


RESEARCH

Open Access



Tissue of origin detection for cancer tumor using low-depth cfDNA samples through combination of tumor-specific methylation atlas and genome-wide methylation density in graph convolutional neural networks

Trong Hieu Nguyen^{1*}, Nhu Nhat Tan Doan^{1†}, Trung Hieu Tran¹, Le Anh Khoa Huynh^{1,2}, Phuoc Loc Doan¹, Thi Hue Hanh Nguyen¹, Van Thien Chi Nguyen¹, Giang Thi Huong Nguyen¹, Hoai-Nghia Nguyen¹, Hoa Giang¹, Le Son Tran¹ and Minh Duy Phan^{1*}

Abstract

Background Cell free DNA (cfDNA)-based assays hold great potential in detecting early cancer signals yet determining the tissue-of-origin (TOO) for cancer signals remains a challenging task. Here, we investigated the contribution of a methylation atlas to TOO detection in low depth cfDNA samples.

Methods We constructed a tumor-specific methylation atlas (TSMA) using whole-genome bisulfite sequencing (WGBS) data from five types of tumor tissues (breast, colorectal, gastric, liver and lung cancer) and paired white blood cells (WBC). TSMA was used with a non-negative least square matrix factorization (NNLS) deconvolution algorithm to identify the abundance of tumor tissue types in a WGBS sample. We showed that TSMA worked well with tumor tissue but struggled with cfDNA samples due to the overwhelming amount of WBC-derived DNA. To construct a model for TOO, we adopted the multi-modal strategy and used as inputs the combination of deconvolution scores from TSMA with other features of cfDNA.

Results Our final model comprised of a graph convolutional neural network using deconvolution scores and genome-wide methylation density features, which achieved an accuracy of 69% in a held-out validation dataset of 239 low-depth cfDNA samples.

[†]Trong Hieu Nguyen and Nhu Nhat Tan Doan contributed equally to this study.

*Correspondence:
Trong Hieu Nguyen
hieunguyen@genesolutions.vn
Minh Duy Phan
pmduy@yahoo.com

Full list of author information is available at the end of the article



Conclusions In conclusion, we have demonstrated that our TSMA in combination with other cfDNA features can improve TOO detection in low-depth cfDNA samples.

Keywords Tissue of origin, cfDNA, Tumor-specific methylation atlas, Genome-wide methylation density, Graph convolutional neural networks

Background

Liquid biopsies based on cell free DNA (cfDNA) have recently emerged as a novel method for early cancer detection owing to their non-invasive, sensitive, and multi-modal characteristics. Multiple features can be derived from cfDNA sequences to reveal various aspects of cancer-specific aberration including fragment length profile [1–3], copy number aberration [4], motif-end [5], genome-wide and targeted methylation profiles [6]. To maximize the capacity to distinguish between cancer patients and healthy individuals, the integration of multiple features in advanced machine learning or deep learning has become a common approach and demonstrated encouraging performance [7–12]. However, predicting tumor of origin (TOO) for multiple cancer types at early stage remains challenging due to the low abundance of cfDNA originating from tumors (ctDNA), further confounded by the presence of various DNA components released from non-tumor sources.

Recent advances have highlighted the increasing importance of DNA methylation in the early cancer detection context [13–15]. DNA methylation is an epigenetic marker that plays a crucial role in regulating gene expression, maintaining genomic stability, and directing cell development. These epigenetic modifications contribute heavily to the dysregulation of multiple pathways, allowing cancer cells to proliferate uncontrollably [16–19]. Most importantly, these DNA methylation patterns are tissue-specific and remain stable during neoplastic transformation, which could allow tumor of origin identification [20–22]. Therefore, characterization of DNA methylation biomarkers in tumor tissues could potentially guide the detection of both cancer characteristics and TOO in cfDNA samples.

Constructing a methylation atlas is an attractive approach in analyzing large methylation data. Moss et al. [23] developed a comprehensive methylation atlas of healthy human cell types using data from 450 K Illumina microarray to decompose of human cell types in bulk samples. Loyer et al. [24] created a human methylation atlas using deep whole-genome bisulfite sequencing (WGBS) from 39 normal cell types of 205 healthy tissue samples to further enhance the understanding of the human normal cell types methylome. The methylation atlas was constructed based on differential fractions of hypo-methylated and hyper-methylated reads at various cell-type specific genome segments. While the authors reported a promising deconvolution resolution of 0.1%,

the study did not demonstrate the application of this methylation atlas to determine TOO of cfDNA samples from cancer patients.

In this study, we investigated the application of methylation atlas to cell type deconvolution of cfDNA samples, with the ultimate goal of creating a model that can detect TOO in low-depth cfDNA samples for early cancer screening (Fig. 1). In this project, we focused on the five most prevalent cancers in Vietnam [25]. We first constructed a tumor-specific methylation atlas (TSMA) using WGBS data from five types of tumor tissues (breast, colorectal, gastric, liver and lung cancer) and paired white blood cells (WBC). We then validated the use of TSMA in deconvolution of tumor tissue samples and cfDNA samples, demonstrated that TSMA worked well for tumor tissues but struggled with cfDNA samples due to the overwhelming amount of DNA fragments derived from WBC. To construct a model for TOO in low-depth cfDNA samples, we adopted the multi-modal strategy and combined deconvolution scores from our TSMA with other features previously explored by our group [9]. Our final model comprised of a graph convolutional neural network (GCNN) using deconvolution scores and genome-wide methylation density features (GWMD) as inputs and achieved an accuracy of 69% in a held-out validation dataset of 239 samples.

Methods

Dataset description

Atlas construction dataset This dataset includes 64 tumor tissue samples from five distinct cancer classes (Liver (11 samples), Breast (24 samples), Lung (11 samples), CRC (11 samples), Gastric (11 samples)) and 24 WBC samples. All samples were whole-genome bisulfite sequenced at 5-15x depth coverage. Metadata is provided in Supplementary Table 2.

Validations were conducted on four different datasets:

- **Dataset 1:** 450 K/850 methylation microarray datasets were downloaded from TCGA. The curated dataset only includes samples from our five group of tissues of interest (4,415 samples), which consists of 886 CRC samples, 1,814 Lung samples, 398 Gastric samples, 888 Breast samples and 429 Liver samples from cancer patients.
- **Dataset 2:** We generated an *in-silico* spike-in dataset using three healthy WGBS cfDNA samples as background. To simulate the presence of circulating

