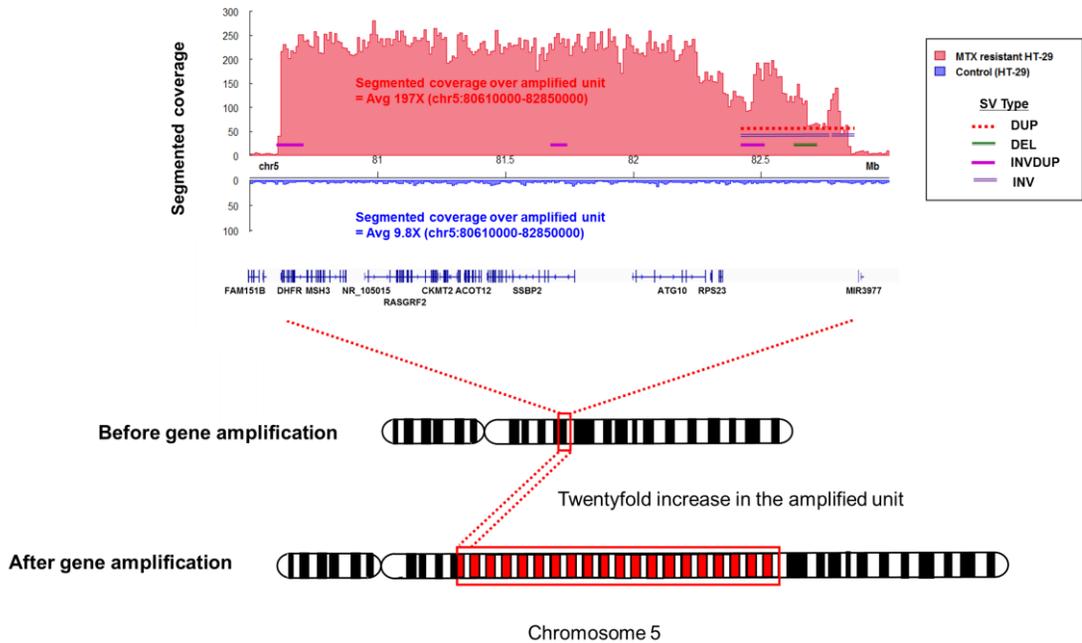


Figure 11. The comparison of segmented coverage and structural variants over amplified region between MTX resistant HT-29 and control samples. The segmented coverage and genomic variants were compared between control and MTX resistant HT-29 over amplified region on chromosome 5: 80,610,000-82,850,000.

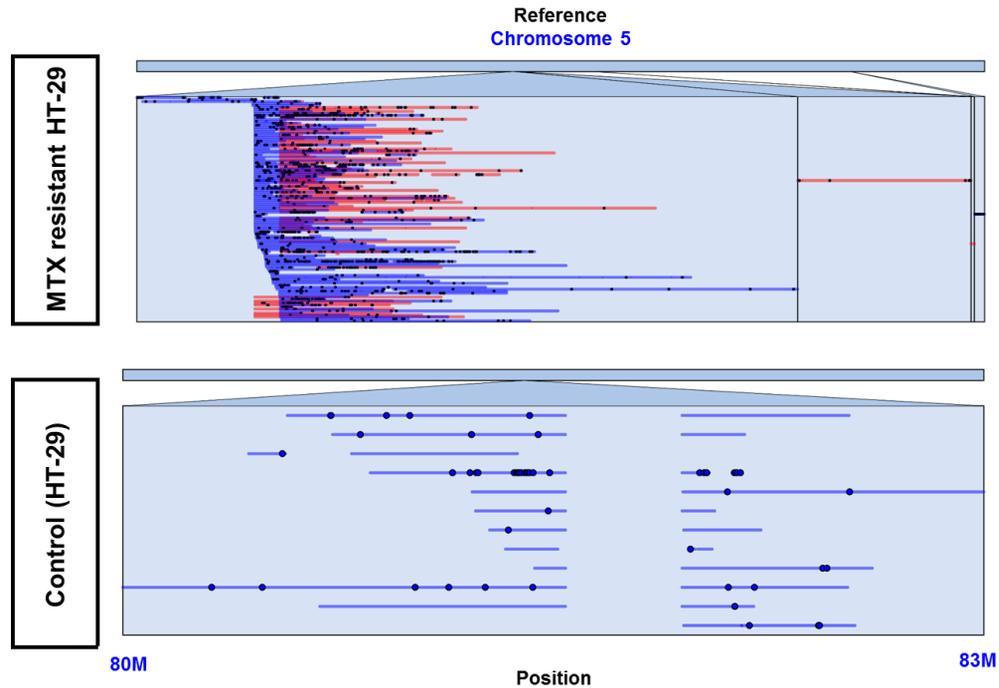


Figure 12. The multiple-read view of alignment over amplified region with Ribbon.

The multi-reads from alignment data (BAM) was visualized and compared between control and MTX resistant sample on chromosome 5: 80,000,000-83,000,000 by using Ribbon. Blue and red lines indicated the forward and reverse strands, respectively.

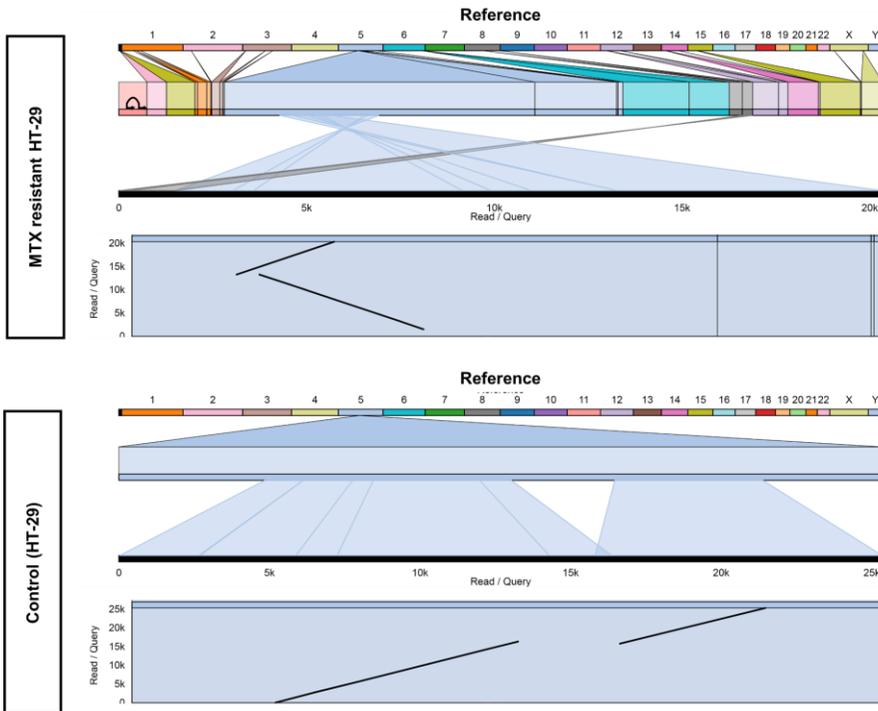


Figure 13. The single-read view and dot plot of the read over the amplified region. The forward read from single-read view was visualized and compared between control and MTX resistant sample on amplified region from alignment data (BAM) by using Ribbon. The dot plot was drawn by the single read.

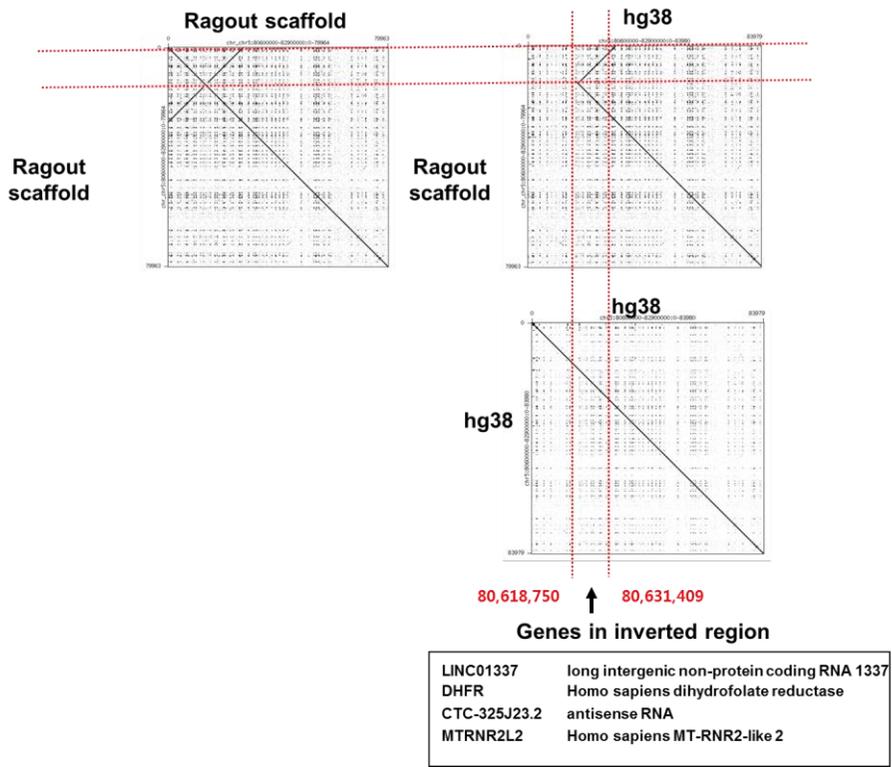


Figure 14. The scaffolding of PacBio long reads over amplified region compared to hg 38. The inversion at the start point of amplified unit was visualized by the three dot plots compared to hg38. The inverted region and its genes were indicated with the description.

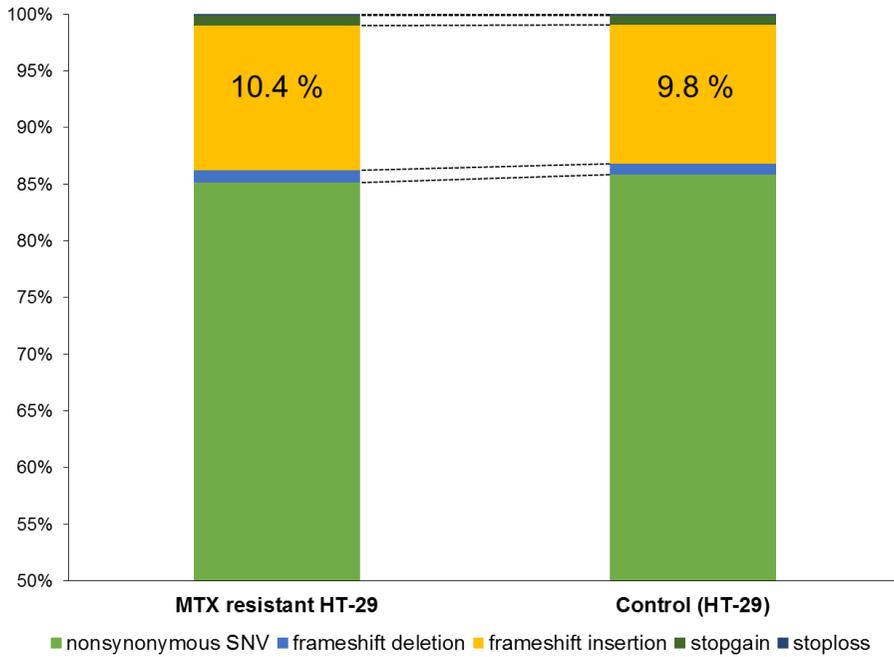


Figure 15. The difference of non-synonymous mutations between MTX resistant HT-29 and control over whole chromosomes. The non-synonymous mutations (non-synonymous SNV, frameshift deletion, frameshift insertion, stop-gain, and stop-loss) were compared between MTX resistant HT-29 and control over whole chromosomes, and the percentage of frameshift insertions was indicated.

