Epigenomic tumor evolution modeling with single-cell methylation data profiling

Xuan Cindy Li[1,3,†], Yuelin Liu[1,4,5,†], Farid Rashidi Mehrabadi[1,7,†], Salem Malikić[1], Stephen M. Mount[6], Eytan Ruppin[1], Kenneth Aldape[2], and S. Cenk Sahinalp[1,*]

[1]Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20892, USA
[2]Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, MD 20892, USA
[3]Program in Computational Biology, Bioinformatics, and Genomics, University of Maryland, College Park, MD 20742, USA
[4]Department of Computer Science, University of Maryland, College Park, MD 20742, USA
[5]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA
[6]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA
[7]Department of Computer Science, Indiana University, Bloomington, IN 47408, USA

[†]Joint first authors
[*]Corresponding author

Abstract

Recent studies on the heritability of methylation patterns in tumor cells, suggest that tumor heterogeneity and progression can be studied through methylation changes. To elucidate methylation-based evolution trajectories in tumors, we introduce a novel computational framework for methylation phylogeny reconstruction, leveraging single cell bisulfite treated whole genome sequencing data (scBS-seq), additionally incorporating copy number information inferred independently from matched single cell RNA sequencing (scRNA-seq) data, when available. Our framework consists of three components: (i) noise-minimizing site selection, (ii) likelihood-based sequencing error correction, and (iii) pairwise expected distance calculation for cells, all designed to mitigate the effect of noise and uncertainty due to data sparsity commonly observed in scBS-seq data. We validate our approach with the scBS-seq data of multi-regionally sampled colorectal cancer cells, and demonstrate that the cell lineages constructed by our method strongly correlate with original sampling regions. Additionally, we show that the constructed phylogeny can be used to impute missing entries, which, in turn, may help reduce sparsity issues in scBS-seq data sets.

Contact: cenk.sahinalp@nih.gov

1 Introduction

The impact of CpG methylation in cancer has been of interest since the emergence of high throughput sequencing. For example, by using the now defunct (Roche) 454 single-molecule sequencing platform, Sottoriva et al. [22] inferred methylation profiles of microdissected colorectal cancer glands and build their "tumor tree" demonstrating a hierarchy of mitotic clones. More recent studies have demonstrated that methylation patterns are highly heritable among tumor cells, e.g. in brain cancers [4], suggesting that tumor heterogeneity and progression can be interpreted and possibly predicted in the context of methylation changes. This premise, along with advances in single-cell sequencing technology, has prompted a recent effort to cluster bisulfite-treated single cells based on their observed methylation patterns, e.g. through Bayesian (non-hierarchical) methods [11, 12]. More interestingly, some recent studies have started to explore how to cluster single-cell methylation patterns in a hierarchical manner towards obtaining a "methylation phylogeny". For example [6] examined heritable epimutations in chronic lymphocytic leukemia (CLL) tumor evolution by reconstructing cell lineages with methylation information binarized from single-cell reduce-representation bisulfite sequencing data (scRRBS-seq), and observed early branching and longer branch length in the CLL phylogeny compared to that of normal B-cells. Interestingly, the study also concluded that, despite the CLL genome being near-diploid, the coverage provided by available data was mostly monoallelic [6].

Constructing phylogeny based on low-depth monoallelic coverage data presents a major challenge: how can we determine the methylation status at a CpG-site given little read information which, most of the time, only comes from one out of potentially multiple alleles? One relatively simple way to address this issue by [6] is to assign each CpG-site the methylation status with which at least 90% of its reads agree. This approach has two major drawbacks. First, a majority of the sites is only each covered by a single read, which offers less than sufficient information required to accurately determine the methylation status of a site in a non-haploid cell. In addition, by binarizing methylation status, this approach entirely excludes potential allele-specific methylations in the genome, which is prevalent in both diploid mammalian normal genome [20] and human cancer genome [5]. Importantly, [6] used IQ-TREE, a maximum-likelihood-based tool, to construct their methylation phylogenies [17]. IQ-TREE is an alternative to those approaches with an explicit infinite-sites assumption [8], appropriate for this particular application, however it is computationally expensive, highly reliant on the underlying model, and inflexible to incorporating information regarding uncertainties or errors in the data. We are also familiar with one other approach to infer the methylation status of a given site, DeepCpG [1], which works through deep-learning based implicit clustering. Unfortunately, DeepCpG is developed for normal (non-tumor) samples and as such is not designed to capture highly variable methylation patterns across cells from the same tumor tissue.

Our Contributions. In this paper, we introduce a novel computational framework for methylation phylogeny reconstruction leveraging single cell bisulfite treated whole genome sequencing data (scBS-seq) [21], which offers the ability to incorporate additional copy number information inferred independently from matched single cell RNA sequencing (scRNA-seq) data, when available.

Our computational approach consists of three components: (i) noise-minimizing site selection, (ii) likelihood-based sequencing error correction, and (iii) pairwise expected distance calculation for cells, all designed to mitigate the effect of noise and uncertainty due to data sparsity commonly observed in scBS-seq data.

Component (i), discussed in Section 3.1, features an integer linear program (ILP) based biclustering formulation to select a set of CpG-sites and cells so that the number of CpG-sites with non-zero coverage in the selected cells is maximized. This procedure filters out cells with read information in too few sites and CpG-sites with read information in too few cells.

Component (ii), discussed in Section 3.2, addresses the sequencing errors commonly encountered in currently available platforms with a maximum log likelihood approach to correct likely sequencing errors in scBS-seq reads, incorporating CpG-site copy number information in case

it can be orthogonally obtained. Given the copy number and read information for a site in a cell, together with the overall sequencing error probability, we compute the log likelihood for all possible underlying allele status. If the mixed read status at the CpG-site for the cell are more likely due to sequencing error on homozygous alleles as opposed to the presence of alleles mixed methylation status, we correct the reads of the minority methylation status to the majority one.

Component (iii), discussed in Section 3.3, introduces a formulation for estimating distances between any pair of cells. Our method incorporates copy number information when available (Section 3.3.2). For each CpG-site in a cell, we compute a probability distribution across all possible *methylation zygosities*. Then, given specific distance values between pairs of distinct zygosities and the likelihood of each possible zygosity for each shared CpG-site between any given pair of cells, we compute the expected total distance between the cells as some function (e.g. normalized L_k, etc.) of expected distances across all shared CpG-sites. We leverage such pairwise distances in methylation phylogeny reconstruction.

As a motivation, we start our discussion in Section 2 on bulk methylation array data from central nervous system tumors [4]. We show that a distance-based hierarchical clustering of patient methylation profiles exhibit strong concordance with cancer (sub)types, suggesting that methylation phylogenies may help identify epigenetic evolutionary trajectories for tumors - akin to those inferred for mutation trees [7]. Such trajectories may help develop predictive models for epigenomic tumor evolution.

Later in Section 3.4, we extend our analysis to scBS-seq data of multiregionally sampled colorectal cancer cells from a recent study [3], which also attempted to analyze the evolutionary landscape of the tumors analyzed from a methylation perspective - only with limited success. We demonstrate that the cell lineages predicted by our method strongly correlates with the tissue of origin of single-cells. For comparison, we also show the tumor single cell phylogeny constructed using mutations profiled via matched scRNA-seq data from Bian et al. [3]. The absence of correlations between tissue of origin of cells and the mutation tumor phylogeny further motivates a methylation-based approach, such as the one we propose in this paper.

Finally, we show in Section 4 that the methylation phylogenies we built for this data set can be used to impute the methylation status of unknown (held-out) sites, which may help address sparsity issues in this and possibly other single cell sequencing data sets.

2 Hierarchical Clustering of Methylation Patterns across Cancer Samples

There is ample evidence that distinct tumor types have distinct methylation profiles, and the stage of the tumor, as well as its metastatic state, could be reliably inferred from its methylation profile [4]. Motivated by this observation we first employ a "naive" neighbor-joining (NJ) hierarchical clustering strategy [18] to build a "phylo-epigenetic" tree of samples of central nervous system (CNS) tumors. For that, we used the genome-wide DNA methylation profiles of tumor samples from [4], obtained through the Infinium HumanMethylation450K BeadChip arrays representing almost all WHO-defined neuroectodermal and sellar region tumors. Overall, the data set was comprised of 2,324 samples from 9 distinct tumor classes, reporting on 31,322 probes with the highest methylation variability (s.d. > 0.228). As shown in Figure 1[1], the cosine distance based NJ tree we built for these CNS tumors can be interpreted as a phylo-epigenetic tree, since not only it clusters distinct tumor types well, but also demonstrates how certain methylation profiles of a particular tumor type may evolve into another distinct tumor type: in particular, observe that Glioma IDH samples (in yellow) and Glioblastoma samples (dark green) appear to be "evolutionarily related". Interestingly Glioma IDH patients have an average survival time twice as long as that of Glioblastoma patients [10].

