Bioinformatics, 36(7), 2020, 2017–2024
doi: 10.1093/bioinformatics/btz885
Advance Access Publication Date: 26 November 2019
Original Paper

Genome analysis

Detection of differentially methylated CpG sites between tumor samples with uneven tumor purities

Weiwei Zhang[1,†], Ziyi Li[2,†], Nana Wei[3], Hua-Jun Wu[4] and Xiaoqi Zheng[3,*]

[1]Department of Mathematics, School of Science, East China University of Technology, Nanchang, Jiangxi 330013, China, [2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA, [3]Department of Mathematics, Shanghai Normal University, Shanghai 200234, China and [4]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA 02215, USA

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Russell Schwartz

Received on August 6, 2019; revised on November 14, 2019; editorial decision on November 20, 2019; accepted on November 23, 2019

Abstract

Motivation: Inference of differentially methylated (DM) CpG sites between two groups of tumor samples with different geno- or pheno-types is a critical step to uncover the epigenetic mechanism of tumorigenesis, and identify biomarkers for cancer subtyping. However, as a major source of confounding factor, uneven distributions of tumor purity between two groups of tumor samples will lead to biased discovery of DM sites if not properly accounted for.

Results: We here propose InfiniumDM, a generalized least square model to adjust tumor purity effect for differential methylation analysis. Our method is applicable to a variety of experimental designs including with or without normal controls, different sources of normal tissue contaminations. We compared our method with conventional methods including minfi, limma and limma corrected by tumor purity using simulated datasets. Our method shows significantly better performance at different levels of differential methylation thresholds, sample sizes, mean purity deviations and so on. We also applied the proposed method to breast cancer samples from TCGA database to further evaluate its performance. Overall, both simulation and real data analyses demonstrate favorable performance over existing methods serving similar purpose.

Availability and implementation: InfiniumDM is a part of R package *InfiniumPurify*, which is freely available from GitHub (https://github.com/Xiaoqizheng/InfiniumPurify).

Contact: xqzheng@shnu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer has long been depicted as a type of genetic disease caused by either activation of oncogenes or inactivation of tumor suppressor genes. In recent years, it is becoming acceptable that the aberrant of epigenetic modification of DNA molecular, mainly including DNA methylation, could also contribute to tumor initiation and progression. For example, as one of the hallmarks of cancer, hyper-methylation in CpG islands (Burbee *et al.*, 2001; Yoon *et al.*, 2001) and global hypo-methylation across the genome were observed for most cancer types (Raddatz *et al.*, 2012). Moreover, the reversibility of DNA methylation can be used as a potential target for therapeutic treatment of cancer (Ahuja *et al.*, 2016). Consequently, it is of vital importance to accurately quantify the methylation difference between two sets of samples, including tumors versus adjacent normal tissues, tumor samples from different subtypes or those respond differently to a given cancer therapy.

The detection of differentially methylated (DM) CpG sites can be implemented by the typical differential expression detection tools such as limma (Ritchie *et al.*, 2015) and edgeR (Robinson *et al.*, 2010). Meanwhile, there are also a number of methods available that were developed specifically for DNA methylation data from bisulfite sequencing (Akalin *et al.*, 2012; Feng *et al.*, 2014; Hansen *et al.*, 2012; Hebestreit *et al.*, 2013; Park *et al.*, 2014; Sun *et al.*, 2014; Wu *et al.*, 2015) and Illumina Infinium microarray (Aryee *et al.*, 2014; Kuan *et al.*, 2010; Morris *et al.*, 2014; Peters *et al.*, 2015; Warden *et al.*, 2013; Zheng *et al.*, 2017).

However, tumor tissues are highly heterogeneous by consisting of different cell populations, e.g. tumor cells, normal cells, blood vessel and immune cells. Among them, the normal cell contamination is a major confounder for downstream analyses. In the past few years, it has been found by many researchers that tumor purity has substantial confounding effect in a number of DNA methylation analyses, e.g. detection of DM sites between tumor and normal samples (Zheng *et al.*, 2017), subtype clustering (Zhang *et al.*, 2017) and epigenome-wide association studies (Jaffe and Irizarry, 2014). A few methods have also been proposed to detect cell type-specific differential signals from DNA methylation data (Li *et al.*, 2019; Zheng *et al.*, 2018). However, all these methods are developed to address differential analysis accounting for cellular composition from general design. None of the methods are tailored for studying tumor samples, nor are they comprehensively tested by cancer data. For tumor-normal comparison, we have previously developed a generalized least square method to infer DM CpG sites by considering tumor purity effect (Zheng *et al.*, 2017). However, DM between different tumor subtypes could be more practically useful for the understanding of biological or clinical processes, and identifying effective biomarkers for diagnoses and treatments. For example, given two sets of tumor samples that show distinct clinical trajectories or outcomes for a cancer therapy, we may want to identify a set of marker genes whose methylation statuses are correlated with clinical phenotypes. An intuitive solution to this problem is to include tumor purities as a covariate into the regression model. However, as will be shown in the Method part, this kind of approach is not statistically rigorous since the relationship between differential methylation and tumor purity is multiplicative rather than additive, thus resulting in many false positive and false negative DM sites.

