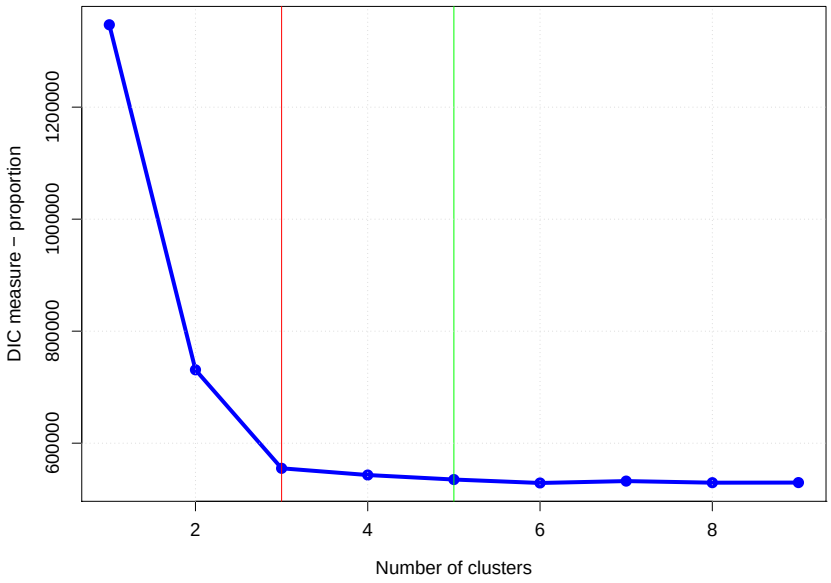
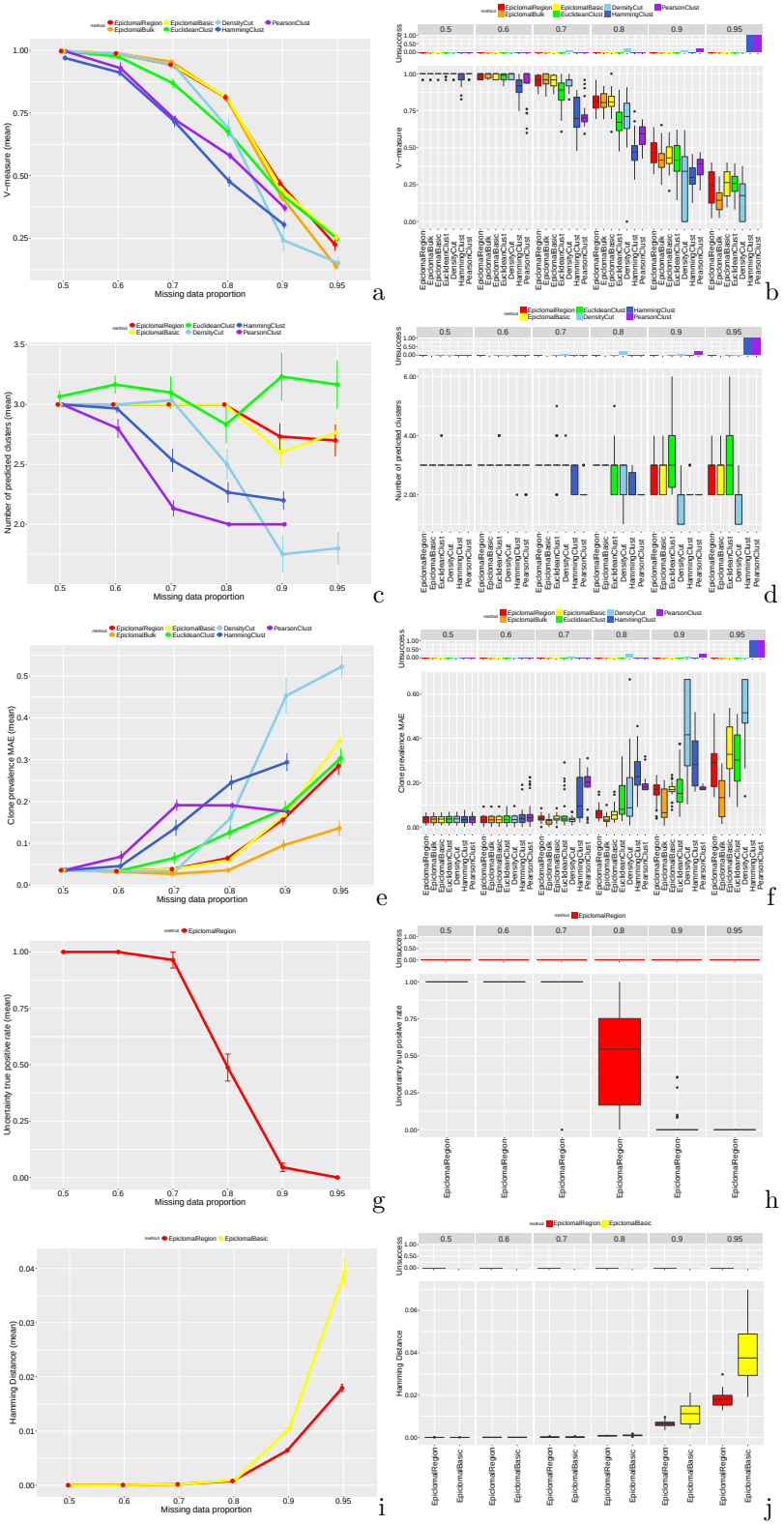
Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data

Supplementary Figures

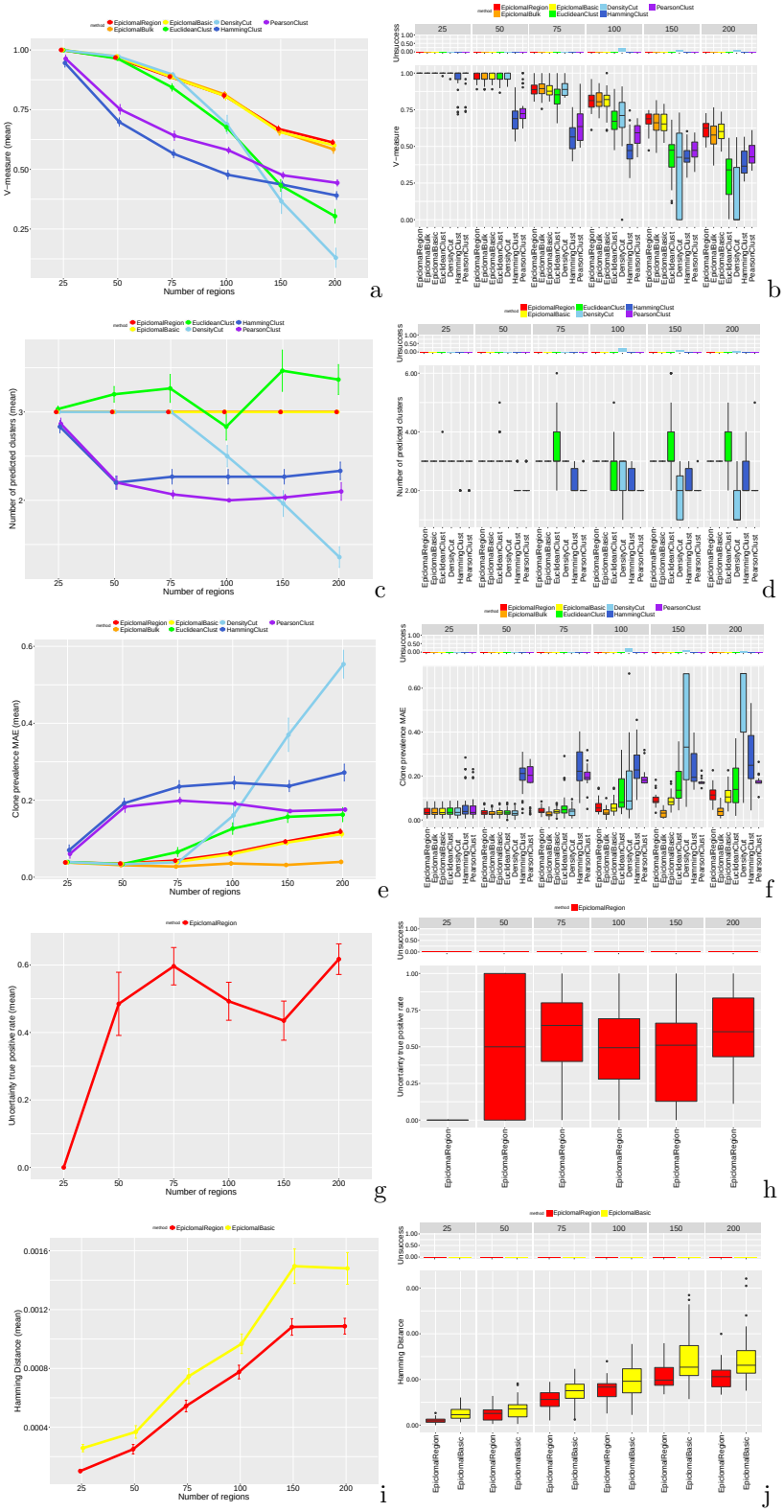
List of Supplementary Figures



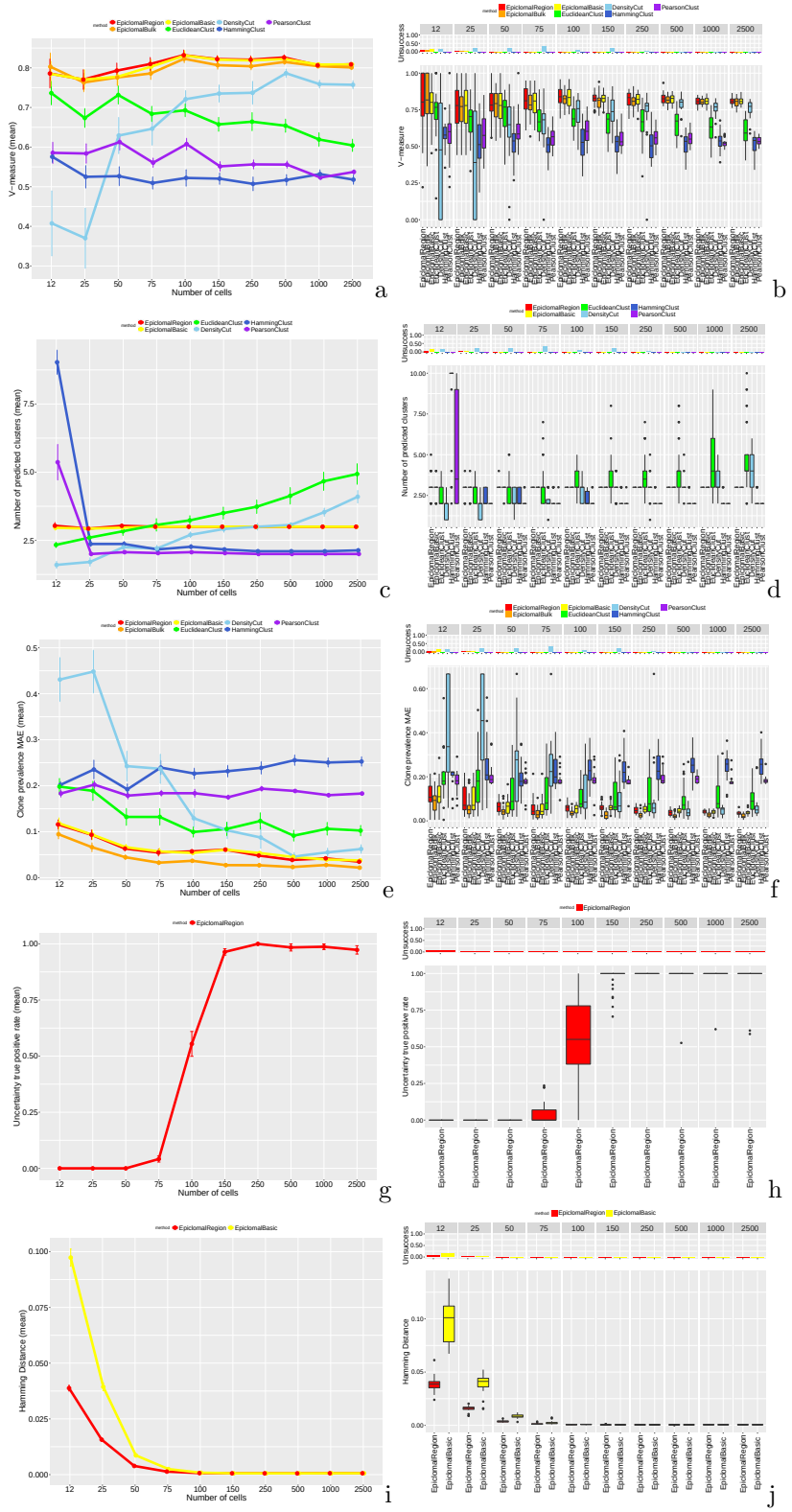
Supplementary Figure 1 Example of a DIC elbow plot for EpiclomalRegion for the InHouse data set with 10 000 loci. Our DIC algorithm selects only number of clusters for which there is a decrease of at least 2% in DIC values, green vertical line. Then, the elbow value is picked, red vertical line.