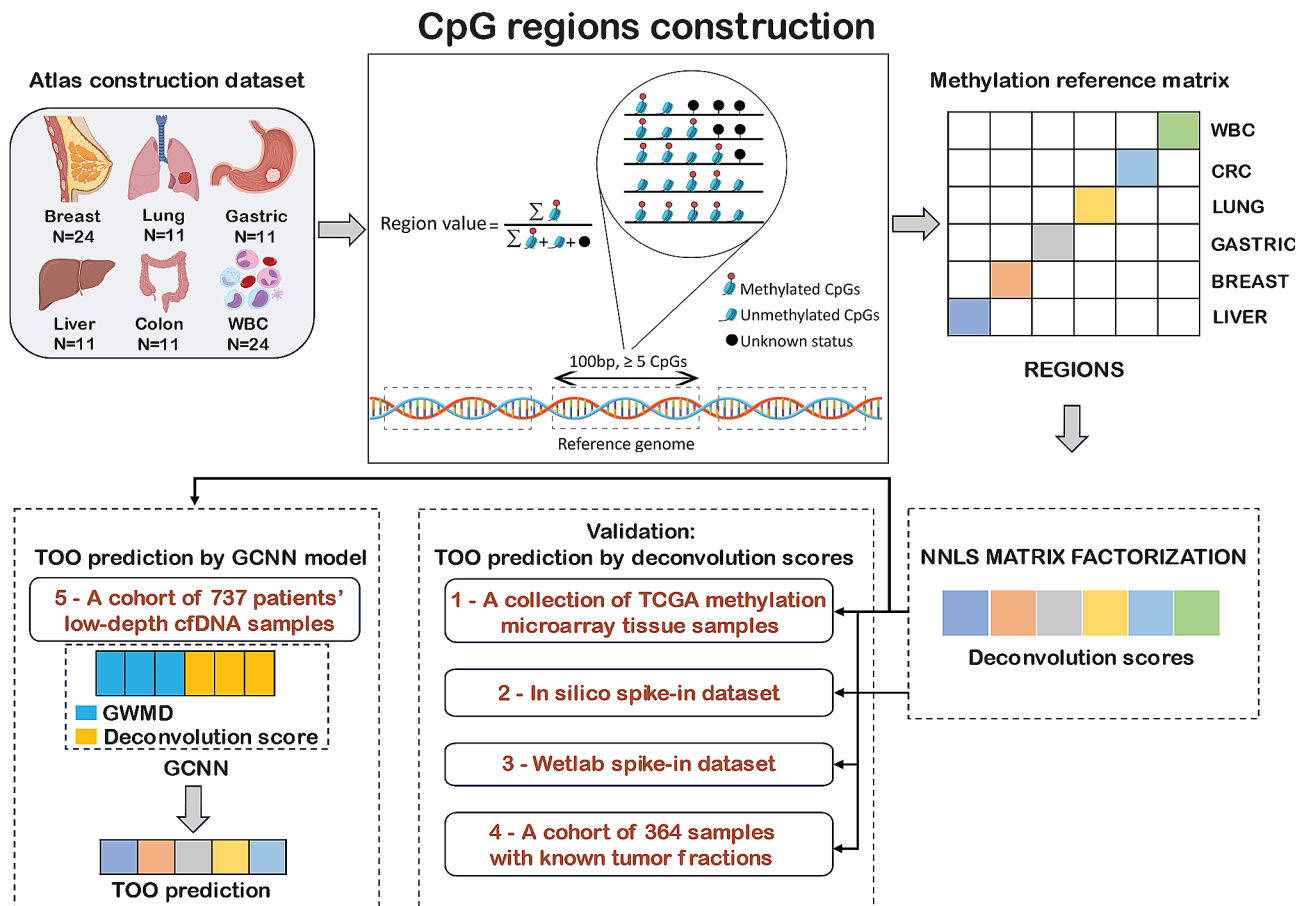


Fig. 1 Schematic overview of this study. We first constructed a tumor-specific methylation atlas (TSMA) using a WGBS dataset (Atlas construction dataset) of 64 tumor tissues comprising five cancer types (breast, colorectal, gastric, liver and lung cancer) and paired white blood cells (WBC). The methylation signals from WGBS data (region value) were calculated for approximately 1.1 million pre-defined CpG regions (regions of 100 bp in length covering at least 5 CpG sites). This large matrix of region values was then used to construct the TSMA, comprising of 2,945 differential regions between five tumor-tissue types and WBC across the entire genome. With the TSMA, deconvolution scores for new input samples were calculated using a non-negative least square (NNLS) matrix factorization. We next validated the use of TSMA and deconvolution scores in 3 datasets, including Dataset 1: tumor tissue methylation microarray data from TCGA, Dataset 2: in silico spike-in samples with known amount of tissue DNA fragments and Dataset 3: wet lab spike-in samples with known amount of tissue DNA fragments. Finally, we implemented a graph convolutional neural network combining deconvolution scores and genome-wide methylation density (GWMD). The model was trained and validated on a cohort of 737 low-depth WGBS cfDNA samples (Dataset 4).

cell-free tumor-derived DNA (ctDNA) at varying abundance levels, spike-in reads were randomly extracted from pooled tumor tissue samples in the Atlas dataset, representing ratios of 0.01%, 0.05%, 0.1%, 1%, 10%, and 25%. These simulated reads were then merged with the background cfDNA samples, generating 3 samples for each cancer type at each spike-in ratio (3 samples x 5 cancer types x 6 ratios = 90 samples). The process was repeated three times, generating in total a dataset of 270 samples. This simulation dataset allows us to assess the sensitivity of our constructed atlas in detecting tumor-related signals within the cfDNA background at different levels of abundance.

- **Dataset 3:** We refer to section “Wet-lab spike-in experiments in Dataset 3” in Materials and methods for a detailed description of this dataset.

- **Dataset 4:** A cohort of 737 low-depth WGBS cfDNA samples (0.5x). The low-depth dataset was previously employed in the construction and validation of an integrated multi-modal model for early cancer detection [9]. 498 samples are used in the training set and 239 samples are served as a held-out validation set.

Metadata tables for all datasets are provided in Supplementary Table 3.

Wet-lab spike-in experiments in dataset 3

This section is devoted to the preparation of Dataset 3. The spike-in experiment was conducted using healthy control and cancerous tissue samples. The healthy control sample was created by pooling cfDNA of multiple healthy, non-cancerous individuals. The extracted cancer

gDNA was subjected to a fragmentation process to create fragmented cancer DNA. These cancer DNA fragments were then spiked in the healthy control sample with four different amounts to create a set of four samples containing different tumor abundances (0.1%, 1%, 10% and 25% of cancer DNA in total amount of DNA). This experiment was repeated twice for each cancer type with extracted gDNA from two different cancer samples, resulting in a total of 40 samples (10 cancer samples in total – two samples for each type of cancer including breast, colorectal, liver, lung and gastric cancer). These spike-in samples then undergone bisulfite conversion and purification by EZ DNA Methylation-Gold Kit (Zymo Research, D5006, USA). DNA library was prepared from bisulfite-converted DNA samples using xGen Methyl-Seq DNA Library Prep Kit (Integrated DNA Technologies, 10,009,824, USA) with Adaptase technology, according to the manufacturer's instructions. All the DNA concentrations were identified by the QuantiFluor dsDNA system (Promega, USA). The library products were sequenced on the DNBSEQ-T7 DNA system (MGI Tech, Shenzhen, China).

Bioinformatics pipeline

FASTQ files were fed to an in-house Bioinformatics pipeline. We performed FASTQ file quality control with fastqc (version 0.11.2 [26]) and trimming adapters and low-quality bases by TrimGalore (version 0.6.7, [27]). We only trimmed the first 15 bp --HEADCROP at the 5' end of Read 1, since it is not possible to exactly locate the 5' end of read 2 in this assay. Sequence alignment and CpG site methylation calling were done by the Bismark suite (version 0.23.1, [28]). We used picard MarkDuplicate (version 2.18.7 [29]) to mark and remove duplicated reads. Reads were then filtered by samtools (version 1.18, [30]) to keep only reads whose quality is greater than Q30. All tools' parameters were kept default unless mentioned here. Other processing steps and data analysis were done by our in-house Python and R scripts.

Data downloaded from TCGA databases were in tab-separated table format and contained processed methylation density at each CpG site. We selected CpG sites that were available in TSMA's regions and discarded the rest. In summary, we retained around 1088 regions on average. Regions' values were calculated by the average methylation density values of their CpG sites. With this transformation, our TSMA is technically interchangeable between the sequencing platform and the microarray platform.

Multi-modal cfDNA feature sets

We adopted a multi-modal set of genome-wide and targeted features from [9] to combine with our deconvolution scores derived from the TSMA. This set included

genome-wide methylation density (GWMD), targeted region methylation density (TMD), genome-wide fragmentation profile (GWFP and Flen), end-motif distribution (EM) and copy number aberration (CNA). We refer to [9] for detailed constructions of these features.