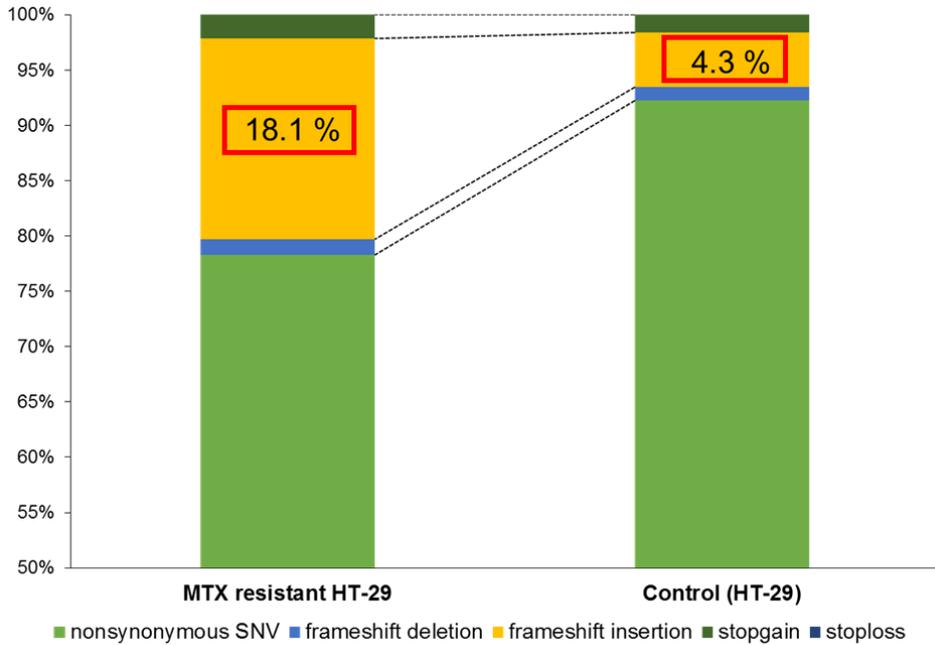


Figure 16. The comparison of non-synonymous mutations between MTX resistant HT-29 and control on chromosome 5. The non-synonymous mutations (non-synonymous SNV, frameshift deletion, frameshift insertion, stop-gain, and stop-loss) were compared between MTX resistant HT-29 and control on chromosome 5, and the percentage of frameshift insertions was indicated by the red rectangle.

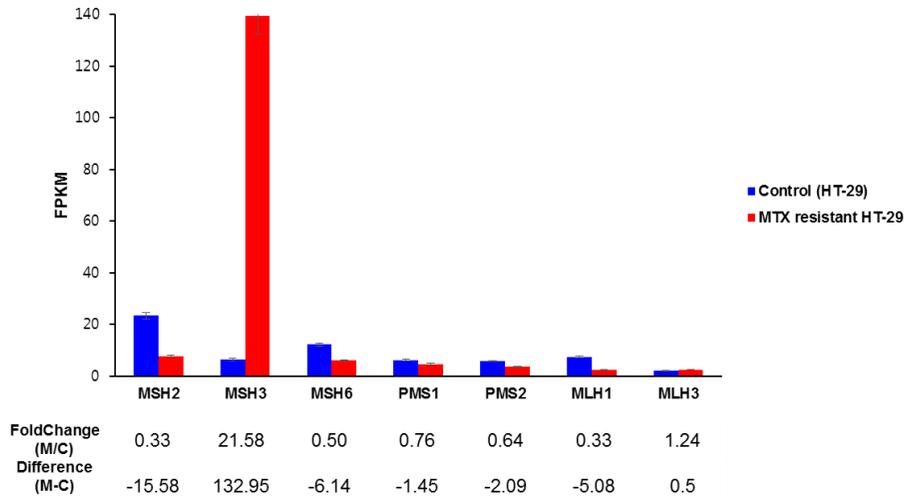


Figure 17. The comparison of MMR expression level. The gene expression (FPKM) of mutS homologs and mutL homologs was computed and compared between MTX resistant HT-29 and control. The foldchange and difference between MTX resistant HT-29 (M) and control (C) were indicated below the bar graph.

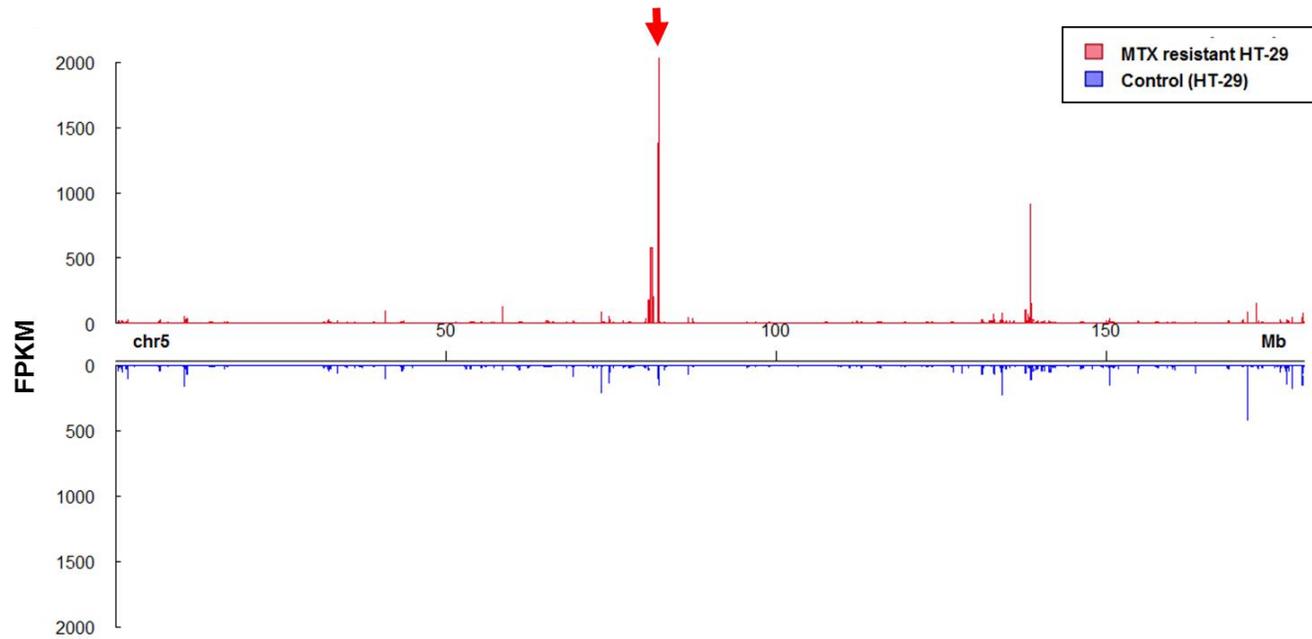


Figure 18. The comparison of expression level between MTX resistant and control samples on chromosome 5. The gene expression (FPKM) was computed and compared between MTX resistant HT-29 and control over chromosome 5. The significant high expression on the specific region was indicated by a red arrow.

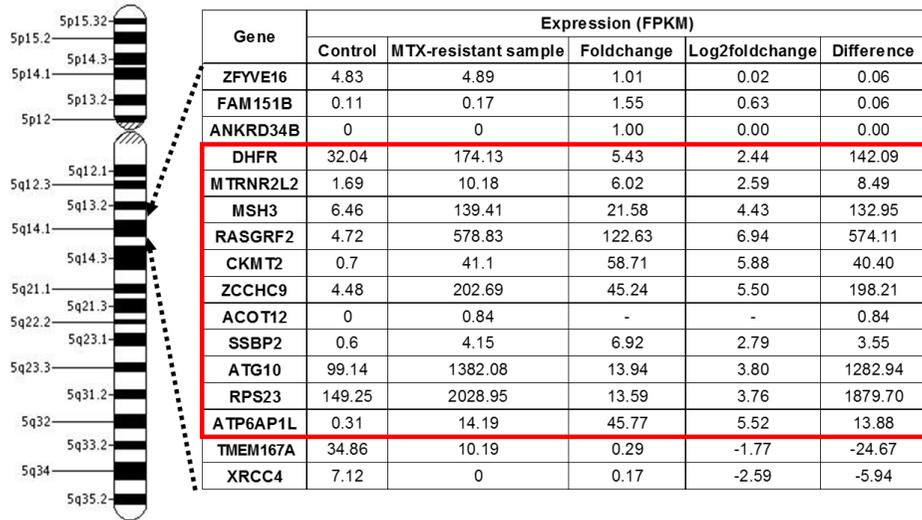


Figure 19. The expression level in 5q 14.2 region. The table depicted the expression level (FPKM) over amplified region, and the fold change and dkmmf of expression were computed between control and MTX resistant HT-29. The fold change, which is bigger than 5, was indicated with the red box. .

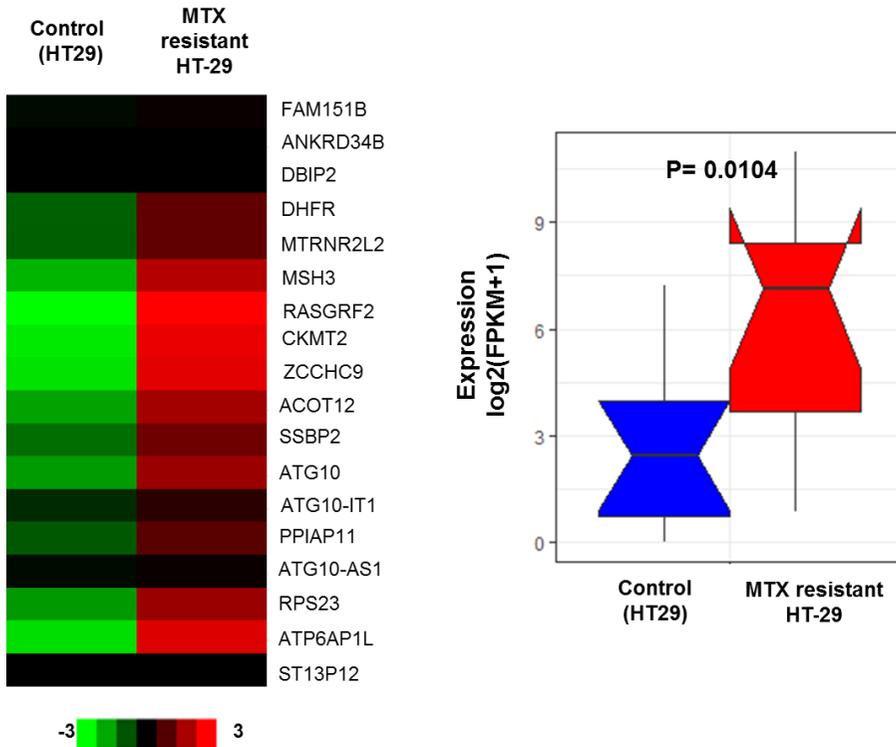


Figure 20. The comparison of expression level in the amplified regions between MTX resistant and control samples. A heatmap depicted the difference of expression level (FPKM) over amplified region between MTX resistant HT-29 and control, and the expression ($\log_2\text{fPKM}+1$) from *DHFR* gene to *ATP6AP1L* was box-plotted with statistical p-value by using Mann-Whitney U test.

