FinaleMe: Predicting DNA methylation by the fragmentation patterns of plasma cell-free DNA

Yaping Liu[1-9 #], Sarah C. Reed[8], Christopher Lo[8,], Atish D. Choudhury[8,10], Heather A. Parsons[10], Daniel G. Stover[10], Gavin Ha[8], Gregory Gydush[8], Justin Rhoades[8], Denisse Rotem[8], Samuel Freeman[8], David Katz[1-3], Ravi Bandaru[1-3], Haizi Zheng[3], Hailu Fu[1-3], Viktor A. Adalsteinsson[6,#], Manolis Kellis[6,7,#]

Affiliations:

1. Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611

2. Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, IL 60611

3. Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229

4. Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229

5. Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229

6. University of Cincinnati Center for Environmental Genetics, Cincinnati, OH 45229

7. University of Cincinnati Cancer Center, Cincinnati, OH 45229

8. Broad Institute of MIT and Harvard, Cambridge, MA 02142

9. Massachusetts Institute of Technology, Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139

10. Dana-Farber Cancer Institute, Boston, MA, USA

Corresponding email: lyping1986@gmail.com (Y.L.), viktor@broadinstitute.org (V.A), and manoli@mit.edu (M.K)

Abstract

Analysis of DNA methylation in cell-free DNA (cfDNA) reveals clinically relevant biomarkers but requires specialized protocols and sufficient input material that limits its applicability. Millions of cfDNA samples have been profiled by genomic sequencing. To maximize the gene regulation information from the existing dataset, we developed FinaleMe, a non-homogeneous Hidden Markov Model (HMM), to predict DNA methylation of cfDNA and, therefore, tissues-of-origin directly from plasma whole-genome sequencing (WGS). We validated the performance with 80 pairs of deep and shallow-coverage WGS and whole-genome bisulfite sequencing (WGBS) data.

Keywords:

Cell-free DNA, fragmentation, DNA methylation, tissues-of-origin, whole-genome sequencing

Background

DNA methylation plays an instrumental role in gene regulation during disease progression and embryonic development[1,2]. Genome-wide DNA methylation level in cfDNA has been extensively studied for disease diagnosis and prognosis[3–7]. The current gold standard to measure DNA methylation from cfDNA molecules is bisulfite sequencing[8]. However, sodium bisulfite treatment causes non-uniformly sequence-dependent degradation of most DNA fragments[9,10]. The substantial loss of input DNA during the bisulfite treatment limits the sensitivity of diagnostic tests and analyses[11]. Recent advances in enzymatic conversion and long-read sequencing approaches have partly mitigated these issues but have yet to be widely applied in clinics[12–16].

Unlike genomic DNA (gDNA), cfDNA is not randomly fragmented and its fragmentation pattern is highly associated with the local epigenetic background[17,18]. Several recent studies have identified significantly different DNA fragmentation patterns between methylated and unmethylated cfDNA molecules[7,19,20]. These findings suggest the possibility of computationally inferring DNA methylation levels from cfDNA fragmentation patterns. One recent study provided a proof-of-concept solution to predict the binary status of DNA methylation in ultra-high-coverage WGBS through a deep-learning model[19]. However, the ability to predict methylation status from cfDNA WGS remains unexplored. The 2020 American College of Obstetricians and Gynecologists (ACOG) guidelines recommend non-invasive prenatal testing (NIPT) for all pregnancies regardless of risk, which will eventually result in millions of shallow-coverage (~0.1X-1X) cfDNA WGS every year in the US. In addition, hundreds of thousands of cfDNA WGS samples have already been sequenced for cancer early detection and other purposes worldwide by academic communities and commercial entities[21]. Given the potential to leverage cfDNA WGS datasets to advance understanding of gene regulation and human health[22], we developed a computational method, named FinaleMe (**F**ragmentat**I**o**N** **A**na**L**ysis of c**E**ll-free DNA **Me**thylation), to predict the DNA methylation status in each CpG at each cfDNA fragment and obtain the continuous DNA methylation level at CpG sites, mostly accurate in CpG rich regions. We further predicted the associated tissues-of-origin status directly from the fragmentation patterns in cfDNA WGS. We validated the predictions of both methylation level and tissues-of-origin status using paired WGS and WGBS of plasma cfDNA from the same tube of blood across different physiological conditions at deep (~16-39X) and shallow (~0.1X) WGS.

Results

Since DNA methylation has been tightly correlated with nucleosome occupancy[23,24], we hypothesized that if the boundaries of cfDNA fragments are biased by their association with nucleosomes, then the fragmentation pattern observed in each cfDNA molecule should indicate its associated DNA methylation pattern and thus its tissue-of-origin. To evaluate this hypothesis, we first studied the correlation between fragment size and mean methylation level of DNA fragments from publicly available WGBS of cfDNA and gDNA of buffy coat samples from two healthy individuals[7] (**Figure 1**). Replicate samples of cfDNA showed waved methylation patterns at mono-nucleosomal lengths that were not present in the gDNA samples. This observation supported the hypothesis that the fragmentation pattern of cfDNA can provide information related to the DNA methylation level.

Next, we built a non-homogeneous Hidden Markov Model, named FinaleMe, to predict the methylation status in cfDNA (details in Methods and Supplementary Methods, **Figure 2**). Since CpG is not evenly distributed in the human genome, we incorporated the distance between CpG sites into the model and utilized the following

three features: fragment length, normalized coverage, and the distance of each CpG to the center of the DNA fragment (**Figure 1b**). We first evaluated the model using high coverage WGBS of cfDNA (non-pregnant healthy individuals), masking the methylation status, and then benchmarked the model performance using the ground truth DNA methylation states at each CpG in each DNA fragment. After sampling an equal number of the methylated and unmethylated CpGs, we observed high performance in predicting the methylation status at each single CpG from each DNA fragment based on the area under the receiver operating characteristic curve (auROC) within CpG-rich regions (auROC=0.91, for CpGs at fragments with ≥ 5 CpGs, **Figure 1c**).

To further benchmark the model performance in cfDNA WGS, we generated our own matched high coverage WGS (~16X-39X) and WGBS (~10-15X) libraries at the plasma cfDNA samples from the same tube of blood in the healthy individuals and a prostate cancer patient (**Figure 3a**, **Supplementary Table 1**). Without using cfDNA WGBS data, we trained the HMM model and predicted the methylation level from the same cfDNA WGS dataset. By comparing the results with the methylation level at CpG sites in the reference genome from matched WGBS, we achieved a high correlation at single-CpGs and 1kb windows in CpG-rich regions (CpG island and CpG island shore regions, **Figure 3b-c**). At differentially methylated regions (DMRs) detected in the cfDNA WGBS between cancer and healthy individuals at CpG-rich regions, we also observed consistent methylation changes in the predicted methylation levels from matched cfDNA WGS (**Figure 3d**). To check the potential overfitting problem of the model, we further trained and decoded the model at gDNA WGS from cancer and normal blood cells, in which the fragments are sonicated and do not have a correlation with the epigenetics status. The predicted results at gDNA WGS did not show any methylation differences between cancer and normal cells in the DMRs detected at the matched gDNA WGBS datasets (**Supplemental Fig. 1a**). This result suggested that the differential methylation we predicted in cfDNA WGS is not driven by the methylation prior we used but indeed the fragmentation features. However, we noticed that, in the CpG-poor regions, the model in cfDNA WGS did not work well as that in CpG-rich regions (**Supplemental Fig. 1b**). We further assessed the methylation level at important regulatory elements, such as CpG island (CGI) promoters (**Figure 3e**), 5'exon boundaries, and CTCF motifs (**Supplemental Fig. 2**). These results showed a high correlation between the ground truth (WGBS) and the prediction (WGS) in cfDNA from both healthy individuals and the cancer patient (**Figure 3e**, **Supplemental Fig. 2-3**), but not in gDNA dataset (**Supplemental Fig. 4**).

Since DNA methylation in CGI and CGI shore regions are often cell-type-specific, we further estimated the tissue-of-origin in cfDNA by using DNA methylation levels that were measured or predicted using WGBS and WGS, respectively. We found similar tissue-of-origin profiles between predicted and measured methylation levels for each of the individuals in both cancer and healthy conditions (**Figure 3f**), which is also largely consistent with other previous tissues-of-origin studies by cfDNA WGBS[3,6].