The above observation may not come as a surprise since it was already known that CNS tumors can be classified using supervised techniques such as random forest, or clustered using

[1] All visualizations of trees in this work were created with `ggtree` [25, 27, 26]

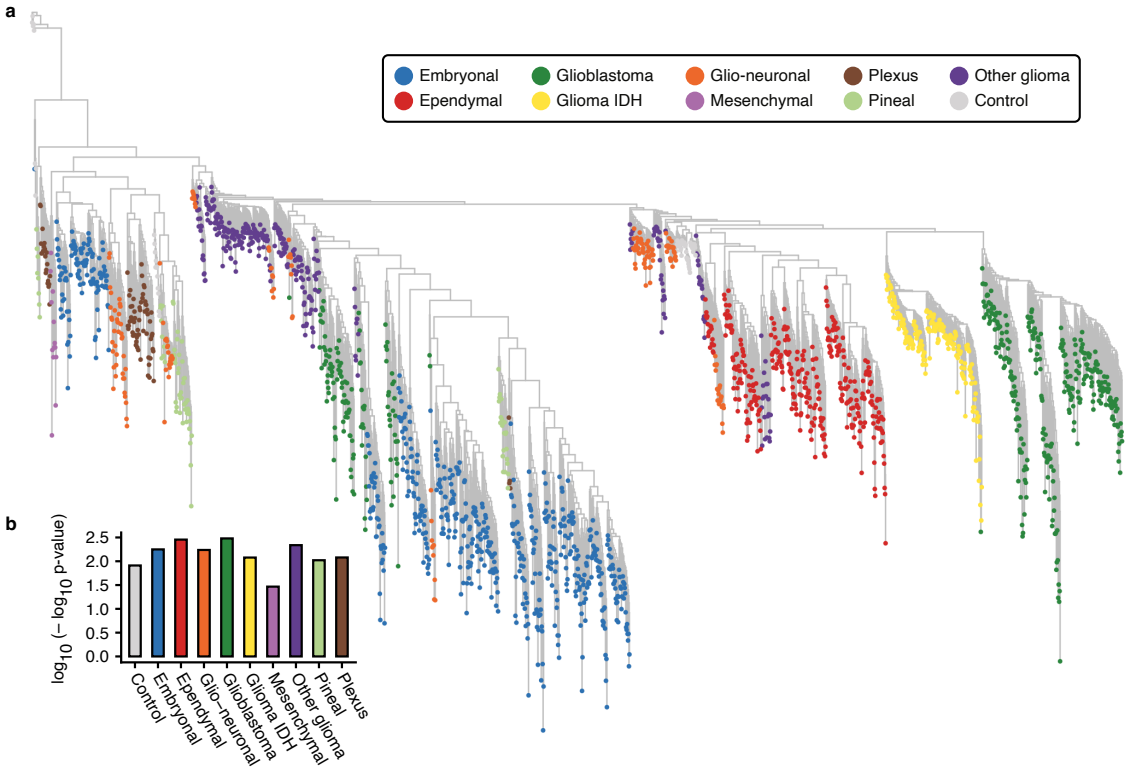

Figure 1: **(a)** NJ clustering of methylation array data from tumor samples with distinct types of central nervous system (CNS) cancers. **(b)** p-values for each cancer type were calculated using one-tailed binomial test. Each sample in the tree is labeled as "success" if all samples in the subtree of its direct parent have the same cancer type. Otherwise it is labled as "failure". Note that the null hypothesis used here assumes the probability of success (in a Bernoulli experiment) to be the proportion of samples with the associated cancer type among all samples.

unsupervised techniques including t-distributed stochastic neighbor embedding (t-SNE) [24]. In fact, as mentioned earlier, methylation profiles of evolving cells are known to exhibit a certain level of heritability [23]. Unfortunately, a recent detailed analysis of scBS-seq data from single cells harvested from colorectal cancer samples suggested that methylation heritability occurs only at a limited degree [14, 3]. As we will show, this conclusion was mostly due to a methodological limitations. As the scBS-seq data is very sparse (the average read coverage of each single cell was less than 1) the study analyzed very coarse methylation "vectors" derived from single cells, based on the premise that the methylation status of proximal CpG-sites are correlated. In such a vector each dimension represents the average methylation level of a fixed length, non-overlapping window. The cells were then hierarchically clustered based on the "distance" between their methylation vectors, as per Figure 1. The resulting tree [3] does not strongly corroborate with where the tumor cells were harvested, suggesting a high level of noise in epigenetic phylogeny reconstruction.

In the remainder of the paper, we will demonstrate that it is in fact possible to infer the methylation phylogeny of a tumor sample through scBS-seq data and that the methylation status of individual CpG-sites are strongly heritable. The window based approach of [3] likely suffers from the observation that methylation status concordance between proximal CpG-sites are much weaker in cancer samples [13]. We achieve this through our computational approach described in Section 3 and show our experimental results on colorectal cancer [3] in Section 4.3.

3 Single Cell Methylation Phylogeny Reconstruction

In this section, we lay out the three key components to our computational approach for *distance-based* (NJ or possibly UPGMA-based) methylation phylogeny reconstruction from scBS-seq data: (i) noise-minimizing site selection (Section 3.1), (ii) likelihood-based sequencing error correction (Section 3.2), and (iii) pairwise expected distance calculation (Section 3.3).

3.1 A Biclustering Formulation for Noise Minimization in Single Cell Methylation Phylogenies

Typical scBS-seq data has a very low (average) read coverage per CpG-site on any given cell: the set of CpG-sites with non-zero read coverage - for which we have methylation status information - is limited and varies substantially across cells. As a consequence, the number of "shared" sites, i.e. those sites that have non-zero coverage in two cells to be compared, is also limited and varies substantially across cell pairs. Since our distance-based methylation phylogeny construction approach requires a fairly accurate estimate of the distance between all cell pairs, we need to filter out those cells for which the number of CpG-sites with non-zero read coverage are low. Similarly, CpG-sites with non-zero coverage in only a few cells need to be filtered out so as to minimize the overall noise. In fact, it is highly desirable to coordinate the reduction in the number of cells and number of CpG sites so as to minimize information loss. For that, we present an ILP based biclustering formulation below.

Given matrix $M_{n\times m}$, where n is the number of cells and m is the number of CpG-sites, and

$$M_{ij} = \begin{cases} 1 & \text{if site } j \text{ is covered in cell } i \\ -1 & \text{otherwise} \end{cases}$$

and the fraction of cells α and fraction of CpG-sites β we would like to keep, we wish to compute a biclustering of M (i.e. a selection of rows of columns of M) so that we maximize the number of CpG-sites covered in the $\lfloor \alpha n \rfloor \times \lfloor \beta m \rfloor$ submatrix. In jointly doing cell and CpG-site selection, we hope to at the same time 1) remove cells that do not have read coverage for a lot of sites, which thus will share little information with other cells and produce noisy distance estimation, and 2) remove sites that are not covered in many cells for dimensionality reduction.

Let $C \in \{0,1\}^n, S \in \{0,1\}^m$ be binary vectors indicating whether a cell or site is kept, and $A \in \{0,1\}^{n\times m}$ be a binary matrix corresponding to whether any site in any cell is kept, we formulate the ILP as follows:

$$\begin{aligned} \text{maximize:} \quad & \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}M_{ij} \\ \text{subject to:} \quad & a_{ij} \in \{0,1\} \\ & c_i \in \{0,1\} \\ & s_j \in \{0,1\} \\ & a_{ij} \le c_i \\ & a_{ij} \le s_j \\ & c_i + s_j - 1 \le a_{ij} \\ & \textstyle\sum_{i=1}^{n} c_i = \lfloor \alpha n \rfloor \\ & \textstyle\sum_{j=1}^{m} s_j = \lfloor \beta m \rfloor \end{aligned}$$

The resulting submatrix will then be $\tilde{M}$, which takes from the original matrix cells $\{i|c_i = 1\}$ and sites $\{j|s_j = 1\}$. These cells and CpG-sites represented in $\tilde{M}$ form the basis for the procedures and analyses that follow.

3.2 Sequencing Error Correction Accounting for Copy Number Variation (CNV)

Any currently available sequencing technology for methylated site identification comes with a nontrivial sequencing error rate, and any sequencing error will have a pronounced effect in

downstream distance computation given the already shallow read coverage in the data. In this section, we present a maximum log likelihood approach to correct likely sequencing errors in scBS-seq reads, incorporating CpG-site copy number information orthogonally obtained.

As mentioned earlier, the single-cell triple omics sequencing (scTrio-seq) protocol presented in [3] offers not only genomic scBS-seq data but also transcriptomic (scRNA-seq) read data from single cells. Even though the original goal of the protocol was to correlate methylation patterns with gene expression at single cell resolution, here we offer a novel use of the matching scRNA-seq reads for CNV calling for the purposes of sequencing error correction, as a pre-processing step for methylation phylogeny reconstruction. Specifically, we use the `inferCNV` tool[2] (from the Trinity CTAT Project) to estimate the number of copies of genomic segments across individual cells through the use of scRNA-seq data. Briefly, these somatic CNAs are estimated by sorting the analyzed genes with respect to their chromosomal location and applying a moving average of the relative expression values, with a sliding window including 100 genes within each chromosome. The average expression of genes in each malignant cell compared to the normal cells (with 2 copies), which are extracted from the normal colon in this data set, gives an estimate for the copy number of each gene in that cell.

Following the above analysis, we outline our sequencing error correction approach in Algorithm 1. For a CpG-site in a cell, given its copy number c and sequencing error probability $0 \leq \epsilon \leq 1$, we can enumerate the probability of drawing from a methylated allele p_M under all possible underlying allele status. The probability of drawing from an unmethylated allele in each case would be $1 - p_M$.[3]

In case all alleles are methylated, we have $p_M = 1-\epsilon$. Similarly, if all alleles are unmethylated, we have $p_M = \epsilon$. If we have a mix of allele status then for each possible value of γ, where $0 < \gamma < c$ denotes the number of methylated alleles, we have $p_M = \frac{\gamma}{c}$. Note that here we assume $p_M, 1 - p_M \gg \epsilon$ for all reasonable values of c and commonly encountered sequencing error rates, ϵ. Thus, we only consider the effect of sequencing error in the case of homozygous allele status.

Now given a CpG-site in the cell, the number of its methylated reads $n \geq 0$ and the number of its unmethylated reads $m \geq 0$, we can compute the likelihood of each possible underlying allele status, and identify the allele status with the maximum log likelihood for the site.

Sequencing error correction takes place when we have $n > 0$ and $m > 0$, but the allele status with the highest log likelihood is a homozygous one. In that case, we correct the reads with the minority methylation status to the majority one. In the case where sequencing error is needed, yet $n = m$, which will likely happen for $c = 1$, the CpG-site is discarded.

In case copy number information is not available, one can assume an appropriate (uninformative) copy number for all sites in all cells (e.g. $c = 2$ for diploid). We use the sequence error-corrected read information for pairwise distance estimation between cells, as described in the following section.

3.3 Computing the Expected Distance between Cell Pairs

Having a good measurement of the pairwise distance among cells based on the methylation status of selected CpG sites is critical to constructing high-quality methylation phylogeny. However, due to the shallow read coverage afforded by the scBS-seq data, we rarely have two or more reads (depending on CNV status) per CpG-site. Since allele-specific methylation has been shown to have increased frequency in cancer tissues [5], given the reads at a CpG-site, it is especially important to consider the possibility of unobserved alleles and their methylation status when determining the CpG-site's possible "methylation zygosities". In this section, we describe a formulation for computing the expected distance between two cells given their respective copy number and (sequence error-corrected) read status information. We consider such formulation

[2] https://github.com/broadinstitute/inferCNV

[3] Note that when $c = 1$ the site will implicitly be identified as homozygous and the methylation status of any read that does not conform to the majority allele will be corrected.

