Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data

Camila P.E. de Souza[1†], Mirela Andronescu[2†], Tehmina Masud[2], Farhia Kabeer[2,7], Justina Biele[2], Emma Laks[2,3] Daniel Lai[2], Jazmine Brimhall[2], Beixi Wang[2] Edmund Su[4,5], Tony Hui[4,5], Qi Cao[4,5], Marcus Wong[4], Michelle Moksa[4], Richard A. Moore[6] Martin Hirst[4,5], Samuel Aparicio[2,7*], Sorab P. Shah[2,7,8*]

[1]Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON, Canada
[2]Department of Molecular Oncology, British Columbia Cancer Research Centre, Vancouver, BC, Canada
[3]Genome Science and Technology Graduate Program, University of British Columbia, Vancouver, BC, Canada
[4]Department of Microbiology and Immunology and Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada
[5]Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada
[6]Genome Sciences Centre, BC Cancer, Vancouver, BC, Canada
[7] Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada
[8] Dept of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

*Correspondence: sshah@bccrc.ca, saparicio@bccrc.ca
† Equal contributor

Abstract

We present Epiclomal, a probabilistic clustering method arising from a hierarchical mixture model to simultaneously cluster sparse single-cell DNA methylation data and infer their corresponding hidden methylation profiles. Using synthetic and published single-cell CpG datasets we show that Epiclomal outperforms non-probabilistic methods and is able to handle the inherent missing data feature which dominates single-cell CpG genome sequences. Using a recently published single-cell 5mCpG sequencing method (PBAL), we show that Epiclomal discovers sub-clonal patterns of methylation in aneuploid tumour genomes, thus defining epiclones. We show that epiclones may transcend copy number determined clonal lineages, thus opening this important form of clonal analysis in cancer.
Keywords: single-cell sequencing; DNA methylation; clustering

Introduction

DNA methylation of the fifth cytosine position (5mC) is a well studied epigenetic mark that has decisive roles in the regulation of cell transcriptional programs [1]. In mammals, 5mC occurs mainly at CpG dinucleotides [2] whose distribution is clustered within regions of the genome called CpG islands (CGIs). Bisulfite mediated conversion of 5mC to uracil, referred to as bisulphite sequencing, has been a key tool for the quantification of genome-wide DNA methylation at single-cytosine resolution. Advances in technology and laboratory protocols have allowed the generation of high-throughput sequencing data

of individual cells [3, 4, 5, 6]. In particular, single-cell whole-genome bisulfite sequencing (sc-WGBS) techniques have been developed to assess the epigenetic diversity of a population of cells [7, 8]. Because of the limited amount of DNA material, the generated sc-WGBS data are usually sparse, that is, data from a large number of CpG sites are missing and/or are subject to measurement error. Therefore, there is a great need for the development of statistical and computational methods to cluster cells according to their DNA methylation profiles taking into account the large sparsity of the data.

An increasing amount of sc-WGBS data has been generated from various cell types including: mouse embryonic stem cells [9, 10], human hematopoietic stem cells [7, 11], human hepatocellular carcinoma [12], mouse hepatocytes and fibroblasts [13], human and mouse brain cells [14] and human cell lines [15]. In order to assess the epigenetic diversity in these different cell populations, a variety of non-probabilistic methods have been considered. Smallwood et al. [9] propose a sliding window approach to compute methylation rates of CpG sites across the genome followed by complete-linkage hierarchical clustering considering Euclidean distances and the most variable sites. Angermueller et al. [10] compute the mean methylation levels across gene bodies and as in [9] cluster the cells using hierarchical clustering and only the most variable genes. Farlik et al. [11] cluster cells based on the average methylation over different sets of transcription factor binding sites using also hierarchical clustering. Gravina et al. [13] consider the sliding window approach of [9] to compute methylation rates and use principal component analysis to visually assess clusters of cells. Considering CpG-based Pearson correlation between pairs of cells, Hou et al. [12] use hierarchical clustering to obtain their clusters. Mulqueen et al. [15] use an NMF (Non-Negative Matrix Factorization) approach for dimensionality reduction, followed by t-SNE for visualization and DBSCAN for clustering. Recently, Hui et al. [7] proposed PDclust, a genome-wide pairwise dissimilarity clustering strategy that leverages the methylation states of individual CpGs.

Despite the considerable diversity in clustering approaches, there is still a great need for probabilistic, model-based approaches to simultaneously cluster sc-WGBS data while also inferring the missing methylation states. Because such methods allow for statistical strength to be borrowed across cells and neighbouring CpGs, we surmise they should result in more robust inference than non-probabilistic methods. Moreover, no comparative studies to rigorously evaluate the performance of different methods on a wide variety of data sets has been undertaken.

In this work, we propose Epiclomal, a probabilistic algorithm to cluster sparse CpG-based DNA methylation data from sc-WGBS. Our approach is based on a hierarchical mixture model (see graphical models in Figure 1), which pools information from observed data across all cells and neighbouring CpGs to infer the cell-specific cluster assignments and their corresponding hidden methylation profiles. Epiclomal is part of a novel comprehensive statistical and computational framework (Figure 2) that includes data pre-processing, different clustering methods corresponding to previously proposed approaches [9, 10, 15, 7, 12], plotting and quantitative performance evaluation measures to analyze the results. We use our framework to present an extensive benchmarking of clustering methods over a wide range of previously published and synthetic data sets, plus a novel large scale sc-WGBS data set from breast cancer xenografts [16, 17] generated using state of the art methodology [7].

Results

Overview of Epiclomal

Epiclomal is a clustering method based on a hierarchical mixture of Bernoulli distributions. We are given a sparse matrix of N rows (cells) and M columns (CpG sites), in which each entry is either 0 (unmethylated), 1 (methylated) or missing. The distribution of the observed data X_{nm} for each CpG site m from cell n depends on the latent cell-specific cluster assignment Z_n and corresponding true hidden methylation state (epigenotype) at that CpG, G_{km} (Figure 1a). We use a Variational Bayes (VB) algorithm with random and informed initalizations to infer not only cell cluster assignments but also the true hidden cluster-specific epigenotypes $G_{k1}, \ldots, G_{kM}$ for each cluster k. Full details of our proposed model and VB algorithm are provided in Section 2.1 of the Supplementary Material. We run Epiclomal considering $K = 1, 2, \ldots, 10$ maximum number of clusters and choose the best K along with the best clustering assignments as the combination that minimizes the deviance information criterion (DIC, [18]) via an elbow plot type of selection procedure (see, for example, Supplementary Figure 1 and Section 2.2 of the Supplementary Material).

Epiclomal has two variants: EpiclomalBasic (Figure 1a) and EpiclomalRegion (Figure 1b). While EpiclomalBasic imposes less structure to the model by assuming the true hidden epigenotypes follow the same distribution across all the genomic functional regions considered, EpiclomalRegion allows their distribution to vary across regions to better reflect what we expect to happen in the real data. Bulk data can be used to reassign cells to the EpiclomalRegion clusters using an algorithm that stochastically reassigns cells to clusters while trying to best match the cumulative CpG states of all cells to the corresponding bulk CpG state. This extension is called EpiclomalBulk (Section 2.3 of the Supplementary Material).

Epiclomal is then incorporated into the framework presented in Figure 2, for which an overview is provided in the Methods Section.

Epiclomal outperforms other methods on synthetic data

To evaluate the accuracy and performance of our proposed methods over a wide range of characteristics, we generated a large number of synthetic data sets following the steps presented in Section 3 of the Supplementary Material. We applied our proposed Epiclomal approaches (EpiclomalRegion, EpiclomalBasic and EpiclomalBulk), as well as the four non-probabilistic methods included in our framework (EuclideanClust, DensityCut, HammingClust and PearsonClust) to each generated data set.

We considered several experiments, where in each one we varied one of seven parameters while keeping the others fixed as indicated in Table 1. For each setting, we generated 30 input data sets and computed our five performance measures.

Figure 3 shows the results when we vary the proportion of data missing. Our proposed probabilistic Epiclomal methods give better or comparable V-measures (panel **a**) than the non-probabilistic methods. PearsonClust and HammingClust fail to produce results in the case of 0.95 missing proportion. The Epiclomal methods overall better estimate the true number of clusters ($K = 3$) than the other methods (panel **b**), which tend to overestimate (EuclideanClust) or underestimate (PearsonClust, HammingClust and DensityCut) the number of clusters. Using bulk data via EpiclomalBulk shows improvement in estimating cluster (epiclone) prevalences, especially when the missing data proportion is large (panel **c**). The cluster assignment uncertainty is well estimated by EpiclomalRegion for up to 0.7 missing proportion;

however, it drops rapidly for 0.8 and 0.9 missing proportion. EpiclomalRegion better infers the vectors of cluster-specific methylation profiles than EpiclomalBasic (see also Supplementary Figure 2).

