

# Identification of key genes by integrating DNA methylation and next-generation transcriptome sequencing for esophageal squamous cell carcinoma

Yang Chen<sup>1,2,\*</sup>, Lian-Di Liao<sup>1,3,\*</sup>, Zhi-Yong Wu<sup>4,\*</sup>, Qian Yang<sup>1,2</sup>, Jin-Cheng Guo<sup>1,2</sup>, Jian-Zhong He<sup>1,3</sup>, Shao-Hong Wang<sup>5</sup>, Xiu-E Xu<sup>1,3</sup>, Jian-Yi Wu<sup>1,2</sup>, Feng Pan<sup>1,2</sup>, De-Chen Lin<sup>6</sup>, Li-Yan Xu<sup>1,3</sup>, En-Min Li<sup>1,2</sup>

<sup>1</sup>The Key Laboratory of Molecular Biology for High Cancer Incidence Coastal Chaoshan Area, Shantou University Medical College, Shantou 515041, Guangdong, P.R. China

<sup>2</sup>Department of Biochemistry and Molecular Biology, Shantou University Medical College, Shantou 515041, Guangdong, P.R. China

<sup>3</sup>Institute of Oncologic Pathology, Shantou University Medical College, Shantou 515041, Guangdong, P.R. China

<sup>4</sup>Departments of Oncology Surgery, Shantou Central Hospital, Affiliated Shantou Hospital of Sun Yat-Sen University, Shantou 515041, Guangdong, P.R. China

<sup>5</sup>Departments of Pathology, Shantou Central Hospital, Affiliated Shantou Hospital of Sun Yat-Sen University, Shantou 515041, Guangdong, P.R. China

<sup>6</sup>Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

\*Equal contribution

**Correspondence to:** En-Min Li, Li-Yan Xu, De-Chen Lin; **email:** [nmli@stu.edu.cn](mailto:nmli@stu.edu.cn), [lyxu@stu.edu.cn](mailto:lyxu@stu.edu.cn), [dchlin11@gmail.com](mailto:dchlin11@gmail.com)

**Keywords:** esophageal squamous cell carcinoma, DNA methylation, next-generation sequencing, histone modification, survival analysis

**Received:** October 12, 2019

**Accepted:** December 25, 2019

**Published:** January 21, 2020

**Copyright:** Chen et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## ABSTRACT

Aberrant DNA methylation leads to abnormal gene expression, making it a significant regulator in the progression of cancer and leading to the requirement for integration of gene expression with DNA methylation. Here, we identified 120 genes demonstrating an inverse correlation between DNA methylation and mRNA expression in esophageal squamous cell carcinoma (ESCC). Sixteen key genes, such as SIX4, CRABP2, and EHD3, were obtained by filtering 10 datasets and verified in paired ESCC samples by qRT-PCR. 5-Aza-dC as a DNA methyltransferase (DNMT) inhibitor could recover their expression and inhibit clonal growth of cancer cells in seven ESCC cell lines. Furthermore, 11 of the 16 genes were correlated with OS (overall survival) and DFS (disease-free survival) in 125 ESCC patients. ChIP-Seq data and WGBS data showed that DNA methylation and H3K27ac histone modification of these key genes displayed inverse trends, suggesting that there was collaboration between DNA methylation and histone modification in ESCC. Our findings illustrate that the integrated multi-omics data (transcriptome and epigenomics) can accurately obtain potential prognostic biomarkers, which may provide important insight for the effective treatment of cancers.

## INTRODUCTION

Esophageal cancer is globally one of the most prevalent cancers which is diagnosed more than 500,000 new cases yearly [1]. Esophageal adenocarcinoma (EAC)

and esophageal squamous cell carcinoma (ESCC) are the two main clinical subtypes of esophageal cancer, with more than 80% of esophageal cancers being ESCC [2]. In Asia, ESCC has higher morbidity and mortality compared with western countries [3]. More than half of

patients with these tumors have distal metastases at the time of diagnosis, and only 10%–20% of sufferers survive for 5 years [1]. Therefore, there is an urgent need to provide effective targets for therapy or early detection of ESCC.

Epigenetic alterations have been suggested as underlying mechanisms of cancer progression [4]. DNA methylation is one of the major epigenetic mechanisms involved in cancer [5], with global DNA hypomethylation accompanied by hypermethylation of tumor suppressor genes being recognized as an epigenetic hallmark of cancer [6]. Altered DNA methylation patterns can change gene expression by silencing or activating genes, so numerous studies aimed at revealing the pathogenesis of tumors have focused on epigenetics, such as circulating tumor DNA (ctDNA) analysis [7–9] and whole-genome bisulfite sequencing (WGBS) [10–12]. In 1999, the feasibility of detecting tumor-associated abnormal ctDNA methylation was initially described [13]. Since then, more and more studies have been conducted to characterize the potential use of ctDNA methylation for early diagnosis and prognosis [14]. Some methylation-associated drugs, such as decitabine, an effective drug for myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML), have been used for the clinical therapy of AML [15]. On the other hand, next-generation transcriptome sequencing (RNA-Seq) provides a method for tracing transcriptional aberrations in diseases [16, 17]. In addition, posttranslational modifications of histones also regulate gene activity in tumors [18, 19]. Nevertheless, there is a little research on integrating DNA methylation and gene expression or histone modification in ESCC.

Here, we systematically define DNA methylation status and mRNA expression level and demonstrate the correlation between DNA methylation and gene expression by using DNA methylation array and RNA-seq data. Furthermore, we perform a comprehensive analysis to determine whether some identified genes could serve as potential prognostic biomarkers and could be applied to the treatment of ESCC.

## RESULTS

### Identification of differentially-methylated genes

We performed an Illumina Infinium HumanMethylation450 BeadChip assay to compare esophageal DNA patterns in fifteen ESCC patients' tumor tissues and paired normal tissues. After the chip was pretreated, the methylation data of the 15 pairs of esophageal squamous cell carcinoma samples was analyzed using the *minfi* package. We detected 80,557

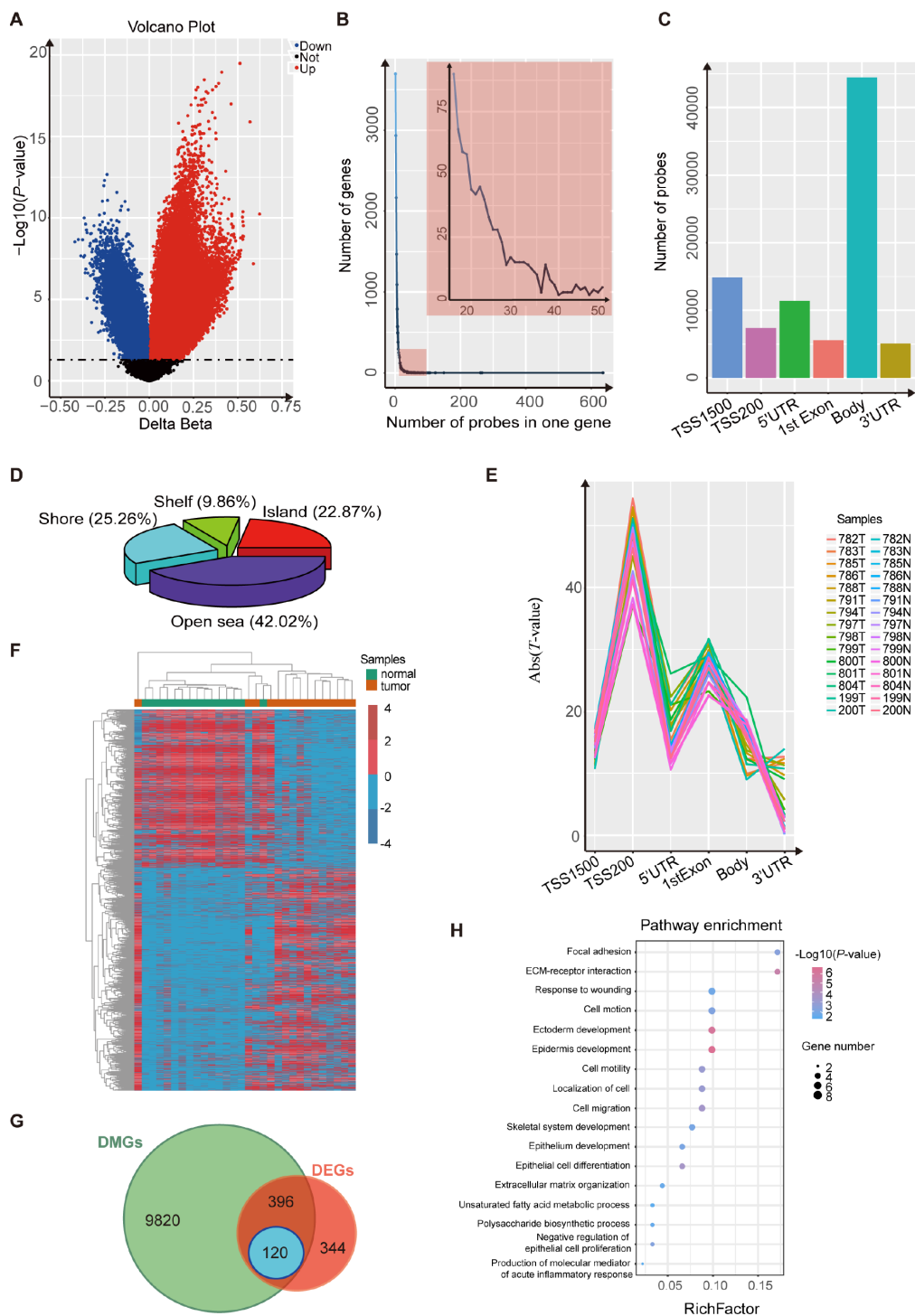
CpG sites, related to 15,882 different genes, that showed significant differences in DNA methylation in ESCC ( $p$ -value $<0.05$ , Figure 1A and Supplementary Table 1). We also found that most of the genes were covered with 1~15 differentially-methylated probes (DMPs) (Figure 1B). Most of the 80,557 differentially-methylated sites were located in the gene body and TSS1500 (Figure 1C), and more than half of the DMPs were located around the CpG islands (Figure 1D). Because several probes can map to a given gene when using the Illumina Infinium 450k bead array, we wanted to evaluate which probes were most relevant to the state of gene expression. To determine this, we averaged the DNA methylation  $\beta$  value of probes mapping to the same gene region (TSS1500, TSS200, 5'UTR, 1st exon, gene body and 3'UTR). Regression analysis revealed that the strongest associations between gene expression and DNA methylation were for TSS200, 1st exon and TSS1500 regions (Figure 1E). Finally, we assigned a unique  $\beta$  value to a given gene by the following scheme: for a gene with probes binding to TSS200, the average  $\beta$  value of such TSS200 probes was used. For a gene with no TSS200 probes but with probes binding to the 1st exon, we used the average over 1st exon probes. For a gene with no TSS200- or 1st exon-binding probes, the average over the TSS1500 probes was used [20]. After these analyses, we obtained 10,336 differentially-methylated genes (Figure 1F).

### Integration of DNA methylation and mRNA expression data to obtain candidate genes

Expression analysis of genes was performed on 15 paired esophageal samples. We identified 860 differentially-expressed genes between tumors and non-tumor matched samples (Figure 1G;  $p$ -value $<0.05$ ,  $|\log_2(\text{FC})|>1$ ). Based on these results, we ultimately obtained 120 candidate genes that were both differentially methylated and expressed, and with the strongest negative associations between gene expression and DNA methylation (Figure 1G;  $p$ -value $<0.05$ ,  $\text{PCC}<-0.5$ ). Then we found that these candidate genes were cancer-associated, by DAVID enrichment analysis, which suggested that abnormal methylation also plays an important role in esophageal squamous cell carcinoma (Figure 1H).

### Use of multiple expression profiles to screen the specific methylated key genes in ESCC

In order to find the specific genes regulated by DNA methylation only in ESCC, we then downloaded 10 datasets (5 ESCC datasets and 5 other cancer datasets) from the GEO database to filter the 120 candidate genes. The genes that showed differential expression in



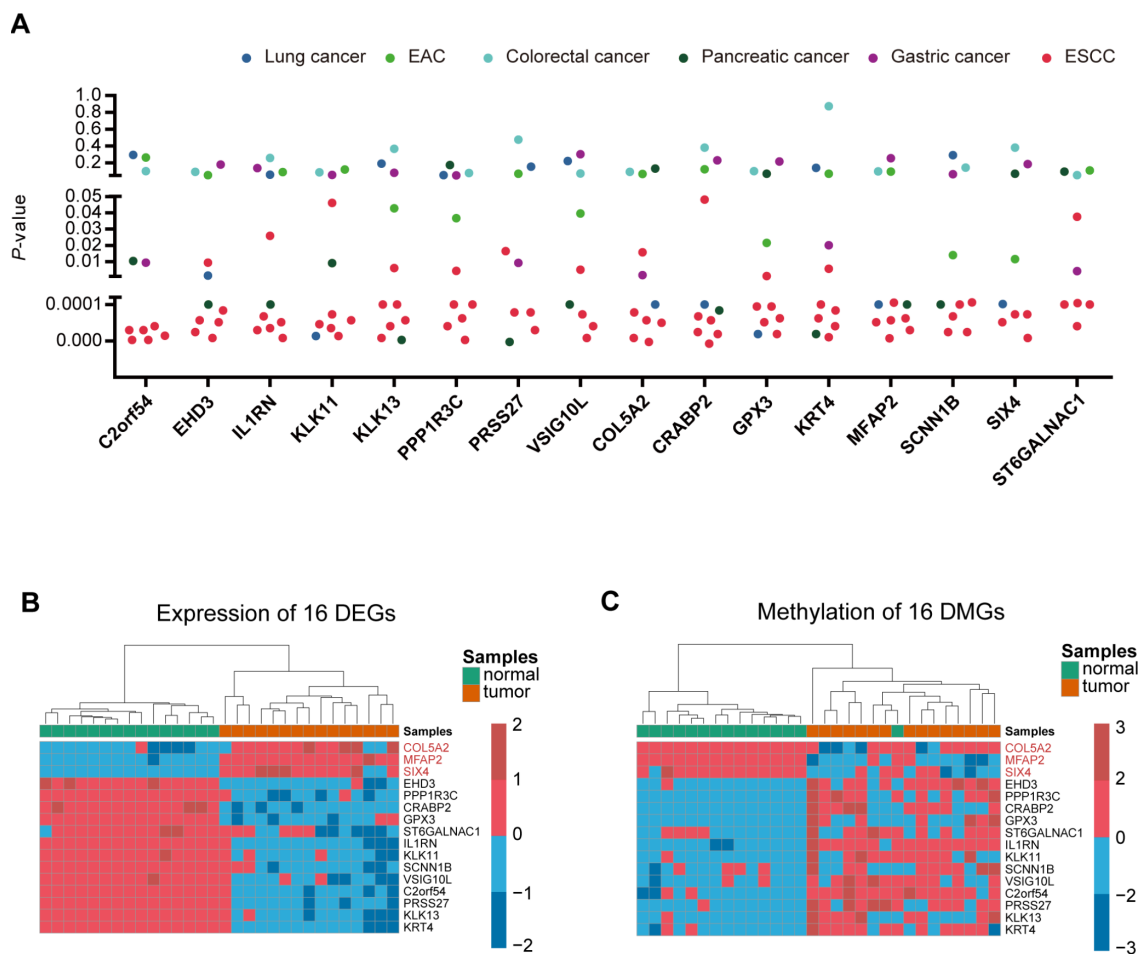
**Figure 1. Analysis using the Illumina 450K bead array and RNA-Seq data for 15 paired ESCC samples. (A)** Volcano plot representing the probes for differentially-methylated genes. The probes for hypermethylated genes are shown in red while probes for hypomethylated genes are shown in blue ( $P < 0.05$ ). **(B)** Frequency line graph shows the coverage rate of differentially-methylated regions (DMPs) in one gene. The results show that most genes have more than twenty DMPs. **(C)** Distribution of differentially-methylated sites in six gene regions (TSS1500, TSS200, 5' UTR, 1st exon, gene body and 3' UTR). **(D)** Proportions of differentially-methylated regions from genes with associated CpG islands (CGI). **(E)** Plot of the regression t-statistics between log-normalized RNA-Seq RPKM values and corresponding average DNA methylation  $\beta$  values for probes, stratified according to six genetic regions. The number of curves equals the number of samples. **(F)** Heatmap shows the differentially-methylated genes in 15 paired ESCC samples. **(G)** Venn plot shows the overlap between differentially-methylated genes and differentially-expressed genes in 15 paired ESCC samples. **(H)** KEGG and GO analysis of 120 candidate genes that are both differentially methylated and differentially expressed.

at least 3 of the 5 ESCC datasets, and differential expression in no more than 2 pan-cancer datasets in 5 other cancers, were selected for key genes. Ultimately, we identified 16 specific genes in ESCC (Figure 2A and Supplementary Table 2). The heatmap of the 16 genes shows the expression level and methylation status in 15 paired ESCC samples (Figure 2B and 2C).