Deep coverage WGBS and WGS remain costly for routine clinical application. Many publicly available cfDNA WGS data sets are sequenced with shallow coverage (0.1X-1X). We sought to determine whether we could predict DNA methylation levels using ultra-low-pass whole-genome sequencing (~0.1X, ULP-WGS). We generated matched ULP-WGS and ultra-low-pass WGBS (~0.1X, ULP-WGBS) of cfDNA from 77 individuals, including healthy donors, breast, and prostate cancer patients (**Supplementary Table 1**). We examined the methylation level globally and at important regulatory elements, such as CGI promoters, and observed similar average methylation profiles in predicted and measured methylation levels from ULP-WGS and WGBS, respectively (**Fig. 4a,b**). We also observed the differential methylation level in ULP-WGS at differentially methylated regions detected in ULP-WGBS (**Supplemental Fig. 5**). Next, we assessed whether methylation

levels from ultra-low-pass sequencing could be utilized for the estimation of tissues-of-origin. We downsampled the deep coverage sequencing results and found largely consistent tissue-of-origin estimates with ultra-low-pass sequencing (**Supplemental Fig. 6**). Finally, we estimated the tissue-of-origin in both ULP-WGS and ULP-WGBS. We found consistent results between the two assays. The fractions of prostate or breast-originated cell types are low in healthy individuals and showed a high correlation with tumor fraction as estimated by copy number variations (ichorCNA) across all samples in both assays (**Fig. 4c**). These results suggested that the application of FinaleMe to ULP-WGS is consistent with the ground truth in terms of both DNA methylation and tissues-of-origin predictions.

Discussions

Our study demonstrates the ability to infer cfDNA methylation level and tissues-of-origin status directly from deep and shallow-coverage cfDNA WGS. This overcomes a major hurdle associated with bisulfite conversion of limited amounts of cfDNA and, more importantly, enables the usage of a large number of existing, publicly available cfDNA genomic datasets for epigenetic analysis. Our predictions are most accurate in CpG-rich regions of the genome but not in CpG-poor regions. Further work is required to improve the predictions in CpG-poor regions for the detection of other disease-related methylation features, such as the partially methylated domains in cancers. Moreover, the Bayesian prior we utilized from genomic DNA methylome may cause overfitting problems and the false positive call of DMRs in cancer WGS. Previous studies have suggested that analysis of tissue-of-origin is possible based on analysis of nucleosome spacing in WGS of cfDNA[17]. However, only the relative rank of most related cell types is estimated in deep WGS. The tissues-of-origin estimation from inferred DNA methylation here can provide the estimation of absolute fractions in each cell type and utilize the rich reference methylome resources. Although we do not expect to replace bisulfite sequencing for direct measurement of methylation levels, we provide a generalizable method that could enable the methylation analysis of cfDNA samples with limited material or samples that would otherwise only undergo genomic profiling.

Methods

1. Clinical samples.

Cancer patient blood samples were obtained from appropriately consented patients as described in Adalsteinsson et al[25]. Healthy donor blood samples were obtained from appropriately consented individuals from Research Blood Components (http://researchbloodcomponents.com/services.html). Samples were collected and fractionated as described in Adalsteinsson et al[25].

2. Whole-genome bisulfite sequencing of cfDNA.

Library construction was performed on 25 ng of cfDNA using the Hyper Prep Kit (Kapa Biosystems) with NEXTFlex Bisulfite-Seq Barcodes (Bioo Scientific) and methylated adapters (IDT) along with HiFi Uracil+ polymerase (Kapa Biosystems) for library amplification. NEXTFlex Bisulfite-Seq Barcodes were used at a final concentration of 7.5 uM and the EZ-96 DNA Methylation-Lightning MagPrep kit (Zymo Research) was used for bisulfite conversion of the adapter-ligated cfDNA prior to library amplification. Libraries were sequenced using paired-end 100bp in the platform of HiSeq2500 (Illumina) with a 20% spike of PhiX.

3. Whole-genome sequencing of cfDNA.

Library construction was performed on 5-20 ng of cfDNA using the Hyper Prep Kit (Kapa Biosystems) and custom sequencing adapters (IDT) on a Hamilton STAR-line liquid handling system. Libraries were sequenced using paired-end 100bp in the platform of the HiSeq2500 (Illumina).

4. Model development and training.

4.1. **Data preprocessing.**

For WGS data, reads were aligned to the human genome (GRCh37) using BWA-MEM 0.7.15[26] with default parameters. Each fragment containing CpGs in the autosomal chromosomes reference genome was used for the analysis. Fragment lengths of more than 500bp or less than 30 bp were discarded. Regions with coverage more than 250X or ENCODE blacklist regions (merged *wgEncodeDukeMapabilityRegionsExcludable* and *wgEncodeDacMapabilityConsensusExcludable*) were also discarded. Only high-quality reads were considered in the following analysis (high quality: uniquely mapped, no PCR duplicates, both of ends are mapped with mapping qualities more than 30 and properly paired). To calculate the methylation status for each CpG in each fragment, only bases with a base quality of more than 5 were used.

For cfDNA WGBS data, a recent study demonstrated that the existence of the jagged-end at the end of cfDNA fragment will affect the estimation accuracy of DNA methylation[27]. We first generated the M-bias plot by using Bismark[28] to map the reads without trimming (see **Supplementary Figure 7**). To avoid the artifact potentially brought by the jagged end for Figure 1a, we trimmed the 40bp from the 5' end and 10bp from 3' end at the R2 reads. The 3'end of R1 reads seems to be not affected by the jagged-end problem. However, in CpG islands (often open chromatin regions), cfDNA fragments are usually very small. To avoid the potential bias at these small fragments, we also trimmed 40bp from 3' end at the R1 reads, and the results were still largely the same. After trimming, reads were aligned to the human genome (GRCh37) using Bismark (v0.22.3) with bowtie2 (v2.3.5)[29]. The methylation status of CpGs was counted from the first converted cytosine in each of the fragments as described in Bis-SNP[30]. Fragment coverage at each CpG site was first normalized by dividing the total number of high-quality reads in the bam file. Further, the three features (fragment length, normalized coverage, and distance to the center of the fragment) were transformed into Z-score by the mean and standard deviation of the features within the same bam file as the input for the HMM model (**Figure 2**). All details are implemented in 'CpgMultiMetricsStats.java' (with parameters "-stringentPaired" for only high-quality fragments and with parameters "-wgsMode" for WGS data). The methylation level from WGBS was called by Bis-SNP v0.90[30].

4.2. **Non-homogeneous Hidden Markov Model.**

The initiation matrix was summarized based on the methylation states of the first CpG in each DNA fragment separately (**Figure 2**). A nonparametric model was used to calculate the initiation and transition matrix by considering the distance with adjacent CpG sites. A gaussian mixture model was applied to model the emission likelihood of each of the three fragmentation features (fragment length, coverage, and distance to the center of the fragment). A weighted DNA methylation prior, estimated from methylation level at genomic DNA (buffy coat) in healthy individuals, was utilized to calculate the posterior emission probability of hidden status only in the decoding (i.e., prediction) step, which models the base DNA methylation differences in different genomic contexts. For example, the probability of observing methylated event *em* given that located at the CpG site with methylation prior *k* is:

$$\Pr(e_m) = \frac{\Pr(e_m \mid k)\Pr(k)}{\Pr(e_m \mid k)\Pr(k) + \Pr(e_u \mid 1-k)(1-\Pr(k))}$$

Two states Hidden Markov Model (HMM) is implemented as described in Rabiner 1990[31] at Jahmm framework with some adaptations to our problem. Baum-Welch algorithm was used to estimate the parameters with a maximum of 50 iterations. The model was trained by all the cfDNA fragments with at least 7 CpGs within the same fragments. The number of CpGs was not limited at the decoding step. In low-coverage data, we utilized an HMM model trained in high-coverage samples (HD_45, a healthy individual) to estimate the model parameters and applied it directly to each ULP-WGS dataset for the decoding. All details are implemented in 'FinaleMe.java' (with parameters "-miniDataPoints 7 -gmm -covOutlier 3" for the training step and parameters "-decodeModeOnly" for the decoding step).

4.2.1. Gaussian Mixture Model (GMM) initialization for HMM model.
GMM algorithm was utilized to estimate the initiation state of each CpG in each fragment by three fragmentation feature vectors with a maximum of 10,000 iterations. After GMM initialization, in WGBS, the methylated and unmethylated states were identified by the mean methylation level of each state. In WGS data, the state with a higher distance to the center was defined as the methylated state. Then the initiation parameters of HMM model were estimated based on the GMM initialization.

4.2.2. Initiation and transition probability.
The initiation probability of each state with the same offset from the start of the fragment was averaged by the states of the first CpGs with the same offset range at all the high-quality fragments. The transition probability matrix between states was also calculated separately for each of the possible distance ranges to the previous CpG.

4.2.3. Emission distributions.
Three features were modeled by Multivariate Mixture Gaussian distribution. Two components mixture of Gaussian distribution was used to model each of the features separately.

$$\Pr(e_m \mid k) = (1-\pi) * N\left(\mu_i, \sigma_i^2\right) + \pi * N\left(\mu_j, \sigma_j^2\right)$$

In the Viterbi decoding step, methylation prior estimated from genomic DNA in buffy coat samples from healthy individuals[7] was only used to calculate the emission probability for each CpG.

