**Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC)**

**Abstract**

**Background**: DNA methylation has been implicated as the promising biomarker for precise cancer diagnosis. However, limited DNA-methylation based biomarkers have been found in esophageal squamous cell carcinoma (ESCC).

**Results**: In our study, high-throughput DNA methylation dataset (100 samples) of ESCC from the cancer genome atlas (TCGA) project was analyzed and validated in another independent dataset (12 samples) from gene expression omnibus (GEO) database. The methylation status from CD4+ and CD8+ T cells from healthy controls was also utilized for biomarker selection. Based on above procedures, five candidate CpGsites as well as its adjacent regions were further validated in another 94 pairs of ESCC tumor tissues and adjacent normal tissues from Chinese Han population using targeted bisulfite sequencing method. Logistic regression model was applied to the methylation status of genomic regions covering the five candidate CpGsites, yielding a robust performance (Sensitivity = 0.75, Specificity=0.88, AUC=0.85). Eight statistical models along with fivefold cross-validation were also applied, in which the SVM model reached the best accuracy in both train and test dataset (Accuracy = 0.82 and 0.80, respectively). In addition, subgroup analyzes revealed strong difference of diagnostic performance in the alcohol use and non-alcohol use subgroups.

**Conclusions**: In summary, based on the high-throughput DNA methylation dataset for biomarker screening, we identified five candidate CpGsites, further validated these CpGsites as well as their nearby regions with another independent ESCC samples with targeted bisulfite sequencing. Methylation profiles of the five genomic regions covering cg05249644 (STK3), cg15830431, cg20655070, cg26671652 (ZNF418) and cg27062795 (ZNF542) would be effective methylation-based testing for ESCC diagnosis.

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

**Background**

Esophageal cancer is one of the most aggressive cancers and is one of the leading causes of cancer death all over the world[1-3]. Esophageal cancer can be classified as the esophageal adenocarcinoma (EAC) and esophageal squamous cell carcinoma (ESCC) by histology[4, 5]. Recently, through the integration of the whole-genome and epigenome datasets from TCGA, EAC and ESCC were found to be significantly differed, which suggested that the EAC and ESCC subtype should be studied separately for precision cancer diagnosis and treatment[6]. Moreover, EAC and ESCC differed greatly in the demographic and geographic characteristics, risk factors as well as the pathogenesis[5]. The incidence of EAC is higher in the Western countries, while ESCC subtype is the predominant type in the Asians, especially in China, suggesting that the studies of ESCC in Chinese population is of great importance[7-10]. Currently, most of ESCCs are diagnosed at advanced stages, and studies have revealed that the 5-year survival rate is much higher in the early stage of ESCC than in the advanced stages of ESCC, indicating the strong need for effective early diagnosis methods[11-13].

DNA methylation is a key epigenetic modification in the mammalian genomes with many essential functions, including the repression of gene expression, genomic imprinting etc.[14-17]. Numerous studies have suggested that the altered DNA methylation patterns in tumor tissues may silence the tumor suppressor genes and activate the oncogenes through the hypo/hyper methylation status[18, 19]. In addition to its key functions in the tumorigenesis, it is also found that DNA methylation patterns were significantly different in each cancer types, indicating that DNA methylation could be applied as the promising biomarkers for cancer early detection[20-22]. However, despite of several diagnostic panels for ESCC detection, these studies were limited by the relatively small sample size, inaccurate methylation detection method and lack of validation dataset etc. The biomarkers with these limitations may pose a burden for the further prospective studies with large sample sizes.

Therefore, in the current study we firstly integrated the ESCC methylation datasets from GEO database and TCGA project for biomarker screening. After stringent selection procedures, several CpGsites were selected. As a result, we then constructed the multiplex PCR reaction system and designed the primers for each CpGsite. Due to the technical limitations of multiplex PCR, we finally selected five candidate CpGsites including cg05249644 (STK3), cg15830431, cg20655070, cg26671652 (ZNF418) and cg27062795 (ZNF542) for further validation. We then applied the targeted bisulfite sequencing method for quantifying the methylation status of these five CpGsites as well as its nearby genomic regions in 94 pairs of ESCC and normal tissues from Chinese Han population.