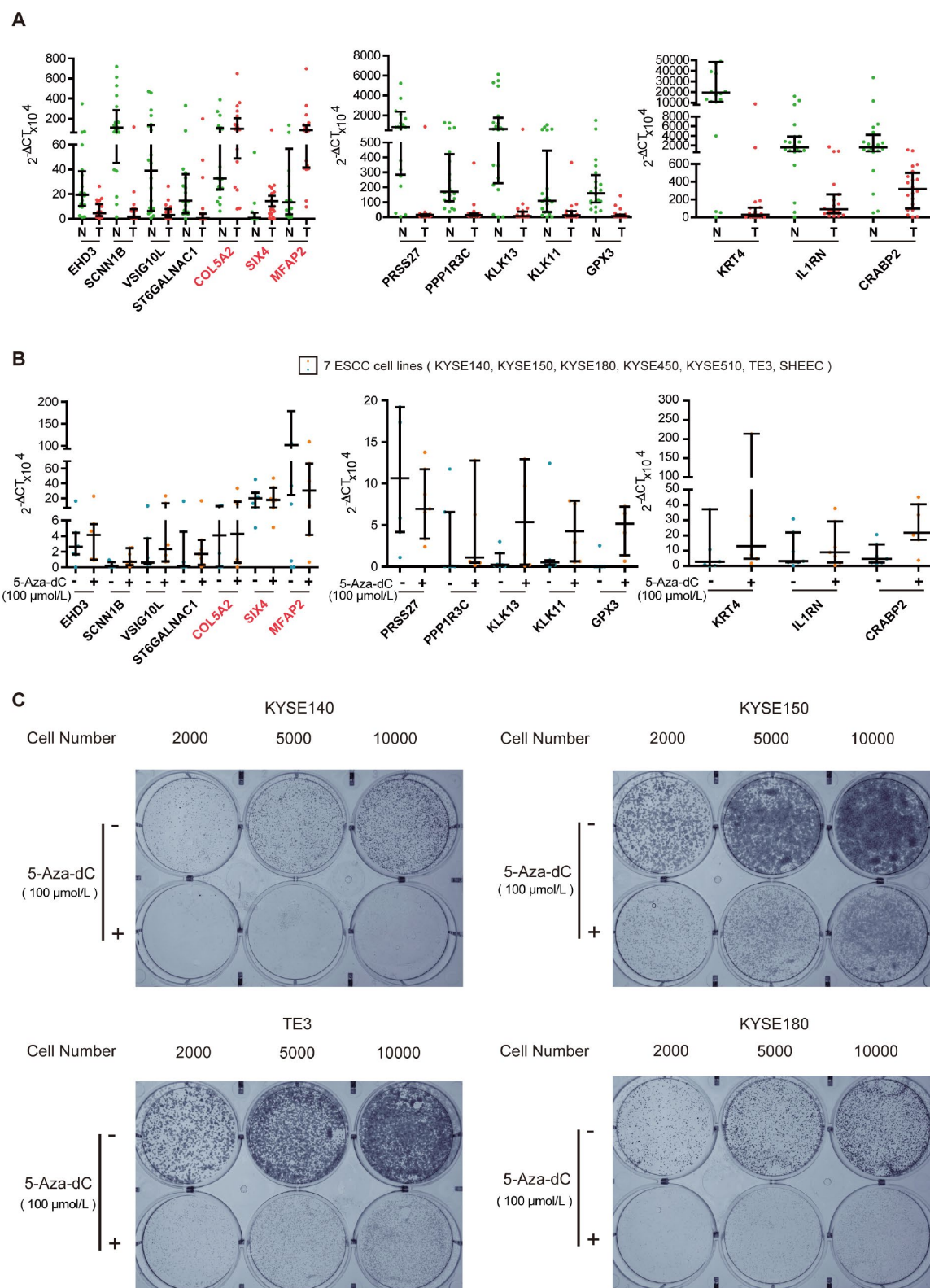
### Key genes regulated by methylation affect the proliferation of ESCC cells

To test the biological functions of the 16 genes, we first analyzed expression of the genes, in twenty ESCC tumor tissues and paired normal esophageal epithelial tissues, using qRT-PCR. Expression of most of the 16 genes could be detected in all of the tumors and normal esophageal epithelium tissues, except for C2orf54. As shown in Figure 3A, expression of the hypermethylated genes (VSIG10L, ST6GALNAC1, SCNN1B, PRSS27, PPP1R3C, KRT4, KLK13,

KLK11, IL1RN, GPX3, EHD3 and CRABP2) in tumor tissues was lower than that of the corresponding genes in normal esophageal epithelial tissues. The hypomethylated genes (SIX4, MFAP2 and COL5A2) showed the reverse trend. In order to know whether the DNA methylation status of these 15 genes was associated with their expression in ESCC, we further detected their mRNA expression in 5-aza-dC-treated ESCC cell lines. The expression of hypermethylated genes increased in seven ESCC cell lines after 5-aza-dC treatment (Figure 3B). Then, we cultured KYSE140, KYSE150, TE3 and KYSE180 cells with or without 5-aza-dC and examined the ability of cells to form colonies. The efficiency of cell colony formation in the presence of 5-aza-dC was significantly decreased, suggesting that aberrant hypermethylation of these key genes could promote cell colony formation and proliferation of ESCC cells (Figure 3C). All of these results illustrate that methylation-related genes play a vital role in ESCC.



**Figure 2. Identification of ESCC-specific genes.** (A), Genes were filtered from ten data sets (5 ESCC data sets and five other cancer data sets) to obtain specific key genes in ESCC. Scatter plot shows the  $p$ -value of 16 differentially-expressed genes in the ten data sets. (B, C) Heatmap of methylation and expression of the 16 key genes, respectively.



**Figure 3. Experimental verification of key genes in ESCC.** (A) qRT-PCR analysis of key genes in tumor (T) and normal (N) tissues of twenty paired ESCC samples. (B) Scatter plot of qRT-PCR analyses for key genes in seven ESCC cell lines. Blue spots represent cells treated with DMSO, whereas the orange spots represent cells treated with 5-aza-dC. (C) Colony formation assays of ESCC cells after 5-aza-dC treatment. ESCC cells were plated in 6-well plates. After 24 h, the cells were treated with 5-aza-dC. Cultures were maintained for six days, and cells were then stained and photographed. DMSO was used as the control. Colony formation assays illustrate that hypermethylation of key genes plays an important role in cell growth.

## Evaluation of the prognostic performance of the key genes

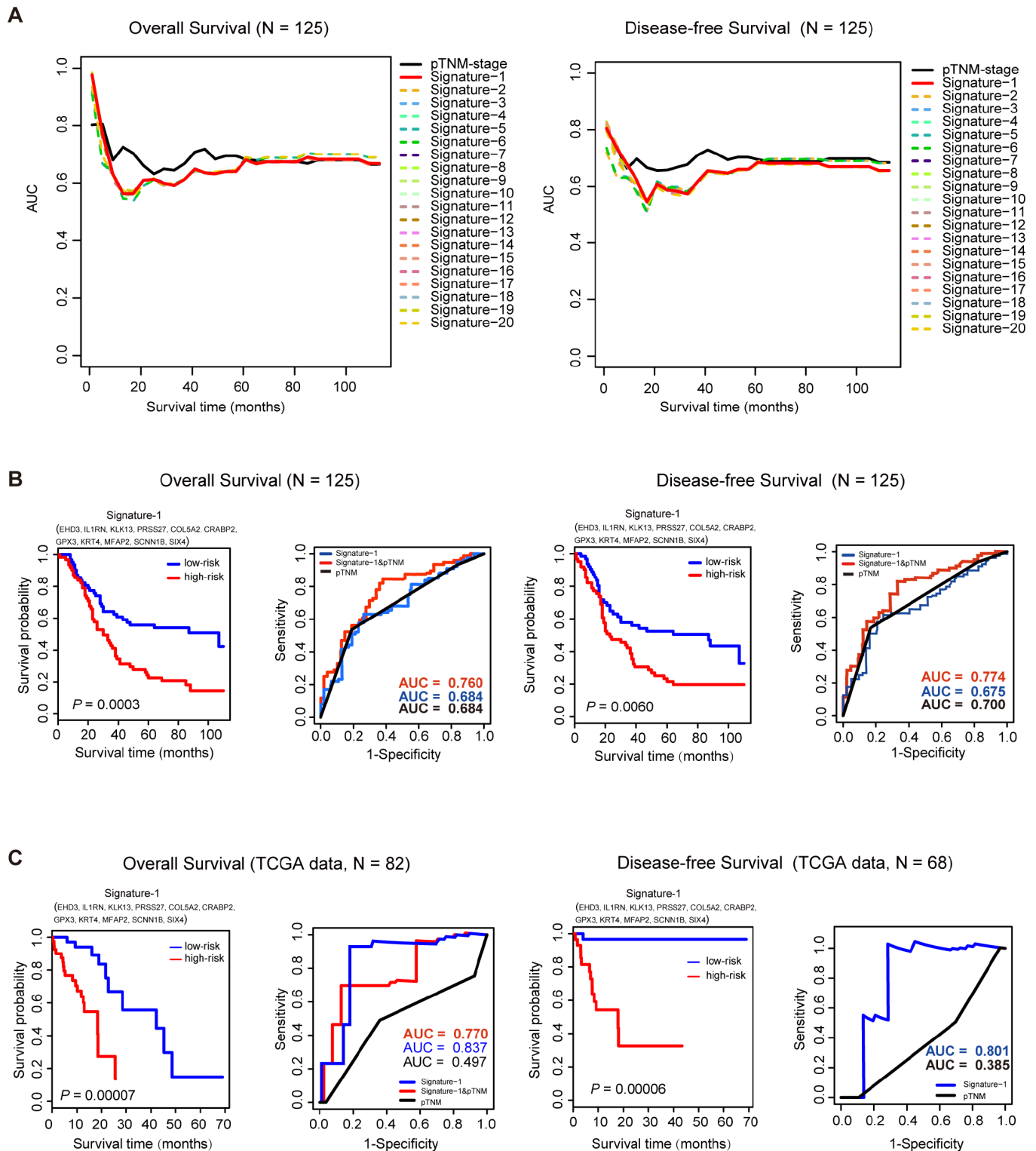
In order to explore the clinical significance of the key genes, an Affymetrix IVT microarray of 125 ESCC patients (GSE121931) was used to find the potential prognostic factors (VSIG10L was not scored because it was not detected by the chip). The clinicopathological characteristics of the 125 ESCC patients are shown in Supplementary Table 3. Then survival analysis was used to evaluate the impact of these key genes on the overall survival (OS) and disease-free survival (DFS) of the 125 ESCC patients. First, we conducted  $16383 (C_{14}^1 + C_{14}^2 + C_{14}^3 + \dots + C_{14}^{14})$  Cox proportional hazard models by permutating and combining the expression of 14 key genes and OS time or DFS time, respectively. Next, we evaluated the efficiency of each survival model and found that 2,923 models in OS and 1,181 models in DFS were survival-associated ( $p$ -value $<0.01$ ). Although pTNM stage is a crucial prognostic indicator of esophageal cancer [21, 22], the analysis lacks auxiliary biomarkers to aid the accuracy of pTNM [23, 24]. We then performed ROC analysis to compare the prognostic efficiency between these signatures and pTNM stage. Figure 4A and 4B shows the top 20 AUCs of these models for OS and DFS. Then, we found a model (Signature-1) composed of 11 key genes (EHD3, IL1RN, KLK13, PRSS27, COL5A2, CRABP2, GPX3, KRT4, MFAP2, SCNN1B and SIX4) with excellent prognostic capability in both the OS and DFS models, and the AUC of the time-dependent ROC curve was similar to pTNM stage (OS:  $AUC_{Signature-1} = 0.684$ ,  $AUC_{pTNM} = 0.684$ , DFS:  $AUC_{Signature-1} = 0.675$ ,  $AUC_{pTNM} = 0.700$ ). In order to understand the clinical significance of Signature-1 better, we associated a series of clinicopathological parameters with this signature in 125 ESCC patients. There was no relevance between Signature-1 and pTNM stage, age and gender (Table 1 and Supplementary Table 4). We found that Signature-1 and pTNM stage were independent of these clinical features and also were independent prognostic factors in OS and DFS (univariate analysis  $p$ -value $<0.05$ ). So we combined Signature-1 with pTNM stage to construct a new model with better prognostic performance (OS:  $AUC = 0.760$ , DFS:  $AUC = 0.774$ ) in our cohort.

To further investigate the prognostic power of this signature, the test ESCC dataset that was downloaded from TCGA was also used to evaluate the Signature-1 prognostic model. There were 82 ESCC patients with OS time and 68 patients with DFS time in the test datasets. Cox regression analysis was performed first and found that Signature-1 was indeed independent of these clinical features, and pTNM stage also was an independent prognostic factor in OS, but not in DFS

(Table 1). In the OS group ( $n=82$ ), the predictive ability of Signature-1 was significantly better than pTNM stage ( $AUC_{Signature-1}=0.837$  vs.  $AUC_{pTNM}=0.497$ ) and the combined model also had a better ability than the TNM stage ( $AUC_{Signature-1\&pTNM}=0.770$ , Figure 4C). In the DFS group ( $n=68$ ), the predictive ability of Signature-1 was significantly better than pTNM stage ( $AUC_{Signature-1}=0.801$  vs.  $AUC_{pTNM}=0.385$ , Figure 4C). These results demonstrate that Signature-1 has important clinical relevance and is a novel prognostic signature with high accuracy.

## Relationship between DNA methylation and histone modification of the key genes in ESCC

DNA methylation usually is considered to repress transcription, which is also influenced by histone modifications [25, 26]. In order to further reveal the role of the 11 survival-associated key genes that have aberrant epigenomic regulation in ESCC, we performed H3K27ac ChIP-seq analysis on five ESCC cell lines and whole-genome bisulfite sequencing (WGBS) in seven primary ESCC samples. We define the H3K27ac histone modification level and the WGBS level of a 100bp region which contain the differentially methylated site. Then, we performed the differential analysis between H3K27ac and DNA methylation (WGBS) in this region. After this analysis, we can see that most of these 14 genes have a significant  $p$ -value between H3K27ac histone modification and methylation (Supplementary Figure 1). Take CRABP2 for example (Figure 5), we examined the methylation  $\beta$  values of all 16 CpG probes that are located on CRABP2 from HM450 array data (Infinium HumanMethylation450) in three cohorts (our data,  $n=30$ ; TCGA ESCC data,  $n=85$  and TCGA EAC data,  $n=87$ ). By plotting the change in methylation ( $\Delta\beta$ ) of ESCC and EAC samples, we identified hypermethylation of promoter CpG loci in ESCC samples, but not in EAC samples. We next calculated the Pearson correlation between each probe's methylation  $\beta$  value and CRABP2 mRNA expression level in all samples. The uniquely hypermethylated region in ESCC was inversely correlated with CRABP2 expression, but was unchanged in EAC samples. These results indicated that abnormally high methylation of the CRABP2 promoter region was most prominently associated with aberrantly low CRABP2 expression in ESCC. Consistent with the 450K array, the CRABP2 promoter region ( $TSS\pm 1.5kb$ ) was intensively methylated in seven ESCC samples and corresponded to both a decrease in H3K27acetylation in the same region in ESCC cells and low CRABP2 expression. Interestingly, a previous study indicated that CRABP2 is associated with super-enhancer activity in healthy esophageal tissue [27]. This shows epigenetically



**Figure 4. Survival analysis of 125 ESCC samples and TCGA data sets. (A),** Time-dependent AUC curves of overall survival (OS) and disease-free survival (DFS) for the top 20 signatures in 125 ESCC samples. **(B)** For the 125 ESCC samples, Kaplan-Meier curves for overall survival (OS) and disease-free survival (DFS) for Signature-1. ROC analysis shows a better prognostic efficiency of Signature-1 combined with pTNM-stage compared with Signature-1 or pTNM-stage. **(C)** In the TCGA dataset, Kaplan-Meier curves of overall survival (OS) and disease-free survival (DFS) for Signature-1. ROC analysis shows better prognostic efficiency of Signature-1 compared with Signature-1 combined with pTNM-stage or pTNM-stage alone.

**Table 1. Univariate and multivariate analysis of factors associated with overall survival and disease-free survival.**

| Variables   | Univariate analysis |        |               |          | Multivariate analysis |        |               |         |
|---|---------------------|--------|---------------|----------|-----------------------|--------|---------------|---------|
|   | P-value*            | HR     | 95% CI for HR |          | P-value*              | HR     | 95% CI for HR |         |
|   |                     |        | Lower         | Upper    |                       |        | Lower         | Upper   |
| <b>Overall survival (N = 125)</b>                   |                     |        |               |          |                       |        |               |         |
| Age (>59 vs. ≤ 59)                                  | 0.079               | 1.524  | 0.953         | 2.438    |                       |        |               |         |
| Gender (Female vs. Male)                            | 0.253               | 0.708  | 0.391         | 1.280    |                       |        |               |         |
| pTNM-stage (III vs. I+II)                           | 0.000               | 2.554  | 1.611         | 4.048    | 0.000                 | 2.526  | 1.621         | 3.939   |
| Signature-1 (High score vs. Low score) <sup>a</sup> | 0.016               | 1.814  | 1.120         | 2.938    | 0.001                 | 2.119  | 1.342         | 3.345   |
| <b>Disease-free survival (N = 125)</b>              |                     |        |               |          |                       |        |               |         |
| Age (>59 vs. ≤ 59)                                  | 0.447               | 1.195  | 0.755         | 1.890    |                       |        |               |         |
| Gender (Female vs. Male)                            | 0.687               | 0.884  | 0.485         | 1.611    |                       |        |               |         |
| pTNM-stage (III vs. I+II)                           | 0.000               | 2.570  | 1.620         | 4.077    | 0.000                 | 2.582  | 1.645         | 4.052   |
| Signature-1 (High score vs. Low score) <sup>a</sup> | 0.017               | 1.777  | 1.110         | 2.847    | 0.008                 | 1.856  | 1.173         | 2.936   |
| <b>Overall survival (TCGA data, N = 82)</b>         |                     |        |               |          |                       |        |               |         |
| Age (>57 vs. ≤ 57)                                  | 0.417               | 1.384  | 0.631         | 3.035    |                       |        |               |         |
| Gender (Female vs. Male)                            | 0.168               | 4.397  | 0.535         | 36.165   |                       |        |               |         |
| pTNM-stage (IV+III vs. I+II)                        | 0.019               | 2.687  | 1.177         | 6.138    | 0.020                 | 2.656  | 1.167         | 6.042   |
| Signature-1 (High score vs. Low score) <sup>b</sup> | 0.000               | 6.148  | 2.220         | 17.023   | 0.000                 | 6.147  | 2.239         | 16.878  |
| <b>Disease-free survival (TCGA data, N = 68)</b>    |                     |        |               |          |                       |        |               |         |
| Age (>57 vs. ≤ 57)                                  | 0.839               | 1.117  | 0.384         | 3.252    |                       |        |               |         |
| Gender (Female vs. Male)                            | 0.244               | 29.251 | 0.100         | 8527.816 |                       |        |               |         |
| pTNM-stage (IV+III vs. I+II)                        | 0.377               | 0.558  | 0.153         | 2.032    |                       |        |               |         |
| Signature-1 (High score vs. Low score) <sup>b</sup> | 0.004               | 19.462 | 2.538         | 149.258  | 0.005                 | 19.018 | 2.422         | 149.333 |

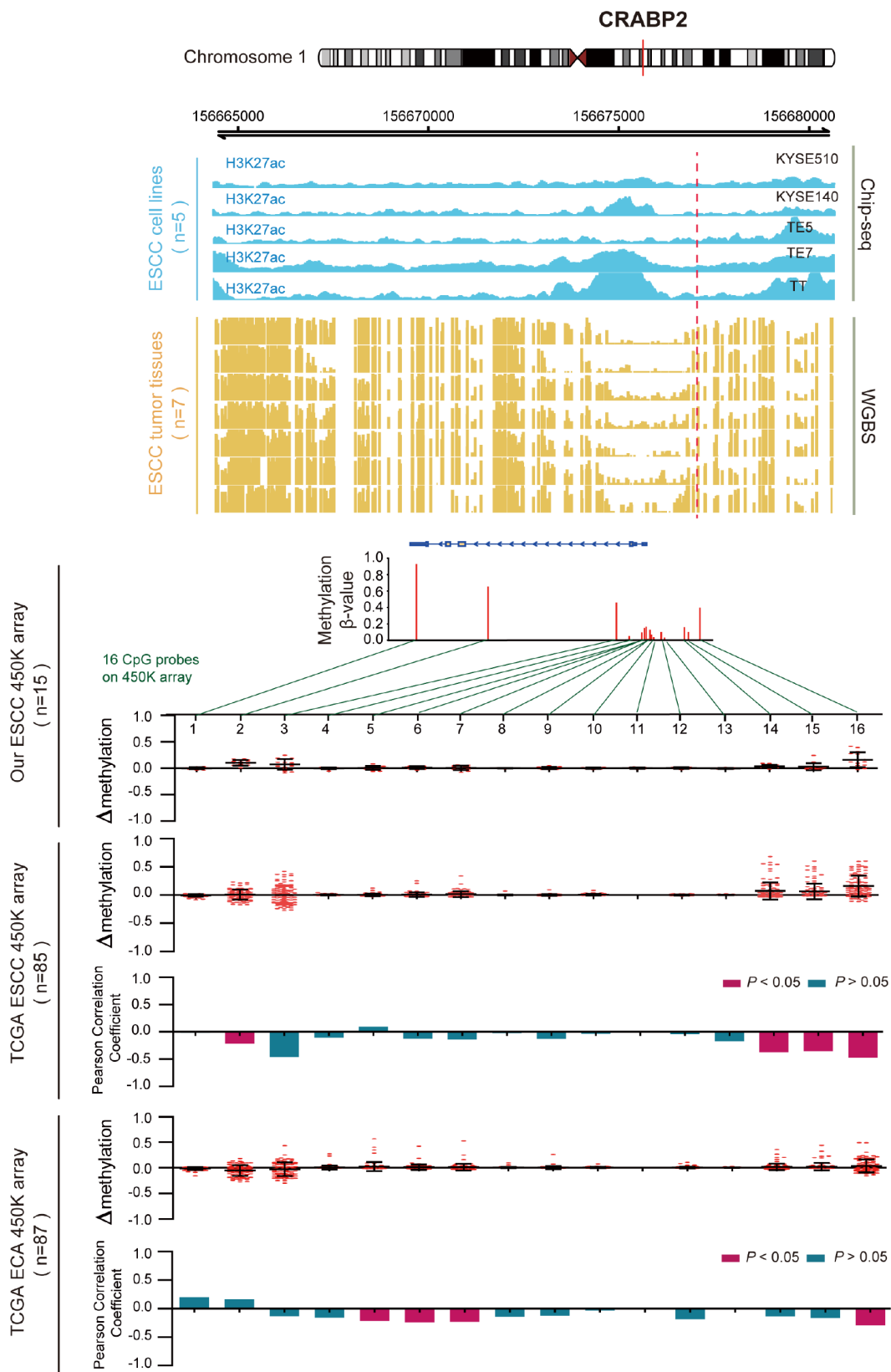
\*Multivariate analysis, Cox proportional hazards regression model. Variables were adopted for their prognostic significance by univariate analysis. <sup>a</sup> Low, score < -2.221; high, score ≥ -2.221. <sup>b</sup> Low, score < -2.066; high, score ≥ -2.066.

modified regions of CRABP2 may affect histone binding and correlates with expression silencing.

