**Identification of hyper-methylated tumor suppressor genes-based diagnostic panel for ESCC in a Chinese Han population**

Chenji Wang1†, Weilin Pu2†, Dunmei Zhao1, Sidi Chen3, Yinghui Zhou1, Zhenglei He1, Xulong Feng1, Ying Wang4, Caihua Li4, Shilin Li3, Li Jin2, Shicheng Guo5\*, Jiucun Wang2\*, Minghua Wang1\*

1Department of Biochemistry and Molecular Biology, Medical College, Soochow University, Suzhou, Jiangsu, China

2State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China

3Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China

4Genesky Biotechnologies Inc., Shanghai, China

5Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, MC0412, La Jolla, CA, US

**Running title:** A panel of DNA methylation biomarkers for ESCC diagnosis

Word Count: 3750

Number of Figures: 3

Number of Tables: 3

Quantity of Supplementary Data: 7

**Corresponding authors:**

Minghua Wang, Department of Biochemistry and Molecular Biology, Medical College, Soochow University, Suzhou, Jiangsu, China, Phone: +86-0512-65880108, Fax: +86-0512-65880103, Email: mhwang@suda.edu.cn

Jiucun Wang, School of Life Sciences, Fudan University, Shanghai 200438, China, Phone: +86-021-51630606, Fax: +86-21-51630607, E-mail: [jcwang@fudan.edu.cn](mailto:jcwang@fudan.edu.cn).

Shicheng Guo, Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, MC0412, La Jolla, CA 92093-0412. Telephone: 281-685-5882, Fax: 858-534-5722, Email: [scguo@ucsd.edu](mailto:scguo@ucsd.edu)

**Abstract**

DNA methylation-based biomarkers were suggested to be promising for early cancer diagnosis. However, DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC), especially in Chinese Han populations have not been identified and evaluated quantitatively. Here, we selected candidate tumor suppressor genes (N=65) through literature searching and four public high-throughput DNA methylation microarray datasets including 136 samples totally were collected for initial confirmation. Furthermore, targeted bisulfite sequencing was applied in an independent cohort of 94 pairs of ESCC and normal tissues from a Chinese Han population for eventual validation. It is found that *ADHFE1, EOMES, SALL1* and *TFPI2* were identified and validated in the ESCC samples from a Chinese Han population. All four candidate regions were validated to be significantly hyper-methylated in ESCC samples (*ADHFE1*, p = 1.70×10-3; *EOMES*, p = 2.90×10-9; *SALL1*, p = 3.90×10-7; *TFPI2*, p = 3.40×10-6). Logistic regression based prediction model shown a robust ESCC classification performance (Sensitivity = 66%, Specificity = 87%, AUC = 0.81). Moreover, advanced classification method had better performances (random forest and naive Bayes). Interestingly, the diagnostic performance could be improved in non-alcohol use subgroup (AUC = 0.84). In summary, methylation panels of *ADHFE1*, *EOMES*, *SALL1* and *TFPI2* could be an effective methylation-based assay for ESCC diagnosis.

Keywords: Esophageal squamous cell carcinoma, DNA methylation, Biomarker, Diagnosis, Targeted bisulfite sequencing

**Introduction**

Esophageal cancer is one of the most aggressive malignant tumors with high prevalence and poor prognosis worldwide [1]. Esophageal cancer usually occurs as two subtypes, esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC), which differed significantly in pathogenesis, pathology, epidemiology and geographical distribution [2]. The regions of the highest occurrence of esophageal cancer stretching from northern China to northwestern Iran, including Japan and India, are localized in the so-called Asian Esophageal Cancer Belt [3, 4]. The prevalence of ESCC and EAC in these regions are significantly unbalanced with 90% of esophageal cancer patients are ESCCs [5]. In addition, the clinical outcomes of ESCC patients depend largely on its diagnosed stage [2]. The majority of ESCCs are diagnosed at advanced stages and the overall 5-year survival rate is relatively poor, while the 5-year survival rate for early-stage diagnosed ESCC patients is significantly higher [6]. Therefore, it is imperative to identify biomarkers for early diagnosis of ESCC patients.

DNA methylation, which usually occurs in CpG dinucleotides, functioning as an epigenetic modification in mammalian genome and is involved in regulating gene and microRNA expression and alternative splicing [7-9]. Global hypo-methylation as well as the hyper-methylation of CpG islands in the tumor suppressor genes have been widely identified in the process of tumorigenesis [10, 11]. Comparing with the biomarkers based on SNP/mutation, copy number variations (CNV) and gene/microRNA expression, DNA methylation is advantageous due to its flexible stability and diagnosis accuracy [12-14]. Numerous studies have suggested that DNA methylation could be one of the most promising early-detection biomarkers for several types of cancers [15-19]. For example, dozens of genes have been reported to be hyper-methylated in ESCC, including *APC, MGMT, CDH1, RASSF1* [20-23]. In addition, due to the heterogeneity of ESCC, a single biomarker could only achieve relatively limited prediction ability, which calling for the comprehensive combinations of these candidate biomarkers.