4.2.4. KL divergence.
Kullback-Leibler distance was used to estimate the divergence of new HMM during Baum-Welch re-estimation. Since methylation prior was used for the decoding step and is different at different CpG site, 10,000 random fragments with a minimum of 5 CpGs is selected to calculate the Kullback-Leibler distance. If the distance between new and old HMM was less than $1e^{-4}$ or the changes of distance were less than 1%, the model was considered converged.

4.2.5. Summary of the model
In cfDNA WGS (**Figure 2**), our HMM model infers the model parameters directly from WGS data without using cfDNA WGBS data. The principle of the model is: we assume that there are two binary states ("u" or "m") in each CpG at each cfDNA fragment. These two states are not observable in WGS (thus "hidden"). We

assume that the states are affected by three fragmentation features. At each CpG in each fragment in the bam file ("CpG point"), we can obtain three features: the fragment's length, the CpG's distance to the center of that fragment, and the fragment coverage at that particular CpG position in the reference genome. We also assume the status of each CpG in each fragment is a Multivariate Gaussian distribution of these three features.

Step 1, we utilized a Gaussian mixture model to classify all the CpG points in WGS into two groups ("u" or "m") to initiate the HMM model (the initial parameters). Given the hypothesis in Figure 1B, we always assume "m" group has a larger average distance to the center of fragments.

Step 2, we applied the initiated parameters to the HMM model and built a Markov chain for each single cfDNA fragment. Due to the Markov process, the status of each CpG point is affected by its adjacent CpG in the same fragment. Then, the Baum-Welch algorithm was used to estimate the maximum likelihood parameters in the WGS dataset. Different from the traditional HMM model that assumes equal transition probability between CpGs, we utilized a non-homogenous model to estimate different transition probability matrices given different distances between CpGs. Kullback-Leibler distance was utilized to estimate whether or not the model converged during the iteration.

Step 3, after the estimation of parameters in step 2 (training), we utilize the Viterbi algorithm to estimate the best state ("u" or "m") in each CpG at each fragment. Different from the traditional HMM model, we add methylation prior from WGBS in a healthy buffy coat to calculate the posterior probability.

Step 4, after the prediction in step 3, we aggregated the methylation status across fragments at each CpG site in the reference genome and calculated the continuous methylation level (0-100%).

4.3. **Performance evaluation.**

4.3.1. Comparison of the binary methylation status of each CpG in each fragment (WGBS).

The equal number of methylated and unmethylated CpGs was randomly sampled at the evaluation step. Prediction results were compared with ground truth methylation binary states at each CpG in each cfDNA fragment of WGBS. The threshold was varied to identify methylated status at the Viterbi decoding step in order to calculate the ROC curve.

4.3.2. Comparison of the continuous methylation level at each CpG or windows in the reference genome (paired WGBS and WGS).

FinaleMe was trained and decoded at WGS data only. The methylation level was calculated by aggregating the binary methylation status across fragments at each CpG in the reference genome. Finally, the continuous methylation level at each CpG or window was compared with the methylation level obtained from matched WGBS in the same blood draw.

4.3.3. Comparison of methylation profiles at important regulatory elements (paired WGBS and WGS).

FinaleMe was trained and decoded at WGS data. The predicted methylation level was calculated as described in 4.2. The average methylation level around CpG island promoters, 5' end of exons, and CTCF motifs were calculated by Bis-Tools as described in Lay & Liu et al. 2015[32]. CpG island definition was downloaded from UCSC genome browser[33]. CpG island shore was defined by the regions within 2kb regions around the CGI.

4.3.4. Benchmark of the speed

We downsampled the high-coverage cfDNA WGS data and calculated the time cost with different numbers of fragments in the bam files (**Supplementary Figure 8**). Benchmark was performed at a single CPU in the computational cluster (Intel(R) Xeon(R) Gold 6338 CPU @ 2.0GHz).

4.4. **Tissue-of-origin deconvolution**.

To infer tissue of origin from measured or inferred DNA methylation data, we modeled patient methylation data as a linear combination of reference methylomes. We constrain the weights to sum up to one so that the weights can be interpreted as tissue contribution to cfDNA. Quadratic programming was utilized to solve the constrained optimization problem. This method and approach closely follow the tissue deconvolution algorithm described in Sun et al PNAS[6]. To reduce the noise, we utilized the methylation density at 1kb non-overlapped windows within the CpG island and CpG island shore regions at autosomes and binarized the methylation level (window with methylation density <0.1 was defined as 0, otherwise 1) in both reference methylomes and cfDNA data. Only windows with at least 10 Cs or Ts across all the reference methylomes were utilized for the analysis. Only windows that were highly variable across reference methylomes (top 1% most variable regions in the reference methylomes) were further utilized for the deconvolution.

We incorporated WGBS from the major immune cell types (Neutrophil, B cell, T cell, Macrophage, Nature Killer cell, Erythroblast cells), blood vessel endothelial cells, and liver hepatocyte cells, as suggested by Moss 2018 Nature Communications[3]. We also incorporated methylomes from mammary epithelial cells (HMEC) and prostate epithelial cells (PrEC) since they are related to the cancer types we analyzed.

In the low pass data, we further relaxed our criteria about the coverage to keep more windows. The top 25% of most variable regions in the reference methylomes were utilized for deconvolution. Windows with less than 5 Cs or Ts in either reference methylome or cfDNA data were marked as NA. Samples or windows with more than 80% NA were filtered. We further imputed the missing data of the windows by K-nearest neighbor (k=5 and maxp="p" in *impute.knn* function at *impute* package, R 4.2.1) and finally binarized the methylation level within the window as that in high coverage data.

4.5. ichorCNA analysis

Estimation of tumor fraction was performed using ichorCNA as described previously in Adalsteinsson et al. Nature Communications 2017[25].

4.6. Differential methylation analysis

Differential methylation regions (predefined non-overlapped 1kb windows in autosomes) in high-coverage WGBS were identified by metilene (v 0.2-8)[34] with q value < 0.05. Data in ULP-WGBS are very sparse and noisy. Therefore, we utilized two-sided Wilcoxon Rank Sum Tests to identify the windows that were different between cancers and healthy controls with a p value cut-off 0.01.

Availability of data and materials:

Code for FinaleMe and associated scripts are publicly available on GitHub under the MIT license for academic researchers: https://github.com/epifluidlab/FinaleMe.git. The raw sequencing data for ULP-WGS is obtained from Adalsteinsson et al[25](dbGap id: phs001417.v1.p1). The newly generated deep WGS, WGBS, and ULP-WGBS data are deposited at Sequence Read Archive with controlled access (dbGap id: phs003287.v1.p1). All the intermediate results and de-identified data are available at zenodo.org (doi: https://doi.org/10.5281/zenodo.7647046).

Acknowledgments

This work was supported by the computational resources from the Broad Institute of MIT and Harvard, the Biomedical Informatics (BMI) high-performance computing cluster in CCHMC, and QUEST computational cluster in Northwestern University. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562. This work used the XSEDE at the Pittsburgh Supercomputing Center (PSC) through allocation MCB190124P and MCB190006P.

Funding

Y.L. is supported by the Broad Next10 grant from the Broad Institute of MIT and Harvard, trustee award from CCHMC, the startup grant to Y.L. from CCHMC, Northwestern University, Robert H. Lurie Comprehensive Cancer Center of Northwestern University, and NHGRI (R56HG012360 to Y.L.). The authors acknowledge the generous support of the Gerstner Family Foundation to V.A., the Wong Family Foundation and DFCI Medical Oncology grant to A.D.C.

Contributions

Y.L., V.A., and M.K. conceived the study. Y.L. implemented the computational method. S.R. performed the library constructions. Y.L., C.L., D.K, R.B, G.H., G.G, J.R, D.R, H.Z., H.F., and S.F. performed the data analysis with the input from A.D.C., H.A.P, D.G.S, V.A., and M.K. A.D.C., H.A.P, and D.G.S. provided the clinic samples and guidance related to the clinic applications. Y.L. and V.A. wrote the manuscript together. All authors read and approved the final manuscript.

Ethics declarations

Ethics approval and consent to participate

This research study was approved by the Broad Institute Institutional Review Board in accordance with the Declaration of Helsinki. De-identified plasma sample collection was approved by the Dana-Farber Cancer Institute and Broad Institute Institutional Review Boards. All participants provided written informed consent to participate.

Competing interests

Y.L., V.A., and M.K. have an approved patent ("Methods for genome characterization", US Patent 11,788,135, 2023, filed by MIT and Broad Institute of MIT and Harvard). Y.L. owns stocks from Freenome Inc. V.A., G.H. and S.F. are inventors on a patent (US20190078232A1) on methods for estimating tumor fraction in cfDNA.