Construction of the methylation reference matrix (atlas)

A CpG region is defined as a region of 100 bp in length covering at least 5 CpG sites. Reads covering at least one CpG site in a region were collected. Since we were interested in the methylation pattern of CpG sites within the region only, information on CpG sites outside of the region carried by the reads was discarded. For a given region j , let us denote by N_j the total number of times a CpG site was covered by a read and M_j the number of times a CpG site was covered by a read and was methylated. The "region value" methylation level β_j was then calculated by the ratio of M_j to N_j . In regions whose depth of coverage were zero, we discarded the calculation and considered the region as missing input data without applying any imputation technique. For each class of sample and each region, a Student t-test was performed. False discovery rate was controlled by a Bonferroni test correction. Top-500 negative logFC and significantly different regions in each one-versus-rest test were selected. The final tumor-specific methylation matrix is constructed by aggregating the average methylation density of all samples having the same label at each region. Missing values are removed before aggregating. We finally obtained a methylation reference matrix of shape $6 \times m$.

Non-negative least square matrix factorization

For a given cfDNA sample, to determine the relative weights (deconvolution scores) of each class that contribute to the sample, we implemented a simple non-negative least square (NNLS) algorithm. Let us denote by X the methylation reference matrix, $X \in \mathbb{R}^{6 \times m}$, representing 6 contributing classes and m selected regions. The given input cfDNA sample was represented by a vector α of shape $1 \times m$. NNLS proceeds to find the weights by solving the following minimization problem.

$$W = \operatorname{argmin} \|X - W\alpha\|_2$$

constrained by $W \geq 0$, where $W \in \mathbb{R}^{6 \times 1}$. W was then normalized to unit sum. To solve this optimization problem, we implemented an in-house Python script based on the Python *scipy* library.

Graph convolutional neural network (GCNN)

Following the same procedure as our previous work [9], we constructed a graph to train the GCNNs model from discovery and validation cohorts, comprised of low-depth WGBS cfDNA samples of five cancer types

(Dataset 4). The overall framework was depicted in [9]. The discovery cohort was then split into two subgroups, including the train dataset and validation dataset for 10-fold cross-validation with stratified sampling to ensure that each cancer class within the train dataset and validation dataset receives the proper representation. The model achieved the highest accuracy among ten folds on the validation dataset chosen and evaluated on an unseen test dataset built from the validation cohort. The undirected input graph $G = (V, E)$ incorporated a node set $V = \{X_i, Y_i | i = 1, \dots, N\}$ ($|V| = N$) and an edge set $E = \{e_{ij}\}$ ($|E| = \epsilon$), where X_i and Y_i denoted a node i and its label, N and ϵ denoted the number of node and edge in the graph, respectively. Each node X_i was represented by a feature vector $x_i \in \mathbb{R}^d$, which was a concatenation of groups of features (e.g. GWMD, Deconvolution score). We constructed interconnections between nodes at the first layer by k -nearest neighbor (k -NN) of $X = \{x_i | i = 1, \dots, N\} \in \mathbb{R}^{N \times d}$, where $k = 5$ in our experiments. We defined an adjacency matrix $A = \{a_{uv}\} \in \mathbb{R}^{N \times N}$ where we initialized $a_{uv} = 1$ if an edge $(u, v) \in E$, and $a_{uv} = 0$ otherwise. While the label $Y_i \in \{traindataset, validationdataset\}$ was seen by the model during the training phase, the label $Y_j \in \{testdataset\}$ was unseen, denoted as unknown nodes, and only used in the evaluation phase. In this way, we adopted a transductive learning [31] setting to analyze the model's performance on the TOO prediction.

Regarding the model's architecture, we based our model on Graph Transformer Networks [32], a graph convolutional neural network with support for learning interconnections between nodes in latent spaces. At every layer $l_i \in \{l_1, \dots, l_L\}$, feature representations of node x_u at layer l_i and edges connected node u and node v were defined by:

$$a_{uv}^l = \frac{\exp\left(\left(W_Q^l z_u^l\right)^T \left(W_k^l z_v^l\right)\right)}{\sum_{w=1}^N \exp\left(\left(W_Q^l z_u^l\right)^T \left(W_k^l z_w^l\right)\right)}$$

$$z_u^{(l+1)} = \sum_{v=1}^N a_{uv}^l \left(W_V^l z_v^l\right) + z_u^l$$

where a_{uv}^l was the connection between node u and node v at the l -th layer, W_Q^l, W_k^l, W_V^l were the learnable parameters at the l -th layer, z_u^l was the latent feature vector of node u at the l -th layer. In this work, we chose $L = 2$, each with the hidden feature size of 64. The output prediction layer was a single Multi Linear Perceptron (MLP).

We dealt with the TOO prediction as a node classification problem, where each node was predicted as

one of the five classes: Breast, CRC, Gastric, Liver, and Lung. To avoid the model bias towards major class for the imbalance dataset problem, we applied the focal loss [33] to guide the model focusing more on hard, misclassified samples. The model was trained using the Adam optimizer [34] with a learning rate of 10^{-3} . A general description of our model's architecture is shown in Supplementary Fig. 3.

Results

Construction of a tumor-specific methylation atlas (TSMA)

The construction of the TSMA relied on an established hypothesis that adjacent CpG sites could be co-methylated and share similar methylation status [18, 35–39]. We therefore defined a CpG region as a region of 100 bp in length covering at least 5 CpG sites. This definition allowed us to capture regions of dense CpG sites and multiple CpG sites within a region were expected to be covered by a single sequencing read. With this definition, we identified approximately 1.1 million CpG regions in the human reference genome hg19 (Supplementary Table 1). We then used WGBS data of 64 tumor tissue samples and 24 WBC samples from the Atlas construction dataset (Fig. 1, Materials and methods) to calculate the “region value”, defined as proportion of total methylated CpGs in the reads mapped to a region over the total CpGs that covered by these reads (Fig. 1), for each CpG region. To construct the TSMA, the regions in which region values were significantly different among five cancer tissues and WBC were captured. This process allowed the selection of the top 500 regions with the highest absolute log2 fold-change (log2FC) using a one-versus-rest test statistics strategy. Only regions with negative log2FC (significantly lower in the test tissue versus rest) were chosen, based on the observation that most cell-type specific differentially methylated CpG regions were unmethylated [24]. These selected regions were then organized into their respective groups and sorted based on their absolute log2FC values. Representative values for each tissue type at each region was calculated by averaging the region value across all respective samples, as illustrated in Fig. 2A. Overall, a TSMA of 2,945 differential CpG regions between 5 tumor tissue types and WBC were constructed.

To assert that our methodology could select for cancer relevant regions, we mapped these regions to genes and conducted a gene set over-representation analysis using GO and KEGG databases to identify pathways that were enriched by our selected regions. Indeed, cancer-related pathways emerged within the top 30 enriched terms (Fig. 2B). This result provided the first layer of evidence that our TSMA could capture tumor-specific methylation signals. Encouraged by these outcomes, we next hypothesize that our TSMA could be used to deconvolute samples from an independent source with methods

