# Hidden Markov Models Characterize Consistency Between Methylation Signals

**Judy Du**
Quantitative and Computational Biology
jtdu@princeton.edu

## Abstract

DNA methylation is a key feature of genomes that regulates gene transcription and cellular function. Human Methylation 450 (HM450) arrays have been extensively used to characterize population methylomes by many large consortia such as The Cancer Genome Atlas (TCGA) and ENCODE, characterizing patterns associated with tissue-specific behaviors and identifying abnormalities that lead to disease. These arrays are often preferred over whole-genome bisulfate sequencing in population studies due to the high cost of the necessary reagent, bisulfate, despite missing many CpG base steps of the genome. These missing sites, however, often occur at biologically and clinically relevant sites such as CTCF binding sites. Here, we characterize the correspondence between nearest neighbor signals by modeling HM450 data with a Gaussian Hidden Markov Model, leveraging the assumption that nearby CpG sites are more likely to have correlated signals. We find that our model is robust to varied initializations of latent parameters and can better characterize the latent structure of our testing data than random noise. Given our fitted models, we find that on average, sites which are nearest neighbors to ”methylated” and ”unmethylated” states are 79% and 71% likely to share the same state, respectively. Moveover, these Hidden Markov Models can annotate methylation states without using arbitrary cutoffs. Inference of sites in between observed probes allows researchers to study a larger set of oncogenic marks, leverage existing data to uncover new oncogenic patterns.

## 1 Introduction

Alongside the sequence of DNA a person is born with, epigenomic signals across the genome define the normal function of a cell, regulating which genes are turned on (expressed) and in what proportion. In particular, the methylation of CpG base steps is a chemical signal that sterically hinders the association of regulatory molecules to DNA, ultimately affecting the levels of a gene and thus the behavior of a cell [1]. Understanding the methylation patterns characteristic of specific cell types, organisms, and disease states is of particular interest in genomics, giving insight to the biological mechanisms that underlie cellular function and the abnromalities that result in human disease.

There are two main types of experiments that characterize the methylome of a given sample: whole-genome bisulfate sequencing (WGBS) and methylation arrays. The former investigates methylation at virtually all sites of the genome while the latter only analyzes a specific set of pre-selected sites. While the diminished cost of of whole-genome sequencing has allowed for large genome-wide studies, methylation studies are still costly because of their dependence on the bisulfite reagent. As a result, when conducting methylomic profiling of a large population, there is a tradeoff between the coverage of the methylation experiments and the number of samples that can be studied.

Many large consortia such as TCGA [2] and ENCODE [**?**] have generated a wealth of methylation data using Illumina Human Methylation 450 (HM450) arrays. However, many sites that are biologicaly relevant such as CTCF binding sites are not directly analyzed from these experiments. Early whole-genome studies of epigenomic marks centered around the methylation around gene bodies: for instance, methylation of promoter regions that an inhibit polymerase binding, or methylation of intronic regions that affect splicing patterns [1]. In the last decade, research on noncoding regions have elucidated their roles in genomic regulation, moving away from the ”junk DNA” model of noncoding DNA. In particular, highly self-interacting compartments called topologically associating domains (TADs) that contain genes modulate interactions with regulatory elements, and methylation at CTCF binding sites at the junctions have been shown to disrupt this interaction, leading to glial cancers [4, 5]. The vast majority of these sites are missed by HM450 loci (Figure 11). Imputation of methylation signals using nearby probes from Illumina HM450 arrays would enrich the existing 10,943 methylation studies conducted by TCGA, allowing researchers to leverage the wealth of already existing data to study patterns between mutations, expression, copy number aberrations, and methylation at a richer set of sites.

Here, we aim to investigate the imputability of methylation sites from nearby signals by fitting TCGA HM450 data to a Hidden Markov Model with a Gaussian generative process. This Markov Model will estimate the transferability of methylation states to nearby sites, with the added bonus of annotating methylation data without employing an arbitrary threshold. We verify the robustness of inferred latent parameters to different initializations of starting parameters and the ability to model sequence-specific patterns as opposed to random values. We explore a crude form of imputation at a locus by averaging the states of flanking data. All in all, these studies confer intuition into the extendability of methylation signals to nearby sites.

# 2 Data

## 2.1 Raw Methylation Data

The TCGA consortium has profiled 12,359 methylome sequencing experiments derived from 10,943 individuals across 33 cancer types via the Illumina Human Methylation 450 Array. Here, we first focus our anayses on colon adenocarcinoma (COAD). This is justified by the observation that CTCF binding sites have been shown to be commonly mutated in colorectal cancers [6], suggesting promising yet unexamined driver roles according to positive selection. There are 353 instances of processed TCGA-COAD Illumina Human Methylation 450 (HM450) array data from 296 individuals, with 315 deriving from tumor samples, 38 from matched normals, and 11 patients with technical replicates of tumor samples. Each HM450 dataset can be represented by a $485,577 \times 11$ matrix, where each row corresponds to a 2-nucleotide CpG base step and each column describes features such as probe IDs, genomic coordinates according to the Hg38 genome build, the estimated methylation fraction, the type of genomic element that occurs at the locus (e.g. gene body, CpG Islands, etc.) and other variables describing nearby genes. This data is widely available at https://portal.gdc.cancer.gov/repository using the search parameters Project ID=TCGA-COAD and Platform = Illumina Human Methylation 450.

