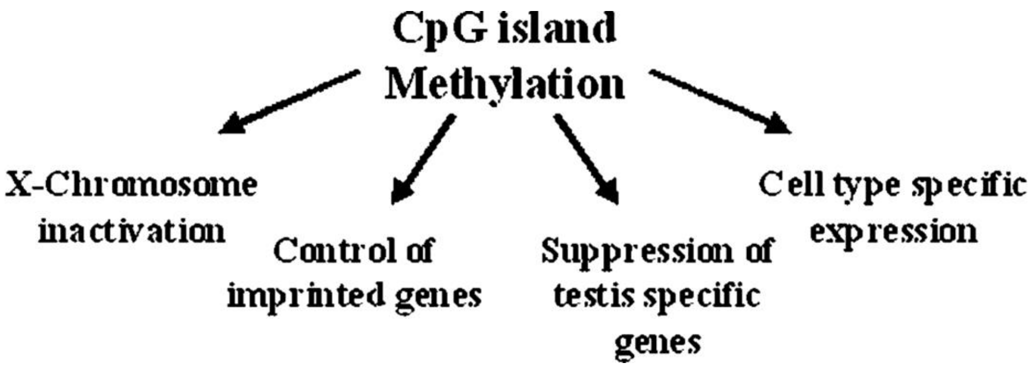
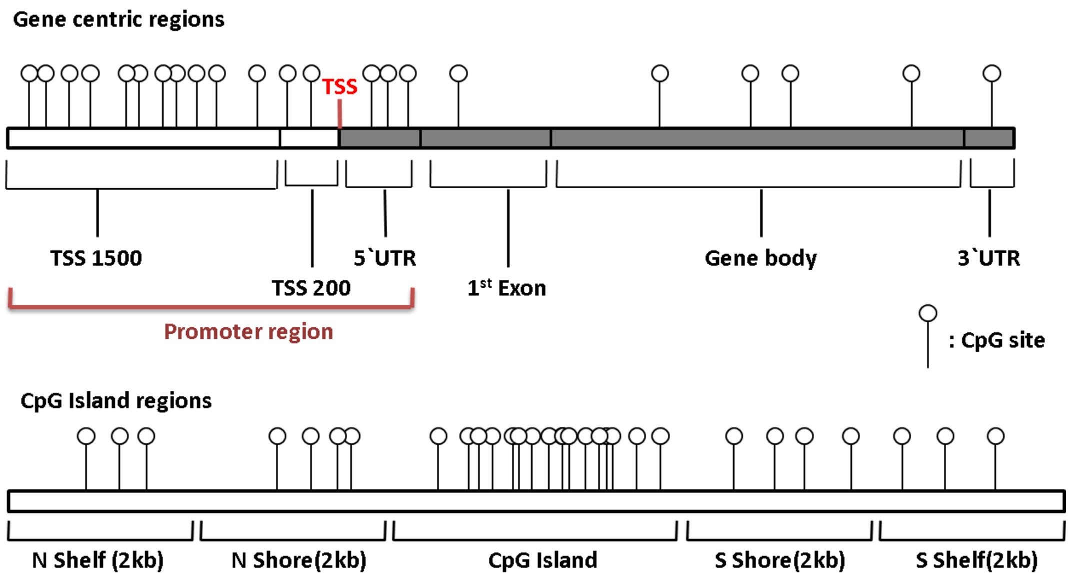
The 16th KJC Bioinformatics Symposium

Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis

Tieliu Shi
tlshi@bio.ecnu.edu.cn
The Center for bioinformatics
East China Normal University
Aug. 30, 2018
Hayama Compus, Japan

Why Do We Focus on DNA Methylation?

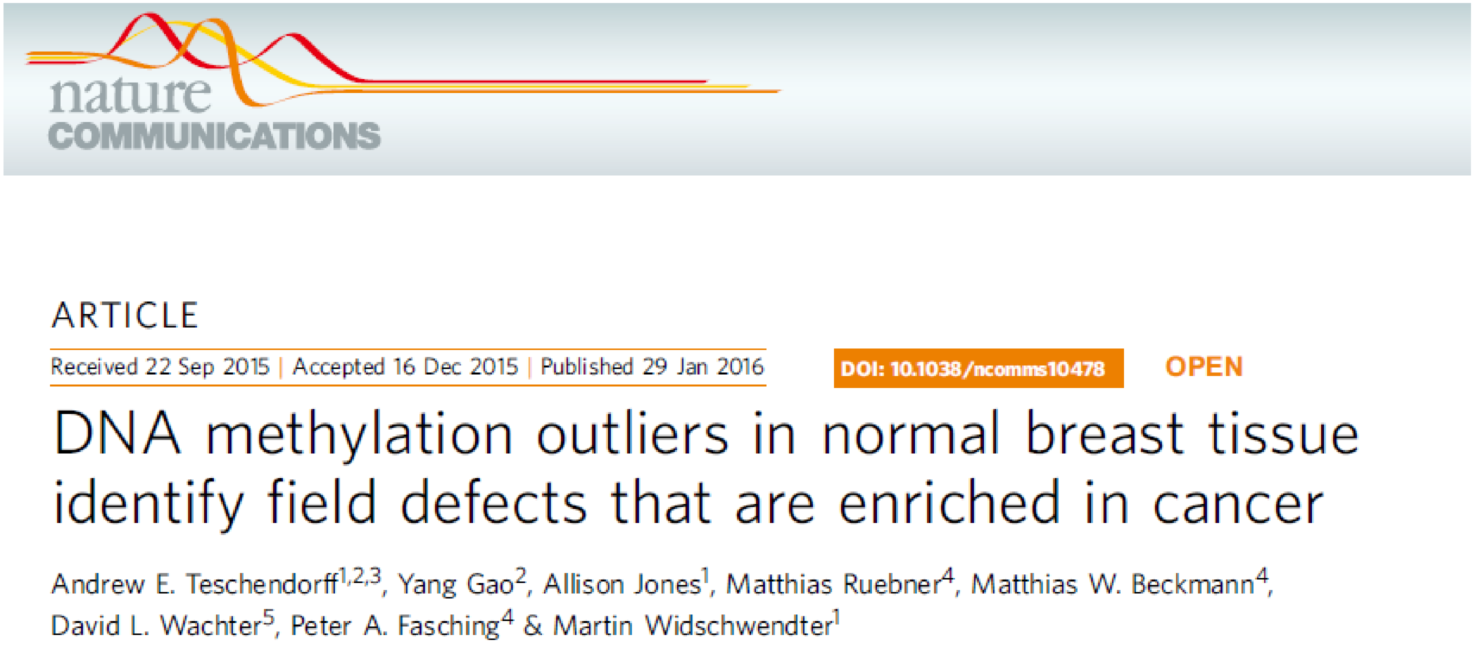

CpG island Methylation
X-Chromosome inactivation
Control of imprinted genes
Suppression of testis specific genes
Cell type specific expression



- DNAm changes in the non-islands regions, such as shores and gene bodies also play important roles in gene expression regulation.
- Aberrant methylation could be used as biomarker for clinical decisions, such as cancer diagnosis and prognosis.
- DNA methylation markers can also be used to predict the origin of tumors with metastases.
- The Cancer Genome Atlas Project (TCGA) now provides unprecedented cancer genomic and methylation data resources for various cancer researches.

Progressive Change in DNA Methylation from Normal Tissue to Breast Cancer Tissue

DNA methylation field defects in cancer tissue



nature COMMUNICATIONS

ARTICLE

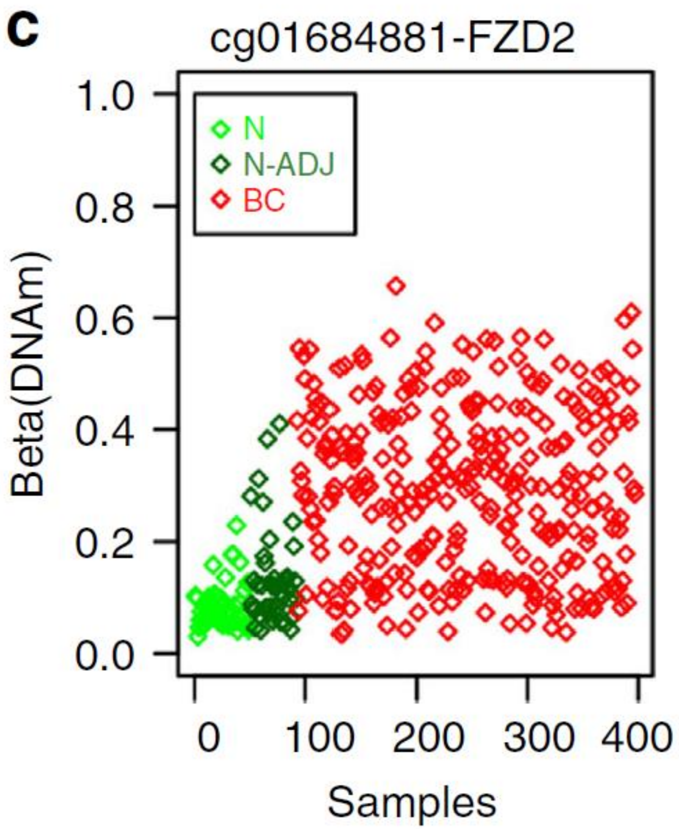
Received 22 Sep 2015 | Accepted 16 Dec 2015 | Published 29 Jan 2016

DOI: 10.1038/ncomms10478

OPEN

DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer

Andrew E. Teschendorff[1,2,3], Yang Gao[2], Allison Jones[1], Matthias Ruebner[4], Matthias W. Beckmann[4], David L. Wachter[5], Peter A. Fasching[4] & Martin Widschwendter[1]



Progression of field defects in breast cancer. (c) Example of a DNAm profile of a hypervariable and hypermethylated DVMC, showing the progressive change in DNA methylation. N represents normal tissue from cancer-free women, NADJ for age-matched normal samples adjacent to breast cancers.