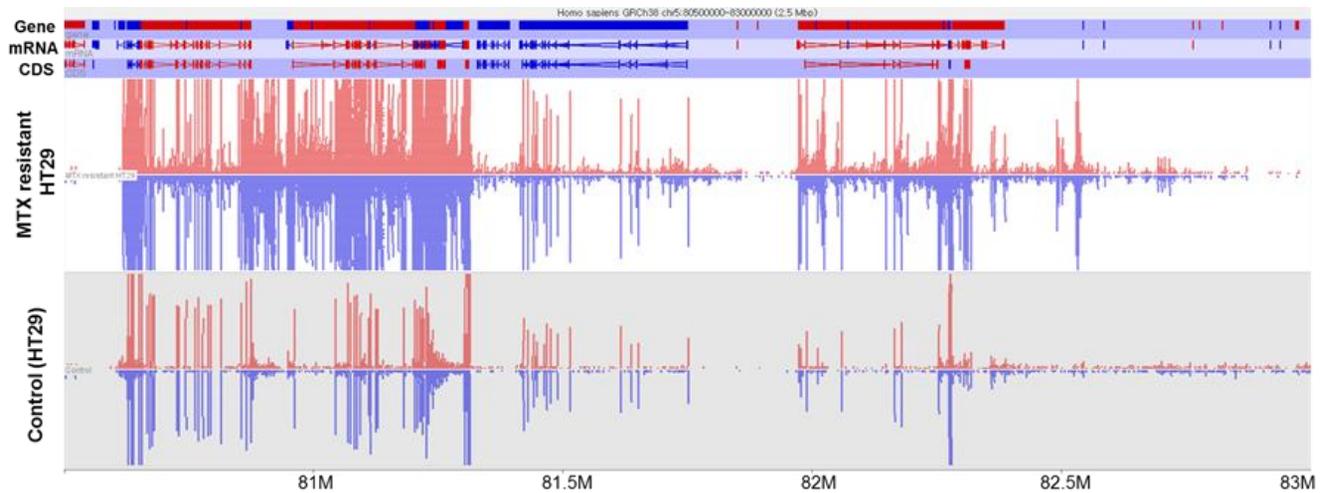


Figure 21. The visualization of mapped reads on the amplified region from transcriptome data.

The mapped reads from transcriptome sequencing data were visualized over amplification unit (chr5:80.5M-83M) by the SeqMonk.

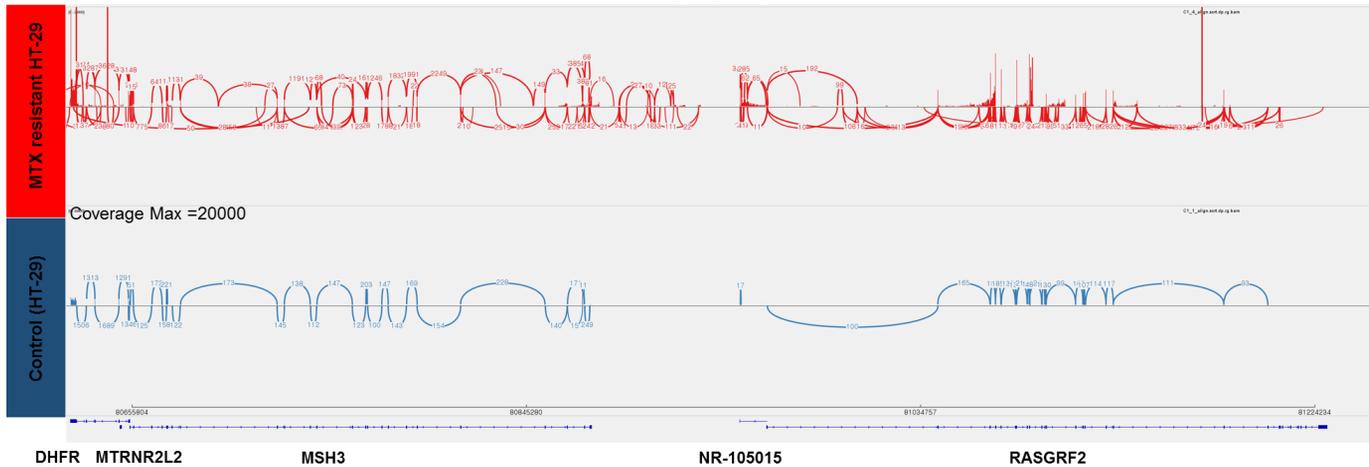


Figure 22. The identification of junctions between exons over amplification unit. The junctions between exons over amplification unit (chr5:80.5M-83M) were displayed by Sashimi plot from IGV.

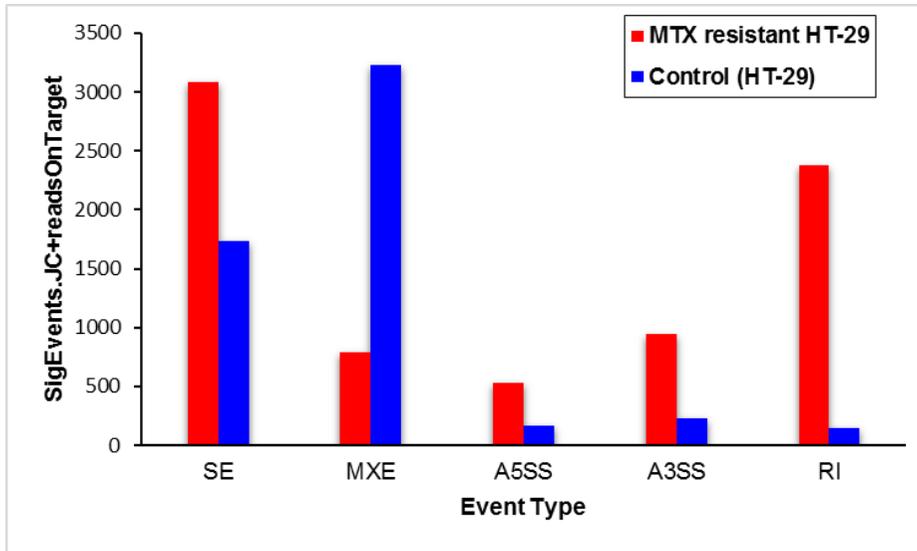


Figure 23. The comparison of five different alternative splicing events between MTX resistant and control samples. The five different types of alternative splicing patterns (SE:Skipped exon, MXE: Mutually exclusive exon, A5SS: Alternative 5' splice site, A3SS: Alternative 3' splice site, RI: Retained intron) were estimated and compared between MTX resistant HT-29 and control sample by using rMAT.

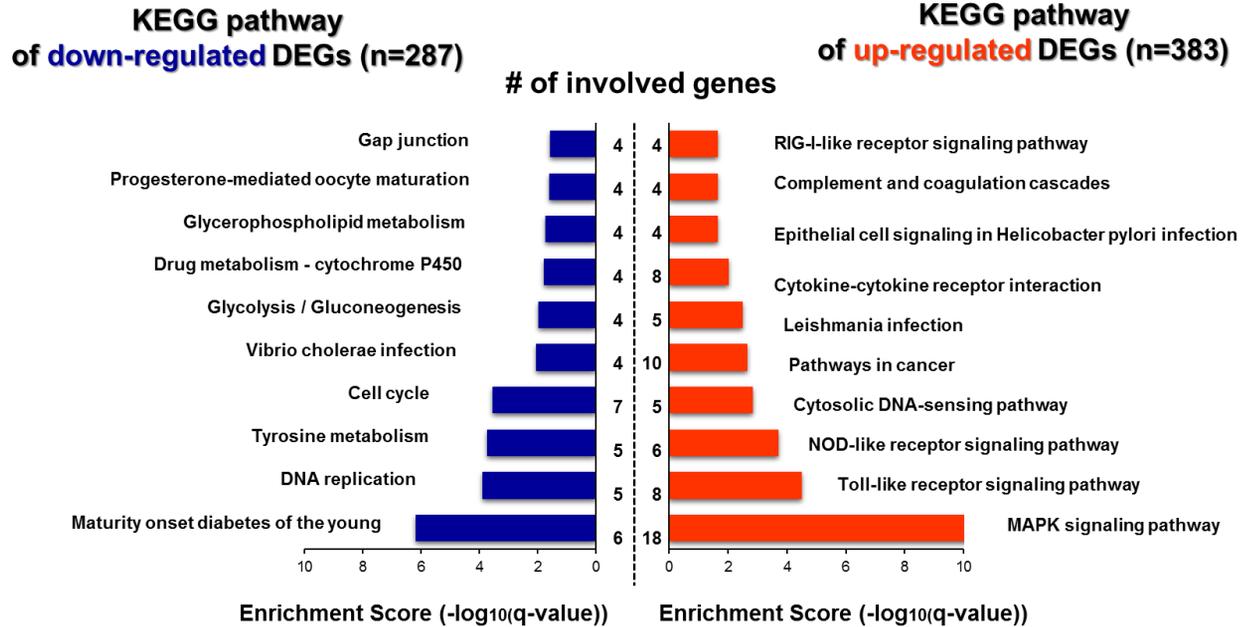


Figure 24. The identification of differentially expressed genes and enrichment with KEGG pathways. Top 10 enriched KEGG gene sets for up-regulated and down-regulated differentially expressed genes were visualized with enrichment score (-log(qvalue)).

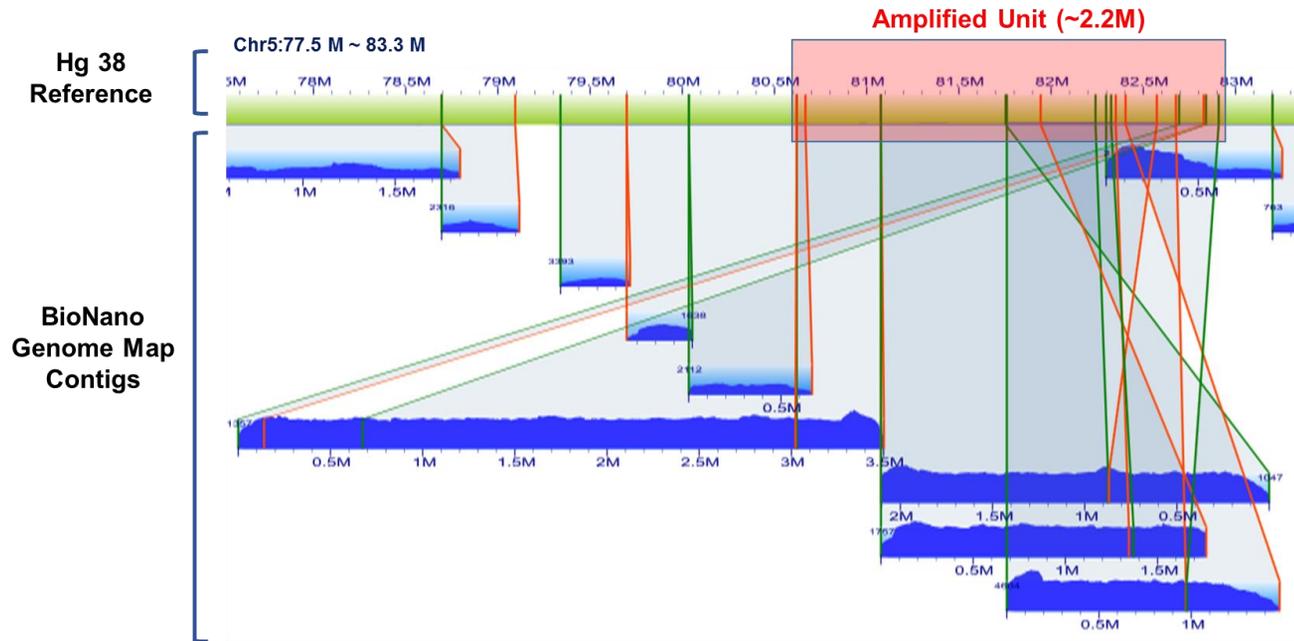


Figure 25. Genome mapping over the amplified region. The BioNano genome map contigs around amplified region were visualized and compared along with the reference (hg38). The start position of contig indicated with the green color, and the end position of contig indicated with the orange color.

