HiTAIC: hierarchical tumor artificial intelligence classifier traces tissue of origin and tumor type in primary and metastasized tumors using DNA methylation

Ze Zhang[1,2], Yunrui Lu[2], Soroush Vosoughi[3], Joshua J. Levy[1,2,4], Brock C. Christensen[1,2,5] and Lucas A. Salas[1,2,5,*]

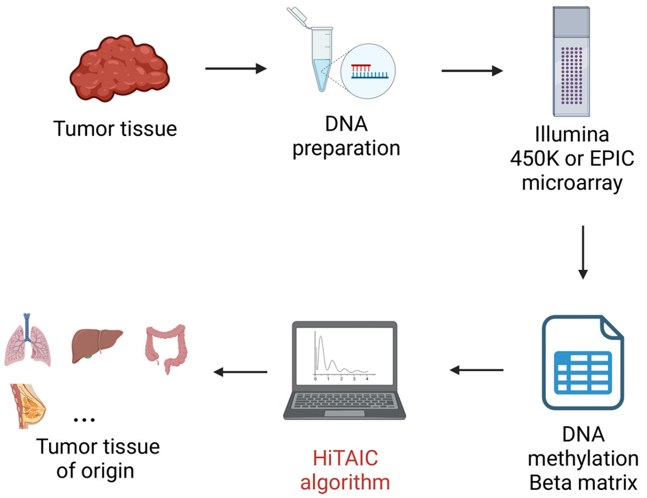
[1]Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA, [2]Quantitative Biomedical Sciences Program, Guarini School of Graduate and Advanced Studies, Dartmouth College, Hanover, NH, USA, [3]Department of Computer Science, Dartmouth College, Hanover, NH, USA, [4]Department of Pathology and Dermatology, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA and [5]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA

Received January 16, 2023; Revised April 04, 2023; Editorial Decision April 07, 2023; Accepted April 13, 2023

ABSTRACT

Human cancers are heterogenous by their cell composition and origination site. Cancer metastasis generates the conundrum of the unknown origin of migrated tumor cells. Tracing tissue of origin and tumor type in primary and metastasized cancer is vital for clinical significance. DNA methylation alterations play a crucial role in carcinogenesis and mark cell fate differentiation, thus can be used to trace tumor tissue of origin. In this study, we employed a novel tumor-type-specific hierarchical model using genome-scale DNA methylation data to develop a multilayer perceptron model, HiTAIC, to trace tissue of origin and tumor type in 27 cancers from 23 tissue sites in data from 7735 tumors with high resolution, accuracy, and specificity. In tracing primary cancer origin, HiTAIC accuracy was 99% in the test set and 93% in the external validation data set. Metastatic cancers were identified with a 96% accuracy in the external data set. HiTAIC is a user-friendly web-based application through https://sites.dartmouth.edu/salaslabhitaic/. In conclusion, we developed HiTAIC, a DNA methylation-based algorithm, to trace tumor tissue of origin in primary and metastasized cancers. The high accuracy and resolution of tumor tracing using HiTAIC holds promise for clinical assistance in identifying cancer of unknown origin.

GRAPHICAL ABSTRACT



INTRODUCTION

Cancer is the second leading cause of death in the United States, following heart disease (1). One thousand six hundred seventy deaths are projected to be caused by cancer per day in 2023, aggregating to an estimated 609 820 cancer deaths in the year (1). Cancer metastasis is the primary cause of cancer mortality, representing around 90% of cancer deaths (2). Metastatic cancer hallmarks a significantly worse prognosis with limited treatment options and a low response rate (2,3). Cancer metastasis happens when advanced tumor cells acquire the ability to detach from the primary tumor tissue, migrate through the blood and lymphatic vessels, invade a distal tissue site, and

[*]To whom correspondence should be addressed. Tel: +1 603 646 5420; Email: lucas.a.salas@dartmouth.edu

proliferate at the new site (2,3). Although primary tumor site for metastatic cancer can usually be recognized by clinical procedures like imaging, immunohistochemistry (IHC) tests and pathological analyses, cancer of unknown primary (CUP) persists for advanced cancer with high tumor cell heterogeneity, atypical morphological patterns, and absence of identifiable features (4). CUP is defined as metastatic cancer for which the primary anatomic origin cannot be identified. CUP accounts for 2–5% of all cancers and marks significantly worse clinical outcomes (5,6). The median survival rate is less than one year for CUP patients (5). Therapeutic strategies today are highly dependent on the clinical, pathological, and molecular profiles of cancer (7,8). CUP poses the challenge of proper and timely treatment for patients, resulting in unfavorable clinical outcomes. Although clinical tools, e.g. IHC, pathology test, are available for identifying CUP, only 25% of CUP can be diagnosed due to the limited sensitivity and specificity of the traditional methods (9). Additional tools to assist CUP identification could be clinically important.

DNA methylation is an epigenetic modification that regulates gene expression and is essential to establishing and preserving cellular identity (10). In recent years, DNA methylation has been widely utilized as a biomarker for cell typing in blood and the tumor microenvironment (11–14). Furthermore, cell-free DNA methylation profiling is beginning to show promise for early detection and classification of cancer (15,16). As a well-established biomarker for cell identity, DNA methylation holds promising value for distinguishing heterogeneous tumor subtypes, especially for CUP. Genome-wide DNA methylation arrays provide a standardized and cost-effective approach to measuring DNA methylation (17). The high-dimensional methylation data in combination with artificial intelligence (AI) technologies promises new opportunities to efficiently trace tumor tissue of origin that may have clinical significance, especially for metastasized cancer and CUP.

The advance of AI in biomedical science enables translational technology from sophisticated computational tasks and high-dimensional data to potential clinical usage (18). AI-powered medicine provides streamlined analysis and efficient processing of complex clinical and biomedical data, especially in pathology and laboratory medicine (19). In the past decade, by virtue of the progress made in computational capacity and new technologies for genomic sequencing, publicly available biomedical data repertoires were established and structured to serve the scientific community with easily accessible data sets, e.g. The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), and Array-Express. Combined with advances in AI, researchers have utilized the enormity of publicly available data to study genomic biology. Perhaps the most popular example of these technologies is the AI-powered protein folding prediction research (20). At the genomic level, studies are beginning to show the application of machine learning modeling to integrate multiomics information for disease diagnosis and prognostication, which is especially relevant for studying cancer biology (21–23).

In recent years, researchers have demonstrated high performance of DNA methylation-based machine learning models tracing the tissue origin of tumor cells (24–26). However, previous works were limited by the number and variation of cancer types and validation data sets considered, especially for metastasized cancers and circulating cell-free DNA from cancer patients. Furthermore, previous models were devised based on tissue site instead of tumor type, generating potential problems of indistinguishable tumor subtypes from the same site, e.g. esophageal squamous cell carcinoma versus esophageal adenocarcinoma. Although there is some work to illustrate the utility and replicability of the DNA methylation-based machine learning models on tracing tissue of origin for cancers, these works lacked tumor-type specificity and were not made readily accessible and user-friendly to the general scientific community. A previous study used a hierarchical modeling approach to address the challenge of deconvolving cell types that are of same lineage in the tumor microenvironment (14). Similarly, to address the limitations of existing methods and enhance the accuracy, utility and accessibility of tumor tracing, we developed a novel DNA methylation-based algorithm that employs a tumor-type-specific hierarchical model and broadens the number of solid tumor types that are traced. Our method, called Hierarchical Tumor Artificial Intelligence Classifier (HiTAIC), uses multilayer perceptron models in combination with the discriminatory CpGs specific to tumor type in each layer in the hierarchy, to trace tumor tissue of origin and subtypes in 27 primary and metastasized cancers. HiTAIC's ability to trace tumor tissue of origin with high resolution promises valuable application to clinical CUP identification. HiTAIC is publicly accessible on a user-friendly web page https://sites.dartmouth.edu/salaslabhitaic/.

MATERIALS AND METHODS

Discovery dataset and quality control

The initial discovery data sets included DNA methylation microarray data from 7932 samples across 30 cancer types with tagged known primary from TCGA, which is a publicly available cancer data repertoire. 194 leukemia samples were removed as we target only solid tumors. Ninety-nine ovarian tumor samples were added to the discovery data set from GEO data set GSE133556 due to limited ovarian tumor sample size on TCGA. 102 samples were excluded from the discovery data set because of the ambiguity of the tumor subtypes. The ambiguous tumors are rare tumor subtypes that do not fall into any category of the HiTAIC hierarchy. Supplementary Table S1 summarized the discovery data set based on cancer site, tumor subtype, and exclusion criteria. In total, 7735 tumor samples from 27 cancer types were included in the discovery data set (Table 1). The discovery data set was then randomly split into 80% training and 20% testing for model training and testing. In the methylation data quality control process, we retained CpGs that measured on both Illumina HumanMethylation450k and HumanMethylationEPIC platforms to accommodate cross-platform applications. The *SeSAMe* (version 1.8.2) pipeline from Bioconductor was used to preprocess the data, including data normalization and quality control (27). Cross-reactive probes, SNP-related probes, sex chromosome probes, non-CpG probes and low-quality probes (pOOBHA > 0.05) were masked in the analysis. 384640 CpGs were retained after this process.

