Submitted to the Annals of Applied Statistics

INTEGRATIVE MODEL-BASED CLUSTERING OF MICROARRAY METHYLATION AND EXPRESSION DATA [1]

BY MATTHIAS KORMAKSSON *, JAMES G. BOOTH*, MARIA E. FIGUEROA†, AND ARI MELNICK†

Cornell University and Weill Cornell Medical College†*

In many fields, researchers are interested in large and complex biological processes. Two important examples are gene expression and DNA methylation in genetics. One key problem is to identify aberrant patterns of these processes and discover biologically distinct groups. In this article we develop a model-based method for clustering such data. The basis of our method involves the construction of a likelihood for any given partition of the subjects. We introduce cluster specific latent indicators that, along with some standard assumptions, impose a specific mixture distribution on each cluster. Estimation is carried out using the EM algorithm. The methods extend naturally to multiple data types of a similar nature, which leads to an integrated analysis over multiple data platforms, resulting in higher discriminating power.

1. INTRODUCTION. Epigenetics refers to the study of heritable characteristics not explained by changes in the DNA sequence. The most studied epigenetic alteration is cytosine (one of the four bases of DNA) methylation, which involves the addition of a methyl group (a hydrocarbon group occurring in many organic compounds) to the cytosine. Cytosine methylation plays a fundamental role in epigenetically controlling gene expression, and studies have shown that aberrant DNA methylation patterning occurs in inflammatory diseases, aging, and is a hallmark of cancer cells (Rodenhiser and Mann, 2006; and Figueroa *et al.*, 2010). Figueroa *et al.* (2010) performed the first large-scale DNA methylation profiling study in humans, where they hypothesized that DNA methylation is not randomly distributed in cancer but rather is organized into highly coordinated and well-defined patterns, which reflect distinct biological subtypes. Similar observations had already been made for expression data (Golub *et al.*, 1999; Armstrong *et al.*, 2002). Identifying such biological subtypes through abnormal patterns is a very important task as some of these malignancies are highly heterogeneous presenting major challenges for accurate clinical classification, risk stratification and targeted therapy. The discovery of aberrant patterns in subjects can identify tumors or disease subtypes and lead to a better understanding of the underlying biological processes, which in turn

[1]Supported in part by NSF Grant DMS-0805865.

Keywords and phrases: Integrative Model-based Clustering, Microarray Data, Mixture Models, EM Algorithm, Methylation, Expression, AML.

can guide the design of more specifically targeted therapies. Due to the biological interaction between methylation and expression, biologists hope to optimize the amount of biological information about cancer malignancies by borrowing strength across both platforms. As an example Figueroa *et al.* (2008) showed that the integration of gene expression and epigenetic platforms could be used to rescue genes that were biologically relevant but had been missed by the individual analyses of either platform separately.

In this article we propose a model-based approach to clustering such high dimensional microarray data. In particular we build finite mixture models that guide the clustering. These types of models have been shown to be a principled statistical approach to practical issues that can come up in clustering (McLachlan and Bashford, 1988; Banfield and Raftery, 1993; Cheeseman and Stutz, 1995; Fraley and Raftery, 1998, 2002). The motivating application is the cluster analysis of Figueroa *et al.* (2010), which focused on patients with Acute Myeloid Leukemia (AML). Both methylation and expression data are available and we develop a clustering method that can be applied to both data types separately. Furthermore we extend our methodology to facilitate an integrated cluster analysis of both data platforms simultaneously. Although the methods are designed for these particular applications we expect that they can be applied to other types of microarray data, such as ChIP-Chip data.

A lot of attention has been given to classification based on gene expression profiles and more recently based on methylation profiles. Siegmund, Laird and Laird-Offringa (2004) give an overview and comparison of several clustering methods on DNA methylation data. They point out that among biologists, agglomerative hierarchical cluster analysis is popular. However, they argue in favor of model-based clustering methods over nonparametric approaches and propose a Bernoulli-lognormal model for the discovery of novel disease subgroups. This model had previously been applied by Ibrahim, Chen and Gray (2002) to identify differentially expressed genes and profiles that predict known disease classes. More recently Houseman *et al.* (2008) proposed a Recursively Partitioned Mixture Model algorithm (RPMM) for clustering methylation data using beta mixture models (Ji *et al.*, 2005). They proposed a beta mixture on the subjects and the objective was to cluster subjects based on posterior class membership probabilities. The RPMM approach is a model-based version of the HOPACH clustering algorithm developed in van der Laan and Pollard (2003).

In high dimensional data clustering is often performed on a smaller subset of the variables. In fact, as pointed out in Tadesse, Sha and Vannucci (2005), using all variables in high-dimensional clustering analysis has proven to give misleading results. There is some literature on the problem of simultaneous clustering and variable selection (Friedman and Meulman, 2003; Tadesse, Sha and Vannucci, 2005;

Kim, Tadesse and Vannucci, 2006). However, most statistical methods cluster the data only after a suitable subset has been chosen. An example of such practice is McLachlan, Bean and Peel (2002), where the selection of a subset involves choosing a significance threshold for the covariates. That is also essentially what Houseman *et al.* (2008) and Figueroa *et al.* (2010) did, but they selected a subset of the most variable DNA fragments. In this paper we present an integrated model-based hierarchical clustering algorithm that clusters samples based on multiple data types on the most variable features. There is of course a clear advantage of automated variable selection methods such as in Tadesse, Sha and Vannucci (2005). However, the implementation of such methods seem far from straightforward and due to the popularity of hierarchical algorithms among biologists (Kettenring, 2006 concluded that hierarchical clustering was by far the most widely used form of clustering in the scientific literature), there is a clear benefit in having a simple hierarchical algorithm that can handle multiple data types.

The article is organized as follows. In Section 2 we describe the features of the motivating data set. In Section 3 we construct the model as a mixture of Gaussian densities, which leads to a specific mixture likelihood that serves as an objective function for clustering. We also introduce individual specific parameters to account for subject to subject variability within clusters (i.e., the array effect). In Section 4 we present two model-based clustering algorithms. The first algorithm is a hierarchical clustering algorithm that can be used to find a good candidate partition. The second clustering algorithm is an iterative algorithm that is designed to improve upon any initial partition. The likelihood model can be applied to classification of new subjects and in Section 5 we describe a discriminant rule for this purpose. In Section 6 we extend the model to account for multiple data platforms and in Sections 7 and 8 we apply the methods to real data sets, which involve both methylation and expression data. We conclude the article with a discussion in Section 9.

2. MOTIVATING DATA. The Erasmus data were obtained from AML samples collected at Erasmus University Medical Center (Rotterdam) between 1990 and 2008 and involve DNA methylation and expression profiles of 344 patient specimens. For each specimen it was confirmed that $> 90\%$ of the cells were blasts (leukemic cells). Description of the sample processing can be found in Valk *et al.* (2004) and data sets are available in GEO, http://www.ncbi.nlm.nih.gov/geo/, with accession numbers GSE18700 for the methylation data and GSE6891 for the expression data. The gene expression profiles of the AML samples were determined using oligonucleotide microarrays (Affymetrix U133Plus2.0 GeneChips) and were normalized using the rma normalization method of Irizarry *et al.* (2003). The processed data involved 54,675 probe sets and demonstrated a right skewed distri-

bution of the expression profiles for each subject, see Supplementary Figure 4. The methylation profiles of the AML samples were determined using high density oligonucleotide genomic HELP arrays from NimbleGen Systems that cover 25,626 probe sets at gene promoters, as well as at imprinted genes. Briefly, genomic DNA is isolated and digested by the enzymes HpaII and MspI, separately. While HpaII is only able to cut the DNA at its unmethylated recognition motif (genomic sequence 5'-CCGG-3'), MspI cuts the DNA at any HpaII site whether methylated or unmethylated. Following PCR, the HpaII and MspI digestion products are labeled with different fluorophores and then cohybridized on the microarray. This results in two average signal intensities that measure the relative abundances (in a population of cells) of MspI and HpaII at each probe set. The data are preprocessed and normalized using the analytical pipeline of Thompson *et al.* (2008) and the final quantity of interest is $\log(\text{HpaII}/\text{MspI})$. Note that although theoretically HpaII should always be less than MspI, complex technical aspects that arise during the preparation and hybridization of these samples may result in an enrichment of the HpaII signal over that of MspI. Therefore, the log-ratio does not have a one-to-one correspondence with percent methylation at a given probe set but rather provides a relative methylation value that correlates with actual percentage value (see lower left panel of Figure 1).

In what follows, notation will be based on the HELP methylation data. However, if we abstract away from this particular application, the terminology can be adapted to other microarray data such as gene expression. Let y_{ij} denote the continuous response $\log(\text{HpaII}_{ij}/\text{MspI}_{ij})$ for subject $i = 1, \ldots, n$, and probe set $j = 1, \ldots, G$. Lower values of y_{ij} indicate that probe set j has high levels of methylation (in a population of cells) for subject i, whereas higher values indicate low levels of methylation. In the upper panel of Figure 1 we see bimodal histograms of the methylation profiles for two patients in the AML data set along with two component Gaussian mixture fits. In Supplementary Figure 5 we see density profiles for all 344 samples stratified by clusters. There is a large microarray effect in the methylation data, but we observe that all profiles are either skewed or exhibit a bimodal behavior. The lower left panel of Figure 1 shows how the HELP assay correlates with methylation percentages obtained using the more accurate, but much more expensive, quantitative single locus DNA methylation validation MASS Array (see Figueroa *et al.*, 2010). It is clear that the HELP values are forming two clusters of relatively low or high methylation levels with some noise in the percentage range $[20\%, 80\%]$. This apparent dichotomization inspires modeling each individual profile, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iG})'$, with a two component mixture distribution and normality is assumed for each component due to its flexibility and ease of implementation. We know of no biological mechanism that would imply normality, however the assumption gives consistent and reasonable fits of the individual

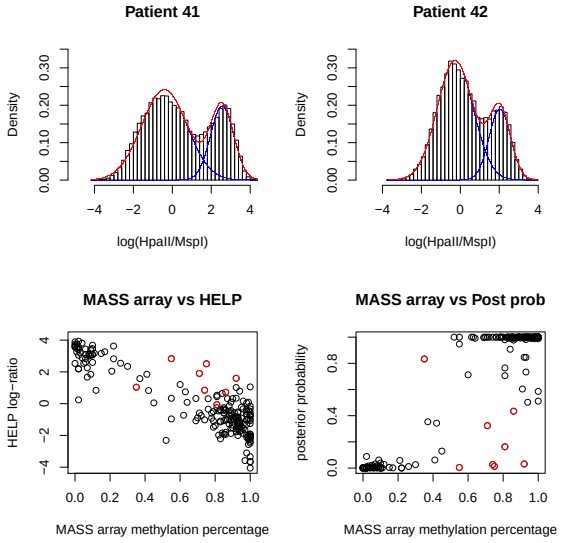

FIG 1. *Upper panel: A histogram of the log signal ratio,* $\log(\textit{HpaII/MspI})$*, for patients number* 41 *and* 42 *in the Erasmus data set, along with two component Gaussian mixture fits. Lower panel: Left graph shows HELP methylation values plotted against more accurate MASS array methylation percentages. Right graph shows that the posterior probabilities from a two component Gaussian mixture classifies probe sets well into low and high methylation.*

methylation profiles (see upper panel of Figure 1).

