Article

Accurate prediction of pan-cancer types using machine learning with minimal number of DNA methylation sites

Wei Ning[1,2,3], Tao Wu[1], Chenxu Wu[1], Shixiang Wang[1], Ziyu Tao[1], Guangshuai Wang[1], Xiangyu Zhao[1], Kaixuan Diao[1], Jinyu Wang[1], Jing Chen[1], Fuxiang Chen[4], and Xue-Song Liu[1,*]

[1] School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China

[2] Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China

[3] University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100049, China

[4] Department of Clinical Immunology, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200011, China

* Correspondence to: Xue-Song Liu, E-mail: liuxs@shanghaitech.edu.cn

Abstract

DNA methylation analysis has been applied to determine the primary site of cancer; however, robust and accurate prediction of cancer types with minimum number of sites is still a significant scientific challenge. To build an accurate and robust cancer type prediction tool with minimum number of DNA methylation sites, we internally benchmarked different DNA methylation site selection and ranking procedures, as well as different classification models. We use The Cancer Genome Atlas (TCGA) dataset (26 cancer types with 8296 samples) to train and test model and use the independent dataset (17 cancer types with 2738 samples) for model validation. A deep neural network (DNN) model using a combined feature selection procedure (named as MethyDeep) can predict 26 cancer types using 30 methylation sites with superior performance compared with known methods for both primary and metastatic cancer in independent validation datasets. In conclusion, MethyDeep is an accurate and robust cancer type predictor with the minimum number of DNA methylation sites; it could potentially help the cost-effective clarification of cancer of unknown primary (CUP) patients and also the liquid biopsy early screening of cancers.

Keywords: DNA methylation, MethyDeep, cancer type prediction, deep neural network (DNN), machine learning

Introduction

The diagnosis of cancer type is necessary for proper treatment planning and management of the disease. Accurate and timely diagnosis can greatly improve patient outcomes, reduce morbidity and mortality, and increase survival rates (T. Chen et al., 2018). However, there are several current problems with cancer-type diagnosis: limited availability of diagnostic tools, misdiagnosis, late diagnosis, etc. Therefore, there is an urgent need to develop a tool that can diagnose a variety of cancers and can give an accurate and rapid diagnosis.

In the human genome, DNA methylation is a selective epigenetic modification of cytosine residues (Fernandez et al., 2012). Selected methylation of cancer-related genes has been used as biomarkers for the early diagnosis and prognosis of cancer (Koch et al., 2018; Locke et al., 2019; M. Li et al., 2020; Hou et al., 2021). For example, methylation of *APC* gene promoter is a biomarker for early diagnosis of prostate cancer (Richiardi et al., 2009), and methylation of *Mgmt* gene promoter can be a prognosis biomarker for triple-negative breast cancer (Yu et al., 2018). Cells of different tissue origins usually have distinct DNA methylation patterns, and DNA methylation has been applied in cancer type prediction (B. Zhang et al., 2013). These known methods need thousands of methylation sites for pan-cancer type prediction (Moran et al., 2016b; Vrba and Futscher, 2018; Zheng and Xu, 2020) or only focus on one or few cancer types (Hlady et al., 2019; Chen et al., 2022). Robust and accurate pan-cancer type prediction based on the least DNA methylation sites is still a significant challenge.

In this study, we systematically internally benchmarked the combination of different DNA methylation site selection and classification models. MethyDeep trained with overlapping differential methylation sites (DMSs) shows the best performance in both held-out dataset and independent validation dataset. Thus, MethyDeep represents the currently available tool that can perform accurate and robust pan-cancer type prediction using as few as 30 DNA methylation sites.

Results

DNA methylation dataset for cancer type prediction model training

We obtained DNA methylation microarray data from TCGA (Figure 1), and most of the

methylation data are generated based on the 450K platform (Figure 2A). In addition, most of the TCGA samples are derived from primary cancer (Figure 2B). Therefore, we chose the methylation data of primary cancer samples from the 450K platform to build the model. Methylation sites that miss values in all TCGA samples are removed from subsequent analysis (Figure 2D). Cancer types that have <50 samples are also removed from further analysis (Figure 2C). In total 26 cancer types with sufficient samples are available for this study (Supplementary Table S1). TCGA samples are divided into 8:2 ratio for model training dataset and held-out dataset, respectively (Figure 1).

Differential analysis of methylation data among paired cancer types

To find effective methylation sites in cancer type classification, we first used ChAMP (Tian et al., 2017) to perform differential analysis of methylation sites among paired cancers. The methylation sites with adjust *P*-values <0.01 are defined as DMSs. For most paired cancers, ChAMP analysis leads to ~200000 to 300000 significant DMSs (Figure 3A).

The DMSs from ChAMP analysis are ranked by |log(fold change)|, and the top 500 DMSs of all cancer type pairs are selected for downstream analysis (Supplementary Table S5). Through principal component analysis (PCA), we can see that these methylation sites can effectively separate different cancer types (Supplementary Figure S1B and D). Results of cancer-type DMS overlapping analysis indicate that most DMSs only exist in <5 pairs of cancer types (Figure 3B).

Effect of DMS selection procedures

To build a robust pan-cancer type predictor with a minimum number of DNA methylation sites, the selection of optimum DNA methylation site for the model is a critical process. Here, the DNA methylation site selection procedure is divided into three consecutive steps (Figure 1). The first step is to select the best overlapping frequency of methylation sites; the second step is to select the best feature sorting method; and the third step is to select the optimal number of methylation sites. We performed an internal benchmark analysis of these three steps to identify the optimum DNA methylation site selection procedure.

First, we compared the model performance of eXtreme Gradient Boosting (XGBoost) (T. Chen and Guestrin, 2016), Extremely randomized trees (Extratree) (Geurts et al., 2006), Random Forest (RF) (Breiman, 2001), K Nearest Neighbors (KNN), Naive Baye (NB), and Decision Tree (DT) under different overlapping DMS frequencies

(Supplementary Figure S1A and Table S6). XGBoost, RF, and Extratree show superior performance compared with other methods (Supplementary Figure S2). Next, we compared the performance of XGBoost, RF, and Extratree under different numbers of methylation sites and different frequencies of DMS. All model evaluation indexes of XGBoost and RF have reached 0.9 when they only use 30 methylation sites (Figure 4A and B; Supplementary Figures S3 and S4). Only when Extratree uses 45 methylation sites, the model evaluation index reaches 0.9 (Supplementary Figure S5). This analysis suggests that XGBoost and RF could be the optimum machine learning algorithm for DMS selection.

To more intuitively evaluate the impact of DMS frequency on the performance of different models, we used the area under the curve when the model index reaches 0.9 to evaluate the performance of models at different DMS frequencies (see Materials and methods for details). XGBoost using methylation sites with DMS frequency of 35 shows advantages over XGBoost using other methylation sites with different DMS frequencies (Figure 4C; Supplementary Figure S6). RF using methylation sites with DMS frequency of 45 shows advantages over RF using other methylation sites with different DMS frequencies (Figure 4D; Supplementary Figure S7). This analysis suggests that the optimal conditions are the same for different machine learning methods.