Algorithm 1 Sequencing Error Correction Accounting for CNV

```
n ← number of methylated reads for a CpG-site in a cell
m ← number of unmethylated reads for the CpG-site in the cell
c ← copy number for the CpG-site in the cell
ϵ ← sequencing error probability

procedure SequencingErrorCorrection(n, m, c, ϵ)
    if c == 0 then
        return None                          ▷ site is discarded in case of deletion event
    end if
    L[1^c] ← ln[(1 − ϵ)^n ϵ^m]
    L[0^c] ← ln[ϵ^n (1 − ϵ)^m]
    for 0 < γ < c do
        L[1^γ 0^(c−γ)] ← ln[(γ/c)^n (1 − γ/c)^m]
    end for
    status ← arg max L
    if [status == 1^c OR status == 0^c]
        AND [n ≠ 0 AND m ≠ 0] then
        if n > m then
            n ← n + m
            m ← 0
        else if n < m then
            m ← n + m
            n ← 0
        else
            return None                      ▷ site is discarded if correction is impossible
        end if
    end if
    return n, m
end procedure
```

in hopes of correcting for the potential bias contributed by low coverage CpG-sites. For clarity, we will first introduce the intuition behind our formulation with the assumption of copy number $c = 2$ for all CpG-sites in all cells. Then, in the following section, we will generalize the formulation to account for CNV information.

3.3.1 Expected Distance Calculation Assuming Copy Number $c = 2$

In this section, we hold the following assumption: in case a cell has a heterozygously methylated site with copy number $= 2$, the probability of drawing the allele with the site methylated is p, and that of drawing the allele with the site unmethylated is $1-p$. In our results section we have used $p = 1 - p = 0.5$, assuming it is equally probable to draw from either allele; however, our model makes it possible to have $p \neq 0.5$ to model any potential bias in methylation status-specific allele sampling rate as has been reported in the literature [9].

In addition, our formulation assumes that the prior probability for all possible allele status are given - i.e. the probability of observing a pair of homozygous methylated, heterozygous methylated, and homozygous unmethylated alleles for any CpG site in any cell is independently and identically distributed and are respectively denoted as P(11), P(10) and P(00).

With these assumptions, we introduce a formulation for calculating the expected distance

between a pair of cells. Key to this formulation is modeling the probability of the alleles at the CpG-site having a particular methylation status, given the observed reads at CpG-site. For the sake of clarity in notation, we first give an example of computing the probability of each allele status having observed methylated reads at a CpG-site. We briefly show the case for observing homozygous unmethylated reads, and reads of mixed methylation status. Lastly, we show how we can compute the expected distance for sites between two cells given the expected site allele status for either cell.

Methylation Status Probabilities of CpG-sites with Only Methylated Reads. Suppose we have observed n reads at a CpG-site and all n of them are methylated; then we know the probability of drawing only methylated reads from a pair of homozygous methylated alleles is:

$$\mathrm{P}(\text{reads} \mid 11) = 1$$

And the probability of drawing only methylated reads from a pair of heterozygous methylated alleles is:

$$\mathrm{P}(\text{reads} \mid 10) = p^n$$

Since we have already corrected for sequencing errors in a previous step as described in Section 3.2, we know that we cannot draw methylated reads from a pair of homozygous unmethylated alleles:

$$\mathrm{P}(\text{reads} \mid 00) = 0$$

By Bayes' Theorem we know that:

$$\mathrm{P}(\text{allele status} \mid \text{reads}) \propto \mathrm{P}(\text{reads} \mid \text{allele status})\mathrm{P}(\text{allele status}) \tag{1}$$

Thus, for each possible allele status, we compute its probability given the observed reads by computing the product on the right hand side of Equation 1, then normalize it by the sum of that of all allele status. For example, let

$$a = \mathrm{P}(\text{reads} \mid 11)\mathrm{P}(11) + \mathrm{P}(\text{reads} \mid 10)\mathrm{P}(10) + \mathrm{P}(\text{reads} \mid 00)\mathrm{P}(00)$$

then we can compute the probability of the reads being drawn from two alleles of any of the three possible combined status as described in Equations 2:

$$\begin{aligned}
\mathrm{P}(11 \mid \text{reads}) =& \frac{\mathrm{P}(\text{reads} \mid 11)\mathrm{P}(11)}{a} = \frac{\mathrm{P}(11)}{\mathrm{P}(11)+p^n\mathrm{P}(10)} \\
\mathrm{P}(10 \mid \text{reads}) =& \frac{\mathrm{P}(\text{reads} \mid 10)\mathrm{P}(10)}{a} = \frac{p^n\mathrm{P}(10)}{\mathrm{P}(11)+p^n\mathrm{P}(10)} \\
\mathrm{P}(00 \mid \text{reads}) =& \frac{\mathrm{P}(\text{reads} \mid 00)\mathrm{P}(00)}{a} = 0
\end{aligned} \tag{2}$$

Methylation Status Probabilities of CpG-sites with Only Unmethylated Reads. Similarly, suppose we observed n reads at a CpG-site and all of them are unmethylated, we have:

$$\begin{aligned}
\mathrm{P}(\text{reads} \mid 11) =& 0 \\
\mathrm{P}(\text{reads} \mid 10) =& (1-p)^n \\
\mathrm{P}(\text{reads} \mid 00) =& 1
\end{aligned}$$

and we can compute the probability of the reads being drawn from alleles of each combined status as described in Equations 3:

$$\begin{aligned}
\mathrm{P}(11 \mid \text{reads}) &= \frac{\mathrm{P}(\text{reads} \mid 11)\mathrm{P}(11)}{a} = 0 \\
\mathrm{P}(10 \mid \text{reads}) &= \frac{\mathrm{P}(\text{reads} \mid 10)\mathrm{P}(10)}{a} = \frac{(1-p)^n \mathrm{P}(10)}{\mathrm{P}(00)+(1-p)^n \mathrm{P}(10)} \\
\mathrm{P}(00 \mid \text{reads}) &= \frac{\mathrm{P}(\text{reads} \mid 00)\mathrm{P}(00)}{a} = \frac{\mathrm{P}(00)}{\mathrm{P}(00)+(1-p)^n \mathrm{P}(10)}
\end{aligned} \tag{3}$$

Methylation Status Probabilities of CpG-sites with Methylated Reads of Mixed status. For completion, we also consider the case where we observe both methylated and unmethylated reads at a CpG-site in a cell. Since we have already corrected for potential sequencing error in a previous step as describe in Section 3.2, if we observe reads of mixed methylation status, $\mathrm{P}(11 \mid \text{reads}) = \mathrm{P}(00 \mid \text{reads}) = 0$. Therefore the probability of the reads being drawn from alleles of each combined status are as described in Equations 4:

$$\begin{aligned}
\mathrm{P}(11 \mid \text{reads}) &= 0 \\
\mathrm{P}(00 \mid \text{reads}) &= 0 \\
\mathrm{P}(10 \mid \text{reads}) &= 1
\end{aligned} \tag{4}$$

Expected Distance Calculation. Now, to compute the expected distance for a CpG-site in two cells, we take the expectation over the possible combinations of allele status between the two cells. Here, we assume:

$$\begin{gathered}
\text{dist}(11,11)=\text{dist}(10,10)=\text{dist}(00,00)=0 \\
\text{dist}(11,10)=\text{dist}(10,11)=\text{dist}(00,10)=\text{dist}(10,00)=0.5 \\
\text{dist}(11,00)=\text{dist}(00,11)=1
\end{gathered}$$

We can then compute the expected distance between some Cell A and Cell B at methylation site s - assuming copy number $c = 2$ in both cells - as follows:

$$\text{dist}(\text{reads}_{\text{A},s}, \text{reads}_{\text{B},s}) = \sum_{\substack{\text{status}_{\text{A},s} \\ \in\{11,10,00\}}} \sum_{\substack{\text{status}_{\text{B},s} \\ \in\{11,10,00\}}} \mathrm{P}(\text{status}_{\text{A},s}|\text{reads}_{\text{A},s})\mathrm{P}(\text{status}_{\text{B},s}|\text{reads}_{\text{B},s}) \text{dist}(\text{status}_{\text{A},s}, \text{status}_{\text{B},s}) \tag{5}$$

The total expected distance between Cell A and B can now be computed via some distance function over the vector of distances across all of their shared sites. For example, one can define the total distance between Cell A and B as the L_1 norm of the distance vector, normalized by the number of shared sites, as described in Equation 6:

$$\text{dist(A,B)} = \frac{\sum_{s \in \text{sites}_\text{A} \cap \text{sites}_\text{B}} \text{dist}(\text{reads}_{\text{A},s}, \text{reads}_{\text{B},s})}{|\text{sites}_\text{A} \cap \text{sites}_\text{B}|} \tag{6}$$

3.3.2 Expected Distance Calculation Accounting for CNV

Section 3.3.1 describes a scheme for expected distance computation between a pair of cells via first computing each cell's expected methylation status, assuming that both cells have copy number $c = 2$ at all sites. In this section, we describe how we can extend the formulation for CpG-sites with CNV information.

Recall from Section 3.3.1 that, if there are 2 alleles that are heterozygously methylated at a particular site, we let p to be the probability of drawing the allele with the CpG-site methylated, and $1 - p$ be the probability of drawing the allele with the CpG-site unmethylated. Suppose

that now, at this CpG-site we have copy number $c > 2$, and we have γ alleles with the CpG-site methylated, and $c - \gamma$ alleles with the CpG-site unmethylated. We can recompute the probability of drawing from an allele with the CpG-site methylated by normalizing over all alleles as described in Equation 7:

$$p_{c,\gamma} = \frac{\gamma p}{\gamma p + (c - \gamma)(1 - p)} \tag{7}$$

It follows that, the probability of drawing from an allele with the CpG-site unmethylated is $1 - p_{c,\gamma}$. Now, we can extend extend Section 3.3.1 to take into account CNV information at CpG-sites. For brevity, we only show results for when we observe n reads that are all methylated: results for when observing all n unmethylated reads and for when observing n mixed methylation status reads can be extended in a similar manner.