Figure 4 shows that Epiclomal results in better V-measure when compared with the non-probabilistic methods in all the remaining six scenarios (see also Supplementary Figures 3 to 8). All methods perform worse when the problem is more difficult, such as when increasing the number of regions (therefore decreasing the number of different loci among clusters - Figure 4 panel **a**) or increasing the cell-to-cell variability (panel **c**). Increasing the number of cells (panel **b**) does not improve the V-measure values, except for DensityCut, but it does reduce their variability. The Epiclomal methods are more robust to the increase in the number of epiclones (panel **d**) and change in epiclone prevalences (panel **e**). When increasing the number of loci (panel **f**) the performance of HammingClust and PearsonClust remains somewhat constant, while the other methods show a decreasing pattern in performance. However, the Epiclomal methods still perform better for all number of loci considered than all the other methods. Therefore, this provides support to a strategy for selecting a smaller number of loci (under 50 000) in order to keep the true signal and eliminate noise when analyzing a real data set.

Epiclomal recapitulates epiclonal subgroups from public datasets

We assess the performance of our methods on four published sc-WGBS data sets [9, 12, 14, 11] considering as ground truth the major clustering results reported in each paper. Experimental validation of epiclones is often difficult, therefore when working with cells from different known types or treatment conditions authors expected their clusters to somewhat reflect the epigenetic diversity of those types [9, 14, 11]. In [12], there were no predefined cell subpopulations, however the authors consider gene expression and copy number changes to further support their findings. Table 2 shows that these data sets display a variety of characteristics, with number of cells varying from 25 to 2 740 and missing data proportions varying from 0.54 to 0.98.

We applied our framework (Figure 2) to these four real data sets considering three filtered inputs with 10 000, 15 000 and 20 000 loci, respectively (see Methods for details on pre-processing real data). Figure 5 presents the barplots with the results (V-measure, cluster prevalence mean absolute error and predicted number of clusters) obtained for these three inputs for each data set and each clustering method. The bars with "Failure" y-axis value in Figure 5 indicate the cases for which the corresponding method failed to return a result. This may occur because of convergence issues (in the case of Epiclomal or DensityCut), or because of the inability to produce a hierarchical clustering or automatically choose the optimal number of clusters.

In addition, we evaluated the usefulness of selecting regions by interquartile range (IQR) of mean methylation levels by running all the non-probabilistic methods on a large input that filters out only regions with methylation IQR < 0.01 and also plotted the results in Figure 5. These larger input data sets (see number of CpGs in the seventh column of Table 2) can be advantageous for the distance-based methods (EuclideanClust, HammingClust and PearsonClust), resulting in more common positions with observed data between every pair of cells, and hence fewer missing entries in the dissimilarity matrix. However, having a large input may obscure the signal that can best separate the clusters. We therefore performed a large input data evaluation for all published data sets except Luo2017 [12] due to its complexity (2 740 cells with missing proportion of 0.94 for the 10 000 loci input data).

In what follows we present the results obtained on each of the four data sets.

The Smallwood2014 data set [9] comprises 32 mouse embryonic stem cells, where 20 cells were cultured

in a regular serum medium and 12 cells in a 2i medium inducing hypomethylation. We consider as ground truth the two clusters predicted by [9]. The first cluster contains all 2i cells plus two "2i-like" regular serum cells (namely Ser3 and Ser6). The second cluster is formed only by regular serum cells. PearsonClust correctly clustered all cells on the filtered input data sets (V-measure = 1, Figure 5a and Supplementary Figure 9), while EpiclomalRegion and the other non-probabilistic methods misclassified one or two cells (see Supplementary Figures 10 to 13).

The Hou2016 data set [12] contains only 25 cells from a human hepatocellular carcinoma tissue sample. We consider as ground truth the two subpopulations identified by [12] based not only on DNA methylation but also on copy number and gene expression data. EpiclomalRegion correctly assigns all cells to their corresponding subpopulation (V-measure = 1, Figure 5**a** and Supplementary Figure 14). Figure 5 also shows that EuclideanClust and PearsonClust performed 64% and 15% better, respectively, on the large input data than on the average V-measure of the filtered data sets. HammingClust achieved perfect V-measures for the large and 20 000 loci inputs. See also Supplementary Figures 15 to 18.

The Luo2017 data set [14] is challenging as it contains the largest number of cells and with very sparse data (2 740 human neurons with missing proportion of 0.94 for the 10 000 loci input data). As ground truth we considered the two major clusters found by [14], one corresponding to inhibitory neurons and the other one to excitatory neurons. EpiclomalRegion was the only method able to correctly find two major clusters with some misclassified cells (V-measures from 0.51 to 0.61, Figure 5**a** and Supplementary Figure 19). A few of these misclassified cells present clustering assignment probabilities close to 0.5, indicating that for EpiclomalRegion those cells could belong to either cluster. The non-probabilistic methods either inferred the wrong number of clusters (in the case of EuclideanClust and DensityCut), or failed to produce results due to the large amount of missing data.

The Farlik2016 data set [11] contains different types of human hematopoietic cells, totalizing 122 cells. We considered as ground truth the three major clusters found by [11]: one of hematopoietic stem cells (HSC) and two comprising progenitor cell types (myeloid, multipotent and lymphoid progenitor cells). We found that there was not enough signal to separate between the latter two clusters and, therefore, all methods performed poorly in terms of V-measure (Figure 5**a** and Supplementary Figures 20 to 24). EpiclomalRegion, however, most correctly estimated the number of clusters and epiclone prevalences (Figures 5**b** and 5**c**). We surmise that obtaining one aggregated methylation measure over a large number of regions, as performed by [11], probably enhanced the signal for clustering in this particular application.

Epiclomal reveals copy number dependent and copy number independent epiclones in breast cancer

Having verified the performance of Epiclomal on synthetic data and public domain datasets, we set out to perform epiclone group discovery on single cell epigenomes generated in house on a range of patient derived breast tumour xenografts. First, to illustrate scalability of Epiclomal with aneuploid single cell cancer epigenomes, we analysed 558 tumour xenograft single epigenomes from three different breast cancer patients (see Supplementary Table 1), sequenced using the PBAL method [7]. We reasoned that inter-patient CpG methylation differences would dominate over intra-patient CpG methylation and therefore anticipated at least three major epigenetic groups. We observed that EpiclomalRegion correctly classified all cells (Figure 5**a**) for the 15 000 and 20 000 loci scenarios (V-measure = 1, Figure 6) and only one cell was misallocated to a different patient for the case with 10 000 loci (V-measure = 0.99). The distance-

based methods (EuclideanClust, HammingClust and PearsonClust) produced a hierarchical clustering of this dataset but failed to automatically choose the best number of clusters for all filtered input data sets. The PearsonClust method failed to obtain the hierarchical clustering for the 10 000 loci scenario (Supplementary Figures 25 and 26). For the larger input data set, EuclideanClust and HammingClust correctly clustered all cells (V-measure = 1) and PearsonClust produced a hierarchical clustering but failed to automatically choose the best number of clusters (Supplementary Figure 27). DensityCut obtained better results (Figure 5) for the filtered input data sets than for the large input (see also Supplementary Figure 28).

We next focused our analysis on one of the three patient-derived xenografts (PDX) above, which was previously characterized with whole-genome sequencing (WGS) [17] and single-cell WGS [4] (patient SA501 in Supplementary Table 1). Breast cancers often exhibit whole chromosome gains and losses (in addition to sub-chromosomal aneuploidy), especially of the X chromosome, which provides a strong methylation signal. As previously described, this PDX undergoes copy number clonal dynamics between passages, with clones losing one copy of X eventually dominating the populations of later passages. Patient tumour cells at diagnosis were mouse xenografted and serially transplanted over generations. Then, sc-WGBS data from passages 2, 7 and 10 were generated using the PBAL protocol [7]. After filtering out cells that did not pass quality control upon alignment (see Methods), we obtained a final sc-WGBS data set of 244 single cells over 3 passages. We considered as initial feature regions the set of differentially methylated CpG islands found when comparing bulk BS-seq data from passages 1 and 10 (see Methods). We then applied non-negative matrix factorization to the region mean methylation data of all 244 cells (NMF - [19, 20, 15]) as a feature selection strategy obtaining a final input set of 94 regions (see Supplementary Table 2 for their coordinates). Over all 94 regions (Supplementary Figures 30**a** and 30**b**) chromosome X contained the most differentially methylated regions of any single chromosome (29 out of 94; Supplementary Figure 29).