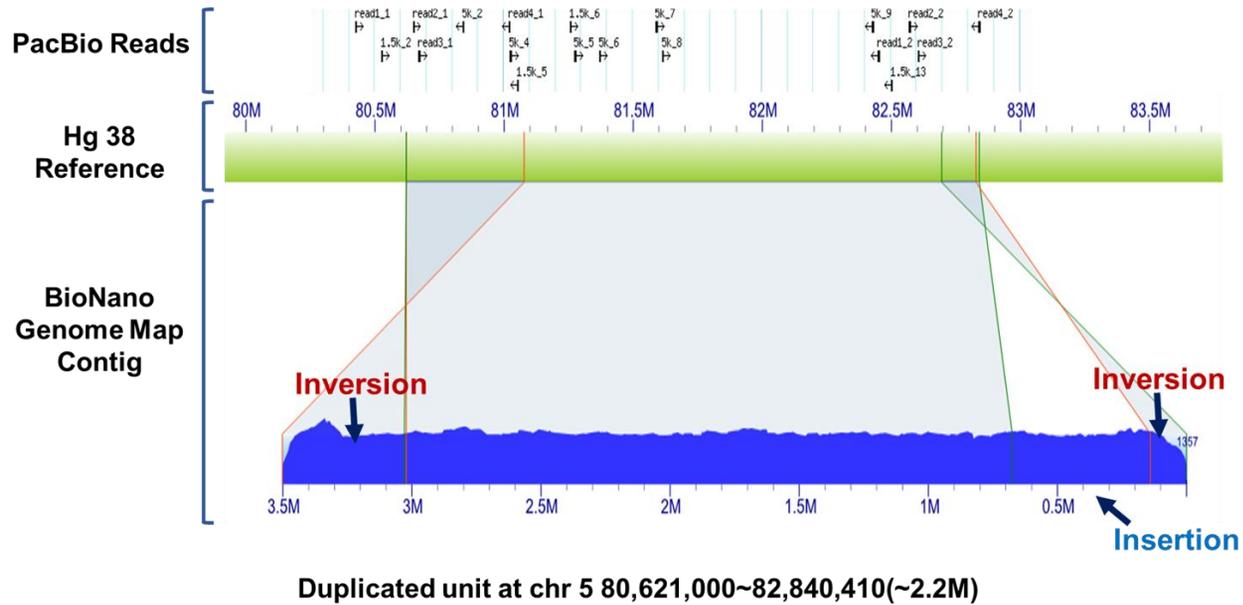


Figure 26. The detection of structural variants over the amplified region in genome mapping. The BioNano contig and the pacbio reads were matched up with the reference (hg38), and the several genomic rearrangements were indicated by blue arrow.

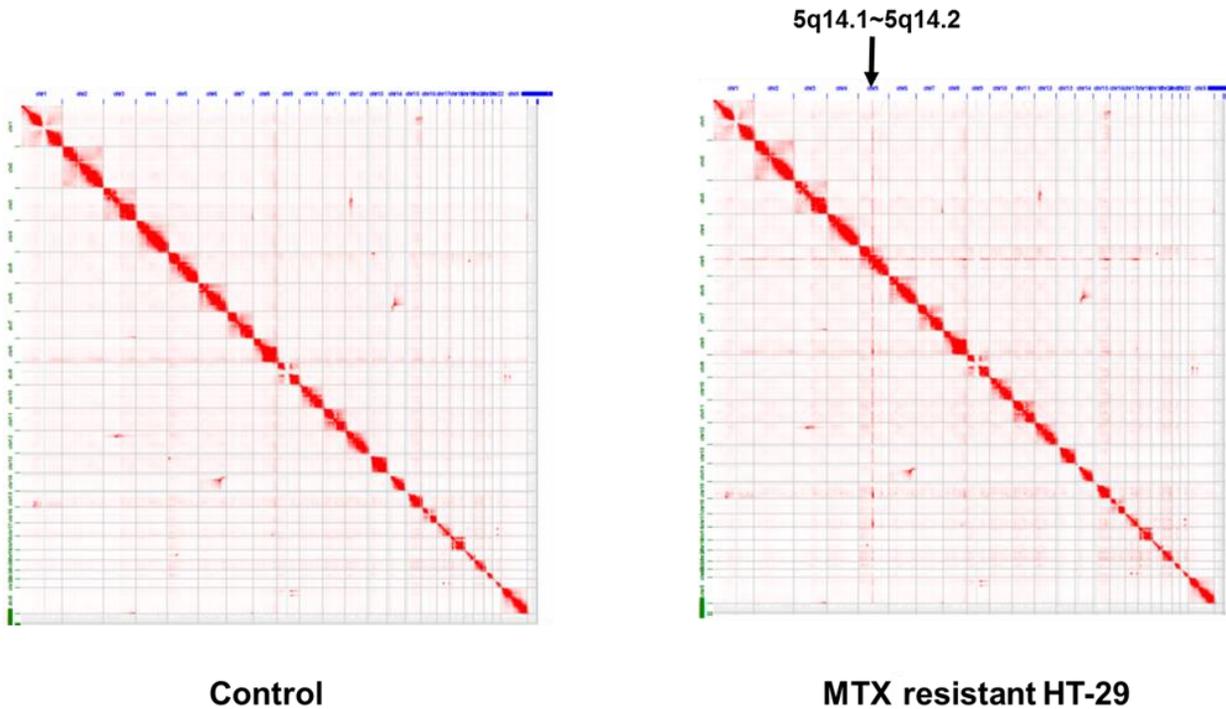


Figure 27. Genome-wide view of intra-chromosomal interactions. The intra-chromosomal interactions over genome-wide view in MTX resistant HT-29 and control were visualized by Juicebox at 5kb resolution, and the amplified region was indicated by the arrow.

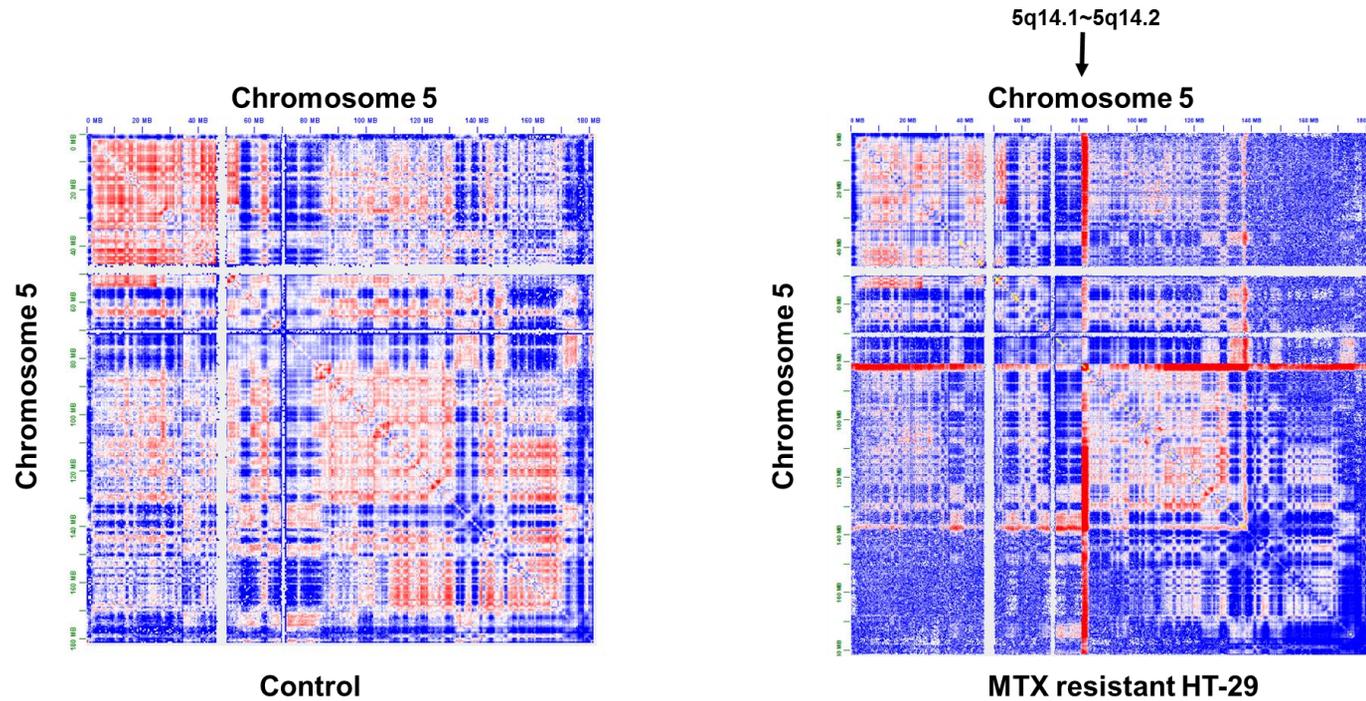


Figure 28. Intra-chromosomal interactions on chromosome 5. The intra-chromosomal interactions (Observed/Expected) over chromosome 5 in MTX resistant HT-29 and control were visualized by Juicebox at 300kb, and the amplified region was indicated by the arrow.

