Molecular Therapy
Nucleic Acids
Original Article

TiSMeD: A tissue-specific methylation and expression database for biomarker and translational applications

Jiao Cheng,[1,6,7] Zhe Lin,[2,3,6] Lvying Wu,[4] Qianying Li,[2] Huajun Yin,[2] Hao Wang,[1] Hao Chen,[2] Xiaoyan Chen,[1] and Zhi-Liang Ji[2,3,5]

[1]Maternal-Fetal Medicine Institute, Department of Obstetrics and Gynaecology, Shenzhen Baoan Women's and Children's Hospital, Shenzhen 518100, China; [2]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, China; [3]National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361102, China; [4]Institute of Clinical Medicine, The Second Affiliated Hospital of Hainan Medical University, Haikou 570311, China; [5]Fujian Provincial Key Laboratory of Chemical Biology, Xiamen University, Xiamen 361005, China

Tissue-specific methylation sites (TSMs) are important epigenetic features associated with gene regulation, tissue development, and disease pathogenesis. However, the lack of comprehensive and reliable resources for TSMs restricts advancements in epigenetic and translational research. We present TiSMeD (http://www.bio-add.org/TiSMeD/), a multi-omics database integrating 6,782 DNA methylation, 16,894 transcriptome, and 241 proteome profiles across 48 normal human tissues. Using a scoring framework based on SPM_{adjust} and Tscore, we identified 67,427 high-confidence TSMs, 4,607 tissue-specific genes, and 2,833 tissue-specific proteins, along with over 11 million housekeeping methylation sites. TiSMeD enables interactive exploration and data retrieval, supporting biomarker discovery and disease research. We demonstrate its utility in tracing the tissue-of-origin of cell-free DNA (cfDNA), prioritizing 1,849 cancer biomarkers from The Cancer Genome Atlas (TCGA), and constructing a multi-cancer tracing and diagnostic model achieving 95.7% accuracy. TiSMeD serves as a robust, user-friendly platform integrating multi-omics data to advance epigenetic research and biomarker translation.

INTRODUCTION

DNA methylation is a widespread and stable epigenetic modification in mammals, regulating gene expression by recruiting gene-silencing proteins or by inhibiting transcription factor binding.[1,2] This modification is highly conserved across species and tissues,[3] and tissue-specific methylation sites (TSMs) may underpin the regulation of tissue-specific genes (TSGs) in primates.[4] Accumulating evidence has linked patterned methylation, particularly TSM, to the pathogenesis of complex diseases such as cancer.[5] To date, numerous DNA methylation biomarkers have been developed for cancer diagnosis, disease management, and post-treatment monitoring.[6–8] However, the diagnostic performance of these methylation biomarkers warrants further optimization, particularly regarding their tissue specificity. For example, aberrant methylation of *SEPTIN9* serves as a biomarker not only for colorectal cancer but also for hepatocellular carcinoma and head and neck squamous cell carcinoma, potentially leading to ambiguous diagnoses.[9–11] Beyond conventional tissue-based testing, TSMs demonstrate significant potential for liquid biopsy applications, particularly for tracing the tissue of origin of circulating cell-free DNA (cfDNA) in cancer patients. This approach enables non-invasive tumor localization and has the potential to improve early detection and disease monitoring.[12–14] However, current cfDNA-based assays are hindered by incomplete and imprecise TSM reference maps, which limit both their diagnostic accuracy and clinical utility.[15,16]

Recent advancements in methylation microarray and sequencing technologies have generated abundant methylation data, spurring the development of numerous databases for methylation data dissemination, storage, and analysis. MethDB is a pioneering repository that focuses on the environmental effects on DNA methylation and experimental details.[17] MethBase,[18] NGSmethDB 2017,[19] and MethBank 4.0[20] host bisulfite-sequencing methylomes of different cell lines, tissues, and species. The EWAS Data Hub provides an open platform to catalog array-based DNA methylation metadata for epigenome-wide association studies.[21,22] The scMethBank specializes in single-cell DNA methylation data and metadata.[23] MethAgingDB is a dedicated resource for age-related DNA methylation patterns across human and mouse tissues.[24] Other specialized repositories, such as

Received 9 September 2025; accepted 26 February 2026;
https://doi.org/10.1016/j.omtn.2026.102884.

[6]These authors contributed equally

[7]Present address: Shenzhen Baoan Women's and Children's Hospital, Shenzhen, 518100, China

Correspondence: Xiaoyan Chen, Maternal-Fetal Medicine Institute, Department of Obstetrics and Gynaecology, Shenzhen Baoan Women's and Children's Hospital, Shenzhen 518100, China.
E-mail: chenxiaoyan@cuhk.edu.hk
Correspondence: Zhi-Liang Ji, State Key Laboratory of Cellular Stress Biology, School of Life Sciences, Xiamen University, Xiamen 361102, China.
E-mail: appo@xmu.edu.cn

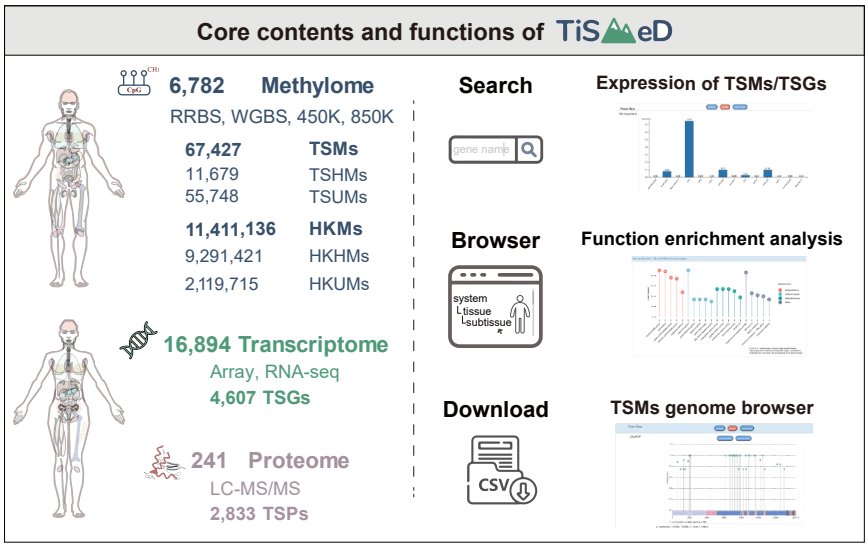


Figure 1. Overview of TiSMeD

TiSMeD is a comprehensive resource that integrates 6,782 methylomes, 16,894 transcriptomes, and 241 proteomes obtained from multiple platforms. All datasets were processed using a standardized pipeline, resulting in the identification of 67,427 TSMs, 4,607 TSGs, and 2,833 TSPs based on tissue specificity and confidence score evaluation. In addition, TiSMeD includes 11,411,136 HKMs. TiSMeD features a user-friendly interface that enables data searching, browsing, downloading, and visualization.

ASMdb and MethmiRbase, focus on allele-specific methylation and methylation-mediated microRNA regulation, respectively.[25,26] Several databases, including PubMeth,[27] MethyCancer,[28] MeInfoText 2.0,[29] MENT,[30] MethHC 2.0,[31] DDMGD,[32] DiseaseMeth 3.0,[33] and DNMIVD,[34] provide extensive information regarding disease-methylation associations. These repositories establish a robust foundation for methylation research across diverse biological contexts. However, only a limited number of databases currently provide information on TSMs. For example, the EWAS Data Hub calculates tissue-specific scores for approximately 850,000 methylation sites determined on Infinium HumanMethylation450 (450K) and MethylationEPIC (850K) arrays. Due to technical limitations inherent in the array platforms, numerous TSMs located within gene body regions remain undetected.[35] Several methods have been proposed for identifying tissue specificity, including the preferential expression measure,[36] Tau,[37] Extended-Tau,[38] Gini coefficient,[39] Hg,[40] expression enrichment,[41] *Z* score,[42] specificity measure (SPM),[43] tissue specificity index,[44] SPECS,[45] and TransTEx.[46] However, the tissue-specific features identified by current methods often exhibit limited concordance. The impact of sample size and tissue diversity has often been overlooked; tissue-specific features identified from datasets covering limited tissue types may lose their specificity when evaluated across a broader range of tissues. This limitation significantly hinders the comprehensive characterization of human tissue-specific methylation patterns. Furthermore, despite the pivotal role of DNA methylation in transcriptional regulation, few databases have integrated methylation data with corresponding gene and protein expression profiles.

To bridge the critical data and technical gaps in epigenetic regulation research, we have developed a database, the Tissue-Specific Methylation and Expression Database (TiSMeD, http://www.bio-add.org/TiSMeD/), which provides integrated and comprehensive information on tissue-specific methylation sites, genes, and proteins (Figure 1). Specifically, we integrated an extensive collection of methylomes, transcriptomes, and proteomes derived from normal human tissues across multiple technological platforms. To ensure robust identification of tissue-specific features, we introduced two statistical metrics to systematically evaluate tissue specificity. We anticipate that TiSMeD will facilitate a deeper understanding of tissue-specific regulation, transcription, and translation, highlighting their crucial roles in tissue differentiation, development, and function, thereby advancing the clinical utility of DNA methylation in diagnostic biomarker discovery and therapeutic targeting.

