Harvard University

Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 80

Model-based Clustering of Methylation Array Data: A Recursive-partitioning Algorithm for High-dimensional Data Arising as a Mixture of Beta Distributions

E. Andres Houseman, *University of Massachusetts Lowell and Harvard School of Public Health*
Brock C. Christensen, *Harvard School of Public Health*
Ru-Fang Yeh, *University of California San Francisco*
Carmen J. Marsit, *Brown University*
Margaret R. Karagas, *Dartmouth-Hitchcock Medical Center*
Margaret Wrensch, *University of California San Francisco*
Heather H. Nelson, *University of Minnesota School of Public Health*
Joseph Wiemels, *University of California San Francisco*
Shichun Zheng, *University of California San Francisco*
John K. Wiencke, *University of California San Francisco*
Karl T. Kelsey, *Brown University*

Model-based clustering of methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions

E. Andres Houseman*[1,2], Brock C. Christensen[3], Ru-Fang Yeh[4], Carmen J. Marsit[5],
Margaret R. Karagas[6], Margaret Wrensch[7], Heather H. Nelson[8], Joseph Wiemels[4],
Shichun Zheng[7], John K. Wiencke[7], and Karl T. Kelsey[5,9]

[1]Department of Work Environment
University of Massachusetts Lowell
Lowell, Massachusetts 01854

[2]Department of Biostatistics
[3]Department of Environmental Health
Harvard School of Public Health
Boston, Massachusetts 02115

[4]Department of Epidemiology and Biostatistics
[7]Department of Neurological Surgery
University of California San Francisco
San Francisco, California 94143

[5]Department of Pathology and Laboratory Medicine
[9]Department of Community Health
Center for Environmental Health and Technology
Brown University
Providence, Rhode Island 02912

[6]Department of Community and Family Medicine
Dartmouth-Hitchcock Medical Center
Lebanon, New Hampshire 03756

[8] Division of Epidemiology and Community Health
University of Minnesota School of Public Health
Minneapolis, Minnesota 55455

Key words

Beta Mixture, DNA Methylation, Cancer, Epigenetics, Mixture Model
*Communicating author

ABSTRACT

Background: Epigenetics is the study of heritable changes in gene function that cannot be explained by changes in DNA sequence. One of the most commonly studied epigenetic alterations is cytosine methylation, which is a well recognized mechanism of epigenetic gene silencing and often occurs at tumor suppressor gene loci in human cancer. In order to understand methylation in normal tissue, we have collected 217 normal tissue samples on 11 types of normal tissue and used the Illumina GoldenGate platform to assess methylation at 1505 loci associated with over 800 cancer-related genes. While model-based cluster analysis is often used to identify methylation subgroups in data, it is unclear how to cluster methylation data from arrays in a scalable and reliable manner.

Results: We propose a novel model-based recursive-partitioning algorithm to navigate clusters in a beta mixture model. We present simulations that show that the method is more reliable than competing nonparametric clustering approaches, and is at least as reliable as conventional mixture model methods. We also show that our proposed method is more computationally efficient than conventional mixture model approaches. We demonstrate our method on the normal tissue samples and show that the clusters are associated with tissue type as well as age.

Conclusions: Our proposed recursively-partitioned mixture model is an effective and computationally efficient method for clustering methylation data.

BACKGROUND

Epigenetics is the study of heritable changes in gene function that cannot be explained by changes in DNA sequence [1]. One of the most commonly studied epigenetic alterations is cytosine methylation, which occurs in the context of a CpG dinucleotide. Concentrations of CpGs known as CpG islands, when sufficiently methylated, are associated with transcriptional gene silencing tantamount to "one hit" as part of Knudson's two hit hypothesis of carcinogenesis [2]. DNA methylation associated gene silencing is a well recognized epigenetic mechanism that often occurs at tumor suppressor gene (TSG) loci in human cancer. Hundreds of reports of methylation induced silencing at TSGs in virtually all types of human cancer have been published [3, 4].

While there has been a tremendous effort to characterize epigenetic alterations in cancer, surprisingly little work has been done in disease-free tissues. There is a basic need for epigenetic profiling of normal tissues to better understand the contribution of these profiles to tissue specificity, especially in the context of phenotypically important CpGs, where deregulation is associated with human diseases such as cancer. While efforts to characterize the methylation profiles of normal tissues in humans and mice have begun, and certain themes are slowly becoming apparent, relatively few reports have emerged [4-11]. Most CpGs or CpG regions have been found to have a bimodal distribution of methylation profiles, either hypo- or hypermethylated. Another theme is the disproportionate location of cell or tissue type dependent differentially methylated regions in non-CpG island sites [4, 8]. Furthermore, the Human Epigenome Project showed that the human major histocompatability loci have differential methylation across various human tissues, but that differential methylation does not necessarily lead to differential expression [8]. It is therefore critical to first outline the basal-state of phenotypically important epigenetic marks that are known to contribute to cancer in order to have a background for comparison to other normal and diseased tissue. This approach is best suited to foster the discovery of epigenetic profiles that are associated with particular disease states or covariates that contribute to pathogenesis.

Cluster analysis is often used to identify methylation subgroups in data [12, 13] and, in particular, Siegmund (2004) argues that model-based clustering techniques are often superior to nonparametric approaches [13]. Large-scale methylation arrays are now available for studying methylation genome-wide; the GoldenGate methylation platform from the manufacturer Illumina (San Diego, CA) simultaneously measures cytosine methylation at 1505 phenotypically-important loci associated with over 800 cancer-related genes. The result of the array is a sequence of “beta” values, one for each locus, calculated as the average of approximately 30 replicates (approximately 30 beads per site per sample) of the quantity $\max(M , 0)/(|U| + |M| + Q)$, where U is the fluorescent signal from an unmethylated allele on a single bead, M is that from a methylated allele, and Q is a constant chosen to ensure that the quantity is well-defined; an absolute value is used in the denominator of the formula to compensate for negative signals due to background subtraction. Note that each beta value is an approximately continuous variable lying between zero and one, where zero represents an unmethylated locus and one represents a methylated locus. As such, the beta value is appropriately modeled with a beta distribution. A data set consisting of such sequences produces a high-dimensional data-analysis problem which poses challenges for traditional clustering approaches. In addition, analysis of heterogeneous tissue data can lead to a large number of clusters, as we demonstrate below, which presents further challenges for clustering techniques. For example, nonparametric approaches rely on a choice of metric, which may be difficult to justify in the context of high dimensions and numerous clusters. On the other hand, in model-based clustering, multi-modality of the data likelihood may lead to numerical instability or difficulty in determining the best solution [14].

We propose a novel method for model-based clustering of data of the type produced by Illumina GoldenGate arrays. Our method makes use of a beta mixture model [15]. Although one could use BIC (or similar quantities) to select the number of clusters in the data set, we propose a recursive-partitioning algorithm that provides the number of clusters and a reliable solution in a shorter amount of time than sequential attempts with different numbers of assumed clusters. This is similar in spirit to the idea of recursive partitioning used in Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH, [16]), in which clusters are recursively partitioned using a nonparametric algorithm such as PAM [17]. Our method is also an unsupervised variant of Hierarchical Mixtures of Experts [18], a fuzzy version of CART [19]. We also propose a method for reducing the number of loci considered in the analysis, and selecting the optimal number using an "augmented" BIC statistic. We also present a simulation study comparing its properties to those of competitor methods. Finally, we demonstrate the methodology on GoldenGate methylation array data obtained from 217 normal tissue samples.