Using these 94 features, EpiclomalRegion clustered the cells (Figure 7**a**) into four epiclones: two primarily containing passage 2 cells, and two containing a mix of passage 7 and 10 cells (EpiclomalBasic produced the same results). The distribution of posterior cluster assignment probabilities (p) indicates most cells were classified with p>0.9, with the exception of two cells that were assigned to Cluster 3 with probabilities of 0.73 and 0.69. For completeness we also attempted to evaluate epiclones with the non-probabilistic clustering methods included in our framework (see Supplementary Material Section 4); however as noted above, these methods either failed to execute completely, or failed to identify an optimal number of clusters due to the large amount of missing data.

Inspection shows that Cluster 1 contains 10/40 passage 2 cells (and 1 passage 7 cell) with unmethylated features, except for chromosome X regions, which are mainly methylated. Cluster 2 contains the remaining 30/40 passage 2 cells (and one passage 10 cell), but with unmethylated chromosome X regions, consistent with X inactivation partitioning of the X chromosomes. At later passages, several autosomal regions became methylated (see, for example, chromosomes 1, 9, 12, 19 and 20, Supplementary Table 3). In addition, we identified three main regions that are methylated only in some of the later passage cells (see also Supplementary Table 3), resulting in two different epiclones each containing a mix of passage 7 and 10 cells (cluster 3 containing 53/98 passage 7 cells and 34/106 passage 10 cells; and cluster 4 containing 44/98 passage 7 cells and 71/106 passage 10 cells).

The observations above suggest that some chromosomal regions, such as X, may show strong copy number influence on CpG states, whereas others may differ in CpG state, but unrelated to copy number

state in aneuploid genomes. Therefore we next investigated possible correlations between methylation and copy number alterations derived from the same sc-WGBS data (see Methods). A systematic comparison shows that indeed the average methylation levels and copy number states across cells for each of the 94 regions (Figure 7**b**), are only highly correlated (Pearson correlation >0.5; Figure 7**c**) for the X chromosome. This implies that epiclones may transcend copy number defined clones.

Indeed, when we compared the four epiclones with the four sc-PBAL copy number (CN) clones we noticed that they can match or transcend each other as follows (Figure 7**b**): epiclone 1 with methylated regions in the X chromosome matches exactly CN clone I having two copies of the X chromosome, which shows a strong relationship between the presence of the second copy of the X chromosome and the methylation pattern. However, 26/31 passage-2 cells with all 94 regions unmethylated from epiclone 2 are found in CN clone III, which also contains 47/87 cells from epiclone 3 and 6/115 cells from epiclone 4, even though these have several regions that are methylated. Finally epiclone 3 transcends CN clones III (47/87 cells from epiclone 3 are in CN clone III) and IV (27/87 cells from epiclone 3 are in CN clone IV); and epiclone 4 transcends CN clones II (72/115 cells from epiclone 4 are in clone II) and IV (29/115 cells from epiclone 4 are in CN clone IV). Taken together, these data show for the first time with single cell methylation analysis that epigenetically defined clones may present a different lineage to that of copy number defined clonal architecture, opening up this form of analysis for cancer genomes.

Discussion

Single cell CpG genome analysis is currently held back by a dearth of principled methods for handling the features of single cell methylation data. To this end, we have developed Epiclomal, a probabilistic CpG-based clustering method for clustering sparse sc-WGBS data and elucidation of epigenetic diversity on different types of cell populations. Our method has produced overall better results than non-probabilistic based methods when tested on synthetic data from seven extensive simulation scenarios (Figures 3 and 4) and five comprehensive real data sets (Figure 5). Importantly, Epiclomal is robust and reliable when the amount of data missing is large and/or varies across cells, and can find the true clusters and epiclone prevalence when the signal is subtle, both limiting features of current sc-WGBS data.

It is well understood that 5mC distribution in the genome is regionally clustered and this has implications for computational methods. EpiclomalRegion is the first approach that considers CpG-based methylation dependencies in functional regions and models errors while simultaneously assigning cells to clusters and imputing the missing data. It is also the first method that can use bulk DNA methylation data to improve the epiclone prevalences, an important measure particularly to the study of cancer tumour composition. Epiclomal works at the CpG level and thus considers the contribution of every sequenced CpG site in the selected regions, without the loss of information by region averaging. Epiclomal does not only run an uninformed clustering method, but uses the clustering results of four other methods (with more easily added) and a robust input data selection strategy to return the best prediction. Epiclomal can provide reliable measures of uncertainty of cell-to-cluster assignments, as exemplified in the large Luo2017 data set. Our probabilistic method Epiclomal is versatile in that it can be easily extended to consider additional data structures and information, such as multiples samples and grouping of different functional region types.

Epiclomal is part of an extensive statistical and computational framework that provides interpretable results and five different performance measures. Since not all CpG sites will exhibit variation and are

therefore uninformative, our framework includes a pre-processing step that allows the selection of specific regions in an effort to increase the signal and eliminate noise in the input data. Even the non-probabilistic methods obtained better results, for example, on the smaller filtered Smallwood2014 input data sets than on the large input, supporting the notion that filtering out the most invariable regions may better separate the true signal. In addition, our synthetic experiments as well as the SA501 intrapatient analysis on a well-designed set of differentially methylated regions showed that pre-processing the initial whole-genome data set in a way that keeps the clone differences and eliminates noise is likely to produce better results overall. Future work includes a region selection strategy that can better achieve this goal.

The Epiclomal methods are more computationally expensive than the non-probabilistic methods, but they can provide some advantages in better prediction of the number of clusters (see, for example, Figure 5**c** and panel **c** of Supplementary Figures 2 to 8), cell to cluster assignment (see V-measure in, for example, Figure 4) or epiclone prevalence prediction (see Figures 3**c** and 5**b**). EpiclomalRegion is more computationally intensive than EpiclomalBasic, but because it imposes more structure into the data modelling it leads to overall better results (see, for example, Supplementary Figures 5**a**, 5**c** and panel **i** of Supplementary Figures 2 to 5).

Although epigenomic states are of importance in cancer biology, to date very few single-cell whole genome bisulfite datasets have been generated on aneuploid cancer genomes. Here we used Epiclomal with a large (598 genomes) new dataset of sc-WGBS generated by the recently published PBAL method, to demonstrate how epiclones and copy number determined clones differ. Epiclomal was able identify known and novel CpG methylation substructure that could not be identified by non-probabilistic distance based methods, due to the missing data inherent in sc-WGBS. Specifically, the separation between the two passage 7/10 subclusters, was not found by any of the non-probabilistic methods we considered, nor when a larger set of regions was used. This demonstrates that sophisticated modeling of the data missingness and appropriate region selection are necessary in order to clearly separate biological signal when this signal is not so abundant.

The ability to identify CpG defined sub-clones, epiclones, allowed us for the first time to compare a copy number determined lineage with an epigenetically defined lineage. It is expected that for certain regions of the genome, for example where allelic hemi-methylation occurs, changes in chromosomal copy number would strongly pattern 5mC CpG status. Indeed, we were able to observe this with subclones of a breast cancer PDX (SA501) where biallelic X chromosome clones present in early passages contain epiclones with and without CpG methylation, whereas for autosomes the copy number relationship is much weaker. In contrast, we observe that clones defined by autosomal copy number aberrations can exhibit quite distinct epiclone structure, leading to the notion that in some cases epiclone defined lineage will transcend that of copy number defined lineage. This has important implications for the study of cancer evolution and clonal states, as a failure to include epigenetic states will under-represent the cellular population structures of interest. Further work is required to define the scope and nature of epiclone versus copy number clone defined cellular lineages in cancer.

Methods

Overview of proposed framework

Figure 2 describes our statistical and computational framework, which can be divided into three main parts: input data and pre-processing, clustering and output and performance measures. In what follows we present an overview of each one of those parts.

Input data and pre-processing
Our framework can take as input either real or synthetic data. For real data we take files with methylation calls across the genome from individual cells and extract data for a certain set of regions of interest (e.g., CpG islands, gene bodies, differentially methylated regions, etc). One can consider the data from all regions of interest or use our region selection pipeline to select a subset of the data. To do this, the proposed pipeline first keeps the regions with a certain amount of coverage across all or in a subset of cells and then selects regions with the most variable methylation levels (via interquartile range - IQR), controlling or not for a desired number of loci. If available, our pipeline also takes as input bulk methylation data that can be used to improve results.

For synthetic data we provide a pipeline that generates single-cell methylation data considering various parameters (e.g., missing proportion, number of cells, number of loci, etc.) assuming that the true cluster methylation profiles arise from a phylogenetic tree with only one region changing methylation states at each new cluster generation (see Supplementary Material Section 4).

This first part of our framework ends with the production of input files in the appropriate format for clustering.