RESULTS

Database content

The TiSMeD database integrates multi-omics data, including 6,782 methylomes generated using reduced-representation bisulfite sequencing (RRBS), whole-genome bisulfite sequencing (WGBS), and 450K and 850K arrays; 16,894 transcriptomes analyzed by microarray and RNA sequencing (RNA-seq); and 241 proteomes characterized by liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Figure 2A). A total of 67,427 high-confidence TSMs, 4,607 TSGs, and 2,833 tissue-specific proteins (TSPs) were identified across multiple tissue levels, encompassing 11 organ systems, 38 tissues, and 15 sub-tissues (Table 1). The specificity and robustness of these features were ensured by applying stringent criteria (adjusted specificity measure score [SPM_{adjust}] $\geq$ 0.8 and tissue-specific confidence score [Tscore] $\geq$ 0.6), requiring support from at least two distinct datasets (Figure 2B). Among the TSMs, 11,679 were classified as tissue-specific hypermethylation sites (TSHMs) and 55,748 as tissue-specific hypomethylation or unmethylation sites (TSUMs). Additionally, the database contains 11,411,136 housekeeping methylation sites (HKMs) exhibiting consistent methylation patterns across all analyzed tissues.

To assess potential biases arising from imbalances in sample sizes, we systematically examined the relationship between sample size per tissue and the number of identified TSMs. Across all tissues, no significant correlation was observed between sample size and the number of identified TSMs (Spearman's ρ = 0.14, p = 0.48; Figure S1A). We further verified the robustness of our identification method using a down-sampling strategy on the large-scale RRBS dataset (GEO: GSE233417). TSM profiles derived from uniformly

Figure 2. Tissue-specific features in TiSMeD

(A) Overview of multi-omics data integrated in TiSMeD. (B) Relationship between the number of datasets supporting the tissue specificity of TSMs, TSGs, TSPs, and their respective Tscore values. A Tscore greater than 0.6 indicates that the tissue specificity of TSMs, TSGs, or TSPs is supported by at least two datasets. (C) Distribution of TSMs, TSGs, and TSPs across different tissues. Yellow blocks indicate whether omics data have been collected for the corresponding tissue. (D) Correlations between TSM methylation and target gene expression levels across genomic regions. (E) Proportions of correlation categories between TSM methylation and target gene expression. (F) Correlation distribution between TSG mRNA levels and corresponding protein expression. (G) Correspondence of tissue specificity between the transcriptomic and proteomic levels.

Table 1. Statistics of tissue-specific features in TiSMeD

Tissue-specific features	Tissue level	Count	Across tissue number
TSHMs	System	7,696	9
TSHMs	Tissue	3,945	22
TSHMs	Sub-tissue	38	5
TSUMs	System	45,418	10
TSUMs	Tissue	10,287	24
TSUMs	Sub-tissue	43	6
TSGs	System	2,003	11
TSGs	Tissue	2,492	34
TSGs	Sub-tissue	112	9
TSPs	System	1,605	10
TSPs	Tissue	1,228	22
TSPs	Sub-tissue	0	0

down-sampled subsets demonstrated high concordance with the complete dataset (TSHMs: Spearman's $\rho = 0.73$, $p < 0.001$; TSUMs: Spearman's $\rho = 0.62$, $p < 0.01$; Figures S1B–S1E). These results indicate that the identified tissue-specific profiles reflect intrinsic biological characteristics rather than artifacts of sample size imbalances.

TSMs exhibited distinct enrichment patterns across specific tissues. At the tissue level, the liver exhibited the highest number of TSHMs, followed by the testis, stomach, skin, and pancreas (Figure 2C). The testis displayed the highest number of TSUMs, followed by the liver, ovary, muscle, and colon; collectively, these tissues accounted for over 84% of all TSUMs. At the sub-tissue level, TSHMs were predominantly enriched in the tibial nerve, right lobe of the liver, and body of the pancreas; meanwhile, TSUMs were primarily concentrated in the transverse colon, right lobe of the liver, and esophageal squamous epithelium. TSGs were detected across all tissues but exhibited an uneven distribution, with the testis, brain, and liver collectively contributing approximately 55% of all TSGs. Similarly, the highest counts of TSPs were identified in the brain, testis, and liver.

To elucidate the regulatory landscape underlying these multi-omics layers, we conducted a systematic correlation analysis. TSMs exhibited significant associations with the expression of their corresponding target genes (Figure 2D). Specifically, TSUMs demonstrated a more pronounced global negative correlation with gene expression than TSHMs. Furthermore, TSMs located in promoter regions were predominantly inversely correlated with expression, whereas those in gene bodies showed mixed correlation patterns (Figure 2E). Assessment of the concordance between transcriptomic and proteomic layers revealed a right-skewed distribution of mRNA-protein correlations. A substantial proportion of TSGs showed strong positive associations (Spearman's $\rho > 0.5$; Figure 2F), indicating that tissue-specific transcriptomic profiles are predominantly preserved at the proteomic level. Notably, 519 TSGs exhibited congruent tissue specificity at the proteomic level (Figure 2G), underscoring a coordinated regulatory landscape across epigenetic, transcriptomic, and proteomic layers.

Database access

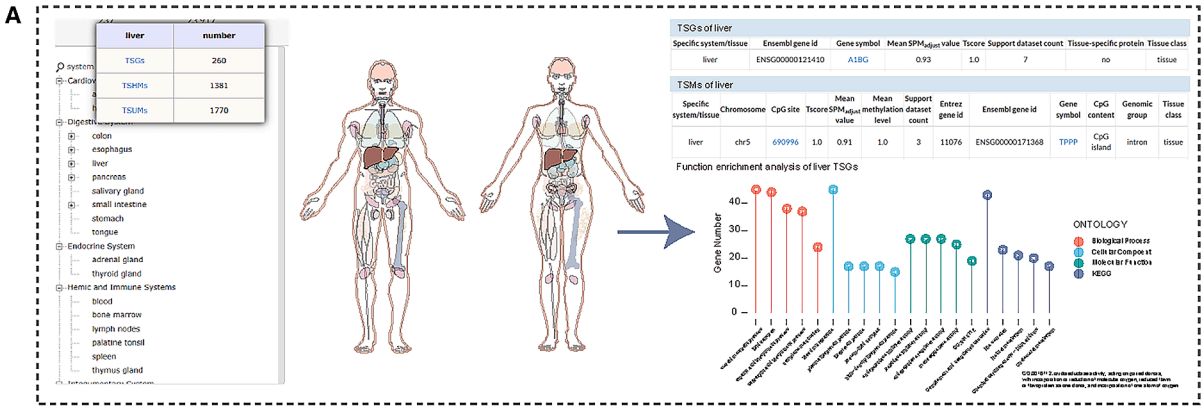
TiSMeD is freely available at http://www.bio-add.org/TiSMeD. User-friendly interfaces were developed to facilitate interactive data search, visualization, and retrieval. TiSMeD provides two primary methods for direct data access: browsing and searching. The browsing module utilizes a hierarchical tissue tree and an anatomical diagram to streamline tissue selection (Figure 3A). Selecting a tissue reveals three distinct categories: "TSGs," "TSHMs," and "TSUMs," each linking to the corresponding data list. On the results page, TSMs, TSGs and supporting information are displayed alongside functional enrichment charts.

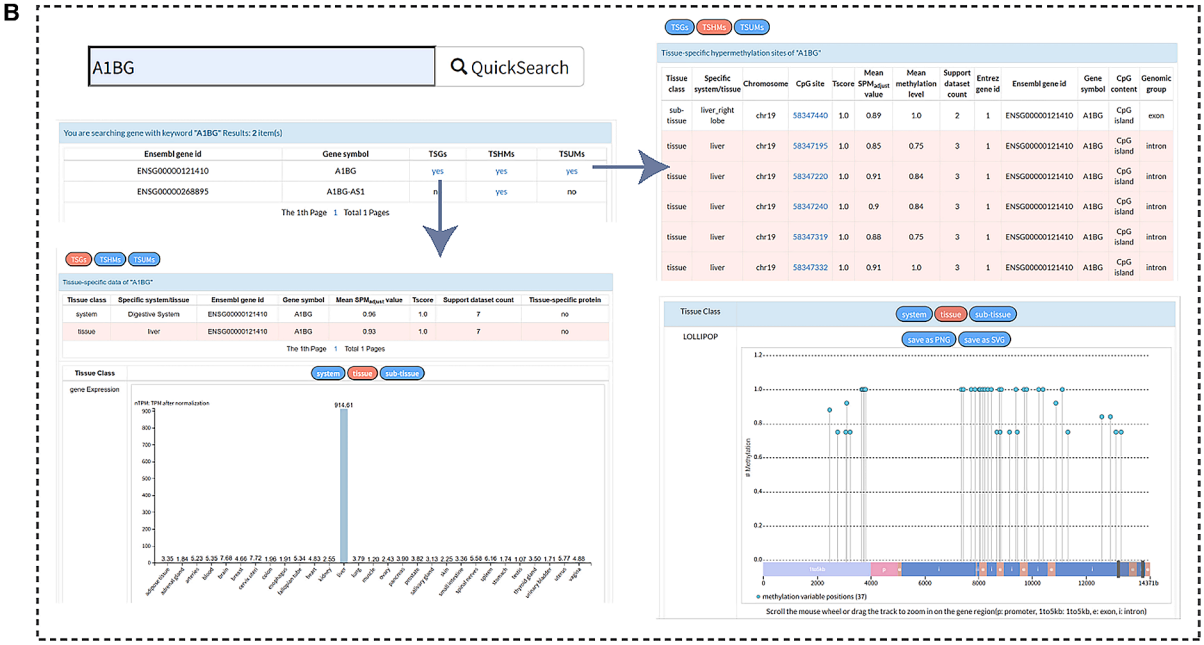
The search function supports multiple query modes to accommodate diverse research needs. Users can perform fuzzy keyword queries, allowing partial or complete gene symbol input, as well as a chromosome-based search enabling the retrieval of TSMs by inputting the chromosome number along with start and end coordinates. Wildcard characters (e.g., "*," "&," "?") are not supported. Upon conducting a keyword search, matching TSGs or TSMs are listed sequentially (Figure 3B). Search results for TSGs, TSHMs, and TSUMs are presented in three separate sections. The TSG detail page displays a table of key information, including the gene symbol, Ensembl gene ID, SPM_{adjust} value, Tscore, number of supporting datasets, and TSP status. The tissue-wide expression profile is illustrated in a bar chart below the table. Similarly, TSMs are displayed in a table providing methylation-specific details, including genomic coordinates, methylation levels, and CpG annotations. Below the table, a lollipop plot visualizes the CpG positions and methylation levels, equipped with interactive zooming functionality along the chromosome. Upon conducting a chromosome-based search, matching TSMs are listed sequentially (Figure 3C).

