PNAS

Supporting Information for

Comprehensive tissue deconvolution of cell-free DNA by deep learning for disease diagnosis and monitoring

Shuo Li, Weihua Zeng, Xiaohui Ni, Qiao Liu, Wenyuan Li, Mary L. Stackpole, Yonggang Zhou, Arjan Gower, Kostyantyn Krysan, Preeti Ahuja, David S. Lu, Steven S. Raman, William Hsu, Denise R. Aberle, Clara E. Magyar, Samuel W. French, Steven-Huy B. Han, Edward B. Garon, Vatche G. Agopian, Wing Hung Wong, Steven M. Dubinett, Xianghong Jasmine Zhou

Wing Hung Wong

Email: whwong@stanford.edu

Steven M. Dubinett

Email: SDubinett@mednet.ucla.edu

Xianghong Jasmine Zhou

Email: XJZhou@mednet.ucla.edu

This PDF file includes:

Supporting Information Text

S1. Genomic DNA RRBS library construction.

The RRBS libraries of the genomic DNA from the 521 tissue samples were constructed following the standard RRBS protocol [1]. 100-200 ng of intact genomic DNA in the volume of 21.5 µl was used as input material. Restriction digestion was done with 2.5 µl 10xCutSmart buffer and 1 µl MspI (NEB) for 18 h at 37 ^{o}C and 20 min at 65 ^{o}C. 0.5 µl 10xCutSmart buffer, 0.3 µl dACGTP mixture (100 mM dATP, 10 mM dCTP, 10 mM dGTP), 1 µl Klenow (exo-, 5U/µl, NEB) and 2.6 µl RT-PCR water, 0.6 µl 50 mM DTT (ThermoFisher) was added to the mixture for end repair and A-overhang addition with the program 30 ^{o}C for 20 min, 37 ^{o}C for 1 h and 75 ^{o}C for 20 min. Adapter ligation was then performed with 1 µl 10xThermoFisher HC T4 ligase buffer, 0.4 µl 100 mM ATP (ThermoFisher), 0.2 µl 50 mM DTT, 1 µl ThermoFisher HC T4 DNA ligase (30 Weiss Unit/µl), 30 ng home-made duplex UMI adapter with all the cytosines methylated (protocol adopted from Kennedy et al. [2]) at 16 ^{o}C for 20 h and 65 ^{o}C for 20 min. Bisulfite conversion of the adapter-ligated product was carried out with QIAGEN EpiTect plus DNA bisulfite kit following their protocol for two rounds of conversion. The converted product was purified with Qiagen MinElute spin column and eluted with 20 µl RT-PCR water. PCR amplification was done using the NEBNext Multiplex Oligos for Illumina (2.5 µl of universal and index primer each) and 25 µl KAPA HiFi HotStart Uracil+ ReadyMix (Roche) with the following cycling conditions: 98 ^{o}C for 45 s, 9 cycles of 98 ^{o}C for 15 s, 60 ^{o}C for 30 s and 72 ^{o}C for 30 s, followed by a final extension at 72 ^{o}C for 5 min. The PCR product was purified with 1x AmpureXP beads and eluted with 30 µl EB buffer. DNA concentration was measured by Qubit 1xdsDNA HS assay. 5% TBE-UREA PAGE and bioanalyzer assay was performed as quality control on each library before sequencing.

S2. Plasma cfDNA cfMethyl-Seq library construction.

The cfMethyl-Seq libraries of the serial plasma cfDNA samples from the four NSCLC patients were constructed following the standard protocol [3]. 10 ng of cfDNA in the volume of 25 µl was used as input material. 5'-end dephosphorylation was done with 3 µl 10xCutSmart buffer and 2 µl quick CIP from NEB (Ipswich, MA) at 37 ^{o}C for 30 min then heat-inactivated at 80 ^{o}C for 5 min. The 3'-end blocking was done with 0.5 µl 10xCutSmart buffer, 3 µl 2.5 mM CoCl2, 1 µl terminal transferase (all from NEB), and 0.5 µl 1 mM ddGTP at 37 ^{o}C for 2 h followed by 75 ^{o}C for 20 min. The mixture was then purified with 2x AmpureXP beads (Beckman Coulter, Indianapolis, IN) and eluted in 21.5 µl RT-PCR grade water (Thermo-Fisher, Waltham, MA). Restriction digestion was done with 2.5 µl 10xCutSmart buffer and 1 µl MspI (NEB) for 18 h at 37 ^{o}C and 20 min at 65 ^{o}C . 0.5 µl 10xCutSmart buffer, 0.3 µl dACGTP mixture (100 mM dATP, 10 mM dCTP, 10 mM dGTP), 1 µl Klenow (exo-, 5U/µl, NEB) and 2.6 µl RT-PCR water, 0.6 µl 50 mM DTT (ThermoFisher) was added to the mixture for end repair and A-overhang addition with the program 30 °C for 20 min, 37 °C for 1 h and 75 °C for 20 min. Adapter ligation was then performed with 1 µl 10xThermoFisher HC T4 ligase buffer, 0.4 µl 100 mM ATP (ThermoFisher), 0.2 µl 50 mM DTT, 1 µl ThermoFisher HC T4 DNA ligase (30 Weiss Unit/µl), 5 ng home-made duplex UMI adapter with all the cytosines methylated (protocol adopted from Kennedy et al. [2]) at 16 °C for 20 h and 65 °C for 20 min. Bisulfite conversion of the adapter-ligated product was carried out with QIAGEN EpiTect plus DNA bisulfite kit following their protocol for two rounds of conversion. The converted product was purified with Qiagen MinElute spin column and eluted with 20 µl RT-PCR water. PCR amplification was done using the NEBNext Multiplex Oligos for Illumina (2.5 µl of universal and index primer each) and 25 µl KAPA HiFi HotStart Uracil+ ReadyMix (Roche) with the following cycling conditions: 98 °C for 45 s, 15 cycles of 98 °C for 15 s, 60 °C for 30 s and 72 °C for 30 s, followed by a final extension at 72 °C for 5 min. The PCR product was purified with 1x AmpureXP beads and eluted with 30 µl EB buffer. DNA concentration was measured by Qubit 1xdsDNA HS assay. 5% TBE-UREA PAGE and bioanalyzer assay was performed as quality control on each library before sequencing.

S3. Data preprocessing and analysis.

We performed three steps to preprocess cfMethyl-Seq data. In Step 1, we removed the UMI sequence and trimmed the raw sequencing reads. Our custom adapters contain an 8 bp random UMI and a 5 bp fixed sequence at the beginnings of both forward and reverse reads. These sequences are removed before adapter trimming (and written into the read name). Then Trimgalore [4] was used to trim the default Illumina adapters from the sequencing reads (using the options --three_prime_clip_R1 15 --three_prime_clip_R2 13 --clip_R2 2 --length 15 --phred33). In Step 2, we performed sequence alignment, deduplication, and methylation calling. We first used Bismark [5] to align the trimmed reads to the reference genome hg19 [6] (GRCh37 (GCA 000001405.1)). Then Umi-Grinder [7] was used to remove PCR duplicates based on the UMI labels (now in the read names), allowing 4 mismatches in the total 16 bp UMI. Bismark methylation extractor was then used to call methylation in the mapped, deduplicated reads. In Step 3, the paired reads R1 and R2 were merged to form one fragment based on their mapping location. Tissue RRBS samples were preprocessed in the same manner as cfMethylSeq data.

S4. Orthogonal validation of the tissue marker atlas

We curated orthogonal validation data from public databases (Supplementary Table 6), including the whole-genome bisulfite sequencing (WGBS) data from the Epigenome Roadmap projects [8], the RNA-seq data from the GTEx project [9], and the chromatin immunoprecipitation sequencing (ChIP-seq) data from the ENCODE project [10]. The tissue sites of these data were matched with our tissue RRBS data. Note that not all 29 tissue types in our analysis can find validation data in these public databases, so only those tissue types with available data were validated.