RESULTS

Simulation

Table 1 displays the classification error and computation time resulting from our simulation study. In both cases simulated, the mixture models outperformed the nonparametric methods in terms of classification error. For Case I, based on normal tissue data and described below in the Methods section, the proposed recursive-partitioned mixture model outperformed all other methods, including the sequentially-fit mixture models. For Case II, based on artificial parameters representing extremes of mean and variability, both mixture models performed equally well. In general, the mixture models had longer computation time than the nonparametric methods; however, we note that the mixture models were implemented as interpreted code in R, while the nonparametric methods were precompiled programs with R interfaces. Note that the recursively-partitioned mixture model was anywhere from 3 to 8 times faster than the sequentially fit mixture model.

For Case I, the median number of classes obtained for HOPACH with the "best" setting ranged from 22 (*J*=1000) to 36.5 (*J*=25), where *J* was the number of loci considered in the analysis. For HOPACH with the "greedy" setting, the median number ranged from 4 (*J*=1000) to 7 (*J*=25), with the correct number 5 being estimated at *J*=500. In contrast, the mixture models estimated a median of 5 classes for all values of *J*. In addition, the mixture models almost always obtained the correct number of classes. For Case II, HOPACH with the "best" setting obtained median number of classes between 17 and 24, HOPACH with the "greedy" setting obtained median number of classes between 4 and 8, (4 classes at *J*=10). For the two lower values of *J*, the mixture models obtained 2 classes, and for the two higher values of *J*, the mixture models obtained the correct number of classes, 4. Thus, for the cases considered, the mixture models almost always found the correct number of classes if *J* was high enough.

In the Methods section we propose an augmented BIC as a comparative measure of model fit for different numbers *J* of loci. For Case I, the mixture models always minimized the augmented BIC at *J*=1000, while for Case II, the mixture models always minimized the augmented BIC at *J*=25. For Case I, nearly all 1413 dimensions were at least somewhat informative; it is interesting to note that *J* was always minimized at its highest value for this case. For Case II, the number of informative dimensions was *J*=20, so the minimum *J* was closest to the true number of informative markers among the *J* considered in this simulation. In additional simulations that used a finer mesh of *J*, *J* was minimized at 20. Similar results were obtained when the classes were less balanced (e.g. Case I with class probabilities respectively 0.15, 0.30, 0.2, 0.25, and 0.1).

Normal Tissue

We applied the recursively-partitioned mixture model algorithm to the normal tissue described below in the Methods section. For this analysis, we attempted to split a node only if the weight assigned to the node was greater than 5. The augBIC_J statistic, which we propose as a comparative measure of model fit for different numbers J of loci, was minimized at $J = 1413$, and the algorithm found 23 classes, whose profiles [mean values calculated using (1)] are depicted in Figure 4. All posterior class membership probabilities were indistinguishable from 0 or 1 within numerical error. Table 2 displays the cross-classification between mixture model latent class and tissue sample type. Blood samples were completely separated from other solid tissue samples. In addition, adult blood samples were completely separated from newborn blood samples obtained from Guthrie cards. Placenta samples were also separated from other tissues aside from a single pleura sample. For the most part, head and neck tissue and brain were separated from other samples, but were poorly distinguished between them. These results were consistent with a Random Forests analysis [24], in which we found blood perfectly classified, low classification error for placenta, and some confusion among head and neck tissue and brain tissue.

Using a permutation test with chi-square statistic, the P value for a hypothesis of no association between class and sample type was less than 0.0001. Thus, our proposed method found clusters relevant to sample type. In addition, a permutation test using a Kruskal-Wallis test statistic produced P<0.0001 for a hypothesis of no difference in mean age among the classes. Interestingly, when the clustering and subsequent hypothesis test was restricted to blood, the P<0.0001 for a hypothesis of no difference in mean age among latent classes. Among the two classes found among the adult liquid blood samples, age was significantly different between them (P<0.0039). These results are consistent with known associations between age and methylation. In the latter two analyses restricted to blood, we found no association between class and gender (P>0.45).

We remark that although there is a temptation to interpret the final classes as being more alike if they were split later in the recursive partitioning process, the fact that some brain samples were classified early with blood but separated later in the process tends to cast suspicion on such interpretations. In particular, there is a band of loci that show high methylation in the three classes at the top of Figure 3, indexed as 11100, 11101, and 1111, corresponding to brain and head and neck tissues, but not in the five class immediately below (starting with 1101), all corresponding to blood samples. This band also occurs at the bottom of Figure 3, in classes such as 01010, which also represent brain and head and neck tissue. Together with the simulations, this result suggests that the final classes are meaningful, but the intermediate node classes are not necessarily so.

We also analyzed the normal tissue methylation data using HOPACH. The greedy version of the algorithm produced only 4 classes. The "best" version produced 9 clusters, which are cross-classified with tissue type in Table 3 and with the latent classes obtained from our proposed method in Table 4. As Table 4 shows, the classes found by our proposed method were, for the most part, subsets of the 9 classes found using HOPACH, with a few exceptions that involve minor disagreements in classification. While the apparent compactness of the HOPACH classification seems, at first glance, more attractive than the classification produced by our model-based method, we remark that is has a few subtle problems. It has three singletons, clusters 6, 8, and 9, which could be grouped together with cluster 7 to comprise a class that entirely represents placental tissue. While a similar criticism could be made of our proposed method with respect to classification of blood, some of the classes have verifiable meaning; for example, the classes indexed 1100 and 1101 distinguish age among blood samples taken from adults. HOPACH classification also associates one head and neck tissue sample with blood, and two cervical samples with numerous other tissues in the 5th class. The 5th class associates numerous tissue types, and comprises 6 different classes produced by our proposed method. Table 4 shows the cross-classification of tissue type with these 6 classes for the subjects comprising the 5th HOPACH class. The classification correctly isolates two of three cervix samples, and has a tendency to distinguish pleura from lung samples. Using a permutation test with chi-square statistic, the P value for a hypothesis of no association between class and sample type in this subset was less than 0.0001, demonstrating that the classes produced by the mixture model have additional information with respect to tissue type. In order to compare the predictive ability of the two classification schemes overall, we applied the Random Forest algorithm to indicator variables representing HOPACH clusters (using all 9 variables for every bootstrap) and to indicator variables representing our model-based classification (using all 23 variables for every bootstrap). In the former case we obtained a misclassification error rate of 17.97%, and in the latter case a misclassification error rate of 18.43%, the difference being the misclassification of one less sample using the model-based method. Employing the Random Forest algorithm in a similar manner to predict age, we obtained a mean-squared-residual of 190.1 for HOPACH and 164.8 for the model-based classification, with variance explained equal to 80.6% and 83.2% respectively. Thus, the model-based classification seems to offer modest improvements over HOPACH in ability to make biological distinctions.

DISCUSSION

Our proposed method is a model-based version of the HOPACH algorithm [16]: we recursively use a beta-mixture model [15] to propose a split of an existing cluster, preserving the split only when it is judged on the basis of BIC to better fit the data. While more refined versions of BIC are available in this context [15], we have found that the simpler versions perform adequately. We remark that the candidate clusters at each node are defined in a "fuzzy" manner, where each subject has the opportunity to be portioned out over multiple clusters. This is a distinction between our method and nonparametric methods such as HOPACH. Siegmund et al. (2004) argue that model-based clustering is preferred in this context over hierarchical clustering [13], a finding that bears out in our simulations. One reason for the superior performance, at least in a high-dimensional context, is that the metric used to characterize the differences in nonparametric contexts may be relatively insensitive to differences in particular dimensions. This may play a role in the apparent differences in classification of normal tissue between our proposed method and HOPACH.