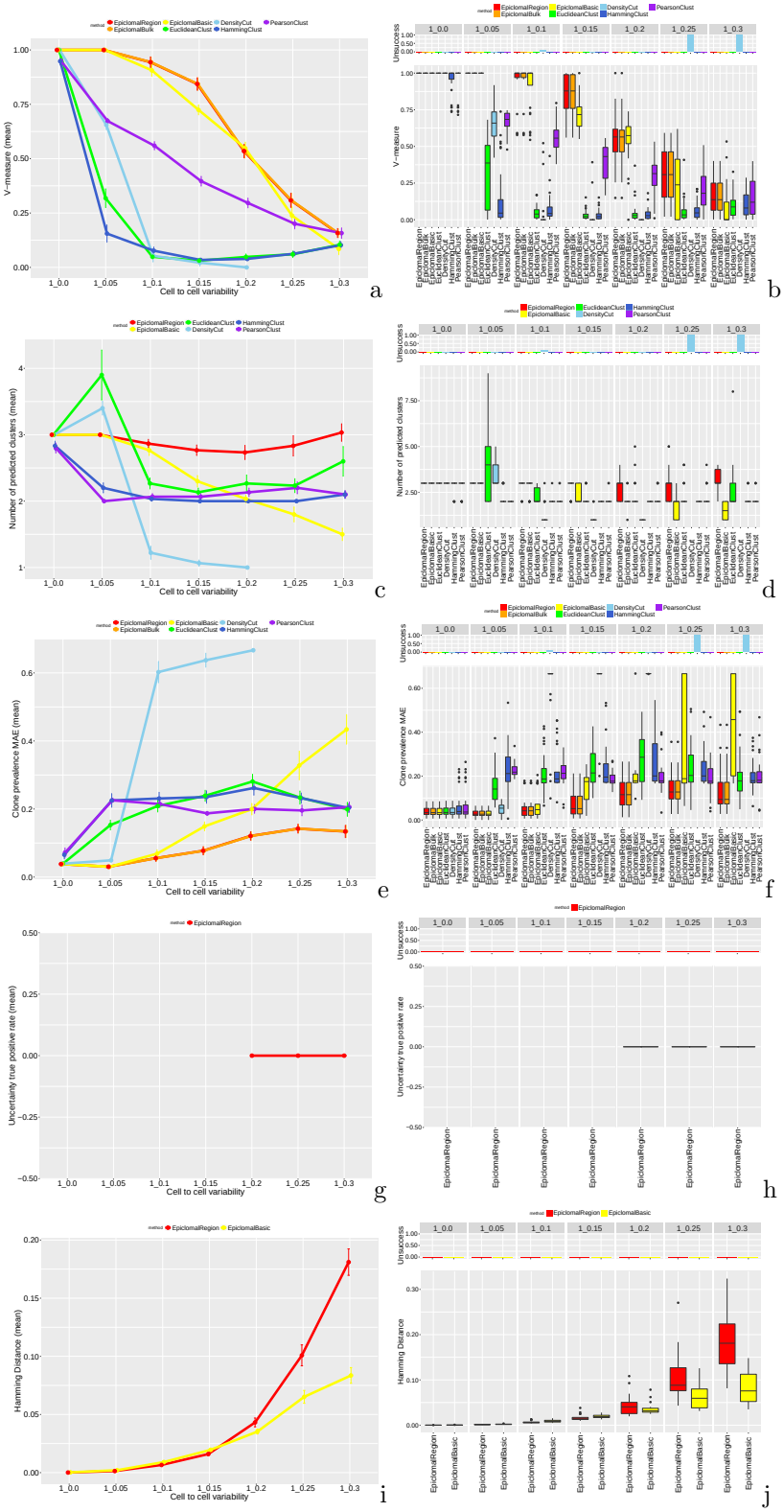
Supplementary Figure 2 Simulation results when varying the missing proportion. The left column shows the mean and error bars for a) V-measure, b) number of predicted epi-clones, c) epi-clone prevalence MAE, d) uncertainty true positive rate, e) hamming distance. The right column presents the corresponding boxplots. The barplots above the boxplots show the proportion of data sets for which a method failed to produce a result.



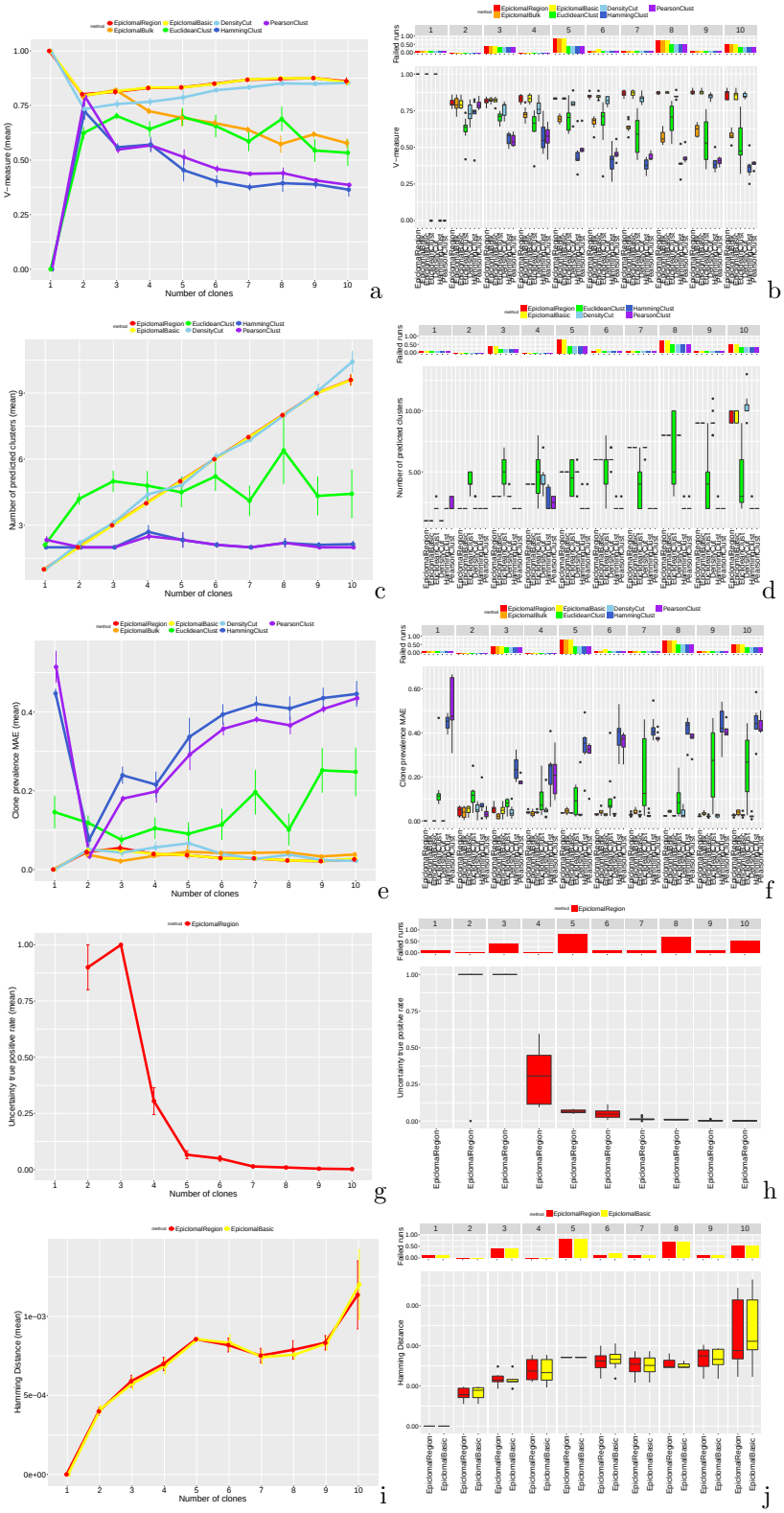
Supplementary Figure 3 Simulation results when varying the number of regions. The larger the number of regions the smaller the differences among epi-clones as the number of loci is fixed and our synthetic data generator only allows for one region to change at each new cluster generation. Plots as in 2. The Epiclomal methods perform better than the other methods and correctly predicts the number of clusters (panel c). As expected, all methods perform worse for the largest number of regions because there is less difference between epi-clones (200 regions correspond to 0.5% of the loci being different, while 25 regions correspond to 4% of the loci being different).



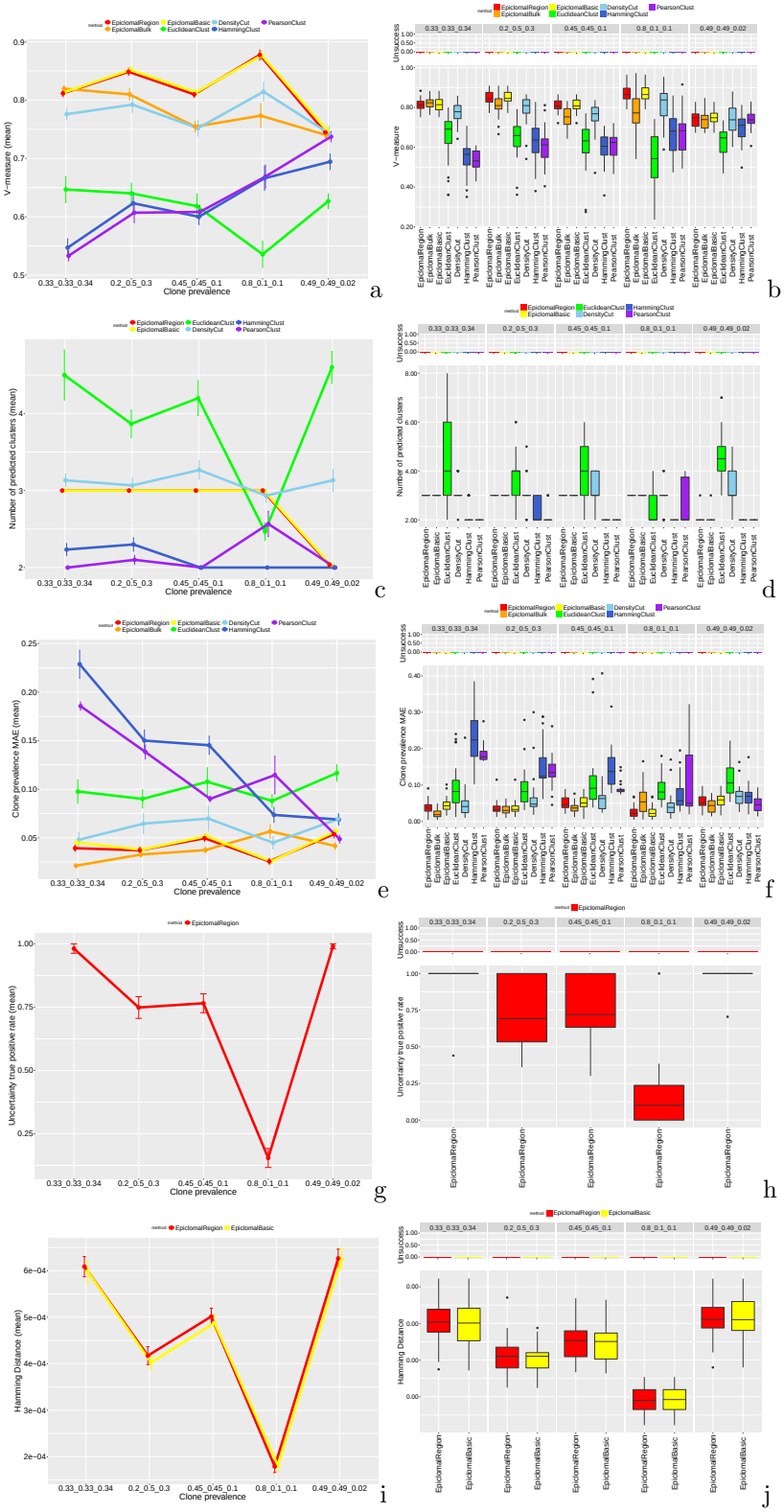
Supplementary Figure 4 Simulation results when varying the number of cells. Plots as in Figure 2. Increasing the number of cells does not improve the overall V-measure, except for DensityCut, but it reduces the variability of V-measure values. Epiclomal methods produce better V-measures in this case than the other methods. Panel e shows that EpiclomalBulk produced the best estimates of epi-clone prevalences. Starting at about 150 cells EpiclomalRegion was able to obtain an uncertainty true positive rate close to one (panel g).



