Modeling Exposures for DNA Methylation Profiles

Kimberly D. Siegmund,[1] A. Joan Levine,[1] Jing Chang,[1] and Peter W. Laird[2]

[1]Department of Preventive Medicine and [2]Norris Cancer Center and Departments of Surgery and Biochemistry and Molecular Biology, Keck School of Medicine, University of Southern California, Los Angeles, California

Abstract

We extend the finite mixture model to estimate the association between exposure and latent disease subtype measured by DNA methylation profiles. Estimates from this model are compared with those obtained from the simpler two-phase approach of first clustering the DNA methylation data followed by associating exposure with disease subtype using logistic regression. The two models are fit to data from a study of colorectal adenomas and are compared in a simulation study. Depending on the analytic approach, we obtain different estimates of the odds ratio (OR) and its 95% confidence interval (95% CI) for the association of RBC folate and DNA methylation subtype in colorectal adenomas (OR, 0.31; 95% CI, 0.08-1.26 from the extended finite mixture model; OR, 0.44; 95% CI, 0.15-1.28 from the two-phase approach; n = 58 case subjects). Although our results could be a chance occurrence due to fluctuations from small sample size, we did a simulation study using larger samples and found that differences between the two approaches emerge when there is noise in the cluster analysis. In the naive two-phase approach, the estimate of the OR is biased towards the null, and its SE is underestimated when there is error in the cluster assignment. Estimates from the extended mixture model are unbiased and have the correct SE estimate but may require larger sample sizes for convergence. Thus, when the clusters are not identified with certainty, the extended mixture model is preferred for valid estimation of the OR and CI. (Cancer Epidemiol Biomarkers Prev 2006;15(3):567–72)

Introduction

Today, researchers study tumor heterogeneity at the molecular level using new high-throughput technologies. Molecular features, such as gene expression, immunohistochemical tumor marker expressions (protein abundances), or DNA methylation, can all be measured. These fingerprints allow investigators to search for novel disease subgroups based on molecular characteristics. Once novel subgroups are identified, they must be validated. External validation is achieved by associating the subgroups with respect to risk factors or outcomes. Several articles have dealt with associating novel classes with outcome (1, 2). We focus on the opposite, treating the novel disease subgroup as the outcome and associating these outcomes with hypothesized risk factors. Thus, we are looking for factors related to the etiology of the subgroups. We use as an example data on folate and DNA methylation subtypes in colorectal adenomas.

DNA methylation is an enzymatic modification of DNA frequently found to be abnormally distributed in cancer. The underlying etiology of abnormal methylation in tumors is currently unknown as is the best measure of tumor methylation class. Several methods have been used to categorize tumors into groups based on DNA methylation profiles. Probably the best known of these is the use of a gene panel to identify tumors with multiple methylated genes (3). This definition uses discrete measurements of DNA methylation; a site is either methylated or unmethylated. In our study, we propose to use a panel of quantitative markers to identify tumor subgroup using cluster analysis.

We compare and contrast two methods for defining DNA methylation subgroups using cluster analysis for etiologic studies. We use both methods to assess the hypothesis that low folate availability is associated with more widespread abnormal methylation in colorectal adenomas. To test this hypothesis, we model folate as the exposure and tumor subgroup as the outcome, where tumors are clustered into groups defined by DNA methylation profiles. The primary variable of interest is the association between folate and tumor subtype.

A standard method to classify samples based on DNA methylation profiles is cluster analysis. To validate the results, one might correlate folate with the newly identified disease subgroups. This describes a two-phase analysis, first clustering the tissue samples into novel disease subgroups followed by a correlation of folate levels with cluster assignment. However, inherent in clustering is the uncertainty in group assignment, and this uncertainty is ignored by this two-phase analysis. A more appropriate analysis would be to take the uncertainty in cluster assignment into account when associating folate with DNA methylation subgroup. This describes a pathway where low folate is associated with a certain likelihood of having aberrant DNA methylation profiles. We fit this second model using an extension to model-based cluster analysis.

Model-based clustering has been proposed for identifying disease subgroups using DNA methylation data (4) and gene expression data (5-8). Furthermore, traditional mixture models have been extended to incorporate clinical covariates (7, 8). We focus on one approach considered by McLachlan et al. (7). Although their focus is on the classification of samples to classes, ours is on the estimation of the association between exposure (e.g., folate level) and disease subgroup defined by classes of DNA methylation profiles. In related articles that have focused on estimation (9-11), results are given for the one-phase model but are not compared with estimates obtained from a two-phase approach. We apply both approaches to a data set on colorectal adenomas and compare results. In a simulation study, we compare the bias and efficiency of the two approaches, showing when they are similar and when they differ.