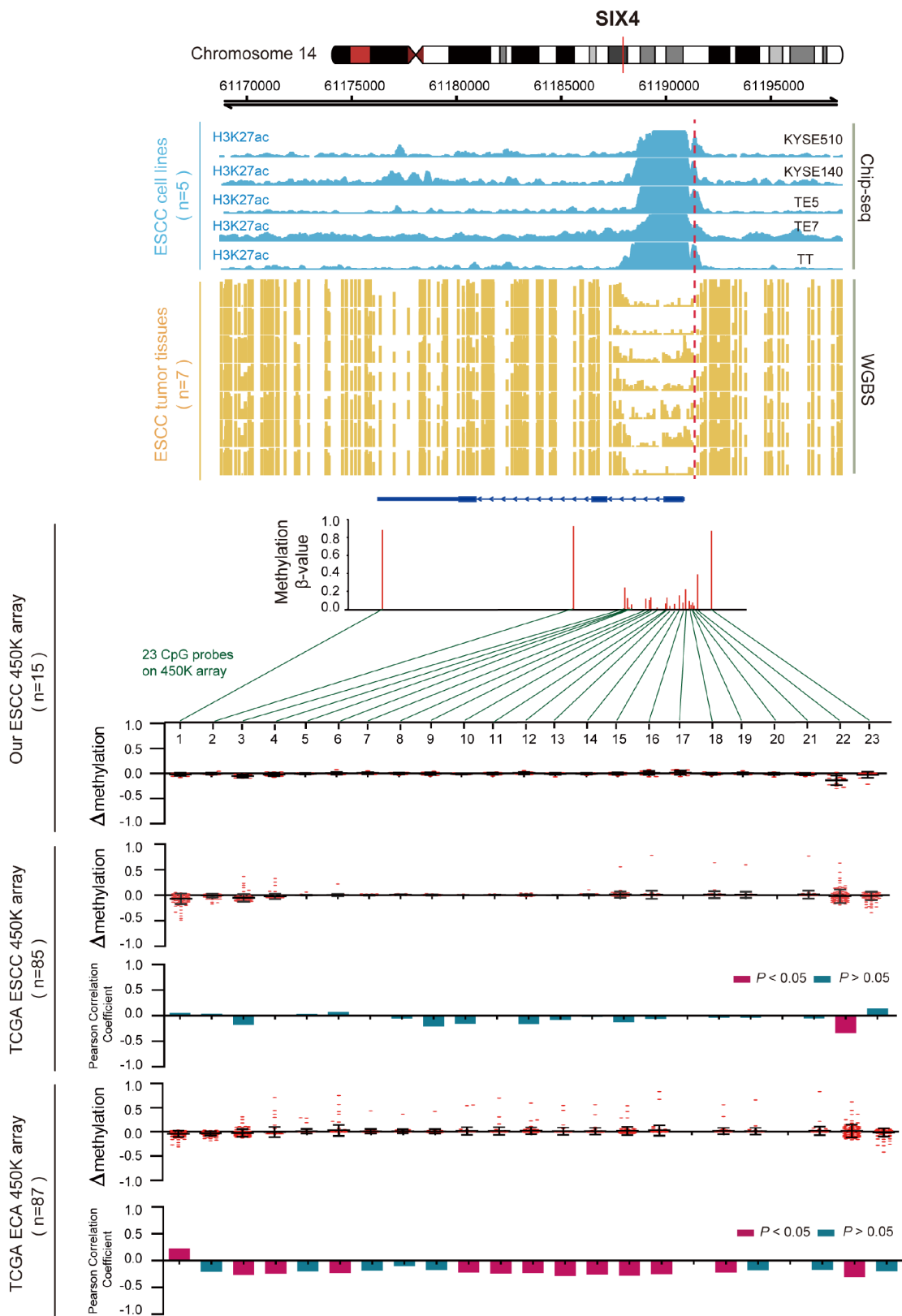
Regarding the hypomethylated gene SIX4 (Figure 6), we also examined the methylation β values of all 23 CpG probes that were located in SIX4 in the cohorts. We found frequent hypomethylation of the CpG loci at the promoter region in ESCC samples, but not in EAC samples. The uniquely hypomethylated probe in ESCC was inversely correlated with SIX4 expression and was hypermethylated in EAC samples. This indicates that

abnormal hypomethylation of the SIX4 promoter was most prominently associated with aberrantly high SIX4 expression in ESCC. ChIP-seq and WGBS data also showed that the SIX4 promoter was weakly methylated in seven ESCC samples, and in the same region, H3K27acetylation was enhanced in ESCC cells and correlated with elevated SIX4 expression (Supplementary Figure 1). Another 12 genes are shown in Supplementary Figure 2. These results further confirmed that aberrant expression of the key genes was regulated by DNA methylation and histone modification.





**Figure 5. Inverse trend between DNA methylation and histone modification of CRABP2 in ESCC.** Blue tracks represent the histone modifications in CRABP2 for five ESCC cell lines, and yellow tracks represent its methylation level, as measured by the WGBS assay. All tracks are on the same scale (0-1). Scatter diagrams show the  $\Delta\beta$  of CRABP2 in ESCC samples compared with normal samples. Histograms show the correlation between DNA methylation and gene expression of CRABP2.



**Figure 6. Inverse trend between DNA methylation and histone modification of SIX4 in ESCC.** Blue tracks represent the histone modifications of SIX4 in five ESCC cell lines, and yellow tracks represent its methylation level, as measured by the WGBS assay, all the tracks are on the same scale (0-1). Scatter diagrams show the  $\Delta\beta$  of SIX4 in ESCC samples compared with normal samples. Histograms show the correlation between DNA methylation and gene expression of SIX4.

## DISCUSSION

ESCC is one of the most fatal carcinomas around the world. However, the application of prognostic biomarkers is scarce [28]. In this study, we established a pipeline to find and investigate potentially aberrant genes in ESCC to correlate DNA methylation and mRNA expression by: (1) identifying both differentially-methylated and expressed genes, (2) keeping the genes whose methylation and expression are highly inversely correlated, (3) performing GO and KEGG pathway analysis on the inversely correlated genes, (4) identifying specific key genes by several datasets and verifying their methylation and expression levels by biological experiments, (5) evaluating the prognostic significance of these genes, and (6) exploring the potential molecular mechanisms by examining histone modification and WGBS data.

DNA methylation plays a major role in expression and disease [29–31]. Many researchers have found that super-enhancers may indicate an important role for the key genes in cancers [31, 32]. RNA-Seq also has been applied to identify the functional mRNAs in ESCC [33, 34]. However, single category data may miss some important information related to tumorigenesis and tumor progression. Recently, integrated analysis between DNA methylation and gene expression has been applied to lung adenocarcinoma [35, 36] and cholangiocarcinoma [37]. Aberrant DNA methylation at functional regulatory elements is often accompanied by alterations of histone modifications [38]. So, a comprehensive analysis of DNA methylation and gene expression or histone modification offers a new basis for the diagnosis and treatment of ESCC.

A total of fourteen DMGs were discovered in our integrated analysis, eleven of which (KLK13, PRSS27, COL5A2, KRT4, MFAP2, SCNN1B, SIX4, CRABP2, IL1RN, EHD3 and GPX3) were associated with prognosis of ESCC and also exhibited a negative correlation between DNA methylation and mRNA expression. KLK13 is a member of the tissue kallikrein (KLK) family, and overexpression of KLK13 has been shown to result in an increase in malignant cell behavior [39]. Marapsin (PRSS27) is a trypsin-like serine protease and is highly expressed in normal esophagus [40]. COL5A2, also known as collagen type V alpha 2 chain, has previously been reported to be involved in the pathology of cancer [41, 42], and the protein encoded by KRT4 is a member of the keratin gene family and is related to differentiation [43]. MFAP2 plays an important role in the extracellular deposition [44], and few studies have explored the role of MFAP2 in cancers.

SCNN1B encodes the  $\beta$  subunit of the epithelial sodium channel (ENaC), which is essential for the maintenance of body salt and water homeostasis [45]. SIX4 encodes a member of the homeobox family who have been reported to be related with tumorigenesis [46]. Importantly, a model (Signature-1) consisting of key genes highly efficient at predicting both OS and DFS, was used to identify SIX4, suggesting SIX4 could be a potential prognostic biomarker. Four of the key genes (CRABP2, IL1RN, EHD3 and GPX3) are associated with super-enhancers in healthy esophageal tissue and are hypermethylated, with correspondingly low expression, in ESCC samples (Supplementary Table 2). CRABP2 is a crucial component of the RAR pathway and can induce apoptosis in MCF-7 mammary carcinoma cells [47, 48]. IL1RN is an important regulator of the inflammatory response [49]. EHD3 regulates endocytic recycling, along with other EHD family members [50, 51], and it has been considered to be a tumor suppressor in gliomas [52]. GPX3 can catalyze the reduction of peroxides and protect cells against oxidative damage and its hypermethylation has been found in esophageal adenocarcinoma. It also involved in the progression and lymph node metastasis of ESCC [53–55]. These results suggest that the present analysis has identified some potential key genes in ESCC, and their aberrant DNA methylation may affect the function of super-enhancers to lead to abnormal gene expression and influence tumor progression. However, there are still some deficiencies in our research. Although we investigated the aberrant coding RNAs in ESCC, many potential non-coding RNAs regulated by methylation may also play key roles in tumor initiation [56–58].

All in all, this is the first integrated study of epigenomics and transcriptomics to identify key genes underlying the tumorigenic processes of ESCC. Our study reveals many genes with aberrant DNA methylation, which may help to understand the tumor progression of ESCC and to develop novel treatment strategies for patients with ESCC. The eleven genes regulated by methylation are correlated with OS and DFS of ESCC patients and may be potential prognostic biomarkers. In addition, our study combines epigenomics and transcriptomics to provide new insight into the molecular basis of ESCC.

## MATERIALS AND METHODS

### Sample collection and preparation

Information of DNA methylation and gene expression was obtained from 15 patients [34], and clinical information was obtained from 125 patients, all from

the Chaoshan District of Guangdong Province, a region of high ESCC prevalence in China. Each sample, comprised of tumor and paired non-tumor tissues, was collected from an individual patient who underwent surgical resection from the Department of Oncological Surgery of the Central Hospital of Shantou City, China. Informed consent was obtained from the participants of this study. This study was approved by the Ethics Committee of the Central Hospital of Shantou City. The 125 ESCC patients' mRNA expression data used for survival analysis are available publicly at the GEO database under accession number GSE121931.

ChIP-seq data from 5 ESCC cell lines (KYSE510, KYSE140, TE5, TT and TE7) and WGBS data of 7 ESCC tumor tissues were generated previously, by Lin et.al. [59], and visualized in R with bigWig files. The level of H3K27ac histone modification and WGBS methylation was defined by the R package *Gviz*. RNA-Seq data of 85 ESCC patients and 87 EAC patients with matched clinical data were downloaded from TCGA (level 3 data). Five whole genome gene expression data of ESCC and five whole genome gene expression data of other cancers (pancreatic, gastric, colorectal, lung and esophagus adenous cancer) were downloaded from the GEO database. The 10 datasets were selected by the following criteria. First, we focused on gastrointestinal carcinomas, so we selected 4 datasets, including pancreatic cancer, gastric cancer, colorectal cancer and esophagus adenous cancer. Then, we selected a lung cancer dataset because the lung is adjacent to the esophagus. In addition, we also selected 5 ESCC datasets to test whether the results of our data were stable. Finally, all of these datasets should contain paired tumor and normal samples. The GEO accession numbers of the five ESCC datasets were GSE53622, GSE53624, GSE20347, GSE17351, and GSE23400, and the GEO accession numbers of the other cancers were GSE15471, GSE19826, GSE32323, GSE27262 and GSE1420. The Robust Multichip Average (RMA) algorithm was used for processing and normalizing the raw gene expression data.

### DNA methylation microarray

The Illumina Human Methylation450K Array was used to analyze the methylation status of 15 paired ESCC samples. This bead chip covers more than 480,000 methylation sites per sample. The probes were distributed across the promoter, first exon (1st exon), 5'UTR, gene body, and 3'UTR regions [60]. DNA methylation data are available publicly at the GEO database under accession number GSE121930. The raw data were processed by the following methods: 1) probes with a null value were removed, 2) probes

located in sex chromosomes were deleted, 3) probes that mapped to multiple genes or were not mapped to genes were removed, and 4) probes containing SNPs were excluded [61].

The Bioconductor R package *minfi* was used for quality control and normalization of the raw data. Probes with a  $p$ -value  $< 0.05$  were considered differentially methylated. Then the linear regression between DNA methylation  $\beta$  value and gene expression RPKM value for each region of a gene (TSS1500, TSS200, 5'UTR, 1st exon, gene body and 3'UTR) was used to assess which probes were most predictive of the gene expression state [20].

### Identification of differentially-expressed mRNAs

Fifteen ESCC sample expression profiles (SRP064894) were downloaded from NCBI [34]. We first used the "estimate" R package [62] to measure the tissues' purity of 15 ESCC patients and found that the purity of these 15 patients is more than 50% and the mean purity is 70%. We extracted this data by TopHat and identified the differentially-expressed mRNAs based on the count number expression profile using DESeq (Heidelberg, Germany). In this study, mRNAs with an absolute log2 fold change  $> 2$ , and  $p$ -value  $< 0.05$  are considered differentially-expressed.

### Integration of DNA methylation with gene expression

We integrated the 450K array and RNA-seq data by following analytical process: (1) identification of differentially-methylated and -expressed genes in 15 paired ESCC samples. The differentially-methylated genes were selected by a  $p$ -value  $< 0.05$ , and (2) testing of both differentially-methylated and differentially-expressed genes for a strong association. Pearson correlation analysis was used to identify their correlations. A negative correlation was considered significant for a Pearson correlation coefficient (PCC)  $< -0.5$  and  $p$ -value  $< 0.05$ .

The DAVID Bioinformatics Tool (version 6.8) [63] was used to infer the potential biological processes of methylation-associated genes. Results with  $p$ -value  $< 0.05$  were considered as significant functional categories.

### Cell culture and cell lines

Eight cell lines (KYSE140, KYSE150, KYSE180, KYSE450, KYSE510, TE3 and SHEEC) were used in this study [64]. All cell lines were incubated at 37°C in a humidified atmosphere containing 5% CO<sub>2</sub>.

## qRT-PCR

Total RNA from ESCC tissue or ESCC cell lines was isolated with TRIzol (Invitrogen) as per the manufacturer's instructions, and the concentration determined with a Nanodrop (Agilent). One microgram total RNA was reverse transcribed into cDNA by a Reverse Transcription System (Promega) according to the manufacturer's protocol [65]. Quantitative real-time PCR (qRT-PCR) was performed using GoTaq® qPCR Master Mix (TaKaRa) in a 7500 Real-Time PCR System (Applied Biosystems). Primer pairs for target genes used in the PCR assay are listed in Supplementary Table 5.

## Colony formation assay

Cell plate colony formation assays were performed as described previously [65, 66]. Briefly, four cell lines (KYSE140, KYSE150, KYSE180 and TE3) were treated with 100  $\mu$ M 5-aza-dC for 36 h before being trypsinized and plated into 6-well plates at a concentration of 2,000, 5,000 and 10,000 cells/well. The same four cell lines without 5-aza-dC treatment were used for the control. After washing with PBS, cultures were fixed with 4°C pre-cooled methanol for 15 min, and stained with hematoxylin for 15 min. Colonies were photographed and their sizes calculated with a FluorChem 8900 image analysis system (Alpha Innotech). Each experiment was performed in triplicate.

## Statistical analysis

SPSS 22.0 software (SPSS, Chicago, IL) and R (<https://www.r-project.org>) were used to perform the statistical analyses. Two-tailed independent sample t tests were used to determine whether differences were significant. Differences were considered statistically significant if the *p*-value < 0.05.

Time-dependent ROC curves were used to assess the prognostic efficiency of the key gene signatures. Kaplan–Meier survival analyses were used to test the survival distributions for ESCC samples. The chi-square test was performed to analyze the association with the clinical signatures, and multivariable Cox regression analysis was used to test whether the signature was independent of other clinical features. R was used to perform all of these analyses. Packages, including *pROC*, *survival* and *survivalROC*, were downloaded from Bioconductor.

## ACKNOWLEDGMENTS

We thank Dr. Stanley Li Lin from the Department of Cell Biology and Genetics of Shantou University Medical College for assistance in revising the manuscript.

## CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

## FUNDING

This work was supported in part by the National Cohort of Esophageal Cancer of China (grant No.2016YFC09014000), the National Science Foundation of China (No.81772532 and 81472613) and the Natural Science Foundation of China-Guangdong Joint Fund (No. U1601229).

