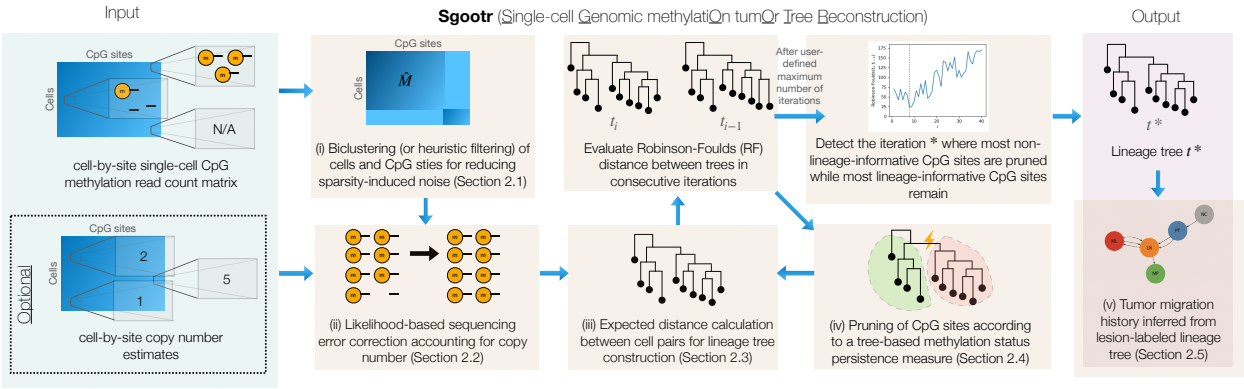
Single-cell methylation sequencing data reveal succinct metastatic migration histories and tumor progression models

Yuelin Liu[1,4,6⋆], Xuan Cindy Li[1,3⋆], Farid Rashidi Mehrabadi[1,7], Alejandro A. Schäffer[1], Drew Pratt[2], David R. Crawford[1,5], Salem Malikić[1], Erin K. Molloy[4], Vishaka Gopalan[1], Stephen M. Mount[5], 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] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, 20742, USA
[6] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, 20742, USA
[7] Department of Computer Science, Indiana University, Bloomington, IN, 47408, USA

Abstract. Recent studies exploring the impact of methylation in tumor evolution suggest that while the methylation status of many of the CpG sites are preserved across distinct lineages, others are altered as the cancer progresses. Since changes in methylation status of a CpG site may be retained in mitosis, they could be used to infer the progression history of a tumor via single-cell lineage tree reconstruction. In this work, we introduce the first principled distance-based computational method, `Sgootr`, for inferring a tumor's single-cell methylation lineage tree and jointly identifying lineage-informative CpG sites which harbor changes in methylation status that are retained along the lineage. We apply `Sgootr` on the single-cell bisulfite-treated whole genome sequencing data of multiregionally-sampled tumor cells from 9 metastatic colorectal cancer patients made available by Bian *et al.*, as well as multiregionally-sampled single-cell reduced-representation bisulfite sequencing data from a glioblastoma patient made available by Chaligne *et al.*. We demonstrate that the tumor lineages constructed reveal a simple model underlying colorectal tumor progression and metastatic seeding. A comparison of `Sgootr` against alternative approaches shows that `Sgootr` can construct lineage trees with fewer migration events and more in concordance with the sequential-progression model of tumor evolution, in time a fraction of that used in prior studies. Interestingly, lineage-informative CpG sites identified by `Sgootr` are in inter-CpG island (CGI) regions, as opposed to CGI's, which have been the main regions of interest in genomic methylation-related analyses. `Sgootr` is implemented as a `Snakemake` workflow, available at `https://github.com/algo-cancer/Sgootr`.



⋆ these authors contributed equally to this work
⋆⋆ corresponding author: cenk.sahinalp@nih.gov

1 Introduction

The recent rise of single-cell sequencing technology empowers more accurate tumor lineage inference by allowing the examination of intratumor heterogeneity at a cellular resolution. However, since single-cell sequencing data is derived from an incredibly limited amount of genetic material, the signals obtained are more scarce and unstable than those from bulk sequencing [23]. DNA methylation, the addition of a methyl group to cytosine, which results in the formation of 5-methylcytosine (5mC) especially in the context of CpG sites, is an epigenetic marker that has been extensively studied for its role in regulating gene expression and maintaining cellular memory [20]. Prior research suggests the changes in methylation status at CpG sites in cancer cells may provide a greater amount of observable evidence for tumor evolution than single-nucleotide variations (SNVs) or copy number aberrations (CNAs) [38, 3].

Recent studies have leveraged CpG methylation in single cells sampled from colorectal cancer (CRC) [2], chronic lymphocitic leukemia (CLL) [14], glioblastoma (GBM) and IDH-mutant glioma [5] tumors to construct lineages, showing CpG methylation to be a valuable signal for lineage inference. However, further examination reveals that data obtained via single-cell bisulfite sequencing (scBS-seq) [2], single-cell reduced representation bisulfite sequencing (scRRBS)[15, 16], or multiplexed scRRBS (MscRRBS) [14] protocols exhibit high level of sparsity. Not only that cells with subpar bisulfite conversion rates have information at few CpG sites, of the millions of unique CpG sites across all cells in a dataset, less than one out of a hundred are sequenced in a sufficient number of cells to be useful in lineage reconstruction; furthermore, even when a CpG site is sequenced in a cell, it is oftentimes covered by less than a few sequencing reads, which is not likely to capture heterozygosity in aneuploid cancers.

Besides sparsity, another key challenge is that not all CpG sites have their methylation statuses stably retained during in tumor evolution. In particular, Meir *et al.* posited two models of methylation dynamics: *mixture*, where the methylation status of a CpG site is resampled (from the parental status) in each cell replication, and *persistency*, where that is an exact copy of the parent cell [28]. It is the CpG sites whose status follows the persistency model would be informative in tumor lineage reconstruction because the maintenance of information from the parental generation is the necessary condition for the infinite sites assumption [21, 27] and Dollo's parsimony [8], which form the basis for tumor lineage inference tools based on mutation profiles. However, as of now, there is no known set of CpG sites that follow either inheritance model, highlighting the necessity to perform intentional CpG site selection while reconstructing tumor lineages by CpG methylation.

In this work, we introduce `Sgootr` (Single-cell Genomic methylatiOn tumOr Tree Reconstruction), the first distance-based computational method to jointly select informative CpG sites and reconstruct tumor lineages from single-cell methylation data. Application of `Sgootr` to a multiregionally-sampled metastatic CRC scBS-seq dataset[2] and a multiregionally-sampled GBM MscRRBS patient sample [5] reveals tumor progression models and metastatic migration histories simpler than previously reported. A comparison of `Sgootr` against alternative lineage-reconstruction and site-selection approaches on the same datasets shows that it infers a simpler migration history in a shorter amount of time. Interestingly lineage-informative CpG sites identified by `Sgootr` appear to be primarily in inter-CpG island (CGI) regions, as opposed to CGI's which have been the main regions of interest in genomic methylation-related analyses.

2 Methods

`Sgootr` consists of five key components: (i) biclustering of cells and sites for reducing sparsity-induced noise (Section 2.1); (ii) likelihood-based sequencing error correction accounting for copy number variation (Section 2.2); (iii) expected distance calculation between cell pairs for tree construction (Section 2.3); (iv) pruning of CpG sites according to a tree-based methylation status persistence measure (Section 2.4); and (v) inference of migration history from the lesion-labeled tree (Section 2.5). Note that components (iii) and (iv) are iteratively applied. Code for this work is available on GitHub as a `Snakemake` [29] workflow. Results from this work are reproducible, and intermediate experiment outputs of `Sgootr` are available for further analysis.

2.1 Biclustering of cells and sites for reducing sparsity-induced noise

`Sgootr` requires a fairly accurate estimate of the differences between the methylation status of distinct CpG sites across all cell pairs. This necessitates filtering out those cells for which the number of CpG sites with

non-zero read coverage is low and CpG sites with non-zero coverage in only a few cells. Ideally, we would like to coordinate the reductions in cells and CpG sites so as to minimize information loss, which can be achieved via an integer linear program (ILP)-based biclustering formulation as follows.

Let the site coverage data be represented in a cell-by-site matrix $M_{v\times u}$, where v is the number of cells and u is the number of CpG sites; $m_{ij} = 1$ if site j is covered (by at least one read) in cell i, and $m_{ij} = -1$ otherwise. Let α and β respectively be the specified fraction of cells and sites to be kept. Given M, α, and β, we wish to compute a biclustering - namely, a selection of rows of columns - of M, $\hat{M}$, so that we maximize the number of CpG sites covered in the resulting $\lfloor \alpha v \rfloor \times \lfloor \beta u \rfloor$ submatrix in which rows and columns respectively correspond to the selected cells and sites.

To solve this biclustering problem, let $C \in \{0,1\}^v, S \in \{0,1\}^u$ be (unknown) binary vectors respectively indicating whether a cell or site is kept. In addition, let $A \in \{0,1\}^{v\times u}$ denote (unknown) binary matrix such that $a_{ij} = 1$ if and only if $c_i = 1$ and $s_j = 1$ (i.e., if cell i and site j are both kept). The ILP formulation is:

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

The output submatrix $\hat{M}$ is obtained by taking cells $\{i|c_i = 1\}$ and sites $\{j|s_j = 1\}$ from M.

Our implementation of the above formulation employs the Gurobi ILP solver, a licensed software which is free to use for academic purposes [17]. The performance of our implementation with respect run time, memory and the accuracy (i.e. optimality gap) with respect to the input size and other parameters, as well as a strategy on how to set α and β are explored in detail in Supplementary Section S1. For problem instances in which our implementation does not lead to a near-optimal solution within reasonable time/memory (with respect to the optimality gap), or if the Gurobi ILP solver is unavailable, we apply the following heuristic filters to the cells and sites in the dataset at hand. (i) Among all input cells, remove low-quality ones with coverage in fewer than 4 million CpG sites; (ii) remove CpG sites covered in less than $\frac{2}{3}$ of the remaining cells. In addition to the (ILP based or heuristic) biclustering, it is advisable to remove CpG sites on sex chromosomes, which may confound the findings, and chromosome 21 peri-centromeric regions (chr21:9825000-9828000), which we discovered to be an overlooked alignment artifact.

After filtering the input for reducing sparsity, we perform sequencing error correction as described in Section 2.2. Since this may further introduce sparsity, in an additional post error correction step, we may remove sites covering $< \frac{2}{3}$ of cells, and then remove cells with coverage in $< \frac{2}{3}$ of the remaining CpG sites. Overall, our filtering approach aims to make sure that the cells remaining are well-covered, and that each cell pair have at least $\frac{1}{3}$ of the selected CpG sites as shared dimensions along which cell-to-cell distance can be measured.

2.2 Likelihood-based sequencing error correction accounting for copy number

Errors in sequencing may provide false evidence for an underlying allele methylation status different from the truth. We present a maximum log-likelihood approach to correct likely errors in methylation reads, incorporating site copy number estimates in each cell (Algorithm 1). For a CpG site in a cell, given $n \ge 0$ reads with a methylated status, $m \ge 0$ reads with an unmethylated status, its copy number c, and sequencing error probability $0 \le \epsilon \le 1$ (which we set to .01 in our experiments to proxy the per-base sequencing error rate in Illumina sequencing instruments), we can enumerate the likelihood of drawing the observed reads from all possible underlying allele statuses, letting $0 \le \gamma \le c$ be the number of alleles with the CpG site methylated. Note that in Algorithm 1, 1^c and 0^c respectively denote the events that all c alleles are methylated and unmethylated; $1^\gamma 0^{c-\gamma}$ denotes the event where γ alleles are methylated and $c - \gamma$ are not.

Sequencing error correction takes place when we have $n > 0$ and $m > 0$, but the allele status with the highest log-likelihood is homozygous. In that case, we correct the reads with the minority methylation status

Algorithm 1 Sequencing error correction accounting for copy number

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

function SequencingErrorCorrection(n, m, c, ϵ)
    if c = 0 then return None                          ▷ site is discarded in case of deletion event
    L[1^c], L[0^c] ← ln[(1 − ϵ)^n ϵ^m], 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, m ← n + m, 0
        else if n < m then m, n ← n + m, 0
        else return None                               ▷ site is discarded if correction is impossible
    end if
    return n, m
end function
```

to the majority one. 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. 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). Sex chromosomes are always excluded from our analysis. We use the sequence error-corrected reads for pairwise distance estimation between cells, as described in the following section.

2.3 Expected distance calculation between cell pairs for tree construction

In this section, we present a formulation for computing the expected distance between two cells given their respective copy number and (sequencing error-corrected) reads. We consider such a formulation with the goal of correcting for the potential bias contributed by low coverage CpG sites. Intuitively, this distance measure serves as an estimate for the expected proportion of methylation status altered alleles across all CpG sites between a cell pair.

Given a cell, let the copy number at a CpG site be c, and we have $0 \leq \gamma \leq c$ alleles with the CpG site methylated, and $c - \gamma$ alleles with the CpG site unmethylated. To allow modeling preferential allele sampling based on site methylation status as previously observed [31], we additionally introduce parameter p, probability of drawing a read from the allele with the CpG site methylated given a pair of alleles heterozygous for the site. We set p to the uninformative .5 in our experiments (see Supplementary Section S6 for when to choose alternative values). Then, we can compute the probability of drawing from an allele with the CpG site methylated by normalizing over all alleles (Equation 1):

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

It follows that the probability of drawing from an allele with the CpG site unmethylated is $1 - p_{c,\gamma}$.

Consider the case where we observe n reads at a CpG site for a cell, and all n reads are methylated for the site. Let the copy number at the site for the cell be c, the probability of sampling all n methylated reads from c alleles with the CpG site methylated (i.e. $\gamma = c$) is: $\mathrm{P}(\text{reads} \mid 1^c) = 1$; here 1^c is the event that in all c copies the CpG site is methylated, and reads is the event that all n sampled reads are methylated. 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;

and when there is a copy number loss (i.e. $c = 1$), the formulation assigns only non-zero probability to a homozygous combined allele status:

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

Given user-defined prior probabilities $\mathrm{P}(1^c)$, $\mathrm{P}(0^c)$, and $\mathrm{P}(\text{mixed})$, which we all set to the uninformative .33 in our experiments (see Supplementary Section S6 for when to choose alternative values), 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 likelihood of any allele status given the observed reads (Equations 2):

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

In the case where the n reads all have the CpG site unmethylated, the three values can be computed similarly with $\mathrm{P}(\text{reads} \mid 1^c) = 0$ and $\mathrm{P}(\text{reads} \mid 0^c) = 1$. In the case we observe among the n reads both ones with the CpG site methylated and ones with the site unmethylated, we have $\mathrm{P}(\text{mixed} \mid \text{reads})=1$, and $\mathrm{P}(1^c \mid \text{reads}) = \mathrm{P}(0^c \mid \text{reads}) = 0$. 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 over the possible combinations of allele status between the two cells at s (Equation 3):

$$\mathrm{dist}(\mathrm{reads}_{\mathrm{A},s}, c_{\mathrm{A},s}, \mathrm{reads}_{\mathrm{B},s}, c_{\mathrm{B},s}) = \sum_{\substack{\mathrm{status}_{\mathrm{A},s} \\ \in\{1^{c_\mathrm{A}}, \\ \text{mixed}, \\ 0^{c_\mathrm{A}}\}}} \sum_{\substack{\mathrm{status}_{\mathrm{B},s} \\ \in\{1^{c_\mathrm{B}}, \\ \text{mixed}, \\ 0^{c_\mathrm{B}}\}}} \mathrm{P}(\mathrm{status}_{\mathrm{A},s} \mid \mathrm{reads}_{\mathrm{A},s})\mathrm{P}(\mathrm{status}_{\mathrm{B},s} \mid \mathrm{reads}_{\mathrm{B},s})\mathrm{dist}(\mathrm{status}_{\mathrm{A},s}, \mathrm{status}_{\mathrm{B},s}) \tag{3}$$

where dist(11,11)=dist(10,10)=dist(00,00)=0, dist(11,10)=dist(10,11)=dist(00,10)=dist(10,00)=0.5, and dist(11,00)=dist(00,11)=1.

The total expected distance between cells 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 4. The use of shared CpG sites between pairs of single cells to estimate their distances was established by Hui *et al.* [19]:

$$\mathrm{dist}(\mathrm{A},\mathrm{B}) = \frac{\sum_{s \in \mathrm{sites}_\mathrm{A} \cap \mathrm{sites}_\mathrm{B}} \mathrm{dist}(\mathrm{reads}_{\mathrm{A},s}, c_{\mathrm{A},s}, \mathrm{reads}_{\mathrm{B},s}, c_{\mathrm{B},s})}{|\mathrm{sites}_\mathrm{A} \cap \mathrm{sites}_\mathrm{B}|} \tag{4}$$

A comparison of our distance formulation against a baseline distance measure commonly used in prior studies [5, 14] on simulated data can be found in Supplementary Section S2.1. After computing the distance between each pair of cells, one can incorporate any distance-based tree reconstruction method into `Sgootr` to obtain the lineage tree and arbitrarily choose a single cell from the normal tissue as the root. We use the `scikit-bio` implementation of the neighbor-joining (NJ) algorithm [35] for our main analysis. It is possible to replace NJ with an alternative distance-based tree construction method in `Sgootr`. We compare `Sgootr`'s performance while using FastME 2.0 [25], a popular alternative, against that using NJ in Supplementary Section S4 and show that they lead to similar results.

2.4 Pruning of CpG sites according to a tree-based methylation status persistence measure

The main body of `Sgootr` consists of an iterative procedure: at each iteration, it (i) computes the pairwise distances among single cells to form the tumor lineage tree using a distance-based algorithm, then (ii)

measures the methylation persistence score of each CpG site along the tumor lineage tree and prunes away a fraction of CpG sites that have the least scores before continuing onto the next iteration, and (iii) outputs the tumor lineage tree of the particular iteration whose distance from the tumor lineage trees obtained in consecutive iterations are minimum possible.

We have described how to compute pairwise distance among single cells in Section 2.3, and we use the widely used Robinson-Foulds (RF) distance [34, 37] to measure the differences among the tumor lineage trees constructed on the same set of cells in consecutive iterations. In this section, we focus on how we measure the methylation persistence of CpG sites at each iteration given the tumor lineage tree obtained for these CpG sites. In particular, we define the *methylation persistence score* for a CpG site given a tumor lineage tree (Equation 6): the higher the methylation persistence score for a CpG site, the more stably maintained its methylation status change along a tumor lineage tree.

To facilitate the analysis, we assign to each CpG site covered in each cell its most likely methylation status: given sequencing error-corrected reads from component (ii) of Sgootr, we call a site 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 remains unknown if there are no reads covering the CpG site. After this step, we measure the persistence of each CpG site independently: first at each branch of the tree, then for the overall tree.

Methylation persistence score for a CpG site at a particular branch in the lineage tree. Each branch in the methylation tumor lineage tree induces a bipartition of the tree and subsequently of the leaf nodes, which represent single cells. In other words, let TN denote the full set of leaf nodes in a methylation tumor lineage tree t, cutting a branch b in the tree creates disjoint subsets of leaf nodes TN_b and $\overline{TN_b}$. Let $Q_{s,tn}$ denote the probability distribution across three possible methylation statuses - homozygously methylated, heterozygously methylated, and homozygously unmethylated - at CpG site s across a subset of cells tn, and let I_s represent the set of cells with status information at site s. Then, we define $\mathcal{MP}_{s,b}$, the methylation persistence score of CpG site s at branch b (Equation 5):

$$\mathcal{MP}_{s,b} = \sqrt{JSD\big(Q_{s,TN_b \cap I_s}, Q_{s,\overline{TN}_b \cap I_s}\big)}. \tag{5}$$

Here $JSD(X, Y)$ denotes the Jensen-Shannon divergence [39, 26] for a pair of distributions X, Y, and is defined as $JSD(X, Y) = \frac{1}{2}D(X\|Z) + \frac{1}{2}D(Y\|Z)$ with $Z = \frac{1}{2}(X + Y)$ and $D(K\|L)$ denoting the Kullback-Leibler divergence measure for any arbitrary distributions K and L [22]. It is worth noting that the square root of the Jensen-Shannon divergence measure, which computes $\mathcal{MP}_{s,b}$ (Equation 5), is metric [32, 12] and is commonly referred to as Jensen-Shannon distance in popular implementations.

The intuition is that, if the change in methylation status of CpG site s has an observed persistent effect in tumor progression - namely, s is lineage-informative - there exists a branch b^* in the methylation tumor lineage tree such that the leaf nodes in the two subtrees induced by b^* show very different distributions of methylation statuses. In other words, $\mathcal{MP}_{s,b^*}$ will be large for such b^*. In contrast, suppose the bipartitions contain the same distributions for methylation statuses for s - either that most cells share the same status, or that both bipartitions are similarly heterogeneous - the score will be low.

Overall methylation persistence score for a CpG site in the lineage tree. We define the overall methylation persistence score of CpG site s in methylation tumor lineage tree t as the maximum methylation persistence score the site s has across all valid bipartitions (Equation 6). We recognize that (i) an extreme difference in the number of cells between the bipartitions, or (ii) a severe lack of cells with status information could both lead to meaningless divergence measurements. Therefore, we only consider bipartitions induced by branch b such that (i) both partitions contain no fewer than a user-defined fraction δ of total number of leaves, and (ii) there is read information in no fewer than a user-defined fraction ω of cells in both partitions. In our experiments, we set $\delta = .05$ and $\omega = .5$.

$$\mathcal{MP}_{s,T} = \max_{\substack{b \in \{b \mid |TN_b| \geq \delta|TN| \wedge |\overline{TN_b}| \geq \delta|TN| \\ \wedge |TN_b \cap I_s| \geq \omega|TN_b| \wedge |\overline{TN_b} \cap I_s| \geq \omega|\overline{TN_b}|\}}} \mathcal{MP}_{s,b} \tag{6}$$

Algorithm 2 Partial order-guided modification to the Fitch's top-down node label assignment