Received 9/13/05; revised 12/7/05; accepted 1/6/06.

Grant support: NIH grant CA097346 and NIEHS grants 5P30 ES07048 and R21 ES011672.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Note: Supplementary data for this article are available at Cancer Epidemiology Biomakers and Prevention Online (http://cebp.aacrjournals.org/).

Requests for reprints: Kimberly D. Siegmund, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, 1540 Alcazar Street, CHP 220, Los Angeles, CA 90089. Phone: 323-442-1310; Fax: 323-442-2349. E-mail: kims@usc.edu

doi:10.1158/1055-9965.EPI-05-0717

Materials and Methods

Colorectal Adenoma Data. The data reported are from a study of colorectal adenomas (12). Cases were individuals diagnosed for the first time with adenomas confirmed by histology. Circulating RBC folate was measured from blood samples obtained for laboratory analysis. DNA methylation was measured using the MethyLight technology as a percent methylated reference (PMR; ref. 13). Measurements were obtained for one CpG region in each of 10 genes: *APC*, *MLH1, MGMT, CDKN2A, PTGS2, ESR1, MYOD1, TIMP3, MTHFR, CALCA*. Measurements for each gene were standardized on the natural log scale using ln(PMR + 1). The data are quantitative, with some loci having an excess of zeros. In previous work, we showed that under a wide range of conditions, one can cluster the log-transformed PMR values reasonably well using normal mixture models (4).

We present results using a subset of nine CpG regions. We chose to exclude *MTHFR* from the analyses because its DNA methylation level did not discriminate between adenoma and adjacent normal tissue. By excluding *MTHFR*, the results from the two analytic approaches showed more differences than had *MTHFR* been included. Thus, we can better highlight the differences that can arise using the two approaches.

We measured DNA methylation on 218 adenoma samples from 132 subjects. The measured RBC folate was available on a subset of 63 subjects (48%). For simplicity, we restrict our analysis to individuals with measured RBC folate and measured DNA methylation in their adenoma tissue at all 10 genes (n = 58 case subjects). For analysis, we select one adenoma at random from individuals having multiple adenomas. In a more comprehensive analysis, the methods would be extended to include samples with missing methylation data, missing covariate data, or multiple tissue samples from the same individual. However, that is beyond the scope of this article.

Statistical Methods. We analyze the colorectal adenoma data using two different approaches. In the first approach, we fit a standard mixture model to classify adenomas into disease subtypes based on their DNA methylation profile. These disease subtypes are then associated with folate level to determine if low folate levels are associated with a subgroup of tumors showing abnormal DNA methylation. This is a two-step approach for characterizing the relationship between folate level and DNA methylation subtype. In a second approach, folate levels are incorporated directly into the mixture model, and the association variable is estimated simultaneously with the clustering procedure. We describe the two mixture models, the estimation procedures, and a simulation study designed to explore the generalizability of our results.

Mixture Model. We assume that after log transformation, the distribution of DNA methylation data follows a mixture of normal distributions. For a review see McLachlan and Basford (1988) and Fraley and Raftery (14). Let $\mathbf{y}_i$ denote the vector of p DNA methylation measurements for sample i ($i = 1, \ldots, n$). We call the vector of measurements the DNA methylation profile for the sample. In our analysis, p = 9, the number of CpG regions selected for cluster analysis. The methylation profiles are considered to be from a mixture of K disease subtypes. The variable denoting to which subtype a sample belongs is given by $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iK})$, where c_{ik} is an indicator denoting membership in the kth subgroup. The probability a sample belongs to the different subtypes is given by $\pi_{i1}, \ldots, \pi_{iK}$. The mixture distribution is written as

$$f(\mathbf{y}_i;\theta) = \sum_{k=1}^{K} \pi_{ik} f(\mathbf{y}_i \mid c_{ik} = 1, \mu_k, \Sigma_k), \qquad (1)$$

where $f(\mathbf{y}_i|c_{ik} = 1, \mu_k, \Sigma_k)$ is a p-dimensional normal distribution for subgroup k, with mean vector μ_k and covariance matrix Σ_k. The average methylation values in μ_k can vary across loci. This permits variable associations of folate level with DNA methylation at individual loci. A discussion of different Σ_k is given by Fraley and Raftery (14). The variable θ denotes the mean vectors, μ_k, the covariance matrices, Σ_k, and the mixing proportions, π_k for $k = 1, \ldots, K$.