**Results**

**Integration of TCGA datasets and GEO datasets for biomarker discovery**

Public DNA methylation microarray datasets of ESCC were carefully searched. The esophageal carcinoma methylation dataset from TCGA was firstly identified, with 84 ESCC tumors and 3 ESCC adjacent normal tissues, as well as 78 EAC tumors and 13 EAC adjacent normal tissues. In order to achieve better statistical power, we combined the ESCC and EAC adjacent normal tissues together as the control samples due to its similarity, which can be validated from PCA analysis (Supplementary Figure 1). As a result, 84 ESCC tumors tissues as well as 16 adjacent normal tissues were included for the discovery stage analysis. In addition, GSE52826 dataset from GEO database, with relatively small sample size (4 ESCC tumors and 8 adjacent normal tissues) was also utilized as the validation dataset[23]. Based on our robust feature selection procedure and the primer design filtering for constructing the multiplex PCR reaction system, which was described in Methods (Figure 1), cg15830431 (*P*=2.20 × 10-4), cg19396867 (*P*=3.60 × 10-4), cg20655070 (*P*=1.71 × 10-3), cg26671652 (*P*=5.77 × 10-4), cg27062795 (*P*=3.60 × 10-4) were selected for further validation. It is shown that these five selected CpGsites were significantly hyper-methylated between the ESCC and adjacent normal tissues (Table 1). Moreover, the methylation status of these 5 CpGsites were also validated in another independent GEO dataset and showed similar results as in the TCGA dataset (Supplementary Table 1). In addition, all of these 5 CpGsites showed hypo-methylated state in the CD4+ and CD8+ T cells from healthy samples (Supplementary Table 1). Based on the above analysis, we believed that these 5 CpGsites would be potential biomarkers for the ESCC non-invasive diagnosis. As a result, we built a prediction model based on the logistic regression using all of the five predictors without adjustment for age, gender and other covariates, and yielding a great discrimination between ESCC and normal tissues (Sensitivity = 0.89, Specificity = 0.81, AUC = 0.87). To further evaluate and validate the diagnostic ability of these five CpGsites, we then conducted the validation study in another independent 94 paired ESCC and adjacent normal tissue samples in Chinese Han population.

**Methylation status validation of the 5 CpGsites with targeted bisulfite sequencing**

In order to validate the results from the previous analysis, we performed the study to detect the methylation status of the above five CpGsites with targeted bisulfite sequencing method in 94 pairs of ESCC and adjacent normal tissues. The characteristics of the ESCC patients were shown in Table 1. Targeted bisulfite sequencing method is an accurate and cost-effective method for detecting the methylation status of one or several selected regions with high coverage and efficiency[24-26]. By using the targeted bisulfite sequencing, we can test not only the five selected CpGsites but many more nearby CpGsites because of the read length. Quality control procedures were firstly applied to the bisulfite sequencing data. We found that the bisulfite conversion rate of each sample was higher than 98%, and no significant difference was found between the tumor and adjacent normal tissues, indicating the bisulfite conversion is efficient and reliable (Figure 3F). In addition, the samples as well as the CpGsites with high missing rate and low coverage were also filtered out as described in Materials and Methods. After the quality control procedures, 174 samples were remained for further study. The differential methylation analyzes were conducted for the five CpGsites as well as its nearby CpGsites, suggesting a strong difference between the ESCC and adjacent normal tissues (Figure 2A-E). Logistic regression model was then applied and showed a significant hyper-methylation status of the five selected CpGsites in the ESCC tissues (Table 2). Previous studies have revealed that the adjacent CpGsites on the same DNA molecules would share similar methylation patterns due to the locally coordinated activities of the DNMTs or TETs, which have been named as methylation haplotypes, epi-alleles or epi-haplotypes[27-29]. Due to the increased CpGsites in the region, the methylation haplotypes may be less susceptible to the complex and random environment stimulus and will be a better representative for methylation quantification[30]. As a result, we then averaged the methylation status of all the nearby CpGsites in a genomic region together as the representatives of the regions for further analysis (Figure 3A-E). Based on the mean methylation status of the five 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 0.64 to 0.74, while the specificity ranges from 0.82 to 0.90, and the AUC ranges from 0.76 to 0.84 (Table 3). Moreover, in the logistic model taking all of the five regions as predictors, we obtained the sensitivity of 0.75 and specificity of 0.88, as well as the AUC of 0.85 (Figure 3G).