```
label_set(node) ← maps a node to the set of labels it is given by the bottom-up phase of Fitch's algorithm
parent(node), children(node) ← returns the parent and the two children of node, respectively
poset_minimal(labels) ← returns the minimal element in subset labels of a poset L, breaking ties arbitrarily
assignment[node] ← records the final label assignments of all nodes
node ← a node in the tree

function LabelAssignment(node)
    candidate_labels ← label_set(node)
    if node is not root AND assignment[parent(node)] is in candidate_labels then
        assignment[node] ← assignment[parent(node)]
    else
        assignment[node] ← poset_minimal(candidate_labels)
    end if
    if node is not a leaf then
        child1, child2 ← children(node)
        LabelAssignment(child1)
        LabelAssignment(child2)
    end if
end function
```

Iterative joint tumor lineage tree reconstruction and lineage-informative site identification. In a given iteration k, `Sgootr` first computes the tumor lineage tree t_k with the distances between pairs of cells (Section 2.3) based on the persistent sites identified in iteration $k-1$ (for iteration 0 the entire set of sites remaining after component (i) and (ii) of `Sgootr` are used). Then, for each CpG site s used in computing t_k, `Sgootr` calculates its overall methylation persistence score $\mathcal{MP}_{s,t_k}$. Among the CpG sites with overall methylation persistence scores, `Sgootr` prunes away a user-defined fraction κ of the CpG sites with the lowest scores, along with those with tying scores at the threshold. It outputs the remaining CpG sites to be used in iteration $k+1$. The process continues for a user-defined maximum number of iterations, and `Sgootr` outputs $t^* = t_i$ where i is the last iteration with $RF(t_i, t_{i-1})$ equals the global minimum across iterations.

Intuitively, we would like to detect an approximate point in the iterative procedure where most non-lineage-informative CpG sites have been pruned out (leading to the initial roughly decreasing trend of RF distance) but most lineage-informative CpG sites still remain (whose further elimination will lead to increasingly inaccurate distance measurements between cell pairs and therefore increasingly unstable tree topologies, which is reflected in a once-again increasing trend of RF distance). We demonstrate empirical observations corroborate with such intuition (Figure 1c.,S9,S15) and provide practical recommendations for choosing κ and the maximum number of iterations in Supplementary Section S7.

2.5 Inference of migration history from lesion-labeled tumor lineage tree

Given a rooted tumor lineage tree t^* produced by `Sgootr` through the procedures described above, provided that the single cells in the tree are labeled by their lesion of origin, we can infer the underlying *tumor migration history*, which we represent as a directed multigraph (without self loops) where each vertex represents a distinct lesion and each edge represents a distinct migration event from the source lesion to the target lesion. We accomplish this by first applying the well-known Fitch's algorithm [13] with slight modification to obtain a unique, maximally parsimonious labeling of the internal nodes of t^*, then identifying the migration events.

One may recall that, the bottom-up phase of the Fitch's algorithm computes in a tree the number of migration events - namely, the change of lesion of origin labels from a parent node to a child - in the most parsimonious labeling of its internal nodes. This phase also produces a set of possible labels for each internal node in the tree. The top-down phase, then, assigns each internal node a label from its set, as described in Algorithm 2, starting with root node of t^*.

The most general formulation of Fitch's algorithm may produce multiple optimal labelings of the internal nodes; however, some of these optimal labelings may adhere to our prior understanding of the sequential-progression (a.k.a. metastatic-cascade) model of tumor evolution [24] better than others. One can encode

such prior understanding as a partial order over the lesion of origin labels of single cells. For example, in this work, we leverage the following partial order:

$$\text{Normal Tissue} \prec \text{Primary Tumor} \prec \text{Lymph Nodes} \prec \text{Distant Metastases}.$$

In our modification to the Fitch's algorithm, during the top-down phase, when a node cannot inherit the label of its parent, we assign it the minimal element in its candidate label set according to the partial order, breaking ties arbitrarily when the elements in the label set cannot be compared (Algorithm 2). This returns a maximally parsimonious labeling of the internal nodes in the tree that also acknowledges the sequential-progression model.

With a fully labeled tree, we produce a directed multigraph that corresponds to the tumor migration history as follows. First, we generate a vertex for each distinct lesion represented in the tree. Then, as we traverse down the tree from the root, for each label change from a parent to a child in the tree, we add one direct edge to the multigraph, from the lesion vertex corresponding to the parent label to that corresponding to the child label. Parallel edges between a pair of vertices correspond to a polyclonal seeding event, and any back edge indicates a reseeding event.

3 Results

The performance of `Sgootr`'s distinct components on simulated data are described in the Supplementary Sections S1 (biclustering), S2.1 (distance calculation), and S2.2 (iterative pruning). As can be seen, each of these components substantially improve on the respective baseline approach on a wide range of settings. Furthermore, the results of the iterative pruning component demonstrate that `Sgootr` can capture the migration history of a tumor accurately.

In the remainder of the paper we focus on the application of `Sgootr` to the scBS-seq dataset generated by Bian *et al.* on 9 metastatic CRC patients[8], among whom 2 (CRC01 and CRC04) have matching scRNA-seq data (via the scTrio-seq2 protocol) from both normal and tumor cells with which copy number calling can be performed [2]. We highlight results from patient CRC01 since in addition to having copy number calls available, they have the largest number of distinct tissue types, sampling locations, and treatment conditions. Full details and results for the other 8 patients can be found in Supplementary Section S3. We additionally present in Supplementary Section S5 `Sgootr`'s results on a GBM patient by Chaligne *et al.*[5], which demonstrate the applicability of our approach to MscRRBS data. To the best of our knowledge, this is the only other single cell bisulfite sequencing dataset with multiregional tumor sampling that is publicly available.

`Sgootr` obtains a simple tumor migration history with scBS-seq and scRNA-seq data from patient CRC01. The single cells of CRC01 are sampled from 4 distinct lesions - primary tumor (PT), lymph node metastasis (LN), liver metastasis (ML), and post-treatment liver metastasis (MP) - as well as normal colon tissue (NC) adjacent to the primary lesion; furthermore, there are multiple sampling locations within each lesion (Figure 1a.). We only include cells with both scBS-seq and scRNA-seq data, and employ `inferCNV`[9] to leverage scRNA-seq data for copy number calling. We apply `Sgootr` with $\kappa = .1$, terminating the iterative procedure after 40 rounds. A unique global minimum among the RF distances is identified at iteration 8 (Figure 1c.), hence $t* = t_8$ (Figure 1b.) is output as the lineage tree. The tumor migration history inferred by `Sgootr` (Figure 1d.) is simpler than previously reported (see Figure S7 by Bian *et al.* [2]): NC grow into PT, followed by a polyclonal migration to LN, which appear to proceed to seed both ML and MP. While the migration history also suggests a polyclonal reseeding from ML to LN, both edges in the graph have low support: they are both due to a subtree of a particular cellular origin harboring a (potentially mislabeled) singular cell of a different origin.

`Sgootr` infers migration histories simpler than alternative methods for Bian *et al.* CRC cohort. We benchmark `Sgootr` against (i) a naïve baseline distance-based tree construction method (column "baseline" in Figure 2a.), (ii) IQ-TREE[30] with the two-state general time reversible model (`GTR2`), an instance

[8] Out of 12 total patients in the Bian *et al.* cohort, we excluded those without scBS-seq data (CRC03 and CRC06), and those without metastasis information (CRC09).

[9] from the Trinity CTAT Project `https://github.com/broadinstitute/inferCNV`

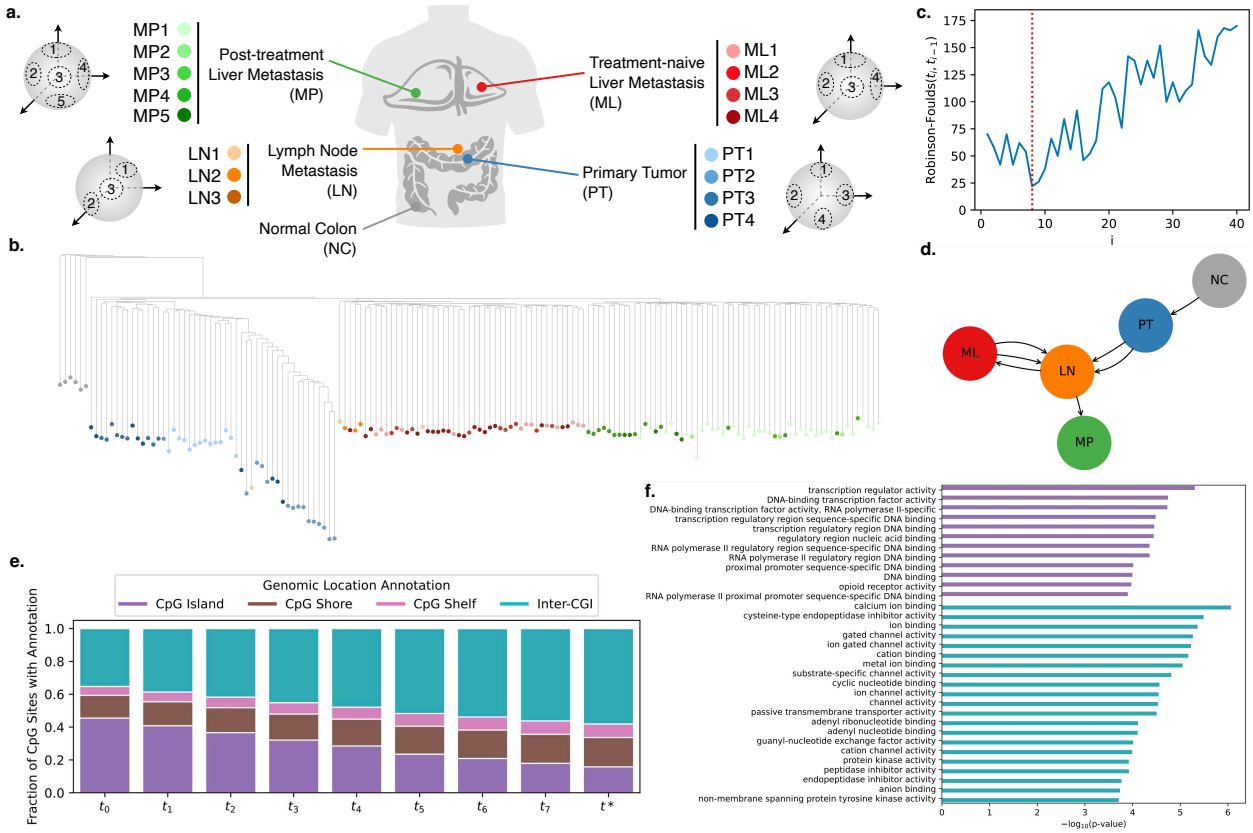


Fig. 1: **Application of `Sgootr` to patient CRC01 scTrio-seq data by Bian *et al.* a.** The multiregional patient data consists of single cells sampled from 4 distinct lesions (in addition to normal tissue) and multiple sampling locations within each lesion. **b.** Tumor lineage tree constructed by `Sgootr` with $\kappa = .1$. Each single cell is represented as a leaf, colored by its sampling location. **c.** RF distances between the trees of consequtive iterations of the pruning procedure, with global minimum occurring at $t* = t_8$. **d.** Tumor migration history inferred by `Sgootr` for patient CRC01. **e.** Fraction of CpG sites located in CpG island, CpG shore, CpG shelf, and inter-CGI regions in the CRC01 trees at each iteration of the pruning procedure of `Sgootr`, from t_0 to $t*$. **f.** GO terms with significant ($< .05$) q-values in enrichment analysis of non-pseudo, protein-coding genes spanning lineage-informative sites in CpG island and inter-CGI regions respectively in CRC01.

of the popular maximum-likelihood-based tool leveraged for inferring lineage trees from single-cell bisulfite sequencing data in prior studies [14, 5] (column "IQ-TREE" in Figure 2a.), and (iii) IQ-TREE (with `GTR2` model) preceded by *four-gamete analysis* (FG), a lineage-informative site-selection method previously described by Gaiti *et al.*, which returns a subset of input CpG sites which lower than expected epimutation rate [14] (column "FG+IQ-TREE" in Figure 2a.). Note that it is IQ-TREE preceded by FG that was employed exactly in previous studies [14, 5]. However, as discussed in more detail below, the FG step is not only very slow, it also leads to highly complex migration histories. These observations lead us to additionally benchmark against IQ-TREE without the FG step, even though it had not been directly applied to single-cell bisulfite sequencing data in prior studies. See Supplementary Section S8 for further experiment details.

For each benchmarking experiment, we take as input the cell-by-site read count matrices resulting from first-pass heuristic filtering step of `Sgootr`, namely, removing low-quality cells, CpG sites with coverage in less than $\frac{2}{3}$ of remaining cells, sites on sex chromosomes, and sites within chromosome 21 peri-centromeric regions. Then, for each CpG site covered in each cell, call whether it is methylated or not by the status receiving support from at least 90% of reads at the site in the cell, discarding entries where such determination cannot be made. This binarization step has been common practice in prior works on single-cell methylation data[2, 14, 5] and is necessary in generating input for the `GTR2` model. For the FG+IQ-TREE experiments, we further

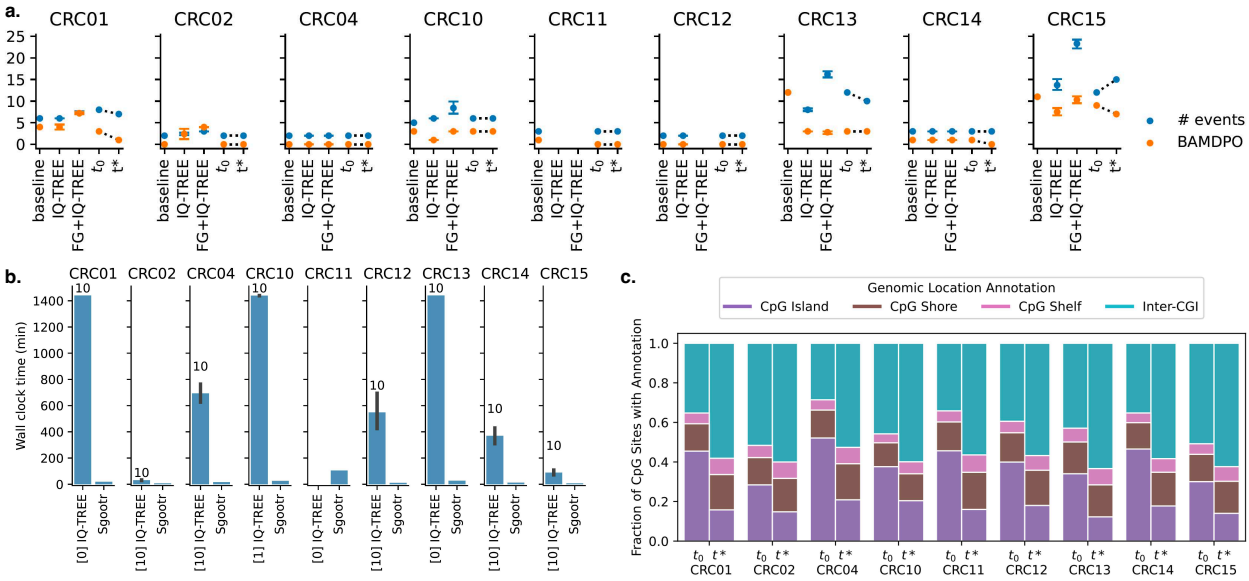


Fig. 2: **Overview of results from the Bian *et al.* metastatic CRC cohort. a.** Number of migration events and BAMDPO values of migration histories obtained via (i) naïve baseline method, (ii) IQ-TREE, (iii) IQ-TREE on FG-selected CpG sites, (iv) `Sgootr` intermediate tree prior to iterative procedure (t_0), and (v) final tree output by `Sgootr` ($t*$). IQ-TREE result is represented by the mean across 10 runs with different random seeds, with the error bars denoting the 95% confidence interval of the mean. The lack of data points in "IQ-TREE" columns means IQ-TREE fail to finish model optimization to start tree search after 24 hours, and that in "FG+IQ-TREE" columns means FG analysis fail to terminate after 100 hours. **b.** Run time comparison between `Sgootr` and IQ-TREE. Each experiment is performed on 4 cores. The number in brackets next to "IQ-TREE" on the x-axis represents the number of IQ-TREE runs (out of 10) that have converged within 24 hours. The number above each IQ-TREE bar represents the total number of runs (out of 10) with any tree search output (converged or intermediate after 24 hours), and hence contributes to the run time and migration history comparison. **c.** Fraction of CpG sites located in CpG island, CpG shore, CpG shelf, and inter-CGI regions, before (t_0) and after ($t*$) the iterative pruning component of `Sgootr`.

