RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Semiparametric Tests for Identifying Differentially Methylated Loci With Case–Control Designs Using Illumina Arrays

Yong Chen,[1]* Yang Ning,[2] Chuan Hong,[1] and Shuang Wang[3]

[1]*Division of Biostatistics, School of Public Health, The University of Texas, Houston, Texas, United States of America;* [2]*Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada;* [3]*Department of Biostatistics, Mailman School of Public Health, Columbia University, New York City, New York, United States of America*

Received 14 June 2013; Revised 13 September 2013; accepted revised manuscript 17 October 2013.
Published online 3 December 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21774

ABSTRACT: DNA methylation plays an important role in the development of many types of cancer. Identifying differentially methylated loci between cancer and normal patients is one of the central tasks to understand the contributions of the methylation process on cancer development. Through investigation of the methylation measurements generated by the Illumina methylation arrays, we notice that the methylation measurements of the cancer and normal groups could differ not only in means but also in variances. Therefore, we propose a generalized exponential tilt model to capture the differences in both means and variances between the cancer and normal groups. We derive the semiparametric tests to obtain model robustness. Through simulation studies, we demonstrate the feasibility of the proposed tests and a much improved power of the proposed tests than that of the *t*-test and the regression-based tests when the cancer and normal groups are different in variances only or in both means and variances. Hence the proposed tests can serve as useful complements to the standard tests that only test differences in means. We also illustrate the proposed methods by applying to a real methylation data from a recent study on ovarian cancer where the proposed methods identified additional methylation loci that were missed by the existing method.

Genet Epidemiol 38:42–50, 2014.

KEY WORDS: Casecontrol design; composite likelihood; conditional likelihood; exponential tilt model; methylation data; pseudolikelihood; semiparametric model

Introduction

The importance of DNA methylation changes on cancer development is now well understood. Many research efforts have been devoted to identifying differentially methylated loci between cancer and normal patients to search for potential contributions of the methylation process on cancer development [Kalari and Pfeifer, 2010; Kerkel et al., 2010; Shen et al., 2013; Slater et al., 2013]. Among several existing high-throughput platforms, the Illumina Infinium methylation arrays have been commonly used to generate genome-wide DNA methylation data. Illumina arrays are based on genotyping bisulfite(BS)-converted DNA, where unmethylated cytosines are converted to uracils and methylated cytosines are protected and remain cytosine. The methylation level of a CpG site, which is called a β-value, is calculated as the average of approximately 30 replicates (with approximately 30 beads per site per sample) of the quantity $\max(M, 0)/(\max(U, 0) + \max(M, 0) + 100)$. Here U (or M) is the fluorescent signal from an unmethylated (or methylated) allele on a single bead. A maximum between signal intensity and 0 is chosen to compensate for negative signals due to background subtraction. The constant 100 is to regularize β-values when both M and U values are small [Bibikova et al., 2006]. This β-values range continuously from 0 (unmethylated) to 1 (completely methylated). Existing methods to select differentially methylated loci such as the *t*-test or the regression-based methods or the method we developed [Wang, 2011] that considers three different components rely on some distributional assumptions that may not be true and mainly focus on testing difference in means between cancer and normal groups. Several recently developed nonparametric tests although are free of distributional assumptions, still test for differences in mean ranks [Chen et al., 2013; Huang et al., 2013]. More recently, several studies and our own group have found that there are differences in methylation variations in different groups [Gervin et al., 2011; Hansen et al., 2011]. Such difference in variations cannot be detected by the existing methods including the method we recently developed.

To account for the potential limitations of the existing methods, we propose to use the parsimonious two-sample generalized exponential tilt model with two parameters [Anderson, 1979; Qin, 1998]. The motivation of the generalized

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Yong Chen, Department of Biostatistics, School of Public Health, The University of Texas, 1200 Pressler Street, Houston, TX 77030, USA. E-mail: yong.chen@uth.tmc.edu.

two-parameter exponential tilt model is its ability to detect the differences in both means and variances between two groups. The exponential tilt model assumes that the log density ratio of the distributions of the methylation measures of the cancer and normal samples is a linear function of the methylation values. Compared to a particular parametric model, an attractive feature of the exponential tilt model is that the distributions are modeled nonparametrically except for a parametric "exponential tilt" that is used to relate one distribution to the other. This is an important property for modeling DNA methylation data because methylation values at different loci may have different distributions [Wang, 2011].

Under the generalized exponential tilt model, testing for the difference between two distributions is equivalent to testing for the two corresponding parameters being zeros. The standard inference procedure for these parameters is based on the empirical likelihood function [Owen, 2001], which eliminates the reference distribution (i.e. the density function of methylation values in the normal group) by a profiling procedure [Anderson, 1972; Fokianos, 2008; Qin, 1998]. However, our simulation studies suggested an inflated Type I error with such empirical likelihood-based method for studies with small sample sizes. Although a permutation method is able to correctly adjust for the Type I errors in such situations, it is usually computationally prohibitive in practice for high-throughput data where thousands of loci are considered. Therefore, we propose a class of pseudolikelihood-based tests where the reference distribution is eliminated by the conditioning technique so that the impact of the unknown reference distribution is minimized [Kalbfleisch and Sprott, 1970]. Such pseudolikelihood-based tests are shown to have better controlled Type I errors than the empirical likelihood-based test for small sample size cases. Both the empirical likelihood-based test and the proposed pseudolikelihood-based tests are substantially more powerful than the *t*-test and the regression-based test when the cancer and normal samples are different in variances only or in both means and variances, but are marginally less powerful than the *t*-test when the two groups are different in means only but not in variances. Therefore, the proposed method serves as a useful complement to the current methods. We also applied the proposed methods to the United Kingdom Ovarian Cancer Population Study (UKOPS) data where the proposed methods identified additional methylation loci that were missed by the existing method.