K-means have been used recently to cluster methylation outcomes (e.g. [12]). However, the work of van der Laan and Pollard (2003) seems to suggest that HOPACH may yield results that are superior to K-means. In particular, with K-means it is difficult to know how many classes are inherent in the data without resampling-based methods such as the gap statistic [25], with implications for scalability. Also, the "curse of dimensionality" would tend to degrade the performance of procedures such as K-means when there are a large number of clusters and the observed data is of high dimension. In general, nonparametric methods such as the *fanny* algorithm [17] rely on tuning parameters that are difficult to optimize without resampling. An additional problem with non-parametric procedures is that they typically consider only the first moment (means) of the underlying distributions, ignoring the second-moment (variance) which for DNA methylation as measured by the GoldenGate assay, may play a critical role in distinguishing tissues.

We propose a dimension-reduction strategy which simply ranks candidate dimensions on the basis of some criterion such as variance, fits the top *J* dimensions in a mixture model, and employs an augmented version of BIC to compare model fit across different values of *J*. This is a departure from the penalized-likelihood methods of the kind described in [22], which would become computationally difficult for truly high-dimensional data. Our approach is similar in sprit to supervised principal components methods such as [26]. Interestingly, for the normal tissue data, all 1413 loci were found to be informative. The implication is that methylation at even the least variable locus, COL6A1_P283_F, contains information about tissue type. In fact, in box-plots showing the distribution of COL6A1_P283_F methylation (not planned for published article – see Supplementary Figure 1 provided for review), there was great heterogeneity in apparent distribution by tissue type, even though all methylation average beta values were less than 0.05. This strongly suggests that the average beta measured by the GoldenGate assay is in fact an average of methylation status over different cell types.

CONCLUSIONS

In summary, our method appears to have good properties both respect to classification error and computation time. It achieves these properties by combining the strengths of model-based and hierarchical methods, navigating the underlying clusters quickly through recursive partitioning, but doing so in a way that makes use of a reasonable probability model. This model is also used to compare different dimensions *J* of input, thus refining the discriminative ability in a scalable manner. Software is available from the authors upon request.

METHODS

Normal Tissue Data

Our proposed method is motivated by methylation array data obtained for normal tissue data. We extracted DNA from 217 normal tissue samples, modified with bisulfite, and processed them on the Illumina GoldenGate methylation platform. Tissue were assembled by a collaborative, multi-institutional network of principal investigators conducting molecular epidemiologic studies of human cancer. Participating institutions include the International Mesothelioma Program at Brigham and Women's Hospital, Brown University, Dartmouth-Hitchcock Medical Center, University of California – San Francisco, Brain Tumor SPORE program, University of Massachusetts – Lowell, and the University of Minnesota. Tissues were obtained through Institutional Review Board approved studies already underway at these institutions, or purchased from the National Disease Research Interchange (NDRI). A variety of normal tissue types were assembled: bladder (*n*=5), blood (*n*=85), brain (*n*=12), cervix (*n*=3), head and neck (*n*=11), kidney (*n*=6), lung (*n*=53), placenta (*n*=19), pleura (*n*=18), and small intestine (*n*=5). All tissue samples were from adults except *n*=55 samples of Guthrie card derived blood samples from newborns. Figure 1 illustrates the methylation pattern for all 217 subjects and 1413 loci passing quality-assurance procedures (median detection p-value < 0.05).

<u>Recursive-partitioning for a Beta Mixture Model</u>

Let $\mathbf{Y}_i = (Y_{i1},...Y_{iJ})$ be a vector of *J* continuous outcomes falling between 0 and 1, and let there be *n* such vectors. We posit a mixture model having *K* classes, such that subject *i* belongs to class $C_i \in \{1,...,K\}$, and conditional on class membership, each outcome is an independent Beta-distributed variable with parameters α_{kj} and β_{kj} depending on both class *k* and dimension *j*. That is,

$$f(Y_{ij}=y \mid C_i=k)=\mathrm{B}(\alpha_{kj},\beta_{kj})^{-1}\,y^{\alpha_{kj}-1}(1-y)^{\beta_{kj}-1},$$

which implies the following identities:

$$E(Y_{ij} \mid C_i=k)=\mu_{jk}=\alpha_{kj}(\alpha_{kj}+\beta_{kj})^{-1} \qquad (1)$$

$$\mathrm{var}(Y_{ij} \mid C_i=k)=\mu_{jk}(1-\mu_{jk})(\alpha_{kj}+\beta_{kj}+1)^{-1}$$

Under the assumption that $C_i=k$ with probability η_k, $\sum_{k=1}^{K}\eta_k=1$, and that methylation at each locus is independent conditional on class membership, the likelihood contribution from subject *i* is

$$f(\mathbf{Y}_i=\mathbf{y}_i)=\sum_{k=1}^{K}\eta_k\prod_{j=1}^{J}\mathrm{B}(\alpha_{kj},\beta_{kj})^{-1}y_{ij}^{\alpha_{kj}-1}(1-y_{ij})^{\beta_{kj}-1}.$$

With observed data $D=\{\mathbf{y}_1,\ldots,\mathbf{y}_n\}$, we then maximize the full-data log-likelihood,

$$\ell(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta})=\sum_{i=1}^{n}\log\{f(\mathbf{Y}_i=\mathbf{y}_i)\}, \qquad (2)$$

with respect to the set of all parameters $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta})$ to be estimated: $\boldsymbol{\alpha}=(\alpha_{11},\ldots,\alpha_{1J},\alpha_{21},\ldots,\alpha_{KJ})$, $\boldsymbol{\beta}=(\beta_{11},\ldots,\beta_{1J},\beta_{21},\ldots,\beta_{KJ})$, and $\boldsymbol{\eta}=(\eta_1,\ldots,\eta_{K-1})$. This is easily achieved using an Expectation-Maximization (EM) algorithm [20]. Briefly, we initialize the procedure with an $n \times K$ matrix of weights $\mathbf{W} = (w_{ik})$ whose rows sum to one. The rows represent initial guesses at class membership probabilities for each subject. For each *k*, we set $\eta_k = n^{-1} \sum_{i=1}^{n} w_{ik}$ and maximize the quantity

$$\begin{aligned} \ell_k^{(w)} &= \sum_{i=1}^{n} w_{ik} \log\{ f(\mathbf{Y}_i = \mathbf{y}_i \mid C_i = k) \} \\ &= Q_k + \sum_{i=1}^{n} w_{ik} \sum_{j=1}^{J} \left[\alpha_{kj} \log(y_{ij}) + \beta_{kj} \log(1 - y_{ij}) - \log\{ \mathrm{B}(\alpha_{kj}, \beta_{kj}) \} \right], \end{aligned} \quad (3)$$

where Q_k is constant with respect to parameters, to obtain the **α** and **β** parameters corresponding to class *k*. We subsequently update the weights as follows:

$$w_{ik} = \frac{\eta_k \prod_{j=1}^{J} \mathrm{B}(\alpha_{kj}, \beta_{kj})^{-1} y_{ij}^{\alpha_{kj}-1} (1 - y_{ij})^{\beta_{kj}-1}}{\sum_{k=1}^{K} \eta_k \prod_{j=1}^{J} \mathrm{B}(\alpha_{kj}, \beta_{kj})^{-1} y_{ij}^{\alpha_{kj}-1} (1 - y_{ij})^{\beta_{kj}-1}}$$

iterating until $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta})$ does not change. The final weight w_{ik} represents the posterior probability that subject *i* belongs to class *k*, i.e. $w_{ik} = P(C_i = k \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. As for most finite-mixture methods, we might decide on the number of classes *K* by fitting mixture models for a range of possible values of *K*, computing the BIC statistic

$$\mathrm{BIC} = \log(n)(2JK + K - 1) - 2 \sum_{i=1}^{n} \log\{ f(\mathbf{Y}_i = \mathbf{y}) \} \quad (4)$$

and selecting the value of *K* corresponding to the minimum BIC. In the context of beta mixture models, slightly modified alternatives to BIC are available [15]. The entire operation has approximate complexity $nJK^2_{\max}$, where $K_{\max}$ is the maximum number of classes attempted. The square term arises under the assumption that for a single model with *K* classes, the complexity will be of order nJK .

Because likelihoods for model-based clustering algorithms can be multi-modal [14, 21], commercial mixture model software packages often use multiple starting values for fitting the model, and subsequently choose the estimates corresponding to the maximum likelihood. However, careful choice of starting values can often minimize the effort [21, 22]. One option is to use hierarchical clustering to find *K* clusters (cutting the clustering dendrogram at the appropriate height), and constructing a weight matrix **W** corresponding to these clusters. Another, similar, option is to use a fuzzy clustering algorithm such as the *fanny* algorithm [17] available in the R package *cluster*.

We now propose a recursive method that, on average, has complexity nK , where *K* is the true number of classes. Consider the following weighted-likelihood version of (1)

$$\ell^{(\omega)}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\eta};\boldsymbol{\omega}) = \sum\nolimits_{i=1}^{n} \omega_i \log\{f(\mathbf{Y}_i = \mathbf{y}_i)\}\,. \qquad (5)$$

When $\omega_i \equiv 1$ for all *i*, (2) and (5) are equivalent. When $0 < \omega_i < 1$, subject *i* represents a partial or pseudo-subject, and when $\omega_i = 0$, subject *i* is excluded entirely from consideration in the model. The EM algorithm described above is easily modified by multiplying each w_{ik} by ω_i in (3), where the interpretation is that the classes under consideration are only a partial set, and that subject *i* belongs to one of these classes only with probability ω_i. If we begin by fitting a 2-class model to the entire data set, the result is two sets of posterior weights representing the posterior probabilities of membership in each of the two classes. Under the assumption that each of these classes can be further split, and that each subject belongs to the subsequent splits only with probability equal to the weight assigned to the un-split class, we apply the weighted-likelihood EM algorithm to obtain the two classes corresponding to the new split.

To make this idea more precise, define a concatenation operation τ on a sequence of binary values $r = (q_1, ..., q_T)$, as $\tau(r, q) = (q_1, ..., q_T, q)$. This provides a natural notation for recursive binary partitioning, where longer sequences represent deeper levels of recursion. The first two-class model, initialized by nonparametric cluster analysis, results in two sets of weights, $\omega_i^{(0)} = w_{i1}$ and $\omega_i^{(1)} = w_{i2}$. For any sequence *r*, a mixture model can be attempted using the weighted EM algorithm with weights $\omega_i^{(r)}$. If the EM algorithm fails, then we terminate the recursion at that point, but if the EM algorithm succeeds, we can set new weights $\omega_i^{(\tau(r,0))} = \omega_i^{(r)} w_{i1}^{(r)}$, $\omega_i^{(\tau(r,1))} = \omega_i^{(r)} w_{i2}^{(r)}$, and continue the recursion. Note that at each level of recursion, the weights become smaller; since a mixture model becomes unstable with small weights (corresponding to small numbers of pseudo-subjects), the recursion ultimately terminates completely at a set of leaf nodes corresponding to un-split classes. We can stabilize this process by terminating the recursion if the sum of the weights is less than some pre-specified value (e.g. 5). We can also terminate early if the split leads to a less parsimonious representation of the data. To this end, we propose the following weighted versions of BIC:

$$\text{wtdBIC}_2(r) = (4J+1)\log\left(\sum\nolimits_{i=1}^{n} \omega_i^r\right) - 2\ell^{(\omega)}(\boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}^{(r)}, \boldsymbol{\eta}^{(r)}; \boldsymbol{\omega}^{(r)})$$

$$\text{wtdBIC}_1(r) = 2J\log\left(\sum\nolimits_{i=1}^{n} \omega_i^r\right) - 2\ell^{(\omega)}(\boldsymbol{\alpha}^{*(r)}, \boldsymbol{\beta}^{*(r)}, \boldsymbol{\eta}^{*(r)}; \boldsymbol{\omega}^{*(r)}),$$

where the first set of parameters, defining wtdBIC$_2$, are obtained from the two-class mixture model and the second set of parameters, defining wtdBIC$_1$, are obtained from a one-class model. If wtdBIC$_2$(r) is greater than wtdBIC$_1$(r), we terminate the recursion at node r. The worst-case complexity of this algorithm is $n\log(n)J$. However, at deeper levels of recursion, two-class models will tend to fit poorly relative to single-class models, and most nodes will terminate before descending to the deepest levels. We demonstrate empirically below that the proposed method tends to terminate with the number of leaf classes equal to the true number of classes, so that the average complexity is typically of approximate order $nJK\log(K)$. Furthermore, in the deeper classes, subjects whose weights are negligible can be dropped from the weighted EM algorithm, so that the complexity of the node-level fit at deeper levels is less than n.

Dimension reduction

Non-informative loci may lead to excessive noise in the solution. Regularization methods may be used to constrain the degrees-of-freedom, leading to more precise solutions [22, 23]. However, in extremely high dimensions, it can also lead to increased computation time and curtail scalability. We propose an alternative, where all L starting loci are ordered with respect to variance, and the J most variable loci are selected for inclusion in the recursive algorithm described above. A final BIC value can be obtained using (4) by considering all leaf-level un-split classes as distinct clusters, with class prevalence parameter vector $\boldsymbol{\eta}$ obtained by summing the final weights $\omega_i^{(r)}$ and dividing by n. However, this BIC is not comparable across different values of J. Note that the exclusion of $L-J$ loci is equivalent to the assumption that all K classes have identical distributions for the excluded loci. Thus, beta distributions can be fit to each excluded locus using maximum-likelihood, and the resulting parameter estimates included in a final BIC statistic. Specifically, the likelihood for the full data $\mathbf{Y}_i^* = (Y_{i1},...Y_{iJ}, Y_{i(J+1)},...,Y_{iL}) = (\mathbf{Y}_i, \tilde{\mathbf{Y}}_i)$, where we assume the dimensions have been ordered by descending variance and $\tilde{\mathbf{Y}}_i$ represents data excluded from the mixture model analysis, can be expressed as

$$\begin{aligned} f^*\left(\mathbf{Y}_i^* = \mathbf{y}_i^*\right) = &\left\{\sum\nolimits_{k=1}^{K} \eta_k \prod\nolimits_{j=1}^{J} \mathrm{B}(\alpha_{kj},\beta_{kj})^{-1} y_{ij}^{\alpha_{kj}-1}(1-y_{ij})^{\beta_{kj}-1}\right\} \\ &\times \prod\nolimits_{l=J+1}^{L} \mathrm{B}(\alpha_l^*,\beta_l^*)^{-1} y_{il}^{\alpha_l^*-1}(1-y_{il})^{\beta_l^*-1} \\ = &f\left(\mathbf{Y}_i = \mathbf{y}_i\right)\tilde{f}\left(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i\right). \end{aligned}$$