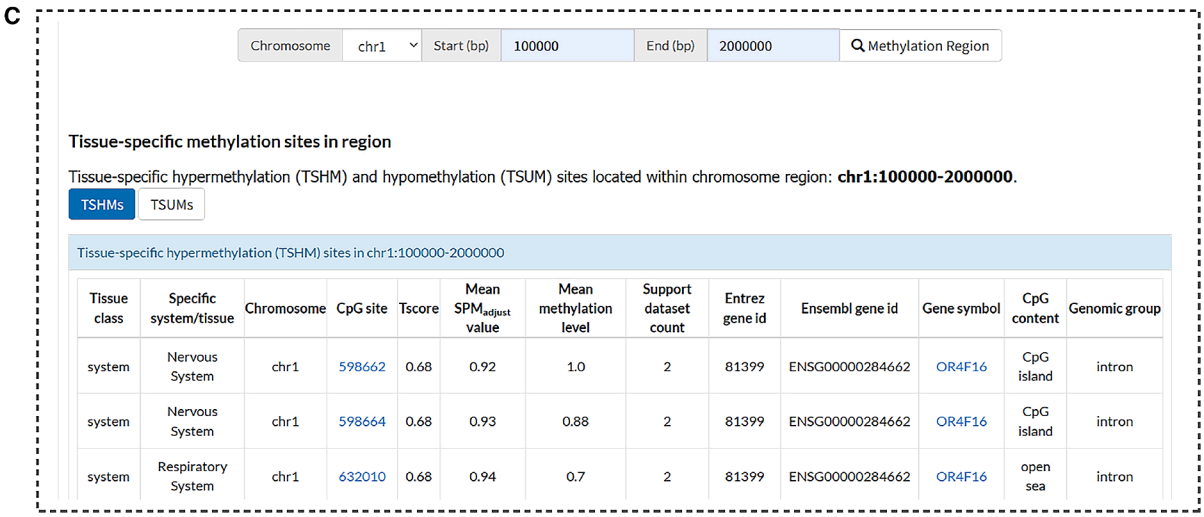
All data on TSMs, TSGs, TSPs, and HKMs can be freely downloaded in batch as CSV files via the Download page without requiring user registration.

Database comparison

To underscore the distinct utility of TiSMeD, we performed a comprehensive comparison with 17 existing methylation databases (Table S1). Most current repositories, such as PubMeth, MethyCancer, DiseaseMeth 3.0, and DNMIVD, primarily function to catalog disease-associated alterations. Other specialized databases, including MethAgingDB, ASMdb, and MethmiRbase, focus on specific biological dimensions such as aging dynamics, allele-specific methylation, or microRNA regulation. In contrast, TiSMeD is conceptually distinct, designed to provide a high-resolution normal reference map of tissue-specific identities. This baseline is critical for distinguishing pathological signals from tissue-of-origin background noise, a capability often limited in disease-centric repositories.







(legend on next page)

EWAS Data Hub represents a valuable effort toward estimating TSMs; however, its scope is constrained by its reliance on array-based methylation detection methods. The number of detectable methylation sites varies widely across different detection platforms. For instance, the 450K and 850K platforms can measure approximately 480,000 and 850,000 methylation sites, respectively, while RRBS and WGBS together provide data for an additional 17,141,905 methylation sites (Figure 4A). Notably, approximately 93.8% of the total 67,427 TSMs identified by TiSMeD fall outside the coverage of the 450K and 850K platforms (Figure 4B). Additionally, TiSMeD introduces the Tscore as a robust metric for assessing reliability and consistency in the identification of tissue-specific features, even when integrating data from diverse and heterogeneous sources. For example, the methylation site chr21:31558892 (TiSMeD_id: 16641573) exhibited liver-specific methylation consistently across three datasets, resulting in a Tscore of 1. In contrast, methylation site chr12:57604534 (TiSMeD_id: 279667) was included in six datasets but demonstrated liver-specific methylation in only one dataset, yielding a Tscore of 0.15 (Figure 4C).

The integrated presentation of TSMs, TSGs, and TSPs provides a unique platform for investigating the regulatory interplay linking methylation modifications to transcription and translation. Furthermore, TiSMeD takes the lead in presenting the information of tissue-specific features according to a hierarchical structure (system, tissue, sub-tissue), enabling the characterization of tissue-specific events across different tissue levels. Collectively, these features establish TiSMeD as a unique, advanced, and reliable resource for research into tissue-specific methylation and expression.

Example applications of TiSMeD

Tracing the tissue of origin of circulating cell-free DNA

cfDNA typically comprises short DNA fragments released into bodily fluids by apoptotic or necrotic cells originating from various tissues. These fragments retain TSM signatures, enabling accurate tracing of their tissue of origin. We performed tissue-of-origin analysis using a publicly available cfDNA methylome dataset consisting of four colon cancer and three breast cancer samples (GEO: GSE122126).[47] Leveraging 14,232 tissue-level TSMs curated in TiSMeD, we utilized CIBERSORT to deconvolute the tissue-of-origin contributions to cfDNA (Figure 5A).[48] Three of the four colon cancer cfDNA samples exhibited a predominant colon-derived signal. In the remaining sample, the majority of the cfDNA originated from blood, with approximately 4% from the colon. Similarly, within the breast cancer cohort, two of the three samples showed a dominant contribution from breast tissue. These findings highlight the utility of TiSMeD for accurately tracing tissue-of-origin using TSM signatures, demonstrating its strong potential for non-invasive diagnostics.

Discovery of methylation biomarkers for disease diagnosis

For effective tumor diagnosis, robust methylation biomarkers should accurately differentiate the target tumor type not only from normal tissues but also from other cancer types. To this end, we identified 468,482 cancer-specific differentially methylated sites (CSDMs) spanning 22 cancer types using 2,908 case-control samples retrieved from The Cancer Genome Atlas (TCGA)[49] database. By intersecting these sites with 14,232 tissue-level TSMs from TiSMeD, we prioritized 1,849 methylation sites exhibiting both tissue specificity and cancer-associated differential methylation, proposing them as candidate biomarkers. We performed unsupervised clustering analysis based on these candidate biomarkers using the uniform manifold approximation and projection (UMAP) algorithm in 1,213 TCGA samples, encompassing 10 cancer types along with their corresponding normal tissues.[50] Samples from the same cancer type clustered into compact and distinct groups, exhibiting clear separation from normal tissue samples. Furthermore, distinct cancer types displayed pronounced inter-group boundaries, facilitating precise classification (Figure 5B). Our findings underscore the pivotal value of incorporating TSM information in optimizing the discovery of methylation-based biomarkers.

Construction of a multi-cancer tracing and diagnosis model

To evaluate the clinical applicability of TiSMeD-derived methylation features, we developed a multi-cancer tracing and diagnosis (MCTD) model using a panel of 80 methylation markers selected from the 1,849 candidates identified above. The model was trained to classify samples into seven distinct categories, comprising colon cancer, kidney cancer, liver cancer, pancreatic cancer, prostate cancer, thyroid cancer, and normal tissues. Ten-fold cross-validation demonstrated an overall accuracy of 95.7%, a sensitivity of 99.2%, a specificity of 93.5%, and an area under the curve (AUC) of 0.995 (Figure 5C). The confusion matrix illustrated the classification performance of the model across different cancer types (Figure 5D). Notably, the diagnostic accuracy for colon cancer reached 98.7%, with 310 of 314 samples correctly identified. To assess the robustness of the model, we validated its performance across six independent external datasets. The model achieved accuracies spanning 88%–100% across these validation cohorts, with sensitivities of 88%–100% and specificities of 91.1%–100%. Furthermore, we assessed the model's performance using the plasma cfDNA dataset GEO: GSE122126,[47] achieving 87.5% accuracy in identifying colon cancer samples (Table S2). Finally, we compared the MCTD model with several representative multi-cancer detection and tissue-of-origin assays

Figure 3. Screenshots of the TiSMeD web interface

(A) Browsing interface for tissue-specific multi-omics data. Users can access TSGs, TSHMs, and TSUMs through a hierarchical tissue tree or an interactive human body schematic. Functional enrichment results for selected tissues are displayed using lollipop charts. (B) Keyword search function and detailed data visualization. The database supports fuzzy keyword queries for gene symbols. The results page provides integrated tables of tissue-specific features, gene expression bar charts, and interactive lollipop plots illustrating the genomic distribution and methylation levels of TSMs relative to target genes. (C) Screenshot of the TiSMeD web interface displaying chromosome-region query results.

