Epiclomal

Probabilistic Clustering of
Sparse DNA Methylation Data

Motivation

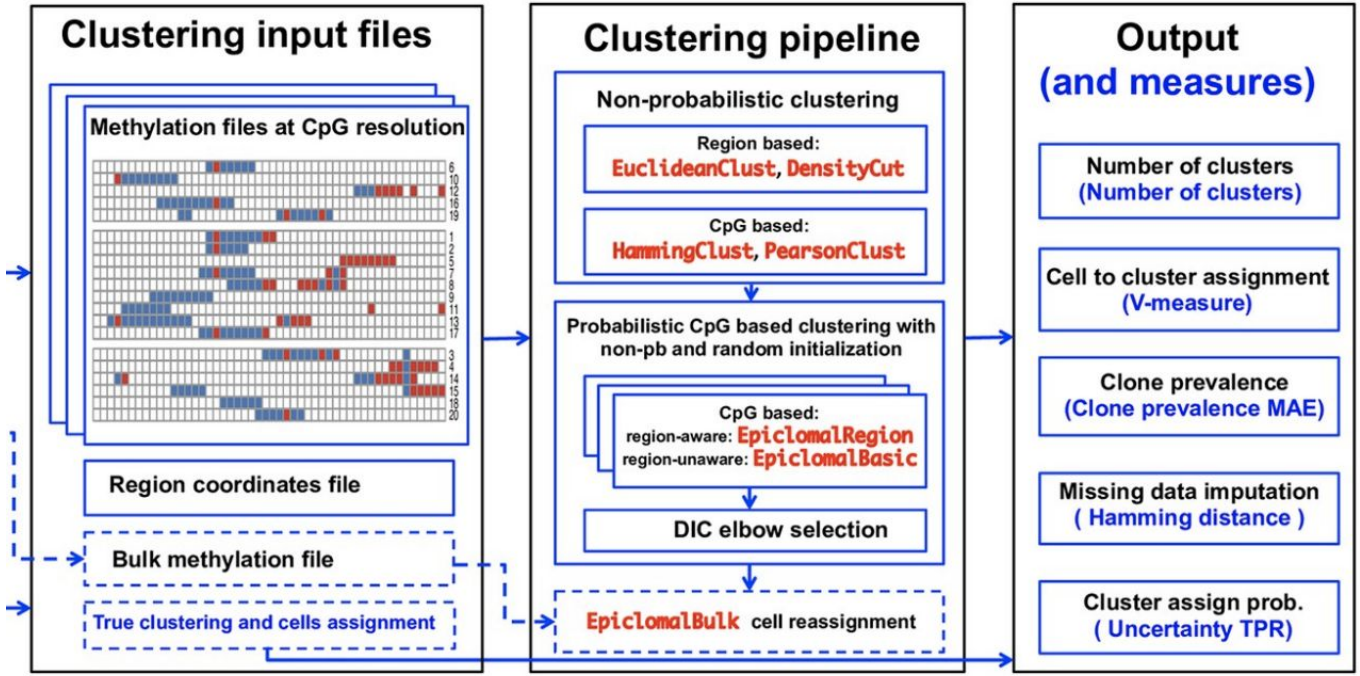
- DNA methylation at 5mC position is important for transcriptional regulation
- Single-cell whole-genome bisulfite sequencing (sc-WGBS) can assess epigenetic diversity of a cell population
- Data is sparse (lots of missing CpG sites) and subject to error
- Need for clustering according to methylation profiles
    - Identification of cancer subtypes
    - Detection of previously unknown cell types, deeper characterization of known ones
    - Imputation of missing CpG data by pooling information across clusters

Previous Work

- A number of non-probabilistic clustering methods for CpG methylation data
  - Hou et al., Farlik et al., others
- Probabilistic approaches to imputation of missing CpG data
  - Kapourani and Sanguinetti, Angermuller et al.
- Until now: no probabilistic method with a focus on clustering based on methylation profiles
- Authors want to simultaneously cluster sc-WGBS data while inferring the missing methylation states