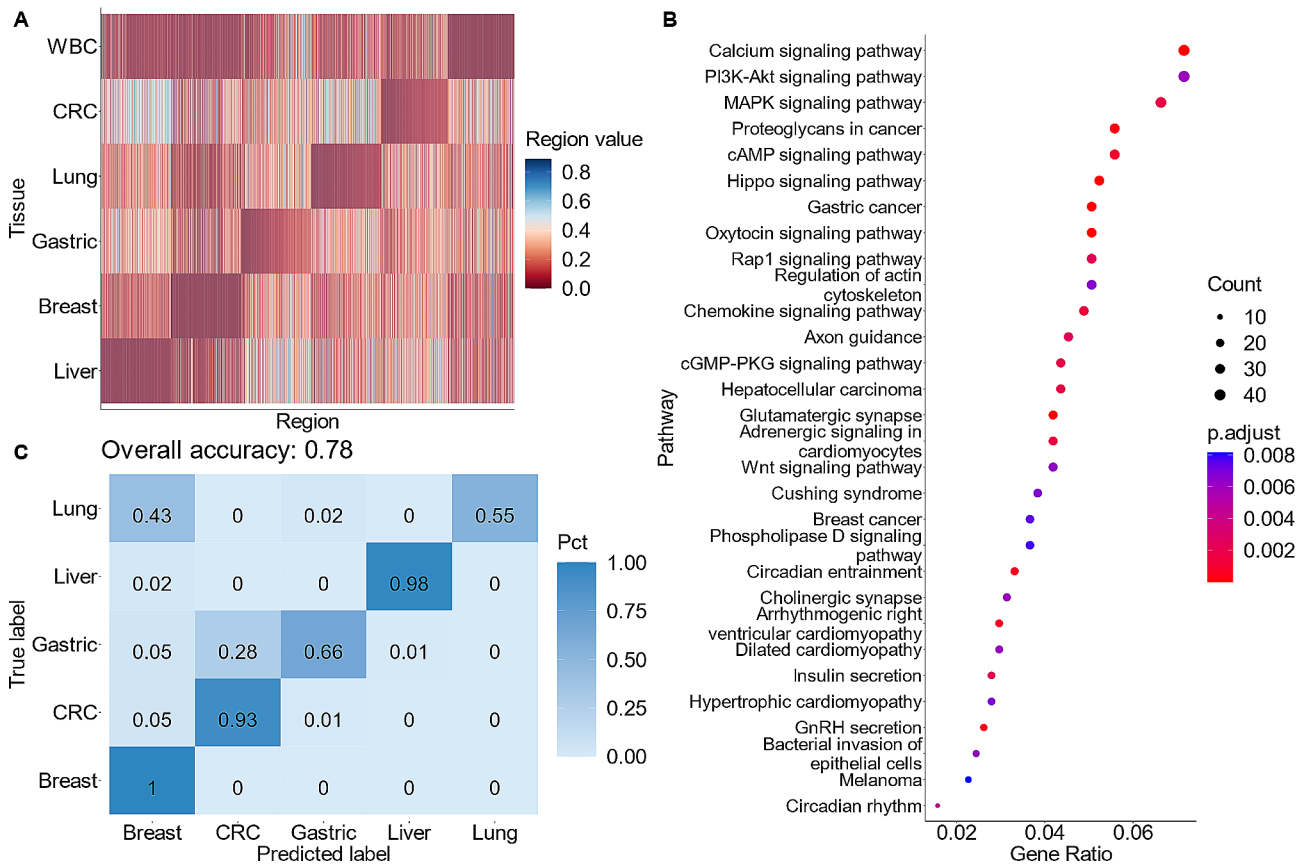


Fig. 2 The tumor-specific methylation atlas. **(A)** Heatmap of average region values in each cancer tissue type or WBC across 2,945 CpG regions included in the TSMA. **(B)** Pathway analysis reveals cancer-related pathways, which were enriched by the set of genes to which TSMA regions were mapped. **(C)** Prediction performance using highest deconvolution score to assign label to samples in Dataset 1 comprising of 888 colorectal samples, 1,814 lung samples, 398 gastric samples, 888 breast samples and 429 liver samples

of measuring methylation signal not limited to WGBS. To test this hypothesis, we obtained the 450 K/850K methylation microarray data comprising of 4,415 samples of cancer tumor tissues from the 5 cancer types of interest from TCGA database [40] (Dataset 1, Materials and methods). We transformed the CpG-wise microarray data into region-wise data to conform with our configuration (Materials and methods). We then performed deconvolution by NNLS and the label of sample was assigned by its highest deconvolution score component. This resulted in an overall accuracy of 78% (Fig. 2C). Specifically, we achieved accuracies of 100%, 98% and 93% for breast, liver and CRC cancer, respectively; while gastric and lung cancer exhibited lower accuracies of 66%, and 55%. These results validated our hypothesis and suggested that our TSMA has successfully captured cancer-specific signals that could be used to determine the TOO of a sample.

Significant correlation between deconvolution scores from TSMA and proportion of tumor DNA

Deconvolution scores derived from a DNA methylation atlas of normal human cell types provided estimates of underlying proportion for each cell type within a sample [41]. In the case of our TSMA, we expected that the TSMA-derived deconvolution scores would estimate the proportion of cancer DNA fragments corresponding to the five specific cancer types used to build TSMA. To confirm this, we first generated a set of samples with known amount of ctDNA by *in silico* mixing DNA fragments (WGBS reads) from tumor tissue into three different cfDNA WGBS samples from healthy donors at various fractions from 0.01 to 25% (Dataset 2, Materials and methods). We then evaluated the correlation between deconvolution scores and the known abundances of tissue-derived DNA fragments. In all samples, the proportions of WBC were consistently reported as the majority and decreased as the amount of tumor tissue DNA increased (Supplementary Fig. 1). We observed that at extremely low tumor fraction ($\leq 0.1\%$) deconvolution scores were mostly 0, except for liver and gastric

tumor where the scores were higher but not changing at 0.01%, 0.05% and 0.1%, suggesting a limit for this method at ~0.1% (Fig. 3). At tumor fraction higher than 0.1%, deconvolution scores showed good correlation with known tumor fractions ($R > 0.78$) (Fig. 3).

The next set of samples that we used to validate the correlation between deconvolution scores and tumor fractions were wet-lab spike-in samples where genomic DNA from each tumor tissue type were mixed with cfDNA from 2 healthy donors at four ratios (0.1%, 1%, 10%, and 25%), the resulting mixed DNA samples were then subjected to WGBS before calculating deconvolution scores (Dataset 3, Materials and methods). Deconvolution scores from our wet-lab spike-in experiment showed similar results with our *in silico* experiment. Across the dataset, we observed correlations in most cancer tissue types with the lowest correlation coefficient ($R = 0.58$) in CRC and highest correlation coefficient ($R = 0.87$) in Lung cancer (Fig. 4). In summary, both *in silico* and wet-lab spike-in experiments demonstrated the ability of

TSMA to deconvolute tumor tissue fragments with high correlation to their abundance, suggesting the potential of this approach in predicting TOO.

Since changes in deconvolution scores were significantly correlated to tumor abundance in cfDNA sample, we further explored the possibility of using deconvolution scores to determine TOO in a given cfDNA sample [41, 42]. Deconvolution scores of all cfDNA samples in Dataset 2 and 3 consistently showed WBC fractions at ~90%, leaving only ~10% for the other five tissue types (Supplementary Fig. 1). Therefore, we removed the WBC fraction and used the top tissue type (out of the 5 remaining types) as the TOO classification label. In Dataset 2, we correctly identified the TOO in 119/270 samples (overall accuracy of 44%, Supplementary Table 3). However, this performance was strongly affected by the spike-in ratio. The accuracy rose to 92% (Supplementary Fig. 2A) for samples with spike-in ratios greater than or equal to 10% (83/90 samples were correctly identified), and dropped to only 20% (Supplementary Fig. 2B) for