DNA Methylation Markers Distinguish Prostate Cancers from Benign Adjacent Tissue

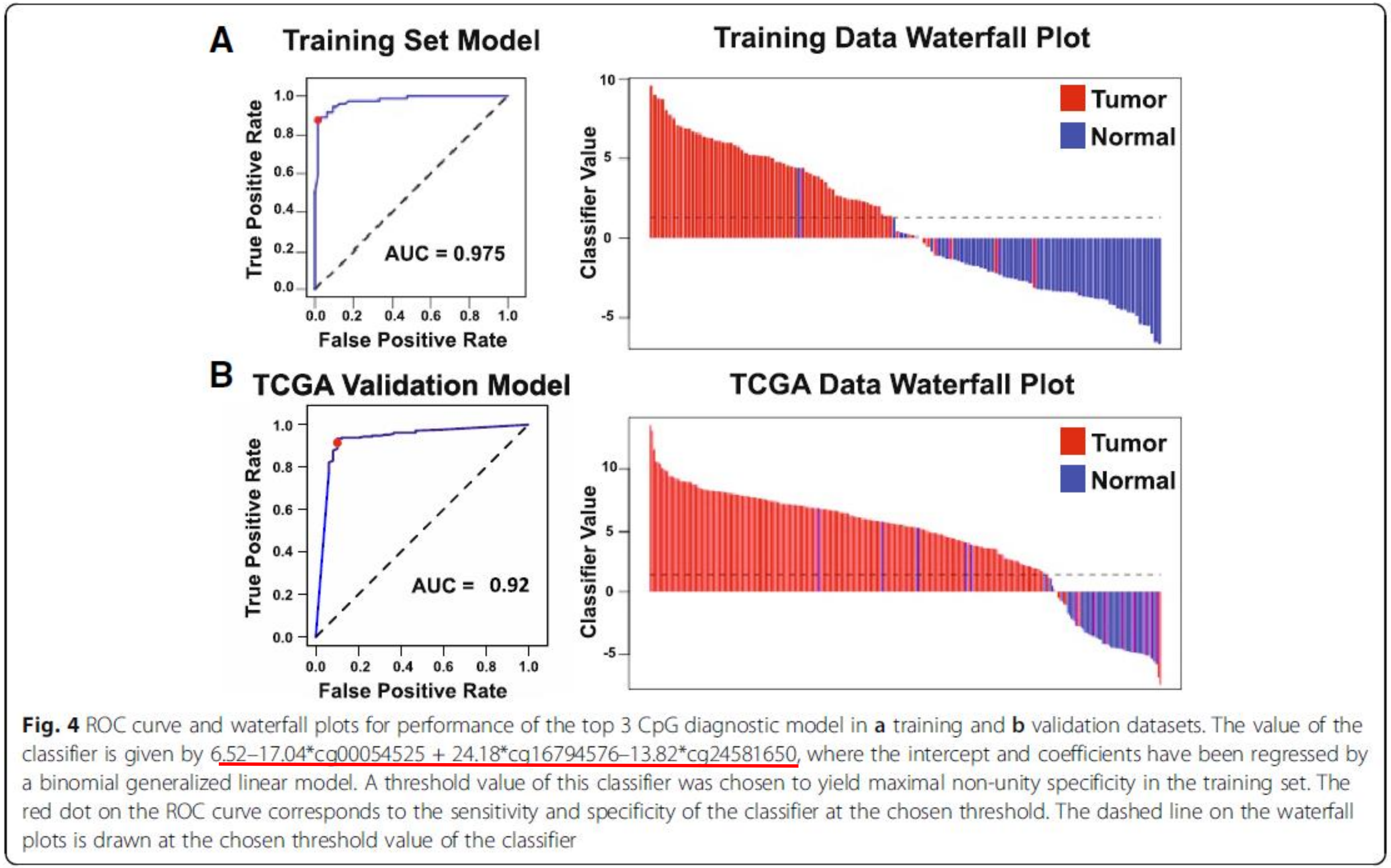


Fig. 4 ROC curve and waterfall plots for performance of the top 3 CpG diagnostic model in **a** training and **b** validation datasets. The value of the classifier is given by 6.52–17.04*cg00054525 + 24.18*cg16794576–13.82*cg24581650, where the intercept and coefficients have been regressed by a binomial generalized linear model. A threshold value of this classifier was chosen to yield maximal non-unity specificity in the training set. The red dot on the ROC curve corresponds to the sensitivity and specificity of the classifier at the chosen threshold. The dashed line on the waterfall plots is drawn at the chosen threshold value of the classifier

BMC Cancer, 17(1): 273, 2017

Methylation Markers can be Used for Diagnosis and Prognosis of Common Cancers

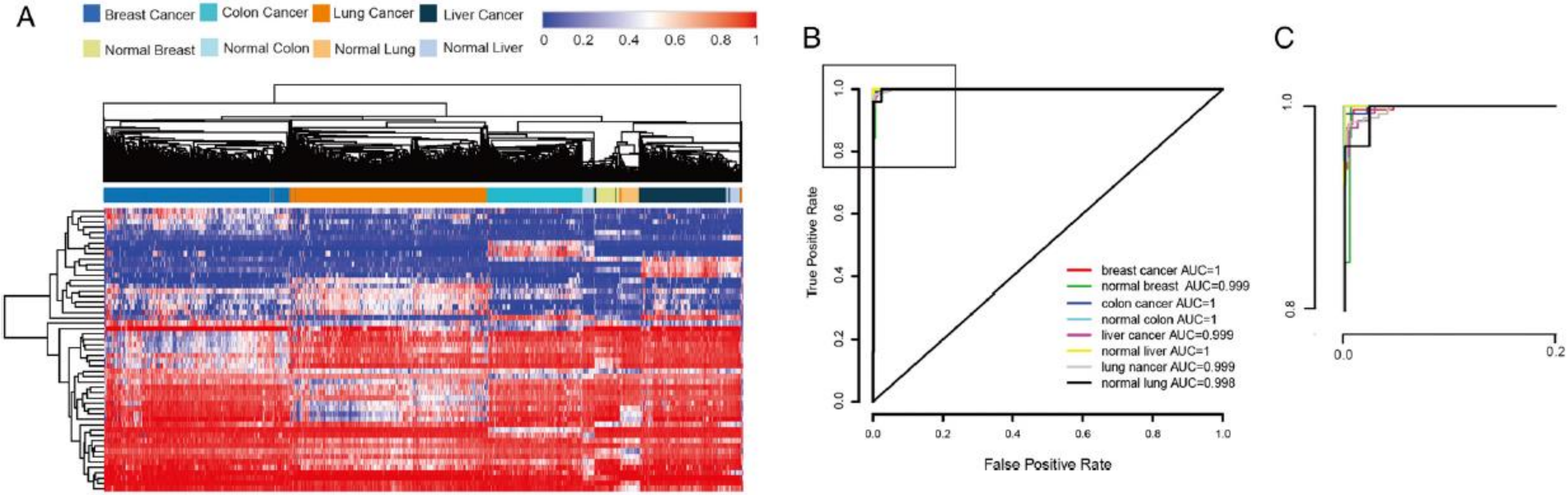
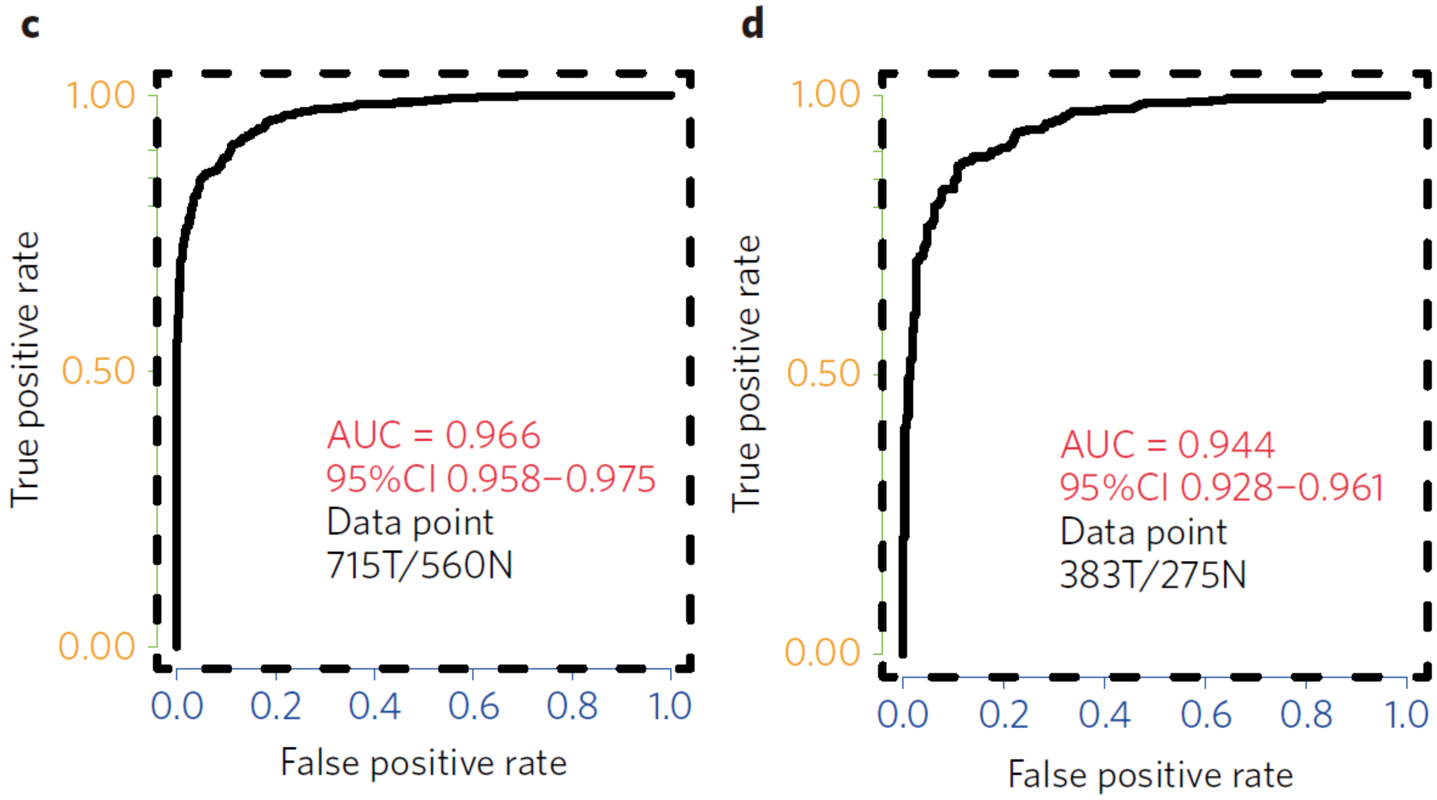