Epiclomal

- Goal: cluster sparse CpG-based DNA methylation data from sc-WGBS



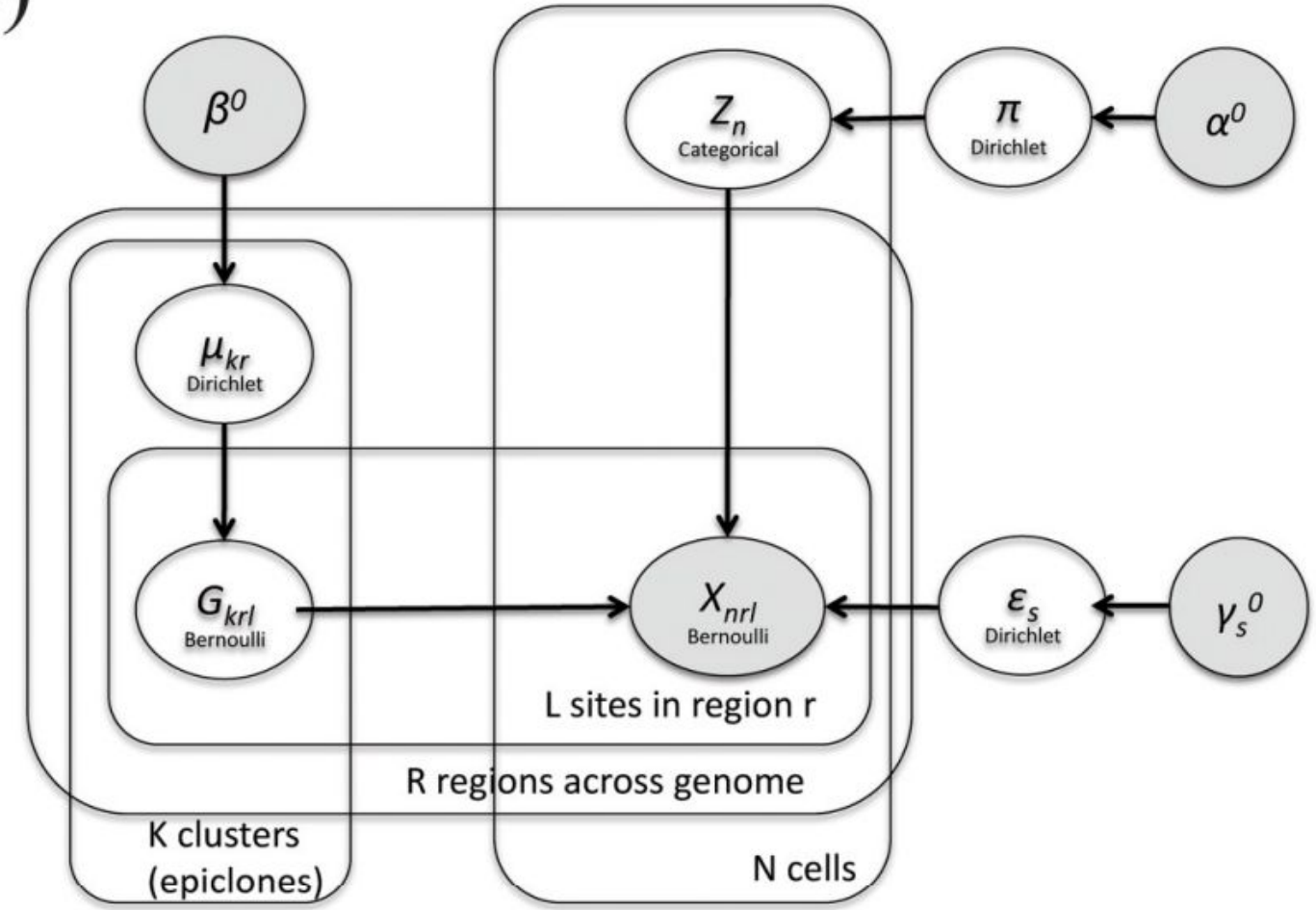
Input

- sparse matrix of *N* rows (cells) and *M* columns (CpG sites)
- each entry is 0 (unmethylated), 1 (methylated), or missing
- partially methylated sites: 1 if methylation fraction is at least 0.5, 0 otherwise
  - median < 1.35% across datasets
- optional: select specific regions
  - EpiclomalBasic vs. EpiclomalRegion

Methods: Non-Probabilistic Methods

- Performance comparisons
- Cluster initialization
- Region-based: EuclideanClust, DensityCut
  - Based on mean methylation for each region
  - DensityCut also infers the optimal number of clusters
- CpG-based: HammingClust, PearsonClust
  - Based on methylation of individual CpG sites
- Epiclomal considers each potential number of clusters up to a chosen *K*
  - Infers optimal number of clusters (at most *K*)

(b)


β^0
Z_n
Categorical
π
Dirichlet
α^0
μ_{kr}
Dirichlet
G_{krl}
Bernoulli
X_{nrl}
Bernoulli
ε_s
Dirichlet
γ_s^0
L sites in region r
R regions across genome
K clusters
(epiclones)
N cells

Let $\mathbf{X}_{nr} = (X_{nr1}, \ldots, X_{nrL_r})^T$ be the vector of observed data for region r in cell n, and let $\mathbf{X}_n = (\mathbf{X}_{nr}^T, \ldots, \mathbf{X}_{nR}^T)^T$ be the vector of all observed data for cell n. Assume that $X_{nr1}, \ldots, X_{nrL_r}$ are independent for all n and r. Suppose that there are $K \ll N$ vectors of true hidden methylation states shared across the cells. Let Z_n with values in $\{1, \ldots, K\}$ be the hidden variable indicating the true cluster (epiclonal) population of cell n. It is assumed that $Z_1, \ldots, Z_N$ are independent with $P(Z_n = k) = \pi_k$ such that $\sum_{k=1}^{K} \pi_k = 1$. If $Z_n = k$, then the distribution of $\mathbf{X}_n$ depends on the k-th vector of true hidden epigenotypes $\mathbf{G}_k = (\mathbf{G}_{k1}^T, \ldots, \mathbf{G}_{kR}^T)^T$, where $\mathbf{G}_{kr} = (G_{kr1}, \ldots, G_{krL_r})^T$. We assume that $G_{kr1}, \ldots, G_{krL_r}$ are independent for all k and r, with $P(G_{krl} = s) = \mu_{krs}$ such that $\sum_{s \in \mathcal{S}} \mu_{krs} = 1$, that is, G_{krl} follows a categorical (Bernoulli) distribution with parameter set $\boldsymbol{\mu}_{kr} = \{\mu_{krs} : s \in \mathcal{S}\}$. Therefore, given the true cluster assignment and the corresponding true hidden methylation states, the observed data $\mathbf{X}_{nr}$ are independent, with X_{nrl} following a categorical distribution with parameters depending on the hidden true state at locus l of region r for cluster population k, that is,

$$P(X_{nrl} = t | Z_n = k, G_{krl} = s) = \epsilon_{st} \text{ with } \sum_{t \in \mathcal{S}} \epsilon_{st} = 1. \tag{1}$$

We can also interpret the probability in (1) as a misclassification error, which in this context is related to sequencing error.

To infer Θ and the hidden states $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ and $\mathbf{G} = \{\mathbf{G}_1, \ldots, \mathbf{G}_K\}$, we adopt a Bayesian approach and derive a Variational Bayes (VB) algorithm [33] to approximate the posterior distribution of Θ, $\mathbf{Z}$, and $\mathbf{G}$ given the observed data $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, $P(\mathbf{Z}, \mathbf{G}, \Theta|\mathbf{X})$ by finding the Variational Distribution (VD), $q(\mathbf{Z}, \mathbf{G}, \Theta)$ with the smallest Kullback-Leibler divergence to the posterior $P(\mathbf{Z}, \mathbf{G}, \Theta|\mathbf{X})$, which is equivalent to maximizing the evidence lower bound (ELBO) given by

$$\mathrm{ELBO}(q) = \mathrm{E}[\log P(\mathbf{X}, \mathbf{Z}, \mathbf{G}, \mathbf{\Theta})] - \mathrm{E}[\log q(\mathbf{Z}, \mathbf{G}, \mathbf{\Theta})]. \tag{2}$$

Methods (cont.)

- ELBO is non-convex optimization - susceptible to local optimum
- ran VB algorithm *T* times (1000 for real data sets, 300 for synthetic data)
- started from different initial cluster assignments for each cell
- K = 10 for all runs (considers 1, 2,..., 10 clusters)
- 10 initializations each from EuclidClust, HammingClust, PearsonClust - 1 from DensityCut
- The rest (*T* - 31) of the initializations are uniformly random
- Selecting the best run
    - Minimum DIC score over runs with *c* clusters
    - DIC elbow plot to select optimal number of clusters, best overall run