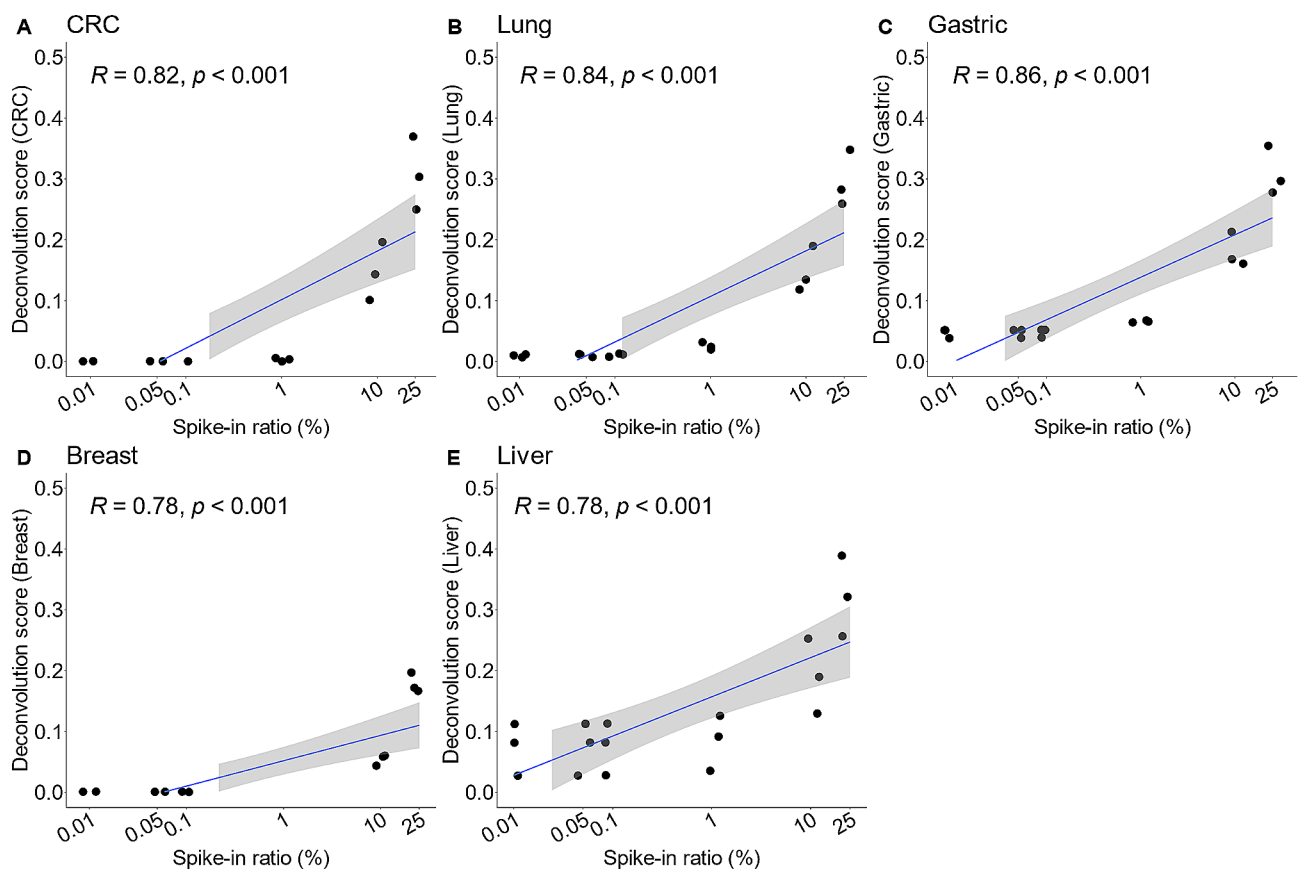


Fig. 3 Correlation between deconvolution scores and the proportion of cancer tissue fragments in *in vitro* spike-in samples. Significant correlation (Pearson correlation $R \geq 0.78$) between spike-in ratios (i.e. percent of tumor DNA), ranging from 0.01–25%, and deconvolution scores (i.e. fractions of specific cell type) was observed in (A) CRC, (B) Lung, (C) Gastric, (D) Breast, and (E) Liver cancer. Spike-in samples were generated by randomly sampling tumor tissue DNA fragments (WGBS reads) and mixing with 3 healthy cfDNA samples at defined ratios (3 replicates at each ratio). Correlation was measured by Pearson coefficient R . Values on x-axis were log-scaled for visualization purpose only. The 95% CI (gray shading region) is not available at spike-in tumor fraction of less than 1% because all deconvolution scores are zero

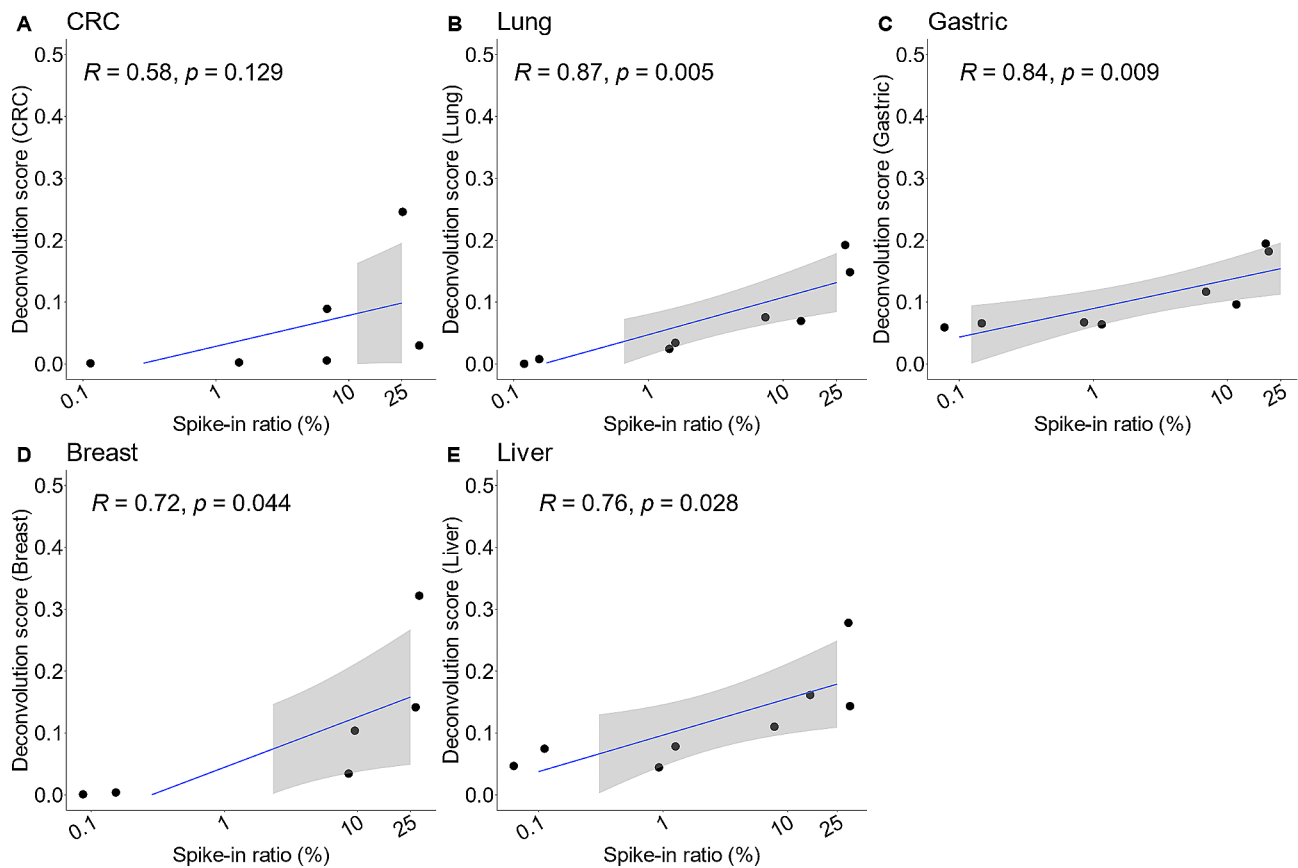


Fig. 4 Correlation between deconvolution scores and the proportion of cancer tissue fragments in wet-lab spike-in samples. Strong correlation between spike-in ratios and deconvolution scores was observed in (A) CRC, (B) Lung, (C) Gastric, (D) Breast, and (E) Liver cancer. Wet-lab spike-in samples were generated by mixing genomic DNA from each tumor tissue type with cfDNA from 2 healthy donors at four ratios (0.1%, 1%, 10%, and 25%), before WGBS and calculating deconvolution scores (2 replicates at each ratio). Correlation was measured by Pearson coefficient R . Values on x-axis were log-scaled for visualization purpose only. The 95% CI (gray shading region) is not available at spike-in tumor fraction of less than 1% because all deconvolution scores are zero