## REFERENCES

1. Enzinger PC, Mayer RJ. Esophageal cancer. *N Engl J Med*. 2003; 349:2241–52.  
<https://doi.org/10.1056/NEJMra035010>  
PMID:[14657432](https://pubmed.ncbi.nlm.nih.gov/14657432/)
2. Pennathur A, Gibson MK, Jobe BA, Luketich JD. Oesophageal carcinoma. *Lancet*. 2013; 381:400–12.  
[https://doi.org/10.1016/S0140-6736\(12\)60643-6](https://doi.org/10.1016/S0140-6736(12)60643-6)  
PMID:[23374478](https://pubmed.ncbi.nlm.nih.gov/23374478/)
3. Brown LM, Devesa SS, Chow WH. Incidence of adenocarcinoma of the esophagus among white Americans by sex, stage, and age. *J Natl Cancer Inst*. 2008; 100:1184–87.  
<https://doi.org/10.1093/jnci/djn211>  
PMID:[18695138](https://pubmed.ncbi.nlm.nih.gov/18695138/)
4. Zerbini LF, Libermann TA. GADD45 deregulation in cancer: frequently methylated tumor suppressors and potential therapeutic targets. *Clin Cancer Res*. 2005; 11:6409–13.  
<https://doi.org/10.1158/1078-0432.CCR-05-1475>  
PMID:[16166414](https://pubmed.ncbi.nlm.nih.gov/16166414/)
5. Felsenfeld G. A brief history of epigenetics. *Cold Spring Harb Perspect Biol*. 2014; 6:a018200.  
<https://doi.org/10.1101/cshperspect.a018200>  
PMID:[24384572](https://pubmed.ncbi.nlm.nih.gov/24384572/)
6. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*. 2012; 13:484–92.  
<https://doi.org/10.1038/nrg3230> PMID:[22641018](https://pubmed.ncbi.nlm.nih.gov/22641018/)
7. Dawson SJ, Tsui DW, Murtaza M, Biggs H, Rueda OM, Chin SF, Dunning MJ, Gale D, Forshew T, Mahler-Araujo B, Rajan S, Humphray S, Becq J, et al. Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med*. 2013; 368:1199–209.  
<https://doi.org/10.1056/NEJMoa1213261>  
PMID:[23484797](https://pubmed.ncbi.nlm.nih.gov/23484797/)
8. Lebofsky R, Decraene C, Bernard V, Kamal M, Blin A, Leroy Q, Rio Frio T, Pierron G, Callens C, Bieche I, Saliou

- A, Madic J, Rouleau E, et al. Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. *Mol Oncol*. 2015; 9:783–90. <https://doi.org/10.1016/j.molonc.2014.12.003> PMID:25579085
9. Schwaederle M, Chattopadhyay R, Kato S, Fanta PT, Banks KC, Choi IS, Piccioni DE, Ikeda S, Talasaz A, Lanman RB, Bazhenova L, Kurzrock R. Genomic Alterations in Circulating Tumor DNA from Diverse Cancer Patients Identified by Next-Generation Sequencing. *Cancer Res*. 2017; 77:5419–27. <https://doi.org/10.1158/0008-5472.CAN-17-0885> PMID:28807936
  10. Heyn H, Vidal E, Sayols S, Sanchez-Mut JV, Moran S, Medina I, Sandoval J, Simó-Riudalbas L, Szczesna K, Huertas D, Gatto S, Matarazzo MR, Dopazo J, Esteller M. Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient. *Epigenetics*. 2012; 7:542–50. <https://doi.org/10.4161/epi.20523> PMID:22595875
  11. Legendre C, Gooden GC, Johnson K, Martinez RA, Liang WS, Salhia B. Whole-genome bisulfite sequencing of cell-free DNA identifies signature associated with metastatic breast cancer. *Clin Epigenetics*. 2015; 7:100. <https://doi.org/10.1186/s13148-015-0135-8> PMID:26380585
  12. Lou S, Lee HM, Qin H, Li JW, Gao Z, Liu X, Chan LL, Kl Lam V, So WY, Wang Y, Lok S, Wang J, Ma RC, et al. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biol*. 2014; 15:408. <https://doi.org/10.1186/s13059-014-0408-0> PMID:25074712
  13. Wong IH, Lo YM, Zhang J, Liew CT, Ng MH, Wong N, Lai PB, Lau WY, Hjelm NM, Johnson PJ. Detection of aberrant p16 methylation in the plasma and serum of liver cancer patients. *Cancer Res*. 1999; 59:71–73. PMID:9892188
  14. De Rubis G, Krishnan SR, Bebawy M. Circulating tumor DNA - Current state of play and future perspectives. *Pharmacol Res*. 2018; 136:35–44. <https://doi.org/10.1016/j.phrs.2018.08.017> PMID:30142423
  15. Hackanson B, Daskalakis M. Decitabine. *Recent Results Cancer Res*. 2014; 201:269–97. [https://doi.org/10.1007/978-3-642-54490-3\\_18](https://doi.org/10.1007/978-3-642-54490-3_18) PMID:24756800
  16. Jardin F. Next generation sequencing and the management of diffuse large B-cell lymphoma: from whole exome analysis to targeted therapy. *Discov Med*. 2014; 18:51–65. PMID:25091488
  17. Jiang T, Tan MS, Tan L, Yu JT. Application of next-generation sequencing technologies in Neurology. *Ann Transl Med*. 2014; 2:125. <https://doi.org/10.3978/j.issn.2305-5839.2014.11.11> PMID:25568878
  18. Greer EL, Shi Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet*. 2012; 13:343–57. <https://doi.org/10.1038/nrg3173> PMID:22473383
  19. Okugawa Y, Grady WM, Goel A. Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers. *Gastroenterology*. 2015; 149:1204–25.e12. <https://doi.org/10.1053/j.gastro.2015.07.011> PMID:26216839
  20. Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*. 2014; 30:2360–66. <https://doi.org/10.1093/bioinformatics/btu316> PMID:24794928
  21. Sawas T, Killcoyne S, Iyer PG, Wang KK, Smyrk TC, Kisiel JB, Qin Y, Ahlquist DA, Rustgi AK, Costa RJ, Gerstung M, Fitzgerald RC, Katzka DA, et al, and OCCAMS Consortium. Identification of prognostic phenotypes of esophageal adenocarcinoma in two independent cohorts. *Gastroenterology*. 2018; 155:1720–1728.e4. <https://doi.org/10.1053/j.gastro.2018.08.036> PMID:30165050
  22. Shao Y, Geng Y, Gu W, Ning Z, Huang J, Pei H, Jiang J. Assessment of Lymph Node Ratio to Replace the pN Categories System of Classification of the TNM System in Esophageal Squamous Cell Carcinoma. *J Thorac Oncol*. 2016; 11:1774–84. <https://doi.org/10.1016/j.jtho.2016.06.019> PMID:27393473
  23. Lagarde SM, ten Kate FJ, Reitsma JB, Busch OR, van Lanschot JJ. Prognostic factors in adenocarcinoma of the esophagus or gastroesophageal junction. *J Clin Oncol*. 2006; 24:4347–55. <https://doi.org/10.1200/JCO.2005.04.9445> PMID:16963732
  24. Peters CJ, Rees JR, Hardwick RH, Hardwick JS, Vowler SL, Ong CA, Zhang C, Save V, O'Donovan M, Rassl D, Alderson D, Caldas C, Fitzgerald RC; Oesophageal Cancer Clinical and Molecular Stratification (OCCAMS) Study Group. A 4-gene signature predicts survival of patients with resected adenocarcinoma of the esophagus, junction, and gastric cardia. *Gastroenterology*. 2010; 139:1995–2004.e15. <https://doi.org/10.1053/j.gastro.2010.05.080> PMID:20621683

25. Hamm CA, Costa FF. Epigenomes as therapeutic targets. *Pharmacol Ther*. 2015; 151:72–86. <https://doi.org/10.1016/j.pharmthera.2015.03.003> PMID:25797698
26. Padmanabhan N, Ushijima T, Tan P. How to stomach an epigenetic insult: the gastric cancer epigenome. *Nat Rev Gastroenterol Hepatol*. 2017; 14:467–78. <https://doi.org/10.1038/nrgastro.2017.53> PMID:28513632
27. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. *Cell*. 2013; 155:934–47. <https://doi.org/10.1016/j.cell.2013.09.053> PMID:24119843
28. He JZ, Wu ZY, Wang SH, Ji X, Yang CX, Xu XE, Liao LD, Wu JY, Li EM, Zhang K, Xu LY. A decision tree-based combination of ezrin-interacting proteins to estimate the prognostic risk of patients with esophageal squamous cell carcinoma. *Hum Pathol*. 2017; 66:115–25. <https://doi.org/10.1016/j.humpath.2017.06.003> PMID:28603065
29. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. *Lancet*. 2018; 392:777–86. [https://doi.org/10.1016/S0140-6736\(18\)31268-6](https://doi.org/10.1016/S0140-6736(18)31268-6) PMID:30100054
30. Koch A, Joosten SC, Feng Z, de Ruijter TC, Draht MX, Melotte V, Smits KM, Veeck J, Herman JG, Van Neste L, Van Criekinge W, De Meyer T, van Engeland M. Analysis of DNA methylation in cancer: location revisited. *Nat Rev Clin Oncol*. 2018; 15:459–66. <https://doi.org/10.1038/s41571-018-0004-4> PMID:29666440
31. Lin DC, Wang MR, Koeffler HP. Genomic and Epigenomic Aberrations in Esophageal Squamous Cell Carcinoma and Implications for Patients. *Gastroenterology*. 2018; 154:374–89. <https://doi.org/10.1053/j.gastro.2017.06.066> PMID:28757263
32. Jiang Y, Jiang YY, Xie JJ, Mayakonda A, Hazawa M, Chen L, Xiao JF, Li CQ, Huang ML, Ding LW, Sun QY, Xu L, Kanojia D, et al. Co-activation of super-enhancer-driven CCAT1 by TP63 and SOX2 promotes squamous cancer progression. *Nat Commun*. 2018; 9:3619. <https://doi.org/10.1038/s41467-018-06081-9> PMID:30190462
33. Guo JC, Wu Y, Chen Y, Pan F, Wu ZY, Zhang JS, Wu JY, Xu XE, Zhao JM, Li EM, Zhao Y, Xu LY. Protein-coding genes combined with long noncoding RNA as a novel transcriptome molecular staging model to predict the survival of patients with esophageal squamous cell carcinoma. *Cancer Commun (Lond)*. 2018; 38:4. <https://doi.org/10.1186/s40880-018-0277-0> PMID:29784063
34. Li CQ, Huang GW, Wu ZY, Xu YJ, Li XC, Xue YJ, Zhu Y, Zhao JM, Li M, Zhang J, Wu JY, Lei F, Wang QY, et al. Integrative analyses of transcriptome sequencing identify novel functional lncRNAs in esophageal squamous cell carcinoma. *Oncogenesis*. 2017; 6:e297. <https://doi.org/10.1038/oncsis.2017.1> PMID:28194033
35. Robles AI, Arai E, Mathé EA, Okayama H, Schetter AJ, Brown D, Petersen D, Bowman ED, Noro R, Welsh JA, Edelman DC, Stevenson HS, Wang Y, et al. An Integrated Prognostic Classifier for Stage I Lung Adenocarcinoma Based on mRNA, microRNA, and DNA Methylation Biomarkers. *J Thorac Oncol*. 2015; 10:1037–48. <https://doi.org/10.1097/JTO.0000000000000560> PMID:26134223
36. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, Siegmund KD, Koss MN, Hagen JA, Lam WL, Lam S, Gazdar AF, Laird-Offringa IA. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res*. 2012; 22:1197–211. <https://doi.org/10.1101/gr.132662.111> PMID:22613842
37. Farshidfar F, Zheng S, Gingras MC, Newton Y, Shih J, Robertson AG, Hinoue T, Hoadley KA, Gibb EA, Roszik J, Covington KR, Wu CC, Shinbrot E, et al, and Cancer Genome Atlas Network. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell Rep*. 2017; 18:2780–94. <https://doi.org/10.1016/j.celrep.2017.02.033> PMID:28297679
38. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol*. 2015; 33:364–76. <https://doi.org/10.1038/nbt.3157> PMID:25690853
39. Chou RH, Lin SC, Wen HC, Wu CW, Chang WS. Epigenetic activation of human kallikrein 13 enhances malignancy of lung adenocarcinoma by promoting N-cadherin expression and laminin degradation. *Biochem Biophys Res Commun*. 2011; 409:442–47. <https://doi.org/10.1016/j.bbrc.2011.05.022> PMID:21596022
40. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpour S, Danielsson A, Edlund K, Asplund A, Sjöstedt E, Lundberg E, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*. 2014; 13:397–406.

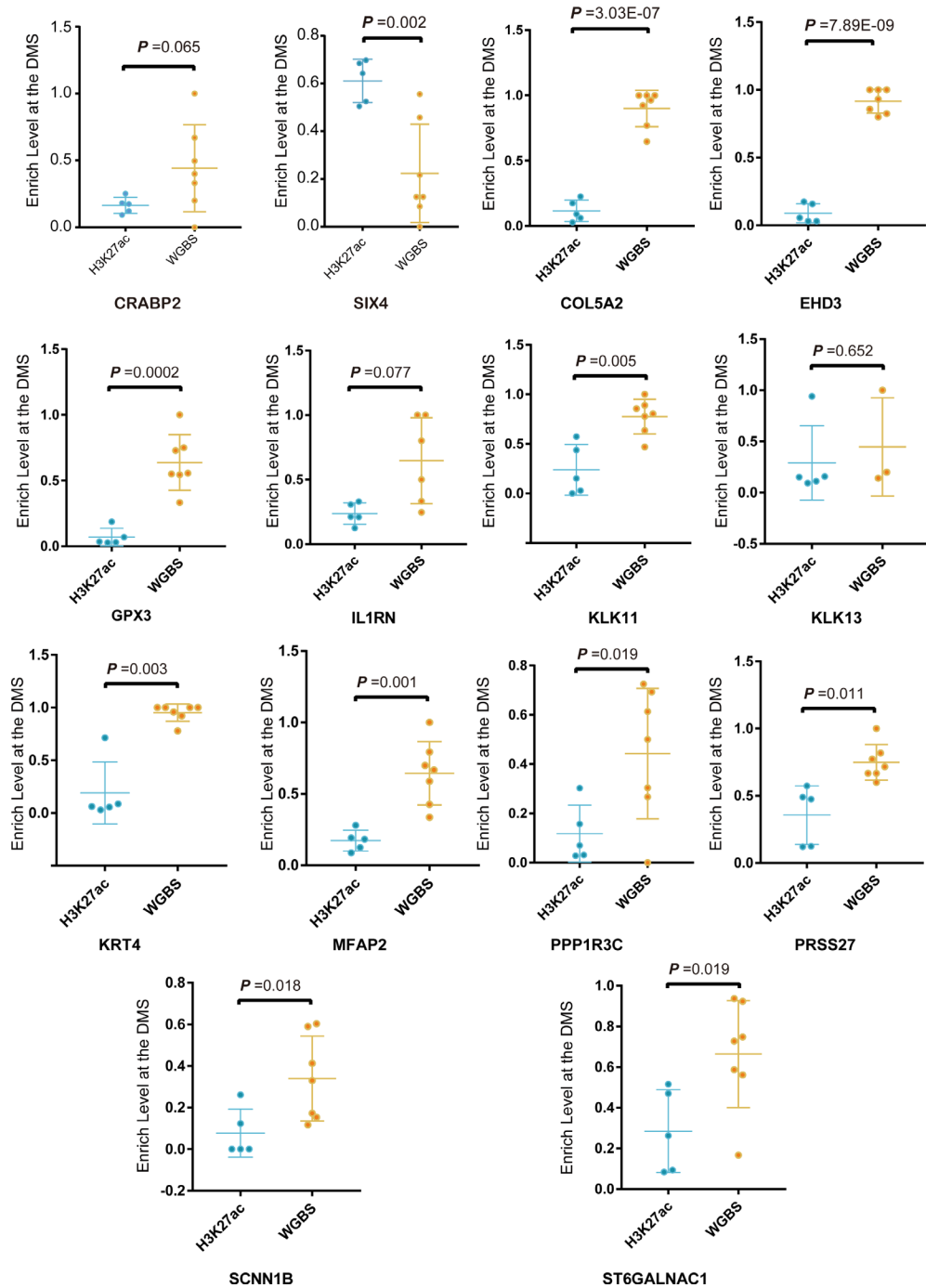
- <https://doi.org/10.1074/mcp.M113.035600>  
PMID:24309898
41. Fischer H, Stenling R, Rubio C, Lindblom A. Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. *Carcinogenesis*. 2001; 22:875–78. <https://doi.org/10.1093/carcin/22.6.875>  
PMID:11375892
42. Guillen-Ahlers H, Buechler SA, Suckow MA, Castellino FJ, Ploplis VA. Sulindac treatment alters collagen and matrilysin expression in adenomas of ApcMin/+ mice. *Carcinogenesis*. 2008; 29:1421–27. <https://doi.org/10.1093/carcin/bgn123>  
PMID:18499699
43. Sakamoto K, Aragaki T, Morita K, Kawachi H, Kayamori K, Nakanishi S, Omura K, Miki Y, Okada N, Katsube K, Takizawa T, Yamaguchi A. Down-regulation of keratin 4 and keratin 13 expression in oral squamous cell carcinoma and epithelial dysplasia: a clue for histopathogenesis. *Histopathology*. 2011; 58:531–42. <https://doi.org/10.1111/j.1365-2559.2011.03759.x>  
PMID:21371075
44. Wang JK, Wang WJ, Cai HY, Du BB, Mai P, Zhang LJ, Ma W, Hu YG, Feng SF, Miao GY. MFAP2 promotes epithelial-mesenchymal transition in gastric cancer cells by activating TGF- $\beta$ /SMAD2/3 signaling pathway. *Onco Targets Ther*. 2018; 11:4001–17. <https://doi.org/10.2147/OTT.S160831> PMID:30034240
45. Turan I, Kotan LD, Tastan M, Gurbuz F, Topaloglu AK, Yuksel B. Molecular genetic studies in a case series of isolated hypoaldosteronism due to biosynthesis defects or aldosterone resistance. *Clin Endocrinol (Oxf)*. 2018; 88:799–805. <https://doi.org/10.1111/cen.13603>  
PMID:29582446
46. Liu Y, Han N, Zhou S, Zhou R, Yuan X, Xu H, Zhang C, Yin T, Wu K. The DACH/EYA/SIX gene network and its role in tumor initiation and progression. *Int J Cancer*. 2016; 138:1067–75. <https://doi.org/10.1002/ijc.29560>  
PMID:26096807
47. Donato LJ, Noy N. Suppression of mammary carcinoma growth by retinoic acid: proapoptotic genes are targets for retinoic acid receptor and cellular retinoic acid-binding protein II signaling. *Cancer Res*. 2005; 65:8193–99. <https://doi.org/10.1158/0008-5472.CAN-05-1177>  
PMID:16166294
48. Donato LJ, Suh JH, Noy N. Suppression of mammary carcinoma cell growth by retinoic acid: the cell cycle control gene Btg2 is a direct target for retinoic acid receptor signaling. *Cancer Res*. 2007; 67:609–15. <https://doi.org/10.1158/0008-5472.CAN-06-0989>  
PMID:17234770
49. Nevalainen T, Kananen L, Marttila S, Jylhä M, Hervonen A, Hurme M, Jylhävä J. Transcriptomic and epigenetic analyses reveal a gender difference in aging-associated inflammation: the Vitality 90+ study. *Age (Dordr)*. 2015; 37:9814. <https://doi.org/10.1007/s11357-015-9814-9>  
PMID:26188803
50. Galperin E, Benjamin S, Rapaport D, Rotem-Yehudar R, Tolchinsky S, Horowitz M. EHD3: a protein that resides in recycling tubular and vesicular membrane structures and interacts with EHD1. *Traffic*. 2002; 3:575–89. <https://doi.org/10.1034/j.1600-0854.2002.30807.x>  
PMID:12121420
51. Naslavsky N, McKenzie J, Altan-Bonnet N, Sheff D, Caplan S. EHD3 regulates early-endosome-to-Golgi transport and preserves Golgi morphology. *J Cell Sci*. 2009; 122:389–400. <https://doi.org/10.1242/jcs.037051> PMID:19139087
52. Chukkapalli S, Amessou M, Dekhil H, Dilly AK, Liu Q, Bandyopadhyay S, Thomas RD, Bejna A, Batist G, Kandouz M. Ehd3, a regulator of vesicular trafficking, is silenced in gliomas and functions as a tumor suppressor by controlling cell cycle arrest and apoptosis. *Carcinogenesis*. 2014; 35:877–85. <https://doi.org/10.1093/carcin/bgt399>  
PMID:24306026
53. Lee OJ, Schneider-Stock R, McChesney PA, Kuester D, Roessner A, Vieth M, Moskaluk CA, El-Rifai W. Hypermethylation and loss of expression of glutathione peroxidase-3 in Barrett's tumorigenesis. *Neoplasia*. 2005; 7:854–61. <https://doi.org/10.1593/neo.05328>  
PMID:16229808
54. Yu YP, Yu G, Tseng G, Cieply K, Nelson J, Defrances M, Zarnegar R, Michalopoulos G, Luo JH. Glutathione peroxidase 3, deleted or methylated in prostate cancer, suppresses prostate cancer growth and metastasis. *Cancer Res*. 2007; 67:8043–50. <https://doi.org/10.1158/0008-5472.CAN-07-0648>  
PMID:17804715
55. Peng DF, Razvi M, Chen H, Washington K, Roessner A, Schneider-Stock R, El-Rifai W. DNA hypermethylation regulates the expression of members of the Mu-class glutathione S-transferases and glutathione peroxidases in Barrett's adenocarcinoma. *Gut*. 2009; 58:5–15. <https://doi.org/10.1136/gut.2007.146290>  
PMID:18664505
56. Diaz-Lagares A, Crujeiras AB, Lopez-Serra P, Soler M, Setien F, Goyal A, Sandoval J, Hashimoto Y, Martinez-Cardús A, Gomez A, Heyn H, Moutinho C, Espada J, et al. Epigenetic inactivation of the p53-induced long noncoding RNA TP53 target 1 in human cancer. *Proc Natl Acad Sci USA*. 2016; 113:E7535–44.