3. MODEL SPECIFICATION. By dichotomizing the methylation process we can cluster the probe sets into high or low methylation for each patient i by applying a two component Gaussian mixture model. Let $\mathcal{C}$ denote the true partition of the subject set, $[n] = \{1, \ldots, n\}$. We assume that on any given probe set j, all subjects sharing a cluster $c \in \mathcal{C}$ have the same relative methylation status (high or low), and introduce for each cluster c a single latent indicator vector, $\mathbf{w}_c = (w_{c1}, \ldots, w_{cG})'$, with

$$w_{cj} = \begin{cases} 1, & \text{if } j \text{ has high methylation for all subjects in cluster } c, \\ 0, & \text{if } j \text{ has low methylation for all subjects in cluster } c. \end{cases} \tag{1}$$

It is well known that methylation does exhibit biological variability from individual to individual. However, it is biologically reasonable to expect consistency in relative methylation patterns for patients that share the same disease subtype. Define $\boldsymbol{\theta}_i = (\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i})'$ and assume that the observed data, $\mathbf{y} = (\mathbf{y}'_1, \ldots, \mathbf{y}'_n)'$, given the unobserved methylation indicators, $\mathbf{w} = (\mathbf{w}'_1, \ldots, \mathbf{w}'_K)'$ (K being the

number of clusters), arise from the following density:

$$f(\mathbf{y}|\mathbf{w},\boldsymbol{\theta}) = \prod_{c\in\mathcal{C}}\prod_{i\in c} f(\mathbf{y}_i|\mathbf{w}_c,\boldsymbol{\theta}_i), \tag{2}$$

with

$$f(\mathbf{y}_i|\mathbf{w}_c,\boldsymbol{\theta}_i) = \prod_{j=1}^{G} \phi(y_{ij}|\mu_{1i},\sigma^2_{1i})^{w_{cj}}\phi(y_{ij}|\mu_{2i},\sigma^2_{2i})^{1-w_{cj}}, \tag{3}$$

where ϕ denotes the normal density and $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1,\ldots,\boldsymbol{\theta}'_n)'$. We refer to the density in (2) as the classification likelihood of the observed data (Scott and Symons, 1971; Symons, 1981; Banfield and Raftery, 1993) and assume $\mu_{1i} < \mu_{2i}$ for all i. We interpret $\boldsymbol{\theta}_i$ as the individual specific means and variances of the high and low methylation probe sets, respectively. Note that this setup is different from the usual model based clustering setup where we have cluster specific parameters only. However, due to array effects it is reasonable and in fact necessary to require different parameters for different subjects. In the upper panel of Figure 1 we see histograms and fits for two patients that both have chromosomal inversions at chromosome 16, inv(16) (inversions refer to when genetic material from a chromosome breaks apart and then, during the repair process, it is re-inserted back but the genetic sequence is inverted from its original sense). These two patients cluster together under various clustering algorithms, including the model based algorithm presented below. However, the two distributions are clearly not equal.

We put a Bernoulli prior on the latent methylation indicators in (1):

$$f(\mathbf{w}) = \prod_{c\in\mathcal{C}}\prod_{j=1}^{G} \pi_{1c}^{w_{cj}}\pi_{0c}^{1-w_{cj}}, \quad \pi_{0c}+\pi_{1c}=1, \tag{4}$$

where π_{1c} and π_{0c} denote the proportions of probe sets that have high and low methylation, respectively, in cluster c. From (2) and (4) it is clear that the complete data density is

$$f(\mathbf{y},\mathbf{w}) = \prod_{c\in\mathcal{C}}\prod_{j=1}^{G} \Big(\pi_{1c}\prod_{i\in c}\phi(y_{ij}|\mu_{1i},\sigma^2_{1i})\Big)^{w_{cj}} \Big(\pi_{0c}\prod_{i\in c}\phi(y_{ij}|\mu_{2i},\sigma^2_{2i})\Big)^{1-w_{cj}}, \tag{5}$$

and if we integrate out the latent variable $\mathbf{w}$ we arrive at the marginal likelihood

$$L_{\mathcal{C}}(\boldsymbol{\Psi}) = \prod_{c\in\mathcal{C}}\prod_{j=1}^{G} \Big(\pi_{1c}\prod_{i\in c}\phi(y_{ij}|\mu_{1i},\sigma^2_{1i}) + \pi_{0c}\prod_{i\in c}\phi(y_{ij}|\mu_{2i},\sigma^2_{2i})\Big), \tag{6}$$

where $\mathbf{\Psi} = \{(\pi_{1c})_{c\in\mathcal{C}}, (\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i})_i\}$ denotes the set of parameters. This likelihood can be used as an objective function for determining the goodness of different partitions and the maximization of (6) is carried out with the EM algorithm of Dempster, Laird and Rubin (1977). Note that $L_{\mathcal{C}}$ can be written as a product, $\prod_{c\in\mathcal{C}} L_c$, where L_c denotes the likelihood contribution of cluster c. Thus, maximizing $L_{\mathcal{C}}$ can be achieved by maximizing L_c independently for all $c \in \mathcal{C}$. Details of the maximization algorithm are provided in Appendix B of the Supplementary Materials.

REMARK 1. *The premise of the clustering algorithm presented in Section 4 is to cluster subjects together that have similar methylation patterns. Similarities across the genome in the posterior probabilities of high/low methylation guide which subjects are clustered together and thus, if the posterior probability predictions reflect the data well, the clustering algorithm should perform well. In the lower right panel of Figure 1 we see that the posterior probabilities of high methylation fit very well with the actual percentage values.*

REMARK 2. *When we allow for unequal variances $\sigma^2_{1i} \neq \sigma^2_{2i}$, the likelihood in (6) is unbounded and does not have a global maximum. This can be seen by setting one of the means equal to one of the data points, say $\mu_{1i} = y_{ij}$, for some i, j. Then the likelihood approaches infinity as $\sigma^2_{1i} \to 0+$. However, McLachlan and Peel (2000) using the results of Kiefer (1978) point out that, even though the likelihood is unbounded, there still exists a consistent and asymptotically efficient local maximizer in the interior of the parameter space. They recommend running the EM algorithm from several different starting values, dismissing any spurious solution (on its way to infinity), and picking the parameter values that lead to the largest likelihood value.*

REMARK 3. *Note that the likelihood in (6) is identifiable except for the standard and unavoidable label switching problem in finite mixture models (see for example McLachlan and Peel, 2000). Furthermore, there exists a sequence of consistent local maximizers, as $G \to \infty$. This becomes more evident if one recognizes that the expression for a single cluster c can be written as a multivariate normal mixture*

$$L_c = \prod_{j=1}^{G} \Big(\pi_{1c}\phi(\mathbf{y}_{cj}|\boldsymbol{\mu}_{1c}, \mathbf{\Sigma}_{1c}) + \pi_{0c}\phi(\mathbf{y}_{cj}|\boldsymbol{\mu}_{2c}, \mathbf{\Sigma}_{2c})\Big),$$

where $\mathbf{y}_{cj} = (y_{1j}, \ldots, y_{n_cj})'$, $\mu_{kc} = (\mu_{k1}, \ldots, \mu_{kn_c})'$ and $\mathbf{\Sigma}_{kc} = diag(\sigma^2_{ki})^{n_c}_{i=1}$, $k = 1, 2$ (assuming for convenience that $i = 1, \ldots, n_c$ are the members of cluster c). Standard theory thus applies (see McLachlan and Peel, 2000).

REMARK 4. *The correlation structure of high dimensional microarray data is complicated and hard to model. Thus we assume independence across variables in the likelihood (6) even though it may not be the absolutely correct model. However, we can view (6) as a composite likelihood (see Lindsay, 1988) which yields consistent parameter estimates but with a potential loss of efficiency. The correlations observed in microarray data are usually mild and involve only a few and relatively small groups of genes that have moderate or high within-group correlations. In Supplementary Appendix A we perform a simulation study to get a sense of how robust our algorithm is to this independence assumption. The results indicate that with a sparse overall correlation structure in which genes tend to group into many small clusters with moderate to high within-group correlations, our algorithm is not affected by assuming independence across variables. However, there is some indication that with larger groups of genes with very extreme within-group correlations the algorithm will break down. In microarray data such extreme correlation structures are not to be expected on a global scale and therefore we believe that the independence assumption is quite reasonable.*

4. MODEL-BASED CLUSTERING. Our clustering criterion involves finding the partition that gives the highest maximized likelihood $L_{\mathcal{C}}$ as given in (6). This provides us with a model selector, as we can compare the maximized likelihoods of any two candidate partitions. In theory we would like to maximize $L_{\mathcal{C}}$ with respect to all possible partitions of $[n]$ and simply pick the one resulting in the highest likelihood. However, as this is impossible for even moderately large n we propose two clustering algorithms. In subsection 4.1 we propose a simple hierarchical clustering algorithm, and in subsection 4.2 we propose an iterative algorithm that is designed to improve upon any initial partition.