Tumor classifier hierarchy and tumor type specific CpG identification

The tumor classifier hierarchy was established based on cancer pathophysiological differences and tissue location by tumor type. Two layers with four categories were established for 27 cancer types in the hierarchy (Figure 1). Layer 1 contains nine major tumor types. Layer 2A contains 16 types of adenocarcinoma. Layer 2B includes three types of squamous cell carcinoma. Layer 2C includes two types of melanoma. The sample labels were generated by examining the primary diagnosis and cancer-type information from TCGA following the HiTAIC hierarchy. Sample labels can be found on FigShare (DOI: 10.6084/m9.figshare.22179089). To reduce the high-dimensionality of the DNA methylation data with 384 640 CpGs, we performed epigenome-wide association study (EWAS) to identity differentially methylated CpGs across the cancer types in each category of the hierarchy as input for machine-learning model training using the whole discovery dataset to maximize the power. We applied the *Meffil* (version 1.1.1) package in R (28), which used *limma* linear regression with empirical Bayes adjustment statistics to reduce methylation profiles to top 100 cell-type-specific hyper- and hypo-methylated CpGs per cancer type. Thus, four libraries of cancer-type discriminatory CpGs were developed. T-distributed stochastic neighbor embedding (T-SNE) was used to visualize the separation of cancers by methylation status from the *Meffil* selected CpGs in the libraries. R version 4.2.0 was used in this study.

Machine learning model development, validation and application

All data mining and machine learning model building were operated in Jupyter Notebook Python 3 with the *scikit-learn 1.1.1* package (29). To select the best model, four different types of multi-class machine learning models were tested on the adenocarcinoma, squamous cell carcinoma, glioblastoma, and melanoma samples splitting to 80% training and 20% testing randomly. The support vector machine (SVM) model was established using the *sklearn.svm.SVC* function with gamma set as 'auto'. The random forest classifier (RFC) was built using the *sklearn.ensemble.RandomForestClassifier* function with the number of trees set as 500. The Gaussian naïve Bayes (GNB) model was established using the *sklearn.naive_bayes.GaussianNB* function. Finally, the multilayer perceptron (MLP) model was built using the *sklearn.neural_network.MLPClassifier* function with the max number of iterations set at 300. The performances of the models were evaluated on the test data set using the *sklearn.metrics.confusion_matrix* and *sklearn.metrics.classification_report* functions, which include stratified and overall precision, recall and $F1$-score. Among four machine learning models applied, the MLP performed the best in the test data set (Supplementary Table S2). Thus, MLP model was selected as the final model for cancer classification. In each layer of the hierarchy, an MLP model was trained using the selected cancer type discriminatory CpGs. In total, four MLPs were developed for the hierarchy. HiTAIC was established based on the four hierarchical MLP models. Next, hyperparameter tuning was conducted to select the best set of hyperparameters for the MLP model in each layer of the hierarchy. The hidden layer parameter iterated through [100], [200], [500], [1000], [1000, 500], [1000, 200], [1000, 100], [500, 200], [500, 100], [200, 100], [1000, 500, 200], [1000, 500, 100], [500, 200, 100]. The optimizer iterated through 'sgd' and 'adam'. The learning rate iterated through 0.0001, 0.0005, 0.001, 0.002, 0.005, 0.01. To ensure that the MLP model is generalizable and the performance of the model is consistency across the discovery data set, we established 5-fold cross-validation, which split the data into 80% training and 20% testing 5 times for model evaluation, in each layer of the hierarchy. For external validation, we identified 1175 samples with DNA methylation data on 24 cancer subtypes from 21 publicly available data sets on GEO and ArrayExpress (30–41) (Supplementary Table S3). To further validate HiTAIC on samples with a low tumor purity, we applied HiTAIC to TCGA adenocarcinoma samples with a tumor purity <30% based on HITIMED DNA methylation tumor cell deconvolution (14). Although developed to trace tumor origin, we hypothesize that HiTAIC could also provide tissue of origin information on normal samples. As a result, we tested HiTAIC on normal tissues, including breast, lung, kidney, liver and colon, from TCGA with five samples per tissue type. We also tested HiTAIC on the 102 rare tumors we excluded for training. To demonstrate the application of the model in cancer metastasis, we identified 175 samples with DNA methylation data on five cancer types with six different metastatic locations from eight data sets on GEO and ArrayExpress (42–45) (Supplementary Table S4). We further identified 266 cfDNA samples with DNA methylation data from cancer patients in five types of cancer on GEO for application (46–50) (Supplementary Table S5). We applied the model to the external validation data sets computing stratified and overall precision, recall and $F1$-score to evaluate the performance. Next, we applied the model to the application data sets and used stratified and overall precision, recall, and F1-score to evaluate model performance in metastasized cancers and cfDNA from cancer patients.

Functional pathway and genomic context enrichment analyses

To explore the potential biological pathways and functions related to the CpGs distinguishing cancers, we used the Genomic Regions Enrichment of Annotations Tool (GREAT) to perform enrichment analysis for cancer type specific CpGs in each layer in the hierarchy. GREAT uses Gene Ontology (GO) database, which includes biological process, cellular component, and molecular function categories. False discovery rate (FDR) was used to select and rank the significantly enriched GO terms ($FDR < 0.05$). Next, genomic context enrichment analyses were conducted to investigate whether the cancer type specific CpGs are enriched in certain genomic locations. The relation of probes to CpG islands and enhancers was identified from the

Table 1. Baseline characteristics of the discovery data set

Cancer type	Location	*N*	Mean age (SD)	Male *N* (%)	Data source
Adrenocortical adenocarcinoma	Adrenal gland	80	47 (15.9)	31 (38.8)	TCGA
Bladder adenocarcinoma	Bladder	410	69 (10.6)	303 (73.9)	TCGA
Breast adenocarcinoma	Breast	761	59 (13.2)	9 (1.2)	TCGA
Cervical adenocarcinoma	Cervix	48	46 (12.3)	0 (0)	TCGA
Cervical squamous cell carcinoma	Cervix	259	49 (14.0)	0 (0)	TCGA
Colorectal adenocarcinoma	Colon and rectum	379	64 (12.9)	202 (52.3)	TCGA
Glioma	Brain	138	60 (12.8)	80 (58.0)	TCGA
Kidney chromophobe adenocarcinoma	Kidney	66	52 (14.3)	39 (59.1)	TCGA
Kidney renal clear cell adenocarcinoma	Kidney	319	61 (11.8)	205 (64.3)	TCGA
Kidney renal papillary cell adenocarcinoma	Kidney	275	62 (12.1)	202 (73.5)	TCGA
Liver hepatocellular adenocarcinoma	Liver	377	59 (13.5)	255 (67.6)	TCGA
Lung adenocarcinoma	Lung	458	65 (10.2)	214 (46.7)	TCGA
Lung squamous cell carcinoma	Lung	370	68 (8.7)	274 (74.1)	TCGA
Pancreatic adenocarcinoma	Pancreas	175	66 (11.0)	98 (56.0)	TCGA
Prostate adenocarcinoma	Prostate	498	61 (6.8)	498 (100.0)	TCGA
Esophageal and head and neck squamous cell carcinoma	Esophagus and head and neck	624	61 (11.7)	467 (75.8)	TCGA
Cutaneous melanoma	Skin	104	65 (13.9)	62 (59.6)	TCGA
Uveal melanoma	Eye	80	62 (14.0)	45 (56.2)	TCGA
Esophageal and stomach adenocarcinoma	Esophagus and stomach	484	66 (10.9)	336 (69.4)	TCGA
Thyroid adenocarcinoma	Thyroid	502	47 (15.8)	133 (26.5)	TCGA
Mesothelioma	Pleura	87	64 (9.8)	71 (81.6)	TCGA
Pheochromocytoma and Paraganglioma	Adrenal gland	179	48 (15.1)	78 (43.6)	TCGA
Sarcoma	Soft tissues	249	61 (14.7)	113 (45.4)	TCGA
Testicular germ cell tumor	Testis	150	32 (9.3)	150 (100)	TCGA
Ovarian adenocarcinoma	Ovary	109	49 (13.5)*	0 (0)	TCGA, GSE133556
Thymoma	Thymus	124	59 (13.0)	64 (51.6)	TCGA
Endometrial adenocarcinoma	Uterus	430	65 (11.2)	0 (0)	TCGA
Total		7735			

Horvath methylation age inferred using the *wateRmelon* package in R for GSE133556.