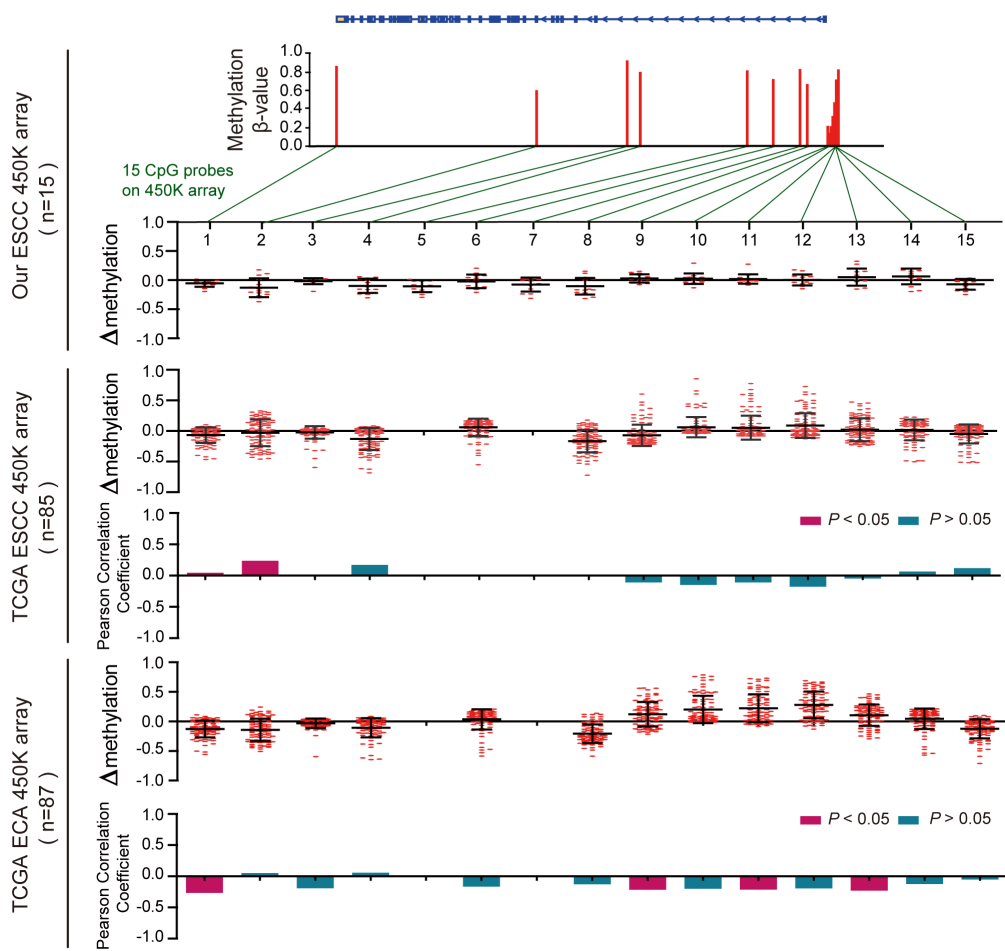
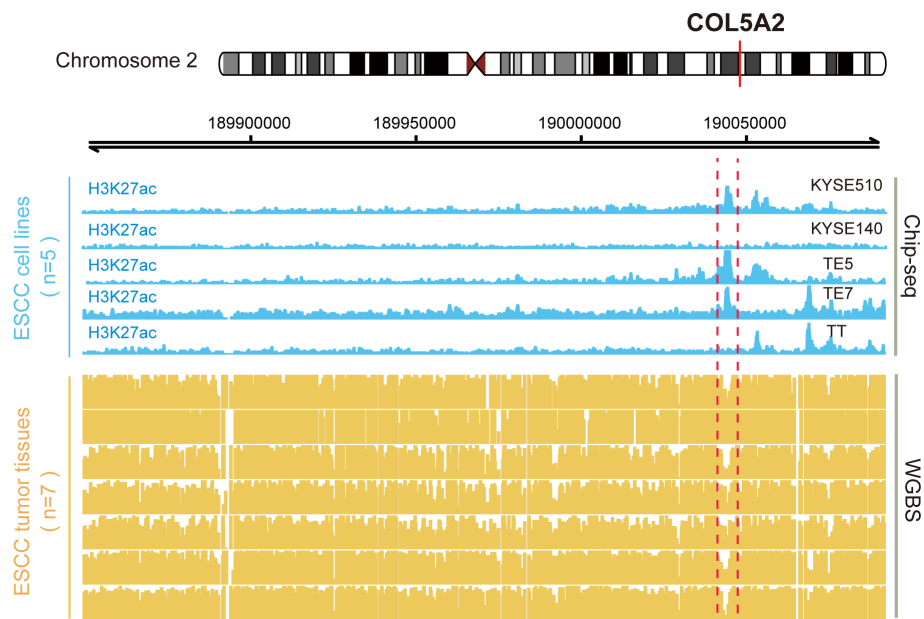
- <https://doi.org/10.1073/pnas.1608585113>  
PMID:[27821766](https://pubmed.ncbi.nlm.nih.gov/27821766/)
57. Yang Y, Chen L, Gu J, Zhang H, Yuan J, Lian Q, Lv G, Wang S, Wu Y, Yang YT, Wang D, Liu Y, Tang J, et al. Recurrently deregulated lncRNAs in hepatocellular carcinoma. *Nat Commun.* 2017; 8:14421. <https://doi.org/10.1038/ncomms14421>  
PMID:[28194035](https://pubmed.ncbi.nlm.nih.gov/28194035/)
58. Guo JC, Fang SS, Wu Y, Zhang JH, Chen Y, Liu J, Wu B, Wu JR, Li EM, Xu LY, Sun L, Zhao Y. CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic Acids Res.* 2019; 47:W516–22. <https://doi.org/10.1093/nar/gkz400>  
PMID:[31147700](https://pubmed.ncbi.nlm.nih.gov/31147700/)
59. Lin DC, Dinh HQ, Xie JJ, Mayakonda A, Silva TC, Jiang YY, Ding LW, He JZ, Xu XE, Hao JJ, Wang MR, Li C, Xu LY, et al. Identification of distinct mutational patterns and new driver genes in oesophageal squamous cell carcinomas and adenocarcinomas. *Gut.* 2018; 67:1769–79. <https://doi.org/10.1136/gutjnl-2017-314607>  
PMID:[28860350](https://pubmed.ncbi.nlm.nih.gov/28860350/)
60. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R. High density DNA methylation array with single CpG site resolution. *Genomics.* 2011; 98:288–95. <https://doi.org/10.1016/j.ygeno.2011.07.007>  
PMID:[21839163](https://pubmed.ncbi.nlm.nih.gov/21839163/)
61. van Dongen J, Ehli EA, Sliker RC, Bartels M, Weber ZM, Davies GE, Slagboom PE, Heijmans BT, Boomsma DI. Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. *Genes (Basel).* 2014; 5:347–65. <https://doi.org/10.3390/genes5020347>  
PMID:[24802513](https://pubmed.ncbi.nlm.nih.gov/24802513/)
62. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-García W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun.* 2013; 4:2612. <https://doi.org/10.1038/ncomms3612>  
PMID:[24113773](https://pubmed.ncbi.nlm.nih.gov/24113773/)
63. Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009; 4:44–57. <https://doi.org/10.1038/nprot.2008.211>  
PMID:[19131956](https://pubmed.ncbi.nlm.nih.gov/19131956/)
64. Lv GQ, Zou HY, Liao LD, Cao HH, Zeng FM, Wu BL, Xie JJ, Fang WK, Xu LY, Li EM. Identification of a novel lysyl oxidase-like 2 alternative splicing isoform, LOXL2 Δe13, in esophageal squamous cell carcinoma. *Biochem Cell Biol.* 2014; 92:379–89. <https://doi.org/10.1139/bcb-2014-0046>  
PMID:[25275797](https://pubmed.ncbi.nlm.nih.gov/25275797/)
65. Long L, Pang XX, Lei F, Zhang JS, Wang W, Liao LD, Xu XE, He JZ, Wu JY, Wu ZY, Wang LD, Lin DC, Li EM, Xu LY. SLC52A3 expression is activated by NF-κB p65/Rel-B and serves as a prognostic biomarker in esophageal cancer. *Cell Mol Life Sci.* 2018; 75:2643–61. <https://doi.org/10.1007/s00018-018-2757-4>  
PMID:[29428966](https://pubmed.ncbi.nlm.nih.gov/29428966/)
66. Zhang HF, Chen Y, Wu C, Wu ZY, Twardy DJ, Alshareef A, Liao LD, Xue YJ, Wu JY, Chen B, Xu XE, Gopal K, Gupta N, et al. The Opposing Function of STAT3 as an Oncoprotein and Tumor Suppressor Is Dictated by the Expression Status of STAT3β in Esophageal Squamous Cell Carcinoma. *Clin Cancer Res.* 2016; 22:691–703. <https://doi.org/10.1158/1078-0432.CCR-15-1253>  
PMID:[26405196](https://pubmed.ncbi.nlm.nih.gov/26405196/)

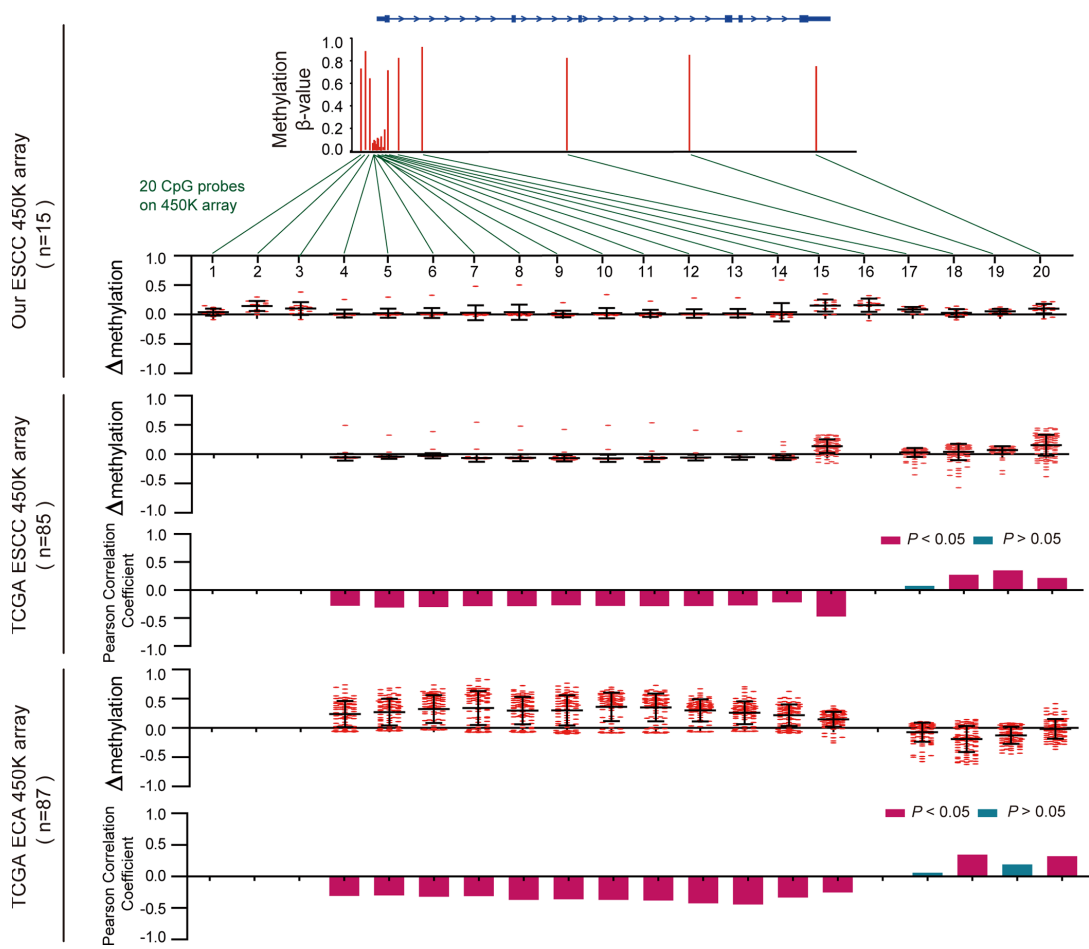
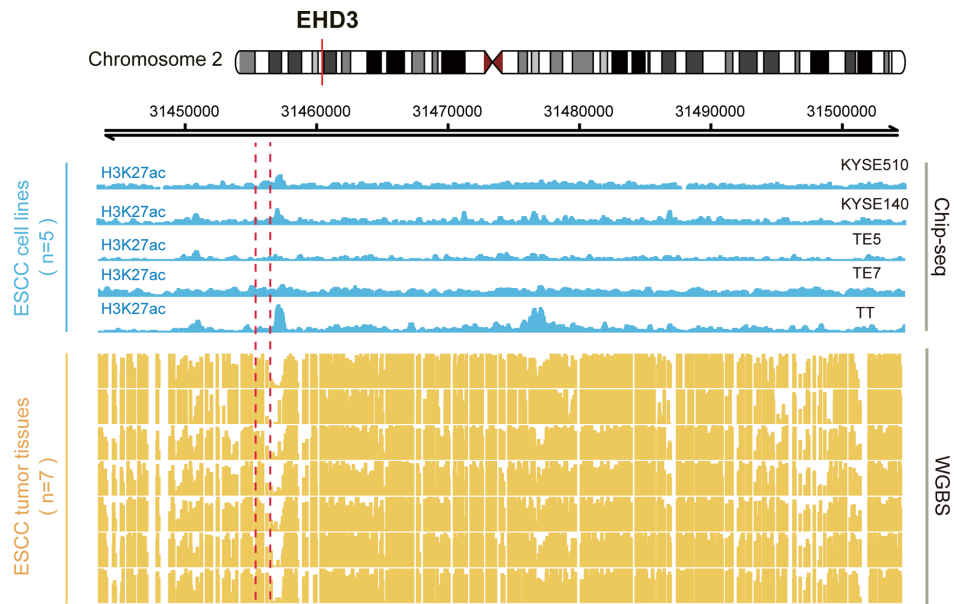
# SUPPLEMENTARY MATERIALS

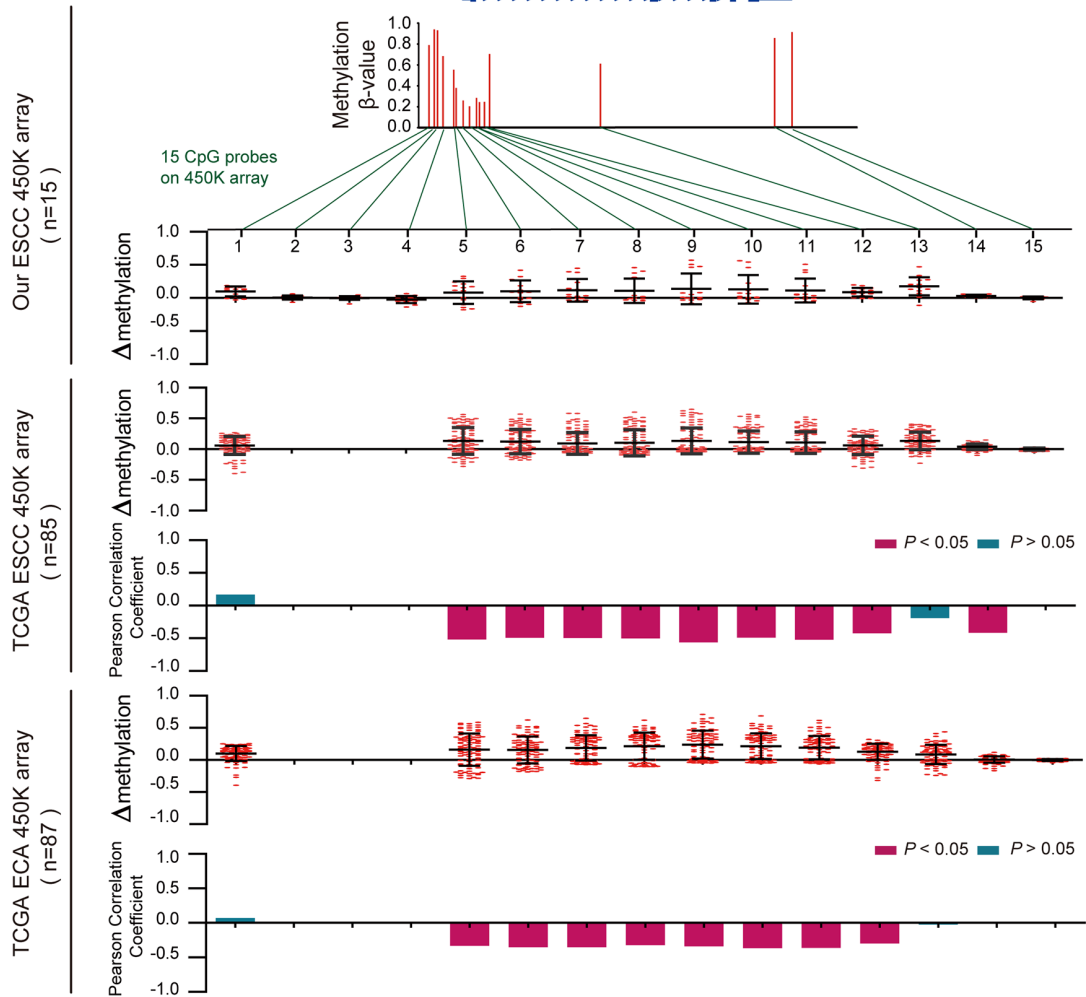
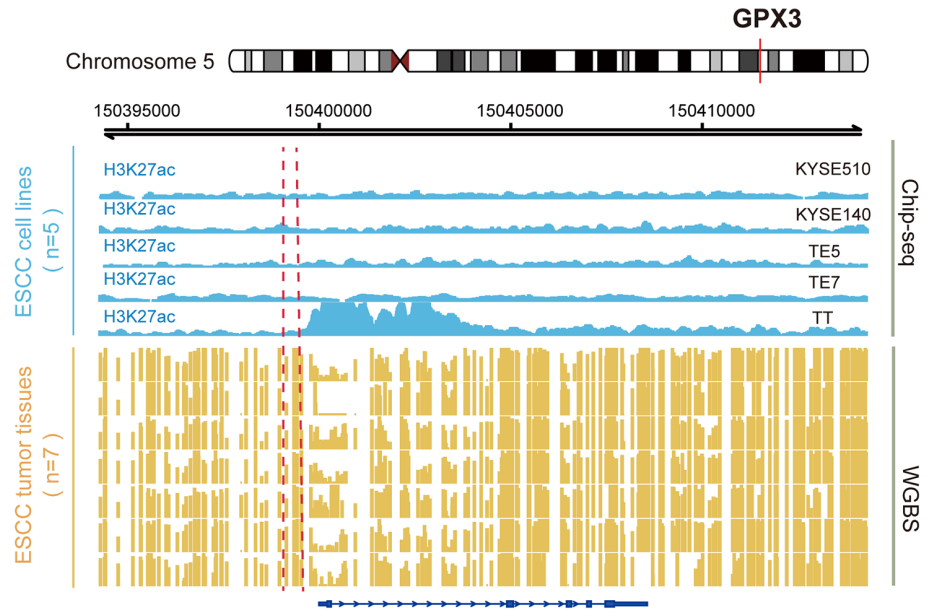
## Supplementary Figures