Comparisons between deep learning model and traditional machine learning model

Furthermore, we explored whether the emerging deep learning neural network model is better than the traditional machine learning model in cancer type classification. We built deep learning models based on the DMSs selected by XGBoost (DNN_XGBoost) and deep learning models based on the DMSs selected by RF (DNN_RF) (Figure 1). DNN_XGBoost shows similar performance to XGBoost in all indicators (Figure 5A; Supplementary Figure S8). DNN_RF shows better performance than RF in all indicators (Figure 5B; Supplementary Figure S9). Similarly, we looked at the area under the curve of different models. DNN_XGBoost also shows the best performance when the frequency of DMS is 35 (Figure 5C; Supplementary Figure S10). DNN_RF shows the best performance when the frequency of DMS is 45 (Figure 5D; Supplementary Figure S11). Next, we compared the area under the curve of the four candidate prediction models. DNN_RF shows superior performance compared with other three methods (Figure 5E). We also calculated the change rate of the performance of DNN_RF with the decreasing number of DMSs. When the number of DMSs in the model exceeds 25, the performance of the model only fluctuates around 0.001 (Figure 5F). Therefore, after comprehensively considering the

performance of the model and the number of methylation sites used, we chose DNN_RF with 30 methylation sites as the final model and named the prediction model as “MethyDeep”.

The performance of MethyDeep on additional primary and metastatic cancer datasets

In order to further evaluate the performance of MethyDeep, we verified it on additional independent datasets. On the additional primary cancer dataset, the recall and precision of glioma (TCGA-GBM), prostate adenocarcinoma (TCGA-PRAD), liver hepatocellular carcinoma (TCGA-LIHC), bladder urothelial carcinoma (TCGA-BLCA), acute myeloid leukemia (TCGA-LAML), and pancreatic adenocarcinoma (TCGA-PAAD) both exceed 0.9 (Figure 6A). For breast invasive carcinoma (TCGA-BRCA), although the recall is only 0.71, the precision exceeds 0.9, probably because the training data does not include all breast cancer subtypes.

On the additional metastatic cancer dataset, the recall of pheochromocytoma and paraganglioma (TCGA-PCPG), TCGA-BLCA, thyroid carcinoma (TCGA-THCA), sarcoma (TCGA-SARC), head and neck squamous cell carcinoma (TCGA-HNSC), TCGA-PAAD, and TCGA-PRAD exceeds 0.95 (Figure 6B). For skin cutaneous melanoma (TCGA-SKCM), although recall is only 0.74, the precision exceeds 0.9. This may also be due to the high heterogeneity of melanoma (Rambow et al., 2019).

Comparison between MethyDeep and existing methods

In order to further evaluate the performance of MethyDeep in identifying unknown primary cancer types, we compared MethyDeep with other published models. First, there is almost no overlap between the 30 methylation sites used by MethyDeep and those used by other methods (Supplementary Figure S12), showing that the methylation sites used by MethyDeep are unique. On the additional primary and metastatic cancer datasets, MethyDeep shows superior performance compared with existing methods (Figure 7A–D; Supplementary Figures S12B–E and S13). Although Matthews and Kappa of MethyDeep are relatively poor, they still have advantages over existing methods. We also used receiver operating characteristic (ROC) to compare the models, and MethyDeep shows the best performance on the primary cancer dataset compared with other models (Figure 7E). The performance of MethyDeep on the metastatic cancer dataset is similar to that of Xia (283), which applied 283 DNA methylation sites for cancer type prediction, and significantly better than remaining methods (Figure 7F). We also checked several other

indicators, including predictive value (PPV), negative predictive value (NPV), sensitivity (SEN), and specificity (SPE). MethyDeep still shows improved performance compared with those known methods (Supplementary Tables S9 and S10).

Potential biology of the selected methylation sites

To explore the potential biology of the 30 methylation sites selected in MethyDeep, we first checked the frequency distribution of these sites. The frequency of most methylation sites is <60, and the most frequent site is cg01979888, which is located on the gene *IFFO1* (Supplementary Figure S14 and Table S11). Previous studies have reported that the nuclear skeleton protein IFFO1 fixes the broken DNA and inhibits chromosome translocation during tumorigenesis (W. Li et al., 2019a). IFFO1 is a tumor suppressor, which can inhibit tumor metastasis and reverse the drug resistance of ovarian cancer (Y. Zhang et al., 2023). We also looked at the methylation site cg00582524, which ranks first in the importance of characteristics, and it is located on the gene *SALL1*. SALL1 has been proven to be a biomarker for many cancer types (Ma et al., 2018; Misawa et al., 2018; Salman et al., 2018; Z. Li et al., 2019b). Then, we carried out gene set enrichment analysis on the genes in which these 30 methylation sites were distributed (Supplementary Table S12). The enriched pathway is insulin signaling pathway, which is known to play a key role in aging and cancer (Anisimov and Bartke, 2013). Insulin PI3K signal transduction is also a driving factor in the evolution and progression of cancer (Hopkins et al., 2020). These additional analyses provide some biological insights for the selected methylation sites.

Discussion

Accurate and robust prediction of multiple cancer types with a cost-effective method is still a significant application and scientific challenge. The traditional method for cancer type determination includes immunohistochemistry or gene expression, and these methods cannot achieve accurate pan-cancer type prediction (Horlings et al., 2008; Alsarraj and Hunter, 2013). DNA mutation patterns derived from whole genome sequencing (WGS) have been applied in pan-cancer type prediction with 83% accuracy in validation dataset (Jiao et al., 2020). However, the high cost of high-coverage WGS would prohibit the clinical application of this method. DNA methylation has also been applied in cancer type prediction; however, these known works need thousands of DNA methylation sites for accurate pan-cancer type prediction, or can only predict one or few types of cancer (Vrba and Futscher, 2018; Hlady et al., 2019; Zheng and Xu, 2020). The MethyDeep tool constructed here represents the first accurate and robust pan-cancer type predictor with

only 30 DNA methylation sites.

Since MethyDeep shows robust cancer type prediction in metastatic cancer, it could enable cost-effective and accurate cancer typing of cancer of unknown primary (CUP) patients. Prospective clinical trials could be initiated to test the clinical applicability of MethyDeep. Early detection of cancer with a cost-effective method is an urgent clinical need. Most currently liquid biopsy methods can detect one type of cancer, such as prostate cancer (J. Wang et al., 2020), colon cancer (Jin et al., 2021), liver cancer (Ye et al., 2019), breast cancer, and head and neck cancer (Yan Zhang et al., 2020), but accurate pan-cancer type prediction is still not achieved. MethyDeep developed in this study could guide the design of robust pan-cancer early detection methods using liquid biopsy, and these need to be tested using liquid biopsy samples.

Although our cancer type prediction model has achieved good results in both held-out and independent validation datasets, there are still limitations. MethyDeep can predict 26 cancer types, but some cancer types with <50 samples have not been included in the prediction model training. In the future, with the availability of additional data, MethyDeep can be improved to include these additional cancer types. The training DNA methylation data are derived from SNP arrays, which contain 450K sites, but some potentially important cancer classification DNA methylation sites may not be included in these SNP arrays (Teh et al., 2016). To fully capture the DMSs for cancer type prediction, whole genome bisulfite sequencing-derived data should be employed; however, this type of DNA methylation data is still very limited (Ziller et al., 2015). The prediction of pan-cancer types with a minimum number of DNA methylation sites will facilitate the cost-effective detection of cancer using qPCR-based method, and thus the final selection of DNA methylation sites need to consider the design of appropriate qPCR primers for detection. In together, the construction of MethyDeep represents an important attempt at DNA methylation site selection and optimization, and it will facilitate the cost-effective pan-cancer type prediction in clinics, including the diagnosis for CUP patients and the early screening for cancer with liquid biopsy.

Materials and methods

Data preprocessing

TCGA Illumina Human Methylation 450K DNA methylation data were obtained from UCSC Xena (https://xena.ucsc.edu/) (Goldman et al., 2020; S. Wang et al., 2021; J. Li et al., 2022). Based on the similarity of cancer types and primary original sites, we combined pleomorphic glioma (TCGA-GBM) and low-grade glioma (TCGA-

GBM), combined rectal cancer (TCGA-READ) and colorectal cancer (TCGA-COAD) into colorectal cancer (TCGA-COAD), excluded cancer types with <50 samples (cholangiocarcinoma (TCGA-CHOL), ovarian serous cystadenocarcinoma (TCGA-OV), stomach adenocarcinoma (TCGA-COAD), and esophageal carcinoma (TCGA-ESCA)), and finally retained 26 cancer types with 8296 cancer samples to build a predictor of cancer type (Figure 1; Supplementary Table S1).