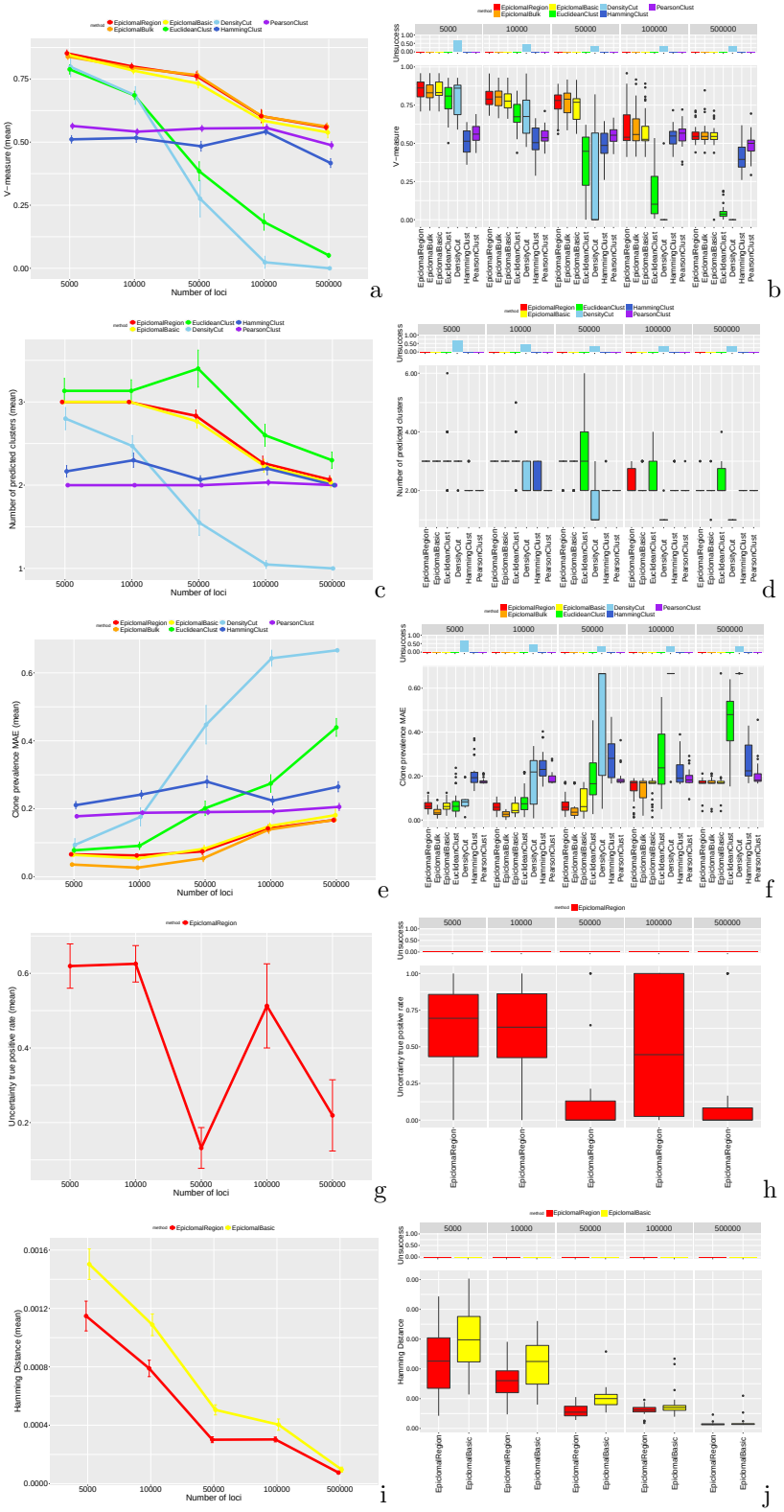
Supplementary Figure 5 Simulation results when varying the cell-to-cell variability. Plots as in Figure 2. The Epiclomal methods perform significantly better than the other methods when the cell-to-cell variability is between 0.05 to 0.20. A cell-to-cell variability of 0.20 means that at each CpG location that is not in the separating regions (regions that are different among clusters), 20% random cells are in the opposite methylation state than the remaining 80%. Hence a variability of 0.50 means that all the non-separating CpGs have completely random methylation states. When the variability is large (≥ 0.25), all methods perform poorly.



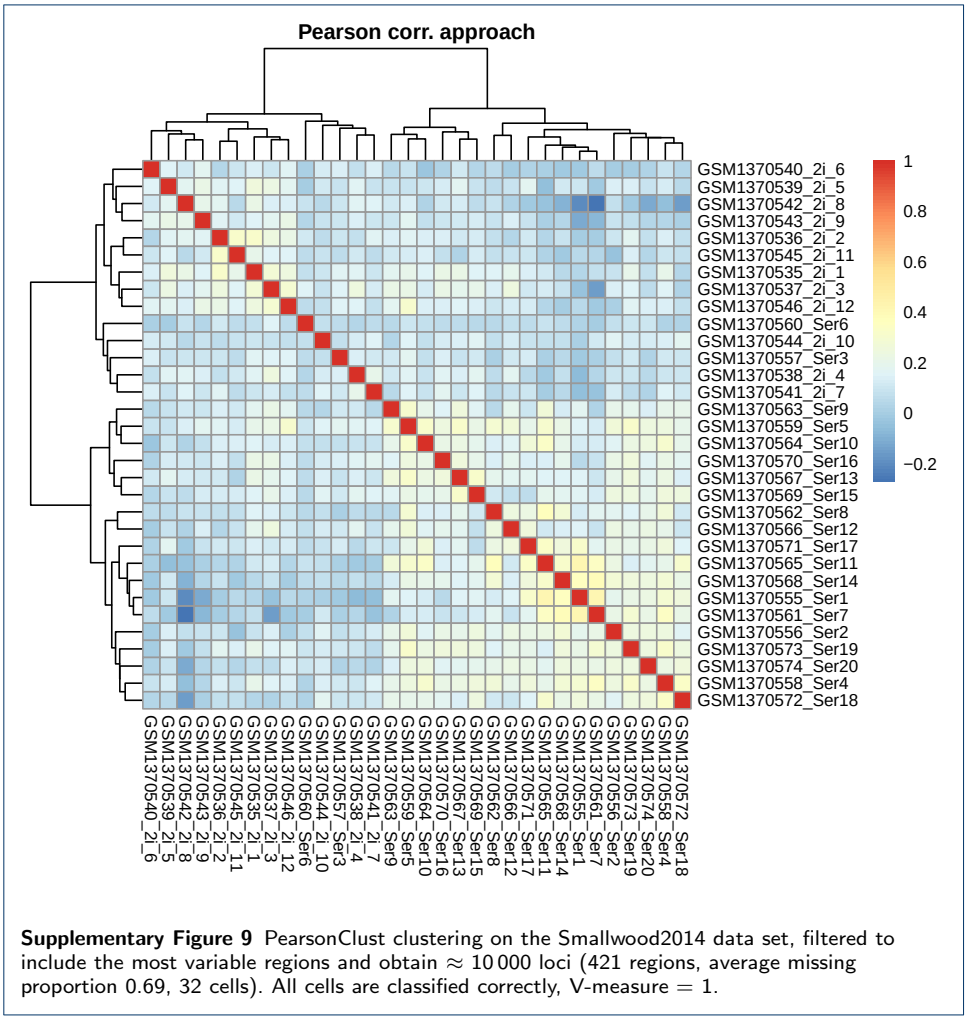
Supplementary Figure 6 Simulation results when varying the number of epi-clones from 1 to 10. Plots as in Figure 2. Epiclomal methods perform slightly better than DensityCut and much better than the other methods in terms of V-measure. Epiclomal and DensityCut are the only methods that can correctly predict the number of clusters (panel c).



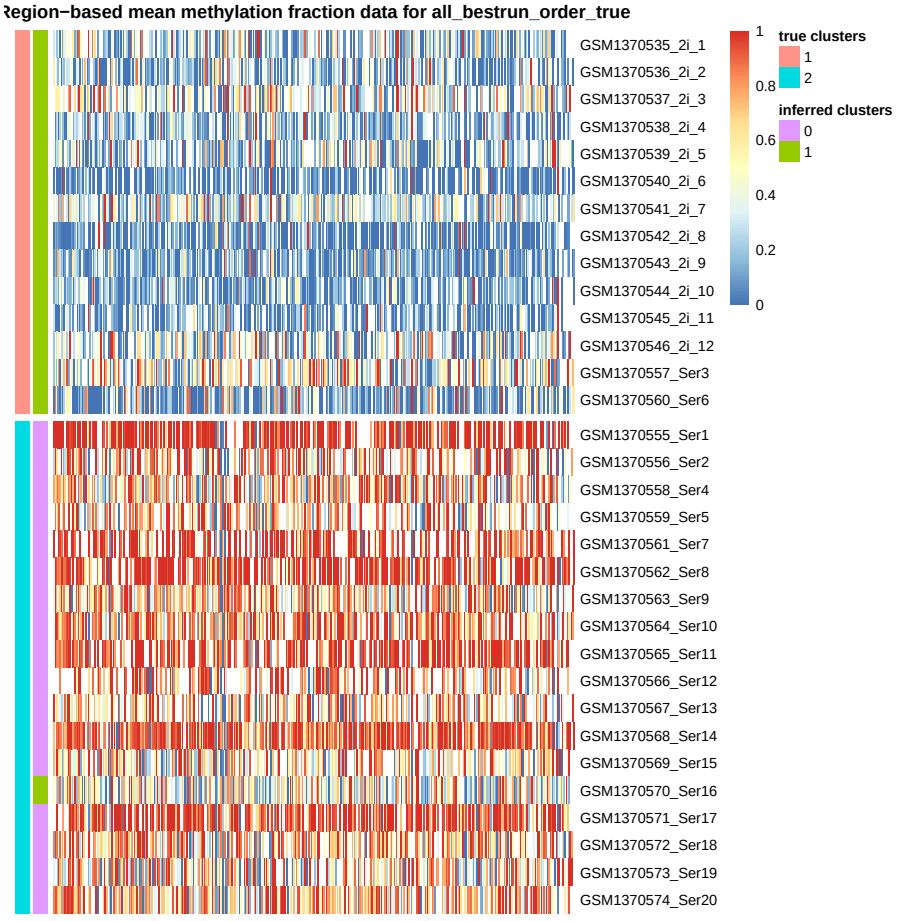
Supplementary Figure 7 Simulation results when varying the cluster prevalences from equal to very imbalanced. The number of clusters is three. Plots as in Figure 2. EpiclomalRegion and EpiclomalBasic give better V-measures than the other methods, and they correctly predict three clusters, except in the case where one of the clusters has only 2% of the cells. Interestingly, DensityCut does predict three clusters for this difficult case. The uncertainty true positive rate for EpiclomalRegion is above 0.75 for all cases except the case where two of the clusters have only 10% of the cells each.



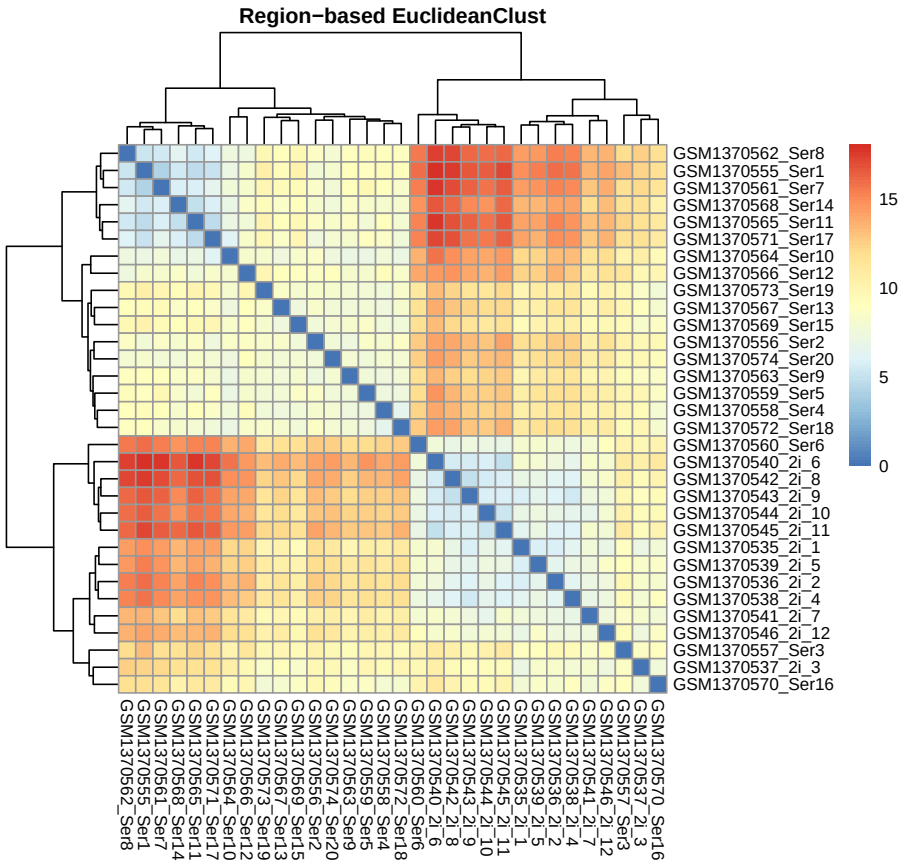
Supplementary Figure 8 Simulation results when varying the number of loci. Plots as in Figure 2. As we increase the number of loci the number of regions also increase, but their sizes remain fixed and so the amount of CpGs that make the clusters different. The performance of HammingClust and PearsonClust remains somewhat constant as we increase the number of loci, while the other methods show a decreasing pattern in performance. However, the Epiclomal methods still perform better in all cases than all the other methods, especially for 5 000, 10 000 and 50 000 loci. Therefore, this provides support to the strategy of selecting a smaller number of loci (under 50 000) in order to keep the true signal an eliminate noise when analyzing a real data set.