samples with spike-in ratios less than 10% (36/180 correctly identified). Similarly, TOO prediction was correct in 22/40 samples of Dataset 3 (overall accuracy of 55%, Supplementary Table 3), of which the samples with spike-in ratios greater than or equal to 10% exhibited 75% accuracy (Supplementary Fig. 2C), compared to accuracy of 35% for samples with spike-in ratios less than 10% (Supplementary Fig. 2D). Thus, our data indicated that the deconvolution approach worked well only in samples where the tumor tissue abundance exceeded 10%.

Combining deconvolution scores and genome-wide methylation density in a graph convolutional neural network enhances prediction performance

In early cancer and TOO detection from cfDNA, the multi-modal approach has become increasingly popular, where different features derived from different characteristics of cfDNA are combined as inputs for a machine learning or deep learning model to improve performance [7–12]. We have previously published a GCNN model using methylomics, fragmentomics,

copy number, and end motifs features for TOO detection with encouraging accuracy of 70% [9]. Here, we used the same dataset (Dataset 5) [9] to explore the possibility of combining deconvolution scores with other cfDNA features to enhance TOO performance. In total, we examined 57 combinations of deconvolution scores (Supplementary Table 4) with genome-wide methylation density (GWMD), targeted methylation density (TMD), copy number aberration (CNA), genome-wide fragment length profile (GWFP), and end motifs (EM) to search for the best performing model. For each combination, we trained a GCNN in a set of 438 cancer patient samples and validated the model in a held-out 239 samples (Dataset 5). GWMD, expressed as the average methylation density across non-overlapping 1 M bins in the entire genome, when combined with deconvolution scores achieved the highest accuracy of 69% (Fig. 5A, D). This was markedly improved from deconvolution scores alone (26% accuracy, Fig. 5B), or GWMD alone (63% accuracy, Fig. 5C). This result highlighted the contribution

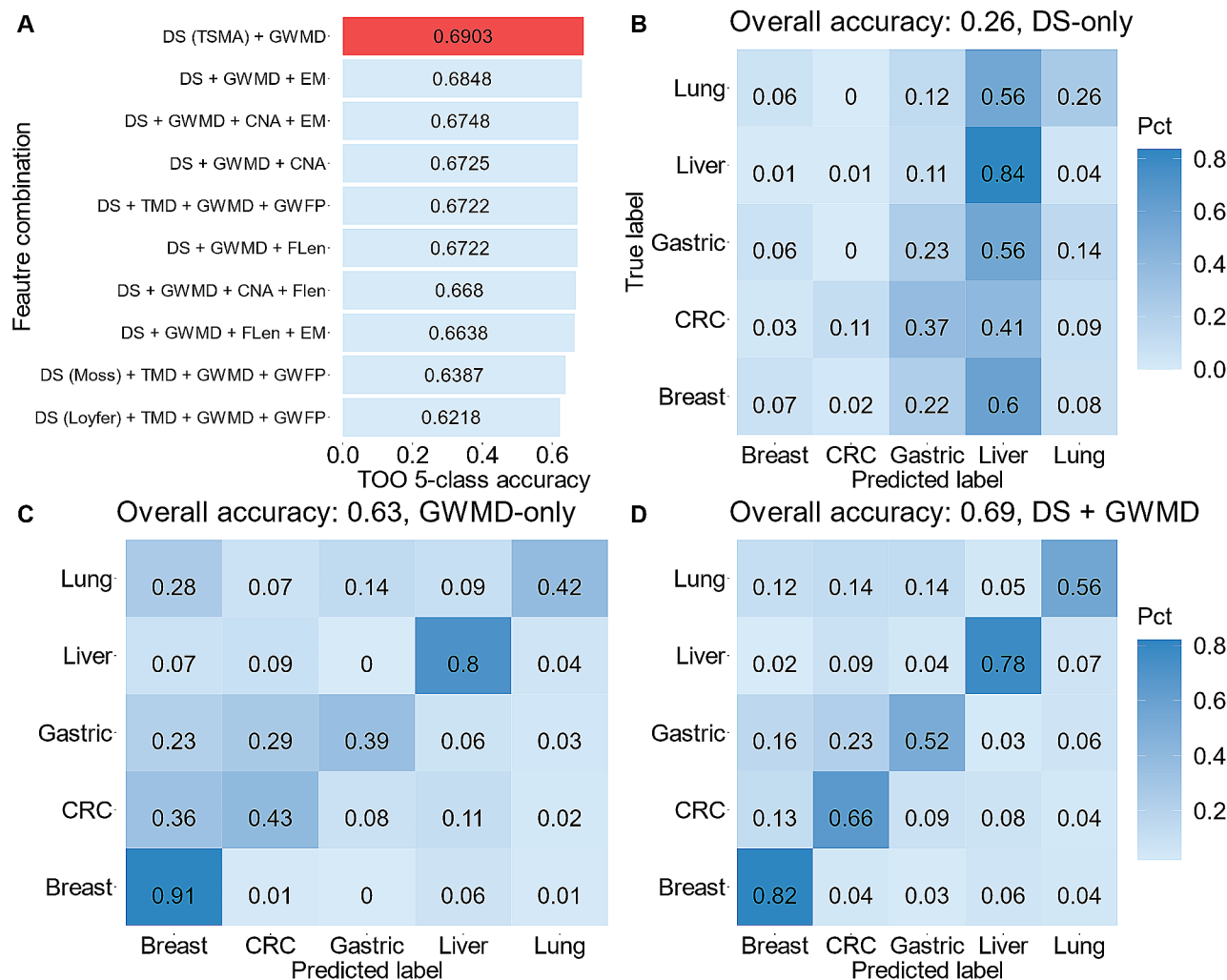


Fig. 5 Multi-modal approach combining TSMA deconvolution scores with other cfDNA features in a graph convolutional neural network. (A) Top 10 feature combinations achieving the highest accuracies. Results from other combinations are given in the Supplementary Table 4. The combination of deconvolution score (DS) and GWMD achieved highest accuracy and is shown in red. (B) Confusion matrix obtained from the deconvolution scores. A sample label was assigned by the highest tumor tissue type as indicated by deconvolution scores. (C) Confusion matrix obtained from a GCNN using GWMD feature only. (D) Confusion matrix obtained from a GCNN using both deconvolution scores and GWMD feature

of TSMA deconvolution scores in the GCNN model for TOO detection, especially when combined with GWMD.

Benchmarking TSMA against other methylation-reference deconvolution-based method

To benchmark our method against other methylation-reference TOO prediction method, we calculated the cell-type fractions (equivalent to our deconvolution scores) of the five aforementioned tissue types for 677 low-depth cfDNA samples using the two well-known methylation-reference-based atlases by Moss et al [23, 24] and Loyfer et al. [24]. We applied the same strategy as for our deconvolution scores (DS): the cell-type fractions from Moss et al. atlas and Loyfer et al. atlas (labelled DS-Moss and DS-Loyfer, respectively) were combined with GWMD, GWFP, TMD, CNA, and EM

features. The combined features were input to the GCNN model with the same set of hyper-parameters. Among all combinations with DS-Moss, the combination of “DS-Moss+TMD+GWMD+GWFP” obtained highest accuracy on the validation cohort (62%, **Supplementary Fig. 4.A**). Similarly, the combination of “DS-Loyfer+TMD+GWMD+GWFP” achieved highest accuracy among all combinations involving DS-Loyfer (64%, **Supplementary Fig. 4.C**). Overall, the combination of our deconvolution scores and GWMD achieved the highest accuracy of 69% (DS+GWMD, Fig. 5.A).