To evaluate the performance of cancer type classifier trained on the TCGA dataset, we obtained additional DNA methylation datasets from the comprehensive gene expression database (GEO) (Barrett et al., 2012), i.e. Illumina 450k array (Weisenberger et al., 2008; Bibikova et al., 2011) and EPIC array (Moran et al., 2016a). The collection criteria are that it must be one of the 26 cancer types in the training data and must meet the model's need for methylation sites. For data whose number of missing methylation sites is <10% of the number of methylation sites required by the model, missing values are replaced with the mean value of other existing methylation sites; otherwise, the data are removed. Finally, the data were divided into metastatic cancer dataset (8 cancer types with 605 samples; Supplementary Table S2) and primary cancer dataset (13 cancer types with 2133 samples; Supplementary Table S3).

Differential analysis of DNA methylation sites

The Bioconductor (https://www.bioconductor.org/) DNA methylation microarray analysis pipeline ChAMP was used to analyze raw DNA methylation data (Tian et al., 2017). The methylation data collected from TCGA project is β value, which reflects the level of DNA methylation. β values range from 0 to 1, where 0 means no methylated molecules detected at these CpG sites and 1 means all molecules been completely methylated. For the normalization of methylation probes, we chose the peak-based correction method (Dedeurwaerder et al., 2011). The differentially methylated probe function uses a linear model to calculate the *P*-value of differential methylation. *P*-values were adjusted with the Benjamini–Hochberg correction method (Benjamini and Hochberg, 1995). DMSs between cancer type pairs were defined as adjusted *P*-value <0.01 (Supplementary Table S4). Other parameters were the default values. To further narrow feature searching space, we selected the union of top 500 DMSs by |log(fold change)| for all cancer type pairs for downstream analysis (Supplementary Table S5).

Cluster analysis

We used "FactoMineR" (Lê et al., 2008) and "factoextra" packages for PCA clustering

and plotting of methylation microarray data. In PCA clustering, the parameter is graph = FALSE. In drawing, the parameters are geom.ind = "point", col.ind = group_list, addEllipses = TRUE, and legend.title = "Groups", where group_list is the sample information used for clustering.

Training data preprocessing

We first divided the processed TCGA methylation data into a training dataset and a held-out dataset using stratified sampling at a ratio of 8:2. For the training set, to solve the problem of sample imbalance, we oversampled cancer types with fewer samples to ensure that the number of samples for each cancer is consistent. The number of samples used for training reaches 16094 (Supplementary Table S7). For hyperparameter adjustment, we additionally divided the training set into a hyperparameter training set and a hyperparameter test set according to the ratio of 8:2, also by applying stratified sampling.

Comparison of machine learning methods

Next, we followed the frequency of DMS, according to the frequency ≥1, ≥2, ≥5, ..., ≥50, to divide the data into 12 datasets (Supplementary Table S6) and compared the effectiveness of XGBoost, Extratree, RF, KNN, NB, and DT on each dataset. The model parameters of XGBoost are max_depth = 5, learning_rate = 0.1, n_estimators = 160, silent = True, and objective = 'multi: softmax'. The model parameters of Extratree and RF are n_estimators = 10, max_depth = None, min_samples_split = 2, and random_state = 0. The model parameters of DT are max_depth = None, min_samples_split = 2, and random_state = 0. The model parameters of KNN and NB are default. To measure the performance of the classifiers, the following conventional performance indicators have been defined: recall, precision, F1 score, accuracy, Matthews, and Kappa. These scores are calculated using TP (true positive), TN (true negative), FP (false positive), and FN (false negative) values as below.

Recall is the proportion of samples of a particular cancer type that are correctly assigned to that type:

$$Recall = \frac{TP}{TP+FN} \#(1)$$

.

Precision is the proportion of samples assigned to a particular type that are truly that type:

$$Precision = \frac{TP}{TP+F} \#(2)$$

.

F1 Score is the harmonic mean of recall and precision:

$$F1 = \frac{2(Recall \times Precision)}{Recall + Precisi} \#(3)$$

.

Accuracy is the proportion of correct assignments. We used this metric only when summarizing the performance of the classifier across all 26 tumor types:

$$Accuracy = \frac{TP+T}{\text{TP+TN+FN+FP}} = \frac{\text{correct assignments}}{\text{total samples}} \#(4)$$

.

Matthews correlation coefficient (MCC) is the geometric mean of the regression coefficient of the problem and its dual:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+F\ \ )(TP+FN)(TN+FP)(TN+FN)}} \#(5)$$

.

Kappa coefficient is a method for evaluating consistency in statistics:

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \#(6)$$

, where P_o is the empirical probability (observed consistency ratio) of labels assigned to any sample, and P_e is the expected consistency when two labels randomly assign labels. P_e is estimated using the prior experience of each annotator on the class label.

Determination of the best classification model

We used XGBoost, Extratree, and RF to rank the feature importance of DNA methylation sites, and then the top 5, top 10, top 15, ..., and top 50 methylation sites were used to construct different models. We used the above model evaluation indicators to evaluate the performance of the models. We also used the area under the curve, when various indicators of the model first reach 0.9, to evaluate the performance of the model under a few methylation sites. The better the performance of the model, the smaller the area under the curve. The area under the curve is calculated as follows:

$$\frac{(x_2 + x_i + x_j - 2 \times x_1) \times 5}{2} \#(7)$$

, where x is the score of the model evaluation index, x_1 is the score of the model evaluation index when the number of methylation sites is 5, x_2 is the score of the model evaluation index when the number of methylation sites is 10, and so on. The value of i is greater than 3 and less than or equal to the number of methylation sites divided by 5, when the model evaluation index reaches 0.9 for the first time. The value of j is equal to i−1.

We also built different neural network models under the above conditions. The optimal hyperparameter is selected by Gridsearch (Pedregosa et al., 2011). Adam optimization (Kingma and Ba, 2014) is used to build DNN model. We used the same indicators to compare machine learning models and neural network models, and finally determined the best prediction model and the best methylation site selection procedure.

The structure and filter conditions of MethyDeep

We determined several screening conditions: (i) DMS frequency ≥45; (ii) RF used to rank the importance of DMS; and (iii) 30 methylation sites used.

MethyDeep is a deep learning network model with full connection architecture. It consists of one input layer, six hidden layers, and one output layer. The number of neurons in the input layer is 30, and the activation function is "Relu". The number of neurons in the hidden layer is (230,192,154,116, 78, 40), and the activation function is also "ReLU". The number of neurons in the output layer is 26, and the activation function is "softmax". Other parameters are batch_size=100, epochs=100，loss='categorical_crossentropy', optimizer='adam', and metrics=['accuracy'].

Hyperparameter optimization by Gridsearch

The number of neurons gradually decreases from the first hidden layer to the last hidden layer, and the reduced step formula is:

$$\frac{first\_layer\_nodes - last\_layer\_nodes}{n\_layers - 1} \#(8)$$

, where first_layer_nodes is the number of nodes in the first hidden layer, last_ layer_nodes is the number of nodes in the last hidden layer, and n_layers is the number of hidden layers.