4.1. *Hierarchical clustering algorithm.* In this subsection we describe a simple hierarchical algorithm that attempts to find the partition that maximizes $L_{\mathcal{C}}$ as given in (6). Heard, Holmes and Stephens (2006) used a similar approach, but they constructed a hierarchical Bayesian clustering algorithm that seeks the clustering leading to the maximum marginal posterior probability. The algorithm can be summarized in the following steps:

1. We start with the partition where each subject represents its own cluster, $\mathcal{C}_1 = \{\{1\}, \ldots, \{n\}\}$, and calculate the maximized likelihood, $L_{\mathcal{C}_1}$. Note that this likelihood can be written as a product $L_{\mathcal{C}_1} = \prod_i L_{\{i\}}$ and thus the first step involves maximizing $L_{\{i\}}$ for each $i = 1, \ldots, n$. This is achieved by fitting a two component Gaussian mixture to each of the n individual profiles. As mentioned in Remark 2 each fit can be obtained by using the EM algorithm starting from several different initial values and finding a local

maximum. It is important that the user verifies these initial individual fits before proceeding with the hierarchical algorithm. For example, by going through the 344 methylation profile fits of the Erasmus data, one by one, we observe pleasing fits. The upper panel of Figure 1 gives examples of two such profile fits.

2. Next we merge the two subjects that leads to the highest value of $L_{\mathcal{C}}$ and denote the maximized likelihood value by $L_{\mathcal{C}_2}$. Note that there are $\binom{n}{2}$ many ways of merging two subjects at this step. However, since we already obtained fits for $L_{\{i\}}$, $i = 1, \ldots, n$, at Step 1, we only need to maximize $L_{\{i,i'\}}$, for all pairs (i, i') and find the pair that maximizes

$$\ell_{\mathcal{C}_2} = \ell_{\mathcal{C}_1} - (\ell_{\{i\}} + \ell_{\{i'\}}) + \ell_{\{i,i'\}},$$

where ℓ denotes the loglikelihood. Even though we are applying several EM algorithms, the complexity of each algorithm is low since it only involves two subjects at a time.

3. We continue merging clusters under this maximum likelihood criteria, at each step making note of the maximized likelihood, until we are left with one cluster containing all n subjects, $\mathcal{C}_n = [n]$. Among the n partitions that are obtained we pick the partition that has the highest value of $L_{\mathcal{C}}$. Note that the likelihood value may either increase or decrease at each step. This provides us with a method that automatically determines the number of clusters.

It is our experience that the individual parameter estimates $(\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i})_i$ do not change much at each merging step of the hierarchical algorithm. Thus, if the initial estimates provide good fits for all the individual profiles the algorithm can be expected to perform well. Furthermore, by using the individual parameter estimates at a previous merging step as initial values at the next step, each EM algorithm converges very quickly, which is essential since the total number of EM algorithms that are conducted is of the order $O(n^2)$. For the data sets that we consider in this article, the hierarchical algorithm takes anywhere from a couple of minutes to run, for the smallest data set in Section 8.1 ($n = 14$), up to a couple of hours for the Erasmus high dimensional data set of Section 7.1 ($n = 344$), using a regular laptop. However, it should be noted that our R code is neither optimized nor precompiled to a lower level programming language at this stage.

4.2. *Iterative clustering algorithm.* The hierarchical algorithm results in a partition that serves as a good initial candidate for the true partition. In this subsection we present an iterative algorithm that is designed to improve upon any initial partition. We introduce cluster membership indicators for the subjects in order to develop an EM algorithm for clustering subjects under the assumption of a fixed

number of clusters. Define for each subject $i = 1, \ldots, n$ and cluster $c \in \mathcal{C}$

$$X_{ic} = \begin{cases} 1, & \text{if subject } i \text{ is in cluster } c, \\ 0, & \text{otherwise,} \end{cases}$$

and let $\mathbf{X}_i = (X_{ic})_{c \in \mathcal{C}}$. Assume $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. Multinom$\{1, \mathbf{p} = (p_c)_{c \in \mathcal{C}}\}$, so the density of $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')'$ is

$$f(\mathbf{X}) = \prod_{i=1}^{n} \prod_{c \in \mathcal{C}} p_c^{X_{ic}}, \quad \sum_{c \in \mathcal{C}} p_c = 1. \tag{7}$$

These cluster membership indicators fully define the partition $\mathcal{C}$ and we note that the classification likelihood in (2) can be written as

$$f(\mathbf{y}|\mathbf{X}) = \prod_{c \in \mathcal{C}} \prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i)^{X_{ic}}. \tag{8}$$

Multiplying (7) and (8) together and integrating out $\mathbf{X}$ we arrive at the marginal likelihood

$$f(\mathbf{y}; \boldsymbol{\Psi}) = \prod_{i=1}^{n} \sum_{c \in \mathcal{C}} p_c f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i), \tag{9}$$

where $\boldsymbol{\Psi} = \{(p_c)_{c \in \mathcal{C}}, (\mathbf{w}_c)_{c \in \mathcal{C}}, \boldsymbol{\theta}_i = (\mu_{1i}, \sigma_{1i}^2, \mu_{2i}, \sigma_{2i}^2)_i\}$ involves both the continuous parameters and the discrete indicators, $\mathbf{w}$, which we now assume are fixed. We make this assumption because if $\mathbf{w}$ is random as in (4) the joint posterior distribution of $(\mathbf{w}, \mathbf{X})$ is highly intractable and an EM algorithm based on (8) would be problematic.

The likelihood in (9) is that of a finite mixture model and can be maximized using the EM algorithm. We detail the maximization procedure in Appendix B of the Supplementary Materials. In short, let $\mathbf{X}^{(0)}$ denote the clustering labels corresponding to a candidate partition. Using $\mathbf{X}^{(0)}$ as an initial partition we run an EM algorithm that converges to a local maximum of (9). Once the mixture model has been fitted, a probabilistic clustering of the subjects can be obtained through the fitted posterior expectations of cluster membership for the subjects, $(E[X_{ic}|\mathbf{y}])_{i,c}$ (see McLachlan and Peel, 2000). Essentially, a subject will be assigned to the cluster to which it has the highest estimated posterior probability of belonging. We have found empirically that the derived partition not only results in a higher value of (9) but also in the objective likelihood (6), but we do not have a theoretical justification for this. A good clustering strategy is to come up with a few candidate partitions, with varying numbers of clusters, and run the EM algorithm using these partitions as initial partitions. Each resulting partition will be a local maximum of (9), but

we choose the partition with the highest value of the original objective function (6). Good initial partitions can be found by running the hierarchical algorithm of subsection 4.1 or applying one of the more standard clustering algorithms.

5. CLASSIFICATION. The construction of a likelihood for any given partition of the subjects also provides a powerful tool for classification. Assume we have methylation data on n subjects and we know which class each subject belongs to, i.e. we know the true $\mathcal{C}$. A by-product of maximizing the likelihood in (6) with the EM algorithm (detailed in the Supplementary Appendix B) is posterior predictions of the latent indicators, $(\hat{\mathbf{w}}_c)_{c\in\mathcal{C}}$, which we round to either 0 or 1. Given these estimated methylation indicators the conditional likelihood of a new observation $(y_{ij})_j$, on the assumption that $i \in c$, is given by

$$L_c(\boldsymbol{\theta}_i) = \prod_{j=1}^{G} \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2)^{\hat{w}_{cj}} \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2)^{1-\hat{w}_{cj}}. \tag{10}$$

The discriminant likelihood, L_c, is maximized with respect to the individual specific parameters at

$$\hat{\mu}_{1i} = \frac{\sum_j \hat{w}_{cj} y_{ij}}{\sum_j \hat{w}_{cj}}, \qquad \hat{\sigma}_{1i}^2 = \frac{\sum_j \hat{w}_{cj}(y_{ij} - \hat{\mu}_{1i})^2}{\sum_j \hat{w}_{cj}},$$
$$\hat{\mu}_{2i} = \frac{\sum_j (1-\hat{w}_{cj}) y_{ij}}{\sum_j (1-\hat{w}_{cj})}, \qquad \hat{\sigma}_{2i}^2 = \frac{\sum_j (1-\hat{w}_{cj})(y_{ij} - \hat{\mu}_{2i})^2}{\sum_j (1-\hat{w}_{cj})}.$$

By substituting these estimates into (10) we arrive at the following discriminant rule:

$$i \in c \quad \text{if} \quad L_c\big(\hat{\boldsymbol{\theta}}_i(\hat{\mathbf{w}}_c)\big) > L_{c'}\big(\hat{\boldsymbol{\theta}}_i(\hat{\mathbf{w}}_{c'})\big) \text{ for all } c' \neq c. \tag{11}$$

6. EXTENSION TO MULTIPLE PLATFORMS. In this section we discuss how to extend the methods of this paper to account for multiple data types as long as each data type can reasonably be modeled by the model described in Section 3. For subject $i = 1, \ldots, n$ let y_{ijk} denote the signal response of the jth variable, $j = 1, \ldots, G_k$, on platform $k = 1, \ldots, m$. As before we let $\mathcal{C}$ denote the true partition of the n subjects. We assume subjects in a given cluster $c \in \mathcal{C}$ have identical activity (methylation, expression, etc.) profiles on each platform $k = 1, \ldots, m$ independently and define a cluster and platform specific indicator for each variable

$$w_{cjk} = \begin{cases} 1, & \text{if variable } j \text{ on platform } k \text{ is active in cluster } c, \\ 0, & \text{if variable } j \text{ on platform } k \text{ is inactive in cluster } c. \end{cases}$$