**The diagnostic ability of the five genomic regions based on machine learning methods**

In order to get a better estimation of the diagnostic ability of the selected biomarker, several machine learning methods, including logistic regression, random forest (RF), supporting vector machine (SVM), neural network (NN), Naïve Bayes (NB), linear discriminant analysis (LDA), mixture discriminant analysis (MDA) and flexible discriminant analysis (FDA), were utilized to build the diagnostic models for ESCC classification. The mean methylation percentage of the CpGsites in each genomic region were utilized for analysis. Fivefold cross-validation method was also conducted to give a robust estimation of the performance of the models. The results were shown in Table 4. In the train stage, the sensitivity of all the models ranged from 0.63 to 0.76, while the specificity ranged from 0.77 to 0.89. It turned out that the logistic regression model and the SVM model both performed well in the accuracy. In the test stage, the sensitivity of the models ranged from 0.63 to 0.73, while the specificity ranged from 0.78 to 0.88. And the SVM model again achieved the highest accuracy, indicating the robustness and effectiveness of the model. In addition, we found that the diagnostic performance was similar between the train and test stage in all the models, suggesting the reliability of our results.

**Evaluation of diagnostic models in ESCC subgroup analysis**

Previous studies have found several risk factors for the incidence of ESCC, including age, gender, smoking status, alcohol status etc[31-34]. As a result, we conducted the subgroup analyzes according to these risk factors. The mean methylation percentage of each region was utilized for subgroup analysis. In the young/old subgroups, the median age of the samples was taken as the criteria for dividing the samples. And we found that there is no significant difference between the sensitivity, specificity and the AUC between the two subgroups (Supplementary Table 2). The overall AUCs using all the variables in the two subgroups were both 0.86 (Supplementary Figure3 A-B). In the male/female subgroups, we found that the diagnostic model performed better in the female subgroup than in the male subgroup (Supplementary Table 3), and the overall AUC of the female subgroup was much higher than that in the male subgroup (AUC: 0.89 vs. 0.84, Supplementary Figure3 C-D). Moreover, in the smoked/non-smoked subgroup analysis, no significant difference of the diagnostic performances was found (Supplementary Table 4 and Supplementary Figure3 E-F). However, strong difference was found in the analysis of the Alcohol/Non-alcohol subgroups. It is found that the AUC in four of the five genomic regions were elevated in the non-alcohol subgroup, especially the two genomic regions covering ZNF418 and ZNF542 (Supplementary Table 5). And the overall AUC of the non-alcohol subgroup was substantially higher than that of the alcohol subgroup (0.89 vs. 0.79 respectively, Supplementary Figure3 G-H). The strong difference of the diagnostic performance in the alcohol/non-alcohol subgroup indicated that the alcohol use may be vital to the epigenetic changes in ESCC, though further studies are urgently needed.