The dropout rate of each hidden layer is also different. Similarly, given the dropout rate of the first layer, the dropout rate of the subsequent hidden layer continues to decrease until it is 0.5%. The reduced step size is:

$$Round\left(\frac{dropout}{n\_layer - 1}, 2\right) \#(9)$$

, where dropout is the proportion of randomly deleted neurons in the first hidden layer. In the follow-up training, we also made some manual adjustments. When the dropout of the first hidden layer is 0, it means that we do not randomly delete the neurons of each layer.

Comparison between different cancer type prediction models

We implemented model architectures based on different methylation sites and

numbers according to the methods described in other literatures and compared the performance of the models on independent datasets. We named the models according to the number of methylation sites used for each model. The three models of Xia et al. (2020) (PMID: 32415265) are named Xia (283), Xia (28), and Xia (53). The model of Liu (2022) (PMCID: PMC8770539) is named P.Liu (6). The model of Liu et al. (2019) (PMID: 31590287) is named B.Liu (12). In addition to the above model evaluation indicators, we used ROC, PPV, NPV, SEN, and SPE for further comparison.

Enrichment analysis

For the 30 methylation sites selected by RF, we used g:profiler (Raudvere et al., 2019) to perform enrichment analysis on the selected genes and used the bubble chart to display the results.

Statistical analyses

Statistical analyses were performed with the statistical programming language R and Python.

Availability of data and material

All datasets are publicly available. The methylation data used for training are available at https:/xena.ucsc.edu/. In addition, the methylation data used for validation are available at https://www.ncbi.nlm.nih.gov/geo/. All descriptions and codes of our experiments are available at https://github.com/XSLiuLab/MethyDeep.

Supplementary material

Supplementary material is available at *Journal of Molecular Cell Biology* online.

Acknowledgements

We thank ShanghaiTech University High Performance Computing Public Service Platform for computing services. We thank multi-omics facility and molecular and cell biology core facility of ShanghaiTech University for technical help.

Funding

This work was supported by Shanghai Science and Technology Commission (21ZR1442400), the National Natural Science Foundation of China (31771373), and the startup funding from ShanghaiTech University.

Conflict of interest: none declared

Author contributions: W.N. collected the data and performed the computational analysis. T.W., C.W., S.W., X.Z., G.W., Z.T., K.D., J.W., J.C., and F.C. participated in critical project discussion. X.-S.L. designed, supervised the study, and wrote the manuscript.

References

Alsarraj, J., and Hunter, K.W. (2013). Chapter 67—Metastatic Cancer. In: Ginsburg, G.S. and Willard, H.F. (eds). Genomic and Personalized Medicine (Second Edition). Academic Press, 776-788.

Anisimov, V.N., and Bartke, A. (2013). The key role of growth hormone–insulin–IGF-1 signaling in aging and cancer. Critical reviews in oncology/hematology *87*, 201-223.

Barrett, T., Wilhite, S.E., Ledoux, P., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research *41*, D991-D995.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B *57*, 289-300.

Bibikova, M., Barnes, B., Tsan, C., et al. (2011). High density DNA methylation array with single CpG site resolution. Genomics *98*, 288-295.

Breiman, L. (2001). Random Forests. Machine Learning *45*, 5-32.

Chen, K., Zhang, F., Yu, X., et al. (2022). A molecular approach integrating genomic and DNA methylation profiling for tissue of origin identification in lung-specific cancer of unknown primary. Journal of translational medicine *20*, 158.

Chen, T., and Guestrin, C.J.A. (2016). XGBoost: A Scalable Tree Boosting System. arXiv, https://doi.org/10.48550/arXiv.1603.02754

Chen, T., Ren, L., Liu, X., et al. (2018). DNA Nanotechnology for Cancer Diagnosis and Therapy. International journal of molecular sciences *19*, 1671.

Dedeurwaerder, S., Defrance, M., Calonne, E., et al. (2011). Evaluation of the Infinium Methylation 450K technology. Epigenomics *3*, 771-784.

Fernandez, A.F., Assenov, Y., Martin-Subero, J.I., et al. (2012). A DNA methylation fingerprint of 1628 human samples. Genome research *22*, 407-419.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Machine Learning *63*, 3-42.

Goldman, M.J., Craft, B., Hastie, M., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. Nature Biotechnology *38*, 675-678.

Hlady, R.A., Zhao, X., Pan, X., et al. (2019). Genome-wide discovery and validation of diagnostic DNA methylation-based biomarkers for hepatocellular cancer detection in circulating cell free DNA. Theranostics *9*, 7239-7250.

Hopkins, B.D., Goncalves, M.D., and Cantley, L.C. (2020). Insulin–PI3K signalling: an evolutionarily insulated metabolic driver of cancer. Nature reviews. Endocrinology *16*, 276-283.

Horlings, H.M., Laar, R.K.v., Kerst, J.-M., et al. (2008). Gene Expression Profiling to Identify the Histogenetic Origin of Metastatic Adenocarcinomas of Unknown Primary. J. Clin. Oncol. *26*, 4435-4441.

Hou, P., Bao, S., Fan, D., et al. (2021). Machine learning-based integrative analysis of methylome and transcriptome identifies novel prognostic DNA methylation signature in uveal melanoma. Briefings in bioinformatics *22*, bbaa371.

Jiao, W., Atwal, G., Polak, P., et al. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nature communications *11*, 728.

Jin, S., Zhu, D., Shao, F., et al. (2021). Efficient detection and post-surgical monitoring of colon cancer with a multi-marker DNA methylation liquid biopsy. Proceedings of the National Academy of Sciences of the United States of America *118*, e2017421118.

Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv, https://doi.org/10.48550/arXiv.1412.6980

Koch, A., Joosten, S.C., Feng, Z., et al. (2018). Analysis of DNA methylation in cancer: location revisited. Nature reviews. Clinical oncology *15*, 459-466.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: AnRPackage for Multivariate Analysis. Journal of Statistical Software *25*, 1–18. https://doi.org/10.18637/jss.v025.i01

Li, J., Miao, B., Wang, S., et al. (2022). Hiplot: a comprehensive and easy-to-use web service for boosting publication-ready biomedical data visualization. Briefings in bioinformatics *23*, bbac261.

Li, M., Zhang, C., Zhou, L., et al. (2020). Identification and validation of novel DNA methylation markers for early diagnosis of lung adenocarcinoma. Molecular oncology *14*, 2744-2758.

Li, W., Bai, X., Li, J., et al. (2019a). The nucleoskeleton protein IFFO1 immobilizes broken DNA and suppresses chromosome translocation during tumorigenesis. Nature cell biology *21*, 1273-1285.

Li, Z., Zhao, S., Wang, H., et al. (2019b). miR-4286 promotes prostate cancer progression via targeting the expression of SALL1. The journal of gene medicine, doi: 10.1002/jgm.3127..

Liu, B., Liu, Y., Pan, X., et al. (2019). DNA Methylation Markers for Pan-Cancer Prediction by Deep Learning. Genes *10*, 778.

Liu, P. (2022). Pan-Cancer DNA Methylation Analysis and Tumor Origin Identification of Carcinoma of Unknown Primary Site Based on Multi-Omics. Front Genet. *12*, 798748.

Locke, W.J., Guanzon, D., Ma, C., et al. (2019). DNA Methylation Cancer Biomarkers: Translation to the Clinic. Frontiers in genetics *10*, 1150.

Ma, C., Wang, F., Han, B., et al. (2018). SALL1 functions as a tumor suppressor in breast cancer by regulating cancer cell senescence and metastasis through the NuRD complex. Molecular cancer *17*, 78.

Misawa, K., Misawa, Y., Imai, A., et al. (2018). Epigenetic modification of SALL1 as a novel biomarker for the prognosis of early stage head and neck cancer. Journal of Cancer *9*, 941-949.

Moran, S., Arribas, C., and Esteller, M. (2016a). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics *8*, 389-399.