Method

Model Setting

For $k = 1, 2, \ldots, K$, at the kth DNA methylation marker, denote $u_{k1}, \ldots, u_{km_k}$ as the methylation β-values of m_k independent subjects in the normal group, and $v_{k1}, \ldots, v_{kn_k}$ as that of n_k independent subjects in the cancer group where K is the total number of markers. Denote $f_k(\cdot)$ and $g_k(\cdot)$ as the distributions of u_{ki} and v_{kj}, respectively. We consider the following parsimonious semiparametric model. Specifically, at the kth marker, the distributions of the methylation β-values in the normal and cancer groups, $f_k(\cdot)$ and $g_k(\cdot)$, are assumed to be related in the following form

$$\frac{g_k(x)}{f_k(x)} = \exp\{\alpha_k + \beta_{k1} h_1(x) + \beta_{k2} h_2(x)\}, \qquad (1)$$

where the marker specific density function $f_k(\cdot)$ describes the distribution of methylation values in the normal group, referred to as the reference distribution, $h_1(\cdot)$ and $h_2(\cdot)$ are prespecified functions (e.g., $h_1(x) = x$ and $h_2(x) = x^2$, or $h_1(x) = \log(x)$, and $h_2(x) = \log(1 - x)$ depending on the distribution of the DNA methylation measures), β_{k1} and β_{k2} are unknown parameters characterizing the differences between the two distributions $f_k(\cdot)$ and $g_k(\cdot)$ that have different interpretations under specific distributional models (more details later in this subsection), and $\alpha_k = -\log\left[\int \exp\{\beta_{k1} h_1(x) + \beta_{k2} h_2(x)\} f_k(x) dx\right]$ is a normalizing constant for the density function $g_k(x)$. For notation simplicity, we hereafter suppress the index k. We note that unlike any particular parametric model, model (1) allows an unspecified marker specific density function $f_k(\cdot)$, which is an important feature to model methylation data due to heterogeneous distributions across loci.

Model (1) is known as the exponential tilt model and has been considered by many researchers [Anderson, 1972, 1979; Fokianos, 2008; Qin, 1998; Tan, 2009]. Popular parametric models such as normal, gamma, log normal distributions, and a mixture of normal distributions for continuous variables and binomial, poisson, and negative binomial distributions for discrete variables are special cases of the exponential tilt model. For methylation measures, two parametric models are often used, the Beta model and the Normal model with unequal variances.

Model 1 (Beta model). To account for the fact that the methylation values are proportions between 0 and 1, a beta distribution is often used [Kuan et al., 2010; Laurila et al., 2011]. Let $f(\cdot)$ and $g(\cdot)$ be the density functions of two beta distributions with shape parameters (a_1, b_1) and (a_2, b_2), respectively. Some calculations yield

$$\log\{g(x)/f(x)\} = \log\frac{B(a_1, b_1)}{B(a_2, b_2)} + (a_2 - a_1)\log x + (b_2 - b_1)\log(1 - x),$$

where $B(\cdot, \cdot)$ is the beta function. Thus, the beta distribution belongs to the exponential tilt model (1) with $h_1(x) = \log x$, $h_2(x) = \log(1 - x)$, $\beta_1 = a_2 - a_1$ and $\beta_2 = b_2 - b_1$. Testing for equivalence between two beta distributions is equivalent to testing for $\beta_1 = \beta_2 = 0$.

Model 2 (Normal model with unequal variance). A common practice with DNA methylation β-values is to take the logit transformation on β-values and then model the transformed data as normally distributed [Chowdhury et al., 2010; Du et al., 2010]. Let $f(\cdot)$ and $g(\cdot)$ be the density functions of two normal distributions with means and variances (μ_1, σ_1^2) and

(μ_2, σ_2^2), respectively. Some calculations yield

$$\log\{g(x)/f(x)\} = \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_2^2} + \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) + \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)x^2.$$

Thus, the normal distribution belongs to the exponential tilt model (1) with $h_1(x) = x$, $h_2(x) = x^2$, $\beta_1 = \mu_2/\sigma_2^2 - \mu_1/\sigma_1^2$, and $\beta_2 = 1/\sigma_1^2 - 1/\sigma_2^2$. Testing for equivalence between two normal distributions is equivalent to testing for $\beta_1 = \beta_2 = 0$.

The exponential tilt model (1) provides a flexible alternative to detecting differentially methylated loci without being limited to any particular parametric model. At each marker, under the exponential tilt model assumption, testing for differential methylation is equivalent to testing for $H_0 : \beta_1 = \beta_2 = 0$. We note that H_0 is testing for any difference between the two distributions $f(\cdot)$ and $g(\cdot)$ for the normal and cancer groups, whereas the conventional *t*-test or regression-based models focus on the difference in means.

Proposed Tests

To test the null hypothesis of $H_0 : \beta_1 = \beta_2 = 0$, the empirical likelihood can be used. As derived by Qin and Zhang [1997] and Fokianos [2008], the empirical log likelihood [Owen, 2001] can be calculated as

$$\begin{aligned}\log L_e(\alpha, \beta_1, \beta_2) &= -\sum_{i=1}^{m}\log\left[1 + \frac{\rho}{1-\rho}\exp\{\alpha + \beta_1 h_1(u_i) + \beta_2 h_2(u_i)\}\right] \\ &\quad - \sum_{j=1}^{n}\log\left[1 + \frac{\rho}{1-\rho}\exp\{\alpha + \beta_1 h_1(v_j) + \beta_2 h_2(v_j)\}\right] \\ &\quad + \sum_{i=1}^{m}\{\alpha + \beta_1 h_1(u_i) + \beta_2 h_2(v_i)\},\end{aligned}$$

where $\rho = m/N$ and $N = n + m$. Denote $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$, $\alpha_0 = 0$, $\boldsymbol{\beta}_0 = (0, 0)^T$ and the maximum empirical likelihood estimator by $(\tilde{\alpha}, \tilde{\boldsymbol{\beta}}) = (\tilde{\alpha}, \tilde{\beta}_1\tilde{\beta}_2) = \text{argmax}\log L_e(\alpha, \beta_1, \beta_2)$. The empirical likelihood ratio test (ELRT) for H_0 can be constructed as

$$\text{ELRT} = -2\left\{\log L_e(\alpha_0, \boldsymbol{\beta}_0) - \log L_e(\tilde{\alpha}, \tilde{\boldsymbol{\beta}})\right\},$$

whose asymptotic distribution has been shown to be a χ_2^2 distribution by Tan [2009].