Define $\mathbf{w}_c = (w_{cjk})_{j,k}$ and let $\mathbf{y}_i = (y_{ijk})_{j,k}$ denote the vector of observed activity profiles of subject i across platforms. Let $\boldsymbol{\theta}_i = (\mu_{1ik}, \sigma^2_{1ik}, \mu_{2ik}, \sigma^2_{2ik})_{k=1}^m$ denote the subject specific mixture parameters. We assume that the observed data, $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$, given the unobserved activity indicators, $\mathbf{w} = (\mathbf{w}_c)_{c \in \mathcal{C}}$, arise from the following density:

$$f(\mathbf{y}|\mathbf{w}, \boldsymbol{\theta}) = \prod_{c \in \mathcal{C}} \prod_{i \in c} f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i),$$

where the conditional density of $\mathbf{y}_i$, on the assumption that $i \in c$, is given by

$$f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i) = \prod_{k=1}^{m} \prod_{j=1}^{G_k} \phi(y_{ijk}|\mu_{1ik}, \sigma^2_{1ik})^{w_{cjk}} \phi(y_{ijk}|\mu_{2ik}, \sigma^2_{2ik})^{1-w_{cjk}}.$$

We can either assume that the activity indicators for cluster c are fixed as in subection 4.2, or independent Bernoullis, both across platforms and variables,

$$f(\mathbf{w}_c) = \prod_{k=1}^{m} \prod_{j=1}^{G_k} \pi_{1ck}^{w_{cjk}} \pi_{0ck}^{1-w_{cjk}}, \quad \pi_{1ck} + \pi_{0ck} = 1,$$

where π_{1ck} represents the proportions of variables on platform k that are active in cluster c. The likelihood in this integrated framework is identical to the one given in (6), except we now have an additional product across platforms k. The methods presented in Sections 4 and 5 thus extend to multiple platforms in a straightforward manner.

7. IDENTIFYING SUBTYPES OF AML. Figueroa *et al.* (2010) performed the first large-scale DNA methylation profiling study in humans using the Erasmus data described in Section 2. They clustered patients using hierarchical correlation based clustering on a subset of the most variable probe sets. Using unsupervised clustering they were able to classify the patients into known and well characterized subtypes as well as discover novel clusters. In subsection 7.1 we report our clustering results on the data and compare to those of Figueroa *et al.* (2010). We ran a cluster analysis on both methylation and expression data separately as well as an integrative cluster analysis on both platforms simultaneously. In subsection 7.2 we present results from a discriminant analysis study in which we classified an independent validation data set using the methods of Section 5.

7.1. *Clustering results.* Figueroa *et al.* (2010) hierarchically clustered the $n = 344$ patients (methylation profiles only) on a subset of the 3,745 most variable probe sets, using 1-correlation distance and Ward's agglomeration method. These

TABLE 1
The sensitivity and specificity of the clustering results.

Sensitivity (# of false negatives in parentheses)

subtype	COR	M	E	ME
inv(16) [$n_1 = 28$]	$0.929(2)$	$0.964(1)$	$0.857(4)$	$0.964(1)$
$t(15;17)$ [$n_2 = 10$]	$0.800(2)$	$0.800(2)$	$1.000(0)$	$1.000(0)$
$t(8;21)$ [$n_3 = 24$]	$0.917(2)$	$0.875(3)$	$0.917(2)$	$0.958(1)$
CEBPA dm [$n_4 = 24$]	$0.792(5)$	$0.917(2)$	$0.75(6)$	$1.000(0)$
CEBPA Sil [$n_5 = 8$]	$0.625(3)$	$0.875(1)$	$1.000(0)$	$1.000(0)$
$11q23$ + FAB M5 [$n_6 = 7$]	$0.857(1)$	$0.857(1)$	$0.714(2)$	$0.714(2)$

Specificity (# of false positives in parentheses)

subtype	COR	M	E	ME
inv(16) [$n - n_1 = 316$]	$1.000(0)$	$0.997(1)$	$0.997(1)$	$1.000(0)$
$t(15;17)$ [$n - n_2 = 334$]	$1.000(0)$	$1.000(0)$	$1.000(0)$	$1.000(0)$
$t(8;21)$ [$n - n_3 = 320$]	$0.972(9)$	$0.994(2)$	$1.000(0)$	$0.991(3)$
CEBPA dm [$n - n_4 = 320$]	$0.988(4)$	$0.988(4)$	$0.991(3)$	$1.000(0)$
CEBPA Sil [$n - n_5 = 336$]	$1.000(1)$	$0.997(1)$	$0.985(5)$	$0.997(1)$
$11q23$ + FAB M5 [$n - n_6 = 337$]	$0.991(3)$	$0.991(3)$	$0.991(3)$	$0.994(2)$

were probe sets that exceeded a standard deviation threshold of 1. We ran the hierarchical algorithm of subsection 4.1 on the same subset to obtain an initial partition. Among the 344 candidate partitions, obtained at each merging step, the loglikelihood was maximized at $K = 17$ clusters but to avoid singletons we chose a partition with $K = 16$, the same number of clusters Figueroa *et al.* (2010) chose. We then applied the iterative algorithm of subection 4.2 in an attempt to improve upon the initial partition. We denote the resulting partition "M" (Methylation). We repeated this process separately for the expression data using the 3,370 most variable probe sets, or those that exceeded a standard deviation threshold of 0.5. This resulted in a partition "E" (Expression) with $K = 17$ clusters. Finally we repeated this process jointly on the 3,745 and 3,370 probe sets from the methylation and expression data, respectively, resulting in the partition "ME" with $K = 14$ clusters.

Figueroa *et al.* (2010) identified 3 robust and well-characterized biological clusters and 8 clusters that were associated with specific genetic or epigenetic lesions. Five clusters seemed to share no known biological features. The three robust clusters corresponded to cases with inversions on chromosome 16, inv(16), and translocations between chromosomes 15 and 17, $t(15;17)$, and chromosomes 8 and 21, $t(8;21)$ (translocations refer to when genetic material from two different chromosomes breaks apart and when being repaired, the material from one chromosome is incorrectly attached to the other chromosome instead and vice versa). The World

Health Organization has identified these subtypes of AML as indicative of favorable clinical prognosis (see for example Figueroa *et al.*, 2010). The remaining 8 clusters included patients with CEBPA double mutations (two different abnormal changes in the genetic code of the CEBPA gene), CEBPA mutations irrespective of type of mutation, silenced CEBPA (abnormal loss of expression of CEBPA which is not due to mutations in the genetic code), one cluster enriched for 11q23 abnormalities (any type of change in the genetic code that affects position 23 of the long arm of chromosome 11) and FAB M5 morphology (specific shape and general aspect of the leukemic cell as defined by the French American British classification system for Acute Leukemias), and four clusters with NPM1 mutations (mutations in the genetic code of the NPM1 gene). A detailed sensitivity and specificity analysis of 6 of the above 11 clusters [sample sizes in brackets], inv(16) [$n_1 = 28$], $t(15;17)$ [$n_2 = 10$], $t(8;21)$ [$n_3 = 24$], CEBPA double mutations [$n_4 = 24$], CEBPA silenced AMLs [$n_5 = 8$], and $11q23$ + FAB M5 [$n_6 = 7$], is given in Table 1 for the different clustering results. We include the correlation based clustering result (COR) on the methylation data to compare with "M". The remaining five of the 11 biological clusters (CEBPA mutations irrespective of type of mutation and the four NPM1 mutation clusters) had sensitivity or specificity below 0.5 for all four clustering results and were thus excluded from the table. We can see that the model based approach, "M", is doing better than the correlation based method, "COR", for the most part. Aside for sensitivity of $t(8;21)$ (1 less false negative) and specificity of inv(16) (1 less false positive) the model based approach has as good or better sensitivity and specificity. The most striking differences are in the numbers of false negatives of CEBPA dm and false positives of $t(8;21)$ where "M" is doing better. Note also that aside for the sensitivity of $11q23$ + FAB M5 and specificity of $t(8;21)$ the integrated analysis "ME" always does better than the analyses "M" and "E" separately, with perfect sensitivity and specificity for many of the clusters. Most notably, the integrative analysis is able to perfectly classify the CEBPA double mutants even though both "M" and "E" have quite a few false positives and false negatives. This demonstrates the increased power to identify clusters by sharing information across multiple platforms. As a side product from our clustering algorithm we obtain posterior probabilities of high methylation/expression, $E[w_{cj}|\mathbf{y}]$, which can be used to order genes in heatmaps to discover subtype specific methylation/expression patterns. In Figure 2 we see heatmaps of the two data sets used for the integrative clustering, "ME", after rows have been ordered by increasing posterior probabilities (one cluster at a time). Such heatmaps are useful for graphically displaying the distinct methylation/expression patterns that characterize the different subtypes of cancer.

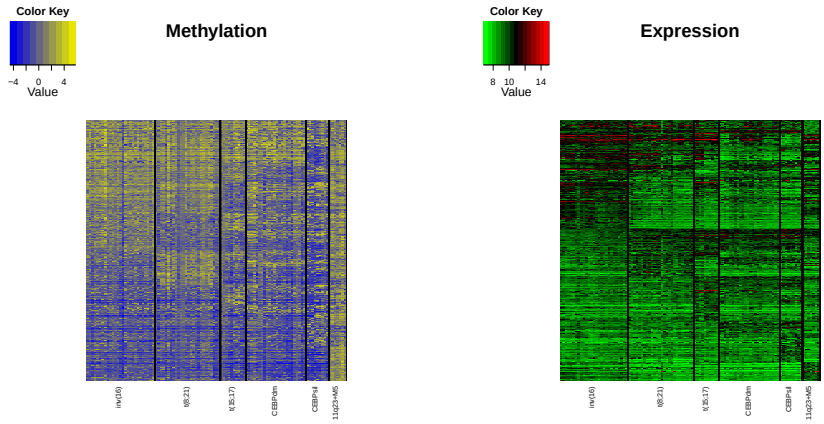

FIG 2. *Heatmaps of the methylation and gene expression values for the 6 well characterized clusters obtained from the integrative clustering ("ME"). Columns correspond to patient samples and rows correspond to genes. On the left heatmap "yellow", "gray", and "blue" represent low, intermediate, and high methylation, respectively. On the right heatmap "green", "black", and "red" represent low, intermediate, and high expression, respectively. Rows are ordered by increasing posterior probabilities of high methylation/expression,* $E[w_{cj}|\mathbf{y}]$*, first for inv*(16)*, then for* $t(8;21)$*,* $t(15;17)$*, CEBPdm, CEBPsil, and finally for 11q23+M5. Rows with equal probabilities for all* 6 *clusters (either equal to* 1 *or* 0*) are excluded to emphasize differences.*