Moran, S., Martínez-Cardús, A., Sayols, S., et al. (2016b). Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. The Lancet. Oncology *17*, 1386-1395.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *12*, 2825–2830.

Rambow, F., Marine, J.C., and Goding, C.R. (2019). Melanoma plasticity and phenotypic diversity: therapeutic barriers and opportunities. Genes & development *33*, 1295-1318.

Raudvere, U., Kolberg, L., Kuzmin, I., et al. (2019). g:Profiler: a web server for functional enrichment analysis

and conversions of gene lists (2019 update). Nucleic Acids Res *47*, W191-W198.

Richiardi, L., Fiano, V., Vizzini, L., et al. (2009). Promoter methylation in APC, RUNX3, and GSTP1 and mortality in prostate cancer patients. Journal of clinical oncology *27*, 3161-3168.

Salman, H., Shuai, X., Nguyen-Lefebvre, A.T., et al. (2018). SALL1 expression in acute myeloid leukemia. Oncotarget *9*, 7442-7452.

Teh, A.L., Pan, H., Lin, X., et al. (2016). Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. Epigenetics *11*, 36-48.

Tian, Y., Morris, T.J., Webster, A.P., et al. (2017). ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics *33*, 3982-3984.

Vrba, L., and Futscher, B.W. (2018). A suite of DNA methylation markers that can detect most common human cancers. Epigenetics *13*, 61-72.

Wang, J., Ni, J., Beretov, J., et al. (2020). Exosomal microRNAs as liquid biopsy biomarkers in prostate cancer. Critical reviews in oncology/hematology *145*, 102860.

Wang, S., Xiong, Y., Zhao, L., et al. (2021). UCSCXenaShiny: An R/CRAN Package for Interactive Analysis of UCSC Xena Data. Bioinformatics *38*, 527-529.

Weisenberger, D., Van Den Berg, D., Pan, F., et al. (2008). Comprehensive DNA methylation analysis on the Illumina Infinium assay platform. https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf

Xia, D., Leon, A.J., Cabanero, M., et al. (2020). Minimalist approaches to cancer tissue-of-origin classification by DNA methylation. Mod Pathol. *33*,1874-1888.

Ye, Q., Ling, S., Zheng, S., et al. (2019). Liquid biopsy in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA. Molecular cancer *18*, 114.

Yu, J., Yu, H., Yan, J., et al. (2018). Methylation of O6-methylguanine DNA methyltransferase promoter is a predictive biomarker in Chinese melanoma patients treated with alkylating agents. Translational Cancer Research *7*, 495-505.

Zhang, B., Zhou, Y., Lin, N., et al. (2013). Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome research *23*, 1522-1540.

Zhang, Y., Bewerunge-Hudler, M., Schick, M., et al. (2020). Blood-derived DNA methylation predictors of mortality discriminate tumor and healthy tissue in multiple organs. Molecular Oncology *14*, 2111-2123.

Zhang, Y., Qiu, J.G., Jia, X.Y., et al. (2023). METTL3-mediated N6-methyladenosine modification and HDAC5/YY1 promote IFFO1 downregulation in tumor development and chemo-resistance. Cancer letters *553*, 215971.

Zheng, C., and Xu, R. (2020). Predicting cancer origins with a DNA methylation-based deep neural network model. PloS one *15*, e0226461.

Ziller, M.J., Hansen, K.D., Meissner, A., et al. (2015). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. Nature Methods *12*, 230-232.

Figure 1

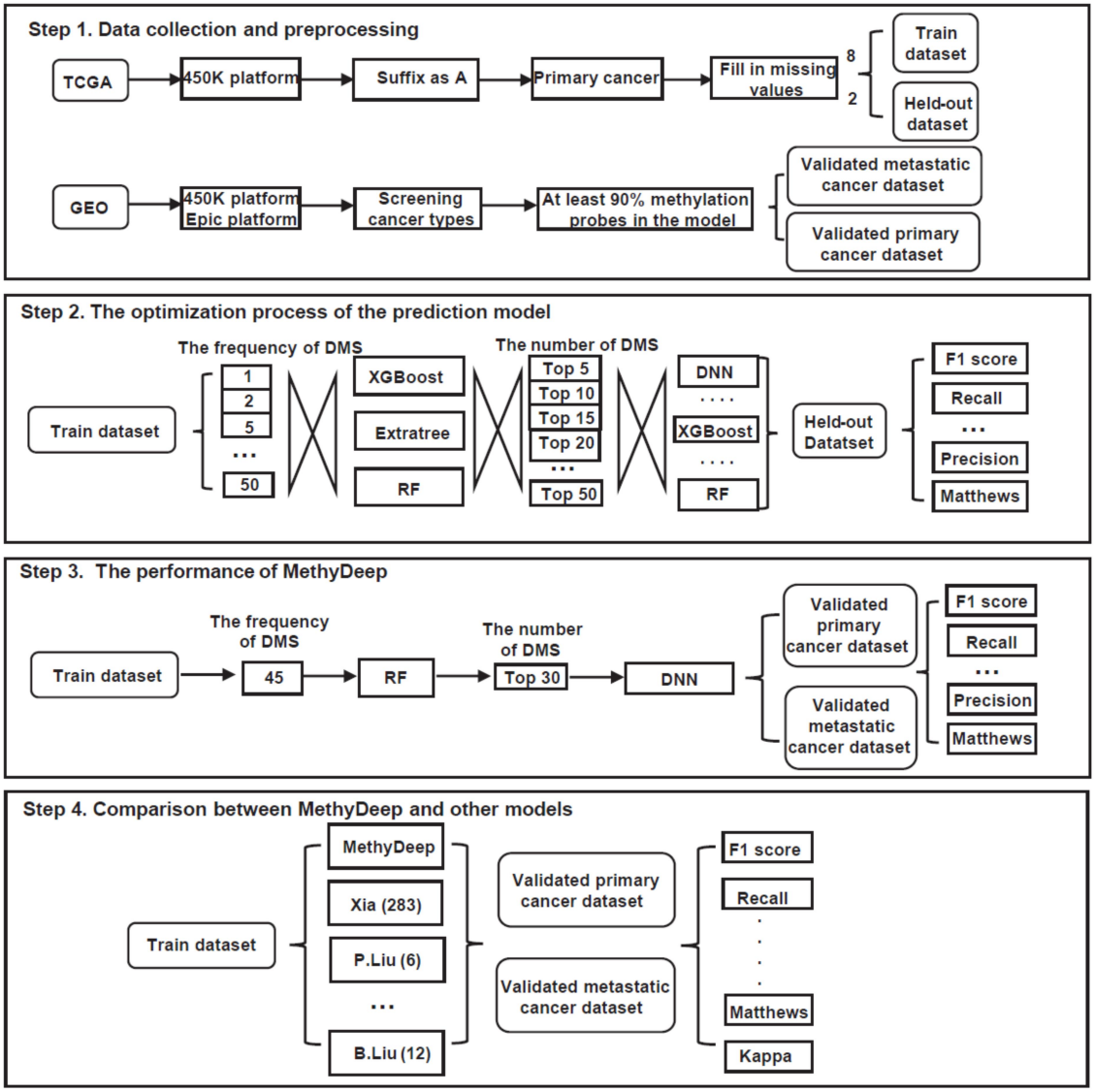

Figure 1 Workflow for the optimization of pan-cancer type prediction.