Fig. **Methylation signatures can differentiate different cancer types from corresponding normal tissues**.
(A) Unsupervised hierarchical clustering and heat map presentation associated with the methylation profile (according to the color scale shown) in different cancer types. (B) ROC curve showing the high sensitivity and specificity in predicting different cancer types. (C) Zoom-in view of the block diagram in B.

PNAS, 114:7414-7419, 2017

ctDNA Methylation Markers for Diagnosis and Prognosis of Hepatocellular Carcinoma



ROC of the diagnostic prediction model with methylation markers in the training (**c**) and validation data sets (**d**).

Nature materials, 16:1155, 2017

All of the studies confirmed that there are distinguished DNA methylation patterns between cancer tissues and their related normal tissues.

Data Resources and Methods

Datasets (TCGA)

- **Gene expression data**:

  Level 3 expression data from TCGA, log2(x+1) transformed.

- **DNA methylation data**:

  (i) The probes mapped to sex chromosomes were removed;

  (ii) The samples with missing data (i.e. NAs) in more than 30% of the probes were excluded;

  (iii) The probes with missing data in more than 30% of the samples were discarded;

  (iv) The rest of the probes with NAs were imputed using the EMimpute algorithm;

  (v) BMIQ was employed to correct for the type II probe bias.

Validation dataset

- GEO: GSE69914, GSE76938, GSE48684, GSE73549, GSE65820, GSE66836, GSE89852, GSE58999 and GSE38240

Feature Selection & Function Analysis

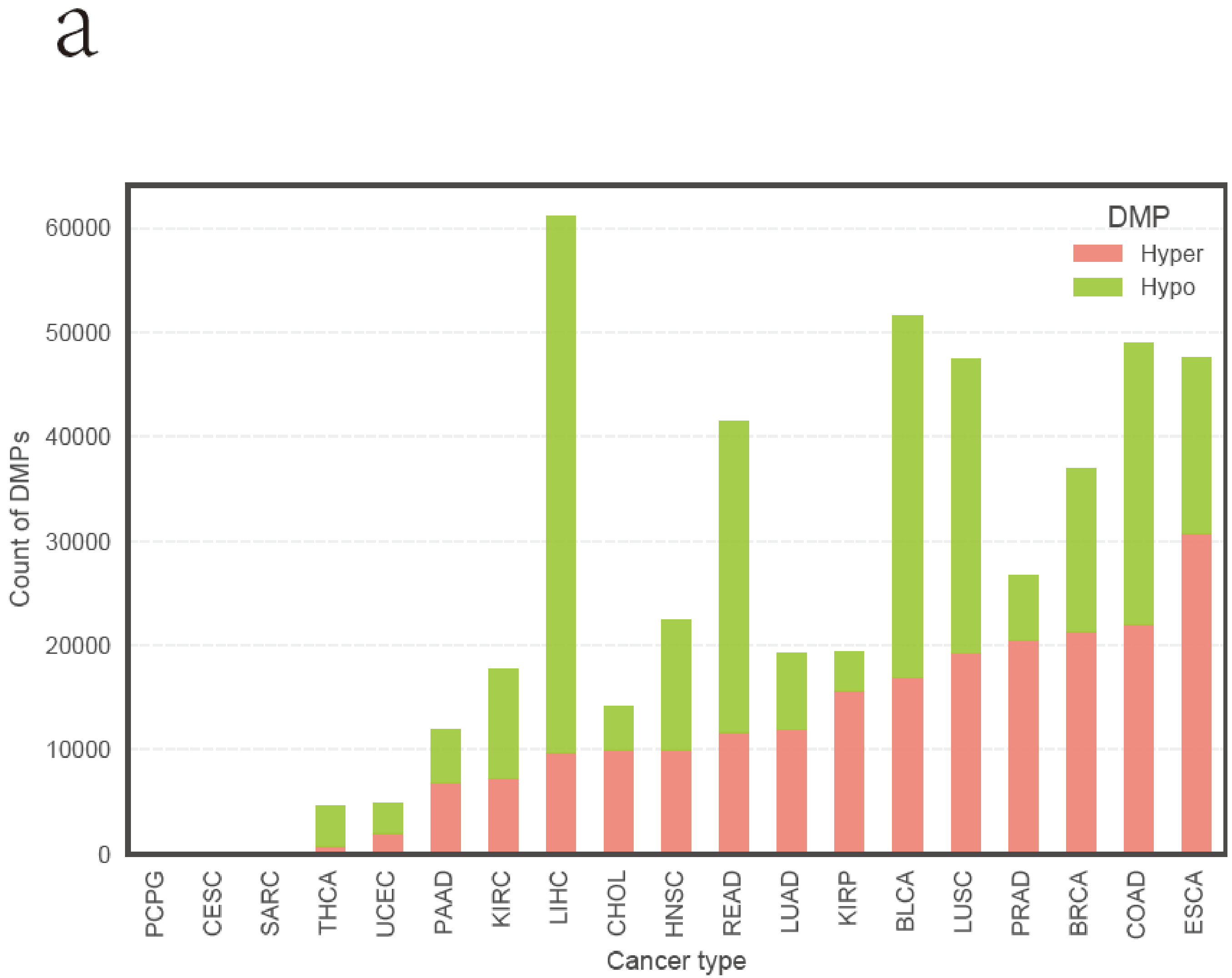
- Definition of differentially methylated probes between tumor and normal samples
  - |β-difference| > 0.2 and a false discovery rate (FDR) corrected P-value (Benjamini/Hochberg) < 0.05
- Definition of differentially methylated probes between different cancer types
  - |β-difference| > 0.3 and FDR < 0.01
- Feature selection & Tumor specific multiclass classifier
  - Recursive feature elimination & logistic regression, OneVsRest Classifier
- Statistical analysis
  - All statistical analyses and visualization were performed with Python3.5.2 on anaconda3-4.0.0.
- Gene ontology enrichment analysis and pathway enrichment analysis
  - DAVID

Data Resources and Methods for Survival Analysis

Cox regression analysis

- (1) Standard deviation (SD) across all tumor samples of 26 cancers should be > 0.2.
- (2) FDR (Benjamini/Hochberg method) for every probe was calculated via univariate cox regression in each cancer, the probes with FDR < 0.05 were retained for further filtration.
- (3) Log-rank test P-value for survival time among tumor samples should be < 0.05.
- (4) Multivariate cox regression was performed for the left probes, and stepwise regression was conducted, the probes of multivariate cox regression p-value < 0.05 were removed from the feature set in each iteration.
- (5) The remaining probes were used to fit the prognostic classifier. Python package lifelines and cox’s proportional hazard model was implemented in cox regression analysis.

Methylation Profiles of Different Cancers Vary Tremendously

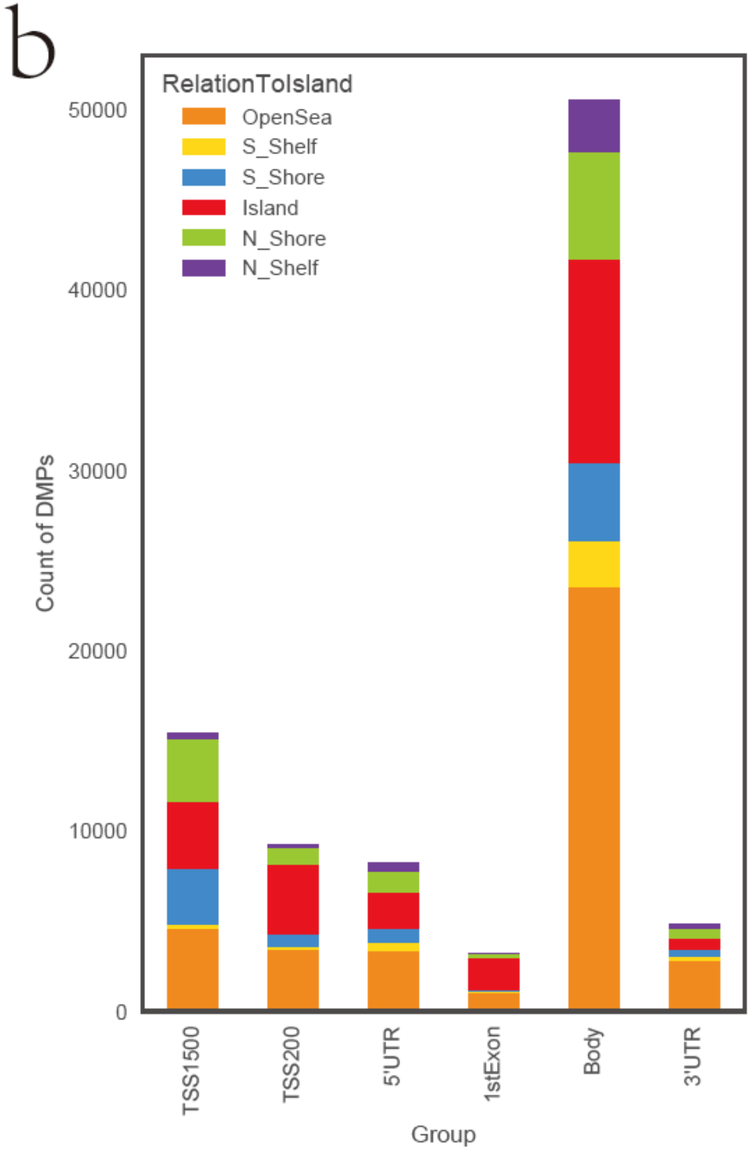


Countplot or differential methylated probes in different cancer types.