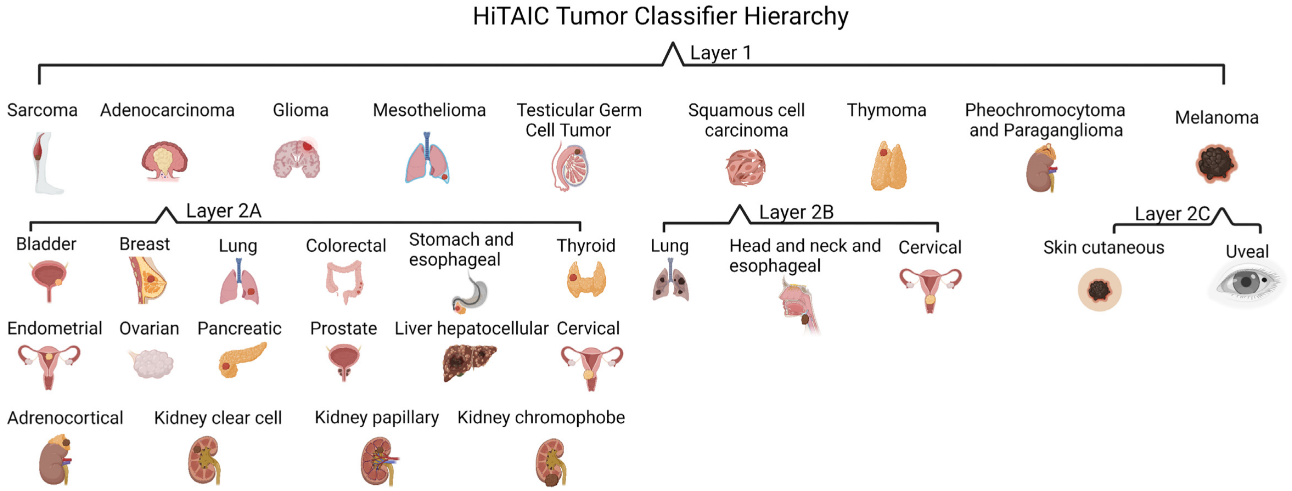


Figure 1. HiTAIC tumor classifier hierarchy with two layers for twenty-seven cancer types (created with BioRender.com).

HumanMethylation450K annotation file. To define the genomic regions as promoters, introns, exons, or intergenic for each probe, the *annotateWithGeneParts* function from the R-package *genomation* and the *UCSC_hg19_refGene* file were used to map the regions to all CpG loci on the Illumina *HumanMethylation450K* array. If a probe mapped to more than a single genomic region, the probe was assigned preferentially with the order: promoters, exons, introns, and intergenic. Fisher's exact tests were conducted to calculate odds ratios (ORs), *P*-values and 95% confidence intervals for genomic context enrichment analysis. For both functional pathway and genomic context enrichment analyses, cancer discerning CpGs in each layer were tested over the background CpGs used ($n = 384\,640$) for EWAS.

HiTAIC web-based application development

We used Python Dash and Heroku to develop a user-friendly web-based HiTAIC tool. Python Dash is a framework developed by Plotly, which is based on Python and used for building and deploying data apps with customized interface. Heroku, which is a cloud platform, was next

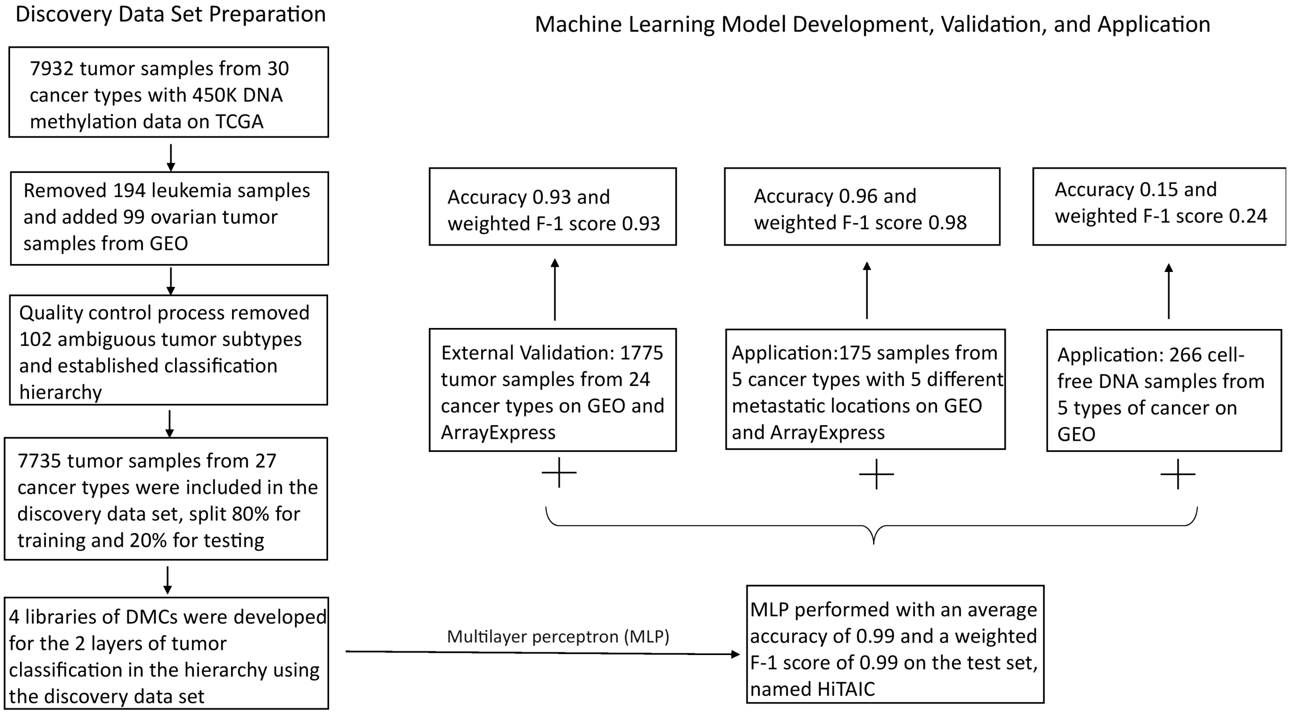

Discovery Data Set Preparation
7932 tumor samples from 30 cancer types with 450K DNA methylation data on TCGA
Removed 194 leukemia samples and added 99 ovarian tumor samples from GEO
Quality control process removed 102 ambiguous tumor subtypes and established classification hierarchy
7735 tumor samples from 27 cancer types were included in the discovery data set, split 80% for training and 20% for testing
4 libraries of DMCs were developed for the 2 layers of tumor classification in the hierarchy using the discovery data set
Multilayer perceptron (MLP)
Machine Learning Model Development, Validation, and Application
MLP performed with an average accuracy of 0.99 and a weighted F-1 score of 0.99 on the test set, named HiTAIC
External Validation: 1775 tumor samples from 24 cancer types on GEO and ArrayExpress
Accuracy 0.93 and weighted F-1 score 0.93
Application:175 samples from 5 cancer types with 5 different metastatic locations on GEO and ArrayExpress
Accuracy 0.96 and weighted F-1 score 0.98
Application: 266 cell-free DNA samples from 5 types of cancer on GEO
Accuracy 0.15 and weighted F-1 score 0.24

Figure 2. Flowchart of HiTAIC model building, validation and application.

employed to host and deploy the HITAIC web application. The HiTAIC web application contains two major parts. The first part is user guide. Users should follow the instructions to finish the prediction process. An exemplary input data csv file was available for demonstration. The second part includes data upload, model running, and output download. After constructing the input data as instructed, users can either click the data upload box to choose the file or drag the file to the box from the local end to upload the input data. Then the algorithm will automatically compute the output and show up on the right side of the panel. Users can also download the output result as a csv file by clicking the export box.

RESULTS

The pipeline of this study is shown in Figure 2. In total, four libraries of cancer type discriminatory CpGs were developed for the hierarchy. In Layer 1, 1641 CpGs were identified to discern nine major cancer types. In Layer 2A, 3225 CpGs were identified to distinguish 18 adenocarcinoma cancer types. In Layer 2B, 767 CpGs were identified to discriminate four squamous cell carcinoma cancer types. In Layer 2C, 200 CpGs were identified to discern two types of melanoma. The heatmaps in Figure 3 demonstrated discriminative methylation status for the cancer type specific CpGs in the libraries. The libraries are relatively unique to each other as 0 overlapped CpGs were identified across the four libraries, one CpG appeared in three out of four libraries, and 80 CpGs in total overlapped in two out of four libraries (Supplementary Figure S1)

T-SNE clustering showed separation of clusters by cancer type using the cancer type discriminatory CpGs. However, certain cancer types did not show clear separation. Esophageal carcinoma was split into clusters with head and neck squamous cell carcinoma and stomach adenocarcinoma. Colon and rectum adenocarcinoma samples were indistinguishable in T-SNE (Supplementary Figure S2A). The two clusters of esophageal carcinoma can be separated by tumor subtypes, i.e. squamous cell carcinoma versus adenocarcinoma. Esophageal squamous cell carcinoma was clustered with head and neck squamous cell carcinoma (Supplementary Figure S2B) while esophageal adenocarcinoma was clustered with stomach adenocarcinoma (Supplementary Figure S2C). To avoid ambiguity and ensure the sensitivity of the model, we collapsed colon adenocarcinoma and rectal adenocarcinoma into colorectal adenocarcinoma in the adenocarcinoma layer, esophageal and head and neck squamous cell carcinoma into one group in the squamous cell carcinoma layer, and esophageal and stomach adenocarcinoma into one group in the adenocarcinoma layer in the hierarchy for MLP model training. Indicating differential methylation status by cancer types and feasibility for machine learning model training the T-SNE clustering showed clear separation of cancer types by using the cancer type discriminatory CpGs following the hierarchical cancer classification regime with minimal outliers. (Supplementary Figure S3).