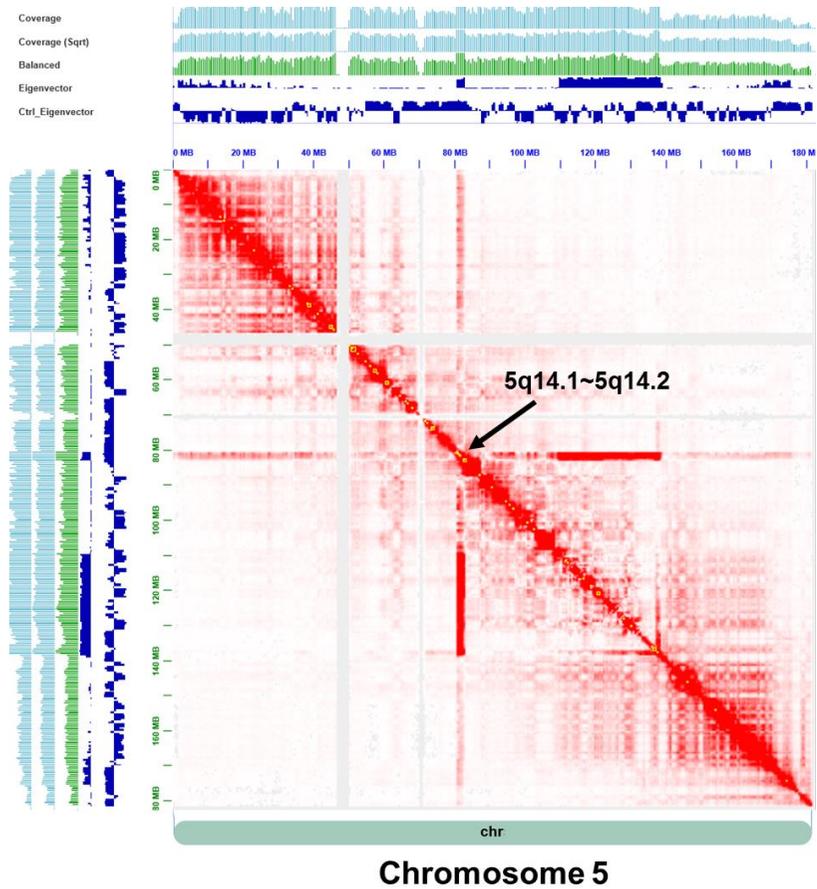


Figure 29. The topologically associating domains (TADs) on **chromosome 5**. The topologically associating domains were identified by Arrowhead algorithms and visualized with the intra-chromosomal interactions on chromosome 5. The TADs were yellow-boxed.

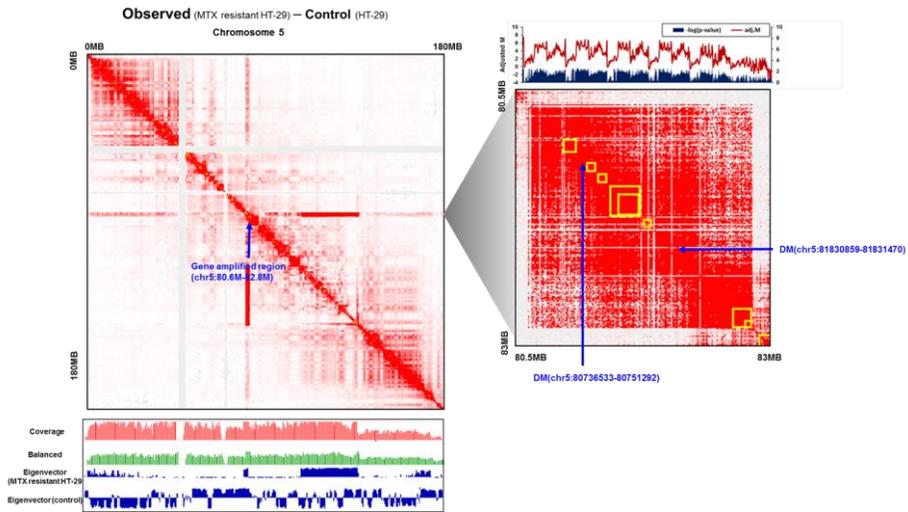


Figure 30. The topologically associating domains (TADs) on chromosome 5 and its adjusted interaction frequencies. The intra-chromosomal interactions at chromosome 5 (MTX resistant HT-29 – control) were visualized with the coverage and eigenvectors, and the newly identified topologically associating domains (TADs) on the amplified region were indicated with the yellow box at 500kb resolution.

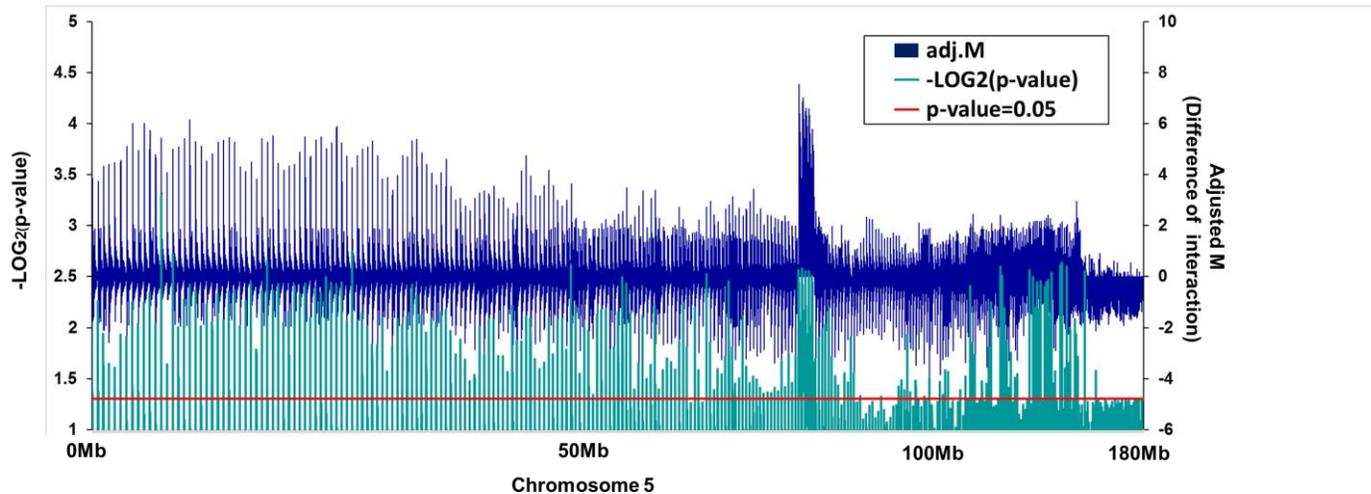


Figure 31. The comparison of adjusted intra-chromosomal interactions between MTX resistant HT-29 and control samples. The difference of adjusted interaction frequencies (adjusted M) between MTX resistant and control samples was computed, and it was plotted with the $-\log_2(p\text{-value})$. The line for statistical significance ($p\text{-value}=0.05$) was drawn by the red line.

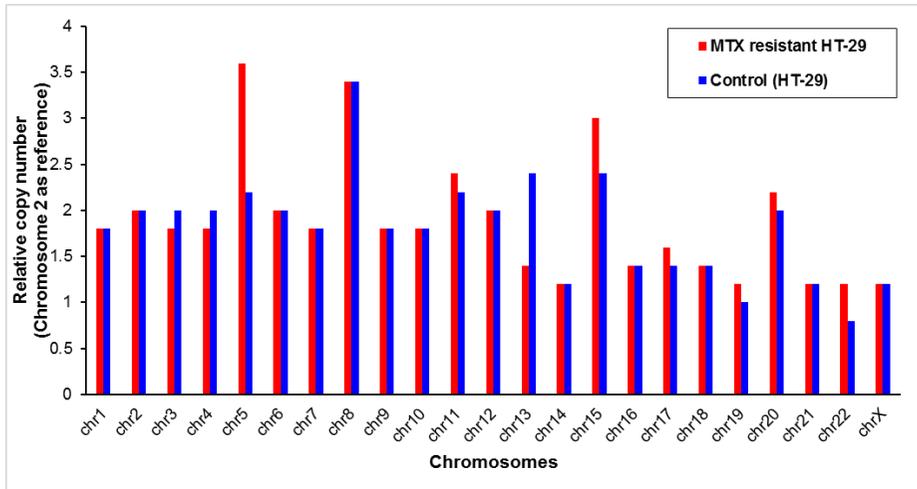


Figure 32. The comparison of relative copy number between MTX resistant and control samples. The relative copy number over whole chromosomes was estimated by HiCnv from Hi-C data. The copy number of chromosome 2 was used as the reference.

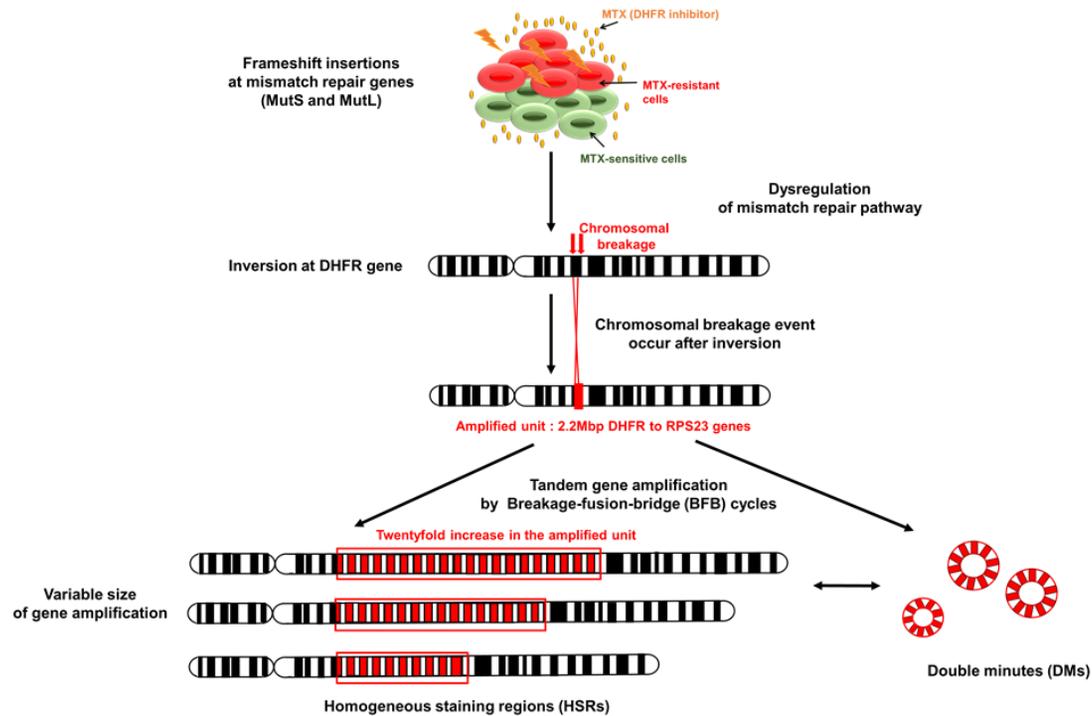


Figure 33. The mechanism of tandem gene amplification under MTX. The frameshift insertion at the family of mismatch repair genes (MutS and MutL homologs) could initiate the dysregulation of mismatch repair pathway. The chromosomal breakage event could occur after the emergence of inversion at *DHFR* gene position. Finally, the variable size of tandem amplification could be done by Breakage fusion-bridge (BFB) cycles.

DISCUSSION

High-throughput sequencing based methods have emerged as an attractive method for structural variant identification, but the short reads sequencing cannot effectively map and capture the full range of the repetitive regions in human genome [84, 85]. Additionally, the FISH as well as short read sequencing have an inherently low resolution and are the low throughput method that cannot adequately characterize extensively cancer genomes, which possess a variety of structural variations such as insertion, deletion, duplication, translocation, and inversion [86, 87].

Also, cancer cells which are evolved under anti-cancer drugs, occasionally have high copies of specific genes, but its mechanism is fully unknown at the molecular level since repetitive DNA sequences possess many technical difficulties and problems for alignment of NGS reads and data processing [88, 89]. Now, the recent developments of NGS technologies by Pacific Biosciences, Oxford Nanopore, and others, which produce the large size reads (~10Kb), are able to generate relatively huge amounts of genomic information at base level resolution, and this provides the right approaches to the analysis of unknown region, which was not described by the previous technologies [90-92].