**Discussion**

DNA methylation plays a key role in the gene expression regulation and of great potential to be the non-invasive biomarker for cancer diagnosis and prognosis. And the early diagnosed ESCC patients will have longer survival time and lower mortality. Previous studies have found several candidate methylation biomarkers for the ESCC detection and prognosis as well as treatment response. In our study, we innovatively integrated the methylation dataset from TCGA project and the GEO dataset for the biomarker discovery, and removing the candidate biomarker with hyper-methylation status in the CD4+ and CD8+ T cells of healthy controls to guarantee its usefulness in the future non-invasive diagnosis. A novel DNA methylation biomarker panel, consisting of five CpGsites were then identified. Moreover, we validated these five CpGsites in another 94 pairs of ESCC tumors and its adjacent normal tissues with targeted bisulfite sequencing method, enabling us to not only detect the methylation status of five CpGsites but the five genomic regions as well. As a result, we then obtained the mean methylation percentage of each region as the representative, which is more robust than the single CpGsite itself. The methylation testing of these five genomic regions has a fair accuracy, sensitivity and specificity in different models, suggesting that the methylation testing of these five genomic regions may be promising biomarkers for the detection of ESCC. In addition, the subgroup analyzes identified that the diagnostic performance of the methylation testing is much better in the non-alcohol use samples than in the ESCC patients with alcohol use, suggesting the importance of taking the clinical data into considerations in ESCC diagnosis. Further studies may be required to explore the association between the methylation status of these five genomic regions and the use of alcohol.

Of the five genomic regions, two genomic regions covering cg15830431 and cg20655070 were not in the coding region of genes. However, the H3k4me3, and H3k4me1 and H3k27ac status of these two regions from ENCODE project showed that these two regions might be associated with the enhancers, indicating that these two regions might also have important regulatory functions in carcinogenesis (data not shown). Despite the two genomic regions in the non-coding regions of the genome, Serine/Threonine Kinase 3 (STK3) gene encodes a serine/threonine protein kinase and functions as a growth suppressor, which is one of the key components of the Hippo signaling pathway and involving in the cell apoptosis. Previous study has found that the deletion of STK3 in mouse liver results in tissue overgrowth and tumor development, demonstrating its importance in suppressing carcinogenesis[35]. And the hyper-methylation status of STK3 has been found in soft tissue sarcoma as well as head and neck squamous cell carcinoma, which is in accordance with our study in ESCC[36, 37]. ZNF418 (Zinc Finger Protein 418) is a kind of zinc finger-containing transcription factors which have been implicated as critical regulators for development and diseases. ZNF418 has been shown to be a transcriptional repressor, which may act as a negative regulator in MAPK signaling pathway[38]. ZNF542 (zinc finger protein 542) is a pseudogene, which also may be involved in transcriptional regulation. Studies have found the hypermethylation of ZNF542 in oropharyngeal squamous cell carcinoma and sporadic colorectal cancer[39, 40]. Moreover, a pan-cancer study analysis based on the TCGA methylation datasets identified the hyper-methylation status of ZNF542 in 12 cancer types[41].

In our study, we innovatively applied the targeted bisulfite sequencing method to explore the methylation status of our candidate CpGsites. And it turned out that the targeted bisulfite sequencing method is cost-effective and reliable to be a promising method in methylation testing. Comparing with the traditional MSP/qMSP method, it could reveal the methylation percentage of each CpGsite in a region. Moreover, the targeted bisulfite sequencing method is also simpler to be utilized and able to reach higher coverage than the BSP (bisulfite sequencing PCR) method. By incorporating all of the CpGsites in a region, we could obtain the overall methylation status of a region with higher confidence and precision for the further analysis.

The early diagnosis of esophageal squamous cell carcinoma is challenging due to its high heterogeneity and variability. And the single biomarker itself may not adequate for accurate diagnosis, which suggesting that a panel consisting of multi-biomarkers is essential. Though our DNA methylation-based biomarkers have reached a fair accuracy in distinguishing the ESCC tumors from normal tissues, part of the ESCC tumor tissues still remained misclassification. Integration analysis of multi-omics datasets, ranging from genomics, epigenomics, as well as proteomics etc. may reveal more heterogeneity in ESCC and identified more biomarkers for accurate non-invasive diagnosis of esophageal squamous cell carcinoma.