In the present study, we first collected 65 candidate tumor suppressor genes and evaluated their methylation status in ESCC and adjacent control tissues from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) datasets. After a stringent biomarker selection procedure, four of the candidate hyper-methylated genes (*ADHFE1, EOMES, SALL1, TFPI2*) were validated with high-throughput datasets from public databases. Moreover, the methylation profiles of these four genes were further validated with targeted bisulfite sequencing method in 94 pairs of ESCC tumor and adjacent control tissues from a Chinese Han population, yielding a robust performance for ESCC diagnosis (Figure 1).

**Patients and Methods**

**Biomarker selection based on publications and public datasets**

Firstly, we collected 65 candidate reported tumor suppressor genes through literature search and the candidate genes were listed in Table S1. In order to test the methylation status of these 65 candidate genes in ESCC patients, we searched high-throughput microarray datasets in TCGA and GEO database to collect the DNA methylation profiles of the ESCC samples. After stringent quality control, we found that TCGA project has quantified the methylation profiles of 84 ESCC and 3 normal tissues, as well as 78 EAC and 13 normal tissues. Due to the similarities of the adjacent control tissues from ESCC and EAC, the 13 normal tissues of EAC were included in our combined dataset as controls equally (Figure. S1). In addition, three datasets in GEO database named GSE52826, GSE74693 and GSE79366 were also retrieved, including 26 ESCC and 10 normal tissues. Eventually, 110 ESCC and 26 normal tissues were included from TCGA/GEO for further study. ComBat was applied for removing the batch effect between the different datasets [24]. Due to the fact that we want to obtain the diagnostic biomarkers which might be applied for liquid biopsy, we then defined the CpG sites with high methylaiton percent (>0.25) in the ESCCs and relatively lower methylation percent (< 0.25) in the adjacent control tissues as the significant CpG sites. Further, it is widely acknowledged that the methylation status of CpG sites was largely variable in different cell types. As a result, we then filtered out the significant CpG sites with high methylation percentage (> 0.15) in either peripheral blood mononuclear cells (PBMC, N = 111) or peripheral blood leucocytes (PBL, N = 527) of the healthy normal samples from the GEO database. The PBMC dataset came from the GSE53045 dataset, and the PBL dataset was the combination of GSE36054 and GSE42861 dataset [25-27]. Moreover, we selected the candidate genes with at least two eligible significant CpG sites for further validation. In summary, six genes were included (*ADHFE1, EOMES, RUNX1, SALL1, TFPI2, WT1,* Table S2). However, due to the technical limitations of the multiplex PCR design, we finally selected four candidate regions located at four genes, for the final validation stage with 94 pairs of Chinese ESCC and control samples (*ADHFE1, EOMES, SALL1, TFPI2*, Table 1).

**Patients, samples and DNA**

ESCC samples and their paired adjacent control tissues were obtained for validation study from the First Affiliated Hospital of Soochow University and Fourth Military Medical University between the years of 2011 and 2015. All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments. The studies were approved by the institutional review boards of Soochow University at Jiangsu Province and Fudan University, Shanghai, China. Written informed consent was obtained from each study subject. In addition, all of the subjects were re-examined and confirmed by professional pathologists for histopathological diagnosis. All tissues were immediately frozen at -80oC after surgical resection. Face-to-face interviews were conducted by professional investigators with a comprehensive questionnaire, including clinical information on tobacco smoking, alcohol consumption and family history.

**DNA extraction, bisulfite conversion and targeted bisulfite sequencing**

Genomic DNA from ESCC tumor tissue and adjacent control tissue samples were extracted by AIIperp DNA/RNA Mini Kit (Qiagen, Duesseldorf, Germany) according to the manufacturer’s protocols. For methylation analysis, 500 ng genomic DNA was subjected to bisulfite conversion using the EpiTect Fast DNA Bisulfite Kit (Qiagen, Duesseldorf, Germany). A multiplex PCR was performed first with optimized primer sets combination (Table S3). PCR amplicons were diluted and amplified using indexed primers and the products (170bp – 270 bp) were separated by agarose electrophoresis and purified by QIAquick Gel Extraction kit (Qiagen, Duesseldorf, Germany). Libraries from different samples were quantified and pooled together equally, sequenced with the Illumina Hiseq 2000 platform according to the manufacturer's protocols. BSseeker2 software was utilized for reads mapping and methylation calling [28]. Samples and CpG sites with high missing rates (> 30%) were removed. In order to make sure the reliability of the technique and analysis pipeline, we take LINE-1 as the technical control, whose methylation rate was decreased in cancer tissues compared with normal tissues. Therefore, LINE-1 methylation status was applied to check the credibility of the experiments. Meanwhile, the conversion ratio of C to T in non-CpG sites were applied to evaluate the bisulfite conversion efficiency.