perform FG (Supplementary Section S8 for detail) to select CpG sites to input to IQ-TREE. Additionally, we further filter the input to the benchmarking experiments so that all experiments (Figure 2a.) on a particular patient act on the same set of input cells. This calibration is to ensure fair comparisons among results from different methods using the two performance measures we will introduce below. To facilitate fair runtime comparison between IQ-TREE and `Sgootr`, we also filter to ensure the input to IQ-TREE for each patient is the same set as those to `Sgootr`, so that the runtime measurements in Figure 2b. are taken on the same set of cells and CpG sites for each patient.

For each patient in the Bian *et al.* cohort, we generate a baseline lineage tree by first computing the Hamming distance between every pair of cells over their binary CpG site status vectors, averaging across the sites covered in both cells, then applying NJ to the distance matrix. As IQ-TREE is a stochastic method, we perform 10 separate runs of each IQ-TREE experiment with different random seeds, and for each instance report the intermediate tree at convergence or after 24 hours if convergence has yet to be achieved by then. If none of the 10 runs of IQ-TREE for a particular experiment have finished model optimization, started tree search, and produced some intermediate result, we do not report IQ-TREE result for that experiment. Similarly, if FG does not terminate in 100 hours, we do not report FG+IQ-TREE result for that experiment. We root the baseline and IQ-TREE trees with the same normal cells we used to root the trees by `Sgootr`, and we infer the tumor migration histories from them via `Sgootr`'s modified Fitch's algorithm.

We compare the migration histories inferred via `Sgootr` and the three alternative approaches by two measures: (i) the total number of migration events, namely the number of edges in our multigraph representation, and (ii) one we name *binary adjacency matrix distance from partial order* (BAMDPO), which is

the Hamming distance between the binary adjacency matrix of the simple directed acyclic graph induced by the lesion partial order (see Section 2.5) and that obtained by collapsing all parallel edges in the migration history multigraph. Intuitively, the first measure (output from Fitch's algorithm) captures the conventional definition of parsimony, and the latter measures the degree the inferred history deviates from the sequential-progression model of tumor evolution. Together, they offer complementary perspectives on the complexity of the inferred histories. For both measures, a lower value would correspond to a simpler history.

Figure 2a compares the tree $t*$ obtained by `Sgootr` against the alternatives with respect to the two measures. The figure also includes results of t_0 for each patient, the tree obtained by `Sgootr` prior to its iterative site pruning procedure. Comparisons of the measures for t_0 against those for $t*$ show `Sgootr`'s iterative procedure almost always reduces the complexity of the inferred migration history, provided such a reduction is at all possible. In the case of patient CRC15, even though the total number of migration events increases from t_0 to $t*$, BAMDPO decreases; this suggests what first appeared to be deviations from the simple sequential-progression model could potentially be resolved as polyclonal seeding events more in concordance with the model, once the tree structure becomes more refined through the iterative procedure. Compared to naïve baseline and the two IQ-TREE-based methods, `Sgootr` typically infers simpler migration histories; the only exception is patient CRC10, where IQ-TREE (without FG site-selection) infers a migration history with the lowest BAMDPO.

`Sgootr`-identified lineage-informative CpG sites construct lineages suggesting simpler migration histories than those identified via FG. Specifically, comparisons between the measures from running IQ-TREE on FG-selected sites and those from `Sgootr` show that `Sgootr` consistently outperforms the former in terms of the number of events and BAMDPO where further improvements were possible. This suggests the sites selected by `Sgootr` may be more meaningful for the purpose of lineage reconstruction than those by FG. FG is also time-consuming: for patient CRC11 and CRC12, FG fail to complete after 100 hours, and no experimental results were reported (Figure 2a.).

`Sgootr` is orders-of-magnitude faster than IQ-TREE. While `Sgootr` produce as simple, if not simpler migration history than IQ-TREE (without FG site-selection) in terms of the two measures in all but one case, `Sgootr` obtain the results in a fraction of time IQ-TREE took given the same input dimensions. Each IQ-TREE run was allotted 24 hours. If the program converge within 24 hours, we report the wall clock time elapsed; if the program did not converge within 24 hours but have produced some intermediate output from tree search, we report the program run time as 24 hours; and if IQ-TREE fail to produce any intermediate result from tree search within 24 hours, we do not include the elapsed time from that run. Within the allotted time, IQ-TREE produce converged results in 6 out of 9 patients, some (intermediate) result for 8 out of 9 patients, and is not able to finish model optimization and start tree search within 24 hour for any of the 10 runs for CRC11 (Figure 2b.). `Sgootr` is deterministic and obtains lineage trees suggesting simpler histories in a mere fraction of the time when given the same number of cores and memory as IQ-TREE, positioning itself as the better option in more time-frugal settings.

`Sgootr`-identified lineage-informative CpG sites are enriched in inter-CGI regions. In all 9 CRC patients in the Bian *et al.* cohort, as well as the multiregional GBM patient MGH105 in the Chaligne *et al.* cohort, the set of lineage-informative sites identified by `Sgootr` showed enrichment in inter-CGI regions (Figures 1e., 2c., S16f.). The median distances between CpG sites and their closest CpG islands also increase as the iterations approach $t*$ across all patients (Figure S12).

Genes associated with `Sgootr`-identified lineage-informative CpG sites are enriched for pan-cancer and CRC-related terms. To examine whether the lineage-informative sites discovered by `Sgootr` in the Bian *et al.* cohort have potential functional implications, we stratified lineage-informative sites by their genomic annotation with respect to CGI, for each group finding a set of non-pseudo, protein-coding genes whose gene bodies and 1000 bp upstream promoter regions contain the sites, and carried out Gene Ontology (GO)[1, 6] enrichment analyses respectively using GOrilla [11, 10]. All annotated non-pseudo, protein-coding genes in the genome were used as the background set. A hypergeometric test was used to measure the overrepresentation of the genes in associated GO terms, and the Benjamini-Hochberg FDR-adjusted p-values (q-values) are presented. In 5 out of 9 CRC patients, genes mapped to lineage-informative sites within CGI

are strongly associated with transcription regulation and DNA binding (Figure 1f., S10,), processes known to be linked to cancer onset[7]. Among informative sites mapped to inter-CGI regions, prominent enrichment in neural activities including ion transport and gated channel activities were observed across all 9 CRC patients (Figure 1f, S10), concurring with recent studies which support the striking cross-talk between the neural system and colorectal cancer development[40, 33]. Enrichment results using an alternative tool, DAVID[36, 18], are also included for comparison (Figure S11). While the GO terms discovered are not identical between the tools, the results highly concur as indicated by the overlap coefficients of the terms (Table S4).

4 Discussion

We present Sgootr, the first distance-based tool to jointly infer tumor lineage using single-cell methylation data and select for lineage-informative CpG sites. We show Sgootr is able to produce simpler tumor migration history, perform CpG site selection, and obtain results orders-of-magnitude faster than alternative methods in applications to real multiregionally sampled single-cell methylation data in CRC and GBM. Sgootr concomitantly tackles two key challenges in tumor migration history inference from single-cell methylation data: sparsity and diverse inheritance dynamics of CpG methylation. Sgootr provides a fast and reliable solution to leverage single-cell methylation profiles and elucidate the complex tumor epigenetic landscape.

On top of tumor lineage inference, Sgootr captures meaningful biological insights of lineage-informative CpG sites. In both cancer types included in our study, CRC and GBM, we discovered the enrichment of lineage-informative sites in inter-CGI regions across different patients. Interestingly, we also observed that the variability of the methylation status of a CpG site among all sequenced cells seems to increase with its distance from the closet CpG island (Table S5). While the underlying mechanisms of these phenomena are yet to be elucidated and beyond the scope of this study, a plausible explanation is that CpG sites in islands are protected by the reduced chromatin accessibility in the region marked by histone modifications[4], whereas inter-CGI sites harbors more stochastic changes due to malfunctioning of DNA methyltransferases and TET enzymes [9]. While CpG islands have attracted much attention in the studies of tumor epigenetic changes and more frequently covered in methylation arrays, our findings suggest that inter-CGI regions should not be overlooked especially in the context of tumor lineage tracing. We are looking to explore the prognostic potential of inter-CGI CpG sites in our subsequent studies.

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25–29 (2000). https://doi.org/10.1038/75556, https://doi.org/10.1038/75556
2. Bian, S., Hou, Y., Zhou, X., Li, X., Yong, J., Wang, Y., Wang, W., Yan, J., Hu, B., Guo, H., Wang, J., Gao, S., Mao, Y., Dong, J., Zhu, P., Xiu, D., Yan, L., Wen, L., Qiao, J., Tang, F., Fu, W.: Single-cell multiomics sequencing and analyses of human colorectal cancer. Science **362**(6418), 1060–1063 (Nov 2018). https://doi.org/10.1126/science.aao3791, https://doi.org/10.1126/science.aao3791
3. Biezuner, T., Spiro, A., Raz, O., Amir, S., Milo, L., Adar, R., Chapal-Ilani, N., Berman, V., Fried, Y., Ainbinder, E., Cohen, G., Barr, H.M., Halaban, R., Shapiro, E.: A generic, cost-effective, and scalable cell lineage analysis platform. Genome Res **26**(11), 1588–1599 (Nov 2016). https://doi.org/10.1101/gr.202903.115
4. Cedar, H., Bergman, Y.: Linking dna methylation and histone modification: patterns and paradigms. Nature Reviews Genetics **10**(5), 295–304 (2009)
5. Chaligne, R., Gaiti, F., Silverbush, D., Schiffman, J.S., Weisman, H.R., Kluegel, L., Gritsch, S., Deochand, S.D., Gonzalez Castro, L.N., Richman, A.R., Klughammer, J., Biancalani, T., Muus, C., Sheridan, C., Alonso, A., Izzo, F., Park, J., Rozenblatt-Rosen, O., Regev, A., Suvà, M.L., Landau, D.A.: Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. Nature Genetics **53**(10), 1469–1479 (2021). https://doi.org/10.1038/s41588-021-00927-7, https://doi.org/10.1038/s41588-021-00927-7
6. Consortium, T.G.O.: The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research **49**(D1), D325–D334 (12 2020). https://doi.org/10.1093/nar/gkaa1113, https://doi.org/10.1093/nar/gkaa1113
7. Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al.: The chromatin accessibility landscape of primary human cancers. Science **362**(6413), eaav1898 (2018)
8. Dollo, L.: The laws of evolution. Bull. Soc. Bel. Geol. Paleontol **7**, 164–166 (1893)
9. Du, J., Johnson, L.M., Jacobsen, S.E., Patel, D.J.: Dna methylation pathways and their crosstalk with histone methylation. Nature reviews Molecular cell biology **16**(9), 519–532 (2015)
10. Eden, E., Lipson, D., Yogev, S., Yakhini, Z.: Discovering motifs in ranked lists of dna sequences. PLoS computational biology **3**(3), e39 (2007)
11. Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z.: Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. BMC bioinformatics **10**(1), 1–7 (2009)
12. Endres, D., Schindelin, J.: A new metric for probability distributions. IEEE Transactions on Information Theory **49**(7), 1858– 1860 (Jul 2003). https://doi.org/10.1109/TIT.2003.813506
13. Fitch, W.M.: Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Systematic Biology **20**(4), 406–416 (12 1971). https://doi.org/10.1093/sysbio/20.4.406, https://doi.org/10.1093/sysbio/20.4.406
14. Gaiti, F., Chaligne, R., Gu, H., Brand, R.M., Kothen-Hill, S., Schulman, R.C., Grigorev, K., Risso, D., Kim, K.T., Pastore, A., Huang, K.Y., Alonso, A., Sheridan, C., Omans, N.D., Biederstedt, E., Clement, K., Wang, L., Felsenfeld, J.A., Bhavsar, E.B., Aryee, M.J., Allan, J.N., Furman, R., Gnirke, A., Wu, C.J., Meissner, A., Landau, D.A.: Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. Nature **569**(7757), 576–580 (May 2019). https://doi.org/10.1038/s41586-019-1198-z, https://doi.org/10.1038/s41586-019-1198-z