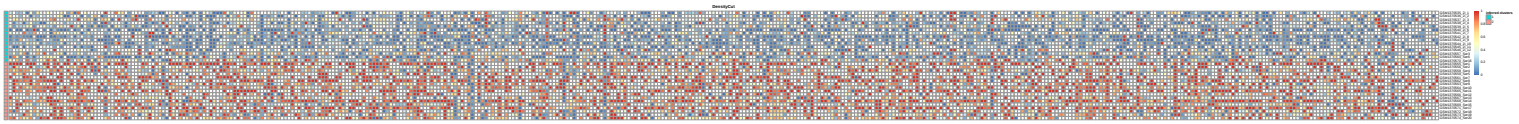
Supplementary Figure 9 PearsonClust clustering on the Smallwood2014 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci (421 regions, average missing proportion 0.69, 32 cells). All cells are classified correctly, V-measure = 1.



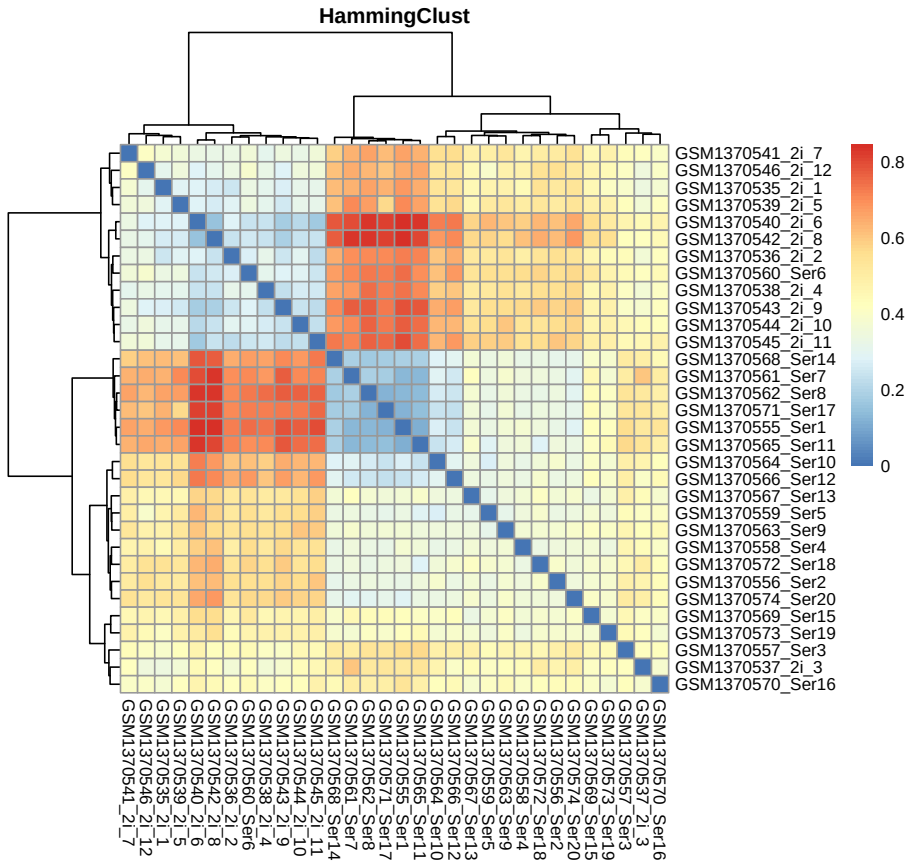
Supplementary Figure 10 EpiclomalRegion clustering on the Smallwood2014 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci (421 regions, average missing proportion 0.69, 32 cells). EpiclomalRegion returned the correct number of clusters (2), but missclassified one cell (GSM1370570_Ser16), V-measure = 0.83. The configuration with the misclassified cell obtained a slightly better DIC score than the ground truth.



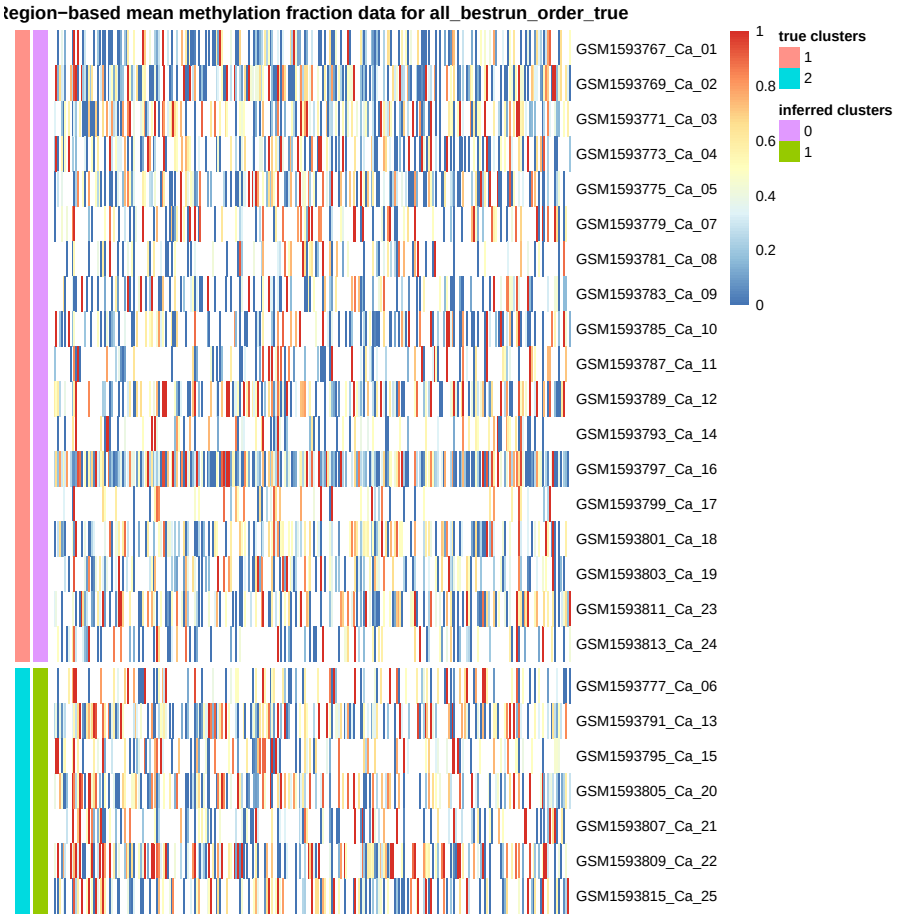
Supplementary Figure 11 EuclideanClust clustering on the Smallwood2014 data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (421 regions, average missing proportion 0.69, 32 cells). Similarly to EpiclomalRegion, EuclideanClust missclassified one cell (GSM1370570_Ser16), V-measure = 0.83.



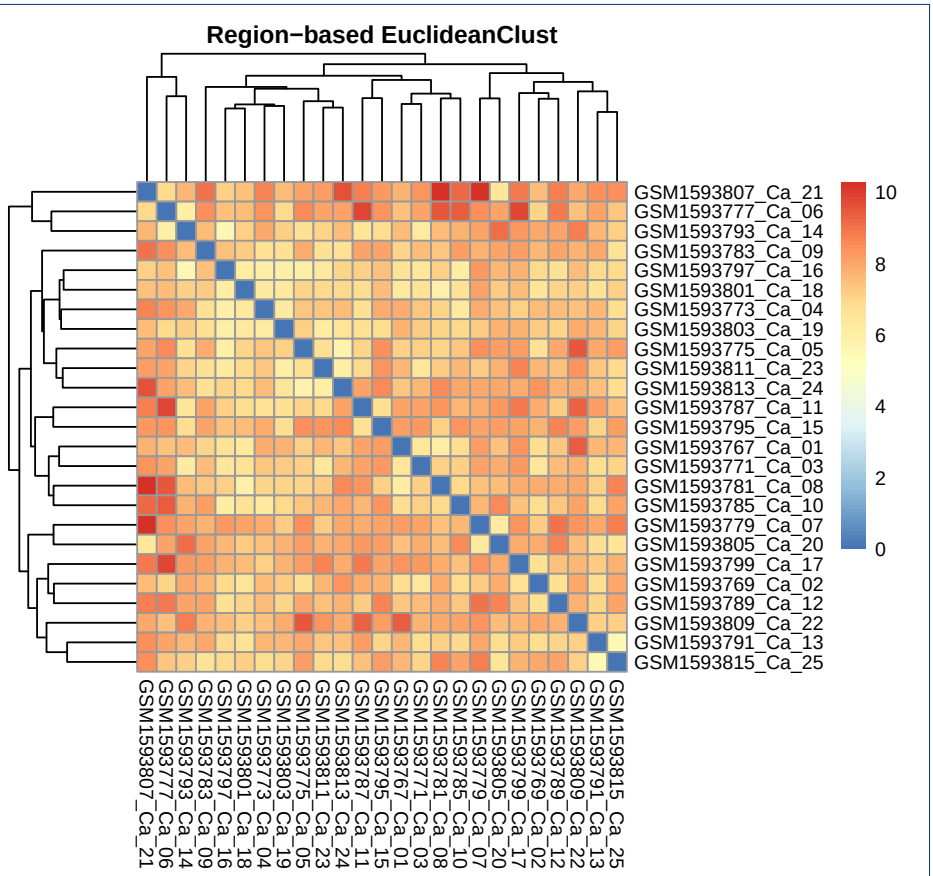
Supplementary Figure 12 DensityCut clustering on the Smallwood2014 data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (421 regions, average missing proportion 0.69, 32 cells). Similarly to EpiclomalRegion, DensityCut missclassified one cell (GSM1370570_Ser16), V-measure = 0.83.



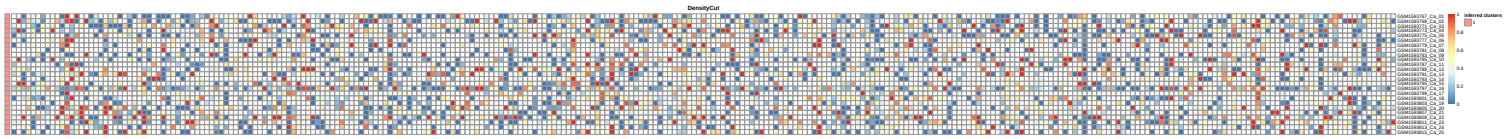
Supplementary Figure 13 HammingClust clustering on the Smallwood2014 data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (421 regions, average missing proportion 0.69, 32 cells). V-measure = 0.72.