**Statistical analysis and machine learning**

In the first and second stage, we tested the differential methylation of the CpG sites between cancer and normal tissues using Wilcoxon rank-sum test. False discovery rate (FDR) correction was conducted for multiple test correction. In order to discriminate the ESCC tumor and normal tissues, we utilized several machine learning methods, including logistic regression (Package stats), support vector machine (SVM, Package e1071), random forest (Package randomForest), naïve Bayes (Package e1071), neural network (Package nnet), linear discriminant analysis (LDA, Package mda), mixture discriminant analysis (MDA, Package mda), as well as the flexible discriminant analysis (FDA, Package mda) followed with five-fold cross-validation. All statistical analyses were conducted using R 3.2.1 [29].

**Results**

**Public datasets collection and CpG sites validation**

In order to quantify the methylation status of these four candidate genes, public DNA methylation microarray datasets of ESCC were carefully searched. In total, 110 ESCC tumor tissues and 26 adjacent control tissues were enrolled [30-32]. Based on the CpG sites selection criteria which was described in Patients and Methods, six significant CpG sites (cg20295442, cg20912169, cg22383888, cg04550052, cg04698114, cg12973591) located at the four candidate genes were selected for validation (Table 1).Integratively, though some of the six CpG sites did not reach the statistical significance threshold due to the limited sample size, we still believed that all of these 6 CpG sites may be of potential as the non-invasive potential biomarkers for ESCC and thus were included for validation. To test the prediction ability based on these six CpG sites, we built a prediction model based on the logistic regression using the methylation status of these 6 CpG sites without adjustment for age, gender and other covariates, which provided a fair good performance to discriminate between ESCC and normal tissues (Sensitivity = 79%, Specificity = 92%, AUC = 0.87). To further evaluate and validate the diagnostic ability of these six CpG sites, we then conducted the validation study in 94 paired ESCC and adjacent control tissue samples obtained from the patients from the Chinese Han population.

**Methylation status validation with targeted bisulfite sequencing**

The characteristics of the ESCC patients are shown in Table S4. In order to give a robust characterization of the methylation status of these 6 CpG sites as well as the four genes, we applied the targeted bisulfite sequencing method, which was based on the next generation sequencing (NGS) platforms. Because the NGS platforms could generate millions of reads with length > 200 bp, we then designed to test four genomic regions for the four candidate tumor suppressor genes for validation (Table 2). In the quality control process, we found that the bisulfite conversion rate (C to T ratio in non-CpG loci) of our samples were higher than 98%, and no significant difference was found between the tumor and adjacent control tissues (Figure 2A). Besides, we used the LINE-1 methylation status as technical control and showed that our study was robust and reliable (Figure 2B). In addition, the samples and the CpG sites with high missing rates were also filtered out as described in Patients and Methods. After quality control, 163 samples remained for further study. Differential methylation analyses were conducted for the four genomic regions, suggesting a major difference between the ESCC and adjacent control tissues (Figure. S2). A logistic regression model was then applied, and showed significant hyper-methylation status of the six selected CpG sites in the ESCC tissues (cg20295442, p= 5.10×10-3; cg20912169, p= 2.10×10-3; cg22383888, p= 3.30×10-9; cg04550052, p= 2.50×10-4; cg04698114, p= 1.10×10-6; cg12973591, p = 3.30×10-5). To better characterize the methylation status of the four genomic regions as well as the four candidate genes, we averaged the methylation status of all the CpG sites in each genomic region as representative for further analysis (Figure 3). Based on the mean methylation status of the four genomic regions, the prediction ability of each region separately was evaluated through logistic regression without adjustment for age, gender and other covariates. The sensitivity of each region ranges from 29% to 69%, while the specificity ranges from 77% to 94%, and the AUC ranges from 0.64 to 0.78 (Table 2). Moreover, in the logistic model taking all of the four regions as predictors, we obtained the sensitivity of 66% and specificity of 87%, as well as the AUC of 0.81 (Figure. S3).

**The prediction performance of the diagnosis panel in different classification models**

Several machine learning methods, including logistic regression model, random forest, support vector machine (SVM), neural network (NN), Naïve Bayes (NB), linear discriminant analysis (LDA), mixture discriminant analysis (MDA), flexible discriminant analysis (FDA) and gradient boosting machine (GBM) following with five-fold cross validation were utilized for ESCC classification based on the targeted bisulfite sequencing regions (Table 3). It turned out that the GBM model achieved the highest classification accuracy among all machine learning methods in train stage, whose sensitivity, specificity and accuracy were 82.6%, 85.6% and 84.0%. The naive bayes model achieved the best specificity (91.6%) in the train stage. In the test stage, the random forest and naive bayes performed with the best sensitivity (72.8%) and specificity (91.0%), respectively. In addition, the linear discriminant analysis and flexible discriminant analysis model both achieved the best accuracy (73.5%).