S4.1. Validation of the reproducibility using WGBS data from the Epigenome Roadmap projects

The WGBS data from the Epigenome Roadmap projects (16 samples from 14 tissue types) were downloaded as the bigwig files containing the total allele count and the methylated allele count at individual CpG sites. We calculated the beta values for the tissue-specific methylation markers of our atlas from the bigwig files. The beta values were calculated as the proportion of the methylated alleles in all alleles across the marker region. To validate the tissue-specific markers in our atlas, on the WGBS data from the Epigenome Roadmap project, we calculated the fold change of the average beta values at a marker between the tissues in the positive and negative groups from which the marker was selected. On the RRBS data of our tissue samples, we performed a one-sided Wilcoxon rank-sum test (comparing less methylated tissues with more methylated tissues) on the count of tissue-specific fragments. We calculated fold changes on the WGBS data due to the very limited sample size. There were only one or two samples per tissue type, which was inadequate for statistical test. On the contrary, our RRBS data contained 6~57 samples (median 15) per tissue type, so we performed statistical test on our RRBS data. From our marker discovery method, the tissues in the positive group were supposed to have a lower methylation level than those in the negative group. Therefore, we treated the markers with a fold change < 1 in the WGBS data as having consistent tissue-specific methylation patterns with our marker atlas. The markers with consistent methylation patterns were considered to be reproducible in the independent dataset. In this independent dataset, 92.9% of the tissue markers showed tissue-specific methylation consistent with our tissue RRBS data (Figure 2b and Supplementary Figure 1a). This indicated that the tissue marker atlas captured real tissue-specific methylation patterns reproducible in other tissue samples.

S4.2. Marker association with tissue-specific transcription activity using RNA-seq data from the GTEx project

The RNA-seq data from the GTEx projects were downloaded as a numeric matrix containing the transcript per million (TPM) for each transcript. Since our tissue samples were also from the GTEx projects, we can find the matched RNA-seq data of the tissue samples from which our tissue RRBS data were generated. Therefore, we included these samples in our validation analysis. These matched RNA-seq data could show whether the tissue-specific differentially methylated regions identified from our tissue RRBS data impacted the tissue-specific transcriptome. To map the transcription data to the tissue-specific markers, we overlapped the promoter regions (defined by GeneHancer [11]) with the tissue-specific marker regions. If a gene promoter was identified to have over 50% overlap with a tissue-specific marker, we mapped the TPM of that gene to the tissue-specific marker. To evaluate the tissue specificity of transcription activity, we performed Wilcoxon rank-sum tests on the transcription levels between the corresponding tissues from which the gene-associated markers were identified. At 63.0% of tissue markers, we observed increased gene transcription levels when the corresponding promoter regions were hypomethylated in the tissue types (Figure 2d). The tissue-specific transcription pattern implied that the methylation level at the tissue markers in our atlas may impact the downstream transcription activity.

S4.3. Marker validation for the association with tissue-specific histone modification using ChIP-seq data from the ENCODE project

We downloaded the ChIP-seq data (72 samples of 19 tissue types) for the histone modification H3K27ac from the ENCODE project as the bed files of replicated peak calls. Specifically, in a tissue marker region, for each tissue type, we calculated the peak frequency among the samples, i.e., the fraction of tissue samples that had H3K27ac peak calls overlapping with the tissue marker region. To evaluate the tissue-specific H3K27ac modification, we calculated the fold change of the peak frequency at a marker between the tissues in the positive and negative groups from which the marker was selected. Note that we calculated fold changes because there were only one peak frequency profile per tissue type, which was not enough for statistical tests. We observed consistent tissue-specific H3K27ac modification at 93.7% of the tissue markers (Figure 2c and Supplementary Figure 1b). A hypomethylated region for a tissue usually corresponds to a tissue-specific elevation of H3K27ac modification, consistent with previous studies [8][12].

S4.4. Validation of tissue-specific markers by their association with tissue-specific transcription regulation

We performed the enrichment analysis for the transcription factor binding motifs at the tissue-specific markers. The enriched transcription factor binding motifs were identified using HOMER [13] findMotifsGenome.pl with the hg19 reference genome and an input bed file of the genomic coordinates of the tissue-specific markers used in the tissue deconvolution. We found that the enriched motifs are mostly related to development, differentiation, and tissue-specific expression (Figure 2e and Supplementary Table 2). For example, HOXD12 was the top 3 enriched motif, which was part of the developmental regulatory system [14][15]. The transcription factors that regulated specific tissue development and differentiation were also enriched, such as MEF2A, MEF2B, and MEF2C for the muscle, GSC for the nervous system, USF2 for the mammary gland (in the breast), COUP-TFI for the adipose tissue, MR2F2 for the ovary, BAPX1 for the stomach, NKX3.1 for the prostate, and GLIS3 for the pancreas, thyroid, liver, and kidney [14][15]. These results indicated that the tissue markers may involve in tissue-specific biological processes.

S5. *cfSort* workflow.

S5.1. Input data preprocessing

After the preprocessing of the raw sequencing reads, we need to convert the methylation information in the aligned DNA fragments to the numerical features at the selected tissue markers. From the marker selection procedure, every selected tissue marker has two associated information: (1) a genomic region and (2) an α-value threshold that reflects the tissue-specific methylation and can be used to determine the tissue-specific DNA fragments. Note that the distribution of cfDNA in a genomic region was impacted by the epigenetics (e.g. the nucleosome positioning) of the cells which the cfDNA originates from. Thus different epigenetics in different tissues can affect the tissue contribution of the cfDNA in a local genomic region, thus affecting the methylation profiles of the cfDNA. In other words, the tissue composition at the small-sized tissue markers can deviate from the overall tissue composition. As a result, the cfDNA methylation profiles in the small-sized tissue markers can be unstably fluctuated, which further impairs the data quality and tissue deconvolution performance. To address the challenge of this data characteristic, we designed a strategy of combining tissue markers of small genomic regions into large-size merged markers that are robust against the local read distribution variation. Specifically, we performed constrained K-means clustering [16] on the individual markers from each comparison (e.g. liver vs. non-liver tissues), allowing four to seven individual markers in a cluster. The clustering was based on the methylation profiles of the training tissue samples. In this way, the markers with similar methylation profiles among tissues will be clustered together. Then we combined the individual markers in a cluster. Because the markers within a cluster have similar methylation profiles across tissues, they share similar tissue-specific methylation signals. Therefore, the tissue-specific methylation signals in individual markers will not be blurred in the merged marker. As a result, the merged markers (approximately 400bp to 1000bp) had a much larger size compared to the DNA wrapped around a nucleosome (approximately 166bp [17][18]), which can make methylation at the merged-marker level effectively robust against the local read distribution variation due to the nucleosome positioning. For every merged tissue marker, we derived a numerical feature, by calculating the fraction of tissue-specific DNA fragments across all individual markers within this merged marker. The tissue-specific DNA fragments were identified at every individual marker as the DNA fragment with α-value above the marker-specific threshold.

Then we transformed the feature values to make the input data more suitable for machine learning. Firstly, we transformed the features to the logarithmic space using the log1p function in the python NumPy package [19]. This transformation accounted for the heteroscedasticity and improved the statistical properties of the features [20]. Secondly, we rescaled every feature to the range [0,1], making the features at the same scale. Because the raw features were the fraction of tissue-specific reads at every marker, these features were not impacted by per-sample variation in the sequencing depth. Therefore, additional per-sample scaling was not necessary. After the two rescaling steps, the data were ready to be used as the input of *cfSort*. Note that the rescaling factors were learned only from the training data. To preprocess the validation, testing, or any new data, we applied the same rescaling factors. Therefore, there was no data leakage in the preprocessing step.

S5.2. *cfSort* model

The *cfSort* model was built by supervised training on the simulated cfMethyl-Seq data of the *in silico* cfDNA samples for deconvolving the fractions of the 29 tissue types in cfDNA. The *cfSort* model is an ensemble of two DNNs, which can effectively reduce the prediction variance as previously reported [21]. The two DNNs have three dense hidden layers with a decreased number of nodes (1024, 512, 128, and 256, 128, 32 respectively), which finally connect to an output layer with 29 nodes. In each dense layer, we will use the rectified linear unit [22] as an activation function to implement the nonlinearity of methylation caused by the tissue epigenetics and generate higher-order latent representations of the tissue signatures. Considering the complexity and diversity of our training data, we applied a batch normalization layer before each dense layer to standardize the contributions of each merged marker and each training batch. This was proven effective to stabilize and accelerate the training process [23]. We also applied a dropout layer after each dense layer (with the dropout rate 0, 0.05, 0, and 0, 0, 0 respectively for the two DNNs) to regularize the DNN to increase model robustness and avoid overfitting [24]. Since the tissue composition naturally adds up to 1, we used the Softmax activation function in the output layer. We used python and the TensorFlow library to implement the *cfSort* model.

S5.3. Model training and prediction

After data preprocessing, the *cfSort* model was trained on the simulated cfMethyl-Seq data generated from the mixtures of RRBS data of real tissue samples with known tissue composition. To train the model, we used the state-of-the-art optimizer Adam [25] with a learning rate of 0.001 and a batch size of 32. We used the mean absolute error between the estimated tissue composition and the ground truth as the loss function. In addition, we utilized the early stopping strategy to further avoid overfitting, which has proven effective in cell-type proportion estimation with gene expression data [20]. Specifically, we evaluated the mean absolute error on the validation data (i.e., the validation loss) after each epoch. If the validation loss did not drop after 10 epochs, we terminated the training process. We also tried waiting for 5 epochs, which did not have much impact on the performance. To predict tissue composition on the real cfDNA sample (or simulated testing cfDNA sample) with the *cfSort*, we can extract the methylation profile of the merged markers from the raw data of a cfDNA sample into the trained *cfSort* model, and a predicted tissue composition consisting of fractions of 29 tissue types will be generated.