References

1. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
2. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
3. Moss, J. *et al.* Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 5068 (2018).
4. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann. Oncol.* **31**, 745–759 (2020).
5. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
6. Sun, K. *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5503-12 (2015).
7. Jensen, T. J. *et al.* Whole genome bisulfite sequencing of cell-free DNA and its cellular contributors uncovers placenta hypomethylated domains. *Genome Biol.* **16**, 78 (2015).
8. Olova, N. *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol.* **19**, 33 (2018).
9. Tanaka, K. & Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg. Med. Chem. Lett.* **17**, 1912–1915 (2007).
10. Yi, S., Long, F., Cheng, J. & Huang, D. An optimized rapid bisulfite conversion method with high recovery of cell-free DNA. *BMC Mol. Biol.* **18**, 1–8 (2017).
11. Sun, K. *et al.* Orientation-aware plasma cell-free DNA fragmentation analysis in open chromatin regions informs tissue of origin. *Genome Res.* **29**, 418–427 (2019).
12. Erger, F. *et al.* cfNOMe - A single assay for comprehensive epigenetic analyses of cell-free DNA. *Genome Med.* **12**,

54 (2020).

13. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* (2021) doi:10.1101/gr.266551.120.
14. Liu, Y. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **37**, 424–429 (2019).
15. Yu, S. C. Y. *et al.* Single-molecule sequencing reveals a large population of long cell-free DNA molecules in maternal plasma. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2114937118 (2021).
16. Choy, L. Y. L. *et al.* Single-molecule sequencing enables long cell-free DNA detection and direct methylation analysis for cancer patients. *Clin. Chem.* **68**, 1151–1163 (2022).
17. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57–68 (2016).
18. Ivanov, M., Baranova, A., Butler, T., Spellman, P. & Mileyko, V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* **16 Suppl 13**, S1 (2015).
19. Zhou, Q. *et al.* Epigenetic analysis of cell-free DNA by fragmentomic profiling. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2209852119 (2022).
20. An, Y. *et al.* DNA methylation analysis explores the molecular basis of plasma cell-free DNA fragmentation. *Nat. Commun.* **14**, 287 (2023).
21. Liu, S. *et al.* Genomic Analyses from Non-invasive Prenatal Testing Reveal Genetic Associations, Patterns of Viral Infections, and Chinese Population History. *Cell* **175**, 347-359.e14 (2018).
22. Liu, Y. At the dawn: cell-free DNA fragmentomics and gene regulation. *Br. J. Cancer* **126**, 379–390 (2022).
23. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).
24. Collings, C. K., Waddell, P. J. & Anderson, J. N. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.* **41**, 2918–2931 (2013).
25. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**,

1754–1760 (2009).

27. Jiang, P. *et al.* Detection and characterization of jagged ends of double-stranded DNA in plasma. *Genome Res.* **30**, 1144–1153 (2020).
28. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
30. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* **13**, R61 (2012).
31. Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Readings in Speech Recognition* 267–296 Preprint at https://doi.org/10.1016/b978-0-08-051584-7.50027-9 (1990).
32. Lay, F. D. *et al.* The role of DNA methylation in directing the functional organization of the cancer epigenome. *Genome Res.* **25**, 467–477 (2015).
33. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
34. Jühling, F. *et al.* metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res.* **26**, 256–262 (2016).

Figures:

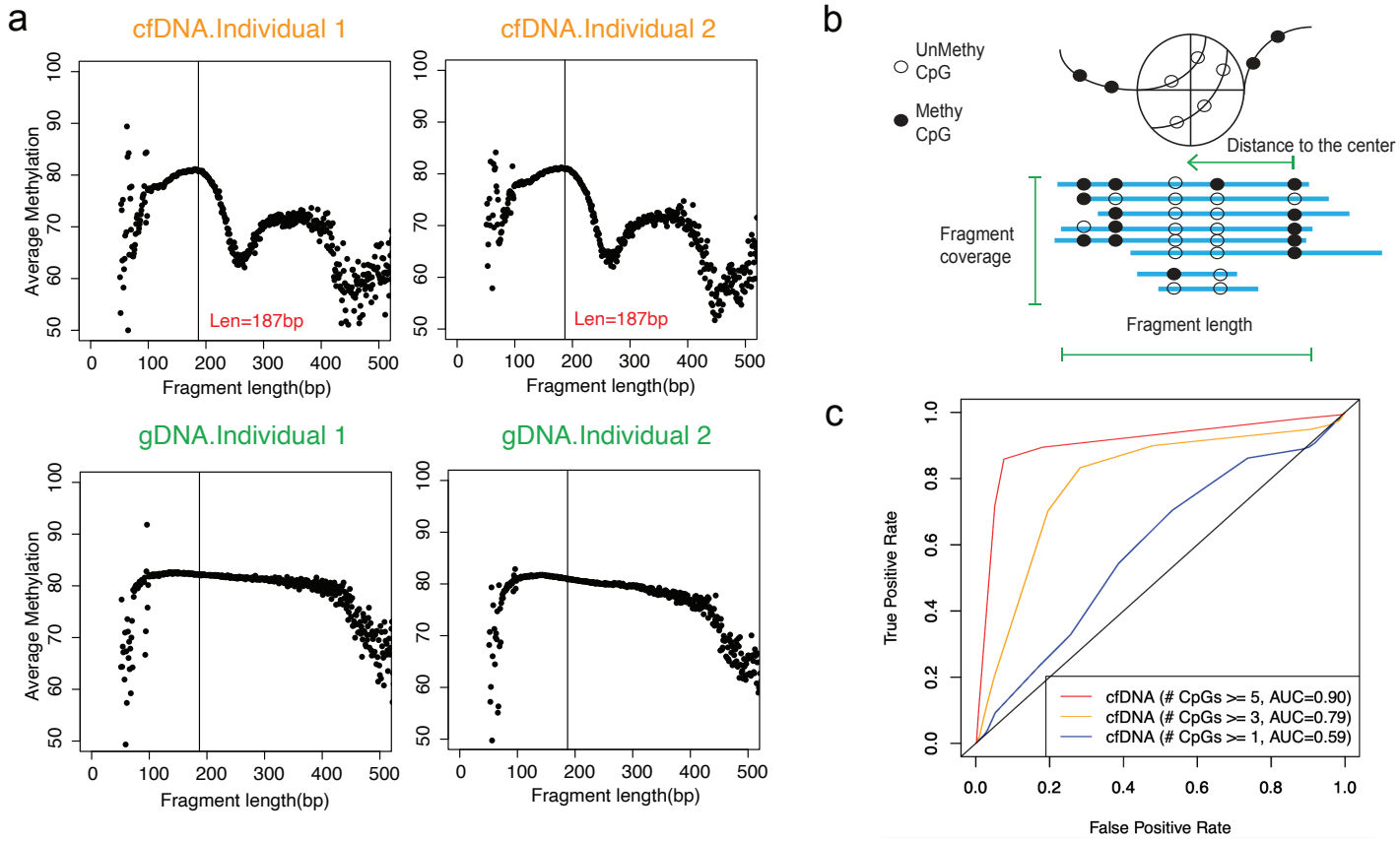


Figure 1. **Inferring DNA methylation from high-coverage whole genome bisulfite sequencing.** a. The correlation between mean DNA methylation and fragment lengths in cfDNA and gDNA WGBS in healthy individuals. b. Diagram of the features utilized for the inference of DNA methylation level. c. ROC curve for the model performance at deep WGBS in fragments with different numbers of CpGs.