Extended Finite Mixture Model. In the extended finite mixture model, we model our latent disease subtype as a function of exposures. In our data set, there is only one exposure of interest, level of circulating RBC folate. However, in general, there can be any number of exposures; thus, our formulae are written for the more general situation. Let $\mathbf{x}$ be a vector of q exposures and $\pi_{ik}(\mathbf{x}_i)$ the probability that sample i belongs in disease subtype k given exposures $\mathbf{x}_i$ [$\pi_{ik}(\mathbf{x}_i) = \Pr(c_{ik} = 1|\mathbf{x}_i)$]. Then $\pi_i(\mathbf{x}_i) = [\pi_{i1}(\mathbf{x}_i), \ldots, \pi_{ik}(\mathbf{x}_i)]$ is a K-dimensional variable vector and $\text{logit}[\pi_i(\mathbf{x}_i)] = (\ln[\pi_{i2}(\mathbf{x}_i)/\pi_{i1}(\mathbf{x}_i)], \ldots, \ln[\pi_{iK}(\mathbf{x}_i)/\pi_{i1}(\mathbf{x}_i)])$ is a $(K - 1)$-dimensional vector. The extended mixture model is

$$f(\mathbf{y}_i;\theta) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) f(\mathbf{y}_i \mid c_{ik} = 1, \mu_k, \Sigma_k).$$

The vector of probabilities $\boldsymbol{\pi_i}(\mathbf{x}_i)$ is fit using polytomous logistic regression,

$$\text{logit}(\pi_i(\mathbf{x}_i)) = \alpha_c + \beta_c \mathbf{x}_i,$$

where α_c is a $(K - 1)$-dimensional variable vector and β_c is a $(K - 1) \times q$-dimensional variable matrix. In the simplest situation of one exposure and two disease subtypes, there is only one variable in α_c and one in β_c. Then, the variable β_c is interpreted as the log odds ratio (OR) of the abnormal subtype for a 1-unit increase in exposure. At this point, we note that the clusters identified by the extended mixture model are conditioned on exposure. Because of this conditioning, the model detects disease subtypes related to exposure so that the association of exposure with clustering is not an *independent* validation of the clustering results.

One could further condition the normal distributions $f(\mathbf{y}_i|c_{ik} = 1, \mu_k, \Sigma_k)$ on the exposures $\mathbf{x}$ as proposed by McLachlan et al. (7), but we do not do this. We introduce the exposure as a predictor of the disease subtype probabilities only. Thus, we assume that conditional on disease subtype, DNA methylation of individual loci no longer depends on folate values.

Estimation. We fit the normal mixture model to the colorectal adenoma data using the function MCLUST in SPLUS version 6.1 (15) specifying equal spherical ($I\sigma^2$) or unequal spherical ($I{\sigma_k}^2$) variance structures. More flexible variance structures cannot be fit due to the limited sample size. The function MCLUST was downloaded from http://www.stat.washington.edu/fraley/mclust (14). The program EMMIX is also available for fitting mixtures of normal distributions (16). The number of components for the mixture model is selected using the Bayesian information criterion [BIC = $-2 \times$ log-likelihood + number of variables $\times$ ln(number of observations)]. The model with the lowest BIC value is considered to have the best fit. Convention states that differences in BIC values that are <2 denote weak evidence for model differences, and differences that are >6 denote strong evidence (17). When faced with BIC values that differ by <2, we opted for the more parsimonious model. After selecting the best-fitting model, adenomas are assigned to the subgroup yielding the greatest posterior probability. The

assigned subgroup is then associated with RBC folate (on the natural log scale) using standard logistic regression.

We calculate the average RBC folate level for each cluster using a weighed average. The (log-transformed) folate level for each subject is multiplied by the probability that they belong to the given subgroup; these terms are summed over all individuals and divided by the sum of the probabilities of belonging to the different subgroups. The exponential of the weighed average (geometric mean) estimates the average folate level for the subgroup.

The mixture model and the extended mixture model are fit using the expectation-maximization algorithm (18). In the two-phase approach, the data are clustered using the mixture distribution in Eq. (1) before assigning samples to subtypes and associating the subtypes with the level of RBC folate. For the one-phase approach, the cluster outcome is associated with RBC folate directly from the extended mixture model. We assume $\Sigma_k = I$ but investigate the robustness of our model to misspecification of the variance matrix. The SE for the association variable is computed using the observed information matrix as described by Louis (19). These models are programmed using the C++ language. We consider the expectation-maximization algorithm converged when the increase in the observed data likelihood is $<1 \times 10^{-5}$. As starting values, we used both a random assignment of class membership and true class assignment, obtaining the same results each time. When the algorithm did not converge, it did not converge for either starting value. The code is available from the first author upon request.