S6. Comparison with two existing methods.

S6.1. Non-negative least squares (NNLS)

Non-negative least squares or quadratic programming was widely used in methylation-based tissue deconvolution [26][27][28]. This method assumes and models cfDNA methylation as a linear combination of the tissue methylation in the reference tissue samples at tissue markers. Then it determines the tissue composition using quadratic programming [26][27][28]. Because the tissue markers used in these methods were identified from a limited number of samples for each tissue type, these markers cannot cover all tissue types in our analysis and may not be representative enough for each tissue type under the inter-individual variance. To fairly compare the non-negative least squares with *cfSort*, we identified the new markers using our training and validation sets of the tissue RRBS samples using the tissue marker selection procedure of Sun et al. [26]. Then we used the training and validation tissue RRBS samples as the reference samples and extracted the reference tissue methylation profiles from them. The reference methylation profile was calculated as the averaged methylation levels of the training and validation RRBS samples for each tissue type. The reference profiles were then used in the quadratic programming to estimate the tissue composition in the testing samples.

S6.2. CelFIE

CelFIE is a recent tissue deconvolution algorithm that uses an expectation maximization (EM) algorithm [29]. CelFIE required a reference methylation profile per tissue type, and then it optimized the estimated tissue fractions by maximizing the likelihood of the observed methylation levels in the cfDNA. CelFIE included a tissue marker selection pipeline. Therefore, we directly applied its marker selection method to the reference methylation profiles of the 29 tissue types. The reference methylation profile was calculated as the average normalized methylated read counts and the average normalized total read counts of the training and validation RRBS samples for each tissue type. Based on the reference tissue profiles and the selected markers, CelFIE was directly applied to the testing samples.

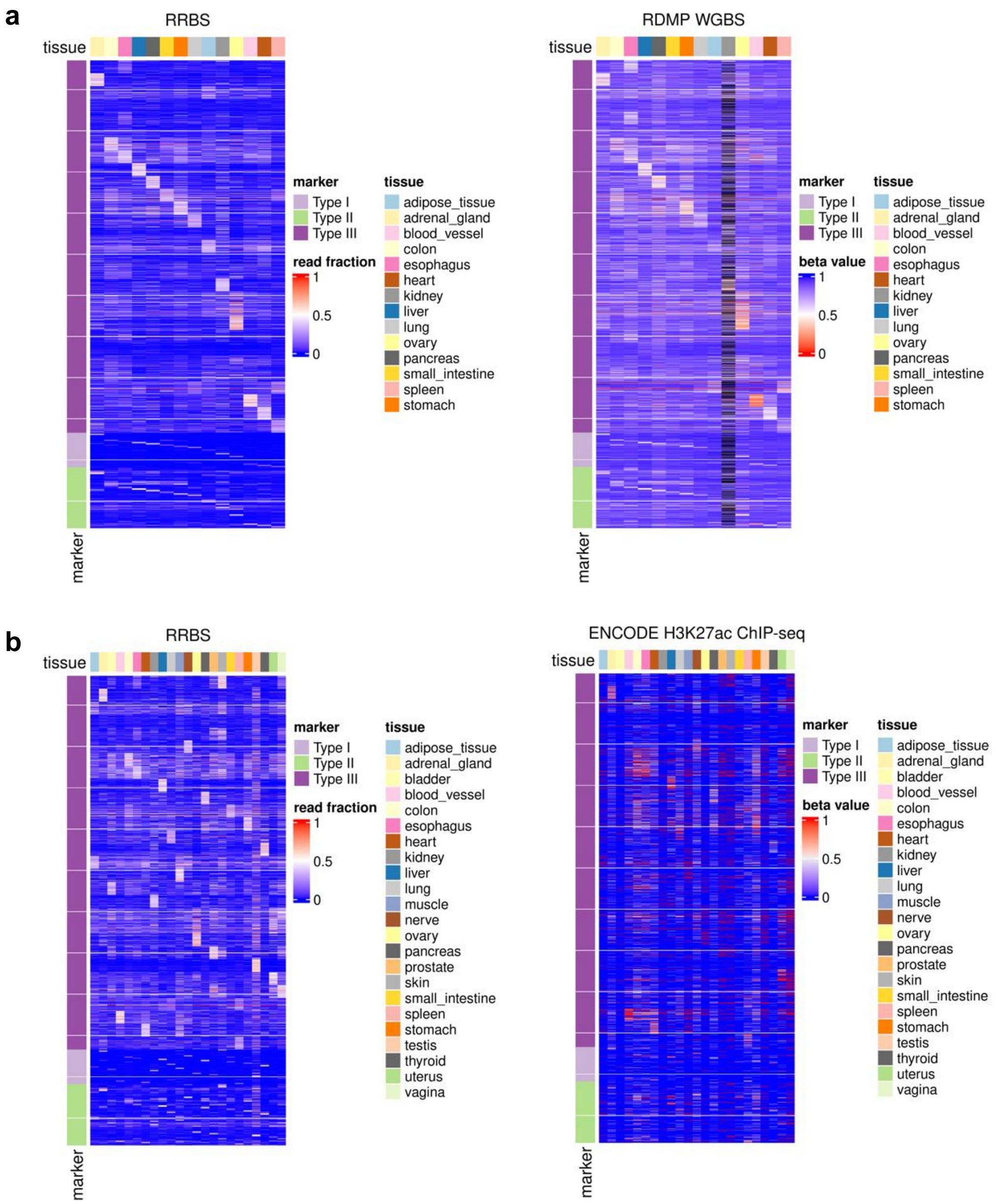

Fig. S1. Marker validation in (a) the Epigenome Roadmap WGBS data and (b) the ENCODE H3K27ac ChIP-seq data. (a) Heatmaps of the average methylation level at the tissue markers in our RRBS data (left) and in the Epigenome Roadmap WGBS data (right). **(b)** Heatmaps of the average methylation level at the tissue markers in our RRBS data (left) and the peak frequency in ENCODE H3K27ac ChIP-seq data (right). The rows in the left and right subfigures correspond to the tissue markers that were sorted in the same order in **(a)** and **(b)**. Only the tissue markers with available data in the Epigenome Roadmap project and the ENCODE project were shown in **(a)** and **(b)** respectively.

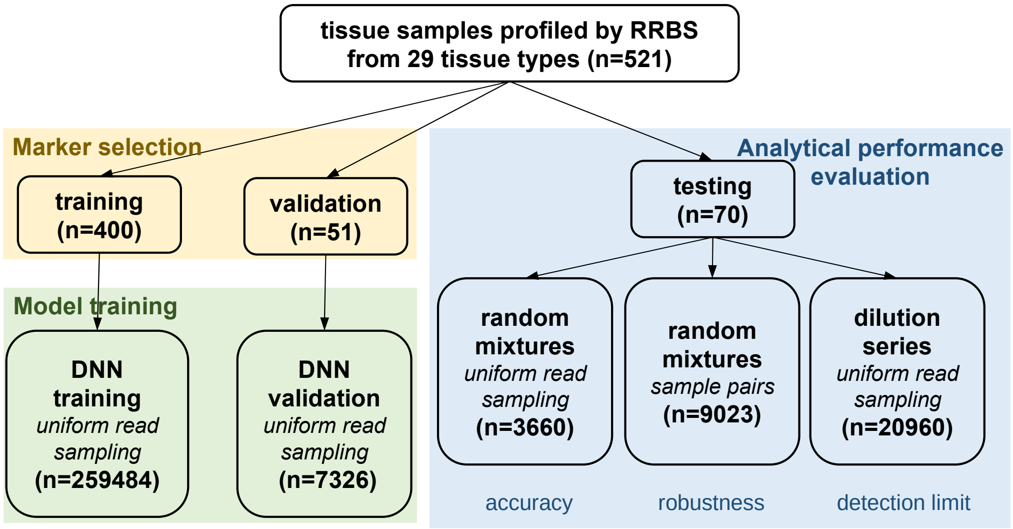
tissue samples profiled by RRBS from 29 tissue types (n=521)
Marker selection
training (n=400)
validation (n=51)
Model training
DNN training uniform read sampling (n=259484)
DNN validation uniform read sampling (n=7326)
Analytical performance evaluation
testing (n=70)
random mixtures uniform read sampling (n=3660)
random mixtures sample pairs (n=9023)
dilution series uniform read sampling (n=20960)
accuracy
robustness
detection limit

Fig. S2. Overview of the tissue sample used for the marker discovery, model training, validation, and testing. All tissue samples were randomly split into three sets, used for the model training (75%), validation (10%), and testing (15%). The training set and the validation set were used in the marker selection.