## 2.2 Properties of the HM450 Data

Each HM450 experiment describes an identical pre-selected set of 2-nucleotide genomic loci (probes) where methylation is characterized. Methylation fractions produced by an HM450 experiment are calculated by summing the signal intensities at a locus and dividing by the total signal intensities at said site. The distribution of the CpG base steps in reference human genome varies, and so the genomic location and the amount of probes per experiment vary from chromosome to chromosome (Figure 10). Within a patient, the distribution of methylation values is typically bimodal and approximately normally distributed, with peaks at large and small methylation fractions (Figures 7,12).

## 2.3 Training and Testing Sets

There is no perfect gold standard with which to test our COAD HM450 methylation fractions (e.g. matched WGBS sequencing experiments). There are, however, replicates of the HM450 experiment, which can be thought of as a realization of an underlying random distribution controlling for locus, patient, tissue type, and cancer state. However, when no other covariates alter the inherent methylation state, one would expect this state to remain consistent between replicates, and a Hidden Markov Model offers a crude estimation of these states behind the random variations. We thus use HM450 data from patients with technical replicates of the HM450 experiment, randomly selecting one experiment from each patient's tumor tissue to be part of the testing set. Then, for each chromosome, we partition our data into the following:

1. training set: $20 \times L$ matrix of methylation fractions for 20 patients at $L$ CpG sites
2. testing set: $11 \times L$ matrix from 11 patients with matched normals in the training set
3. shuffled training set: $20 \times L$ matrix obtained by shuffling the entries of dataset 2 at random seed 7
4. shuffled testing set: $11 \times L$ matrix obtained by shuffling the entries of dataset 2 at random seed 7

The total number of probes ($L$) characteristic of each chromosome is given by Figure 8. The majority of chromatin signals between the training and testing sets have tend to have a small difference, but are not completely identical (Figure 8).

## 3 Methods and Evaluation Metrics

### 3.1 Hidden Markov Model Implementation

Here, we employ a Hidden Markov Model to estimate transitions between methylation states across a chromosome using the colon adenocarcinoma (TCGA-COAD) HM450 training data (type 1), fitting a separate model for each of the 24 chromosome types. For a given chromosome, we ordered the pre-selected chromosome loci by their genomic coordinates and indexed the loci by their cardinal order $i$.

Let $x_i$ denote the observed methylation fraction at locus $i$ across COAD patients and $z_i \in \{0, 1, 2\}$ denote their hidden methylation state. Each $z_i$ is thought to generate the corresponding observed $x_i$ according to a Gaussian distribution. The joint probability of our observed and latent variables is given by

$$p(x_0, \dots x_L, z_0, \dots, z_L) = p(z_0)p(x_0|z_0) \prod_{i=1}^{L} p(z_i|z_{i-1})p(x_i|z_i)$$

where $p(z_i|z_{i-1})$ is the state transition probability that describes the propensity for nearest neighbor probes to either share states or differ, and $p(x_i|z_i)$ is the emmisson probability that describes the probability of the observed value given the inferred state.

To infer the emission probabilities, state, transition matrix, and most likely chromatin states $z_i$ for each patient using the observed methylation data, we used the GaussianHMM function implemented by the HmmLearn libraries [10] in Python with the following parameters: $n\_components$ = 3, $n\_iter$ = 100, $covariance\_type$ = "tied", random seed = 7, and default parameters otherwise. High, intermediate, and low fractions are commonly interpreted as methylated, hemimethylated, and unmethylated, and so choosing 3 components is a natural choice for annotating methylation data. The labels of the inferred states were reassigned such that $mean(x_i|z_i = 0) < mean(x_i|z_i = 1) < mean(x_i|z_i = 2)$.

We note that a core assumption to HMMs is that $p(z_i|z_{i-1}) = p(z_i|z_{i-1}, z_{i-2}, ...z_0)$. In other words, the state of a given probe is only dependent on the upstream probe in the model, even though in reality, $x_i$ is correlated with many upstream or downstream probes a certain distance away. We also note that a core assumption to HMMs is that each step $i \rightarrow i + 1$ is a discrete uniform step, even though there is variation in the distances between loci $i$ and $i + 1$ (Figure 10). Still, imputation of missing sites are often estimated from nearest neighbors irrespective of the variation in distances from the nearest neighbors. Despite these caveats, this model is still informative in examining the similarity in states of nearest neighbor sites, pooling information across all probes within a chromosome from multiple patients.

### 3.2 Evaluation

The fit of the model to the observed data is estimated by the log probability of the testing set given the inferred parameters. The accuracy of the model is estimated by the absolute difference between the most likely decoded states of of the training set and predicted states of the testing set for pairs of loci $i$ from the same patient, cancer tissue, and chromosome. Because we reassigned the states so that a greater state label corresponds to a greater mean methylation fraction, the range of the difference function, [0,1,2], indicates correctly predicted states, similar states, to vastly different predicted states between the training and testing set.

