creative commons

COMMONS DEED

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

dCollection

의학박사 학위논문

# Multi-omics Analysis of Cancers with Epithelial Origin Reveals Multi-faceted Dysregulation of Genes Associated with Tumorigenesis

# 상피성 세포암의 다중오믹스 분석을 통한 종양 발달 관련 유전자의 다층적 조절 장애에 관한 연구

2022 년 7 월

서울대학교 대학원

의과학과 의과학 전공

손 민 환

# 상피성 세포암의 다중오믹스 분석을 통한 종양 발달 관련 유전자의 다층적 조절 장애에 관한 연구

지도교수 김 종 일

이 논문을 의학박사 학위논문으로 제출함
2022년 4월

서울대학교 대학원

의과학과 의과학전공

손 민 환

손민환의 의학박사 학위논문을 인준함
2022년 7월

위 원 장 ____________________ (인)

부위원장 ____________________ (인)

위 원 ____________________ (인)

위 원 ____________________ (인)

위 원 ____________________ (인)

# Multi–omics Analysis of Cancers with Epithelial Origin Reveals Multi–faceted Dysregulation of Genes Associated with Tumorigenesis

by

Min–Hwan Sohn

A thesis submitted to the Department of Biomedical Sciences
in partial fulfilment of the requirement of the
Degree of Doctor of Philosophy in Biomedical Science
at Seoul National University College of Medicine

July 2022

Approved by Thesis Committee:

Professor ____________________ Chairman

Professor ____________________ Vice chairman

Professor ____________________

Professor ____________________

Professor ____________________

# ABSTRACT

## Multi-omics Analysis of Cancers with Epithelial Origin Reveals Multi-faceted Dysregulation of Genes Associated with Tumorigenesis

Min-Hwan Sohn
Major in Biomedical Science
Department of Biomedical Science
Seoul National University Graduate School

A high-throughput sequencing technology, so called next-generation sequencing (NGS) has enabled the simultaneous interrogation of thousands and even millions of molecular targets consituting numerous biological processes and progression of a variety of diseases in multi-omics fashion. Particularly by its exceptional capability in clinical application, NGS has made it possible at last to resolve the multi-layered hallmarks of cancer, which have been indicated universally in different types of the diseases.

Here, using NGS technology, we investigated the distinct molecular characteristics and multi-faceted dysfunction of gene expression program from two types of cancers with epithelial origin: High-grade serous ovarian cancer (HGSOC) and stomach adenocarcinoma (STAD).

In the first part of the thesis, we characterized molecular profiles of HGSOC through comprehensive analysis of multi-layered data made up of whole-exome sequencing (WES) and RNA sequencing (RNA-seq). Investigation of genomic and transcriptomic landscapes of the HGSOC demonstrated that genomic scars and epithelial-to-msenchymal transition (EMT) play an important role in our cancer cohorts and that they can be divided into two distinct molecular subtypes: homologous recombination repair (HRR)-activated type and mesenchymal type. Patients with activated EMT transcriptional program showing low genomic alteration and diverse cell type properties, exhibited poor prognosis compared to HRR-activated type HGSOC did. Further validation of our findings using the cancer genome atlas (TCGA) HGSOC data verified significant worse overall survival of patients with high EMT transcriptional profiles.

In the second part of the thesis, we utilized integrative, high-dimensional multi-omics approaches to outline the DNA methylome landscape and to describe the putative oncogenic drivers of STAD using whole-genome bisulfite sequencing (WGBS) and RNA-seq. We discovered that almost 95% of cytosine-phosphate-guanine (CpG) sites were hypomethylated in STAD, while remaining hypermethylated CpGs were enriched in promoters, super enhancers and polycomb repressive complex (PRC) binding sites. Altered methylation in these elements were associated with cancer-specific gene dysregulation. Speicifically, as a putative STAD oncogenic driver, hypermethylation-mediated stimulation of canonical WNT/β-catenine/MMP signaling is discovered. Moreover, we could identify the relationship of downregulation of super enhancer related genes and re-activation of homeobox cluster genes with DNA hypermethylation. Thus, beyond classical genomic and trasncriptomic ablation-driven STAD formation, we demonstrated STAD tumorigenesis owing to epigenetic dysregulation through multi-factorial mechanisms.

Considering the high incidence rate and mortality of epithelial-origin cancers, it is imperative to elucidate the complex lanscape of their molecular disruption that bring about tumor formation. These studies provide a comprehensive multi-omics-centered analysis and a resource for novel diagnostic and therapeutic targets to epithelial-origin cancers.

* The first part of this thesis was published in Genes [1].

---

**Keywords:** Epithelial neoplasm; High-grade serous ovarian carcinoma; Whole-exome sequencing; Transcriptome sequencing; Epithelial-to-Mesenchymal Transition; Stomach adenocarcinoma; Whole-genome bisulfite sequencing; Epigenome; DNA methylation

**Student number:** 2014-25063

# CONTENTS

# LIST OF TABLES

## Part 1

## Part 2

# LIST OF FIGURES

## Part 1

## Part 2

# LIST OF ABBREVIATIONS

NGS: Next-generation sequencing

WES: Whole-exome sequencing

RNA-seq: RNA sequencing

WGBS: Whole-genome bisulfite sequencing

GRCh37: Genome reference consortium human genome build 37

GRCh38: Genome reference consortium human genome build 38

HGSOC: High-grade serous ovarian carcinoma

PARP: Poly (adenosine diphosphate-ribose) polymerase

STAD: Stomach adenocarcinoma

TF: Transcription factor

EMT: Epithelial-to-mesenchymal transition

HRR: Homologous recombination repair

BRCA1: Breast cancer type 1

BRCA2: Breast cancer type 2

TFEA: Transcription factor enrichment analysis

DMR: Differentially methylated region

PMD: Partially methylated domain

SNV: Single nucleotide variation

Indel: Insertion and deletion

SCNA: Somatic copy number alteration

TMB: Tumor mutational burden

TPM: Transcript per million

VST: Variance stabilized transformation

CAF: Cancer-associated fibroblast

PFS: Progression-free survival

OS: Overall survival

TCGA: The cancer genome atlas

ACRG: Asian cancer research group

TEMTIA: The EMT international association

ENCODE: The encyclopedia of DNA elements

GENCODE: The encyclopedia of genes and gene variants

WHO: World health organization

CIN: Chromosome instability

EBV: Ebstein-Barr virus

GS: Genomically stable

MSI: Microsatellite instability

MSS: Microsatellite stability

H.pylori: Helicobacter pylori

ANOVA: Analysis of variance

cCRE: Candidate cis-regulatory element

PLS: Promoter-like Signature

TSS: Transcription start site

TES: Transcription end site

CGI: 5’-cytosine-phosphate-guanine-3’ island

CIMP: CpG island methylator phenotype

H3K4me3: Trimethylation of lysine 4 on histone H3 protein subunit

H3K4me1: Monomethylation of lysine 4 on histone H3 protein subunit

H3K27ac: Acetylation of lysine 27 on histone H3 protein subunit

H3K27me3: Trimethylation of lysine 27 on histone H3 protein subunit

H3K36me3: Trimethylation of lysine 36 on histone H3 protein subunit

HC: Hierarchical clustering

PCA: Principal component analysis

# General Introduction

## The era of routine clinical application of next-generation sequencing

A high-throughput and massively parallel sequencing technology, so called next-generation sequencing (NGS) [2, 3] has enabled the simultaneous interrogation of many targets on the range of hundreds of thousands and even remarkably millions of targets since it is invented. These advancements have permitted read lengths as long as some complete genomes [2], lowered the cost of sequencing a human genome to almost under US $1,000, and allowed sequencing to be used as a clinical tool. Several limitations for this new technology exist, such as slightly higher error rates and generally shorter read lengths than those of conventional sanger sequencing platforms [2] however, in fact, NGS has been utilized or is being developed in routine clinical setting for genetic screening, diagnostics, and clinical evaluation [4] especially for cancer specimen.

## Multi-omics analysis as a powerful tool to investigate the hallmarks of cancer

Hallmarks of cancer [5] were presented as a set of functional capabilities acquired by human cells as they progress from normality to neoplastic development states, with a focus on characteristics critical for the formation of malignant tumors [6]. Originally, the hallmarks of cancers were initially divided into four subcategories [5] and as our understanding of knowledge of cancer mechanisms has immensely progressed, extended into eight hallmarks and two enabling features [6]: (1) sustaining proliferative signaling (2) evading growth suppressors (3) avoiding immune destruction (4) enabling replicative immortality (5) tumor-promoting inflammation (6) activating invasion and metastasis (7) inducing or accessing vasculature (8) genome instability and mutation (9) resisting cell death (10) deregulating cellular metabolism. Furthermore, following continued efforts to incorporating additional features of prospective brand-new hallmarks and enabling characteristics of cancer, ″polymorphic microbiomes″, ″senescent cells″, ″unlocking phenotypic plasticity″ and ″nonmutational epigenetic reprogramming″ are proposed to be capable of positioning in the parameters of hallmarks [6]. Herein, we mainly addressed the latter two features.

First, the repression of genes associated with the previous cell type, as well as the activation of genes associated with the new cell type, are both the examples of phenotypic and cellular plasticity [6]. While undergoing de-differentiation, blocked differentiation or transdifferentiation, cells may inhabit out-of-cells-of-origin identity states. Such modifications may be reversible and implicated in various cancer types [7]. Secondly, the global erasure and remodeling of epigenetic markers throughout the normal development is referred to as epigenetic reprogramming [8] and it is also studied thoroughly around cancer studies. For these two newly suggested hallmarks and enabling features of cancers, it is cumbersome to detect them with a single univariate marker, because they exhibit fairly multi-layered features [6]. In this regard, elucidation of cancers in multi-omics fashion cannot be too emphasized.

## Carcinoma – a malignant neoplasm of cells with epithelial origin

Carcinoma is a cancer that begins in epithelial tissue cells throughout the body that make up the external surface and lining of cavities of internal organs and form glandular tissues, such as the ovary or the stomach [9, 10]. Carcinoma develops when mutations and other alterations in the DNA, histones, and other biological substances that make up the cell's genome accumulate in a single progenitor cell [9]. The structure of the cell's biochemical components, the biochemical events that occur within the cell, and the cell's biological relationships with other cells are all controlled by the genomic, transcriptomic and epigenomic states of the cells and various environmental factors. Certain mutations in a progenitor cell (also known as a cancer stem cell [11]) eventually cause that to exhibit a number of aberrant, malignant cellular features that, when combined, are considered indicative of carcinoma. There are various different subtypes of carcinoma [12], notably including adenocarcinoma, basal cell carcinoma, squamous cell carcinoma and adenosquamous carcinoma, and anaplastic carcinoma each by the cell of origin from which they arise respectively. Given the fact that the rapid surge in burden of cancer incidence and mortality rate is reflected globally [13] and that as many as 90% of all human cancers are accounted for carcinomas arising from epithelial tissue cells [10], one should note that it is crucial to delineate their complex network of organization and function.

## Objective of the studies

Cancers originated form epithelial cell have highly heterogeneous properties and have multifaceted features, which act as obstacles for clinicians to treat them. With the advent of the $1,000 genome, there is no doubt that various multilayer analyses will be feasible and suitable strategies to tackle the heterogeneous daunting nature of epithelial-cell-of-origin cancers. Therefore, two types of epithelial cell-derived cancers were thoroughly investigated in terms of aforementioned emerging hallmarks of cancer by means of multi-omics, first by understanding ovarian cancers with homologous recombination repair (HRR) and epithelial-to-mesenchymal transition (EMT) by leveraging genomic and transcriptomic information, and second by analyzing stomach adenocarcinoma (STAD) through epigenomic and transcriptomic analysis from whole-genome bisulfite sequencing (WGBS) and RNA sequencing.

# Part 1

# Classification of High-Grade Serous Ovarian Carcinoma by Epithelial-to-Mesenchymal Transition Signature

# Abstract

High-grade serous ovarian cancer (HGSOC) is one of the deadliest cancers that can occur in women. This study aimed to investigate the molecular characteristics of HGSOC through integrative analysis of multi-omics data. We used fresh-frozen, chemotherapy-naive primary ovarian cancer tissues and matched blood samples of HGSOC patients and conducted next-generation whole-exome sequencing (WES) and RNA sequencing (RNA-seq). Genomic and transcriptomic profiles were comprehensively compared between patients with germline *BRCA1/2* mutations and others with wild-type *BRCA1/2*. HGSOC samples initially divided into two groups by the presence of germline *BRCA1/2* mutations showed mutually exclusive somatic mutation patterns, yet the implementation of high-dimensional analysis of RNA-seq and application of epithelial-to-mesenchymal (EMT) index onto the HGSOC samples revealed that they can be divided into two subtypes; homologous recombination repair (HRR)-activated type and mesenchymal type. Patients with mesenchymal HGSOC, characterized by the activation of the EMT transcriptional program, low genomic alteration, and diverse cell-type compositions, exhibited significantly worse overall survival than those with HRR-activated HGSOC did ($p = 0.002$). In validation with the cancer genome atlas (TCGA) HGSOC data, patients with a high EMT index (≥the median) showed significantly worse overall survival than did those with a low EMT index (<the median) ($p = 0.030$). In conclusion, through a comprehensive multi-omics approach towards our HGSOC cohorts, two distinctive types of HGSOC (HRR-activated and mesenchymal) were identified. Our novel EMT index could be a potential prognostic biomarker for HGSOC.

* This work was published in Genes [1].

---

**Keywords:** Epithelial neoplasm; ovarian cancer; High-grade serous ovarian carcinoma; Next-generation sequencing; epithelial-to-mesenchymal transition; homologous recombination repair

**Student number:** 2014-25063

# Introduction

Ovarian cancer, one of the deadliest gynecologic malignancies, is a global burden with an estimated 313,959 new cases and 207,252 cancer deaths in 2020 alone [14]. The majority of ovarian cancers are epithelial ovarian cancers, and high-grade serous ovarian carcinoma (HGSOC) is the most prevalent histologic type [15]. In patients with HGSOC, germline or somatic mutations in *BRCA1* or *BRCA2* gene are frequently observed, and women harboring germline *BRCA1/2* mutations are at high risk of developing HGSOC [16].

The patients' *BRAC1/2* mutational status is of high interest because several poly (adenosine diphosphate-ribose) polymerase (PARP) inhibitors are currently available for the treatment of primary and recurrent HGSOC, based on the phase 3 clinical trials, which have demonstrated the significant survival benefit brought by PARP inhibitors [17-21]. However, beyond focusing on *BRCA1/2* gene mutations, there is an urgent need to discover other genetic mutations and altered gene expression programs that might potentially be prognostic biomarkers or therapeutic targets.

One important feature of HGSOCs is that they are commonly diagnosed at an advanced stage, therefore showing high disease recurrence and mortality rates despite the primary treatment [22]. Researchers have noted epithelial-to-mesenchymal transition (EMT), a process referring to the conversion of an epithelial cell to a mesenchymal cell, as the mechanism for invasion and metastasis of ovarian cancer cells [23], as well as for achieving chemoresistance [24]. Interestingly, in breast cancer, loss of BRCA1 protein is associated with EMT [25]. However, such a relationship has been poorly investigated in ovarian cancer. Broadening the molecular understanding of HGSOC and elucidating the underlying mechanisms for EMT as well as *BRCA1/2* gene alterations is expected to open a new horizon in the treatment of HGSOC [26].

In this regard, we carried out next-generation whole-exome sequencing (WES) and RNA sequencing (RNA-seq) to find the causal variants that bring about HGSOC in terms of homologous recombination repair (HRR) and EMT.

# Materials and methods

## Study Population

Inclusion criteria for the study population were as follows: (1) diagnosed with HGSOC between January 2013 and December 2016; (2) having undergone primary debulking surgery; (3) having donated their blood samples, obtained one day before surgery, and fresh-frozen primary ovarian cancer tissues, obtained at the time of surgery, for scientific purposes after providing written informed consent; and (4) having an identifiable germline *BRCA1/2* mutational status. In addition, patients were excluded if (1) they had any malignancy other than HGSOC; (2) received neoadjuvant chemotherapy; or (3) had insufficient clinical data or were lost to follow-up.

Among patients who met these criteria, we further selected patients referring to their germline *BRCA1/2* genetic test results as follows: (1) five patients harboring germline deleterious *BRCA1* mutations and wild-type *BRCA2* (g*BRCA1*mut); (2) five patients harboring germline deleterious *BRCA2* mutations and wild-type *BRCA1* (g*BRCA2*mut); and (3) 10 patients with wild-type *BRCA1/2* genes (g*BRCA1/2*wt). Details of the germline *BRCA1/2* gene testing methods at our institution were described in a previous study [27].

We collected the patients' baseline clinicopathologic characteristics, such as age at diagnosis, International Federation of Gynecology and Obstetrics (FIGO) stage, initial serum CA-125 levels, and residual tumor size after surgery. In terms of survival outcomes, progression-free survival (PFS) was defined as the time interval between the date of diagnosis to the date of disease progression, while overall survival (OS) was defined as the time interval between the date of diagnosis to the date of cancer-related death or last visit.

**Whole-exome library preparation, sequencing, and data analysis**

The fresh-frozen, primary ovarian cancer tissues and blood samples of 20 patients were retrieved from Seoul National University Hospital Human Biobank. One expert gynecologic pathologist (Cheol Lee) in Seoul National University Hospital reviewed and confirmed all the HGSOC cases in our study population according to the World Health Organization Classification of Tumors, 5th edition.

We obtained DNA from primary ovarian cancer tissue samples and matched normal blood samples using Puregene Core kit and QIAamp DNA Blood Mini kit, respectively. Then, we captured human exon regions using SureSelect Human All Exon V6 kit following standard protocols. Subsequently, we performed 101X2 paired-end WES using HiSeq2500 instrument (Illumina, San Diego, CA, USA), according to the manufacturer's instruction. Raw

FASTQ files were aligned onto GRCh37 using Burrow-wheeler Aligner (BWA) mem algorithms [28], and the resulting bam files were subjected to duplicate removal using the Genome Analysis Tool Kit (GATK) version 4.1.4.1 [29]. After base quality score recalibration and applying it to each bam file using GATK, we proceeded to the discovery of somatic single nucleotide variants (SNVs) and small insertions and deletions (indels) for each tumor sample using Strelka2 [30] and a paired-normal sample as a control. For germline variant discovery of the WES from normal samples, we applied GATK's HaplotypeCaller. All of the above variants were annotated by Oncotator [31] and ANNOVAR [32]. To accurately pinpoint the actually harmful ones, we only retained exonic variants (i.e. Missense, Nonsense, Frameshift insertion, Frameshift deletion, In-frame insertion, In-frame deletion and Splice site mutation) with at least 10x coverage of alternate allele, predicted to be deleterious by SIFT [33] and having minor allele frequency below 0.1% in 1000 Genomes Project phase 3 data [34], EXome Aggregation Consortium data [35], and Northeast Asian Reference Database [36]. Then, we only kept genes that are overlapped with cancer consensus genes from the Catalogue of Somatic Mutations in Cancer database [37]. Tumor mutational burden (TMB) was estimated by the total somatic mutations for each sample divided by the length of the captured exon regions (61 Mb).

### Copy number alteration detection from Whole-exome sequencing data

In order to discover somatic copy number alterations (SCNAs), we used CNVkit with default parameters [38]. Specifically, bin-level log2 ratio (.cnr) and segmented log2 ratio (.cns) files, generated from bam files by a separate reference for each matched tumor-normal pair, were processed into the residual bin-level log2 ratio estimates (segmetrics command). Then, GISTIC2 [39] was implemented to identify frequently altered chromosomal regions with a confidence level of 0.90 and a Q-value threshold of 0.05. The purity of the tumor samples was calculated via Sequenza algorithm [40]. We also used seqz files generated by Sequenza as an input to scarHRD [41] for HRD score estimation. To discover germline SCNAs by using the CNVkit, we constructed a pooled reference from 20 normal blood samples and followed the same approach as that used for detecting somatic SCNAs using the CNVkit.

**RNA-seq library preparation, sequencing, and general analysis**

We extracted RNA from primary ovarian cancer tissue samples and prepared sequencing library using TruSeq RNA Access Library Prep Kit under standard protocol. Then we conducted RNA-seq on 20 HGSOC tumor samples by 101X2 paired-end mode using Illumina HiSeq2500 (Illumina, San Diego, California), in accordance with the manufacturer's instruction. For RNA-seq data analysis, each transcript expression was first quantified by pseudoalignment algorithm implicated in kallisto [42] version 0.46.1 using

RefSeq annotation release 105 for GRCh37. Quantified transcript-level transcripts per million (TPM) values were collapsed to give gene-level expression, and only the protein-coding genes were processed for the rest of the analysis. TPM values were implemented for comparison among different groups and for inputs to cell type enrichment analysis. With regard to discovering differentially expressed genes (DEG) among sample groups ((i) g*BRCA1*mut, g*BRCA2*mut and g*BRCA1/2*wt and (ii) homologous recombination repair (HRR)-activated and mesenchymal), we used DESeq2 [43] version 1.24.0. Raw counts of the RNA-seq were transformed using variance stabilizing transformations (vst), which were later used as inputs to principal component analysis (PCA), unsupervised hierarchical clustering (HC) and identification of gene co-expression modules and interaction networks [44]. With respect to PCA, the top 5,000 variable genes among 19,023 genes were used as inputs. Samples were then grouped into two clusters according to K-means clustering with k=2. For HC, we used euclidean distance measure and uncentered correlation measure for epithelial-to-mesenchymal transition transcription factors (EMT-TFs) and HRR genes, respectively. Each gene expression was centered to the average intensity of samples, along with pair-wise complete-linkage for clustering of both samples and genes using Cluster 3.0 [45]. We visualized the resulting distance measures and dendrograms through Java Treeview [46].

**Transcription Factor Enrichment Analysis**

Adding to the DEG analysis, PCA, K-means clustering, and unsupervised hierarchical clustering, we performed transcription factor enrichment analysis (TFEA) for a particular set of genes by using ChIP-X Enrichment Analysis version 3 [47]. Particularly, we used a complete list of transcription factors and their target gene-set libraries from ARCHS4 [48], which is a compendium of publicly available, processed RNA-seq data (https://maayanlab.cloud/chea3/assets/tflibs/ARCHS4_Coexpression.gmt, accessed on 14 April 2021). We only used the top 10 enriched TFs with false discovery rate <0.05 for subsequent analyses.

### Calculation of EMT Index

To analyze RNA-seq data in relation to EMT, we manually coined an index, the "EMT index". Specifically, the EMT index was calculated for each sample based on the geometric mean of TPM values for five core EMT-TFs (*TWIST1*, *SNAI1*, *SNAI2*, *ZEB1*, and *ZEB2*) and 33 EMT-related TFs (*KLF4, GSC, TCF7L2, ALX1, GATA6, RUNX2, TCF3, SOX4, FOXC2, NFKB1, KLF2, KLF6, TBX3, TCF4, PRRX1, HOXB7, JUN, FOS, TAZ, TGIF1, ATF1, ERG, ETS1, ID1, TEAD1, YAP1, NFYA, KLF8, SOX9, SIX1, TBXT, GATA4,* and *TWIST2*) according to the consensus statement on EMT led by the EMT International Association (TEMTIA) [49].

## Deriving other EMT intensity measures

EMT index from Cristescu et al. [50] which derive its EMT index from Loboda et al. [51], was calculated for each sample based on the geometric mean of TPM values for total of 149 EMT-related genes, similar to our calculation of EMT index. EMT score from Guo et al. [52] was calculated in terms of the expression signature of 76 EMT-related genes as follow:

EMT score of sample n = $\sum_{m=1}^{76} WmGm, n$ where, Wm represents the pearson correlation coefficient between the expression of m$^{th}$ gene and that of *CDH1*, while Gm,n represents the expression of m$^{th}$ gene of sample n.

## Identification of Co-Expressed Gene Modules and Interaction Networks

To identify gene co-expression modules and interaction networks from RNA-seq data, we used CEMiTool [44] version 1.14.0. In total, 19,023 genes, upon which was applied variance-stabilizing transformation implemented in DESeq2 [43], were used as inputs and samples were divided into two pre-annotated clusters by K-means clustering, namely, cluster A and cluster B, with the following settings: corr_method = “spearman”, network type = “signed”, tom_type = “signed”, rank_method = “mean”, gsea_max_size = 2000. Calculated modules were considered significant only if the absolute value of normalized enrichment scores (NES) for both cluster A and cluster B was above 4 and with a Benjamini–Hochberg adjusted p value < 0.0001. For the input-constructing interaction network of each co-expressed gene module, we retrieved TFs target gene-set libraries from ARCHS4 [48] as a Gene Matrix Transposed (gmt) file format with a minor modification, putting TF genes and their target genes in the first column and the second column, respectively (https://github.com/ryansohny/HGSOC/blob/main/RNA-seq/ARCHS4_Coexpression_interaction.csv). Then, we performed overrepresentation analysis implemented in CEMiTool using HALLMARK gene sets from the Molecular Signature Database (MSigDB) [53].

### Cell-Type Enrichment Analysis

To further validate our findings regarding classification of our samples into two groups based on their genomic and transcriptomic profiles, we performed cell-type enrichment analysis from gene expression data. An expression profile of samples was uploaded to XCell [54] web interface with default parameters using “xCell (N = 64)” gene signature.

### Analysis of TCGA Data

We downloaded The Cancer Genome Atlas (TCGA) RNA-seq data of 376 HGSOC samples and corresponding clinicopathological profiles from the National Cancer Institute Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/, accessed on 22 February 2018) and cBioPortal for Cancer Genomics (https://www.cbioportal.org, accessed on 22 February 2018) website. TPM values were calculated by dividing each gene's fragments per kilobase per million (FPKM) value with the sum of FPKM of that particular sample. To divide the TCGA cohort in terms of EMT index, the median value of the EMT indices of all samples was used; samples having a higher EMT index than the median value (11.999) were classified as EMT-high, while the remainders were classified as EMT-low. Similar approach was applied to sample classification based on EMT index from Cristescu et al. [50]. For EMT score-based classification of samples, samples with negative EMT score were classified as "Mesenchymal" and with positive score as "Epithelial".

**Statistical Analysis**

Differences in baseline characteristics and genomic or transcriptomic profiles between two groups (g*BRCA1*mut and g*BRCA1/2*wt) or among three (g*BRCA1*mut, g*BRCA2*mut, and g*BRCA1/2*wt) were assessed: Pearson's chi-square or Fisher's exact tests were used for categorical variables, while Student's t-, Mann–Whitney U, analysis of variance (ANOVA), or Kruskal–Wallis tests were used for continuous variables. Tukey's HSD was used for multiple comparisons. Pearson correlation coefficients were calculated between patient characteristics and somatically mutated genes. Survival outcomes were compared using Kaplan-Meier analysis with log-rank test. R statistical software version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria) was used for the statistical analyses. P values < 0.05 were considered statistically significant unless otherwise noted.

**Code Availability**

The codes to reproduce our results and algorithms implemented in this study are available in Github repository at https://github.com/ryansohny/HGSOC.

# Results

## Characteristics and Survival Outcomes of Patients with HGSOC

Between the g*BRCA1/2*mut and g*BRCA1/2*wt groups, no differences were observed in baseline clinicopathologic characteristics (**Table 1-1**). None of the study population received PARP inhibitors at their primary treatment, whereas three patients in the g*BRCA1/2*mut group received PARP inhibitor maintenance therapy to treat relapsed disease. A median observation period was 63.4 months. The two groups showed a similar PFS (median, 26.0 vs. 24.6 months; p = 0.895) and OS (mean, 76.8 vs. 71.6 months; p = 0.519; **Figure 1-1**).

## Genomic Profiling of HGSOC

WES of 20 blood samples revealed the same germline *BRCA1/2* mutations as those identified by our in-house gene testing (**Figure 1-2**). In detail, samples from the g*BRCA1*mut group had a frameshift insertion (g*BRCA1*mut_1), a frameshift deletion (g*BRCA1*mut_3, g*BRCA1*mut_4), and a stop-gain SNV (g*BRCA1*mut_2) in the *BRCA1* gene, which were all heterozygous, and a hemizygous deletion of exon 1 through 14 of the *BRCA1* gene (g*BRCA1*mut_5). All samples from the g*BRCA2*mut group had the frameshift deletion of a single *BRCA2* gene in five different sites (g*BRCA2*mut_1 through g*BRCA2*mut_5). Next, we investigated somatic mutations and putative drivers of HGSOC progression from tumor–normal pairs (**Figure 1-3**). Interestingly, we observed a mutually exclusive variants pattern with few co-occurring somatic single nucleotide variants (SNVs) and indels across our samples, except for the *TP53* mutation (pairwise Fisher's exact test $p > 0.05$). The lack of *TP53* somatic mutations in some of our samples, which is rare in HGSOC, might originate from their low tumor purity. In particular, two g*BRCA1/2*wt samples lacked any apparent driver mutations of SNVs or indels. Tumor mutational burden (TMB) was assessed for each sample, but no significant difference was detected among the g*BRCA1*mut, g*BRCA2*mut, and g*BRCA1/2*wt groups (one-way ANOVA test $p = 0.313$) (**Figure 1-4**). In terms of somatic copy number alterations (SCNAs), we observed amplification of genes, such as *CSF3R*, *LCK*, *MPL*, *MUTYH*, *SFPQ*, *STIL*, and *TAL1*, and loss of genes, such as *GNA11*, *MLLT1*, *MAP2K2*, and *SH3GL1* (**Figure 1-5**).

## Transcriptomic Profiling of HGSOC in terms of HRR and EMT

Based on the RNA-seq data from 20 HGSOC samples, we conducted PCA to cluster the samples on the basis of the top 5000 variable genes out of 19,023 genes and observed highly similar transcriptomic profiles between the g*BRCA1*mut and g*BRCA2*mut groups (**Figure 1-6**). Six out of 10 samples in the g*BRCA1/2*wt group were clustered into "cluster A" together with the g*BRCA1*mut and g*BRCA2*mut groups, with the exception of one g*BRCA2*mut sample. Meanwhile, the remaining four samples in the g*BRCA1/2*wt group and the g*BRCA2*mut sample were segregated into "cluster B" (**Figure 1-6**). To determine the causal or regulatory variants for clusters A and B, we first performed TFEA [47] for genes exhibiting a negative correlation ($r < -0.9$, $n = 60$) with the first principal component (PC1) and that were upregulated in cluster A rather than in cluster B. The most significantly enriched TF gene was *GRHL2* (**Table 1-2**), known as an EMT suppressor in various cancers [55-57].