The number of hypermethylated and hypomethylated CpG sites vary greatly in 18 different cancer types.

1. Esophageal carcinoma (ESCA) has the largest count of hypermethylated CpG sites, whereas pheochromocytoma and paraganglioma (PCPG) has the least.
2. Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) has the least count of hypomethylated CpG sites, while liver hepatocellular carcinoma (LIHC) shows the highest number of hypomethylated CpG sites

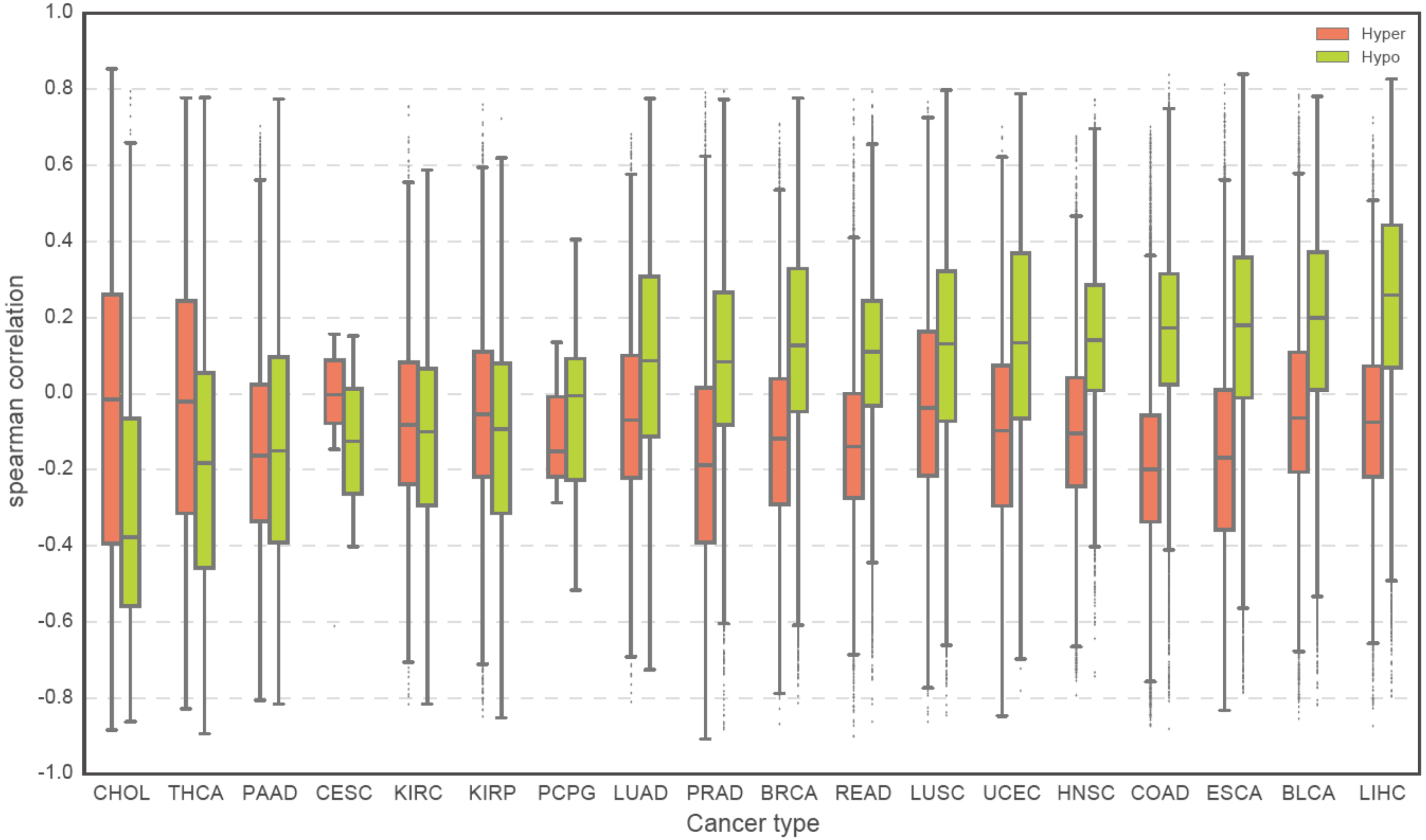
Methylation Profiles of Different Cancers Vary Tremendously



The distribution of differential methylated probes based on Relation to Island vary significantly in different groups.

1. Differential methylated CpG sites (DMCs) located at gene body are far more than that of other regions, and OpenSea holds a large proportion DMCs among all different Relation To Island (OpenSea, S_Shelf, S_Shore, Island, N_Shore and N_Shelf).
2. The number of differential methylated CpG sites located in the gene body regions is the highest among different genomic regions (TSS1500, TSS200, 5'UTR, 1stExon, Body and 3'UTR).

Methylation Profiles of Different Cancers Vary Tremendously



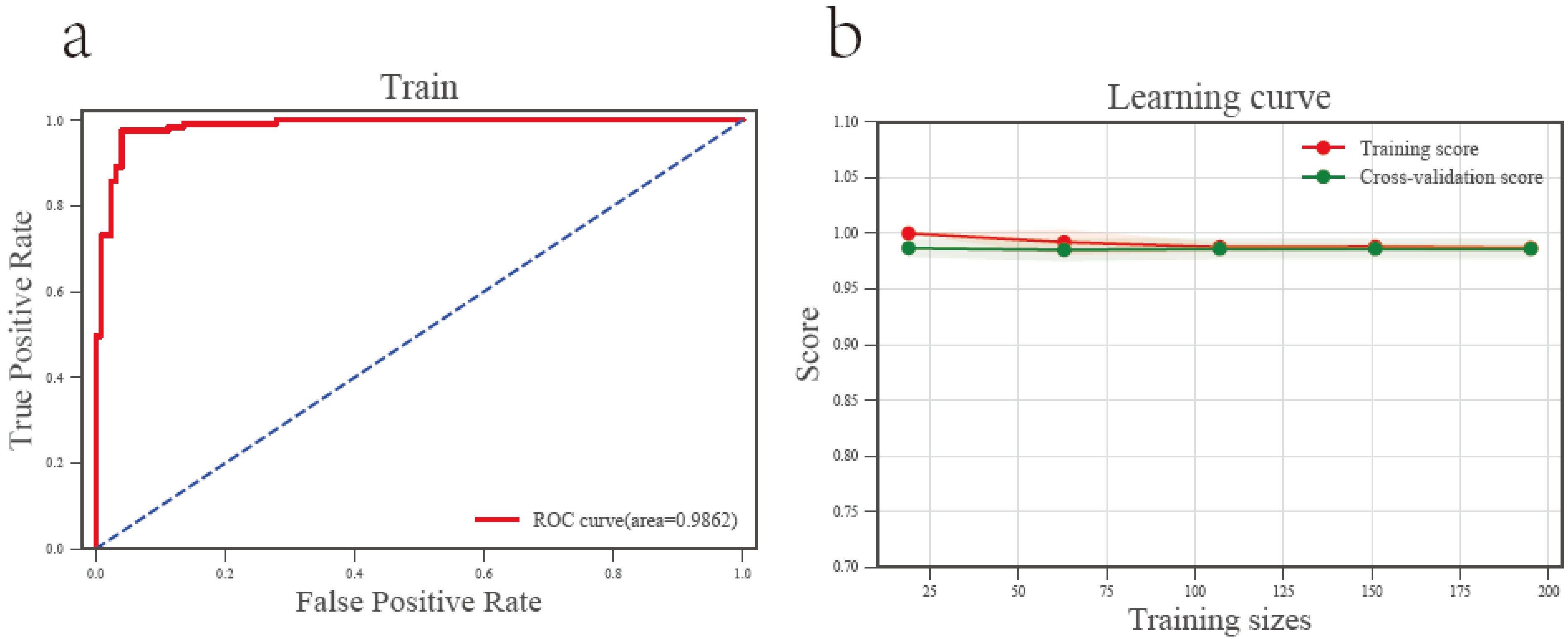
Spearman’s correlation analysis between the methylation level of CpG sites and the expression of their corresponding genes for each cancer.

Indicating that aberrant DNAms in different tumors may have different functions.

Boxplot of Spearman’s correlation among different cancer types.

