**Table 1 The methylation status of the 6 CpG sites in the TCGA dataset and the validation dataset**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **CpGsite** | **Gene** | **Position** | **Relation to**  **CpG\_Island** | **McaMa** | **McoMa** | **P valueb** | **Sensc** | **Specc** | **AUCc** |
| TCGA | cg20295442 | ADHFE1 | chr8:67344665 | Island | 0.26 | 0.15 | 0.18 | 0.42 | 0.85 | 0.61 |
| cg20912169 | ADHFE1 | chr8:67344720 | Island | 0.26 | 0.14 | 0.22 | 0.46 | 0.85 | 0.60 |
| cg22383888 | EOMES | chr3:27764816 | N\_shore | 0.53 | 0.22 | **3.10×10-7** | 0.77 | 0.92 | 0.87 |
| cg04550052 | SALL1 | chr16:51184355 | Island | 0.46 | 0.22 | **7.10×10-5** | 0.79 | 0.85 | 0.78 |
| cg04698114 | SALL1 | chr16:51184379 | Island | 0.47 | 0.22 | **1.90×10-4** | 0.77 | 0.85 | 0.77 |
| cg12973591 | TFPI2 | chr7:93519473 | Island | 0.33 | 0.15 | 0.06 | 0.63 | 0.88 | 0.65 |
|  |  |  |  |  |  |  |  |  |  |  |
| Validation | cg20295442 | ADHFE1 | chr8:67344665 | Island | 0.18 | 0.09 | **5.10×10-3** | 0.28 | 0.95 | 0.63 |
| cg20912169 | ADHFE1 | chr8:67344720 | Island | 0.17 | 0.07 | **2.10×10-3** | 0.30 | 0.94 | 0.64 |
| cg22383888 | EOMES | chr3:27764816 | N\_shore | 0.31 | 0.11 | **3.30×10-9** | 0.55 | 0.94 | 0.77 |
| cg04550052 | SALL1 | chr16:51184355 | Island | 0.29 | 0.13 | **2.50×10-4** | 0.44 | 0.91 | 0.67 |
| cg04698114 | SALL1 | chr16:51184379 | Island | 0.34 | 0.16 | **1.10×10-6** | 0.47 | 0.96 | 0.72 |
| cg12973591 | TFPI2 | chr7:93519473 | Island | 0.25 | 0.08 | **3.30×10-5** | 0.49 | 0.89 | 0.69 |

aMcaM represents the mean methylation percentage of the cases, and the McoM represents the mean methylation percentage of the controls. bP value is calculated through the Wilcoxon rank-sum test followed by FDR (false discovery rate) adjustment for multiple correction. cSens = sensitivity, while Spec = specificity, AUC = area under curve. 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 2 The mean methylation status of the 4 genomic regions in the validation datasets**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Genomic**  **Regiona** | **No.**  **CpG sitesb** | **CpGsite**  **Included** | **Gene** | **McaMc** | **McoMc** | **P valued** | **log10(OR)e** | **95% CIe** | **Sensf** | **Specf** | **AUCf** |
| chr8:67344610-67344805 | 24 | cg20295442,  cg20912169 | ADHFE1 | 0.24 | 0.15 | **1.70×10-3** | 2.20 | 1.00-3.72 | 0.29 | 0.94 | 0.64 |
| chr3:27764697-27764940 | 8 | cg22383888 | EOMES | 0.38 | 0.24 | **2.90×10-9** | 3.88 | 2.51-5.51 | 0.69 | 0.77 | 0.78 |
| chr16:51184268-51184468 | 18 | cg04550052,  cg04698114 | SALL1 | 0.37 | 0.19 | **3.90×10-7** | 2.41 | 1.51-3.51 | 0.53 | 0.90 | 0.74 |
| chr7:93519367-93519503 | 13 | cg12973591 | TFPI2 | 0.28 | 0.13 | **3.40×10-6** | 3.82 | 2.26-5.89 | 0.50 | 0.91 | 0.71 |

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.CpG sites represents the number of the CpG sites in each region. cMcaM represents the mean methylation percentage of the cases in each region, which consisting of several CpG sites, while the McoM represents the mean methylation percentage of the controls in each region. dP value 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. fSens = sensitivity, while Spec = specificity, AUC = area under curve. 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 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 | 0.683 | 0.873 | 0.773 | 0.645 | 0.830 | 0.732 |
| Random Forest | 0.726 | 0.739 | 0.732 | **0.728** | 0.741 | 0.734 |
| Supporting Vector Machine | 0.635 | 0.907 | 0.764 | 0.599 | 0.881 | 0.731 |
| Naive Bayes | 0.539 | **0.916** | 0.718 | 0.532 | **0.910** | 0.709 |
| Neural Network | 0.701 | 0.841 | 0.768 | 0.667 | 0.794 | 0.726 |
| Linear Discriminant Analysis | 0.617 | 0.906 | 0.754 | 0.594 | 0.894 | **0.735** |
| Mixture Discriminant Analysis | 0.618 | 0.868 | 0.736 | 0.564 | 0.843 | 0.695 |
| Flexible Discriminant Analysis | 0.616 | 0.907 | 0.754 | 0.594 | 0.894 | **0.735** |
| Gradient Boosting Machine | **0.826** | 0.856 | **0.840** | 0.699 | 0.728 | 0.713 |

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, smoking status and alcohol status. Sensitivity, specificity and classification accuracy were the mean value in fiv-efold cross-validations with 1,000 replications.