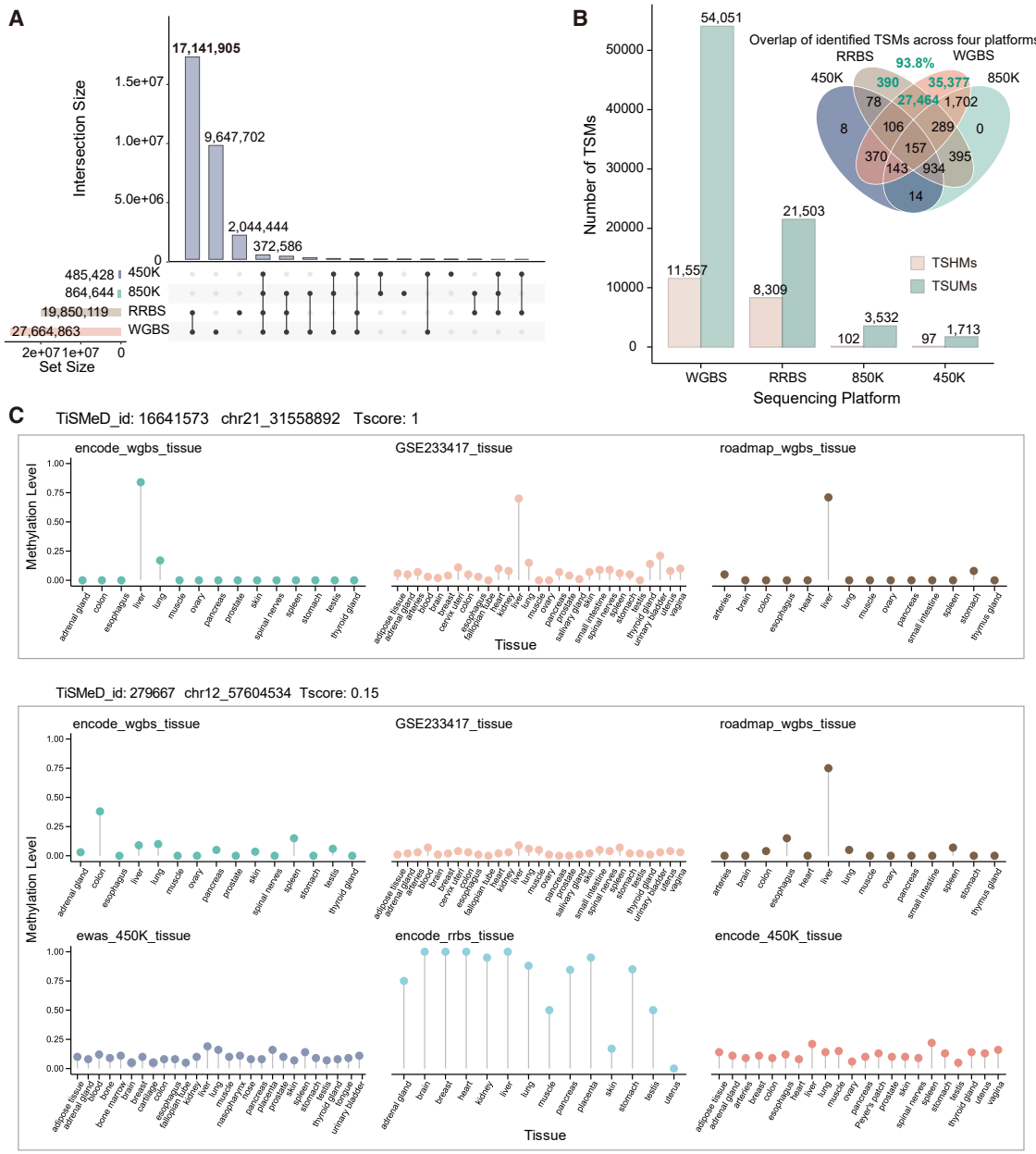


Figure 4. Comprehensive coverage and cross-platform robustness of TSM identification in TiSMeD

(A) Comparison of CpG site coverage across diverse DNA methylation platforms. (B) Identification and overlap of TSMs across different platforms. The bar chart summarizes the number of TSMs identified within each platform's datasets. The accompanying Venn diagram illustrates the high degree of overlap among TSMs detected across the four platforms. (C) Examples illustrating the methylation levels of two CpG sites in different datasets and their corresponding Tscore values.

reported between 2018 and 2025, including THUNDER-I, GutSeer, Galleri, PanSeer, and Spot-MAS (Table S3).[51–55] The MCTD model demonstrated competitive diagnostic performance while utilizing significantly fewer markers and leveraging explicit tissue specificity. These results underscore the utility of TiSMeD-derived markers and the applicability of the MCTD model for both tissue-based diagnosis and non-invasive cancer detection.

DISCUSSION

In this study, we present TiSMeD, a comprehensive database characterizing tissue-specific methylation and expression, derived from the systematic analysis of an extensive collection of methylomes, transcriptomes, and proteomes from normal human tissues. The TSMs, TSGs, and TSPs were identified on a genome-wide scale, encompassing nearly all major human tissues

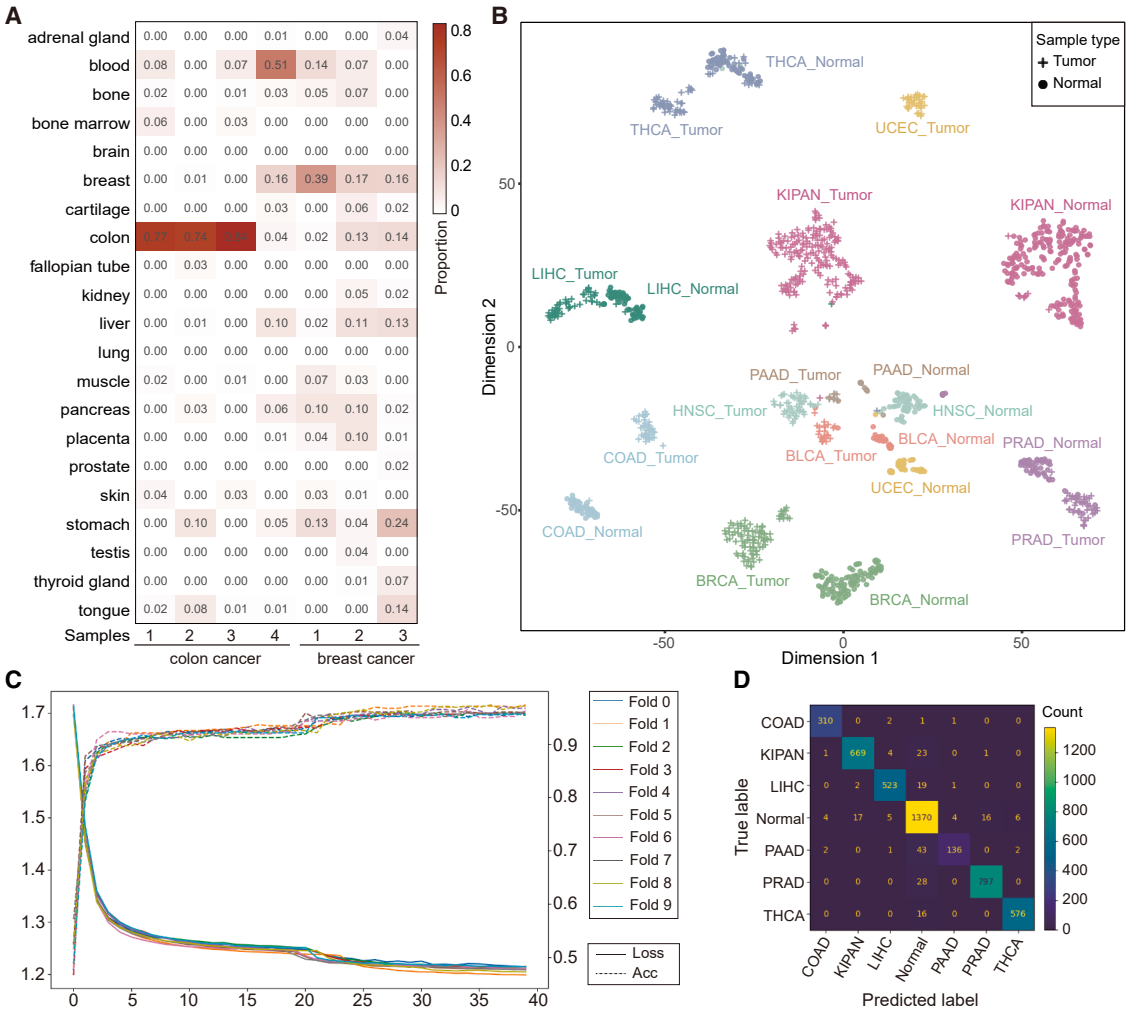


Figure 5. Example applications of TiSMeD

(A) Heatmap illustrating tissue-of-origin predictions for four colon cancer and three breast cancer cfDNA samples. (B) Results of unsupervised clustering using 1,849 candidate methylation markers across 1,213 TCGA samples, encompassing 10 cancer types and their corresponding normal tissues. (C) Ten-fold cross-validation performance of the MCTD model. (D) Confusion matrix of the MCTD model.

and sub-tissues. The reliability and consistency of tissue-specific feature identification are ensured by the rigorous assessment of both data heterogeneity and tissue coverage. To our knowledge, TiSMeD is the most comprehensive database to date specifically dedicated to TSM profiling. It complements existing methylation-related resources by providing valuable and reliable information for in-depth epigenetic research. Moreover, TiSMeD demonstrates significant utility for prioritizing candidate methylation features in biomarker discovery for disease diagnosis, identifying targets for precision therapy, and tracing the tissue of origin of cfDNA. Notably, TiSMeD distinguishes itself from existing databases by offering a unified platform that integrates tissue-specific methylation and expression data, thereby facilitating deeper insights into tissue-specific transcriptional regulation, tissue development, and functional differentiation.