Simulation Study. We conduct a simulation study to evaluate the bias and SE estimate for the variable associating folate with disease subtype. We assume a disease with two subtypes denoting normal and abnormal DNA methylation. The association variable of interest is the coefficient β from a logistic regression model (β = ln OR). We evaluate the effect of different model parameters and sample size on our ability to estimate β. We vary the strength of association between exposure and disease subtype, the proportion of samples assigned to each subtype, and the difference in mean methylation level between the two subtypes. We simulate data when $\beta = -1$ and -2. For $\beta = -1$, a one-SD decrease in folate (on the ln scale) results in a 2.7-fold increased odds of having abnormal DNA methylation. For $\beta = -2$, it corresponds to a 7.4-fold increased odds. The value -1 is chosen to reflect the association we find in our data. The value -2, although strong for an epidemiologic study, shows the affect that the strength of association has on the properties of the β estimate. Log-transformed folate levels are generated using a random normal distribution with variance one. The mean folate level is selected to control the proportion of observations assigned to abnormal methylation subgroup ($\sim$23% or 50%).

Once the subgroup is known, we generate the DNA methylation data. First, we consider the simple scenario where the average DNA methylation level is the same at each of the nine genes and is independent of all other genes within the subgroup. Later, we allow for correlation among genes within a single disease subtype. In the first simulation setting, DNA methylation level is generated from a multivariate normal distribution, $N_9(\mu_k, I)$, $k = 1, 2$ for the two classes. The larger the distance between the mean methylation levels for the two clusters, μ_1 and μ_2, the more distinct the disease subtypes. We assume differences of 0.75, 1.0, and 1.5 in the log methylation value. The discrimination between the two groups is greatest in the last scenario, where the difference in means is greater than the SD of the measures in each subgroup. The other two scenarios have a smaller signal that results in higher uncertainty in cluster assignment. The second and third simulation settings allow us to compare estimates of association from the two mixture-model approaches under misspecification of the variance matrix. In the second simulation setting, we generate data under three models that do not assume conditional independence among loci within disease subtype. In the first, all loci share a pair wise correlation of 0.1. We call this the exchangeable correlation design. In the second and third models, we have two subsets of loci, one size four and the second size five, where pairs of loci from the same subset share a constant pairwise correlation and pairs of loci from different subsets are uncorrelated. This describes a diagonal correlation structure for two gene subsets each having an exchangeable correlation design. We consider correlations of 0.1 and 0.2 for the two exchangeable correlation designs in the diagonal correlation structure. For the third simulation setting, we consider a different type of variance misspecification; we assume conditional independence of loci within subgroup but a higher variance in one subgroup than the other, $\Sigma_1 = I$ and $\Sigma_2 = 2I$.

Results

Cluster Colorectal Adenoma Tissue

Two-Phase Approach. Using BIC for model selection, we select two clusters under the equal spherical variance structure as our best model (Supplementary Fig. S1). This model shows strong evidence for the existence of more than one cluster; only weak evidence favors it over a two-cluster model with unequal spherical variance structure. Using the more parsimonious model, 32 subjects are assigned to one subgroup and 26 to the other. Table 1 shows the average methylation values for each gene by disease subgroup. The profiles for the mean DNA methylation levels show that methylation is higher in subgroup one compared with subgroup two for all CpG regions except *CALCA*. For *CALCA*, there is very little difference in average methylation value between the two subgroups.

Using logistic regression we find that RBC folate (on the ln scale) is inversely associated with DNA methylation subgroup [OR, 0.44; 95% confidence interval (95% CI), 0.15-1.28]. The CI includes 1, suggesting that the association is not statistically significant. However, as expected, the subgroup with the higher folate intake shows less DNA methylation. The geometric mean levels of folate are 267 and 218 ng/dL in the clusters with low and high DNA methylation profiles, respectively.