For the proposed pseudolikelihood-based tests, the novelty is the elimination of the reference distribution $f(\cdot)$ by a conditioning technique, thus the impact of the unknown reference distribution is minimized [Kalbfleisch and Sprott, 1970]. This is an important feature for maintaining correct Type I errors, especially at low nominal levels. Consider the *i*th observation u_i from the normal group and the *j* th observation v_j from the cancer group. The conditional density of (u_i, v_j) given their order statistics $(t^{(1)}, t^{(2)})$ can be calculated as

$$\begin{aligned}\Pr(u_i, v_j | t^{(1)}, t^{(2)}) &= \frac{f(u_i)g(v_j)}{f(u_i)g(v_j) + f(v_j)g(u_i)} \\ &= \left\{1 + R(u_i, v_j; \beta_1, \beta_2)\right\}^{-1},\end{aligned}$$

where $R(u_i, v_j; \beta_1, \beta_2) = \exp[\beta_1\{h_1(u_i) - h_1(v_j)\} + \beta_2\{h_2(u_i) - h_2(v_j)\}]$ and the derivation of such conditional density is described in the supplementary material. Note that both the reference distribution $f(\cdot)$ and the parameter α are eliminated by conditioning on the order statistic $(t^{(1)}, t^{(2)})$.

By multiplying all the possible pairwise conditional densities, we then construct the pseudolikelihood of the methylation values at this marker for all samples

$$L_p(\beta_1, \beta_2) = \left[\prod_{i=1}^{m}\prod_{j=1}^{n}\left\{1 + R(u_i, v_j; \beta_1, \beta_2)\right\}^{-1}\right]^{2/N}.$$

Note that the product of conditional densities is raised to the power of $2/N$ to form the pseudolikelihood. As we show in the supplementary material, such modification is necessary so that the conventional pseudolikelihood ratio test has a simple χ_2^2 distribution. It is also worth mentioning that the pseudolikelihood $L_p(\beta_1, \beta_2)$ is not a true likelihood function because the conditional densities are multiplied together as if they are independent. In fact, the proposed pairwise pseudolikelihood is a type of composite likelihood where (weighted) marginal or conditional densities are multiplied together to form the composite likelihood [Besag, 1974; Cox and Reid, 2004; Lindsay, 1988; Varin et al., 2011]. For more discussion on the composite likelihood methods, please see the excellent review paper by Varin et al. [2011] and the references therein.

Denote the log pseudolikelihood function as $\log L_p(\boldsymbol{\beta})$, the pseudo score function as $S(\boldsymbol{\beta})$, and the maximum pseudolikelihood estimator as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2) = \text{argmax}\log L_p(\beta_1, \beta_2)$. We can construct the Wald, score and pseudolikelihood ratio tests as follows:

$$\begin{aligned}\text{T}_W &= \hat{\boldsymbol{\beta}}^T\left\{\widehat{\text{var}}(\hat{\boldsymbol{\beta}})\right\}^{-1}\hat{\boldsymbol{\beta}}, \\ \text{T}_S &= S(\boldsymbol{\beta}_0)^T\left[\widehat{\text{var}}\{S(\boldsymbol{\beta}_0); H_0\}\right]^{-1}S(\boldsymbol{\beta}_0), \\ \text{T}_L &= -2\left\{\log L_p(\boldsymbol{\beta}_0) - \log L_p(\hat{\boldsymbol{\beta}})\right\},\end{aligned}$$

where $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ is a consistent variance estimator of $\hat{\boldsymbol{\beta}}$ and $\widehat{\text{var}}\{S(\boldsymbol{\beta}_0); H_0\}$ is a consistent variance estimator of the pseudo score function $S(\boldsymbol{\beta}_0)$ under H_0.

For the limited space, we only describe the pseudolikelihood ratio test T_L and leave the detailed construction of Wald T_W and pseudo score test T_S in the supplementary material. The asymptotic distribution of T_L is generally a weighted sum of χ_1^2 distributions which requires the calculations of the information matrix to obtain estimates of the weights [Chen and Liang, 2010; Liang and Self, 1996]. However, the asymptotic distribution of the pseudolikelihood ratio test takes a simple distribution under H_0 as described by the following theorem.

Theorem 1 Under H_0, the pseudolikelihood ratio test T_L converges to χ_2^2.

An outline of proof is described in the supplementary material. The simple asymptotic distribution in theorem 1 avoids the calculation of the weights and greatly simplifies the computation of P-values.

Asymptotic Power and Design Consideration

The power of different tests in finite samples will be investigated in Simulation Study section. Here we provide a useful result on the asymptotic power of T_W, T_S, and T_L under a sequence of local alternatives, which provides some insights in maximizing the statistical power on experimental design. Specifically, consider a sequence of alternatives:

$$H_a : f(\cdot) = f_0(\cdot), \quad \boldsymbol{\beta} = N^{-1/2}\tau_0,$$

where τ_0 is a vector of constant. Note that under H_a, we do not explicitly specify the local alternative for α, because α is a function of $\boldsymbol{\beta}$ and $f_0(\cdot)$. With LeCam's third lemma [Van der Vaart, 2000], we can establish the following results.

Theorem 2 Under the alternatives H_a, the limiting distributions of T_W, T_S, and T_L are $\chi^2_2((1-\rho)\rho\boldsymbol{\beta}\tau_0^T\Sigma_U\boldsymbol{\beta}\tau_0)$, where Σ_U is a 2×2 covariance matrix calculated as $\mathrm{cov}(h_1(U_i), h_2(U_i))^T$.

A sketch of the proof is given in the supplementary material. Theorem 2 implies that T_W, T_S and T_L have the same asymptotic power under H_a and the power increases as $\boldsymbol{\beta}$ further departs from $(0, 0)^T$. More importantly, given the total number of subjects N, the asymptotic power is maximized when the design is balanced (i.e., $\rho = 0.5$ or $m = n$).