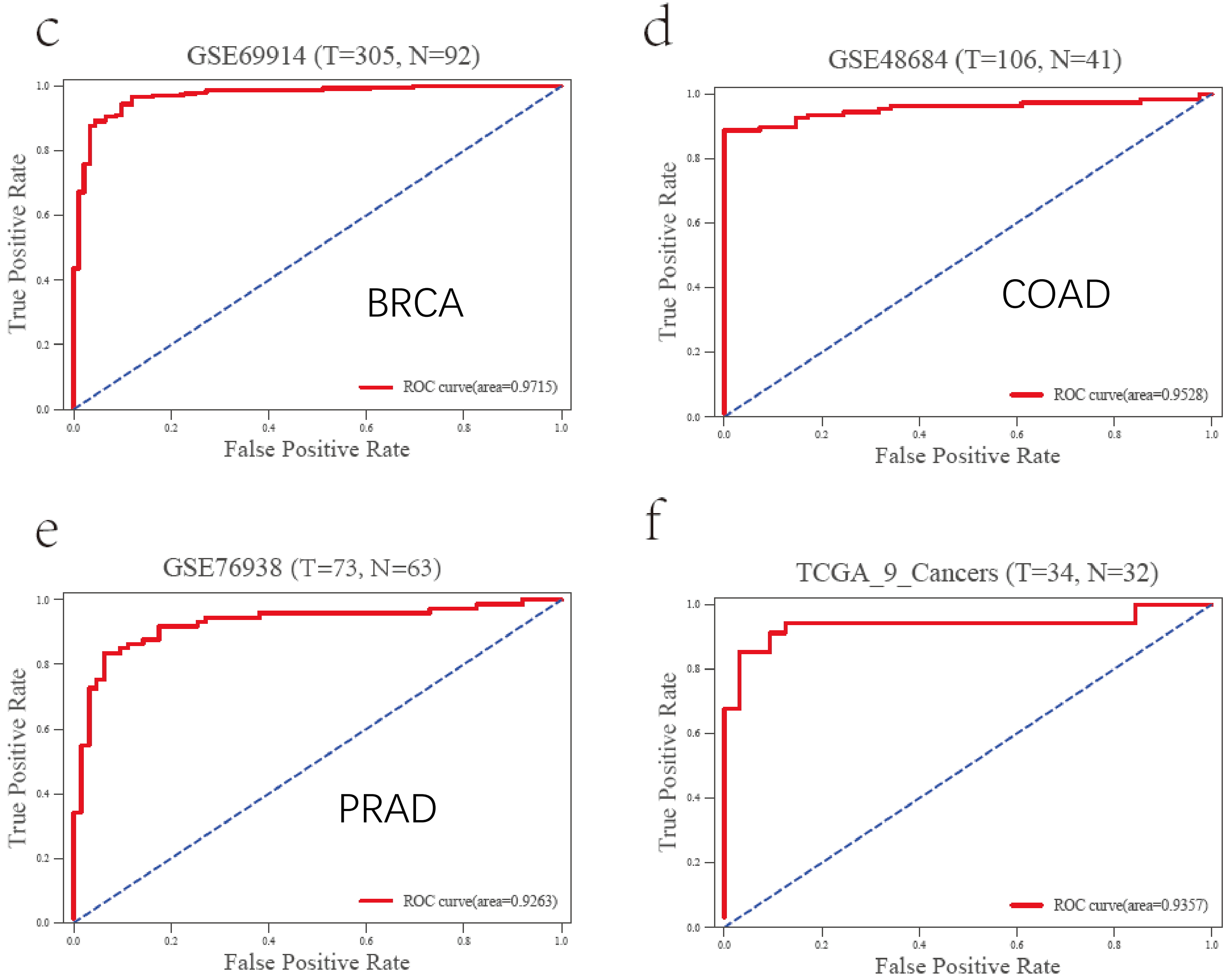
The hypermethylated CpG sites tends to negatively correlate with the expression of their corresponding genes in almost all different tumor types. The methylation level of most hypomethylated CpG sites are positively correlated with the expression of their corresponding genes in cancers, but some of them are negatively correlated with the expression of their relevant genes in other cancers

Seven Probes Classifier has Good Performance in Distinguishing Tumor and Normal



a, ROC curve of training (12 cancer types and 1216 samples) showing the high sensitivity and specificity in predicting different cancer types from corresponding normal tissues. **b** Learning curve of 5-fold cross validation in training (ROC = 0.979).

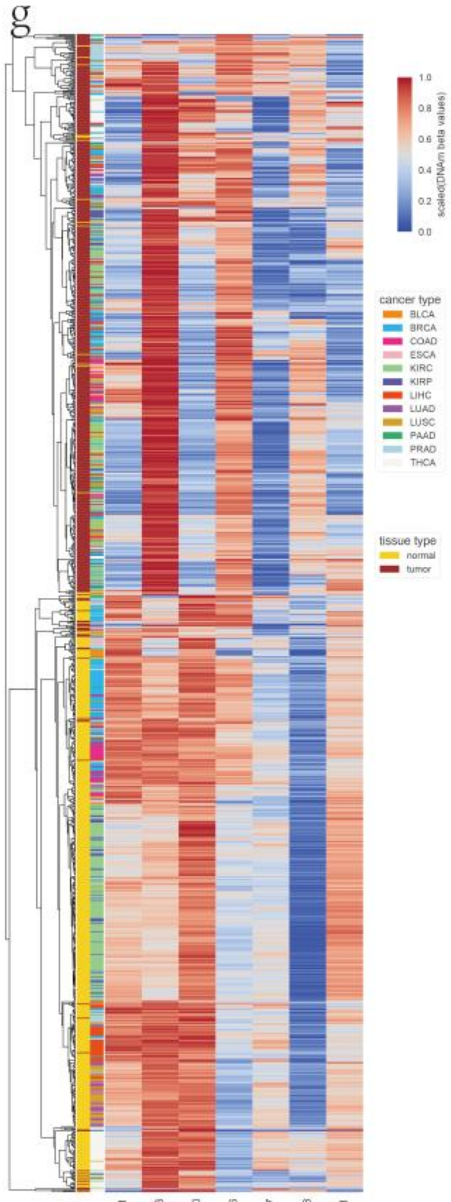
Seven Probes Classifier has Good Performance in Distinguishing Tumor and Normal



c, d, e ROC curve for the validation data set of GSE69914(Breast Cancer), GSE48684(Colorectal Cancer) and GSE76938 (Prostate Cancer).

f ROC curve for the independent validation data set of the remaining 9 TCGA cancers not included in the training set.
T and N indicate numbers of tumor and normal samples

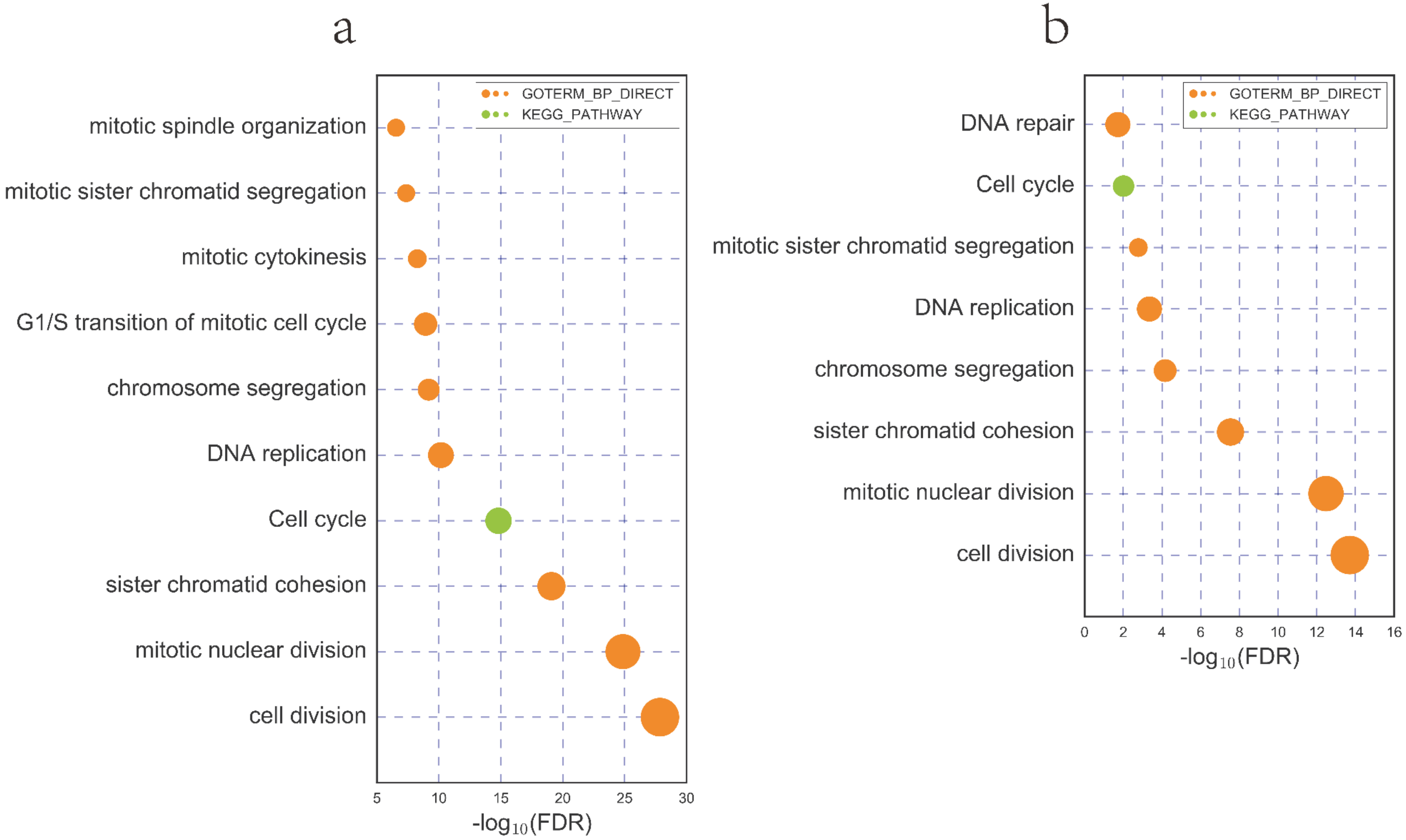
Seven Probes Classify Those Samples of 12 Cancer Types into Two Distinguished Groups