15. Guo, H., Zhu, P., Guo, F., Li, X., Wu, X., Fan, X., Wen, L., Tang, F.: Profiling dna methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. Nature Protocols **10**(5), 645–659 (2015). https://doi.org/10.1038/nprot.2015.039, https://doi.org/10.1038/nprot.2015.039
16. Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., Tang, F.: Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome research **23**(12), 2126–2135 (2013)
17. Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2022), https://www.gurobi.com
18. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc **4**(1), 44–57 (2009). https://doi.org/10.1038/nprot.2008.211
19. Hui, T., Cao, Q., Wegrzyn-Woltosz, J., O'Neill, K., Hammond, C.A., Knapp, D.J., Laks, E., Moksa, M., Aparicio, S., Eaves, C.J., Karsan, A., Hirst, M.: High-resolution single-cell dna methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. Stem Cell Reports **11**(2), 578–592 (2018). https://doi.org/https://doi.org/10.1016/j.stemcr.2018.07.003, https://www.sciencedirect.com/science/article/pii/S2213671118303084
20. Kim, M., Costello, J.: DNA methylation: an epigenetic mark of cellular memory. Experimental & Molecular Medicine **49**(4), e322–e322 (Apr 2017). https://doi.org/10.1038/emm.2017.10, http://www.nature.com/articles/emm201710

21. Kimura, M.: The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. Genetics **61**(38), 14254–14261 (2008)
22. Kullback, S., Leibler, R.A.: On Information and Sufficiency. The Annals of Mathematical Statistics **22**(1), 79 – 86 (1951). https://doi.org/10.1214/aoms/1177729694, https://doi.org/10.1214/aoms/1177729694
23. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B.d., Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.H., Lelieveldt, B.P.F., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J.d., Saliba, A.E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A.: Eleven grand challenges in single-cell data science. Genome Biology **21**(1), 31 (2020). https://doi.org/10.1186/s13059-020-1926-6, https://doi.org/10.1186/s13059-020-1926-6
24. Lambert, A.W., Pattabiraman, D.R., Weinberg, R.A.: Emerging biological principles of metastasis. Cell **168**(4), 670–691 (2017). https://doi.org/https://doi.org/10.1016/j.cell.2016.11.037, https://www.sciencedirect.com/science/article/pii/S0092867416316099
25. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. Molecular Biology and Evolution **32**(10), 2798–2800 (06 2015). https://doi.org/10.1093/molbev/msv150, https://doi.org/10.1093/molbev/msv150
26. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991). https://doi.org/10.1109/18.61115
27. Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Miller, W., Haussler, D.: The infinite sites model of genome evolution. Proceedings of the National Academy of Sciences **105**(38), 14254–14261 (2008)
28. Meir, Z., Mukamel, Z., Chomsky, E., Lifshitz, A., Tanay, A.: Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. Nature Genetics **52**(7), 709–718 (Jul 2020). https://doi.org/10.1038/s41588-020-0645-y, https://doi.org/10.1038/s41588-020-0645-y
29. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J.: Sustainable data analysis with snakemake. F1000Res **10**, 33 (2021). https://doi.org/10.12688/f1000research.29032.2
30. Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q.: IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution **32**(1), 268–274 (Nov 2014). https://doi.org/10.1093/molbev/msu300, https://doi.org/10.1093/molbev/msu300
31. Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R.V., Branco, M.R., Reik, W.: Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting dna methylation data. Genome Biology **19**(1), 33 (2018). https://doi.org/10.1186/s13059-018-1408-2, https://doi.org/10.1186/s13059-018-1408-2
32. Österreicher, F., Vajda, I.: A new class of metric divergences on probability spaces and its applicability in statistics. Annals of the Institute of Statistical Mathematics **55**, 639–653 (2003)
33. Rademakers, G., Massen, M., Koch, A., Draht, M.X., Buekers, N., Wouters, K.A., Vaes, N., De Meyer, T., Carvalho, B., Meijer, G.A., et al.: Identification of dna methylation markers for early detection of crc indicates a role for nervous system-related genes in crc. Clinical epigenetics **13**(1), 1–12 (2021)
34. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical biosciences **53**(1-2), 131–147 (1981)
35. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**(4), 406–425 (07 1987)
36. Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T., Chang, W.: David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res **50**(W1), W216–21 (Mar 2022). https://doi.org/10.1093/nar/gkac194
37. Sul, S.J., Williams, T.L.: An experimental analysis of robinson-foulds distance matrix algorithms. In: European Symposium on Algorithms. pp. 793–804. Springer (2008)
38. Ushijima, T., Watanabe, N., Okochi, E., Kaneda, A., Sugimura, T., Miyamoto, K.: Fidelity of the methylation pattern and its variation in the genome. Genome Res **13**(5), 868–874 (May 2003). https://doi.org/10.1101/gr.969603
39. Wong, A.K.C., You, M.: Entropy and distance of random graphs with application to structural pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-7**(5), 599–609 (1985). https://doi.org/10.1109/TPAMI.1985.4767707
40. Zhu, P., Lu, T., Chen, Z., Liu, B., Fan, D., Li, C., Wu, J., He, L., Zhu, X., Du, Y., et al.: 5-hydroxytryptamine produced by enteric serotonergic neurons initiates colorectal cancer stem cell self-renewal and tumorigenesis. Neuron (2022)

Supplementary Material

S1 Biclustering: benchmarking on simulated data and a case study on CRC01

As can be seen in Figure S2 (left plot) on the colorectal cancer patient CRC01 [2], the distribution of the number of sites covered by at least one read across cells is bimodal and a relatively small fraction of cells have very few sites covered. In contrast, the distribution of the number of cells with at least one read covering a site (right plot) is unimodal across the sites and only a relatively small fraction of sites are covered in a large number of cells. Provided the presence of those well-covered sites are distributed over the cells independently and identically (i.e. in i.i.d. fashion), the heuristic filtering described in Section 2.1 would work well. However, if the distribution is not i.i.d., fixing the output size, the submatrix obtained by ILP-based biclustering formulation would have higher density , improving the overall reliability of the tree constructed by `Sgootr`. Since, for given α and β values, the ILP formulation can only produce a submatrix no sparser than that produced by the heuristic filters, it is our preferred choice. Below we explore how the running time and memory usage of the ILP formulation vary with key problem parameters, and identify under what parameter settings of the problem instance it remains feasible.

S1.1 Running time and memory benchmark on simulated data

We first demonstrate how the expected running time (in other words, time the Gurobi solver takes to obtain an optimal solution) and memory behaviors of the Gurobi solver on our formulation vary as a function of the input size.

For the problem at hand, the input size is parameterized by v, the number of cells, and u, the number of CpG sites in the input matrix M. The number of variables (N_{var}), the number of constraints (N_{constr}), and the number of non-zero entries in the constraint matrix (N_{nz}) are some features of the Gurobi representation of the linear program that may influence its memory usage and running time; note that all three features are $\mathcal{O}(vu)$ (Table S1). Also note that Gurobi uses the sparse representation for the constraint matrix; as a result, even though the naive representation of the constraint matrix requires $\mathcal{O}(v^2u^2)$ space, the size of its representation by Gurobi is proportional to N_{nz}, which is $\mathcal{O}(vu)$.

Below we proceed to simulate datasets with varying sizes and study their running time and memory usage.

Feature	$f(v,u)$	**Asymptotics**
# Variables (N_{var})	$v \times u + v + u$	$\mathcal{O}(vu)$
# Constraints (N_{constr})	$3 \times v \times u + 2$	$\mathcal{O}(vu)$
Size of constraint matrix	$N_{var} \times N_{constr}$	$\mathcal{O}(v^2u^2)$
# Non-zeros (N_{nz})	$7 \times v \times u + v + u$	$\mathcal{O}(vu)$
Density of constraint matrix	$N_{nz}/(N_{var} \times N_{constr})$	$\mathcal{O}((vu)^{-1})$

Table S1: **Feature sizes of the biclustering formulation model as a function of the number of cells, v, and the number of CpG sites, u in the input.**

Simulation details For simplicity, we set $v = u$ in our simulations. For each (square) input binary matrix M_{sim}, we set a "dense" submatrix of size $.7v \times .7u$ such that each entry in the submatrix is set to 1 independently with probability q, while each entry outside of the submatrix is set to 1 with probability $1 - q$. (An entry set to 1 indicates that the corresponding cell has at least one read covering the corresponding site.) We vary the simulation parameters as follows:

$$v = u = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$
$$q = \{1, .9, .8, .7\}$$

For $q = 1$, we have a $.7v \times .7u$ submatrix with all 1's and 0's everywhere else. As $q > .5$ decreases, while we still have a dense $.7v \times .7u$ submatrix, its density difference with the rest of the matrix is not as pronounced, making it more difficult to achieve our objective. With each set of parameters (v, u, q), 5 matrices are simulated (for investigating the stochasticity introduced by $q < 1$), resulting in a total of $10 \times 4 \times 5 = 200$ simulated matrices.

Results We perform biclustering on each of the 200 simulated binary matrices with $\alpha = \beta = \{.6, .7, .8\}$, resulting in 600 trials total. (We remind the reader that α and β are respectively the proportion of the rows and columns of the densest submatrix we aim to identify.) The trials are performed serially on the same machine with a single thread. As the dense submatrix in each simulated matrix has dimensions $.7v \times .7u$, performing biclustering with $\alpha = \beta = .6$ and $\alpha = \beta = .8$ will respectively correspond to an underestimation and overestimation of the two parameters. We profile (i) running time and (ii) peak memory usage for all trials.

As a general pattern, both wall-clock running time (Figure S1a., where the y-axis is $\log_2$-transformed to better visualize the differences) and peak memory usage (Figure S1b.) grow linearly with $v \times u$. For any given input size, the running time

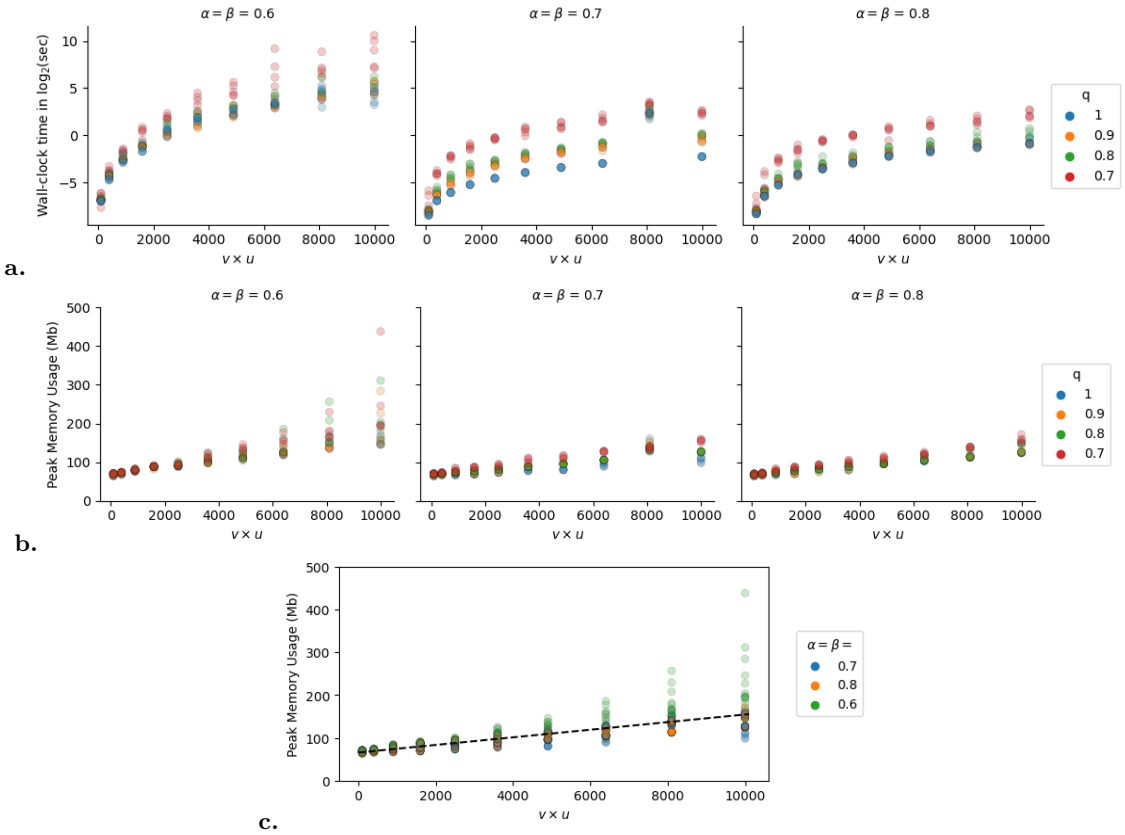


Fig. S1: **Running time and memory behavior on 200 simulated square binary matrices. (a.)** Wall-clock time in $\log_2$ of seconds. Measures were $\log_2$-transformed to better visualize the difference. **(b.)** Peak Memory consumption in Mb. **(c.)** Linear regression on peak memory usage across all 600 trials. Dashed line describes $y = .0089x + 65.69$.

increases with decreasing value of q, as the objective becomes more difficult to achieve (as demonstrated within each column in Figure S1a.). Notably, the running time also seems to be impacted by the choice of α and β: while both underestimation and overestimation lead to an increase of running time possibly due to the existence of several equally good solutions, underestimation seems to lead to a more substantial penalty (as can be observed across columns corresponding to the same input size in Figure S1a.). The peak memory usage seems to be similar for a given size input across choices of α and β, though an underestimation seems to occasionally lead to considerably higher usage and hence higher variance across trials (as can be observed across columns corresponding to the same input size in Figure S1b.).

To allow extrapolation of peak memory usage for any given input size, we perform linear regression using the ordinary least squares method on the data from the 600 trials, with $v \times u$ as the independent variable and peak memory usage in Mb as the dependent variable, obtaining a y-intercept of 65.59 and slope of .0089.

We do not extrapolate running time in the same manner, as it varies drastically with problem difficulty, which we proxy with q in our simulation study, but is hard to estimate in real data. Running time will also depend on the specific optimization strategies employed by Gurobi, a proprietary software, that may depend on other features of the problem instance, such as the density of the constraint matrix. Furthermore, one can always specify a time limit for a Gurobi run and obtain an intermediate output and its associated optimality gap, meanwhile Gurobi fails without output if memory limit is reached; therefore, memory is more likely to present a bottleneck for large-scale problems than time in practice.