Figure 2

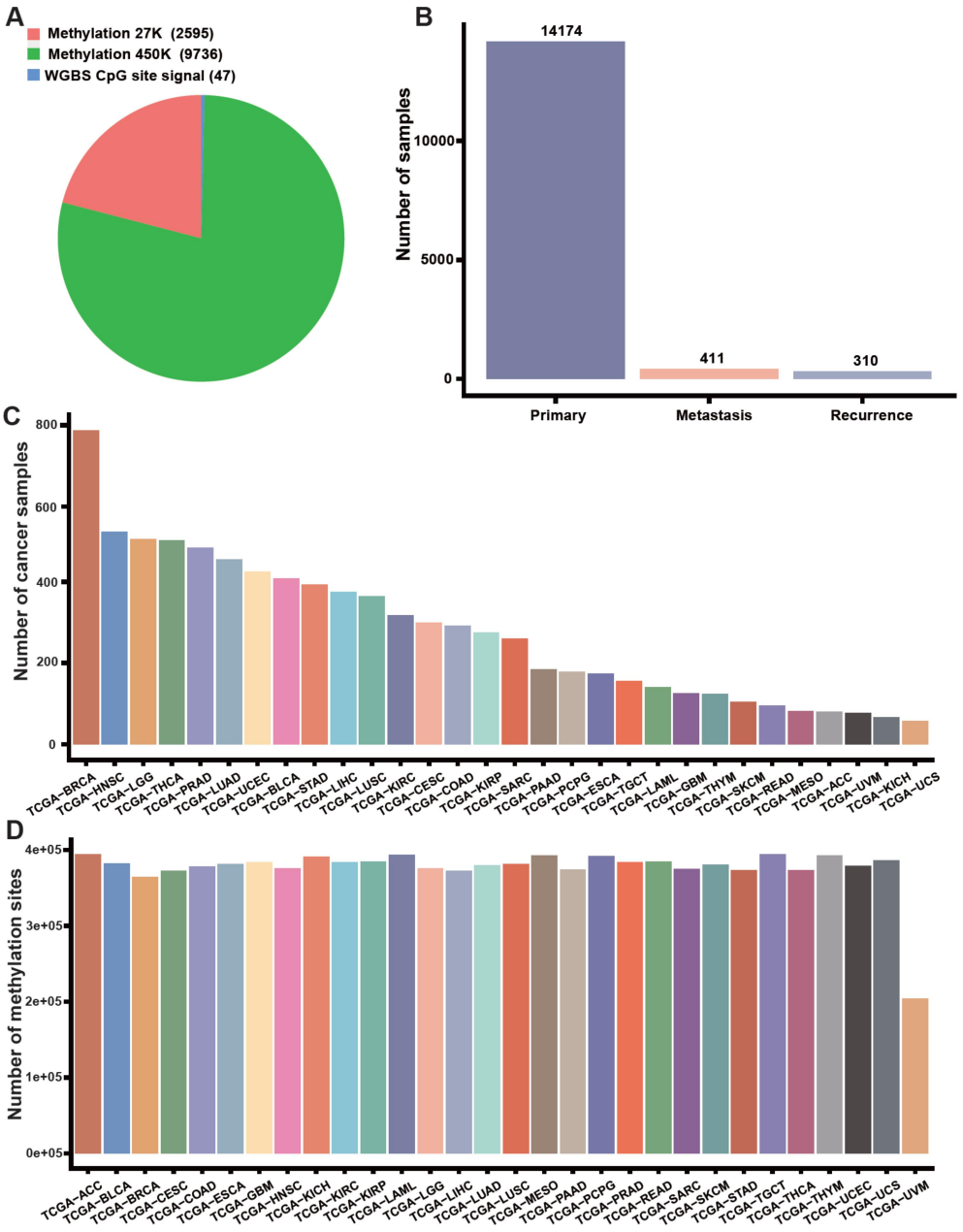

Figure 2 Overview of TCGA methylation data. (**A**) Proportion of methylation data of various platforms in TCGA database. (**B**) Number of samples of various cancer types in TCGA database. (**C**) After preliminary screening, the number of samples for each cancer. (**D**) After preliminary screening, the number of methylation sites of each cancer.

Figure 3

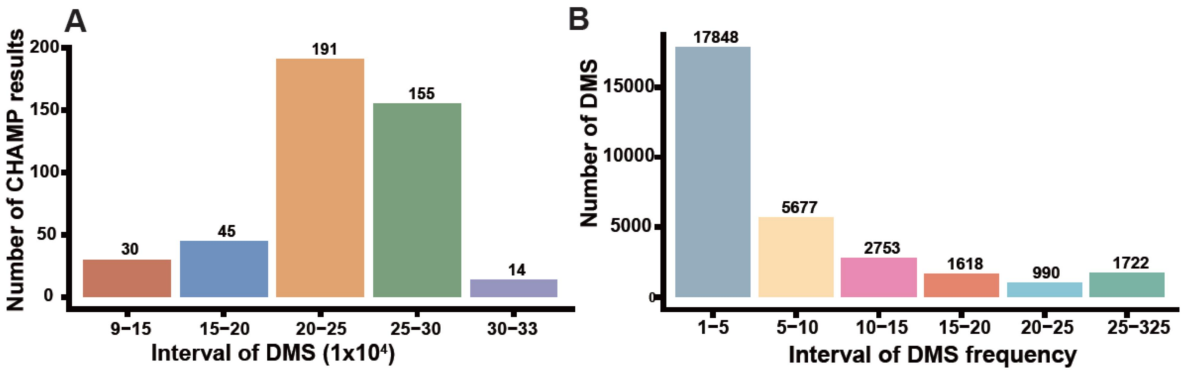

Figure 3 DMS analysis with ChAMP. (**A**) Distribution of DMS number in ChAMP results. The abscissa unit is ten thousand. '9–15' represents the range from 90000 to 150000. (**B**) Distribution of overlapping DMS number for ChAMP-selected DMS. '1–5' means the frequency of DMS is between 1 and 5.