Next, HiTAIC was trained using the MLP models and cancer type discriminatory CpGs using the training data set following the hierarchical structure. HiTAIC integrated four MLP models for tracing tumor tissue of origin. With the 156 sets of hyperparameters examined, the accuracy ranged from 96–98%, 82–98%, 87–99% and 100% for Layer 1 (Supplementary Table S6), Layer 2A (Supplementary Table S7), Layer 2B (Supplementary Table S8) and Layer 2C

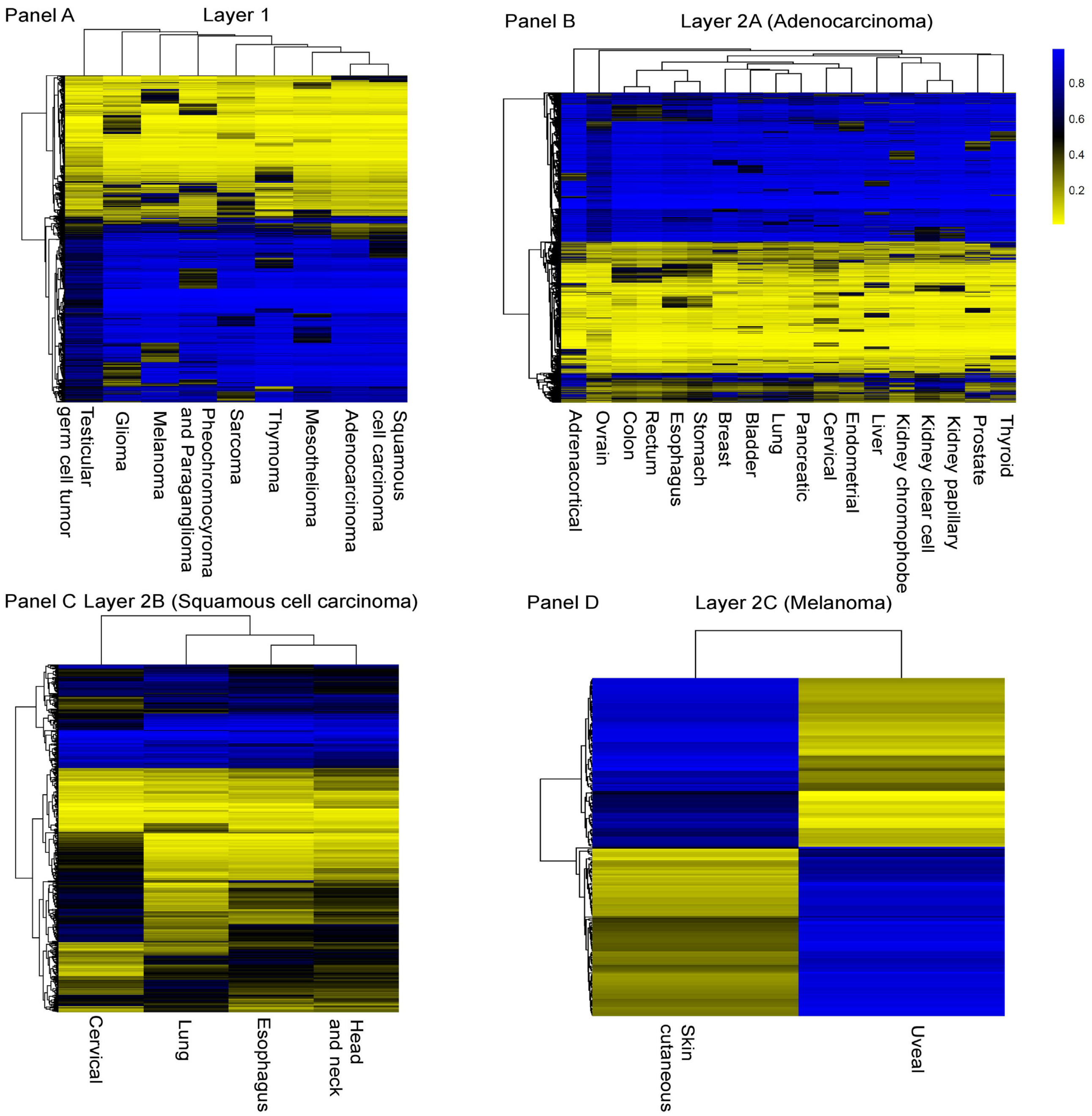


Figure 3. DNA methylation states of cancer discriminating CpGs for Panel **A.** Layer 1 major cancer types; Panel **B.** Layer 2A adenocarcinoma subtypes; Panel **C.** Layer 2B squamous cell carcinoma subtypes; Panel **D**. Layer 2C melanoma subtypes.

(Supplementary Table S9) respectively. As a result, we selected 1 hidden layer with 100 nodes, 'adam' optimizer, and 0.001 learning rate as the hyperparameters for the final model. The architecture of the HiTAIC model is shown in Supplementary Table S10. Specifically in the testing data set, in Layer 1, HiTAIC performed with 98% accuracy and 98% weighted average F1-score (Table 2). In Layer 2A with adenocarcinoma, HiTAIC performed with 98% accuracy and 98% weighted average F1-score (Table 3). In Layer 2B with squamous cell carcinoma, HiTAIC performed with 99% accuracy and 99% weighted average F1-score (Table 4). In Layer 2C with melanoma, HiTAIC performed with 100% accuracy and 100% weighted average *F*1-score (Table 5). For 5-fold cross-validation, the model performed consistently well in every layer of the hierarchy. In Layer 1, the accuracy and weighted average *F*1-score are all 98% (Supplementary Table S11). In Layer 2A, the accuracy and weighted average F1-score ranged from 98% to 99% (Supplementary Table S12). In Layer 2B, the accuracy and weighted average *F*1-score ranged from 96% to 99% (Supplementary Table S13). In Layer 2C, the accuracy and weighted average *F*1-score ranged from 97% to

Table 2. HiTAIC performance on the test data set for Layer 1 cancer types

Layer 1	Precision	Recall	*F*1-score	Sample size
Adenocarcinoma	0.99	0.99	0.99	1075
Glioma	1	1	1	27
Melanoma	0.97	0.97	0.97	37
Mesothelioma	1	0.94	0.97	17
Pheochromocytoma and paraganglioma	1	1	1	36
Sarcoma	0.92	0.96	0.94	50
Squamous cell carcinoma	0.97	0.94	0.96	251
Testicular germ cell tumor	1	1	1	30
Thymoma	1	1	1	25
Accuracy			0.98	1548
Macro average	0.98	0.98	0.98	1548
Weighted average	0.98	0.98	0.98	1548

Table 3. HiTAIC performance on the test data set for Layer 2A (adenocarcinoma) cancer subtypes

Layer 2A (adenocarcinoma)	Precision	Recall	*F*1-score	Sample size
Adrenocortical	1	1	1	16
Bladder	0.96	1	0.98	82
Breast	0.99	1	1	152
Cervical	0.88	0.7	0.78	10
Colorectal	0.95	1	0.97	76
Endometrial	0.99	0.97	0.98	86
Kidney chromophobe	0.93	1	0.96	13
Kidney clear cell	0.98	0.92	0.95	64
Kidney papillary cell	0.96	0.95	0.95	55
Liver hepatocellular	1	1	1	75
Esophageal and stomach	1	1	1	97
Lung	0.99	0.98	0.98	92
Ovarian	0.96	1	0.98	22
Pancreatic	0.94	0.94	0.94	35
Prostate	1	1	1	100
Thyroid	1	1	1	100
Accuracy			0.98	1075
Macro average	0.97	0.97	0.97	1075
Weighted average	0.98	0.98	0.98	1075

100% (Supplementary Table S14). HiTIMED identified 101 TCGA adenocarcinoma samples with a tumor purity below 30% (14) (Supplementary Table S15). HiTAIC achieved 97% accuracy and 98% weighted average F1-score on those samples (Supplementary Table S16). When applied to normal tissues, HiTAIC classified them into their corresponding tumor types, i.e. normal breast as breast adenocarcinoma, normal lung as lung adenocarcinoma, normal kidney as kidney adenocarcinoma, normal liver as liver adenocarcinoma and normal colon as colorectal adenocarcinoma (Supplementary Table S17). When applied to the 102 rare tumors that excluded for model training, HiTAIC achieved 81% accuracy to trace tissue of origin (Supplementary Table S18). Although the accuracy is lower relative to its performance on more common tumor subtypes, HiTAIC provides useful information that might be clinically useful for the rare tumors. For example, two bladder squamous cell carcinoma samples were accurately classified as squamous cell carcinoma in the first layer of HiTAIC classification despite that HiTAIC does not have a bladder squamous cell carcinoma category.