The "augmented" BIC is now

$$\begin{aligned} \mathrm{augBIC}_J &= (2JK + K - 1 + 2L - 2J)\log(n) - 2\sum\nolimits_{i=1}^{n} \log f(\mathbf{Y}_i^* = \mathbf{y}_i^*) \\ &= \mathrm{BIC}_J + 2(L-J)\log(n) - 2\sum\nolimits_{i=1}^{n} \log \tilde{f}(\tilde{\mathbf{Y}}_i = \tilde{\mathbf{y}}_i), \end{aligned}$$

where BIC$_J$ is the BIC computed for just the J selected loci. The augmented BIC is now comparable across different values of J. As we demonstrate below, augBIC$_J$ leads to sensible dimension reduction.

Simulation

We conducted simulations to compare the properties of our proposed method with similar competing methods. For our first case (Case I), each simulated data set consisted of n=100 subjects, each having 1413 continuous responses lying in the unit interval. Each subject was a member of one of 5 classes, each class occurring with 0.2 probability. The classes were defined by beta-distribution parameters for each of L=1413 methylation loci that were autosomal and passed quality-assurance, obtained by fitting a beta model on each locus to one of five data sets from our normal data: adult blood, newborn blood, placenta, lung/pleura, and everything else. Figure 2A illustrates a typical data set generated from these parameters. For each data set, we conduct 5 analyses, each using the J most variable loci, $J \in \{25,50,500,1000\}$. The first analysis used hierarchical clustering, implemented using *hclust* in the R *cluster* package, with Euclidean metric and average linkage, and assigned 5 classes by cutting the resulting dendrogram at the appropriate height using the *cutree* function in the same package. The second analysis used HOPACH (R *hopach* package) to select the "best" classes as defined in the function settings. The third analysis used HOPACH with classes obtained by the "greedy" version of the algorithm. The fourth analysis fit 6 sequential mixture models ($1 \le K \le 6$), each initialized two different ways (hierarchical clustering and the *fanny* algorithm), selecting the value of K producing the lowest BIC. The fifth analysis was an application of our proposed method. In the latter two types of analysis, we recorded the value of J that produced the best augmented BIC.

For our second case (Case II), which represented a lower-dimensional setting (L=200) with greater variation in variance of individual beta distributions, we considered

100 subjects from 4 classes, described as follows. Five sets of 10 “informative” beta parameters were drawn randomly at the beginning of the simulation study: $a_{1j} \sim Gamma(10{,}10)$, $b_{1j} \sim Gamma(10{,}10)$; $a_{2j} \sim Gamma(400{,}10)$, $b_{2j} \sim Gamma(100{,}10)$; $a_{3j} \sim Gamma(100{,}10)$, $b_{3j} \sim Gamma(400{,}10)$; and $a_{4j} \sim Gamma(100{,}1)$, $b_{4j} \sim Gamma(250{,}1)$. These were used to construct four classes of 20 informative dimensions: $\boldsymbol{\alpha}_1 = (\mathbf{a}_2, \mathbf{a}_1)$, $\boldsymbol{\alpha}_2 = (\mathbf{a}_2, \mathbf{a}_4)$, $\boldsymbol{\alpha}_1 = (\mathbf{a}_3, \mathbf{a}_1)$, $\boldsymbol{\alpha}_1 = (\mathbf{a}_3, \mathbf{a}_4)$, where $\mathbf{a}_l = (a_{lj})$, and similarly for the $\boldsymbol{\beta}_k$ parameters with $\mathbf{b}_l = (b_{lj})$. Each such 20-dimensional parameter was augmented with a set of 180 “noninformative” parameters, constructed as 60 copies of the vector (100,1,50) for $\boldsymbol{\alpha}_k$ and 60 copies of the vector (1,100,50) for $\boldsymbol{\beta}_k$. The class probabilities were respectively 0.2, 0.3, 0.2, and 0.3. Although the pattern corresponding to this collection of parameters may be difficult to visualize at first glance, Figure 2B shows a typical data set generated under these conditions, and reveals a small set of informative markers, some having distinctions in mean and others in variability. Similar analyses were conducted for this simulation, except with $J \in \{5,10,25,50\}$, and 4 classes assumed for hierarchical clustering.

Misclassification error was assessed for all simulated data sets and analyses. Each estimated class was matched to true class by minimizing the distance between the *J* means calculated according to (1). When the number of estimated classes was greater than the true number, multiple estimated classes were assigned to a single matching true class, thus generating no misclassification error when the estimated class merely partitioned the true class. When the number of estimated classes was fewer than the true number, subjects within true classes that failed to match to an estimated class were considered misclassified. In the latter case, coarsening of the true classes would lead to the smaller absorbed class being judged as misclassified. In the Results section below, we show that HOPACH tends to overestimate the number of classes for the cases we considered, so our strategy, which favors inappropriate partitioning over inappropriate coarsening, is conservative with respect to comparison with HOPACH in this set of simulations.

AUTHORS' CONTRIBUTIONS

EAH developed statistical methodology, with critical intellectual input from RFY. BCC prepared laboratory samples and coordinated analysis with Illumina; BCC and CJM participated in testing software on methylation data obtained from real tissues. CJM, MRK, MW, HHN, JW, SZ, JKW, and KTK contributed tissue samples. All eleven authors participated in weekly conference calls and discussed the biological implications of the analytical methods.

ACKNOWLEDGEMENTS

Supported by: Mesothelioma Applied Research Foundation grant, NIH/NIEHS Training grant in environmental health sciences T32ES007155, NIH/NIEHS grant P42ES05947, and NCI grants CA48061, CA126939, CA078609, CA100679, CA121147, and FAMRI.

REFERENCES

1. Russo V, Martienssen RA, Riggs AD: **Epigenetic mechanisms of gene regulation**: Cold Spring Harbor Laboratory Press; 1996.
2. Knudson AG: **Chasing the cancer demon**. *Annu Rev Genet* 2000, **34**:1-19.
3. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer**. *Nat Rev Genet* 2002, **3**:415-428.
4. Sakamoto H, Suzuki M, Abe T, Hosoyama T, Himeno E, Tanaka S, Greally JM, Hattori N, Yagi S, Shiota K: **Cell type-specific methylation profiles occurring disproportionately in CpG-less regions that delineate developmental similarity**. *Genes Cells* 2007, **12**:1123-1132.
5. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA *et al*: **DNA methylation profiling of human chromosomes 6, 20 and 22**. *Nat Genet* 2006, **38**:1378-1385.
6. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J *et al*: **Epigenetic differences arise during the lifetime of monozygotic twins**. *Proc Natl Acad Sci U S A* 2005, **102**:10604-10609.