One-Phase Approach. Assuming the same two-subgroup model, we estimate a slightly stronger inverse association between RBC folate level and DNA methylation profile (OR, 0.31; 95% CI, 0.08-1.26). The model coefficients (ln OR estimates) for the one-phase and two-phase approaches are compared in Table 2. The SE estimate is greater in the

Table 1. Average DNA methylation measurements in colorectal adenoma tissue by disease subtype identified using a normal mixture model (n = 58 case subjects)

Gene	Subtype 1 (n = 32)	Subtype 2 (n = 26)	Difference
APC	1.08	−0.12	1.20
MLH1	0.29	−0.21	0.50
MGMT	0.45	−0.14	0.58
CDKN2A	0.74	−0.60	1.34
PTGS2	0.20	−0.12	0.31
ESR1	0.51	−0.22	0.73
MYOD1	0.79	−0.65	1.45
TIMP3	0.63	−0.37	1.00
CALCA	−0.05	0.06	−0.11
Average (overall)	0.52	−0.26	0.78

NOTE: DNA methylation is measured as PMR. For each gene, measurements are log transformed [ln(PMR + 1)] and standardized across all samples.

Table 2. Coefficient and standard error estimates for the association of RBC folate with abnormal DNA methylation subtype in a set of colorectal adenomas using the mixture model with post hoc validation (two phase) and extended mixture model (one phase) approach (n = 58 case subjects)

Approach	$\hat{\beta}$	SE($\hat{\beta}$)	P	Total uncertainty
Two phase	−0.81	0.54	0.14	3.7
One phase	−1.16	0.71	0.10	3.8

one-phase than the two-phase approach. The total uncertainty, a measure describing uncertainty in the assignment of disease subtype, is also higher in the one-phase approach. This could be a chance occurrence due to small sample size as we would expect a lower total uncertainty from the model that incorporates exposure information in the estimation of disease subtypes.

The one-phase approach is not an independent validation of the clustering results as the cluster probabilities are conditioned on the observed exposures. A result of this conditioning is illustrated by three samples that are assigned to different subgroups using the one-phase and two-phase approaches. Two samples with high folate values (448 ng/dL, 89th percentile and 736 ng/dL, 98th percentile) are assigned to the subgroup that has the higher average folate measure using the one-phase approach. One sample with low folate (129 ng/dL, 11th percentile) is assigned to the subgroup with the lower average folate measure. These samples are assigned to the opposite subgroups by the mixture model that did not incorporate information on folate.

In summary, the variable estimates and their SEs differ depending upon the modeling approach we use (OR, 0.44; 95% CI, 0.15-1.28 two-phase analysis versus OR, 0.31; 95% CI, 0.08-1.26 one-phase analysis). We also find a difference in the measure of uncertainty of group assignment. We investigate the cause of these differences in a simulation study.

Simulation Study. The models we consider result in univariate correlations of DNA methylation with folate levels ranging from −0.11 to −0.37. These correlations are similar to those observed in our set of colorectal adenomas. Overall, we find that when the disease subtypes have a large separation in the average DNA methylation values at each locus, the two approaches give similar results. However, when the separation is less distinct, there is measurement error in the outcome (disease subtype) that is ignored by the two-phase model. This results in biased estimates of association and low coverage probabilities that are not seen in the one-phase approach when the association variable is modeled jointly with the clustering of disease subtypes.

Table 3 compares the bias and SE of the estimate of β under the two-phase and one-phase approaches for a sample size of 200 subjects. For the two-phase approach, the average bias in the estimate of β increases as the mean methylation levels for the two subgroups get closer together; the bias is towards underestimation of the true regression coefficient. At the same time, the mean SE estimate decreases. For the one-phase approach, the average bias of the regression coefficient seems to increase as the distance between average methylation values in the two subtypes decreases; however, this increase is due to a slight skewness in the estimate of β in the simulation. Overall, there is no increase in the median estimate of bias. In contrast to what was observed for the two-phase approach, the mean SE estimate of β increases as the distance between the two clusters decreases. This accurately reflects the higher uncertainty (measurement error) in the cluster assignment.

Table 3. Bias and standard error for the mixture model with post hoc validation and extended mixture model (n = 200 observations; 500 replicates)