Next, considering that cluster A included most samples of the g*BRCA1/2*mut group, we investigated transcriptomic aberration of the HRR genes (**Table 1-3**) [21]. Unsupervised hierarchical clustering of 30 HRR genes recapitulated the PCA result, and 18 out of 30 HRR genes (e.g., *ATR*, *FANCA*, and *FANCD2*) were significantly upregulated in cluster A rather than in cluster B (**Figure 1-7**). The activation of HRR pathways might be explained by a genetic compensation for the dysfunction of *BRCA1* or *BRCA2* in the g*BRCA1/2*mut group, which accounts for a large part of cluster A. Furthermore, six samples from the g*BRCA1/2*wt group that fell into cluster A had several somatic alterations in HRR genes: missense mutations in *BRCA1*, *ATRX*, and *ATR*, copy number loss of *BRCA2*, *FANCC*, *FANCG*, and *RAD50*, and copy number gain of *RAD51B* and *RAD54L* (**Figure 1-8**). Then, in order to find specific TFs regulating the expression of HRR genes, we again conducted TFEA for the 18 upregulated HRR genes and discovered that *E2F8*, *E2F2*, *E2F3*, *PRDM9*, *CENPA*, and *TGIF* were the core regulators or components of the gene networks overexpressed in cluster A (**Table 1-4**).

Focusing on genes upregulated in cluster B compared to their expression in cluster A, we also performed TFEA for genes exhibiting a positive correlation ($r > 0.9$, $n = 180$) with PC1. Interestingly, among the enriched TFs (**Table 1-5**), *TCF21, TWIST2, MEOX2, OSR1, PRRX1, PRRX2*, and *TWIST1* were associated with EMT [58]. Investigation of the RNA expression of these TFs indicated that 6 out of 7 genes were upregulated (Mann-Whitney U test *P* value < 0.05) in cluster B rather than in cluster A (**Figure 1-9**).

Analyzing RNA-seq data in relation to EMT, we manually coined the method and term "EMT index" (**Table 1-6**) which is defined as a geometric mean of gene expression values (TPM) across 5 core EMT transcription factor genes and 33 EMT-related transcription factor genes from the The EMT International Association (TEMTIA) [49]. First, unsupervised hierarchical clustering of samples with these 38 TFs accurately separated 20 HGSOC tissue samples into clusters A and B (**Figure 1-10**). Between the two clusters, the EMT index was significantly higher in cluster B than in cluster A ($p = 0.001$; **Figure 1-10**).

In addition to the 38 genes used to calculate the EMT index, *CDH1* (coding E-cadherin), known to be highly expressed in epithelial tissue and downregulated in mesenchymal tissue [49], was downregulated in cluster B (**Figure 1-11**, left). In contrast, *VIM* (coding vimentin), another key indicator of EMT highly expressed in mesenchymal rather than in epithelial tissue [59], was upregulated in cluster B (**Figure 1-11**, middle). In addition, *TGFB1* (TGFβ), known as a key accelerator of EMT [60], was also upregulated in cluster B (**Figure 1-11**, right).

Interestingly, homologous recombination deficiency (HRD) score [41], a genomic scar estimate combining three measures (loss of heterozygosity, telomeric allelic imbalance, and large-scale state transitions) was higher in cluster A, compared to that of cluster B (**Figure 1-12**, left). Moreover, EMT index was found to be negatively correlated with the genomic scar estimate (**Figure 1-12**, right).

To dissect variation in the transcriptional network of our samples and further validate the transcriptional nature of two groups, cluster A and cluster B, we performed gene co-expression network analysis [44]. With this approach, we were able to identify one module (Co-expression Module 1) enriched in samples from cluster B, and two modules (Co-expression Modules 2 and 3) enriched in samples from cluster A (**Figure 1-13**). Co-expression Module 1 had EMT-TFs (e.g., *KLF2* and *PRRX1*) as interaction hub genes, consistent with the finding that EMT gene signature was enriched in cluster B. Co-expression Modules 2 and 3 were characterized by distinctive hub genes such as *SLC2A1*, which is known to be regulated by estrogens [61], and *MYBL2*, a core regulator of cellular differentiation [62], was among the main components of the complex network of gene expression in cluster A.

Meanwhile, we found a negative correlation between PC1 and tumor purity, derived from WES data ($r = -0.84$, $p < 0.001$; **Figure 1-14**), consistent with the finding that mesenchymal-type ovarian cancers tend to have lower tumor purity than do other types [63, 64]. Using the gene expression data, we also conducted cell-type enrichment analysis [54], and the mesenchymal stromal cell, the intra-tumoral cancer-associated fibroblast (CAF), and epithelial cell signature were investigated (**Figure 1-15**). Samples in cluster B were enriched in mesenchymal stromal cells and CAFs compared to samples in cluster A enriched in epithelial cells. Consistently, we also observed that two CAF marker genes, *DCN* and *PDPN*, were significantly upregulated in cluster B compared to their expression in cluster A (**Figure 1-16**). Taken together, we could classify 20 HGSOC tissue samples into two categories: (1) HRR-activated HGSOC (cluster A) and (2) mesenchymal HGSOC (cluster B).

### EMT Index and Survival Outcomes

We performed survival analysis between patients with mesenchymal HGSOC (n = 5) and those with HRR-activated HGSOC (n = 15). While the two groups showed similar PFS (Log-rank *P* value = 0.708), patients with mesenchymal HGSOC exhibited significantly worse OS than those with HRR-activated HGSOC (Log-rank *P* value = 0.002) (**Figure 1-17**).

Next, we investigated the reproducibility of our study findings using TCGA HGSOC data [65]. Processing 379 RNA-seq samples, we calculated each sample's EMT index (**Figure 1-18,** left) and examined its correlation with known EMT markers (**Figure 1-18,** right). Although the expression of *CDH1*, which was expected to be decreased with the increasing EMT index, had a weak positive correlation with the EMT index (Pearson r = 0.177, *P* < 0.001), its presence in EMT-high samples might indicate epithelial/mesenchymal intermediate states or reflect transient activation and repression of the EMT program [66, 67]. *CDH2*, encoding N-cadherin and serving as an indicator of EMT [68], was positively correlated with the EMT index (Pearson r = 0.255, *P* < 0.001), suggesting the possibly increased mesenchymal population within the EMT-high samples. *VIM* and *TGFB1* also increased in direct proportion as EMT index increased (Pearson r = 0.582, *P* < 0.001; and r = 0.591, *P* < 0.001, respectively).

Then, we analyzed the survival outcomes by the level of EMT index in TCGA HGSOC samples for which survival data were available (n = 374) (**Figure 1-**

**19**). The OS of patients whose samples had a high EMT index (≥the median, n = 187) was significantly worse than that of patients whose samples had a low EMT index (<the median, n = 187) (median, 44.0 vs. 47.4 months; Log-rank *P* = 0.030). As we checked how the EMT-high and -low groups were distributed in the four subtypes of TCGA HGSOC (**Figure 1-20**), we observed that the EMT-high samples were mostly enriched in the TCGA-defined mesenchymal HGSOC subtype (Chi-square test *P* < 0.001; Benjamini-Hochberg corrected *P* < 0.001 for all pairwise Fisher's Exact test between mesenchymal and others). Moreover, among the four subtypes of TCGA HGSOC, the mesenchymal subtype exhibited the highest level of EMT index (one-way ANOVA test *P* < 0.001; adjusted *P* < 0.05 for all Tukey's HSD).

### Comparing EMT index method with other EMT intensity measures

To quantify the degree and intensity of state transition that cells in a given sample have gone through, a number of different scoring schemes using transcriptome data have been devised [69]. Since our newly developed EMT index is one of the transcriptome-based scoring systems, we compared our EMT index method with other EMT intensity measures. Specifically, we selected the method from Guo et al. [52] termed "EMT score" which is a weighted sum of 76 EMT-related genes, and the method from Cristescu et al. [50] which uses similar approach to our EMT index in terms of using geometric mean of EMT-related gene expression, yet using their own different set of 149 genes (see Materials and methods). First, we applied these two EMT intensity measures to 20 HGSOC. As a result, we found out that both measures were capable of discriminating Cluster A (HRR-activated) and Cluster B (Mesenchymal) similar to our EMT index from 38 EMT-TF method (**Figure 1-22**). Next, extending our analysis further, we applied these two methods to TCGA HGSOC cohort (**Figure 1-23**). We observed that there was a weak positive correlation between EMT score and EMT index (PCC=0.231, *P*=5.50E−06). In light of the fact that sample with a negative EMT score is classified as "Mesenchymal" according to Guo et al. method [52], this positive correlation is quite unexpected. Conversely, our EMT index and EMT index from Cristescu et al. showed high similarity in terms of correlation (PCC=0.886, P=6.76E−128). EMT score and EMT index from Cristescu et al. showed no correlation as expected (PCC=0.100, *P*=0.053). Next, in order to evaluate the capability of each EMT intensity measure of inferring epithelial or mesenchymal marker gene expression, we examined the correlation of known EMT marker gene expression with each EMT scoring system. EMT score was positively correlated with epithelial cell marker, *CDH1* expression (PCC=0.687, P=3.61E−54), which means it was possible for EMT score to infer the intensity of *CDH1* expression of each TCGA HGSOC sample. This is largely due to intrinsic nature of EMT score derived from computing it based on *CDH1* expression (**Figure 1-24**, left, see Materials and methods). Nevertheless, it didn't capture the expression of mesenchymal marker genes such as *CDH2* (PCC=0.018, P=0.722), *VIM* (PCC=−0.021, *P*=0.722) and

*TGFB1* (PCC=0.272, P=7.74E−8), all of which should be negatively correlated because a sample with mesenchymal feature should have a negative EMT score value by the method from Guo et al. [52]. In contrast, EMT index from Cristescu et al. performed relatively well in terms of inferring the expression of known mesenchymal marker genes (**Figure 1-24**, right). Additionally, we observed that "Mesenchymal" types defined by using EMT score were not enriched in the TCGA-defined mesenchymal subtype (**Figure 1-25**, Chi-square test *P*=0.034; Benjamini-Hochberg corrected P > 0.05 for all pairwise Fisher's exact test between TCGA mesenchymal-type and others). When we analyzed the OS of patients in the "Mesenchymal" groups and of those in the "Epithelial" group, "Epithelial" group exhibit worse OS than "Mesenchymal" group (**Figure 1-25**). Given the fact that tumor with mesenchymal cell-like feature generally has worse prognosis in HGSOC compared to that with different molecular feature [70, 71], EMT score from Guo et al. failed to show the expected prognosis of the mesenchymal-type HGSOC. On the other hand, EMT index from Cristescu et al. was largely similar to our results based on EMT index in terms of recapitulating the TCGA 4 subtypes, especially the TCGA-defined mesenchymal subtype (**Figure 1-26**, Chi-square test *P* < 0.005; Benjamini-Hochberg corrected *P* < 0.05 for all pairwise Fisher's exact test between TCGA mesenchymal-type and others). Yet, it also failed to show the tendency of worse OS of samples with mesenchymal features (Log-rank P=0.093). Overall, we showed, through multiple approaches, that our EMT index had relatively better capability in examining the mesenchymal feature of the ovarian cancer sample than other methods did.

**Table 1-1. Patients′ clinicopathologic characteristics**

| Characteristics | All<br>(*n*=20, %) | *BRCA* mutation<br>(n=10, %) | *BRCA* wild-type<br>(n=10, %) | *P* |
|---|---|---|---|---|
| Age, years | | | | |
| Mean±SD | 52.8±8.4 | 54.2±9.4 | 51.4±7.4 | 0.705 |
| Family History of breast cancer | 1 (5.0) | 1 (10.0) | 0 | >0.999 |
| Family History of ovarian cancer | 1 (5.0) | 1 (10.0) | 0 | >0.999 |
| FIGO stage | | | | 0.779 |
| IIIA | 2 (10.0) | 1 (10.0) | 1 (10.0) | |
| IIIB | 1 (5.0) | 1 (10.0) | 0 | |
| IIIC | 11 (55.0) | 5 (50.0) | 6 (60.0) | |
| IV | 6 (30.0) | 3 (30.0) | 3 (30.0) | |
| CA-125, IU/ml | | | | |
| Median (range) | 798.5 (5.1-3545.0) | 798.0 (5.1-3545.0) | 798.5 (47.0-2433.0) | 0.940 |
| Lymphovascular space invasion | 16 (80.0) | 8 (80.0) | 8 (80.0) | >0.999 |
| Lymph node metastasis | 12 (60.0) | 6 (60.0) | 6 (60.0) | >0.999 |
| Residual tumor after surgery | | | | 0.139 |
| No gross | 14 (70.0) | 9 (90.0) | 5 (50.0) | |
| <1 cm | 5 (25.0) | 1 (10.0) | 4 (40.0) | |
| ≥1 and <2 cm | 1 (5.0) | 0 | 1 (10.0) | |

Abbreviations: CA-125, cancer antigen 125; FIGO, International Federation of Gynecology and Obstetrics; SD, standard deviation.

## Table 1-1. continued

| Characteristics | All<br>(*n*=20, %) | *BRCA* mutation<br>(n=10, %) | *BRCA* wild-type<br>(n=10, %) | *P* |
|---|---|---|---|---|
| Chemotherapy at primary treatment | | | | 0.628 |
| 6 cycles of paclitaxel-carboplatin | 14 (70.0) | 6 (60.0) | 8 (80.0) | |
| 9 cycles of paclitaxel-carboplatin | 6 (30.0) | 4 (40.0) | 2 (20.0) | |
| Recurrence | 16 (80.0) | 9 (90.0) | 7 (70.0) | 0.582 |
| Treatment-free interval, months | | | | |
| Median (range) | 20.4 (3.0-73.0) | 20.9 (13.5-73.0) | 19.6 (3.0-67.9) | 0.496 |
| Germline *BRCA1* mutational status | | | | 0.033 |
| Wild-type | 15 (75.0) | 5 (50.0) | 10 (100.0) | |
| Mutation | 5 (25.0) | 5 (50.0) | 0 | |
| Germline *BRCA2* mutational status | | | | 0.033 |
| Wild-type | 15 (75.0) | 5 (50.0) | 10 (100.0) | |
| Mutation | 5 (25.0) | 5 (50.0) | 0 | |

Abbreviations: CA-125, cancer antigen 125; FIGO, International Federation of Gynecology and Obstetrics; SD, standard deviation.

**Table 1-2. TFEA results from genes negatively correlated with PC1 (Pearson r < -0.9).**

| Transcription Factor | Overlapping Genes | FDR *Q* value |
|---|---:|---:|
| *GRHL2* | 28 | 4.02E-27 |
| *SPDEF* | 27 | 4.85E-26 |
| *FOXA1* | 26 | 7.58E-25 |
| *OVOL1* | 24 | 1.91E-22 |
| *ELF3* | 24 | 1.91E-22 |
| *IRF6* | 24 | 1.91E-22 |
| *EHF* | 23 | 3.50E-21 |
| *KLF5* | 22 | 5.07E-20 |
| *GATA3* | 22 | 5.07E-20 |
| *TFCP2L1* | 22 | 5.07E-20 |

**Table 1-3. List of 30 homologous recombination repair genes used in this study**

| Homologous recombination repair genes |
|:---:|
| *BRCA1* |
| *BRCA2* |
| *ATM* |
| *ATR* |
| *ATRX* |
| *BARD1* |
| *BLM* |
| *BRIP1* |
| *CHEK1* |
| *CHEK2* |
| *FANCA* |
| *FANCC* |
| *FANCD2* |
| *FANCE* |
| *FANCF* |
| *FANCG* |
| *FANCI* |
| *FANCL* |
| *FANCM* |
| *MRE11* |
| *NBN* |
| *PALB2* |
| *RAD50* |
| *RAD51* |
| *RAD51B* |
| *RAD51C* |
| *RAD51D* |
| *RAD52* |
| *RAD54L* |
| *RPA1* |

**Table 1-4. TFEA results from 18 out of 30 HRR genes upregulated in cluster A compared to those in cluster B**

| Transcription Factor | Overlapping Genes | FDR *Q* value |
|---|---:|---:|
| *E2F8* | 8 | 4.18E-06 |
| *E2F2* | 6 | 8.68E-04 |
| *ZNF227* | 4 | 4.10E-02 |
| *ZNF107* | 4 | 4.10E-02 |
| *PRDM9* | 4 | 4.10E-02 |
| *ZNF45* | 4 | 4.10E-02 |
| *E2F3* | 4 | 4.10E-02 |
| *CENPA* | 4 | 4.10E-02 |
| *ZNF689* | 4 | 4.10E-02 |
| *TGIF2* | 4 | 4.10E-02 |

**Table 1-5. TFEA results from genes negatively correlated with PC1 (Pearson r < -0.9)**

| Transcription Factor | Overlapping Genes | FDR *Q* value |
|---|---:|---:|
| *TCF21* | 20 | 1.86E-08 |
| *TWIST2* | 17 | 3.11E-06 |
| *MEOX2* | 16 | 1.09E-05 |
| *OSR2* | 16 | 1.09E-05 |
| *OSR1* | 15 | 3.41E-05 |
| *PRRX1* | 15 | 3.41E-05 |
| *PRRX2* | 15 | 3.41E-05 |
| *BCL6B* | 14 | 1.35E-04 |
| *ATOH8* | 14 | 1.35E-04 |
| *TWIST1* | 14 | 1.35E-04 |

**Table 1-6. List of 38 genes used to calculate the EMT index**

| Core EMT-TFs | Other EMT-related TFs |
|---|---|
| *SNAI1* | *KLF4* |
| *SNAI2* | *GSC* |
| *ZEB1* | *TCF7L2* |
| *ZEB2* | *ALX1* |
| *TWIST1* | *GATA6* |
| | *RUNX2* |
| | *TCF3* |
| | *SOX4* |
| | *FOXC2* |
| | *NFKB1* |
| | *KLF2* |
| | *KLF6* |
| | *TBX3* |
| | *TCF4* |
| | *PRRX1* |
| | *HOXB7* |
| | *JUN* |
| | *FOS* |
| | *TAZ* |
| | *TGIF1* |
| | *ATF1* |
| | *ERG* |
| | *ETS1* |
| | *ID1* |
| | *TEAD1* |
| | *YAP1* |
| | *NFYA* |
| | *KLF8* |
| | *SOX9* |
| | *SIX1* |
| | *TBXT* |
| | *GATA4* |
| | *TWIST2* |

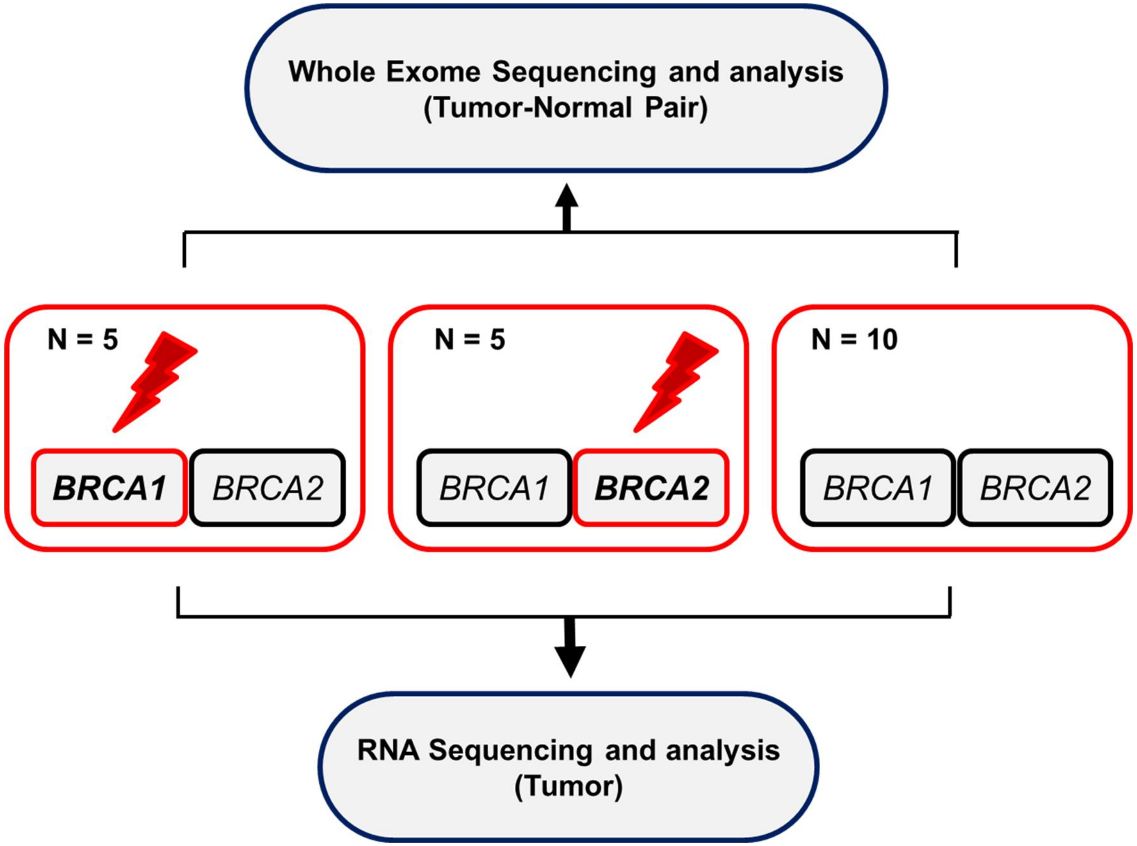
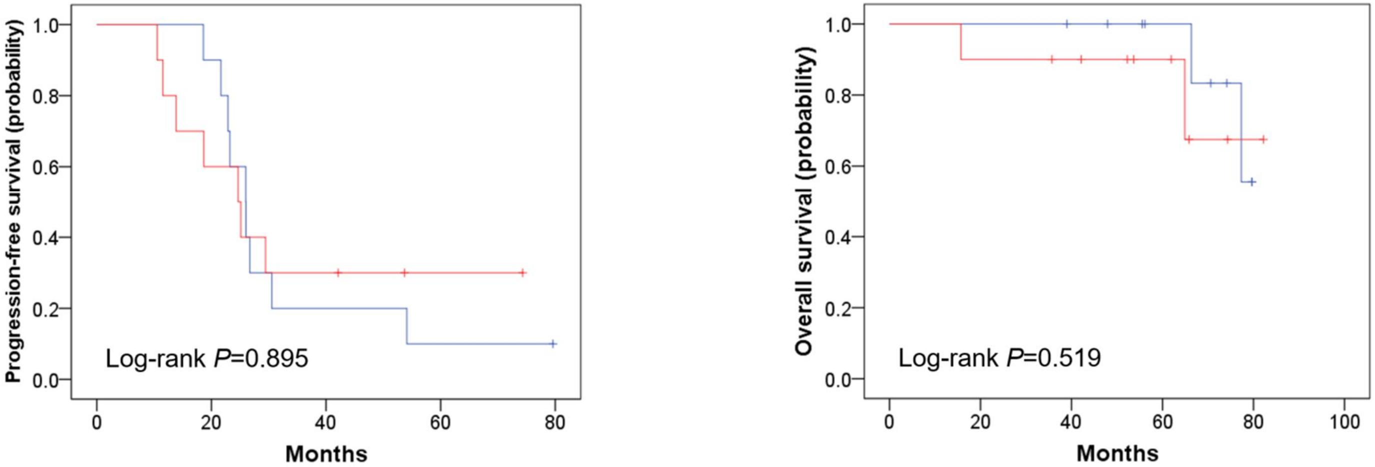


**Figure 1-1. Schematic diagram of our study design.**



| Germline *BRCA1/2* | N | Events | Median (months) | 95% CI |
|---|---|---|---|---|
| 🟥 Wild-type | 10 | 7 | 24.6 | 14.6–34.7 |
| 🟦 Mutation | 10 | 9 | 26.0 | 21.7–30.3 |

| Germline *BRCA1/2* | N | Events | Mean (months) | Standard error |
|---|---|---|---|---|
| 🟥 Wild-type | 10 | 2 | 71.6 | 6.8 |
| 🟦 Mutation | 10 | 2 | 76.8 | 2.0 |

**Figure 1-2. Comparisons of survival outcomes between germline *BRCA1/2* mutation and wild-type groups in terms of progression-free survival (PFS, left), and overall survival (OS, right).**

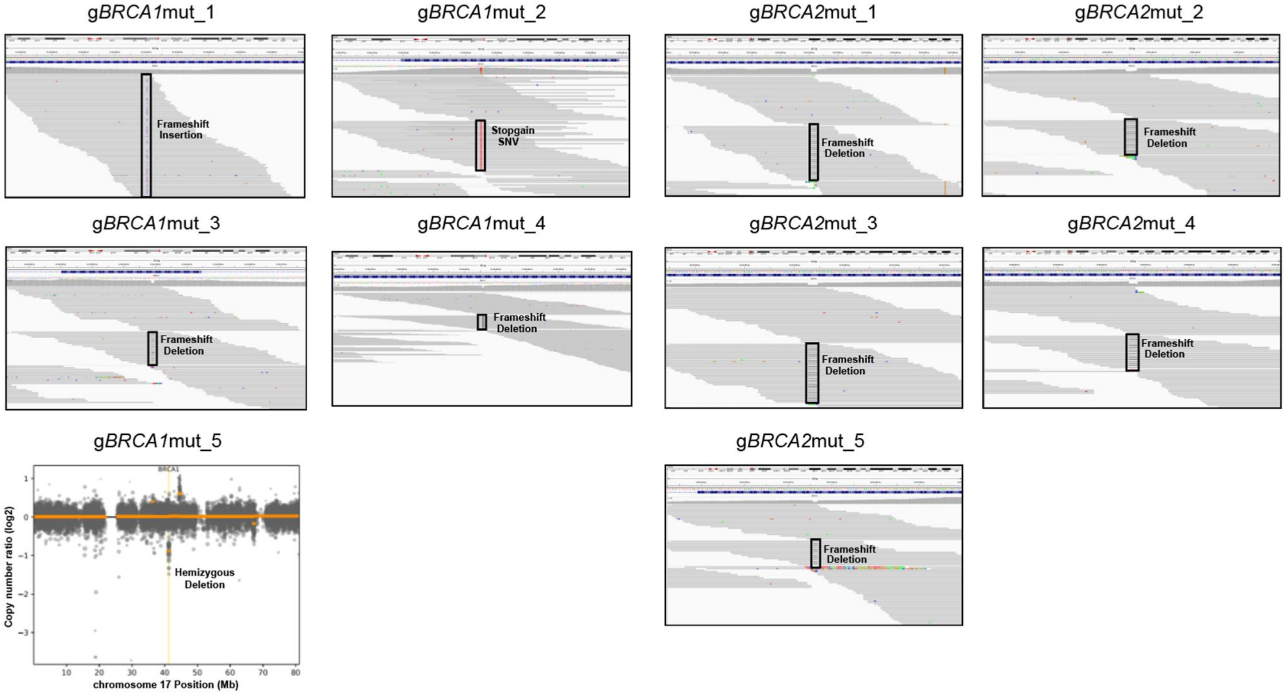


**Figure 1-3. Germline *BRCA1/2* mutations across g*BRCA1/2*mut samples validated by whole-exome sequencing**. Integrative Genomics Viewer alignment views (a copy number scatterplot for g*BRCA1/2*mut_5) of germline mutations across 10 g*BRCA1/2*mut samples show next-generation-sequencing-validated hemizygous mutations in the *BRCA1* or *BRCA2* gene.

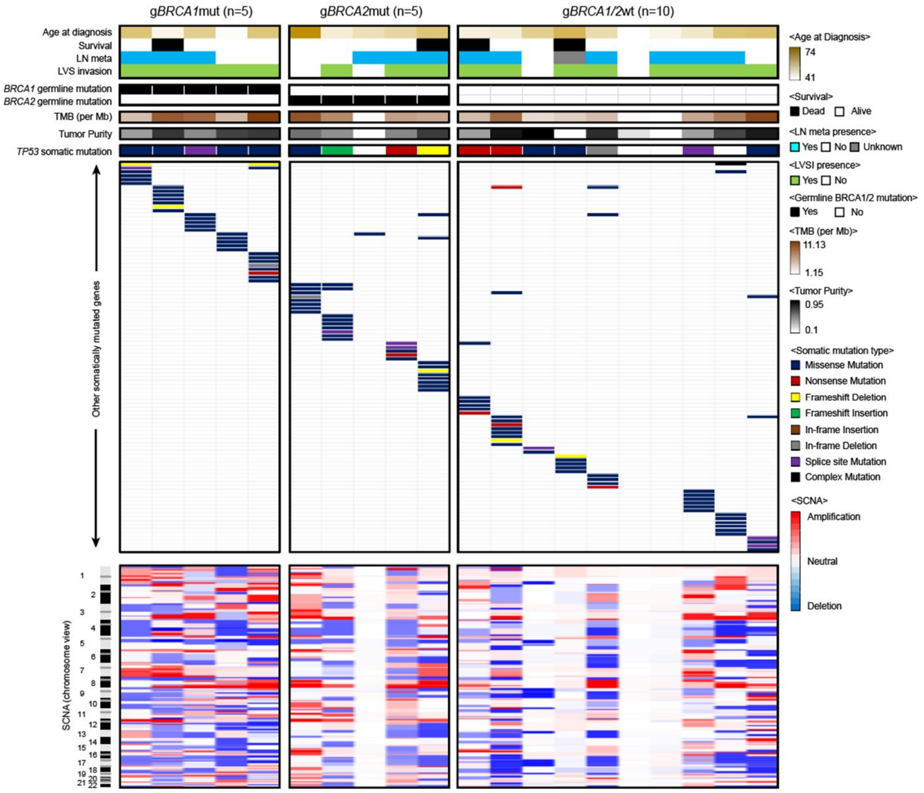


**Figure 1-4. Genomic mutational characterization of 20 HGSOC samples.** The distribution of somatic mutations among three categories of samples is presented here as oncoplot. Each column displayed here represents an individual case. LN, LVSI, TMB, and SCNA stand for lymph node, lymphovascular space invasion, tumor mutational burden, and somatic copy number alteration, respectively.