Unsupervised hierarchical clustering and heatmap associated with the methylation profile of the seven probes across all **1216 samples of 12 cancers**. Those samples were classified into two distinct classes by the 7 CpG sites

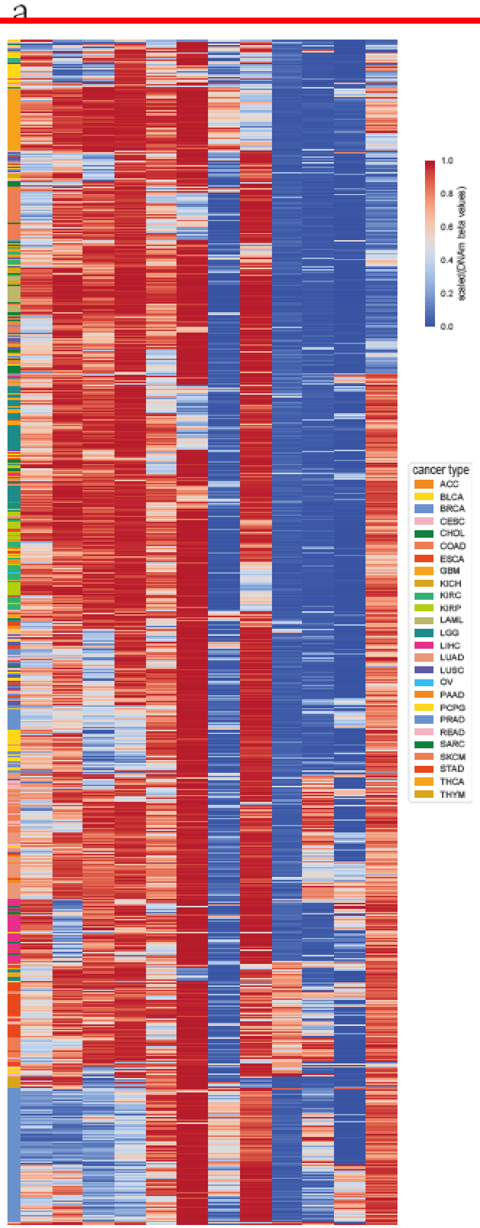
Right color bars mark the tissue type and cancer type.

GO Biological Process and KEGG Pathway Enrichment Analysis Results



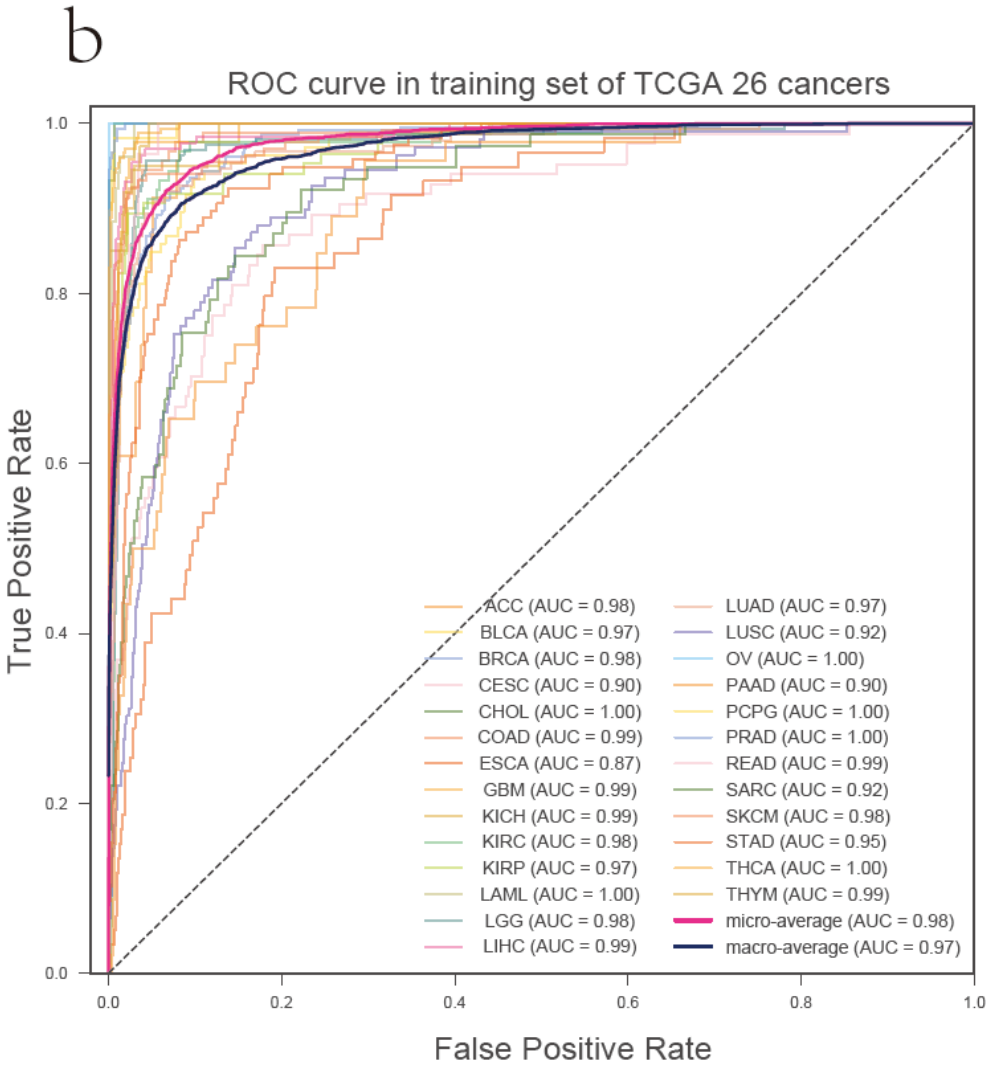
Enrichment analyses for those genes that their expression levels were significantly correlated with 4 probes, highly associated with tumor biogenesis.

Tumor-Specific Classifier with 12 Probes Effectively Distinguishes Different Cancers



Unsupervised hierarchical clustering and heatmap showing the methylation profile of the selected **12 probes across 7605 tumor samples of 26 cancer types** reals that those 12 probes complement with each other to distinguish different cancer types.

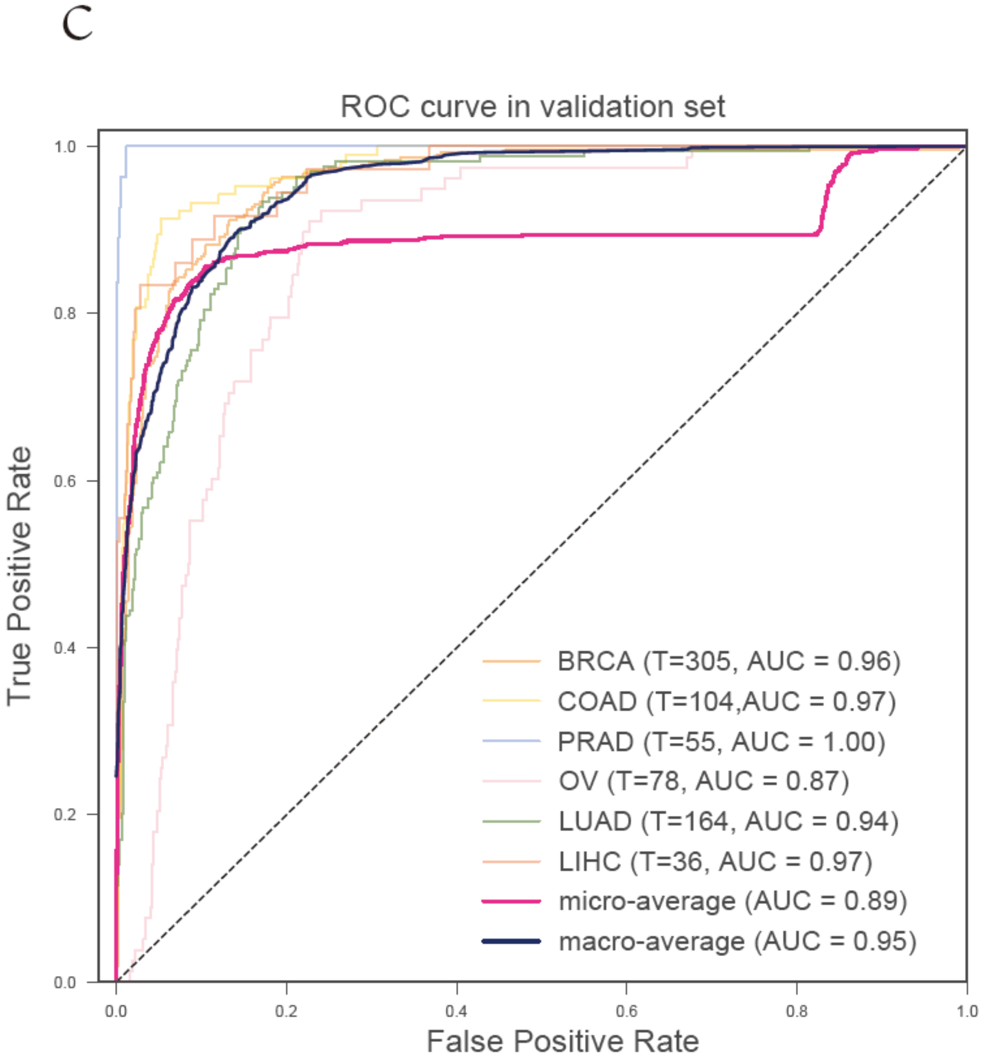
Tumor-Specific Classifier Effectively Distinguishes Different Cancers



The micro-average **AUC** of OneVsRestClassifier based on those 12 selected CpG sites reaches to **0.98** for those 26 different cancers.