### 3.3 Additional HMM Treatments

#### 3.3.1 Consistency of Model Fit Across Different Starting Positions

We fit an HMM with the above parameters to the chromosome 12 data (type 1) at 10 different sets of starting positions for all latent variables. This allows us to examine the variation in model fit and estimated latent parameters, e.g., whether or not our inferred parameters converged to a consistent optimum.

#### 3.3.2 Model Fit of Randomized Data

We tested the fit of the model to sequences of randomized data (type 4). Unlike its unshuffled counterpart (type 2), randomizing the data destroys any pattern between methylation fractions and states of sequential loci. Observing a poor model fit using the shuffled testing data (in comparison to the fit on data type 2) would indicate that the fitted HMM is in fact modeling patterns between nearest neighbor methylation fractions.

#### 3.3.3 Model Fit to Randomized Data

For every chromosome, we tested the effect of training the Hidden Markov Model with a randomized sequence of methylation fractions $x_i$ (Type 3 data). The transition matrix inferred from this model can act as a baseline probability of state transitions simply based on the enrichment of methylation fractions characteristic of the data (Figure 7).

### 3.4 Imputation of Methylation States

Lastly, we can use our models to predict the methylation state of regions inbetween probes, particularly at sites of interest such as CTCF binding sites. We find CTCF binding sites by applying FIMO [7] to the Hg38 genome using the CTCF PSWM from JASPAR2.0 [8]. At each CTCF binding site, we take the imputed state $z_i*$ to be the mean state between flanking regions $0.5 \cdot (z_{i-1} + z_{i+1})$. This imputation is evaluated by taking its difference from the true annotated state $(z_i * - z_i)$. We note that the Hidden Markov Models should not be trained on the sites we aim to impute, though they are at the point of this submission.

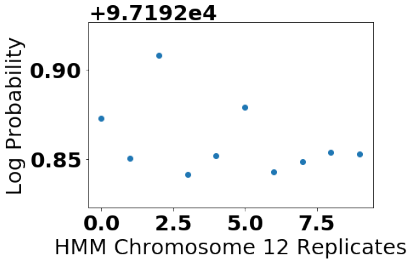
## 4 Results

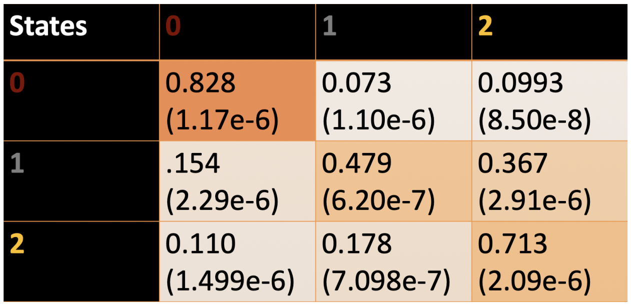
#### 4.0.1 Consistency of Model Fit Across Different Starting Positions

For chromosome 12, the log probability of the training set was very consistent across the 10 different initializations of the latent variables (Figure 1). The variation in the log probabilities are very low in comparison to the magnitude of the total log probabilities. The variation in estimated transition probabilities is also fairly low across replicates, ranging from $8.5 \times 10^{-8}$ to $2.9 \times 10^{-6}$ (Figure 1). Together, this verifies that the model did, in fact, converge to a global minimum when inferring the latent parameters (emission probability, transition matrix, states).

### 4.1 Model Fit of Randomized Data

Upon fitting the Hidden Markov Model to unshuffled methylation data (type 1), the log probabilities of our unshuffled testing set (type 2) are much larger than the log probabilities applied to the shuffled





| States | 0 | 1 | 2 |
|---|---|---|---|
| 0 | 0.828<br>(1.17e-6) | 0.073<br>(1.10e-6) | 0.0993<br>(8.50e-8) |
| 1 | .154<br>(2.29e-6) | 0.479<br>(6.20e-7) | 0.367<br>(2.91e-6) |
| 2 | 0.110<br>(1.499e-6) | 0.178<br>(7.098e-7) | 0.713<br>(2.09e-6) |

Figure 1: (Left) Log Probability of testing set (Type 2) across 10 HMM models fit on the Chromosome 12 training set (Type 1) at 10 different initializations of latent variables. (Right) Mean and standard deviation (in parentheses) of each transition probability across the 10 chromosome 12 replicates