Results from simulated data lead us to two key conclusions: (i) memory usage may be a critical limiting factor to the feasibility on large input, and (ii) well-chosen α and β values will not only yield a more informative submatrix for the remainder of the `Sgootr` workflow, they will also (typically) lead to faster converge than ill-chosen parameters.

S1.2 Biclustering case study on CRC01 from the Bian cohort

Next, we investigate the performance of our biclustering formulation on the CRC01 dataset from the Bian cohort as we vary α and β values.

Data descriptions The dataset has in total 409 single cells with scBS-seq data from patient CRC01, among which 241 also has matching scRNA-seq data. As we will call copy number from scRNA-seq data to input to `Sgootr`, we only include those 241 cells with both scRNA-seq and scBS-seq data for the biclustering procedure. Collectively, the 241 cells cover 52,343,774

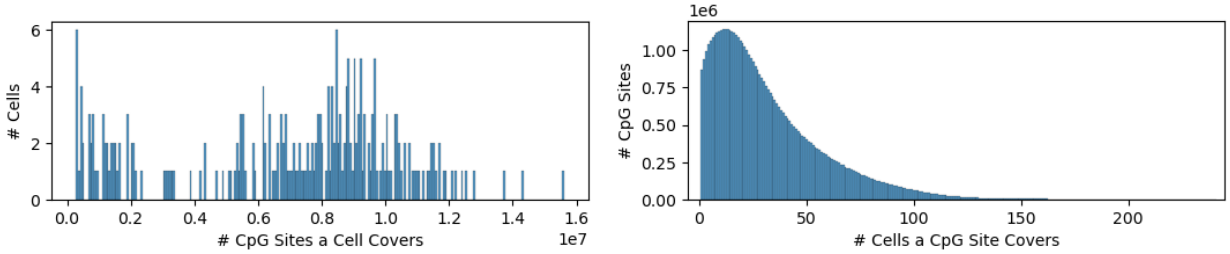


Fig. S2: **Patient CRC01 cell (left) and site (right) coverage level distribution.**

Original	**Subsamples** $(24 \times 52,343)$											
$(241 \times 52,343,774)$	0	1	2	3	4	5	6	7	8	9	**mean**	**std**
13.21%	14.53%	12.22%	14.58%	11.90%	12.17%	12.74%	12.09%	15.11%	16.31%	10.34%	**13.20%**	**1.74%**

Table S2: **Densities of the original binary matrix and the subsampled matrices.**

unique autosomal CpG sites. Figure S2 shows per-cell CpG site coverage distribution and per-site cell coverage distribution. The induced $241 \times 52,343,774$ binary coverage matrix has a density of 13.21%.

We aim for the resulting submatrix to have a density of at least 66%. The intuition here is that Sgootr estimates the distance between any pair of cells using the set of CpG sites which are covered by at least one read in each cell; provided 2/3 of the picked sites are covered by a read in each cell, the distance between any pair of cells will be based on at least 1/3 of the picked sites, i.e. at least half of the sites covered by a read in each cell. This ensures that the number of CpG sites employed in distance estimations involving one particular cell do not vary by more than a factor of 2. While we do not ensure that each row (or for that matter each column) has a density of at least 2/3, an overall density of 2/3 (or $\sim 66\%$) in the submatrix identified by our ILP formulation should provide a reasonable distance estimate between cell pairs.

Estimating α and β values To estimate α and β, we generate 10 disjoint subsamples of the dataset, each containing 10% of the 241 cells and 0.1% of the 52,343,774 CpG sites. For each subsample, we generate the corresponding $24 \times 52,343$ binary matrix (Table S2), on which we perform biclustering with all possible combinations of $\alpha = \{.4, .5, .6, .7, .8, .9\}$ and $\beta = \{.001, .005, .01, .05, .1, .15\}$ in parallel, resulting in $10 \times 6 \times 6 = 360$ runs of biclustering in total. The choices for the ranges of α and β values are informed by the density of the original matrix, the desired density of roughly 66%, and the cell and site coverage distributions (Figure S2). We will determine α and β based on i) time-to-convergence, ii) optimality gap at time limit, and iii) density of resulting submatrix in the subsampled trials.

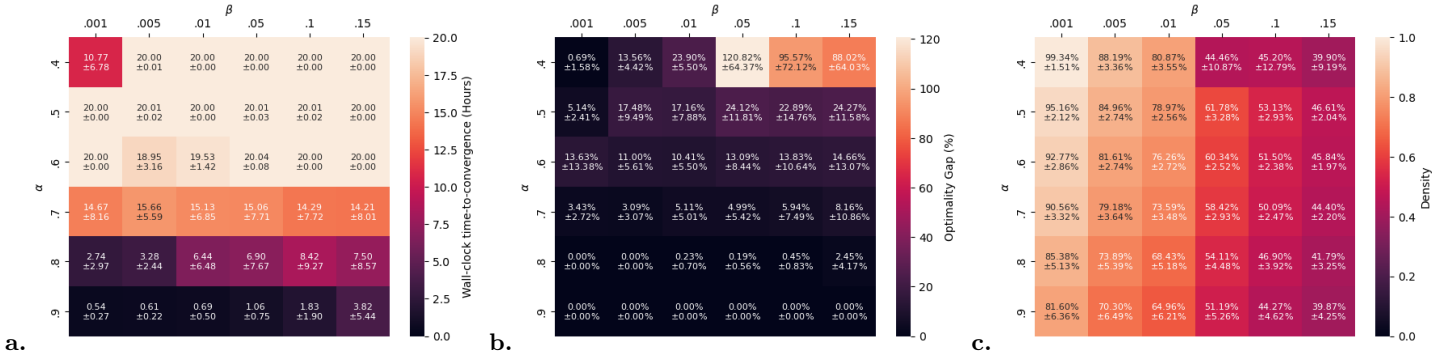


Fig. S3: **Mean and standard deviation of (a.) wall clock time-to-convergence in hours, (b.) optimality gap at 20 hours, and (c.) resulting submatrix density across the 10 subsamples, over a range of (α, β) value pairs.** Note that we set the time limit for the biclustering program to 20 hours, and therefore wall clock time-to-convergence is at most 20 hours.

Results A total of 360 runs of biclustering were performed across the 10 subsamples over a range of α and β values. Each run was given 4 threads and 20 hours. Runs with $\alpha = \{.8, .9\}$ were able to achieve negligible optimality gap in a matter of several hours, and smaller values of α leads to substantially higher time-to-convergence and optimality gap within the allotted time (Figure S1a.,b.). Fixing $\alpha = \{.8, .9\}$, looking at the other parameter, runs with $\beta = \{.001, .005\}$ all solved to convergence

with no optimality gap (Figure S1a.,b.). With $\alpha = \{.8, .9\}$ and $\beta = \{.001, .005\}$, the resulting submatrices have all achieved a density within a neighborhood of 66%. We choose $\alpha = .8$ and $\beta = .005$ to i) avoid over-pruning, and ii) considering variance in the different subsamples, have some confidence that results from the full original binary matrix will be at least 66% dense. (Figure S1c.).

The next step would be running biclustering on the original $241 \times 52,343,774$ binary matrix with the chosen parameters $\alpha = .8$ and $\beta = .005$. However, for input of this scale, the peak memory usage is predicted to be 112 Tb according to our linear regression result from simulated data, and is not feasible on the resources available to us and likely most researchers. Therefore, we opted for performing heuristic filtering in generating the results in the main text. Note that the submatrix we obtained with heuristic filtering on CRC01 is of a similar scale as that from performing biclustering on a $241 \times 52,343,774$ matrix with $\alpha = .8$ and $\beta = .005$.

For future iterations of the tool, we aim to optimize the implementation of the biclustering procedure to improve memory efficiency, and a combination of heuristic filtering and biclustering may be needed for best solving the problem at hand. In particular we aim to utilize the ILP formulation on the entire set of cells but only a subset of sites so as to stay within the limits of Gurobi's ability to solve the problem. Given the subset of cells identified by the ILP formulation, we will then be able to employ the heuristic filtering to identify the subset of sites with high coverage on these cells.

S2 Benchmarking Sgootr on simulated data

In this section we evaluate the performance of distinct components of Sgootr on simulated data, in comparison to baseline approaches. All Sgootr-related parameter settings will be the default configuration. Furthermore, for simplicity, we will assume copy number = 2 in our simulations.

S2.1 Expected distance calculation on simulated data

First we focus on our formulation for the expected distances between cell pairs and evaluate its performance against a simple baseline approach that was utilized in earlier studies [14, 5] which ignores sites that are potentially heterozygously methylated.

Simulation details Our simulations generate pairs of diploid cells with length-normalized ground truth distance of $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$. For each value of π we simulate 50 cell pairs, producing in total $9 \times 50 = 450$ simulated cell pairs. Each cell starts with 2 (CpG) alleles of 1000 unmethylated CpG sites. For any simulated pair, given π, we uniformly sample the number of sites where one cell is homozygously methylated but the other is homozygously unmethylated (each contributing 1 to the total distance) and the number of sites where one cell is homozygous but the other is heterozygously methylated (contributing .5 to the total distance) from all possible combinations. For example, with $\pi = .6$, by pigeonhole principle we have at least 200 and at most 600 CpG sites each of which contributing 1 to the total distance. After sampling those sites, the remainder of the ground truth distance need to be contributed by sites where one cell is heterozygous but the other is homozygous. After sampling this second set of sites, the ground truth distance is already achieved and the remaining sites will be of the same methylation status across the two cells.

Next, for each of the 450 cell pairs with known normalized ground truth distances, we simulate methylated and unmethylated read count vectors for either cell (i) for a range of possible depth of coverage to study how that may impact the measurements, and (ii) using the empirical per-site read coverage distributions from patient CRC01, the key patient we analyze in the main text. For (i), specifically, we model the number of reads covering a CpG site by a zero-truncated Poisson distribution (ZTPois) parameterized by $\lambda = \{1, 2, 3, 4, 5, 6\}$ (Figure S4a.). For (ii), we obtain the empirical per-site read coverage from patient CRC01 after first-pass filtering (Figure S4b.). The empirical distribution is truncated at 20, as probability substantially diminishes for larger values. For each simulated cell pair, we simulate 5 sets of methylated and unmethylated read count vectors for every depth of coverage distribution, obtaining $450 \times 7 \times 5 = 15,750$ sets of read count vectors.

We compare the distances between the read count vectors of a cell pair obtained by (i) a baseline distance measure, which we will describe in detail in the following section, and (ii) our expected distance formulation with default parameters. This comparison will help us better understand examine (a) how the depth of coverage affects the two distance measures, and (b) given the empirical read coverage distribution, which measure better models the ground truth distance.

Results The baseline distance measure we leverage in the comparison is one commonly used in prior studies [14, 5]. For each CpG site in a single cell, this approach first calls its status as methylated or unmethylated depending on which has with support from at least 90% of the observed reads at that site for the cell; it then discards all sites for which such binarization cannot be done. As a result, each cell will have a binary vector (likely with some missing entries due to failed binarization) indicating whether each site is methylated. Then, for each pair of cells, this approach computes the Hamming distance

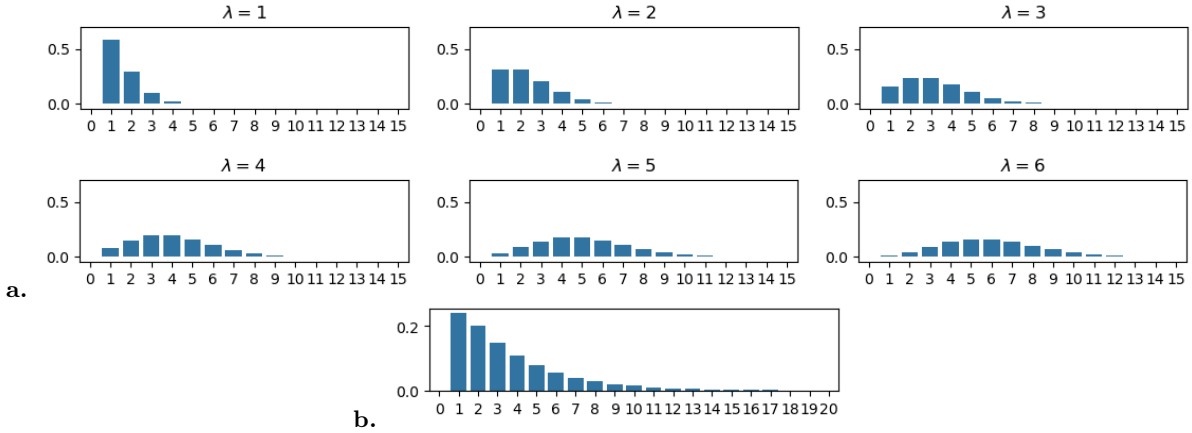
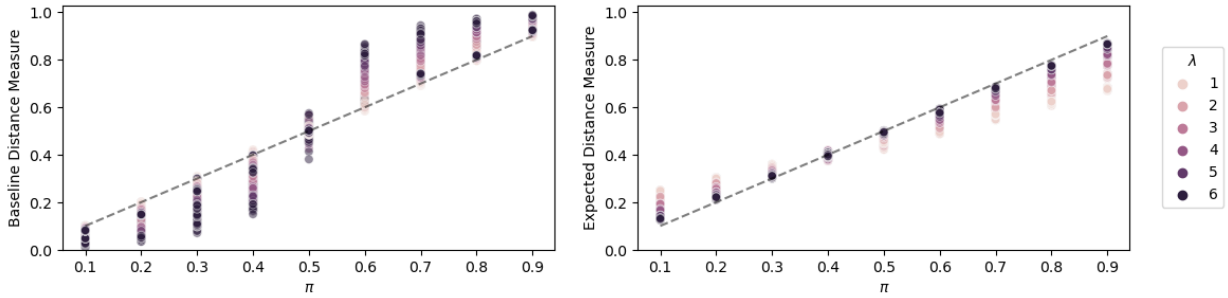


Fig. S4: **Probability mass functions of (a.) ZTPois(λ), where $\lambda = \{1, 2, 3, 4, 5, 6\}$ and (b.) the empirical per-site read coverage distribution from patient CRC01.**

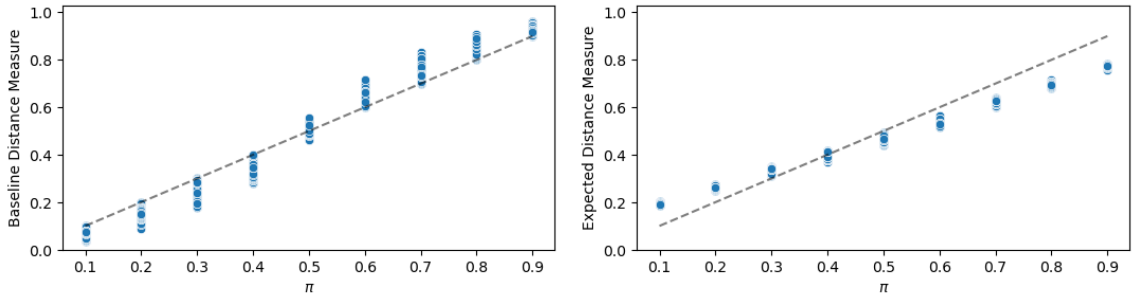


a.

b.

λ	1	2	3	4	5	6
$\mathbf{R^2}$**(**π**, Baseline Distance Measure)**	.998	.994	.987	.978	.970	.962
$\mathbf{R^2}$**(**π**, Expected Distance Measure)**	.997	.997	.997	.998	.999	.999

Fig. S5: **Baseline distance measure and expected distance measure on** $13,500$ **sets of read count vectors, simulated from cell pairs with known normalized ground truth distance** $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$**, with level of read coverage depth modeled by ZTPois(**λ**), where** $\lambda = \{1, 2, 3, 4, 5, 6\}$**. (a.)** Measured distances across increasing value of π for various λ. The dashed lines describe $y = x$, and are drawn to highlight the direction of deviation in the measured distances. **(b.)** For a given level of read coverage depth parameterized by λ, the R^2 values of linear regressions with the base line v.s. our expected distance measures as the independent variable and the true distance π as the dependent variable.



a.

b.

$\mathbf{R^2}$**(**π**, Baseline Distance Measure)**	.987
$\mathbf{R^2}$**(**π**, Expected Distance Measure)**	.997

Fig. S6: **Baseline distance measure and expected distance measure on** $2,250$ **sets of read count vectors, simulated from cell pairs with known normalized ground truth distance** $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$**, with level of read coverage depth modeled by the empirical read coverage distribution gathered from the CRC01 sample. (a.)** Measured distances across increasing value of π. The dashed lines describe $y = x$, and are drawn to highlight the direction of deviation in the measured distances. (Note that the expected distance measure is based on the default configuration with the probability of a site being heterozygous set to $1/3$.)**(b.)** R^2 values of linear regressions with the measured distances being independent variable and the true distance π being the dependent variable.

between the respective binary vectors over all sites at which both have a binarized status, then normalizes the distance by the number of such sites. This distance measure essentially excludes heterozygosity from consideration.