Table 4. HiTAIC performance on the test data set for Layer 2B (squamous cell carcinoma) cancer subtypes

Layer 2B (squamous cell carcinoma)	Precision	Recall	*F*1-score	Sample size
Cervical	1	0.98	0.99	52
Esophageal and head and neck	0.99	0.99	0.99	125
Lung	0.97	0.99	0.98	74
Accuracy			0.99	251
Macro average	0.99	0.99	0.99	251
Weighted average	0.99	0.99	0.99	251

Table 5. HiTAIC performance on the test data set for Layer 2C (melanoma) cancer subtypes

Layer 2C (melanoma)	Precision	Recall	F1-score	Sample size
Eye uveal	1	1	1	16
Skin cutaneous	1	1	1	21
Accuracy			1	37
Macro average	1	1	1	37
Weighted average	1	1	1	37

For external validation, we observed 93% accuracy and 93% weighted average F1-score across 25 cancer types (Table 6). Specifically, all cancer types showed a *F*1-score over 80% except for endometrial adenocarcinoma (F1-score 51%) and ovarian adenocarcinoma (F1-score 77%). Among 65 ovarian adenocarcinomas in the validation data set, 16 were misclassified. All of them were misclassified as endometrial adenocarcinoma, resulting in a compromised *F*1-score for endometrial adenocarcinoma and ovarian adenocarcinoma. Although the *F*1-scores were relatively low in endometrial adenocarcinoma and ovarian adenocarcinoma, the misclassification were contained within the gynecologic cancer types, which still provides valuable information for tracing tumor tissue of origin.

In metastasized cancer, HiTAIC demonstrated 96% accuracy and 98% weighted average *F*1-score across five cancer types with six different metastatic locations (colon to liver, colon to lung, lung to brain, prostate to bone, prostate to liver, prostate to lymph node, breast to lymph node, testis to lung, testis to lymph node) (Table 7). In cfDNA from cancer patients, the model has low performance with 15% accuracy and a weighted *F*1-score of 24% (Table 8). Taken together, HiTAIC traces tumor tissue of origin and cancer subtype with a high accuracy in primary and metastasized cancer but not in cfDNA from cancer patients.

To investigate the biological pathways enriched for the cancer-discerning CpGs, we conducted pathway enrichment analysis using GREAT. We identified significantly enriched GO biological processes, cellular components, and molecular functions for each layer in the hierarchy. In Layer 1, the top 10 enriched biological pathways involve substantially cell differentiation and morphogenesis (Figure 4A). In Layer 2A, which is designed for adenocarcinoma classification, the top 10 enriched biological pathways include majorly inositol phosphate metabolism (Figure 4B). In Layer 2B, which is designed for squamous cell carcinoma classification, the top 10 enriched biological pathways contain mostly lung cell differentiation and development (Figure

Table 6. HiTAIC performance on the external validation data set

Cancer type	Precision	Recall	*F*1-score	Sample size
Adrenocortical adenocarcinoma	1	0.72	0.84	18
Bladder adenocarcinoma	0.78	1	0.88	25
Breast adenocarcinoma	0.96	1	0.98	188
Cervical adenocarcinoma	1	1	1	3
Cervical squamous cell carcinoma	1	0.67	0.8	3
Colorectal adenocarcinoma	1	0.89	0.94	54
Endometrial adenocarcinoma	0.35	1	0.51	9
Eye uveal melanoma	1	0.78	0.88	23
Glioma	0.97	1	0.99	70
Esophageal and head and neck squamous cell carcinoma	0.73	1	0.84	8
Kidney clear cell carcinoma	1	1	1	17
Liver hepatocellular carcinoma	0.99	1	0.99	66
Esophageal and stomach adenocarcinoma	0.92	1	0.96	12
Lung adenocarcinoma	0.96	0.98	0.97	47
Lung squamous cell carcinoma	0.98	0.86	0.92	57
Mesothelioma	1	0.85	0.92	79
Ovarian adenocarcinoma	0.79	0.75	0.77	65
Pancreatic adenocarcinoma	0.8	1	0.89	12
Pheochromocytoma and Paraganglioma	0.85	1	0.92	22
Prostate adenocarcinoma	0.89	1	0.94	25
Sarcoma	0.99	0.91	0.94	158
Skin cutaneous melanoma	0.9	1	0.95	46
Testicular germ cell tumor	0.94	0.93	0.93	130
Thymoma	1	0.82	0.9	11
Thyroid adenocarcinoma	1	1	1	27
Accuracy			0.93	1175
Macro average	0.91	0.93	0.91	1175
weighted average	0.95	0.93	0.93	1175

Table 7. HiTAIC performance on metastasized tumors

Original cancer type	Metastasized location	Precision	Recall	*F*1-score	Sample size
Breast	Lymph node	1	0.98	0.99	44
Colon	Liver, Lung	1	0.93	0.96	29
Lung	Brain	0.94	0.85	0.89	20
Prostate	Bone, Liver, Lymph node	1	0.99	0.99	76
Testicular germ cell tumor	Lung, Lymph node	1	1	1	6
Accuracy				0.96	175
Macro average		0.99	0.95	0.97	175
Weighted average		0.99	0.96	0.98	175

Table 8. HiTAIC performance on cell-free DNA from cancer patients

Cancer type	Precision	Recall	*F*1-score	Sample size
Breast cancer	0.33	1	0.5	3
Colorectal cancer	1	0.75	0.86	4
Liver cancer	1	0.09	0.17	22
Lung cancer	0	0	0	4
Prostate cancer	1	0.13	0.23	233
Accuracy			0.15	266
Macro average	0.67	0.39	0.36	266
Weighted average	0.98	0.15	0.24	266

4C). In Layer 2C, which is designed for melanoma classification, the top 10 enriched biological pathways involve skeletal system development, regionalization process and embryonic morphogenesis (Figure 4D). Multiple genomic context enrichment analyses were conducted to investigate whether the cancer specific CpGs are enriched in certain genomic locations. The Layer 1 and Layer 2A CpGs were significantly enriched for open sea, exon, intron and enhancer regions (Supplementary Figure S4A, Supplementary Figure S4B). The Layer 2B CpGs were significantly over-represented in CpG island and promoter regions (Supplementary Figure S4C). The Layer 2C CpGs were significantly enriched for open sea, intron, and enhancer regions (Supplementary Figure S4D).

DISCUSSION

DNA methylation is well studied to show significant alteration in carcinogenesis, particularly hypermethylation in tumor suppressor genes and hypomethylation in oncogenes (51). Although methylation alteration is a generic phenomenon in cancer, the across-cancer-type heterogeneity of genetic and epigenetic landscape enables distinguishable methylation patterns by tumor type (52). Furthermore, DNA methylation retains tissue and cell identities as it marks cell fate determination, which we hypothesized would enable identification of tumor tissue of origin and tumor type (10). We developed and validated DNA methylation-based machine learning model, HiTAIC, to trace tissue of origin and tumor type in 27 primary tumor types. We also demonstrated its high performance in metastasized cancers.