Discussion

The multi-modal approach has become increasingly popular for the construction of a machine learning or deep learning model in early cancer and TOO detection from

cfDNA [7–12]. This approach relies on bioinformatic analysis and feature engineering of WGS or WGBS data to reveal different characteristics of cfDNA, including but not limited to methylomics, fragmentomics, copy number, and end motifs [9]. In this study, we aimed to engineer a new feature that can be used either alone or more likely in combination with other features to enhance the performance of TOO detect from shallow WGBS of cfDNA samples. We constructed the TSMA to distinguish five tumor tissue types and white blood cells using WGBS of tumor tissue and paired WBC samples (Fig. 1). This novel approach set our atlas apart from the methylation atlas previously built using healthy human cell types [23, 24]. With this TSMA, methylation data (either WGBS or methylation microarray) can be deconvoluted by non-negative least square matrix factorization into deconvolution scores, which represent the proportion of each tissue type in a sample (Fig. 1). Given an input sample, deconvolution scores could be directly utilized to predict its cell type composition, limited to the five tumor tissues and WBC components in our atlas, or used in combination of other features in a model for TOO detection.

In our *in silico* and wet-lab spike-in experiments, we observed strong correlation between deconvolution scores and the known percentage of spike-in tumor-tissue DNA fragments in cfDNA samples (Figs. 3 and 4). However, the applications of deconvolution scores in real cfDNA samples, especially in low-depth WGBS samples, posed a greater challenge. In most samples, we found that WBC accounted for nearly 90% of the composition, leaving the sum of all potential tumor-specific signals to less than 10%. This finding aligned with the biological notion that tumor-derived DNA fraction could account for at most 1–10% in cfDNA context [43]. Validations on *in silico* and wet-lab spike-in datasets (Dataset 2, 3) indicated that the limit of detection was 10% tumor abundance to accurately predict a sample TOO (accuracy of 75–92%, Figs. 3 and 4), which made deconvolution scores alone unsuitable for TOO detection for low-depth cfDNA samples.

Alternatively, deconvolution scores can be used in combination with other cfDNA features readily available from previous works [7–9]. Using the same dataset that was presented previously [9], we calculated the deconvolution scores using our newly built TSMA and concatenated them to 57 different combinations of feature vectors (Fig. 5A). This combined feature vector was fed to a graph convolutional neural network to achieve the final prediction of the tissue of origin. Training was done on a cohort of 498 low-depth WGBS cfDNA samples (0.5x) and validation on a cohort of 239 low-depth WGBS cfDNA samples (0.5x) of cancer patients. We achieved a 5-class accuracy of 69% for TOO prediction, which is

comparable to the results obtained in our previous study [9]. Specifically, compared to the GCNN previously published [9], this new GCNN model achieved higher accuracy for breast and liver (82% vs. 78%, and 78% vs. 76%), similar accuracy for CRC (66%) and lower accuracy for gastric and lung (52% vs. 55%, and 56% vs. 63%). This result highlighted that the GCNN using deconvolution scores and GWDM achieved comparable performance to a GCNN built with 9 different sets of features, suggesting that the contribution of deconvolution scores is equal to 8 other feature sets.

To benchmark our method, we have compared the use of TSMA deconvolution scores against cell-type fractions estimated by Moss et al. atlas [23] and Loyfer et al. atlas [24] in combination with other features in the GCNN model (**Supplementary Fig. 4**). TSMA deconvolution scores when combined with GWMD achieved the highest accuracy among all combinations (Fig. 5A). A possible reason for this is because both atlases by Moss et al. and Loyfer et al. were constructed using healthy primary cell types, hence reflecting the cell-type specific signal but not the cancerous tumor tissue signal. This further supports our approach of using cancer tissue samples for the construction of TSMA for the purpose of TOO prediction.

There are several limitations in this work. First, due to time and budget constraints, we focused only on the five most prevalent cancers in the Vietnamese population [25]. Expanding the atlas to a broader panel of cancers to include other common cancers, such as thyroid, pancreatic or kidney cancer, is of interest for our future project. A matching set of cfDNA, WBC and tumor tissue samples from the same patient is required and this process is challenging in Vietnam due to the rarity of samples. Attention should also be paid to the ctDNA shedding characteristics of the cancer types [44–47], since our approach relies on the presence of tumor-derived DNA fragments in the bloodstream at certain level of abundance. Second, the two spike-in experiments (Figs. 3 and 4) indicate the influence of tumor abundance on our method. At tumor-fraction less than 1%, deconvolution scores are almost zero for all samples, which suggests that the limit of detection for ctDNA is approximately 1%. However, to achieve adequate TOO prediction performance, the tumor-fraction must exceed 10% (**Supplementary Fig. 2**). Increasing the sequencing depth of TSMA regions via targeted sequencing, incorporating various type of features with the TSMA deconvolution scores, or implementing a multi-omics approach might improve the sensitivity of our TSMA method. Third, the development and validation of our method were conducted on a relatively small dataset. Utilizing public dataset could objectively evaluate real-world applicability of our method. However, in the construction of TSMA,

we wish to focus on reads overlapping regions that have five CpG sites in a 100 bp window. This approach, along with the calculation of genome-wide and targeted features [9], requires the input data to be in read-based format (FASTQ, BAM). Unfortunately, to the best of our knowledge, most public methylation data are deposited as tables or matrices of methylation density at CpG sites only, and thus do not meet this requirement.

Conclusions

In conclusion, we have developed a TSMA depicting differential methylated regions across five cancer tumor types and white blood cells. The deconvolution scores from our atlas correlated well with the tumor fraction in cfDNA samples although this is limited mostly to tumor fraction of more than 10%. However, the combination of the deconvolution scores and genome-wide methylation density features significantly enhanced the TOO detection performance when applying to low-depth WGBS cfDNA samples. In summary, our study has paved the way for the application of tumor-specific atlas in TOO detection. Future development of such an atlas might hold the key to significant improvement in TOO detection of all cancer types in low-depth cfDNA samples.

Abbreviations

cfDNA	Cell free DNA
TOO	Tissue-of-origin
TSMA	tumor-specific methylation atlas
WBC	white blood cells
NNLS	non-negative least square matrix factorization
GCNN	graph convolutional neural network
GWMD	genome-wide methylation density
TMD	targeted region methylation density
GWFP	genome-wide fragmentation profile
EM	end-motif
CNA	copy number aberration

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12967-024-05416-z>.

Supplementary Material 1
Supplementary Material 2
Supplementary Material 3
Supplementary Material 4
Supplementary Material 5

Authors' contributions

Conceptualization: Trong Hieu Nguyen, Hoa Giang, Le Son Tran, Minh Duy Phan.
 Formal analysis: Nhu Nhat Tan Doan, Trung Hieu Tran, Van Thien Chi Nguyen.
 Data analysis: Thi Hue Hanh Nguyen.
 Acquisition: Hoai-Nghia Nguyen.
 Writing-original draft: Trong Hieu Nguyen.
 Writing-review and editing: Trong Hieu Nguyen, Giang Thi Huong Nguyen, Hoa Giang, Le Son Tran, Minh Duy Phan.

Funding

The study was funded by Gene Solutions.