**Conclusion**

Integration analysis of ESCC high-throughput DNA methylation datasets from TCGA project and GEO database identified five CpGsites to be candidate biomarkers for ESCC diagnosis, which were further validated in an independent 94 pairs of ESCC tumors and normal tissues using targeted bisulfite sequencing method. Methylation profiles of the five genomic regions covering cg05249644 (STK3), cg15830431, cg20655070, cg26671652 (ZNF418) and cg27062795 (ZNF542) would be effective DNA methylation-based testing for ESCC diagnosis.

**Materials and Methods**

**Biomarker discovery based on the public datasets**

Public high-throughput DNA methylation microarray datasets were searched and obtained from the TCGA project and the GEO database. After careful search, a GSE record named GSE52826 was found. And we also downloaded the comprehensive methylation dataset of esophageal cancer from TCGA project. There is 4 ESCC and 8 normal tissue samples, while 84 ESCC and 3 normal tissues were included in TCGA dataset. In addition, we also found that there are 78 EAC and 13 adjacent normal tissues in TCGA dataset. To increase the sample size for a more robust biomarker discovery, the adjacent normal tissues of the EAC and ESCC were combined for analysis. Finally, 84 EACC as well as 16 normal tissues from TCGA were obtained for further analysis.

We then conducted the wilcoxon rank-sum test to find the most significantly differentially methylated CpGsites in the datasets derived from TCGA project. Simultaneously, the methylation differences of the candidate CpGsites in the GEO dataset were also obtained for further validation. It is of great importance that the methylation rate of the candidate biomarker should be very low in the normal tissues as well as in the peripheral blood in case of its future application in the non-invasive cancer diagnosis. As a result, we then filtered the candidate CpGsite with high methylation percentage in the CD4+ and CD8+ T cells of the healthy normal samples from our unpublished dataset. After that, we then constructed the multiplex PCR reaction system for the candidate CpGsites. Due to the technical limitations of the multiplex PCR, we then designed the primers and filtered some of the CpGsites which may in conflict with the system. Finally, we selected five of our candidate biomarkers for the further validation, including cg05249644, cg15830431, cg20655070, cg26671652 and cg27062795.

**Patients, samples and DNA**

ESCC samples and the paired adjacent normal tissues for validation study were obtained from xx Hospital between the years of xx and xx. 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 study was also approved by the ethics committee of Fudan University and the informed consent was obtained from all of the patients. In addition, all of the subjects were re-examined and confirmed by professional pathologists for histopathological diagnosis. All tissues were immediately frozen at -80 oC after surgical resection. Face-to-face interviews were conducted by professional investigators with a comprehensive questionnaire, including the clinical information on tobacco smoking, alcohol drinking and family history etc.

**Targeted bisulfite sequencing assay**

DNA extraction and bisulfite conversion were performed as previously described. Based on the genomic coordinates of the five candidate CpGsites, we carefully designed the primers in order to detect them in a panel (Supplementary Table 6). The net-PCR was performed firstly to amplify the targeted DNA sequence. Then the designed DNA fragments were sequenced by Illumina Hiseq 2000. BSseeker2 is one of the most commonly used tools for analyzing the bisulfite sequencing results and was applied in our study for mapping bisulfite treated reads as well as for methylation calling[42]. After methylation calling, we obtained the bisulfite conversion rate for each sample, and the samples with bisulfite conversion rate < 98% were firstly filtered out. After the preliminary analysis, we then calculated the average coverage as well as the missing rate for each CpGsite. The CpGsites with average coverage less than 20X and/or with missing rate > 0.20 were further filtered out. In addition, the samples with missing rate > 0.30 were filtered out finally.