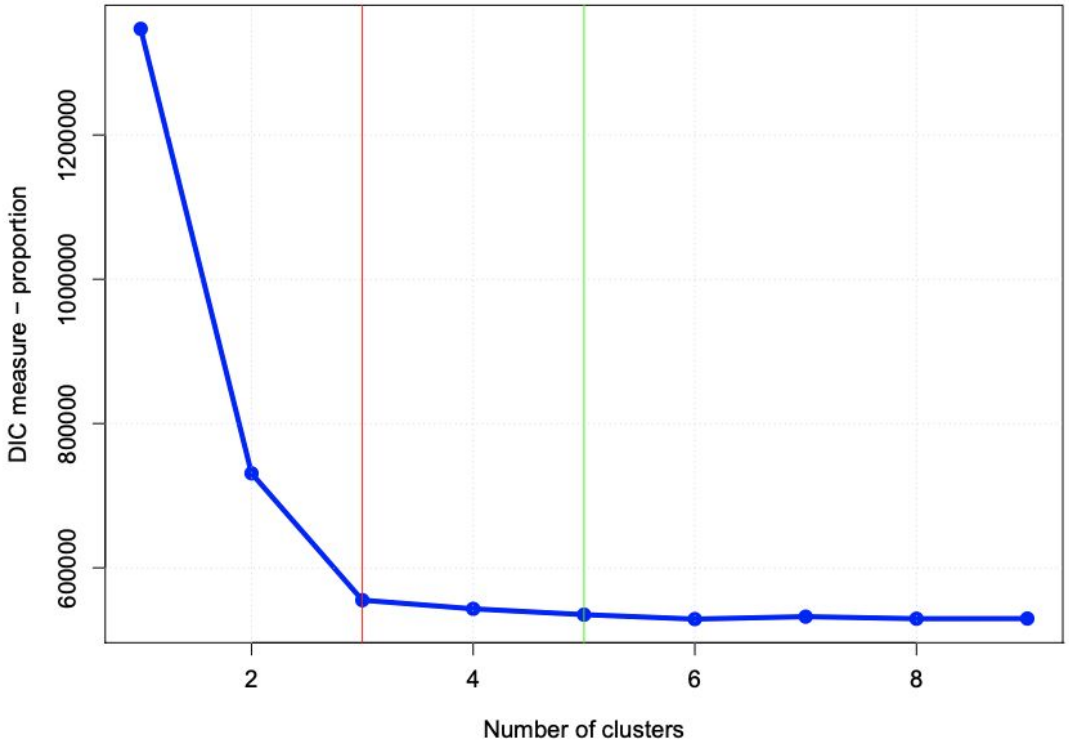
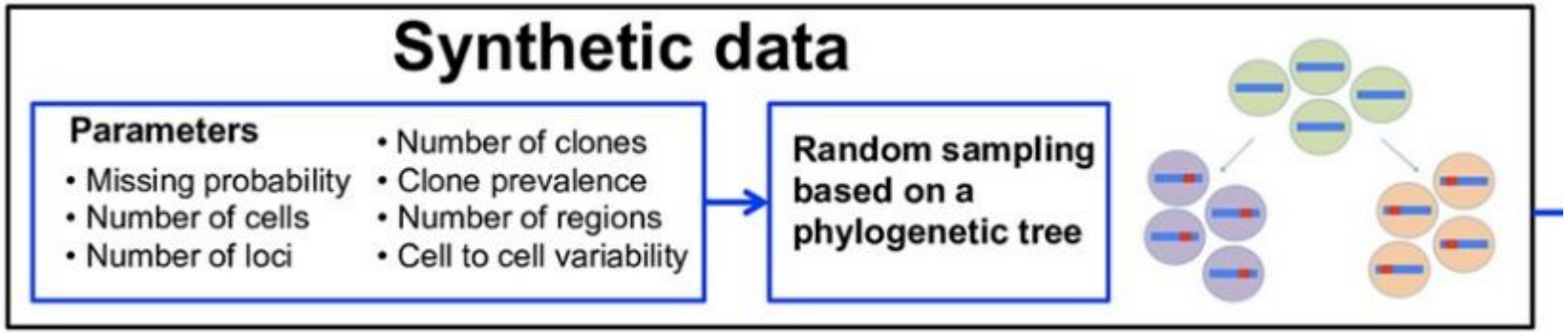


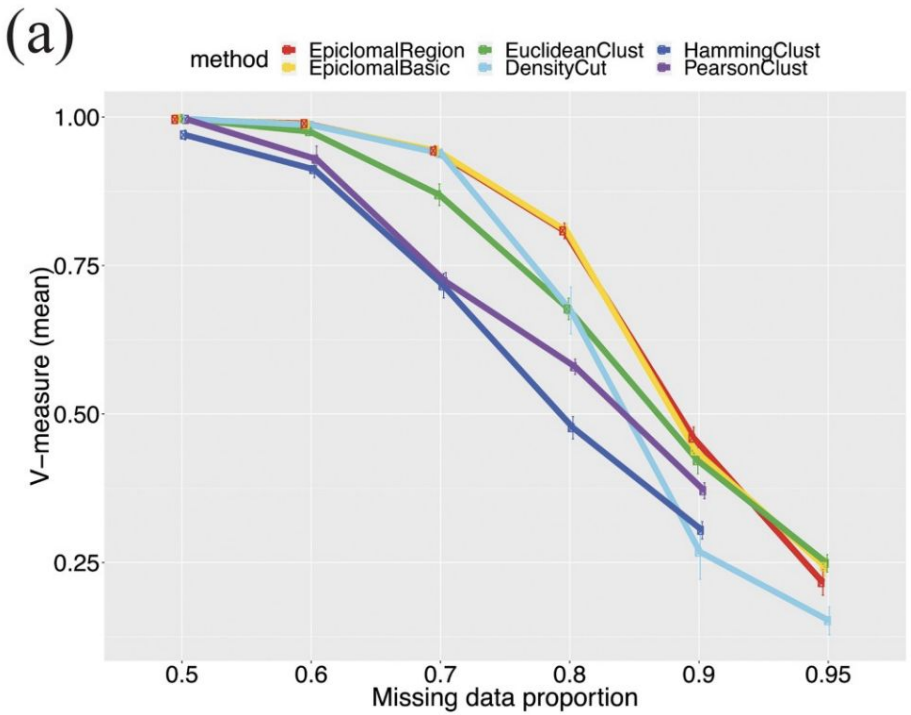
Figure A Example of a DIC elbow plot for EpiclomalRegion for the InHouse data set with 10 000 loci. Our DIC algorithm selects only number of clusters for which there is a decrease of at least 2% in DIC values, green vertical line. Then, the elbow value is picked, red vertical line.