Despite these strengths, several limitations warrant mention. First, certain tissues remain underrepresented or absent due to the scarcity of high-quality or multi-source data. Second, because our reference was built from bulk profiles of normal adult tissues, the database does not currently cover fetal tissues, pathological samples, or single-cell data. These limitations may constrain its applicability in studies focusing on developmental biology, intra-tumoral heterogeneity, or immune microenvironments.

Future iterations of TiSMeD aim to incorporate underrepresented tissues and extend the atlas to include cell-type-specific methylation features, leveraging emerging single-cell and spatial methylome datasets. To strictly validate our findings, we prioritize the experimental validation of selected TSMs in representative tissues using targeted bisulfite amplicon sequencing and pyrosequencing. This validation

Table 2. Summary of datasets integrated in TiSMeD

Omics	Data source	Year	Dataset number	Tissue number	Sample number	Data type	PMID
Transcriptome	GEO: GSE96	2002	1	17	50	Array	11904358
Transcriptome	GEO: GSE803	2003	5	10	24	Array	15388519
Transcriptome	GEO: GSE1133	2004	2	25	102	Array	15075390
Transcriptome	GEO: GSE2361	2005	1	25	32	Array	15950434
Transcriptome	GEO: GSE7905	2007	1	25	69	Array	19014478
Transcriptome	GEO: GSE14938	2011	1	28	32	Array	19534766
Transcriptome	GEO: GSE30611	2011	1	15	26	RNA-seq	NA
Transcriptome	RNA Seq Atlas	2012	1	11	11	RNA-seq	22345621
Transcriptome	HPA	2021	2	34	1,506	RNA-seq	25613900/32139519
Transcriptome	GTEx	2017	1	29	15,042	RNA-seq	26484569
Methylome	ENCODE	2021	1	22	85	450K	22955616/31713622
Methylome	ENCODE	2021	1	14	37	RRBS	22955616/31713622
Methylome	ENCODE	2021	1	15	25	WGBS	22955616/31713622
Methylome	EWAS Data Hub	2021	210	27	5,023	450K	31584095
Methylome	ROADMAP	2015	1	7	17	RRBS	25693563
Methylome	ROADMAP	2015	1	14	17	WGBS	25693563
Methylome	GEO: GSE50192	2014	1	12	70	450K	24690455/27004446
Methylome	GEO: GSE213478	2022	1	9	987	850K	36510025
Methylome	GEO: GSE233417	2023	1	29	521	RRBS	37399400
Proteome	PRIDE: PXD016999 (GTEx)	2020	1	23	194	LC-MS/MS	32916130
Proteome	PRIDE: PXD010154 (HPA)	2019	1	27	29	LC-MS/MS	30777892
Proteome	PRIDE: PXD000561	2014	1	17	18	LC-MS/MS	24870542

will provide locus-specific confirmation and further support the biological reliability of the computationally predicted TSMs. In parallel, we will advance from threshold-based scoring toward machine learning and deep learning approaches capable of identifying tissue-specific patterns directly from high-dimensional multi-omics data. These models may help identify regulatory features beyond linear associations and better resolve tissue identity in complex or ambiguous samples.

MATERIALS AND METHODS

Data collection

Methylome datasets of normal human tissues were primarily obtained from the Encyclopedia of DNA Elements (ENCODE),[56–58] the Roadmap Epigenomics Project,[59] the EWAS Data Hub,[21,22] and the Gene Expression Omnibus (GEO).[60,61] These methylomes were generated using four platforms: RRBS, WGBS, 450K, and 850K arrays. Nine distinct methylome datasets were collected, comprising 6,782 samples from 42 normal tissues and 59 sub-tissues. The transcriptome datasets were obtained from GEO,[60,61] the Genotype-Tissue Expression (GTEx) database,[62] the Human Protein Atlas (HPA),[63] and the RNA-seq Atlas.[64] Ten transcriptomic datasets were collected, including five generated using microarrays and five using RNA-seq. These datasets encompassed 16,894 samples derived from 43 normal tissues and 191 sub-tissues. Additionally, three proteome datasets were obtained from GTEx, HPA, and PRoteomics IDEntifications (PRIDE) databases,[65–68] covering 241 samples from 34 distinct normal tissues and 25 sub-tissues. Detailed information on these datasets is summarized in Table 2.

Methylome datasets of cancers and corresponding controls, utilized for the construction of the MCTD model and cancer biomarker discovery, were primarily obtained from the EWAS Data Hub,[21,22] TCGA,[49] and GEO.[60,61] These datasets collectively encompassed 21,238 samples representing 39 distinct cancer types. Detailed information on these datasets is summarized in Table S4.

Data processing

Unification of tissue nomenclature

Prior to data integration, tissue nomenclature was standardized to ensure consistency across datasets. A hierarchical structure consisting of system, tissue, and sub-tissue levels was established based on the anatomical hierarchy defined by Medical Subject Headings (MeSH). Tissue names were manually curated and standardized within this hierarchical framework. For example, the term “transverse colon” was standardized as “colon_transverse colon.” A complete list of standardized tissue names used in TiSMeD is available in the supplemental information (Table S5). Fetal tissues were excluded from the study.

Processing and integration of multi-omics data

For 450K and 850K array data, all IDAT files were processed using the minfi package (v.1.38.0).[69] Probes with fewer than three beads in ≥5% of samples, probes containing common SNPs at the CpG site, cross-reactive or multi-mapping probes, non-CpG probes, and probes located on sex chromosomes were removed. Only probes with detection p values <0.01 were retained for downstream analysis. The DNA methylation level at each CpG is referred to as a beta value, calculated as (M/(M + U)), where M represents the methylated allele intensity and U represents the unmethylated allele intensity. Beta values range from 0 to 1, with values near 0 indicating low methylation levels and those near 1 indicating high methylation levels. Gaussian Mixture Quantile Normalization (GMQN) was applied to remove batch effects and other unwanted technical variations.[70] The annotation files (HM450.hg38.manifest.gencode.v36 and EPIC.hg38.manifest.gencode.v36) were downloaded from the Genomic Data Commons project (https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files,[71] accessed on February 3, 2024) to facilitate methylation site annotation.

For RRBS and WGBS data, we followed the ENCODE uniform workflow based on gemBS. Reads were aligned to the GRCh38.95 genome using GEM-Mapper (v.3) with default parameters. Prior to the calling step, BScall (v.2.02) was used to remove duplicated reads, collapse overlapping paired-end reads, and trim read-end cytosines known to be artificially generated during library preparation. Cytosine methylation levels were calculated using BScall. Samples with C-to-T conversion rates below 98% were excluded. For integration, we focused exclusively on methylation sites located on autosomes. Replicate tissue samples from the same individual were merged, and the methylation level for each CpG site was calculated as the average methylation value across the corresponding replicates. Measurement sites from WGBS and RRBS samples were integrated by taking the union, and sites with a missing rate greater than 50% were removed. Missing methylation values were treated as 0. CpG and genic features of the methylated sites were annotated using the annotate_regions function from the Bioconductor package annotatr (v.1.1.3).[72] The CpG features included islands, shores, shelves, and open sea regions, while genic features encompassed regions located 1–5 kb upstream of the transcription start site (TSS), promoter regions (<1 kb upstream of TSS), 5′ UTRs, exons, introns, and 3′ UTRs.

For RNA-seq datasets with raw sequencing data, initial quality assessment was performed using FastQC (v.0.11.6, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapter sequences were trimmed using NGSQC (v.2.3.3),[73] and low-quality reads were filtered out using Trimmomatic (v.0.38).[74] Potential microbial contamination was filtered using Kraken (v.1.0).[75] The remaining clean reads were aligned to the human reference genome (GRCh38.95 assembly) using HISAT2 (v.2.1.0) with default parameters.[76] Mapped reads were assembled into transcripts using StringTie (v.1.3.4d) with reference to the genome annotation file (hg38_ucsc.annotated.gtf).[77] Transcript abundance was quantified as transcripts per kilobase million (TPM). For RNA-seq transcriptomes with available expression matrices, gene expression levels were converted into TPM values. We performed trimmed mean of M values normalization to account for batch differences between RNA-seq datasets.[78] For microarray transcriptomes with expression matrices, signal intensity values were log-transformed and missing values were imputed as 0. If multiple probes corresponded to the same gene, the average intensity of these probes was used to represent the gene's expression level in the transcriptome. Replicate tissue samples from the same individual were merged, and the average expression value was calculated to represent the gene's expression level. In this study, genes with TPM <1.0 or intensity <1.0 were defined as not expressed.

For proteomes with available expression matrices, proteins were annotated according to the GRCh38.95 genome. Replicate tissue samples from the same individual were averaged to determine the final protein expression levels.

To avoid overestimation of tissue specificity due to limited expression coverage, features detected exclusively in a single tissue within any dataset were removed from the expression matrix before proceeding with tissue-specificity evaluation.