For a cell with an arbitrary copy number c at a particular CpG-site and n observed reads for that CpG-site that are all methylated, we know the probabilities of drawing all n methylated reads from c alleles with the CpG-site methylated (i.e. $\gamma = c$) is:

$$\mathrm{P}(\text{reads} \mid 1^c) = 1$$

The probability of drawing all n methylated reads from c alleles with the CpG-site unmethylated (i.e. $\gamma = 0$) is:

$$\mathrm{P}(\text{reads} \mid 0^c) = 0$$

To compute the probability of drawing n methylated reads from c alleles with mixed methylation status for the site, we need to sum the probabilities over all other possible values of γ, the number of alleles with the site methylated, assuming that all other possible values of γ are equally likely:

$$\mathrm{P}(\text{reads} \mid \text{mixed}) = \begin{cases} \frac{1}{c-1} \sum_{\gamma=1}^{c-1} p_{c,\gamma}^n & , c \geq 2 \\ 0 & , c = 1 \end{cases}$$

It is worth noting that, in the case of copy number loss (i.e. $c = 1$), the formulation assigns only non-zero probability to a homozygous combined allele status. Given $\mathrm{P}(1^c)$, $\mathrm{P}(0^c)$, and $\mathrm{P}(\text{mixed})$, let $a = \mathrm{P}(\text{reads} \mid 1^c)\mathrm{P}(1^c) + \mathrm{P}(\text{reads} \mid \text{mixed})\mathrm{P}(\text{mixed}) + \mathrm{P}(\text{reads} \mid 0^c)\mathrm{P}(0^c)$; then we apply Bayes' Theorem to get the probability of any allele status given the observed reads, as described in Equations 8:

$$\begin{aligned} \mathrm{P}(1^c \mid \text{reads}) &= \frac{\mathrm{P}(\text{reads} \mid 1^c)\mathrm{P}(1^c)}{a} = \frac{\mathrm{P}(1^c)}{\mathrm{P}(1^c) + \frac{\mathrm{P}(\text{mixed})}{c-1} \sum_{\gamma=1}^{c-1} p_{c,\gamma}^n} \\ \mathrm{P}(\text{mixed} \mid \text{reads}) &= \frac{\mathrm{P}(\text{reads} \mid \text{mixed})\mathrm{P}(\text{mixed})}{a} = \frac{\frac{\mathrm{P}(\text{mixed})}{c-1} \sum_{\gamma=1}^{c-1} p_{c,\gamma}^n}{\mathrm{P}(1^c) + \frac{\mathrm{P}(\text{mixed})}{c-1} \sum_{\gamma=1}^{c-1} p_{c,\gamma}^n} \\ \mathrm{P}(0^c \mid \text{reads}) &= \frac{\mathrm{P}(\text{reads} \mid 0^c)\mathrm{P}(0^c)}{a} = 0 \end{aligned} \tag{8}$$

Then, for a given CpG site s in Cell A and Cell B, the respective copy numbers $c_{\mathrm{A},s}$, $c_{\mathrm{B},s}$, and respective reads, we can compute the expected distance between the cells at s as in Equation 9:

$$\begin{aligned} \mathrm{dist}(\text{reads}_{\mathrm{A},s}, c_{\mathrm{A},s}, \text{reads}_{\mathrm{B},s}, c_{\mathrm{B},s}) = \sum_{\substack{\text{status}_{\mathrm{A},s} \\ \in \{1^{c_\mathrm{A}}, \\ \text{mixed}, \\ 0^{c_\mathrm{A}}\}}} \sum_{\substack{\text{status}_{\mathrm{B},s} \\ \in \{1^{c_\mathrm{B}}, \\ \text{mixed}, \\ 0^{c_\mathrm{B}}\}}} \mathrm{P}(\text{status}_{\mathrm{A},s} | \text{reads}_{\mathrm{A},s}) \mathrm{P}(\text{status}_{\mathrm{B},s} | \text{reads}_{\mathrm{B},s}) \\ \mathrm{dist}(\text{status}_{\mathrm{A},s}, \text{status}_{\mathrm{B},s}) \end{aligned} \tag{9}$$

The total expected distance between Cell A and B can now be computed with some distance function over the vector of expected distances over all shared sites. The L_1 norm normalized by

the number of shared sites is computed via Equation 10:

$$\text{dist(A,B)} = \frac{\sum_{s \in \text{sites}_\text{A} \cap \text{sites}_\text{B}} \text{dist}(\text{reads}_{\text{A},s}, c_{\text{A},s}, \text{reads}_{\text{B},s}, c_{\text{B},s})}{|\text{sites}_\text{A} \cap \text{sites}_\text{B}|} \tag{10}$$

After computing the distance between each pair of cells, we can leverage any distance-based phylogeny reconstruction method to obtain the final methylation phylogeny.

3.4 Applications to scBS-seq Data from Colorectal Cancer Samples

3.4.1 Data Overview

We applied our methylation phylogeny reconstruction pipeline to the colorectal cancer data set generated using the scTrio-seq protocol [3]. Data from patient CRC01 is the primary source of input in our analysis, as CRC01 has, by far, the greatest number of cells sequenced, largest number of distinct tissue types, sampling locations, and treatment conditions. The cells in CRC01 fall into the following main categories: primary colorectal tumor cells (PT), normal colon cells adjacent to tumor (NC), lymph node metastasis cells (LN), liver metastasis cells (ML), and post-treatment liver metastatsis cells (MP). For each category, there are one or more sampling locations: PT has 4, NC has 1, LN has 3, ML has 4, and MP has 5 sampling locations. In total these sampling locations have 409 cells with scBS-seq read data; 102 of them have very low total read coverage and were excluded from our analysis.

To demonstrate that tumor methylation profiles contain richer signal of tumor evolution than mutation profiles, we also construct mutation-based phylogeny using scRNA-seq data from Bian et al. [3] as a comparison (Figure 2). Method for constructing mutation-based phylogeny is described in Section 3.4.3.

3.4.2 Reconstructing the methylation phylogeny

Our biclustering approach was used to select maximally informative CpG-sites and cells from a binary matrix indicating read availability for each CpG-site, as described in Section 3.1. Specifically, this binary matrix contained the read availability information for 133,458 sites shared by at least 65% of 307 cells in patient CRC01. We performed biclustering with parameters $\alpha = 0.95, \beta = 0.5$, which resulted in 291 cells and 66,729 sites. Of those 291 biclustering-selected cells, we focused only on 188 which had available CNV calls generated from matching scRNA-seq read data through the use of the inferCNV tool. The methylated and unmethylated read counts of these cells were then corrected based on CNV calls as described in Section 3.2 with a sequencing error rate of $\epsilon = 0.01$.

For any pair of cells, the mean expected distance across all of their shared sites were computed as their total distance with parameters $\text{P}(1^c)$=P(mixed)=$\text{P}(0^c)$=0.33 and $p = 1 - p = 0.5$. The tree was then obtained from the pairwise distance calculated for cells with the NJ approach described in [18] using the scikit-bio package [19].

3.4.3 Reconstructing the mutation phylogeny

Constructing mutation-based phylogenies calls for an approach different from the methylation phylogeny reconstruction framework proposed in this work, because there are typically notably fewer observed mutation events than methylation events. The mutation phylogeny we will construct is based on single-nucleotide variant (SNV) and insertion-deletion (indel) mutations derived from matching scRNA-seq data [3] through the use of GATK HaplotypeCaller pipeline followed by filtering germline variants (from matched bulk whole genome sequencing (WGS) data), and SNPs from dbSNP, and 1000 Genomes database. We used ANNOVAR to annotate genetic variants, and filtered those variants whose status was unknown (i.e. only synonymous and nonsynonymous were kept). After that private mutations (i.e only appeared in one cell) were excluded. Next, to reduce the artifacts and false positive errors of RNA sequencing, the

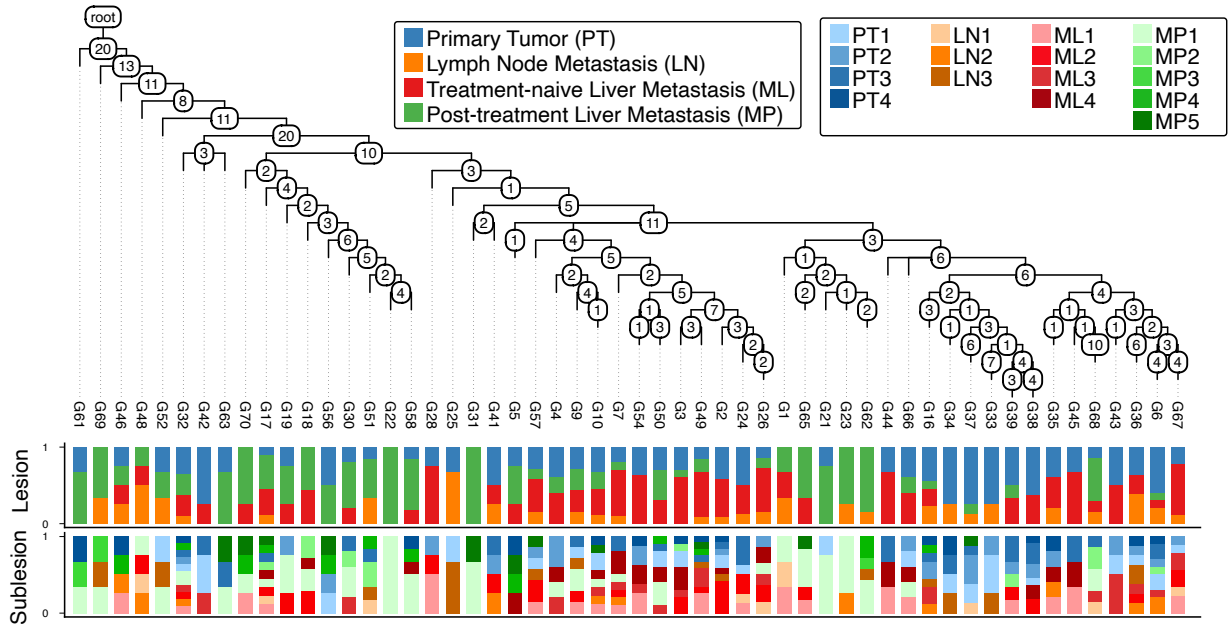

Figure 2: Tumor phylogeny based on SNVs/Indels obtained from scRNA-seq data from patient CRC01 Bian et al. [3]. Leaves are labeled by cells and numbers shown in rounded rectangular nodes represent number of *de novo* mutations shared only by the set of cells that are descendants of these nodes. The middle and bottom panels are stacked plots of the proportion of cells in leaves with respect to the tissue and sampling location, respectively. Note that normal adjacent colon cells (NC) contain no mutations and therefore are part of the root in this phylogeny.

mutations that were present in the bulk WGS of the same lesion were kept for downstream analysis resulting in 290 mutations in total. Finally, a local clustering of cells was performed for reducing the sparsity of the data - as a result each leaf depicted in the tree represents a combination of multiple cells with similar mutational profiles. The final phylogeny on these cell clusters was obtained by applying PhISCS [15, 2, 16].