**The diagnostic ability in the ESCC subgroups**

Previous studies have found several risk factors for the incidence of ESCC, including age, gender, smoking status, and alcohol status [33-36]. In order to explore the effects of these risk factors on the ESCC diagnosis, we conducted the subgroup analyses. Similarly, the mean methylation percentage of each genomic region was utilized. To explore the diagnostic ability in the young/old samples, we first divided the samples according to the median age of our patients. No significant difference between the sensitivity, specificity and the AUC between the two subgroups (Table S5). The AUCs in the two subgroups was 0.82 and 0.80 for the young and old subgroups, respectively (Figure. S4). When it comes to the gender, the difference was still quite limited (AUC: 0.79 vs. 0.82 for male and female subgroups, Table S6). Similarly, no significant difference of the diagnostic performances was found between smoker/non-smoker subgroup analysis (Table S7). However, when concentrating on the effect of alcohol use, we found that the non-alcohol use subgroup showed obviously higher AUC than that of the alcohol use subgroup (0.84 vs. 0.77 respectively, Table S8). The significant difference in the diagnostic performance between the alcohol use and non-alcohol use subgroup indicates that alcohol use may contribute to the epigenetic changes in ESCC as well as to the pathogenesis of ESCC [34].

**Discussion**

Although epigenetic aberrations were reported having critical role in cancer progression and prognosis in many studies, limited DNA methylation based biomarkers have been utilized for the diagnosis and prognosis of ESCC. Moreover, numerous studies have found that a number of tumor suppressor genes have been inhibited through hyper-methylation and contribute significantly to the carcinogenesis [37-40]. In this study, we retrieved 65 reported tumor suppressor genes in several kinds of tumor types from literature screening and tested their methylation status from the TCGA and GEO datasets. Specifically, six candidate CpG sites, located at the four candidate genes (*ADHFE1, EOMES, SALL1, TFPI2*) were found to be hyper-methylated in ESCC tissues while hypo-methylated in the adjacent control tissues as well as the peripheral blood samples. In order to further confirm their methylation status and quantify their potential to be the ESCC diagnostic biomarkers, we validated these six CpG sites in an independent 94 pairs of ESCC and adjacent control tissues from Chinese Han population. Targeted bisulfite sequencing method was conducted to detect the methylation status of the candidate genes. Because that the targeted bisulfite sequencing is based on the next generation sequencing (NGS) technology, we could not only detect the methylation status of these six CpG sites but also their genomic regions as well. To give a robust estimation of the methylation status of the candidate regions, we averaged the methylation percent of all the CpG sites on each region, and applied several machine learning methods to assess the diagnostic ability of these DNA methylation-based biomarkers. In addition, the subgroup analyses were also conducted and the alcohol use was found to be associated with the diagnostic ability, indicating the importance of taking the epidemiological information into account when performing the ESCC diagnosis.

All four candidate genes have been reported to be associated with the tumorigenesis in several tumor types. Alcohol dehydrogenase, iron containing 1 (*ADHFE1*) encodes hydroxyacid-oxoacid transhydrogenase, which is responsible for the oxidation of 4-hydroxybutyrate in mammalian tissues [41]. *ADHFE1* promoter hyper-methylation was found in [colorectal cancer](https://www.ncbi.nlm.nih.gov/pubmed/24886599/) (CRC) and the alcohol could down-regulate the expression of *ADHFE1* through hyper-methylation and further induce the proliferation of CRC cells [42, 43]. *EOMES* belongs to the TBR1 (T-box brain protein 1) sub-family of T-box genes, encoding a transcription factor which is necessary for the embryonic development. It has been reported that *EOMES* promoter methylation could serve as a promising biomarker for the prediction of occurrence, recurrence and prognosis of bladder cancer [44-46]. In addition, EOMES has also been confirmed to have potential anti-cancer functions through siRNA experiments, and was regarded as candidate tumor suppressor gene for human hepatocellular carcinoma [47]. Spalt like transcription factor 1(*SALL1*) encodes a zinc finger transcriptional repressor, which has recently been identified as a tumor suppressor gene, whose expression was in positive correlation with *CDH1* and associated with the survival of patients in breast cancer [48]. In addition, *SALL1* hyper-methylation has already been confirmed as the diagnostic biomarker for breast cancer and other epithelial cancers, especially for the colorectal cancer [49]. Tissue factor pathway inhibitor 2 (*TFPI2*) encodes a member of the Kunitz-type serine proteinase inhibitor family, and was found to be a tumor suppressor gene in several types of cancer [50-53]. Moreover, numerous studies have suggested that the *TFPI2* promoter hyper-methylation could be of potential as a biomarker for cancer progression and prognosis in hepatocellular carcinoma, pancreatic adenocarcinoma, epithelial ovarian cancer and melanomas [54-57]. In order to confirm the association between gene expression and methylation of these four candidates, we also conducted the study to demethylase the human esophageal squamous carcinoma cell line (CaES-17) with 5-aza-2’-deoxycytidine and quantified the gene expression of these candidate genes. Concordantly, three of these four genes (*EOMES, SALL1, TFPI2*) showed significant down-regulation after 5-aza-2’-deoxycytidine treatment, while *ADHFE1* showed down-regulation but did not reach the significance threshold (Figure. S5). In summary, our results validated the inverse correlations between gene expression and methylation of these four genes, and suggesting that these four genes might be vital to the ESCC carcinogenesis.