7. Frigola J, Song J, Stirzaker C, Hinshelwood RA, Peinado MA, Clark SJ: **Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band**. *Nat Genet* 2006, **38**:540-549.
8. Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, Andrews TD, Howe KL, Otto T, Olek A *et al*: **DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project**. *PLoS Biol* 2004, **2**:e405.
9. Schilling E, Rehli M: **Global, comparative analysis of tissue-specific promoter CpG methylation. Genomics**. 2007, **90**:314-323.
10. Shann YJ, Cheng C, Chiao CH, Chen DT, Li PH, Hsu MT: **Genome-Wide Mapping and Characterization of Hypomethylated Sites in Human Tissues and Breast Cancer Cell Lines**. *Genome Res* 2008.
11. Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, Held WA: **Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression**. *Proc Natl Acad Sci U S A* 2005, **102**:3336-3341.
12. Shen L, Kondo Y, Guo Y, Zhang J, Zhang L, Ahmed S, Shu J, Chen X, Waterland RA, Issa J-PJ: **Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters**. *PLOS Genetics* 2007, **3**:e181.
13. Siegmund KD, Laird PW, Laird-Offringa IA: **A comparison of cluster analysis methods using DNA methylation data**. *Bioinformatics* 2004, **20**:1896-1904.
14. Stephens M: **Dealing with label switching in mixture models**. *Journal of the Royal Statistical Society Series B* 2000, **62**:795-809.
15. Ji Y, Wu C, Liu P, Wang J, Coombes KR: **Applications of beta-mixture models in bioinformatics**. *Bioinformatics* 2005, **21**:2118-2122.
16. van der Laan MJ, Pollard KS: **A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap**. *Journal of Statistical Planning and Inference* 2003, **117**:275-303.
17. Kaufman L, Rousseeuw PJ: **Finding Groups in Data: An Introduction to Cluster Analysis**. New York: Wiley; 1990.
18. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. New York: Springer; 2001.
19. Breiman L, Friedman JH, Olshen RA, Stone CJ: **Classification and Regression Trees**. Boca Raton, Florida: Chapman & Hall; 1984.
20. Dempster A, Laird N, Rubin D: **Maximum likelihood from incomplete data via the EM algorithm (with discussion)**. *J R Statist Soc B* 1977, **39**:1-38.
21. Leroux BG, Puterman ML: **Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models**. *Biometrics* 1992, **48**:545-558.
22. Houseman EA, Coull BA, Betensky RA: **Feature-specific penalized latent class analysis for genomicdata**. *Biometrics* 2006, **62**:1062-1070.
23. Fraley C, Raftery AE: **Bayesian regularization for normal mixture estimation and model-based clustering**. In.: Department of Statistics, University of Washington; 2005.
24. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**: 5-32.

25. Tibshirani R, Walther G, Hastie T: **Estimating the number of clusters in a dataset via the gap statistic**. *J Royal Statist Soc B* 2001, **63**:411-423.

26. Bair E, Tibshirani R: **Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data**. *PLoS Biol* 2004, **2**:1544-9173.

TABLES

Table 1. Classification error and computation time for various clustering methods applied to simulated data.

Classification Error		*J*	HC	HOPACH(1)	HOPACH(2)	MM(1-6)	RPMM
	Case 1	25	33.2	12.5	18.5	12.6	4.6
		50	32.5	7.4	13.6	7.1	0.4
		500	33.9	10.4	14.1	1.9	0.1
		1000	34.0	15.6	16.9	1.7	0.0
		J	HC	HOPACH(1)	HOPACH(2)	MM(1-6)	RPMM
	Case 2	5	60.6	65.7	66.7	60.6	60.6
		10	59.2	67.2	67.9	60.1	59.6
		25	29.2	5.0	8.4	0.0	0.0
		50	29.1	4.0	7.9	0.2	0.0
Computation Time (s)		*J*	HC	HOPACH(1)	HOPACH(2)	MM(1-6)	RPMM
	Case 1	25	0.01	4.95	1.57	46.94	12.29
		50	0.01	4.26	1.50	59.56	15.41
		500	0.06	4.36	1.48	505.70	118.92
		1000	0.12	3.84	1.60	995.57	223.86
		J	HC	HOPACH(1)	HOPACH(2)	MM(1-6)	RPMM
	Case 2	5	0.00	3.33	1.35	40.83	5.09
		10	0.00	2.76	1.32	63.77	8.44
		25	0.01	3.95	1.44	29.69	9.55
		50	0.01	3.33	1.36	45.38	12.92

HC = Hierarchical clustering
HOPACH(1) = HOPACH with ‘best’ number of classes
HOPACH(2) = HOPACH with ‘greedy’ number of classes
MM(1-6) = Beta mixture model fitting 1-6 classes sequentially
RPMM = Recursively partitioned mixture model
J = Number of loci considered in analysis

Table 2. Cross-classification of sample type with latent classes obtained from proposed method

Class	bladder	blood (ad)	blood (nb)	brain	cervical	H & N	kidney	lung	placenta	pleura	sm intestine	Total
000	3						2	12		8	3	28
0010								19		5		24
0011								20		2	1	23
0100	2			2	1		4	2		2	1	14
01010				1		4						5
0101100						3						3
0101101						3						3
010111					2							2
01100									1	1		2
01101									5			5
0111									13			13
1000			3									3
100100			2									2
100101			4									4
1001100			3									3
1001101			4									4
100111			5									5
101			34									34
1100		18										18
1101		12										12
11100				5								5
11101				3								3
1111				1		1						2
Total	5	30	55	12	3	11	6	53	19	18	5	217

Classes are labeled with the sequence vector representing the terminal node from which the class was derived.

Table 3. Cross-classification of sample type with clusters obtained from HOPACH

Class	bladder	blood (ad)	blood (nb)	brain	cervical	H & N	kidney	lung	placenta	pleura	sm intestine	Total
1		30				1						31
2			55									55
3				10								10
4				2	1	10						13
5	5				2		6	53		18	5	89
6									1			1
7									16			16
8									1			1
9									1			1
Total	5	30	55	12	3	11	6	53	19	18	5	217

Table 4. Cross-classification of latent classes obtained latent classes obtained from proposed method with clusters obtained from HOPACH

Class	1	2	3	4	5	6	7	8	9	Total
000					28					28
0010					24					24
0011					23					23
0100			2		12					14
01010				5						5
0101100				3						3
0101101				3						3
010111				1	1					2
01100					1	1				2
01101							4		1	5
0111							12	1		13
1000		3								3
100100		2								2
100101		4								4
1001100		3								3
1001101		4								4
100111		5								5
101		34								34
1100	18									18
1101	12									12
11100			5							5
11101			3							3
1111	1			1						2
Total	31	55	10	13	89	1	16	1	1	217

Rows represent classes from proposed method, labeled with the sequence vector representing the terminal node from which the class was derived. Columns represent clusters from HOPACH.