Cluster distance*	β†	Average frequency in group 2‡ (%)	Average bias	Median bias	SD($\hat{\beta}$)	Average SE($\hat{\beta}$)	Empirical coverage 95%	Average total uncertainty§
Mixture model with validation (two phase)								
1.5	−1	50	−0.02	−0.01	0.20	0.19	95.0	2.4
		22	−0.01	0.01	0.23	0.22	95.2	1.9
	−2	50	0.04	0.07	0.30	0.28	92.0	2.4
		23	0.05	0.07	0.33	0.32	93.0	1.9
1.0	−1	50	0.15	0.15	0.18	0.18	82.0	12.7
		22	0.16	0.16	0.21	0.21	84.4	10.1
	−2	50	0.50	0.50	0.25	0.23	43.0	12.7
		23	0.52	0.53	0.30	0.27	47.8	10.1
0.75	−1	50	0.31	0.33	0.17	0.17	49.6	23.8
		22	0.35	0.36	0.22	0.20	53.8	18.6
	−2	50	0.88	0.90	0.21	0.20	4.0	23.7
		23	0.92	0.94	0.29	0.23	8.8	18.6
Extended finite mixture model (one phase)								
1.5	−1	50	−0.05	−0.05	0.20	0.20	95.6	2.1
		22	−0.05	−0.04	0.24	0.23	94.8	1.6
	−2	50	−0.08	−0.06	0.32	0.31	94.2	1.6
		23	−0.09	−0.05	0.37	0.35	95.0	1.3
1.0	−1	50	−0.06	−0.05	0.24	0.22	94.6	11.0
		22	−0.06	−0.03	0.28	0.26	95.8	8.7
	−2	50	−0.11	−0.07	0.39	0.37	95.0	8.6
		23	−0.11	−0.05	0.46	0.42	94.8	6.7
0.75	−1	50	−0.07	−0.05	0.29	0.27	95.0	20.5
		22	−0.08	−0.05	0.38	0.33	95.0	16.0
	−2	50	−0.18	−0.10	0.60	0.48	94.6	15.7
		23‖	−0.17‖	−0.07‖	0.70‖	0.56‖	94.8‖	12.2‖

NOTE: Nine CpG regions are simulated for each observation.
*Cluster distance is the difference in average methylation level between disease subtypes for each CpG region.
†β is the ln OR associating exposure with subgroup having abnormal methylation in a logistic regression model.
‡Group 2 is the subgroup having a DNA methylation defect.
§Average total uncertainty is the average across replicates of the sum of uncertainties from the mixture model.
‖Estimates not obtained for one data set.

We also compare the two approaches in terms of the empirical coverage for the estimate of β. We find the two-phase approach does not attain the proper 95% empirical coverage for settings in which the clusters are not determined with a high level of certainty; the coverage probability decreases as the two clusters become closer together. In addition, the coverage decreases the stronger the association between exposure and outcome ($\beta = -2$ versus -1). In general, the one-phase model attains the proper 95% empirical coverage regardless of the distance between disease subtypes or the association between exposure and outcome. If we decrease the sample size, we find the empirical coverage of the one-phase model is low on occasion (Supplementary Table S1). Interestingly, decreasing the sample size improved the empirical coverage estimates for the two-phase model; however, they are still well below the nominal 95% level (Supplementary Table S1).

Overall, we see that the average total uncertainty is lower for the one-phase than the two-phase approach (Table 3). This can be explained by the fact that information from the exposure variable is being exploited by the one-phase approach. Thus, we see that for the one-phase approach, the average total uncertainty decreases as the disease-exposure association increases. This same phenomenon is not seen for the two-phase model where clustering is carried out without use of the exposure information.

We also compare the bias and SE estimates of the two-phase and one-phase approaches under a misspecified covariance matrix. In general, the one-phase approach shows less bias and better empirical coverage than the two-phase approach (Table 4). This suggests that by using exposure information, the one-phase model was more robust for estimating β under a misspecified covariance structure. A small amount of correlation among genes within a disease subtype led to a slight overestimate of the association variable for the two-phase approach. The bias was less noticeable if the correlation was not among all loci but among two smaller subgroups of loci. When the genes were independent within disease subtype but the variance differed in the two subtypes, the bias was larger and the empirical coverage worse the more unequal the size of the two disease subgroups. At the same time, more samples were assigned to the subgroup with the larger variance than should have been (data not shown).

Discussion

We present an extended mixture model to study the association between exposures and latent subgroups of disease. In a simulation study, we show that a naive two-phase approach of a cluster analysis followed by association analysis yields estimates similar to a one-phase approach when the clusters are distinct and can be determined with certainty from the methylation data alone. When there is uncertainty in the cluster assignments, the naive approach can lead to underestimates of the association variable and its SE. On the other hand, the extended mixture model provides unbiased estimates of association and valid estimates of precision. The larger average SE estimate from the one-phase approach can be explained by taking uncertainty of the cluster assignment into account when simultaneously estimating the cluster assignment and the association variable. However, this one-phase approach is no longer unsupervised. The latent classes are generated conditional on the association with the exposure variables.