Figure 4

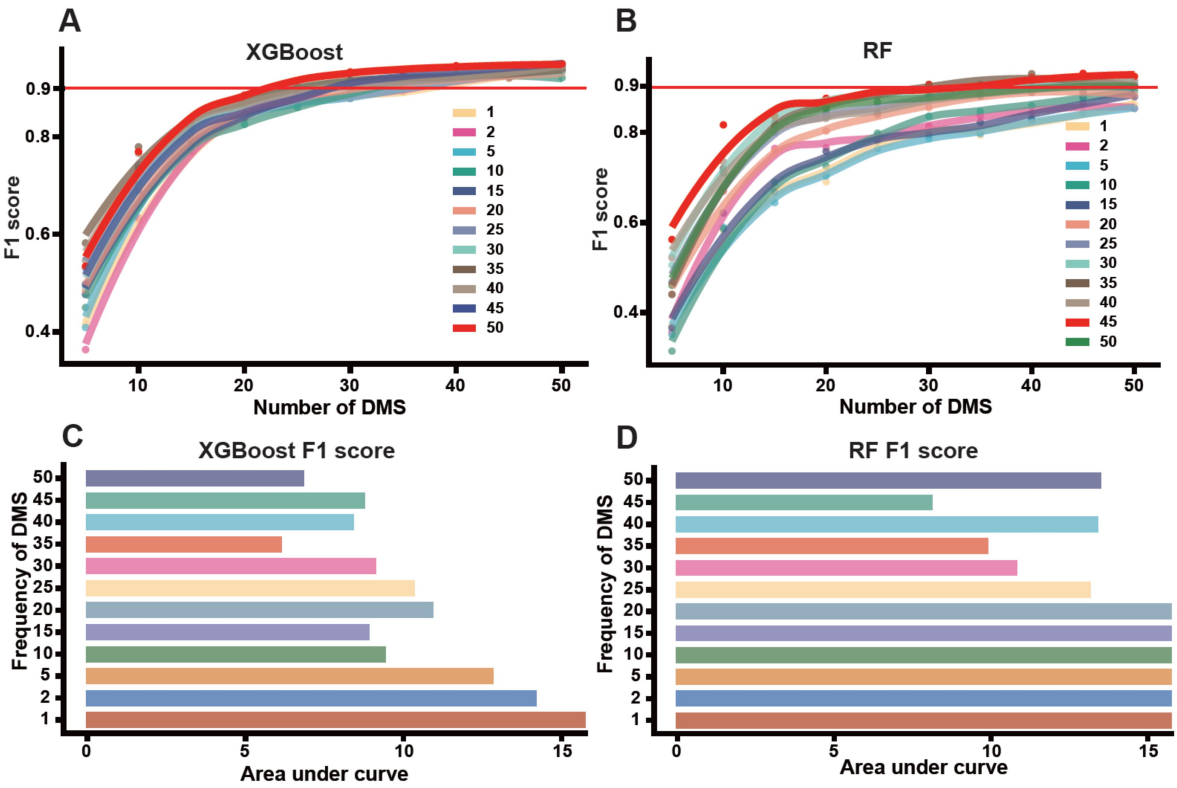

Figure 4 Comparison of machine learning models. (**A**) Comparison of F1 scores of XGBoost under different numbers of methylation sites and different DMS frequencies. The red line represents 0.9. (**B**) Comparison of F1 scores of RF under different numbers of methylation sites and different DMS frequencies. The red line represents 0.9. (**C**) Comparison of the area under the curve when the F1 score of XGBoost reaches 0.9 for the first time under different DMS frequencies. The red line represents 0.9. (**D**) Comparison of the area under the curve when the F1 score of RF reaches 0.9 for the first time under different DMS frequencies.

Figure 5

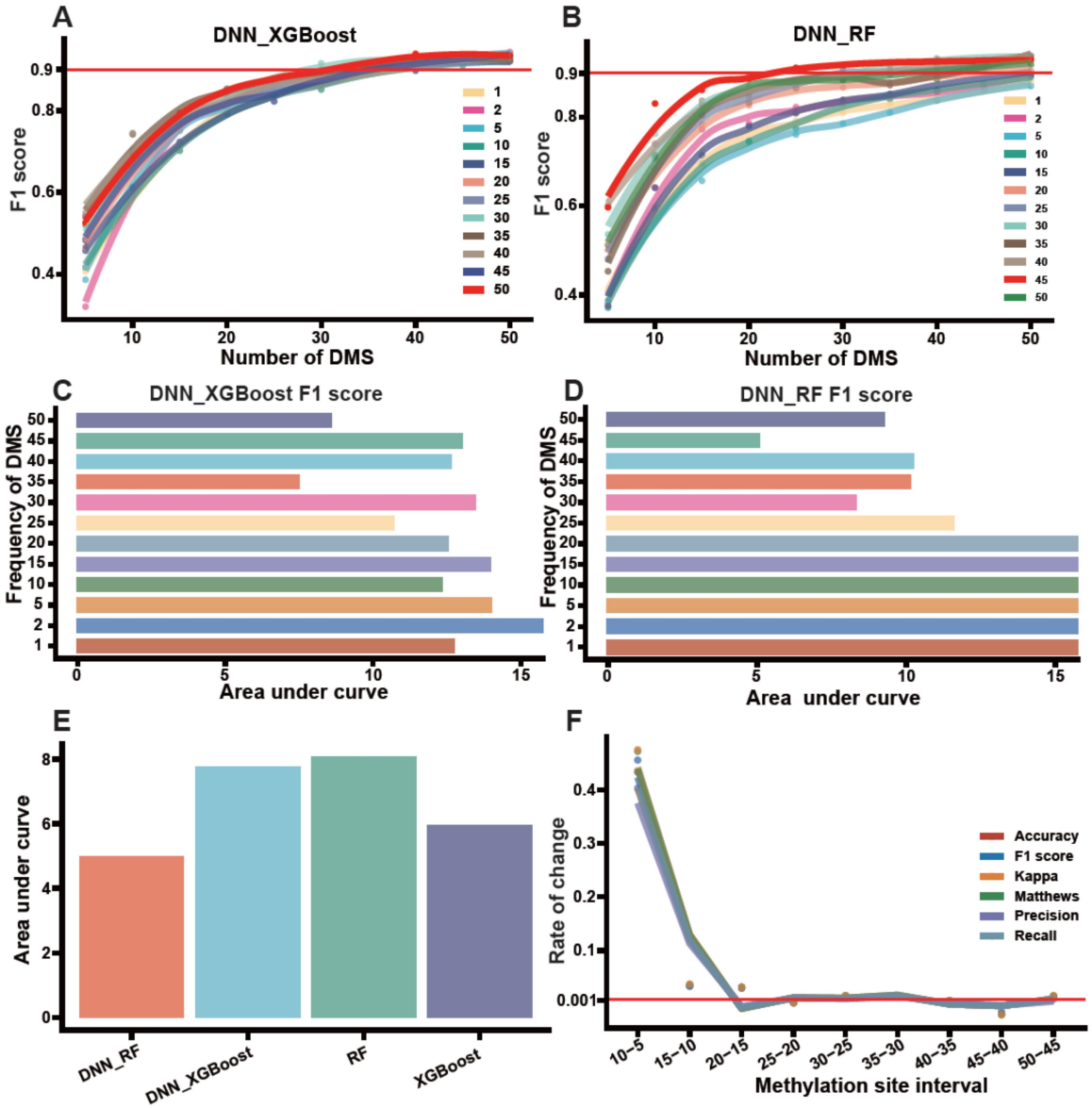

Figure 5 Comparison of deep learning models based on the methylation sites selected by different machine learning methods. (**A**) Comparison of F1 scores of DNN_XGBoost under different numbers of methylation sites and different DMS frequencies. DNN_ XGBoost is a deep learning neuron network model based on the DMSs selected with XGBoost. The red line represents 0.9. (**B**) Comparison of F1 scores of DNN_RF under different numbers of methylation sites and different DMS frequencies. DNN_RF is a deep learning neuron

network model based on the DMSs selected with RF. The red line represents 0.9. (**C**) Comparison of the area under the curve when the F1 score of DNN_XGBoost reaches 0.9 for the first time under different DMS frequencies. (**D**) Comparison of the area under the curve when the F1 score of DNN_RF reaches 0.9 for the first time under different DMS frequencies. (**E**) Comparison of the area under the curve of the four best performing models. (**F**) The relationship between the variation range of six model-evaluating indicators and the number of methylation sites based on the DNN_RF model.