3.4.4 Results

Tumor phylogeny based on scRNA-seq SNVs/Indels mutations. Before we show the methylation phylogenies constructed using our proposed framework, for comparison, we provide the tumor phylogeny we obtained by applying tools such as PhISCS [15, 2, 16] from mutation calls through the use of scRNA-seq data. Even though mutation data is, in principle, better suited for cellular lineage identification and phylogeny reconstruction, the sparsity of scRNA-seq data makes it very difficult to infer a reliable tumor phylogeny. Thus it is not surprising that the resulting tumor phylogeny poorly corroborates with the tissue origin and sampling locations of the tumor cells (Figure 2). This motivates phylogeny reconstruction using single cell methylation profiles as shown next.

Tumor phylogeny based on single-cell methylation. First, we randomly selected 188 cells and 66,729 CpG-sites from the original data set from Bian et al. [3], and naïvely constructed a baseline methylation tree to compare against the methylation phylogeny constructed with our proposed framework (Figure S1). For each pair of cells, we computed the L1 distance over the read methylation rates of their shared CpG-sites, and normalized over the number of shared sites. We then constructed trees using such pairwise distances via NJ [18]. As we observe, while there is some noticeable degree of clustering among cells from the same tissue of

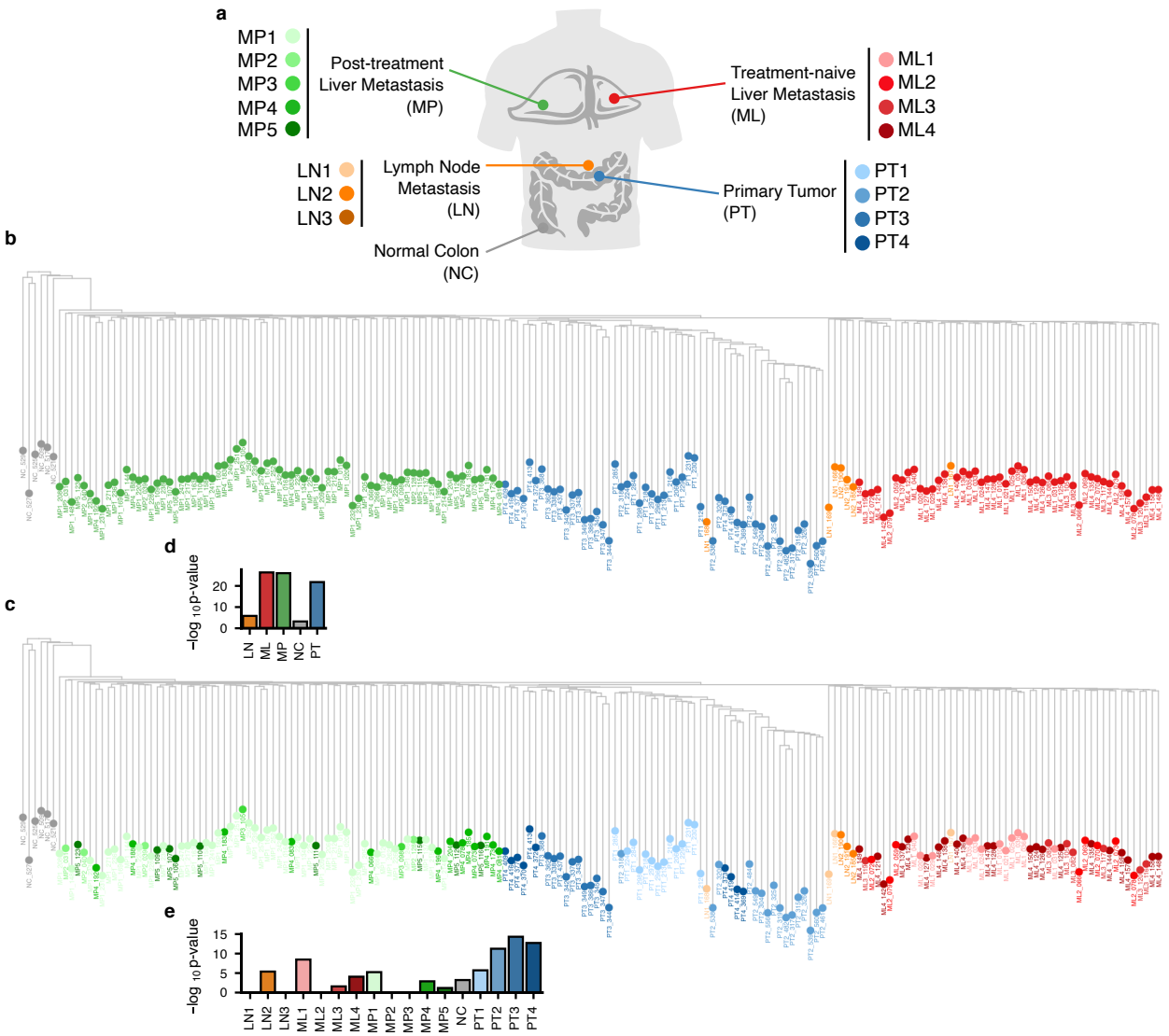

Figure 3: Methylation phylogeny of tumor cells from colorectal cancer patient CRC01 [3]. **(a)** Tissue origins and sampling locations of tumor cells sampled from colorectal cancer patient CRC01 in scTrio-seq data set [3]. **(b)** Phylogeny reconstructed following biclustering with $\alpha = 0.95$ and $\beta = 0.5$ with nodes colored according to the tissue origin of cells. **(c)** Phylogeny reconstructed following biclustering with $\alpha = 0.95$ and $\beta = 0.5$ with nodes colored according to the (more detailed) sampling locations of cells. **(d)** P-values for each tissue type were calculated using one-tailed binomial test. **(e)** P-values for each sampling location were calculated using one-tailed binomial test. (See Table S1 for p-values of these phylogenies.)

origin (Figure S1(a)) or sampling location (Figure S1(b)), there is also nontrivial mixing among the groups.

In contrast, the phylogeny in Figure 3(b), reconstructed following biclustering with $\alpha = 0.95, \beta = 0.5$, reflect the tissue origins of cells very well. We also observe such correspondence, although to a lesser extent, between the phylogeny in Figure 3(c) and distinct sampling locations of cells within each tissue. The distribution of all pairwise distances used to construct the phylogeny is visualized in Figure S4.

We also considered how biclustering parameters α and β affect the quality of constructed methylation phylogenies. Lowering α and/or β eliminates those rows and columns with limited signal, though at the expense of losing some valuable information. To illustrate, we alternatively performed biclustering by maintaining $\alpha = 0.95$ but setting $\beta = 0.1$, a substantially smaller value, to reduce the number of CpG-sites in the data set. The effect of such over-reduction in the number of CpG-sites can be seen in Figure 3: trees built with over-reduced sites show poorer correlation with tissues of origin for cells (Figure 3(b) vs S2(a)) as well as their sampling locations within the tissues (Figure 3(c) vs S2(b)). The difference is not only visibly evident, it is also supported by the corresponding p-values for the phylogenies (Figure 3(d),(e), S2(c)(d), Table S1).

We would like to also point out that there is negligible correlation between the coverage information and the tissue or sampling locations of the cells: the proximity of a pair of cells (indicated by whether they have been extracted from the same tissue or from the same sampling location within the tissue) does not correlate with the number of CpG sites where they both have reads. This implies that the correspondence between our reconstructed phylogeny and cellular origins is not the result of tissue or sampling locus specific site sequencing bias. In order to demonstrate this, we "binarized" the cell-by-site matrix used for phylogeny reconstruction based on the presence (corresponding to a 1) and absence (corresponding to a 0) of reads, and constructed a normalized L_1 distance-based NJ tree using this binary matrix. The resulting trees labeled by cellular tissue of origin and specific sampling location within the tissue are respectively given in Figures S3(a) and (b). As we can see, cells of distinct origins are only weakly clustered.

4 Methylation Status Imputation in scBS-seq Data

As we have shown strong concordance between the methylation phylogeny constructed by our proposed framework and the tissue origin of single cells, we now demonstrate that such constructed methylation phylogeny can be effective in imputing missing methylation status of CpG-sites.

4.1 Data Preparation

In order to generate data sets that can be validated, we leveraged the cell-by-site matrix of read information (after one round of biclustering and corrected for sequencing error) used to reconstruct our methylation phylogeny in Section 3.4.2. For each CpG-site, we randomly selected one cell with at least two reads at that CpG-site, and held out its read information. We did not hold out entries with only one read because we believe it is very unlikely to capture the true methylation status at the CpG-stie, and therefore is unreliable to use as validation.

In the end, we obtained two matrices: (i) the now further sparsified cell-by-site read information matrix, and (ii) the validation matrix with the ground-truth read information for the held-out sites in cells. We used the first matrix to construct distance-based phylogeny using our pipeline Section 3.3.2, imputed the most likely methylation status at held-out sites with the constructed phylogeny with an approach described below (Section 4.2), and used the second ground-truth matrix to validate the accuracy of imputation (Section 4.3).

4.2 Phylogeny-based Methylation Status Imputation via Plurality Vote

Our imputation approach is based on the methylation phylogeny constructed by our proposed framework, and leverages the fact - observable from Figure 3 - that the methylation profile of a cell is most likely similar to those of its neighboring cells in the phylogeny.

In a methylation phylogeny reconstructed via a distance-based approach, the leaves are single cells whose read information in the original matrix serves as the basis of the phylogeny, and the internal nodes do not have any information. Given a phylogeny constructed from read information, for each leaf node, we assign each CpG-site its most likely methylation status. For

sequencing error-corrected reads, a site is homozygously methylated if all of its reads have the site methylated, homozygously unmethylated if all unmethylated, and heterozygous if the read status for the site is mixed. The status is unknown if there is no read information available.