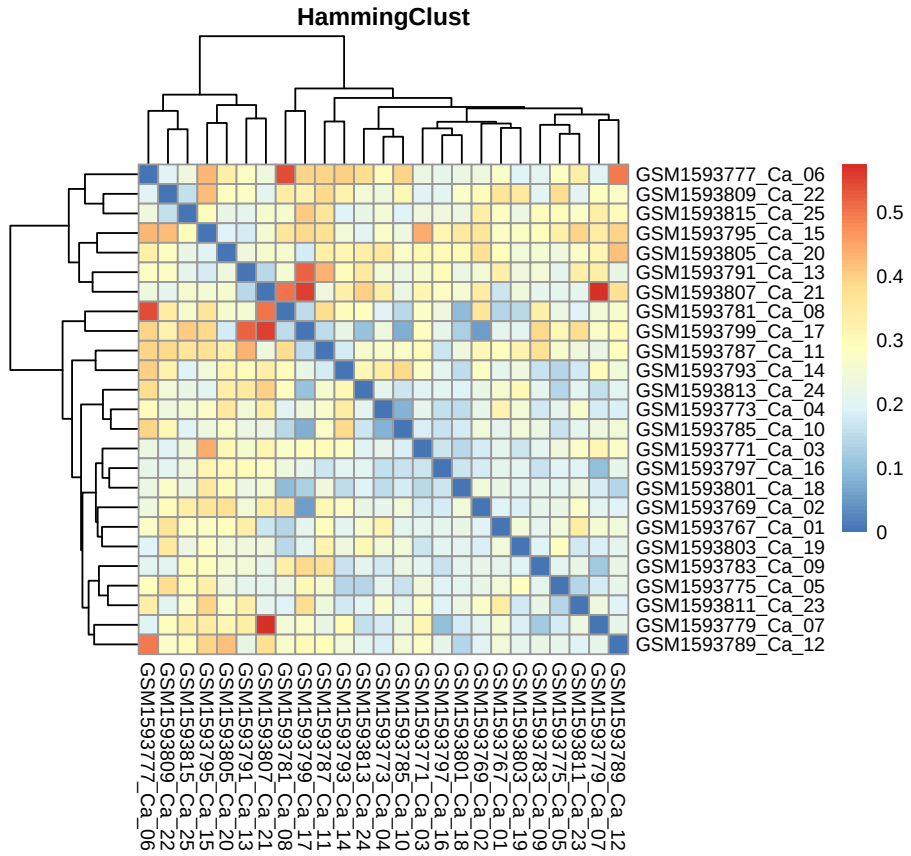
Supplementary Figure 14 EpiclomalRegion clustering on the Hou2016 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci (287 regions, average missing proportion 0.87, 25 cells). EpiclomalRegion correctly classified all cells, V-measure = 1.



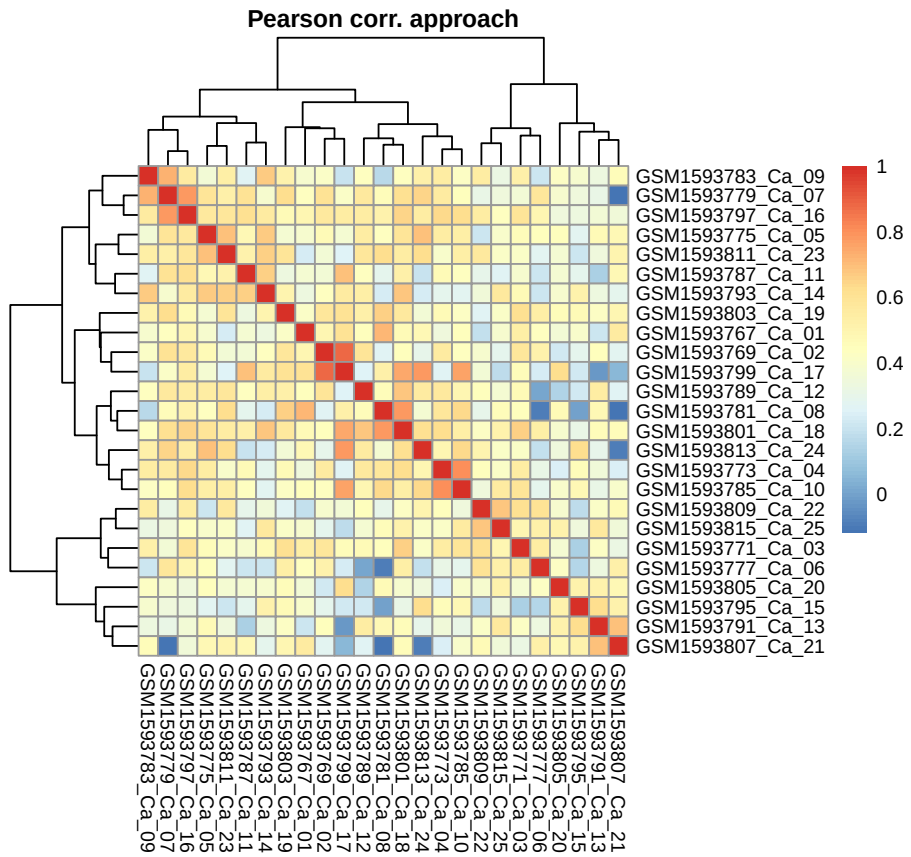
Supplementary Figure 15 EuclideanClust clustering on the Hou2016 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci (287 regions, average missing proportion 0.87, 25 cells). EuclideanClust predicted three clusters (correct is two), V-measure = 0.2.



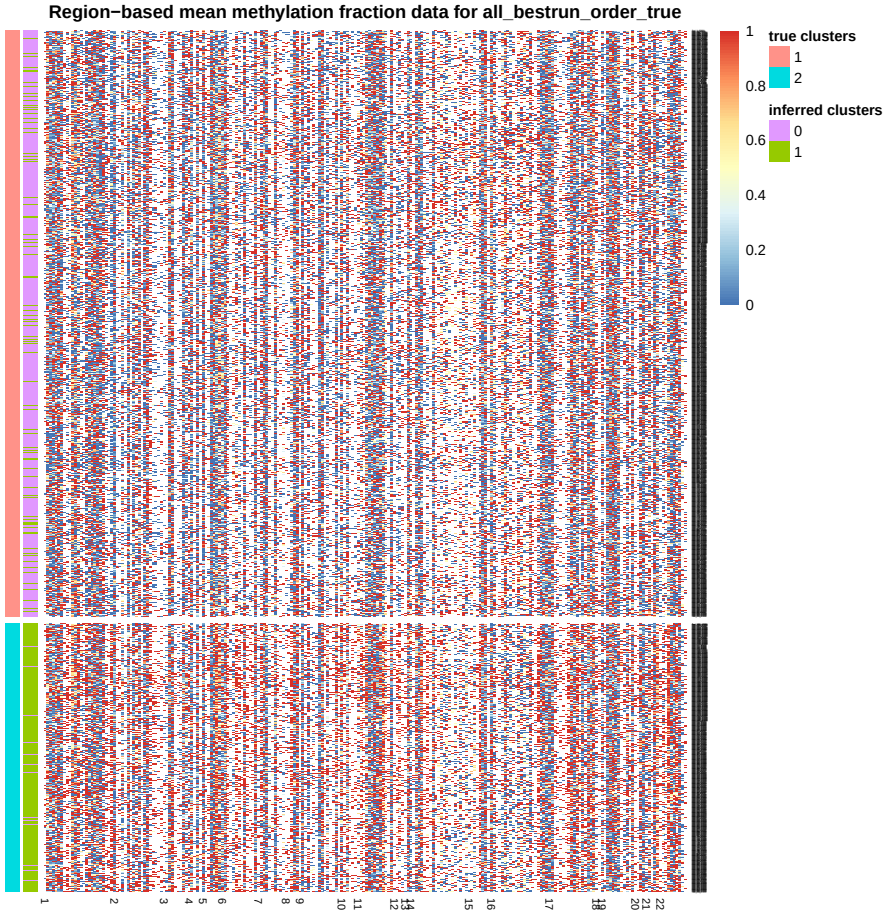
Supplementary Figure 16 DensityCut clustering on the Hou2016 data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (287 regions, average missing proportion 0.87, 25 cells). DensityCut predicts only one cluster (correct is two), V-measure = 0.



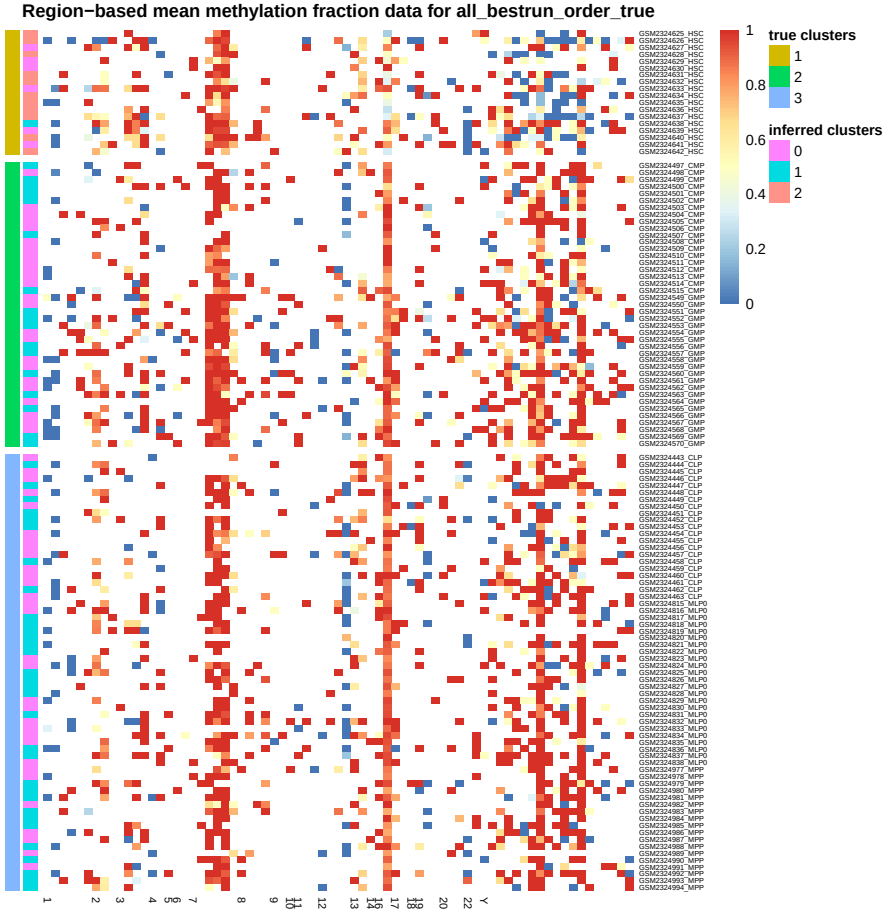
Supplementary Figure 17 HammingClust clustering on the Hou2016 data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (287 regions, average missing proportion 0.87, 25 cells). HammingClust correctly classified all cells, V-measure = 1.



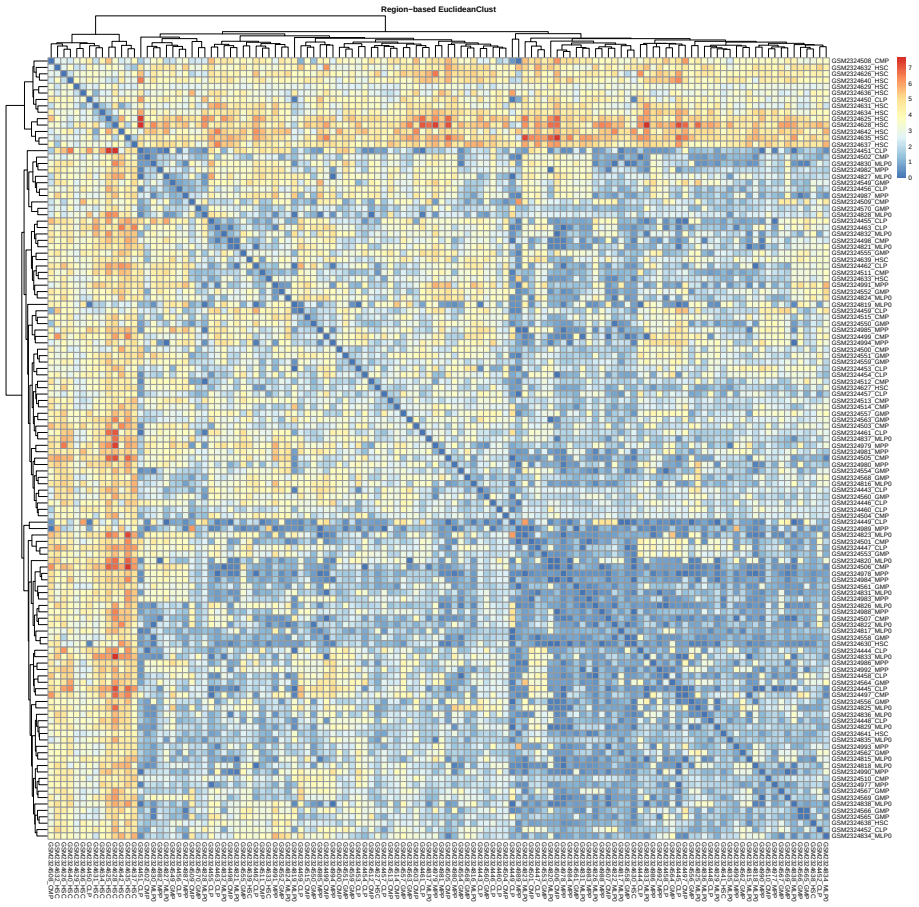
Supplementary Figure 18 PearsonClust clustering on the Hou2016 data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (287 regions, average missing proportion 0.87, 25 cells). V-measure = 0.77.