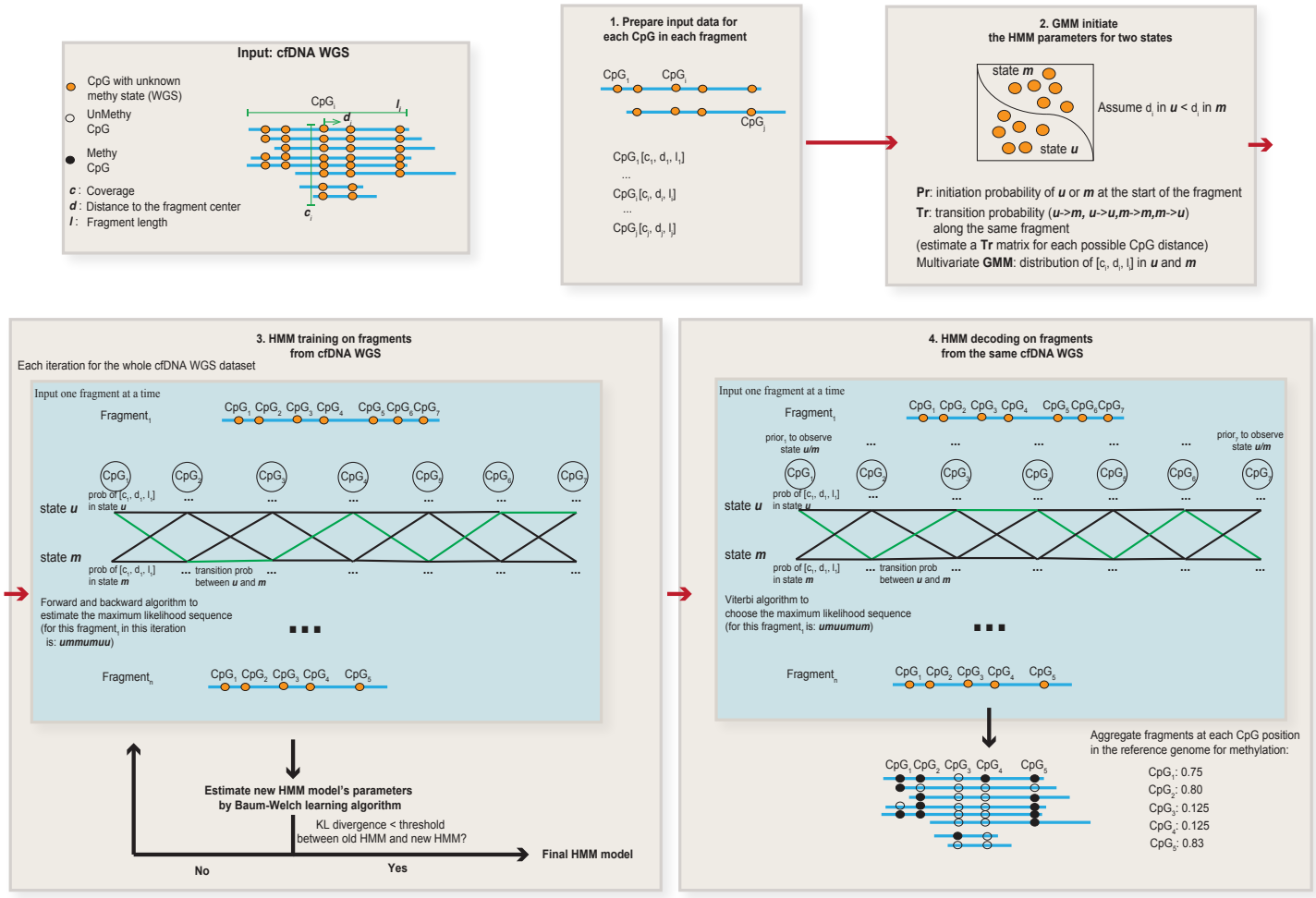
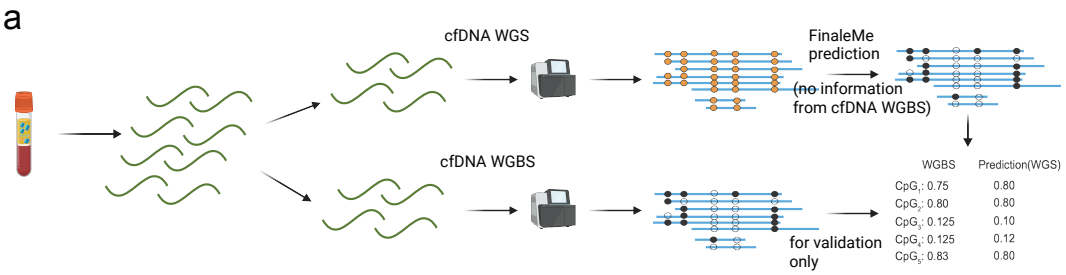
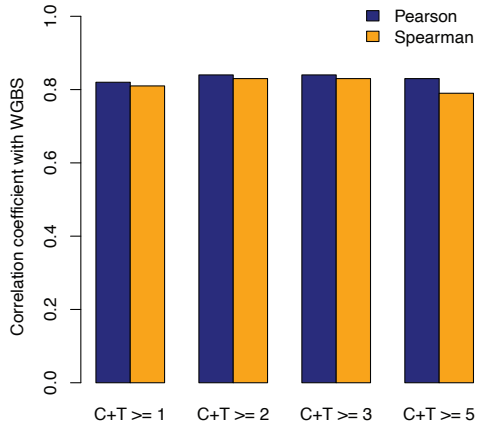
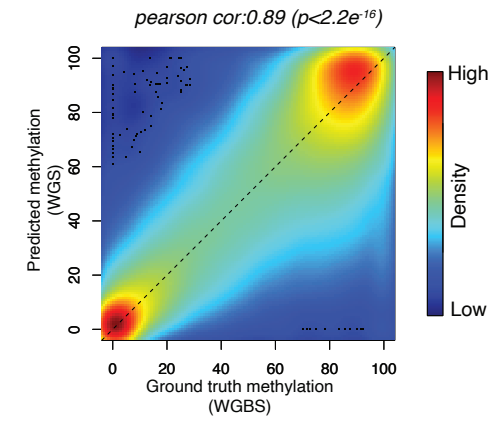
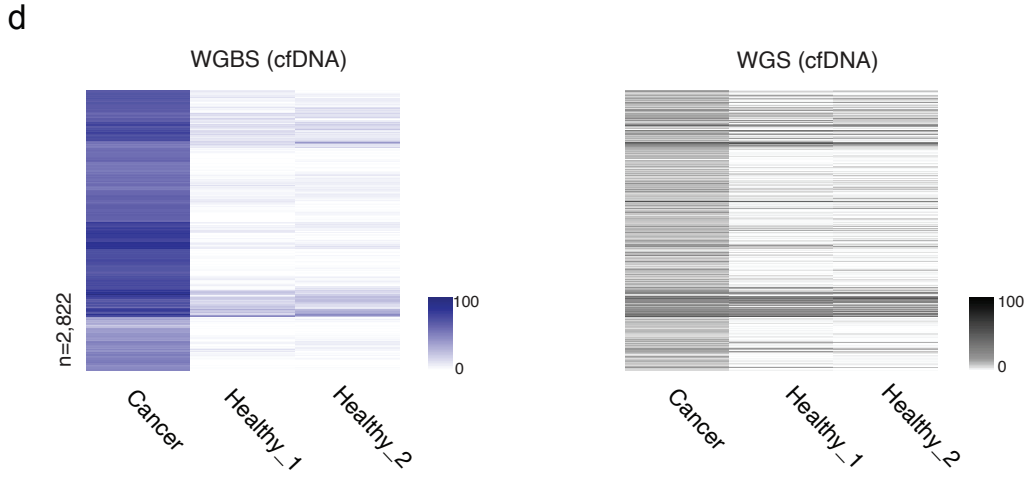


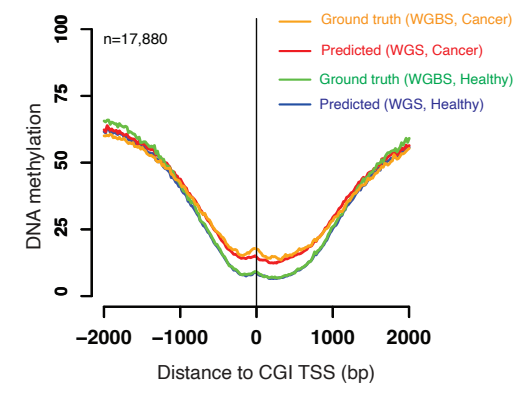
Figure 2. Summary of the HMM model to infer DNA methylation status.











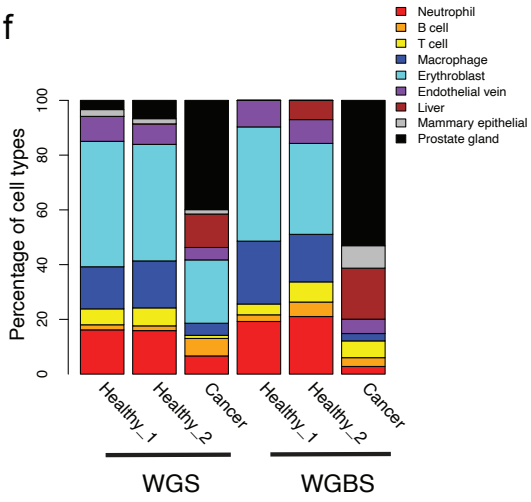


Figure 3. **Inferring DNA methylation from high-coverage whole genome sequencing**. a. workflow to benchmark the model performance. b. Pearson and Spearman correlation of DNA methylation at single CpGs with different coverages at CpG island and CpG island shore regions between matched cfDNA WGBS and WGS. c. Scatterplot of DNA methylation level within 1kb non-overlapped bins at CpG island and CpG island shore regions between matched cfDNA WGBS and WGS. d. Heatmap of measured (left panel, cfDNA WGBS) and predicted (right panel, matched cfDNA WGS) DNA methylation level at hypermethylated differentially methylated windows (1kb) characterized in CGI and CGI shore regions. The row orders in both WGBS and WGS datasets were based on the clustering of DNA methylation levels in WGBS only. e. Average ground truth (WGBS) and predicted (WGS) DNA methylation level at CpG island promoter region from cancer and healthy individuals. f. The fraction of cell types that contributed to cfDNA was estimated by matched WGS and WGBS.