Identification of tissue-specific methylation sites, genes, and proteins

Evaluation of tissue specificity

In our previous study,[43] we introduced a statistical parameter termed the SPM, which demonstrated promising performance in evaluating tissue specificity of genes. In the present study, we adapted the SPM method to identify tissue-specific features across multiple omics layers (Figure 6). Taking DNA methylation as an example, the methylation profile of a specific site x across a dataset was transformed into a high-dimensional vector X_p:

$$X_p = (x_1, x_2, x_3, \ldots, x_i, \ldots, x_n), \quad \text{(Equation 1)}$$

where x_i represents the average methylation level of site x in tissue i, and n represents the total number of tissues analyzed. Consequently, the component corresponding to a specific tissue i can be viewed as a vector X_i projected onto the tissue axis within this space:

$$X_i = (0, 0, 0, \ldots, x_i, \ldots, 0). \quad \text{(Equation 2)}$$

The tissue specificity of site x in tissue i is quantified as the ratio of the scalar projection of vector X_i onto the profile vector X_p ($||X||$) to the magnitude of X_p ($||X_p||$):

$$||X|| = \left| \frac{(X_i \cdot X_p)}{||X_p||} \right|, \quad \text{(Equation 3)}$$

$$\mathrm{SPM} = \frac{||X||}{||X_p||} = \frac{{x_i}^2}{\sum_{i=1}^{n} {x_i}^2}. \quad \text{(Equation 4)}$$

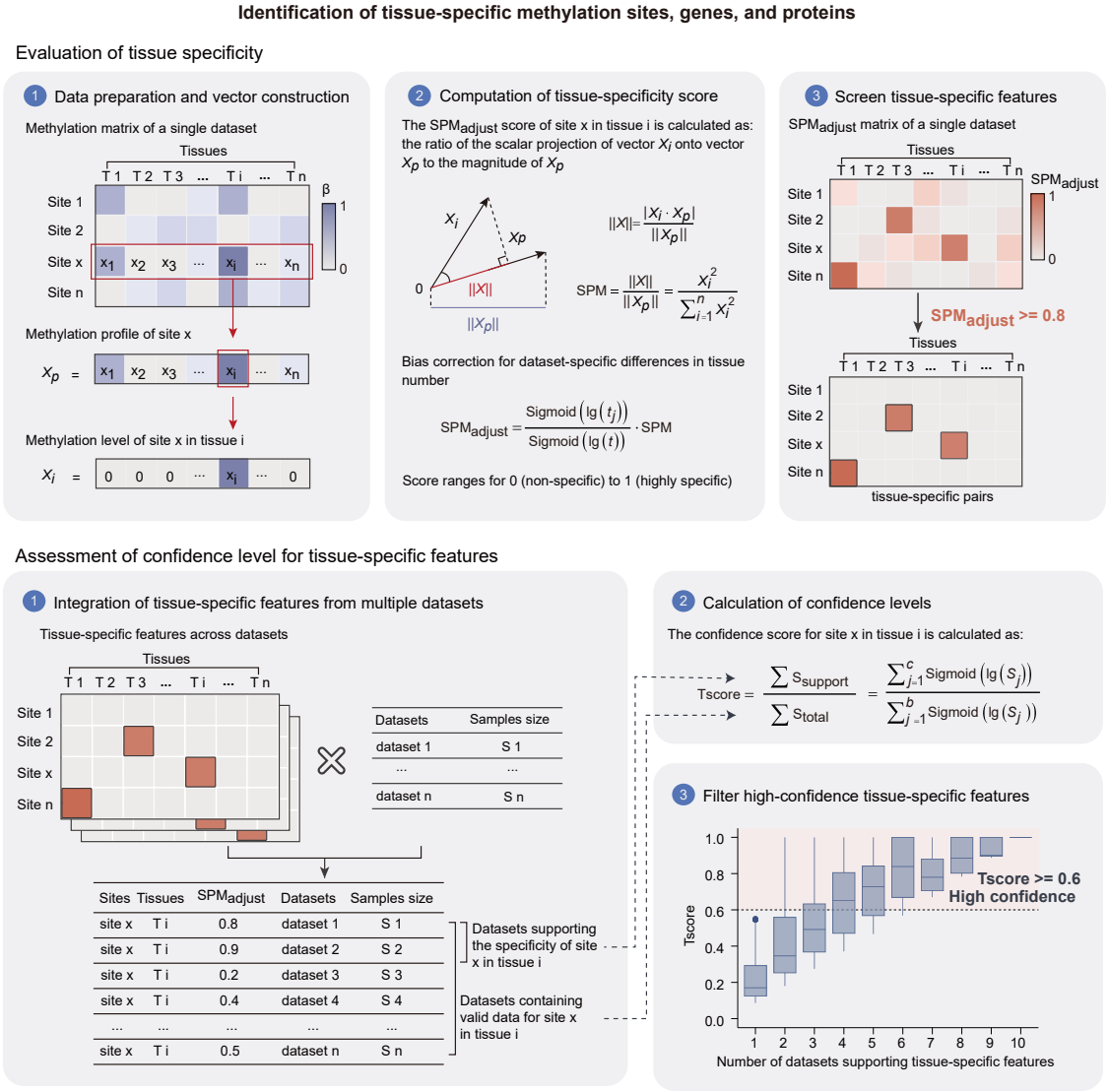


Figure 6. Workflow for identifying high-confidence tissue-specific features

Upper panel: Evaluation of tissue specificity. A threshold of $SPM_{adjust} \geq 0.8$ is utilized to screen for initial tissue-specific features. Lower panel: Confidence-level assessment for tissue-specific features. Features exhibiting a Tscore ≥0.6 are categorized as high-confidence tissue-specific features.

To identify TSUMs, the methylation profile of site x was inverted to quantify unmethylation specificity:

$$X'_p = ((1 - x_1), (1 - x_2), (1 - x_3), \ldots, (1 - x_i), \ldots, (1 - x_n)). \quad \text{(Equation 5)}$$

Theoretically, the SPM value ranges from 0 to 1.0, with values approaching 1.0 indicating higher specificity to the target tissue. However, variations in the number of tissues included across different datasets can introduce bias into the raw SPM scores. To correct for this dimensionality bias, we introduced an adjusted metric termed SPM_{adjust}:

$$\text{Sigmoid}(z) = \frac{1}{(1+e^{-z})}, \quad \text{(Equation 6)}$$

$$SPM_{adjust} = \frac{\text{Sigmoid}(\lg(t_j))}{\text{Sigmoid}(\lg(t))} SPM, \quad \text{(Equation 7)}$$

where t_j denotes the number of tissues included in dataset j and t represents the total number of unique tissues integrated across all datasets. Similar to the raw score, the SPM_{adjust} value ranges from 0 to 1.0.

Assessment of confidence level

Integrating data from heterogeneous sources requires addressing discrepancies in sample sizes and dataset quality. To address these discrepancies in the evaluation of tissue specificity, we introduced

a metric termed the Tscore to quantify the confidence level of the identified features. The Tscore is calculated as:

$$\text{Tscore} = \frac{\sum_{j=1}^{c} \text{Sigmoid}\left(\lg\left(s_j\right)\right)}{\sum_{j=1}^{b} \text{Sigmoid}\left(\lg\left(s_j\right)\right)}, \quad \text{(Equation 8)}$$

where c denotes the number of datasets supporting the tissue specificity of the methylation site x in tissue i; b denotes the total number of datasets that containing valid data for site x in tissue i, and s_j represents the sample size of datasets. The Tscore ranges from 0 to 1.0, with higher values indicating greater consensus and confidence.

Identification of tissue-specific methylation sites, genes, and proteins

Based on these metrics, we established specific thresholds to categorize tissue-specific features. TSMs were stratified into two groups: TSHMs and TSUMs. TSHMs were identified by an $SPM_{adjust} \geq 0.8$, an average methylation level ≥ 0.7, and a Tscore ≥ 0.6, whereas TSUMs were defined by an $SPM_{adjust} \geq 0.8$, an average methylation level ≤ 0.3, and a Tscore ≥ 0.6. Similarly, TSGs and TSPs were defined using the criteria of $SPM_{adjust} \geq 0.8$ and Tscore ≥ 0.6.

Additionally, HKMs were identified, including housekeeping hypermethylation sites, defined as sites consistently hypermethylated across all tissues (methylation levels ≥ 0.7 and Tscore ≥ 0.6), and housekeeping unmethylation sites, defined as sites consistently hypomethylated or unmethylated across all tissues (methylation levels ≤ 0.3 and Tscore ≥ 0.6).

Down-sampling analysis for evaluating sample size robustness

To assess the potential bias introduced by uneven sample sizes among different tissues, we performed a resampling analysis using the large-scale RRBS dataset GEO: GSE233417, where tissue sample sizes ranged from 6 to 57. We standardized the dataset to a uniform sample size ($n = 6$) by conducting random subsampling across all eligible tissues. This procedure was repeated for 100 iterations to ensure statistical reliability. For each iteration, we identified TSMs and calculated the mean abundance of TSHMs and TSUMs. The robustness of the identified patterns was subsequently evaluated by calculating the Spearman's rank correlation coefficient between the vector of TSM counts derived from the down-sampled analysis and the vector obtained from the complete dataset.

Database implementation

TiSMeD was developed based on a Linux-Apache-JSP server architecture. MySQL (v.5.7.25) was utilized for data storage, access, and management. User-friendly interfaces were developed using JavaScript to enable interactive data retrieval and visualization. Tissue diagrams were generated using the R package gganatogram (v.1.1.1).[79]

Multi-omics correlation analysis

To evaluate the regulatory associations between DNA methylation and gene expression, TSMs were mapped to their nearest genes based on genomic annotations encompassing promoters (defined as ±2 kb from the TSS), gene bodies, and flanking regions (1–5 kb upstream or downstream). Spearman rank correlation coefficients were calculated between DNA methylation and gene expression levels for each TSM-gene pair across matched tissues. p values were adjusted using the Benjamini-Hochberg procedure, and only correlations with a false discovery rate (FDR) < 0.05 were retained. To assess transcriptome-proteome concordance, Spearman correlations were computed between mRNA expression and protein abundance for identified TSGs. Additionally, the cross-layer consistency of tissue specificity was evaluated by analyzing the overlap between TSGs and TSPs.