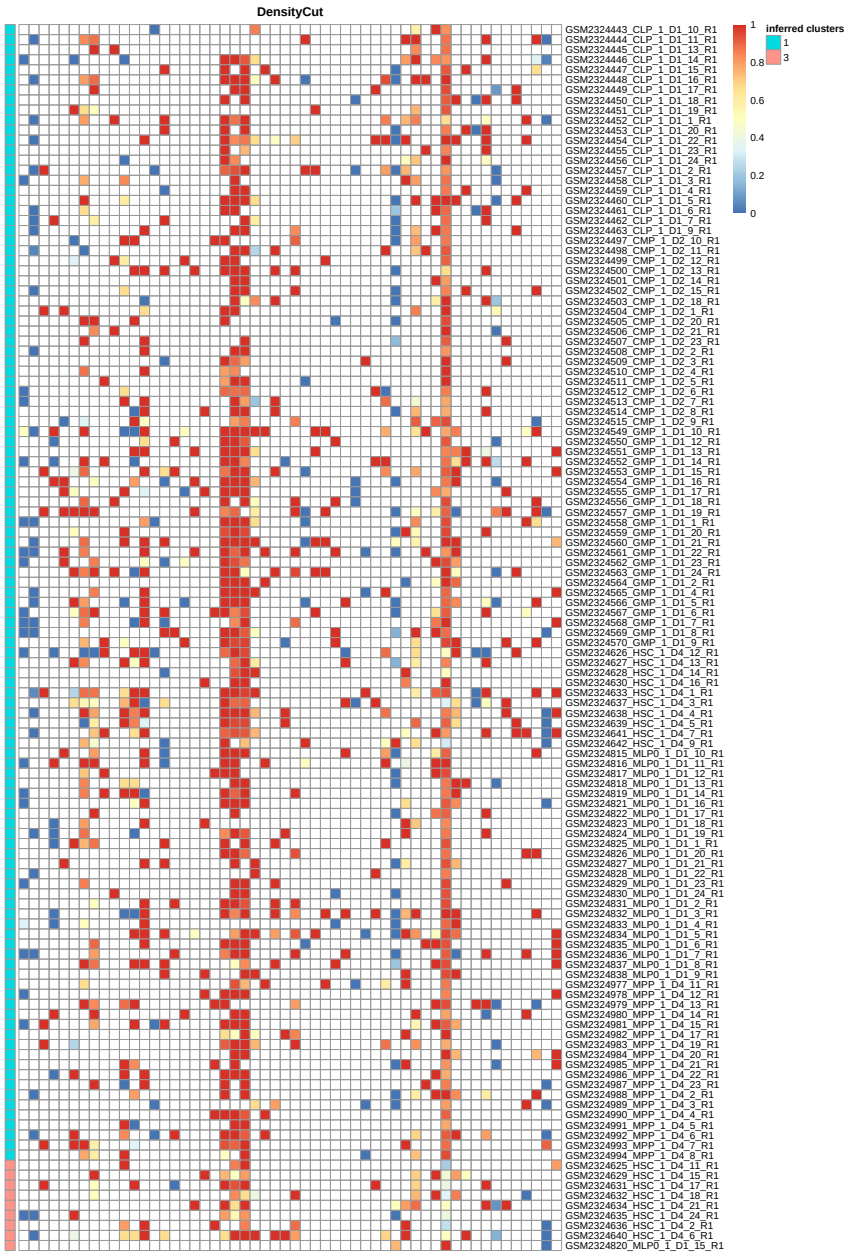
Supplementary Figure 19 EpiclomalRegion clustering on the Luo2017 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci (232 regions, average missing proportion 0.94, 2740 cells). EpiclomalRegion predicted the correct number of clusters (two), and misclassified some of the cells, V-measure = 0.52. Note that a number of cells are also misclassified in the Luo et al. paper.



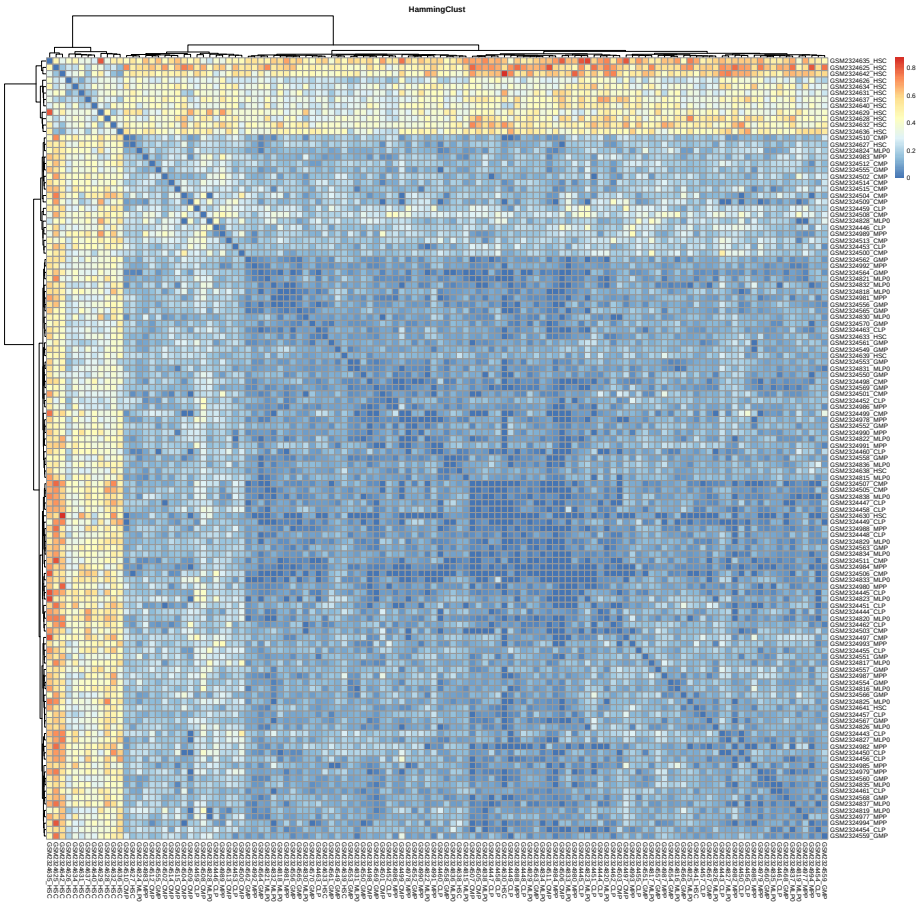
Supplementary Figure 20 EpiclomalRegion clustering on the Farlik2016 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci. For this set, due to the very large missing proportion of 0.98, there were much fewer loci in the final set than the desired number: 650 loci, 54 regions, average missing proportion 0.89, 122 cells. EpiclomalRegion correctly predicted the number of clusters (three), but misclassified a large number of cells, V-measure = 0.23.



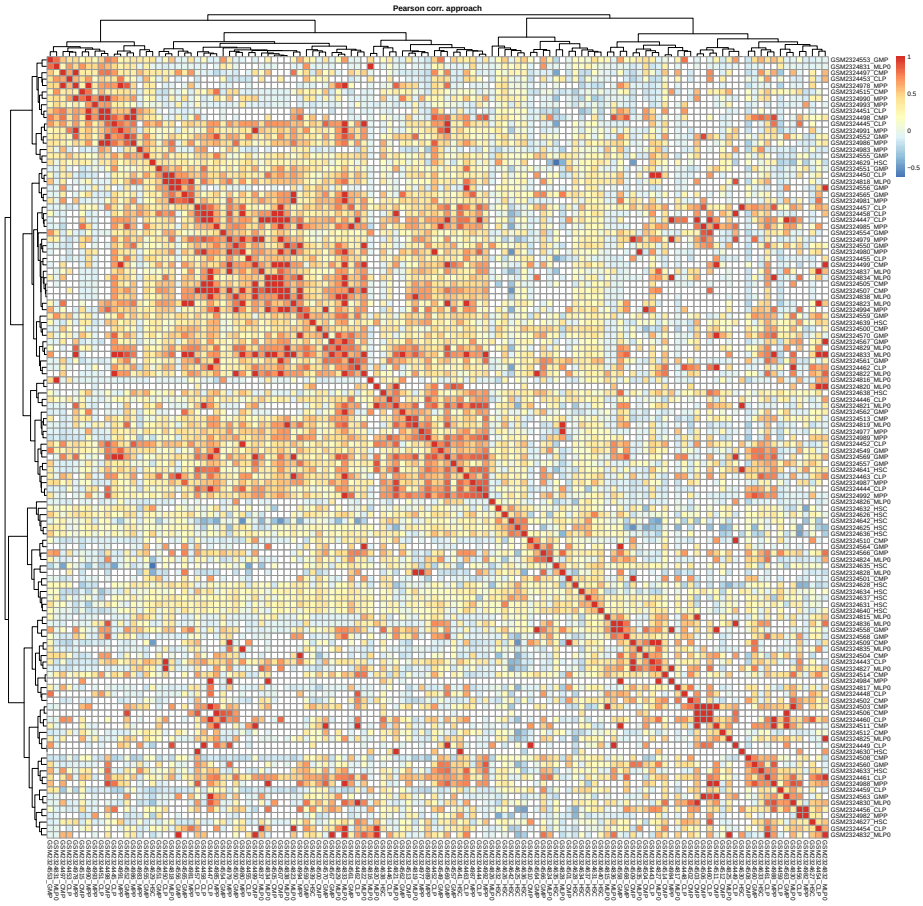
Supplementary Figure 21 EuclideanClust clustering on the Farlik2016 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci. For this set, due to the very large missing proportion of 0.98, there were much fewer loci in the final set than the desired number: 650 loci, 54 regions, average missing proportion 0.89, 122 cells. EuclideanClust predicted 2 clusters, V-measure = 0.27.



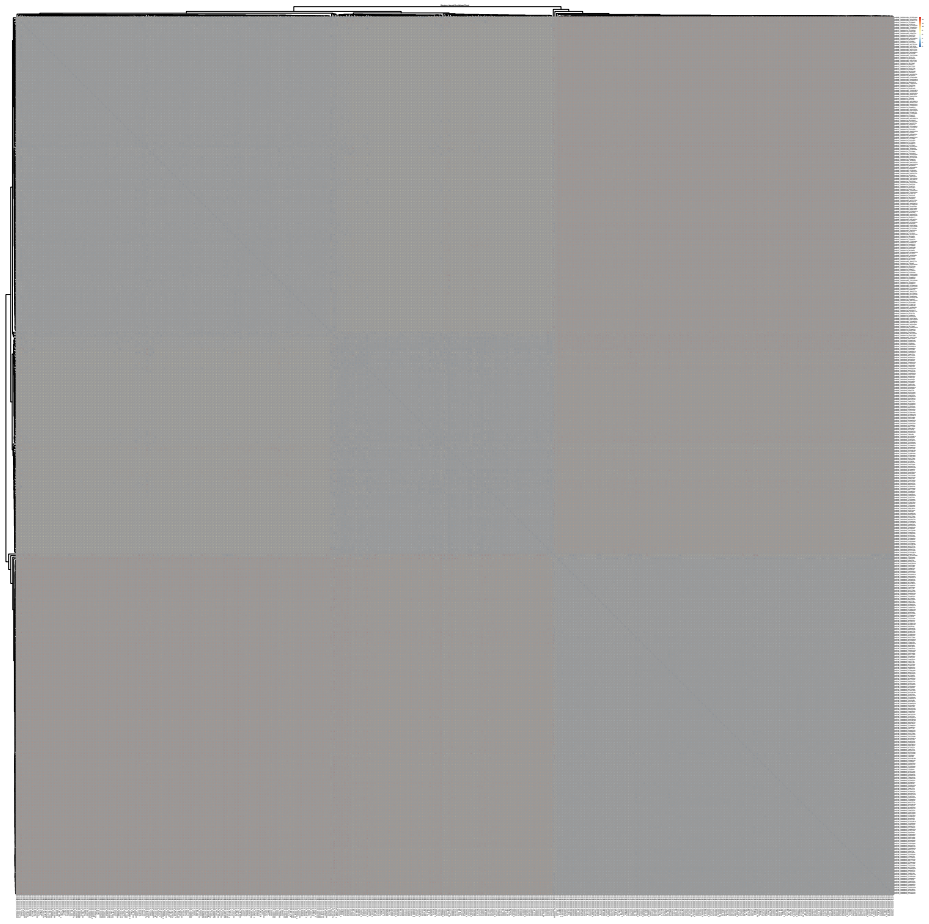
Supplementary Figure 22 DensityCut clustering on the Farlik2016 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci. For this set, due to the very large missing proportion of 0.98, there were much fewer loci in the final set than the desired number: 650 loci, 54 regions, average missing proportion 0.89, 122 cells. DensityCut predicted 2 clusters, V-measure = 0.19.