Figure 6

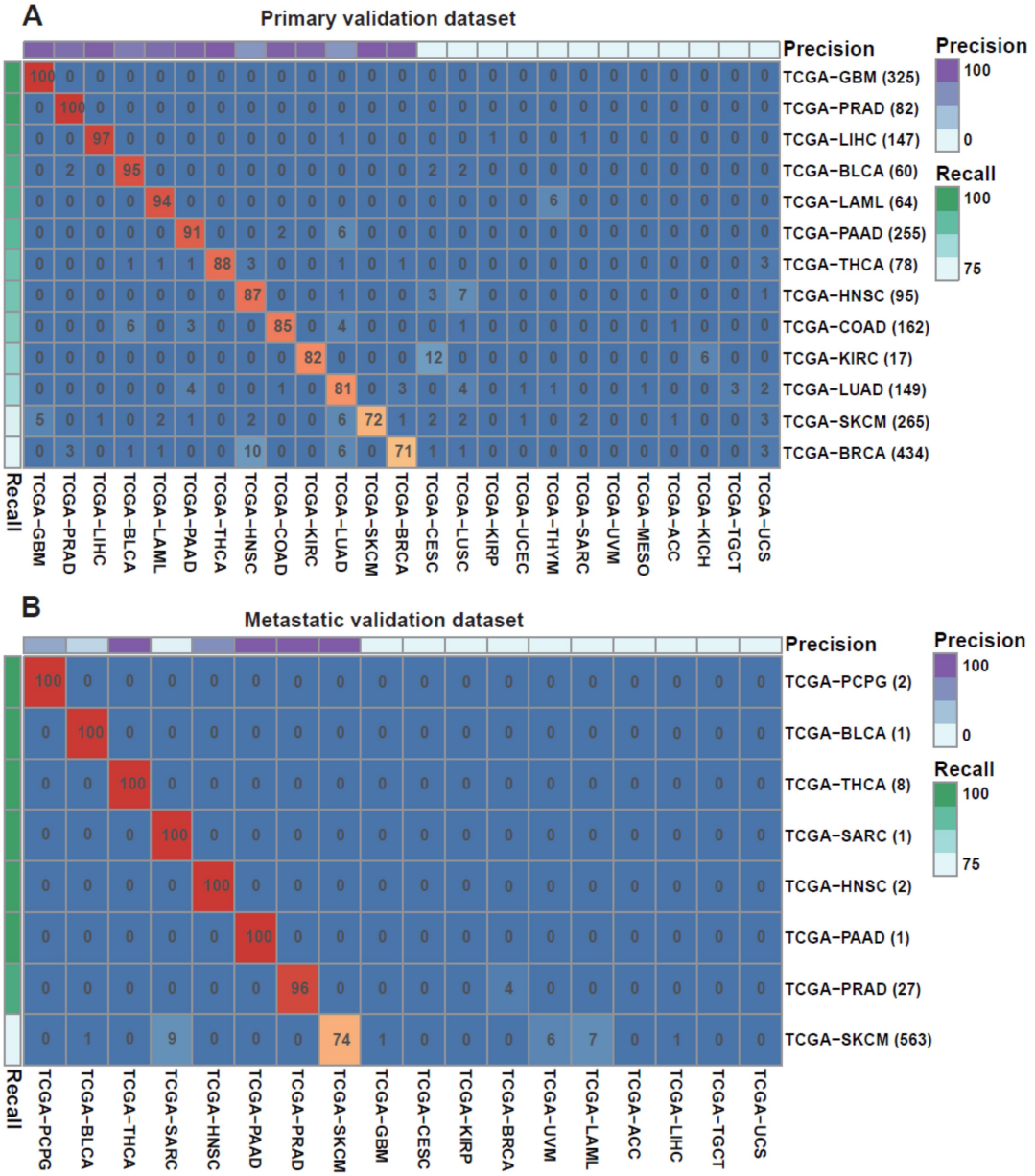

Figure 6 The performance of MethyDeep on independent validation datasets. Each row corresponds to the true tumor type; each column corresponds to the forecast category of the model. The values in the matrix represent the recall rate for each cancer type. The value in the bracket of the label represents the number of cancer samples. The recall and precision of each classifier are displayed in the color bar at the top and left of the matrix.

The total value of some lines is slightly greater or less than 100% due to rounding. (**A**) Primary cancer validation dataset. (**B**) Metastatic cancer validation dataset.

Figure 7

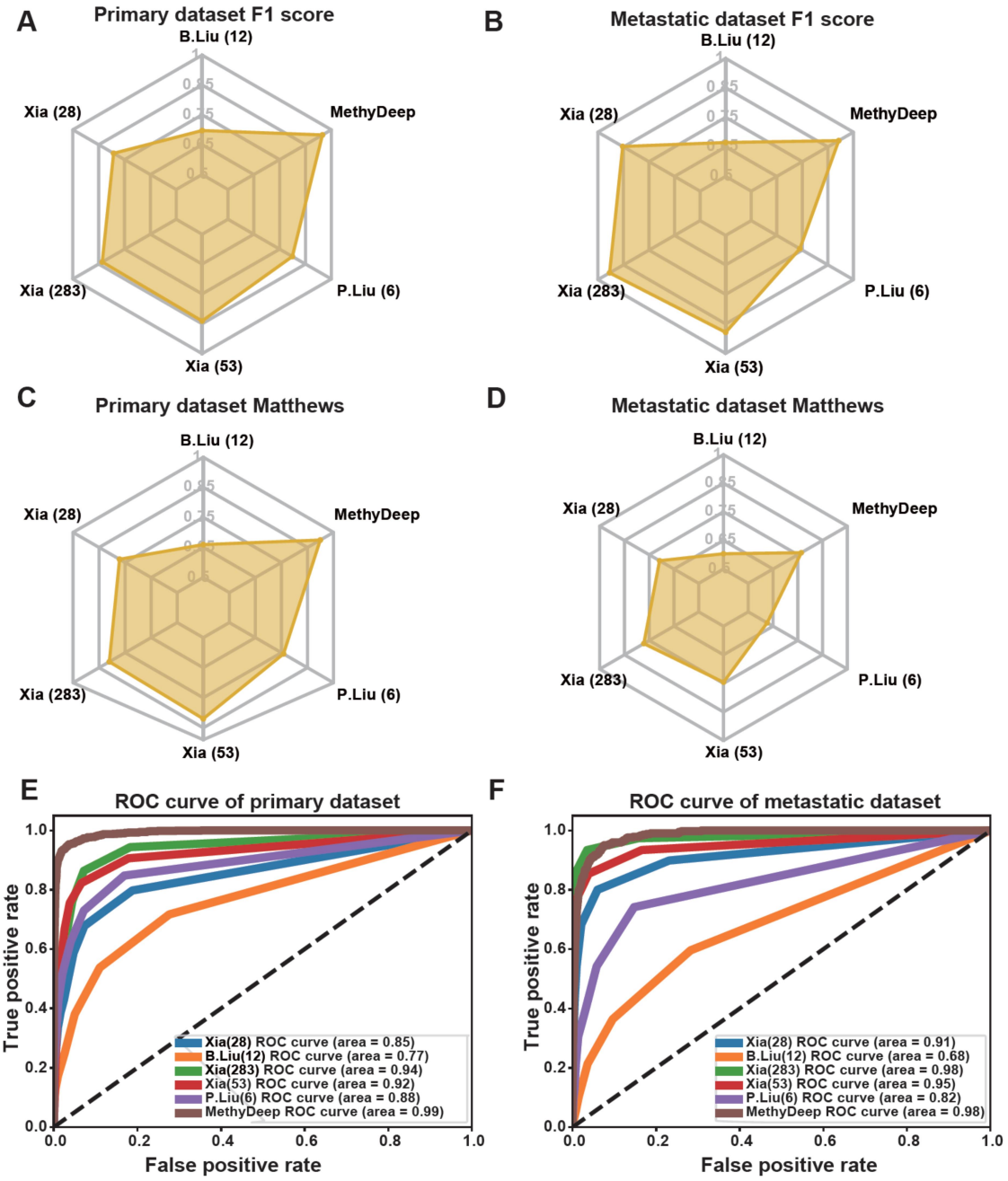

Figure 7 Performance comparison of MethyDeep and other methods. (**A**) F1 score comparison between MethyDeep and other five methods on additional primary cancer dataset. (**B**) F1 score comparison between MethyDeep and other five methods on

additional metastatic cancer dataset. (**C**) Matthews comparison between MethyDeep and other five methods on additional primary cancer dataset. (**D**) Matthews comparison between MethyDeep and other five methods on additional metastatic cancer dataset. (**E**) Comparison of ROC curves of MethyDeep and other five methods on additional primary cancer dataset. (**F**) Comparison of ROC curves of MethyDeep and other five methods on additional metastatic cancer dataset.