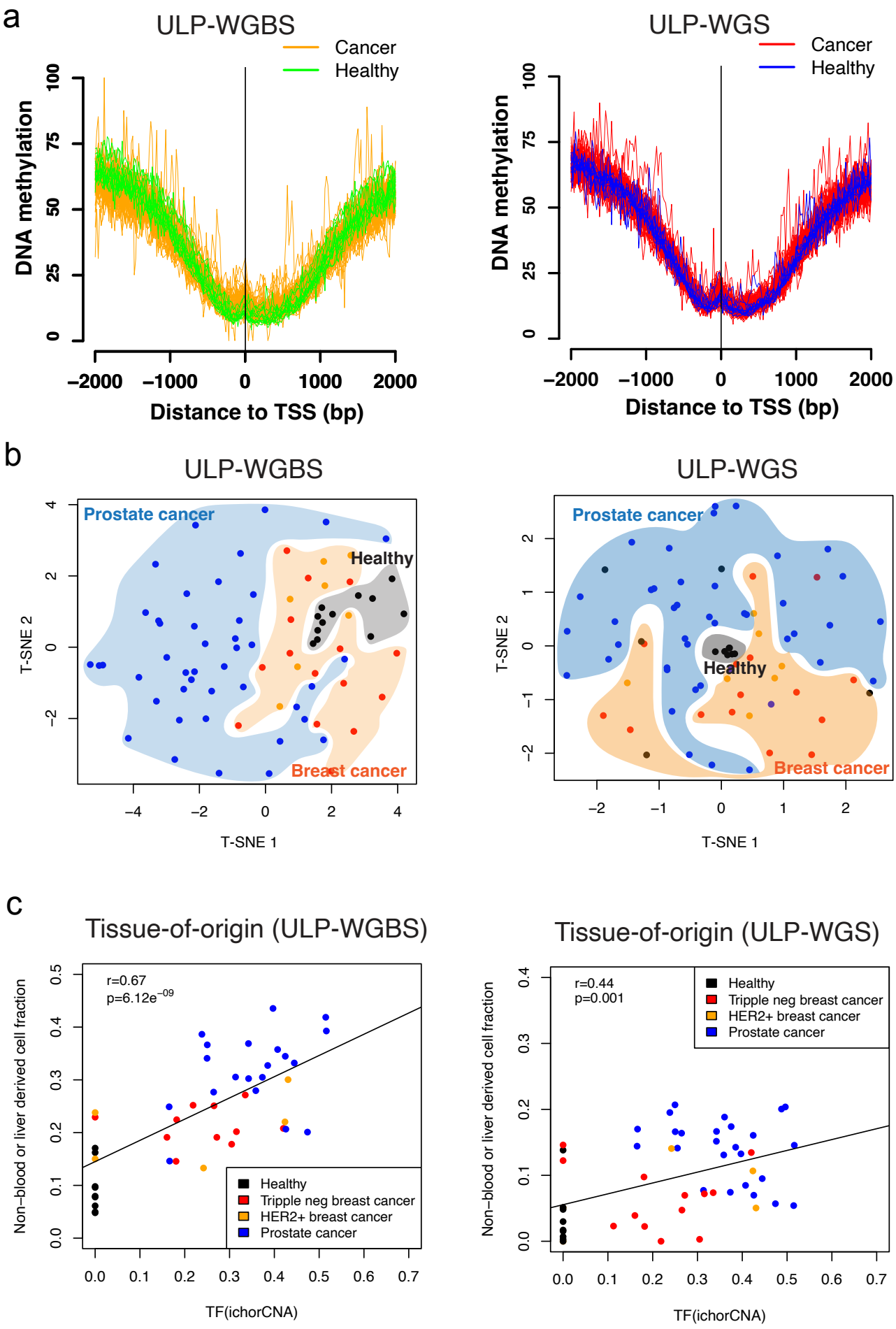
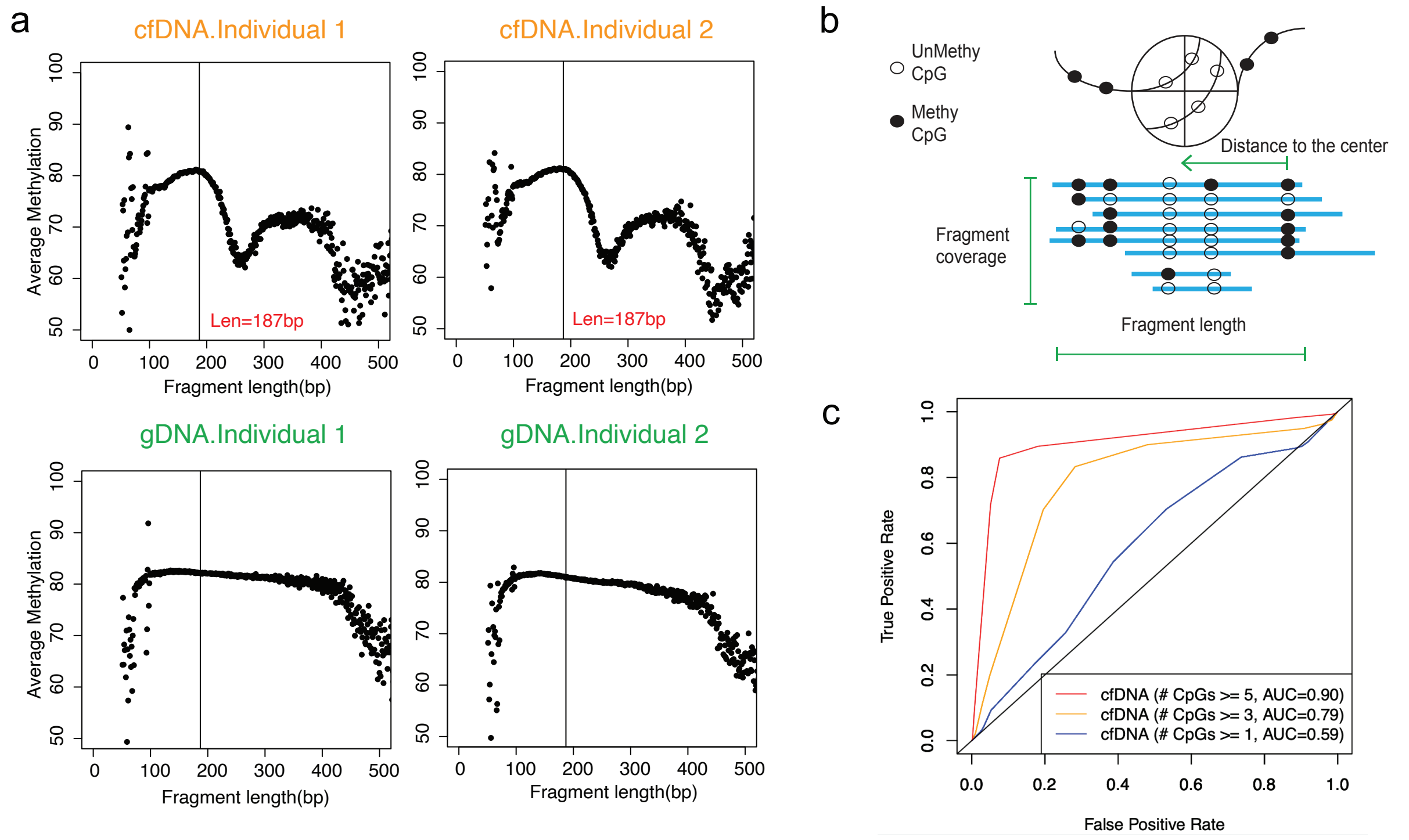


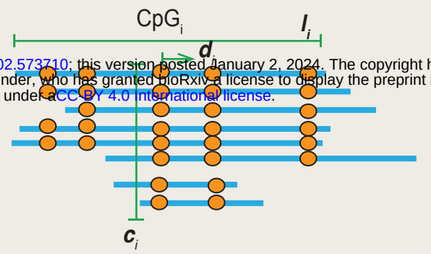
Figure 4. **Inferring DNA methylation and tissues-of-origin from cfDNA ULP-WGS**. a. Average ground truth (ULP-WGBS) and predicted (ULP-WGS) DNA methylation level from cancer and healthy individuals at CpG island promoter regions. b. T-SNE plot by using the DNA methylation level in the 100kb non-overlapped window in autosomes but only summarized from CGI and CGI shore regions in the ground truth (ULP-WGBS) and predicted (ULP-WGS) results from cancer and healthy individuals. c. the concordance of prostate or breast related cell type fractions with tumor fraction estimated by ichorCNA in both healthy and cancers.