Simulation Study

We conducted simulation studies to evaluate the finite sample performance of the proposed tests and compared to that of the existing tests. The following six tests were considered: the Wald, score and likelihood ratio tests based on the pseudolikelihood (T_W, T_S, and T_L, respectively), the empirical likelihood ratio test (ELRT), the Wald test based on the logistic regression of cancer status on methylation level (Logistic), and the t-test. We note that the exponential tilt model can be represented as a logistic regression of cancer status on $h_1(x)$ and $h_2(x)$ (e.g., $h_1(x) = x$ and $h_2(x) = x^2$), where x is the methylation measure [Qin, 1998]. In this case, testing H_0 that a specific locus is not associated with the cancer status is equivalent to testing coefficients of $h_1(x)$ and $h_2(x)$ being 0. Note that this is different from the conventional regression-based test when only the coefficient of $h_1(x)$ is tested. We considered a relatively large sample case with 100 normal and 100 cancer samples, a moderate sample case with 50 normal and 50 cancer samples, and a relatively small sample case with 20 normal and 20 cancer samples. Three parametric models were considered to generate the methylation data, namely, the beta distribution (model 1), the normal distribution (model 2) and the mixture of two normal distributions (model 3).

The model parameters were carefully chosen based on a real methylation dataset. Specifically, we randomly sampled 500 markers from the 22,951 markers of the United Kingdom Ovarian Cancer Population Study (UKOPS) data, and then calculated the means and standard deviations (sds) of the methylation values at all markers. The median of the 500 means is close to 0.80, and the median of the 500 sds is close to 0.04. Therefore, the shape parameters of the beta distribution in the normal group (model 1) were set at $a_1 = 100$ and $b_1 = 25$ such that the corresponding mean and sd are 0.80 and 0.04, respectively. Similarly, the mean and variance of the normal distribution in the normal group (model 2) were set at $\mu_1 = -1$, and $\sigma_1^2 = 0.5$ based on the *logit-transformed* methylation values of the 500 randomly sampled markers from the UKOPS data. The parameters of the mixture of two normal distributions (model 3), $\pi N(\mu_{11}, \sigma_{11}^2) + (1-\pi)N(\mu_{12}, \sigma_{12}^2)$, were set at $(\pi, \mu_{11}, \sigma_{11}^2, \mu_{12}, \sigma_{12}^2) = (0.6, -3, 0.5, 2, 0.5)$. We denote the standardized difference between the normal and cancer samples as $\Delta = (\mu_2 - \mu_1)/sd_1$ where μ_1 and μ_2 are the means of the normal and cancer samples, and sd_1 is the sd of the normal samples. We also denote the ratio of sds for cancer and normal samples as $r_{21} = sd_2/sd_1$, where sd_2 is the sd of the cancer samples. We evaluated the Type I error and power of the six tests under different scenarios: (1) Type I error scenario ($\Delta = 0$, $r_{21} = 1$): distributions of normal and cancer samples are the same; (2) Power scenario I ($\Delta = 0.6$, $r_{21} = 1$): distributions of cancer and normal samples are different in means only; (3) Power scenario II ($\Delta = 0$, $r_{21} = 0.7$) or ($\Delta = 0$, $r_{21} = 0.9$): distributions of cancer and normal samples are different in sds only; and (4) Power scenario III ($\Delta = 0.3$, $r_{21} = 0.8$) or ($\Delta = 0.2$, $r_{21} = 0.9$): distributions of cancer and normal samples are different in both means and SDs. We conducted 5, 000 simulations for each scenario.

Table 1 summarizes the Type I error and power of the relatively large sample case for the six tests considered when methylation β-values were generated based on the beta distribution (model 1). All tests except the Wald test using the

Table 1. The type I error (×100) and power (×100) for the six tests considered at 5%, 1%, 0.5%, and 0.1% significance levels when methylation β-values were generated based on the beta distribution (model 1). The numbers of cancer and normal samples are (100, 100)

(Δ, r_{21})	Level (%)	T_W	T_S	T_L	ELRT	Logistic	t-test
(0, 1)	5	4.6	4.7	5.1	5.3	3.7	5.4
	1	1.0	0.7	0.8	1.0	0.5	1.1
	0.5	0.5	0.4	0.5	0.6	0.3	0.5
	0.1	0.1	0.1	0.1	0.1	0.0	0.1
(0.6, 1)	5	97.0	97.1	97.2	97.3	96.6	98.5
	1	89.3	90.2	90.7	91.2	88.6	94.2
	0.5	84.1	85.8	86.7	87.5	82.6	91.2
	0.1	68.3	72.6	73.8	75.7	64.7	81.5
(0, 0.7)	5	85.8	88.6	90.2	90.5	84.8	5.1
	1	63.4	67.4	73.8	75.1	55.7	1.3
	0.5	51.0	56.4	65.0	66.3	41.2	0.7
	0.1	27.4	31.5	44.4	46.8	15.8	0.1
(0.3, 0.8)	5	72.7	74.8	76.3	77.1	70.2	65.0
	1	44.2	48.0	52.4	53.7	38.7	39.8
	0.5	34.0	37.1	42.2	43.9	27.7	31.1
	0.1	15.6	18.9	23.9	26.0	9.8	16.2

Table 2. The type I error (×100) and power (×100) for the six tests considered at 5%, 1%, 0.5%, and 0.1% significance levels when methylation β-values were generated based on the normal distribution (model 2). The numbers of cancer and normal samples are (100, 100)

(Δ, r_{21})	Level (%)	T_W	T_S	T_L	ELRT	Logistic	*t*-test
(0, 1)	5	5.0	4.9	5.5	5.7	3.7	5.1
	1	0.8	0.8	1.0	1.1	0.4	1.0
	0.5	0.4	0.4	0.5	0.6	0.2	0.5
	0.1	0.1	0.1	0.1	0.1	0.0	0.1
(0.6, 1)	5	96.7	96.9	97.1	97.2	96.6	98.7
	1	88.2	89.2	89.6	90.2	87.3	94.8
	0.5	82.6	84.0	84.7	85.7	81.1	91.9
	0.1	65.9	69.2	70.5	72.9	63.0	80.9
(0, 0.7)	5	86.1	88.0	89.9	90.2	83.8	5.1
	1	61.5	65.5	71.9	73.6	53.0	1.0
	0.5	49.8	53.9	62.9	64.5	38.9	0.4
	0.1	26.6	29.1	42.0	44.8	13.5	0.2
(0.3, 0.8)	5	78.9	80.6	82.3	82.8	77.5	64.4
	1	52.6	56.6	61.1	63.0	47.0	40.3
	0.5	41.0	45.7	51.4	53.2	34.7	31.6
	0.1	20.3	24.6	31.4	33.9	12.9	16.3