Now that we have heuristically called methylation statuses for all the single cells at the leaf nodes, we can proceed with a plurality-vote imputation scheme. For each CpG-site:

(i) **Tallying the Votes.** Each single cell that has read information available for that CpG-site holds a single vote for the methylation status called for that CpG-site. If a cell does not have read information for a CpG-site, it holds no vote for the CpG-site. With votes current held by the leaf nodes, we report their votes to their parents. Each parent node tallies the votes received from its children for each methylation status - homozygously methylated, heterozygously methylation, or homozygously unmethylated - before reporting the votes to their parents. As such, starting with the immediate parents of leaf nodes, we continue merging vote counts upwards in the phylogeny. Once the merging finishes, at any internal node, the number of votes for a particular methylation status is exactly the number of leaf nodes with that methylation status in the subtree rooted at the internal node.

(ii) **Answering a Query.** Given a threshold for the minimum number of total votes needed to determine the methylation status of a CpG-site with sufficient confidence, for any query regarding the methylation status of a single cell at a particular CpG-site, we travel upwards in the phylogeny until we find the closest ancestor that has tallied a total vote count greater than or equal to the threshold. We take the methylation status with the most number of votes at that ancestor as the methylation status for the query cell at that CpG-site. Note that if there is a tie, we randomly select a methylation status from those that received the most number of votes.

4.3 Results

Here we compare the plurality-vote phylogeny imputation scheme proposed above against a baseline approach, which for a queried CpG-site randomly selecting a cell with more than two reads at that CpG-site and answer the query with the heuristically called methylation status of that cell at the site. It is apparent that the performance of this baseline approach depends on the homogeneity of methylation statuses at a CpG-site: when 80% of the cells is homozygously methylated at a CpG-site, if one queries the CpG-site methylation status of a held-out cell homozygously methylated at the site, the baseline approach would get the site right 80% of the time. Therefore, to mitigate the effect of the variation in the homogeneity of CpG-site methylation status on the performance of the baseline approach, we consider accuracy results for CpG-sites with different methylation status homogeneity levels separately (Table 1).

We prepared 10 different experimental data sets according to the hold-out procedures described in Section 4.1, imputed the methylation status for the held-out sites with the plurality voting scheme described in Section 4.2 and with the baseline approach, and calculated the mean and standard deviation of the imputation accuracy for sets of CpG-sites within each homogeneity level (Table 1). An imputed methylation status counts as correct only if it is the same as that heuristically called from the held-out read information. The phylogeny-based plurality vote imputation approach outperforms the baseline at all CpG-site homogeneity levels.

Homogeneity Level	Plurality Vote Approach	Baseline Approach
35-45%	$\mathbf{0.4011 \pm 0.005160}$	0.3419 ± 0.009225
45-55%	$\mathbf{0.4796 \pm 0.006316}$	0.3984 ± 0.006067
55-65%	$\mathbf{0.5742 \pm 0.008917}$	0.4605 ± 0.007197
65-75%	$\mathbf{0.6869 \pm 0.006649}$	0.5503 ± 0.010389

Table 1: Accuracy values (mean ± standard deviation) for methylation phylogeny-based imputation v.s. baseline in held-out experiments, varying across probability values.

5 Conclusion

Motivated by the observation that the NJ tree constructed with bulk methylation array data of CNS tumors shows strong clustering of tumor categories (Section 2), we set out to reconstruct phylogenies from single-cell methylation profiles, which we hope will offer a clearer view of tumor evolution on a single cell level. To this end, we introduced a computational framework (Section 3) that constructs phylogenies from single-cell methylation profiles that could additionally leverage orthogonally obtained CNV information, if available. The three main components of our framework - integer linear programming-based noise-minimizing site selection (Section 3.1), maximum likelihood-based sequencing error correction (Section 3.2), and pairwise expected distance calculation for single cell methylation read profiles (Section 3.3.2) - are all designed to mitigate the effect of noise and uncertainty due to data sparsity commonly observed in scBS-seq data.

Leveraging scBS-seq data and CNV called from scRNA-seq data by Bian et al. [3], we demonstrated that, in contrast to the mutation-based single-cell phylogeny constructed from scRNA-seq data, the methylation-based single-cell phylogeny constructed by our proposed framework shows a strong correlation with the tissue of origin of the cells (Section 3.4.4); furthermore, the inferred phylogeny can facilitate the imputation of missing methylation status in the data (Section 4). Such single cell methylation phylogenies show potential in facilitating the identification of key methylation events in tumor evolution and the discovery of novel methylation markers for overall survival in cancer patients - we have great interests in exploring these aspects in future work.

Acknowledgements

This work is supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute and utilized the computational resources of the NIH Biowulf high performance computing cluster (http://hpc.nih.gov) and Gurobi (http://www.gurobi.com) to solve the optimization problems. F.R.M. was supported in part by Indiana U. Grand Challenges Precision Health Initiative.

References

[1] Christof Angermueller, Heather J. Lee, Wolf Reik, and Oliver Stegle. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1), April 2017. doi: 10.1186/s13059-017-1189-z. URL https://doi.org/10.1186/s13059-017-1189-z.

[2] Erfan Sadeqi Azer, Farid Rashidi Mehrabadi, Salem Malikić, Xuan Cindy Li, Osnat Bartok, Kevin Litchfield, Ronen Levy, Yardena Samuels, Alejandro A Schäffer, E Michael Gertz, Chi-Ping Day, Eva Pérez-Guijarro, Kerrie Marie, Maxwell P Lee, Glenn Merlino, Funda Ergun, and S Cenk Sahinalp. PhISCS-BnB: a fast branch and bound algorithm for the perfect tumor phylogeny reconstruction problem. *Bioinformatics*, 36(Supplement_1):i169–i176, July 2020. doi: 10.1093/bioinformatics/btaa464. URL https://doi.org/10.1093/bioinformatics/btaa464.

[3] Shuhui Bian, Yu Hou, Xin Zhou, Xianlong Li, Jun Yong, Yicheng Wang, Wendong Wang, Jia Yan, Boqiang Hu, Hongshan Guo, Jilian Wang, Shuai Gao, Yunuo Mao, Ji Dong, Ping Zhu, Dianrong Xiu, Liying Yan, Lu Wen, Jie Qiao, Fuchou Tang, and Wei Fu. Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science*, 362(6418):1060–1063, November 2018. doi: 10.1126/science.aao3791. URL https://doi.org/10.1126/science.aao3791.

[4] David Capper et al. DNA methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, March 2018. doi: 10.1038/nature26000. URL https://doi.org/10.1038/nature26000.

[5] Catherine Do et al. Allele-specific DNA methylation is increased in cancers and its dense mapping in normal plus neoplastic cells increases the yield of disease-associated regulatory SNPs. *Genome Biology*, 21(1), June 2020. doi: 10.1186/s13059-020-02059-3. URL https://doi.org/10.1186/s13059-020-02059-3.

[6] Federico Gaiti, Ronan Chaligne, Hongcang Gu, Ryan M. Brand, Steven Kothen-Hill, Rafael C. Schulman, Kirill Grigorev, Davide Risso, Kyu-Tae Kim, Alessandro Pastore, Kevin Y. Huang, Alicia Alonso, Caroline Sheridan, Nathaniel D. Omans, Evan Biederstedt, Kendell Clement, Lili Wang, Joshua A. Felsenfeld, Erica B. Bhavsar, Martin J. Aryee, John N. Allan, Richard Furman, Andreas Gnirke, Catherine J. Wu, Alexander Meissner, and Dan A. Landau. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature*, 569(7757):576–580, May 2019. doi: 10.1038/s41586-019-1198-z. URL https://doi.org/10.1038/s41586-019-1198-z.

[7] Ermin Hodzic, Raunak Shrestha, Salem Malikic, Colin C Collins, Kevin Litchfield, Samra Turajlic, and S Cenk Sahinalp. Identification of conserved evolutionary trajectories in tumors. *Bioinformatics*, 36(Supplement_1):i427–i435, July 2020. doi: 10.1093/bioinformatics/btaa453. URL https://doi.org/10.1093/bioinformatics/btaa453.

[8] Richard R. Hudson and Norman L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of dna sequences. *Genetics*, 111(1):147–164, 1985. ISSN 0016-6731.

[9] Lexiang Ji, Takahiko Sasaki, Xiaoxiao Sun, Ping Ma, Zachary A. Lewis, and Robert J. Schmitz. Methylated dna is over-represented in whole-genome bisulfite sequencing data. *Frontiers in Genetics*, 5:341, 2014. ISSN 1664-8021.

[10] Bozena Kaminska et al. Consequences of IDH1/2 Mutations in Gliomas and an Assessment of Inhibitors Targeting Mutated IDH Proteins. *Molecules*, 24(5):968, March 2019.

[11] Chantriolnt-Andreas Kapourani and Guido Sanguinetti. Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biology*, 20(1), March 2019. doi: 10.1186/s13059-019-1665-8. URL https://doi.org/10.1186/s13059-019-1665-8.

[12] Chantriolnt-Andreas Kapourani, Ricard Argelaguet, Guido Sanguinetti, and Catalina A. Vallejos. scMET: Bayesian modelling of DNA methylation heterogeneity at single-cell resolution. *bioRxiv*, July 2020. doi: 10.1101/2020.07.10.196816. URL https://doi.org/10.1101/2020.07.10.196816.

[13] Dan A. Landau et al. Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell*, 26(6):813–825, December 2014. doi: 10.1016/j.ccell.2014.10.012. URL https://doi.org/10.1016/j.ccell.2014.10.012.

[14] Chongyuan Luo, Christopher L. Keown, Laurie Kurihara, Jingtian Zhou, Yupeng He, Junhao Li, Rosa Castanon, Jacinta Lucero, Joseph R. Nery, Justin P. Sandoval, Brian Bui, Terrence J. Sejnowski, Timothy T. Harkins, Eran A. Mukamel, M. Margarita Behrens, and Joseph R. Ecker. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604, August 2017. doi: 10.1126/science.aan3351. URL https://doi.org/10.1126/science.aan3351.

[15] Salem Malikic, Farid Rashidi Mehrabadi, Simone Ciccolella, Md. Khaledur Rahman, Camir Ricketts, Ehsan Haghshenas, Daniel Seidman, Faraz Hach, Iman Hajirasouliha, and S. Cenk

Sahinalp. PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research*, 29(11):1860–1877, October 2019. doi: 10.1101/gr.234435.118. URL https://doi.org/10.1101/gr.234435.118.