7.2. *Classification results.* A second cohort of patients with AML was available with which we could test the performance of the classification method of Section 5. This second cohort of $n = 383$ cases consisted of samples, obtained from patients enrolled in a clinical trial from the Eastern Cooperative Oncology Group (ECOG) (Data are available at http://www.ncbi.nlm.nih.gov/geo/, accession number pending). These patients were similar in characteristics to the Erasmus cohort with only one exception, all patients were younger than 60 years of age. Samples were processed in the same way as the Erasmus cohort, and their methylation was used to blindly predict their molecular diagnosis. Using the 3,745 most variable probe sets and the clustering result "M" of the previous section we fit the model (6) on the Erasmus cohort with the EM algorithm. By using the posterior predictions of the methylation indicators we applied the discriminant rule (11) on each patient in the ECOG data set. Since CEBPA and NPM1 mutation status have not yet been made available for this cohort, only the performance for the prediction of the inv(16), $t(8;21)$, CEBPA silenced, $t(15;17)$ and $11q23$ + FAB M5 clusters could be tested. Inv(16) cases were predicted with 100% sensitivity and specificity. The predicted $t(8;21)$ cluster contained 100% of cases positive for this abnormality, and only one $t(8;21)$ case was misclassified to another cluster. Two cases, which had previously been unrecognized as CEBPA silenced AMLs were predicted by

TABLE 2
The sensitivity and specificity of the classification result. False negatives and false positives are in parentheses.

subtype	Sensitivity	Specificity
inv(16) [$n = 32$]	$1.000(0)$	$1.000(0)$
$t(15;17)$ [$n = 1$]	$1.000(0)$	$1.000(0)$
$t(8;21)$ [$n = 28$]	$0.964(1)$	$1.000(0)$
CEBPA Sil [$n = 1$]	$1.000(0)$	$0.997(1)$
$11q23$ + FAB M5 [$n = 14$]	$0.643(5)$	$0.986(5)$

the classification method. One of them was later confirmed to indeed correspond to this molecular subtype by an alternative methylation measurement method. Similarly one case was believed to have been misclassified as $t(15;17)$ since there were no molecular data confirming the presence of the PML-RARA gene fusion (the abnormal combination of the PML and RARA genes) resulting from this translocation. However, it was later confirmed that both the morphology and the immune diagnosis corresponded to that of an acute promyelocytic leukemia with $t(15;17)$. Finally, the $11q23+$FAB M5 cluster included 9 of the 14 patients in the cohort that met these two criteria. There were also 5 false positives, 3 of them were M5 cases but did not have $11q23$ abnormalities, 1 of them harbored the $11q23$ abnormality but corresponded to an M1, and the remaining case corresponded to an M4 case with a hyperdiploid karyotype. Summary of these results is provided in Table 2.

8. OTHER APPLICATIONS. The clustering method presented in this paper is not restricted to the microarray platforms that the AML samples were processed on. In this section we demonstrate the versatility of our method by applying it to other microarray platforms and show that our algorithm does well in clustering subjects. We also provide a comparison with other existing methods for clustering microarray data.

8.1. *Expression in endometrial cancer.* In this subsection we analyze the microarray expression data set in Tadesse, Sha and Vannucci (2005). Endometrioid endometrial adenocarcinoma is a gynecologic malignancy typically occuring in postmenopausal women. Identifying distinct subtypes based on common patterns of gene expression is an important problem as different clinicopathologic groups may respond differently to therapy. Such subclassification may lead to discoveries of important biomarkers that could become targets for therapeutic intervention and improved diagnosis. High density microarrays (Affymetrix Hu6800 chips) were used to study expression of 4 normal and 10 endometrioid adenocarcinomas on 7,070 probe sets. Probe sets with at least one unreliable reading (limits of reliable detection were set to 20 and 16,000) were removed from the analysis, which resulted in $G = 762$ variables. Finally, the data were log-transformed, however

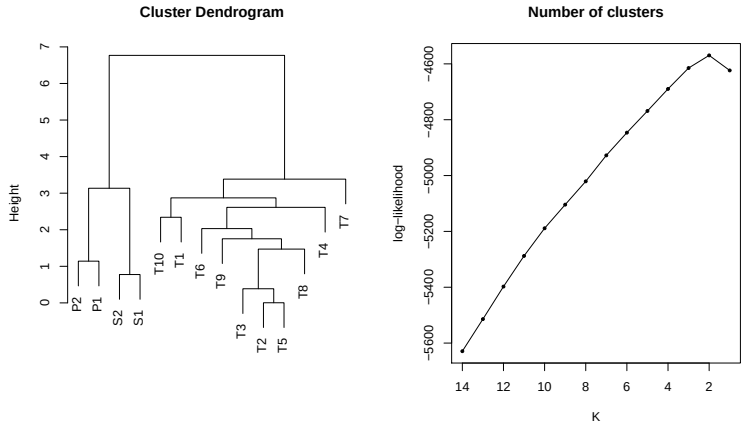

FIG 3. *Left panel: Cluster dendrogram for the* 2 *normal proliferative ("P") endometria,* 2 *normal secretory ("S") endometria, and* 10 *endometrioid adenocarcinomas ("T"). Right panel: Plot of the number of clusters against log-likelihood.*

unlike Tadesse, Sha and Vannucci (2005) we chose not to rescale the variables by their range. More details about the data set can be found in Tadesse, Ibrahim and Mutter (2003) and is publically available at http://endometrium.org.

We hierarchically clustered the samples using the 300 most variable probe sets and plotted the results in Figure 3. We successfully separated the four normal tissues from the endometrial cancer tissues and the log-likelihood plot suggests that $K = 2$. The dendrogram consistently separated the normals and the cancer into two branches for different variance thresholds. However, if we lowered the threshold too much the loglikelihood was maximized at $K = 1$. This makes sense as including many low variability probe sets might mask the true clustering structure. For comparison, Tadesse, Sha and Vannucci (2005) concluded that $K = 3$ and commented that there could possibly be 2 subtypes of endometrial cancer. However, to our best knowledge this subclustering has not been verified. They also reported clustering results using the COSA algorithm of Friedman and Meulman (2003), which like our analysis seemed more suggestive of $K = 2$.

8.2. *Methylation in normal tissues.* Houseman *et al.* (2008) used the RPMM algorithm to cluster a methylation data set consisting of 217 normal tissues and compared the performance to that of the HOPACH algorithm of van der Laan and Pollard (2003). The RPMM analysis was discussed in more detail in Christensen *et al.* (2009) and the data made publically available at the GEO depository with accession number GSE19434 (http://www.ncbi.nlm.nih.gov/geo/). Briefly, DNA was extracted from the tissues, modified with sodium bisulfite, and processed on the

Illumina GoldenGate methylation platform. Average fluorescence for methylated (M) and unmethylated (U) alleles were derived from raw data at $1,505$ loci. However, in this study $1,413$ loci passed the quality assurance procedures. A total of 11 tissue types were available, bladder ($n = 5$), adult blood ($n = 30$), infant blood ($n = 55$), 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 ($n = 5$). Houseman *et al.* (2008) constructed an average "beta" value from raw data, which they claimed was very close to the quantity $M/(M+U) \in (0,1)$, rendering a beta distributional assumption (assuming M and U follow a gamma distribution) with class and locus specific parameters. We however chose to work with the quantity $\log(M/U)$, which fits better with our two component mixture distributional assumption. In order to get a direct comparison with RPMM and HOPACH we used all $1,413$ loci. In Supplementary Figure 6 we see a plot of cluster number versus loglikelihood, which is maximized at $K = 6$ clusters. However, given the relatively small difference between the loglikelihood values at $K = 6$, $K = 7$, and $K = 8$ one could argue that all three clustering results are worthy of consideration. The clusters are cross-classified with tissue type in Table 3.

If we compare these results with those obtained with the RPMM algorithm our result favors few and concise clusters, whereas RPMM is indicative of a total of 23 subclasses of tissues. The HOPACH clustering algorithm was suggestive of $K = 9$ clusters with 3 of those clusters representing placenta singletons separated from the main placenta cluster. We present a cross-classification table for both RPMM and HOPACH in Supplementary Table 6 for comparison (borrowed from Houseman *et al.*, 2008). Our method perfectly classifies blood, brain, infant blood, kidney, and placenta. For comparison, after bundling subclusters together, RPMM classifies blood and infant blood perfectly, and HOPACH classifies infant blood and placenta perfectly. All three methods have problems distinguishing between bladder, cervical, lung, pleural, and small intestine tissues. Overall, our approach seems to outperform HOPACH, and although Houseman *et al.* (2008) have demonstrated that a few of their tissue-specific subclusters (obtained by RPMM) have verifiable meanings, such as through age difference, it seems that without a further justification of such finer substructure in the data our clustering result is more desirable. As a side note, under the assumption of Houseman *et al.* (2008) that M and U follow a gamma distribution it is clear that $\log(M/U)$ will not be a mixture of two Gaussian distributions. The favorable clustering result for this data set suggests that the normality assumption on each mixture component provides a robust and flexible modeling distribution.

9. DISCUSSION. We have proposed a model-based method for clustering microarray data. The methods have been demonstrated to work well on expression

TABLE 3
Cross-classification of our clustering result ($K = 8$) and tissue type. By merging the top two clusters, and clusters 3 and 4 we obtain the clustering corresponding to $K = 6$.

Class	Blad	Bl	Br	Cerv	Inf bl	HN	Kid	Lung	Plac	Pleu	Sm int
1	5			2		1		53		18	4
2				1							
3							6				
4						10					1
5		30									
6			12								
7					55						
8									19		

data and methylation data separately. An integrated cluster analysis has further shown the power of combining platforms in a joint analysis. We believe this method can be applied to a variety of microarray data types. However, further research is needed to validate the method on different types of data such as ChIP-chip data.