Clustering
The clustering step of our framework takes as input the methylation calls for the loci in regions selected in the pre-processing step along with the genomic coordinates of each region. We first cluster the cells considering different non-probabilistic clustering methods, whose results will be then used as initial values for Epiclomal as well as for comparison purposes. There are two types of non-probabilistic clustering methods: region and CpG based. In the region-based approaches we cluster cells considering as input the mean methylation level of each region, whereas in the CpG-based approaches we consider the methylation status of each individual CpG. EuclideanClust is one of the region-based approaches and consists of a hierarchical clustering procedure with Euclidean distances. Another region-based approach is DensityCut [21], which is a density-based clustering method. HammingClust and PearsonClust are both CpG-based approaches using hierarchical clustering with dissimilarities based on Hamming distances and Pearson correlation values, respectively. Section 1 of the Supplementary Material presents more details on these methods.

After running the non-probabilistic based methods, our pipeline applies Epiclomal to the data taking the results of the previous methods as initial values along with a set of random initials (total of 1000 initializations). We run Epiclomal considering $K = 1, 2, \ldots, 10$ maximum number of clusters and choose the best K and cluster assignments using the DIC-elbow criterion. If available, bulk data can be used by EpiclomalBulk to improve the results.

Output and performance measures

For all clustering methods we obtain as results the inferred number of clusters, the cell-to-cluster assignments and the cluster (or epiclone) prevalences (i.e., the proportion of cells in each cluster). In addition, for Epiclomal we obtain the estimated cluster-specific vectors of epigenotyes and the cell-to-cluster assignment probabilities.

To fairly evaluate the performance of each method when true cluster assignments are available, our framework includes several metrics. For number of clusters we consider the mean predicted number of clusters. For cluster prevalence we calculate the mean absolute error (MAE), that is, the mean absolute difference between true and inferred values. For cell-to-cluster assignments we consider the V-measure [22]. In addition, for synthetic data and Epiclomal we consider the hamming distance between true and inferred vectors of methylation states. We also compute for Epiclomal the uncertainty true positive rate of cluster assignment probabilities, that is, how well the uncertainty is estimated for cells whose membership is unclear due to missing data (Section 2.4 of the Supplementary Material).

Pre-processing of the real data

We pre-process real data sets using the first part of our proposed framework. For each data set, we started considering all regions of corresponding type presented in the fifth column of Table 2. Then, after eliminating the empty regions across all cells, we also removed regions with an average missing proportion across all cells greater than or equal to 95%. Next, we kept the most variable regions (as measured by IQR of mean methylation levels) that would produce three filtered inputs with 10 000, 15 000 and 20 000 loci, respectively.

In-house sc-WGBS data generation

Biospecimen collection and ethical approval

Tumour fragments from women diagnosed with breast lump undergoing surgery or diagnostic core biopsy were collected with informed consent, according to procedures approved by the Ethics Committees at the University of British Columbia. Patients in British Columbia were recruited and samples collected under tumor tissue repository (TTR-H06-00289) protocol that falls under UBC BCCA Research Ethics Board.

Tissue processing

The tumor materials were processed as mentioned in [17]. Briefly, the tumor fragments were minced finely with scalpels then mechanically disaggregated for one minute using a Stomacher 80 Biomaster (Seward Limited, Worthing, UK) in 1-2 mL cold DMEM-F12 medium. Aliquots from the resulting suspension of cells and clumps were used for xenotransplants.

Xenografting

Xenograft samples were transplanted and passaged as described in [17]. Female immune compromised, NOD/SCID interleukin-2 receptor gamma null (NSG) and NOD Rag-1 null interleukin-2 receptor gamma null (NRG) mice were bred and housed at the Animal Resource Centre (ARC) at the British Columbia Cancer Research Centre (BCCRC) supervised by Aparicio lab. Surgery was carried out on mice between the ages of 8-12 weeks. The animal care committee and animal welfare and ethical review committee,

the University of British Columbia (UBC), approved all experimental procedures. For subcutaneous transplants, mechanically disaggregated cells and clumps of cells were resuspended in 100-200μl of a 1:1 v/v mixture of cold DMEM/F12: Matrigel (BD Biosciences, San Jose, CA, USA). 8-12 weeks old mice were anesthetised with isoflurane, then the cell/clumps suspension was injected under the skin on the flank using a 1ml syringe and 21gauge needle.

Histopathological review

On histopathological review, two out of three, i.e. SA501 and SA609, patient derived xenografts (PDX) used in this study are triple negative breast cancers (TNBC). On immunohistochemistry, they were found to be receptor negative breast cancer subtype. SA532 is ER-PR-HER2+ xenografts. A pathologist reviewed the slides.

Cell preparation and dispensing

Xenograft tissues were dissociated to cells as described in [4] before dispensing single cells into the wells of 384 well plates using a contactless pieoelectric dispenser (sciFLEXArrayer S3, Scienion) with real-time cell detection in the glass capillary nozzle (CellenOne).

sc-WGBS experimental protocol

The Post-Bisulfite Adapter Ligation (PBAL) protocol described in [7] was used to obtain our in-house sc-WGBS data.

Data alignment and methylation calls

One lane of paired end sequencing was used to create each single cell library. Trim Galore (v0.4.1) and Cutadapt(v1.10) was used for quality and adapter trimming. Libraries were aligned to GRCh37-lite reference using Novoalign (v3.02.10) in bisulphite mode, and converted to BAM format and sorted using Sambamba (v0.6.0). Bam files were annotated for duplicates using Picard Tools MarkDuplicates.Jar (v1.92). Novomethyl (v1.10) was used in conjunction with in house scripts (samtools v1.6 and bedtools v2.25.0) to determine methylation of each CpG as described in Section "NovoMethyl - Analysing Methylation Status" Section of the Novoalign documentation (http://www.novocraft.com/documentation/novoalign-2/novoalign-user-guide/bisulphite-treated-reads/novomethyl-analyzing-methylation-status/).

Quality control

Using an in-house script, libraries were filtered according to a delta CT and 100K read count threshold to account for suitable library depth. Libraries over the expected number of CNV, were filtered out to control for chromothripsis and shattered cells.

Copy number calling

Copy number changes for SA501 were called using the same sc-WGBS DNA methylation data (copy number calling from the DLP protocol [4] largely match the sc-WGBS copy number calling for passage 2, results to appear). Control Free-c (v7.0) was used to copy number variant call on processed BAMs. The

following settings were used: ploidy : 2, window and telocentromeric : 500000, sex : XY, minExpectGC : 0.39, and maxExpectedGC: 0.51.

In-house bulk whole genome bisulfite sequencing (SA501 passages 1 and 10)

Whole genome bisulfite library construction for Illumina sequencing

To track the efficiency of bisulfite conversion, 10 ng lambda DNA (Promega) was spiked into 1 μg genomic DNA quantified using Qubit fluorometry and arrayed in a 96-well microtitre plate. DNA was sheared to a target size of 300 bp using Covaris sonication and the fragments end repaired using DNA ligase and dNTPs at 30°C for 30 min. Repaired DNA was purified using a 2:1 AMPure XP beads to sample ratio and eluted in 40 μL elution buffer in preparation for A-tailing; the addition of adenosine to the 3 end of DNA fragments using Klenow fragment and dATP incubated at 37°C for 30 min. Following reaction clean-up with magnetic beads, cytosine methylated paired-end adapters (5'-A^{m}CA^{m}CT^{m}CTTT^{m}C^{m}C^{m}CTA^{m}CA^{m}CGA^{m}CG^{m}CT^{m}CTT^{m}C^{m}CGAT^{m}CT-3' and 3'-GAG^{m}C^{m}CGT-AAGGA^{m}CGA^{m}CTTGG^{m}CGAGAAGG^{m}CTAG-5') were ligated to the DNA at 30°C, 20 min and adapter flanked DNA fragments bead purified. Bisulfite conversion of the methylated adapter-ligated DNA fragments was achieved using the EZ Methylation-Gold kit (Zymo Research) following the manufacturers protocol. Seven cycles of PCR using HiFi polymerase (Kapa Biosystems) was used to enrich the bisulfite converted DNA and introduce fault tolerant hexamer barcode sequences. Post-PCR purification and size-selection of bisulfite converted DNA was performed using 1:1 AMPure XP beads. To determine final library concentrations, fragment sizes were assessed using a high sensitivity DNA assay (Agilent) and DNA quantified by Qubit fluorometry. Where necessary, libraries were diluted in elution buffer supplemented with 0.1% Tween-20 to achieve a concentration of 8nM for Illumina HiSeq2500 flowcell cluster generation.

Data alignment and methylation calls

FASTQ files were trimmed with TrimGalore (0.4.1) and then input into Bismark (0.14.4) aligning with bowtie2 (2.2.6). With the output BAM, we use samtools (1.3) to sort by name, fix mates, sort by position, remove duplicates, then finally sort by name once again and filter out reads with a mapping quality of 10 or less. We then ran the resulting BAM file through the bismark_methylation_extractor script that accompanies Bismark, to call methylation sites. All tools were run on all default settings, with changes made only to increase run speed.