Table 3. The type I error (×100) and power (×100) for the six tests considered at 5%, 1%, 0.5%, and 0.1% significance levels when methylation β-values were generated based on the mixture of two normal distributions (model 3). The numbers of cancer and normal samples are (100, 100)

(Δ, r_{21})	Level (%)	T_W	T_S	T_L	ELRT	Logistic	*t*-test
(0, 1)	5	4.9	4.9	5.0	5.2	4.6	5.0
	1	0.9	0.9	0.9	1.1	0.7	1.0
	0.5	0.5	0.5	0.5	0.6	0.4	0.5
	0.1	0.1	0.1	0.1	0.2	0.1	0.2
(0.6, 1)	5	99.2	98.5	98.5	98.6	98.4	98.2
	1	96.4	93.6	93.6	94.1	93.0	93.4
	0.5	94.6	90.6	90.5	91.3	89.2	90.2
	0.1	88.3	78.9	78.5	80.6	75.7	79.5
(0, 0.9)	5	59.6	60.1	60.6	61.6	58.4	5.3
	1	33.5	34.7	35.4	36.8	31.1	1.3
	0.5	24.3	25.6	26.6	28.0	22.5	0.7
	0.1	11.0	12.4	13.1	14.6	9.3	0.1
(0.2, 0.9)	5	90.2	89.4	89.7	90.1	88.4	32.2
	1	74.0	72.2	72.7	74.1	69.1	13.3
	0.5	64.8	63.4	64.3	66.1	59.0	9.0
	0.1	44.1	42.5	44.0	47.0	36.0	3.5

logistic regression (Logistic) control the Type I errors well at all nominal levels. Although the true distribution of the methylation values is not normal, the rejection rates of the *t*-test under Type I error scenario are very close to the nominal levels, suggesting that the *t*-test is quite robust for this beta distribution setting. When the distributions of two groups are different in means only and differ by 0.6 SD, the *t*-test has about 4% to 22% more power than the other five tests, whereas the other five tests have a similar power (except for the Logistic test which has 15% less power than T_L at the lower nominal level of 0.1%). This is because the *t*-test has one degree of freedom whereas the other five tests have two degrees of freedom. When the two distributions are different in SDs only, T_L and ELRT are the most powerful tests. Note that the *t*-test has essentially no power beyond Type I errors in this scenario because it can only detect differences in means, not in SDs. When both means and SDs are different, T_L and ELRT remain to be the most powerful tests among the six tests, and the *t*-test is grossly underpowered. Table 2 and Table 3 summarize the results when methylation β-values were generated based on the normal distribution (model 2) and the mixture of normal distributions (model 3), respectively. Similar findings were observed as in the beta distribution setting. If we compare the performance of different tests under the Power scenario I across Tables 1, 3, we found that the superiority of the *t*-test over the proposed tests depends on the underlying true distributions. When the true distribution is beta or normal, the *t*-test has 4% to 22% more power. While when the true distribution is the mixture of two normal distributions, the *t*-test has similar power as the other tests.

To evaluate the robustness of the proposed tests to model mis-specification, we conducted additional simulation studies. We generated data from *t* distributions, which do not satisfy the exponential tilt model assumption described in equation (1). For the normal samples, a *t* distribution with 3 degrees of freedom with mean 0 and variance of 3 was chosen,

Table 4. The type I error (×100) and power (×100) for the six tests considered at 5%, 1%, 0.5%, and 0.1% significance levels when methylation β-values were generated based on the *t*-distribution (mis-specification model). The numbers of cancer and normal samples are (100, 100)

(Δ, r_{21})	Level (%)	T_W	T_S	T_L	ELRT	Logistic	*t*-test
(0, 1)	5	5.8	3.7	6.0	6.2	2.3	4.4
	1	1.1	0.4	1.3	1.5	0.2	1.0
	0.5	0.5	0.2	0.6	0.7	0.1	0.6
	0.1	0.1	0.0	0.1	0.1	0.0	0.1
(0, 0.6)	5	50.0	31.9	58.7	59.7	25.2	4.8
	1	18.5	7.5	29.7	31.2	4.6	1.0
	0.5	11.1	3.6	20.5	22.0	1.8	0.5
	0.1	2.7	0.6	7.7	8.7	0.2	0.1
(0.5, 1)	5	87.1	92.7	94.0	94.2	91.2	93.1
	1	68.1	78.9	83.1	84.0	74.1	82.1
	0.5	58.6	71.4	77.1	78.4	65.1	75.5
	0.1	38.5	51.0	59.4	62.2	40.9	57.6
(0.2, 0.6)	5	75.1	68.3	83.8	84.3	65.2	43.1
	1	43.0	35.6	61.5	63.1	29.3	21.6
	0.5	30.8	25.0	51.5	53.3	18.9	15.8
	0.1	11.8	9.4	29.9	32.6	5.5	6.5

and the non-centrality parameter was set at 0. For the cancer group, different degrees of freedom, mean, variance and non-centrality parameters were chosen, so that the distributions of methylation values in cancer group and normal group are identical, different in means only, different in SDs only, or different in both means and SDs, respectively. The results on Type I errors and power under model mis-specifications when sample size is 100 in each group are shown in Table 4. We can see that T_W, T_S, T_L, and ELRT tests control the Type I errors well at all nominal levels, suggesting a certain degree of model robustness to distributions with heavier tails. T_L and ELRT tests are still the most powerful tests among the six. Similar patterns were observed for the *t*-test as in previous three parametric models.