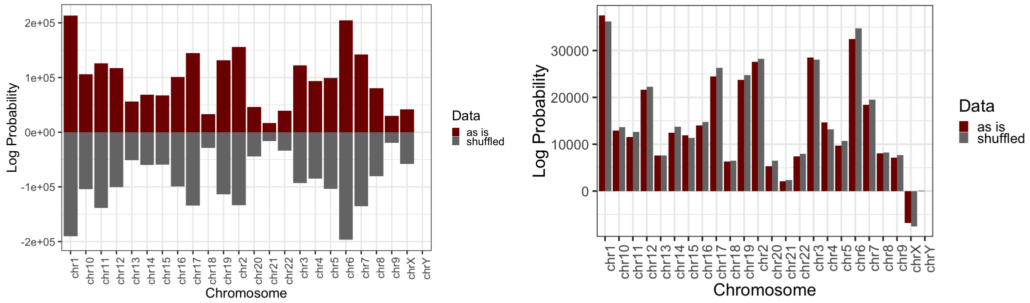


Figure 2: Log probabilities of our models across all 23 chromosome types when fitted on (left) unshuffled training data (Type 1) and (right) on shuffled training data (Type 2). Log probabilities of the testing sets 2 and 4 are shown in red and grey, respectively.

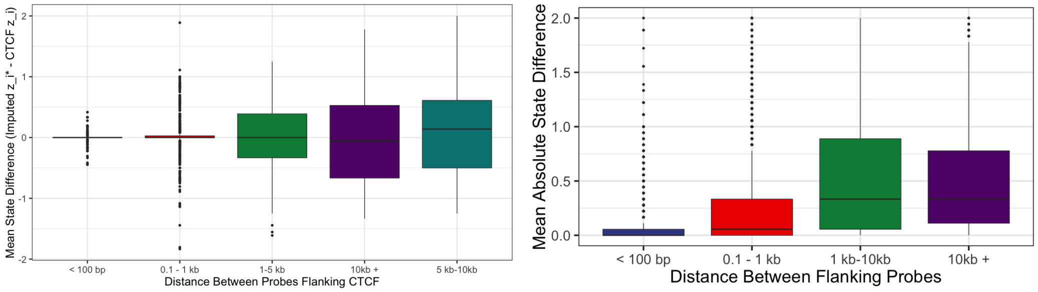


Figure 3: Left: Accuracy of HMM to impute CTCF binding sites across all Chromosomes. For all HM450 locus $i$ that fall within a CTCF binding site, the x-axis corresponds to the distance between flanking probes $(i-1, i+1)$. The y-axis corresponds to the mean state predicted by flanking probes minus the true state at said probe $(mean(z_{i-1}, z_{i+1}) - z_i)$

. Right: Absolute difference in MAP state annotations as a function of distance in chromosome between flanking loci $i$ and $i+2$

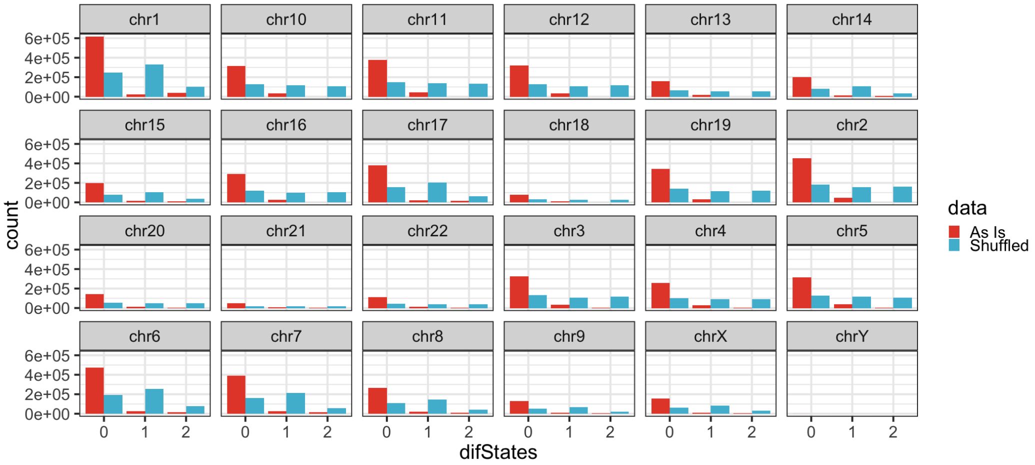


Figure 4: Accuracy of Prediction. The x-axis shows the absolute difference between decoded states of the training set and the predicted states of the testing set for all pairs of matched patient experiments. The color corresponds to the type of testing data used for prediction, where ”As Is” indicates the unshuffled (type 2) testing data, and ”Shuffled” indicates the randomized testing set (type 4)

testing set (type 4) (Figure 2). In other words, upon interrupting the relationship between nearest neighbor sites along a linear sequence, the transition matrix and emission probabilities inferred by the HMM no longer described the observed data as accurately. This trend is consistent over all chromosomes. The magnitude of the log probabilities observed in Figure 2 simply mirrors the number of sites in the testing set (Figure 8).

Figure 4 shows the absolute difference between the testing and training states across all 23 HMM models fit to the 23 chromosomes. When predicting on type 2 data (red), a large majority of probes have an absolute difference of 0 (x-axis), a few have an absolute difference of 1. Very rarely is a state mistaken for its extreme (i.e. rarely does have an absolute difference of 2). This trend is consistent across all chromosomes. When predicting on type 4 data, a different pattern is observed. The absolute difference between states is more or less uniform across the support of the function.