Differentially methylated CpG Islands

Differentially methylated CpG islands between bulk samples from tumour xenograft passages 1 and 10 were obtained via Fisher's exact test considering all CpG islands with coverage greater or equal than five reads. The Benjamini-Hochberg procedure was used to correct for multiple testing.

Availability of data and materials

Our computational code is available online at https://github.com/shahcompbio/Epiclomal. The in-house data is currently available upon request.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

C.P.E.d.S. and M.A. developed and implemented the proposed methodology, performed all the analyses, and wrote the manuscript. T.M., F.K., J.B., E.M., J.B. and B.W. prepared the in-house single cells and bulk samples. Q.C. and M.W. generated the PBAL libraries. M.M. supervised library generation. R.M. supervised library sequencing. E.S., T.H. and D.L. processed the sequencing data and performed QC analyses. S.A. contributed to the manuscript text. M.H. and S.A. contributed ideas during the method development and analysis of the results. S.A. and S.P.S conceived and oversaw the project.

Acknowledgements

We acknowledge generous funding support provided by the BC Cancer Foundation. In addition, SPS receives operating funds from Terry Fox Research Institute (grant 1082) and the Canadian Cancer Society (grant 705636).

References

[1] Smith, Z.D., Meissner, A.: Dna methylation: roles in mammalian development. Nature Reviews Genetics **14**(3), 204 (2013)

[2] Feng, S., Jacobsen, S.E., Reik, W.: Epigenetic reprogramming in plant and animal development. Science **330**(6004), 622–627 (2010)

[3] Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., *et al.*: Tumour evolution inferred by single-cell sequencing. Nature **472**(7341), 90 (2011)

[4] Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S.P., Aparicio, S., Hansen, C.L.: Scalable whole-genome single-cell library preparation without preamplification. Nature methods **14**(2), 167 (2017)

[5] Shapiro, E., Biezuner, T., Linnarsson, S.: Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics **14**(9), 618 (2013)

[6] Gawad, C., Koh, W., Quake, S.R.: Single-cell genome sequencing: current state of the science. Nature Reviews Genetics **17**(3), 175 (2016)

[7] Hui, T., Cao, Q., Wegrzyn-Woltosz, J., O'Neill, K., Hammond, C.A., Knapp, D.J.H.F., 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 (2018). doi:10.1016/j.stemcr.2018.07.003

[8] Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G., Reik, W.: Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome biology **17**(1), 72 (2016)

[9] Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., Kelsey, G.: Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nature methods **11**(8), 817–820 (2014)

[10] Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S.A., Ponting, C.P., Voet, T., *et al.*: Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nature methods **13**(3), 229 (2016)

[11] Farlik, M., Halbritter, F., Müller, F., Choudry, F.A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., *et al.*: Dna methylation dynamics of human hematopoietic stem cell differentiation. Cell stem cell **19**(6), 808–822 (2016)

[12] Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., *et al.*: Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. Cell research **26**(3), 304 (2016)

[13] Gravina, S., Dong, X., Yu, B., Vijg, J.: Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. Genome biology **17**(1), 150 (2016)

[14] Luo, C., Keown, C.L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J.R., Sandoval, J.P., *et al.*: Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science **357**(6351), 600–604 (2017)

[15] Mulqueen, R.M., Pokholok, D., Norberg, S.J., Torkenczy, K.A., Fields, A.J., Sun, D., Sinnamon, J.R., Shendure, J., Trapnell, C., O'Roak, B.J., et al.: Highly scalable generation of dna methylation profiles in single cells. Nature Biotechnology (2018)

[16] DeRose, Y.S., Wang, G., Lin, Y.-C., Bernard, P.S., Buys, S.S., Ebbert, M.T., Factor, R., Matsen, C., Milash, B.A., Nelson, E., *et al.*: Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. Nature medicine **17**(11), 1514 (2011)

[17] Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., *et al.*: Dynamics of genomic clones in breast cancer patient xenografts at single cell resolution. Nature **518**(7539), 422 (2015)

[18] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A.: Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(4), 583–639 (2002)

[19] Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (2007)

[20] Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788 (1999)

[21] Ding, J., Shah, S., Condon, A.: densitycut: an efficient and versatile topological approach for automatic clustering of biological data. Bioinformatics **32**(17), 2567–2576 (2016)

[22] Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)

Varying parameter	Varying range
Missing proportion	0.5 to 0.95
Number of regions	25 to 200
Number of cells	12 to 2500
Cell-to-cell variability	0 to 0.3
Number of clusters (epiclones)	1 to 10
epiclone prevalence	balanced to very unbalanced
Number of loci	5 000 to 500 000

Table 1: The varying parameters and their ranges for the synthetic data simulation. For each experiment, we varied one parameter and kept the other ones fixed. Note that varying the number of regions is equivalent to varying region size as the total number of loci is fixed. Unless otherwise specified, the fixed parameters are: missing proportion 0.8, region size 100, number of cells 100, proportion of cell to cell variability 0, number of epiclones 3, equal epiclone prevalences (0.33) and number of loci 10 000. For the cell-to-cell variability experiment (Figure 4c) we used 25 regions in order to have a larger number of loci that differ between clusters. For the number of epiclones experiment (Figure 4**d**) and the epiclone prevalence experiment (Figure 4**e**) we used 500 cells in order to allow for enough cells to be represented in each case.

Data set	Cell type	# cells	# clusters	Regions	Miss 10K	Nloci IQR $\geq$ .01	Miss IQR $\geq$ .01
Smallwood2014 [9]	mouse embryonic stem cells	32	2	CpG Islands	0.69	786 620	0.54
Hou2016 [12]	human hepatocellular carcinomas	25	2	CpG Islands	0.87	255 136	0.90
Luo2017 [14]	human brain cells	2740*	2	Gene bodies	0.94	15 788 633	-
Farlik2016 [11]	human hematopoietic stem cells	122	3	TFBS	0.89	512 153	0.98
InHouse (3 patients)	human xenografted cancer cells	558	3	CpG Islands	0.82	1 019 956	0.79

Table 2: A summary of the real data sets used in this work. Column descriptions (in order of appearance) are as follows: (1) data set names corresponding to four published data sets and one new in-house data set; (2) the type of cells in each data set; (3) the number of cells considered for each data set, these vary from tens to thousands of cells; (4) the number of clusters, as reported in the respective published papers - only major clusters were considered, or as expected for the InHouse data set; (5) the genomic functional regions considered for each data set, these were the same as in the original papers when applicable, TFBS stands for Transcription Factor Binding Sites; (6) the missing proportion for each data set considering the 10 000 loci filtered input, varying from 0.69 to 0.94; (7) the number of loci for the largest input data sets obtained by including all regions with methylation IQR $\geq$ 0.01, these vary from 1/4 million to 1 million CpG sites; (8) the missing proportion for the largest input data sets, these vary from 0.54 to 0.98 (not performed for the Luo2017 data set). *44 cells with unknown true clustering assignments were excluded.