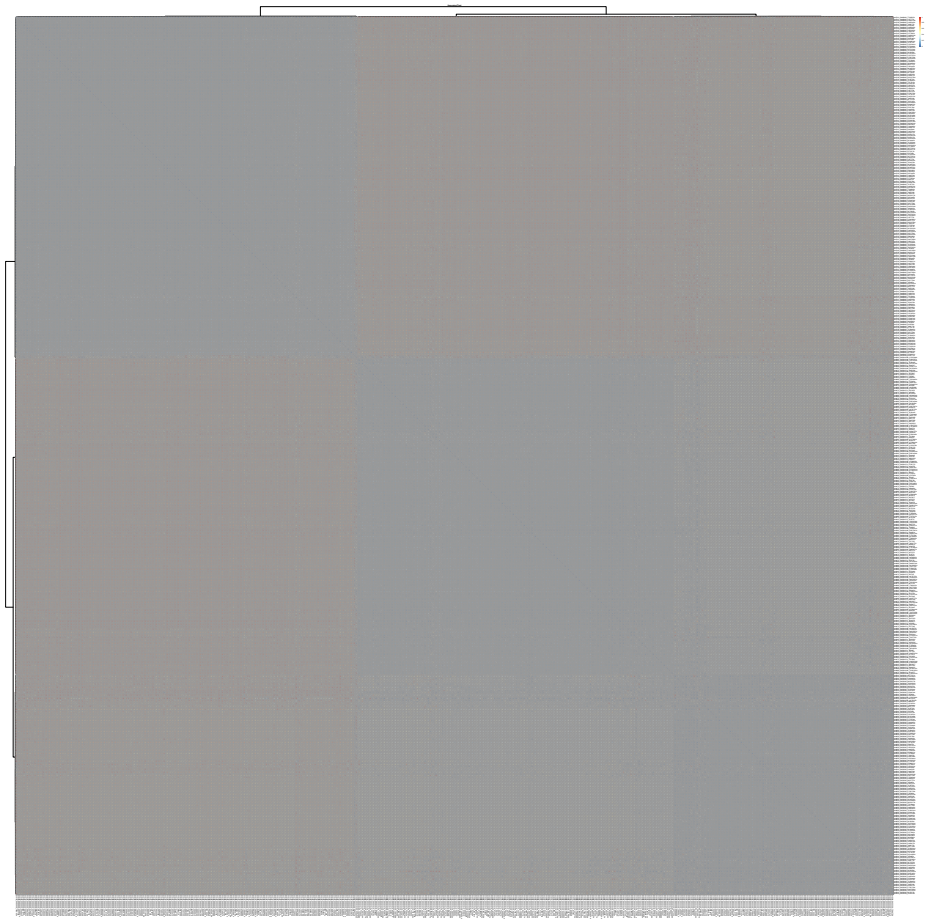
Supplementary Figure 23 HammingClust clustering on the Farlik2016 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci. For this set, due to the very large missing proportion of 0.98, there were much fewer loci in the final set than the desired number: 650 loci, 54 regions, average missing proportion 0.89, 122 cells. HammingClust predicted 2 clusters, V-measure = 0.35.



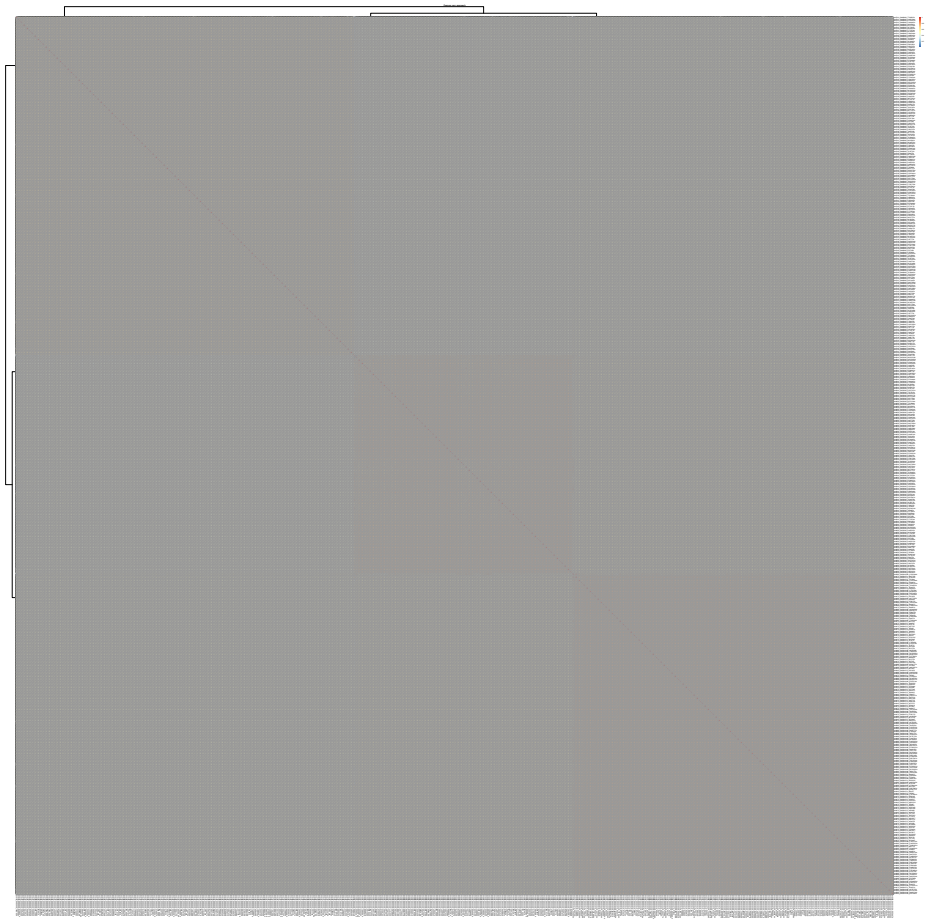
Supplementary Figure 24 PearsonClust clustering on the Farlik2016 data set, filtered to include the most variable regions and obtain ≈ 10 000 loci. For this set, due to the very large missing proportion of 0.98, there were much fewer loci in the final set than the desired number: 650 loci, 54 regions, average missing proportion 0.89, 122 cells. The CH-index cut failed in this case.



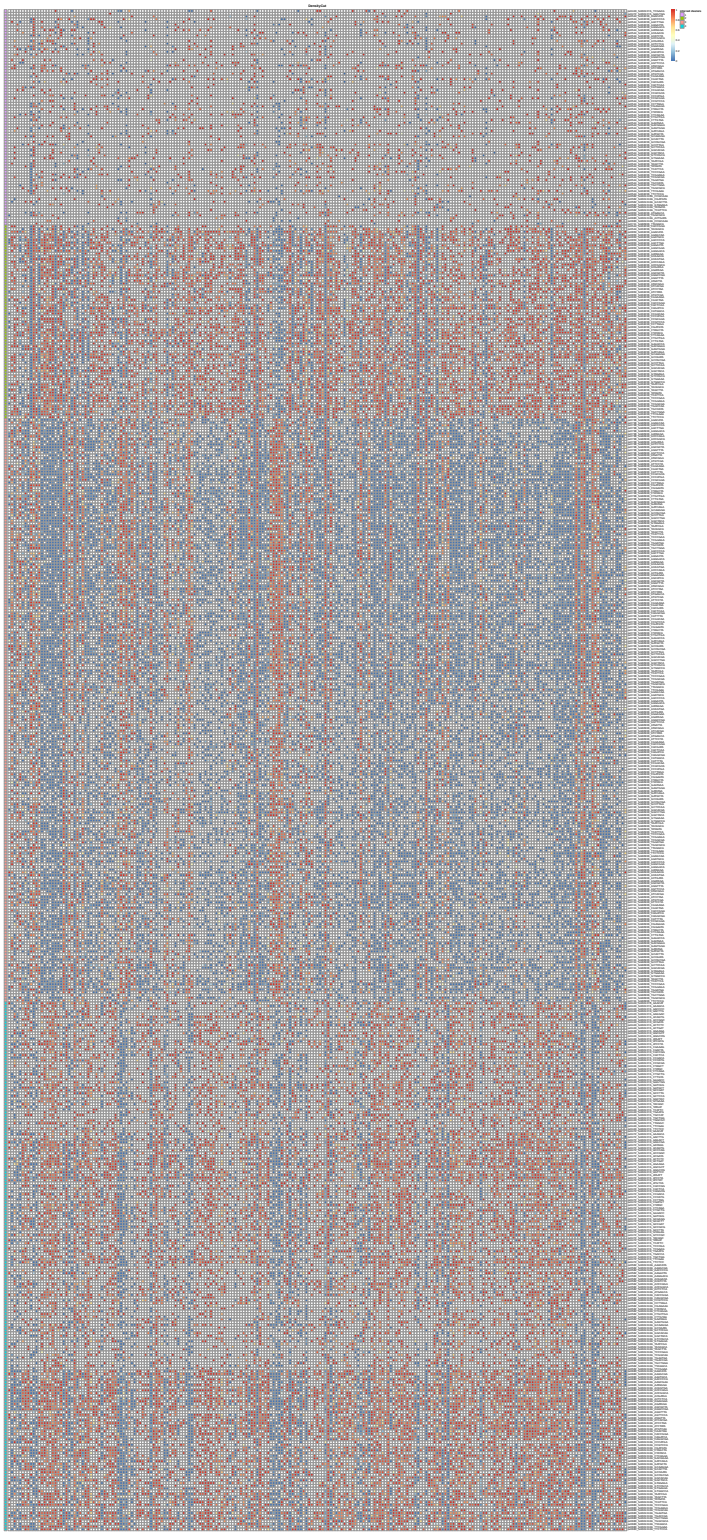
Supplementary Figure 25 EuclideanClust clustering on the InHouse data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (227 regions, average missing proportion 0.82, 558 cells). The CH-index cut of the hierarchical tree failed due to the large amount of missing data; however, the three clusters can be visually observed.



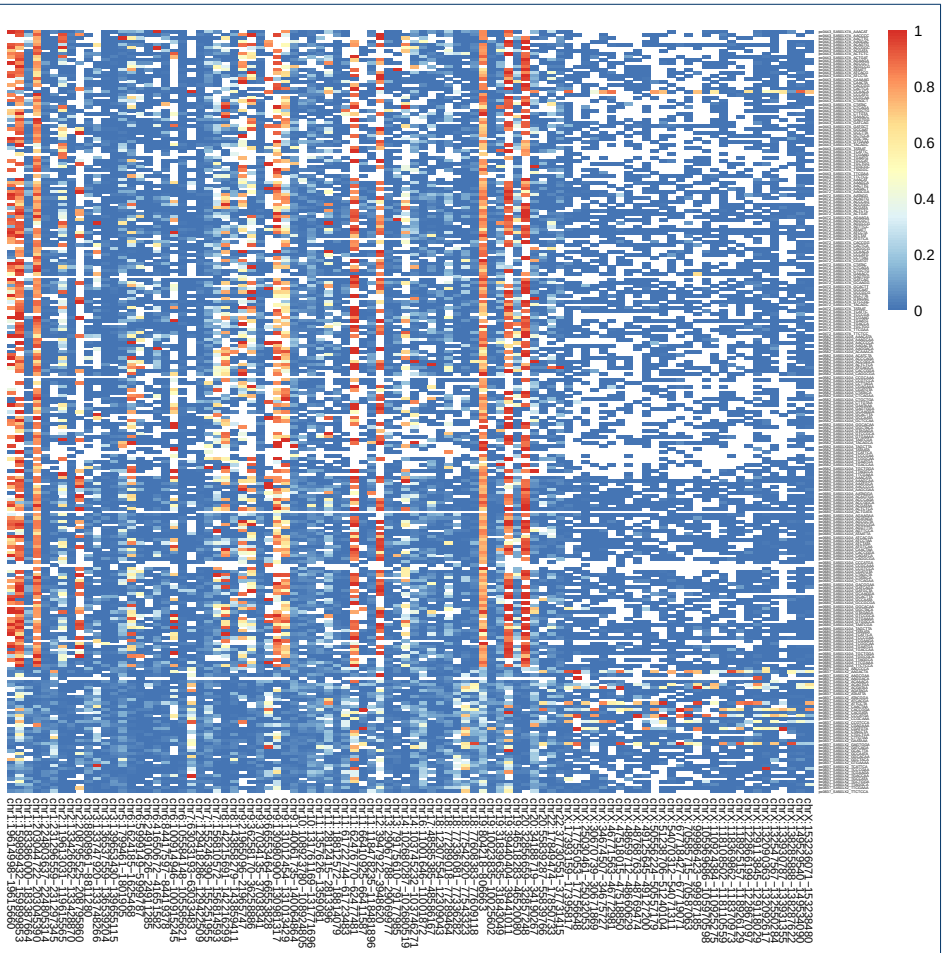
Supplementary Figure 26 HammingClust clustering on the InHouse data set, filtered to include the most variable regions and obtain $\approx$ 10 000 loci (227 regions, average missing proportion 0.82, 558 cells). The CH-index cut of the hierarchical tree failed due to the large amount of missing data; however, the three clusters can be visually observed.