The accurate early diagnosis of cancer is a great challenge due to the cancer heterogeneity. In our study, we selected four candidate tumorigenesis genes and applied the targeted bisulfite sequencing method to explore the methylation status of our candidate CpG sites as well as their adjacent genomic regions, thus yielding a robust estimation of the methylation status of the candidate genes. With the fast development of NGS technology, the targeted bisulfite sequencing method is becoming more and more popular for methylation detection because of high accuracy, high-throughput and cost-effective. In the past studies, we found the single DNA methylation biomarker usually cannot provide enough prediction power in cancer diagnosis. According to our results, the panel consisting of these four candidate genes could distinguish the ESCC tumors with higher specificity and sensitivity compared with single biomarker.

In summary, a panel with four genes was identified and achieved a fair good accuracy in classifying ESCC from normal tissues. However, according to diagnosis performance, our prediction model still has more space to be improved when we introduce more biomarkers. Multi-omics datasets, including genomics, epigenomics and proteomics, which could provide biomarkers in different biological layers, could contribute to the accurate non-invasive diagnosis of esophageal squamous cell carcinoma in the future.

Integrated analysis of public literatures and multiples high-throughput DNA methylation microarray datasets were conducted and discovered four tumor suppressor genes (*ADHFE1, EOMES, SALL1, TFPI2*) as the candidate biomarkers for ESCC diagnosis. All four tumor suppressor genes were then successfully validated in an independent cohort including 94 pairs of ESCC and adjacent control tissues. Moreover, the EOMES showed the highest sensitivity (0.69) and AUC (0.78), while the ADHFE1 showed the best specificity (0.94). Methylation profiles of *ADHFE1, EOMES, SALL1, TFPI2* could be an effective methylation-based assay (Sensitivity = 0.66, Specificity = 0.87, AUC = 0.81) for the ESCC diagnosis with high specificity.

**Acknowledgements**

Weilin Pu† and Chenji Wang† contributed equally on this work

**Funding**

The study was supported by research grants from the National Natural Science Foundation of China(81572923, 31521003,81071957), the Jiangsu Province Postdoctoral Research Funding (7131708615), the Key Project in the National Science & Technology Pillar Program (2011BAI09B00), a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Suzhou City Science and Technology Program (SYS 201419), 111 Project (B13016) and China Medical Foundation (CMF). Computational support was provided by High-End Computing Center located at Fudan University.

**Availability of data and materials**

The datasets used and analyzed in this study are available form the corresponding author on request.

**Consent for publication**

Not applicable

**Disclosure Statement**

The authors declare that they have no competing interests.

**Reference**

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA: a cancer journal for clinicians. 2016; 66: 7-30. doi:

2. Enzinger PC, Mayer RJ. Esophageal cancer. N Engl J Med. 2003; 349: 2241-52. doi: 10.1056/NEJMra035010.

3. Khuroo MS, Zargar SA, Mahajan R, Banday MA. High incidence of oesophageal and gastric cancer in Kashmir in a population with special personal and dietary habits. Gut. 1992; 33: 11-5. doi:

4. Kmet J, Mahboubi E. Esophageal cancer in the Caspian littoral of Iran: initial studies. Science. 1972; 175: 846-53. doi:

5. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011; 61: 69-90. doi: 10.3322/caac.20107.

6. Besharat S, Jabbari A, Semnani S, Keshtkar A, Marjani J. Inoperable esophageal cancer and outcome of palliative care. World J Gastroenterol. 2008; 14: 3725-8. doi:

7. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007; 128: 669-81. doi: 10.1016/j.cell.2007.01.033.

8. He Y, Cui Y, Wang W, Gu J, Guo S, Ma K, Luo X. Hypomethylation of the hsa-miR-191 locus causes high expression of hsa-mir-191 and promotes the epithelial-to-mesenchymal transition in hepatocellular carcinoma. Neoplasia. 2011; 13: 841-53. doi:

9. Flores K, Wolschin F, Corneveaux JJ, Allen AN, Huentelman MJ, Amdam GV. Genome-wide association between DNA methylation and alternative splicing in an invertebrate. BMC Genomics. 2012; 13: 480. doi: 10.1186/1471-2164-13-480.

10. Baylin SB, Esteller M, Rountree MR, Bachman KE, Schuebel K, Herman JG. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. Hum Mol Genet. 2001; 10: 687-92. doi:

11. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med. 2003; 349: 2042-54. doi: 10.1056/NEJMra023075.

12. Guo S, Wang YL, Li Y, Jin L, Xiong M, Ji QH, Wang J. Significant SNPs have limited prediction ability for thyroid cancer. Cancer Med. 2014; 3: 731-5. doi: 10.1002/cam4.211.

13. Jiang F, Todd NW, Li R, Zhang H, Fang H, Stass SA. A panel of sputum-based genomic marker for early detection of lung cancer. Cancer Prev Res (Phila). 2010; 3: 1571-8. doi: 10.1158/1940-6207.CAPR-10-0128.

14. Zhu J, Yao X. Use of DNA methylation for cancer detection: promises and challenges. Int J Biochem Cell Biol. 2009; 41: 147-54. doi: 10.1016/j.biocel.2008.09.003.