Construction and evaluation of the multi-cancer tracing and diagnosis model

Feature selection

Robust biomarkers for simultaneous cancer diagnosis and tissue-of-origin tracing were prioritized based on tissue specificity and differential methylation in cancers. CSDMs were identified using the limma package (v.3.64.3),[80] with significance defined as a Benjamini-Hochberg adjusted p value <0.05 and an absolute difference in β values >0.2. The marker candidate pool was defined as the intersection of identified TSMs and CSDMs. From this intersection, the top 10 markers with the highest Tscores for each of the six cancer types were selected. Additionally, 24 methylation sites exhibiting significant differences between pan-cancer samples and normal tissues were included to enhance the distinction between malignant and non-malignant states. These selection steps resulted in a final panel of 80 methylation sites for model construction.

Model architecture and training

A multi-layer perceptron model was constructed using the PyTorch framework (v.2.1.0) and Python (v.3.11.6).[81] The architecture comprised an input layer corresponding to the 80 selected markers, a single hidden layer of 1,024 neurons with a ReLU activation function, and an output layer for classifying samples into seven categories (six cancer types and healthy controls). Weights were initialized using the Kaiming uniform method. To mitigate overfitting and improve generalization, a dropout layer with a probability of 0.1 was implemented. The model was trained using the cross-entropy loss function and the Adam optimizer with a learning rate of 0.00005.

Model evaluation

Model robustness and generalizability were assessed using stratified 10-fold cross-validation. The dataset was randomly partitioned into 10 subsets with balanced class distributions. In each iteration, nine subsets were utilized for training and one for validation; this process was repeated 10 times to calculate average performance. The comprehensive performance of the MCTD model was evaluated using accuracy, sensitivity, specificity, and AUC.

DATA AND CODE AVAILABILITY

TiSMeD is accessible at http://www.bio-add.org/TiSMeD/. The source codes for analysis are publicly available in the GitHub repository at https://github.com/BADD-XMU/TiSMeD.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (grant no. 2024YFF1206802), National Natural Science Foundation of China (grant no. 82201851), Shenzhen Key Medical Discipline Construction Fund (grant no. SZXK028), and Shenzhen Science and Technology Program (grant no. RCYX20210609104608036).

AUTHOR CONTRIBUTIONS

J.C. conducted the investigation, performed data curation, contributed to formal analysis and visualization, and drafted the original manuscript; Z.L. constructed the database; L.W. and H.Y. contributed to data curation; Q.L. and H.C. assisted with database visualization; H.W., X.C., and Z.-L.J. contributed to manuscript review and editing; X.C. and Z.-L.J. were responsible for funding acquisition. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2026.102884.

REFERENCES

1. Kaluscha, S., Domcke, S., Wirbelauer, C., Stadler, M.B., Durdu, S., Burger, L., and Schübeler, D. (2022). Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. Nat. Genet. *54*, 1895–1906.
2. Rimoldi, M., Wang, N., Zhang, J., Villar, D., Odom, D.T., Taipale, J., Flicek, P., and Roller, M. (2024). DNA methylation patterns of transcription factor binding regions characterize their functional and evolutionary contexts. Genome Biol. *25*, 146.
3. Klughammer, J., Romanovskaia, D., Nemc, A., Posautz, A., Seid, C.A., Schuster, L.C., Keinath, M.C., Lugo Ramos, J.S., Kosack, L., Evankow, A., et al. (2023). Comparative analysis of genome-scale, base-resolution DNA methylation profiles across 580 animal species. Nat. Commun. *14*, 232.
4. Blake, L.E., Roux, J., Hernando-Herraez, I., Banovich, N.E., Perez, R.G., Hsiao, C.J., Eres, I., Cuevas, C., Marques-Bonet, T., and Gilad, Y. (2020). A comparison of gene expression and DNA methylation patterns across tissues and species. Genome Res. *30*, 250–262.
5. Yang, Y., Chen, Y., Xu, S., Guo, X., Jia, G., Ping, J., Shu, X., Zhao, T., Yuan, F., Wang, G., et al. (2024). Integrating muti-omics data to identify tissue-specific DNA methylation biomarkers for cancer risk. Nat. Commun. *15*, 6071.
6. García-Calzón, S., Maguolo, A., Eichelmann, F., Edsfeldt, A., Perfilyev, A., Maziarz, M., Lindström, A., Sun, J., Briviba, M., Schulze, M.B., et al. (2025). Epigenetic biomarkers predict macrovascular events in individuals with type 2 diabetes. Cell Rep. Med. *6*, 102290.
7. Mo, S., Ye, L., Wang, D., Han, L., Zhou, S., Wang, H., Dai, W., Wang, Y., Luo, W., Wang, R., et al. (2023). Early Detection of Molecular Residual Disease and Risk Stratification for Stage I to III Colorectal Cancer via Circulating Tumor DNA Methylation. JAMA Oncol. *9*, 770–778.
8. Baca, S.C., Seo, J.H., Davidsohn, M.P., Fortunato, B., Semaan, K., Sotudian, S., Lakshminarayanan, G., Diossy, M., Qiu, X., El Zarif, T., et al. (2023). Liquid biopsy epigenomic profiling for cancer subtyping. Nat. Med. *29*, 2737–2741.
9. Truong, T.T., Mikloska, K., Sum, J., Oberländer, M., von Bubnoff, N., Christiansen, L., Tornow, S., Derer, S., Janke, F., Sültmann, H., et al. (2025). Coupling Immunoprecipitation with Multiplexed Digital PCR for Cell-Free DNA Methylation Detection in Small Plasma Volumes of Early-Onset Colorectal Cancer. Anal. Chem. *97*, 11259–11268.
10. Saeki, I., Suehiro, Y., Yamauchi, Y., Hoshida, T., Tanabe, N., Oono, T., Kawamoto, D., Nishimura, T., Matsumoto, T., Ishikawa, T., et al. (2023). Methylated SEPT9 assay-based liquid biopsy as a biomarker in molecular targeted agent-treated hepatocellular carcinoma. Hepatol. Int. *17*, 1289–1299.
11. Dietrich, D., Weider, S., de Vos, L., Vogt, T.J., Färber, M., Zarbl, R., Hunecke, A., Glosch, A.K., Gabrielpillai, J., Bootz, F., et al. (2023). Circulating Cell-Free SEPT9 DNA Methylation in Blood Is a Biomarker for Minimal Residual Disease Detection in Head and Neck Squamous Cell Carcinoma Patients. Clin. Chem. *69*, 1050–1061.
12. Jahan, M.I., Inenaga, T., Makimoto, S., Hossain, M.B., Matsuoka, Y., Sithi, S.N., Rajib, S.A., Ansori, A.N.M., Sugata, K., Imakawa, K., et al. (2025). Circulating Cell-Free DNA of Bovine Leukemia Virus: A Promising Biomarker for Enzootic Bovine Leukosis. Microbiol. Immunol. *69*, 466–476.
13. Conway, A.M., Pearce, S.P., Clipson, A., Hill, S.M., Chemi, F., Slane-Tan, D., Ferdous, S., Hossain, A.S.M.M., Kamieniecka, K., White, D.J., et al. (2024). A cfDNA methylation-based tissue-of-origin classifier for cancers of unknown primary. Nat. Commun. *15*, 3292.
14. Bie, F., Wang, Z., Li, Y., Guo, W., Hong, Y., Han, T., Lv, F., Yang, S., Li, S., Li, X., et al. (2023). Multimodal analysis of cell-free DNA whole-methylome sequencing for cancer detection and localization. Nat. Commun. *14*, 6042.
15. Nguyen, T.H., Doan, N.N.T., Tran, T.H., Huynh, L.A.K., Doan, P.L., Nguyen, T.H.H., Nguyen, V.T.C., Nguyen, G.T.H., Nguyen, H.N., Giang, H., et al. (2024). 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. J. Transl. Med. *22*, 618.
16. Li, D., Zhang, L., Fu, J., Huang, H., Liu, Y., Zhu, L., Sun, H., Sun, S., Zhang, D., Tian, T., et al. (2022). Discovery and validation of tissue-specific DNA methylation as noninvasive diagnostic markers for colorectal cancer. Clin. Epigenetics *14*, 102.
17. Grunau, C., Renault, E., Rosenthal, A., and Roizes, G. (2001). MethDB–a public database for DNA methylation data. Nucleic Acids Res. *29*, 270–274.
18. Song, Q., Decato, B., Hong, E.E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J., and Smith, A.D. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One *8*, e81148.
19. Lebrón, R., Gómez-Martín, C., Carpena, P., Bernaola-Galván, P., Barturen, G., Hackenberg, M., and Oliver, J.L. (2017). NGSmethDB 2017: enhanced methylomes and differential methylation. Nucleic Acids Res. *45*, D97–D103.
20. Zhang, M., Zong, W., Zou, D., Wang, G., Zhao, W., Yang, F., Wu, S., Zhang, X., Guo, X., Ma, Y., et al. (2023). MethBank 4.0: an updated database of DNA methylation across a variety of species. Nucleic Acids Res. *51*, D208–D216.
21. Xiong, Z., Li, M., Yang, F., Ma, Y., Sang, J., Li, R., Li, Z., Zhang, Z., and Bao, Y. (2020). EWAS Data Hub: a resource of DNA methylation array data and metadata. Nucleic Acids Res. *48*, D890–D895.
22. Xiong, Z., Yang, F., Li, M., Ma, Y., Zhao, W., Wang, G., Li, Z., Zheng, X., Zou, D., Zong, W., et al. (2022). EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. Nucleic Acids Res. *50*, D1004–D1009.
23. Zong, W., Kang, H., Xiong, Z., Ma, Y., Jin, T., Gong, Z., Yi, L., Zhang, M., Wu, S., Wang, G., et al. (2022). scMethBank: a database for single-cell whole genome DNA methylation maps. Nucleic Acids Res. *50*, D380–D386.
24. Li, S., Tang, S., Ma, H., Wang, H., and Chen, S. (2025). MethAgingDB: a comprehensive DNA methylation database for aging biology. Sci. Data *12*, 1216.
25. Zhou, Q., Guan, P., Zhu, Z., Cheng, S., Zhou, C., Wang, H., Xu, Q., Sung, W.K., and Li, G. (2022). ASMdb: a comprehensive database for allele-specific DNA methylation in diverse organisms. Nucleic Acids Res. *50*, D60–D71.
26. Agustriawan, D., Wijaya, E.B., Huang, C.-H., Lim, E., Hsueh, I.-C., Kurubanjerdjit, N., Tzeng, K.-R., and Ng, K.-L. (2016). MethmiRbase: a database of DNA methylation and miRNA expression in human cancer. Lect. Notes Eng. Comput. Sci. *2221*, 73–76.
27. Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S., and Van Criekinge, W. (2008). PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Res. *36*, D842–D846.
28. He, X., Chang, S., Zhang, J., Zhao, Q., Xiang, H., Kusonmano, K., Yang, L., Sun, Z.S., Yang, H., and Wang, J. (2008). MethyCancer: the database of human DNA methylation and cancer. Nucleic Acids Res. *36*, D836–D841.
29. Fang, Y.C., Lai, P.T., Dai, H.J., and Hsu, W.L. (2011). MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. BMC Bioinf. *12*, 471.
30. Baek, S.J., Yang, S., Kang, T.W., Park, S.M., Kim, Y.S., and Kim, S.Y. (2013). MENT: methylation and expression database of normal and tumor tissues. Gene *518*, 194–200.