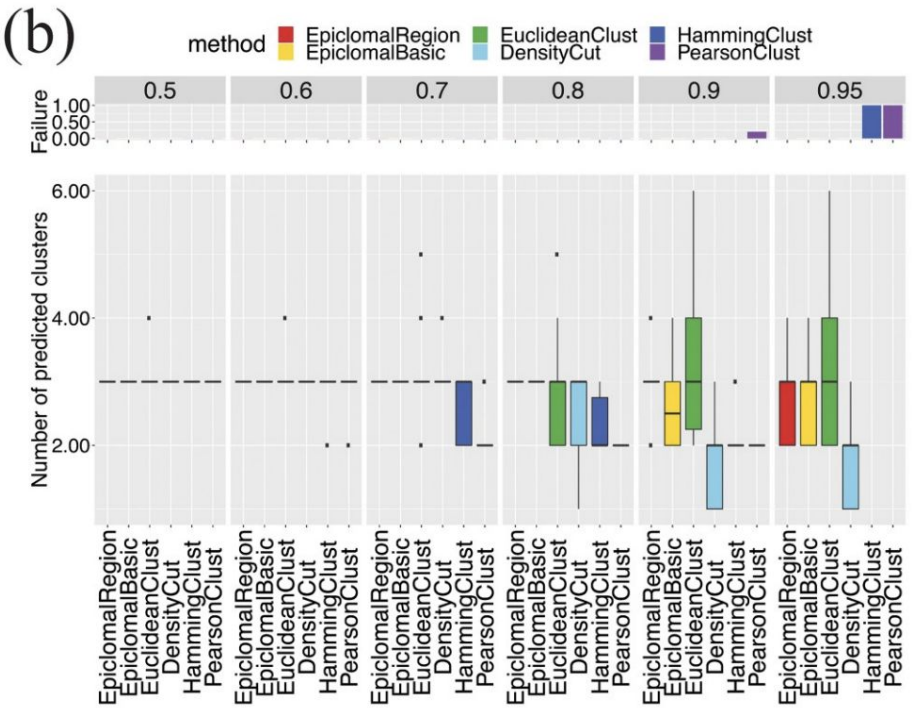
Results

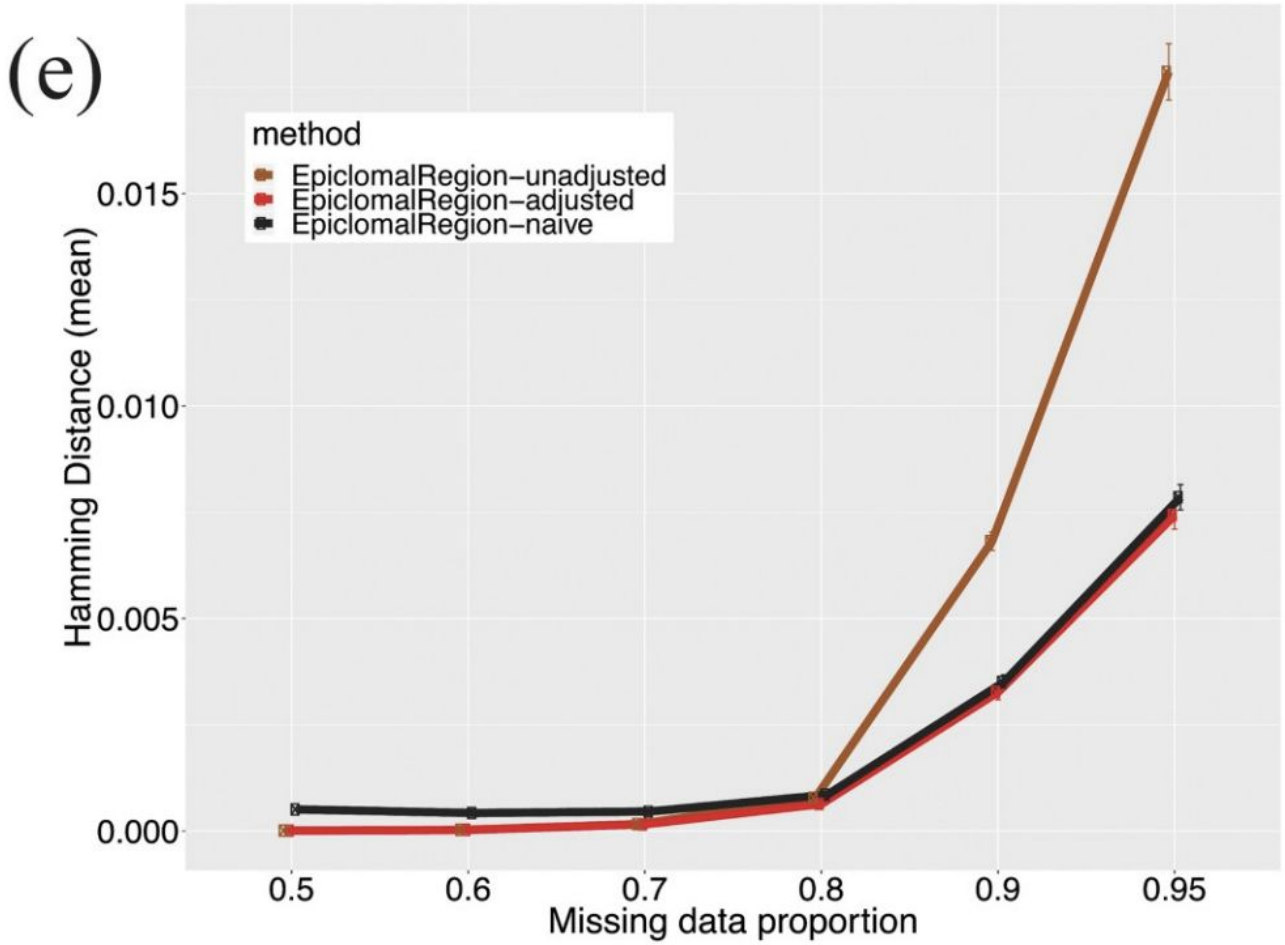
- V-measure is calculated as the harmonic mean between homogeneity (h) and completeness (c)
  - A predicted clustering has homogeneity 1 if all of the predicted clusters contain only data points which are members of a single true class.
  - A predicted clustering result has completeness 1 if it assigns all of those data points that are members of a single true class to a single predicted cluster.