15. Zhao Y, Sun J, Zhang H, Guo S, Gu J, Wang W, Tang N, Zhou X, Yu J. High-frequency aberrantly methylated targets in pancreatic adenocarcinoma identified via global DNA methylation analysis using methylCap-seq. Clin Epigenetics. 2014; 6: 18. doi: 10.1186/1868-7083-6-18.

16. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer. 2003; 3: 253-66. doi: 10.1038/nrc1045.

17. Pu W, Geng X, Chen S, Tan L, Tan Y, Wang A, Lu Z, Guo S, Chen X, Wang J. Aberrant methylation of CDH13 can be a diagnostic biomarker for lung adenocarcinoma. J Cancer. 2016; 7: 2280-9. doi: 10.7150/jca.15758.

18. Geng X, Pu W, Tan Y, Lu Z, Wang A, Tan L, Chen S, Guo S, Wang J, Chen X. Quantitative assessment of the diagnostic role of FHIT promoter methylation in non-small cell lung cancer. Oncotarget. 2017; 8: 6845-56. doi: 10.18632/oncotarget.14256.

19. Guo S, Tan L, Pu W, Wu J, Xu K, Wu J, Li Q, Ma Y, Xu J, Jin L, Wang J. Quantitative assessment of the diagnostic role of APC promoter methylation in non-small cell lung cancer. Clin Epigenetics. 2014; 6: 5. doi: 10.1186/1868-7083-6-5.

20. Kawakami K, Brabender J, Lord RV, Groshen S, Greenwald BD, Krasna MJ, Yin J, Fleisher AS, Abraham JM, Beer DG, Sidransky D, Huss HT, Demeester TR, et al. Hypermethylated APC DNA in plasma and prognosis of patients with esophageal adenocarcinoma. J Natl Cancer Inst. 2000; 92: 1805-11. doi:

21. Chen J, Huang ZJ, Duan YQ, Xiao XR, Jiang JQ, Zhang R. Aberrant DNA methylation of P16, MGMT, and hMLH1 genes in combination with MTHFR C677T genetic polymorphism and folate intake in esophageal squamous cell carcinoma. Asian Pac J Cancer Prev. 2012; 13: 5303-6. doi:

22. Takeno S, Noguchi T, Fumoto S, Kimura Y, Shibata T, Kawahara K. E-cadherin expression in patients with esophageal squamous cell carcinoma: promoter hypermethylation, Snail overexpression, and clinicopathologic implications. Am J Clin Pathol. 2004; 122: 78-84. doi: 10.1309/P2CD-FGU1-U7CL-V5YR.

23. Kuroki T, Trapasso F, Yendamuri S, Matsuyama A, Alder H, Mori M, Croce CM. Allele loss and promoter hypermethylation of VHL, RAR-beta, RASSF1A, and FHIT tumor suppressor genes on chromosome 3p in esophageal squamous cell carcinoma. Cancer Res. 2003; 63: 3724-8. doi:

24. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012; 28: 882-3. doi:

25. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, Monick M, Brody GH, Tan K, Beach SR, Philibert RA. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014; 15: 151. doi: 10.1186/1471-2164-15-151.

26. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST. Age-associated DNA methylation in pediatric populations. Genome Res. 2012; 22: 623-32. doi: 10.1101/gr.125187.111.

27. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013; 31: 142-7. doi: 10.1038/nbt.2487.

28. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen P-Y, Pellegrini M. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC genomics. 2013; 14: 774. doi:

29. Dessau RB, Pipper CB. [''R"--project for statistical computing]. Ugeskr Laeger. 2008; 170: 328-30. doi:

30. Li X, Zhou F, Jiang C, Wang Y, Lu Y, Yang F, Wang N, Yang H, Zheng Y, Zhang J. Identification of a DNA methylome profile of esophageal squamous cell carcinoma and potential plasma epigenetic biomarkers for early diagnosis. PLoS One. 2014; 9: e103162. doi: 10.1371/journal.pone.0103162.

31. Kishino T, Niwa T, Yamashita S, Takahashi T, Nakazato H, Nakajima T, Igaki H, Tachimori Y, Suzuki Y, Ushijima T. Integrated analysis of DNA methylation and mutations in esophageal squamous cell carcinoma. Mol Carcinog. 2016; 55: 2077-88. doi: 10.1002/mc.22452.

32. Hao JJ, Lin DC, Dinh HQ, Mayakonda A, Jiang YY, Chang C, Jiang Y, Lu CC, Shi ZZ, Xu X, Zhang Y, Cai Y, Wang JW, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. Nat Genet. 2016; 48: 1500-7. doi: 10.1038/ng.3683.

33. Morita M, Kumashiro R, Kubo N, Nakashima Y, Yoshida R, Yoshinaga K, Saeki H, Emi Y, Kakeji Y, Sakaguchi Y, Toh Y, Maehara Y. Alcohol drinking, cigarette smoking, and the development of squamous cell carcinoma of the esophagus: epidemiology, clinical findings, and prevention. Int J Clin Oncol. 2010; 15: 126-34. doi: 10.1007/s10147-010-0056-7.