A minor drawback of our method is that it does not allow for automated selection of variables, but rather relies on pre-filtering the data. However, most biologists are still relying on simple clustering algorithms such as K-means or standard agglomerative algorithms, due to their simplicity in implementation and interpretation. Thus having a relatively simple and easily implemented hierarchical algorithm that can integrate multiple platforms and further utilizes the bimodal or skewed structure of the individual profiles, in a model-based manner, has its advantages. For example, the hierarchical algorithm automatically determines the numbers of clusters and provides an easily interpretable dendrogram. Also, as a side product we obtain posterior probabilities of high methylation/expression, $E[w_{cj}|\mathbf{y}]$, for each cluster c and probe set j. By ordering the probe sets with respect to these posterior probabilities and excluding probe sets that are identical across all clusters we can explore patterns in heatmaps such as in Figure 2.

One of the novelties of our clustering algorithm is the inclusion of individual specific parameters, $(\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i})_i$, into the model of Section 3, which facilitates the use of our algorithm even in the presence of extreme microarray effects. Since the amount of data we have to estimate these parameters (G observations per subject) highly exceeds the number of subjects (n) the estimation of these parameters has not been a problem. However, it is common practice to treat such individual specific parameters as random effects. We have established that with conjugate normal and inverse gamma priors on the above parameters we arrive at a marginal likelihood intractable for maximization. However, we have verified that through

such prior specifications we could easily calculate full conditionals in a Bayesian analysis. A Bayesian approach would also prevent us from having to assume the methylation indicators, $\mathbf{w}$, are fixed as in the iterative algorithm of subection 4.2. The reason for that assumption was that the joint posterior distribution of $(\mathbf{w}, \mathbf{X})$ is highly intractable. However, full conditionals for each variable separately are easily obtained and are that of Bernoulli and Multinomial, respectively. Running a fully Bayesian analysis might also facilitate an extended algorithm that could include all variables. One might assume that some variables are informative and follow the mixture in (6) with prior probability p, wheras other variables are non-informative with prior probability $(1-p)$ and follow the mixture in (6) with $\mathcal{C} = [n]$. There are some challenges that arise in implementing such a fully Bayesian model and those require further research.

ACKNOWLEDGEMENTS. The methods presented in this paper have been implemented as an R-package that is available at http://www.stat.cornell.edu/imac/.

References.

ARMSTRONG, S. A., STAUNTON, J. E., SILVERMAN, L. B., PIETERS, R., DEN BOER, M. M. D.M. L., SALLAN, S. E., LANDER, E. S., GOLUB, T. R. and KORSMEYER, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics* **30** 41-47.

BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803-821.

CHEESEMAN, P. and STUTZ, J. (1995). Bayesian Classification (AutoClass): Theory and Results. *in Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press* **49** 153-180.

CHRISTENSEN, B. C., HOUSEMAN, E. A., MARSIT, C. J., ZHENG, S., WRENSCH, M. R., WIEMELS, J. L., NELSON, H. H., KARAGAS, M. R., PADBURY, J. F., BUENO, R., SUGARBAKER, D. J., YEH, R.-F., WIENCKE, J. K. and KELSEY, K. T. (2009). Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLOS Genetics*.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society* **39** 1-38.

FIGUEROA, M. E., REIMERS, M., THOMPSON, R. F., YE, K., LI, Y., SELZER, R. R., FRIDRIKSSON, J., PAIETTA, E., WIERNIK, P., GREEN, R. D., GREALLY, J. M. and MELNICK, A. (2008). An Integrative Genomic and Epigenomic Approach for the Study of Transcriptional Regulation. *PLoS One* **3** e1882.

FIGUEROA, M. E., LUGTHART, S., LI, Y., ERPELINCK-VERSCHUEREN, C., DENG, X., CHRISTOS, P. J., SCHIFANO, E., BOOTH, J., VAN PUTTEN, W., SKRABANEK, L., CAMPAGNE, F., MAZUMDAR, M., GREALLY, J. M., VALK, P. J. M., LOWENBERG, B., DELWELSEND, R. and MELNICK, A. (2010). Epigenetic signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* **17**.

FRALEY, C. and RAFTERY, A. E. (1998). How Many clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *The Computer Journal* **41** 578-588.

FRALEY, C. and RAFTERY, A. E. (2002). Model-Based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association* **97** 611-631.

FRIEDMAN, J. H. and MEULMAN, J. J. (2003). Clustering Objects on Subsets of Attributes. *technical report. Stanford University. Dept. of Statistics and Stanford Linear Accelerator Center.*

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R., CALIGIURI, M. A., BLOOMFIELD, C. D. and LANDER, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286** 531-537.

HEARD, N. A., HOLMES, C. C. and STEPHENS, D. A. (2006). A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **101** 18-29.

HOUSEMAN, E. A., CHRISTENSEN, B. C., YEH, R.-F., MARSIT, C. J. and OTHERS, (2008). Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *Bioinformatics* **9**.

IBRAHIM, J. G., CHEN, M. H. and GRAY, R. J. (2002). Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association* **97** 88-99.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **2** 249-264.

JI, Y., WU, C., LIU, P., WANG, J. and COOMBES, K. R. (2005). Applications of beta-mixture models in bioinformatics. *Bioinformatics* **21**.

KETTENRING, J. R. (2006). The practice of cluster analysis. *Journal of Classification* **23** 3-30.

KIEFER, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica.* **46** 427-434.

KIM, S., TADESSE, M. G. and VANNUCCI, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93** 877-893.

LINDSAY, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80** 221-240.

MCLACHLAN, G. J. and BASHFORD, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

MCLACHLAN, G. J., BEAN, R. W. and PEEL, D. (2002). A Mixture Model-Based Approach to the Clustering of Microarray Expression Data. *Bioinformatics* **18** 413-422.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley.

RODENHISER, D. and MANN, M. (2006). Epigenetics and human disease: translating basic biology into clinical applications. *Canadian Medical Association Journal* **174** 341-348.

SCOTT, A. J. and SYMONS, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27** 387-397.

SIEGMUND, K. D., LAIRD, P. W. and LAIRD-OFFRINGA, I. A. (2004). A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics* **20** 1896-1904.

SYMONS, M. J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* **37** 35-43.

TADESSE, M. G., IBRAHIM, J. G. and MUTTER, G. L. (2003). Identification of Differentially Expressed Genes in High-Density Oligonucleotide Arrays Accounting for the Quantification Limits of the Technology. *Biometrics.* **59** 542-554.

TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100** 602-617.

THOMPSON, R. F., REIMERS, M., KHULAN, B., GISSOT, M., RICHMOND, T. A., CHEN, Q., ZHENG, X., KIM, K. and GREALLY, J. M. (2008). An analytical pipeline for genomic representations used for cytosine methylation studies. *Bioinformatics.* **24** 1161-1167.

VALK, P. J., VERHAAK, R. G., BEIJEN, M. A., ERPELINCK, C. A., VAN WAALWIJK VAN DOORN-KHOSROVANI, S. B., BOER, J. M., BEVERLOO, H. B., MOORHOUSE, M. J., VAN DER SPEK, P. J., LOWENBERG, B. and DELWEL, R. (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med* **350** 1617-1628.

VAN DER LAAN, M. J. and POLLARD, K. S. (2003). A new algorithm for hybrid hierarchical clus-

tering with visualization and the bootstrap. *Journal of Statistical Planning and Inference* **117** 275-303.

SUPPLEMENTARY MATERIAL

Supplement: SIMULATION AND DETAILS OF EM ALGORITHMS (http://lib.stat.cmu.edu/aoas/???/???). We perform a simulation study to assess the performance of our clustering algorithm in the presence of sparse correlation structure. We also derive the steps involved in maximizing the likelihoods of the several models presented in this paper.

APPENDIX A: SIMULATION STUDY

In microarray data one sometimes observes groups of genes that are highly correlated with each other and can sometimes form functionally meaningful pathways. However, these gene pathways tend to be limited to a small number of genes leading to a sparse overall correlation structure. In this appendix we perform a simulation study under the assumption that all of the observed genes form such groups of genes that have high within-group correlations. The simulation scenarios presented below are designed to give a sense of how far the correlation structure can be pushed to the extreme before the algorithm breaks down. All of these scenarios would be extreme in a biological setting and therefore the results give indication of how robust the clustering algorithm is to the independence assumption. We end with a simple simulation under the empirical correlation structure of the HELP methylation data presented in Section 7.1.

We used as baseline the individual specific estimates of $(\mu_{1i}, \sigma^2_{1i}, \mu_{2i}, \sigma^2_{2i})_i$ obtained from the 59 samples that made up the three robust clusters, inv(16), $t(8; 21)$, and $t(15; 17)$, of our methylation clustering result ("M") from Section 7 (See panels $(1, 1)$, $(1, 3)$, and $(2, 3)$ of Supplementary Figure 5). For each cluster $c = 1, 2, 3$ we used the mean of the individual specific proportions estimates $(\pi_{1i})_{i\in c}$ to represent the underlying cluster specific probability of high methylation, π_{1c}. These calculations led to proportions estimates 0.61, 0.59, and 0.59, for the clusters inv(16), $t(8; 21)$, and $t(15; 17)$, respectively. We let $G = 500$ be the number of simulated genes and assumed that the data comprised of a total of G/m groups of m genes each that exhibited moderate to high within-group correlations. For simplicity the correlation structure was taken to be the same in all groups of m genes. We performed the simulation for two sceniarios, $m = 10$ and $m = 50$, and the simulation steps are described as follows

1. Simulate for each group $c = 1, 2, 3$, and each gene $j = 1, \ldots, 500$, independently $w_{cj} \sim Bern(\pi_{1c})$.
2. Simulate for each sample $i \in c$, in all clusters $c = 1, 2, 3$, and each group of m genes

$$(y_{i,j_1}, \ldots, y_{i,j_m})' \sim N(\mathbf{w}_{c.}\mu_{1i} + (1 - \mathbf{w}_{c.})\mu_{2i}, \mathbf{D}_i^{1/2}\mathbf{R}\mathbf{D}_i^{1/2}), \tag{12}$$

where $\mathbf{w}_{c.} = (w_{c,j_1}, \ldots, w_{c,j_m})'$, $\mathbf{D}_i = \text{diag}(\mathbf{w}_{c.}\sigma^2_{1i} + (1 - \mathbf{w}_{c.})\sigma^2_{2i})$, and the correlation matrix $\mathbf{R}$ takes on either an AR(1) structure or compound symmetry structure with correlation coefficient $\rho = 0.7, 0.8, 0.9$, or 0.99.