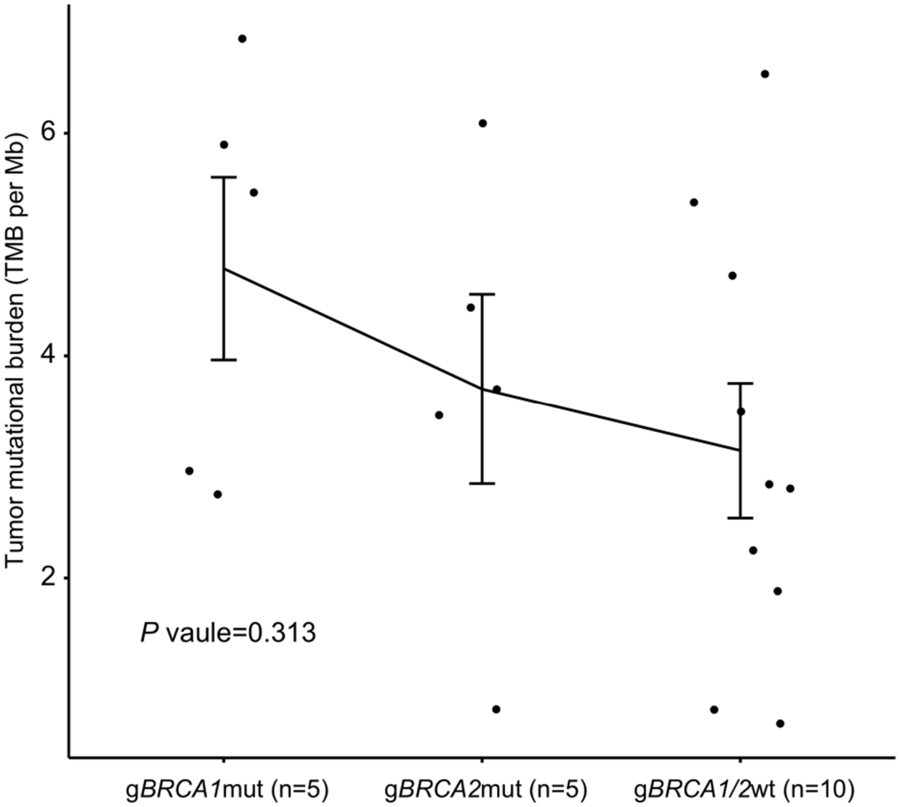


**Figure 1-5. Boxplots showing TMBs across different groups of patients.** There were no statistical differences in TMBs (one-way ANOVA, p = 0.313) among g*BRCA1*mut, g*BRCA2*mut, and g*BRCA1/2*wt samples. Each dot represents each TMB value of an HGSOC sample, while the average TMB values for each group are connected with a line. Boxplots show the 95% confidence interval for each group

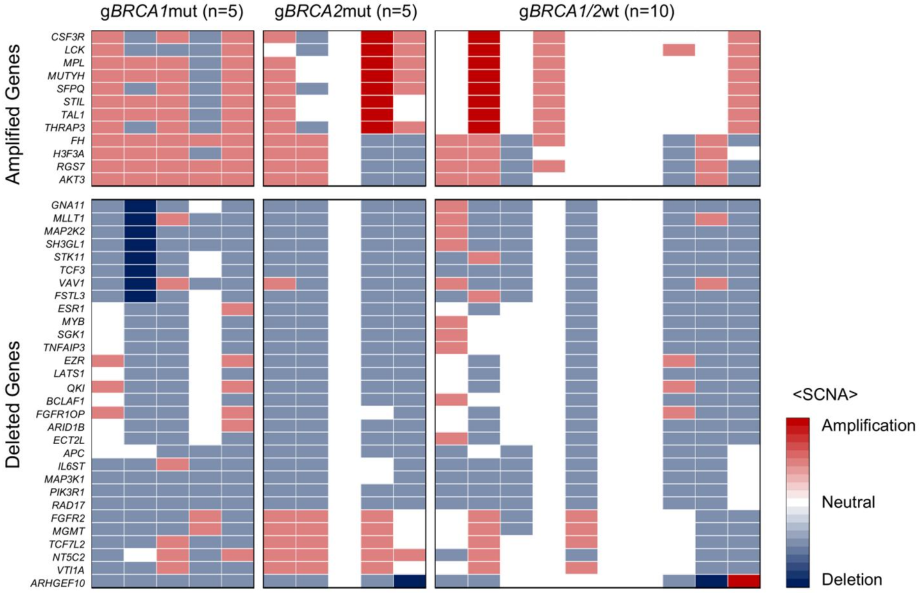


**Figure 1-6. Somatic copy number alteration profiles of 20 HGSOC samples.** Highly amplified or deleted genes are presented here as a heatmap. Each column represents an individual patient.

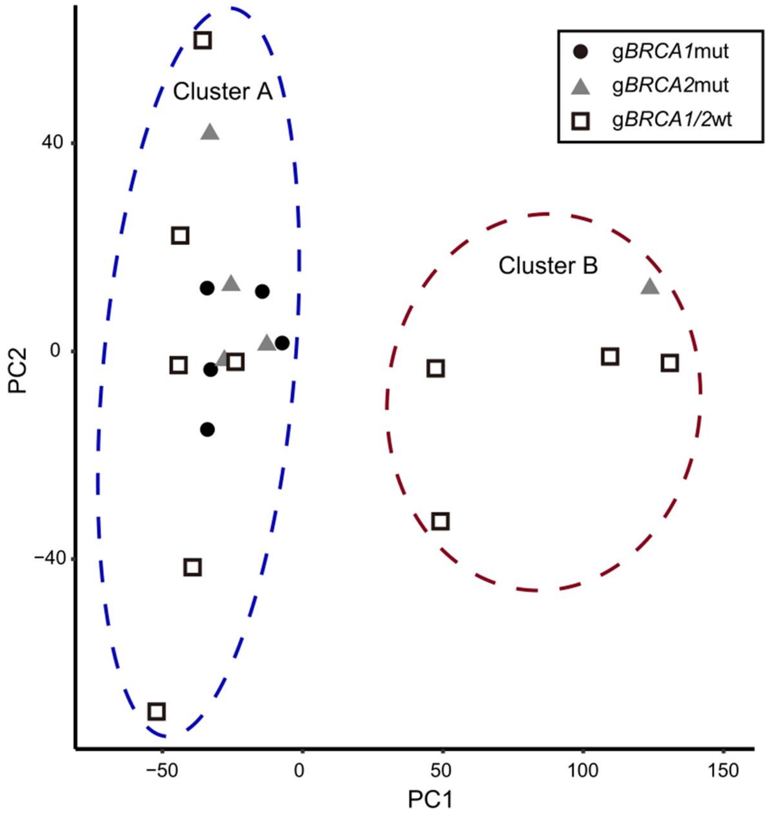


**Figure 1-7. Transcriptional landscape of HGSOC samples through principal component analysis.** Samples are represented by different shapes and colors by their origin and grouped according to K-means clustering with k = 2 (cluster A and cluster B).

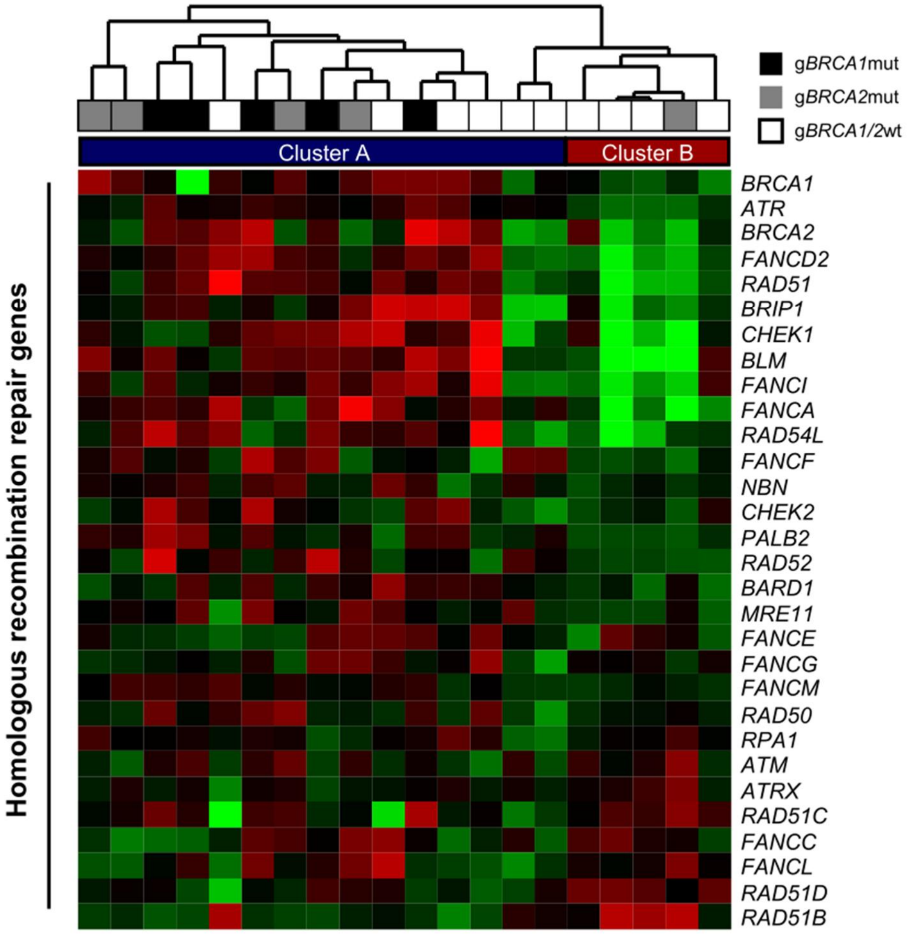


**Figure 1-8. Hierarchical clustering of samples represents the expression profile of 30 HRR genes.**

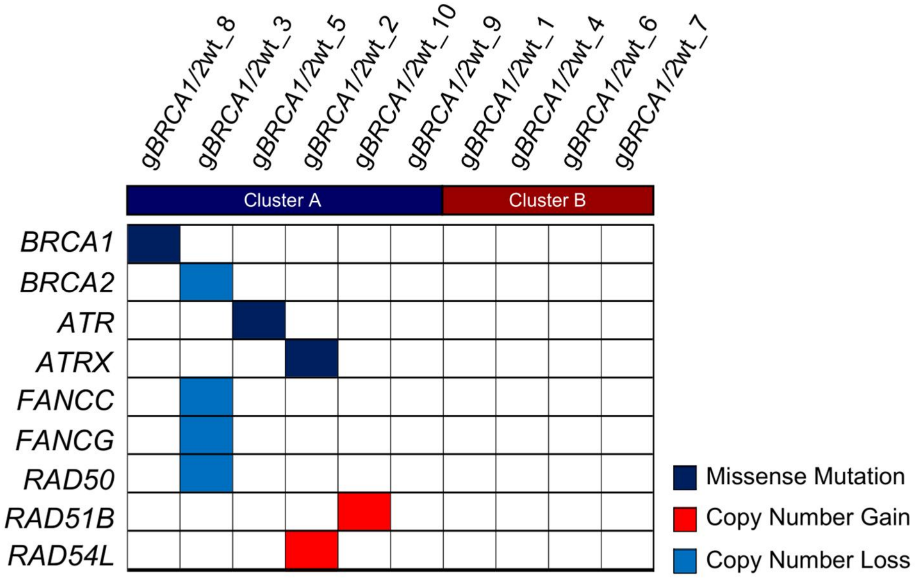


**Figure 1-9. Aberration of HRR genes across g*BRCA1/2*wt samples.** The distribution of HRR gene alterations across 10 g*BRCA1/2*wt tumor samples is represented. Each row corresponds to each tumor sample, and each row corresponds to an altered HRR gene.

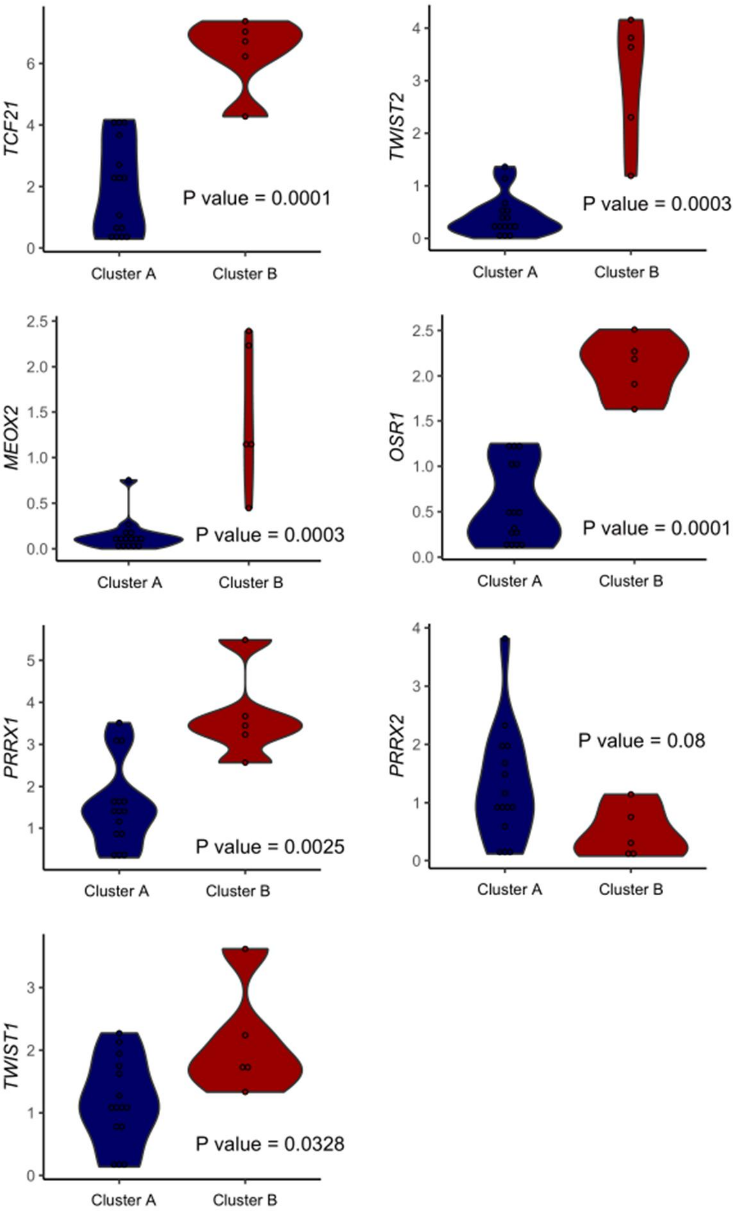


**Figure 1-10. Boxplot showing the expression of TFs related to EMT across cluster A and cluster B.** Boxplot shows the expression of EMT-related TF genes derived from TF enrichment analysis of genes displaying positive correlation (Pearson r > 0.9) with the PC1 value of the principal component analysis.

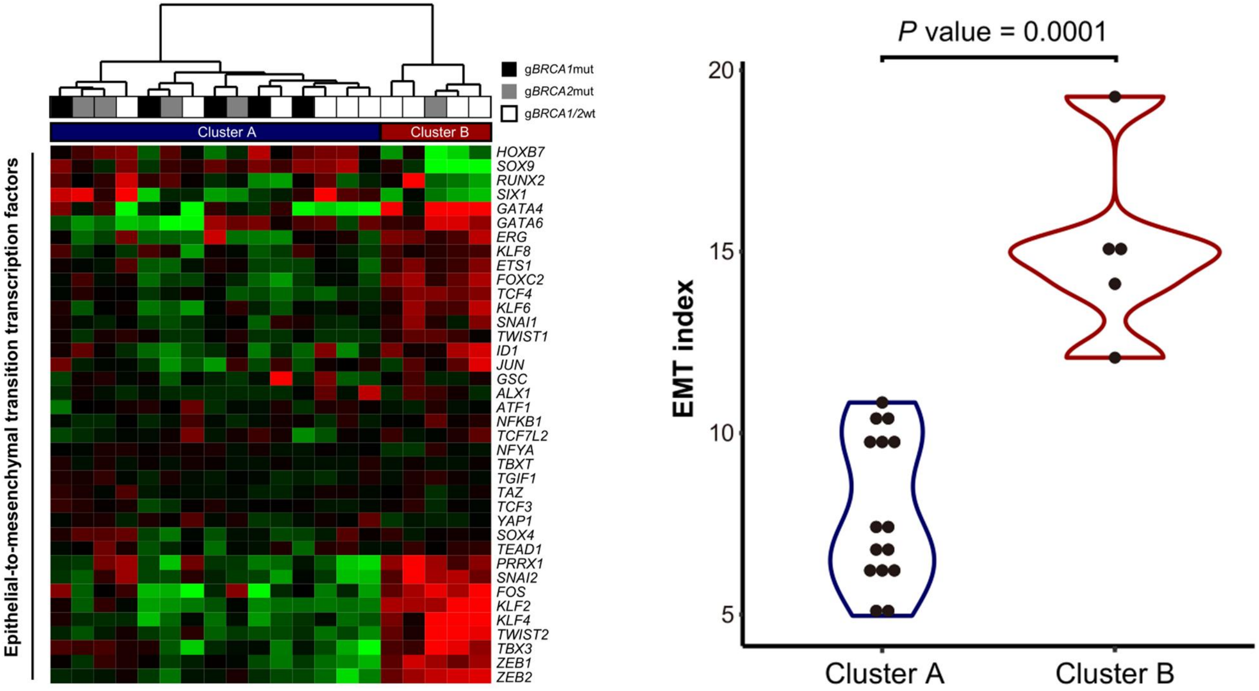


**Figure 1-11. Expression dynamics of EMT-TFs represented by Hierarchical clustering (left) and distribution of EMT index between cluster A and cluster B (right).** Hierarchical clustering of samples with the expression profile of 38 EMT-TFs recapitulated the result from PCA analysis. Violin plot shows the difference in EMT index between cluster A and cluster B.

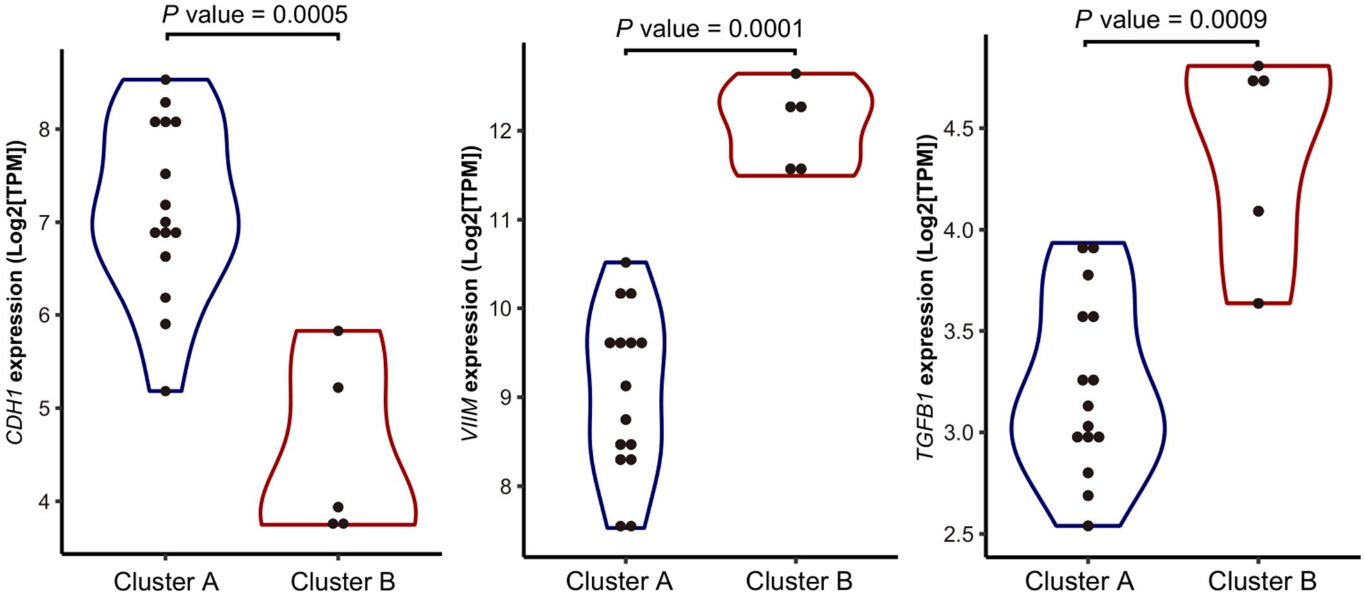


**Figure 1-12. Violin plots showing differences in gene expression of *CDH1* (epithelial cell marker), *VIM* and *TGFB1* (mesenchymal cell markers).** Each *P* value was calculated via Mann-Whitney U test.

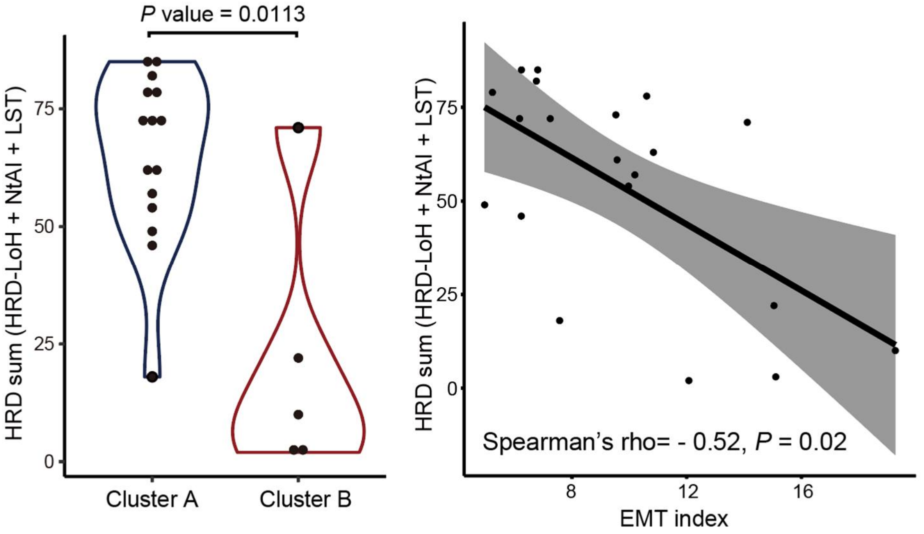


**Figure 1-13. A violin plot-view of HRD score distribution between cluster A and cluster B and relationship between EMT-index and HRD sum scores.** HRD scores between cluster A and cluster B (left) were compared using Mann–Whitney U test. Statistical dependence between EMT index and HRD scores (right) were computed through Spearman's rank correlation coefficients. LoH, NtAI, and LST stand for loss of heterozygosity, number of telomeric allelic imbalances, and large-scale transition, respectively.

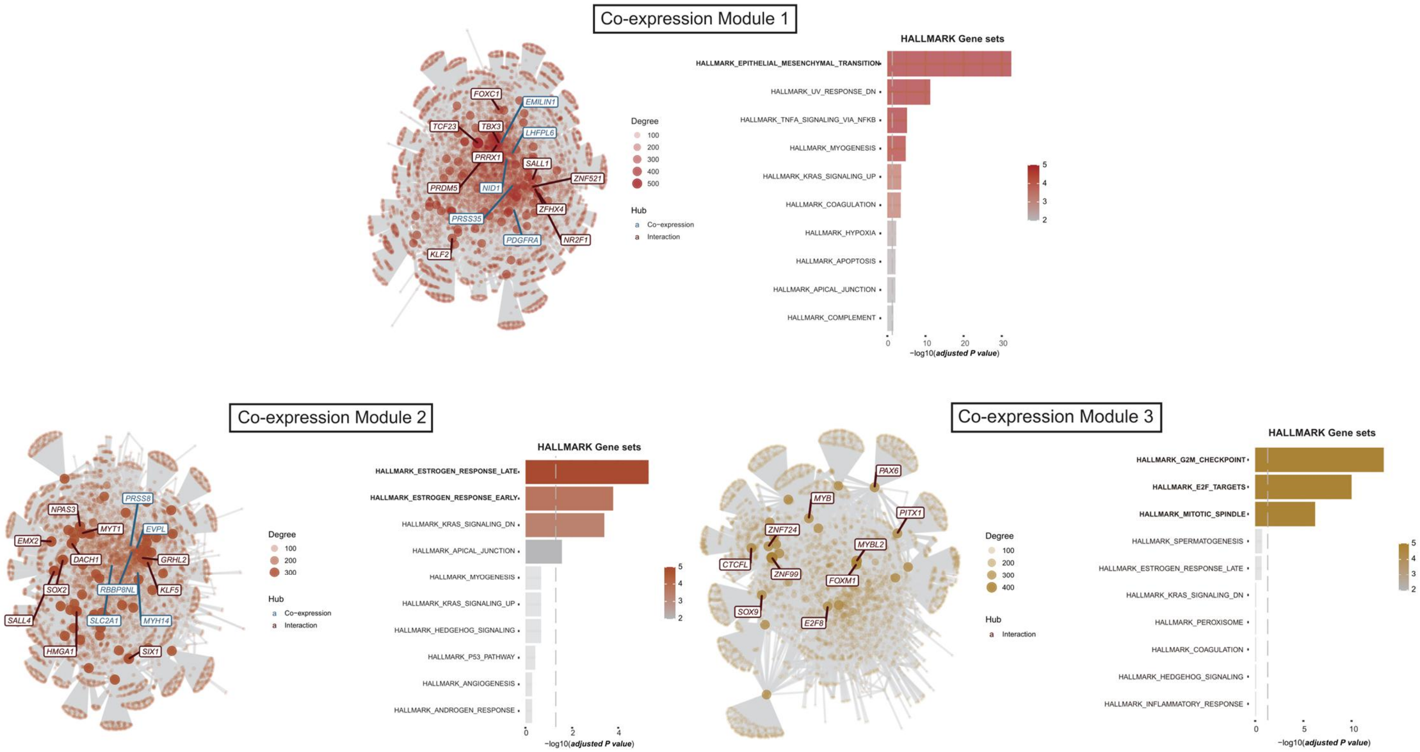


**Figure 1-14. Co-expression gene module idenfication for cluster A and cluster B.** Network of identified gene modules for cluster B (top) and cluster A (bottom) and gene-set enrichment analysis results for module genes in each network displayed on the right panel respectively.

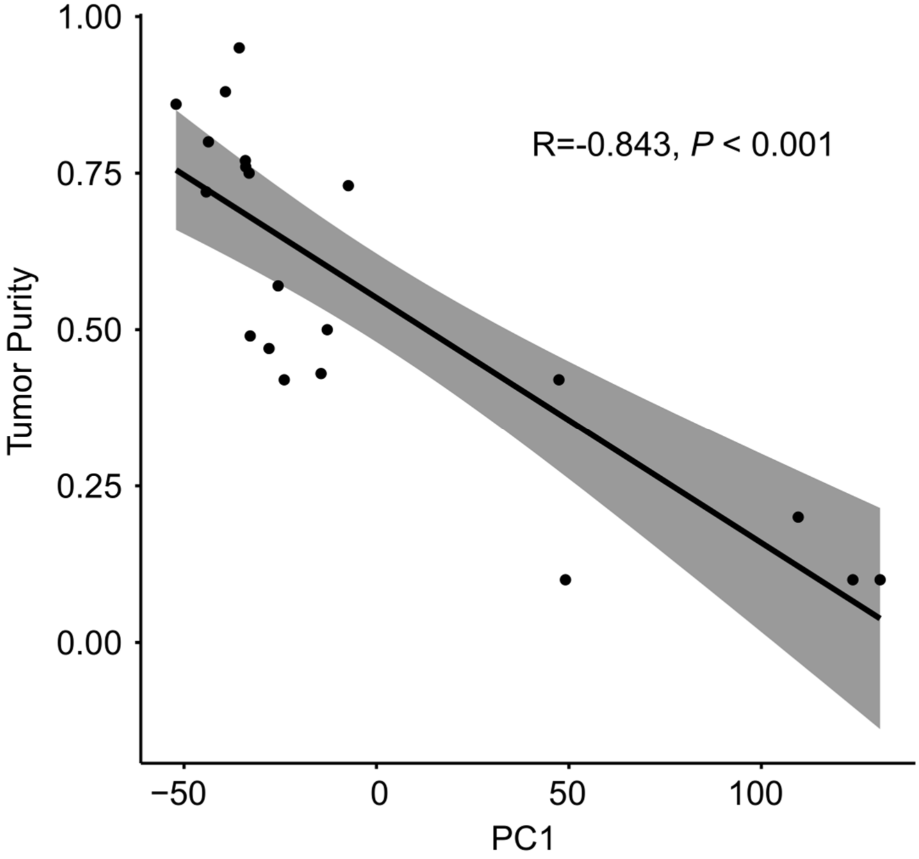


**Figure 1-15. Correlation between PC1 and tumor purity.** Significant negative correlation between PC1 from RNA-seq and tumor purity derived from whole-exome sequencing (Pearson r = −0.843 and $P < 0.001$).

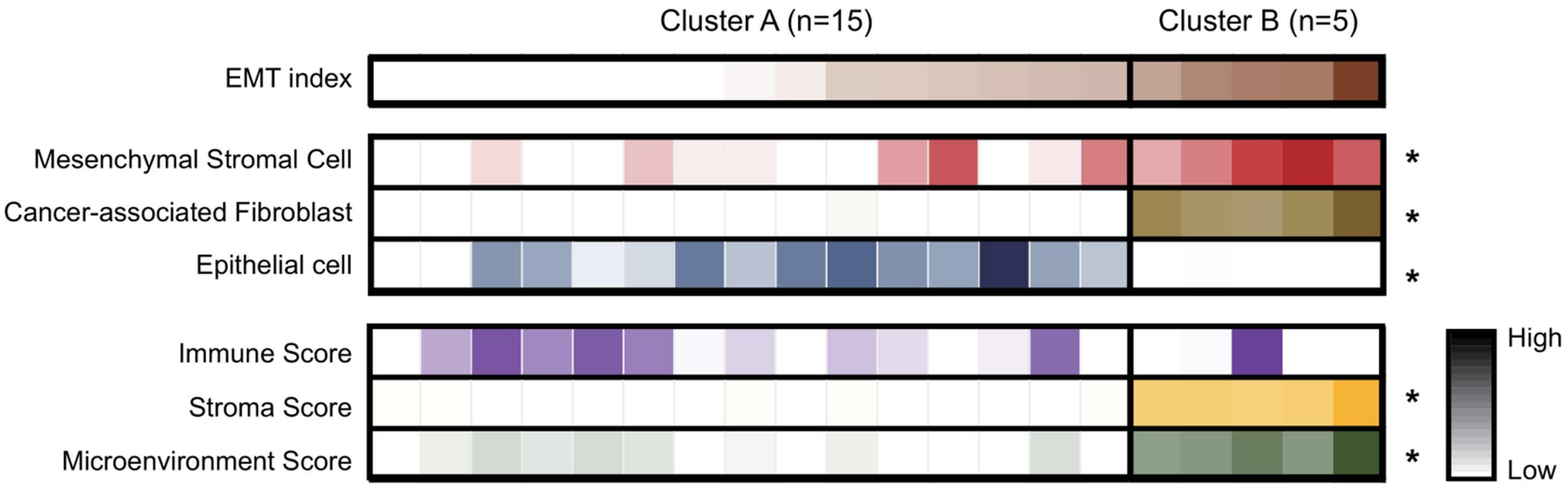


**Figure 1-16. Cell-type enrichment analysis results.** Heatmap of EMT index and cell-type enrichment analysis results across 20 HGSOC samples divided by cluster A and cluster B by order of increasing EMT index. *Mann-Whitney U test $P < 0.05$ between cluster A and cluster B.