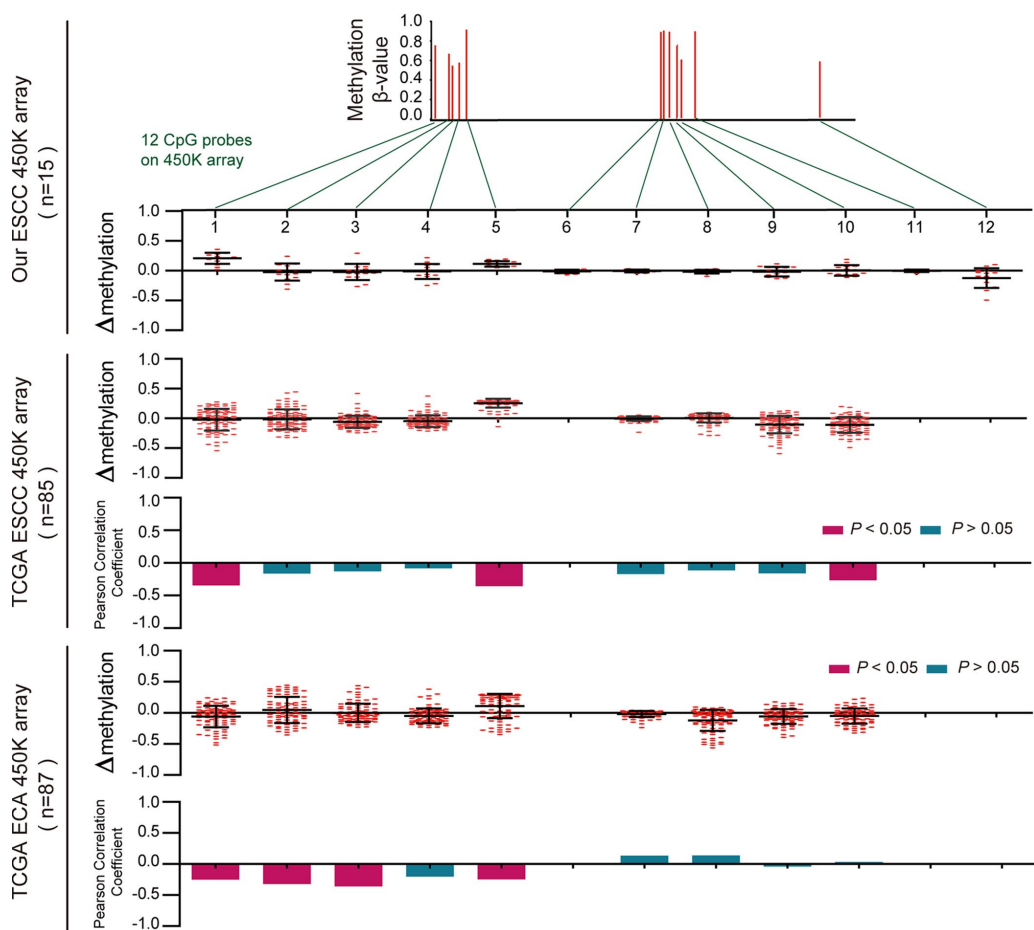
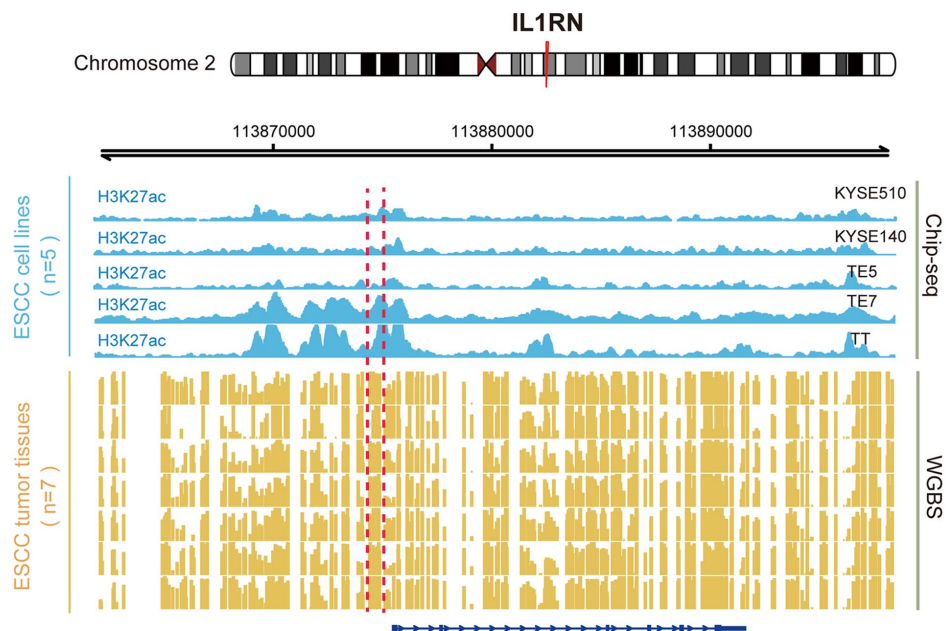


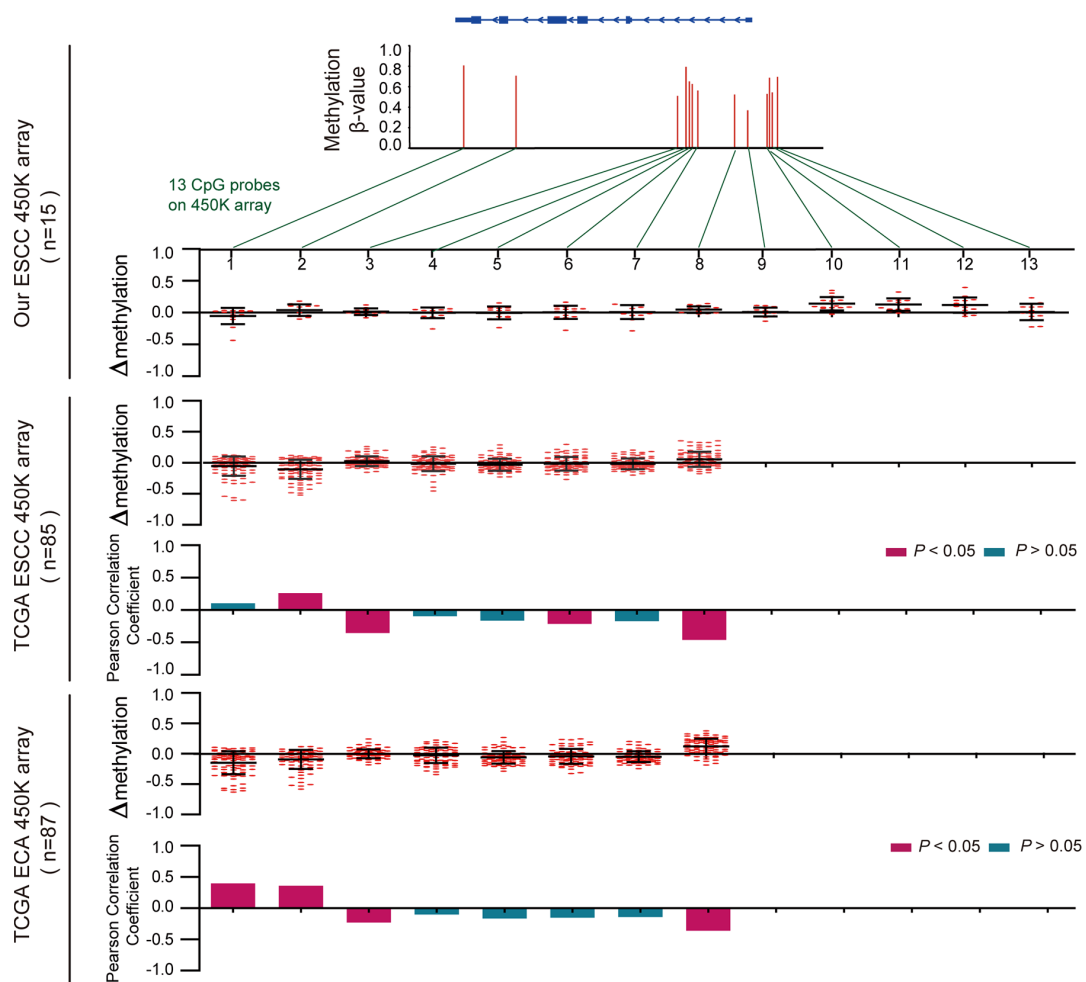
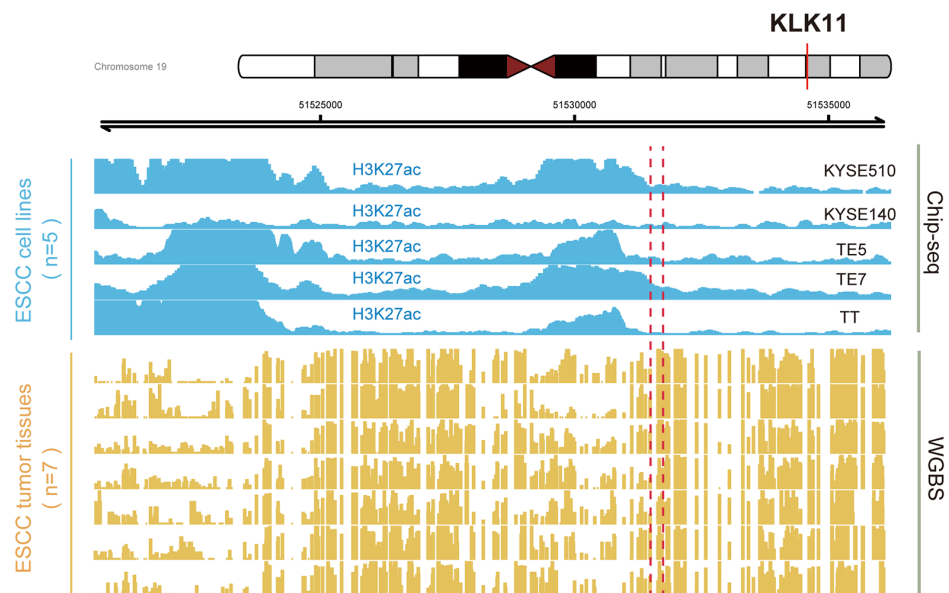
**Supplementary Figure 1. Differential analysis between H3K27ac histone modification and WGBS in the differentially methylated site (DMS) of 14 key genes.** We detected the enrich level of a 100bp region which contain the DMS. Scatter plot showed the p-value of the differential analysis. Blue spots represent H3K27ac enrich level in 5 ESCC cell lines, whereas the yellow spots represent WGBS level in 7 ESCC samples. T-test were used to determine whether differences were significant. The statistically significant were considered with p-value < 0.05.

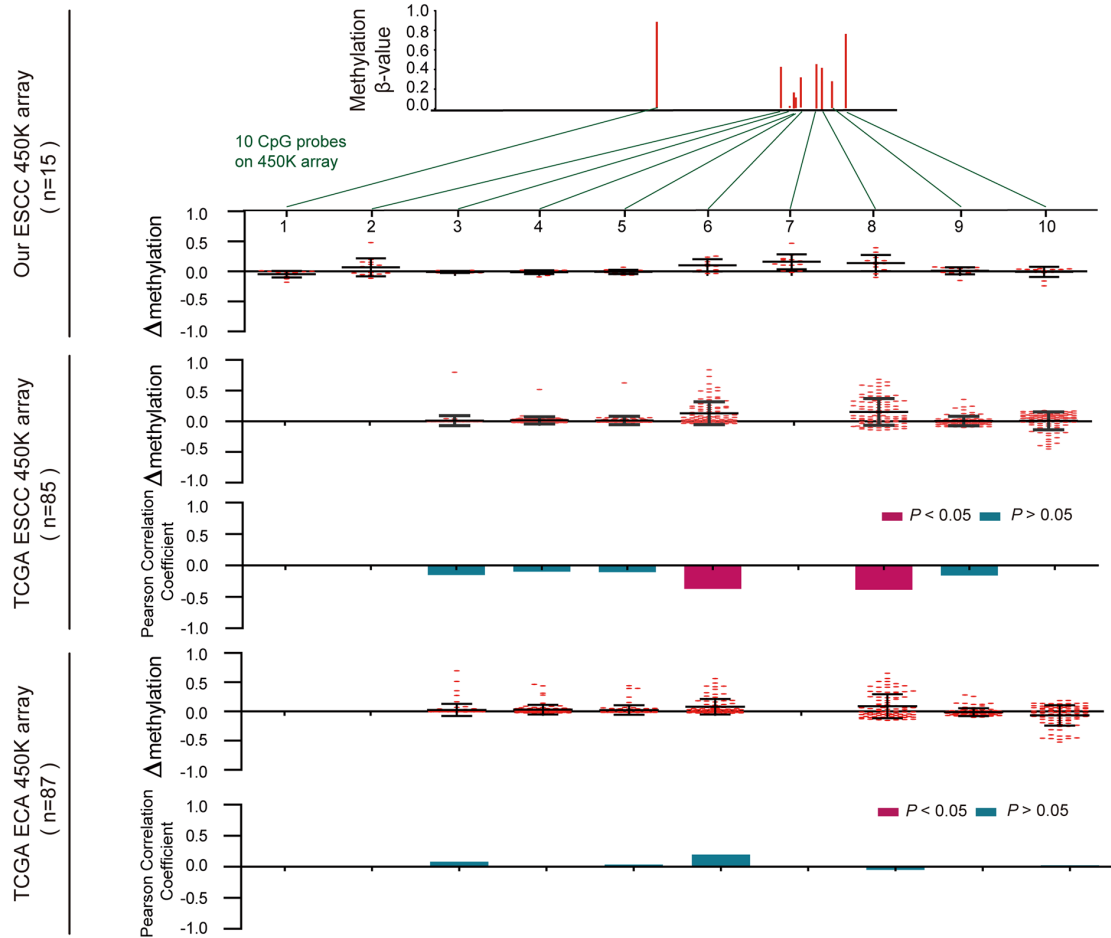
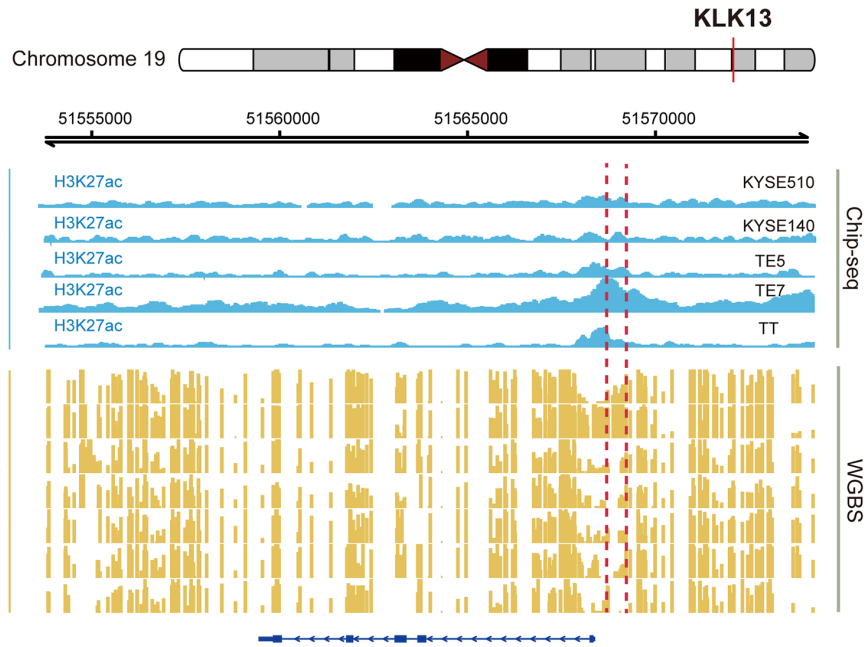