We also conducted simulation studies when sample size is moderate (50, 50). The results are included in the supplementary materials due to limited space. We noticed that the

Table 5. The type I error (×100) for the six tests considered at 5%, 1%, 0.5%, and 0.1% significance levels when methylation β-values were generated based on models 1-3 (upper panel: under no mis-specification scenarios), and t-distribution (lower panel: under mis-specification scenario). The numbers of cancer and normal samples are (20, 20)

Model	Level (%)	T_W	T_S	T_L	ELRT	Logistic	t-test
Under no mis-specification scenarios							
Beta	5	7.3	4.0	5.8	7.2	1.5	4.7
	1	1.4	0.5	1.2	1.9	0.0	1.0
	0.5	0.5	0.2	0.5	1.1	0.0	0.6
	0.1	0.1	0.0	0.1	0.3	0.0	0.1
Normal	5	7.3	4.1	5.9	7.6	1.4	5.4
	1	1.5	0.6	1.1	1.8	0.0	1.2
	0.5	0.7	0.1	0.5	1.0	0.0	0.6
	0.1	0.1	0.0	0.1	0.2	0.0	0.1
Mixture	5	6.3	4.7	4.9	6.2	2.9	5.5
	1	1.0	0.6	0.9	1.7	0.2	1.2
	0.5	0.5	0.3	0.4	0.7	0.1	0.5
	0.1	0.1	0.1	0.1	0.1	0.0	0.1
Under mis-specification scenario							
t-distribution	5	8.4	2.9	6.9	8.6	0.8	4.7
	1	1.8	0.3	1.1	1.9	0.0	0.9
	0.5	0.9	0.1	0.5	1.1	0.0	0.5
	0.1	0.2	0.0	0.1	0.3	0.0	0.1

results on Type I errors and power when sample size is moderate are similar to those when sample size is large. That is, the t-test has more power than the other five tests when the two groups are different in means only, while T_L and ELRT are the most powerful tests when the two distributions are different in sd only. All six tests maintain Type I errors well.

Table 5 summarizes the rejection rates under the Type I error scenario when sample size is small. The upper panel of Table 5 reports the results under no mis-specification scenarios. The power results were included in the supplementary material due to limited space. The test T_L and the t-test maintain the correct empirical Type I errors while the tests T_W and ELRT have inflated Type I errors. The test T_S has relatively conservative Type I errors while the Wald test using logistic regression (Logistic) has very conservative Type I errors and is hence not recommended when sample sizes are small. Under model mis-specification scenario, T_L and T_S tests still maintain the Type I errors well.

We also considered an unbalanced design when there are 10 cancer samples and 30 normal samples. The results are similar to the balanced design when sample sizes are (20, 20) and are included in the supplementary material. We note that the power of T_W, T_S and T_L under sample size (10, 30) is lower than that under sample size (20, 20). This agrees with the results in Theorem 2 that implies that a balanced design has higher power than an unbalanced design given the same total number of samples.

In summary, our simulation studies suggest that both the pseudolikelihood ratio test (T_L) and the empirical likelihood ratio test (ELRT) perform well compared to the t-test and logistic regression with well controlled Type I errors and improved power when two groups differ in variances only, or in both means and variances and when sample sizes are relatively large. The test T_L shows some degree of robustness under model mis-specifications. Moreover, the test T_L is recommended when the sample sizes are relatively small for a better controlled Type I error.

Real Data Application

We applied the proposed tests to the data from the United Kingdom Ovarian Cancer Population Study (UKOPS) to select differentially methylated loci between ovarian cancer cases and age-matched healthy controls using Illumina Infinium Human Methylation27 Beadchip [Teschendorff et al., 2010]. The original data have 266 ovarian cancer cases with 131 pretreatment cases and 135 posttreatment cases, and 274 age-matched healthy controls. Since age and having received treatment or not when blood samples were taken are known factors to affect DNA methylation levels, we chose to use a more homogenous population with 131 ovarian cancer cases who gave their blood at the time of their diagnosis prior to treatment and with age-matched controls to illustrate the feasibility and power of the proposed method. We have previously worked on this data set thus we refer readers to Wang [2011] for the detailed quality control steps. We ended up with 96 cancer samples and 136 normal samples with methylation β-values on 22, 951 loci. Since our simulation studies have suggested that the pseudolikelihood ratio test (T_L) and the empirical likelihood ratio test (ELRT) have similar power and Type I errors but better power than the other proposed tests when sample size is relatively large. We therefore focused on the comparison between the pseudolikelihood ratio test (T_L) and the t-test in this real data application. We also only focused on the original methylation β-values instead of the logit transformed ones. We adjusted for multiple comparisons using Q-values [Storey and Tibshirani, 2003].

Of the total 22, 951 loci tested, 2, 691 loci have q-values $<$ 0.05 using the T_L test and 2, 699 loci have q-values $<$ 0.05 using the t-test. Among those, there are overlapping 2, 324 loci that have q-values $<$ 0.05 using both methods. We further examined the 367 loci that were identified by the T_L test but not by the t-test to see if those are the loci that differ mostly in variances but not means between the cancer and normal groups, and 375 loci that were identified by the t-test but not by the T_L test to see if those are the loci that differ mostly in means but not variances between the cancer and normal groups.

The upper left panel of Figure 1 displays the distribution of Δ and r_{21} for the 367 loci that were identified by the T_L test but not the t-test. The upper right panel of Figure 1 displays the distribution of Δ and r_{21} for the 375 loci that were identified by the t-test but not the T_L test. The lower panel of Figure 1 displays the distribution of Δ and r_{21} for the 2, 324 overlapping loci that were identified by both the T_L test and the t-test.

It is clear that these loci that were identified by the T_L test only but not the t-test have more significant variance differences than mean differences between the cancer and normal groups, scenarios the proposed tests are designed for. In contrast, those loci that were identified by the t-test only have more significant mean differences than variance