Data Availability

Analytic data are available on request to the corresponding authors (THN, MDP). Raw FASTQ data are not publicly available due to ethical and regulatory restrictions. We confirm that this does not alter our adherence to Cancer Investigation policies on sharing data and materials.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of the Medic Medical Center, University of Medicine and Pharmacy and Medical Genetics Institute, Ho Chi Minh city, Vietnam. Written informed consent was obtained from each participant in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Conflict of interest

THN, HNN, HG, LST, MDP receive compensation and have an equity interest in Gene Solutions. NNTD, THT, LAKH, PLD, VTCN, THHN, GTHN are employees of Gene Solutions. The authors ensure that this does not alter the accuracy or integrity of the manuscript. The study was funded by Gene Solutions. The sponsor has no role in the analysis of the data and the preparation of the manuscript.

Author details

¹Medical Genetics Institute, Gene Solutions, Ho Chi Minh, Vietnam
²Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, USA

Received: 3 January 2024 / Accepted: 19 June 2024

Published online: 03 July 2024

References

- Cristiano S, Leal A, Phallen J, Fiksel J, Adleff Vilmos, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nat*. 2019;570:385–9.
- Nguyen VC, Nguyen TH, Phan TH, Tran T-HT, Pham TTT, Ho TD, et al. Fragment length profiles of cancer mutations enhance detection of circulating tumor DNA in patients with early-stage hepatocellular carcinoma. *BMC Cancer*. 2023;23:233.
- Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*. 2018;10:eaat4921.
- Harbers L, Agostini F, Nicos M, Poddighe D, Bienko M, Crossetto N. Somatic copy number alterations in human cancers: An analysis of publicly available data from the cancer genome atlas. *Front Oncol*. 2021;11:700568.
- Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, et al. Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov*. 2020;10:664–73.
- Phan TH, Chi Nguyen VT, Thi Pham TT, Nguyen V-C, Ho TD, Quynh Pham TM, et al. Circulating DNA methylation profile improves the accuracy of serum biomarkers for the detection of nonmetastatic hepatocellular carcinoma. *Future Oncol*. 2022;18:4399–413.
- Nguyen HT, Khoa Huynh LA, Nguyen TV, Tran DH, Thu Tran TT, Khang Le ND, et al. Multimodal analysis of ctDNA methylation and fragmentomic profiles enhances detection of nonmetastatic colorectal cancer. *Future Oncol*. 2022;18:3895–912.
- Pham TMQ, Phan TH, Jasmine TX, Tran TTT, Huynh LAK, Vo TL, et al. Multimodal analysis of genome-wide methylation, copy number aberrations, and end motif signatures enhances detection of early-stage breast cancer. *Front Oncol*. 2023;8:1127086.
- Nguyen VTC, Nguyen TH, Doan NNT, Pham TMQ, Nguyen GTH, Nguyen TD, et al. Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization. *eLife*. 2023;12:RP89083.

10. Jamshidi A, Liu MC, Klein EA, Venn O, Hubbell E, Beausang JF, et al. Evaluation of cell-free DNA approaches for multi-cancer early detection. *Cancer Cell*. 2022;40:1537–49.
11. Kim SY, Jeong S, Lee W, Jeon Y, Kim Y-J, Park S, et al. Cancer signature ensemble integrating cfDNA methylation, copy number, and fragmentation facilitates multi-cancer early detection. *Exp Mol Med*. 2023;55:2445–60.
12. Li Y, Jiang G, Wu W, Yang H, Jin Y, Wu M, et al. Multi-omics integrated circulating cell-free DNA genomic signatures enhanced the diagnostic performance of early-stage lung cancer and postoperative minimal residual disease. *EBioMedicine*. 2023;91:104553.
13. Kang S, Li Q, Chen Q, Zhou Y, Park S, Lee G, et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol*. 2017;18:53.
14. Li W, Li Q, Kang S, Same M, Zhou Y, Sun C, et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res*. 2018;46:e89.
15. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol*. 2020;31:745–59.
16. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*. 2004;429:457–63.
17. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nat Rev Cancer*. 2004;4:143–53.
18. Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003;33:245–54.
19. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. *Lancet*. 2018;392:777–86.
20. Lehmann-Werman R, Neiman D, Zemmour H, Moss J, Magenheimer J, Vaknin-Dembinsky A, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci U S A*. 2016;113:E1826–34.
21. Sun K, Jiang P, Chan KA, Wong J, Cheng YK, Liang RH, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci*. 2015;112:E5503–12.
22. Shen SY, Singhania R, Fehring G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature*. 2018;563:579–83.
23. Moss J, Magenheimer J, Neiman D, Zemmour H, Loyer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun*. 2018;9:5068.
24. Loyer N, Magenheimer J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. *Nature*. 2023;613:355–64.
25. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209–49.
26. Andrews S. FastQC A quality control tool for high throughput sequence data. In: Quantitative Undergraduate Biology Education and Synthesis. National Science Foundation. 2010. <https://qubeshub.org/resources/fastqc>. Accessed 25 Jun 2024.
27. Krueger F, James F, Ewels P, Apyounian E, Weinstein M, Schuster-Boeckler B, et al. FelixKrueger/TrimGalore: v0.6.10 - add default decompression path (0.6.10). In: Zenodo. 2023. <https://zenodo.org/record/7598955>. Accessed 25 Jun 2024.
28. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011;27:1571–1572.
29. Picard toolkit. Broad Institute, Massachusetts. 2018. <https://broadinstitute.github.io/picard>. Accessed 25 Jun 2024.
30. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10:giab008.
31. Joachims T. Transductive learning via spectral graph partitioning. *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. 2003;3:290–7.
32. Yun S, Jeong M, Kim R, Kang J, Kim HJ. Graph transformer networks. *Advances in Neural Information Processing Systems*. 2019;32:9d63484a.
33. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;42:318–327.
34. Kingma DP, Ba J. A method for stochastic optimization. *Third International Conference on Learning Representations*. 2015;1412:6980.
35. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*. 2006;38:1378–85.
36. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, et al. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res*. 2008;18:780–90.
37. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nat*. 2009;462:315–22.
38. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*. 2009;41:178–86.
39. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotypes blocks AIDs in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet*. 2017;49:635–42.
40. Weinstein JN, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45:1113–20.
41. Titus AJ, Gallimore RM, Salas LA, Christensen BC. Cell-type deconvolution from DNA methylation: A review of recent applications. *Hum Mol Genet*. 2017;26:R216–24.
42. Lubotzky A, Zemmour H, Neiman D, Gotkine M, Loyfer N, Piyanzin S, et al. Liquid biopsy reveals collateral tissue damage in cancer. *JCI Insight*. 2022;7:e153559.
43. Neumann MHD, Bender S, Krahn T, Schlange T. ctDNA and CTCs in liquid biopsy – current status and where we need to progress. *Comput Struct Biotechnol J*. 2018;16:190–5.
44. Rhrissorrakrai K, Utro F, Levovitz C, Parida L. Lesion Shedding model: Unraveling site-specific contributions to ctDNA. *Brief Bioinform*. 2023;24:bbad059.
45. Lam VK, Zhang J, Wu CC, Tran HT, Li L, Diao L, et al. Genotype-specific differences in circulating tumor DNA levels in advanced NSCLC. *J Thorac Oncol*. 2021;16:601–9.
46. Pascual J, Attard G, Bidard FC, Curigliano G, De Mattos-Arruda L, Diehn M, et al. ESMO recommendations on the use of circulating tumour DNA assays for patients with cancer: a report from the ESMO Precision Medicine Working Group. *Ann Oncol*. 2022;33:750–68.
47. Zhang Y, Yao Y, Xu Y, Li L, Gong Y, Zhang K, et al. Pan-cancer circulating tumor DNA detection in over 10,000 Chinese patients. *Nat Commun*. 2021;12:11.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.