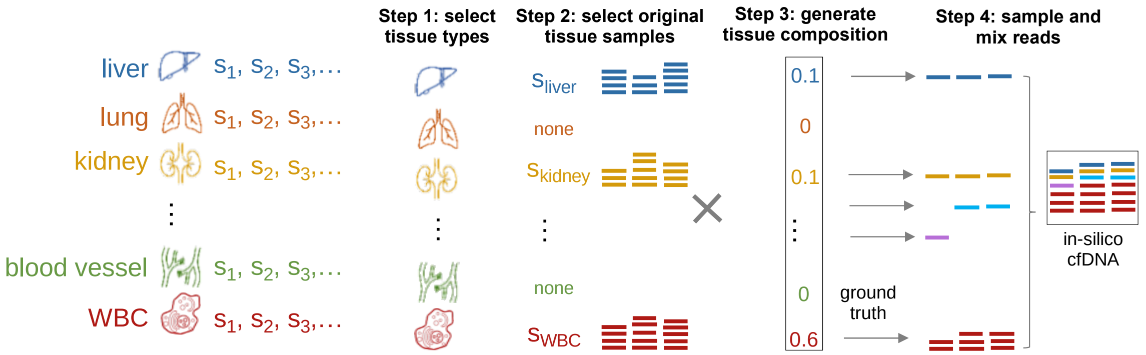
Step 1: select tissue types
Step 2: select original tissue samples
Step 3: generate tissue composition
Step 4: sample and mix reads
liver
lung
kidney
blood vessel
WBC
s1, s2, s3,…
s_liver
none
s_kidney
none
s_WBC
0.1
0
0.1
0
0.6
ground truth
in-silico cfDNA

Fig. S3. The detailed procedure of the generation of a simulated cfMethylSeq sample. In four steps, we generated a simulated cfMethylSeq sample. In Step 1, we first selected the tissue types that contributed positive fractions to the simulated sample. WBC always contributed positively to the final mixture. In Step 2, we chose an original tissue sample at random for each selected tissue type and WBC. In Step 3, we created a random tissue composition for the simulated sample. We set the tissue fraction to zero if a tissue type was not chosen in Step 1, and we required WBC to always have the highest tissue fraction. In Step 4, we sampled sequencing reads at random from the selected samples (from Step 2) based on tissue composition (generated in Step 3).

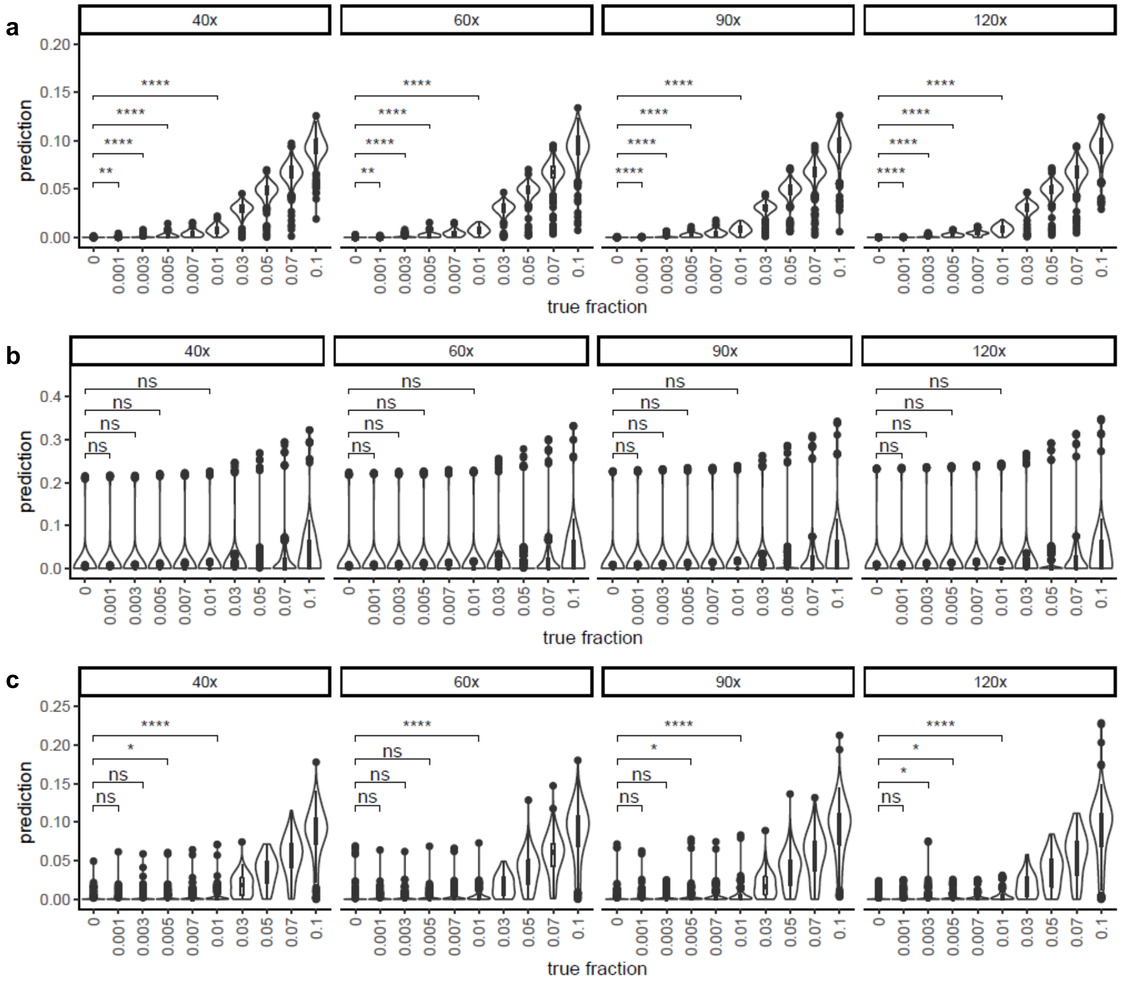

Fig. S4. The detection limit of *cfSort* (a), NNLS (b), and CelFiE (c) on dilution series at 40x, 60x, 90x, and 120x. The detection limit was measured by the statistical significance of a one-sided Student's *t*-test between the estimated tissue fractions of the samples at every dilution level and the control samples (i.e., 0% tissue fraction). The statistical significance in the figures indicated the p-values of the one-sided Student's t-tests at 0.5% and 1%: "ns" means not statistically significant (p-value > 0.05); "*" means p-value < 0.05; "**" means p-value < 0.01; "***" means p-value < 0.001; "****" means p-value < 0.0001.