[16] Salem Malikić, Farid Rashidi Mehrabadi, Erfan Sadeqi Azer, Mohammad Haghir Ebrahimabadi, and S. Cenk Sahinalp. Studying the history of tumor evolution from single-cell sequencing data by exploring the space of binary matrices. *bioRxiv*, July 2020. doi: 10.1101/2020.07.15.204081. URL https://doi.org/10.1101/2020.07.15.204081.

[17] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, November 2014. doi: 10.1093/molbev/msu300. URL https://doi.org/10.1093/molbev/msu300.

[18] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 07 1987. ISSN 0737-4038.

[19] The scikit-bio development team. scikit-bio: A bioinformatics library for data scientists, students, and developers, 2020. URL http://scikit-bio.org.

[20] R. Shoemaker, J. Deng, W. Wang, and K. Zhang. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, 20(7):883–889, April 2010. doi: 10.1101/gr.104695.109. URL https://doi.org/10.1101/gr.104695.109.

[21] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820, July 2014. doi: 10.1038/nmeth.3035. URL https://doi.org/10.1038/nmeth.3035.

[22] Andrea Sottoriva, Inmaculada Spiteri, Darryl Shibata, Christina Curtis, and Simon Tavaré. Single-Molecule Genomic Data Delineate Patient-Specific Tumor Profiles and Cancer Stem Cell Organization. *Cancer Research*, 73(1):41–49, October 2012. doi: 10.1158/0008-5472.can-12-2273. URL https://doi.org/10.1158/0008-5472.can-12-2273.

[23] Marco Trerotola, Valeria Relli, Pasquale Simeone, and Saverio Alberti. Epigenetic inheritance and the missing heritability. *Human Genomics*, 9(1), July 2015. doi: 10.1186/s40246-015-0041-3. URL https://doi.org/10.1186/s40246-015-0041-3.

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[25] Guangchuang Yu. Using ggtree to visualize data on tree-like structures. *Current Protocols in Bioinformatics*, 69(1):e96, 2020.

[26] Guangchuang Yu et al. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 2017.

[27] Guangchuang Yu et al. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. *Molecular Biology and Evolution*, 35(12):3041–3043, 10 2018. ISSN 0737-4038.

Supplementary Material

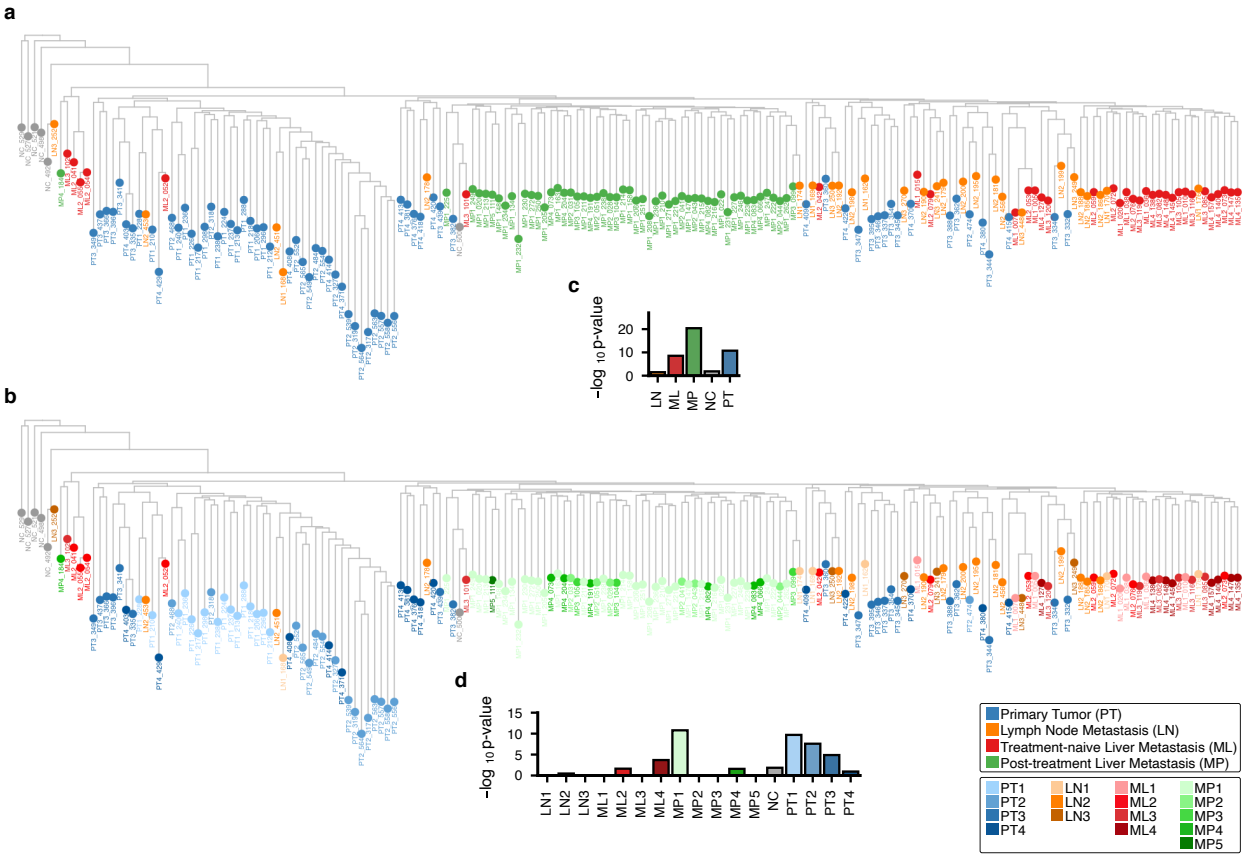

Figure S1: **(a)** Naïve NJ tree for patient CRC01 from Bian et al. [S3] on 188 cells and 66,729 sites, which are selected uniformly at random in contrast to the tree in Figure 3 on the same number of cells and sites but selected by our pipeline; nodes are again colored according to the tissue origin of cells. **(b)** Naïve NJ tree with the same 188 cells and 66,729 sites used in panel (a), with nodes colored according to the (more detailed) sampling locations of cells. **(c)** P-values for each tissue type were calculated using one-tailed binomial test. **(d)** P-values for each sampling location were calculated using one-tailed binomial test. (See Table S1 for p-values of these phylogenies.)

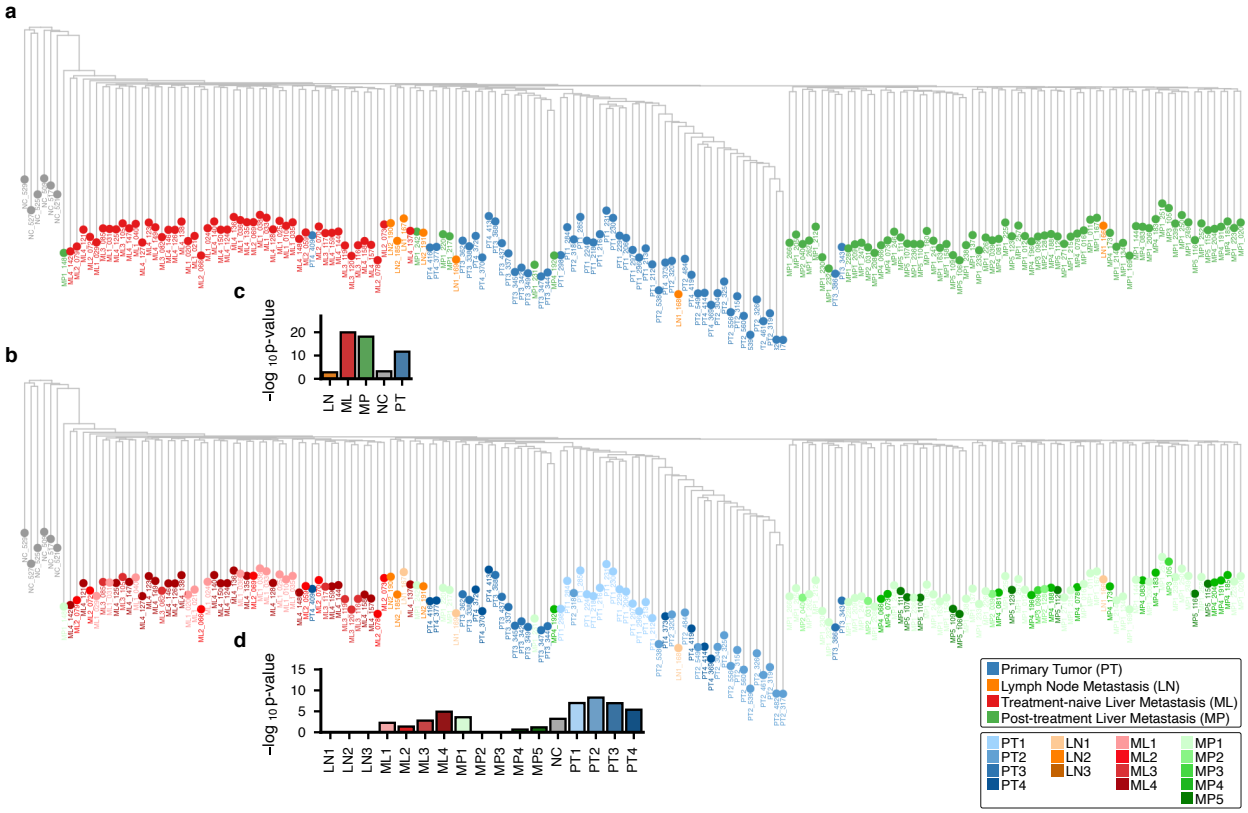

Figure S2: **(a)** Phylogeny reconstructed for patient CRC01 from Bian et al. [S3], following biclustering with $\alpha = 0.95$ and $\beta = 0.1$ with nodes colored according to the tissue origin of cells. **(b)** Phylogeny reconstructed following biclustering with $\alpha = 0.95$ and $\beta = 0.1$ with nodes colored according to the (more detailed) sampling locations of cells. **(c)** P-values for each tissue type were calculated using one-tailed binomial test. **(d)** P-values for each sampling location were calculated using one-tailed binomial test. (See Table S1 for p-values of these phylogenies.)