In 1966, the fact that the cancer cells were resistant to MTX anti-cancer

drug began to be known, and characteristics of MTX resistant cells from the high *DHFR* enzyme activity to the raised *DHFR* gene copies started to be described from 1978 [93, 94]. Finally, the new types of chromosomal abnormalities such as HSR, which is usually longer than any single band in the karyotype and represent site of *DHFR* gene amplification, and DM which is extrachromosomal elements, have been emerged and received the attention of scientists [95-97]. However, the detection and recognition of abnormal structures were limited because of the involved repetitive rearrangements and tumor heterogeneity in cancer genome, and it was difficult to analyze the amplified region through previous technologies and its bioinformatics skills [98, 99].

In this study, the complicated cancer genomics and abnormal structures were detected and analyzed by using a combination of advanced technologies including PacBio SMRT, optical mapping, and Hi-C analysis as well as previous technologies such as short-read sequencing and FISH analysis, and an integrative framework to detect and decipher structural variations in MTX resistant HT-29 cells was proposed, which had the large repeat and complex DNA segments. Finally, the amplified region was characterized and evaluated by its quantified size and involved genetic defects and the possible gene amplification mechanism was suggested at the end.

In order to analyze the complicated repetitive sequence in gene amplification under MTX, the several MTX resistant clones were

selected and generated from HT-29 cell line by increasing MTX concentration through MTX sensitization study. However, when the amplified *DHFR* gene from several clones was detected by using the FISH technology, the cells in each clone had very a large variation of the 5q12 region, and *DHFR* gene amplified in only low portion of cells since the response of the tumor cells were heterogeneous to MTX drug and the effects of drug would be dissimilar even if the cells were under a same condition [100, 101]. This could produce the difficulties in bioinformatics analysis of accurate amplified patterns and repetitive sequence in the previous as well as future study.

Therefore, in order to overcome the limited analysis provided by the heterogeneously amplified patterns, *DHFR* gene amplified patterns in the specific clone were optimized by using the serial dilution and repeated single cell selection to generate the homogenous amplified patterns of *DHFR* gene. The patterns of amplified *DHFR* genes in the optimized resistant HT-29 cells (C1-2-4) were confirmed and identified by FISH analysis, and finally the optimized clone was obtained, which had total 96 % *DHFR* gene amplified cells, and the 75% of cells in clone were converged to a homogeneous *DHFR* gene amplified pattern.

After obtaining the optimized resistant clone, the amplified unit and tandem gene amplification of 11 genes from *DHFR* gene to *ATP6AP1L* gene on chr5 (2.2Mbp) was identified through the long-range genomic information from long read sequencing and optical genome mapping.

The amplified unit had the high coverage (~197X) compared to the control (~10X), and this inferred that the amplified region was about twenty-fold tandemly amplified to original sequence. This result was confirmed from the high gene expression pattern and alternatively splicing patterns. The gene expression on amplified unit was highly overexpressed from 5-fold change to 122-fold change compared to control, and this variable expression patterns could be affected by the complexity of splicing patterns even if these genes were amplified simultaneously.

Also, the inversion at both start and end points of the amplified unit was detected in long read sequencing data, and this structure variant was cross-validated by optical genome mapping data. Previous studies have also shown that the inversion in gene amplification could lead to the initiation of gene amplification by stimulating chromosomal breakage and possibly represent the transformation from circular DNA segments DMs to intra-chromosomal HSRs [102, 103].

Using the most advance technology, Hi-C analysis, the high intra-chromosomal interactions on the amplified region and the significant interactions at the novel TADs level were identified. Also, the long ranged chromosomal interaction from 109Mbp to 138Mb more than amplified region was detected as unexpected. The amplified unit in Hi-C was exactly match up with the amplified unit of long-range genomic information and RNA sequencing result, and this could explain that

amplified position was more closely located and had the frequent contact each other, which could make represent the chromatin architecture for gene amplification [104].

Additionally, using the chromosomal interactions, this study found that the relative copy number in chromosome 5 was significantly higher in MTX resistant HT-29 compared to control and the DMs, which were extrachromosomal DNA and another type of chromosomal abnormality were detected in the amplified region, which was not detected by FISH analysis. This could indicate that this structure on highly compacted chromosome position harbor the amplification of several genes by generating the new chromosomal structure DM and dynamically converting to another type of chromosomal structure HSR, and this could provide the explanation of the involvement of chromosomal abnormalities in MTX drug resistance at the end.

Still, the involved mechanism of conversion from HSRs to DMs was not explained because of the unknown time point for transition between two structures, its low coverage (10X) of sequencing data, and a large variation in amplification size, which produced the difficulties for detection and investigation.

From the identification of amplified unit and intra-chromosomal interactions, this study seems to emphasize that *DHFR* gene could be not the only target for the MTX associated mechanism since 11

unsuspected genes from *DHFR* to *ATP6AP1L* were tandemly amplified, and expanded region on 5q arm over the amplified region was interacted each other. The highest gene expression in amplified unit was *RASGRF2*, which was 122 fold increased to control, not *DHFR* gene, which was only five fold increased to control. Therefore, the future study is needed to target and investigate the role of *RASGRF2* gene instead of *DHFR*, which might utilize the promoter of *DHFR* gene to obtain their high efficiency under MTX condition.

In addition, the novel frameshift insertions in *MSH* and *MLH* genes were identified in the MTX resistant sample, which could play an important role on the rapid progression of gene amplification as well as being resistant to MTX. The microsatellite instability (MSI), which has been known as the effect of deficient DNA mismatch repair in colon cancer, was tested for the status of MMR in MTX resistant HT-29, but there was no significant difference between MTX resistant and control sample [105].

This inferred that this could not affect the genetic instability and entire system of MMR over whole chromosomes whereas the MTX toxicity could produce the mutations on MMR genes and genetic predisposition on chromosome 5 in MTX resistant HT-29 by inserting adenine or thymine nucleotide on *MSH* and *MLH* genes and finally harbor gene amplification. This would provide the additional molecular explanations for possible tandem gene amplification mechanism, which was

progressed through Breakage-fusion bridge cycles.

Therefore, the future study is needed to find whether the frameshift insertion in *MSH3*, might be caused by the co-amplification with *DHFR* gene, as well as other frameshift insertions result in the malfunction of MutS homologs and hyper-mutability under the high MTX condition as previously described [106, 107]. However, it seems that the frameshift insertions in *MSH* and *MLH* genes could not be generated by the gene amplification process unlike *MSH3* since all mutated genes except for *MSH3* are located outside of the amplified unit. Therefore, the frameshift mutations in each genes would be originated from the different cause under MTX toxicity and have different impacts on the MTX resistant HT-29 cells.

In addition, the mismatch repair location is recognized by two MutS homologs such as MutS-alpha (*MSH2* and *MSH6*), which is known for repair of single nucleotide mismatches and MuS-beta (*MSH2* and *MSH3*), which is known for repair the large size of indels [108, 109]. The MutS homologs should need MutL homologs (*MLH1* and *PMS2*) for binding to recognition site.

Therefore, the mutations in MutS and MutL homologs could prevent repairing from the single nucleotide mismatch and large size indels in MTX resistant HT-29. The prevention of mutations in MutS and MutL homologs and inversion could increase the sensitivity of MTX and finally

inhibit genes from amplifying. For the future study, the methods for preventing these mutations and structural variants could be developed for overcoming the drug resistance and optimal clinical use of this drug.

Most of all, the additional validation for identified SVs and repetitive sequences would be required for interpreting all steps in gene amplification mechanism and impact of MTX on the gene amplification, and finally for offering a clue for cancer adaptation. Also, it would be discussed how the inversions at start and end position on the amplified unit affect the chromosomal breakage and formation of the DMs from repetitive sequence. Additionally, the effects of MTX associated mechanism on frameshift insertions in *MMR* and *MLH* genes as well as inversion on amplified region would be described and validated through the comprehensive analysis of involved genes and pathways,

Although several limitations exist in this study, this findings may give new insight into the mechanism underlying the amplification process and evolution of resistance to MTX in colon cancer as well as in leukemia. Also, the possible use of Hi-C data to detect the unsuspected chromosomal rearrangements like DMs and HSRs has been proved for the future analysis.

Overall, the complex NGS technologies could be a clever approach to identify the complicated genomic sequence and novel structural variants that are difficult to detect with previous technologies. Whenever possible,

a comprehensive approach of different technologies is required for interpretation of repetitive sequence and structural variations in gene amplification, and this will help to identify the most important therapeutic mechanism and the new targets of anti-cancer drug, which affect various intracellular pathways at many levels.

Finally, it will support the basis of clinical cancer study such as diagnosis, management, and treatment and provide the depth of insight toward the pharmacology of the anti-cancer drugs and a step towards personalized medicine.

References

1. LeBlanc, V.G. and M.A. Marra, *Next-Generation Sequencing Approaches in Cancer: Where Have They Brought Us and Where Will They Take Us?* *Cancers* (Basel), 2015. **7**(3): p. 1925-58.
2. Meldrum, C., M.A. Doyle, and R.W. Tothill, *Next-generation sequencing for cancer diagnostics: a practical perspective*. *Clin Biochem Rev*, 2011. **32**(4): p. 177-95.
3. Nakagawa, H. and M. Fujita, *Whole genome sequencing analysis for cancer genomics and precision medicine*. *Cancer Sci*, 2018. **109**(3): p. 513-522.
4. Zhang, J., et al., *The impact of next-generation sequencing on genomics*. *J Genet Genomics*, 2011. **38**(3): p. 95-109.
5. Marino, P., et al., *Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study*. *Eur J Hum Genet*, 2018. **26**(3): p. 314-323.
6. Holcomb, I.N., et al., *Genomic alterations indicate tumor origin and varied metastatic potential of disseminated cells from prostate cancer patients*. *Cancer Res*, 2008. **68**(14): p. 5599-608.
7. Kamps, R., et al., *Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification*. *Int J Mol Sci*, 2017. **18**(2).
8. Bieg-Bourne, C.C., et al., *Next-Generation Sequencing in the Clinical Setting Clarifies Patient Characteristics and Potential Actionability*. *Cancer Res*, 2017. **77**(22): p. 6313-6320.
9. Marotta, M., et al., *A common copy-number breakpoint of ERBB2 amplification in breast cancer colocalizes with a complex block of segmental duplications*. *Breast Cancer Res*, 2012. **14**(6): p. R150.
10. Mathew, P., et al., *Detection of MYCN gene amplification in neuroblastoma by fluorescence in situ hybridization: a pediatric oncology group study*. *Neoplasia*, 2001. **3**(2): p. 105-9.
11. Schwab, M., *Oncogene amplification in solid tumors*. *Semin Cancer*