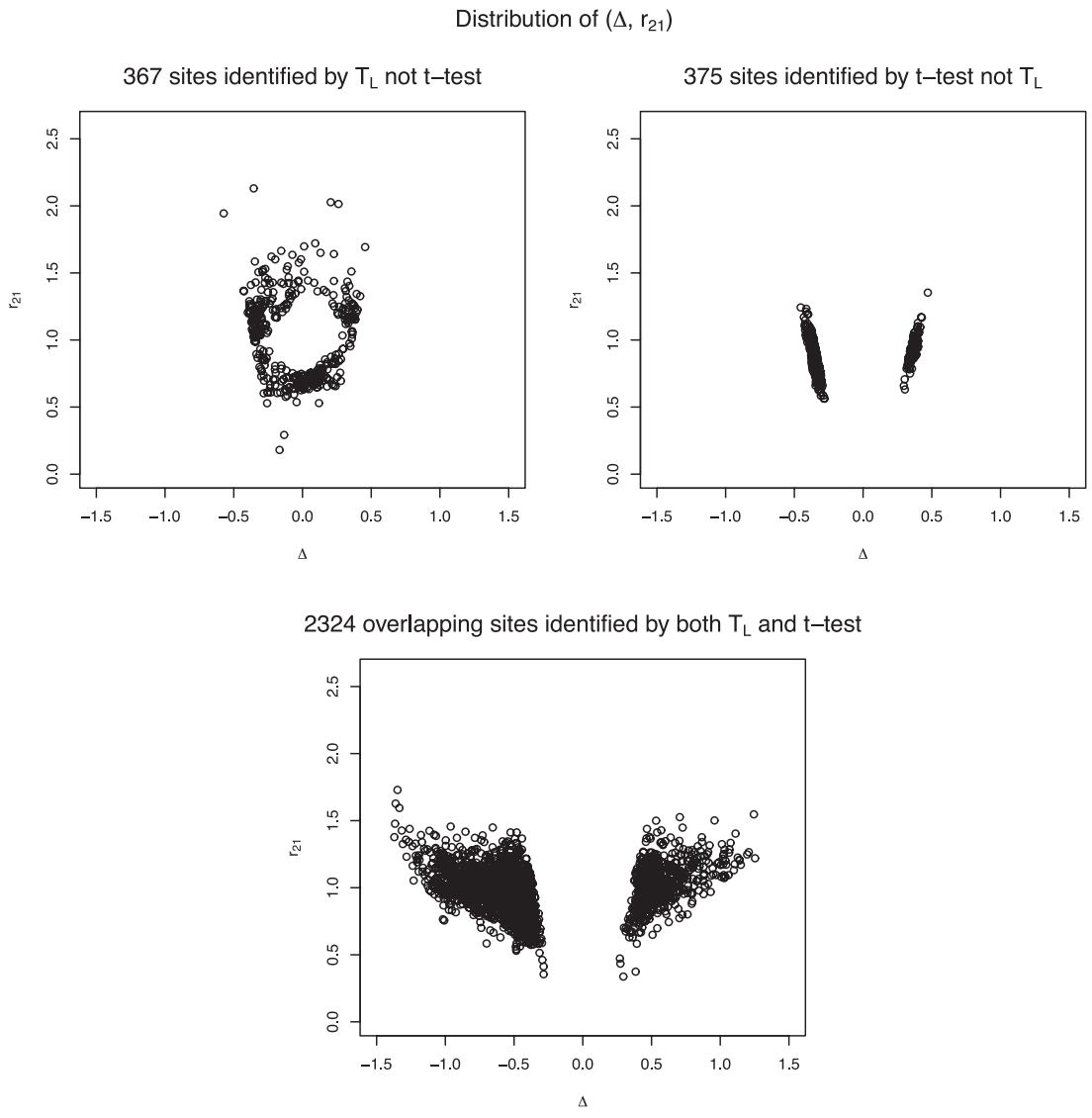

Figure 1. Upper left panel: distribution of Δ and r_{21} for the 367 loci that were identified by the T_L test but not the *t*-test; Upper right panel: distribution of Δ and r_{21} for the 375 loci that were identified by the *t*-test but not the T_L test; Lower panel: distribution of Δ and r_{21} for the 2, 324 overlapping loci that were identified by both the T_L test and the *t*-test.

differences between the cancer and normal groups in general. For the overlapping 2, 324 loci that were identified by both the T_L test and the *t*-test, the majority have much larger differences in means than the loci in the above two categories. Thus these loci are relatively easier to be identified by either methods focusing on mean differences or methods focusing on differences in both means and variances. Moreover, for those loci that have relatively small mean differences (but still have larger mean differences than loci that were identified by T_L test only), they have large differences in variance in general. Thus T_L test is able to identify them although simulation studies suggested a slightly lower power for T_L test in such scenarios than the t-test.

In summary, the proposed test T_L has identified additional loci of potential interest that were missed by the standard *t*-test. We also compared the loci identified by ELRT with those by the *t*-test and the results are displayed in Supporting Information Fig. S1. A similar pattern was observed in that the loci identified by ELRT only but not the t-test have more significant variance differences than mean differences, while the loci identified by the t-test only have more significant mean differences. In addition, we compared the loci identified by T_L with those by ELRT (Supporting Information Fig. S2), where the majority overlapped (2,592 loci were identified by both T_L and ELRT, i.e., 96% of the total 2,702 loci identified by either T_L or ELRT). There were 11 loci identified by ELRT but not T_L, and 99 loci identified by T_L but not ELRT. These loci share similar patterns in mean and variance differences between the cancer and the normal groups except for two loci "cg04498032" and "cg06866862" that were identified by

ELRT but not T_L (upper left panel of Supporting Information Fig. S2). Note that the q-values based on T_L at these two loci are 0.056 and 0.054, respectively, which are very close to the significance level of 0.05. Thus, the real data application results agree with the simulation results that the performance of T_L and ELRT are very similar when sample sizes are relatively large. In conclusion, the proposed tests can serve as useful complements of the standard tests.

Discussion

DNA methylation plays an important role in the development of many types of cancer. Identifying differentially methylated loci between cancer and normal patients is one of the central tasks to understand the contributions of the methylation process on cancer development. Several recent studies and our own group have noticed that methylation values in different groups could differ not only in means but also in variations [Gervin et al., 2011; Hansen et al., 2011]. Current methods do not fully take these features into account. One solution might be to conduct a standard mean test (e.g., the t-test) and a standard variance test (e.g., the Levene's test, Levene [1961]) at each marker. However, correction has to be made to account for multiple testing, which may lead to a sizable loss of power in detecting differential methylation. Therefore, to propose a new test that captures differences in both means and variances is necessary. Recently, Ahn and Wang [2013] proposed a generalized linear model that simultaneously tests differences in means and variance of methylation levels between cases and controls where they used a joint score test for both mean and variance to improve statistical efficiency. Here we proposed a generalized exponential tilt model that compared the two distributions to better account for the special features observed in the methylation measurements. We derived the semiparametric tests to obtain model robustness. Through simulation studies, we demonstrated the feasibility and power of the proposed tests. The proposed tests are much more powerful than the existing methods such as the t-test and the regression-based test when the cancer and normal groups are different in variances only or in both means and variances. When the cancer and normal groups are differ in means only, the t-test is marginally more powerful than the proposed tests. Thus, we believe the proposed class of tests are nice complements to the existing ones.