**Statistical analysis and machine learning**

In the discovery stage, we applied the wilcoxon rank-sum test for testing the differential methylation status between cancer and normal tissues of each CpGsite. Further, differential methylation status in tumor and normal tissues of the candidate CpGsites were tested with logistic regression method. False discovery rate (FDR) correction was used for multiple test correction. In addition, the 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) were used for classifying the ESCC and normal tissues. To obtain the robust evaluation of the prediction ability with these biomarkers and methods, fivefold cross-validation was also applied. In addition, sensitivity, specificity and accuracy were obtained from the logistic regression model. All statistical analyses were all conducted using R 3.2.1[43].

**Abbreviations**

**Acknowledgements**

**Competing interests**

**Authors’ contributions**

**Reference**

1. De Angelis, R., et al., *Cancer survival in Europe 1999-2007 by country and age: results of EUROCARE--5-a population-based study.* Lancet Oncol, 2014. **15**(1): p. 23-34.

2. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2016.* CA Cancer J Clin, 2016. **66**(1): p. 7-30.

3. Torre, L.A., et al., *Global cancer statistics, 2012.* CA Cancer J Clin, 2015. **65**(2): p. 87-108.

4. Siewert, J.R. and K. Ott, *Are squamous and adenocarcinomas of the esophagus the same disease?* Semin Radiat Oncol, 2007. **17**(1): p. 38-44.

5. Enzinger, P.C. and R.J. Mayer, *Esophageal cancer.* N Engl J Med, 2003. **349**(23): p. 2241-52.

6. Cancer Genome Atlas Research, N., et al., *Integrated genomic characterization of oesophageal carcinoma.* Nature, 2017. **541**(7636): p. 169-175.

7. Brown, L.M., S.S. Devesa, and W.H. Chow, *Incidence of adenocarcinoma of the esophagus among white Americans by sex, stage, and age.* J Natl Cancer Inst, 2008. **100**(16): p. 1184-7.

8. Shibata, A., et al., *Trend in incidence of adenocarcinoma of the esophagus in Japan, 1993-2001.* Jpn J Clin Oncol, 2008. **38**(7): p. 464-8.

9. Kamangar, F., et al., *Esophageal cancer in Northeastern Iran: a review.* Arch Iran Med, 2007. **10**(1): p. 70-82.

10. Lin, Y., et al., *Epidemiology of esophageal cancer in Japan and China.* J Epidemiol, 2013. **23**(4): p. 233-42.

11. Zeng, H., et al., *Cancer survival in China, 2003-2005: a population-based study.* Int J Cancer, 2015. **136**(8): p. 1921-30.

12. Besharat, S., et al., *Inoperable esophageal cancer and outcome of palliative care.* World J Gastroenterol, 2008. **14**(23): p. 3725-8.

13. Coleman, M.P., et al., *EUROCARE-3 summary: cancer survival in Europe at the end of the 20th century.* Ann Oncol, 2003. **14 Suppl 5**: p. v128-49.

14. Bernstein, B.E., A. Meissner, and E.S. Lander, *The mammalian epigenome.* Cell, 2007. **128**(4): p. 669-81.

15. Guo, H., et al., *The DNA methylation landscape of human early embryos.* Nature, 2014. **511**(7511): p. 606-10.

16. Robertson, K.D., *DNA methylation and human disease.* Nat Rev Genet, 2005. **6**(8): p. 597-610.

17. Elhamamsy, A.R., *Role of DNA methylation in imprinting disorders: an updated review.* J Assist Reprod Genet, 2017.

18. Jones, P.A., *Cancer. Death and methylation.* Nature, 2001. **409**(6817): p. 141, 143-4.

19. Aran, D. and A. Hellman, *DNA methylation of transcriptional enhancers and cancer predisposition.* Cell, 2013. **154**(1): p. 11-3.