- Biol, 1999. **9**(4): p. 319-25.
12. Kaufman, R.J. and R.T. Schimke, *Amplification and loss of dihydrofolate reductase genes in a Chinese hamster ovary cell line*. Mol Cell Biol, 1981. **1**(12): p. 1069-76.
 13. Ohta, J.I., et al., *Fluorescence in situ hybridization evaluation of c-erbB-2 gene amplification and chromosomal anomalies in bladder cancer*. Clin Cancer Res, 2001. **7**(8): p. 2463-7.
 14. Dacic, S., et al., *Significance of EGFR protein expression and gene amplification in non-small cell lung carcinoma*. Am J Clin Pathol, 2006. **125**(6): p. 860-5.
 15. Kamel, H.F.M. and H. Al-Amodi, *Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine*. Genomics Proteomics Bioinformatics, 2017. **15**(4): p. 220-235.
 16. Schimke, R.T., *Gene amplification in cultured animal cells*. Cell, 1984. **37**(3): p. 705-13.
 17. Kang, J.U., *Characterization of amplification patterns and target genes on the short arm of chromosome 7 in early-stage lung adenocarcinoma*. Mol Med Rep, 2013. **8**(5): p. 1373-8.
 18. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions*. Nat Rev Genet, 2011. **13**(1): p. 36-46.
 19. Schneeberger, K., et al., *Simultaneous alignment of short reads against multiple genomes*. Genome Biol, 2009. **10**(9): p. R98.
 20. Brown, P.C., T.D. Tlsty, and R.T. Schimke, *Enhancement of methotrexate resistance and dihydrofolate reductase gene amplification by treatment of mouse 3T6 cells with hydroxyurea*. Mol Cell Biol, 1983. **3**(6): p. 1097-107.
 21. Goker, E., et al., *Amplification of the dihydrofolate reductase gene is a mechanism of acquired resistance to methotrexate in patients with acute lymphoblastic leukemia and is correlated with p53 gene mutations*. Blood, 1995. **86**(2): p. 677-84.
 22. Morales, C., et al., *Genetic determinants of methotrexate responsiveness and resistance in colon cancer cells*. Oncogene,

2005. **24**(45): p. 6842-7.
23. Meng, X., et al., *Novel role for non-homologous end joining in the formation of double minutes in methotrexate-resistant colon cancer cells*. J Med Genet, 2015. **52**(2): p. 135-44.
24. Singer, M.J., et al., *Amplification of the human dihydrofolate reductase gene via double minutes is initiated by chromosome breaks*. Proc Natl Acad Sci U S A, 2000. **97**(14): p. 7921-6.
25. Hayes, M. and J. Li, *An integrative framework for the identification of double minute chromosomes using next generation sequencing data*. BMC Genet, 2015. **16 Suppl 2**: p. S1.
26. Sedlazeck, F.J., et al., *Accurate detection of complex structural variations using single-molecule sequencing*. Nat Methods, 2018. **15**(6): p. 461-468.
27. Arnoldo, A., et al., *A genome scale overexpression screen to reveal drug activity in human cells*. Genome Med, 2014. **6**(4): p. 32.
28. Mencia, N., et al., *Overexpression of S100A4 in human cancer cell lines resistant to methotrexate*. BMC Cancer, 2010. **10**: p. 250.
29. Saez-Ayala, M., et al., *Melanoma coordinates general and cell-specific mechanisms to promote methotrexate resistance*. Exp Cell Res, 2012. **318**(10): p. 1146-59.
30. Salzano, A., et al., *Enhanced gene amplification in human cells knocked down for DNA-PKcs*. DNA Repair (Amst), 2009. **8**(1): p. 19-28.
31. Morales, C., et al., *Dihydrofolate reductase amplification and sensitization to methotrexate of methotrexate-resistant colon cancer cells*. Mol Cancer Ther, 2009. **8**(2): p. 424-32.
32. Ham, R.G., *Clonal Growth of Mammalian Cells in a Chemically Defined, Synthetic Medium*. Proc Natl Acad Sci U S A, 1965. **53**: p. 288-93.
33. Bleiker, E.M., et al., *100 years Lynch syndrome: what have we learned about psychosocial issues?* Fam Cancer, 2013. **12**(2): p. 325-39.
34. Rio, D.C., et al., *Purification of RNA using TRIzol (TRI reagent)*. Cold Spring Harb Protoc, 2010. **2010**(6): p. pdb prot5439.

35. Parra, I. and B. Windle, *High resolution visual mapping of stretched DNA by fluorescent hybridization*. Nat Genet, 1993. **5**(1): p. 17-21.
36. Trask, B.J., *Fluorescence in situ hybridization: applications in cytogenetics and gene mapping*. Trends Genet, 1991. **7**(5): p. 149-54.
37. Nagy, A., et al., *Karyotyping mouse cells*. CSH Protoc, 2008. **2008**: p. pdb prot4706.
38. Ju, Y.S., et al., *Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals*. Nat Genet, 2011. **43**(8): p. 745-52.
39. Seo, J.S., et al., *Whole exome and transcriptome analyses integrated with microenvironmental immune signatures of lung squamous cell carcinoma*. Cancer Immunol Res, 2018.
40. Seo, J.S., et al., *Comprehensive analysis of the tumor immune micro-environment in non-small cell lung cancer for efficacy of checkpoint inhibitor*. Sci Rep, 2018. **8**(1): p. 14576.
41. Wang, M., et al., *Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication*. Nat Genet, 2017. **49**(4): p. 579-587.
42. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data*. Genome Res, 2010. **20**(9): p. 1297-303.
43. Seo, J.S., et al., *The transcriptional landscape and mutational profile of lung adenocarcinoma*. Genome Res, 2012. **22**(11): p. 2109-19.
44. Shen, S., et al., *rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data*. Proc Natl Acad Sci U S A, 2014. **111**(51): p. E5593-601.
45. Seo, J.S., et al., *De novo assembly and phasing of a Korean human genome*. Nature, 2016. **538**(7624): p. 243-247.
46. Nattestad, M., et al., *Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line*. Genome Res, 2018. **28**(8): p. 1126-1135.
47. Nattestad, M., C.-S. Chin, and M.C. Schatz, *Ribbon: Visualizing complex genome alignments and structural variation*. bioRxiv, 2016.

48. Chan, E.K.F., et al., *Optical mapping reveals a higher level of genomic architecture of chained fusions in cancer*. *Genome Res*, 2018. **28**(5): p. 726-738.
49. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome*. *Science*, 2009. **326**(5950): p. 289-93.
50. Shindo, Y., et al., *Predictive biomarkers for the efficacy of peptide vaccine treatment: based on the results of a phase II study on advanced pancreatic cancer*. *J Exp Clin Cancer Res*, 2017. **36**(1): p. 36.
51. Servant, N., et al., *HiC-Pro: an optimized and flexible pipeline for Hi-C data processing*. *Genome Biol*, 2015. **16**: p. 259.
52. Chakraborty, A. and F. Ay, *Identification of copy number variations and translocations in cancer cells from Hi-C data*. *Bioinformatics*, 2017.
53. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping*. *Cell*, 2014. **159**(7): p. 1665-80.
54. Stansfield, J.C., et al., *HiCcompare: an R-package for joint normalization and comparison of Hi-C datasets*. *BMC Bioinformatics*, 2018. **19**(1): p. 279.
55. Feder, J.N., et al., *The pattern of dihydrofolate reductase expression through the cell cycle in rodent and human cultured cells*. *J Biol Chem*, 1989. **264**(34): p. 20583-90.
56. Baumer, C., et al., *Exploring DNA quality of single cells for genome analysis with simultaneous whole-genome amplification*. *Sci Rep*, 2018. **8**(1): p. 7476.
57. Ulahannan, D., et al., *Technical and implementation issues in using next-generation sequencing of cancers in clinical practice*. *Br J Cancer*, 2013. **109**(4): p. 827-35.
58. Schildhaus, H.U., et al., *Definition of a fluorescence in-situ hybridization score identifies high- and low-level FGFR1 amplification types in squamous cell lung cancer*. *Mod Pathol*, 2012. **25**(11): p. 1473-80.

59. Schimke, R.T., *Gene amplification and drug resistance*. Sci Am, 1980. **243**(5): p. 60-9.
60. Potapova, T.A., J. Zhu, and R. Li, *Aneuploidy and chromosomal instability: a vicious cycle driving cellular evolution and cancer genome chaos*. Cancer Metastasis Rev, 2013. **32**(3-4): p. 377-89.
61. Ashktorab, H., et al., *Distinct genetic alterations in colorectal cancer*. PLoS One, 2010. **5**(1): p. e8879.
62. Perera, L., et al., *Requirement for transient metal ions revealed through computational analysis for DNA polymerase going in reverse*. Proc Natl Acad Sci U S A, 2015. **112**(38): p. E5228-36.
63. Kimura, S., et al., *Template-dependent nucleotide addition in the reverse (3'-5') direction by Thg1-like protein*. Sci Adv, 2016. **2**(3): p. e1501397.
64. Lu, S., et al., *Short Inverted Repeats Are Hotspots for Genetic Instability: Relevance to Cancer Genomes*. Cell Rep, 2015.
65. Reams, A.B. and J.R. Roth, *Mechanisms of gene duplication and amplification*. Cold Spring Harb Perspect Biol, 2015. **7**(2): p. a016592.
66. Miller, A., et al., *High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma*. Blood Cancer J, 2017. **7**(9): p. e612.
67. Edelmann, W., et al., *The DNA mismatch repair genes Msh3 and Msh6 cooperate in intestinal tumor suppression*. Cancer Res, 2000. **60**(4): p. 803-7.
68. Karahan, B., et al., *Relationship between MLH-1, MSH-2, PMS-2, MSH-6 expression and clinicopathological features in colorectal cancer*. Int J Clin Exp Pathol, 2015. **8**(4): p. 4044-53.
69. Edelmann, W., et al., *Mutation in the mismatch repair gene Msh6 causes cancer susceptibility*. Cell, 1997. **91**(4): p. 467-77.
70. Silva, F.C., et al., *Mismatch repair genes in Lynch syndrome: a review*. Sao Paulo Med J, 2009. **127**(1): p. 46-51.
71. Kariola, R., et al., *Functional analysis of MSH6 mutations linked to kindreds with putative hereditary non-polyposis colorectal cancer syndrome*. Hum Mol Genet, 2002. **11**(11): p. 1303-10.

72. Drummond, J.T., et al., *DHFR/MSH3 amplification in methotrexate-resistant cells alters the hMutSalpha/hMutSbeta ratio and reduces the efficiency of base-base mismatch repair*. Proc Natl Acad Sci U S A, 1997. **94**(19): p. 10144-9.
73. Parra, M.K., et al., *Alternative 5' exons and differential splicing regulate expression of protein 4.1R isoforms with distinct N-termini*. Blood, 2003. **101**(10): p. 4164-71.
74. Wessagowit, V., et al., *Normal and abnormal mechanisms of gene splicing and relevance to inherited skin diseases*. J Dermatol Sci, 2005. **40**(2): p. 73-84.
75. Guo, Y., et al., *Identification of Key Candidate Genes and Pathways in Colorectal Cancer by Integrated Bioinformatical Analysis*. Int J Mol Sci, 2017. **18**(4).
76. Yang, X.R., et al., *Identification of genes associated with methotrexate resistance in methotrexate-resistant osteosarcoma cell lines*. J Orthop Surg Res, 2015. **10**: p. 136.
77. Uchiyama, H., et al., *Cyclin-dependent kinase inhibitor SU9516 enhances sensitivity to methotrexate in human T-cell leukemia Jurkat cells*. Cancer Sci, 2010. **101**(3): p. 728-34.
78. Sramek, M., J. Neradil, and R. Veselska, *Much more than you expected: The non-DHFR-mediated effects of methotrexate*. Biochim Biophys Acta Gen Subj, 2017. **1861**(3): p. 499-503.
79. Windle, B., et al., *A central role for chromosome breakage in gene amplification, deletion formation, and amplicon integration*. Genes Dev, 1991. **5**(2): p. 160-74.
80. Tanaka, H., et al., *Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells*. Proc Natl Acad Sci U S A, 2002. **99**(13): p. 8772-7.
81. Harewood, L., et al., *Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours*. Genome Biol, 2017. **18**(1): p. 125.
82. Matsui, A., et al., *Gene amplification: mechanisms and involvement in cancer*. Biomol Concepts, 2013. **4**(6): p. 567-82.
83. Fisher, M.J., et al., *2016 Children's Tumor Foundation conference on*