Previous research has demonstrated the application of machine learning models on genetic and epigenetic data to develop clinical biomarkers for disease diagnosis and prognosis, including cancer (21,53). DNA methylation-based machine learning models have demonstrated the capability to infer the location of unknown primary site from the site of metastasis (24–26). However, previous models heavily favored inference of tissue site instead of tumor type, generating potential problems of indistinguishable tumor subtypes from the same site. Tumor cells originating from the same organ could exhibit widely varying histology and pathogenesis. For example, we observed that esophageal carcinoma can be dichotomized into two clusters by DNA methylation profile, one with head and neck squamous cell carcinoma and the other with stomach adenocarcinoma. Thus, a single-layer classification of 'esophagus cancer' would conflate head and neck squamous cell carcinoma and stomach adenocarcinoma. To address the issue, we used a multilayer hierarchical approach with differential DNA patterns by tumor type to achieve high-resolution tumor tissue of origin and subtype tracing. Our previous publication using DNA methylation data with hierarchical modeling achieved high-resolution deconvolution of the tumor microenvironment (14). In HiTAIC, adenocarcinoma and squamous cell carcinoma were initially distinguished, eliminating the potential issue of confusing esophageal adenocarcinoma with esophageal squamous cell carcinoma. Furthermore, previous work has limited cancer types and validation data sets, especially for metastasized cancers and cell-free DNA from

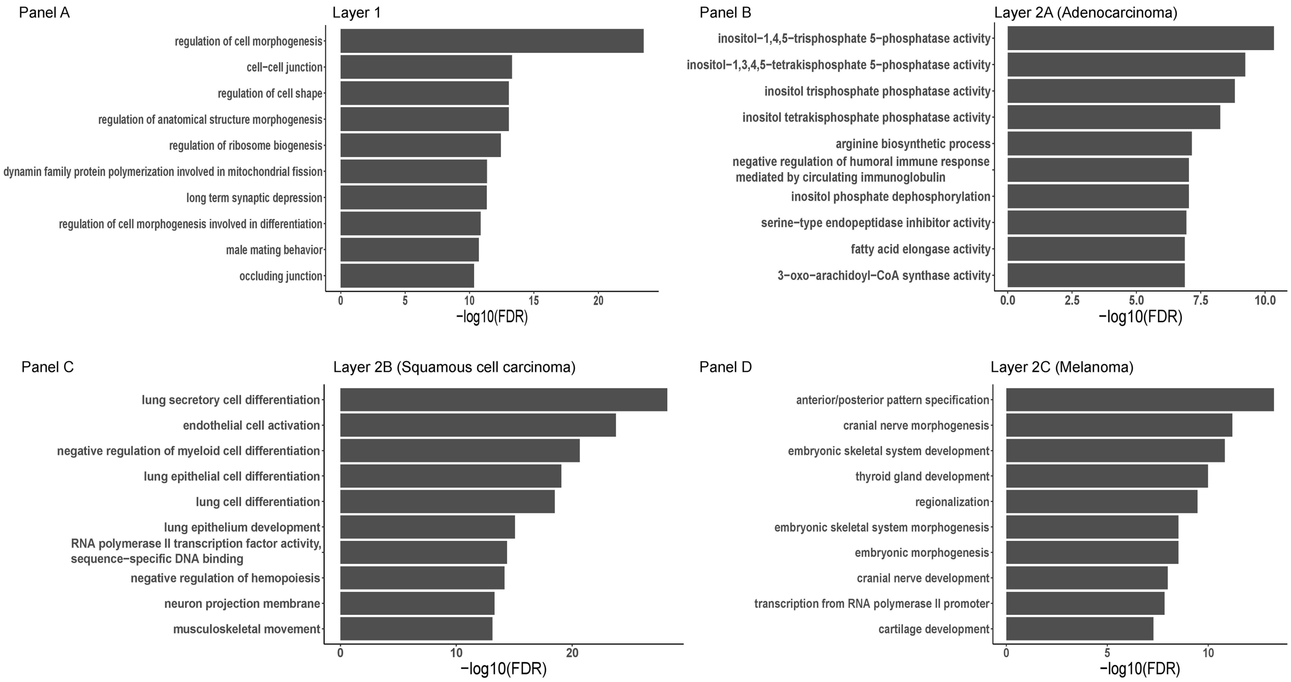


Figure 4. GREAT Gene Ontology pathways enriched for CpGs discerning: Panel **A:** major cancer types; Panel **B:** adenocarcinoma subtypes; Panel **C:** squamous cell carcinoma subtypes; Panel **D**: melanoma subtypes.

cancer patients. Noticeably, previous work did not provide a user-friendly and accessible tool for the generic scientific community. HiTAIC, on the other hand, included 27 cancer types from 23 tissue sites in the discovery data sets, validated using external data sets, and demonstrated utility in external metastasized cancers. For easy accessibility of the algorithm and accommodate users without coding experience, a web-based app is developed and now available through https://sites.dartmouth.edu/salaslabhitaic/. Our effort emphasizes the computational biology research translatability from machine to the scientific and clinical community.

Importantly, HiTAIC traces tissue of origin and tumor type in metastasized cancers with a high accuracy, providing insight into the identification of CUP. Previous research developed a methylation-based assay for tracing CUP, EPICUP, showed significantly better prognosis for CUP cases that were treated with site-specific therapy based on the assay compared to the CUP cases treated with empiric therapy, emphasizing the clinical significance of CUP diagnosis and directing the future of CUP diagnosis towards precision medicine (9). Although not yet clinically deployable as EPICUP, the hierarchical modeling approach employed by HiTAIC maximizes the power of detecting most differentially methylated CpGs in granular tumor subtypes and is a more research friendly bioinformatic tool that can serve the research community and facilitate research which involves inferring the primary tumor site of unknown origin.

The predictability of the HiTAIC model is based on four libraries of CpGs discerning different layers of cancer subtypes. Interestingly, the four CpG libraries are almost completely distinct from each other with very low overlap. The biological pathways and genomic context enriched across the libraries are also diverse, indicating differential pathways involved in carcinogenesis and morphogenesis for different organ sites. In Layer 1, top enriched pathways are associated with cell differentiation, morphogenesis, and function. We posit that the distinguishable epigenetic regulation in cell differentiation resulted from the substantially distinctive cancer types and tumor sites in Layer 1. In Layer 2A, which is designed for adenocarcinoma classification, inositol phosphate metabolism related pathways were highly enriched. Inositol phosphate metabolic pathways are crucial for regulating cell migration, proliferation, apoptosis, and phosphatidylinositol-3-kinase (PI3K)/Akt signaling under normal physiological conditions. Studies have shown that dysregulation in inositol phosphate metabolism plays a key role in carcinogenesis. Tan et al. demonstrated gene variants in the inositol phosphate metabolism pathways are associated with risk of four types of cancer, including lung, esophageal, stomach, and kidney(54). Two studies showed the association between inositol phosphate metabolism and cancer aggressiveness in both human and mouse models (55,56). The biological implications of distinguishability for adenocarcinoma by differential epigenetic patterns regulating inositol phosphate metabolic pathways is intriguing and promises further investigation. Top pathways enriched for squamous cell carcinoma classification in Layer 2B involve lung cell differentiation, emphasizing tumor originating site specification to discern squamous cell carcinoma. To distinguish between cutaneous and uveal melanoma, top pathways were enriched for embryonic development. Differential biological pathways identified for distinguishing different cancer subtypes promote further investigation of differential epigenetic regulation in carcinogenesis and morphogenesis by tumor subtype and tumor originating site.

While our study developed a solid pipeline and user-friendly algorithm, HiTAIC, for cancer tracing, we recognize some limitations. First, esophageal carcinoma cannot be discerned by its own category. The esophageal squamous cell carcinoma was collapsed into head and neck squamous cell carcinoma whereas the esophageal adenocarcinoma was collapsed into stomach adenocarcinoma. Previous studies revealed disparities in the classification of esophageal carcinoma with stomach adenocarcinoma and head and neck squamous cell carcinoma (57,58). Although esophageal cancer cannot be distinguished from stomach or head and neck cancer, HiTAIC still provides a general location of the tumor of interest, which is clinically important. While previous tools treated esophageal carcinoma as an individual cancer type, our study revealed dichotomized methylation patterns of esophageal carcinoma subtypes. HiTAIC addressed the issue by employing a hierarchical modeling approach. We believe this is critical to enhance the accuracy of the model to trace esophageal carcinoma subtypes. Second, for external validation, kidney papillary and chromophobe cancers were not included as no other resources were available for those tumors other than TCGA. However, in the test data set, HiTAIC demonstrated high accuracy in predicting those two kidney tumor subtypes. Third, due to the limited data source, the performance of HiTAIC in metastasized tumor and cell free DNA were evaluated only in five cancer types respectively. Future analyses are necessary on more metastasized cancer type and cell free DNA. Finally, HiTAIC does not work well in tracing cancer in cfDNA based on the data analyzed. As HiTAIC was developed based on differential DNA methylome in solid tumor tissues, it is not optimized to work in a noisy cfDNA environment mixed with blood DNA signals and low abundance of tumor DNA. Therefore, we emphasize the importance of developing environment-specific tools, e.g. in body fluids like peripheral blood serum, menstrual blood, human milk, cerebrospinal fluid, for tracing tumor DNA to correct for the background noise.

HiTAIC provides a comprehensive tracing of major tumor sites and types. Future studies should focus on the development of tumor-specific subtype libraries to expand the classification hierarchy as molecular and anatomic subtypes of certain tumors have been showed to be differentially regulated by DNA methylation (59,60). Although HiTAIC can provide information on tissue of origin, we do not encourage the use of HiTAIC to non-tumor tissues as that deviates the purpose of HiTAIC. Future research on the development of DNA methylation-based normal tissue classifier is necessary.

CONCLUSION

We developed HiTAIC, a DNA methylation-based multilayer perceptron classifier, to trace tissue of origin and tumor type in primary and metastasized tumors. The capability of the model tracing the tumor origin and subtype with high resolution and accuracy promises potential clinical use in identifying cancer of unknown origin and thus strategizing treatment plan to achieve precision medicine. HiTAIC can be easily deployed in a web-based application, which transformed the computational sophistication to a more user-friendly tool for public.