The results indicate the high sensitivity and specificity of our multiclass tumor specific classifier in predicting different cancers.

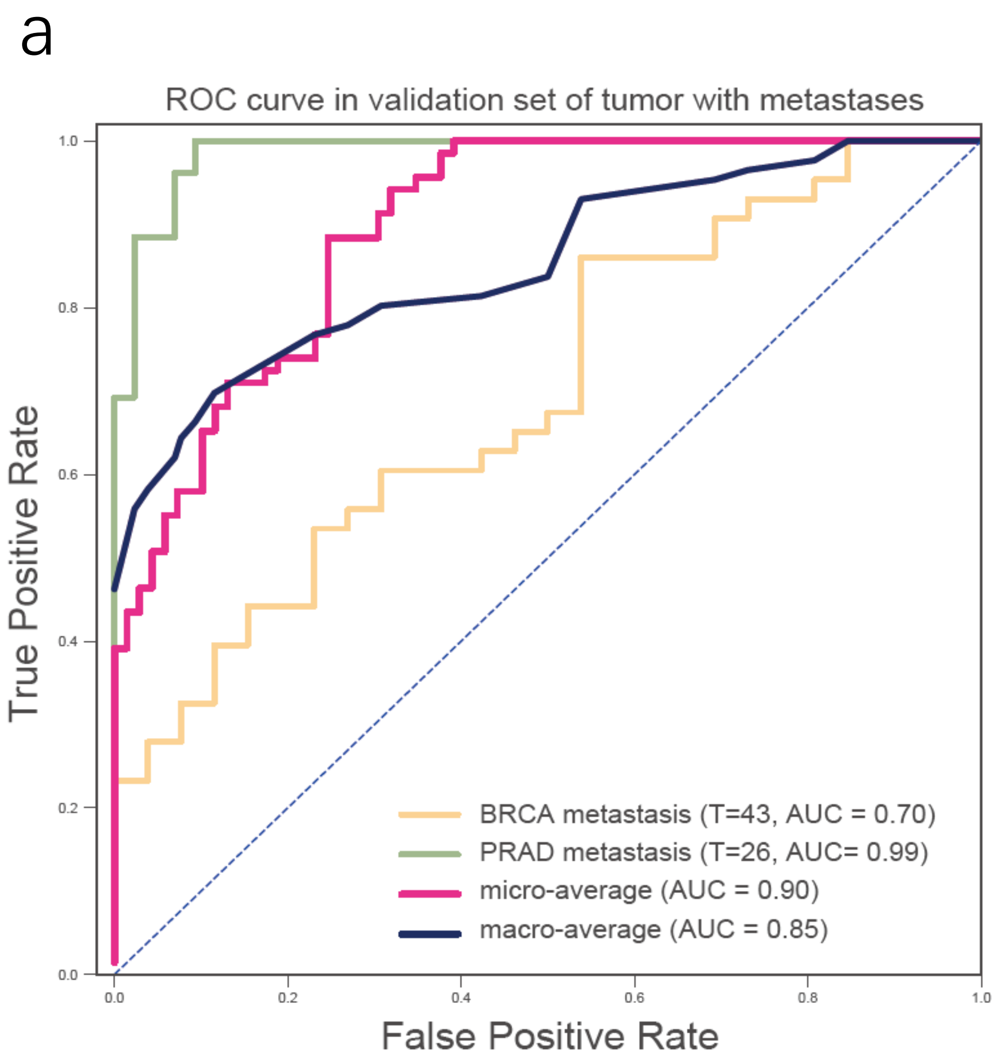
Tumor-Specific Classifier Effectively Distinguishes Different Cancers



Validation based on GEO datasets of breast cancer (BRCA), colorectal cancer (COAD), prostate cancer (PRAD), ovarian cancer (OV), lung adenocarcinoma (LUAD) and hepatocellular carcinomas (LIHC). **micro-average AUC > 0.89**

ROC curve in six independent validation dataset of different cancers. T indicates the numbers of tumor samples used in each dataset.

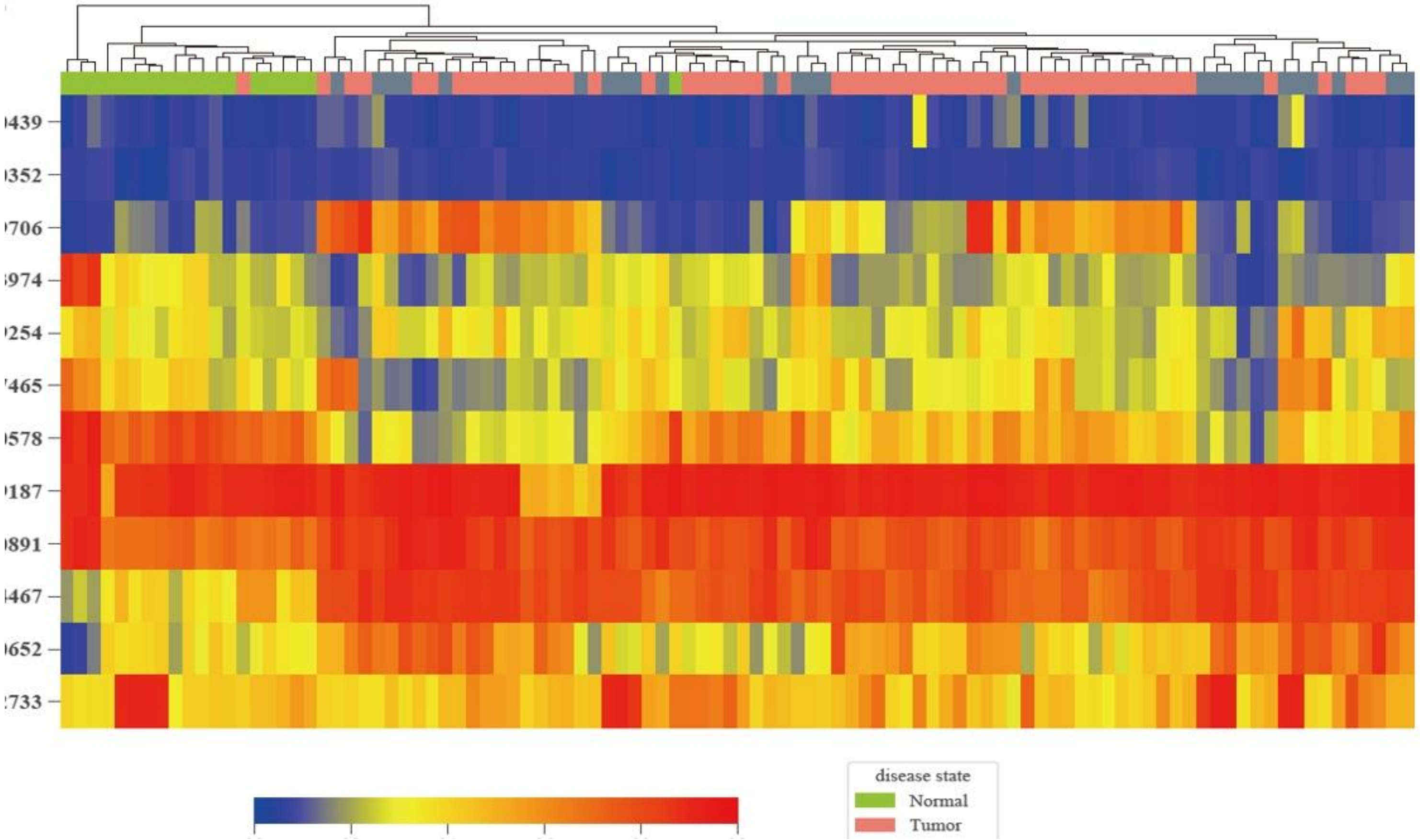
Tumor-Specific Classifier Effectively Predict the Origin of Tumors with Metastases



GSE58999: Breast cancer with metastases to lymph node.
GSE73549 and GSE38240: Prostate cancer with metastases to bone or lymph node (26 metastatic samples in total).

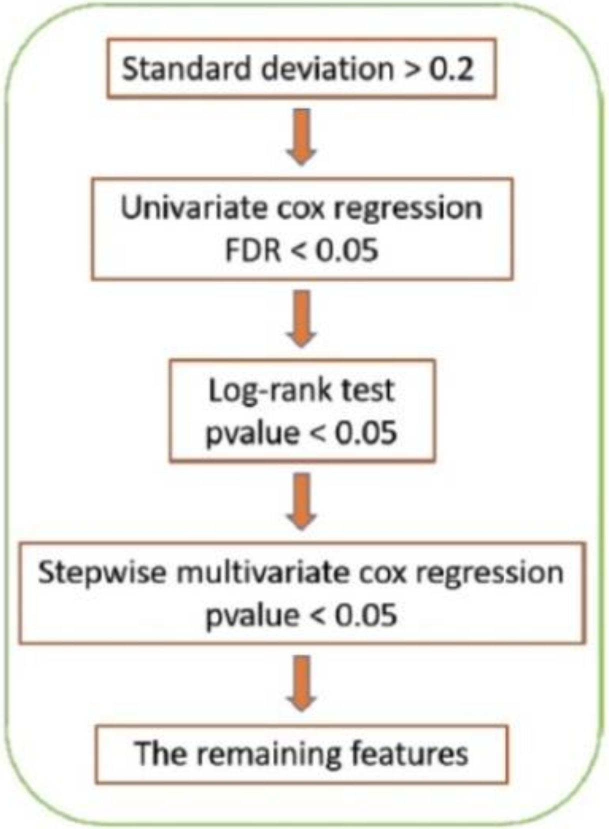
ROC curve of multiclass tumor specific classifier in metastatic breast cancer and metastatic prostate cancer.
T indicates the numbers of tumor samples with metastasis used in each dataset.