### 4.2 Model Fit to Randomized Data

For every chromosome, we tested the effect of training the Hidden Markov Model with a randomized sequence of methylation fractions $x_i$ (Type 3 data). According to Figure 2 (right), the randomized data has a markedly increased model fit. This is not surprising given that the sequential patterns of methylation fractions between the training and testing sets are now very similar on average. The log probabilities of the unshuffled Type 2 data are slightly lower than the their randomized counterparts across chromosomes, and are an order of magnitude less than the model trained on unshuffled data. In short, we have quantified that the model trained on unrandomized Type 1 data better fits the unrandomized testing data.

The latent transition probabilities inferred from the randomized training set (Type 3) (Figure 9) represent the underlying likelihood of the states given the distribution of methylation fractions (Figure 7) when sequential patterns in the data are ignored. In Figure 9, $p(z_i = 0a|z_{i-1} = s) \approx 0.4$, $p(z_i = 1a|z_{i-1} = s) \approx 0.2$, and $p(z_i = 2a|z_{i-1} = s) \approx 0.4$ irrespective of the value of the starting state $s$. These probailities more or less mirror the fraction of loci annotated as each of the states: 44.26% state 0, 16.67% state 1, and 39.05% state 2 with slight bias to the state with the largest peak ( Figure 7).

### 4.3 Imputation of Methylation States

For all HM450 loci that occur in CTCF binding sites, we predicted the methylation state using the MAP state of the flanking loci. In Figure 3, the y-axis describes the mean difference between

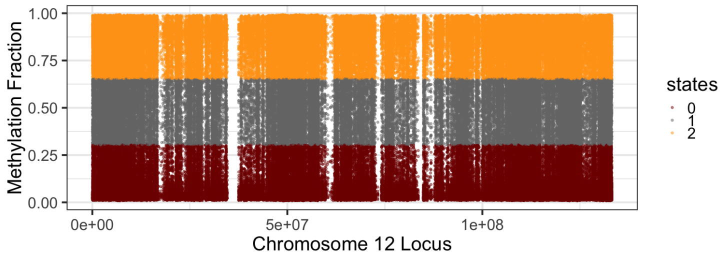


Figure 5: Chromosome 12 Model Representation. Methylation fraction $x_i$ as a function of genomic locus for ordered sequences of probes $i$. The color represents $z_i$, the most likely state inferred by the model.

imputed and true states across our testing set. Across distances, this value tends to be less than +/- 0.75, signaling an accurate imputation, though our models should be re-trained to omit these probes. This accuracy drops off with increasing distance, with much greater inaccuracy by 5 kb.

## 5 Analysis

We have found that the Hidden Markov Models do in fact converge (Figure 1), and that there is, in fact, a linear structure to the sequences of nearest neighbor methylation fractions (Figure 2).

Randomizing the testing sequences destroys any linear relationship between adjacent methylation sites. Thus, the inferred transition probabilities $p(z_i|z_{i-1})$ no longer appropriately describe the sequence of methylation fractions $x_i$ across patients in the testing set. As a result, randomized methylation sequences have vastly smaller log probabilities in comparison to the unshuffled methylation sequences. This verifies that the transition probabilities inferred by our model do actually represent a nearest neighbor relationships between methylation states.

### 5.1 Model Exploration

Figure 5 represents the state annotations inferred at each locus for a given patient. Note that there is not a clear linear separation between inferred states $z_i$ according to methylation fraction; instead, there is much overlap in methylation fraction between states (colors) at the interfaces around 0.66 and 0.25. This is because MAP states are in part inferred by the neighboring methylation patterns surrounding each locus.

We also note that there are large white areas where no probes occur, the largest of which occurring at the centromere around the locus chr12:50,000,000. This highlights the fact that loci pairs (i, i+1) sometimes occur across vastly large chromosome regions. To quantify this, we calculate the distance between genomic coordinates of all such pairs in the HM450 array (Figure 10. The majority of pairs are relatively proximal to each other, with only 12% occuring 10 kb away, and 35% occuring 1 kb away. For reference, 30kb is a distance threshold previously used as a threshold for correlation of methylation signals [11].

Figure 6 shows the mean transition probailities $p(z_i|z_{i-1})$ across all chromosomes as well as their standard deviation. Overall, the variation between estimated probabilities across chromosomes tends to be relatively low. We find that the consistency of sequential probes, i.e. their propensity to share the same state, tends to be around .79 for methylated probes and 0.71 for unmethylated probes, with the former exhibiting higher variaton across chromosomes. At a mean probability of 0.513, hemi-methylated loci are much less likely to stay in their same state. Moreover, $p(z_i = 0|z_{i-1} = 1) > p(z_i = 0|z_{i-1} = 1)$. All in all the transition matrix shows that the intermediate state 1 acts as a transitory state between states 0 and 2. The exception to this rule is that $p(z_i = 2|z_{i-1} = 0) > p(z_i = 2|z_{i-1} = 1)$, which may be the case simply because low methylation fractions are the most common values in our training set (Figure 7).