The simulation study investigated a simple scenario where a methylation defect had the same effect on methylation at all CpG regions. In this situation, a repeated-measures analysis of covariance (RMANCOVA) would be a valid method for testing the hypothesis of an association between folate level and DNA methylation level across multiple CpG regions. In fact, for this simulation study, this approach would give smaller Ps than a test of the log OR from the extended mixture model. This might be explained by the fact that the RMANCOVA model can assume that the effects are the same at each locus where the extended mixture model models the effects at each locus separately. In a more realistic situation, we might expect different CpG regions to hypermethylate at different rates under a methylation defect. To consider this, we tried a few simulations where we allowed the difference in mean methylation level between the two groups to vary across the nine genes while holding the overall average difference constant. These scenarios resulted in smaller Ps using the extended mixture model compared with simple RMANCOVA (data not shown). The situation where we would expect the extended mixture model to be superior to the RMANCOVA model is when the association between exposure and methylation go in different directions for different CpG regions. For example, in an earlier study, we found different

Table 4. Bias and standard error for the mixture model with post hoc validation and extended mixture model under misspecification of the correlation structure (n = 200 observations; 500 replicates)

True correlation within group*	Average frequency in group 2† (%)	Estimated average frequency in group 2 (%)	Average bias	Mean SE($\hat{\beta}$)	Empirical coverage 95%	Average total uncertainty‡
Mixture model with validation (two phase)						
$\text{Exch}(0.1)_{9\times9}$	50	50	0.12	0.18	86.2	4.3
	22	24	0.15	0.20	86.0	3.9
$\text{Diag}[\text{Exch}(0.1)_{4\times4}, \text{Exch}(0.1)_{5\times5}]$	50	50	0.07	0.18	93.6	3.5
	22	23	0.06	0.21	91.4	2.9
$\text{Diag}[\text{Exch}(0.2)_{4\times4}, \text{Exch}(0.2)_{5\times5}]$	50	50	0.10	0.18	87.0	4.2
	22	23	0.13	0.21	88.0	3.8
Extended finite mixture model (one phase)						
$\text{Exch}(0.1)_{9\times9}$	50	50	0.05	0.19	91.4	4.0
	22	24	0.06	0.22	92.0	3.5
$\text{Diag}[\text{Exch}(0.1)_{4\times4}, \text{Exch}(0.1)_{5\times5}]$	50	50	0.01	0.19	97.0	3.1
	22	23	0.00	0.23	95.0	2.6
$\text{Diag}[\text{Exch}(0.2)_{4\times4}, \text{Exch}(0.2)_{5\times5}]$	50	50	0.02	0.20	92.4	3.8
	22	23	0.04	0.23	94.6	3.4

NOTE: Nine CpG regions are simulated. At each CpG region, the difference in average methylation value between the two subgroups is 1.5. The ln OR associating exposure with subgroup having abnormal methylation in a logistic regression model is -1. All models are fit assuming an independence variance matrix (I).

*True correlation structures within group are as follows: (a) exchangeable with correlation 0.1 [$\text{Exch}(0.1)_{9\times9}$]. ($b$) Two subgroups of loci with exchangeable correlation structure in each subgroup, correlation = 0.1. Subgroups of loci are size 4 and size 5 {$\text{Diag}[\text{Exch}(0.1)_{4\times4}, \text{Exch}(0.1)_{5\times5}]$}, and ($c$) same as ($b$) but with correlation of 0.2 in each subgroup of loci {$\text{Diag}[\text{Exch}(0.2)_{4\times4}, \text{Exch}(0.2)_{5\times5}]$}.

†Group 2 is the subgroup having a DNA methylation defect.

‡Average total uncertainty is the mean sum of uncertainties from the mixture model.

DNA methylation profiles in non–small cell and small cell lung cancer (20). In a subset of CpG regions DNA methylation was higher in non–small cell than in small cell lung cancer. In another subset of regions, the reverse was true. If we were to find an exposure that correlated with lung cancer cell type, we would have a situation where the exposure and methylation might also be in different directions for different CpG regions. In such extreme scenarios, the RMANCOVA model will not find any association with exposure unless one was to limit the analysis to CpG regions with associations in the same direction.

Many variables will affect the comparison of the extended mixture model and RMANCOVA. Another variable that we found important is the proportion of observations in the different clusters. The more imbalanced the clusters, the more likely the extended mixture model will compute smaller Ps (data not shown). In practice, one might try all of the methods described to try and understand the complex relationships in the data.