We introduced two parameters β_1 and β_2 in this paper to capture the differences in means and variances between the normal and cancer groups. An extension to a generalized exponential tilt model with more than two parameters could be made to capture the higher order differences (such as kurtosis) between the normal and cancer groups [Teschendorff et al., 2006]. However, the corresponding tests will have more degrees of freedom, leading to a potential power loss. Thus, we consider the current model with two parameters a parsimonious model that captures the major differences between two groups, that is, the differences in means and variances. Note that the functions $h_1(\cdot)$ and $h_2(\cdot)$ have to be specified in the proposed tests. The effect of mis-specifying these functions on empirical likelihood-based estimation has been considered by Fokianos and Kaimi [2006]. We are currently investigating the effect of mis-specification on the Type I errors and power of the proposed pseudolikelihood-based tests and developing testing procedures that are robust to such mis-specification. The results will be reported in the future.

Acknowledgments

Yong Chen was partially supported by the start-up fund and the PRIME award from the University of Texas School of Public Health. Shuang Wang was partially supported by R03 CA150140-01.

The authors have no conflict of interest to declare.

References

Ahn S, Wang T. 2013. A powerful statistical method for identifying differentially methylated markers in complex diseases. *Pages 69–79 of: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing.* NIH Public Access.

Anderson JA. 1979. Multivariate logistic compounds. *Biometrika* 66:17–26.

Anderson JA. 1972. Separate sample logistic discrimination. *Biometrika* 59:19–35.

Besag J. 1974. Spatial interaction and the statistical analysis of lattice systems. *J Royal Stat Soc Series B* 36:192–236.

Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E and others. 2006. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* 16:383–393.

Chen Y, Liang K-Y. 2010. On the asymptotic behavior of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika* 97:603–620.

Chen Z, Huang H, Liu J, HK TN, Nadarajah S, Huang X, Deng Y. 2013. Detecting differentially methylated loci for Illumina Array methylation data based on human ovarian cancer data. *BMC Medical Genomics* 6(Suppl 1):S9–S9.

Chowdhury S, Erickson SW, MacLeod SL, Cleves MA, Hu P, Karim MA, Hobbs CA. 2010. Maternal genome-wide DNA methylation patterns and congenital heart defects. *PLoS One* 6:e16506.

Cox DR, Reid N. 2004. A note on pseudolikelihood constructed from marginal densities. *Biometrika* 91:729–737.

Du P, Zhang X, Huang CC, Jafari N, Kibbe W, Hou L, Lin S. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 12:215.

Fokianos K. 2008. Comparing two samples by penalized logistic regression. *Electronic J Stat* 2:564–580.

Fokianos K, Kaimi I. 2006. On the effect of misspecifying the density ratio model. *Annal Institute Stat Math* 58:475–497.

Gervin K, Hammerø M, Akselsen HE, Moe R, Nygård H, Brandt I, Gjessing HK, Harris JR, Undlien DE, Lyle R. 2011. Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res* 21:1813–1821.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D and others. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 43:768–775.

Huang H, Chen Z, Huang X. 2013. Age-adjusted nonparametric detection of differential DNA methylation with case–control designs. *BMC Bioinformatics* 14:86.

Kalari S, Pfeifer GP. 2010. Identification of driver and passenger DNA methylation in cancer by epigenomic analysis. *Adv Genet* 70:277–308.

Kalbfleisch JD, Sprott DA. 1970. Application of likelihood methods to models involving large numbers of parameters. *J Royal Stat Soc Series B* 32:175–208.

Kerkel K, Schupf N, Hatta K, Pang D, Salas M, Kratz A, Minden M, Murty V, Zigman WB, Mayeux RP and others. 2010. Altered DNA methylation in leukocytes with trisomy 21. *PLoS Genet* 6:e1001212.

Kuan PF, Wang A, Zhou X, Chu HT. 2010. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics* 26:2849–2855.

Laurila K, Oster B, Andersen CL, Lamy P, Orntoft T, Yli-Harja O, Wiuf C. 2011. A Beta-mixture model for dimensionality reduction, sample classification and analysis. *BMC Bioinformatics* 12:215.

Levene H. 1961. Robust tests for equality of variances. *Contrib Prob Stat: Essays Honor Harold Hotel* 2:279–292.

Liang K-Y, Self SG. 1996. On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J Royal Stat Soc Series B* 59:785–796.

Lindsay BG. 1988. Composite likelihood methods. *Contemporary Math* 80:221–39.

Owen AB. 2001. *Empirical Likelihood.* Vol. 92. Boca Raton, FL: Chapman & Hall/CRC.

Qin J. 1998. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika* 85(3), 619–630.

Qin J, Zhang B. 1997. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* 84:609–618.

Shen J, Wang S, Zhang YJ, Wu HC, Kibriya MG, Jasmine F, Ahsan H, Wu DP, Siegel AB, Remotti H and others. 2013. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics* 8:34–43.

Slater AA, Alokail M, Gentle D, Yao M, Kovacs G, Maher ER, Latif F. 2013. DNA methylation profiling distinguishes histological subtypes of renal cell carcinoma. *Epigenetics* 8:252–267.

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide studies. *Proc Natl Acad Sci* 100:9440–9445.

Tan Z. 2009. A note on profile likelihood for exponential tilt mixture models. *Biometrika* 96:229–236.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, et al. 2010. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res* 20:440–446.

Teschendorff AE, Naderi A, Barbosa-Morais NL, Caldas C. 2006. PACK: profile analysis using clustering and Kurtosis to find molecular classifiers in cancer. *Bioinformatics* 22:2269–2275.

Van der Vaart AW. 2000. *Asymptotic Statistics.* Cambridge: Cambridge Univ. Press.

Varin C, Reid N, Firth D. 2011. An overview of composite likelihood methods. *Statistica Sinica* 21:5–42.

Wang S. 2011. Method to detect differentially methylated loci with case-control designs using Illumina arrays. *Genetic Epidemiol* 35:686–694.