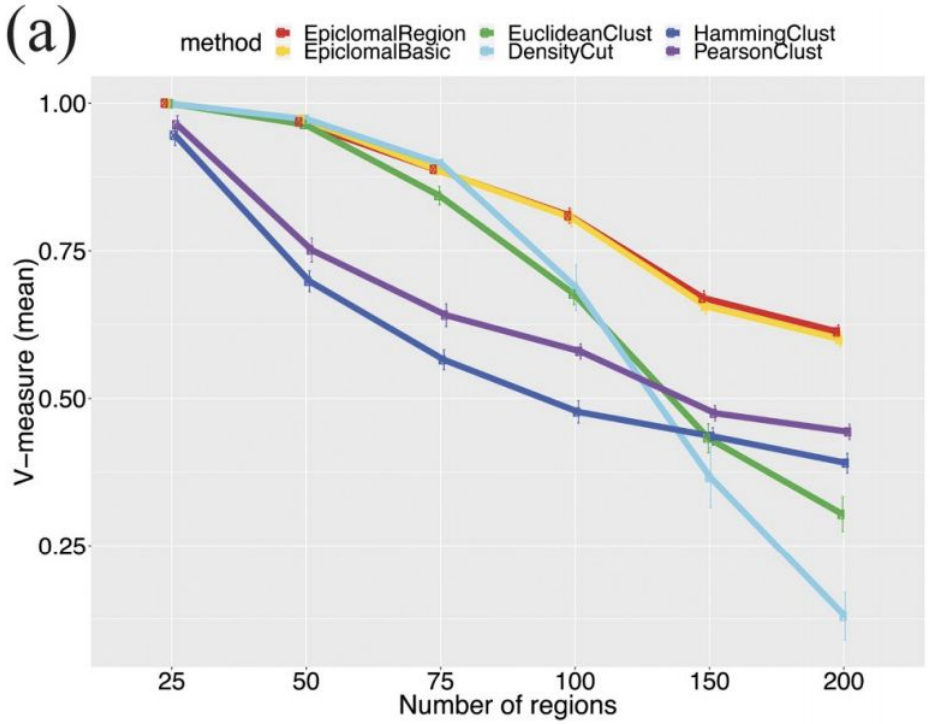


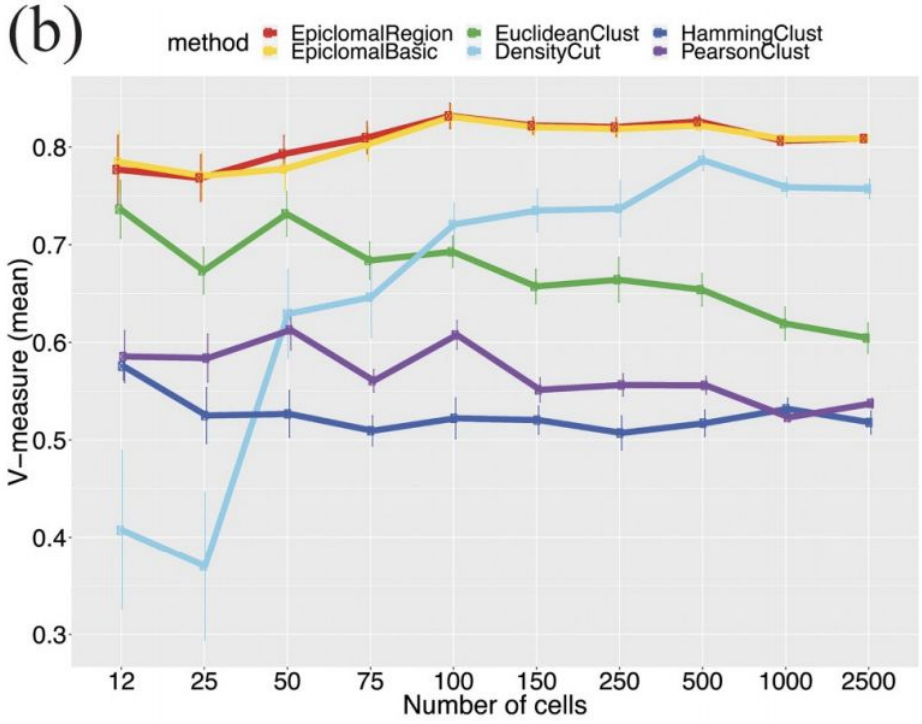
Varying parameter	Varying range
Missing proportion	0.5 to 0.95
Number of regions	25 to 200
Number of cells	12 to 2500
Cell-to-cell variability	0 to 0.3
Number of clusters (epiclones)	1 to 10
Epiclone frequencies	balanced to very unbalanced
Number of loci	5 000 to 500 000
Number of regions different between clusters	1 to 6

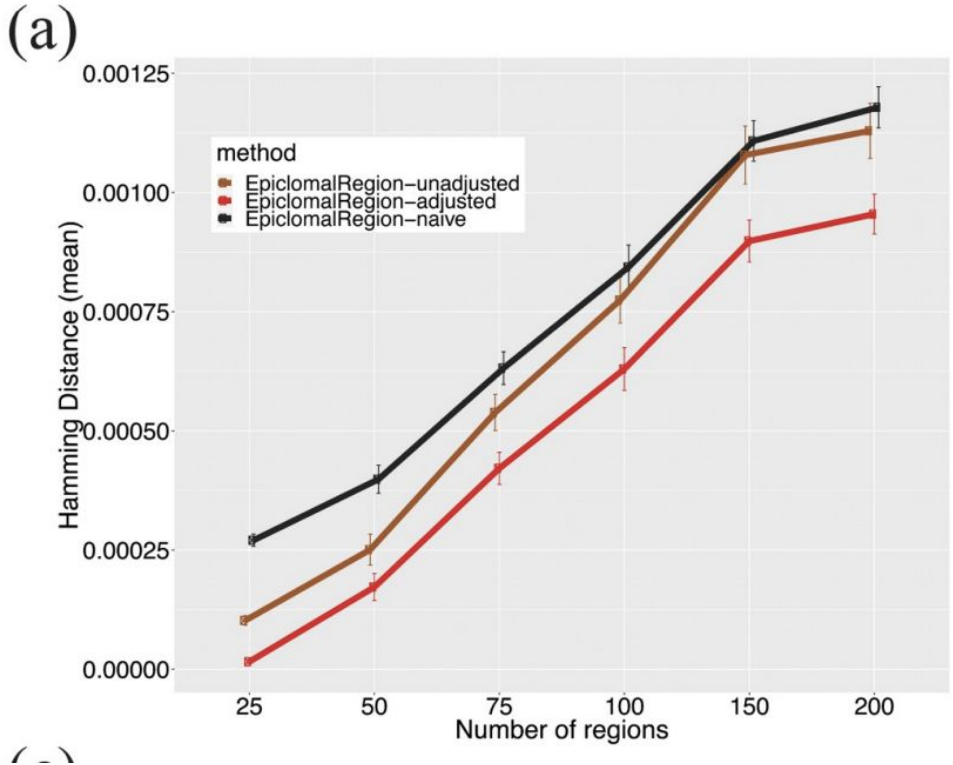

(a)
method
EpiclomalRegion
EpiclomalBasic
EuclideanClust
DensityCut
HammingClust
PearsonClust
V−measure (mean)
1.00
0.75
0.50
0.25
0.5
0.6
0.7
0.8
0.9
0.95
Missing data proportion

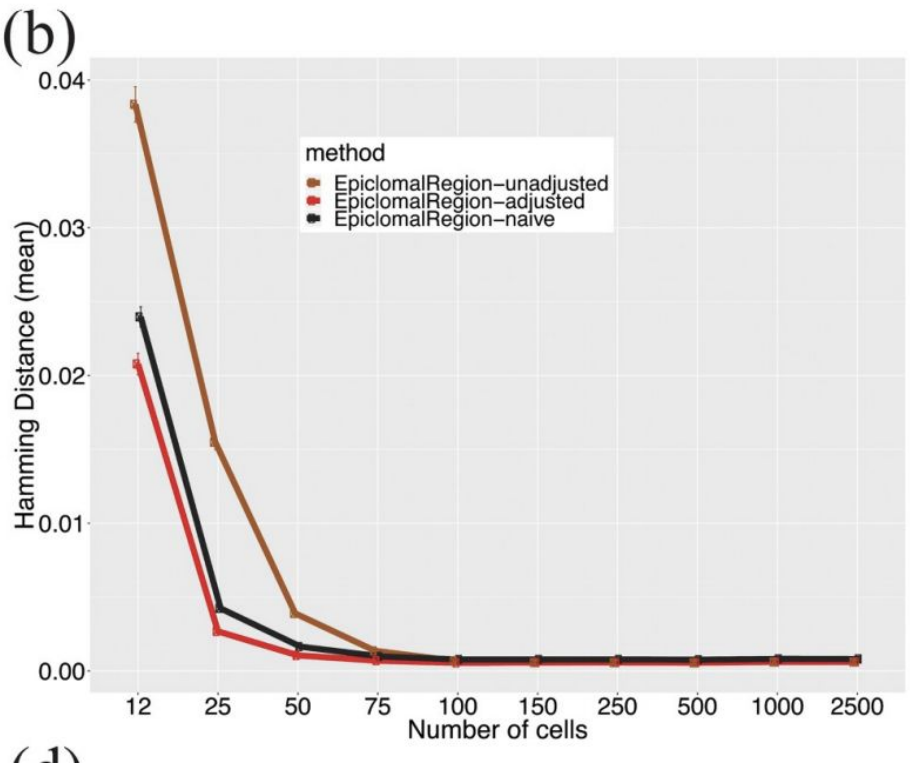

(b)
method
EpiclomalRegion
EpiclomalBasic
EuclideanClust
DensityCut
HammingClust
PearsonClust
0.5
0.6
0.7
0.8
0.9
0.95
Failure
1.00
0.50
0.00
Number of predicted clusters
6.00
4.00
2.00
EpiclomalRegion
EpiclomalBasic
EuclideanClust
DensityCut
HammingClust
PearsonClust