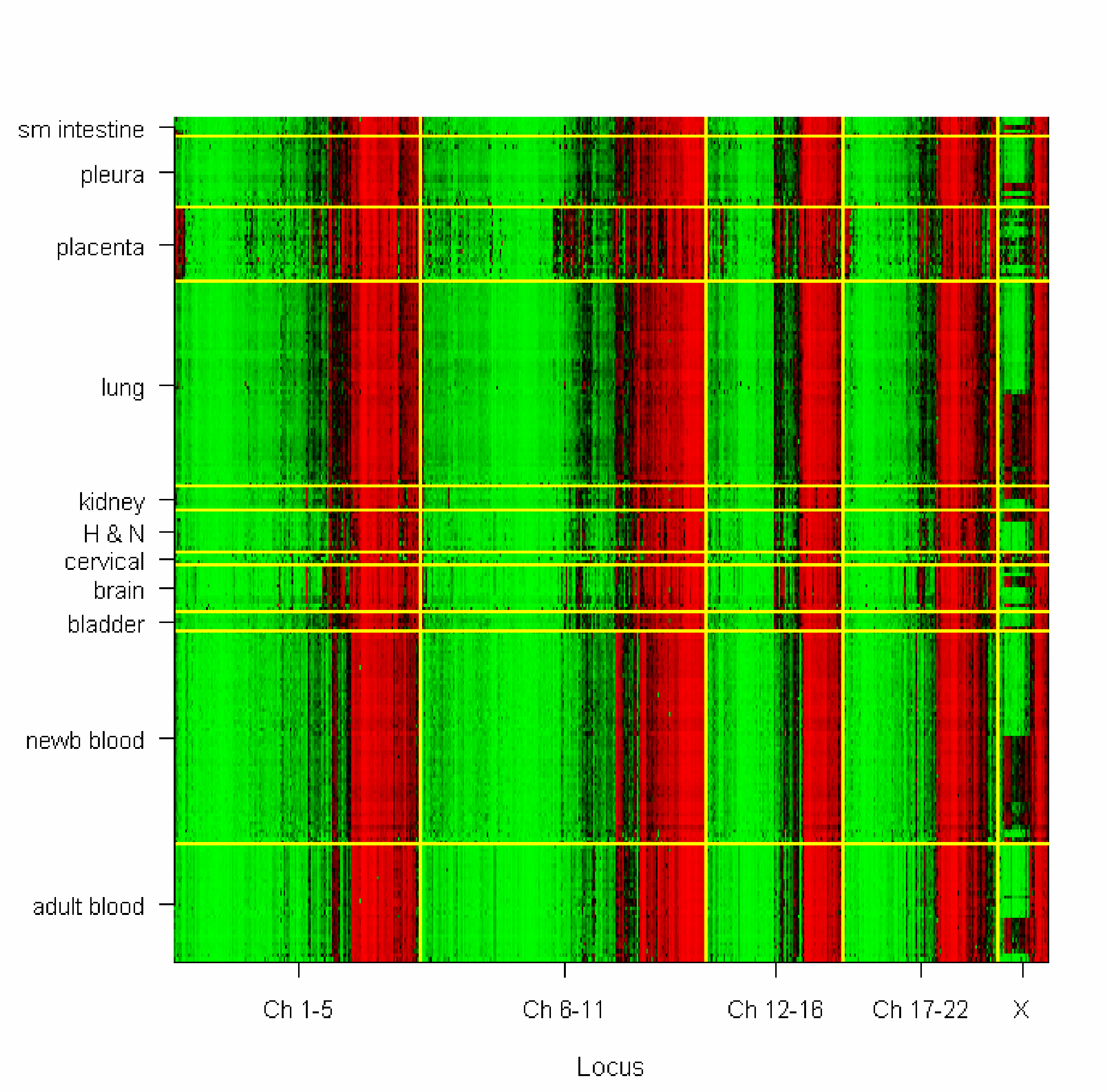
Table 5. Cross-classification of sample type with latent classes obtained from proposed method among subjects within the 5th class obtained by HOPACH

Class	bladder	cervical	kidney	lung	pleura	sm intestine	Total
000	3		2	12	8	3	28
0010				19	5		24
0011				20	2	1	23
0100	2	1	4	2	2	1	12
010111		1					1
01100					1		1
Total	5	2	6	53	18	5	89

Classes are labeled with the sequence vector representing the terminal node from which the class was derived.

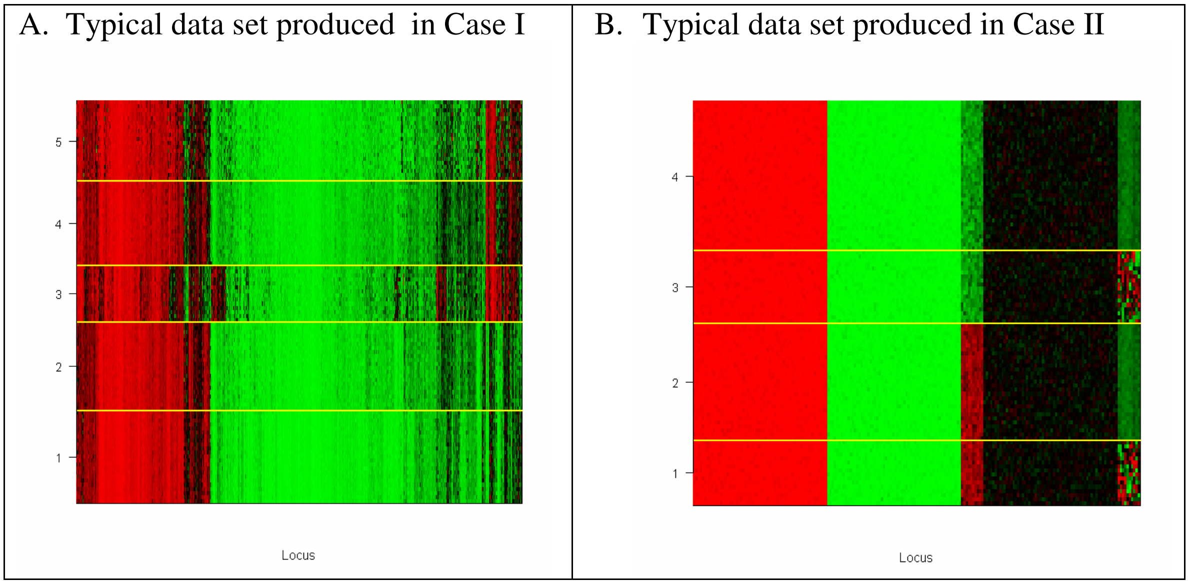
FIGURES

Figure 1. Unadjusted Average Beta values obtained from Illumina GoldenGate methylation platform for 1413 tumor suppressor loci on 217 normal tissue samples.



Red = 1.0, black = 0.5, green = 0.0. Autosomal chromosomes are grouped to aid visualization. For each chromosome group, loci are ordered by their position in a dendrogram produced by hierarchical clustering. Similarly, within tissue sample groups, samples are ordered by their position in a hierarchical clustering dendrogram.