We first examine how the depth of coverage (whose distribution is as per Figure S4 panel a) affects the baseline in comparison to our expected distance measure. For each of the $450 \times 6 \times 5 = 13,500$ sets of read count vectors sampled according to ZTPois(λ), we measure both distances. The baseline distance measure tend to underestimate the ground truth distance when $\pi < .5$ and overestimate it when $\pi > .5$ (Figure S5a., left). This is because $\pi = \frac{\text{measured difference}}{\text{\# shared sites}}$ and the exclusion of each unbinarized (i.e. observed to be heterozygous) site subtracts .5 from the numerator, and subtracts 1 from the denominator. This directional deviation from true π becomes more substantial as read depth increases, witnessed by the decreasing R^2 values resulting from performing linear regression of π on the measured baseline distance (Figure S5b., top row). Intuitively, for small λ, likely only one allele at the heterozygous site will be sampled, and (assuming there is no preferential sampling of alleles of a particular status) its status will with 50% chance be called methylated and 50% unmethylated, and on expectation, a heterozygous status and a homozygous status will still contribute an expected distance of .5; however, for larger λ, heterozygosity is more likely to be reflected in the observed reads, therefore a larger portion of true heterozygous sites will be excluded, leading to a larger deviation of measured distance from π.

In contrast, for any given λ, our expected distance measurement has persistently high R^2 values in a linear regression of π (Figure S5a., right; S5b., bottom row). In fact, the R^2 values slightly improve with increasing level of coverage. We acknowledge that the expected distance measure, too, deviates from the ground truth π, especially at low coverage. As described by the intuition previously provided in Section 2.3, with few reads at a CpG site, it is unlikely (quantified by the

user-defined parameters) that they have sampled from all alleles present; therefore, if a CpG site in a cell only has reads of one methylation status, and the number of reads is low, our expected distance measure will assign some probability to it being heterozygous. This will lead to overestimation of the distance at a CpG site if two cells have the same homozygous status but are low-coverage for the site; similarly, this will lead to underestimation of the distance at a CpG site if two cells have the opposite homozygous statuses but are low-coverage for the site. There are proportionally more of the former type of sites for lower values of π and more of the latter type of sites for higher values of π, explaining the direction of deviation from $y = x$ (Figure S5a., right). That said, as the coverage level increases, the expected distance measure converges to the ground truth distance π.

Since it is reasonable to assume that the per-site read count distribution across all single cells sequenced with the same technology is close to being identical, our expected distance formulation is preferable to the baseline measure, as it tracks the true distance well for any given read coverage level. This is the case especially for higher coverage levels increases, which should be desirable if not destined as sequencing technology improves.

Next, we examine the performance of the two distance measures on data simulated using the empirical read coverage distribution (as per Figure S4 panel b). For each of the $450 \times 1 \times 5 = 2,250$ sets of read count vectors, we again measure distance using (i) the baseline measure and (ii) our expected distance measure. The trends previously observed with read counts sampled from ZTPois(λ) still stand (Figure S6a.), and our expected distance measure again tracks the ground truth distance π better at this level of coverage, resulting a higher R^2 value (Figure S6b.).

S2.2 Iterative pruning on simulated data

Next, we evaluate our iterative pruning approach on simulated data against a baseline approach as described in Section 3.

Simulation details We simulate 10 clone trees with different topologies, various degrees of branching, and various depths (Figure S7a.). Each clone tree has 1 normal clone (grey "ROOT" nodes in Figure S7a.) and 20 tumor clones (numbered nodes in Figure S7a.). Each clone is represented by a pair of ground truth lineage-informative CpG-alleles with 9,000 CpG sites. 3000 of these are infinite-sites-abiding, another 3,000 "invariant", and the final 3,000 those of the "mixture" model [28]. For each clone tree, we first focus on the 3,000 infinite-sites-abiding CpG sites. At "ROOT", all of these sites are set to be homozygously unmethylated, and at each branch in the clone tree, a subset of these CpG sites that have not been previously selected are selected to have their statuses change from homozygously unmethylated to (i) homozygously methylated with 50% chance and (ii) heterozygously methylated with 50% chance. The subsets are roughly equal in size. The remaining 6,000 CpG sites are simulated at the single cell level, as described below.

Given the clone tree with known (sub)clonal methylation status changes for the "infinite-sites-abiding" CpG sites, we simulate single-cell datasets with added noise and sparsity. To simulate one single-cell dataset from a clone tree, we simulate 10 diploid cells from each clone ($21 \times 10 = 210$ cells for each dataset), each with 6,000 additional CpG sites - the methylation status of the 3,000 "infinite-sites-abiding" sites are copied from their clone of origin. Among the additional CpG sites, the first 3,000 are "invariant" whose methylation status in each cell is with 80% chance homozygous (half of these homozygous sites are methylated by default and the other half are unmethylated), 10% chance heterozygous, and 10% chance the alternative homozygous. The remaining 3,000 sites are of the mixture model of methylation dynamics[28]: each is with 50% chance heterozygous, 25% chance homozygously methylated and 25% homozygously unmethylated.

Given such a set of 210 pairs of CpG-alleles, we sample the corresponding $210 \times 9,000$ methylated and unmethylated read count matrices according to the empirical read coverage distribution (Figure S4b.) as previously described in Section S2.1. Furthermore, we randomly mask out the same 33% of the coordinates in either $210 \times 9,000$ matrix, introducing sparsity observed in real datasets while mirroring the desired input matrix density (previously justified in Section S1.2).

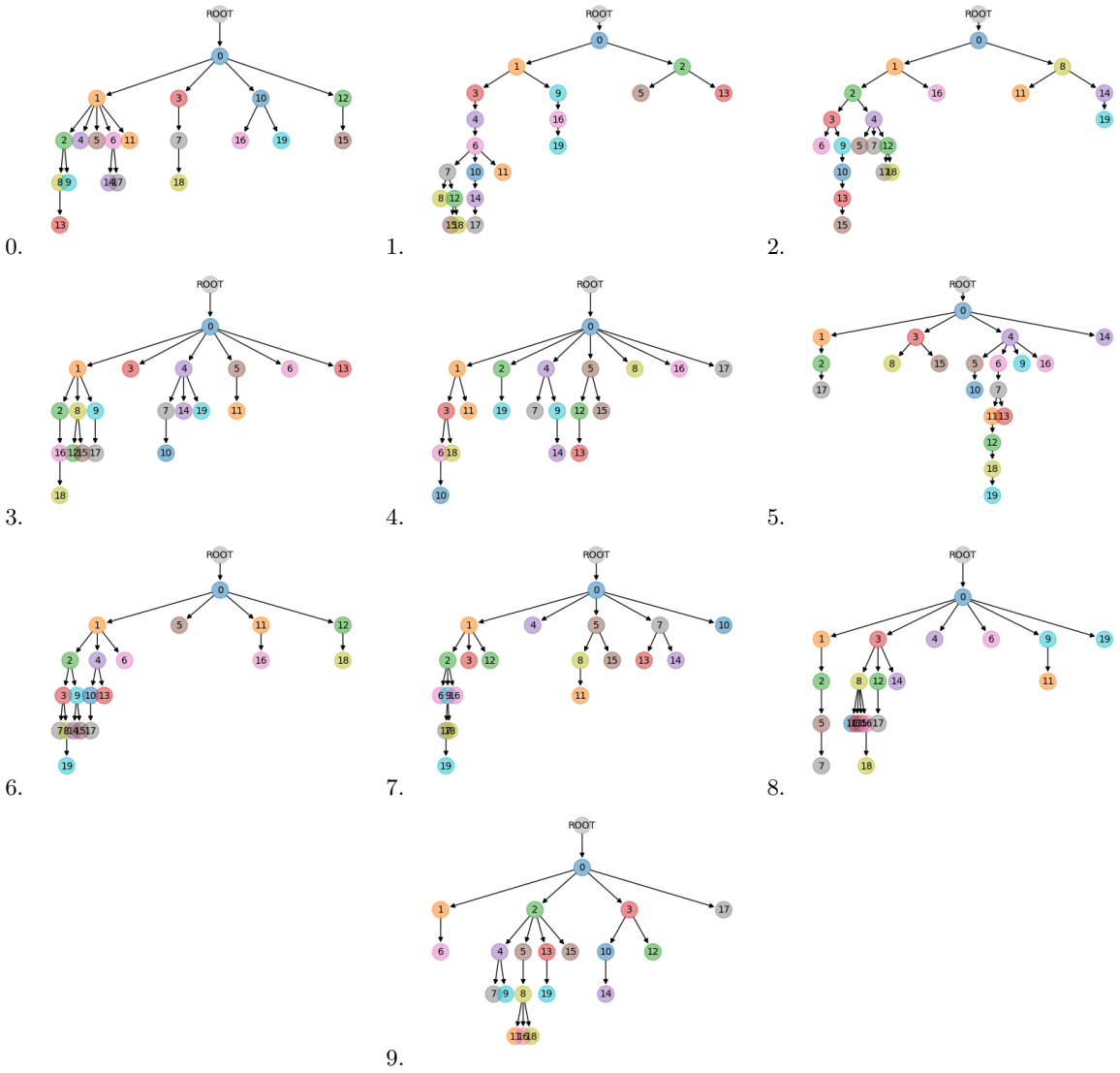
We simulate 5 such single-cell datasets for each clone tree, resulting in $10 \times 5 = 50$ total pairs of (un)methylated read count matrices that are 66% dense. We leverage those as input to our simulation experiments.

Results We compare the performance of `Sgootr` (no biclustering or error correction, $\kappa = .1$, for a maximum of 20 iterations) on simulated datasets against a baseline approach as described in Section 3 with respect to the trees constructed. Briefly, given the read count matrices, the baseline approach computes the Hamming distance between the binarized methylation status vectors of each cell pair, normalize by the number of sites covered in both cells, and applies NJ on the pairwise distance matrix to obtain a tree. We root the `Sgootr`- and baseline-generated trees with the same set of arbitrarily picked single-cell from the normal clone, and infer clonal evolution histories from the constructed trees using our Fitch's algorithm-based approach (Section 2.5) with the ground truth topologies (Figure S7a.) as the guiding partial orders. We use the same two evaluation metrics as in Section 3: (i) number of methylation events, and (ii) BAMDPO, which measures the difference between the inferred and the ground truth topologies.

In terms of the two performance measures, `Sgootr` outperforms the baseline approach in datasets generated from 9 out of 10 simulated clone trees, and performs comparably in datasets generated from simulated clone tree 4 (Figure S7b., comparing between "baseline" columns and "$t*$" columns). Furthermore, despite that `Sgootr` many times already has superior performance compared to baseline prior to its iterative procedure, highlighting the benefits of using the expected

distance formulation as opposed to the baseline distance measure (Figure S7b., comparing between "baseline" columns and "t_0" columns), the iterative procedure of `Sgootr` seems to lead to additional boosts in performance by refining the signal (Figure S7b., comparing between "t_0" columns and "$t*$" columns).

a.



b.

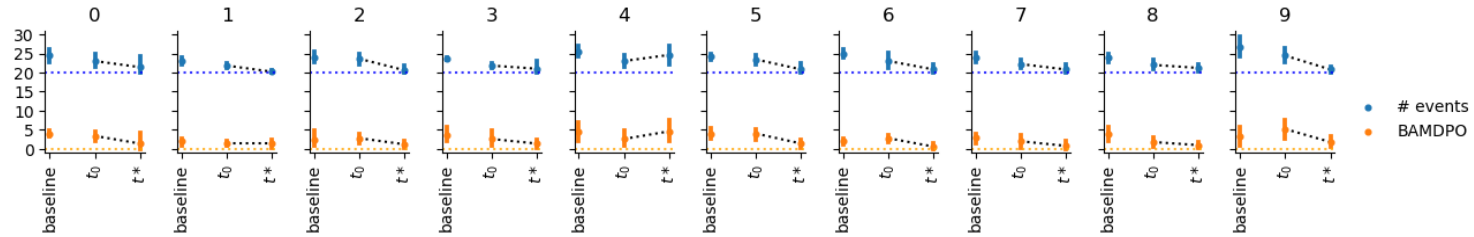
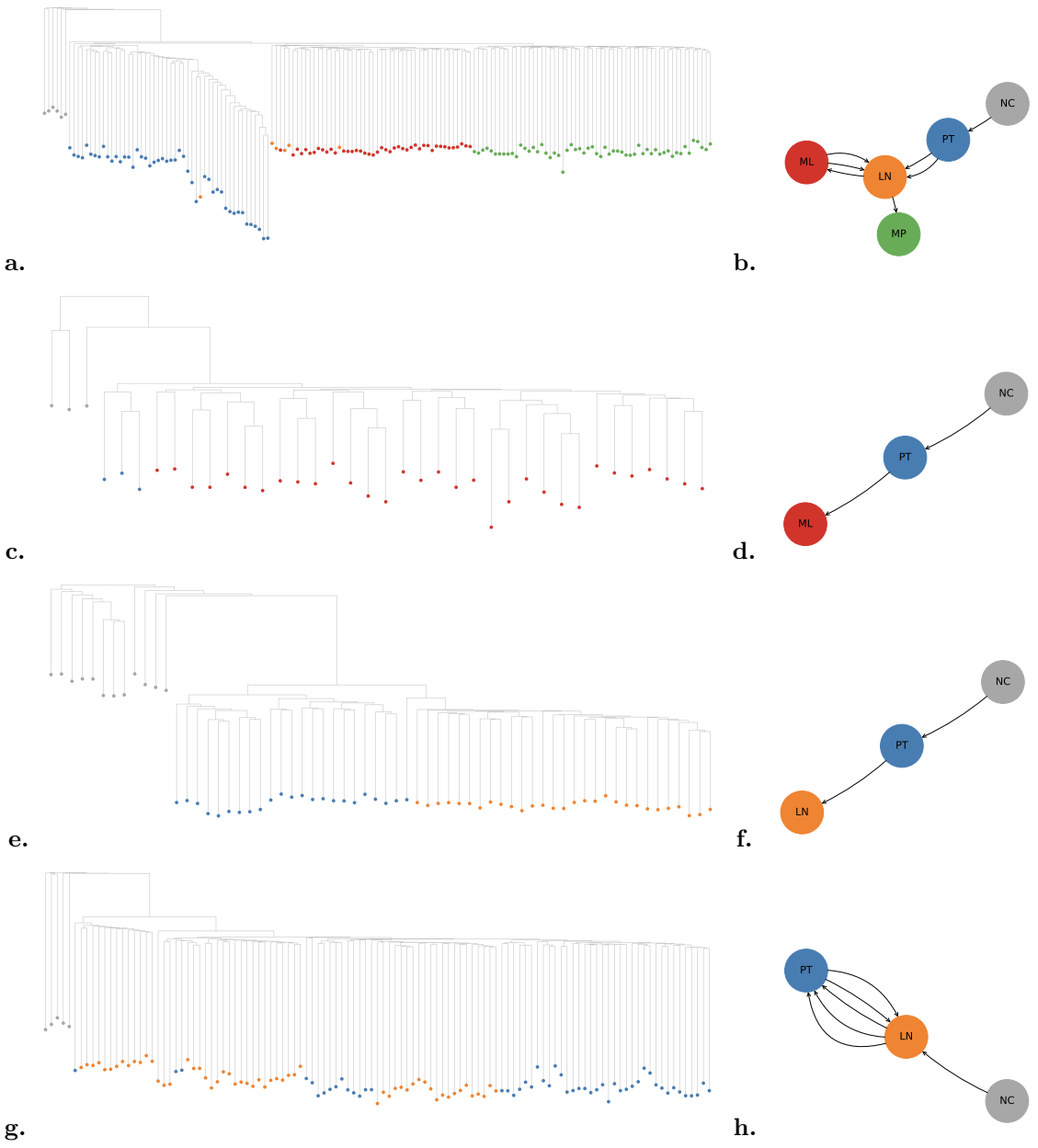


Fig. S7: **Results on 10 simulated clone trees from `Sgootr` and baseline method. (a.)** Topologies of the simulated clone trees. Each clone tree has 1 normal clone and 20 tumor (sub)clones. A total of 3,000 CpG sites abiding by the infinite-sites assumption - and therefore may be lineage-informative - were simulated for each tree. **(b.)** Number of migration events and BAMDPO values of migration histories obtained via (i) naïve baseline method and (ii) `Sgootr` intermediate tree prior to iterative procedure (t_0), and (iii) final tree output by `Sgootr` ($t*$). Results are represented by the mean across 5 single-cell methylation datasets generated from each clone tree, with the error bars denoting the 95% confidence interval of the mean. Dotted lines in blue and orange mark the ground truth (best possible) values for number of events and BAMDPO, respectively.

S3 Application of Sgootr on the complete set of 9 metastatic CRC patients made available by Bian *et al.*[2]

Patient	Primary site	Sampled lesions	Number of cells	Number of sites	CNA calls available
CRC01	Left Colon	NC, PT, LN, ML, MP	160	140,801	✓
CRC02	Right Colon	NC, PT, ML	39	49,633	
CRC04	Right Colon	NC, PT, LN	65	303,849	✓
CRC10	Left Colon	NC, PT, LN	114	275,111	
CRC11	Left Colon	NC, PT, LN	228	579,272	
CRC12	Rectum	NC, PT, LN	22	650,140	
CRC13	Left Colon	NC, PT, LN	185	204,153	
CRC14	Rectum	NC, PT, LN	29	511,588	
CRC15	Right Colon	NC, PT, LN, ML, MO	50	42,824	

Table S3: **Data summary of the Bian *et al.* metastatic CRC cohort [2].** The number of cells and number of sites are those obtained after component (i) and (ii) of Sgootr before entering the iterative procedure.



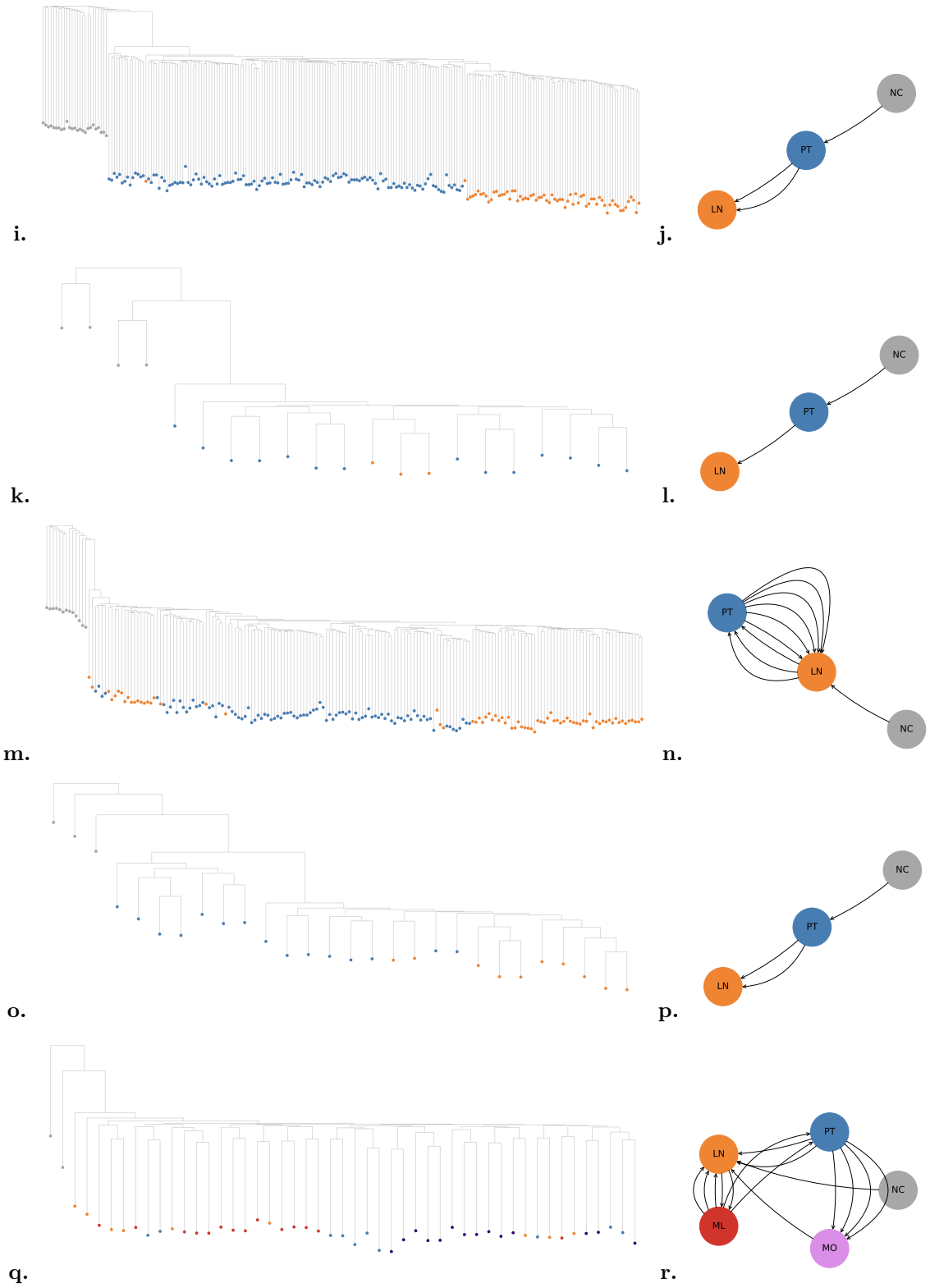


Fig. S8: **`Sgootr`-constructed ($\kappa = .1$) tumor lineage trees (t^*s) and -inferred tumor migration histories for patients in the Bian *et al.* cohort:** CRC01 (**a.**, **b.**), CRC02 (**c.**, **d.**), CRC04 (**e.**, **f.**), CRC10 (**g.**, **h.**), CRC11 (**i.**, **j.**), CRC12 (**k.**, **l.**), CRC13 (**m.**, **n.**), CRC14 (**o.**, **p.**), CRC15 (**q.**, **r.**)

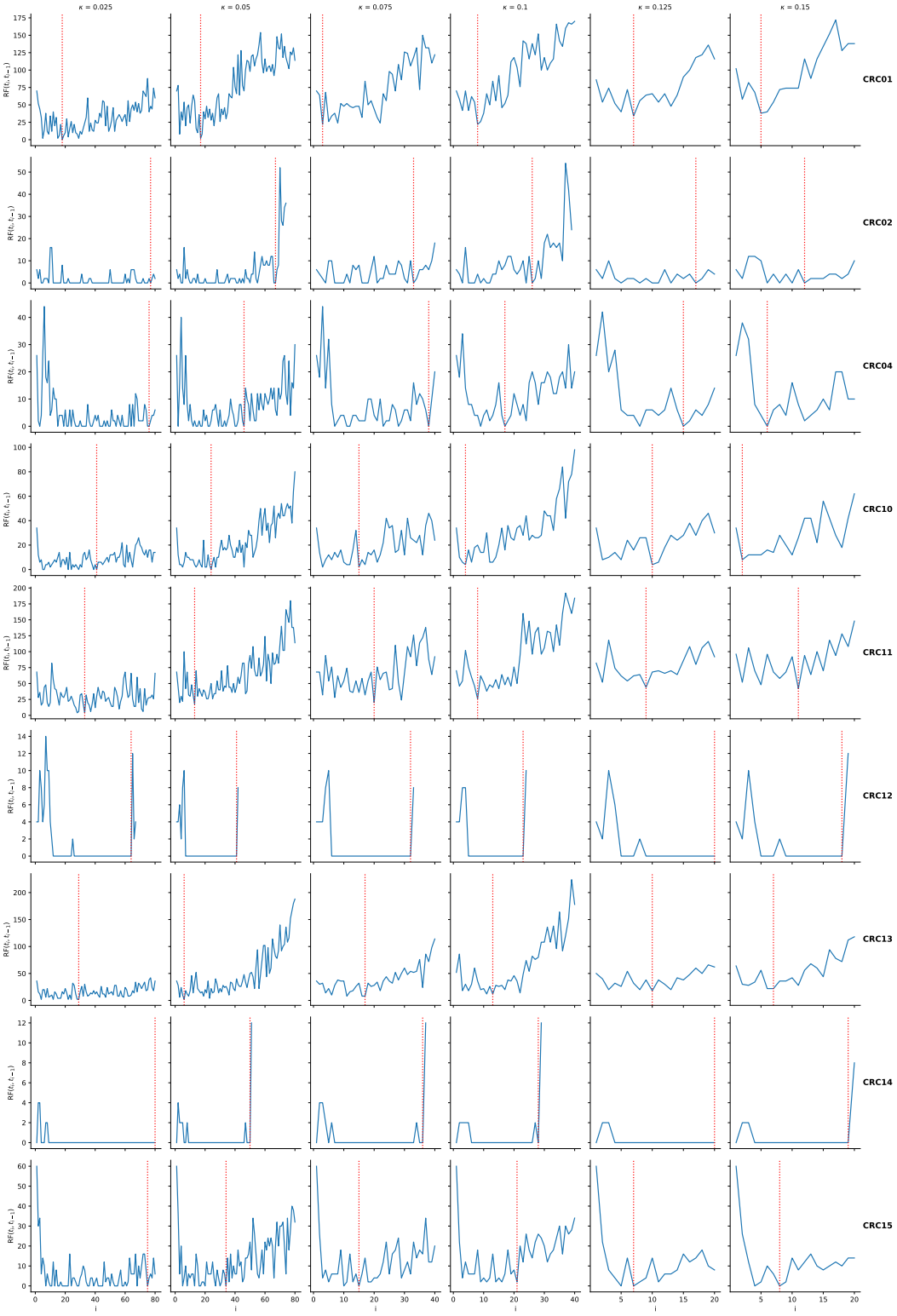
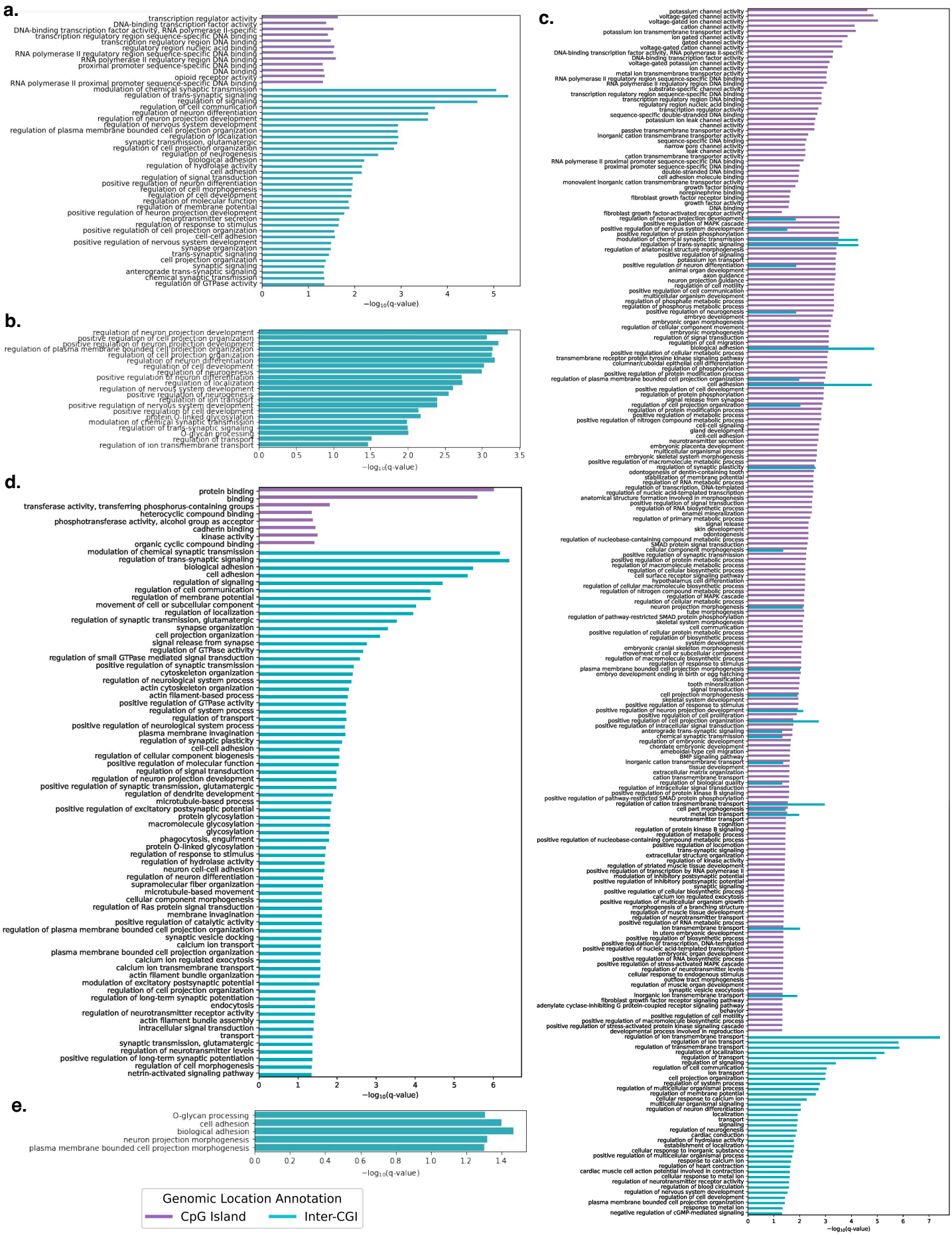


Fig. S9: **RF distances across iterations for each patient in the Bian *et al.* metastatic CRC cohort applying Sgootr with different κ values. RF distances across iterations for each patient in the Bian *et al.* metastatic CRC cohort with different κ values.** Each row represents each of the 9 patients, and each column represents κ=.025, .05, .075, .1, .125, and .15. For κ=.025, .05, the iterative procedure was run for a maximum of 80 rounds; for κ = .075, .1, maximum 40 rounds; and for κ=.125,.15, maximum 20 rounds. Red vertical lines marks the iteration Sgootr identifies as optimal. If the blue line ends before the maximum iteration, it means at the last iteration of pruning, the distance between some cells pairs can no longer be computed as all their shared CpG sites have been pruned out. Note that the maximum iteration is arbitrarily chosen, and Sgootr does not need to finish running the maximum iteration in order to detect whether the optimal condition has been met.


a.
transcription regulator activity
DNA-binding transcription factor activity
DNA-binding transcription factor activity, RNA polymerase II-specific
transcription regulatory region sequence-specific DNA binding
transcription regulatory region DNA binding
regulatory region nucleic acid binding
RNA polymerase II regulatory region sequence-specific DNA binding
RNA polymerase II regulatory region DNA binding
proximal promoter sequence-specific DNA binding
DNA binding
opioid receptor activity
RNA polymerase II proximal promoter sequence-specific DNA binding
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
regulation of signaling
regulation of cell communication
regulation of neuron differentiation
regulation of neuron projection development
regulation of nervous system development
regulation of plasma membrane bounded cell projection organization
regulation of localization
synaptic transmission, glutamatergic
regulation of cell projection organization
regulation of neurogenesis
biological adhesion
regulation of hydrolase activity
cell adhesion
regulation of signal transduction
positive regulation of neuron differentiation
regulation of cell morphogenesis
regulation of cell development
regulation of molecular function
regulation of membrane potential
positive regulation of neuron projection development
neurotransmitter secretion
regulation of response to stimulus