a
cfDNA.Individual 1
cfDNA.Individual 2
gDNA.Individual 1
gDNA.Individual 2
Average Methylation
Fragment length(bp)
Len=187bp
b
UnMethy CpG
Methy CpG
Distance to the center
Fragment coverage
Fragment length
c
True Positive Rate
False Positive Rate
cfDNA (# CpGs >= 5, AUC=0.90)
cfDNA (# CpGs >= 3, AUC=0.79)
cfDNA (# CpGs >= 1, AUC=0.59)

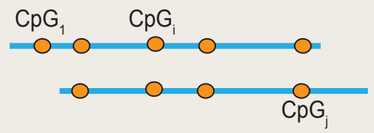
Input: cfDNA WGS

- CpG with unknown methy state (WGS)
- UnMethy CpG
- Methy CpG

c : Coverage
d : Distance to the fragment center
l : Fragment length

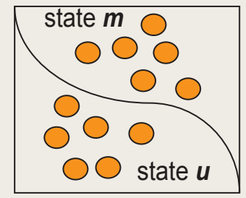


1. Prepare input data for each CpG in each fragment



CpG$_1$ [c$_1$, d$_1$, l$_1$]
...
CpG$_i$ [c$_i$, d$_i$, l$_i$]
...
CpG$_j$ [c$_j$, d$_j$, l$_j$]

2. GMM initiate the HMM parameters for two states



Assume d$_i$ in ***u*** < d$_i$ in ***m***

Pr: initiation probability of ***u*** or ***m*** at the start of the fragment

Tr: transition probability (***u*->*m*, *u*->*u*,*m*->*m*,*m*->*u***)
along the same fragment
(estimate a **Tr** matrix for each possible CpG distance)

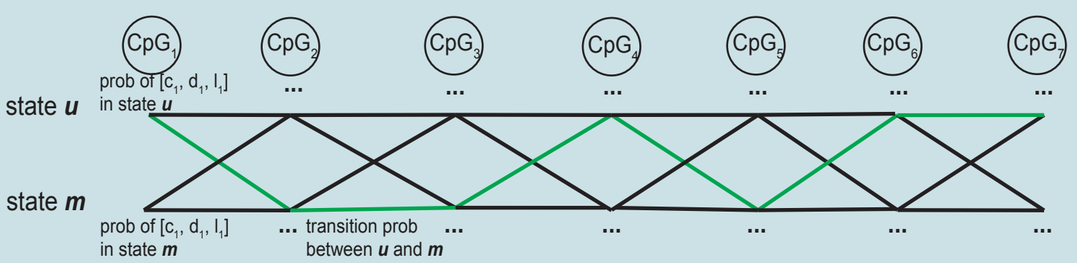
Multivariate **GMM**: distribution of [c$_i$, d$_i$, l$_i$] in ***u*** and ***m***

3. HMM training on fragments from cfDNA WGS

Each iteration for the whole cfDNA WGS dataset

Input one fragment at a time

Fragment$_1$ CpG$_1$ CpG$_2$ CpG$_3$ CpG$_4$ CpG$_5$ CpG$_6$ CpG$_7$



Forward and backward algorithm to estimate the maximum likelihood sequence (for this fragment$_1$ in this iteration is: ***ummumuu***)

Fragment$_n$ CpG$_1$ CpG$_2$ CpG$_3$ CpG$_4$ CpG$_5$

Estimate new HMM model's parameters by Baum-Welch learning algorithm

KL divergence < threshold between old HMM and new HMM?

No

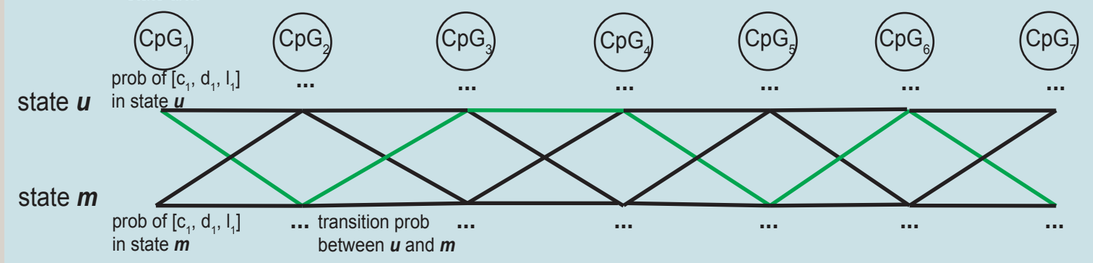
Yes

Final HMM model

4. HMM decoding on fragments from the same cfDNA WGS

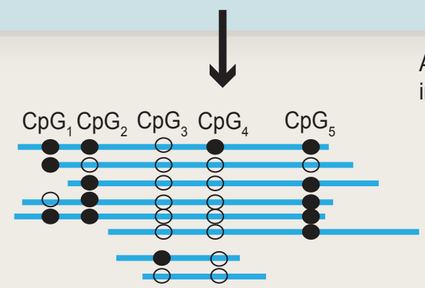
Input one fragment at a time

Fragment$_1$ CpG$_1$ CpG$_2$ CpG$_3$ CpG$_4$ CpG$_5$ CpG$_6$ CpG$_7$



Viterbi algorithm to choose the maximum likelihood sequence (for this fragment$_1$ is: ***umuumum***)

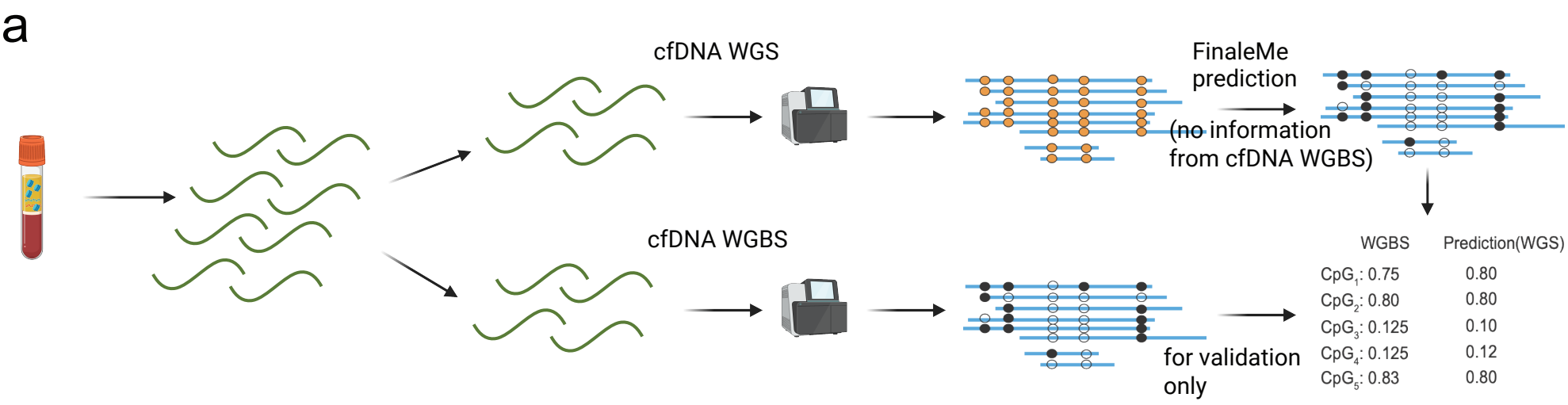
Fragment$_n$ CpG$_1$ CpG$_2$ CpG$_3$ CpG$_4$ CpG$_5$



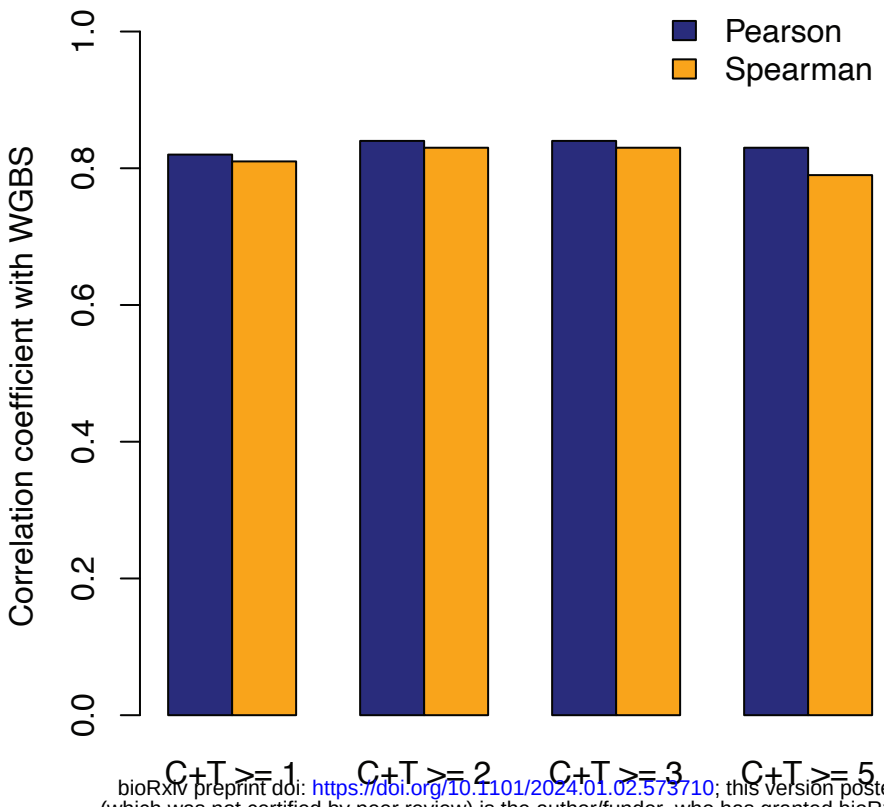
Aggregate fragments at each CpG position in the reference genome for methylation:

CpG$_1$: 0.75
CpG$_2$: 0.80
CpG$_3$: 0.125
CpG$_4$: 0.125
CpG$_5$: 0.83

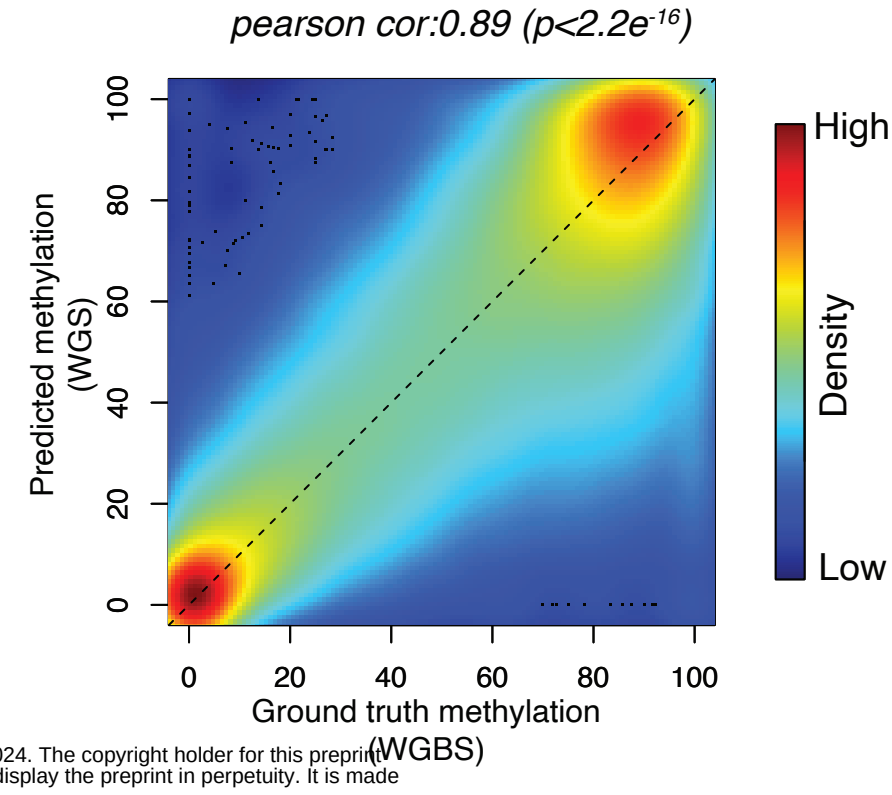
a



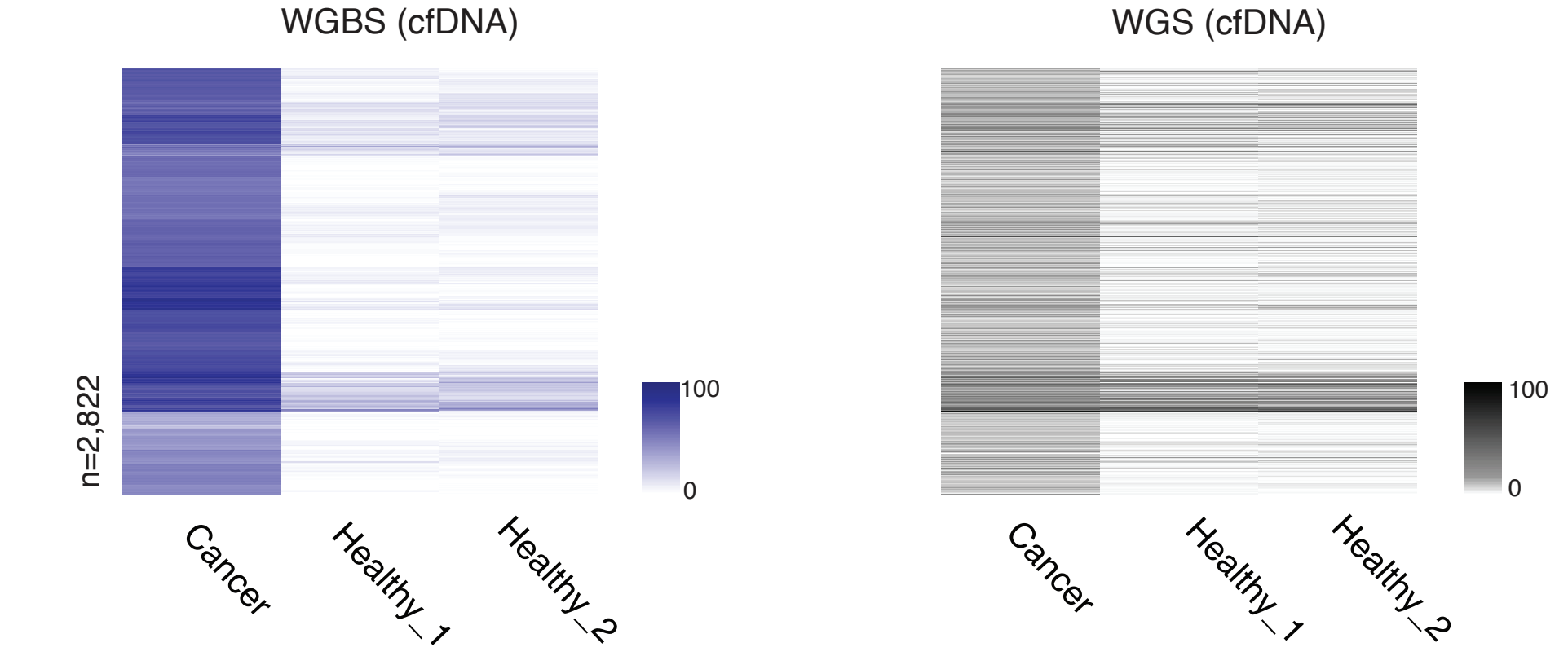
b



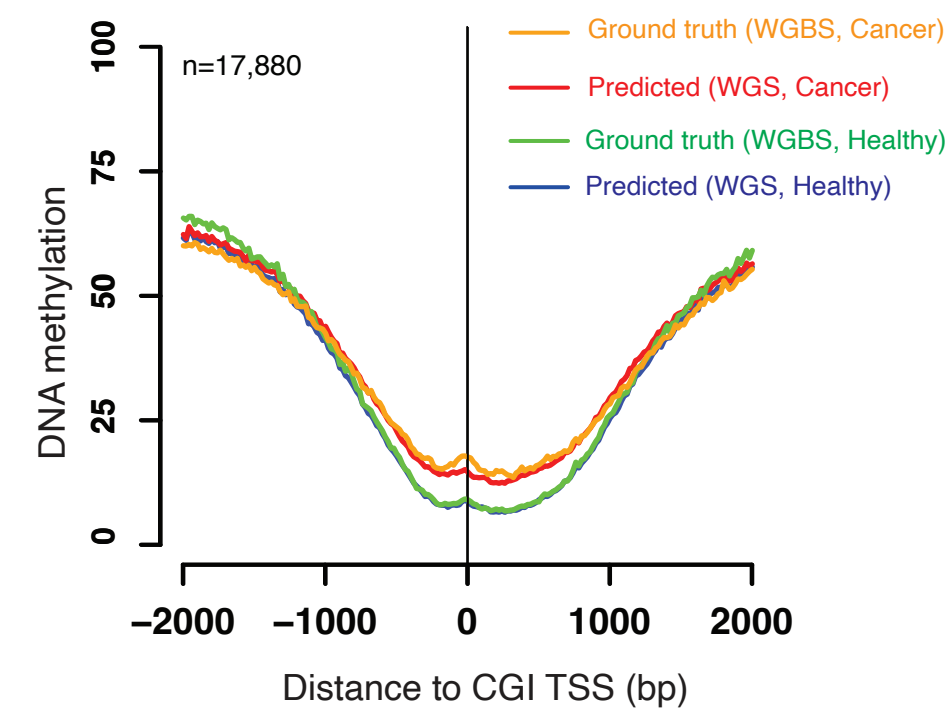
c



d



e



f