31. Huang, H.Y., Li, J., Tang, Y., Huang, Y.X., Chen, Y.G., Xie, Y.Y., Zhou, Z.Y., Chen, X.Y., Ding, S.Y., Luo, M.F., et al. (2021). MethHC 2.0: information repository of DNA methylation and gene expression in human cancer. Nucleic Acids Res. *49*, D1268–D1275.

32. Bin Raies, A., Mansour, H., Incitti, R., and Bajic, V.B. (2015). DDMGD: the database of text-mined associations between genes methylated in diseases from different species. Nucleic Acids Res. *43*, D879–D886.

33. Xing, J., Zhai, R., Wang, C., Liu, H., Zeng, J., Zhou, D., Zhang, M., Wang, L., Wu, Q., Gu, Y., and Zhang, Y. (2022). DiseaseMeth version 3.0: a major expansion and update of the human disease methylation database. Nucleic Acids Res. *50*, D1208–D1215.

34. Ding, W., Chen, J., Feng, G., Chen, G., Wu, J., Guo, Y., Ni, X., and Shi, T. (2020). DNMIVD: DNA methylation interactive visualization database. Nucleic Acids Res. *48*, D856–D862.

35. Peters, T.J., Meyer, B., Ryan, L., Achinger-Kawecka, J., Song, J., Campbell, E.M., Qu, W., Nair, S., Loi-Luu, P., Stricker, P., et al. (2024). Characterisation and reproducibility of the HumanMethylationEPIC v2.0 BeadChip for DNA methylation profiling. BMC Genom. *25*, 251.

36. Huminiecki, L., Lloyd, A.T., and Wolfe, K.H. (2003). Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. BMC Genom. *4*, 31.

37. Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics *21*, 650–659.

38. Lüleci, H.B., and Yılmaz, A. (2022). Robust and rigorous identification of tissue-specific genes by statistically extending tau score. BioData Min. *15*, 31.

39. Ceriani, L., and Verme, P. (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. J. Econ. Inequal. *10*, 421–443.

40. Schug, J., Schuller, W.P., Kappen, C., Salbaum, J.M., Bucan, M., and Stoeckert, C.J., Jr. (2005). Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. *6*, R33.

41. Yu, X., Lin, J., Zack, D.J., and Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. Nucleic Acids Res. *34*, 4925–4936.

42. Vandenbon, A., and Nakai, K. (2010). Modeling tissue-specific structural patterns in human and mouse promoters. Nucleic Acids Res. *38*, 17–25.

43. Xiao, S.J., Zhang, C., Zou, Q., and Ji, Z.L. (2010). TiSGeD: a database for tissue-specific genes. Bioinformatics *26*, 1273–1275.

44. Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., Daish, T., Grützner, F., and Kaessmann, H. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol. *10*, e1001328.

45. Everaert, C., Volders, P.J., Morlion, A., Thas, O., and Mestdagh, P. (2020). SPECS: a non-parametric method to identify tissue-specific molecular features for unbalanced sample groups. BMC Bioinf. *21*, 58.

46. Surana, P., Dutta, P., and Davuluri, R.V. (2024). TransTEx: novel tissue-specificity scoring method for grouping human transcriptome into different expression groups. Bioinformatics *40*, btae475.

47. 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.

48. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

49. Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

50. Healy, J., and McInnes, L. (2024). Uniform manifold approximation and projection. Nat Rev Method Prime *4*, 82.

51. 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.

52. Gao, Q., Li, B., Xu, J., Fang, S., Qiu, F., Su, J., Wang, G., Cai, S., Zhao, H., Zhang, W., et al. (2021). Analysis of epigenomic signatures in cell-free DNA (cfDNA) from cancer patients and high-risk controls: A blinded test cohort of THUNDER-II study. J. Clin. Oncol. *39*, e22518.

53. Klein, E.A., Richards, D., Cohn, A., Tummala, M., Lapham, R., Cosgrove, D., Chung, G., Clement, J., Gao, J., Hunkapiller, N., et al. (2021). Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann. Oncol. *32*, 1167–1177.

54. Nguyen, V.T.C., Nguyen, T.H., Doan, N.N.T., Pham, T.M.Q., Nguyen, G.T.H., Nguyen, T.D., Tran, T.T.T., Vo, D.L., Phan, T.H., Jasmine, T.X., et al. (2023). Multimodal analysis of methylomics and fragmentomics in plasma cell-free DNA for multi-cancer early detection and localization. eLife *12*, RP89083.

55. Tran, L.S., Nguyen, T.H.H., Nguyen, L.H.D., Nguyen, T.H., Nguyen, T.D., Giang, H., Phan, M.D., Nguyen, D.S., Tang, H.S., and Nguyen, H.-N. (2024). Comprehensive validation of a cell-free DNA-based assay for multi-cancer detection: A Vietnamese longitudinal prospective cohort study of 9,024 participants. J. Clin. Oncol. *42*, 102.

56. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

57. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*, D882–D889.

58. Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., Kaul, R., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

59. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

60. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995.

61. Clough, E., Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., et al. (2024). NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. Nucleic Acids Res. *52*, D138–D144.

62. Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

63. Sjöstedt, E., Zhong, W., Fagerberg, L., Karlsson, M., Mitsios, N., Adori, C., Oksvold, P., Edfors, F., Limiszewska, A., Hikmet, F., et al. (2020). An atlas of the protein-coding genes in the human, pig, and mouse brain. Science *367*, eaay5947.

64. Krupp, M., Marquardt, J.U., Sahin, U., Galle, P.R., Castle, J., and Teufel, A. (2012). RNA-Seq Atlas–a reference database for gene expression profiling in normal tissue by next-generation sequencing. Bioinformatics *28*, 1184–1185.

65. Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

66. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D.J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. *50*, D543–D552.

67. Perez-Riverol, Y., Bandla, C., Kundu, D.J., Kamatchinathan, S., Bai, J., Hewapathirana, S., John, N.S., Prakash, A., Walzer, M., Wang, S., and Vizcaíno, J.A. (2025). The PRIDE database at 20 years: 2025 update. Nucleic Acids Res. *53*, D543–D553.

68. Deutsch, E.W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J.J., Mendoza, L., Kundu, D.J., Wang, S., Bandla, C., Kamatchinathan, S., et al. (2022). The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res. *51*, D1539–D1548.

69. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.

70. Xiong, Z., Li, M., Ma, Y., Li, R., and Bao, Y. (2021). GMQN: A Reference-Based Method for Correcting Batch Effects and Probe Bias in HumanMethylation BeadChip. Front. Genet. *12*, 810985.

71. Jensen, M.A., Ferretti, V., Grossman, R.L., and Staudt, L.M. (2017). The NCI Genomic Data Commons as an engine for precision medicine. Blood *130*, 453–459.

72. Cavalcante, R.G., and Sartor, M.A. (2017). annotatr: genomic regions in context. Bioinformatics *33*, 2381–2383.

73. Dai, M., Thompson, R.C., Maher, C., Contreras-Galindo, R., Kaplan, M.H., Markovitz, D.M., Omenn, G., and Meng, F. (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genom. *11*, S7.

74. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

75. Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

76. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

77. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

78. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

79. Maag, J.L.V. (2018). gganatogram: An R package for modular visualisation of anatograms and tissues based on ggplot2. F1000Res. *7*, 1576.

80. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

81. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inform Process Syst *32*, 8024–8035.