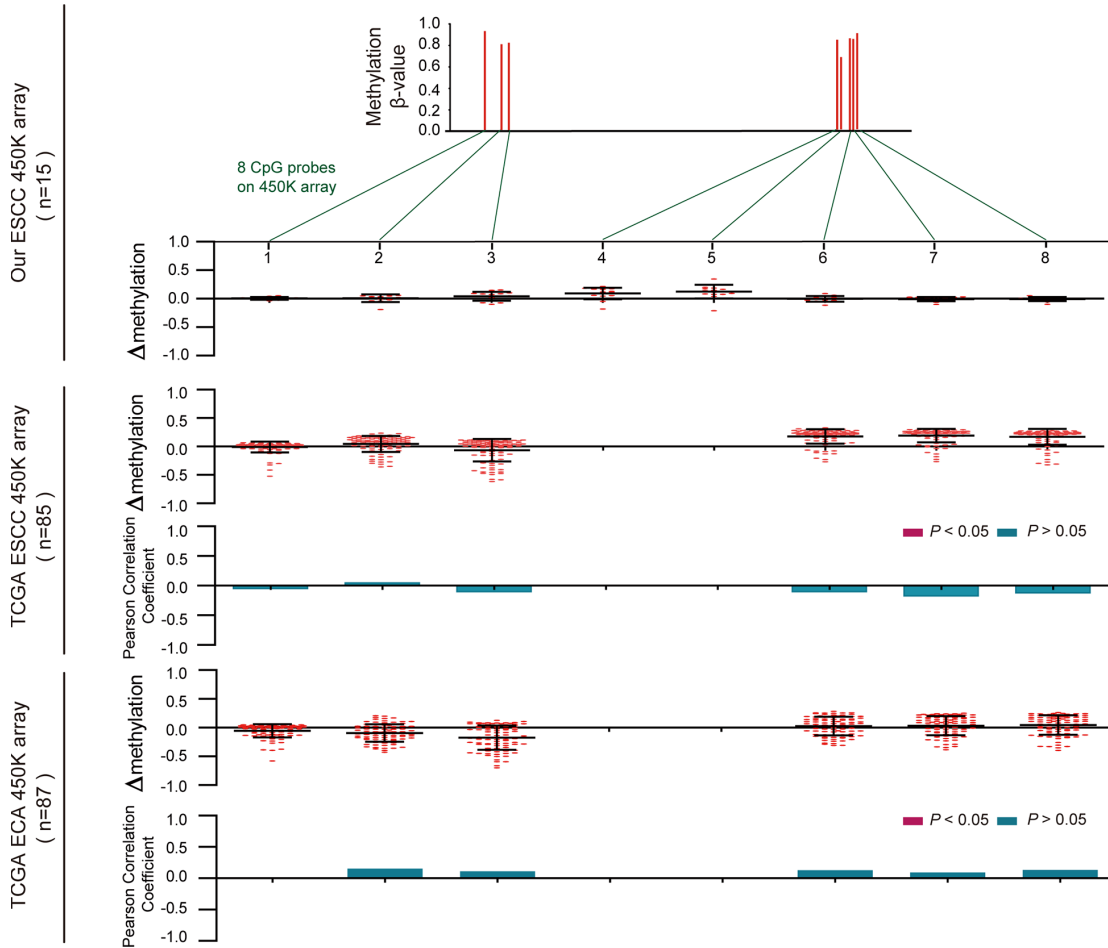
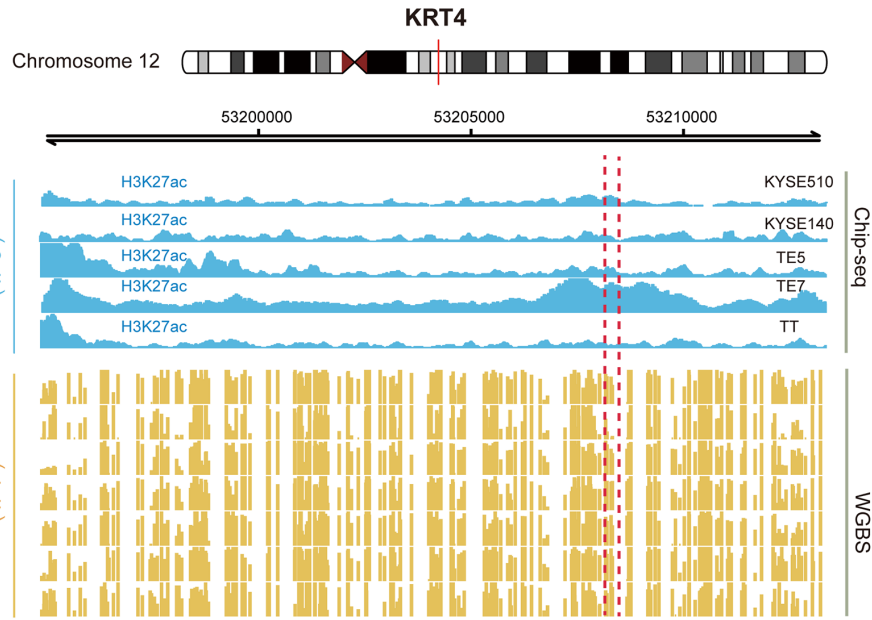


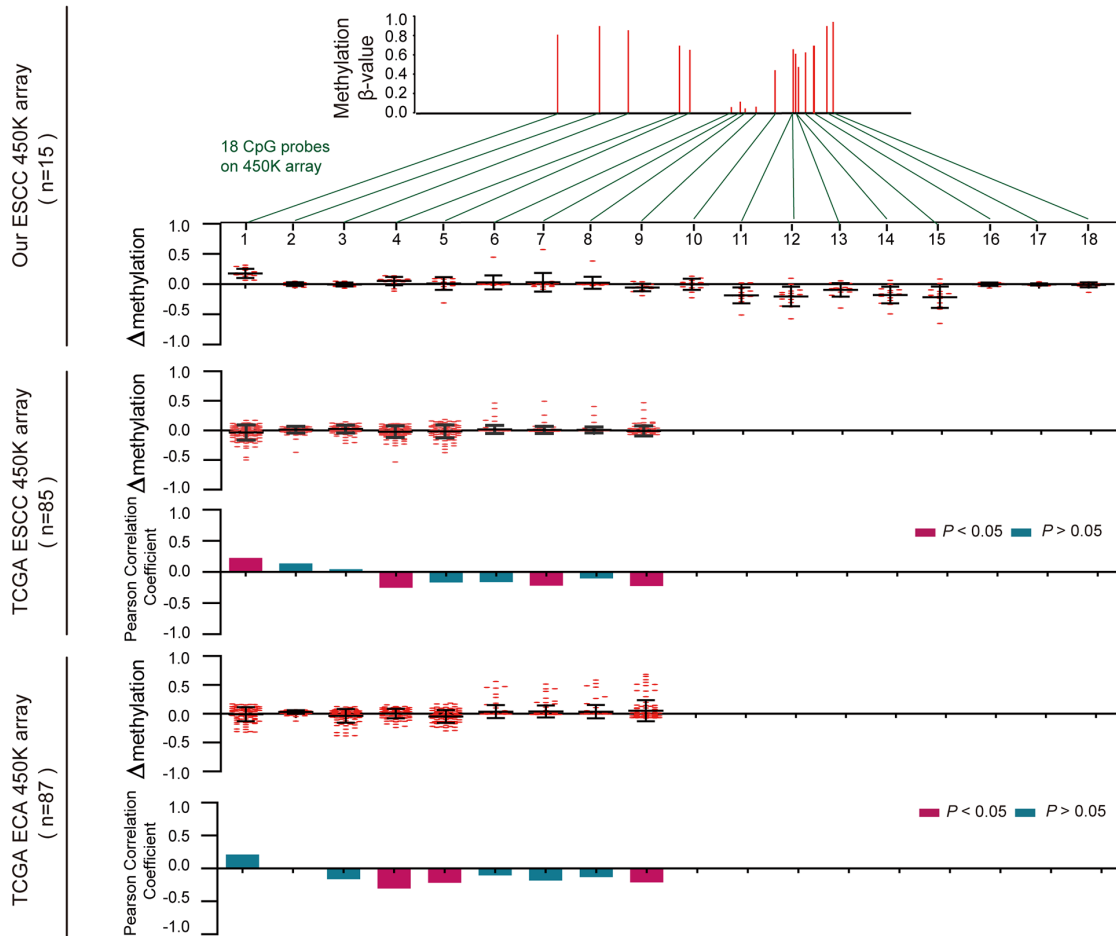
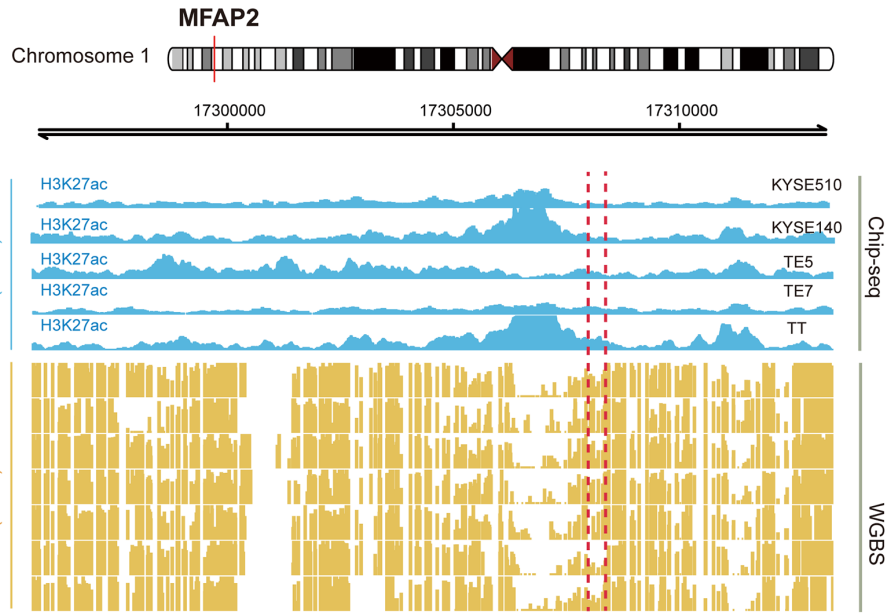


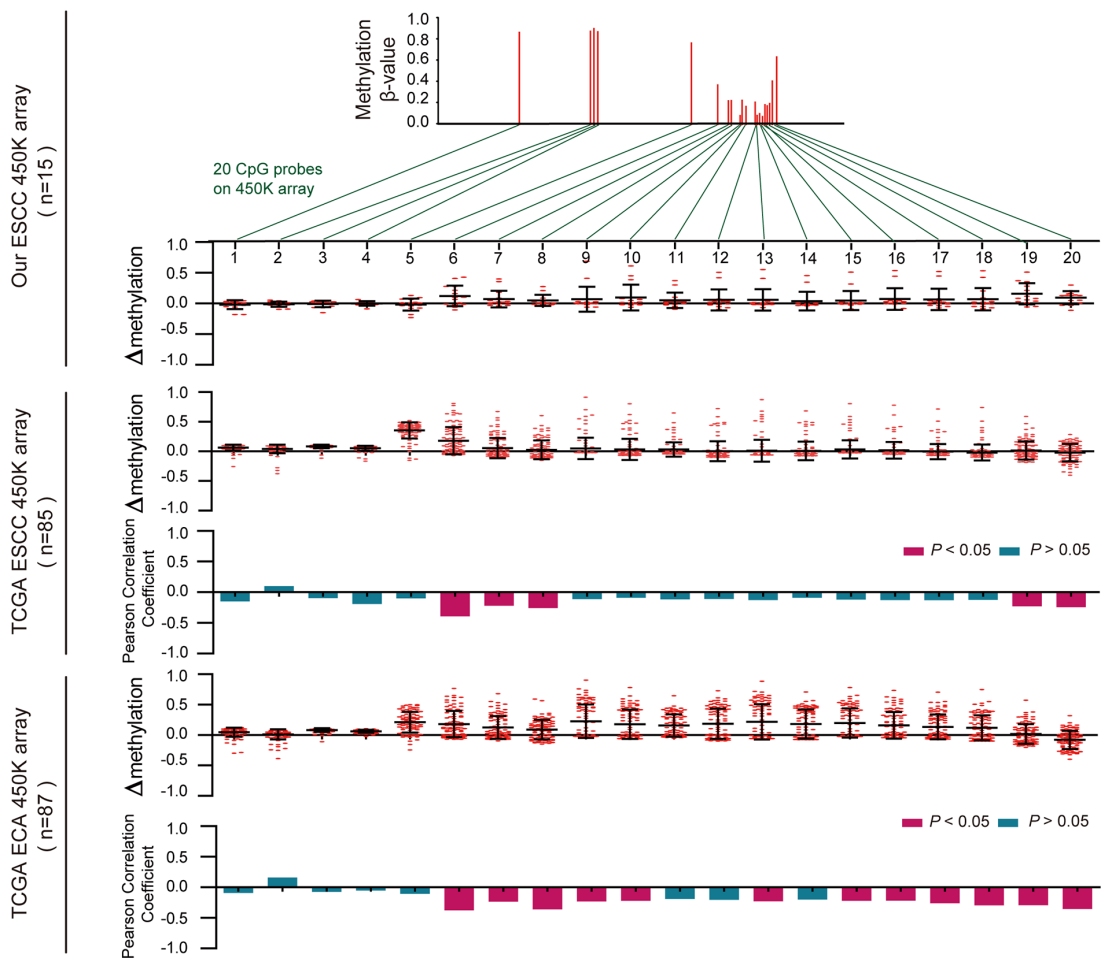
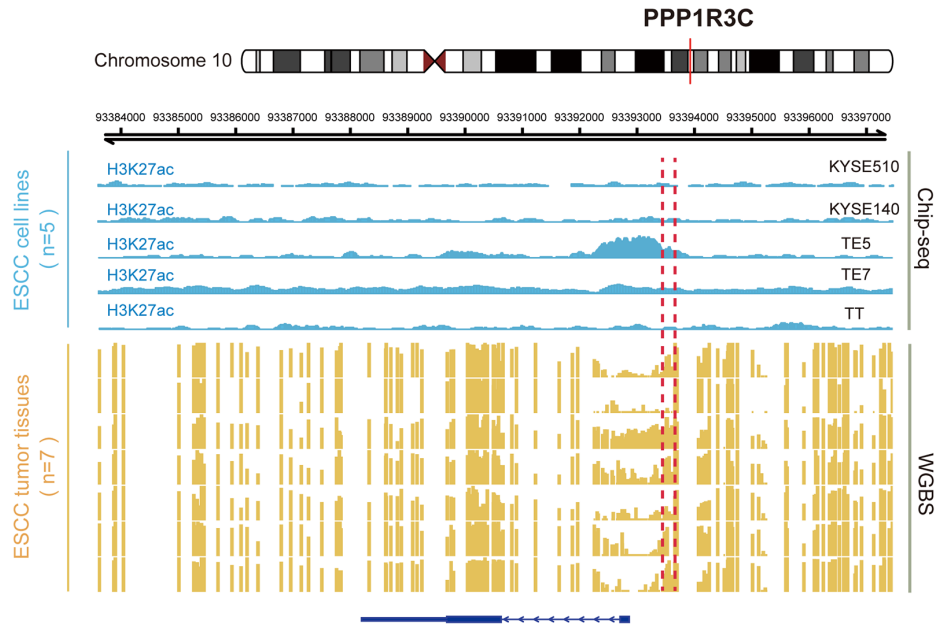


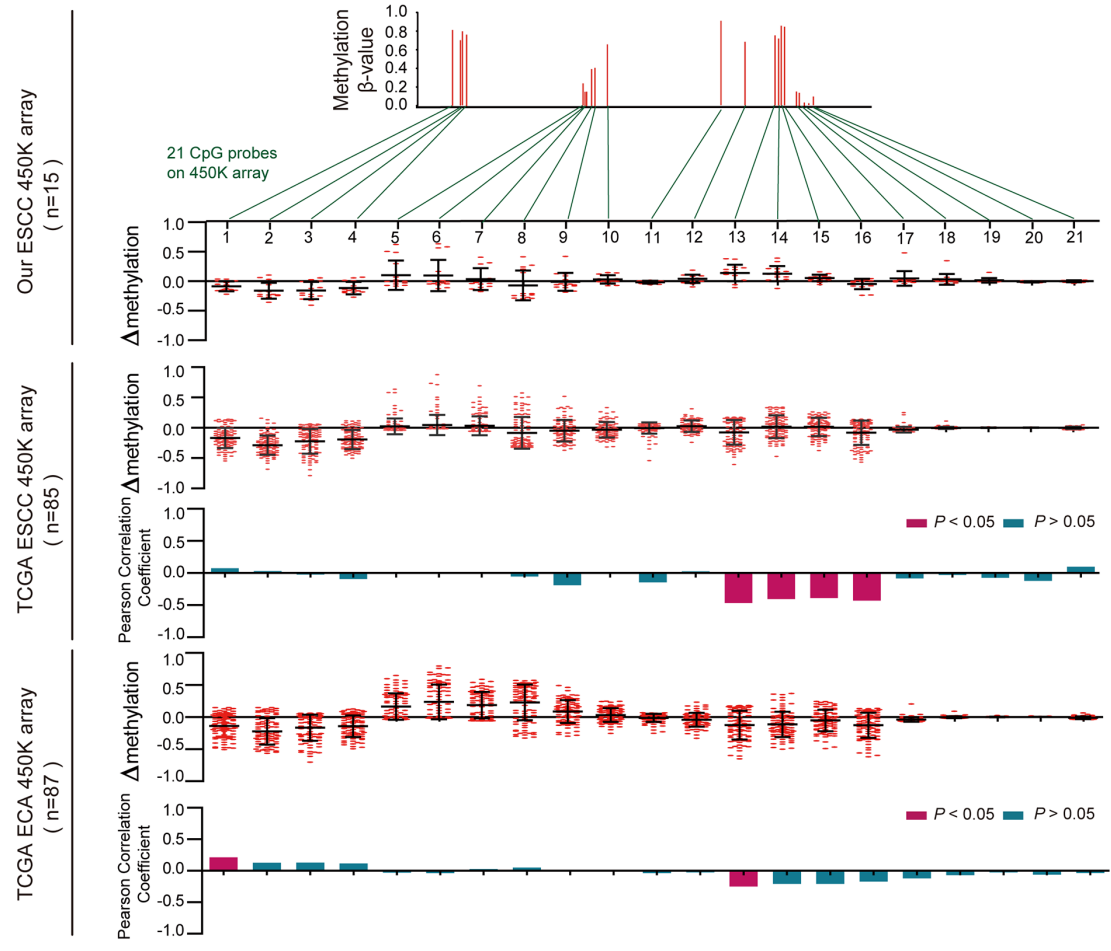
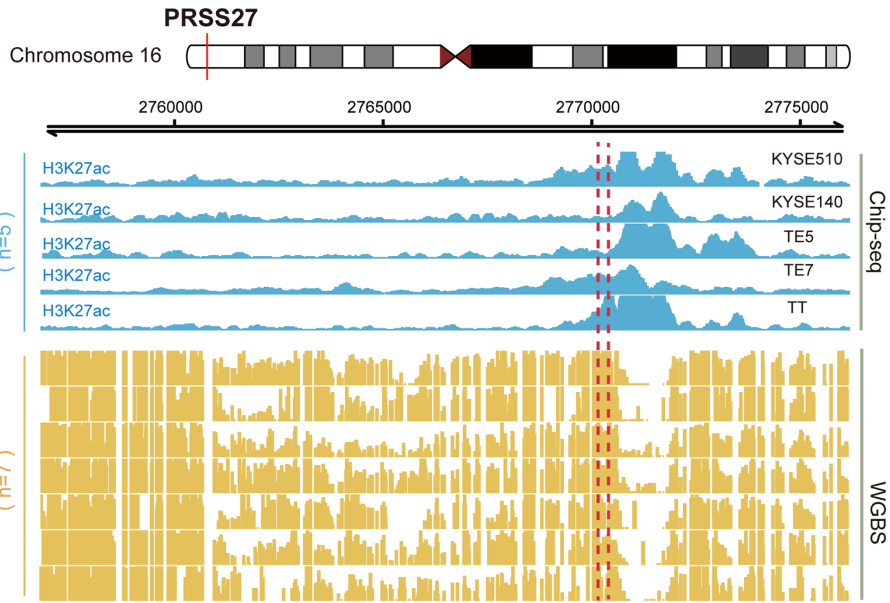