(e)
method
EpiclomalRegion−unadjusted
EpiclomalRegion−adjusted
EpiclomalRegion−naive
Hamming Distance (mean)
0.015
0.010
0.005
0.000
0.5
0.6
0.7
0.8
0.9
0.95
Missing data proportion

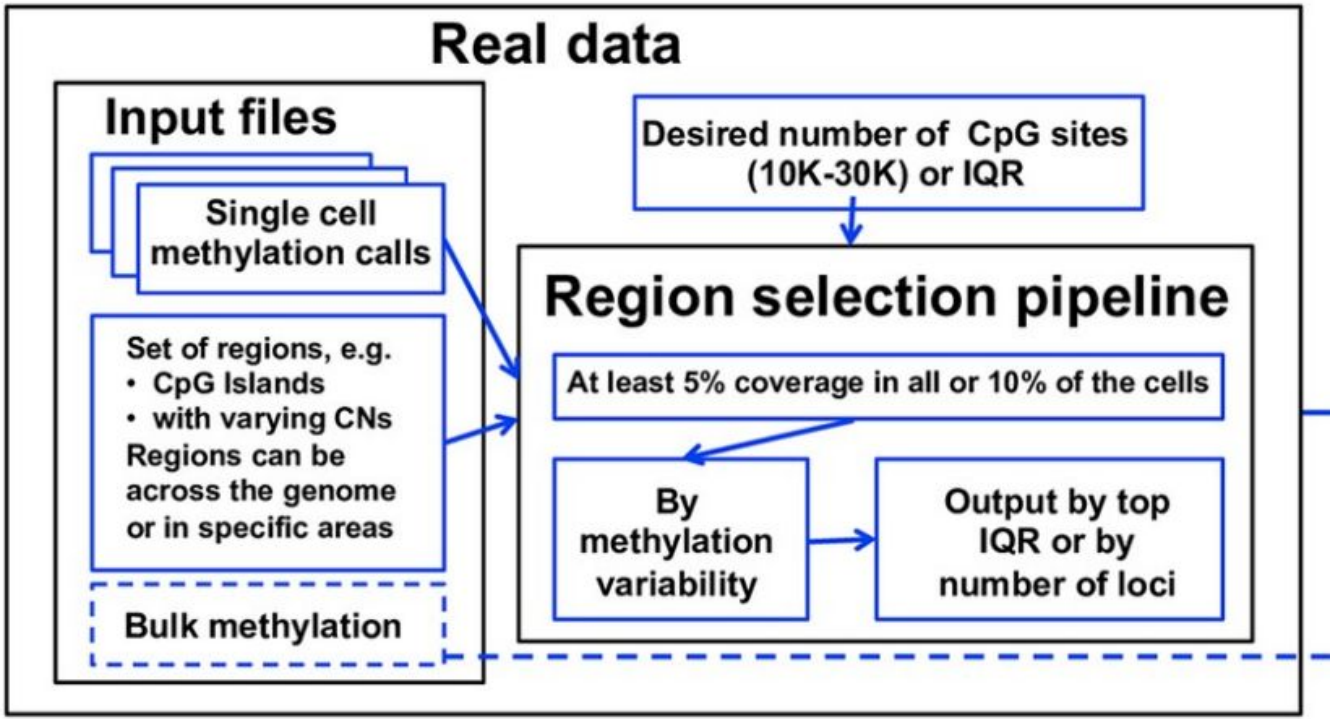

(a)
method
EpiclomalRegion
EpiclomalBasic
EuclideanClust
DensityCut
HammingClust
PearsonClust
1.00
0.75
0.50
0.25
V−measure (mean)
25
50
75
100
150
200
Number of regions