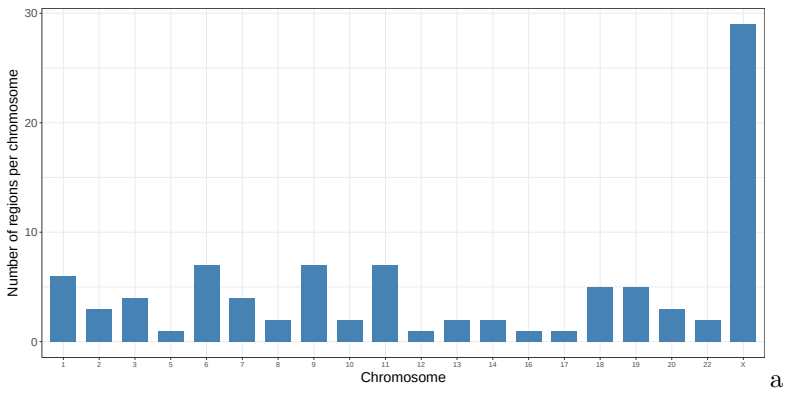
Supplementary Figure 27 PearsonClust clustering on the InHouse data set. For input, all the region with IQR < 0.01 were eliminated (1 019 956 loci in 14 585 regions, average missing proportion 0.79, 558 cells). The CH-index cut of the hierarchical tree failed due to the large amount of missing data; however, the three clusters can be visually observed.



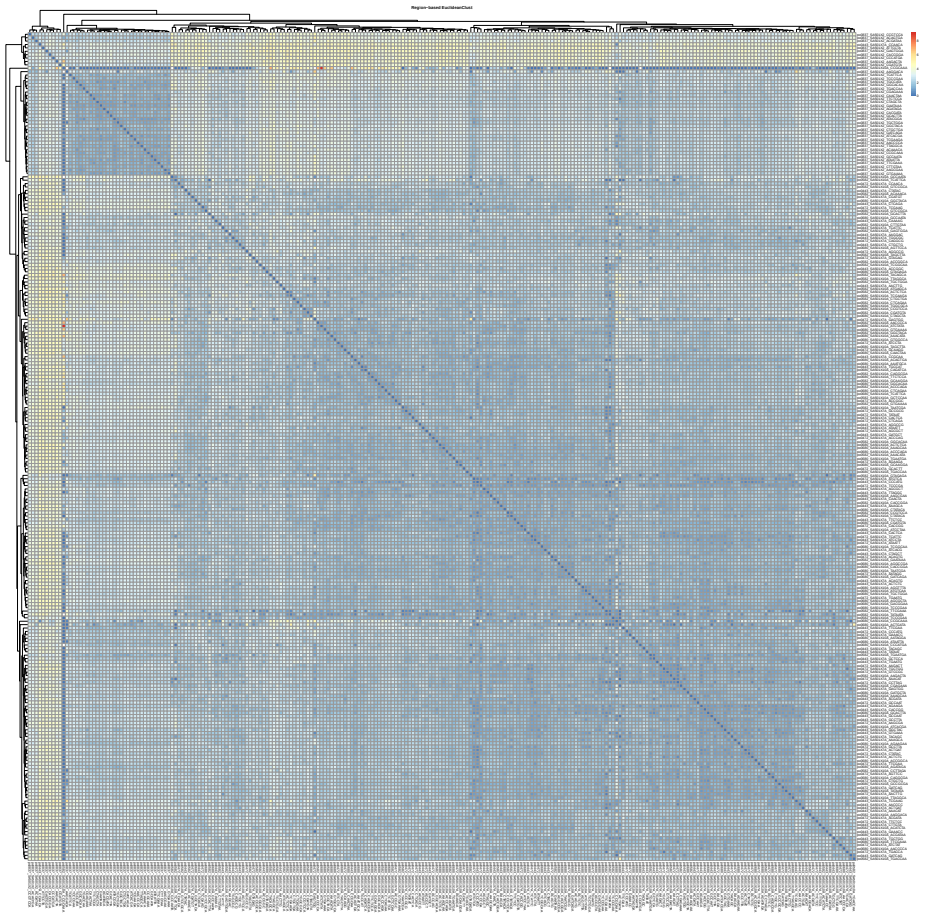
Supplementary Figure 28 DensityCut clustering on the InHouse data set, filtered to include the most variable regions and obtain ≈ 10 000 loci (227 regions, average missing proportion 0.82, 558 cells). DensityCut predicted 4 clusters (correct is 3), V-measure = 0.86.



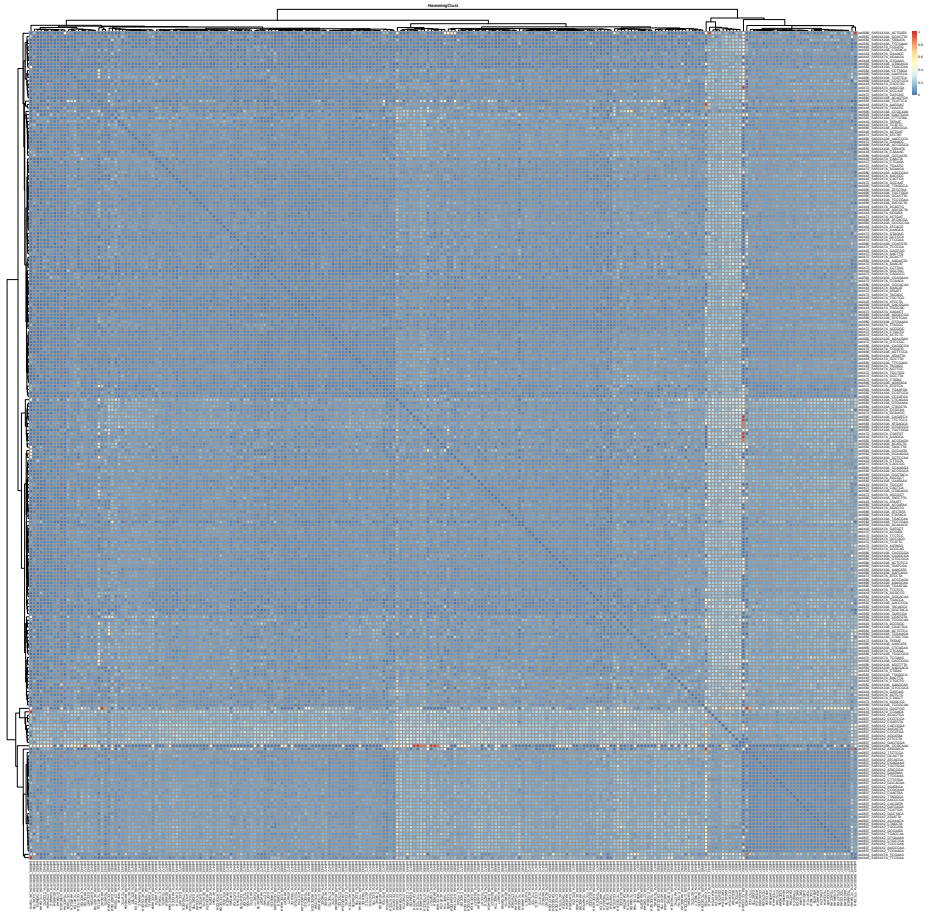
Supplementary Figure 29 Average methylation for the all the 244 SA501 cells and the 94 NMF-selected regions (CGIs), unclustered. The x-axis shows the region chromosome, start and end.



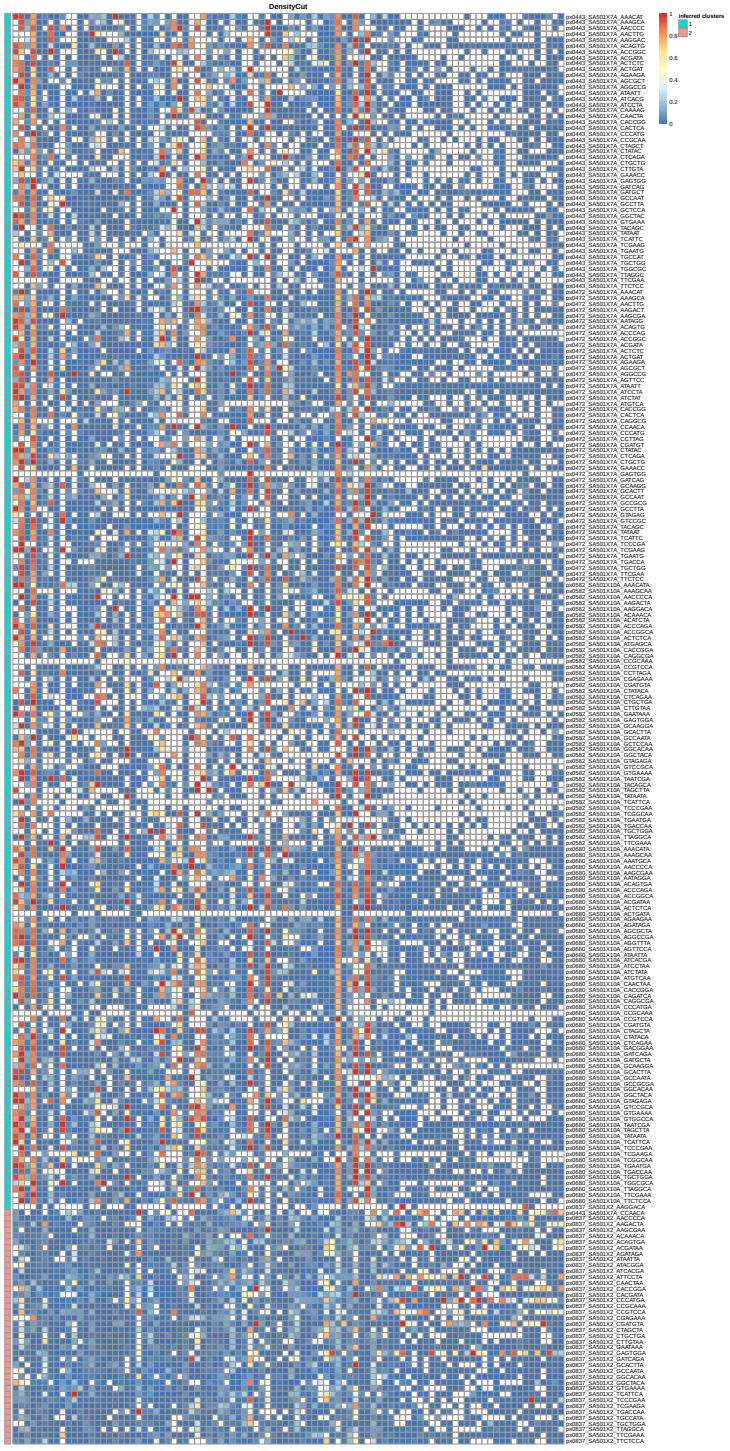
Supplementary Figure 30 a) Distribution of the 94 NMF-selected regions (CGIs) for SA501 data across the genome. b) Average number of CpGs across regions per chromosome.



Supplementary Figure 31 Hierarchical clustering obtained by EuclideanClust on the SA501 data set with the 94 selected regions. The CH-index that selects the number of clusters failed due to empty entries in the matrix. We can visually distinguish on the left/top side the two passage 2 clusters (1 and 2) obtained by EpiclomalRegion. Note the isolated 1-cell cluster containing cell px0582_SA501X10A_CCGCAAA. Although EuclideanClust considers this cell as being very similar to most other cells, this is only due to the fact that this cell has very high amount of missing data, as can be seen from the EpiclomalRegion plot in Figure 7a. For cells from passages 7 and 10, we can visually distinguish two to four clusters that do not match the EpiclomalRegion findings.



Supplementary Figure 32 Hierarchical clustering obtained by HammingClust on the SA501 data set with the 94 selected regions. The CH-index that selects the number of clusters failed due to empty entries in the matrix. We can visually distinguish on the right/bottom side the two passage 2 clusters (1 and 2) obtained by EpiclomalRegion. Again, as for EuclideanClust, note the isolated 1-cell cluster containing cell px0582_SA501X10A_CCGCAAA due to its high missingness. For the passage 7/10 cells, we can visually distinguish two clusters; however, they do not match the EpiclomalRegion findings.



Supplementary Figure 33 Cluster prediction by DensityCut on the SA501 data set with the 94 selected regions. DensityCut finds only two clusters, essentially separating passage 2 from the later passages.