positive regulation of cell projection organization
cell-cell adhesion
positive regulation of nervous system development
synapse organization
trans-synaptic signaling
cell projection organization
synaptic signaling
anterograde trans-synaptic signaling
chemical synaptic transmission
regulation of GTPase activity
0 1 2 3 4 5
−log10(q-value)
b.
regulation of neuron projection development
positive regulation of cell projection organization
positive regulation of neuron projection development
regulation of plasma membrane bounded cell projection organization
regulation of cell projection organization
regulation of neuron differentiation
regulation of cell development
regulation of neurogenesis
positive regulation of neuron differentiation
regulation of localization
regulation of nervous system development
positive regulation of neurogenesis
regulation of ion transport
positive regulation of nervous system development
positive regulation of cell development
protein O-linked glycosylation
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
O-glycan processing
regulation of transport
regulation of ion transmembrane transport
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
−log10(q-value)
d.
protein binding
binding
transferase activity, transferring phosphorus-containing groups
heterocyclic compound binding
phosphotransferase activity, alcohol group as acceptor
cadherin binding
kinase activity
organic cyclic compound binding
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
biological adhesion
cell adhesion
regulation of signaling
regulation of cell communication
regulation of membrane potential
movement of cell or subcellular component
regulation of localization
regulation of synaptic transmission, glutamatergic
synapse organization
cell projection organization
signal release from synapse
regulation of GTPase activity
regulation of small GTPase mediated signal transduction
positive regulation of synaptic transmission
cytoskeleton organization
regulation of neurological system process
actin cytoskeleton organization
actin filament-based process
positive regulation of GTPase activity
regulation of system process
regulation of transport
positive regulation of neurological system process
plasma membrane invagination
regulation of synaptic plasticity
cell-cell adhesion
regulation of cellular component biogenesis
positive regulation of molecular function
regulation of signal transduction
regulation of neuron projection development
positive regulation of synaptic transmission, glutamatergic
regulation of dendrite development
microtubule-based process
positive regulation of excitatory postsynaptic potential
protein glycosylation
macromolecule glycosylation
glycosylation
phagocytosis, engulfment
protein O-linked glycosylation
regulation of response to stimulus
regulation of hydrolase activity
neuron cell-cell adhesion
regulation of neuron differentiation
supramolecular fiber organization
microtubule-based movement
cellular component morphogenesis
regulation of Ras protein signal transduction
membrane invagination
positive regulation of catalytic activity
regulation of plasma membrane bounded cell projection organization
synaptic vesicle docking
calcium ion transport
plasma membrane bounded cell projection organization
calcium ion regulated exocytosis
calcium ion transmembrane transport
actin filament bundle organization
modulation of excitatory postsynaptic potential
regulation of cell projection organization
regulation of long-term synaptic potentiation
endocytosis
regulation of neurotransmitter receptor activity
actin filament bundle assembly
intracellular signal transduction
transport
synaptic transmission, glutamatergic
regulation of neurotransmitter levels
positive regulation of long-term synaptic potentiation
regulation of cell morphogenesis
netrin-activated signaling pathway
0 1 2 3 4 5 6
−log10(q-value)
e.
O-glycan processing
cell adhesion
biological adhesion
neuron projection morphogenesis
plasma membrane bounded cell projection morphogenesis
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
−log10(q-value)
Genomic Location Annotation
CpG Island
Inter-CGI
c.
potassium channel activity
voltage-gated channel activity
voltage-gated ion channel activity
cation channel activity
potassium ion transmembrane transporter activity
ion gated channel activity
gated channel activity
voltage-gated cation channel activity
DNA-binding transcription factor activity, RNA polymerase II-specific
DNA-binding transcription factor activity
voltage-gated potassium channel activity
ion channel activity
metal ion transmembrane transporter activity
RNA polymerase II regulatory region sequence-specific DNA binding
RNA polymerase II regulatory region DNA binding
substrate-specific channel activity
transcription regulatory region sequence-specific DNA binding
transcription regulatory region DNA binding
regulatory region nucleic acid binding
transcription regulator activity
sequence-specific double-stranded DNA binding
potassium ion leak channel activity
channel activity
passive transmembrane transporter activity
inorganic cation transmembrane transporter activity
sequence-specific DNA binding
narrow pore channel activity
leak channel activity
cation transmembrane transporter activity
RNA polymerase II proximal promoter sequence-specific DNA binding
proximal promoter sequence-specific DNA binding
double-stranded DNA binding
cell adhesion molecule binding
monovalent inorganic cation transmembrane transporter activity
growth factor binding
norepinephrine binding
fibroblast growth factor receptor binding
growth factor activity
DNA binding
fibroblast growth factor-activated receptor activity
regulation of neuron projection development
positive regulation of MAPK cascade
positive regulation of nervous system development
positive regulation of protein phosphorylation
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
regulation of anatomical structure morphogenesis
positive regulation of signaling
potassium ion transport
positive regulation of neuron differentiation
animal organ development
axon guidance
neuron projection guidance
regulation of cell motility
positive regulation of cell communication
multicellular organism development
regulation of phosphate metabolic process
regulation of phosphorus metabolic process
positive regulation of neurogenesis
embryo development
embryonic organ morphogenesis
regulation of cellular component movement
embryonic morphogenesis
regulation of signal transduction
regulation of cell migration
biological adhesion
positive regulation of cellular metabolic process
transmembrane receptor protein tyrosine kinase signaling pathway
columnar/cuboidal epithelial cell differentiation
regulation of phosphorylation
positive regulation of protein modification process
regulation of plasma membrane bounded cell projection organization
cell adhesion
positive regulation of cell development
regulation of protein phosphorylation
signal release from synapse
regulation of cell projection organization
regulation of protein modification process
positive regulation of metabolic process
positive regulation of nitrogen compound metabolic process
cell-cell signaling
gland development
cell-cell adhesion
neurotransmitter secretion
embryonic placenta development
multicellular organismal process
embryonic skeletal system morphogenesis
positive regulation of macromolecule metabolic process
regulation of synaptic plasticity
odontogenesis of dentin-containing tooth
stabilization of membrane potential
regulation of RNA metabolic process
regulation of transcription, DNA-templated
regulation of nucleic acid-templated transcription
anatomical structure formation involved in morphogenesis
positive regulation of signal transduction
regulation of RNA biosynthetic process
enamel mineralization
regulation of primary metabolic process
signal release
skin development
odontogenesis
regulation of nucleobase-containing compound metabolic process
SMAD protein signal transduction
cellular component morphogenesis
positive regulation of synaptic transmission
positive regulation of protein metabolic process
regulation of macromolecule metabolic process
regulation of cellular biosynthetic process
cell surface receptor signaling pathway
hypothalamus cell differentiation
regulation of cellular macromolecule biosynthetic process
regulation of nitrogen compound metabolic process
regulation of MAPK cascade
regulation of cellular metabolic process
neuron projection morphogenesis
tube morphogenesis
regulation of pathway-restricted SMAD protein phosphorylation
skeletal system morphogenesis
cell communication
positive regulation of cellular protein metabolic process
regulation of biosynthetic process
system development
embryonic cranial skeleton morphogenesis
movement of cell or subcellular component
regulation of macromolecule biosynthetic process
regulation of response to stimulus
plasma membrane bounded cell projection morphogenesis
embryo development ending in birth or egg hatching
ossification
tooth mineralization
signal transduction
cell projection morphogenesis
skeletal system development
positive regulation of response to stimulus
positive regulation of neuron projection development
positive regulation of cell proliferation
positive regulation of cell projection organization
positive regulation of intracellular signal transduction
anterograde trans-synaptic signaling
chemical synaptic transmission
regulation of embryonic development
chordate embryonic development
ameboidal-type cell migration
BMP signaling pathway
inorganic cation transmembrane transport
tissue development
extracellular matrix organization
cation transmembrane transport
regulation of biological quality
regulation of intracellular signal transduction
positive regulation of protein kinase B signaling
positive regulation of pathway-restricted SMAD protein phosphorylation
regulation of cation transmembrane transport
cell part morphogenesis
metal ion transport
neurotransmitter transport
cognition
regulation of protein kinase B signaling
regulation of metabolic process
positive regulation of nucleobase-containing compound metabolic process
positive regulation of locomotion
trans-synaptic signaling
extracellular structure organization
regulation of kinase activity
regulation of striated muscle tissue development
positive regulation of transcription by RNA polymerase II
modulation of inhibitory postsynaptic potential
positive regulation of inhibitory postsynaptic potential
synaptic signaling
positive regulation of cellular biosynthetic process
calcium ion regulated exocytosis
positive regulation of multicellular organism growth
morphogenesis of a branching structure
regulation of muscle tissue development
regulation of neurotransmitter transport
positive regulation of RNA metabolic process
ion transmembrane transport
in utero embryonic development
positive regulation of biosynthetic process
positive regulation of transcription, DNA-templated
positive regulation of nucleic acid-templated transcription
embryonic organ development
positive regulation of RNA biosynthetic process
positive regulation of stress-activated MAPK cascade
regulation of neurotransmitter levels
cellular response to endogenous stimulus
outflow tract morphogenesis
regulation of muscle organ development
synaptic vesicle exocytosis
inorganic ion transmembrane transport
fibroblast growth factor receptor signaling pathway
adenylate cyclase-inhibiting G protein-coupled receptor signaling pathway
behavior
positive regulation of cell motility
positive regulation of macromolecule biosynthetic process
positive regulation of stress-activated protein kinase signaling cascade
developmental process involved in reproduction
regulation of ion transmembrane transport
regulation of ion transport
regulation of transmembrane transport
regulation of localization
regulation of transport
regulation of signaling
regulation of cell communication
ion transport
cell projection organization
regulation of system process
regulation of multicellular organismal process
regulation of membrane potential
cellular response to calcium ion
multicellular organismal signaling
regulation of neuron differentiation
localization
transport
signaling
regulation of neurogenesis
cardiac conduction
regulation of hydrolase activity
establishment of localization
cellular response to inorganic substance
positive regulation of multicellular organismal process
response to calcium ion
regulation of heart contraction
cardiac muscle cell action potential involved in contraction
cellular response to metal ion
regulation of neurotransmitter receptor activity
regulation of blood circulation
regulation of nervous system development
regulation of cell development
plasma membrane bounded cell projection organization
response to metal ion
negative regulation of cGMP-mediated signaling
0 1 2 3 4 5 6 7
−log10(q-value)

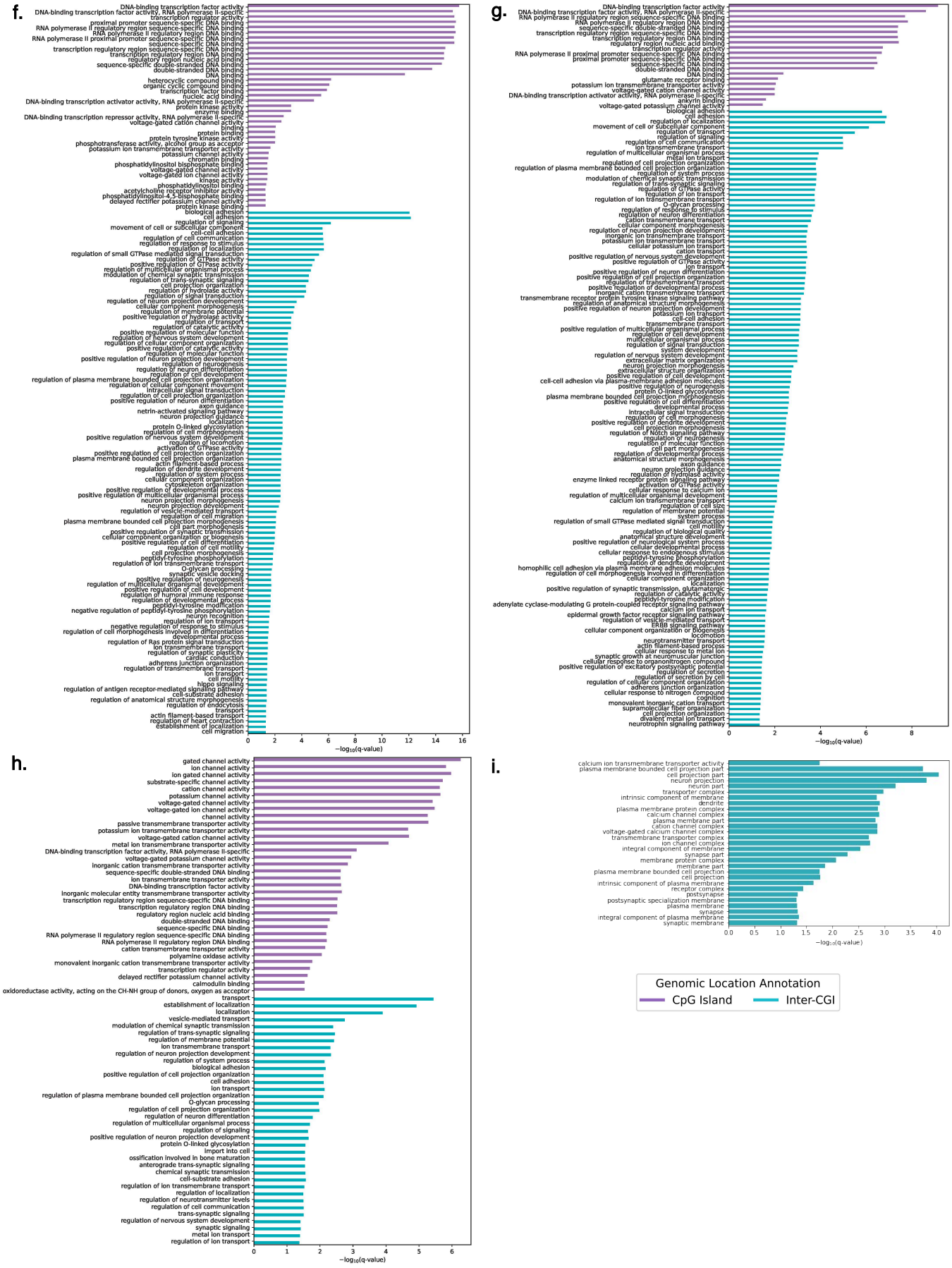