20. Witte, T., C. Plass, and C. Gerhauser, *Pan-cancer patterns of DNA methylation.* Genome Med, 2014. **6**(8): p. 66.

21. Yang, X., L. Gao, and S. Zhang, *Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns.* Brief Bioinform, 2016.

22. Guo, S., et al., *Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.* Nat Genet, 2017. **advance online publication**.

23. Li, X., et al., *Identification of a DNA methylome profile of esophageal squamous cell carcinoma and potential plasma epigenetic biomarkers for early diagnosis.* PLoS One, 2014. **9**(7): p. e103162.

24. Hebestreit, K., M. Dugas, and H.U. Klein, *Detection of significantly differentially methylated regions in targeted bisulfite sequencing data.* Bioinformatics, 2013. **29**(13): p. 1647-53.

25. Lee, E.J., et al., *Analyzing the cancer methylome through targeted bisulfite sequencing.* Cancer Lett, 2013. **340**(2): p. 171-8.

26. Deng, J., et al., *Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming.* Nat Biotechnol, 2009. **27**(4): p. 353-60.

27. Shoemaker, R., et al., *Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome.* Genome Res, 2010. **20**(7): p. 883-9.

28. Jones, B., *DNA methylation: Switching phenotypes with epialleles.* Nat Rev Genet, 2014. **15**(9): p. 572.

29. Schwartzman, O. and A. Tanay, *Single-cell epigenomics: techniques and emerging applications.* Nat Rev Genet, 2015. **16**(12): p. 716-26.

30. Guo, S., et al., *Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA.* Nat Genet, 2017. **49**(4): p. 635-642.

31. Morita, M., et al., *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**(2): p. 126-34.

32. Pandeya, N., et al., *Alcohol consumption and the risks of adenocarcinoma and squamous cell carcinoma of the esophagus.* Gastroenterology, 2009. **136**(4): p. 1215-24, e1-2.

33. Wang, J.M., et al., *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**(2): p. 171-6.

34. Nasrollahzadeh, D., 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**(11): p. 1857-63.

35. Zhou, D., et al., *Mst1 and Mst2 maintain hepatocyte quiescence and suppress hepatocellular carcinoma development through inactivation of the Yap1 oncogene.* Cancer Cell, 2009. **16**(5): p. 425-38.

36. Seidel, C., et al., *Frequent hypermethylation of MST1 and MST2 in soft tissue sarcoma.* Mol Carcinog, 2007. **46**(10): p. 865-71.

37. Steinmann, K., et al., *Frequent promoter hypermethylation of tumor-related genes in head and neck squamous cell carcinoma.* Oncol Rep, 2009. **22**(6): p. 1519-26.

38. Li, Y., et al., *ZNF418, a novel human KRAB/C2H2 zinc finger protein, suppresses MAPK signaling pathway.* Mol Cell Biochem, 2008. **310**(1-2): p. 141-51.

39. Lleras, R.A., et al., *Hypermethylation of a cluster of Kruppel-type zinc finger protein genes on chromosome 19q13 in oropharyngeal squamous cell carcinoma.* Am J Pathol, 2011. **178**(5): p. 1965-74.

40. Chen, C.H., et al., *DNA Methylation Identifies Loci Distinguishing Hereditary Nonpolyposis Colorectal Cancer Without Germ-Line MLH1/MSH2 Mutation from Sporadic Colorectal Cancer.* Clin Transl Gastroenterol, 2016. **7**(12): p. e208.

41. Gevaert, O., R. Tibshirani, and S.K. Plevritis, *Pancancer analysis of DNA methylation-driven genes using MethylMix.* Genome Biol, 2015. **16**: p. 17.

42. Guo, W., et al., *BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.* BMC Genomics, 2013. **14**: p. 774.

43. Dessau, R.B. and C.B. Pipper, *[''R"--project for statistical computing].* Ugeskr Laeger, 2008. **170**(5): p. 328-30.