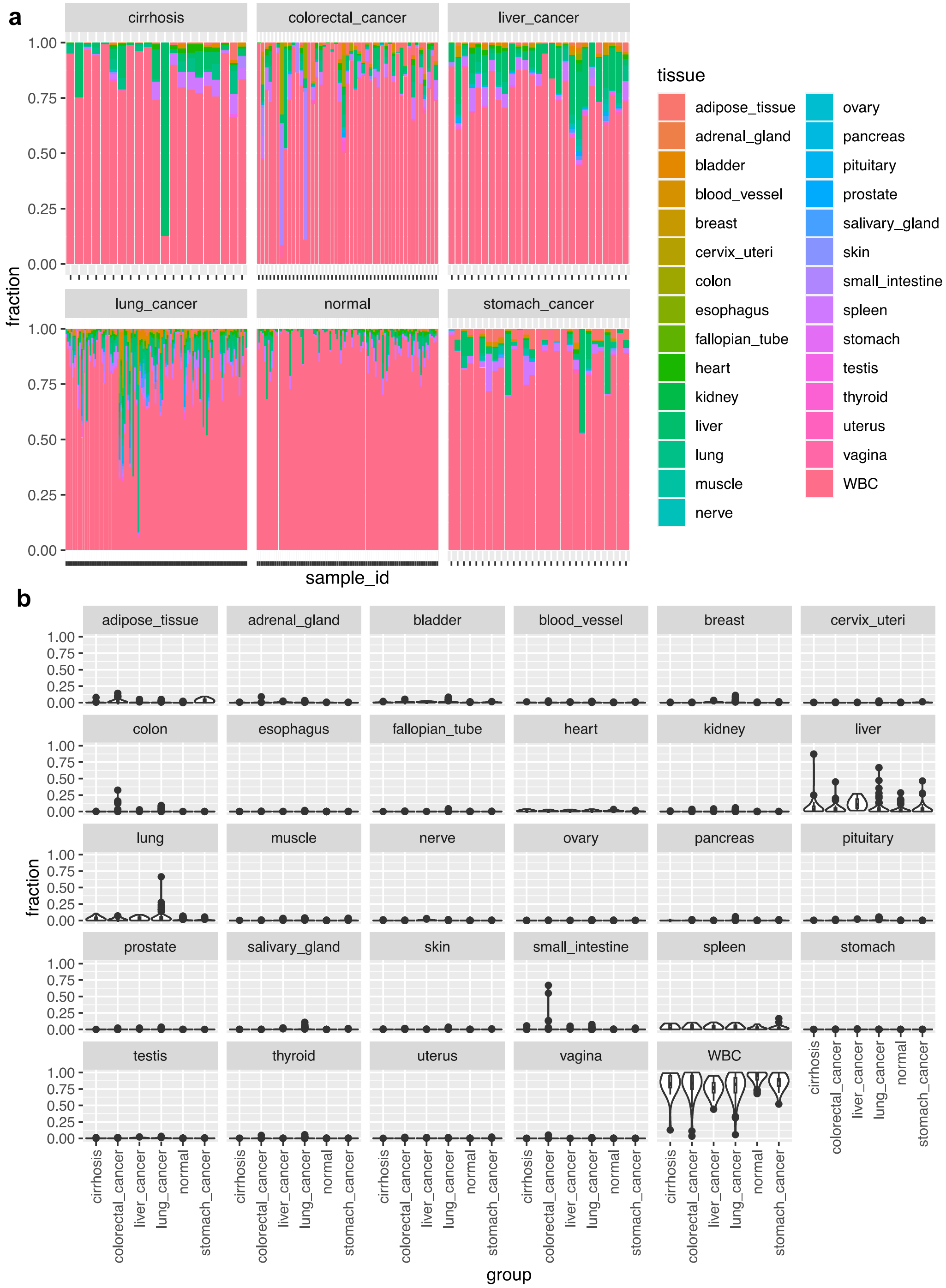

Fig. S5. The tissue fraction of all tissue types estimated from the cfMethyl-Seq data of the plasma samples in the main text Figures 6 and 7. (a) Bar plots of the tissue composition in the cfDNA samples. **(b)** Violin plots of the tissue fractions in each clinical cohorts for individual tissue types.

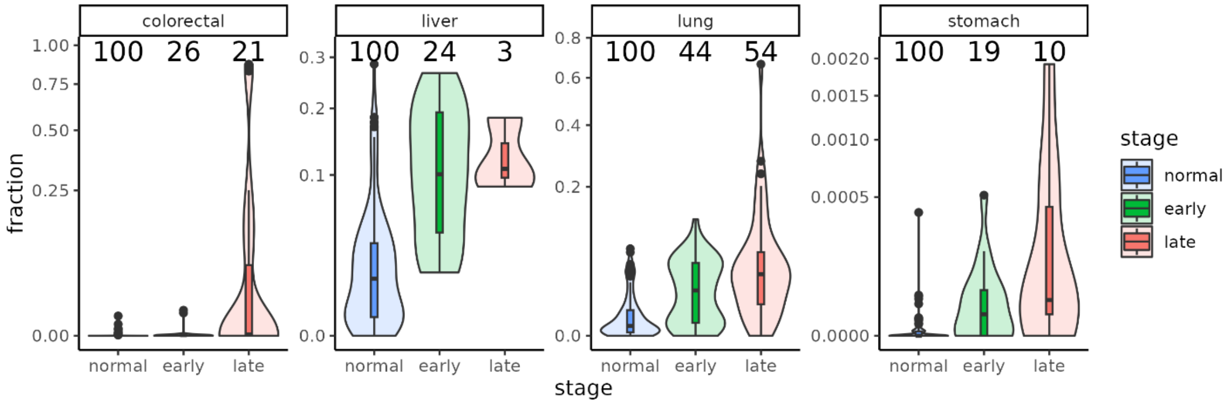

Fig. S6. The tissue-derived cfDNA fractions of the host tissue in the cancer patients at different stages. The plasma samples from cancer patients were divided based on the cancer stages. The early stage included plasma samples from stage I and II patients; the late stage included plasma samples from stage III and IV patients. The title of the subfigures indicated the cancer types. For each cancer type, the value on the y-axis showed the host tissue fraction in the cfDNA estimated from *cfSort*. The number on the top of each violin showed the number of samples within the violin.

Table S1. The number of tissue samples for the 29 tissue types.

tissue type	total	training	validation	test
adipose tissue	22	18	2	2
adrenal gland	14	11	2	1
bladder	6	4	1	1
blood vessel	30	23	3	4
breast	15	11	2	2
cervix uteri	11	8	1	2
colon	29	22	3	4
esophagus	44	34	4	6
fallopian tube	6	4	1	1
heart	24	17	3	4
kidney	13	10	1	2
liver	12	10	1	1
lung	16	12	2	2
muscle	10	7	1	2
nerve	13	10	1	2
ovary	16	11	2	3
pancreas	14	11	1	2
pituitary	15	10	2	3
prostate	13	9	2	2
salivary gland	14	11	1	2
skin	23	18	2	3
small intestine	16	12	2	2
spleen	18	13	2	3
stomach	14	11	1	2
testis	16	12	2	2
thyroid	11	9	1	1
uterus	12	9	1	2
vagina	17	12	2	3
WBC	57	51	2	4

Table S2. The enriched transcription factor binding motifs at the tissue markers.

Motif Name	P-value	q-value Benjamini
Mef2a(MADS)/HL1-Mef2a.biotin-ChIP-Seq(GSE21529)/Homer	1e-663	0
Mef2c(MADS)/GM12878-Mef2c-ChIP-Seq(GSE32465)/Homer	1e-631	0
Hoxd12(Homeobox)/ChickenMSG-Hoxd12.Flag-ChIP-Seq(GSE86088)/Homer	1e-628	0
GSC(Homeobox)/FrogEmbryos-GSC-ChIP-Seq(DRA000576)/Homer	1e-529	0
Mef2b(MADS)/HEK293-Mef2b.V5-ChIP-Seq(GSE67450)/Homer	1e-518	0
EAR2(NR)/K562-NR2F6-ChIP-Seq(Encode)/Homer	1e-400	0
Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)/Homer	1e-346	0
COUP-TFII(NR)/K562-NR2F1-ChIP-Seq(Encode)/Homer	1e-338	0
CRX(Homeobox)/Retina-Crx-ChIP-Seq(GSE20012)/Homer	1e-325	0
Nkx3.1(Homeobox)/LNCaP-Nkx3.1-ChIP-Seq(GSE28264)/Homer	1.00E-301	0
Bapx1(Homeobox)/VertebralCol-Bapx1-ChIP-Seq(GSE36672)/Homer	1.00E-290	0
Npas4(bHLH)/Neuron-Npas4-ChIP-Seq(GSE127793)/Homer	1.00E-280	0
COUP-TFII(NR)/Artia-Nr2f2-ChIP-Seq(GSE46497)/Homer	1.00E-272	0
RARa(NR)/K562-RARa-ChIP-Seq(Encode)/Homer	1.00E-185	0
THRb(NR)/Liver-NR1A2-ChIP-Seq(GSE52613)/Homer	1.00E-153	0
Erra(NR)/HepG2-Erra-ChIP-Seq(GSE31477)/Homer	1.00E-149	0
GLIS3(Zf)/Thyroid-Glis3.GFP-ChIP-Seq(GSE103297)/Homer	1.00E-140	0
Pitx1(Homeobox)/Chicken-Pitx1-ChIP-Seq(GSE38910)/Homer	1.00E-132	0
ZNF711(Zf)/SHSY5Y-ZNF711-ChIP-Seq(GSE20673)/Homer	1.00E-111	0
Mef2d(MADS)/Retina-Mef2d-ChIP-Seq(GSE61391)/Homer	1.00E-106	0
Reverb(NR),DR2/RAW-Reverba.biotin-ChIP-Seq(GSE45914)/Homer	1.00E-104	0
SF1(NR)/H295R-Nr5a1-ChIP-Seq(GSE44220)/Homer	1.00E-86	0
Nr5a2(NR)/Pancreas-LRH1-ChIP-Seq(GSE34295)/Homer	1.00E-56	0
HIF-1b(HLH)/T47D-HIF1b-ChIP-Seq(GSE59937)/Homer	1.00E-56	0
HIF-1a(bHLH)/MCF7-HIF1a-ChIP-Seq(GSE28352)/Homer	1.00E-44	0
Atf1(bZIP)/K562-ATF1-ChIP-Seq(GSE31477)/Homer	1.00E-44	0
ARE(NR)/LNCAP-AR-ChIP-Seq(GSE27824)/Homer	1.00E-44	0
p73(p53)/Trachea-p73-ChIP-Seq(PRJNA310161)/Homer	1.00E-41	0
HIF2a(bHLH)/785_O-HIF2a-ChIP-Seq(GSE34871)/Homer	1.00E-38	0
Esrrb(NR)/mES-Esrrb-ChIP-Seq(GSE11431)/Homer	1.00E-37	0
Cdx2(Homeobox)/mES-Cdx2-ChIP-Seq(GSE14586)/Homer	1.00E-35	0

Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1.00E-34	0
p53(p53)/mES-cMyc-ChIP-Seq(GSE11431)/Homer	1.00E-33	0
HIC1(Zf)/Treg-ZBTB29-ChIP-Seq(GSE99889)/Homer	1.00E-32	0
ERRg(NR)/Kidney-ESRRG-ChIP-Seq(GSE104905)/Homer	1.00E-31	0
MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1.00E-31	0
Smad2(MAD)/ES-SMAD2-ChIP-Seq(GSE29422)/Homer	1.00E-29	0
EBNA1(EBV-virus)/Raji-EBNA1-ChIP-Seq(GSE30709)/Homer	1.00E-28	0
Atf2(bZIP)/3T3L1-Atf2-ChIP-Seq(GSE56872)/Homer	1.00E-28	0
Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	1.00E-28	0
Tlx?(NR)/NPC-H3K4me1-ChIP-Seq(GSE16256)/Homer	1.00E-27	0
Atf7(bZIP)/3T3L1-Atf7-ChIP-Seq(GSE56872)/Homer	1.00E-27	0
LXRE(NR),DR4/RAW-LXRb.biotin-ChIP-Seq(GSE21512)/Homer	1.00E-27	0
Smad4(MAD)/ESC-SMAD4-ChIP-Seq(GSE29422)/Homer	1.00E-24	0
p53(p53)/Saos-p53-ChIP-Seq(GSE15780)/Homer	1.00E-24	0
p53(p53)/Saos-p53-ChIP-Seq/Homer	1.00E-24	0
THRa(NR)/C17.2-THRa-ChIP-Seq(GSE38347)/Homer	1.00E-22	0
Hoxd13(Homeobox)/ChickenMSG-Hoxd13.Flag-ChIP-Seq(GSE86088)/Homer	1.00E-22	0
CDX4(Homeobox)/ZebrafishEmbryos-Cdx4.Myc-ChIP-Seq(GSE48254)/Homer	1.00E-22	0
Tbx5(T-box)/HL1-Tbx5.biotin-ChIP-Seq(GSE21529)/Homer	1.00E-22	0
Nr5a2(NR)/mES-Nr5a2-ChIP-Seq(GSE19019)/Homer	1.00E-21	0
Ascl2(bHLH)/ESC-Ascl2-ChIP-Seq(GSE97712)/Homer	1.00E-20	0
p63(p53)/Keratinocyte-p63-ChIP-Seq(GSE17611)/Homer	1.00E-20	0
Hoxa13(Homeobox)/ChickenMSG-Hoxa13.Flag-ChIP-Seq(GSE86088)/Homer	1.00E-17	0
GRE(NR),IR3/A549-GR-ChIP-Seq(GSE32465)/Homer	1.00E-16	0
HOXB13(Homeobox)/ProstateTumor-HOXB13-ChIP-Seq(GSE56288)/Homer	1.00E-15	0
CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	1.00E-15	0
Elk4(ETS)/Hela-Elk4-ChIP-Seq(GSE31477)/Homer	1.00E-15	0
Atoh1(bHLH)/Cerebellum-Atoh1-ChIP-Seq(GSE22111)/Homer	1.00E-15	0
ZFX(Zf)/mES-Zfx-ChIP-Seq(GSE11431)/Homer	1.00E-15	0
ZNF136(Zf)/HEK293-ZNF136.GFP-ChIP-Seq(GSE58341)/Homer	1.00E-15	0
FoxD3(forkhead)/ZebrafishEmbryo-Foxd3.biotin-ChIP-seq(GSE106676)/Homer	1.00E-14	0
NeuroG2(bHLH)/Fibroblast-NeuroG2-ChIP-Seq(GSE75910)/Homer	1.00E-14	0
Fos(bZIP)/TSC-Fos-ChIP-Seq(GSE110950)/Homer	1.00E-13	0
TEAD1(TEAD)/HepG2-TEAD1-ChIP-Seq(Encode)/Homer	1.00E-13	0
Slug(Zf)/Mesoderm-Snai2-ChIP-Seq(GSE61475)/Homer	1.00E-13	0
FOXA1(Forkhead)/MCF7-FOXA1-ChIP-Seq(GSE26831)/Homer	1.00E-13	0
PGR(NR)/EndoStromal-PGR-ChIP-Seq(GSE69539)/Homer	1.00E-12	0
GATA(Zf),IR4/iTreg-Gata3-ChIP-Seq(GSE20898)/Homer	1.00E-12	0
BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1.00E-12	0

Hnf1(Homeobox)/Liver-Foxa2-Chip-Seq(GSE25694)/Homer	1.00E-11	0
CREB5(bZIP)/LNCaP-CREB5.V5-ChIP-Seq(GSE137775)/Homer	1.00E-11	0
JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	1.00E-11	0
Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1.00E-11	0
TEAD3(TEA)/HepG2-TEAD3-ChIP-Seq(Encode)/Homer	1.00E-10	0
Foxa2(Forkhead)/Liver-Foxa2-ChIP-Seq(GSE25694)/Homer	1.00E-10	0
TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	1.00E-10	0
HOXA2(Homeobox)/mES-Hoxa2-ChIP-Seq(Donaldson_et_al.)/Homer	1.00E-10	0
Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1.00E-09	0
MyoG(bHLH)/C2C12-MyoG-ChIP-Seq(GSE36024)/Homer	1.00E-09	0
FOXA1(Forkhead)/LNCAP-FOXA1-ChIP-Seq(GSE27824)/Homer	1.00E-09	0
Hoxd11(Homeobox)/ChickenMSG-Hoxd11.Flag-ChIP-Seq(GSE86088)/Homer	1.00E-09	0
Max(bHLH)/K562-Max-ChIP-Seq(GSE31477)/Homer	1.00E-09	0
Otx2(Homeobox)/EpiLC-Otx2-ChIP-Seq(GSE56098)/Homer	1.00E-09	0
c-Myc(bHLH)/mES-cMyc-ChIP-Seq(GSE11431)/Homer	1.00E-08	0
USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer	1.00E-08	0
TCF4(bHLH)/SHSY5Y-TCF4-ChIP-Seq(GSE96915)/Homer	1.00E-08	0
Arnt:Ahr(bHLH)/MCF7-Arnt-ChIP-Seq(Lo_et_al.)/Homer	1.00E-08	0
TEAD2(TEA)/Py2T-Tead2-ChIP-Seq(GSE55709)/Homer	1.00E-08	0
FXR(NR),IR1/Liver-FXR-ChIP-Seq(Chong_et_al.)/Homer	1.00E-08	0
E2A(bHLH)/proBcell-E2A-ChIP-Seq(GSE21978)/Homer	1.00E-08	0
ZNF519(Zf)/HEK293-ZNF519.GFP-ChIP-Seq(GSE58341)/Homer	1.00E-07	0
c-Jun-CRE(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1.00E-07	0
Hoxa11(Homeobox)/ChickenMSG-Hoxa11.Flag-ChIP-Seq(GSE86088)/Homer	1.00E-07	0
THRb(NR)/HepG2-THRb.Flag-ChIP-Seq(Encode)/Homer	1.00E-07	0
Hoxc9(Homeobox)/Ainv15-Hoxc9-ChIP-Seq(GSE21812)/Homer	1.00E-07	0
Unknown(Homeobox)/Limb-p300-ChIP-Seq/Homer	1.00E-07	0
AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1.00E-07	0
HEB(bHLH)/mES-Heb-ChIP-Seq(GSE53233)/Homer	1.00E-07	0
Olig2(bHLH)/Neuron-Olig2-ChIP-Seq(GSE30882)/Homer	1.00E-07	0
Ptf1a(bHLH)/Panc1-Ptf1a-ChIP-Seq(GSE47459)/Homer	1.00E-06	0
Foxa3(Forkhead)/Liver-Foxa3-ChIP-Seq(GSE77670)/Homer	1.00E-06	0
FoxL2(Forkhead)/Ovary-FoxL2-ChIP-Seq(GSE60858)/Homer	1.00E-06	0
NPAS2(bHLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer	1.00E-06	0
Stat3(Stat)/mES-Stat3-ChIP-Seq(GSE11431)/Homer	1.00E-06	0
Myf5(bHLH)/GM-Myf5-ChIP-Seq(GSE24852)/Homer	1.00E-05	0
Tcf12(bHLH)/GM12878-Tcf12-ChIP-Seq(GSE32465)/Homer	1.00E-05	0
MafB(bZIP)/BMM-Mafb-ChIP-Seq(GSE75722)/Homer	1.00E-05	0
Tcf21(bHLH)/ArterySmoothMuscle-Tcf21-ChIP-Seq(GSE61369)/Homer	1.00E-05	0

Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1.00E-05	0
Oct4:Sox17(POU,Homeobox,HMG)/F9-Sox17-ChIP-Seq(GSE44553)/Homer	1.00E-05	0
Brachyury(T-box)/Mesoendoderm-Brachyury-ChIP-exo(GSE54963)/Homer	1.00E-05	0
Twist2(bHLH)/Myoblast-Twist2.Ty1-ChIP-Seq(GSE127998)/Homer	1.00E-05	0
Pit1+1bp(Homeobox)/GCrat-Pit1-ChIP-Seq(GSE58009)/Homer	1.00E-05	0
bHLHE40(bHLH)/HepG2-BHLHE40-ChIP-Seq(GSE31477)/Homer	1.00E-05	0
NeuroD1(bHLH)/Islet-NeuroD1-ChIP-Seq(GSE30298)/Homer	1.00E-04	0.0001
IRF4(IRF)/GM12878-IRF4-ChIP-Seq(GSE32465)/Homer	1.00E-04	0.0001
Ap4(bHLH)/AML-Tfap4-ChIP-Seq(GSE45738)/Homer	1.00E-04	0.0001
Hoxd10(Homeobox)/ChickenMSG-Hoxd10.Flag-ChIP-Seq(GSE86088)/Homer	1.00E-04	0.0002
BHLHA15(bHLH)/NIH3T3-BHLHB8.HA-ChIP-Seq(GSE119782)/Homer	1.00E-04	0.0002
KLF5(Zf)/LoVo-KLF5-ChIP-Seq(GSE49402)/Homer	1.00E-03	0.0004
FOXK1(Forkhead)/HEK293-FOXK1-ChIP-Seq(GSE51673)/Homer	1.00E-03	0.0004
Ascl1(bHLH)/NeuralTubes-Ascl1-ChIP-Seq(GSE55840)/Homer	1.00E-03	0.0006
Rfx6(HTH)/Min6b1-Rfx6.HA-ChIP-Seq(GSE62844)/Homer	1.00E-03	0.0008
Foxf1(Forkhead)/Lung-Foxf1-ChIP-Seq(GSE77951)/Homer	1.00E-03	0.001
Bach1(bZIP)/K562-Bach1-ChIP-Seq(GSE31477)/Homer	1.00E-03	0.0017
FOXM1(Forkhead)/MCF7-FOXM1-ChIP-Seq(GSE72977)/Homer	1.00E-03	0.0023
MafF(bZIP)/HepG2-MafF-ChIP-Seq(GSE31477)/Homer	1.00E-03	0.0031
Snail1(Zf)/LS174T-SNAIL1.HA-ChIP-Seq(GSE127183)/Homer	1.00E-02	0.0033
PRDM9(Zf)/Testis-DMC1-ChIP-Seq(GSE35498)/Homer	1.00E-02	0.006
Tcfcp2l1(CP2)/mES-Tcfcp2l1-ChIP-Seq(GSE11431)/Homer	1.00E-02	0.0073
Hoxb4(Homeobox)/ES-Hoxb4-ChIP-Seq(GSE34014)/Homer	1.00E-02	0.0075
Elk1(ETS)/Hela-Elk1-ChIP-Seq(GSE31477)/Homer	1.00E-02	0.0075
Pitx1:Ebox(Homeobox,bHLH)/Hindlimb-Pitx1-ChIP-Seq(GSE41591)/Homer	1.00E-02	0.0075
PAX6(Paired,Homeobox)/Forebrain-Pax6-ChIP-Seq(GSE66961)/Homer	1.00E-02	0.0081
PBX2(Homeobox)/K562-PBX2-ChIP-Seq(Encode)/Homer	1.00E-02	0.0137
ZNF322(Zf)/HEK293-ZNF322.GFP-ChIP-Seq(GSE58341)/Homer	1.00E-02	0.0145
ZNF165(Zf)/WHIM12-ZNF165-ChIP-Seq(GSE65937)/Homer	1.00E-02	0.0167
Foxo3(Forkhead)/U2OS-Foxo3-ChIP-Seq(E-MTAB-2701)/Homer	1.00E-02	0.021
REST-NRSF(Zf)/Jurkat-NRSF-ChIP-Seq/Homer	1.00E-02	0.0224
Stat3+il21(Stat)/CD4-Stat3-ChIP-Seq(GSE19198)/Homer	1.00E-02	0.0227
GRHL2(CP2)/HBE-GRHL2-ChIP-Seq(GSE46194)/Homer	1.00E-02	0.0245
TFE3(bHLH)/MEF-TFE3-ChIP-Seq(GSE75757)/Homer	1.00E-01	0.0342
Barx1(Homeobox)/Stomach-Barx1.3xFlag-ChIP-Seq(GSE69483)/Homer	1.00E-01	0.0361
ZNF143|STAF(Zf)/CUTLL-ZNF143-ChIP-Seq(GSE29600)/Homer	1.00E-01	0.0371
NF-E2(bZIP)/K562-NFE2-ChIP-Seq(GSE31477)/Homer	1.00E-01	0.0443

Table S3. The p-values of the Student's t-tests in the detection limit analysis. For the testing dilution series, a one-sided Student's t-test was performed between the estimated tissue fractions in the samples at each dilution level and the control samples (i.e., 0% tissue fraction).

depth	ground truth tissue fraction	*cfSort*	NNLS	CelFiE
20x	0	0.5	0.5	0.5
20x	0.001	0.028402	0.510893	0.301254
20x	0.003	0.000292	0.562574	0.205946
20x	0.005	3.47E-17	0.383126	0.010409
20x	0.007	1.77E-25	0.367383	0.007062
20x	0.01	3.57E-82	0.406064	0.000184
20x	0.03	1.89E-246	0.165466	2.76E-78
20x	0.05	0	0.009471	1.48E-146
20x	0.07	0	4.90E-08	2.37E-177
20x	0.1	0	3.59E-21	4.06E-218
20x	0.13	0	1.16E-46	9.05E-245
20x	0.15	0	6.60E-58	1.46E-234
20x	0.17	0	3.94E-73	2.95E-241
20x	0.2	0	3.44E-101	2.68E-266
20x	0.23	0	3.66E-114	7.81E-252
20x	0.25	0	1.69E-136	1.58E-261
20x	0.27	0	1.90E-147	2.51E-255
20x	0.3	0	4.04E-173	1.16E-269
40x	0	0.5	0.5	0.5
40x	0.001	0.002906	0.482749	0.385675
40x	0.003	1.40E-17	0.47991	0.19932
40x	0.005	1.51E-41	0.456067	0.03948
40x	0.007	2.51E-60	0.446784	0.006553
40x	0.01	2.68E-85	0.42277	5.77E-06
40x	0.03	2.95E-174	0.241485	1.83E-60
40x	0.05	6.08E-193	0.041948	3.25E-92
40x	0.07	2.47E-204	7.81E-05	5.94E-107
40x	0.1	1.50E-224	2.95E-13	2.51E-124
60x	0	0.5	0.5	0.5
60x	0.001	0.004528	0.488055	0.722332
60x	0.003	2.27E-22	0.478513	0.506807
60x	0.005	1.13E-47	0.467371	0.206796
60x	0.007	1.75E-65	0.452665	0.007831
60x	0.01	4.27E-89	0.437801	4.33E-05
60x	0.03	3.93E-180	0.238676	3.24E-53
60x	0.05	3.96E-203	0.043386	2.10E-83
60x	0.07	6.47E-191	9.03E-05	3.01E-110
60x	0.1	3.81E-202	5.91E-13	1.02E-122
90x	0	0.5	0.5	0.5
90x	0.001	1.25E-05	0.493773	0.43938
90x	0.003	2.16E-21	0.47513	0.200315
90x	0.005	6.67E-56	0.465318	0.021783