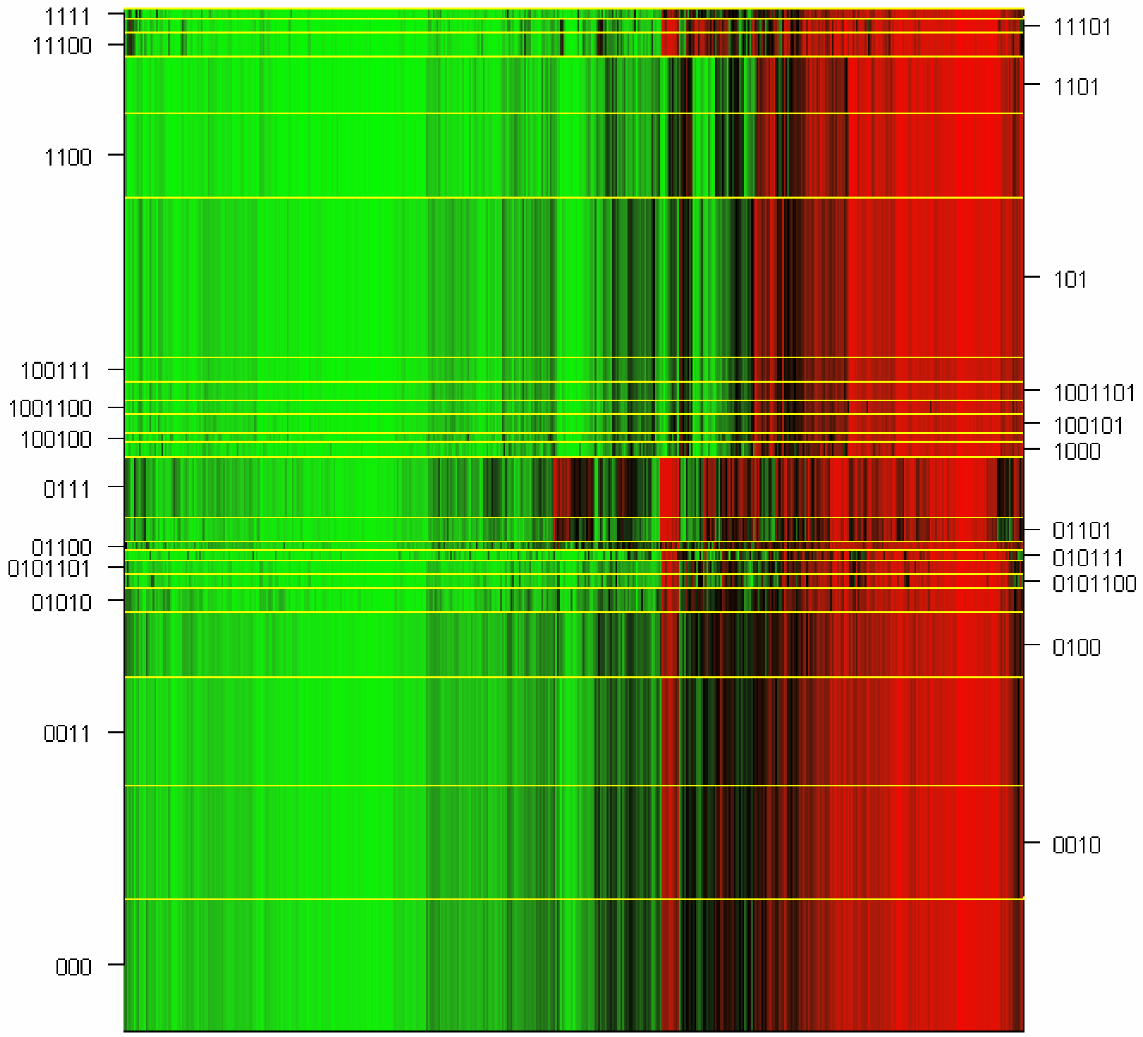
Figure 2. Examples of simulated data


A. Typical data set produced in Case I

B. Typical data set produced in Case II

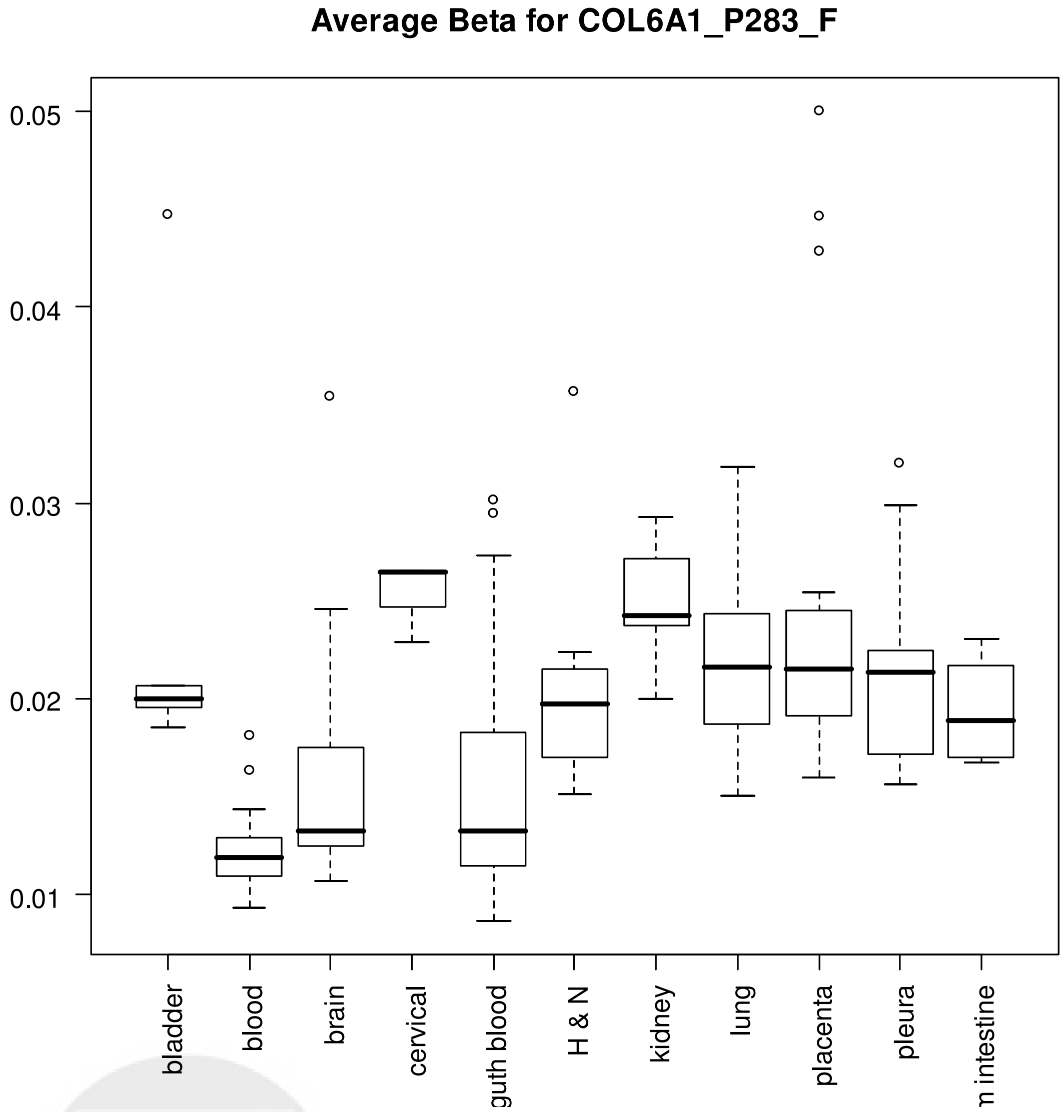
Red=1.0, black = 0.5, green = 0.0. True classes indicated and separated by yellow dividing line. Height of region indicates the relative number of subjects in each class.

Figure 3. Profiles of latent classes among normal tissue samples.



Average value (equation 1) depicted by color: red=1.0, black = 0.5, green = 0.0. Classes are separated by yellow dividing line, with height indicating the relative proportion of subjects within each class. Loci are ordered by their position in a dendrogram obtained via hierarchical clustering.

Figure 4. Distribution of DNA methylation average beta values by tissue type at least variable locus.


Average Beta for COL6A1_P283_F
bladder
blood
brain
cervical
guth blood
H & N
kidney
lung
placenta
pleura
n intestine