DATA AVAILABILITY

All data sets used in this study are publicly available on The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), and ArrayExpress. The accession numbers are GSE133556, GSE77871, GSE75067, GSE169622, GSE193535, GSE123678, GSE67114, GSE54503, E-MTAB-10156, GSE74071, GSE52955, GSE140169, GSE164988, GSE121377, GSE156876, GSE164269, GSE167059, GSE43293, GSE65820, GSE74104, GSE94769, GSE93589, GSE53051, GSE77954, GSE116699, E-MTAB-8660, GSE174613, GSE116338, GSE58999, GSE156512, GSE122126, GSE129374, GSE108462, GSE119260, GSE157273. HiTAIC is publicly accessible on a user-friendly web page https://sites.dartmouth.edu/salaslabhitaic/. Code and documents are shared on FigShare (DOI: 10.6084/m9.figshare.22179089).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Cancer Online.

ACKNOWLEDGEMENTS

Author contributions: Conceptualization, Z.Z., L.A.S., B.C.C., S.V.; Methodology, Z.Z., L.A.S., B.C.C., S.V., J.L.; Software, Z.Z., J.L., Y.L.; Validation, Z.Z.; Formal Analysis, Z.Z..; Investigation, Z.Z., L.A.S.; Resources, Z.Z.; Data Curation, Z.Z.; Writing – Original Draft Preparation, Z.Z., L.A.S.; Writing – Review & Editing, Z.Z., L.A.S., B.C.C., S.V., J.L., Y.L.; Visualization, Z.Z., L.A.S., Y.L., J.L.; Supervision, L.A.S., B.C.C., J.L.; Project Administration, Z.Z., L.A.S.; Funding Acquisition, L.A.S., B.C.C.

FUNDING

Department of Defense [W81XWH-20-1-0778]; National Institute of General Medical Sciences [P20GM104416/8299]; National Cancer Institute [R01CA253976, R01CA216265, R01CA275974, P30CA023108].

Conflict of interest statement. HiTAIC has a pending patent application through Dartmouth College Technology Transfer Office with the invention disclosure number 2023–021. The inventors listed on the patent application are Ze Zhang, Lucas A Salas and Brock C Christensen. This tool is provided ‘as-is’ and is intended FOR RESEARCH USE ONLY, and it must not be used for any clinical purposes, including diagnosis or treatment of diseases. The authors are not responsible for the accuracy of analysis results.

REFERENCES

1. Siegel,R.L., Miller,K.D., Fuchs,H.E. and Jemal,A. (2022) Cancer statistics, 2022. *CA Cancer J. Clin.*, **72**, 7–33.
2. Guan,X. (2015) Cancer metastases: challenges and opportunities. *Acta Pharm. Sin. B*, **5**, 402–418.

3. Fares,J., Fares,M.Y., Khachfe,H.H., Salhab,H.A. and Fares,Y. (2020) Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduct. Target. Ther.*, **5**, 28.
4. Kolling,S., Ventre,F., Geuna,E., Milan,M., Pisacane,A., Boccaccio,C., Sapino,A. and Montemurro,F. (2019) “Metastatic cancer of unknown primary” or “primary Metastatic cancer”?*Front. Oncol.*, **9**, 1546.
5. Qaseem,A., Usman,N., Jayaraj,J.S., Janapala,R.N. and Kashif,T. (2019) Cancer of unknown primary: a review on clinical guidelines in the development and targeted management of patients with the unknown primary site. *Cureus*, **11**, e5552.
6. Massard,C., Loriot,Y. and Fizazi,K. (2011) Carcinomas of an unknown primary origin–diagnosis and treatment. *Nat. Rev. Clin. Oncol.*, **8**, 701–710.
7. Varadhachary,G.R. and Raber,M.N. (2014) Cancer of unknown primary site. *N. Engl. J. Med.*, **371**, 757–765.
8. Rassy,E. and Pavlidis,N. (2020) Progress in refining the clinical management of cancer of unknown primary in the molecular era. *Nat. Rev. Clin. Oncol.*, **17**, 541–554.
9. Niazi,M.K.K., Parwani,A.V. and Gurcan,M.N. (2019) Digital pathology and artificial intelligence. *Lancet Oncol.*, **20**, e253–e261.
10. Bogdanovic,O. and Lister,R. (2017) DNA methylation and the preservation of cell identity. *Curr. Opin. Genet. Dev.*, **46**, 9–14.
11. Salas,L.A., Zhang,Z., Koestler,D.C., Butler,R.A., Hansen,H.M., Molinaro,A.M., Wiencke,J.K., Kelsey,K.T. and Christensen,B.C. (2022) Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat. Commun.*, **13**, 761.
12. Arneson,D., Yang,X. and Wang,K. (2020) MethylResolver-a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. *Commun. Biol.*, **3**, 422.
13. Chakravarthy,A., Furness,A., Joshi,K., Ghorani,E., Ford,K., Ward,M.J., King,E.V., Lechner,M., Marafioti,T., Quezada,S.A. *et al.* (2018) Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.*, **9**, 3220.
14. Zhang,Z., Wiencke,J.K., Kelsey,K.T., Koestler,D.C., Christensen,B.C. and Salas,L.A. (2022) HiTIMED: hierarchical tumor immune microenvironment epigenetic deconvolution for accurate cell type resolution in the tumor microenvironment using tumor-type-specific DNA methylation data. *J. Transl. Med.*, **20**, 516.
15. Chen,X., Gole,J., Gore,A., He,Q., Lu,M., Min,J., Yuan,Z., Yang,X., Jiang,Y., Zhang,T. *et al.* (2020) Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat. Commun.*, **11**, 3475.
16. Li,W. and Zhou,X.J. (2020) Methylation extends the reach of liquid biopsy in cancer detection. *Nat. Rev. Clin. Oncol.*, **17**, 655–656.
17. Schumacher,A., Kapranov,P., Kaminsky,Z., Flanagan,J., Assadzadeh,A., Yau,P., Virtanen,C., Winegarden,N., Cheng,J., Gingeras,T. *et al.* (2006) Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res.*, **34**, 528–542.
18. Noorbakhsh-Sabet,N., Zand,R., Zhang,Y. and Abedi,V. (2019) Artificial intelligence transforms the future of health care. *Am. J. Med.*, **132**, 795–801.
19. Cui,M. and Zhang,D.Y. (2021) Artificial intelligence and computational pathology. *Lab. Invest.*, **101**, 412–422.
20. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Zidek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
21. Huang,S., Cai,N., Pacheco,P.P., Narrandes,S., Wang,Y. and Xu,W. (2018) Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics*, **15**, 41–51.
22. Poirion,O.B., Jing,Z., Chaudhary,K., Huang,S. and Garmire,L.X. (2021) DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med.*, **13**, 112.
23. Wang,T., Shao,W., Huang,Z., Tang,H., Zhang,J., Ding,Z. and Huang,K. (2021) MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.*, **12**, 3445.
24. Zheng,C. and Xu,R. (2020) Predicting cancer origins with a DNA methylation-based deep neural network model. *PLoS One*, **15**, e0226461.
25. Modhukur,V., Sharma,S., Mondal,M., Lawarde,A., Kask,K., Sharma,R. and Salumets,A. (2021) Machine learning approaches to classify primary and metastatic cancers using tissue of origin-based DNA methylation profiles. *Cancers (Basel)*, **13**, 3768.
26. Moran,S., Martinez-Cardus,A., Sayols,S., Musulen,E., Balana,C., Estival-Gonzalez,A., Moutinho,C., Heyn,H., Diaz-Lagares,A., de Moura,M.C. *et al.* (2016) Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol.*, **17**, 1386–1395.
27. Zhou,W., Triche,T.J. Jr, Laird,P.W. and Shen,H. (2018) SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.*, **46**, e123.
28. Min,J.L., Hemani,G., Davey Smith,G., Relton,C. and Suderman,M. (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, **34**, 3983–3989.
29. Fabian Pedregosa,G.V., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., Vanderplas,J. *et al.* (2011) Scikit-learn: machine learning in Python. *JMLR*, **12**, 2825−2830.
30. Legendre,C.R., Demeure,M.J., Whitsett,T.G., Gooden,G.C., Bussey,K.J., Jung,S., Waibhav,T., Kim,S. and Salhia,B. (2016) Pathway implications of aberrant global methylation in adrenocortical cancer. *PLoS One*, **11**, e0150629.
31. Ramalho-Carvalho,J., Graca,I., Gomez,A., Oliveira,J., Henrique,R., Esteller,M. and Jeronimo,C. (2017) Downregulation of miR-130b∼301b cluster is mediated by aberrant promoter methylation and impairs cellular senescence in prostate cancer. *J. Hematol. Oncol.*, **10**, 43.
32. Oltra,S.S., Pena-Chilet,M., Vidal-Tomas,V., Flower,K., Martinez,M.T., Alonso,E., Burgues,O., Lluch,A., Flanagan,J.M. and Ribas,G. (2018) Methylation deregulation of miRNA promoters identifies miR124-2 as a survival biomarker in breast cancer in very young women. *Sci. Rep.*, **8**, 14373.
33. Baharudin,R., Ishak,M., Muhamad Yusof,A., Saidin,S., Syafruddin,S.E., Wan Mohamad Nazarie,W.F., Lee,L.H. and Ab Mutalib,N.S. (2022) Epigenome-wide DNA methylation profiling in colorectal cancer and normal adjacent colon using infinium human methylation 450K. *Diagnostics (Basel)*, **12**, 198.
34. Court,F., Le Boiteux,E., Fogli,A., Muller-Barthelemy,M., Vaurs-Barriere,C., Chautard,E., Pereira,B., Biau,J., Kemeny,J.L., Khalil,T. *et al.* (2019) Transcriptional alterations in glioma result primarily from DNA methylation-independent mechanisms. *Genome Res.*, **29**, 1605–1621.
35. Worsham,M.J., Chen,K.M., Datta,I., Stephen,J.K., Chitale,D., Gothard,A. and Divine,G. (2016) The biological significance of methylome differences in human papilloma virus associated head and neck cancer. *Oncol. Lett.*, **12**, 4949–4956.
36. Ramalho-Carvalho,J., Goncalves,C.S., Graca,I., Bidarra,D., Pereira-Silva,E., Salta,S., Godinho,M.I., Gomez,A., Esteller,M., Costa,B.M. *et al.* (2018) A multiplatform approach identifies miR-152-3p as a common epigenetically regulated onco-suppressor in prostate cancer targeting TMEM97. *Clin Epigenetics*, **10**, 40.
37. Shen,J., Wang,S., Zhang,Y.J., Wu,H.C., Kibriya,M.G., Jasmine,F., Ahsan,H., Wu,D.P., Siegel,A.B., Remotti,H. *et al.* (2013) Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics*, **8**, 34–43.
38. Mirhadi,S., Tam,S., Li,Q., Moghal,N., Pham,N.A., Tong,J., Golbourn,B.J., Krieger,J.R., Taylor,P., Li,M. *et al.* (2022) Integrative analysis of non-small cell lung cancer patient-derived xenografts identifies distinct proteotypes associated with patient outcomes. *Nat. Commun.*, **13**, 1811.
39. Yamamoto,Y., Matsusaka,K., Fukuyo,M., Rahmutulla,B., Matsue,H. and Kaneda,A. (2020) Higher methylation subtype of malignant melanoma and its correlation with thicker progression and worse prognosis. *Cancer Med.*, **9**, 7194–7204.
40. Park,J.L., Jeon,S., Seo,E.H., Bae,D.H., Jeong,Y.M., Kim,Y., Bae,J.S., Kim,S.K., Jung,C.K. and Kim,Y.S. (2020) Comprehensive DNA methylation profiling identifies novel diagnostic biomarkers for thyroid cancer. *Thyroid*, **30**, 192–203.
41. Trimarchi,M.P., Yan,P., Groden,J., Bundschuh,R. and Goodfellow,P.J. (2017) Identification of endometrial cancer methylation features using combined methylation analysis methods. *PLoS One*, **12**, e0173242.