90x	0.007	7.99E-71	0.452403	0.000801
90x	0.01	2.64E-95	0.427204	3.03E-06
90x	0.03	1.50E-188	0.25716	7.49E-51
90x	0.05	3.53E-195	0.035521	8.22E-82
90x	0.07	1.79E-194	5.19E-05	4.55E-97
90x	0.1	2.82E-212	4.54E-13	1.68E-124
120x	0	0.5	0.5	0.5
120x	0.001	8.13E-08	0.494404	0.172485
120x	0.003	2.73E-28	0.478959	0.042327
120x	0.005	1.26E-62	0.469915	0.014322
120x	0.007	7.99E-82	0.452151	1.61E-05
120x	0.01	1.17E-97	0.433203	2.40E-10
120x	0.03	1.71E-193	0.248487	1.25E-52
120x	0.05	8.28E-199	0.032637	9.31E-78
120x	0.07	4.28E-198	5.65E-05	2.55E-96
120x	0.1	9.17E-217	7.16E-13	3.51E-113

Table S4. Pearson's correlation between the biochemical test results and the corresponding tissue fraction in the serial cfDNA of the NSCLC patients under anti-PD-1 immunotherapy. Note that the date that the patient took a biochemical test may not be exactly the same as the blood collection date. Therefore, we matched the biochemical test result of the nearest date to each plasma sample. The Pearson's correlation was calculated between the tissue fractions from *cfSort* and the matched biochemical test results.

patient	tissue	biochemical test	Pearson correlation with tissue fraction	test results
plasma-304	liver	ALP(LDQ)	0.7786278	abnormal
plasma-317	kidney	BUN(LDQ)	0.9993758	abnormal
plasma-317	kidney	CREATININE(LDQ)	0.9993758	abnormal
plasma-318	liver	ALP(LDQ)	0.6909728	abnormal
plasma-318	liver	ALT(LDQ)	0.9869658	abnormal
plasma-318	liver	AST(LDQ)	0.9950926	abnormal
plasma-319	liver	ALP(LDQ)	0.8451881	abnormal
plasma-304	liver	ALT(LDQ)	0.6970593	normal
plasma-304	liver	AST(LDQ)	0.9877352	normal
plasma-304	liver	DIRECT BILIRUBIN(LDQ)	-0.5285949	normal
plasma-304	liver	TOTAL BILIRUBIN(LDQ)	NA	normal
plasma-304	kidney	BUN(LDQ)	-0.3750709	normal
plasma-304	kidney	CREATININE(LDQ)	NA	normal
plasma-318	liver	DIRECT BILIRUBIN(LDQ)	0.99288	normal
plasma-318	liver	TOTAL BILIRUBIN(LDQ)	-0.9771741	normal
plasma-318	kidney	BUN(LDQ)	NA	normal
plasma-318	kidney	CREATININE(LDQ)	NA	normal
plasma-319	liver	ALT(LDQ)	0.2862293	normal
plasma-319	liver	AST(LDQ)	0.6544237	normal
plasma-319	liver	DIRECT BILIRUBIN(LDQ)	0.5288485	normal
plasma-319	liver	TOTAL BILIRUBIN(LDQ)	0.9553871	normal
plasma-319	kidney	BUN(LDQ)	NA	normal
plasma-319	kidney	CREATININE(LDQ)	NA	normal

Table S5. Overlap between RRBS and the cell-type specific markers identified in a recent study [28].

cell-type markers in [28]	total number of markers	number of overlapped markers	fraction of overlapped markers
top25	1246	787	63.2%
top250	11713	6046	51.6%
top1000	50286	23548	46.8%

Table S6. The source of the orthogonal validation data for the tissue marker atlas.

consortium	data	URL	number of samples
Epigenome Roadmap	WGBS	https://egg2.wustl.edu/roadmap/data/byDataType/dnamethylation/WGBS/	1 adipose tissue, 1 adrenal gland, 1 blood vessel, 2 colon, 1 esophagus, 2 heart, 1 kidney, 1 liver, 1 lung, 1 ovary, 1 pancreas, 1 small intestine, 1 spleen, 1 stomach
ENCODE	ChIP-seq	https://www.encodeproject.org/	1 adipose tissue, 9 adrenal gland, 1 bladder, 5 blood vessel, 4 colon, 5 esophagus, 9 heart, 1 kidney, 2 liver, 6 lung, 1 muscle, 4 nerve, 1 ovary, 8 pancreas, 1 prostate, 1 skin, 2 small intestine, 9 spleen, 6 stomach, 1 testis, 2 thyroid, 2 uterus, 1 vagina
GTEx	RNA-seq	https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/	25 adipose tissue, 15 adrenal gland, 6 bladder, 32 blood vessel, 16 breast, 11 cervix uteri, 29 colon, 45 esophagus, 7 fallopian tube, 26 heart, 13 kidney, 14 liver, 16 lung, 11 muscle, 13 nerve, 17 ovary, 14 pancreas, 17 pituitary, 15 prostate, 14 salivary gland, 23 skin, 16 small intestine, 18 spleen, 14 stomach, 17 testis, 14 thyroid, 14 uterus, 17 vagina

SI References

[1] Meissner, Alexander, et al. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis." *Nucleic acids research* 33.18 (2005): 5868-5877.

[2] Kennedy, Scott R., et al. "Detecting ultralow-frequency mutations by Duplex Sequencing." *Nature protocols* 9.11 (2014): 2586-2606.

[3] Stackpole, Mary L., et al. "Cost-effective methylome sequencing of cell-free DNA for accurately detecting and locating cancer." *Nature communications* 13.1 (2022): 1-12.

[4] Andrews, S., et al. "Trim Galore." *Trim Galore* (2015).

[5] Krueger, Felix, and Simon R. Andrews. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications." *bioinformatics* 27.11 (2011): 1571-1572.

[6] Kent, W. James, et al. "The human genome browser at UCSC." *Genome research* 12.6 (2002): 996-1006.

[7] Krueger, F. Unique Molecule Identifiers (UMIs) based sequencing deduplication software. *https://github.com/FelixKrueger/Umi-Grinder.*

[8] Kundaje, Anshul, et al. "Integrative analysis of 111 reference human epigenomes." *Nature* 518.7539 (2015): 317-330.

[9] Lonsdale, John, et al. "The genotype-tissue expression (GTEx) project." *Nature genetics* 45.6 (2013): 580-585.

[10] Sloan, Cricket A., et al. "ENCODE data at the ENCODE portal." *Nucleic acids research* 44.D1 (2016): D726-D732.

[11] Fishilevich, Simon, et al. "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards." *Database* 2017 (2017).

[12] Tsankov, Alexander M., Hongcang Gu, Veronika Akopian, Michael J. Ziller, Julie Donaghey, Ido Amit, Andreas Gnirke, and Alexander Meissner. "Transcription factor binding dynamics during human ES cell differentiation." *Nature* 518, no. 7539 (2015): 344-349.

[13] Heinz, Sven, et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." *Molecular cell* 38.4 (2010): 576-589.

[14] Stelzer, Gil, et al. "The GeneCards suite: from gene data mining to disease genome sequence analyses." *Current protocols in bioinformatics* 54.1 (2016): 1-30.

[15] Safran, Marilyn, Naomi Rosen, Michal Twik, Ruth BarShir, Tsippi Iny Stein, Dvir Dahary, Simon Fishilevich, and Doron Lancet. "The genecards suite." *Practical guide to life science databases* (2021): 27-56.

[16] Bradley, Paul S., Kristin P. Bennett, and Ayhan Demiriz. "Constrained k-means clustering." *Microsoft Research, Redmond* 20.0 (2000): 0.

[17] Yu, Stephanie CY, et al. "Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing." *Proceedings of the National Academy of Sciences* 111.23 (2014): 8583-8588.

[18] Mouliere, Florent, and Nitzan Rosenfeld. "Circulating tumor-derived DNA is shorter than somatic DNA in plasma." *Proceedings of the National Academy of Sciences* 112.11 (2015): 3178-3179.

[19] Harris, Charles R., et al. "Array programming with NumPy." *Nature* 585.7825 (2020): 357-362.

[20] Menden, Kevin, et al. "Deep learning–based cell composition analysis from tissue expression profiles." *Science advances* 6.30 (2020): eaba2619.

[21] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[22] Hahnloser, Richard HR, et al. "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit." *nature* 405.6789 (2000): 947-951.

[23] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *International conference on machine learning*. PMLR, 2015.

[24] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15.1 (2014): 1929-1958.

[25] Kingma, Diederik P., and Jimmy Ba. "Adam: a method for stochastic optimization 3rd int." *Conf. for Learning Representations, San*. 2014.

[26] Sun, Kun, et al. "Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments." *Proceedings of the National Academy of Sciences* 112.40 (2015): E5503-E5512.

[27] Moss, Joshua, et al. "Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease." *Nature communications* 9.1 (2018): 1-12.

[28] Loyfer, Netanel, et al. "A DNA methylation atlas of normal human cell types." *Nature* (2023): 1-10.

[29] Caggiano, Christa, et al. "Comprehensive cell type decomposition of circulating cell-free DNA with CelFiE." *Nature communications* 12.1 (2021): 1-13.