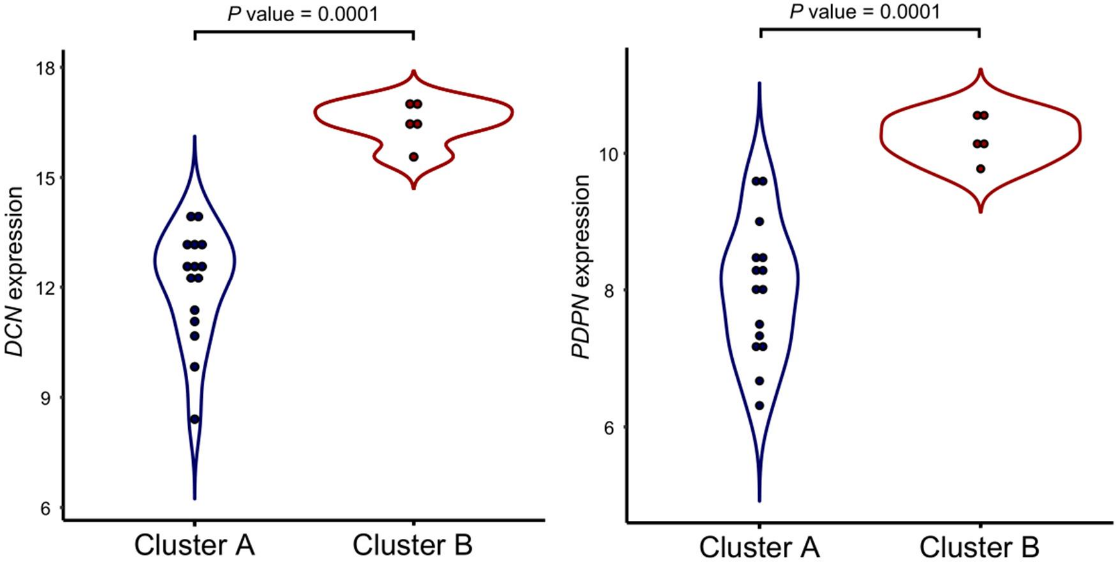
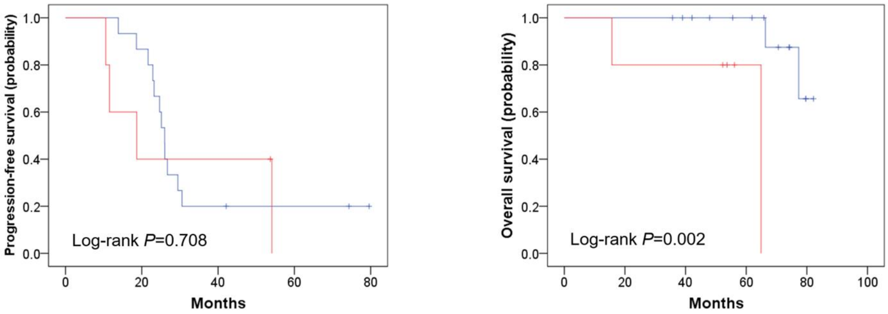


**Figure 1-17. Expression of two CAF marker genes for cluster A and cluster B.** Mann-Whitney U test *P* value for each observation is represented above each violin plot.



| Cluster | N | Events | Median (months) | 95% CI |
|---|---|---|---|---|
| — HRR-activated | 15 | 12 | 26.0 | 24.2–27.7 |
| — Mesenchymal | 5 | 4 | 7.8 | 3.4–34.0 |

| Cluster | N | Events | Mean (months) | Standard error |
|---|---|---|---|---|
| — HRR-activated | 15 | 2 | 79.1 | 1.9 |
| — Mesenchymal | 5 | 2 | 55.1 | 12.4 |

**Figure 1-18. Kaplan-Meier curves of Progression-free and overall survival for patients between HRR-activated and Mesenchymal type.**

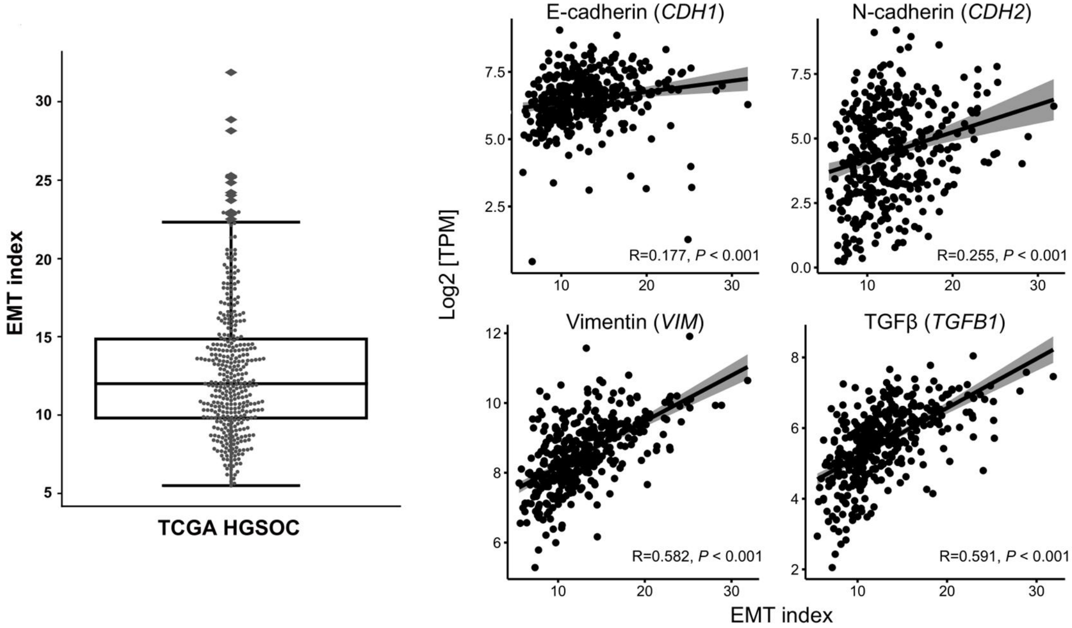


**Figure 1-19. Application of EMT index onto TCGA HGSOC data and association between EMT index and known markers of EMT.** Distribution of EMT index of TCGA HGSOC is displayed on a box plot (left). Scatter plots illustrates the relationship between EMT index and EMT-related gene expression in the TCGA HGSOC cohort. Each dot represents each sample analyzed, and linear trend between EMT index and each marker gene expression is shown respectively.

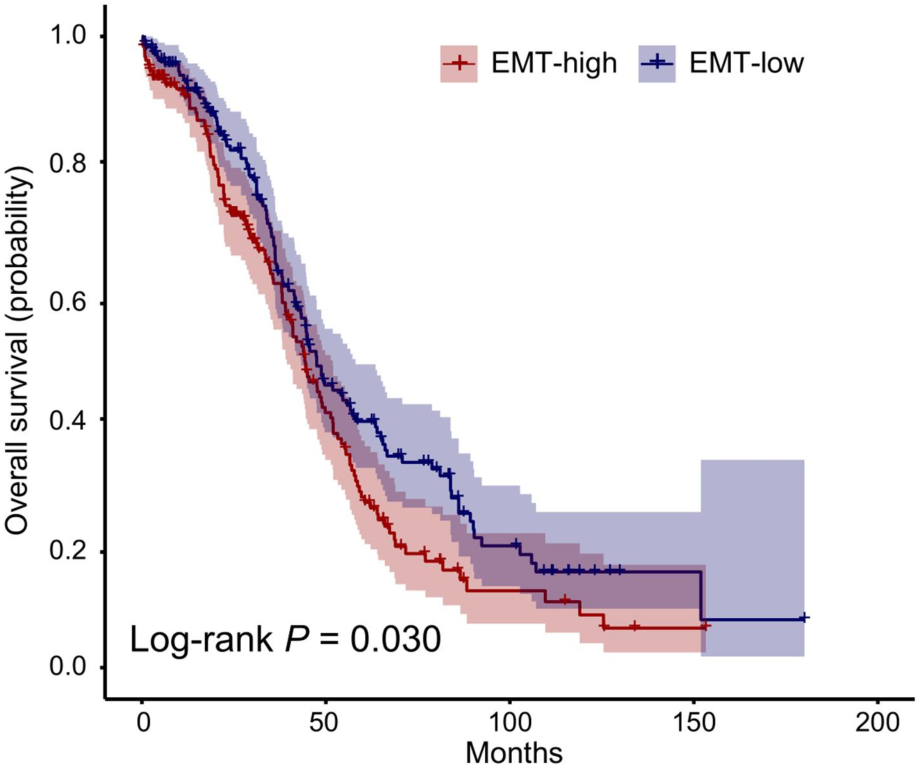


**Figure 1-20. Kaplan-Meier plot depicting overall survival of TCGA HGSOC samples falling into EMT-high and EMT-low groups.**

EMT-high groups show worse prognosis compared to EMT-low groups (Log-rank *P* = 0.030)

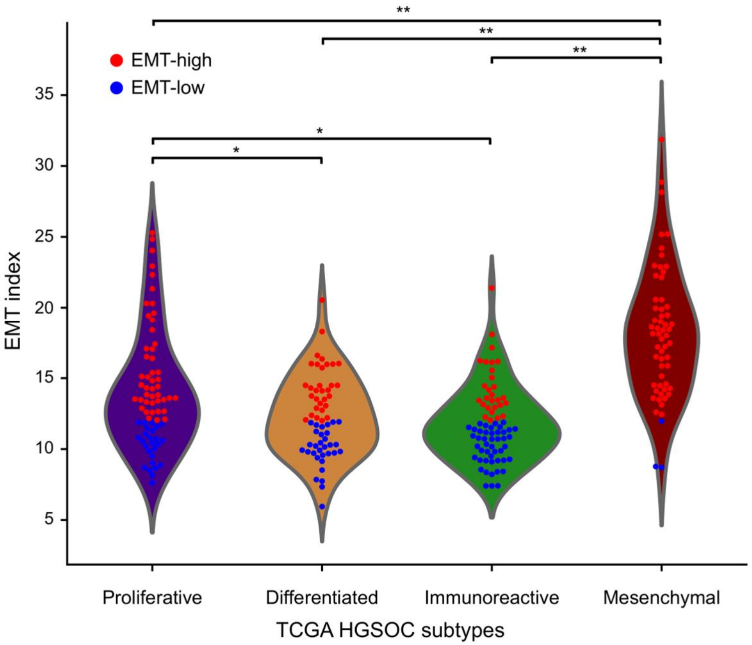


**Figure 1-21. EMT index distribution for four TCGA HGSOC subtypes.** EMT index for four different TCGA-defined HGSOC molecular subtypes was compared, and the TCGA mesenchymal subtype exhibited the highest EMT index (one-way ANOVA test $P < 0.001$; Tukey's HSD adjusted P < 0.005** and P < 0.05*). Red dots and blue dots inside the violin plots represent EMT-high and EMT-low samples, respectively.

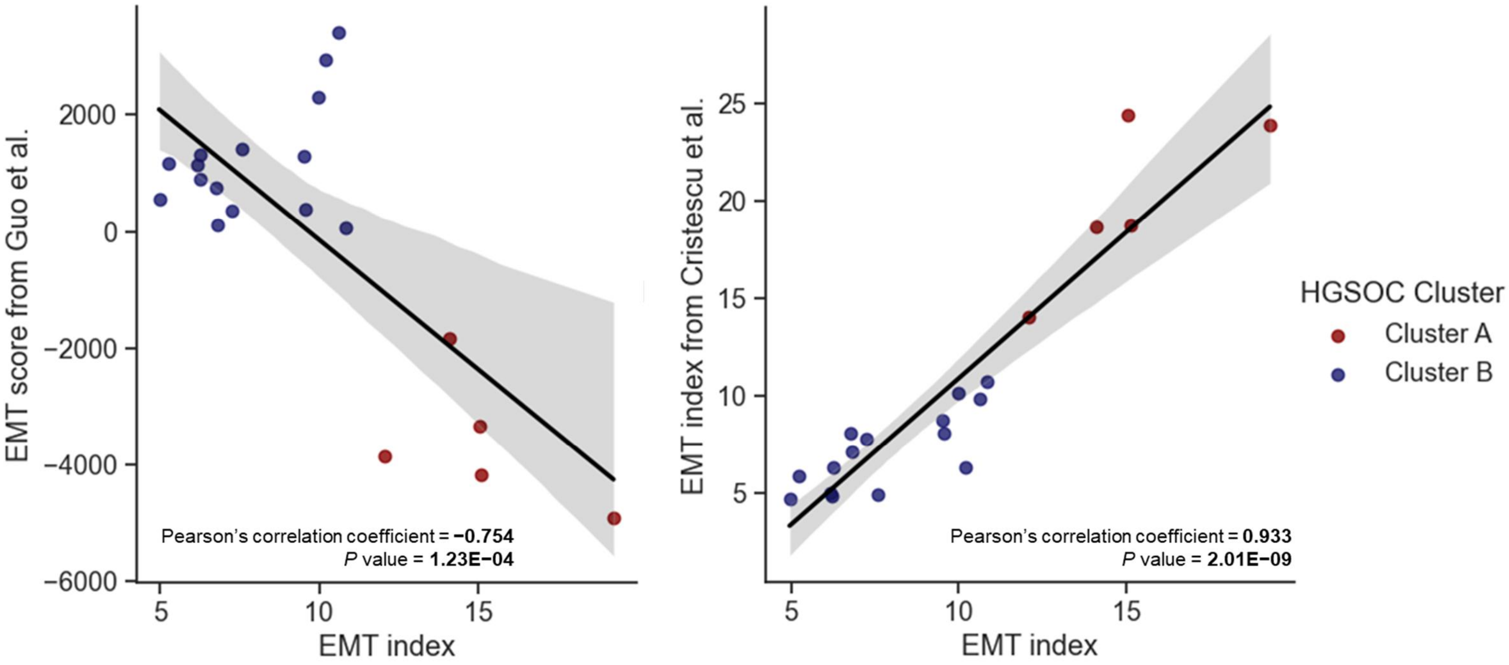


**Figure 1-22. Association between the EMT index and other EMT intensity measures identified in 20 HGSOC cohort.** All three EMT intensity measures were performing well in terms of discriminating Cluster A and Cluster B.

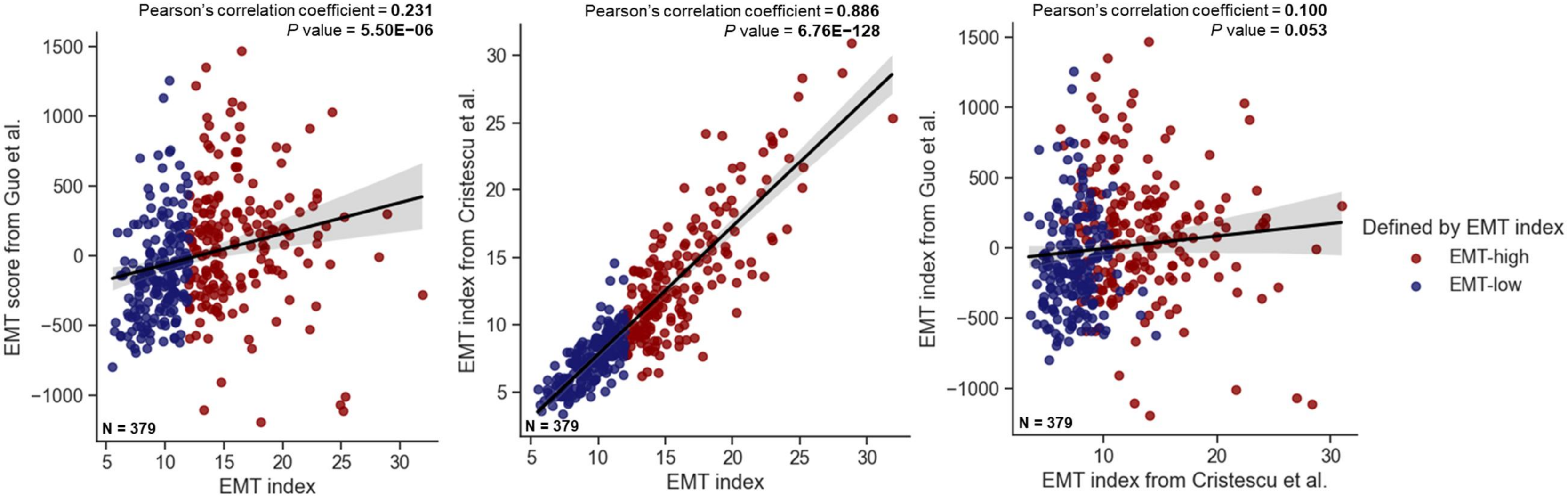


**Figure 1-23. Association between the EMT index and other EMT intensity measures identified in each sample in TCGA.**

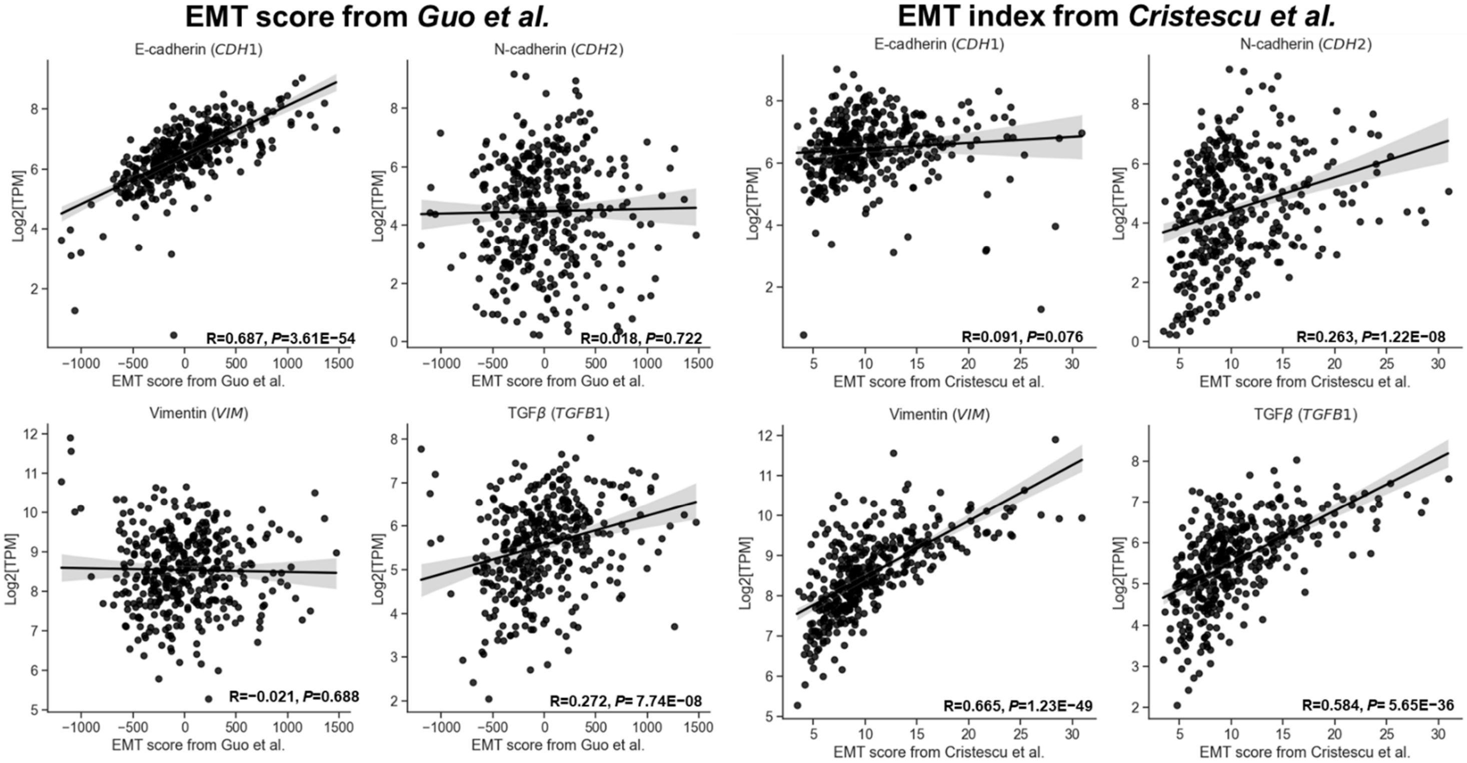


**Figure 1-24. Scatter plots illustrating the relationship between two distinctive EMT intensity measures and the expression of known markers of EMT.**

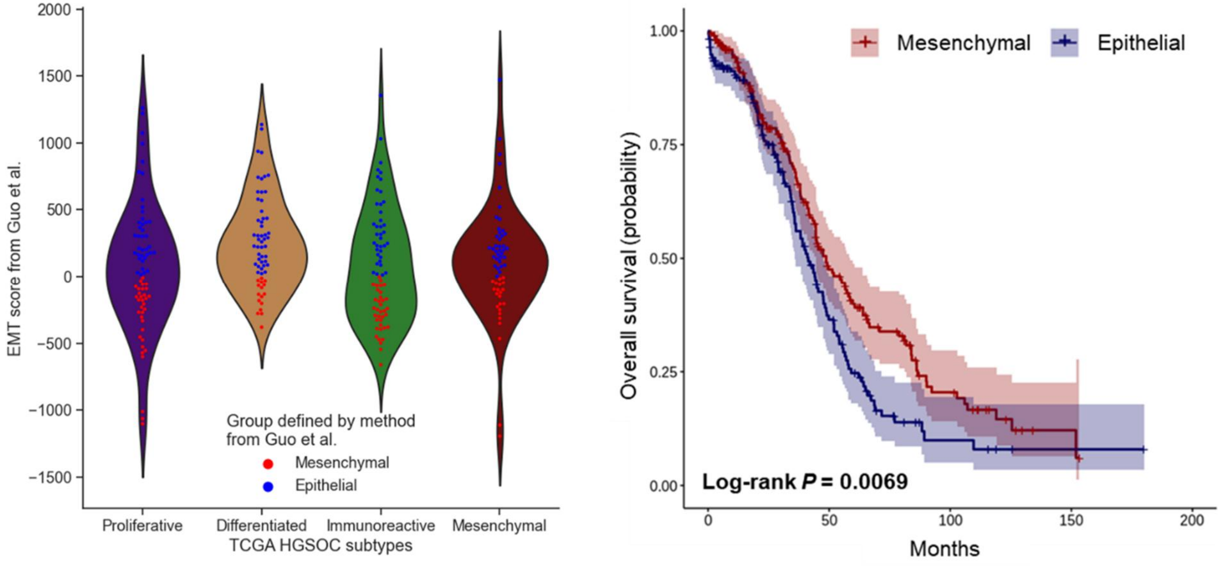


**Figure 1-25. Distribution of EMT score from Guo et al. for four TCGA HGSOC subtypes and Kaplan-Meier plot of overall survival between two groups of TCGA HGSOC falling into EMT score-based categories.** EMT index for four different TCGA-defined HGSOC molecular subtypes was compared (left). Red dots and blue dots inside the violin plots represent Mesenchymal and Epithelial samples, respectively. Epithelial group shows worse prognosis compared to Mesenchymal group (Log-rank $P = 0.0069$).

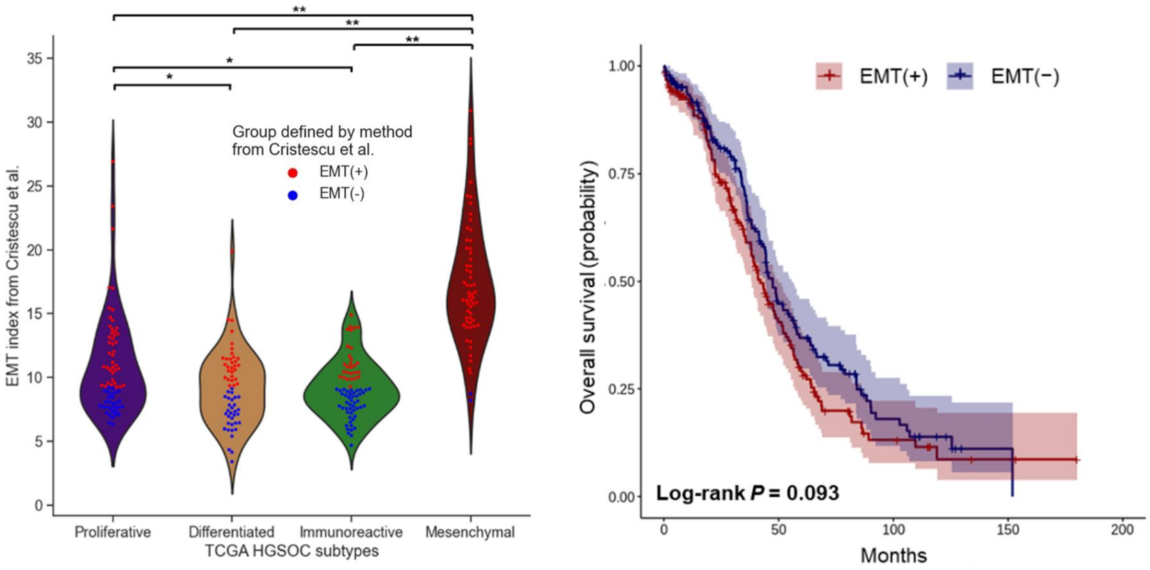


**Figure 1-26. Distribution of EMT index from Cristescu et al. for four TCGA HGSOC subtypes and Kaplan-Meier plot of overall survival between two groups of TCGA HGSOC falling into EMT index (from Gou et al.)-based categories.** EMT index from Cristescu et al. for four different TCGA-defined HGSOC molecular subtypes was compared (left), and the TCGA mesenchymal subtype exhibited the highest EMT index (one-way ANOVA test $P < 0.001$; Tukey's HSD adjusted $P < 0.005$** and $P < 0.05$*). Red dots and blue dots inside the violin plots represent EMT(+) and EMT(−) samples, respectively. There were no significant OS difference between the two groups (right).

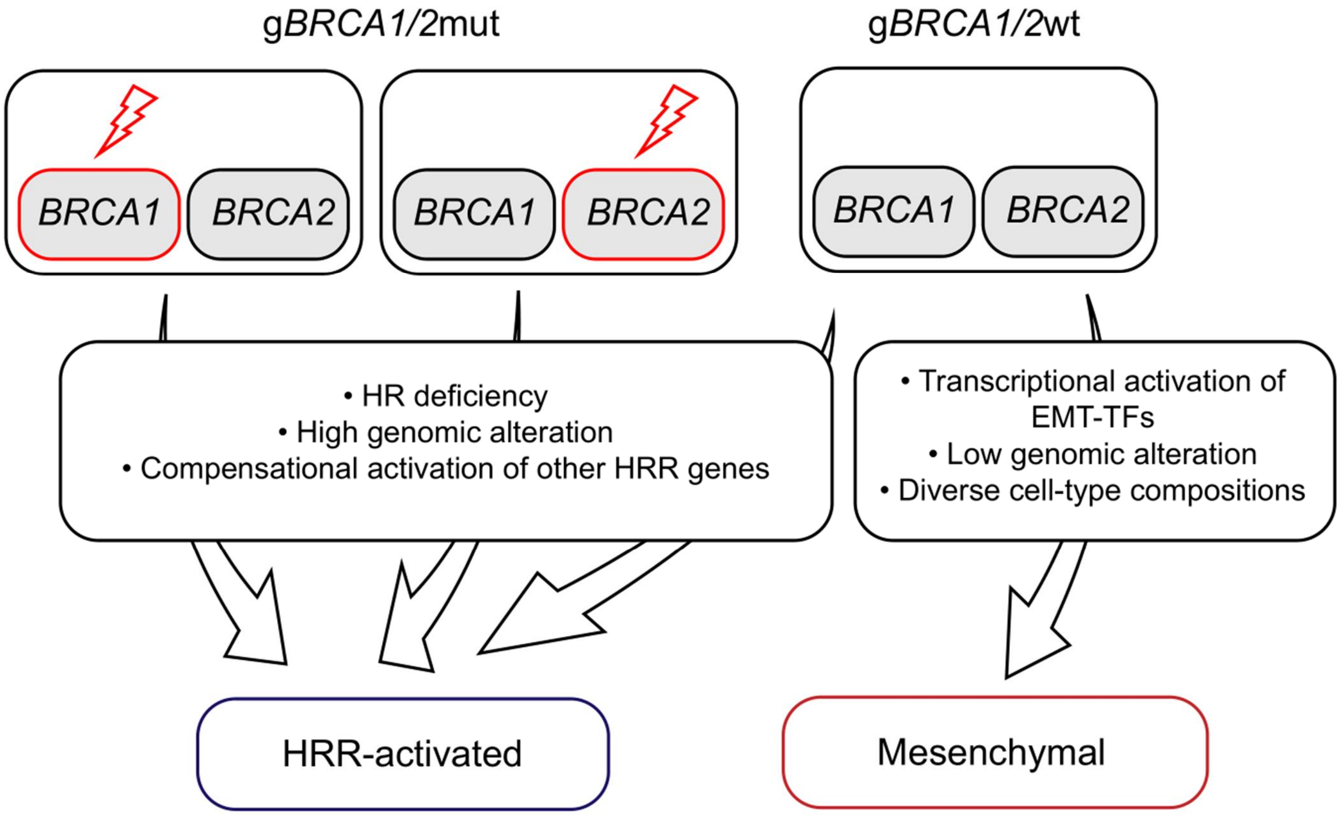

gBRCA1/2mut
gBRCA1/2wt
BRCA1
BRCA2
BRCA1
BRCA2
BRCA1
BRCA2
• HR deficiency
• High genomic alteration
• Compensational activation of other HRR genes
• Transcriptional activation of EMT-TFs
• Low genomic alteration
• Diverse cell-type compositions
HRR-activated
Mesenchymal

**Figure 1-27. Schematic diagram of our findings.**

# Discussion

In this study, we investigated the molecular characteristics of HGSOC through an integrative analysis of genomic and transcriptomic data obtained from chemotherapy-naive primary HGSOC tissues. Consequently, we could simplify the molecular classification of HGSOC to HRR-activated and mesenchymal types (**Figure 1-27**). The prognostic value of the EMT index was also validated using TCGA HGSOC data. Our study results demonstrate that the EMT index would be a potential prognostic biomarker for HGSOC.

