**Table 1 Characteristics of the included ESCC patients**

| **Characteristics** | **Patient Distribution** |
| --- | --- |
| **N = 94** |
| **Age** | 64 (IQR = 57 to 70) |
| **Sex** |  |
| Male | 69 |
| Female | 25 |
| **Cigarette usea** |  |
| Yes | 58 |
| No | 36 |
| **Alcohol useb** |  |
| Yes | 34 |
| No | 58 |
| **T stagec** |  |
| T2 | 14 |
| T3 | 72 |
| T4 | 5 |
| **N stagec** |  |
| N0 | 44 |
| N1 | 38 |
| N2 | 7 |
| N3 | 3 |
| **M stagec** |  |
| M0 | 90 |
| M1 | 1 |

ESCC, esophageal squamous cell carcinoma; aYes represent the former and current smoker individuals. bYes represent individuals take alcohol formerlly and currently. cTNM Stages were assessed by the seventh edition of TNM classification criteria.

**Table 2 The methylation status of the 5 CpGsites in the TCGA dataset and the validation dataset**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CpGsite** | **Gene** | **Position** | **Relation to**  **CpG\_Island** | **McaMb** | **McoMb** | **Pvaluec** | **log10(OR)d** | **95% CId** | **Sense** | **Spece** | **AUCe** |
| TCGA | cg15830431 | STK3 | chr8:99952591 | Island | 0.28 | 0.09 | 2.20E-04 | 4.11 | 1.91-7.43 | 0.65 | 0.94 | **0.82** |
| cg19396867 | NAa | chr19:40314862 | N\_Shore | 0.45 | 0.20 | 3.60E-04 | 1.85 | 0.78-3.21 | 0.85 | 0.75 | 0.79 |
| cg20655070 | NAa | chr19:40315011 | Island | 0.44 | 0.19 | 1.71E-03 | 1.61 | 0.67-2.72 | 0.64 | 0.88 | 0.75 |
| cg26671652 | ZNF418 | chr19:58446312 | N\_Shore | 0.35 | 0.16 | 5.77E-04 | 1.95 | 0.67-3.61 | **0.86** | 0.75 | 0.78 |
| cg27062795 | ZNF542 | chr19:56879613 | Island | 0.43 | 0.17 | 3.60E-04 | 2.93 | 1.65-4.44 | **0.86** | 0.81 | 0.80 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Validation | cg15830431 | STK3 | chr8:99952591 | Island | 0.20 | 0.07 | 1.25E-06 | 3.04 | 1.82-4.53 | 0.66 | 0.77 | 0.71 |
| cg19396867 | NA | chr19:40314862 | N\_Shore | 0.37 | 0.12 | 2.71E-11 | 2.83 | 1.93-3.91 | 0.65 | 0.88 | 0.80 |
| cg20655070 | NA | chr19:40315011 | Island | 0.31 | 0.09 | 8.04E-10 | 3.01 | 2.02-4.22 | 0.62 | 0.89 | 0.77 |
| cg26671652 | ZNF418 | chr19:58446312 | N\_Shore | 0.32 | 0.11 | 4.82E-11 | 3.20 | 2.18-4.39 | 0.58 | **0.93** | 0.79 |
| cg27062795 | ZNF542 | chr19:56879613 | Island | 0.43 | 0.12 | 1.23E-12 | 2.55 | 1.77-3.50 | **0.72** | 0.82 | **0.83** |

a NA indicated that the CpGsite is located outside of the coding region of the genes. bMcaM represents the mean methylation percentage of the cases, and the

McoM represents the mean methylation percentage of the controls. cPvalue is calculated through the wilcoxon rank-sum test following with FDR (false discovery rate) adjustment for multiple correction. d OR and 95% CI were conducted through logistic regression. eSens is short for sensitivty, while Spec is short for specificity, and the AUC is short for area under cuve. The sensitivity, specificity as well as the AUC were both with a logistic regression prediction model without adjustment for gender, age and smoking status and alcohol status.

**Table 3 The mean methylation status of the 5 genomic regions in the validation datasets**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Genomic**  **Regiona** | **No.**  **CpGsitesb** | **CpGsite**  **Included** | **Gene** | **McaMc** | **McoMc** | **Pvalued** | **log10(OR)e** | **95% CIe** | **Sensf** | **Specf** | **AUCf** |
| chr8:99952469-99952722 | 19 | cg15830431 | STK3 | 0.35 | 0.16 | 4.20E-09 | 2.82 | 1.83-4.03 | 0.64 | 0.82 | 0.76 |
| chr19:40314817-40314928 | 6 | cg19396867 | NA | 0.36 | 0.12 | 9.60E-11 | 2.90 | 1.97-4.03 | 0.61 | **0.90** | 0.79 |
| chr19:40314939-40315133 | 17 | cg20655070 | NA | 0.31 | 0.12 | 1.80E-09 | 3.61 | 2.42-5.06 | 0.60 | **0.90** | 0.77 |
| chr19:58446187-58446437 | 19 | cg26671652 | ZNF418 | 0.50 | 0.26 | 1.10E-13 | 3.46 | 2.52-4.54 | **0.74** | 0.86 | **0.84** |
| chr19:56879517-56879735 | 25 | cg27062795 | ZNF542 | 0.41 | 0.14 | 5.20E-13 | 2.81 | 1.94-3.86 | 0.71 | 0.84 | 0.83 |