34. Pandeya N, Williams G, Green AC, Webb PM, Whiteman DC, Australian Cancer S. Alcohol consumption and the risks of adenocarcinoma and squamous cell carcinoma of the esophagus. Gastroenterology. 2009; 136: 1215-24, e1-2. doi: 10.1053/j.gastro.2008.12.052.

35. Wang JM, Xu B, Rao JY, Shen HB, Xue HC, Jiang QW. Diet habits, alcohol drinking, tobacco smoking, green tea drinking, and the risk of esophageal squamous cell carcinoma in the Chinese population. Eur J Gastroenterol Hepatol. 2007; 19: 171-6. doi: 10.1097/MEG.0b013e32800ff77a.

36. Nasrollahzadeh D, Kamangar F, Aghcheli K, Sotoudeh M, Islami F, Abnet CC, Shakeri R, Pourshams A, Marjani HA, Nouraie M, Khatibian M, Semnani S, Ye W, et al. Opium, tobacco, and alcohol use in relation to oesophageal squamous cell carcinoma in a high-risk area of Iran. Br J Cancer. 2008; 98: 1857-63. doi: 10.1038/sj.bjc.6604369.

37. Baylin SB. DNA methylation and gene silencing in cancer. Nat Clin Pract Oncol. 2005; 2 Suppl 1: S4-11. doi: 10.1038/ncponc0354.

38. Baylin SB, Herman JG. DNA hypermethylation in tumorigenesis: epigenetics joins genetics. Trends Genet. 2000; 16: 168-74. doi:

39. Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, Berman DM, Jenuwein T, Pruitt K, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet. 2007; 39: 237-42. doi: 10.1038/ng1972.

40. Esteller M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. Oncogene. 2002; 21: 5427-40. doi: 10.1038/sj.onc.1205600.

41. Kardon T, Noel G, Vertommen D, Schaftingen EV. Identification of the gene encoding hydroxyacid-oxoacid transhydrogenase, an enzyme that metabolizes 4-hydroxybutyrate. FEBS Lett. 2006; 580: 2347-50. doi: 10.1016/j.febslet.2006.02.082.

42. Moon JW, Lee SK, Lee YW, Lee JO, Kim N, Lee HJ, Seo JS, Kim J, Kim HS, Park SH. Alcohol induces cell proliferation via hypermethylation of ADHFE1 in colorectal cancer cells. BMC Cancer. 2014; 14: 377. doi: 10.1186/1471-2407-14-377.

43. Tae CH, Ryu KJ, Kim SH, Kim HC, Chun HK, Min BH, Chang DK, Rhee PL, Kim JJ, Rhee JC, Kim YH. Alcohol dehydrogenase, iron containing, 1 promoter hypermethylation associated with colorectal cancer differentiation. BMC Cancer. 2013; 13: 142. doi: 10.1186/1471-2407-13-142.

44. Reinert T, Borre M, Christiansen A, Hermann GG, Orntoft TF, Dyrskjot L. Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation. PLoS One. 2012; 7: e46297. doi: 10.1371/journal.pone.0046297.

45. Reinert T, Modin C, Castano FM, Lamy P, Wojdacz TK, Hansen LL, Wiuf C, Borre M, Dyrskjot L, Orntoft TF. Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. Clin Cancer Res. 2011; 17: 5582-92. doi: 10.1158/1078-0432.CCR-10-2659.

46. Kim YJ, Yoon HY, Kim JS, Kang HW, Min BD, Kim SK, Ha YS, Kim IY, Ryu KH, Lee SC. HOXA9, ISL1 and ALDH1A3 methylation patterns as prognostic markers for nonmuscle invasive bladder cancer: Array‐based DNA methylation and expression profiling. International journal of cancer. 2013; 133: 1135-42. doi:

47. Gao F, Xia Y, Wang J, Lin Z, Ou Y, Liu X, Liu W, Zhou B, Luo H, Zhou B, Wen B, Zhang X, Huang J. Integrated analyses of DNA methylation and hydroxymethylation reveal tumor suppressive roles of ECM1, ATF5, and EOMES in human hepatocellular carcinoma. Genome Biol. 2014; 15: 533. doi: 10.1186/s13059-014-0533-9.

48. Wolf J, Muller-Decker K, Flechtenmacher C, Zhang F, Shahmoradgoli M, Mills GB, Hoheisel JD, Boettcher M. An in vivo RNAi screen identifies SALL1 as a tumor suppressor in human breast cancer with a role in CDH1 regulation. Oncogene. 2014; 33: 4273-8. doi: 10.1038/onc.2013.515.

49. Hill VK, Hesson LB, Dansranjavin T, Dallol A, Bieche I, Vacher S, Tommasi S, Dobbins T, Gentle D, Euhus D, Lewis C, Dammann R, Ward RL, et al. Identification of 5 novel genes methylated in breast and other epithelial cancers. Molecular Cancer. 2010; 9. doi: Artn 51

10.1186/1476-4598-9-51.

50. Zhu B, Zhang P, Zeng P, Huang Z, Dong TF, Gui YK, Zhang GW. Tissue Factor Pathway Inhibitor-2 Silencing Promotes Hepatocellular Carcinoma Cell Invasion In Vitro. Anatomical Record-Advances in Integrative Anatomy and Evolutionary Biology. 2013; 296: 1708-16. doi: 10.1002/ar.22789.