Of two distinctive types of HGSOC, HRR-activated HGSOC was characterized by a malfunction of the HRR program caused by deficient *BRCA1/2* or HRR genes and the transcriptomic aberration of other HRR genes. Furthermore, we revealed that genes regulating or co-expressed with HRR genes are members of the E2F family (*E2F8*, *E2F2*, and *E2F3*), known as cell cycle regulators [72]; *PRDM9*, related to the process of meiosis and responsible for directing the positions of HRR [73]; *CENPA*, involved in accurate chromosome segregation [74]; and *TGIF*, reported to be over-expressed among ovarian cancer cell lines [75].

The other type, mesenchymal HGSOC, was characterized by low genomic alteration, transcriptional activation of EMT-TFs, decreased epithelial cell marker expression, increased mesenchymal cell marker expression, and diverse cell type composition. Regarding activation of EMT-TFs, a previous study in colorectal cancer reported that *ZEB1*, one of the core EMT-TFs, was activated through the β-catenin/TCF4 complex [76]. Similarly, we also observed upregulation of both β-catenin (*CTNNB1*) and *TCF4* and of their target *ZEB1* in mesenchymal HGSOCs. However, we could only infer the association of these three genes, but not their causal relationship.

EMT is currently known as one of the cancer hallmarks, being involved in tumorigenesis, metastasis, and obtaining chemoresistance [5, 24, 26, 77]. In our understanding, unlike in breast cancer, the link between *BRCA1* and EMT has not been thoroughly investigated in HGSOC. The relationship between expression profiles of HRR and EMT genes might be explained by the following hypotheses: (1) the co-existence of deficient *BRCA1/2* or HRR genes and altered expression of EMT genes together lead cancer cells to extinction; or (2) altered expression of EMT genes may contribute to the tumor microenvironment being nonviable for cancer cells with defects in *BRCA1/2* or HRR genes. To confirm these hypotheses, additional experiments using ovarian cancer cell lines are warranted.

In the current study, we leveraged the EMT index, composed of 38 genes—

five for core EMT-TFs and 33 for EMT-related TFs—which can be utilized in identifying mesenchymal HGSOC. In addition, it may be used as a prognostic marker in HGSOC; both in our samples and TCGA HGSOC data, a high EMT index was associated with significantly worse OS. At the same time, it should be noted that the proportion of stromal cells within samples might be reflected in the EMT index. Indeed, a higher proportion of stromal cells in HGSOC is known to be associated with worse OS [78].

Furthermore, various molecules, such as E-cadherin, N-cadherin, EpCAM, and vimentin, are involved in the EMT process [24]. A complex network of TFs is known to regulate EMT, leading to the downregulation of epithelial genes and the upregulation of mesenchymal genes [24, 79]. We also observed various molecules or genes related to the EMT index and regulators of EMT, including *VIM* (vimentin) and *TGFB1* (TGFβ), which were differentially expressed between the two types of HGSOC.

In terms of anti-EMT therapy, TGFβ is one of the best-studied therapeutic targets in cancer. Phase I and II clinical trials of fresolimumab (a monoclonal anti-TGFβ antibody) have been conducted in renal cell carcinoma, melanoma, mesothelioma, and breast cancer [80-82]. In ovarian cancer, blockade of TGFβ signaling with antibodies reversed EMT in epithelial ovarian cancer ascites-derived cell spheroids [83] and increased platinum sensitivity in a xenograft mouse model [84]. More research is needed to elucidate the therapeutic strategy of anti-EMT therapies in HGSOC.

Based on our study results, if an individual is identified to have a high-EMT-index HGSOC, so poor prognosis is expected, clinicians might prescribe additional targeted agents (e.g., bevacizumab) more actively. Clinicians might also consider dose-dense chemotherapy or extended chemotherapy cycles. After primary treatment, a more intensive surveillance schedule might be administered for an individual. Incorporating the EMT index with the well-known clinicopathologic risk factors of HGSOC, researchers might develop models predicting treatment response and prognosis more accurately. In this manner, we believe that precision cancer medicine can be facilitated in ovarian cancer with a relatively poorer prognosis than any other cancer.

Our study has several limitations. First, the small sample size might be one of the most problematic issues. In survival analysis, we could not conduct multivariate analysis adjusting for clinicopathologic factors. Thus, our study results should be validated in a large, multi-institutional HGSOC cohort. Second, our study results were only derived from bulky specimens composed of various malignant and non-malignant cells. Therefore, specific gene signatures of the mesenchymal HGSOC samples might be a mixed result originating from malignant epithelial or mesenchymal cells and non-malignant cells, such as CAFs, endothelial cells, and immune cells [64] To elucidate the exact cellular compositions and heterogeneity in tumor cells, as well as the cell-to-cell interactions within the tumor microenvironment, further singe-cell-level studies should be conducted. Such studies might supplement and enhance our study results. Nevertheless, we believe that the methodology of our study, especially the step-by-step integrative analysis methods, can be also used in other malignant types of cancers.

# Part 2

# DNA Methylation-driven Dysregulation of Gene Expression in 84 Stomach Adenocarcinoma Revealed by Multi-omics Analysis

# Abstract

Stomach adenocarcinoma (STAD) is a leading contributor to global cancer incidence and mortality and is responsible for over 700,000 deaths annually. In spite of the mapping of multiple molecular aberrations over STAD, our understanding of this deadly disease remains poor owing to its heterogeneous and multi-dimensional nature. Here, we addressed this challenge by using an integrated multi-omics analysis to delineate the molecular alterations of STAD in terms of genome-wide cytosine-phosphate-guanine (CpG) DNA methylation profiles and RNA expression dynamics. DNA methylation-centric analysis combined with public regulatory element data enabled us to identify multiple ablated gene pathways including activation of canonical WNT/β-catenin signaling, downregulation of super-enhancer proximal genes and reactivation of pattern specification genes in homeobox clusters upon DNA hypermethylation. This study underscores the context-dependent epigenetic dysregulation in STAD, and pinpoints various epigenetically dysregulated sites for potential biomarkers and therapeutic targets.

---

**Keywords:** Epithelial neoplasm; Stomach adenocarcinoma; Transcriptome sequencing; Whole-genome bisulfite sequencing; WNT/β-catenin signaling

**Student number:** 2014-25063

# Introduction

Stomach adenocarcinoma (STAD), defined as neoplasia of glandular epithelial cells of the gastric mucosa, is one of the major histological types of stomach cancer, and is responsible for > 750,000 deaths worldwide [14]. STAD has a high incidence rate particularly in East Asia, with South Korea having the second highest incidence rate behind Japan [14].

Manifestation of STAD has been ascribed to various environmental risk factors such as smoking, obesity, and a diet high in smoked, salted foods. Furthermore, viral or bacterial infection is another risk factor for this disease. In fact, long-term infection with the bacteria helicobacter pylori (H. pylori) in the stomach has been proven to be one of the main determinants of stomach cancer, as this bacterium can cause inflammation and pre-cancer growth [85]. Infection with the Epstein-Barr virus, a herpes virus best known for causing mononucleosis, has also been linked to stomach cancer.

Prior to the advent of genomics era, histological classification of stomach adenocarcinoma was first proposed by Lauren et al., in the name of Lauren's classification [86] which classified STAD into intestinal and diffuse categories.

In addition, other histopathological categorization of STAD have been proposed, one of which being WHO classification [87], which is comprised of papillary, tubular, mucinous, poorly cohesive and mixed adenocarcinoma. Nonetheless, unlike clinical staging (e.g., TNM stage) routinely applied to treatment and management of STAD, utilization of aforementioned histopathological variabilities is currently challenging [88].

In the middle of 2014, amidst extensive application of NGS into identifying STAD biology, TCGA [89] categorized the stomach adenocarcinoma, through one of the pioneering multi-omics studies, into four distinct genomic subtypes which comprised of the Epstein-Barr virus (EBV) subtype, the microsatellite instability (MSI) subtype, the genomically stable (GS) subtype, and lastly a chromosomal instability (CIN) subtype. In other group, the Asian cancer research group (ACRG), by using RNA expression profile from microarray data of STAD, classified gastric tumors with four distinct subtypes including MSI, microsatellite stability with the characteristics of epithelial-to-mesenchymal transition (MSS/EMT) and MSS/epithelial tumors further divided into two by the presence of TP53 signatures (MSS/TP53+ and MSS/TP53−) [50]. Although the above mentioned studies are sufficiently valuable, they somewhat lacked epigenetic-centered analysis. Particularly, ACRG subtypes were focused on RNA expression level and lacked complete information on epigenetics in terms of DNA methylation as they inferred its signature using pre-defined sets of gene expression, and TCGA group concerted their efforts only on CpG island methylator phenotype (CIMP, which often associated with transcriptional gene silencing) assayed through DNA methylation microarray technology.

The term “epigenome” refers to the epigenetic information in a cell, which includes DNA methylation, post-translational modifications of histones, and higher-order chromatin structure [8]. Addressed as one of the key epigenetic elements governing the cellular identity, DNA methylation and its changes have long been recognized as a crucial factor in cancer formation [8].

Here, to elucidate the pivotal role of epigenetic change and subsequent aberration of gene expression in the process of STAD tumorigenesis, the tumor and matched normal samples from 84 Korean gastric adenocarcinoma patients were collected, and both WGBS and RNA-seq were performed. In addition to this bi-modality, to highlight the exertion of complex regulatory elements governing STAD formation, we obtained publicly available candidate cis-regulatory element (cCRE) and five histone modification datasets of normal from a healthy individual and putative transcription factor binding sites (TFBS) from ENCODE [90-92] and super enhancer elements from SEdb [93]. Along with these multi-modal data, we set out to extend our understanding of comprehensive epigenetic alteration and their consequent aberration of gene expression in STAD.

# Materials and methods

## Primary STAD specimens

Matched clinical samples of STAD and their adjacent normal stomach mucosa tissues are obtained from 84 patients as fresh frozen specimens at Bundang Seoul National University Hospital. Written informed consents were obtained from patients before surgery. These specimens were used for two different sequencing assays; WGBS (n=168) and RNA-seq (n=168).

## RNA-seq library preparation and sequencing

We prepared RNA libraries using the Illumina TruSeq Stranded Total RNA kit according to the manufacturer's protocol. First, ribosomal RNA (rRNA) depletion was performed followed by fragmentation and conversion to RNA molecule to complementary DNA (cDNA) using reverse transcriptase. Single-stranded cDNAs were then converted to double-stranded cDNAs and end-repaired and adenosine-tailed, and adaptor-ligated. The libraries constructed were amplified using polymerase chain reaction (PCR). Subsequently we sequenced cDNA libraries using NovaSeq 6000 (Illumina, San Diego, CA, USA) with 151bp paired-end configuration.

### WGBS library preparation and sequencing

WGBS libraries were prepared using Accel-NGS Methyl-Seq DNA Library kit (Swift Biosciences) according to manufacturer's instructions. The DNA was treated with the sodium bisulfite using EZ DNA methylation GOLD kit (Zymo Research). After the adapter ligation step, prepared libraries were amplified using PCR. Reads were sequenced using NovaSeq 6000 (Illumina, San Diego, CA, USA) 151bp paired-end configuration.

### RNA-seq data processing

RNA-seq reads were first checked for low-quality bases and adapter contamination and those reads were trimmed using TrimGalore, followed by pseudoalignment by Kallisto v0.46.2 [42] (with default option except for --rf-stranded --bias) onto transcriptome indices generated from GENCODE [94] V24 transcripts fasta. After read trimming and pseudoalignment, we checked for QC metrics using percentage of pseudoaligned reads, uniquely pseudoaligned reads and reads mapped onto rRNA for each individual sample.

Raw read counts were batch adjusted by “ComBat_seq” function in sva [95] package version 3.42.0 using sequencing run information as batch covariates, leaving tumor and normal sample information untouched. Briefly, sva package, by using negative binomial regression to model batch effects, provides the adjusted data by comparing the original count data distribution to an expected distribution if there were no batch effects in the first place.

Next, we performed discovery of differentially expressed genes (DEGs) using DESeq2 1.34.0 [43] onto raw count matrix after filtering genes of which sum of counts across samples were below 10. We modeled a raw count matrix using patient and condition (~patient + condition). Then, log fold change shrinkage was performed from apeglm package version 1.16.0 [96]. Normalized counts transformed using variance stabilized transformation (vst) were analyzed for downstream analysis such as PCA, hierarchical clustering.

### WGBS data processing

Owing to the nature of WGBS library preparation technology using adaptase, in order to accurately identify DNA methylation information from WGBS reads, it is imperative to perform read trimming procedure off the first 10~15bp as well as adapter and low-quality bases. Thus, we implemented trimming using TrimGalore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/; --clip_R1 10 --clip_R2 15 --three_prime_clip_R1 12 --three_prime_clip_R2 13 --illumina --paired). WGBS reads were then aligned onto C>T and G>A converted hg38 reference genome (bismark_genome_preparation) using bismark v0.21.0 [97] (bismark --maxins 700 --dovetail) with bowtie2 v2.4.2 [98]. After removal of duplicate reads (deduplicate_bismark --paired), methylation states under CpG, CHH (where H stands for one of cytosine (C), thymine (T) and Adenine (A) base) and CHG contexts were extracted (bismark_methylation_extractor --paired-end --no_overlap --cutoff 1 --cytosine_report). After selecting CpGs from 22 autosomes and X chromosome, we combined the CpG calls on each Watson and Crick strand and only the CpG site where total coverage is 5 or more is considered valid in the downstream analysis. Quality of the DNA methylation data was evaluated by bisulfite conversion rate inferred from lambda genome spike-in, non-CpG methylation (e.g., cytosines in CHH and CHG contexts) percentage and genomic coverage of CpG sites.

### Subtyping based on TCGA STAD molecular classification

Based on molecular classification for STAD from TCGA consortium [89], we attempted to categorize our 84 STAD samples into 4 distinct subtypes. With the lack of somatic DNA copy number alteration profiles needed to distinguish GS type from CIN type, we only managed to classify our 84 STAD samples into 3 categories of which comprise EBV, MSI and GS/CIN.

Briefly, EBV positivity of specific sample based on clinical information were first used to identify EBV type (N=4) and subsequently MSI groups were categorized using the level of MSI (N=10). Remaining samples were categorized into GS/CIN types (N=70).

### Epimutation Burden calculation

The epimutation burden calculation was adopted from previously described method [99]. Briefly, we count the number of CpG sites, where read coverage above 5 and the difference between tumor DNA methylation and normal DNA methylation is 20% or more. Then we divided this by the total number of CpG sites, resulting in an epimutation burden measure for each sample. Difference of epimutation burden between samples belonging to the TCGA classification were identified using Mann-Whitney U test.

### Discovery of differentially methylated regions (DMR) and downstream analysis

Combined CpG coverage files for each site for each sample were converted to bedGraph format followed by identification of differentially methylated regions (DMR) between tumor and normal using metilene v0.2-8 [100] from the merged bedGraph file as an input. We restricted our downstream DMR analysis on those which absolute methylation difference between tumor and normal is above 10%, 15% or on those which adjusted *P* value is below 0.05 according to the different circumstances. Then, smooth DNA methylation percentage for each sample inside DMRs was calculated using BSmooth algorithm [101]. Hierarchical clustering was performed using clustermap function in Seaborn version 0.11.2. We transformed DNA methylation bedGraph files to bigWig then to tdf format for visualizing CpG methylation information using Integrative Genomics Viewer (IGV) [102]. Genomic coordinate bed file of each hyper-DMR (Tumor-Normal >15%) and of hypo-DMR (Tumor-Normal< −15%) was uploaded to GREAT tool [103] version 4.0.4 to perform functional analysis with a default parameter.

### DMR annotation using various publicly available regulatory elements

We first downloaded the various histone modification call sets retrieved from a reference epigenome for a healthy human stomach (ENCSR949WGV, a 53 year old female) in ENCODE portal (**Table 2-2**) [91] with the following identifiers: H3K4me3 (ENCFF588TFE), H3K27ac (ENCFF910HDI), H3K27me3 (ENCFF313RCC), H3K4me1 (ENCFF712YGW), H3K9me3 (ENCFF152CYD) and H3K36me3 (ENCFF927MLI). In order to annotate DMR with promoter element, we opted not to define arbitrarily sized promoter according to distance with TSS. Instead, we defined promoter utilizing promoter-like signatures (PLS) subcategory constituting cis-regulatory elements (cCREs) [90]. These elements (PLS, PLS-CTCF-bound) were downloaded from the Search Candidate cis-Regulatory Elements by ENCODE (SCREEN) website (screen.encodeproject.org). If there were transcripts with no PLS element present in the vicinity of TSS (NO-PLS), only then we applied arbitrarily defined promoter, upstream 500bp and downstream 500bp of TSS. If there were identical gene name present in both PLS and NO-PLS, we filtered out NO-PLS element and used only the assigned PLS. Next, we annotated PLS with histone modification marks with the following criteria: (1) PLS were merged if they were close to 100bp each other using bedtools merge -d 100 (2) merged PLS region were extended 100bp both upstream and downstream using bedtools slop -b 100 (3) annotate extended PLS with 5 different histone marks (H3K4me3, H3K27ac, H3K27me3, H3K4me1, H3K9me3) using bedtools intersect. PLSs were further filtered according to their assigned gene_biotype in GENCODE. We only used the if they are one of 'protein_coding', 'lincRNA', '3prime_overlapping_ncRNA', 'antisense', 'bidirectional_promoter_lncRNA', 'macro_lncRNA', 'non_coding', 'processed_transcript', 'sense_intronic', and 'sense_overlapping'. Then, PLS overlapped with DMR (adjusted *P* value < 0.05) were selected.

Uniform TFBS from ENCODE (https://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredV3.bed.gz) were lifted over from hg19 to hg38 using LiftOver tool [104].

### Calculating DNA methylation inside gene body, SE element and HOX clusters

The DNA methylation levels inside each of the gene bodies, ENCODE stomach tissue SE element [93] were calculated using the weighted average DNA methylation using a custom Python script. In order to generate heatmap of the DNA methylation inside HOX clusters, we segmented four HOX clusters (HOXA, HOXB, HOXC and HOXD) with the 1000bp window from the start to the end of each cluster and calculated weighted DNA methylation using a custom Python script.

### Discovery and analysis of partially methylated domains (PMD)

PMDs were identified using MethPipe v4.1.1 [105] (pmd -i 1000 -b 1000 -s 42) separately for each sample. Methylated and unmethylated CpG read counts were used as input as previously described [99]. Only the CpG sites with total coverage above 5 were used to identify PMDs. Identified PMDs were further filtered by the score given by MethPipe (below 100) and by the length (below 100kb). Then, PMDs from individual tumor were subsequently merged using bedtools unionbedg to make a union set of 2,282 PMDs. The average PMD methylation level was calculated using the weighted average DNA methylation using a custom Python script.

### CpG island methylator phenotype (CIMP) detection and analysis

The CpG island methylator phenotype was investigated based on recurrently methylated promoter CGIs similar to previously described method [106]. Briefly, the average methylation level for 27,755 CGIs was calculated for each sample using smoothed CpG methylation level using BSmooth [101]. The CGIs determining the CIMP (CIMP-CGIs) were selected based on the following three criteria: (1) CGI overlapped with promoter-like signatures (PLS) (2) mean methylation level across normal samples below 40%; (3) difference in CpG methylation level between tumor and matched normal samples above 10%. We identified total of 239 CIMP-CGIs in this manner and performed hierarchical clustering using CpG DNA methylation difference between tumor and normal samples. CIMP (+) tumors were then selected using hierarchical clustering tree structure.

### Code Availability

All codes used in this study are publicly available at https://github.com/ryansohny/STAD.

# Results

## DNA methylation landscape in STAD revealed by WGBS

Initially, we collected 84 pairs of primary STAD tissues and their adjacent normal tissues (**Table 2-1**) and performed WGBS and RNA-seq (**Figure 2-1**). We achieved WGBS with 30X coverage per sample, and over 99% bisulfite conversion rates across all samples and moreover, the average non-CpG methylation in our samples were 0.66% and 0.64% for CHG context and CHH context respectively, ascertaining the great quality of our WGBS samples (**Figure 2-2**). Comparison of the average DNA methylation of the individual CpG between tumors and normal samples revealed global CpG hypomethylation is present in STAD (**Figure 2-3**, paired t-test *P* value = 3.87E-08) which is one of the general characteristics of tumor [107]. Consistent with this finding, partially methylated domains (PMD) which refer to long stretches of genomic areas of hundreds of kilobase pairs (kb) with highly disorganized methylation levels and which were known to frequently observed in various types of tumors and cultured cell lines [108], were recurrently detected in our STAD samples compared to normal counterparts (**Figure 2-4**). The overall global hypomethylation could be accounted for by these recurring PMD present in our tumor samples. In terms of detected PMD, several tumor samples were affected by PMD up to almost half of their cancer genomes (**Figure 2-5**). Furthermore, we assayed CIMP-CGI (**Figure 2-6**) which is implicated in numerous cancers and especially well characterized in STAD [88, 89]. We found out that the DNA methylation level of CIMP-CGIs displayed negative correlation with the PMD methylation level (Spearman's rho=−0.60, **Figure 2-7**). CIMP(+) tumors (N=35) showed higher tendency to exhibit global hypomethylation compared to CIMP(−) tumors (N=49).

### Applying TCGA classification method to our STAD cohort

Using six different molecular assays, TCGA consortium characterized stomach cancers into four subtypes of which comprise of EBV, MSI, GS and CIN [89]. As MSI-type, and particularly EBV-type stomach cancers exhibited extreme abnormalities in DNA methylation compared to that of GS and CIN, we checked if DNA methylation aberration of the two aforementioned types, EBV and MSI, was relatively pronounced compared to that of GS and CIN. We first categorized 14 out of 84 samples into two categories based on EBV-positivity (EBV; N=4) and MSI-high status (MSI; N=10). Due to the lack of somatic copy number alteration information which is needed to discern GS-type from CIN-type, we labeled the remaining samples GS/CIN (N=70). A flowchart outlining how our 84 samples were categorized into these 3 subtypes is illustrated in **Figure 2-8**. Then, epimutation burden [99], which accounts for the intensity of DNA methylation aberration for a given sample, was calculated and compared among these 3 subtypes (**Figure 2-9**). No statistical difference was observed among the epimutation burden of these TCGA subtypes, meaning the classification method proposed by TCGA study, is failed to be replicated in our STAD cohort in terms of ablation of DNA methylation.

**Differentially methylated regions (DMR) associated with chromatin modifications**

Hence, to better understand the distinct characteristics of our STAD cohort compared to TCGA cohort in terms of DNA methylation, we extended our analysis into differentially methylated regions (DMR) defined as the regions of DNA methylation-dysregulated sites in tumors compared to those of normal tissues (**Figure 2-10**). We found 263,337 number of DMR (FDR q value < 0.05) between tumor and normal tissues, 5.13% (N=13,503) of which were hypermethylated DMR (hyper-DMR) while 94.87% (N=249,834) were hypomethylated DMR (hypo-DMR), consistent with the global decrease of average DNA methylation in tumors. Using principal component analysis (PCA, **Figure 2-11**) and unsupervised hierarchical clustering (**Figure 2-12**), we discovered that DNA methylation inside the DMR (absolute DNA methylation difference between Tumor and Normal > 10%) can distinguish normal tissues from malignant tissues. Examining the CpG density in the region where DMR occurred, we discovered that hyper-DMRs were located in areas with higher CpG density than hypo-DMR were (**Figure 2-13**), in accordance with the fact that the sites where gene regulation occurs are CpG-rich regions which typically characterized by low DNA methylation in normal cells [109]. Next, we performed functional enrichment analysis [103] onto each hyper-DMR (difference in methylation[Tumor-Normal] > 20%) and hypo-DMR (difference in methylation[Tumor-Normal] < −20%) using gene-sets in gene ontology biological process (**Figure 2-14**). We found out that hyper-DMRs were enriched in terms such as "digestive tract development" and "digestive system development", suggesting hyper-DMRs generally occurred on sites where nearby genes govern normal stomach physiology. Also, a "canonical WNT signaling pathway" term of which its dysregulation implicated in human cancer malignancies [110] was significantly enriched in hyper-DMR regions.

When we evaluated the enrichment of regulatory elements in DMR relative to the background whole genome by fold enrichment test similar to method from Lee et al [111]. and Kundaje et al. [112], we found out that the trimethylation at the 4th lysine residue of the histone H3 protein (H3K4me3) and trimethylation at the 27th residue of the H3 protein (H3K27ac), both of which are promoter marks of active genes, were enriched in Hyper-DMR more than 10-fold compared to the expected values (**Figure 2-15**). Super enhancer element was the third-highest category in hyper-DMR followed by promoter

(promoter-like signature (PLS) in ENCODE cis-regulatory elements (cCREs)), and CpG islands. H3K27me3, which is a repressive gene mark showed ~6 fold enrichment, followed by transcription factor binding sites (TFBS). In terms of TFBS enrichment, we found that binding sites for polycomb group genes (e.g. *SUZ12*, *EZH2* and *CTBP2*) responsible for transcriptional repression were enriched in hyper-DMR regions (**Figure 2-16**). Hypo-DMR enriched TFBS (**Figure 2-17**) consisted mainly of ATP-dependent chromatin remodeler, SWI/SNF components [113, 114] (e.g. *SMARCC1*, *SMARCC2* and *SMARCB1*) implicated in transcriptional regulation [115] and tumor formation [116].

### Gain of DNA methylation at the *WNT2* Promoter region associated with *WNT2* activation

As we observed the apparent enrichment of histone modification marks demarcating promoter from other elements (H3K4me3, H3K27ac and H3K27me3) within DMR, we set out to investigate the association between the change of promoter DNA methylation level and the respective gene dysregulation. Notably, among this association, we found a wingless-type MMTV integration site family member 2 (*WNT2*), a member of the WNT gene family which is made up of structurally related genes that code for signaling proteins that are part of the canonical WNT signaling pathway [110, 117]. WNT2 proteins have been linked to tumorigenesis and a variety of developmental processes. Normally, *WNT2* expression is silenced by EZH2-mediated H3K27me3 of the respective gene promoter. In our data, *WNT2* promoter hypermethylation was associated with the increased expression in STAD (**Figure 2-18**, left). Interestingly, the *CTNNB1* which together with *WNT2*, makes up the WNT/β-catenin signaling, was also upregulated in STAD, and known targets of β-catenin, extracellular metalloproteins *MMP3* and *MMP9* were significantly upregulated in STAD as well. Furthermore, the hypermethylation of *WNT2* promoter and upregulation of *WNT2* is implicated in esophageal cancer and colorectal cancer [118, 119]. Moreover, *WNT5A* which is also one of the components of canonical WNT pathways, exhibited promoter hypermethylation and respective gene upregulation (**Figure 2-18**, right), implicating activation of general WNT pathways and downstream signaling in our STAD cohorts.

### Ablation of DNA methylation in Super Enhancers and HOX clusters

Increasing evidence suggests that the DNA methylation aberration of the super enhancer [120, 121] which is defined as large clusters of active transcription enhancer histone marks (e.g. H3K27ac) is prevalent in tumorigenesis [122]. As we observe over ~6-fold enrichment of hyper-DMR in super enhancer region, this led us to explore the hyper-methylated super enhancers and concomitant dysregulation of their nearby genes (**Figure 2-19**). Among DEGs between tumor and normal, 19 out of 30 (~63%) super-enhancer hypermethylated genes were upregulated in tumors and negatively correlated with the DNA methylation of respective super enhancer. Notably, super-enhancer nearby *FOXA2*, one of the pioneer transcription factors identified to recruit coactivators and interact with other transcription factors at functional enhancers [123] and implicated in lung cancers as a putative tumor suppressor [124], was marked by significant DNA hypermethylation, and *FOXA2* expression was significantly reduced in STAD compared to in normal tissues (**Figure 2-20**). Another notable gene associated with super-enhancer hypermethylation, and subsequent downregulation was *CASZ1* (**Figure 2-21**), a candidate tumor suppressor gene in neuroblastoma [125]. Among genes upregulated upon super enhancer hypermethylation, we found out that several HOXA genes (*HOXA1, HOXA3, HOXA6, HOXA9, HOXA10, HOXA11, HOXA13*) were significantly upregulated in STAD. Intriguingly we observed that four homeobox gene clusters (HOXA, HOXB, HOXC and HOXD clusters) involved in developmental patterning process [126, 127] were all marked with significant DNA hypermethylation across H3K27me3 repressive histone modification marks and associated with the respective gene upregulation (**Figure 2-22**). Given that these homeobox clusters were enriched with binding sites for polycomb repressive complex 2 such as EZH2 and SUZ12, we assumed that DNA hypermethylation-associated detachment of repressive factors is related to homeobox gene activation and STAD manifestation in our cohort.