We repeated the above simulation steps 1,000 times and each time we ran the hierarchical clustering algorithm of Section 4.1 on the simulated data.

For each simulation run we estimated the number of clusters K (by comparing loglikelihoods of the merging step candidate partitions) and noted how many times the clustering algorithm correctly estimated $K = 3$. Additionally, for each simulation run we cut the dendrogram at $K = 3$ (regardless of the estimated value of K) and observed whether or not the resulting clustering perfectly classified the $n = 59$ samples into the 3 known clusters. In Table 4 we report the percentages of simulation runs resulting in $K = 3$ (out of $1,000$) for the different scenarios. We note that for $m = 10$ the clustering algorithm correctly estimated K in vast majority of cases. However, for the scenario where $m = 50$ the algorithm broke down when the correlations became too extreme. In Table 5 we report the percentages of simulation runs resulting in perfect classification when the dendrogram was cut at $K = 3$. We observed perfect classification when $m = 10$ in all cases, but again observed that the algorithm broke down when the correlations became too extreme in the scenario where $m = 50$.

To summarize, our clustering algorithm seems to be robust to the independence assumption in situations where we have a sparse overall correlation structure in which genes tend to group into many small clusters with moderate to high within-group correlations. However, there is indication that with larger groups of genes with very extreme within-group correlations the algorithm will break down. In microarray data such extreme correlation structures are not to be expected on a global scale. In fact when we estimated an empirical correlation matrix (using the cor.shrink() function from the R-package "corpcor") for the HELP methylation data in Section 7.1 we observed that only 0.1% of gene pairs have correlation > 0.7. As a final simulation we repeated steps 1. and 2. above 1,000 times with $m = 500$ and the correlation matrix $\mathbf{R}$ taken as the empirical correlation matrix obtained from 500 randomly chosen columns of the HELP data. The clustering algorithm correctly estimated $K = 3$ and perfectly classified the subjects into the 3 clusters in all 1,000 cases. With these and the above results we conclude that the independence assumption in our model seems quite reasonable.

APPENDIX B: EM ALGORITHMS

In this appendix we describe the EM algorithms of subsections 4.1 and 4.2.

TABLE 4
Percentage of simulation runs (out of 1,000) that estimated correctly $K = 3$.

	Scenario 1 ($m = 10$)		Scenario 2 ($m = 50$)	
ρ	AR(1)	Compound Symmetry	AR(1)	Compound Symmetry
0.7	100%	100%	100%	99.5%
0.8	100%	100%	100%	94.8%
0.9	100%	99.9%	100%	76.9%
0.99	99.6%	99.0%	87.1%	14.8%

TABLE 5
Percentage of simulation runs (out of 1,000) that perfectly classified the $n = 59$ subjects when dendrogram was cut at $K = 3$.

	Scenario 1 ($m = 10$)		Scenario 2 ($m = 50$)	
ρ	AR(1)	Compound Symmetry	AR(1)	Compound Symmetry
0.7	100%	100%	100%	100%
0.8	100%	100%	100%	100%
0.9	100%	100%	100%	99.8%
0.99	100%	100%	99.2%	41.0%

B.1. EM algorithm for hierarchical clustering. It is easy to verify that the complete data loglikelihood of cluster c under the model in (6) is given by

$$\begin{aligned}\log f(\mathbf{y}, \mathbf{w}; \mathbf{\Psi}) &= \sum_{c\in\mathcal{C}}\sum_{j=1}^{G} w_{cj}\Big\{\log\pi_{1c} + \sum_{i\in c}\log\phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})\Big\} \\ &\quad +(1-w_{cj})\Big\{\log\pi_{0c} + \sum_{i\in c}\log\phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})\Big\},\end{aligned}$$

and the EM algorithm simply involves iterating between the two steps below:

1. *E-step*. After t iterations we take the expectation of the complete data loglikelihood with respect to the density $f(\mathbf{w}|\mathbf{y}, \mathbf{\Psi}^{(t)})$. But this density is easily derived from (5), and (6), and is that of independent Bernoullis with posterior expectations

$$\tau^{(t)}_{cj} = \frac{\pi^{(t)}_{1c}\prod_{i\in c}\phi\big(y_{ij}|\mu^{(t)}_{1i}, \sigma^{2(t)}_{1i}\big)}{\pi^{(t)}_{1c}\prod_{i\in c}\phi\big(y_{ij}|\mu^{(t)}_{1i}, \sigma^{2(t)}_{1i}\big) + \pi^{(t)}_{0c}\prod_{i\in c}\phi\big(y_{ij}|\mu^{(t)}_{2i}, \sigma^{2(t)}_{2i}\big)}.$$

Substituting into the complete data loglikelihood we arrive at the Q-function

$$\begin{aligned}Q(\mathbf{\Psi}|\mathbf{\Psi}^{(t)}) &= \sum_{c\in\mathcal{C}}\sum_{j=1}^{G}\tau^{(t)}_{cj}\Big\{\log\pi_{1c} + \sum_{i\in c}\log\phi(y_{ij}|\mu_{1i}, \sigma^2_{1i})\Big\} \\ &\quad +\big(1-\tau^{(t)}_{cj}\big)\Big\{\log\pi_{0c} + \sum_{i\in c}\log\phi(y_{ij}|\mu_{2i}, \sigma^2_{2i})\Big\}.\end{aligned}$$

2. *M-step*. The M-step involves maximizing the Q-function with respect to $\mathbf{\Psi}$. This turns out to be a straightforward task and closed form maximizers exist. By differentiating with respect to π_{1c} for each $c \in \mathcal{C}$ and setting to zero we get the updates

$$\pi_{1c}^{(t+1)} = \frac{1}{G}\sum_{j=1}^{G} \tau_{cj}^{(t)}.$$

Differentiating with respect to μ_{1i} and μ_{2i} for each $i \in c$ and setting to zero leads to

$$\begin{aligned} \mu_{1i}^{(t+1)} &= \frac{\sum_j \tau_{cj}^{(t)} y_{ij}}{\sum_j \tau_{cj}^{(t)}}, \\ \mu_{2i}^{(t+1)} &= \frac{\sum_j (1-\tau_{cj}^{(t)}) y_{ij}}{\sum_j (1-\tau_{cj}^{(t)})}, \end{aligned}$$

and finally, by differentiating with respect to σ_{1i}^2 and σ_{2i}^2 for each $i \in c$ and setting to zero we get

$$\begin{aligned} \sigma_{1i}^{2(t+1)} &= \frac{\sum_j \tau_{cj}^{(t)} (y_{ij} - \mu_{1i}^{(t+1)})^2}{\sum_j \tau_{cj}^{(t)}}, \\ \sigma_{2i}^{2(t+1)} &= \frac{\sum_j (1-\tau_{cj}^{(t)}) (y_{ij} - \mu_{2i}^{(t+1)})^2}{\sum_j (1-\tau_{cj}^{(t)})}. \end{aligned}$$

B.2. Two way EM algorithm. In this section we provide details for the EM algorithm for maximizing the likelihood in (9). From (8) and (7) we arrive at the complete data loglikelihood

$$\begin{aligned} \log f(\mathbf{y}, \mathbf{X}; \mathbf{\Psi}) &= \sum_{c\in\mathcal{C}} \sum_{i=1}^{n} X_{ic} \log p_c \\ &+ \sum_{c\in\mathcal{C}} \sum_{j=1}^{G} w_{cj} \sum_{i=1}^{n} X_{ic} \log \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \\ &+ \sum_{c\in\mathcal{C}} \sum_{j=1}^{G} (1-w_{cj}) \sum_{i=1}^{n} X_{ic} \log \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2), \end{aligned} \tag{13}$$

and the EM algorithm simply involves iterating between the two steps below

1. *E-step*. From (8) and (7) it is clear that the posterior distribution of $\mathbf{X}$ is a product of multinomials

$$f(\mathbf{X}|\mathbf{y}) = \prod_{i=1}^{n} \prod_{c \in \mathcal{C}} \left(\frac{p_c f(\mathbf{y}_i|\mathbf{w}_c, \boldsymbol{\theta}_i)}{\sum_k p_k f(\mathbf{y}_i|\mathbf{w}_k, \boldsymbol{\theta}_i)} \right)^{X_{ic}}.$$

At a current iterate of the parameter estimates, $(w_{cj}^{(t)})_{c,j}$, $(\boldsymbol{\theta}_i^{(t)})_i$, and $(p_c^{(t)})_c$, we obtain the posterior expectation of X_{ic}

$$\kappa_{ic}^{(t)} = \frac{p_c^{(t)} f(\mathbf{y}_i|\mathbf{w}_c^{(t)}, \boldsymbol{\theta}_i^{(t)})}{\sum_k p_k^{(t)} f(\mathbf{y}_i|\mathbf{w}_k^{(t)}, \boldsymbol{\theta}_i^{(t)})}, \tag{14}$$

and arrive at the Q-function

$$\begin{aligned} Q(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(t)}) = & \sum_{c \in \mathcal{C}} \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log p_c \\ & + \sum_{c \in \mathcal{C}} \sum_{j=1}^{G} w_{cj} \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij}|\mu_{1i}, \sigma_{1i}^2) \\ & + \sum_{c \in \mathcal{C}} \sum_{j=1}^{G} (1 - w_{cj}) \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij}|\mu_{2i}, \sigma_{2i}^2). \end{aligned}$$

1^* *Modified E-step.* The implementation of the EM algorithm is simplified by calculating $\kappa_{ic}^{(t)}$ as defined in (14) for each i, c, but modify the E-step such that $\kappa_{ic}^{(t)} = 1$ if $\kappa_{ic}^{(t)} > \kappa_{ic'}^{(t)}$ for all $c' \neq c$ and $\kappa_{ic}^{(t)} = 0$ otherwise. Iterating between the modified E-step and the M-step below results in the so called classification EM (CEM) algorithm (McLachlan and Peel (2000)), where instead of maximizing the marginal likelihood we maximize the complete data likelihood.