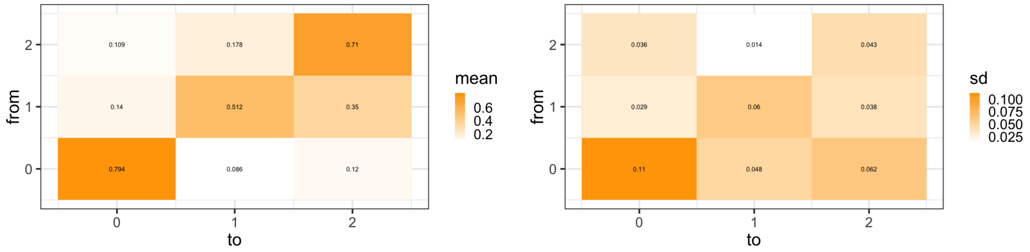


Figure 6: Inferred transition matrix from states $z_i$ (rows) to states $z_{i+1}$. Mean (left) and Standard Deviation (right) of transition probabilities across HMMs trained on all chromosomes.

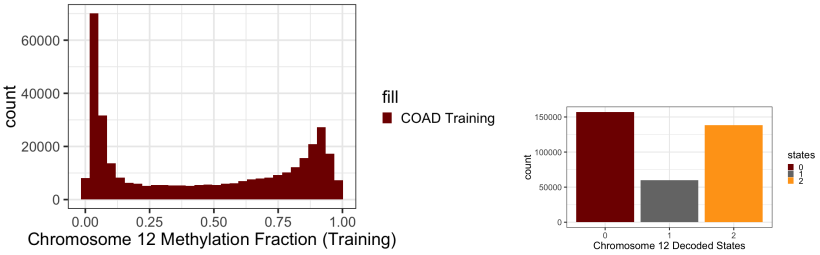


Figure 7: Distributions of methylation fractions $x_i$ (left) and MAP states $z_i$ (right) of the training set (Type 1) for the Hidden Markov Model trained on Chromosome 12

## 6 Conclusion and Possible Extensions

In this work, we have modeled Human Methylation 450 data from TCGA colon adenocarcinoma patients using a Hidden Markov Model, finding annotations for methylated, unmethylated, and hemimethylated states without employing an arbitrary cutoff, and estimating the propensity for a nearby probe to have a similar methylation state to its neighbor. We have found that our model converges to find true sequential trends in methylation fractions across patients, and we have explored the accuracy of imputation at CTCF binding sites represented by HM450 probes, defining a reasonable distance threshold for which imputation of missing sites can be analyzed.

This work can be scaled to describe a larger set of COAD patients to verify that the estimation of transition probabilities remain consistent. This work can also be scaled to the 32 TCGA cancer types, all of which have HM450 experiments widely available. Of particular interest is lower grade glioma (TCGA-LGG), for which a subset of patients are IDH mutants that display hypermethylation across their CpG sites(Figure 12, which may affect transition probabilities from and to state 2.

We have assumed that all transition probabilities between 2 nearest-neighbor probes are equal despite the variation in distances between these pairs (Figure 10). To analyze the accuracy of the model as a function of distance, we can extend this model by training our Hidden Markov Models using only sequences of probes whose nearest-neighbor distances are smaller than a range of thresholds. According to Figure **??**, distances between 1kb and 10kb would be a reasonable thresholds to test HMM model fits.

In addition, HM450 probes are constructed with two types of chemical probes, each with their own bias towards higher or lower methylation. Furthermore, each probes falls within a specific type of genomic element. A Hidden Markov Model can also be fit with these two additional features as covariates to analyze the effect of these chemical probes on the estimation of the MAP state.

Further studies of the accuracy of imputation should be conducted before relating patterns between imputed methyation sites and other omics data. The distance threshold with which to impute unknown values can be further fine-tuned, and the accuracy of imputations can be quanified using matched WGBS experiments. All in all, we have shown the utility in using Hidden Markov Models to characterize HM450 data despite distance caveats (and learned a ton about HMMS in the process:)

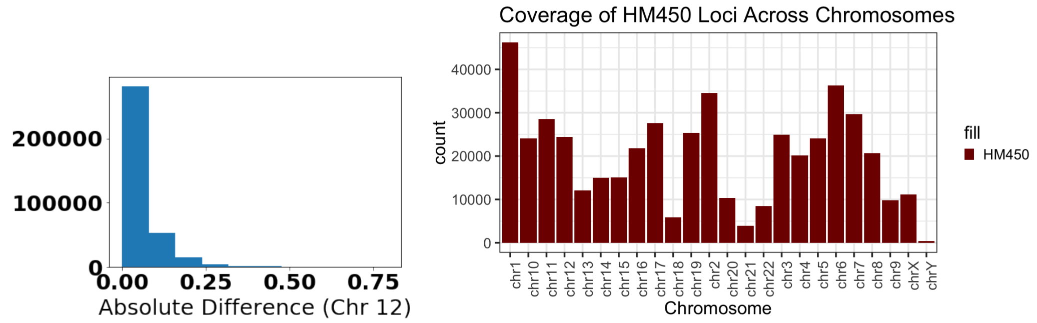


Figure 8: Left: Absolute difference between methylation fractions between training and testing sets for pairs sharing the same tissue sample and genomic locus. Right: Per-chromosome number of CpG sites whose methylation fraction are estimated by the HM450 array. These values represent $L$, the per-patient size of the observed data used to train each Hidden Markov Model.