## Table 2-1. The STAD patients' clinical characteristics

| **Characteristics** | ***N*** |
|---|---|
| *Number of subjects* | 84 |
| **Sex** | |
| Male | 65 (77.38%) |
| Female | 19 (22.62%) |
| **Median age** | 63.0 (37-93) |
| **WHO classification** | |
| Poorly cohesive | 27 (32.14%) |
| Well-differentiated and Moderately-differentiated tubular | 27 (32.14%) |
| Poorly-differentiated tubular | 16 (19.04%) |
| Papillary | 5 (5.95%) |
| Mucinous | 1 (1.19%) |
| Mixed | 2 (2.38%) |
| Others | 6 (7.14%) |
| **Lauren's classification** | |
| Intestinal | 39 (46.43%) |
| Diffuse | 37 (44.05%) |
| Indeterminate | 8 (9.52%) |
| **pT stage** | |
| T1a | 5 (5.95%) |
| T1b | 13 (15.48%) |
| T2 | 18 (21.43%) |
| T3 | 21 (25.00%) |
| T4a | 25 (29.76%) |
| T4b | 2 (2.38%) |
| **pN stage** | |
| N0 | 26 (30.95%) |
| N1 | 11 (13.10%) |
| N2 | 17 (20.24%) |
| N3a | 13 (15.48%) |
| N3b | 17 (20.24%) |
| **pM stage** | |
| M0 | 79 (94.05%) |
| M1 | 5 (5.95%) |
| **Lymphatic invasion** | |
| Positive | 58 (69.05%) |
| Negative | 26 (30.95%) |
| **Vascular invasion** | |
| Positive | 23 (27.38%) |
| Negative | 61 (72.62%) |
| **Perineural invasion** | |
| Positive | 43 (51.19%) |
| Negative | 41 (48.81%) |
| **EBV** | |
| Positive | 4 (4.76%) |
| Negative | 80 (95.24%) |
| **MSI** | |
| High | 10 (11.90%) |
| Low | 74 (88.10%) |
| **HER2 immunohistochemistry** | |
| 0 | 45 (53.57%) |
| 1+ | 25 (29.76%) |
| 2+ | 9 (10.71%) |
| 3+ | 4 (4.76%) |

Abbreviations: pT stage, pathological assessment of the primary tumor; pN stage, pathological assessment of the regional lymph nodes; pM stage, pathological assessment of metastasis; EBV, Epstein Barr Virus; HER2, Human epidermal growth factor receptor 2

**Table 2-2. Public regulatory element data used in this study**

| Assay | Reference Epigenome | Experiment | Accession | Links |
|---|---|---|---|---|
| H3K4me3 | ENCSR949WGV | ENCSR489ZLL | ENCFF588TFE | https://www.encodeproject.org/experiments/ENCSR489ZLL/ |
| H3K27ac | ENCSR949WGV | ENCSR133NBJ | ENCFF910HDI | https://www.encodeproject.org/experiments/ENCSR133NBJ/ |
| H3K27me3 | ENCSR949WGV | ENCSR357ROS | ENCFF313RCC | https://www.encodeproject.org/experiments/ENCSR357ROS/ |
| H3K4me1 | ENCSR949WGV | ENCSR903QBX | ENCFF712YGW | https://www.encodeproject.org/experiments/ENCSR903QBX/ |
| H3K9me3 | ENCSR949WGV | ENCSR546HZF | ENCFF152CYD | https://www.encodeproject.org/experiments/ENCSR546HZF/ |
| H3K36me3 | ENCSR949WGV | ENCSR166CNR | ENCFF927MLI | https://www.encodeproject.org/experiments/ENCSR166CNR/ |

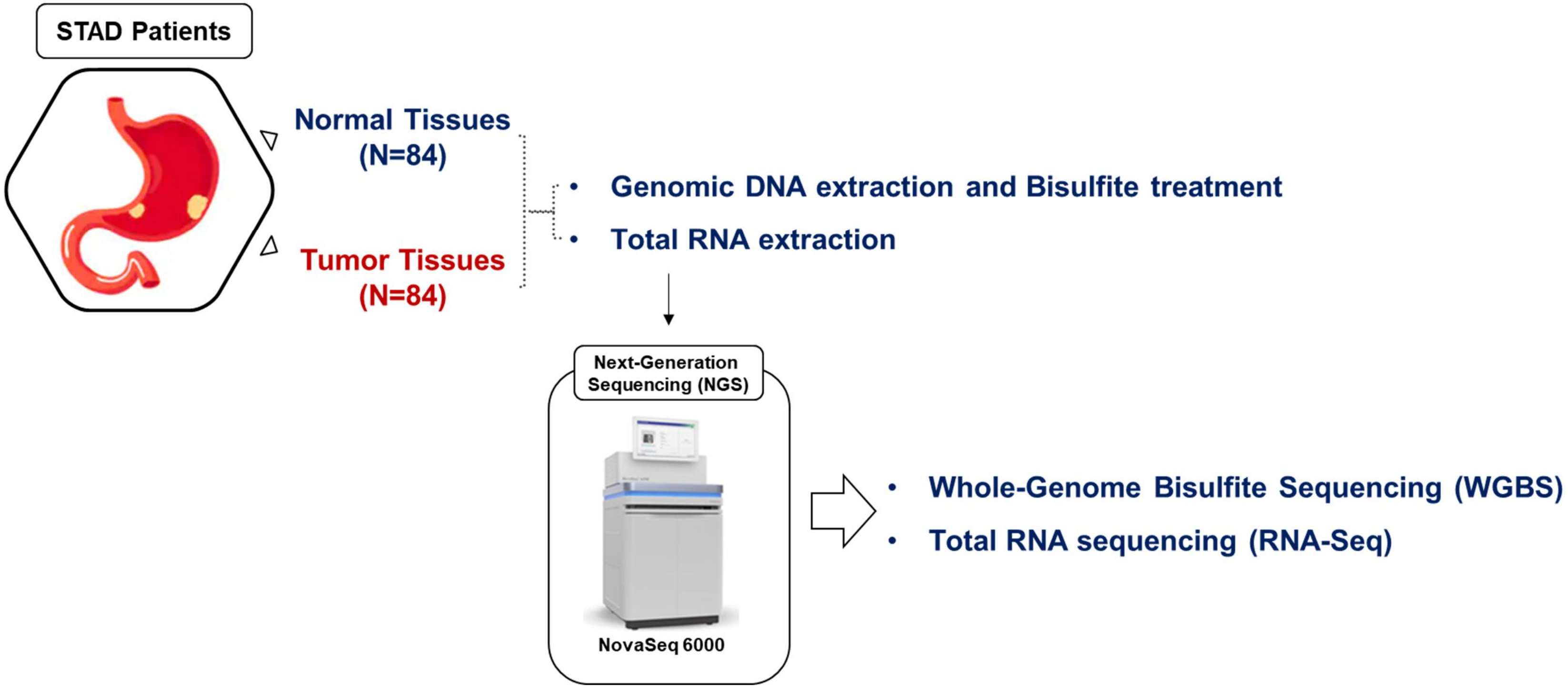

STAD Patients
Normal Tissues (N=84)
Tumor Tissues (N=84)
Genomic DNA extraction and Bisulfite treatment
Total RNA extraction
Next-Generation Sequencing (NGS)
NovaSeq 6000
Whole-Genome Bisulfite Sequencing (WGBS)
Total RNA sequencing (RNA-Seq)

**Figure 2-1. A summary of the study design.** Total of 168 WGBS and 168 RNA-seq (84 each for tumor tissues and normal tissues) were performed.

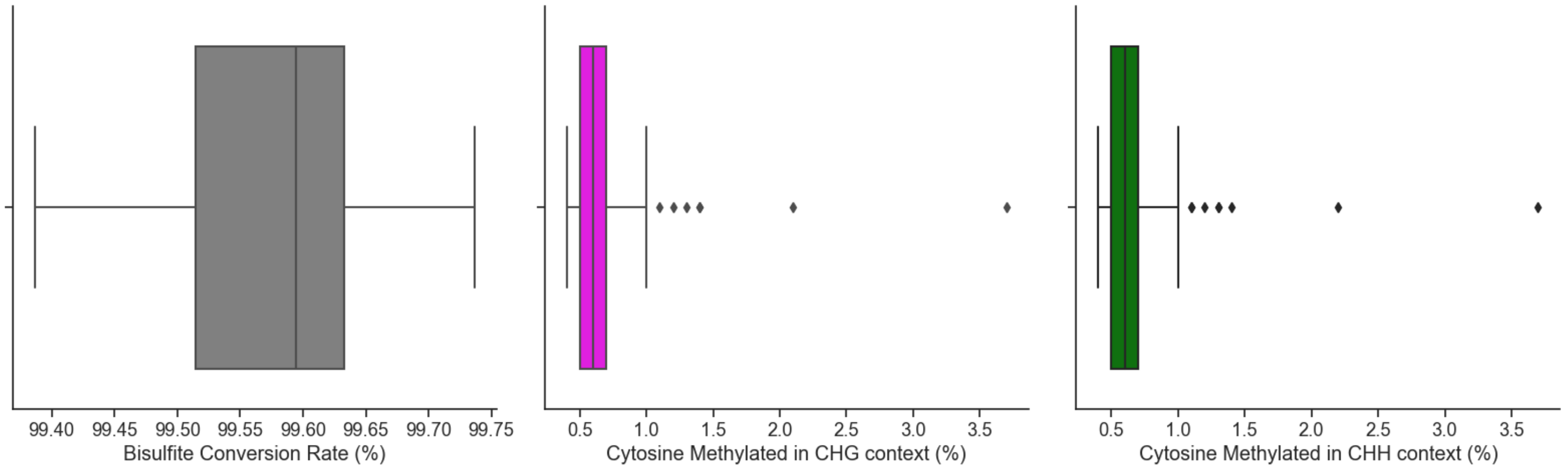


**Figure 2-2. Quality metrics of WGBS samples.** Boxplot representation of percentage of bisulfite conversion rate inferred from lambda genome spike-in (left), cytosine methylated in CHG context (middle) and CHH context (right).

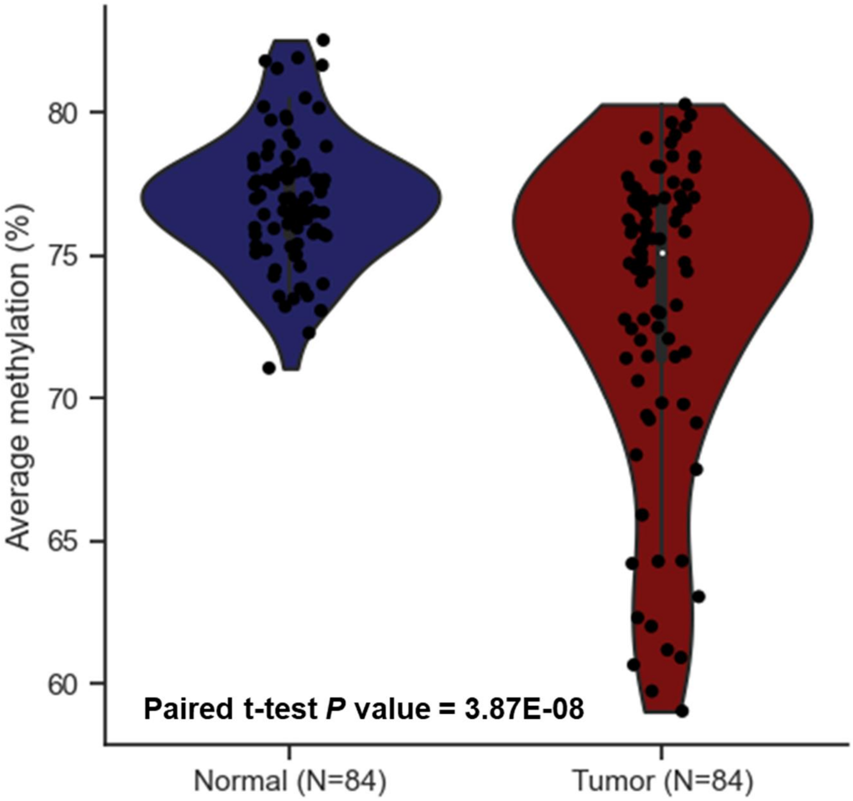


**Figure 2-3. Global hypomethylation in STAD.** Comparison of average CpG DNA methylation in each sample show that tumor samples exhibit lower DNA methylation profile compared to normal counterparts.

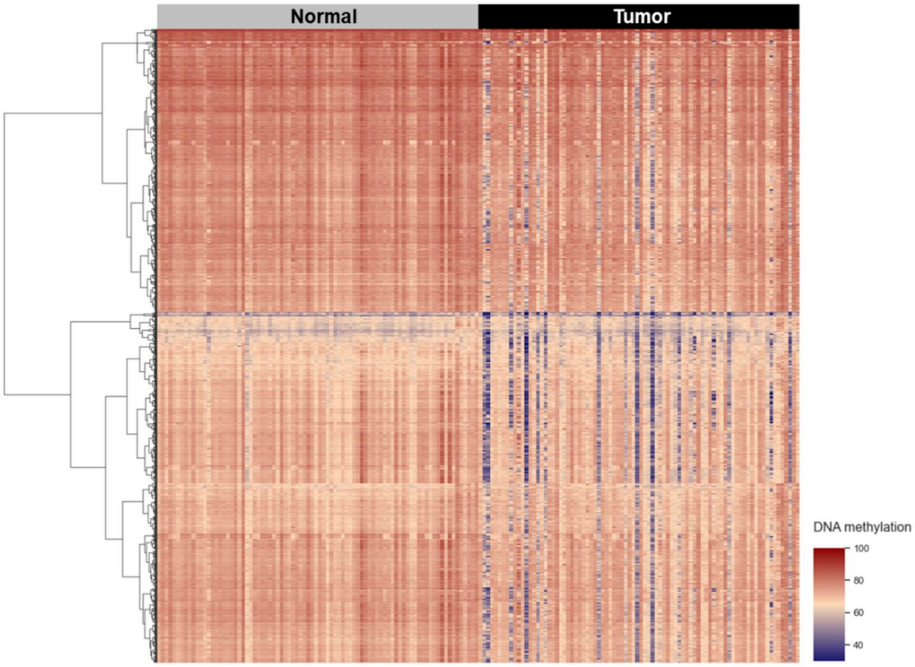


**Figure 2-4. Hierarchical clustering of 2,282 union PMD.** Highly disorganized methylation levels in terms of PMD were detected recurrently in tumor tissues.

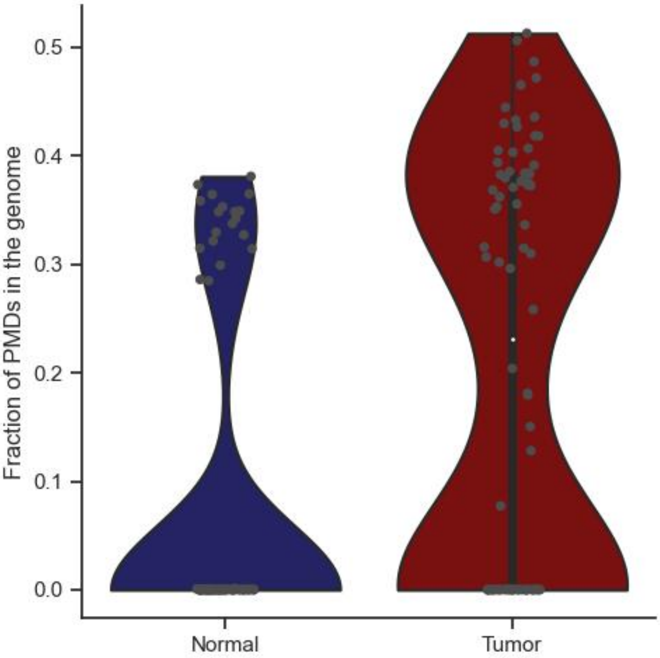


**Figure 2-5. Fraction of PMDs in the genome for normal tissues and tumor tissues.**

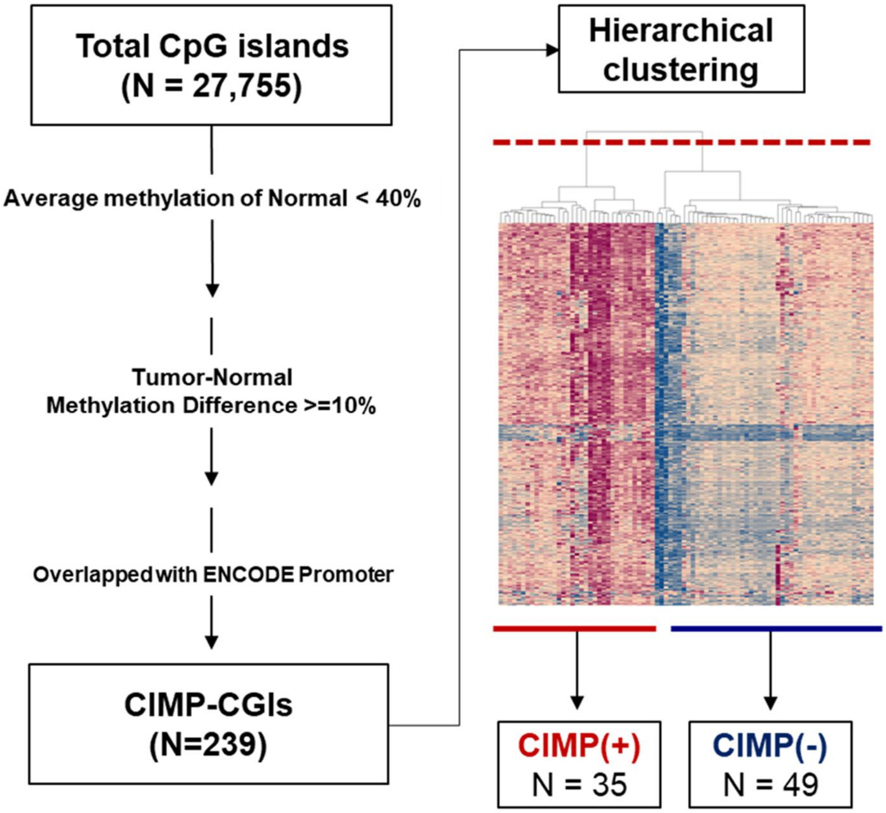


**Figure 2-6. CIMP-CGI detection and analysis.**

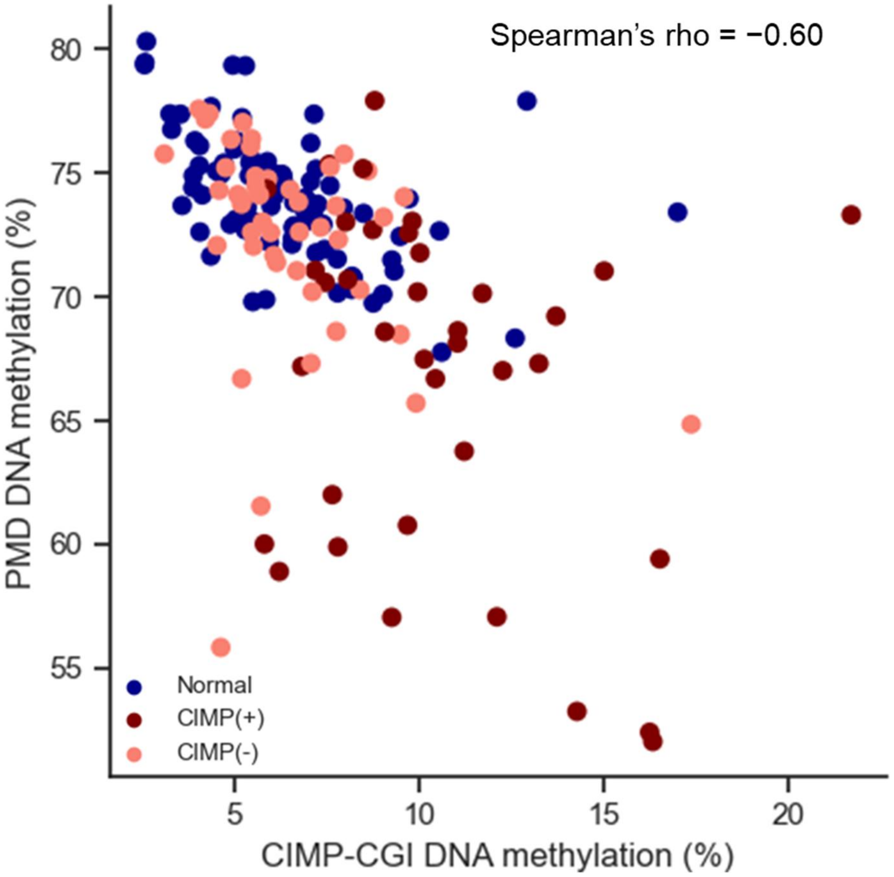


**Figure 2-7. Correlation of CIMP-CGI DNA methylation with PMD DNA methylation.** Each dot represents a sample (Blue; Normal samples, Red; CIMP-positive samples and Salmon; CIMP-negative samples)

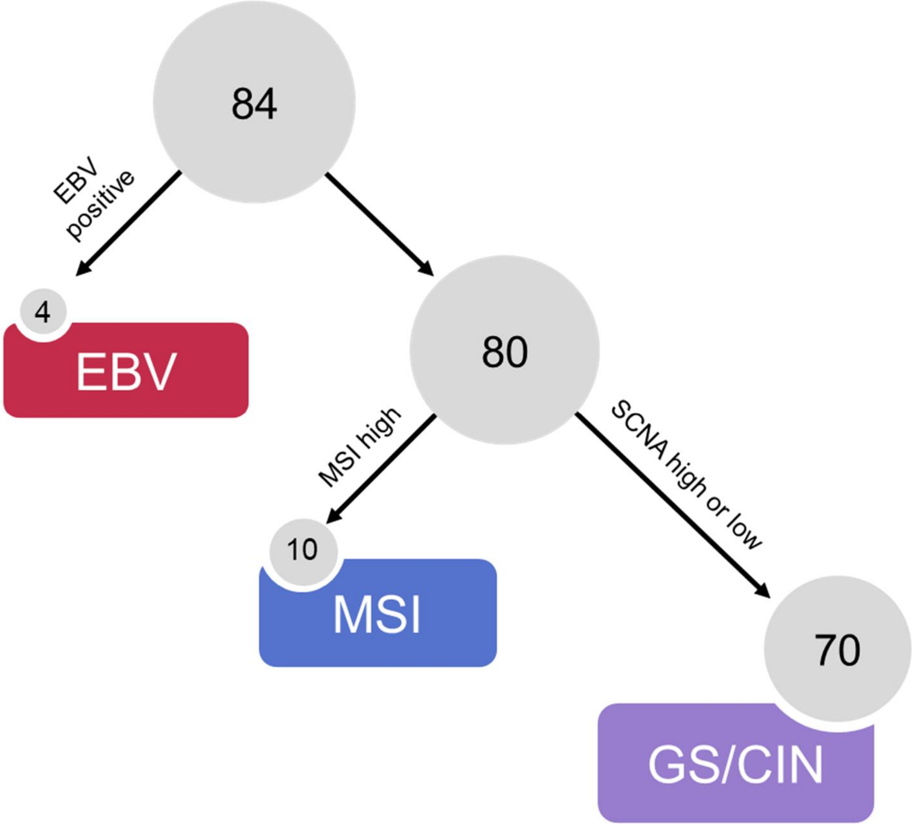


**Figure 2-8. The schematic flowchart outlining how 84 STAD tumors were classified into TCGA molecular subtypes.**

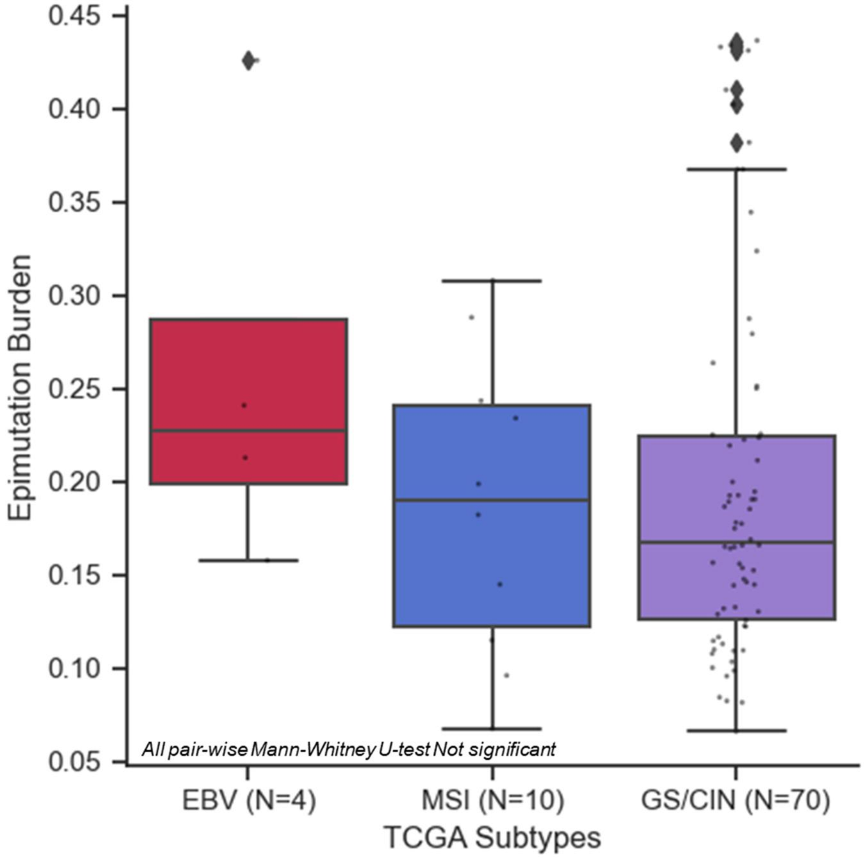


**Figure 2-9. Distribution of epimutation burden among TCGA STAD subtypes.** No significant difference of epimutation burden was observed for all pair-wise comparison. Each dot represents the individual tumor sample corresponding to TCGA subtypes assigned.

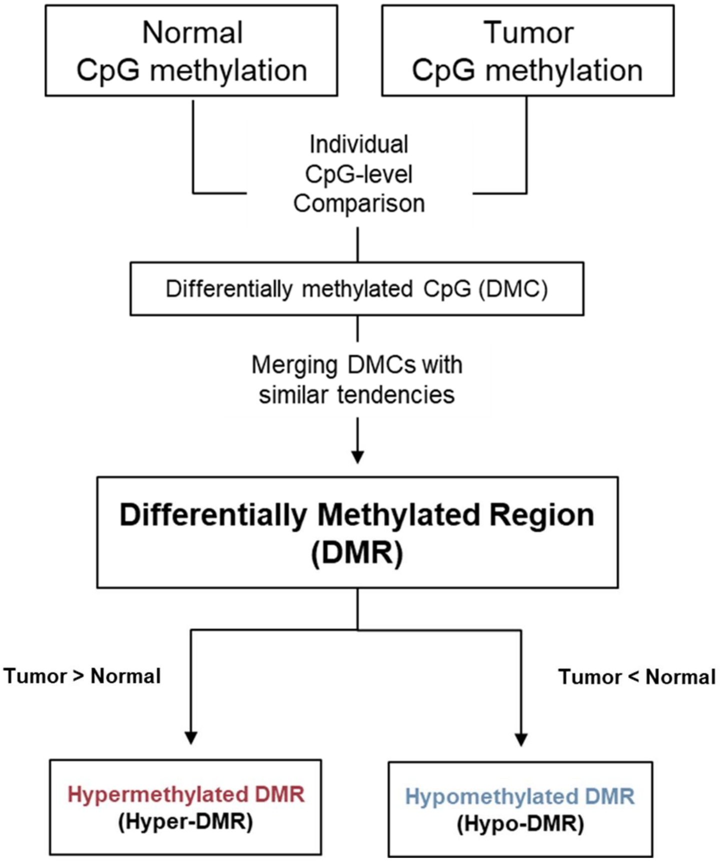


**Figure 2-10. A flowchart depicting identification of differentially methylated regions (DMR) between tumor tissues and normal counterparts.**

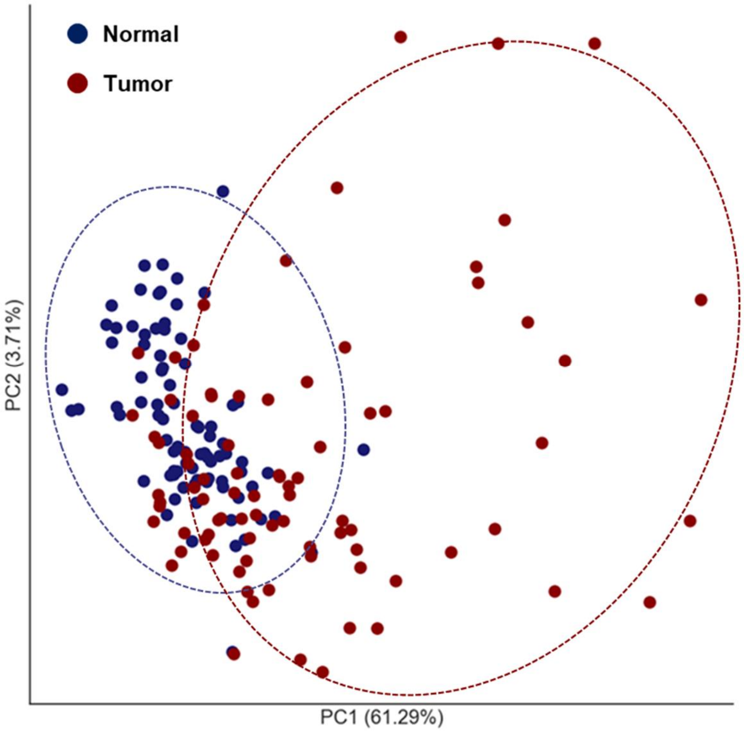


**Figure 2-11. Principal component analysis using DNA methylation inside DMR for each sample.** Normal and malignant tissues can be distinguished by DNA methylation inside the DMR (absolute DNA methylation difference between Tumor and Normal > 10%).

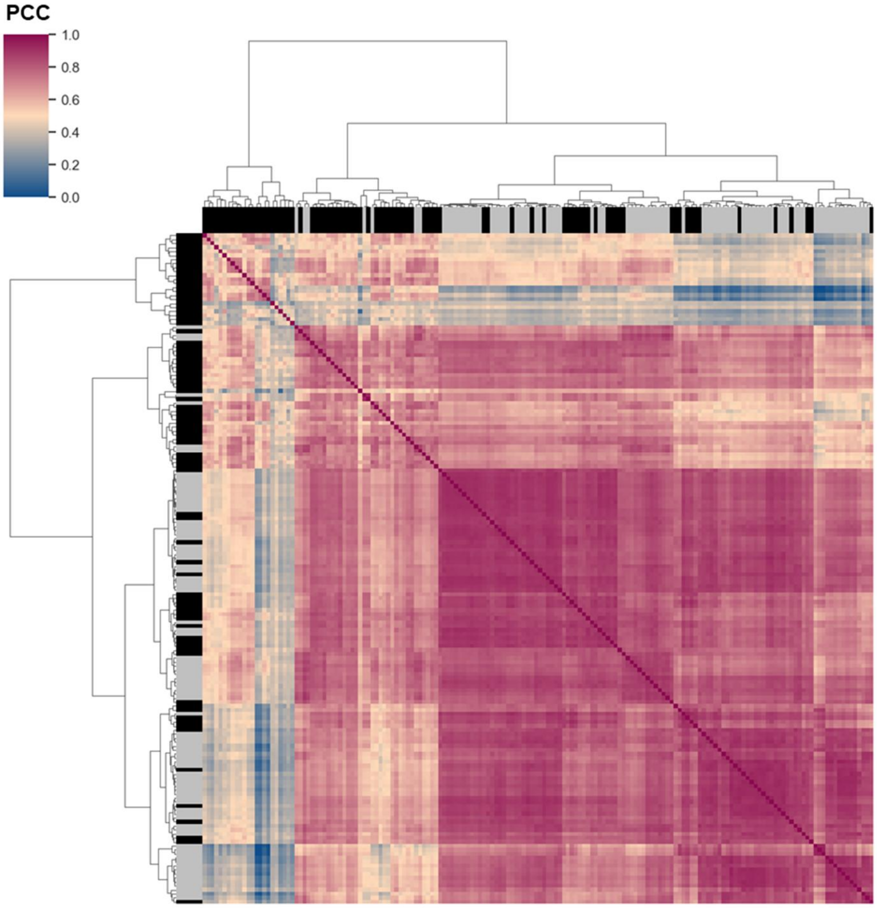


**Figure 2-12. Hierarchical clustering sample-to-sample distance by DNA methylation within DMR using Pearson correlation coefficients.** Black; Tumor samples, Grey; Normal samples. PCC; Pearson correlation coefficient.