- neurofibromatosis type 1, neurofibromatosis type 2, and schwannomatosis*. Am J Med Genet A, 2018. **176**(5): p. 1258-1269.
84. Inaki, K., et al., *Systems consequences of amplicon formation in human breast cancer*. Genome Res, 2014. **24**(10): p. 1559-71.
85. Merker, J.D., et al., *Long-read genome sequencing identifies causal structural variation in a Mendelian disease*. Genet Med, 2018. **20**(1): p. 159-163.
86. Abel, H.J. and E.J. Duncavage, *Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches*. Cancer Genet, 2013. **206**(12): p. 432-40.
87. Yang, L., et al., *Diverse mechanisms of somatic structural variations in human cancer genomes*. Cell, 2013. **153**(4): p. 919-29.
88. Gupta, S., et al., *Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine*. Sci Rep, 2016. **6**: p. 23857.
89. Hillcoat, B.L., V. Swett, and J.R. Bertino, *Increase of dihydrofolate reductase activity in cultured mammalian cells after exposure to methotrexate*. Proc Natl Acad Sci U S A, 1967. **58**(4): p. 1632-7.
90. Kulkarni, P. and P. Frommolt, *Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows*. Comput Struct Biotechnol J, 2017. **15**: p. 471-477.
91. Tattini, L., R. D'Aurizio, and A. Magi, *Detection of Genomic Structural Variants from Next-Generation Sequencing Data*. Front Bioeng Biotechnol, 2015. **3**: p. 92.
92. Sedlazeck, F.J., et al., *Piercing the dark matter: bioinformatics of long-range sequencing and mapping*. Nat Rev Genet, 2018. **19**(6): p. 329-346.
93. Schimke, R.T., et al., *Gene amplification and drug resistance in cultured murine cells*. Science, 1978. **202**(4372): p. 1051-5.
94. Hall, T.C., D. Roberts, and D.H. Kessel, *Methotrexate and folic reductase in human cancer*. Eur J Cancer, 1966. **2**(2): p. 135-42.
95. Schimke, R.T., *Gene amplification, drug resistance, and cancer*. Cancer Res, 1984. **44**(5): p. 1735-42.
96. Schimke, R.T., P.C. Brown, and R.J. Kaufman, *Gene amplification and*

- drug resistance in mammalian cells*. Natl Cancer Inst Monogr, 1982. **60**: p. 79-86.
97. Wahl, G.M., *The importance of circular DNA in mammalian gene amplification*. Cancer Res, 1989. **49**(6): p. 1333-40.
98. Kruglyak, K.M., E. Lin, and F.S. Ong, *Next-generation sequencing in precision oncology: challenges and opportunities*. Expert Rev Mol Diagn, 2014. **14**(6): p. 635-7.
99. Lee, H.C., et al., *Bioinformatics tools and databases for analysis of next-generation sequence data*. Brief Funct Genomics, 2012. **11**(1): p. 12-24.
100. Lengauer, C., K.W. Kinzler, and B. Vogelstein, *Genetic instabilities in human cancers*. Nature, 1998. **396**(6712): p. 643-9.
101. Stapf, M., et al., *Heterogeneous response of different tumor cell lines to methotrexate-coupled nanoparticles in presence of hyperthermia*. Int J Nanomedicine, 2016. **11**: p. 485-500.
102. L'Abbate, A., et al., *Genomic organization and evolution of double minutes/homogeneously staining regions with MYC amplification in human cancer*. Nucleic Acids Res, 2014. **42**(14): p. 9131-45.
103. Cheng, C., et al., *Whole-Genome Sequencing Reveals Diverse Models of Structural Variations in Esophageal Squamous Cell Carcinoma*. Am J Hum Genet, 2016. **98**(2): p. 256-74.
104. Carty, M., et al., *An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data*. Nat Commun, 2017. **8**: p. 15454.
105. Kawakami, H., A. Zaanan, and F.A. Sinicrope, *Microsatellite instability testing and its role in the management of colorectal cancer*. Curr Treat Options Oncol, 2015. **16**(7): p. 30.
106. Matheson, E.C., et al., *DHFR and MSH3 co-amplification in childhood acute lymphoblastic leukaemia, in vitro and in vivo*. Carcinogenesis, 2007. **28**(6): p. 1341-6.
107. Marra, G., et al., *Mismatch repair deficiency associated with overexpression of the MSH3 gene*. Proc Natl Acad Sci U S A, 1998. **95**(15): p. 8568-73.
108. Willems, R.J., et al., *Mutations in the DNA mismatch repair proteins*

MutS and MutL of oxazolidinone-resistant or -susceptible Enterococcus faecium. Antimicrob Agents Chemother, 2003. **47**(10): p. 3061-6.

109. Junop, M.S., et al., *In vitro and in vivo studies of MutS, MutL and MutH mutants: correlation of mismatch repair and DNA recombination.* DNA Repair (Amst), 2003. **2**(4): p. 387-405.

국문초록

유전체 및 전사체 분석을 활용한
항암제 (MTX) 내성 HT-29 세포주의
tandem *DHFR* 유전자 증폭
특성 및 기전 연구

서울대학교 대학원 의과학과 의과학 전공

김 아 름

차세대 시퀀싱 (next generation sequencing; NGS)으로 알려진 대량 병렬 시퀀싱 기술은 암 유전체 내의 질병의 분자 현미경 수준의 새로운 발견 및 치료법을 얻기 위해 개발되고 발전해 왔다. 현재 차세대 시퀀싱 분석을 위한 시간과 비용이 크게 줄어들었으며, 인간 진화의 기본 메커니즘에서 항암제 내성을 보이는 암 세포의 유전자 변형에 관련된 복잡한 메커니즘에 이르기까지 차세대 시퀀싱 분석의 발전을 통하여 종합적으로 분석되어왔다. 따라서 이러한 차세대 시퀀싱 분석 기술들의 조합은 분자 수준의 종양

프로파일을 규명하고 밝혀줌으로써 진단, 관리 및 치료를 위한 암 연구에 기여했으며, 암 치료 및 암 연구에서의 맞춤 의학의 미래에 중요한 역할을 할 것이다.

DHFR 유전자 증폭 현상은 항암제 메토티렉세이트(*methotrexate*; *MTX*)에 내성을 보이는 결장암 세포에 존재하며 또한 급성 림프구성 백혈병에 존재한다. **5q14** 염색체의 영역은 많은 유전자를 포함하고 있으며 대장 암 세포가 메토티렉세이트 상태에서 저항을 보일 때 유전자 증폭 현상의 근원이 되는 것으로 알려져 있으나, 실제 유전체의 변화에 대해서는 거의 알려져 있지 않았다. 이전에는 짧은 염기 서열 분석 기술을 사용해서 분석하였지만, 제공된 짧은 서열은 반복서열 영역 (*repetitive region*)을 분석 할 수 없고 접합 서열 (*junction reads*)를 식별 할 수 없기 때문에 증폭 된 영역의 전체 구조를 조합 (*assemble*) 할 명확한 방법이 없었다.

예외적으로 긴 서열을 제공하는 단일 분자 실시간 (*PacBio SMRT*) 시퀀싱은 이러한 한계를 극복하고 반복 영역의 유전체 서열의 완벽한 조립 (*assembly*) 을 가능하게 한다. 본 연구에서는 단일 분자 실시간 시퀀싱, 차세대 제한효소 광학 지도 (*next generation optical mapping*) 및 DNA 의 3 차원(3D) 구성을 측정하는 분석법 (*high throughput chromosome conformation capture; Hi-C*)과 같은 새로운 유전자 분석 기술을 사용하여 메토티렉세이트에 내성을

보이는 결장암 세포주(HT-29)내의 유전체 복제 과정을 파악하였고, 크고 복잡한 DNA 단편을 갖는 반복 서열의 구조적 변이(structural variations)를 검출하는 통합적인 프레임워크를 제안하였다.

단일 분자 실시간 시퀀싱과 광학 지도를 활용하여, 유전체 반복서열을 완벽하게 조립하고자 하였고, 5 번 염색체의 *DHFR* 유전자에서 *ATP6AP1L* 유전자까지 2.2Mbp 에 이르는 11 개의 유전자가 복제 단위이자 그 유전자들이 그 일렬 순서대로 대조군에 비해 20 배 정도 길게 복제됨을 확인하였다. 또한, 유전자 발현량 및 RNA 유전자 접합 패턴(splicing pattern)을 대조군과 비교 분석한 결과, 유전체 복제 단위에서 작게는 5 배에서 크게는 122 배까지 비정상적인 유전자 발현량이 측정되었으며, 복잡한 RNA 접합 패턴이 동반되는 것을 확인하였다.

또한, 염색체 구조를 파악하는 DNA 의 3 차원(3D) 구성을 측정된 분석 결과를 토대로, 염색체 내의 유전자가 얼마만큼 상호 작용을 하는가 확인하였을 때, 대조군에 비하여 몇몇의 위상 학적 연관 도메인 (topologically associating domains; TADs)이 매토티렉세이트에 내성을 지닌 결장암 세포주(HT-29)의 유전자가 증폭된 영역의 중앙 및 중단점에서 새롭게 발견되었으며, 이 부분에서는 조정된 상호 작용 정도 값이 높고, 그 값이 통계학적으로 유의함($p < 0.05$)을

확인하였다. 더불어, 발견하기 힘든 이중극미염색체(double minute)가 발견되었다.

흥미롭게도, *MSH* 와 *MLH* 유전자의 틀이동 삽입 돌연변이(frameshift insertion)가 메토티렉세이트 (methotrexate) 조건 하에서 염기 쌍의 잘못 짝지음을 수복하는 분자기전(mismatch repair pathway)의 유전적 불안정성과 조절 장애를 일으켰으며, *DHFR* 유전자 위치에서 역위되어 중복된 경우(inverted duplication)으로 인해 5 번 염색체 상의 *DHFR* 유전자 위치에서 염색체 절단(chromosome breakage)이 발생하였고, 다양한 크기의 유전자가 증폭된 균질염색부위(homogeneously staining region; HSR)가 절단융합가교환(breakage-fusion-bridge cycle; BFB cycle)로 생산됨을 유추할 수 있었다.

종합적으로, 본 연구는 5 번 염색체 내에서의 보다 복잡한 염색체 상호 작용 및 복제 단위 내의 역위는 유전체 재배열 (genomic rearrangement) 의 기전을 확인하는 중요한 요소가 될 수 있으며, 이러한 발견은 유전자 증폭 과정의 기초가 되는 메커니즘뿐만 아니라 암세포의 항암제 내성 원리에 대한 새로운 통찰력을 제공할 수 있을 것이라 판단하였다. 따라서 차세대 염기 분석법과 다양한 새로운 첨단 기술을 결합한 분석법은 암 유전체의 해석을 위한 강력한 도구이며, 암 치료의 핵심적인 치료 메커니즘을

파악하여 항암제의 새로운 목표를 설정할 수 있다는 점에서 정밀의학의 발전에 큰 영향을 미칠 것으로 기대한다.

주요어: 유전자 증폭; 구조적 변이; 매토타렉세이트; 차세대 염기서열 분석; 항암제 내성; *DHFR* 유전자

학번: 2015-22027