51. Zhu B, Zhang P, Zeng P, Huang Z, Dong TF, Gui YK, Zhang GW. Tissue factor pathway inhibitor-2 silencing promotes hepatocellular carcinoma cell invasion in vitro. Anat Rec (Hoboken). 2013; 296: 1708-16. doi: 10.1002/ar.22789.

52. Huisman C, van der Wijst MG, Schokker M, Blancafort P, Terpstra MM, Kok K, van der Zee AG, Schuuring E, Wisman GB, Rots MG. Re-expression of Selected Epigenetically Silenced Candidate Tumor Suppressor Genes in Cervical Cancer by TET2-directed Demethylation. Mol Ther. 2016; 24: 536-47. doi: 10.1038/mt.2015.226.

53. Chu X, Zhao P, Lv Y, Liu L. Decreased expression of TFPI-2 correlated with increased expression of CD133 in cholangiocarcinoma. Int J Clin Exp Pathol. 2015; 8: 328-36. doi:

54. Sun FK, Sun Q, Fan YC, Gao S, Zhao J, Li F, Jia YB, Liu C, Wang LY, Li XY, Ji XF, Wang K. Methylation of tissue factor pathway inhibitor 2 as a prognostic biomarker for hepatocellular carcinoma after hepatectomy. J Gastroenterol Hepatol. 2016; 31: 484-92. doi: 10.1111/jgh.13154.

55. Henriksen SD, Madsen PH, Larsen AC, Johansen MB, Drewes AM, Pedersen IS, Krarup H, Thorlacius-Ussing O. Cell-free DNA promoter hypermethylation in plasma as a diagnostic marker for pancreatic adenocarcinoma. Clin Epigenetics. 2016; 8: 117. doi: 10.1186/s13148-016-0286-2.

56. Wang B, Yu L, Yang GZ, Luo X, Huang L. Application of multiplex nested methylated specific PCR in early diagnosis of epithelial ovarian cancer. Asian Pac J Cancer Prev. 2015; 16: 3003-7. doi:

57. Nazarian R, Jazirehi AR. TFPI2 methylation can serve as an epigenetic biomarker of metastatic melanoma. Epigenomics. 2013; 5: 252-53. doi:

**Figure** **legends**

**Figure 1: Flow diagram of the study design**

Candidate tumor suppressor genes were selected based on literature screening, and their methylation status in ESCC and adjacent control tissues were tested with the ESCC methylation data from the TCGA/GEO datasets. Moreover, the PBMC and PBL methylation datasets from healthy controls from GEO database were also included for further confirmation. Finally, due to the limitations of the multiplex PCR design, four of the six candidate tumor suppressor genes were then selected and validated with targeted bisulfite sequencing in an independent Chinese Han ESCC patients.

**Figure 2: Quality control and the methylation status of these four candidate genomic regions.**

Panels A represent the bisulfite conversion rate calculated by using the number of transformed C to T divided by the number of C of non-CpGs in each sample. Panel B represent the methylation status of the technical control LINE-1, which has been shown to be hypo-methylated in several different kinds of tumors. Panel C-F represents the CpG sites in regions covering *ADHFE1, EOMES, SALL1, TFPI2*, respectively. The x axis represents actual position of each CpG sites in the hg19 reference genome. The y axis represents the mean methylation percentage in the ESCC tumor tissues as well as the normal tissues for each of the CpG sites.

**Figure 3: The mean methylation status of each genomic region in tumor and normal tissues**

Panels A-D represent the mean methylation status of the genomic regions covering *ADHFE1, EOMES, SALL1, TFPI2*, respectively. Each point represents mean methylation percentage in a genomic region of a sample. The boxplot showed the overall methylation percentage of different groups in each genomic region.

**Figure. S1** PCA analysis of the ESCC and EAC adjacent normal tissues

**Figure. S2** PCA analysis for the ESCC and adjacent normal tissues in the validation dataset

**Figure. S3** The ROC (Receiver Operating characteristics) curve for the ESCC validation dataset

**Figure. S4** The ROC (Receiver Operating characteristics) curve for the subgroup analyzes

**Figure. S5** The expression profiles for the four genes after 5’-aza treatment

**Table S1** The methylation status of the CpG sites of the candidate tumor suppressor genes in the TCGA/GEO combined dataset

**Table S2** The methylation status of the candidate significant CpG sites in the combined datasets of TCGA/GEO ESCC samples along with the PBMC and PBL datasets from healthy samples

**Table S3** The designed primers of the five genomic regions for targeted bisulfite sequencing

**Table S4** Characteristics of the ESCC patients included in this study

**Table S5** The methylation status of the 4 genomic regions in the Young/Old subgroups

**Table S6** The methylation status of the 4 genomic regions in the Male/Female subgroups

**Table S7** The methylation status of the 4 genomic regions in the Smokers/Non-smokers subgroups

**Table S8** The methylation status of the 4 genomic regions in the Alcohol/ Non-alcohol subgroups