2. *M-step*. It is easy to verify that if we differentiate the Q-function with respect to p_c (keeping in mind that $\sum_c p_c = 1$) and set to zero we arrive at the following updating formula for the cluster membership proportions:

$$p_c^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \kappa_{ic}^{(t)},$$

for all $c \in \mathcal{C}$. However, for maximizing the Q-function with respect to $(w_{cj})_{c,j}$ and $\boldsymbol{\theta}_i = (\mu_{1i}, \mu_{2i}, \sigma_{1i}^2, \sigma_{2i}^2)_i$ we need an iterative procedure. This iterative procedure is identical in nature to the Classification EM mentioned above. The idea is quite simple, we iterate between the two steps below until convergence is reached and update the parameters accordingly to $(\boldsymbol{\theta}_i^{(t+1)})_i$ and $(w_{cj}^{(t+1)})_{c,j}$.

(a) For a current value of $(w_{cj})_{c,j}$ the Q-function is maximized at the following values of the parameters:

$$\hat{\mu}_{1i} = \frac{\sum_c \kappa_{ic}^{(t)} \sum_j w_{cj} y_{ij}}{\sum_c \kappa_{ic}^{(t)} \sum_j w_{cj}}, \qquad \hat{\sigma}^2_{1i} = \frac{\sum_c \kappa_{ic}^{(t)} \sum_j w_{cj} (y_{ij} - \hat{\mu}_{1i})^2}{\sum_c \kappa_{ic}^{(t)} \sum_j w_{cj}},$$

$$\hat{\mu}_{2i} = \frac{\sum_c \kappa_{ic}^{(t)} \sum_j (1 - w_{cj}) y_{ij}}{\sum_c \kappa_{ic}^{(t)} \sum_j (1 - w_{cj})}, \quad \hat{\sigma}^2_{2i} = \frac{\sum_c \kappa_{ic}^{(t)} \sum_j (1 - w_{cj}) (y_{ij} - \hat{\mu}_{2i})^2}{\sum_c \kappa_{ic}^{(t)} \sum_j (1 - w_{cj})}.$$

(b) For current values of $(\hat{\mu}_{1i}, \hat{\mu}_{2i}, \hat{\sigma}^2_{1i}, \hat{\sigma}^2_{2i})_i$ the Q-function is maximized at $w_{cj} = 1$ or 0 according to whether

$$\sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij} | \hat{\mu}_{1i}, \hat{\sigma}^2_{1i}) > \sum_{i=1}^{n} \kappa_{ic}^{(t)} \log \phi(y_{ij} | \hat{\mu}_{2i}, \hat{\sigma}^2_{2i})$$

holds or not.

Note that iterating between steps 1^* (modified E-step) and 2 (M-step) above we are essentially maximizing the complete data loglikelihood in (13) jointly with respect to $(\mathbf{w}, \mathbf{X})$ and the parameters. The resulting MLEs provide us with information about the cluster memberships of the subjects and the activity status of the variables simultaneously. Hence, the name "Two way EM algorithm".

MATTHIAS KORMAKSSON
JAMES G. BOOTH
DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853
E-MAIL: mk375@cornell.edu
jb383@cornell.edu

MARIA E. FIGUEROA
ARI MELNICK
DEPARTMENT OF MEDICINE
HEMATOLOGY ONCOLOGY DIVISION
WEILL CORNELL MEDICAL COLLEGE
NEW YORK, NY 10065
E-MAIL: maf2049@med.cornell.edu
amm2014@med.cornell.edu

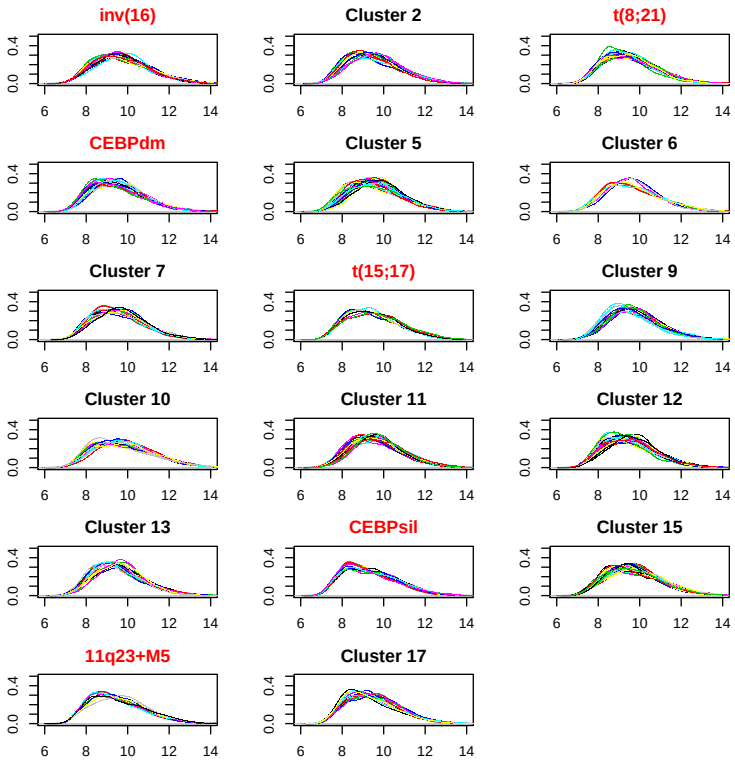

FIG 4. *Gene expression density profiles of all* 344 *patients stratified by the expression clustering result, "E", of Section 7.*

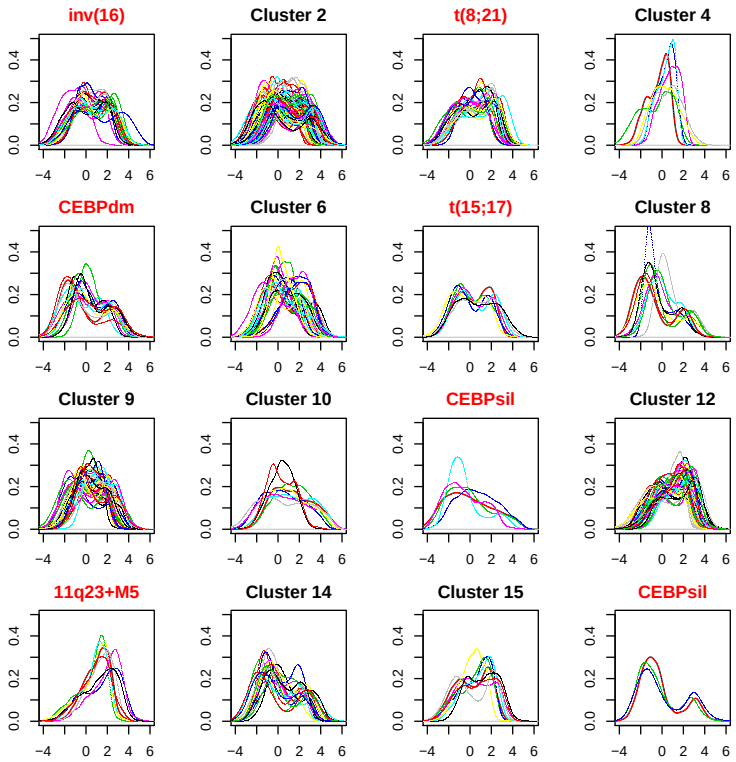

FIG 5. *HELP Methylation density profiles of all* 344 *patients stratified by the methylation clustering result, "M", of Section 7.*

TABLE 6
Cross-classification of the RPMM and HOPACH clustering results and tissue type.

HOPACH

Class	Blad	Bl	Br	Cerv	Inf bl	HN	Kid	Lung	Plac	Pleu	Sm int
1		30				1					
2					55						
3			10								
4			2	1		10					
5	5			2			6	53		18	5
6									1		
7									16		
8									1		
9									1		

RPMM

Class	Blad	Bl	Br	Cerv	Inf bl	HN	Kid	Lung	Plac	Pleu	Sm int
000	3						2	12		8	3
0010								19		5	
0011								20		2	1
0100	2		2	1			4	2		2	1
01010			1			4					
0101100						3					
0101101						3					
010111				2							
01100									1	1	
01101									5		
0111									13		
1000					3						
100100					2						
100101					4						
1001100					3						
1001101					4						
100111					5						
101					34						
1100		18									
1101		12									
11100			5								
11101			3								
1111			1			1					

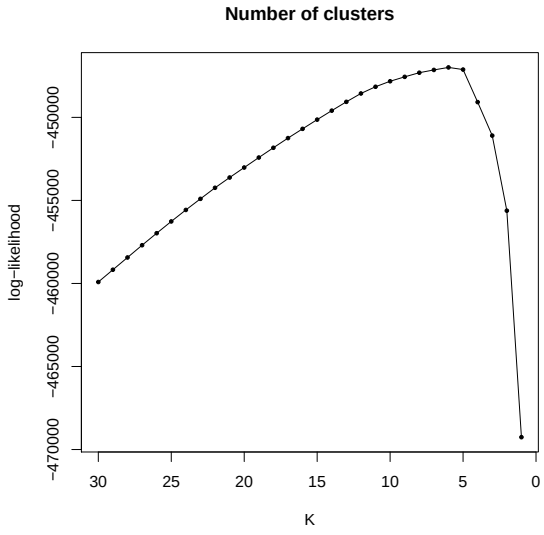

FIG 6. *Plot of log-likelihood values versus numbers of clusters for the normal tissue methylation data of Section 8.2.*