Fig. S10: **GOrilla[11]-identified GO terms with significant (<.05) q-values in enrichment analysis of non-pseudo, protein coding genes whose gene bodies and 1000bp upstream promoter regions contain the Sgootr-identified lineage-informative CpG sites, stratified by whether they locate in CpG island or inter-CGI regions for a. CRC01, b. CRC02, c. CRC04, d. CRC10, e. CRC13, f. CRC11, g. CRC12, h. CRC14, and i. CRC15.**

a.
Genomic Location Annotation
CpG Island
Inter-CGI
b.
−log10(q-value)
c.
−log10(q-value)
d .
−log10(q-value)
e.
−log10(q-value)

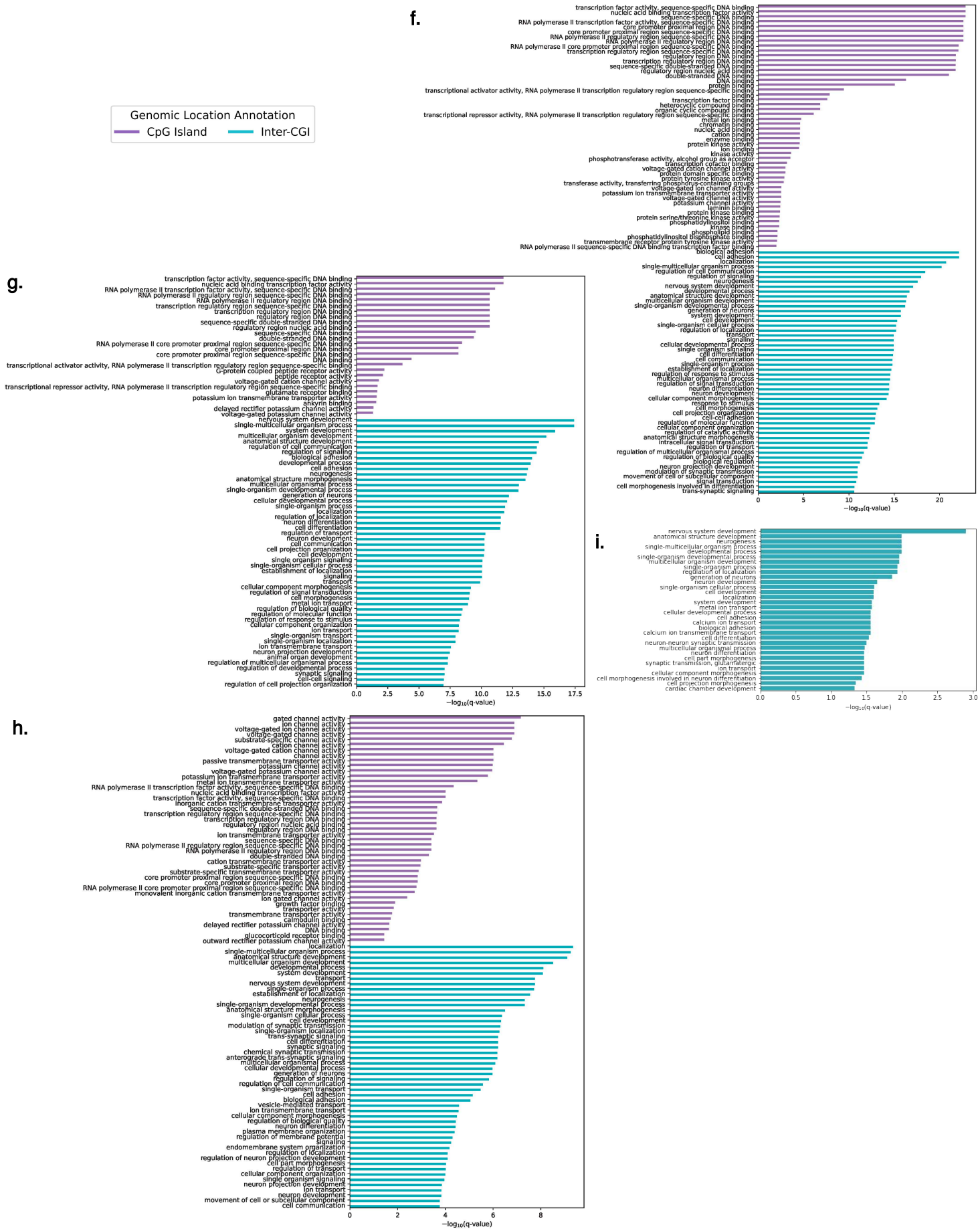


Fig. S11: **DAVID[36, 18]-identified GO terms with significant (<.05) q-values in enrichment analysis of non-pseudo, protein coding genes whose gene bodies and 1000bp upstream promoter regions contain the Sgootr-identified lineage-informative CpG sites, stratified by whether they locate in CpG island or inter-CGI regions for a. CRC01, b. CRC02, c. CRC04, d. CRC10, e. CRC13, f. CRC11, g. CRC12, h. CRC14, and i. CRC15.**

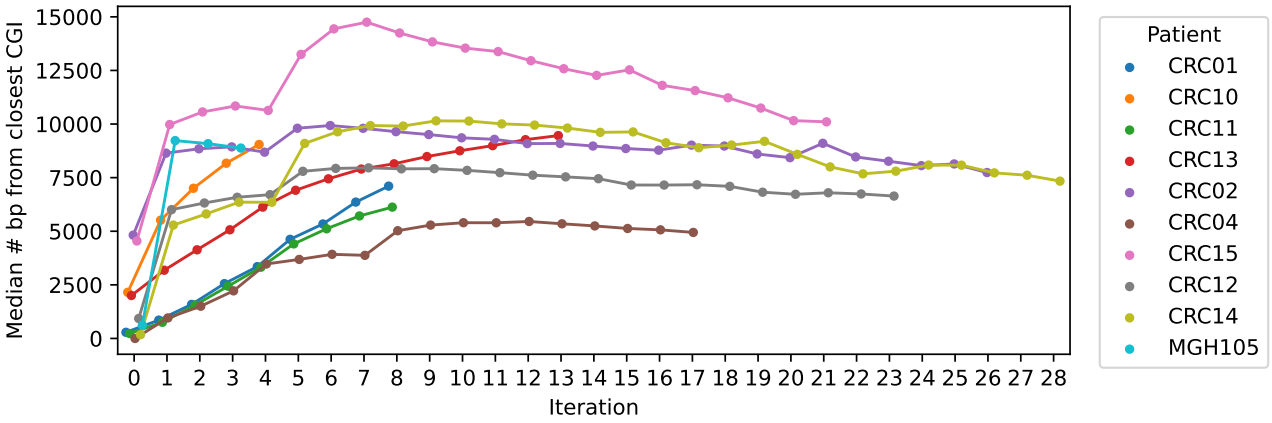


Fig. S12: **Median number of bp to the closest CGI from CpG sites used to construct the methylation lineage tree at each iteration for all CRC patients and GBM patient MGH105.** If a CpG site is within an CGI, it is 0 bp away from the closest CGI.

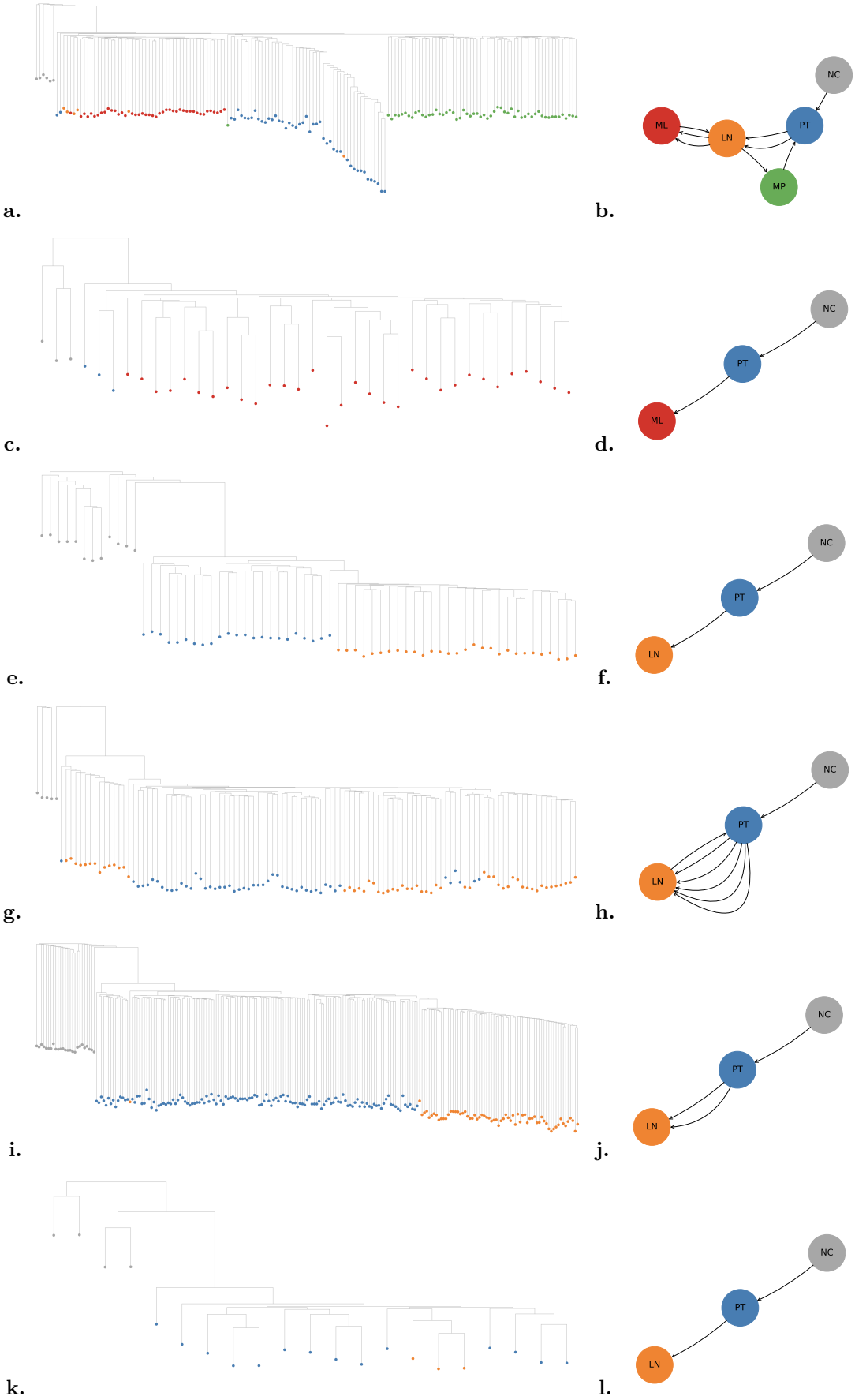
Patient	CpG Island	Inter-CGI
CRC01	0.8182	0.9412
CRC02	NA	0.7620
CRC04	0.7500	0.8689
CRC10	1.0000	0.8571
CRC11	0.9730	0.9200
CRC12	0.95	0.9160
CRC13	NA	0.6
CRC14	0.8824	0.8333
CRC15	NA	NA

Table S4: **Overlap coefficients of GO terms associated with different genomic locations (CpG island and inter-CGI) discovered by GOrilla and DAVID.** CRC02, CRC13, and CRC15 don't have any terms discovered by either GOrilla or DAVID in CpG island. The GO terms associated with inter-CGI regions in CRC15 discovered by GOrilla and DAVID belong to different sub-ontologies (molecular function and biological process, respectively), and are thus not comparable, despite having common themes in neural activities (Figure S10i. and S11i.).

Patient	Spearman's Rho	p-value
CRC01	-0.5977	0.0
CRC02	-0.3006	0.0
CRC04	-0.6224	0.0
CRC10	-0.5827	0.0
CRC11	-0.6369	0.0
CRC12	-0.4700	0.0
CRC13	-0.5354	0.0
CRC14	-0.5922	0.0
CRC15	-0.4431	0.0
MGH105	-0.3907	0.0

Table S5: **Spearman correlation of CpG site distance from closest CpG island and homogeneity**

S4 Application of Sgootr on the Bian *et al.* cohort [2], replacing NJ with FastME 2.0



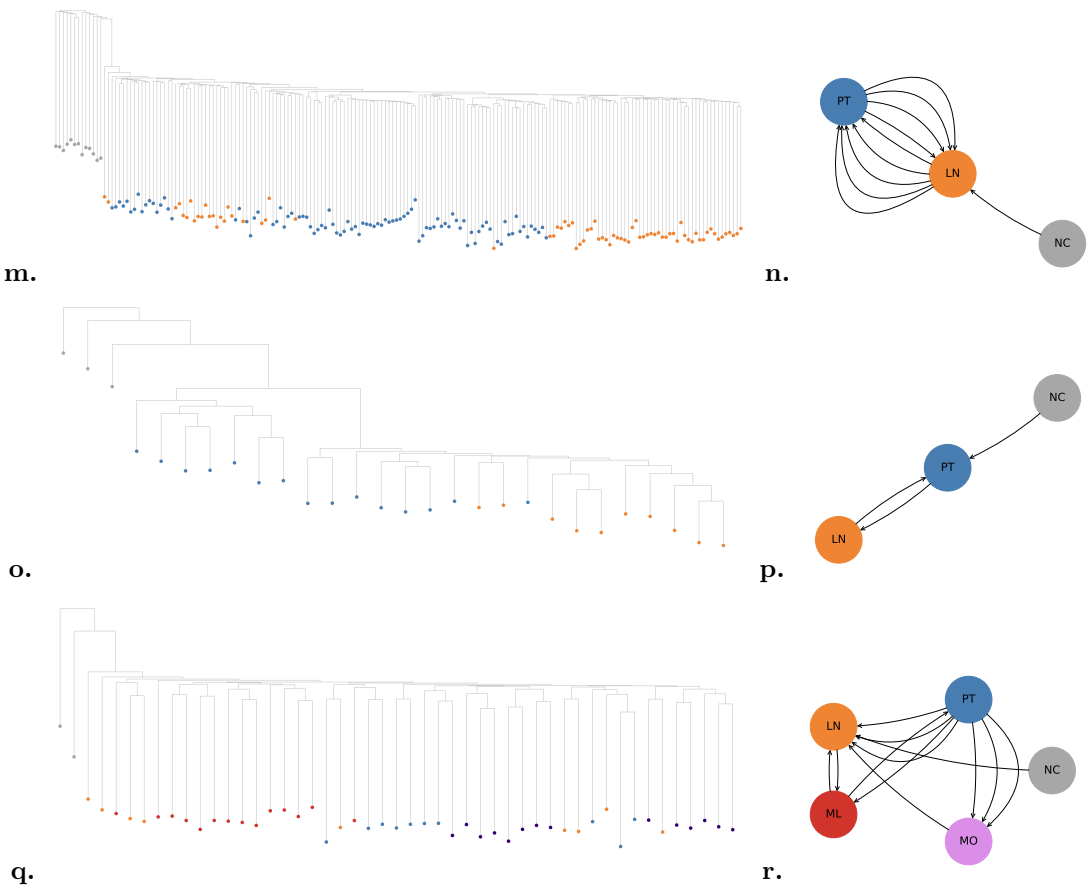


Fig. S13: **Sgootr-constructed ($\kappa = .1$) tumor lineage trees (t^*s) and -inferred tumor migration histories for patients in the Bian *et al.* cohort, replacing NJ with FastME 2.0:** CRC01 (**a.**, **b.**), CRC02 (**c.**, **d.**), CRC04 (**e.**, **f.**), CRC10 (**g.**, **h.**), CRC11 (**i.**, **j.**), CRC12 (**k.**, **l.**), CRC13 (**m.**, **n.**), CRC14 (**o.**, **p.**), CRC15 (**q.**, **r.**)

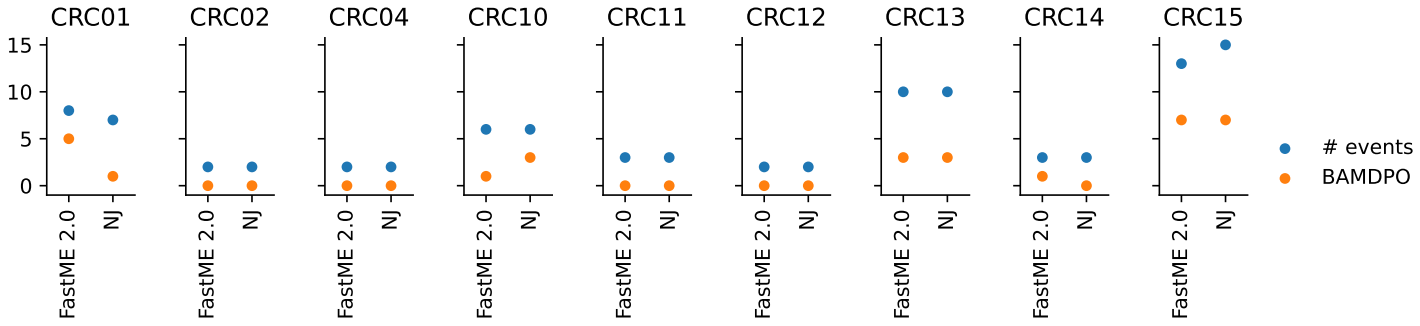


Fig. S14: **Number of migration events and BAMDPO of migration histories obtained via applying Sgootr with FastME 2.0 in comparison to NJ.**

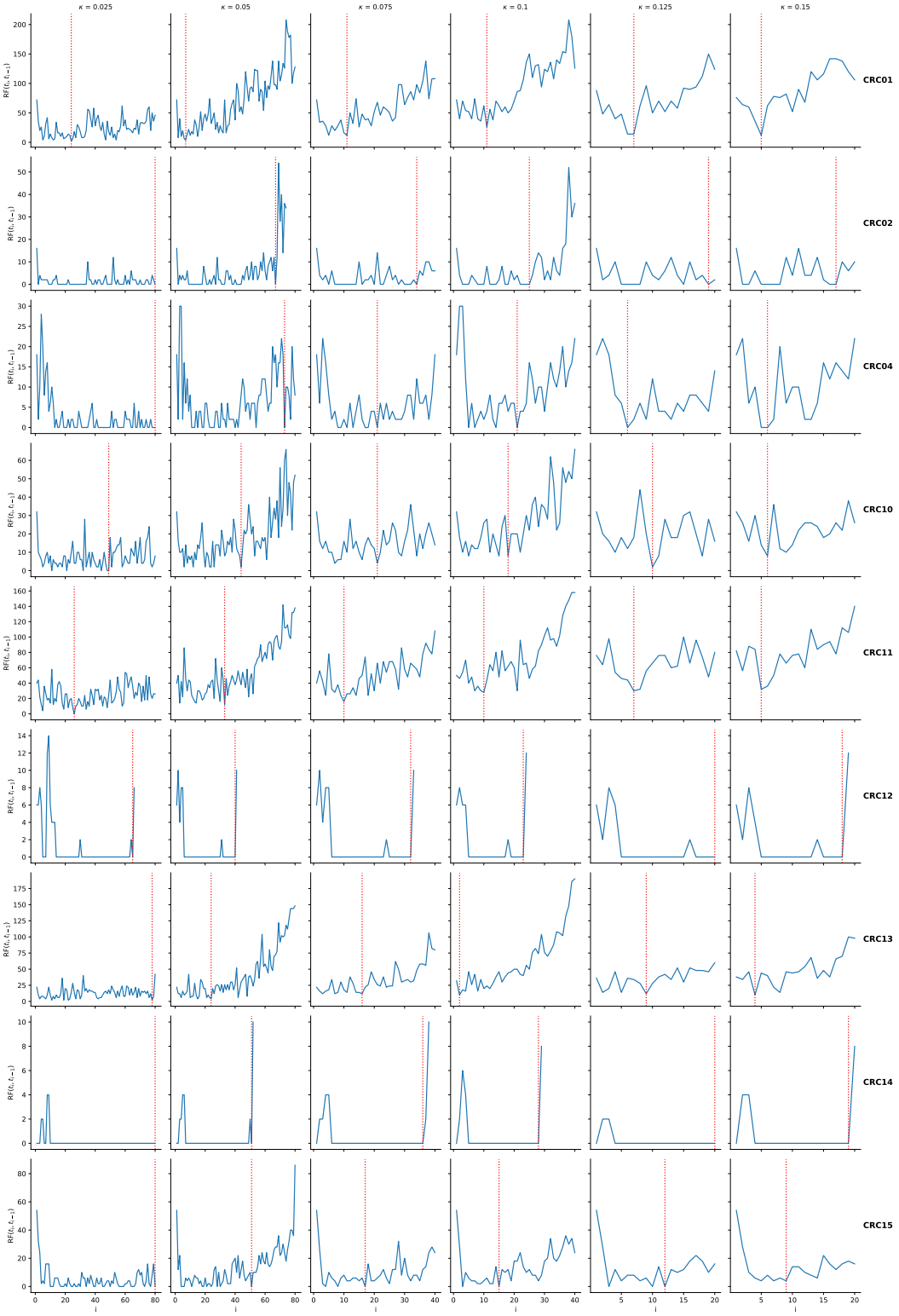


Fig. S15: **RF distances across iterations for each patient in the Bian *et al.* metastatic CRC cohort applying Sgootr - substituting NJ with FastME 2.0 - with different κ values.** Each row represents each of the 9 patients, and each column represents κ=.025, .05, .075, .1, .125, and .15. For κ=.025, .05, the iterative procedure was run for a maximum of 80 rounds; for κ = .075, .1, maximum 40 rounds; and for κ=.125,.15, maximum 20 rounds. Red vertical lines marks the iteration Sgootr identifies as optimal.

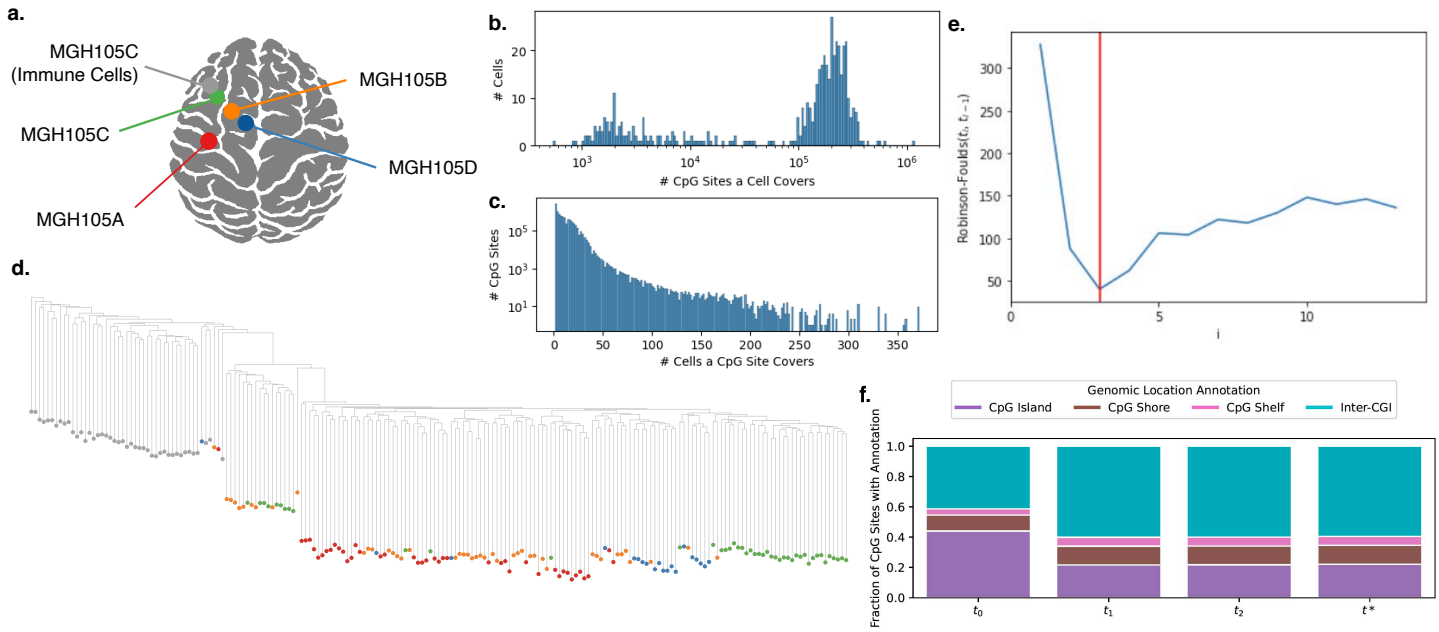


Fig. S16: **Application of `Sgootr` to patient MGH105 MscRRBS data by Chaligne *et al.*** **a.** The multiregional patient data include single cells from 4 distinct sampling locations (in addition to immune cells from sampling location MGH105C) **b.** Cell coverage distribution among autosomal CpG sites in patient MGH105. Note that the x-axis is in log scale. **c.** Autosomal CpG site coverage distribution in patient MGH105. Note that the y-axis is in log scale. **d.** Tumor lineage tree constructed by `Sgootr` with $\kappa = .1$ and each single cell at the leaves colored by its sampling location. **e.** RF distances across 14 iterations, with global minimum occurring at $i = 3$, the tree at which is shown above. **f.** Fraction of CpG sites located in CpG island, CpG shore, CpG shelf, and inter-CGI regions in the MGH105 trees at each iteration of the pruning procedure of `Sgootr`, from t_0 to $t*$.

S5 Application of `Sgootr` to multiregional GBM patient MGH105 MscRRBS data by Chaligne *et al.*[5]

We apply `Sgootr` to the MscRRBS data from MGH105, the only multiregionally sampled GBM patient from the Chaligne cohort [5]. Single tumor cells are obtained from 4 sampling locations within the same lesion: MGH105A, MGH105B, MGH105C, and MGH105D; additionally, non-tumor immune cells are sampled from MGH105C (Figure S16). Since both per cell and total CpG coverage in MscRRBS dataset are lower in comparison to scBS-seq (Figure S16b.,c.), we modify the parameters of `Sgootr` such that it: selects cells covering no fewer than 200,000 CpG sites, selects autosomal sites covered in no less than 5% of the remaining cells, bypasses the post-filtering step, and relaxes the pruning parameters so that $\delta = .02, \omega = .05$. We root the trees with an immune cell from sampling location MGH105C.

`Sgootr` obtains the optimal lineage tree $t* = t_3$ (Figure S16d.,e.), and the clustering implied by the tree topology seems to highly correspond to the sampling locations of the single cells. We attempted to construct a lineage tree on the same set of cells and sites as we use here, with the set of methods Chaligne *et al.* leveraged in the original work[5], namely FG+IQ-TREE; however, FG fail to terminate within 100 hours hence no result is reported.

S6 How to choose p, P(1^c), P(0^c), and P(mixed) for Sgootr

While in our experiments, we default to $p = .5$, P(1^c)=P(0^c)=P(mixed)=.33, it is possible for users to choose other values with additional pieces of prior knowledge. Here we offer examples of when the users may choose other values for those parameters.

The user may set $p > .5$ if there is reasonable belief that, given a pair of alleles truly heterozygous for a CpG site, the sequencing technology will sample from the methylated allele with higher probability than from the unmethylated one. This may manifest in a given dataset as a consistent observation of a higher number of methylated reads than unmethylated reads at diploid heterozygous sites. Similarly, the user may set $p < .5$ if there is reasonable belief that the unmethylated allele is more likely sampled given a heterozygous pair. If strong evidence for such sequencing bias is present, the user should adjust the value of p accordingly.

The user may have P(1^c)=P(0^c)»P(mixed), if there is reasonable belief that true heterozygous CpG sites are rare. In this case, at low-coverage CpG sites, the observation of reads of a single status will give diminished probability that the site is truly heterozygous in the expected distance computation. In another potentially extreme example, the user may set P(1^c)»P(0^c)=P(mixed), if there is reasonable belief that the homozygously methylated status is prevalent, and homozygously unmethylated and heterozygous statuses are equally rare. In this case, at a particular low-coverage CpG site, heterozygous status will only be given diminished probability if all observed reads are methylated; however, if all observed reads are unmethylated, some larger-than-diminished probability will still be given to the heterozygous status.

S7 How to choose κ for Sgootr

In the main text, we present Sgootr results with $\kappa = .1$. Having also experimented with a range of κ values, namely $\kappa = \{.025, .05, .075, .1, .125, .15\}$ (Figure S9,S15), we provide some guidance on how to choose κ for future applications.

Recall that our goal is to detect a point in the iterative procedure where most non-lineage-informative sites have been pruned out but most lineage-informative sites still remain. Translating the objective into measurable terms, we return the last iteration where the global minimum RF distance occurs, after the initial decrease and before the ultimate increase.

When the RF distances across iterations appear flat, or have a region where RF distances have small fluctuations, and iterations where global minimum occur are far from unique - as one can observe across many patients for $\kappa = .025, .05, .075$ columns in Figure S9,S15 - κ is likely too small. With small κ, Sgootr prunes out fewer sites at each iteration. Intuitively, RF distances between trees constructed from more similar sets of sites will be lower, possibly explaining the over all low RF distances in many instances. In those cases, the turning point in the RF distances becomes difficult, if not impossible, to observe, and a higher κ is called for.

While higher κ often lead to clear identification of a unique global minimum in RF distances across iterations - as one can observe across many patients for $\kappa = .125, .15$ columns in Figure S9,S15 - it bares the risk of over-pruning. One may consider attempting lower values of κ while the V-shaped RF distance trend can still be observed.

Of course, determining the maximum number of iterations to run goes hand-in-hand with determining κ. If resource availability allows, we recommend choosing a number of iterations such that a consistent increase in RF distance towards the later iterations can be observed.

S8 Experiment details

In this section, we provide details and parameters used to generate the results presented in this work.

S8.1 Sgootr

While we specifically mention in main text the parameters used to generate the Sgootr results presented in this work, see `config.yaml` at the GitHub repository `https://github.com/algo-cancer/Sgootr` for the full set of default Sgootr configurations in one place.

S8.2 FastME 2.0

In constructing a tumor lineage tree from a distance matrix using FastME 2.0 instead of NJ, we ran FastME 2.0 with the following command: `fastme -i {.mat distance matrix file} -o {.nwk output} -n -s`

S8.3 IQ-TREE

We ran iqtree with the following command, referencing what was done in prior work by Gaiti *et al.* [14] and Chaligne *et al.* [5]: `iqtree2 -s {.phy file} -st BIN -m TESTNEW -mset GTR2 -nt AUTO -bb 1000 -wbtl -abayes -alrt 1000 -nstop 500 -wsr -nm 2000 -v -alninfo -seed {1, 2, 3, 4, 5}`

S8.4 FG

For FG, we follow the method descriptions by Gaiti *et al.*[14]. Given the cell-by-site matrices for methylated and unmethylated reads, we first obtain the binarized methylation status for each covered site in a cell via the procedure described in Section 3. We further prune out sites with coverage in less than 5 cells. For each remaining site s, we randomly downsample the other sites by 100x, and count the number of times s forms four gametes with sites in the downsampled set, disregarding the pair if they are within a read length of each other (150bp for Bian *et al.*, and 50bp for Chaligne *et al.*) or do not have at least 4 cells with coverage for both sites. For each site, we divide the number of occurrences of four gametes by the total number of pairings, obtaining the FG frequency.

The sites are binned into intervals of width 0.05 with regard to its methylation rate (number of cells with methylated status over total number of cells with read coverage for that site), disregarding bins with rates <0.1 or >0.9, and for each bin compute the median FG ratio. Within each bin, sites whose FG frequency is 1.5x median absolute deviation away from the bin median frequency were selected. The aggregate set of sites across all bins form the output of FG.

While random downsampling is necessary for problems of the scale in this work, it also introduces stochasticity. Like Gaiti *et al.*, we perform 5 separate runs of FG for each experiment, and found high concordance of the per site FG frequencies across the runs (Figure S18). For each site, we compute the mean FG frequency across 5 runs to use for the selection process described above (Figure S17).

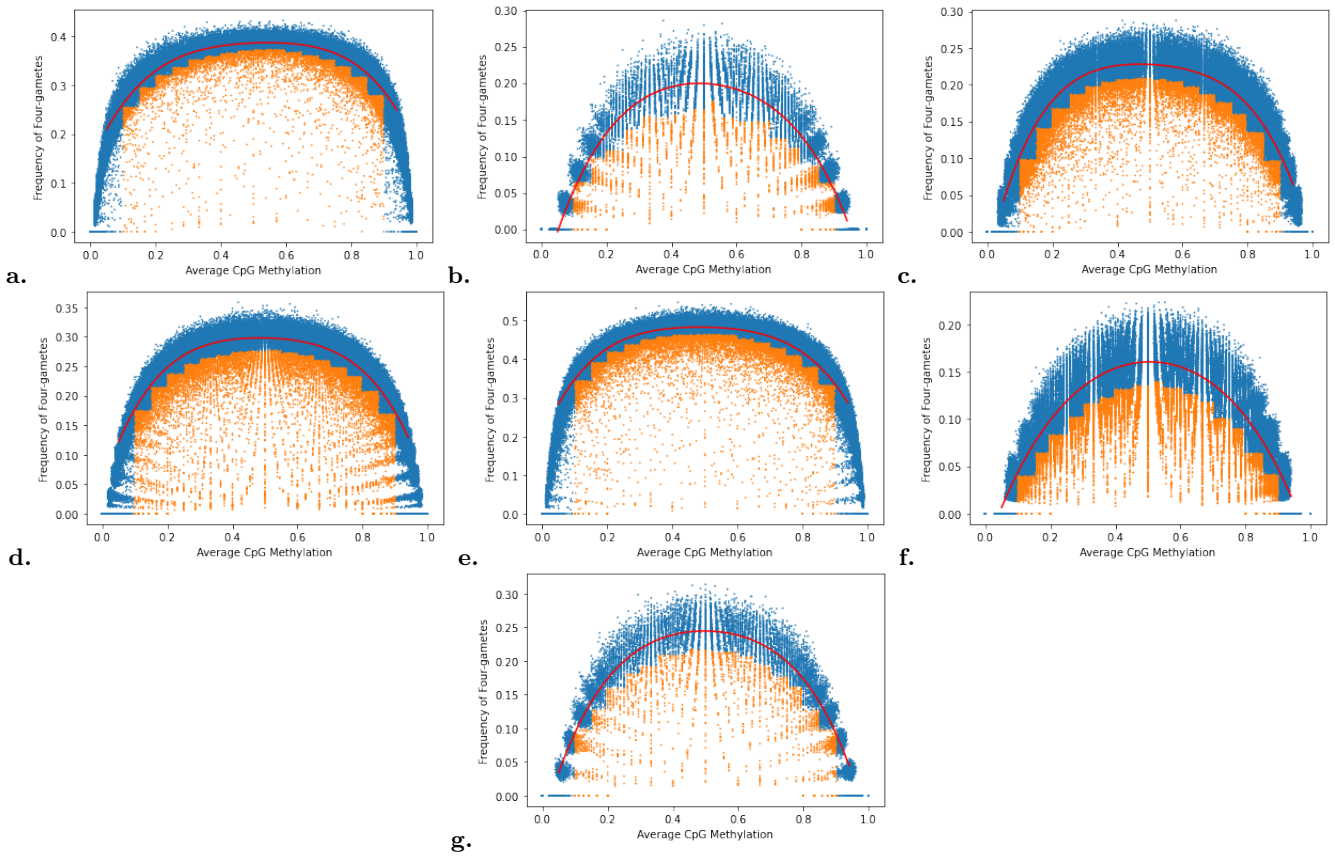


Fig. S17: **FG frequencies (averaged across 5 runs) for a. CRC01, b. CRC02, c. CRC04, d. CRC10, e. CRC13, f. CRC14, g. CRC15.** Selected sites are colored with orange. The red line denote the smoothed line connecting the median FG frequency within each bin.

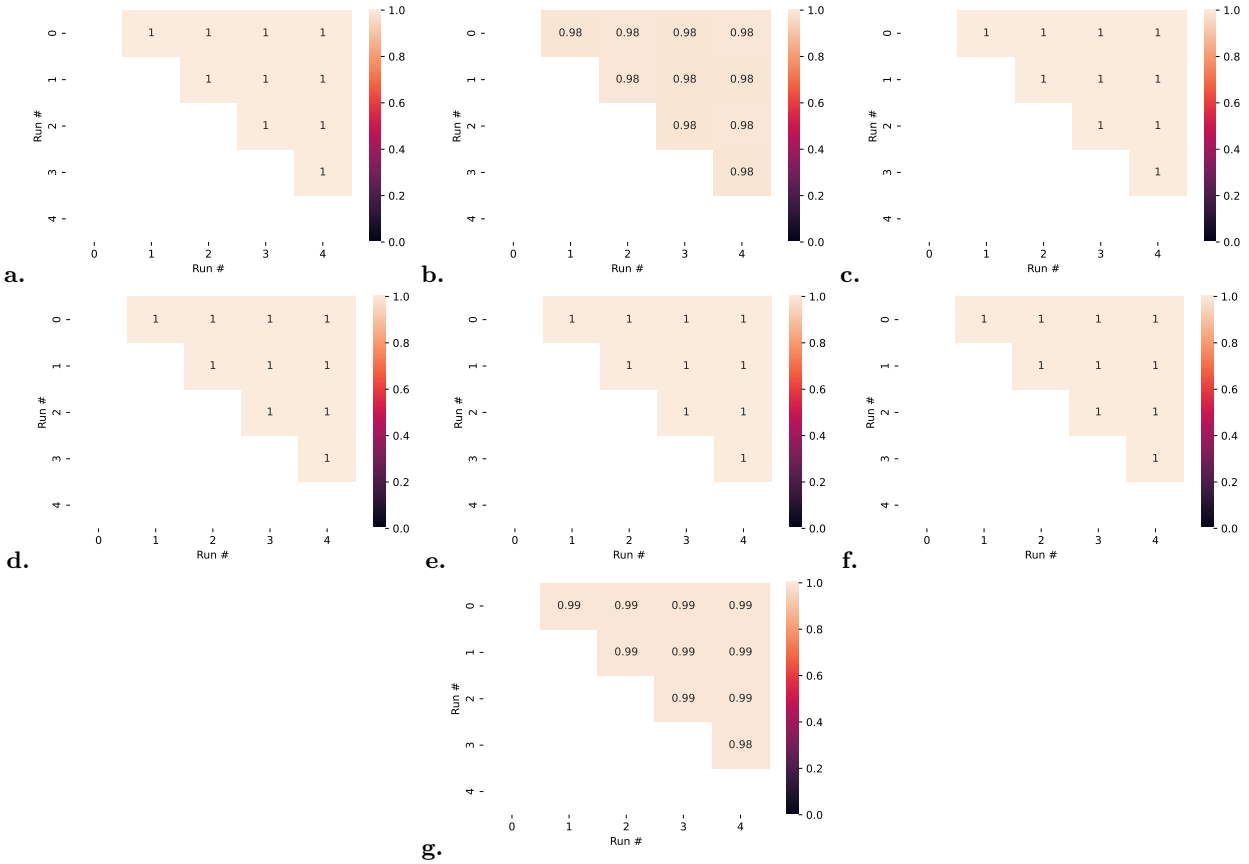


Fig. S18: **Pearson product-moment correlation coefficients between FG frequencies across sites between down-sampled runs for a. CRC01, b. CRC02, c. CRC04, d. CRC10, e. CRC13, f. CRC14, g. CRC15.** Results from each pair of runs from a single experiment exhibit high concordance.