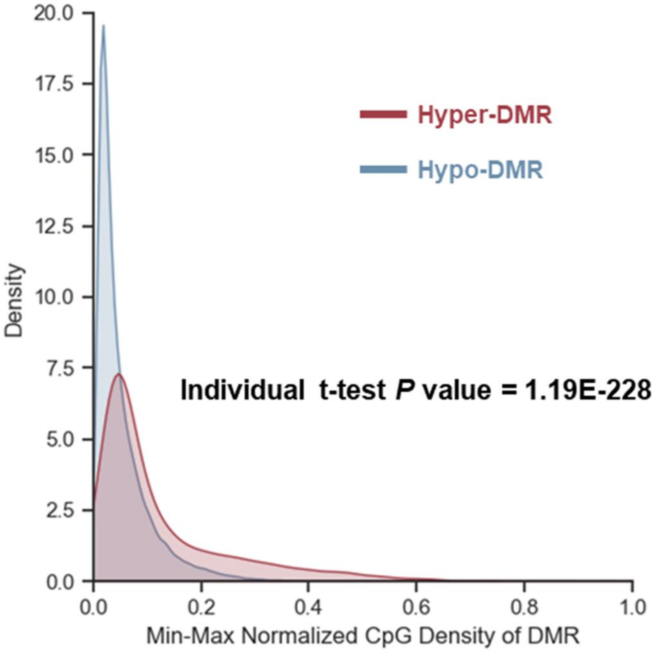


**Figure 2-13. Distribution of normalized CpG density between hyper-DMR and hypo-DMR.**

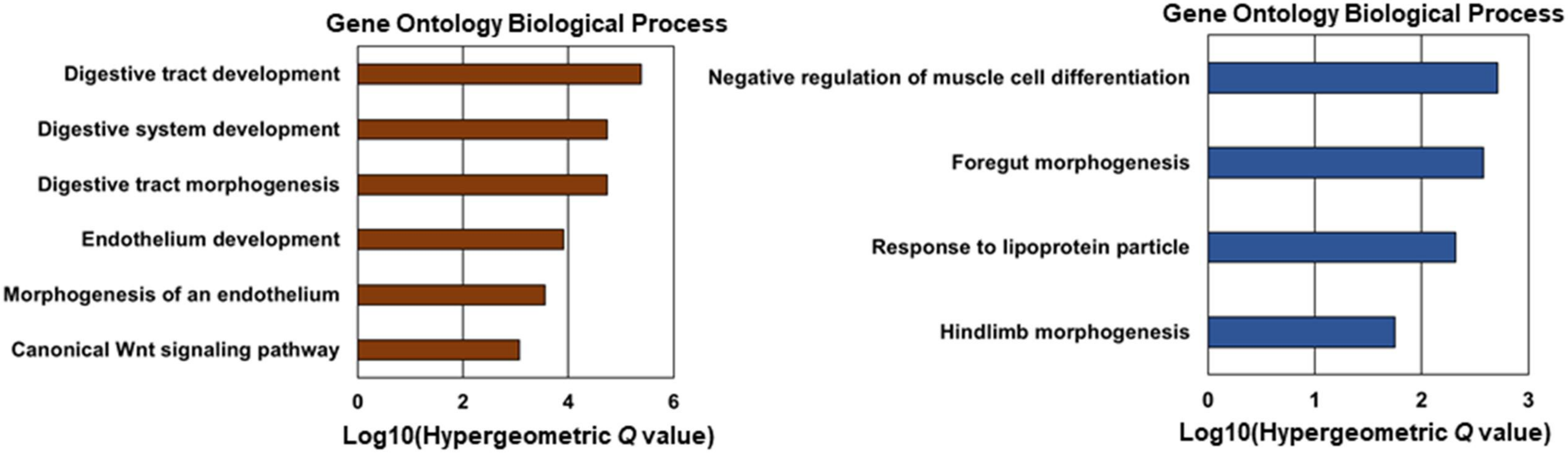


**Figure 2-14. Functional enrichment of hyper-DMR and hypo-DMR using Gene ontology Biological Process.**

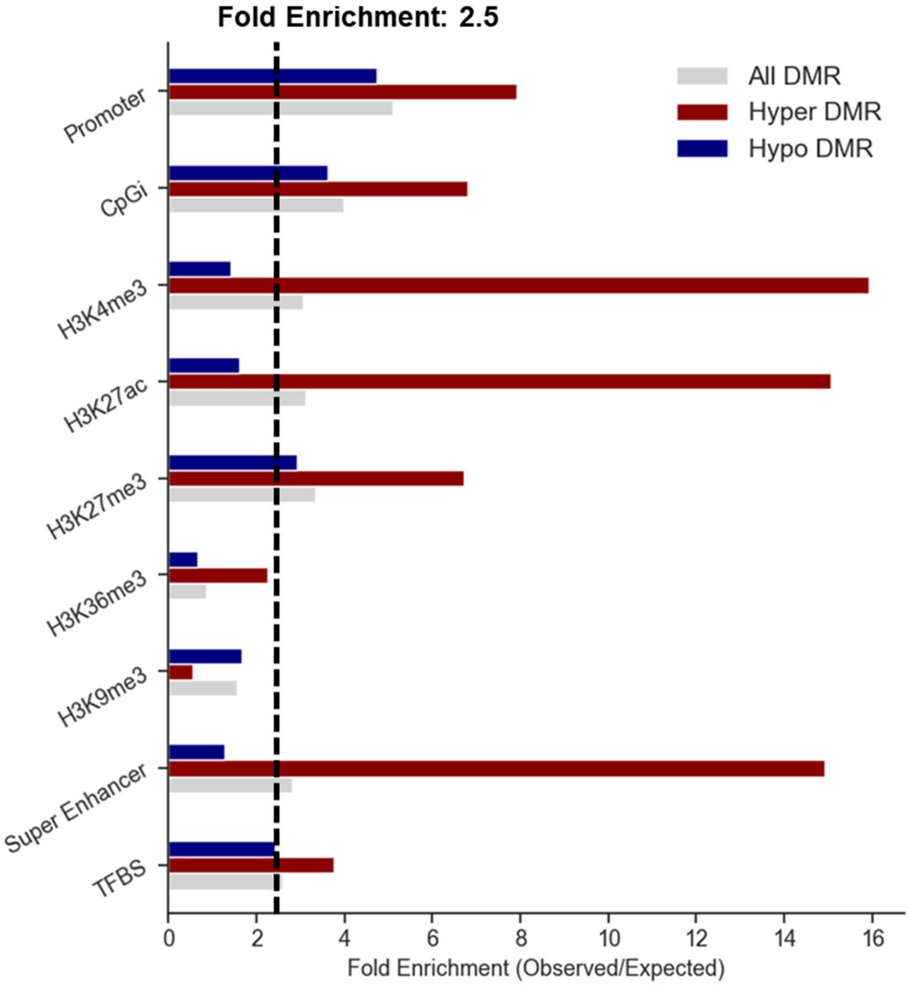


**Figure 2-15. Fold enrichment of the chosen regulatory elements within DMR.**

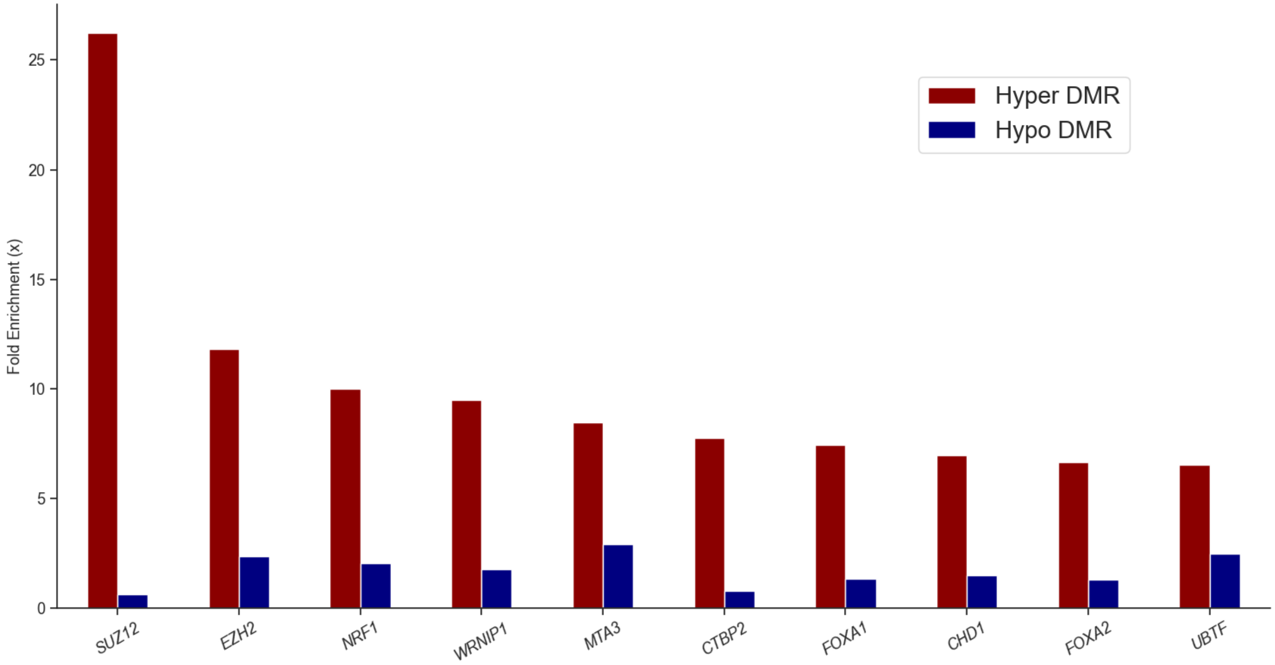


**Figure 2-16. Fold enrichment of the top 10 most enriched transcription factor binding sites within hyper-DMR.**

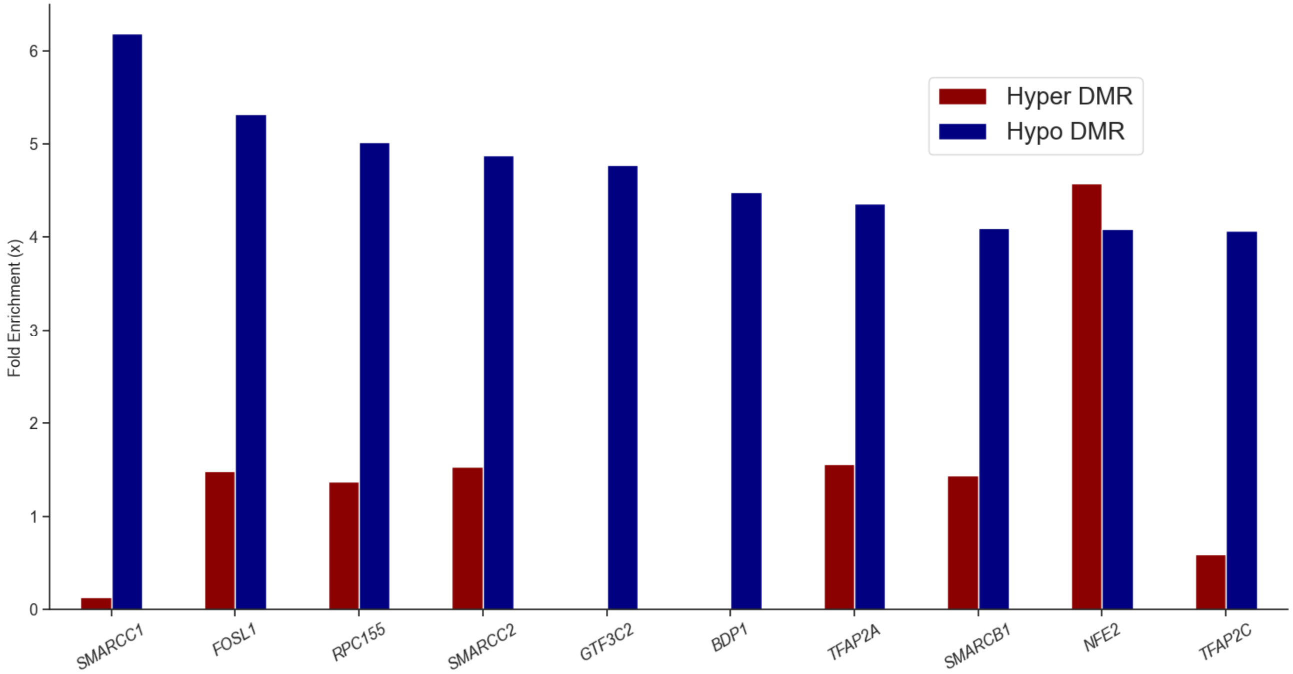


**Figure 2-17. Fold enrichment of the top 10 most enriched transcription factor binding sites within hypo-DMR.**

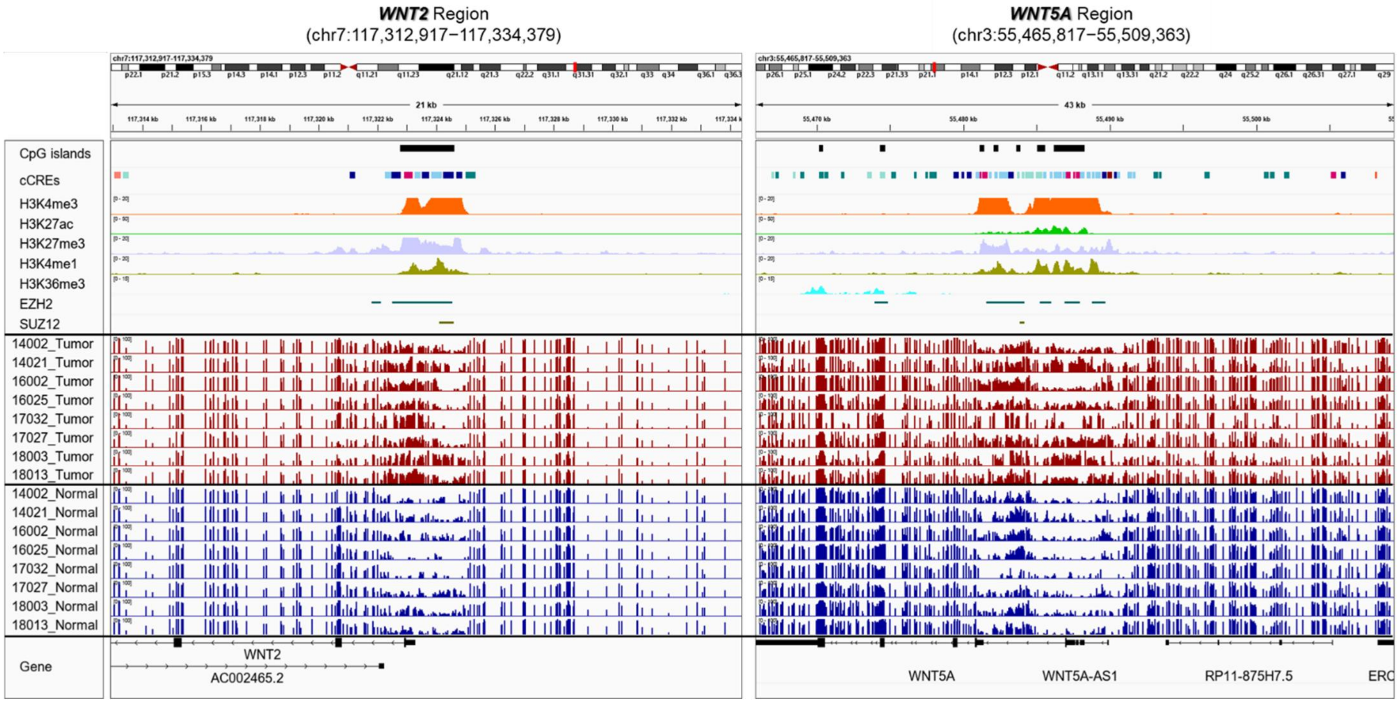


**Figure 2-18. Promoter hypermethylation of *WNT2* and *WNT5A*.** Integrative genomic viewer's view of DNA methylation and other regulatory elements around *WNT2* and *WNT5A* respectively. Representative 8 samples each with tumor and normal counterparts were used as an example to show promoter hypermethylation (red bars: tumor DNA methylation, blue bars; normal DNA methylation)

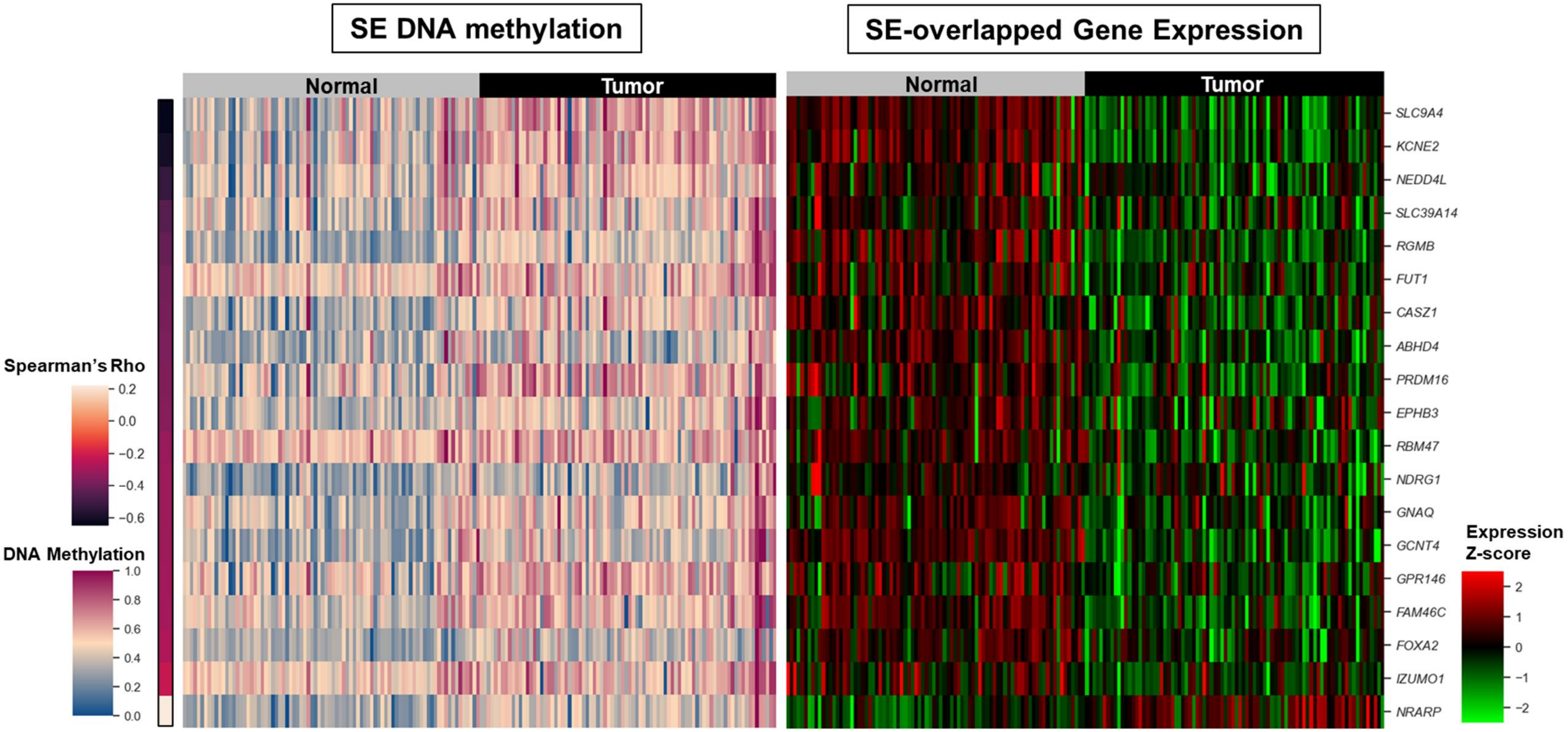


**Figure 2-19. Heatmap depicting super enhancer DNA methylation profiles across samples and respective gene expression dynamics.**

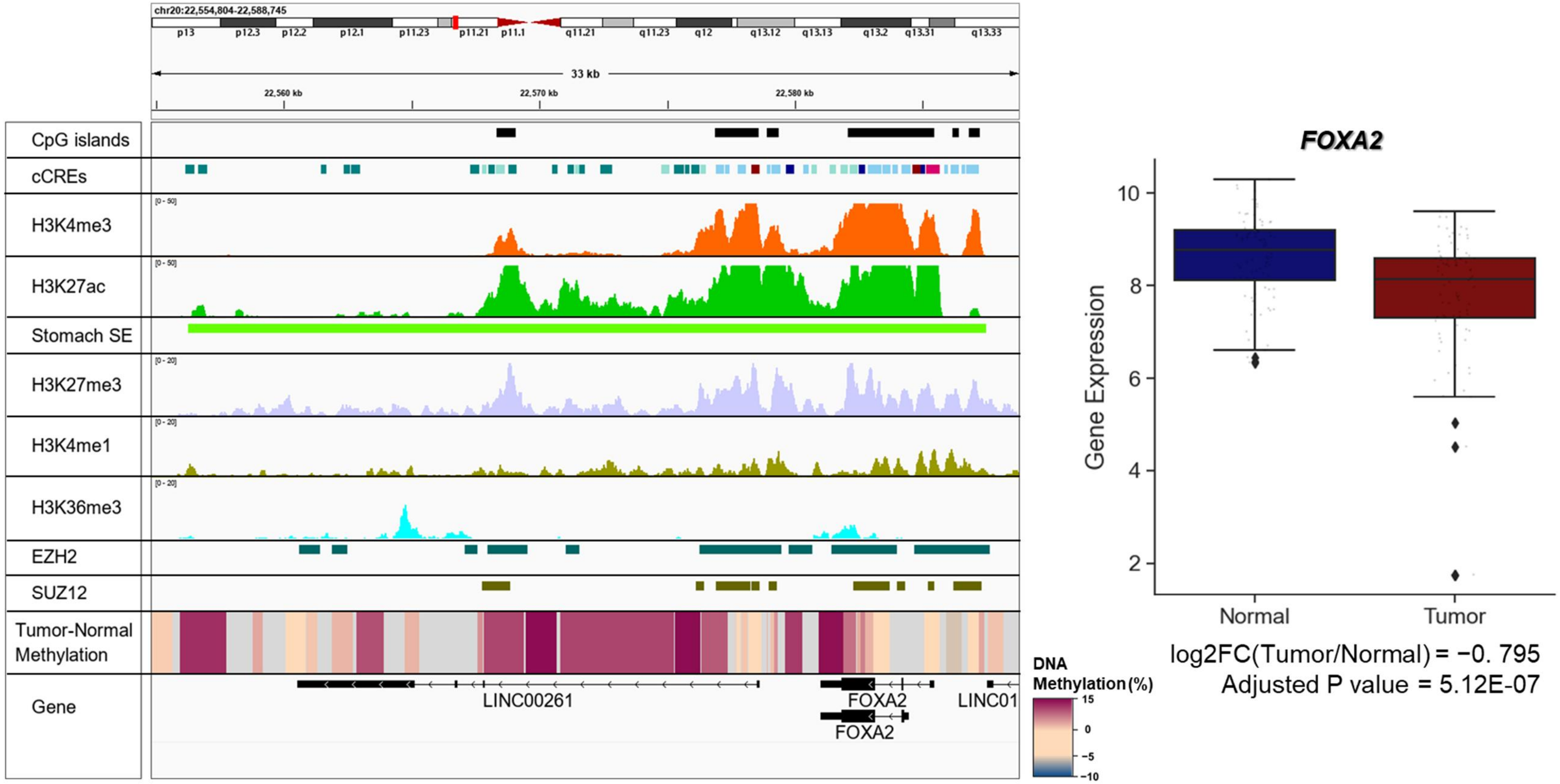


**Figure 2-20. A long stretches of DNA hypermethylation in SE element related to *FOXA2*.**

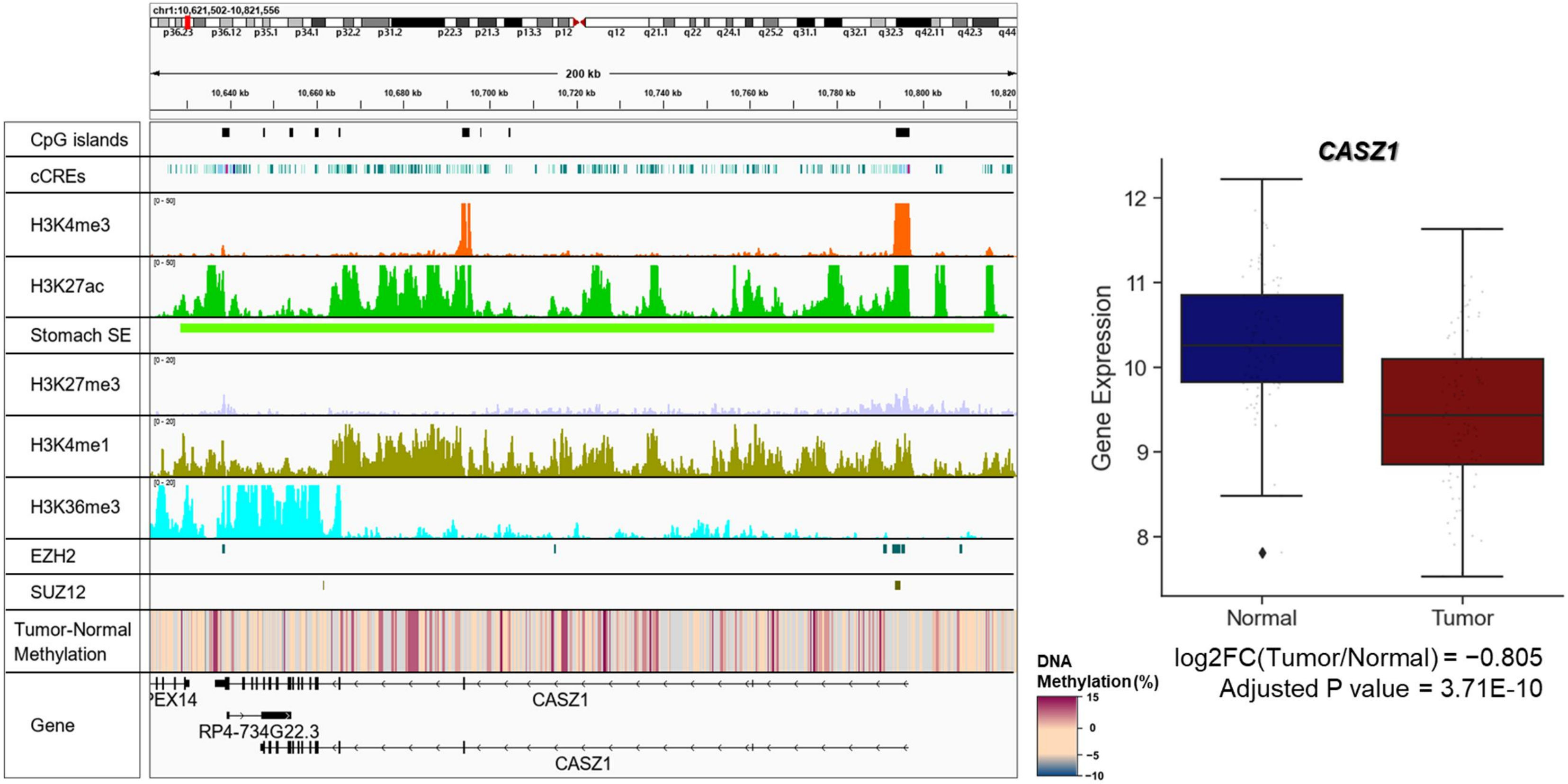


**Figure 2-21. A long stretches of DNA hypermethylation in SE element related to *CASZ1*.**

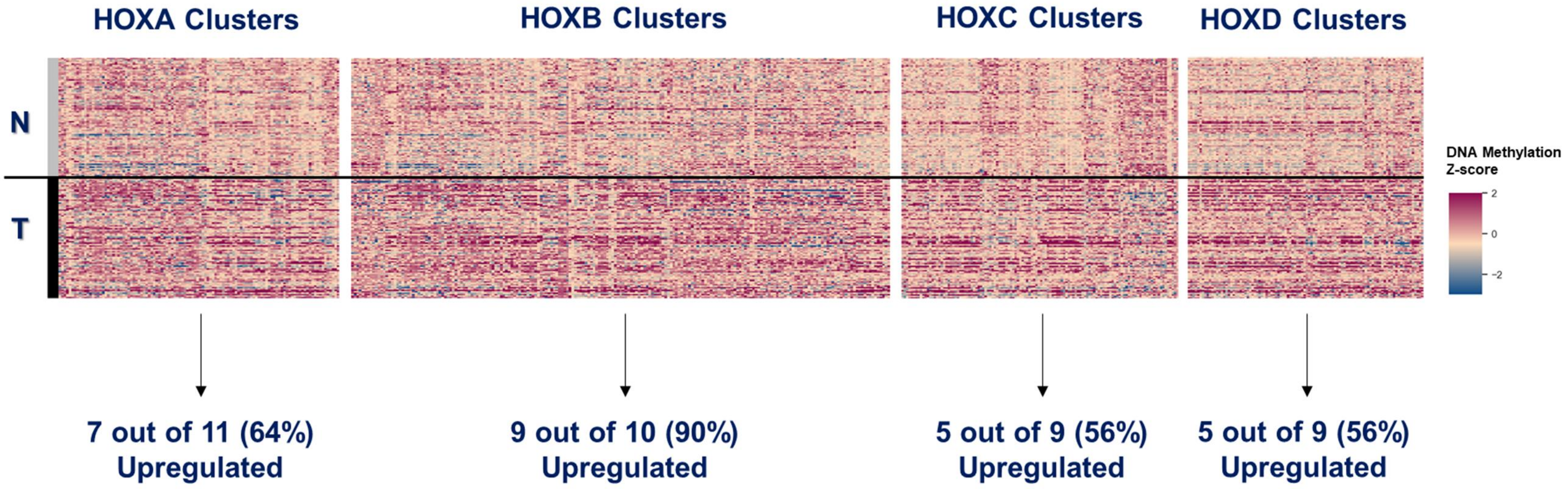


**Figure 2-22. Heatmaps showing hypermethylation of the 4 homeobox clusters.** Most respective genes in the hypermethylated homeobox clusters were upregulated upon DNA hypermethylation. A color bar on the right-hand side represents a percentage of DNA methylation.