## 7 Supplemental Figures

## References

[1] Singer, M., Kosti, I., Pachter, L., Mandel-Gutfreund, Y. (2015). A diverse epigenetic landscape at human exons with implication for expression. Nucleic Acids Research, 43(7), 34983508. https://doi.org/10.1093/nar/gkv153

[2] Wang Z., Jensen M.A., Zenklusen J.C. (2016) A Practical Guide to The Cancer Genome Atlas (TCGA). In: Math E., Davis S. (eds) Statistical Genomics. Methods in Molecular Biology, vol 1418. Humana Press, New York, NY

[3] Feingold et al. The ENCODE (ENCyclopedia of DNA Elements) Project. Science, 306(5696):636640,2004.

[4] Rao et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell,159(7):16651680, 2014.

[5] Flavahan et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nat Publ. Gr. 529(7584):110114, 2016.

[6] Katainen R, Dave K, Pitkanen E, Palin K, Kivioja T, Valimaki N, Gylfe AE, Ristolainen H, Hanninen UA, Cajuso T et al.: CTCF cohesin-binding sites are frequently mutated in cancer. Nat Genet 2015, 47:818-821.

[7] Grant, C. E., Bailey, T. L., Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. Bioinformatics, 27(7), 10171018. https://doi.org/10.1093/bioinformatics/btr064

[8] Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Lee, R. Van Der, Bessy, A., Lenhard, B. (2018). JASPAR 2018 : update of the open-access database of transcription factor binding profiles and its web framework, 46(November 2017), 260266. https://doi.org/10.1093/nar/gkx1126

[9] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/technology/Illumina-HumanMethylation450-Data-Sheet

[10] HMMlearn: https://hmmlearn.readthedocs.io/en/latest/,https://github.com/hmmlearn/hmmlearn

[11] Fan, S., Tang, J., Li, N., Zhao, Y., Ai, R., Zhang, K., Wang, W. (2019). Integrative analysis with expanded DNA methylation data reveals common key regulators and pathways in cancers. Npj Genomic Medicine, 4(1). https://doi.org/10.1038/s41525-019-0077-8