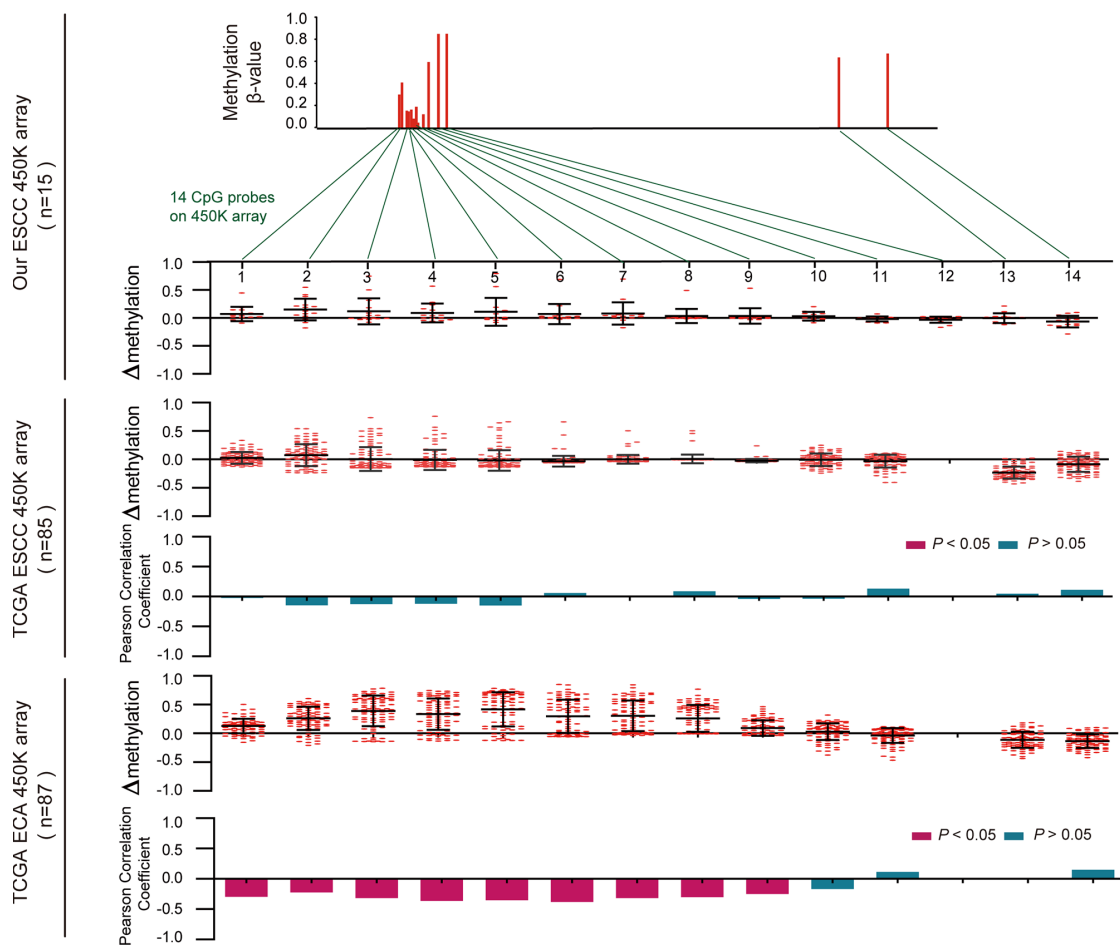
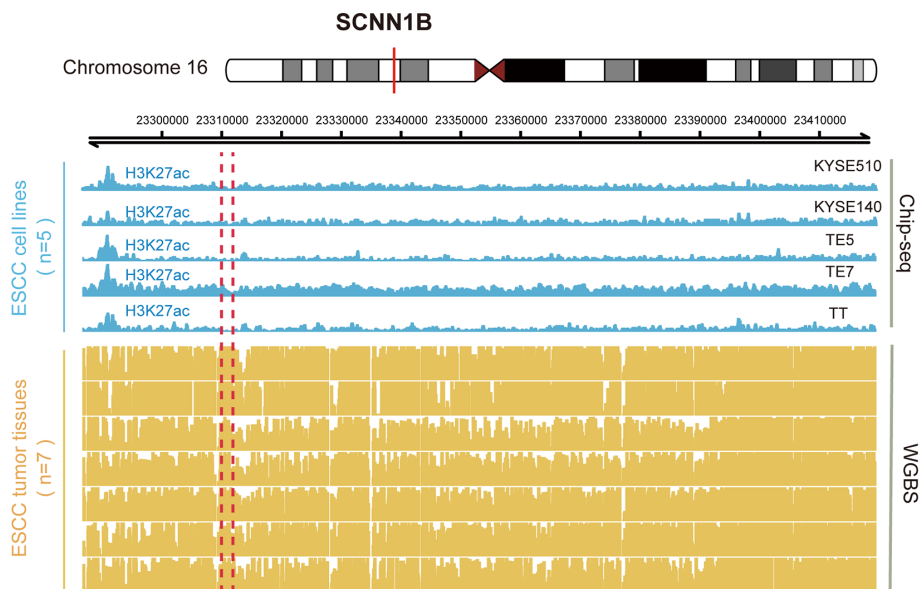


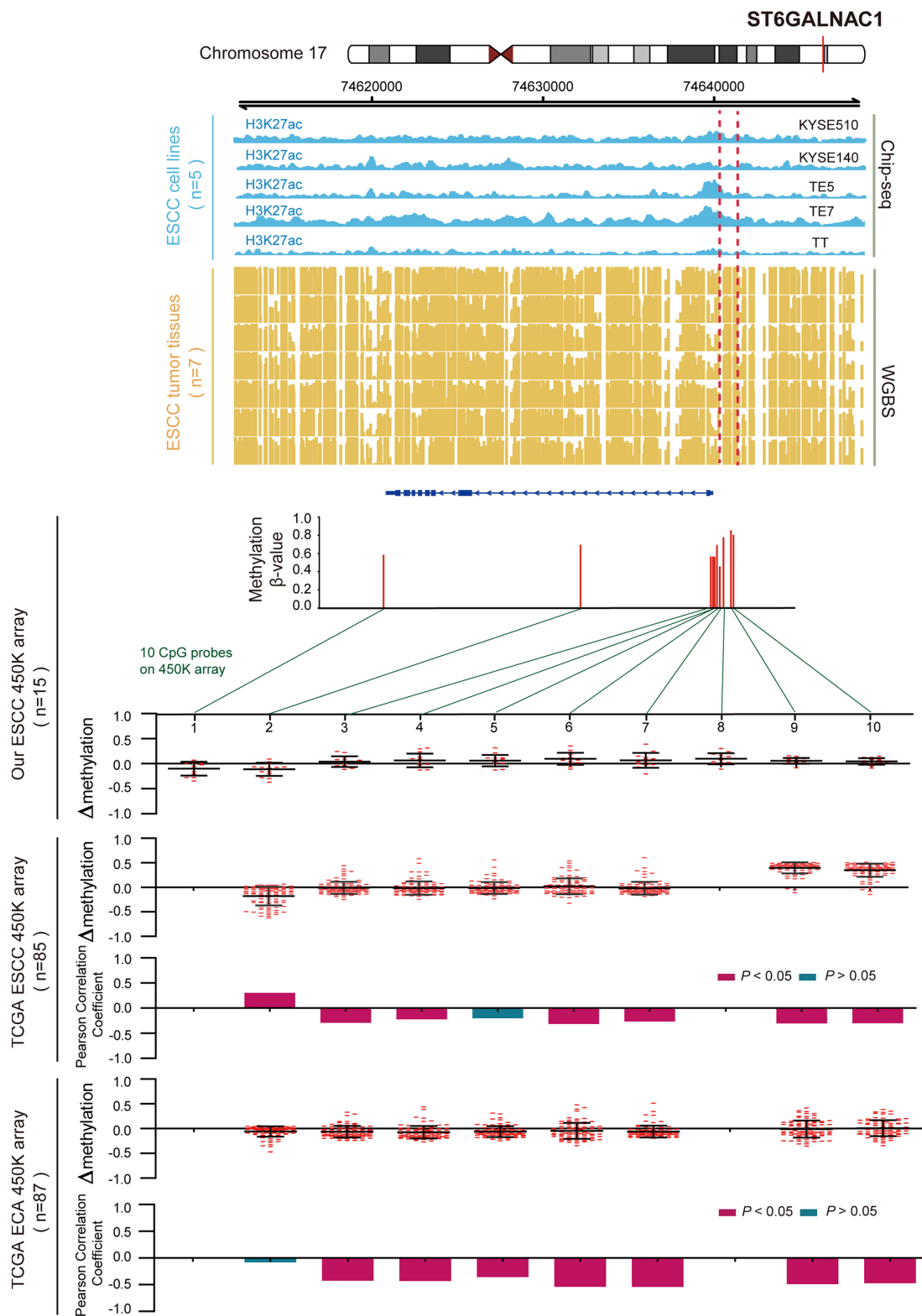












**Supplementary Figure 2. Inverse trend between DNA methylation and histone modification of key genes in ESCC cells.** Blue tracks represent the histone modifications of these genes in five ESCC cell lines, and yellow tracks represent its methylation level, as measured by the WGBS assay. All tracks are on the same scale (0-1). Scatter diagrams show the  $\Delta\beta$  of the genes in ESCC samples compared with normal samples. Histograms show the correlation between DNA methylation and gene expression.

## Supplementary Tables

Please browse Full Text version to see the data of Supplementary Table 1

**Supplementary Table 1. Basic information of 80557 DMPs.**

**Supplementary Table 2. Basic information of 16 genes.**

| Gene Symbol | Full Title  | Chromosomal Position             | Methylation      |          | Expression |                    | Pearson Correlation Coefficient (PCC) | P-value of PCC | SE*      |   |
|-------------|---|----------------------------------|------------------|----------|------------|--------------------|---------------------------------------|----------------|----------|---|
|             |   |                                  | Status           | P-value  | Level      | Log2 (Fold-change) |                                       |                |          |   |
| CRABP2      | Cellular retinoic acid binding protein 2                    | chr1:<br>156,667,400-156,677,608 | Hypermethylation | 2.40E-02 | Low        | -3.045             | 6.63E-07                              | -0.638         | 1.47E-04 | Y |
| IL1RN       | Interleukin 1 receptor antagonist                           | chr2:<br>113,873,470-113,893,593 | Hypermethylation | 3.34E-06 | Low        | -4.308             | 1.34E-11                              | -0.641         | 1.37E-04 | Y |
| EHD3        | EH domain containing 3                                      | chr2:<br>31,454,880-31,493,260   | Hypermethylation | 3.48E-02 | Low        | -2.971             | 1.77E-06                              | -0.749         | 1.96E-06 | Y |
| GPX3        | Glutathione peroxidase 3                                    | chr5:<br>150,397,999-150,408,554 | Hypermethylation | 2.78E-02 | Low        | -2.159             | 7.02E-04                              | -0.521         | 3.17E-03 | Y |
| ST6GALNAC1  | ST6 N-acetylgalactosaminidase alpha-2,6-sialyltransferase 1 | chr17:<br>74,620,839-74,639,900  | Hypermethylation | 7.03E-03 | Low        | -2.601             | 6.33E-03                              | -0.669         | 5.38E-05 | Y |
| SCNN1B      | Sodium channel epithelial 1 beta subunit                    | chr16:<br>23,311,591-23,394,620  | Hypermethylation | 1.00E-02 | Low        | -3.830             | 6.99E-09                              | -0.544         | 1.87E-03 | N |
| KLK11       | Kallikrein related peptidase 11                             | chr19:<br>51,523,487-51,533,290  | Hypermethylation | 2.90E-04 | Low        | -3.524             | 1.21E-07                              | -0.551         | 1.59E-03 | N |
| PPP1R3C     | Protein phosphatase 1 regulatory subunit 3C                 | chr10:<br>93,386,197-93,394,858  | Hypermethylation | 7.88E-03 | Low        | -4.137             | 4.45E-11                              | -0.684         | 3.04E-05 | N |
| KRT4        | Keratin 4   | chr12:<br>53,198,327-53,209,900  | Hypermethylation | 4.45E-03 | Low        | -7.039             | 1.18E-23                              | -0.535         | 2.30E-03 | N |
| SIX4        | SIX homeobox 4  | chr14:<br>61,174,256-61,192,852  | Hypomethylation  | 2.00E-02 | High       | 2.904              | 1.35E-06                              | -0.562         | 1.24E-03 | N |
| MFAP2       | Microfibril associated protein 2                            | chr1:<br>17,298,999-17,310,081   | Hypomethylation  | 3.61E-07 | High       | 3.578              | 7.51E-03                              | -0.756         | 1.17E-06 | N |
| COL5A2      | Collagen type V alpha 2 chain                               | chr2:<br>189,894,641-190,046,605 | Hypomethylation  | 9.12E-03 | High       | 2.941              | 2.28E-06                              | -0.557         | 1.39E-03 | N |
| KLK13       | Kallikrein related peptidase 13                             | chr19:<br>51,557,463-51,570,367  | Hypermethylation | 9.10E-03 | Low        | -4.901             | 5.85E-14                              | -0.596         | 5.13E-04 | N |
| VSIG10L     | V-Set and immunoglobulin domain containing 10 like          | chr19:<br>51,834,784-51,845,375  | Hypermethylation | 2.07E-02 | Low        | -3.749             | 7.09E-10                              | -0.516         | 3.54E-03 | N |
| C2orf54     | Chromosome 2 open reading frame 54                          | chr2:<br>241,825,469-241,835,569 | Hypermethylation | 8.63E-04 | Low        | -5.320             | 2.90E-15                              | -0.717         | 8.20E-06 | Y |
| PRSS27      | Protease, serine 27   | chr16:<br>2,762,419-2,770,556    | Hypermethylation | 2.69E-03 | Low        | -4.543             | 8.65E-12                              | -0.616         | 2.86E-04 | N |

\* SE: super-enhancer around the genes in esophageal tissues or cell lines<sup>[27]</sup>



**Supplementary Table 3. Clinicopathological characteristics of patients with ESCC.**

| Clinical and pathological indices | Cases | 5-year OS (%) | <i>P</i> -value <sup>a</sup> | 5-year DFS (%) | <i>P</i> -value <sup>a</sup> |
|-----------------------------------|-------|---------------|------------------------------|----------------|------------------------------|
| Specimens                         | 125   |               |                              |                |                              |
| Mean age                          | 59    |               |                              |                |                              |
| Age (years)                       |       |               |                              |                |                              |
| ≤59                               | 68    | 48.5          | 0.069                        | 44.1           | 0.156                        |
| >59                               | 57    | 29.8          |                              | 29.8           |                              |
| Gender                            |       |               |                              |                |                              |
| Male                              | 102   | 41.2          | 0.554                        | 37.3           | 0.970                        |
| Female                            | 23    | 31.8          |                              | 39.1           |                              |
| pTNM stage                        |       |               |                              |                |                              |
| I                                 | 13    | 61.5          | 0.000                        | 61.5           | 0.000                        |
| II                                | 62    | 53.2          |                              | 45.2           |                              |
| III                               | 50    | 18.0          |                              | 22.0           |                              |

<sup>a</sup>Log-rank test using the Kaplan Meier method; *P*-value <0.05 was considered significant.

OS, overall survival

DFS, disease-free survival

**Supplementary Table 4. Correlation between Signature-1 and clinicopathological characteristics in ESCC (N = 125).**

| Variables   | Signature-1 <sup>a</sup> |      | Chi-square value | <i>R</i> | <i>P</i> -value |
|-------------|--------------------------|------|------------------|----------|-----------------|
|             | Low                      | High |                  |          |                 |
| Age (years) |                          |      | 2.354            | 0.137    | 0.125*          |
| ≤ 59        | 38                       | 30   |                  |          |                 |
| > 59        | 24                       | 33   |                  |          |                 |
| Gender      |                          |      | 2.476            | -0.141   | 0.116*          |
| Male        | 54                       | 48   |                  |          |                 |
| Female      | 8                        | 15   |                  |          |                 |
| pTNM-stage  |                          |      | 0.970            | -0.041   | 0.616*          |
| I           | 7                        | 6    |                  |          |                 |
| II          | 28                       | 34   |                  |          |                 |
| III         | 27                       | 23   |                  |          |                 |

<sup>a</sup>Low, score<-2.221; high, score≥-2.221. \* Chi-squared test. *P*-value<0.05 was considered significant.

**Supplementary Table 5. qRT-PCR primers used in this study.**

| <b>Primers</b> | <b>Sequences (5'–3')</b> | <b>Positions</b> |
|----------------|--------------------------|------------------|
| EHD3-qF        | CCTTGGAAGAGCATTACCGC     | exon 1           |
| EHD3-qR        | ACCGCAATGAAGGAGTCTGT     | exon 2           |
| KLK11-qF       | GAAGACGCGGCTACTCTGTG     | exon 3           |
| KLK11-qR       | TTGTTGGGGAGGCTGTTGTT     | exon 4           |
| KRT4-qF        | CTTGGGCAATGACAAAGGGC     | exon 2           |
| KRT4-qR        | ACCTTG TTCAGGTAGGCAGC    | exon 4           |
| ST6GALNAC1-qF  | GATATACCGCCCCACCACTG     | exon 9           |
| ST6GALNAC1-qR  | GTGTAGCCGCTTCCAGACTT     | exon 10          |
| COL5A2-qF      | AAACTGGGCGGAAGCAAGA      | exon 1           |
| COL5A2-qR      | GCCATTCTGAGTGCAGGCTA     | exon 2           |
| PRSS27-qF      | ATTACATCCTCCCCGTGTGC     | exon 3           |
| PRSS27-qR      | TTGCACTTGGGTGTGTCGAT     | exon 4           |
| VSIG10L-qF     | CTCAGTCAAGATGGGCGGAA     | exon 6           |
| VSIG10L-qR     | AGCCTCCACGAGGATATGGA     | exon 7           |
| SIX4-qF        | TGTCAGTGGCAGCTTCACAA     | exon 2           |
| SIX4-qR        | GCTCCTTTCCAAGCCTTCCT     | exon 3           |
| SCNN1B-qF      | CACGAGCAGAGGTCATACCC     | exon 6           |
| SCNN1B-qR      | CGGGGACCTCAGAACCATTC     | exon 7           |
| MFAP2-qF       | GTGAGGAACAGTACCCGTGC     | exon 7           |
| MFAP2-qR       | TAATGACGTACACACGGCGG     | exon 8           |
| GPX3-qF        | TACGGAGCCCTCACCATTGA     | exon 2           |
| GPX3-qR        | AGGGAAAGCCCAGAATGACC     | exon 3           |
| C2orf54-qF     | CAGCCAGTGACCCCACTTAC     | exon 4           |
| C2orf54-qR     | ATGGCAGAGATCCGGTCCTT     | exon 5           |
| IL1RN-qF       | TGTGCCTGTCCTGTGTCAAG     | exon 4           |
| IL1RN-qR       | AAGCGCTTGTCTGCTTTCT      | exon 5           |
| KLK13-qF       | TGGCCCTAGTGATCGCCT       | exon 1           |
| KLK13-qR       | CTGGGAGAAACCCACTGGTC     | exon 2           |
| PPP1R3C-qF     | TTCGAATTTGTGCAGGCAGC     | exon 1           |
| PPP1R3C-qR     | TGAATGTGCCAAGCAAAGCC     | exon 2           |
| CRABP2-qF      | CCGATCGGAAAACCTTCGAGGA   | exon 2           |
| CRABP2-qR      | TTGATCTCCACTGCTGGCTT     | exon 3           |

F: forward primer; R: reverse primer.