These 12 Probes can Effectively Distinguish Metastatic Tumors From Normal Tissues



Unsupervised clustering of the DNA methylation levels of the 12 CpGs between normal prostate samples, prostate tumor and prostate tumor with metastases from GSE73549 and GSE38240.
Annotations of the column of the heatmap indicate disease states of patients.

Construction of the Probe-based Prognostic Classifier



Standard deviation > 0.2

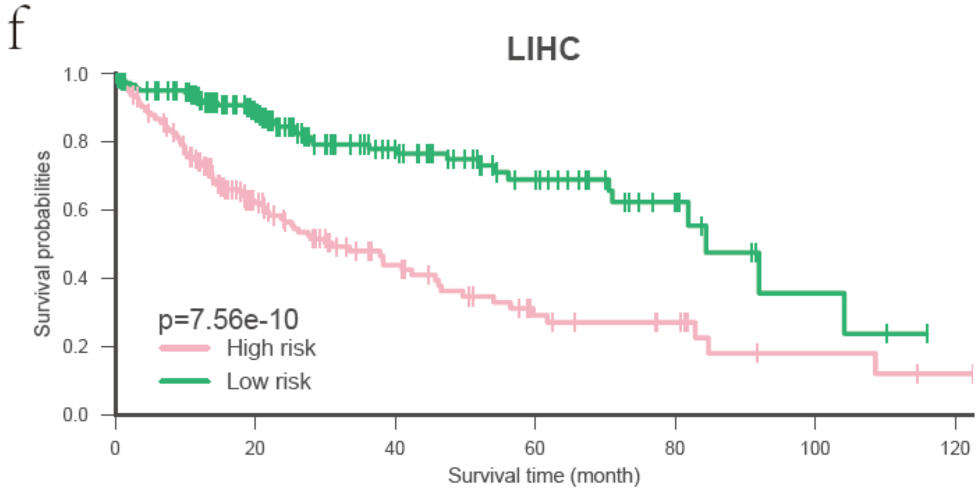
Univariate cox regression
FDR < 0.05

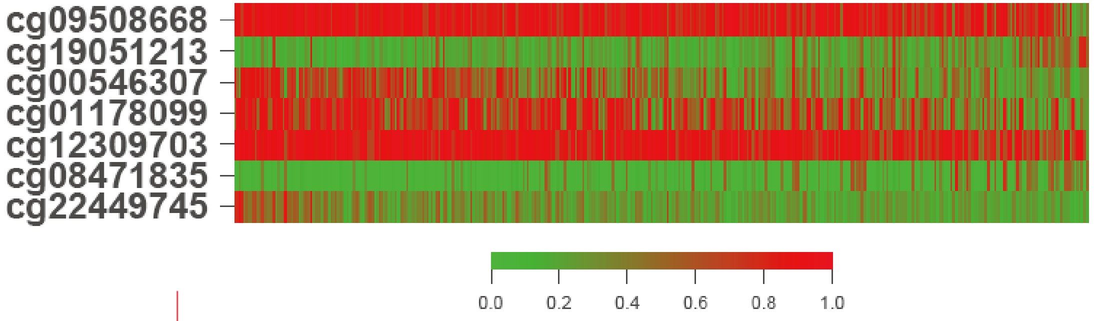
Log-rank test
pvalue < 0.05

Stepwise multivariate cox regression
pvalue < 0.05

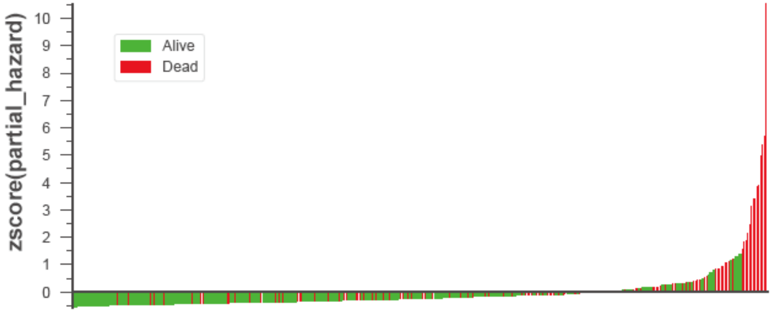
The remaining features

on 28 cancer types



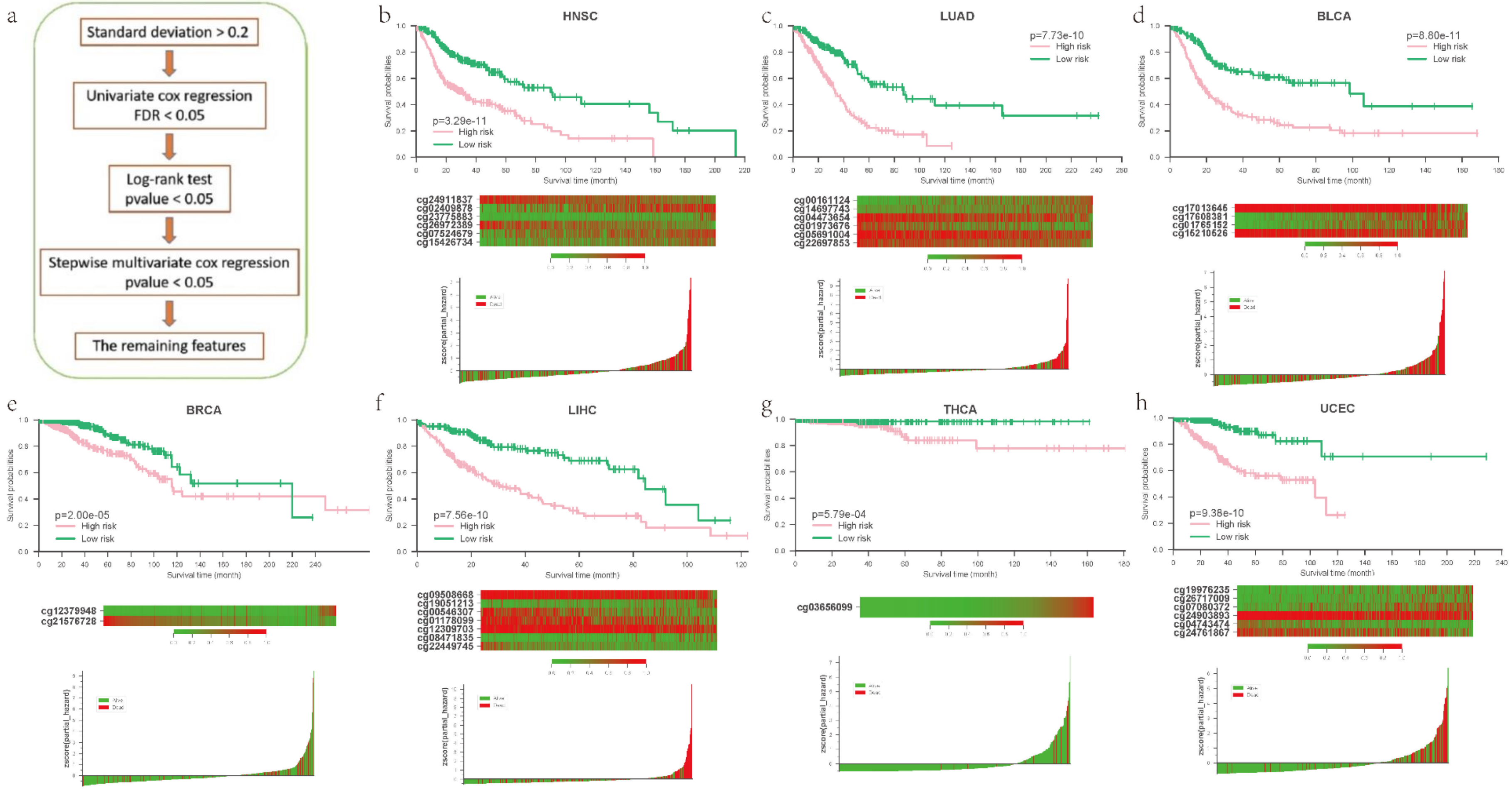


Methylation profile of the probes used by the prognostic classifier in patients



Z-score distribution of the prognostic classifier and patient survival status

The Performance of Probe-based Prognostic Classifier



Construction of the probe-based prognostic classifier.

a The pipeline of feature selection for prognostic classifier. **b-h** The result of Prognostic classifier for 7 cancer types

Upper panel: Kaplan-Meier survival analysis for the patients in each of the 7 cancers

Middle panel: heatmap showing methylation of the CpGs used by the prognostic classifier in patients.

. Lower panel: Z-score distribution of the prognostic classifier and patient survival status.

The patients were divided into low-risk and high-risk groups using the median value of the partial hazard as cutoff. p-value were calculated by the log-rank test.

Conclusions

1. Methylation profiles of different cancers vary tremendously, which can be used to distinguish the cancer from normal as well as different cancer types.
2. Results with independent validation datasets of various cancers demonstrate the performance and robustness of our DNAm site-based tumor-normal classifier.
3. Tumor-specific classifier can effectively distinguish different cancers. The DNAm markers for tumor-specific classifier can also be used to predict the origin of tumor with metastases or with unknown primary origin.
4. Prognostic classifier with DNAm pattern successfully divides patients into high-risk or low-risk group in different cancer types.
5. We identified potential diagnostic and prognostic biomarkers based on the methylation changes of DNAm patterns in diverse TCGA cancers, which have the potential application in clinical practices.

Thank You for Your Attention

Questions?