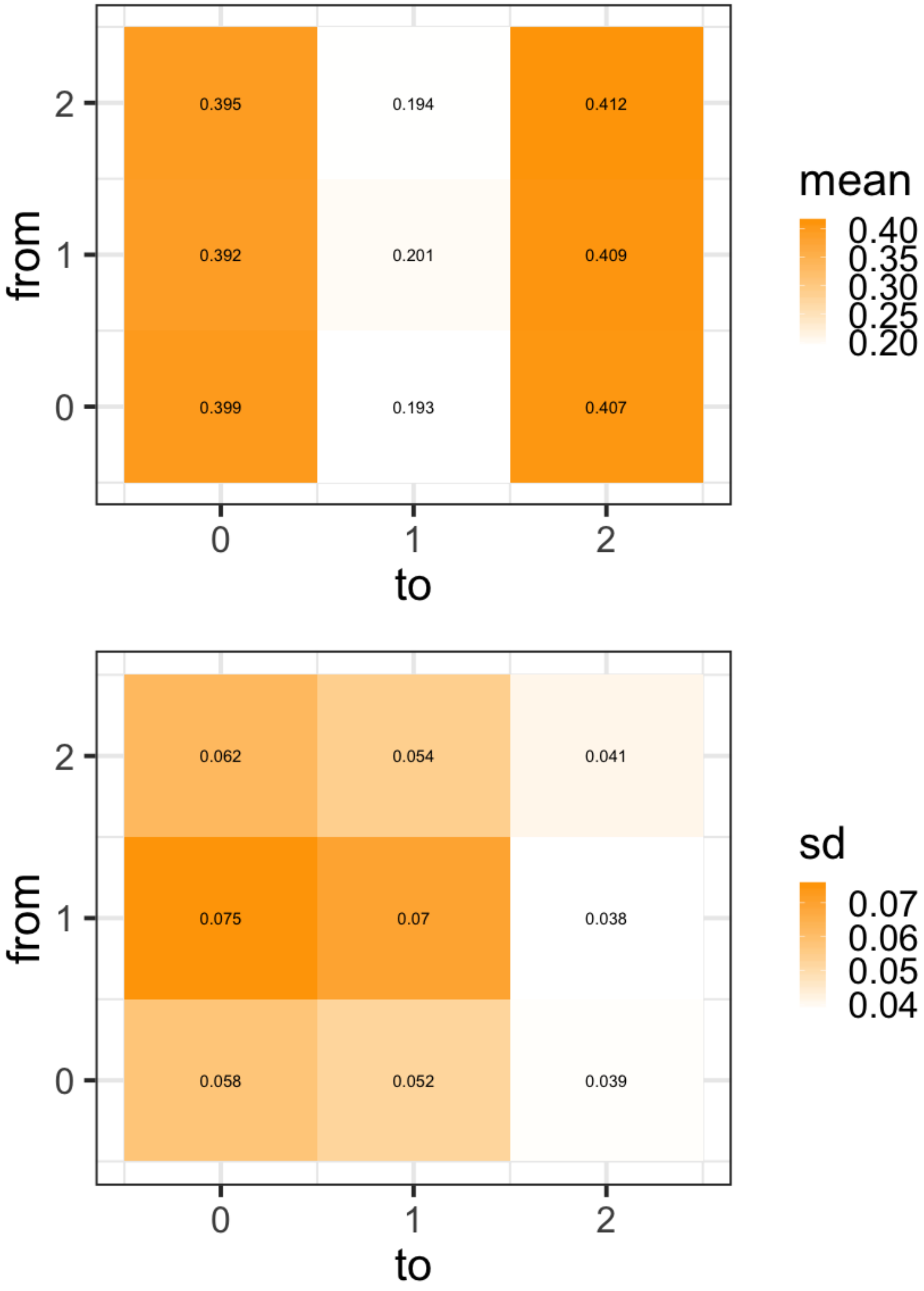


Figure 9: Inferred transition matrix from states $z_i$ (rows) to states $z_{i+1}$ from Hidden Markov Model trained on randomized training data (Type 3). Mean (left) and standard deviation (right) of transition probabilities across HMMs trained on all chromosomes

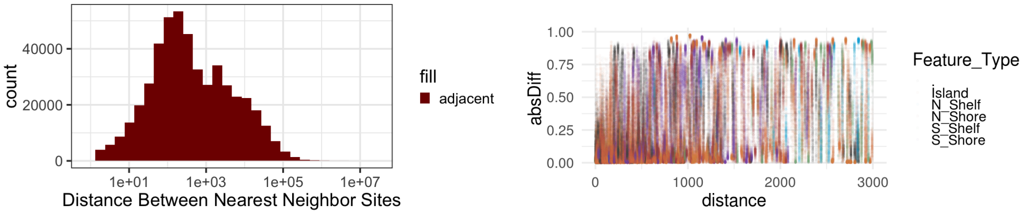


Figure 10: (Left) Distribution of genomic distances between each pair nearest neighbors sites (i, i+1). (Right) Correspondence between signals of neighboring CpG base setps assayed by the Illumina Human Methylation 450 array. The y-axis decribes the difference between methylation fractions for each of 313 COAD patients across all pairs of CpG sites within 3 kb of each other. The x-axis describes the genomic distance between sites, and the color is the genomic element where each site falls.

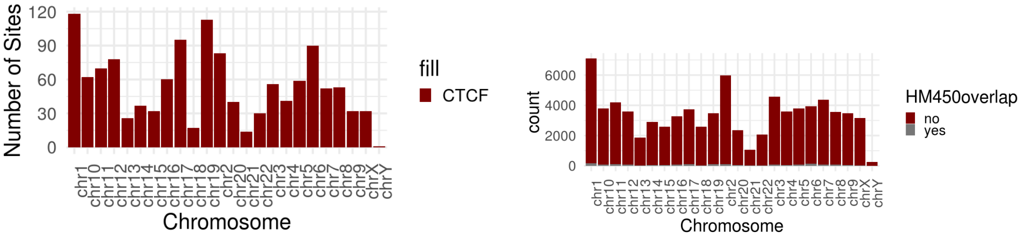


Figure 11: (Left) Distribution of Probes that overlap CTCF binding sites (Right) Overall number of CTCF binding sites across the genome. These sites were calculated using FIMO to find likely CTCF Binding sites according to the PSWM on Jaspar2.0. All probable hits that overlap any ENCODE CTCF ChIP peak are taken to be true CTCF binding sites.

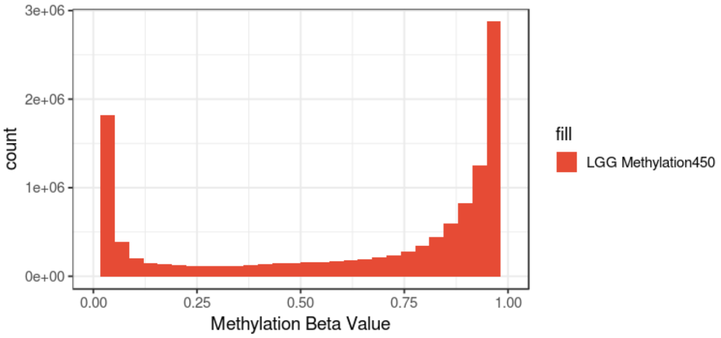


Figure 12: Distribution of HM450 Methylation Fractions for TCGA-LGG (lower-grade glioma).