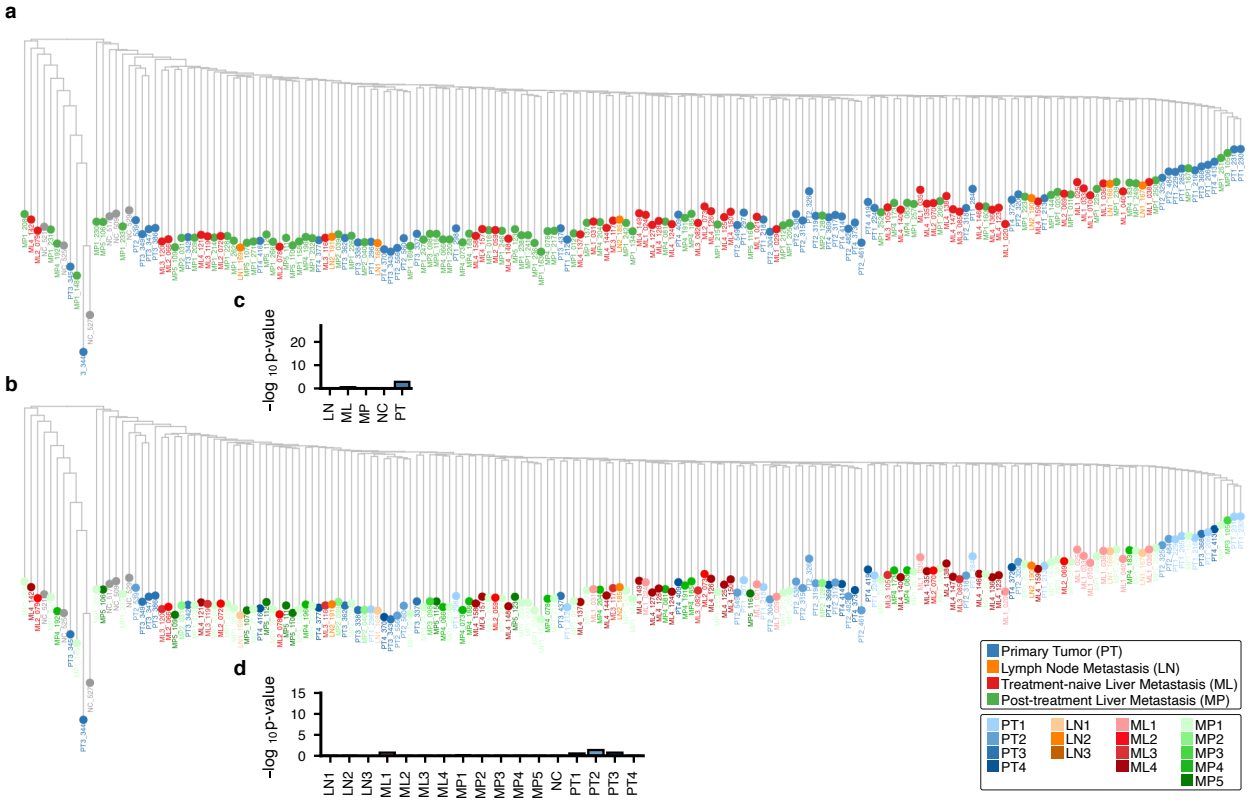

Figure S3: **(a)** Normalized L_1 distance-based NJ tree for patient CRC01 from Bian et al. [S3], from the binary matrix used to perform biclustering with $\alpha = 0.95$ and $\beta = 0.5$ with nodes colored according to the sampling locations of cells. **(b)** Normalized L_1 distance-based NJ tree from the binary matrix used to perform biclustering with $\alpha = 0.95$ and $\beta = 0.5$ with nodes colored according to the tissue origin of cells. **(c)** P-values for each tissue type were calculated using one-tailed binomial test. **(d)** P-values for each sampling location were calculated using one-tailed binomial test.(See Table S1 for p-values of these phylogenies.)

Tissue	Figure S1(a)	Figure 3(b)	Figure S2(a)	Figure S3(a)
MP	3.9855e-21	**9.8321e-27**	7.7612e-19	7.6252e-01
PT	1.8280e-11	**1.5850e-22**	2.2015e-12	1.4880e-03
ML	2.8703e-9	**4.3892e-27**	1.1103e-20	2.8869e-01
LN	3.2514e-02	**1.4487e-06**	1.6388e-03	1.0000e+00
NC	1.4168e-02	**6.1415e-04**	**6.1415e-04**	1.0000e+00

(a) Phylogenies specifying tissue origins of cells

Sampling location	Figure S1(b)	Figure 3(c)	Figure S2(b)	Figure S3(b)
LN1	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00
LN2	3.4149e-01	**4.1289e-06**	1.0000e+00	1.0000e+00
LN3	1.0000e+00	N/A	N/A	N/A
ML1	1.0000e+00	**3.3964e-09**	5.5256e-03	1.7755e-01
ML2	**2.3504e-02**	1.0000e+00	4.3148e-02	1.0000e+00
ML3	1.0000e+00	2.5954e-02	**1.6388e-03**	1.0000e+00
ML4	2.0405e-04	8.5230e-05	**1.2085e-05**	1.0000e+00
MP1	**1.6190e-11**	5.7820e-06	2.5923e-04	8.7996e-01
MP2	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00
MP3	1.0000e+00	1.0000e+00	1.0000e+00	1.0000e+00
MP4	2.5954e-02	**1.3004e-03**	2.2742e-01	1.0000e+00
MP5	1.0000e+00	6.6562e-02	6.6562e-02	1.0000e+00
PT1	**2.0165e-10**	1.9625e-06	1.0442e-07	2.8247e-01
PT2	2.5758e-08	**5.7544e-12**	5.2264e-09	4.2310e-02
PT3	1.3068e-05	**4.8761e-15**	1.1257e-07	1.7755e-01
PT4	1.1365e-01	**1.9124e-13**	4.0718e-06	1.0000e+00
NC	1.4168e-02	**6.1415e-04**	**6.1415e-04**	1.0000e+00

(b) Phylogenies specifying sampling locations of cells within a tissue

Table S1: The p-values for each methylation phylogeny in Figure S1,3,S2, and S3, calculated using the one-tailed binomial test. Each cell in the tree is labeled as "success" if all cells in the subtree of its direct parent have the same cancer type. Otherwise it is labled as "failure". Note that here the null hypothesis assumes the probability of success for a particular tissue/sampling location (in a Bernoulli experiment) to be its proportion of cells among the entire collection of cells. Also note that there are no LN3 cells in the phylogenies in Figure 3,S2, and S3 as cells from this sampling location all have too small numbers of total non-zero read coverage to be included after the biclustering step of our computational framework. Highlighted values represent the most significant p-value (< 0.05) in a given row.

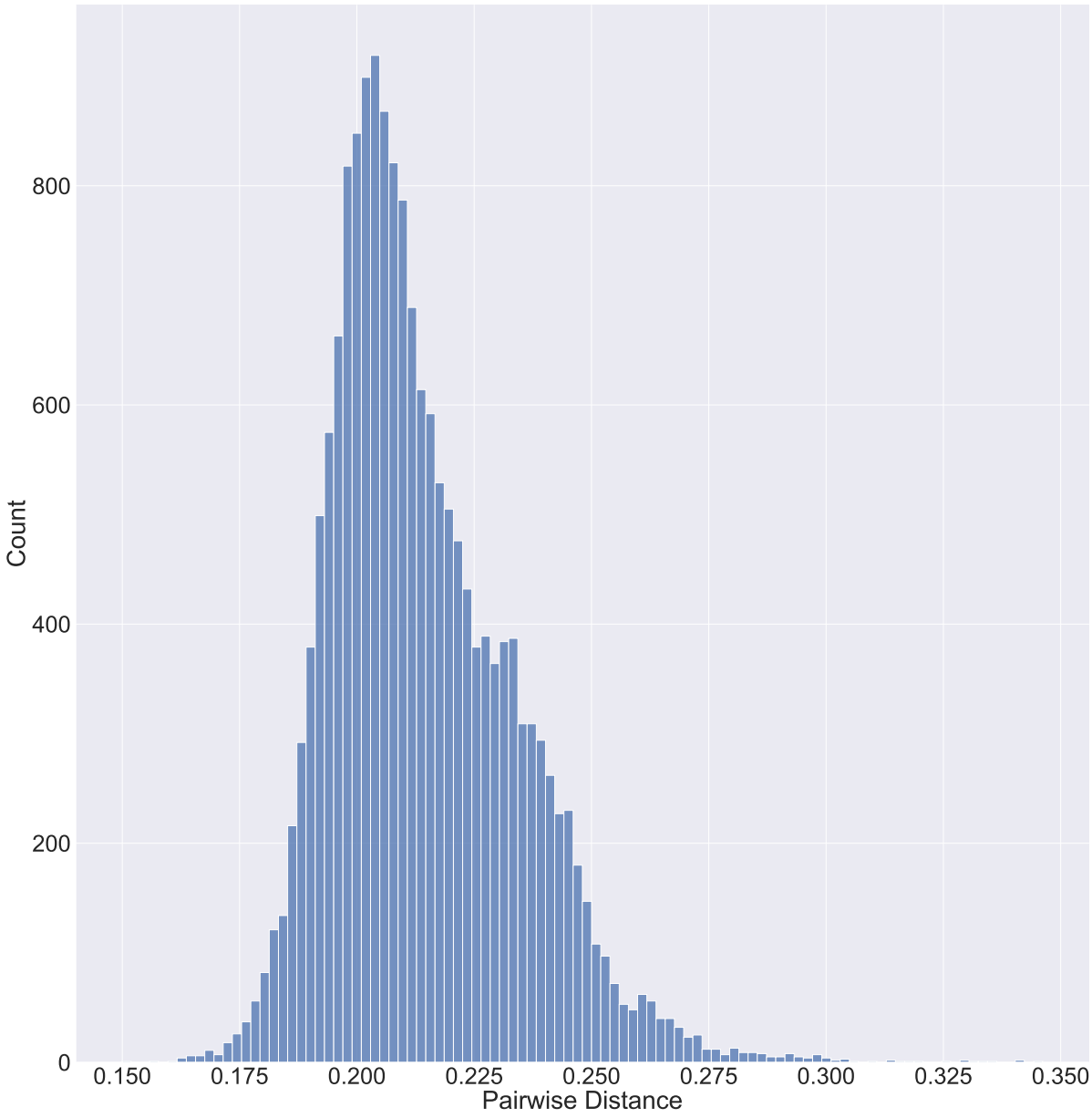

Figure S4: Distribution of pairwise expected distances used to construct the methylation phylogeny in Figure 3.