42. Timp,W., Bravo,H.C., McDonald,O.G., Goggins,M., Umbricht,C., Zeiger,M., Feinberg,A.P. and Irizarry,R.A. (2014) Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med.*, **6**, 61.
43. Qu,X., Sandmann,T., Frierson,H. Jr, Fu,L., Fuentes,E., Walter,K., Okrah,K., Rumpel,C., Moskaluk,C., Lu,S. *et al.* (2016) Integrated genomic analysis of colorectal cancer progression reveals activation of EGFR through demethylation of the EREG promoter. *Oncogene*, **35**, 6403–6415.
44. Jurmeister,P., Scholer,A., Arnold,A., Klauschen,F., Lenze,D., Hummel,M., Schweizer,L., Blaker,H., Pfitzner,B.M., Mamlouk,S. *et al.* (2019) DNA methylation profiling reliably distinguishes pulmonary enteric adenocarcinoma from metastatic colorectal cancer. *Mod. Pathol.*, **32**, 855–865.
45. Ylitalo,E.B., Thysell,E., Landfors,M., Brattsand,M., Jernberg,E., Crnalic,S., Widmark,A., Hultdin,M., Bergh,A., Degerman,S. *et al.* (2021) A novel DNA methylation signature is associated with androgen receptor activity and patient prognosis in bone metastatic prostate cancer. *Clin Epigenetics*, **13**, 133.
46. Moss,J., Magenheim,J., Neiman,D., Zemmour,H., Loyfer,N., Korach,A., Samet,Y., Maoz,M., Druid,H., Arner,P. *et al.* (2018) Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.*, **9**, 5068.
47. Hlady,R.A., Zhao,X., Pan,X., Yang,J.D., Ahmed,F., Antwi,S.O., Giama,N.H., Patel,T., Roberts,L.R., Liu,C. *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.
48. Gordevicius,J., Krisciunas,A., Groot,D.E., Yip,S.M., Susic,M., Kwan,A., Kustra,R., Joshua,A.M., Chi,K.N., Petronis,A. *et al.* (2018) Cell-free DNA modification dynamics in abiraterone acetate-treated prostate cancer patients. *Clin. Cancer Res.*, **24**, 3317–3324.
49. Silva,R., Moran,B., Russell,N.M., Fahey,C., Vlajnic,T., Manecksha,R.P., Finn,S.P., Brennan,D.J., Gallagher,W.M. and Perry,A.S. (2020) Evaluating liquid biopsies for methylomic profiling of prostate cancer. *Epigenetics*, **15**, 715–727.
50. Silva,R., Moran,B., Baird,A.M., O'Rourke,C.J., Finn,S.P., McDermott,R., Watson,W., Gallagher,W.M., Brennan,D.J. and Perry,A.S. (2021) Longitudinal analysis of individual cfDNA methylome patterns in metastatic prostate cancer. *Clin Epigenetics*, **13**, 168.
51. Ehrlich,M. (2009) DNA hypomethylation in cancer cells. *Epigenomics*, **1**, 239–259.
52. Zhang,C., Zhao,H., Li,J., Liu,H., Wang,F., Wei,Y., Su,J., Zhang,D., Liu,T. and Zhang,Y. (2015) The identification of specific methylation patterns across different cancers. *PLoS One*, **10**, e0120361.
53. Sammut,S.J., Crispin-Ortuzar,M., Chin,S.F., Provenzano,E., Bardwell,H.A., Ma,W., Cope,W., Dariush,A., Dawson,S.J., Abraham,J.E. *et al.* (2022) Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, **601**, 623–629.
54. Tan,J., Yu,C.Y., Wang,Z.H., Chen,H.Y., Guan,J., Chen,Y.X. and Fang,J.Y. (2015) Genetic variants in the inositol phosphate metabolism pathway and risk of different types of cancer. *Sci. Rep.*, **5**, 8473.
55. Benjamin,D.I., Louie,S.M., Mulvihill,M.M., Kohnz,R.A., Li,D.S., Chan,L.G., Sorrentino,A., Bandyopadhyay,S., Cozzo,A., Ohiri,A. *et al.* (2014) Inositol phosphate recycling regulates glycolytic and lipid metabolism that drives cancer aggressiveness. *ACS Chem. Biol.*, **9**, 1340–1350.
56. Rao,F., Xu,J., Fu,C., Cha,J.Y., Gadalla,M.M., Xu,R., Barrow,J.C. and Snyder,S.H. (2015) Inositol pyrophosphates promote tumor growth and metastasis by antagonizing liver kinase B1. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 1773–1778.
57. Lindblad,M., Ye,W., Lindgren,A. and Lagergren,J. (2006) Disparities in the classification of esophageal and cardia adenocarcinomas and their influence on reported incidence rates. *Ann. Surg.*, **243**, 479–485.
58. Businello,G., Fassan,M., Degasperi,S., Traverso,G., Scarpa,M., Angriman,I., Kotsafti,A., Castagliuolo,I., Sbaraglia,M., Bardini,R. *et al.* (2021) Esophageal squamous cell carcinoma metachronous to head and neck cancers. *Pathol. Res. Pract.*, **219**, 153346.
59. Holm,K., Hegardt,C., Staaf,J., Vallon-Christersson,J., Jonsson,G., Olsson,H., Borg,A. and Ringner,M. (2010) Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Res.*, **12**, R36.
60. Nakagawa,T., Kurokawa,T., Mima,M., Imamoto,S., Mizokami,H., Kondo,S., Okamoto,Y., Misawa,K., Hanazawa,T. and Kaneda,A. (2021) DNA methylation and HPV-associated head and neck cancer. *Microorganisms*, **9**, 801.