aGenomic region represents the genomic coverage of the reads with targeted bisulfite sequencing, and the genomic coordinates shown here is based on the hg19 version of the genome. bNo.CpGsites represents the number of the CpGsites in each region. cMcaM represents the mean methylation percentage of the cases in each region, which consisitng of several CpGsites, while the McoM represents the mean methylation percentage of the controls in each region. dPvalue is calculated through the wilcoxon rank-sum test following with FDR (false discovery rate) adjustment for multiple correction. e OR and 95% CI were conducted through logistic regression. eSens is short for sensitivty, while Spec is short for specificity, and the AUC is short for area under cuve. The sensitivity, specificity as well as the AUC were both with a logistic regression prediction model without adjustment for gender, age and smoking status and alcohol status.

**Table 4 Diagnosis accuracy, sensitivity and specificity of different classification models with five fold cross-validation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Methods** | **Train** | | | **Test** | | |
| **Sensitivity** | **Specificity** | **Accuracy** | **Sensitivity** | **Specificity** | **Accuracy** |
| Logistic Regression | 0.75 | **0.89** | **0.82** | **0.73** | 0.86 | 0.79 |
| Random Forest | 0.73 | 0.77 | 0.75 | **0.73** | 0.78 | 0.75 |
| Supporting Vector Machine | 0.74 | **0.89** | **0.82** | **0.73** | 0.87 | **0.80** |
| Naive Bayes | 0.63 | **0.89** | 0.76 | 0.63 | **0.88** | 0.75 |
| Neural Network | **0.76** | 0.87 | 0.81 | 0.72 | 0.81 | 0.76 |
| Linear Discriminant Analysis | 0.73 | 0.88 | 0.80 | 0.71 | 0.87 | 0.79 |
| Mixture Discriminant Analysis | 0.74 | **0.89** | 0.81 | 0.71 | 0.84 | 0.77 |
| Flexible Discriminant Analysis | 0.73 | 0.88 | 0.80 | 0.71 | 0.87 | 0.79 |

The mean methylation percentage of each genomic region was considered as the independent variable for constructing the models, which means that all of the models were based on these five independent variables without adjustment for gender, age and smoking status and alcohol status. Sensitivity, specificity and classification accuracy were the mean value in five fold cross-validations with 1,000 replications.

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

Candidate biomarkers were selected from the high-throughput DNA methylation microarrays from TCGA project, and further validatated with the ESCC methylation data in GEO dataset, as well as the CD4+/CD8+ T cells from healthy controls. Histone methylation status of the candidate biomarkers were also downloaded and checked from the ENCODE project. Based on our preliminary screening, the candidate methylation biomarkers for ESCC were then further validated with targeted bisulfite sequencing in an independent Chinese Han ESCC patients.

**Figure 2: The methylation status of the CpGsites in the five genomic regions.**

A-E represents the methylation status of the CpGsites in regions covering STK3, cg19396867, cg20655070, ZNF418 and ZNF542, respectively. The x axis represents the distance (bp) to the candidate CpGsite of the discovery stage. The negative value represents the upstream of the selected CpGsite, while the postive value of the x axis represents the downstream of the selected CpGsite. The y axis represents the mean methylation percentage in the ESCC tumor tissues as well as the normal tissues in each of the CpGsite.

**Figure 3: The mean methylation status of each genomic region and bisulfite conversion efficiency between ESCC tumors and normal tissues as well as the overall ROC curve.**

Figure A-E represents the mean methylation status of the genomic regions covering STK3, cg19396867, cg20655070, ZNF418 and ZNF542, respectively. Each point represents mean methylation percentage in a genomic region of a sample. The boxplot showed overall methylation percentage of different groups in each genomic region. Figure F represents the bisulfite conversion efficiency between ESCC and adjacent normal tissues. Bisulfite conversion efficiency was calculated by using the number of transformed C to T dividing all the number of C in each sample. Figure G represents the overall ROC (Receiver Operating characterstics) curve, which was calculated through a logistic regression model, incorporating the mean methylation percentage of the five genomic regions as the variables, and without the adjustment for gender, age and smoking status and alcohol status.