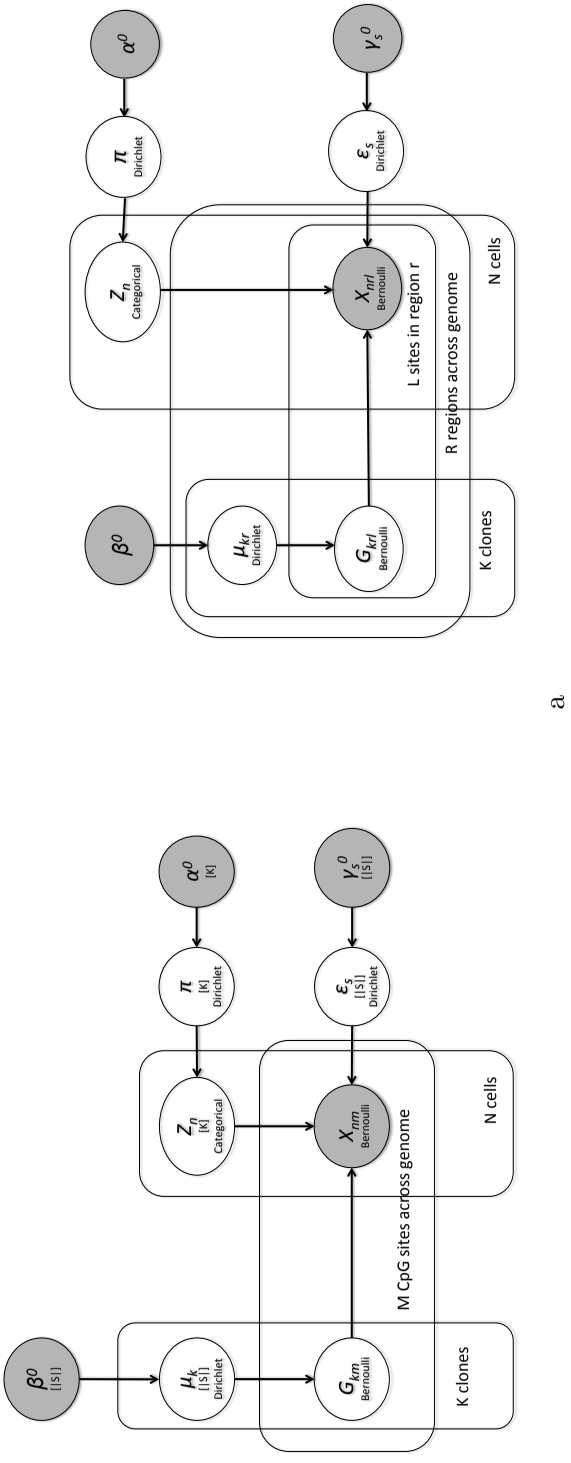

Figure 1: *Our proposed Epiclomal graphical models.* The shaded X nodes are observed variables. The unshaded Z and G nodes are latent variables corresponding to the hidden cell-to-cluster assignments and true epigenotypes, respectively. The unshaded μ, π and ϵ nodes correspond to the unknown model parameters, which under a Bayesian approach have prior distributions with fixed hyperparameters described in the shaded nodes with the 0 superscript. The distribution assumed for each variable/parameter is written within its node. The edges of the graphs depict dependencies. The plates depict repetitions. (a) EpiclomalBasic, true hidden epigenotypes share the same probability distribution across all CpG sites in the same epiclone. (b) EpiclomalRegion, true hidden epigenotypes for CpG sites from different regions share different probability distributions.

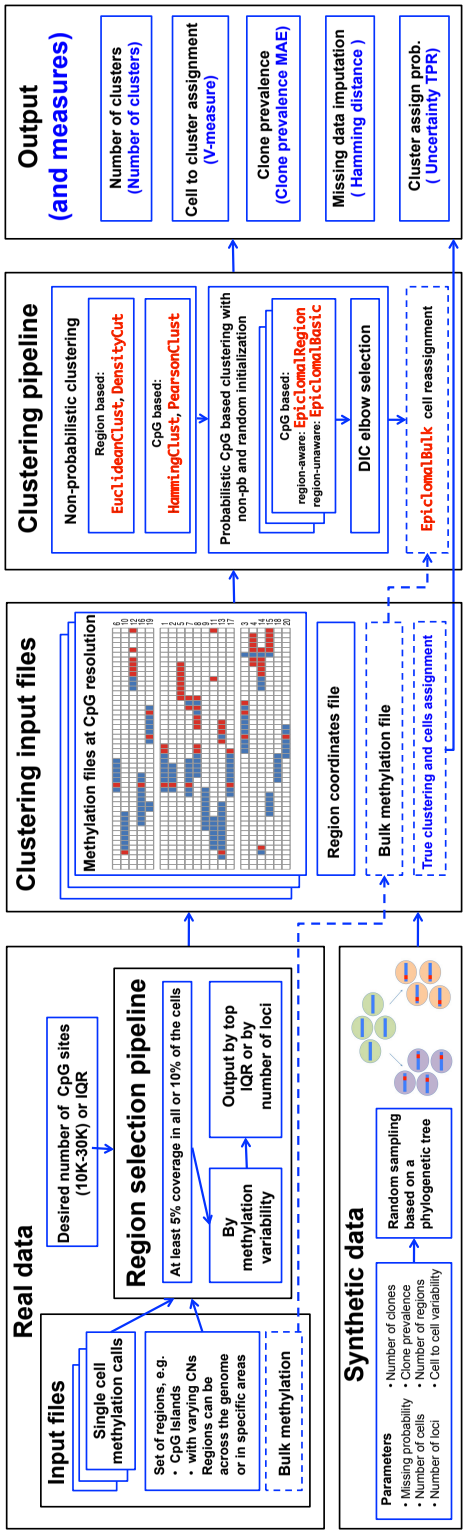

Figure 2: *Our proposed framework consists of three parts. Input data and pre-processing:* data from regions of interest are extracted from methylation call files, which can be filtered keeping only data from regions with a desired amount of missing data and methylation level IQR. A synthetic data pipeline is also provided to simulate data under different parameters. *Clustering:* cells are clustered using different non-probabilistic clustering methods, whose results will be then used as initial values for Epiclomal methods. *Output and performance measures:* different metrics are provided to evaluate the output of each method when true cluster assignments are known.

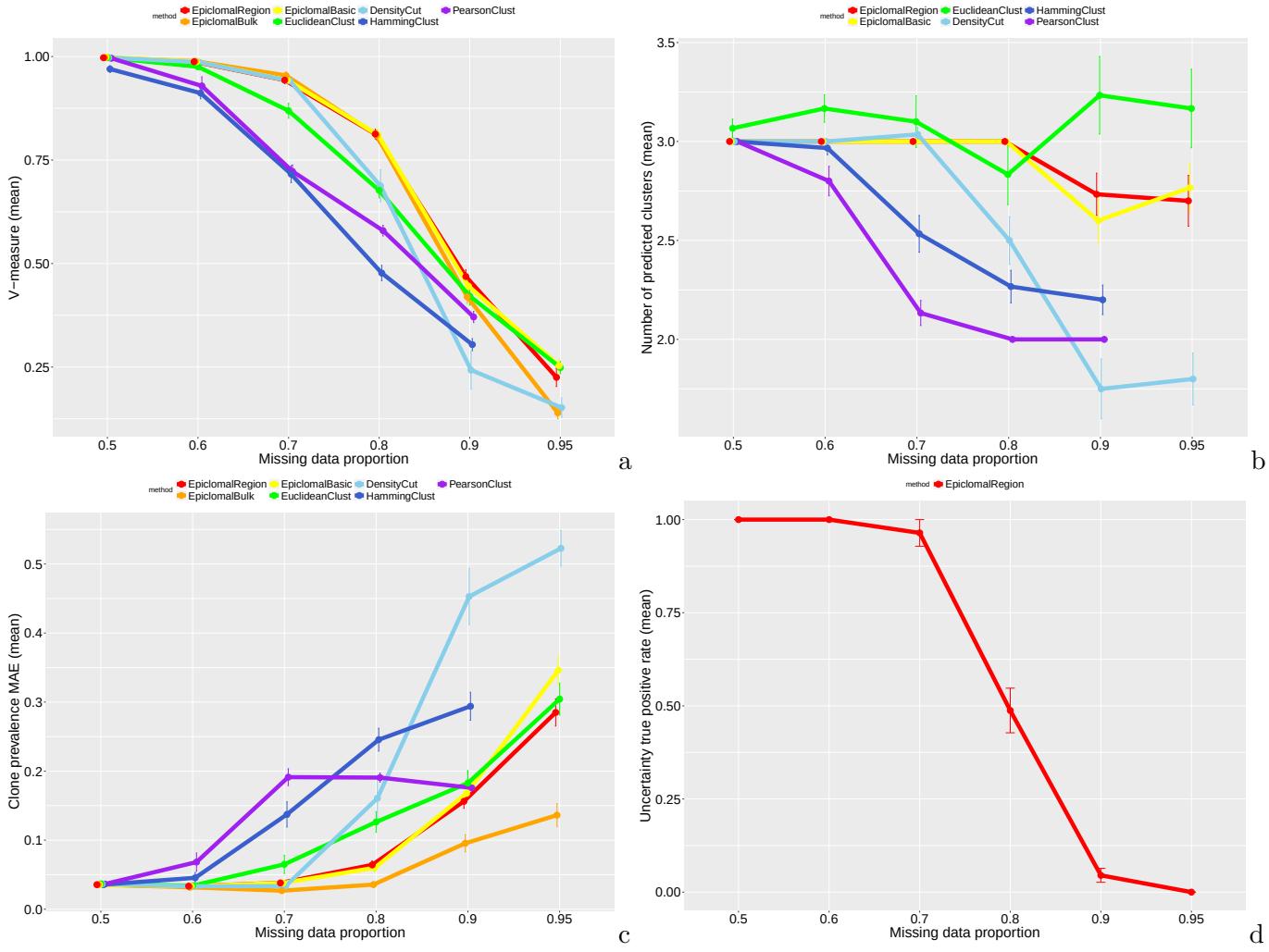

Figure 3: *Results for synthetic data when we vary the missing proportion.* a) Mean V-measure, b) Mean number of predicted clusters (true $K = 3$), c) Mean epiclone prevalence MAE, d) Mean uncertainty true positive rate. The vertical bars correspond to one standard deviation above and below the mean value.