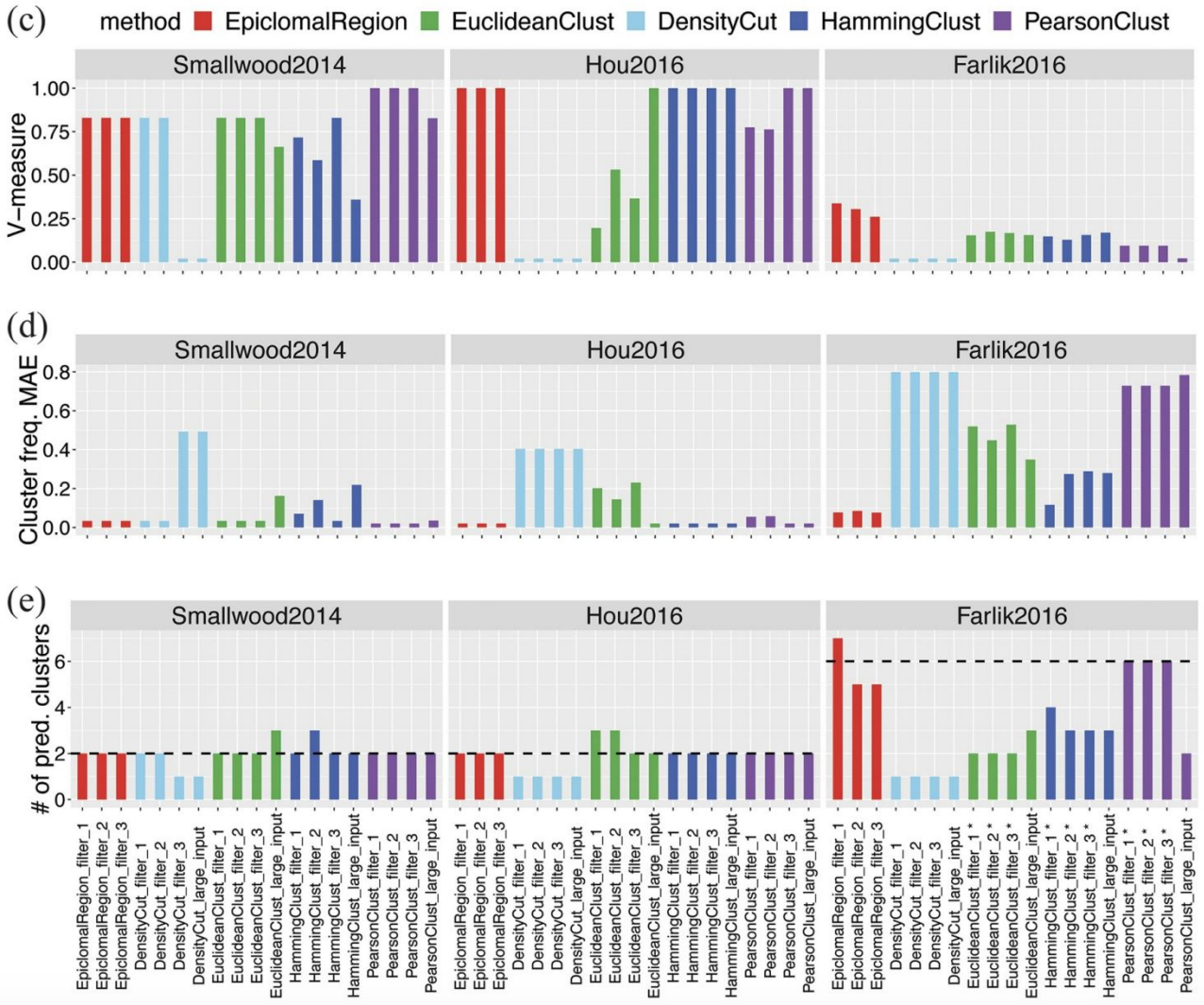

(b)
method
EpiclomalRegion
EpiclomalBasic
EuclideanClust
DensityCut
HammingClust
PearsonClust
0.8
0.7
0.6
0.5
0.4
0.3
V−measure (mean)
12
25
50
75
100
150
250
500
1000
2500
Number of cells

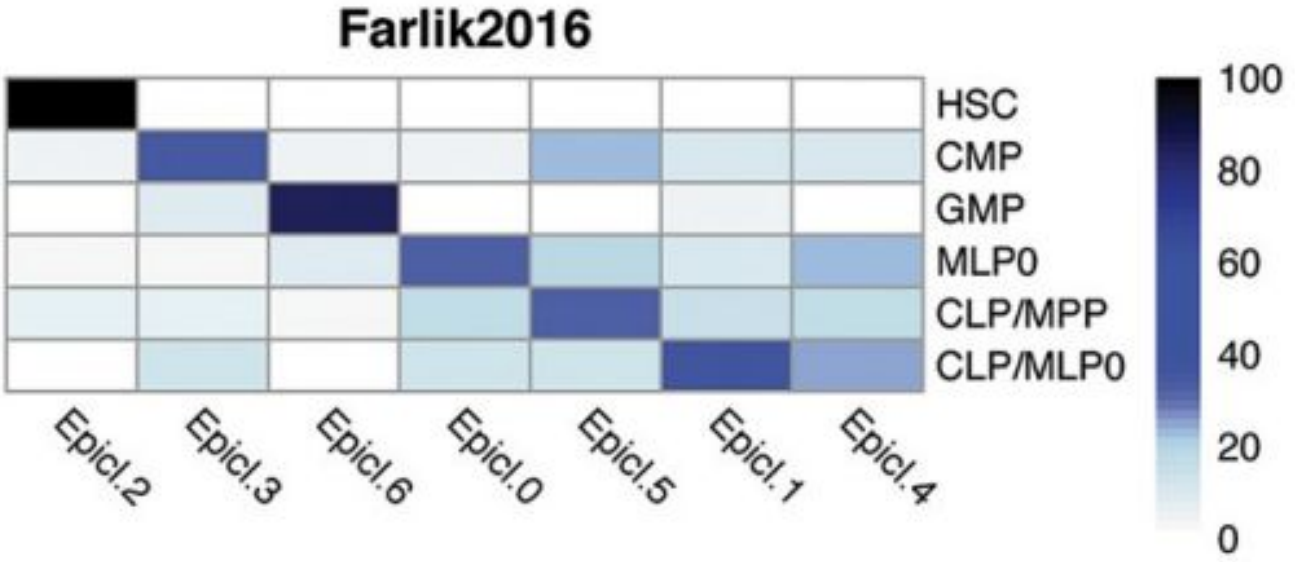

(a)
method
EpiclomalRegion−unadjusted
EpiclomalRegion−adjusted
EpiclomalRegion−naive
Hamming Distance (mean)
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
25
50
75
100
150
200
Number of regions


(b)
method
EpiclomalRegion−unadjusted
EpiclomalRegion−adjusted
EpiclomalRegion−naive
Hamming Distance (mean)
0.04
0.03
0.02
0.01
0.00
12
25
50
75
100
150
250
500
1000
2500
Number of cells

Data set	Cell type	# cells	# clusters	Regions	Miss 10K	Nloci IQR ≥ .01	Miss IQR ≥ .01
Smallwood2014 [16]	mouse embryonic stem cells	32	2	CGI	0.69	786 620	0.54
Hou2016 [11]	human hepatocellular carcinomas	25	2	CGI	0.87	255 136	0.90
Farlik2016 [12]	human hematopoietic cells	122	6	TFBS	0.89	512 153	0.98
InHouse	human xenografted cancer cells (3 patients)	558	NA	CGI	0.82	1 019 956	0.79