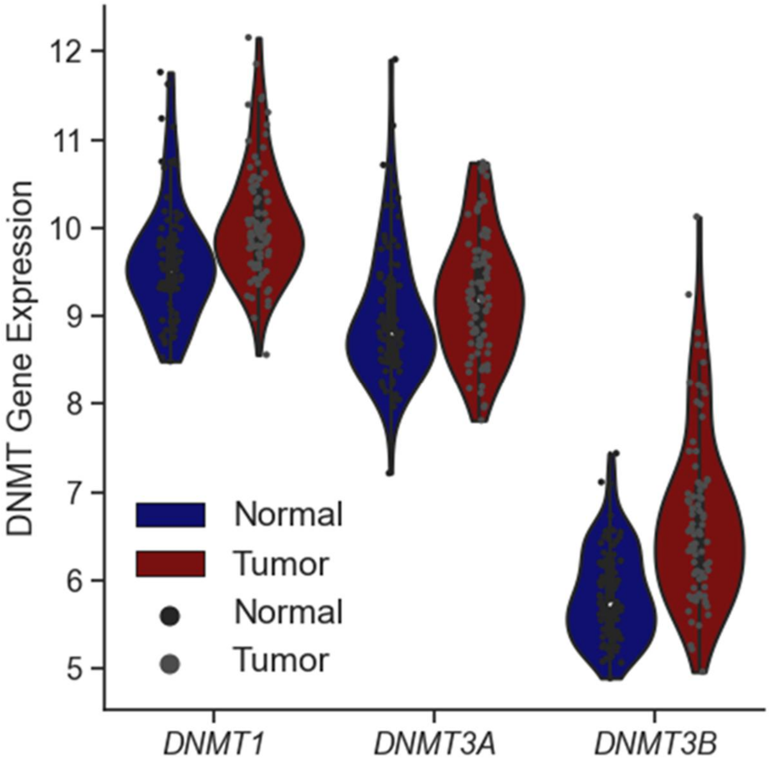


**Figure 2-23. Comparison of the gene expression of DNMTs between STAD and normal tissues.** Maintenance (*DNMT1*) and de novo (*DNMT1*, *DNMT3A* and *DNMT3B*) methyltransferase genes were all upregulated in STAD compared to in normal tissues.

# Discussion

Understanding of complex and dynamic molecular process in the progression of STAD in terms of genetic, epigenetic and transcriptomic regulations has been delineated in numerous studies. However, the DNA methylation landscape of STAD manifestation, despite its defining attribute of regulating cellular identity in formation of the cancer [8], remained obscure.

Here, we performed WGBS, RNA-seq on 84 STAD samples and their matched normal samples to elucidate how the epigenetic landscape contributes to STAD tumorigenesis. We observed global DNA hypomethylation and local hypermethylation across the epigenomes of STAD, consistent with studies in other types of cancers [99, 107, 108, 118]. Interestingly, in our STAD cohorts, three DNA methyltransferase (DNMT) genes, coding enzymes that catalyze DNA methylation of CpG sites (*DNMT1*, *DNMT3A* and *DNMT3B*), were all upregulated (**Figure 2-23**). The activation of DNMT genes is in fact implicated in several other cancers [128]. These observations seem paradoxical, given DNMTs' ability to methylate cytosines, because our STAD cohorts exhibited global hypomethylation, that is, demethylation. This phenomenon may be present in part due to the fact that tumors have infinite replicative potentials [5, 6, 129] causing passive DNA demethylation upon replication and consequent activation of compensational expression of DNMT program. Furthermore, it hints at the possibility of selective exertion of DNMT on several tumor-associated gene promoter that brings their hypermethylation considering that there have been a few evident studies that DNMTs are recruited by transcriptional regulators [130].

Moreover, with multi-omics analysis approach, we observed a highly methylated promoter region and upregulated expression of *WNT2* constituting canonical WNT/β-catenin/MMP signaling. Interestingly, in the study of esophageal squamous cell carcinoma [118], they found out that *WNT2* promoter hypermethylation and upregulation of its expression and then, they experimentally validated that decreased binding of EZH2 to the hypermethylated promoter region of *WNT2* was associated with higher expression of *WNT2* in cancer compared to that in normal counterpart. Our study also found that this process was applicable to *WNT5A* re-activation. In fact, upon querying TFBS within DNA methylation-affected gene promoters, we observed that binding sites for EZH2 and SUZ12 proteins, subunits of PRC2 were present in hypermethylated *WNT2* and *WNT5A* promoters respectively. This suggests that higher expression of *WNT2* and *WNT5A* upon dysregulation of respective promoter methylation can be putative biomarkers for STAD progression.

Furthermore, super enhancer hypermethylation and its nearby gene dysregulation was evident in our STAD cohorts. We proposed that local changes in transcription factor binding acted on DNA methylation profiles of super enhancer element with subsequent effects on target gene dysregulation. Of note, downregulation of *FOXA2*, and *CASZ1* which individually play a role as a tumor-suppressing factor, was associated with nearby super-enhancer epigenetic dysregulation. We also found out that the hypermethylation of homeobox gene clusters was related to their downregulation. Given that binding signatures of polycomb group proteins, a set of transcriptional repressors that recognize and bind to H3K27me3 repressive marks and best known for restricting homeobox gene expression [131], were enriched in these areas, we believe that targeted DNA methylation assaying approach in homeobox clusters could be a potential STAD detection method. Indeed, homeobox genes have been shown to play their parts in oncogenesis [132].

Overall, our study on STAD using DNA methylation and RNA expression provides a roadmap for delineating the functional roles of epigenetic dysregulation. Owing to the bulk nature of our samples, we believe that further single-cell level analysis would provide a better understanding of the impact on epigenetic dysregulation in STAD.

# General Discussion

In the first part of the thesis, we characterized molecular profiles of HGSOC using multi-omics data. Investigation of genomic and transcriptomic landscapes of the HGSOC demonstrated that alteration of genome and epithelial-to-mesenchymal transition (EMT) play an important role in our cancer cohorts and that they can be divided into two distinct molecular subtypes; homologous recombination repair (HRR)-activated type and mesenchymal type.

In the second part of the thesis, we used integrative, high-dimensional multi-omics approaches to outline the DNA methylome landscape and describe the putative oncogenic drivers of STAD using whole-genome bisulfite sequencing (WGBS) and RNA-seq. Altered DNA methylation were associated with cancer-specific gene dysregulation including canonical WNT/β-catenin/MMP signaling, super enhancer related genes such as *FOXA2* and *CASZ1* and genes for four homeobox clusters. We showed that epigenetic dysregulation promotes STAD tumorigenesis through multi-factorial mechanisms.

In summary, these studies advances our understanding of how multi-faceted molecular landscapes shape cancer pathogenesis and provides a resource for biomarker and target discovery.

# References

1. Sohn, M.H., et al., *Classification of High-Grade Serous Ovarian Carcinoma by Epithelial-to-Mesenchymal Transition Signature and Homologous Recombination Repair Genes.* Genes (Basel), 2021. **12**(7).
2. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.
3. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
4. Kircher, M. and J. Kelso, *High-throughput DNA sequencing--concepts and limitations.* Bioessays, 2010. **32**(6): p. 524-36.
5. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation.* Cell, 2011. **144**(5): p. 646-74.
6. Hanahan, D., *Hallmarks of Cancer: New Dimensions.* Cancer Discov, 2022. **12**(1): p. 31-46.
7. Yuan, S., R.J. Norgard, and B.Z. Stanger, *Cellular Plasticity in Cancer.* Cancer Discov, 2019. **9**(7): p. 837-851.
8. Feinberg, A.P., *The Key Role of Epigenetics in Human Disease Prevention and Mitigation.* N Engl J Med, 2018. **378**(14): p. 1323-1334.
9. Weinberg, R.A., *The biology of cancer*. Second edition. ed. 2014, New York: Garland Science, Taylor & Francis Group. xx, 876, A 6, G 30, I 28 pages.
10. Hinck, L. and I. Nathke, *Changes in cell and tissue organization in cancer of the breast and colon.* Curr Opin Cell Biol, 2014. **26**: p. 87-95.
11. Ayob, A.Z. and T.S. Ramasamy, *Cancer stem cells as key drivers of*

*tumour progression.* J Biomed Sci, 2018. **25**(1): p. 20.

12. Berman, J.J., *Tumor taxonomy for the developmental lineage classification of neoplasms.* BMC Cancer, 2004. **4**: p. 88.
13. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
14. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
15. Cho, K.R. and M. Shih Ie, *Ovarian cancer.* Annu Rev Pathol, 2009. **4**: p. 287-313.
16. Kuchenbaecker, K.B., et al., *Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers.* Jama, 2017. **317**(23): p. 2402-2416.
17. Moore, K., et al., *Maintenance Olaparib in Patients with Newly Diagnosed Advanced Ovarian Cancer.* N Engl J Med, 2018. **379**(26): p. 2495-2505.
18. González-Martín, A., et al., *Niraparib in Patients with Newly Diagnosed Advanced Ovarian Cancer.* N Engl J Med, 2019. **381**(25): p. 2391-2402.
19. Pujade-Lauraine, E., et al., *Olaparib tablets as maintenance therapy in patients with platinum-sensitive, relapsed ovarian cancer and a BRCA1/2 mutation (SOLO2/ENGOT-Ov21): a double-blind, randomised, placebo-controlled, phase 3 trial.* Lancet Oncol, 2017. **18**(9): p. 1274-1284.
20. Del Campo, J.M., et al., *Niraparib Maintenance Therapy in Patients With Recurrent Ovarian Cancer After a Partial Response to the Last Platinum-Based Chemotherapy in the ENGOT-OV16/NOVA Trial.* J Clin Oncol, 2019. **37**(32): p. 2968-2973.
21. Coleman, R.L., et al., *Bevacizumab and paclitaxel-carboplatin chemotherapy and secondary cytoreduction in recurrent, platinum-sensitive ovarian cancer (NRG Oncology/Gynecologic Oncology Group study GOG-0213): a multicentre, open-label, randomised, phase 3 trial.* Lancet Oncol, 2017. **18**(6): p. 779-791.

22. Vaughan, S., et al., *Rethinking ovarian cancer: recommendations for improving outcomes.* Nat Rev Cancer, 2011. **11**(10): p. 719-25.

23. Vergara, D., et al., *Epithelial-mesenchymal transition in ovarian cancer.* Cancer Lett, 2010. **291**(1): p. 59-66.

24. Dongre, A. and R.A. Weinberg, *New insights into the mechanisms of epithelial-mesenchymal transition and implications for cancer.* Nat Rev Mol Cell Biol, 2019. **20**(2): p. 69-84.

25. Sengodan, S.K., et al., *Regulation of epithelial to mesenchymal transition by BRCA1 in breast cancer.* Crit Rev Oncol Hematol, 2018. **123**: p. 74-82.

26. Loret, N., et al., *The Role of Epithelial-to-Mesenchymal Plasticity in Ovarian Cancer Progression and Therapy Resistance.* Cancers (Basel), 2019. **11**(6).

27. Kim, S.I., et al., *Effect of BRCA mutational status on survival outcome in advanced-stage high-grade serous ovarian cancer.* J Ovarian Res, 2019. **12**(1): p. 40.

28. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

29. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.

30. Kim, S., et al., *Strelka2: fast and accurate calling of germline and somatic variants.* Nat Methods, 2018. **15**(8): p. 591-594.

31. Ramos, A.H., et al., *Oncotator: cancer variant annotation tool.* Hum Mutat, 2015. **36**(4): p. E2423-9.

32. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.

33. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.

34. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

35. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706*

*humans.* Nature, 2016. **536**(7616): p. 285-91.

36. Yoo, S.K., et al., *NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants.* Genome Med, 2019. **11**(1): p. 64.

37. Tate, J.G., et al., *COSMIC: the Catalogue Of Somatic Mutations In Cancer.* Nucleic Acids Res, 2019. **47**(D1): p. D941-D947.

38. Talevich, E., et al., *CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing.* PLoS Comput Biol, 2016. **12**(4): p. e1004873.

39. Mermel, C.H., et al., *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol, 2011. **12**(4): p. R41.

40. Favero, F., et al., *Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data.* Ann Oncol, 2015. **26**(1): p. 64-70.

41. Sztupinszki, Z., et al., *Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer.* NPJ Breast Cancer, 2018. **4**: p. 16.

42. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.

43. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

44. Russo, P.S.T., et al., *CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses.* BMC Bioinformatics, 2018. **19**(1): p. 56.

45. de Hoon, M.J., et al., *Open source clustering software.* Bioinformatics, 2004. **20**(9): p. 1453-4.

46. Saldanha, A.J., *Java Treeview--extensible visualization of microarray data.* Bioinformatics, 2004. **20**(17): p. 3246-8.

47. Keenan, A.B., et al., *ChEA3: transcription factor enrichment analysis by orthogonal omics integration.* Nucleic Acids Res, 2019. **47**(W1): p. W212-W224.

48. Lachmann, A., et al., *Massive mining of publicly available RNA-seq*

*data from human and mouse.* Nat Commun, 2018. **9**(1): p. 1366.

49. Yang, J., et al., *Guidelines and definitions for research on epithelial-mesenchymal transition.* Nat Rev Mol Cell Biol, 2020. **21**(6): p. 341-352.

50. Cristescu, R., et al., *Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes.* Nat Med, 2015. **21**(5): p. 449-56.

51. Loboda, A., et al., *EMT is the dominant program in human colon cancer.* BMC Med Genomics, 2011. **4**: p. 9.

52. Guo, C.C., et al., *Dysregulation of EMT Drives the Progression to Clinically Aggressive Sarcomatoid Bladder Cancer.* Cell Rep, 2019. **27**(6): p. 1781-1793 e4.

53. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.

54. Aran, D., Z. Hu, and A.J. Butte, *xCell: digitally portraying the tissue cellular heterogeneity landscape.* Genome Biol, 2017. **18**(1): p. 220.

55. Chung, V.Y., et al., *GRHL2-miR-200-ZEB1 maintains the epithelial status of ovarian cancer through transcriptional regulation and histone modification.* Sci Rep, 2016. **6**: p. 19943.

56. Werner, S., et al., *Dual roles of the transcription factor grainyhead-like 2 (GRHL2) in breast cancer.* J Biol Chem, 2013. **288**(32): p. 22993-3008.

57. Chung, V.Y., et al., *The role of GRHL2 and epigenetic remodeling in epithelial-mesenchymal plasticity in ovarian cancer cells.* Commun Biol, 2019. **2**: p. 272.

58. Kalluri, R. and R.A. Weinberg, *The basics of epithelial-mesenchymal transition.* J Clin Invest, 2009. **119**(6): p. 1420-8.

59. Zeisberg, M. and E.G. Neilson, *Biomarkers for epithelial-mesenchymal transitions.* J Clin Invest, 2009. **119**(6): p. 1429-37.

60. Katsuno, Y., S. Lamouille, and R. Derynck, *TGF-beta signaling and epithelial-mesenchymal transition in cancer progression.* Curr Opin Oncol, 2013. **25**(1): p. 76-84.

61. Wang, D.Y., et al., *Identification of estrogen-responsive genes by complementary deoxyribonucleic acid microarray and*

*characterization of a novel early estrogen-induced gene: EEIG1.* Mol Endocrinol, 2004. **18**(2): p. 402-11.

62. Musa, J., et al., *MYBL2 (B-Myb): a central regulator of cell proliferation, cell survival and differentiation involved in tumorigenesis.* Cell Death Dis, 2017. **8**(6): p. e2895.

63. Aran, D., M. Sirota, and A.J. Butte, *Systematic pan-cancer analysis of tumour purity.* Nat Commun, 2015. **6**: p. 8971.

64. Izar, B., et al., *A single-cell landscape of high-grade serous ovarian cancer.* Nat Med, 2020.

65. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.

66. Thiery, J.P., et al., *Epithelial-mesenchymal transitions in development and disease.* Cell, 2009. **139**(5): p. 871-90.

67. Yu, M., et al., *Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition.* Science, 2013. **339**(6119): p. 580-4.

68. Mrozik, K.M., et al., *N-cadherin in cancer metastasis, its emerging role in haematological malignancies and potential as a therapeutic target in cancer.* BMC Cancer, 2018. **18**(1): p. 939.

69. Chakraborty, P., et al., *Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum.* Front Bioeng Biotechnol, 2020. **8**: p. 220.

70. Winterhoff, B., et al., *Molecular classification of high grade endometrioid and clear cell ovarian cancer using TCGA gene expression signatures.* Gynecol Oncol, 2016. **141**(1): p. 95-100.

71. Konecny, G.E., et al., *Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer.* J Natl Cancer Inst, 2014. **106**(10).

72. Attwooll, C., E. Lazzerini Denchi, and K. Helin, *The E2F family: specific functions and overlapping interests.* EMBO J, 2004. **23**(24): p. 4709-16.

73. Cheung, V.G., S.L. Sherman, and E. Feingold, *Genetics. Genetic control of hotspots.* Science, 2010. **327**(5967): p. 791-2.

74. Regnier, V., et al., *CENP-A is required for accurate chromosome*

*segregation and sustained kinetochore association of BubR1.* Mol Cell Biol, 2005. **25**(10): p. 3967-81.

75. Imoto, I., et al., *Amplification and overexpression of TGIF2, a novel homeobox gene of the TALE superclass, in ovarian cancer cell lines.* Biochem Biophys Res Commun, 2000. **276**(1): p. 264-70.

76. Sanchez-Tillo, E., et al., *beta-catenin/TCF4 complex induces the epithelial-to-mesenchymal transition (EMT)-activator ZEB1 to regulate tumor invasiveness.* Proc Natl Acad Sci U S A, 2011. **108**(48): p. 19204-9.

77. Skovierova, H., et al., *Molecular regulation of epithelial-to-mesenchymal transition in tumorigenesis (Review).* Int J Mol Med, 2018. **41**(3): p. 1187-1200.

78. Schwede, M., et al., *The Impact of Stroma Admixture on Molecular Subtypes and Prognostic Gene Signatures in Serous Ovarian Cancer.* Cancer Epidemiol Biomarkers Prev, 2020. **29**(2): p. 509-519.

79. De Craene, B. and G. Berx, *Regulatory networks defining EMT during cancer initiation and progression.* Nat Rev Cancer, 2013. **13**(2): p. 97-110.

80. Morris, J.C., et al., *Phase I study of GC1008 (fresolimumab): a human anti-transforming growth factor-beta (TGFbeta) monoclonal antibody in patients with advanced malignant melanoma or renal cell carcinoma.* PLoS One, 2014. **9**(3): p. e90353.

81. Stevenson, J.P., et al., *Immunological effects of the TGFbeta-blocking antibody GC1008 in malignant pleural mesothelioma patients.* Oncoimmunology, 2013. **2**(8): p. e26218.

82. Formenti, S.C., et al., *Focal Irradiation and Systemic TGFbeta Blockade in Metastatic Breast Cancer.* Clin Cancer Res, 2018. **24**(11): p. 2493-2504.

83. Rafehi, S., et al., *TGFbeta signaling regulates epithelial-mesenchymal plasticity in ovarian cancer ascites-derived spheroids.* Endocr Relat Cancer, 2016. **23**(3): p. 147-59.

84. Newsted, D., et al., *Blockade of TGF-beta signaling with novel synthetic antibodies limits immune exclusion and improves chemotherapy response in metastatic ovarian cancer models.*

Oncoimmunology, 2019. **8**(2): p. e1539613.

85. Tan, P. and K.G. Yeoh, *Genetics and Molecular Pathogenesis of Gastric Adenocarcinoma.* Gastroenterology, 2015. **149**(5): p. 1153-1162 e3.

86. Lauren, P., *The Two Histological Main Types of Gastric Carcinoma: Diffuse and So-Called Intestinal-Type Carcinoma. An Attempt at a Histo-Clinical Classification.* Acta Pathol Microbiol Scand, 1965. **64**: p. 31-49.

87. Nagtegaal, I.D., et al., *The 2019 WHO classification of tumours of the digestive system.* Histopathology, 2020. **76**(2): p. 182-188.

88. Yeoh, K.G. and P. Tan, *Mapping the genomic diaspora of gastric cancer.* Nat Rev Cancer, 2022. **22**(2): p. 71-84.

89. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of gastric adenocarcinoma.* Nature, 2014. **513**(7517): p. 202-9.

90. Consortium, E.P., et al., *Expanded encyclopaedias of DNA elements in the human and mouse genomes.* Nature, 2020. **583**(7818): p. 699-710.

91. Davis, C.A., et al., *The Encyclopedia of DNA elements (ENCODE): data portal update.* Nucleic Acids Res, 2018. **46**(D1): p. D794-D801.

92. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

93. Jiang, Y., et al., *SEdb: a comprehensive human super-enhancer database.* Nucleic Acids Res, 2019. **47**(D1): p. D235-D243.

94. Frankish, A., et al., *GENCODE reference annotation for the human and mouse genomes.* Nucleic Acids Res, 2019. **47**(D1): p. D766-D773.

95. Zhang, Y., G. Parmigiani, and W.E. Johnson, *ComBat-seq: batch effect adjustment for RNA-seq count data.* NAR Genom Bioinform, 2020. **2**(3): p. lqaa078.

96. Zhu, A., J.G. Ibrahim, and M.I. Love, *Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences.* Bioinformatics, 2019. **35**(12): p. 2084-2092.

97. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and*

*methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-2.

98. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

99. Li, J., et al., *A genomic and epigenomic atlas of prostate cancer in Asian populations.* Nature, 2020. **580**(7801): p. 93-99.

100. Juhling, F., et al., *metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data.* Genome Res, 2016. **26**(2): p. 256-62.

101. Hansen, K.D., B. Langmead, and R.A. Irizarry, *BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions.* Genome Biol, 2012. **13**(10): p. R83.

102. Robinson, J.T., et al., *Integrative genomics viewer.* Nat Biotechnol, 2011. **29**(1): p. 24-6.

103. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nat Biotechnol, 2010. **28**(5): p. 495-501.

104. Lee, B.T., et al., *The UCSC Genome Browser database: 2022 update.* Nucleic Acids Res, 2022. **50**(D1): p. D1115-D1122.

105. Song, Q., et al., *A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics.* PLoS One, 2013. **8**(12): p. e81148.

106. Weisenberger, D.J., et al., *CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer.* Nat Genet, 2006. **38**(7): p. 787-93.

107. Ehrlich, M., *DNA hypomethylation in cancer cells.* Epigenomics, 2009. **1**(2): p. 239-59.

108. Salhab, A., et al., *A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains.* Genome Biol, 2018. **19**(1): p. 150.

109. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.

110. Katoh, M., *Canonical and non-canonical WNT signaling in cancer stem cells and their niches: Cellular heterogeneity, omics*

*reprogramming, targeted therapy and tumor plasticity (Review).* Int J Oncol, 2017. **51**(5): p. 1357-1369.

111. Lee, D.S., et al., *An epigenomic roadmap to induced pluripotency reveals DNA methylation as a reprogramming modulator.* Nat Commun, 2014. **5**: p. 5619.

112. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

113. Phelan, M.L., et al., *Reconstitution of a core chromatin remodeling complex from SWI/SNF subunits.* Mol Cell, 1999. **3**(2): p. 247-53.

114. Chen, G., et al., *A heterotrimeric SMARCB1-SMARCC2 subcomplex is required for the assembly and tumor suppression function of the BAF chromatin-remodeling complex.* Cell Discov, 2020. **6**: p. 66.

115. Kowenz-Leutz, E. and A. Leutz, *A C/EBP beta isoform recruits the SWI/SNF complex to activate myeloid genes.* Mol Cell, 1999. **4**(5): p. 735-43.

116. Alver, B.H., et al., *The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers.* Nat Commun, 2017. **8**: p. 14648.

117. Klaus, A. and W. Birchmeier, *Wnt signalling and its impact on development and cancer.* Nat Rev Cancer, 2008. **8**(5): p. 387-98.

118. Cao, W., et al., *Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma.* Nat Commun, 2020. **11**(1): p. 3675.

119. Jung, Y.S., et al., *Wnt2 complements Wnt/beta-catenin signaling in colorectal cancer.* Oncotarget, 2015. **6**(35): p. 37257-68.

120. Flam, E.L., et al., *Differentially Methylated Super-Enhancers Regulate Target Gene Expression in Human Cancer.* Sci Rep, 2019. **9**(1): p. 15034.

121. Heyn, H., et al., *Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer.* Genome Biol, 2016. **17**: p. 11.

122. Hnisz, D., et al., *Super-enhancers in the control of cell identity and disease.* Cell, 2013. **155**(4): p. 934-47.

123. Iwafuchi-Doi, M., et al., *The Pioneer Transcription Factor FoxA*

*Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation.* Mol Cell, 2016. **62**(1): p. 79-91.

124. Tang, Y., et al., *FOXA2 functions as a suppressor of tumor metastasis by inhibition of epithelial-to-mesenchymal transition in human lung cancers.* Cell Res, 2011. **21**(2): p. 316-26.

125. Liu, Z., et al., *CASZ1, a candidate tumor-suppressor gene, suppresses neuroblastoma tumor growth through reprogramming gene expression.* Cell Death Differ, 2011. **18**(7): p. 1174-83.

126. Luo, Z., S.K. Rhie, and P.J. Farnham, *The Enigmatic HOX Genes: Can We Crack Their Code?* Cancers (Basel), 2019. **11**(3).

127. Duverger, O. and M.I. Morasso, *Role of homeobox genes in the patterning, specification, and differentiation of ectodermal appendages in mammals.* J Cell Physiol, 2008. **216**(2): p. 337-46.

128. Subramaniam, D., et al., *DNA methyltransferases: a novel target for prevention and therapy.* Front Oncol, 2014. **4**: p. 80.

129. Fouad, Y.A. and C. Aanei, *Revisiting the hallmarks of cancer.* Am J Cancer Res, 2017. **7**(5): p. 1016-1036.

130. Hervouet, E., et al., *Specific or not specific recruitment of DNMTs for DNA methylation, an epigenetic dilemma.* Clin Epigenetics, 2018. **10**: p. 17.

131. Simon, J.A., *Polycomb group proteins.* Curr Biol, 2003. **13**(3): p. R79-80.

132. Shah, N. and S. Sukumar, *The Hox genes and their roles in oncogenesis.* Nat Rev Cancer, 2010. **10**(5): p. 361-71.

# 국문 초록

## 상피성 세포암의 다중오믹스 분석을 통한 종양 발달 관련 유전자의 다층적 조절 장애에 관한 연구

서울대학교 대학원 의과학과 의과학 전공

손 민 환

차세대 서열분석 기술 (NGS) 이라고 일컬어지는 대용량 염기서열분석법은 수많은 생물학적 과정 및 다양한 질병의 발현을 구성하는 수천, 수백만 가지의 분자 표적을, 다중오믹스 (multi-omics) 방식을 통하여 동시에 발굴하는 것을 가능케 하였다. 특히, 임상적 효용 면에서 특출 난 능력을 가진 차세대 서열분석 기술을 이용하여, 다양한 유형의 암 종에서 보편적으로 나타나는 암의 다층적 특징을 규명하는 것이 비로소 가능하게 되었다.

본 연구에서는 대규모 병렬 차세대 서열분석 기술을 사용하여, 상피성 세포를 기원으로 가지는, 고등급 장액성 난소암 (HGSOC)과 위선암 (STAD) 두 가지 유형의 암종에서, 각기 특이적인 유전자 발현 프로그램과 다면적인 기능 조절 장애를 분석하였다.

첫 번째 연구에서는 전장 엑솜 서열분석 (WES) 데이터와 전사체 서열분석 (RNA-seq) 데이터의 포괄적인 분석을 통하여 고등급 장액성 난소암의 분자적 프로파일을 특성화하였다. 고등급 장액성 난소암의 유전체 및 전사체적 환경에 대한 분석은, 유전체 손상 (genome scar)과 상피 간엽 이행 (EMT)이 본 코호트에서 각기 중추적인 역할을 하며, 결과적으로 우리의 고등급 장액성 난소암 코호트가 상동 재조합 복구 (HRR) 활성화 유형과 중간엽 (mesenchymal) 유형이라는, 두 가지의 분자적 하위 유형으로 나누어질 수 있음을 보여주었다. 특히, 낮은 유전체적 변화와 다양한 세포 유형으로 구성되어 있다는 특성을 보이는 상피 간엽 이행 전사 프로그램이 활성화된 환자군은, 상동재조합 복구 활성화 환자군에 비해 예후가 좋지 않다는 것을 밝혔다. 마지막으로, 암 유전체 아틀라스 (TCGA)의 난소암 공개 데이터를 분석하여 우리의 발견을 추가로 검증한 결과 실제로, 높은 상피간엽이행 전사체 프로파일을 가진 환자의 전반적인 생존률이 악화되어 있음을 확인할 수 있었다.

두 번째 연구에서는, 전장 유전체의 중아황산염 처리 염기서열 분석 (WGBS)과 전사체 서열분석을 이용한, 통합적이고 고차원적인 다중체 방식으로 접근하여, 위선암 (STAD)의 DNA 메틸화 양상을 기술하고 발암 추정 요인을 설명하였다. 우리는 본 코호트의 위선암 환자 샘플에서 DNA 메틸화 감소를 보이는 지역이 95% 이상인 것을 발견하는 한편, 나머지 DNA 메틸화 증가를 보이는 지역이 프로모터 (promoter), 슈퍼 인핸서 (super enhancer)및 폴리콤브 억압 복합체

(PRC) 결합 부위에서 풍부하다는 것을 발견하였다. 위 요소에서 변형된 DNA 메틸화는 암 특이적 유전자 조절 장애와 관련이 있었다. 특히 DNA 메틸화 증가로 매개된 WNT/β-카테닌/MMP 신호의 활성화는 잠재적인 위선암의 발암 원인이라는 것을 알 수 있었다. 또한, 슈퍼 인핸서 관련 유전자의 하향 조절과 호메오박스 군집 유전자들의 재활성화를 DNA 메틸화 증가와의 관계를 통하여 확인할 수 있었다. 이를 통하여, 고전적인 유전체 및 전사체적 조절장애로 말미암아 생기는 위선암 발생을 넘어, 후성체 유전학적 조절이 다중 인자 메커니즘을 통해 위선암 발생을 촉진한다는 것을 보여주었다.

상피성 세포 유래 암의 높은 발병률 및 사망률을 고려할 때, 해당 종양의 발생을 초래하는 분자 수준의 복잡한 상호작용을 이해하는 것이 필수적이라고 할 수 있다. 본 연구는 상피성 세포 유래 암에 대한 포괄적인 다중오믹스 중심 분석 방식과 새로운 진단 및 표적 치료 대상에 대한 귀중한 자원을 제공한다는 데 의의가 있다고 할 수 있다.

* 본 논문의 첫 번째 연구는 Genes 에 출판된 내용임 [1].

---

**주요어:** 상피성 세포암; 고등급 장액성 난소암; 전장엑솜 서열분석; 전사체 서열분석; 상피간엽이행; 위선암; 전장후성유전체 서열분석; 다중오믹스 서열분석; 후성유전체; DNA 메틸화

**학번:** 2014-25063