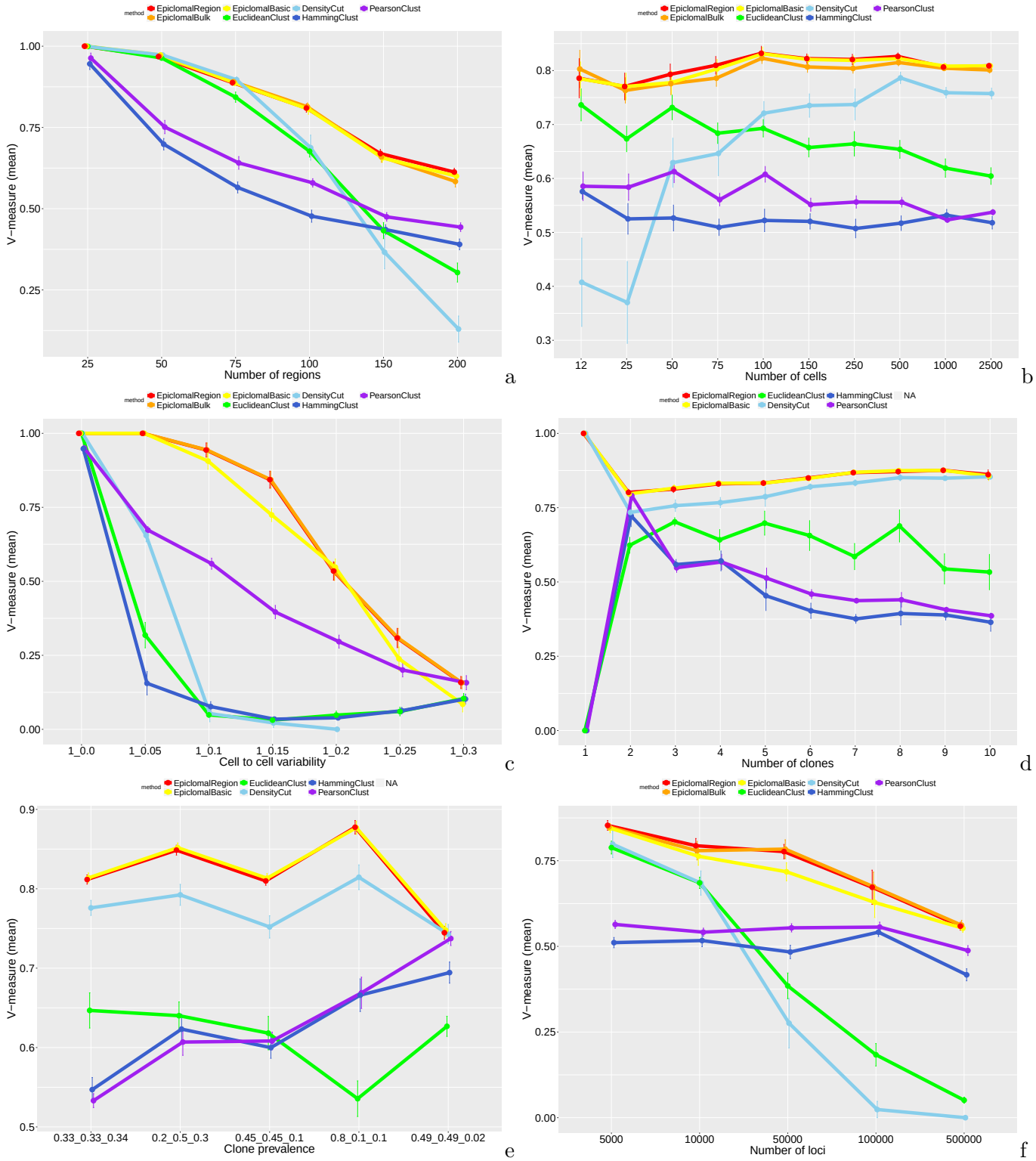

Figure 4: *Mean V-measure for synthetic data when we vary:* a) the number of regions, b) the number of cells, c) the cell-to-cell variability, d) the number of clones, e) the cluster prevalence, f) the number of loci. The vertical bars correspond to one standard deviation above and below the mean value. The Epiclomal methods outperform the other methods in all cases.

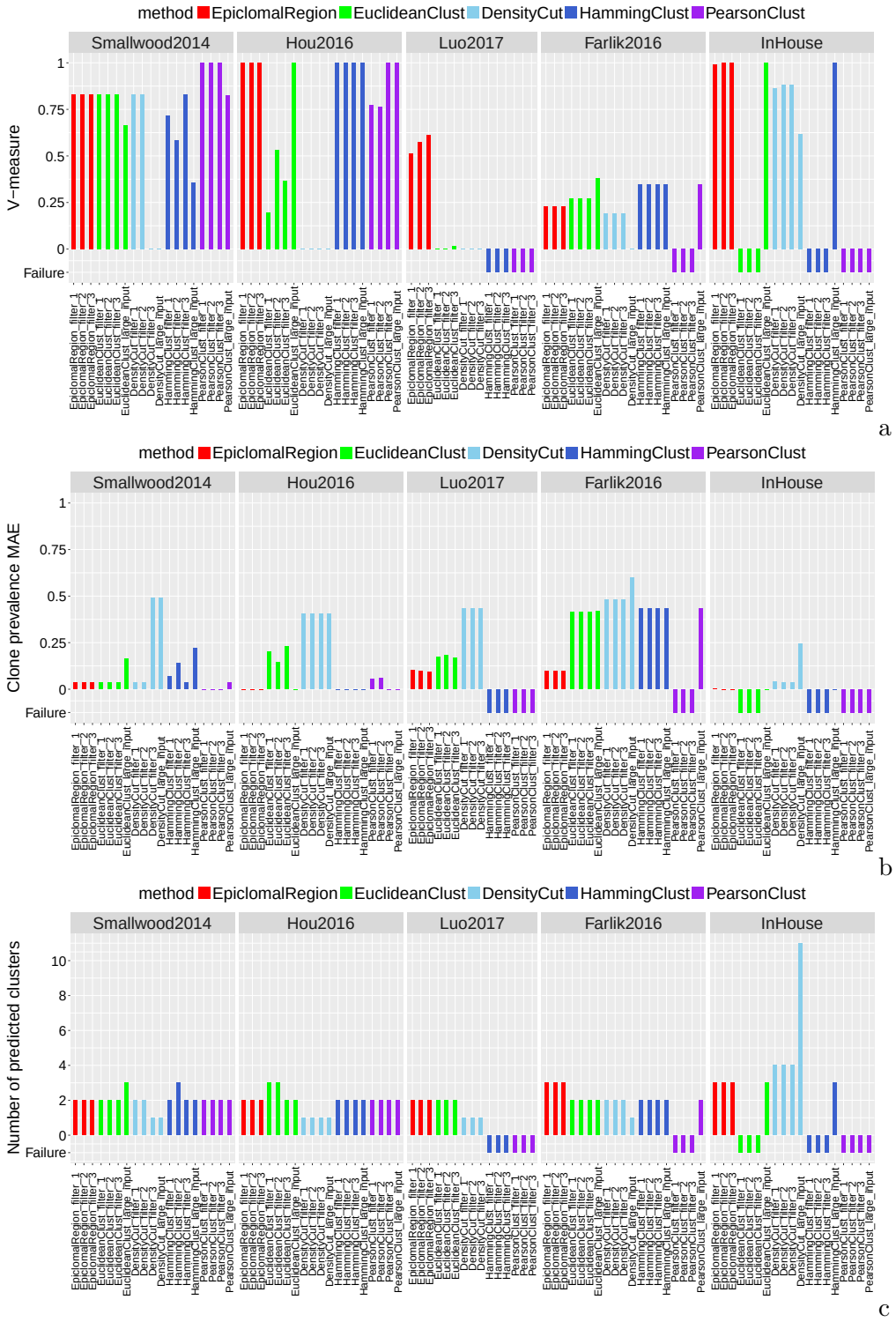

Figure 5: *Results for real data sets.* a) V-measure, b) epiclone prevalence MAE and c) number of predicted clusters for three filtered scenarios for each data set: filter 1, 2 and 3 corresponding to 10 000, 15 000 and 20 000 loci, respectively. For all data sets, except Luo2017, we also present the results for a large input data set, in which only the regions with average IQR < 0.01 were eliminated. The bars with "Failure" y-axis value correspond to the cases where a method failed to return a result. The ground-truth number of clusters are 3 for InHouse and Farlik2016; 2 for Smallwood2014, Hou2016 and Luo2017. Here we consider as true epiclone prevalences the cluster proportions calculated from the clustering reported in the corresponding papers. See also Table 2.