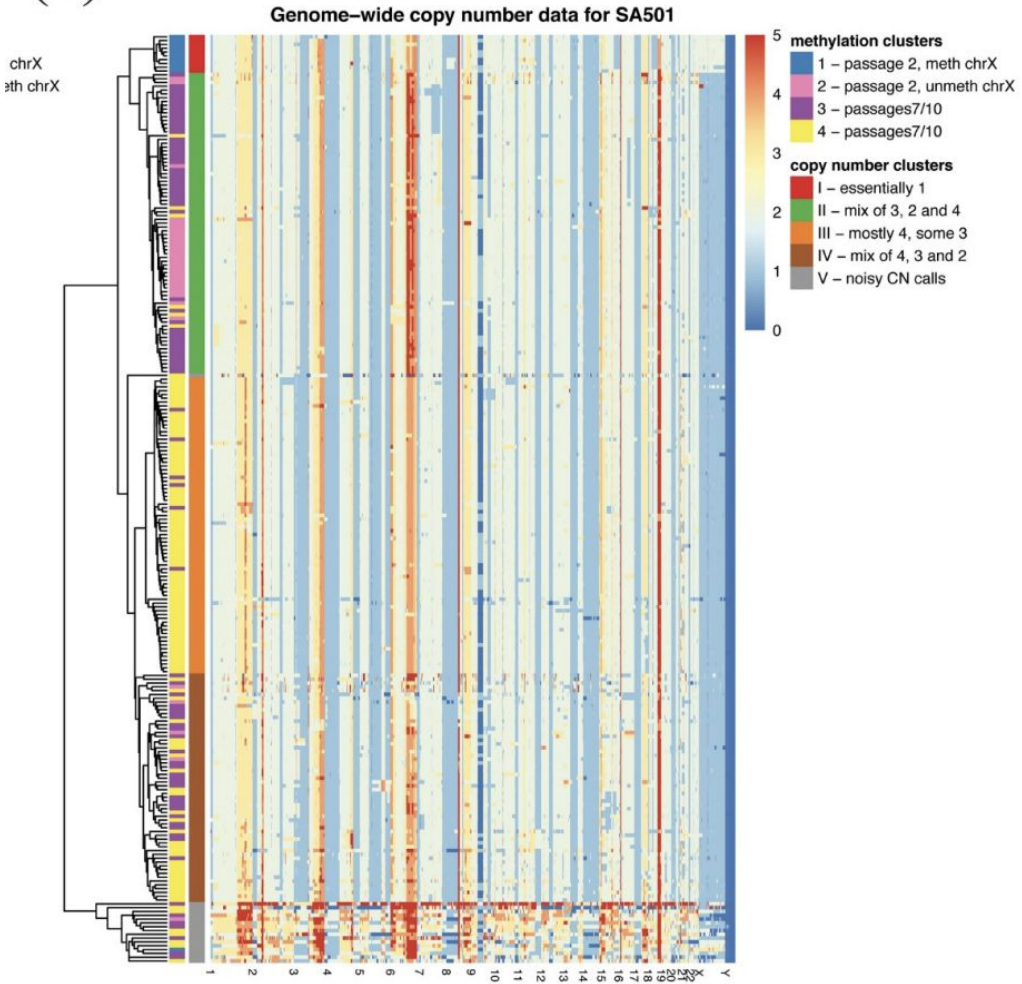

(c)
method
EpiclomalRegion
EuclideanClust
DensityCut
HammingClust
PearsonClust
Smallwood2014
Hou2016
Farlik2016
V−measure
1.00
0.75
0.50
0.25
0.00
(d)
Cluster freq. MAE
0.8
0.6
0.4
0.2
0.0
(e)
# of pred. clusters
6
4
2
0
EpiclomalRegion_filter_1
EpiclomalRegion_filter_2
EpiclomalRegion_filter_3
DensityCut_filter_1
DensityCut_filter_2
DensityCut_filter_3
DensityCut_large_input
EuclideanClust_filter_1
EuclideanClust_filter_2
EuclideanClust_filter_3
EuclideanClust_large_input
HammingClust_filter_1
HammingClust_filter_2
HammingClust_filter_3
HammingClust_large_input
PearsonClust_filter_1
PearsonClust_filter_2
PearsonClust_filter_3
PearsonClust_large_input
EuclideanClust_filter_1 *
EuclideanClust_filter_2 *
EuclideanClust_filter_3 *
HammingClust_filter_1 *
HammingClust_filter_2 *
HammingClust_filter_3 *
PearsonClust_filter_1 *
PearsonClust_filter_2 *
PearsonClust_filter_3 *


Farlik2016
HSC
CMP
GMP
MLP0
CLP/MPP
CLP/MLP0
Epicl.2
Epicl.3
Epicl.6
Epicl.0
Epicl.5
Epicl.1
Epicl.4
100
80
60
40
20
0

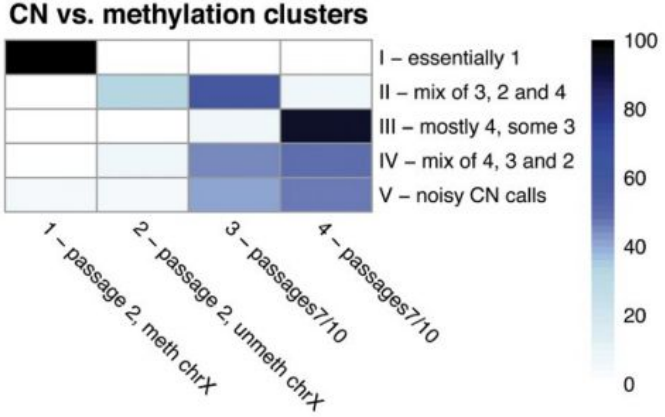
InHouse Data

- Single-cell epigenomes generated in-house on a range of patient-derived breast tumour xenografts
- 2 patients with triple-negative breast cancer, 1 with ER+PR-Her2+ breast cancer
- For one patient: compare epiclone clusters with copy-number clones
- Some chromosomal regions may show strong copy-number influence on CpG states (X chromosome), while others may not
- Epiclones and CN clones can match or transcend each other

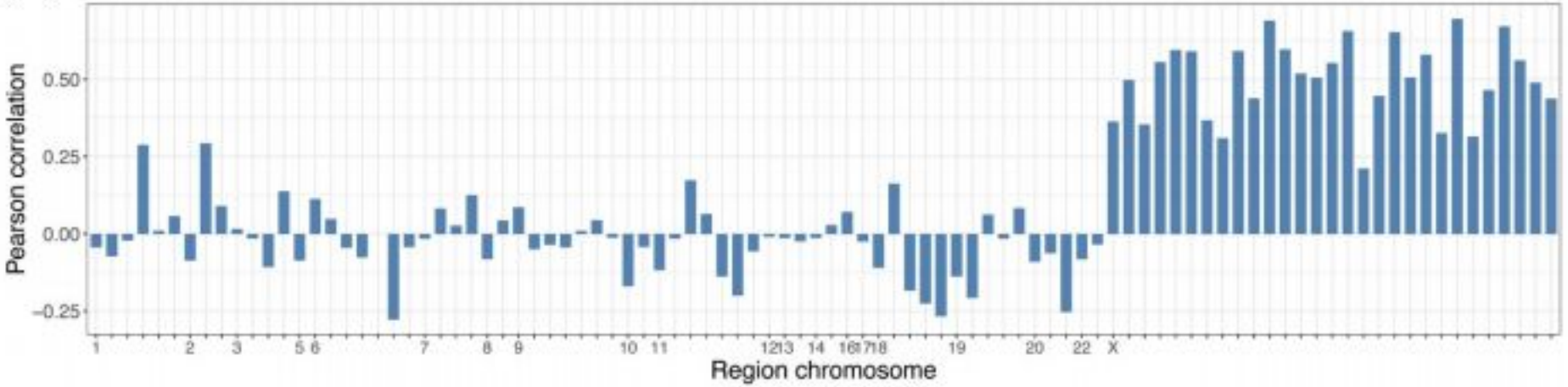
(b)


Genome–wide copy number data for SA501
chrX
eth chrX
methylation clusters
1 – passage 2, meth chrX
2 – passage 2, unmeth chrX
3 – passages7/10
4 – passages7/10
copy number clusters
I – essentially 1
II – mix of 3, 2 and 4
III – mostly 4, some 3
IV – mix of 4, 3 and 2
V – noisy CN calls
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

(d)


CN vs. methylation clusters
I – essentially 1
II – mix of 3, 2 and 4
III – mostly 4, some 3
IV – mix of 4, 3 and 2
V – noisy CN calls
1 – passage 2, meth chrX
2 – passage 2, unmeth chrX
3 – passages7/10
4 – passages7/10
100
80
60
40
20
0

(c)


Pearson correlation
0.50
0.25
0.00
−0.25
1
2
3
5
6
7
8
9
10
11
12
13
14
16
17
18
19
20
22
X
Region chromosome