One simplification we made in our real data example was to select one adenoma at random from patients with multiple adenomas. This could introduce bias to a cluster analysis if individuals having multiple adenomas tended to have higher or lower methylation values on average than individuals having only one adenoma. Then the clusters we identify might simply be related to the number of adenomas in the patient. In our study, we found no differences in the average methylation values at any of the nine loci studied between subjects having one or multiple adenomas. Furthermore, having multiple adenomas was not associated with our final disease subgroups.

Other articles have considered how to correlate clinical covariates with gene expression profiles (7, 21). Shannon et al. (21) present an approach that does not require a fixed number of clusters. They correlate the distance matrix of gene expression profiles with a distance matrix of clinical covariate profiles using a Mantel Statistic. The distance matrix is computed by measuring the distance between all pairs of observations (gene expression profiles or clinical covariate profiles). Their method allows valid statistical inference correlating the two types of data. McLachlan et al. (7) fit a model with a distinct number of disease subtypes similar to ours. However, their focus is on the classification of tissue samples and not variable estimation.

Finally, a popular method to validate clusters using an external criterion is the adjusted Rand index (22). This index compares two classification schemes by comparing how pairs of observations are classified by the two schemes. The index equals one when there is perfect agreement and zero when it equals its expected value under random partitioning. The index requires the external variable to be categorical, which may be limiting when using exposures for validation in studying disease etiology. In addition, it does not take cluster uncertainty into account.

In summary, we find the mixture model approach is a useful tool for identifying potentially homogeneous subgroups of disease. When exposure data are available, the mixture model can be extended to condition on the exposure data. When the results obtained from the model resemble those obtained from a two-phase analysis, it suggests that the clusters are distinct based on the outcome data alone. In this situation, the exposure might be used to validate the identified clusters. When the results differ, only the one-phase approach provides unbiased estimates of association with valid estimates of precision.

References

1. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.
2. Shi T, Seligson D, Belldegrun AS, et al. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol 2005;18:547–57.
3. Toyota M, Ahuja N, Ohe-Toyota M, et al. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci U S A 1999;96:8681–6.
4. Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. Bioinformatics 2004;20:1896–904.
5. Yeung KY, Fraley C, Murua A, et al. Model-based clustering and data transformations for gene expression data. Bioinformatics 2001;17:977–87.
6. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 2002;18:413–22.
7. McLachlan GJ, Change SU, Mar J, et al. On the simultaneous use of clinical and microarray expression data in the cluster analysis of tissue samples. In Y.P. Chen, editors. 2nd Asia-Pacific Bioinformatics Conference (APBC2004): Conferences in Research and Practice in Information Technology. vol. 29. Dunedin (New Zealand): the Australian Computer Society, Inc.; 2004. p. 167–71.
8. McLachlan GJ, Chang SU. Mixture modelling for cluster analysis. Stat Methods Med Res 2004;13:347–61.
9. Muthen B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics 1999;55:463–9.
10. Lin H, McCulloch CE, Turnbull BW, et al. A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations. Stat Med 2000;19:1303–18.
11. Lin H, Turnbull BW, McCulloch CE, et al. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal PSA readings and prostate cancer. J Am Stat Assoc 2002;97:53–65.
12. Haile RW, Witte JS, Longnecker MP, et al. A sigmoidoscopy-based case-control study of polyps: macronutrients, fiber and meat consumption. Int J Cancer 1997;73:497–502.
13. Uhlmann K, Rohde K, Zeller C, et al. Distinct methylation profiles of glioma subtypes. Int J Cancer 2003;106:52–9.
14. Fraley C, Raftery AE. Mclust: software for model-based cluster analysis. Journal of Classification 1999;16:297–306.
15. SPLUS. http://www.insightful.com/products/default.asp, version 6.1; 2002.
16. McLachlan GJ, Peel D, Basford KE, et al. The EMMIX software for the fitting of mixtures of normal and t-components. J Stat Software 1999;4:1–14.
17. Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc 1995;90:773–95.
18. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Statist Soc Ser B 1977;39:1–38.
19. Louis TA. Finding the observed information matrix when using the EM algorithm. J Roy Statist Soc Ser B 1982;44:226–33.
20. Virmani AK, Tsou JA, Siegmund KD, et al. Hierarchical clustering of lung cancer cell lines using DNA methylation markers. Cancer Epidemiol Biomarkers Prev 2002;11:291–7.
21. Shannon WD, Watson MA, Perry A, et al. Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. Genet Epidemiol 2002;23:87–96.
22. Hubert L, Arabie P. Comparing Partitions. Journal of Classification 1985;2:193–218.