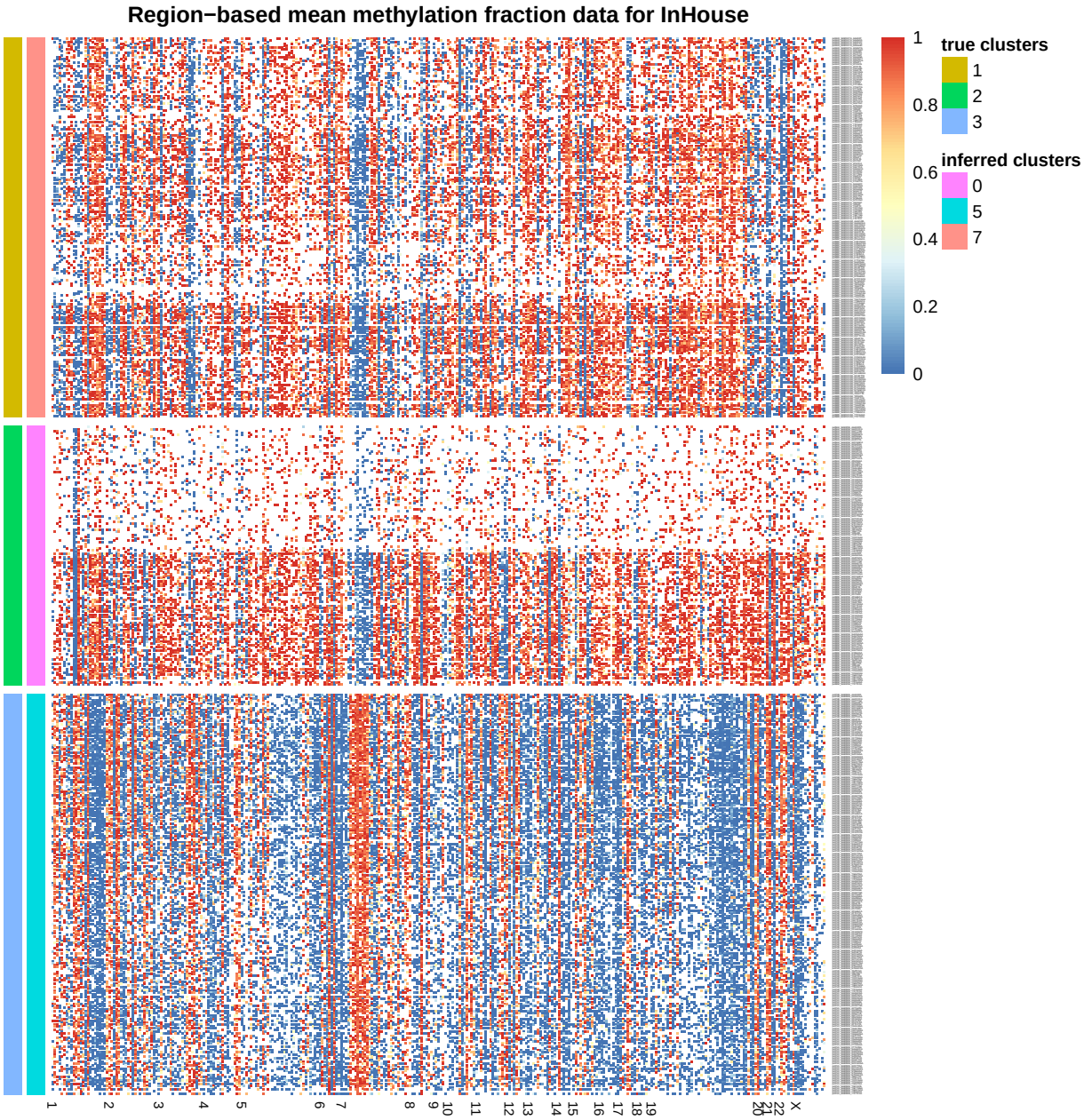

Figure 6: EpiclomalRegion clustering on the InHouse data set, filtered to include the most variable regions and obtain $\approx$ 15 000 loci (327 regions, cell average missing proportion 0.82, 558 cells). EpiclomalRegion obtained 3 clusters, V-measure = 1.

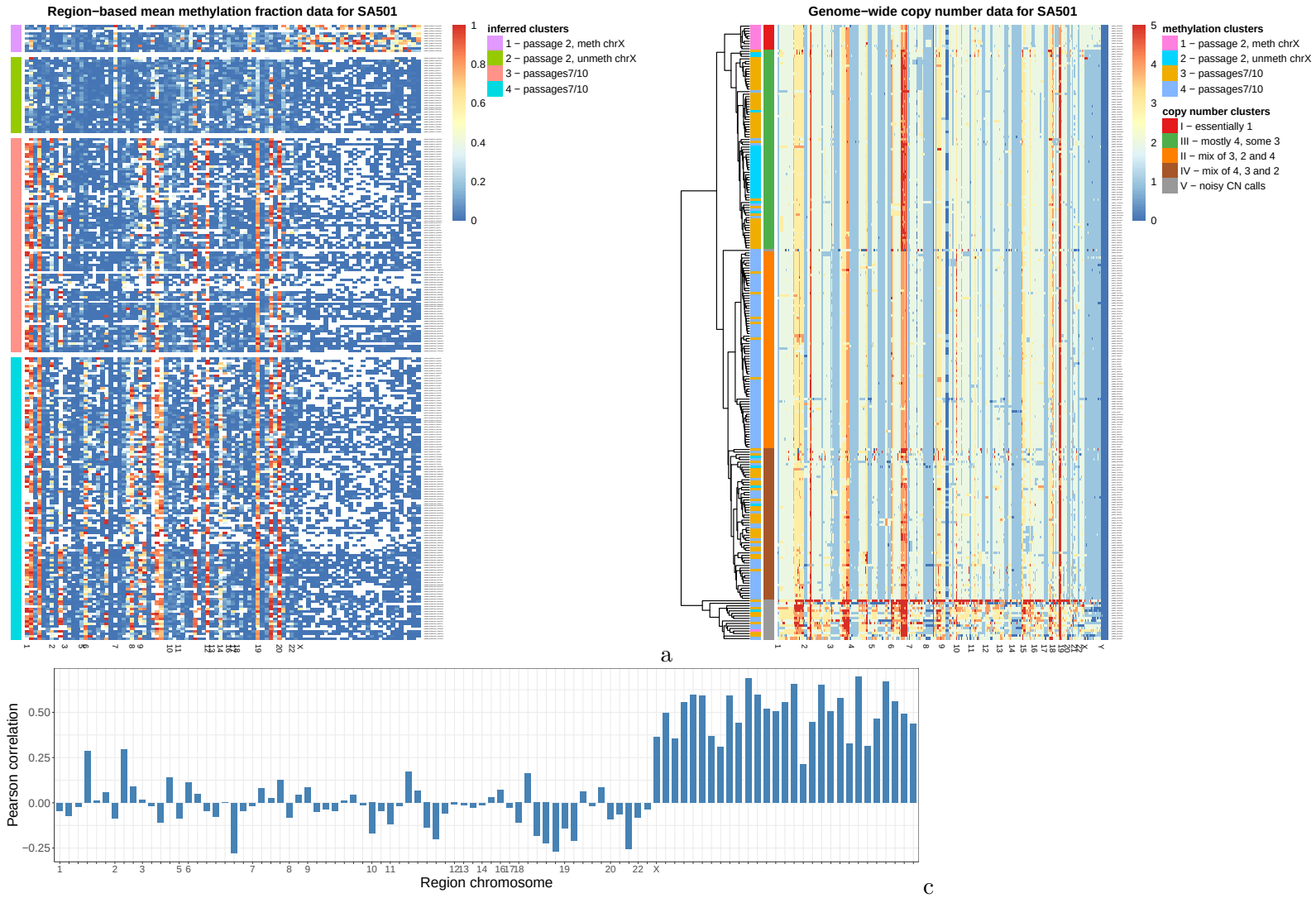

Figure 7: *Results for patient SA501.* a) Mean methylation level for each of the 94 NMF-selected regions (CGIs) for patient SA501 across all cells ordered according to the four methylation clusters found using EpiclomalRegion. b) The inferred copy numbers genome-wide for the same cells as in a) clustered using a ward.D2 hierarchical clustering method of the Euclidean copy number distances. Note that copy number 5 actually means 5 or more copies. To call copy number changes, we used the methylation sc-WGBS data. Only one epiclone and one copy number clone match, the remaining clones transcend each other. c) Pearson correlation between the mean methylation data and the copy number data in each of the 94 regions. There is correlation in chromosome X, but not in the autosomal chromosomes.