For bisulfite sequencing data, Hakkinen *et al.* proposed a statistical method to deconvolve bisulfite reads of tumor samples by maximum likelihood estimation (Hakkinen *et al.*, 2018). Their model shows superior and robust results compared with typical DM detection tools. However, bisulfite-sequencing techniques are costly and not used extensively in clinical research. As another popular technique to measure whole genome methylation profile at single CpG site resolution, Infinium 450k array is more affordable and the generated data are easier to analyze and interpret. It thus becomes the primary choice to study DNA methylation in cancer research, and was widely adopted in many cohorts, e.g. TCGA and ENCODE. In the present work, we propose a novel and statistically rigorous framework, based on a generalized linear model, for DM analysis using Infinium 450k array data from two groups of tumor samples. The performance of the proposed method was comprehensively evaluated by simulation studies as well as real data analyses using TCGA breast tumor samples.

2 Materials and methods

2.1 Data and preprocessing

DNA methylation beta values range from 0 to 1, and mainly locate at the boundary regions (0 and 1). As a result, raw beta values of a CpG site cannot be modeled as normal distribution in the original scale. To fully embrace the powerful and flexible linear regression model which often assumes the normality of data distribution, we first transformed the raw beta values using an arcsine transformation: $f(x) = \arcsin(2x - 1)$. Such transformation has been previously used in differential methylation analysis and shows reliable results (Zheng *et al.*, 2017).

2.2 Adding tumor purities as regression covariate is problematic

Suppose we have two groups of tumor samples and presumably a set of normal control samples from the same cancer type. For CpG site i in dataset, let X_i be the transformed beta values for pure normal samples, we assume that $X_i \sim N(m_i, \sigma_i^2)$. Let $Y_{1,i}$ and $Y_{2,i}$ be the transformed beta values for pure cancer cells in subtypes 1 and 2, respectively. We assume that $Y_{1,i} = X_i + \delta_{1,i}$, and $Y_{2,i} = X_i + \delta_{2,i}$, where $\delta_{1,i}$ and $\delta_{2,i}$ are the differences between pure cancer and pure normal cells in subtypes 1 and 2, which are also assumed to follow normal distributions, i.e. $\delta_{1,i} \sim N\left(\mu_{1,i}, \tau_{1,i}^2\right)$, $\delta_{2,i} \sim N\left(\mu_{2,i}, \tau_{2,i}^2\right)$. Due to the additivity of normal distribution, $Y_{1,i}$ and $Y_{2,i}$ also follow normal distributions, i.e. $Y_{1,i} \sim N\left(m_i + \mu_{1,i}, \sigma_i^2 + \tau_{1,i}^2\right)$ and $Y_{2,i} \sim N\left(m_i + \mu_{2,i}, \sigma_i^2 + \tau_{2,i}^2\right)$. Our goal here is to test, for each CpG site, whether the mean methylation levels of pure cancer samples are identical in two subtypes, i.e. $m_i + \mu_{1,i} = m_i + \mu_{2,i}$.

In real clinical scenarios, however, the data from pure cancer samples in two subtypes are not available. Instead, we only observe methylation profiles of tumor tissues, representing mixture signals from pure normal and cancer cells at different proportions. The proposed method assumes tumor purities to be known, which can be inferred using state-of-the-art methods (Ahn *et al.*, 2013; Bao *et al.*, 2014; Carter *et al.*, 2012; Yoshihara *et al.*, 2013; Zheng *et al.*, 2017). If a sample s comes from the subtype k, the observed methylation level of a CpG site i, denoted as $Y'_{k,is}$, is the mixed signals from pure normal and cancer cells and can be expressed as $Y'_{k,is} = (1 - \lambda_{k,s})X_{is} + \lambda_{k,s}Y_{k,is} = X_{is} + \lambda_{k,s}\delta_{k,i}$. Thus $Y'_{1,is} \sim N(m_i + \lambda_{1,s}\mu_{1,i}, \sigma'^2_{i1})$ and $Y'_{2,is} \sim N(m_i + \lambda_{2,s}\mu_{2,i}, \sigma'^2_{i2})$, where σ'^2_{i1} and σ'^2_{i2} are the variances for $Y'_{1,is}$ and $Y'_{2,is}$ respectively and $\sigma'^2_{i1} \neq \sigma_i^2$, $\sigma'^2_{i2} \neq \sigma_i^2$. We further let $\mu_{2,i} = \mu_{1,i} + b_i$, then the hypothesis test becomes: $H_0: \ b_i = 0$. Under this parameterization, the distributions of $Y'_{1,is}$ and $Y'_{2,is}$ are now $Y'_{1,is} \sim N\left(m_i + \lambda_{1,s}\mu_{1,i}, \sigma'^2_{i1}\right)$ and $Y'_{2,is} \sim N\left(m_i + \lambda_{2,s}\mu_{1,i} + \lambda_{2,s}b_i, \sigma'^2_{i2}\right)$. From the derivations of $Y'_{1,is}$ and $Y'_{2,is}$, we found that the DM analysis between tumor samples is primarily to test $H_0: \ \lambda_{2,s}\mu_{1,i} - \ \lambda_{1,s}\mu_{1,i} + \lambda_{2,s}b_i = 0$, which is not equivalent to test $H_0: \ b_i = 0$. We further found that uneven level of tumor purities will seriously affect the differential methylation analysis, and more importantly, tumor purity has multiplicative effect on differential methylation, instead of additive. Therefore, the common practice by adding tumor purity as a continuous covariate in addition to other experimental factors fails to incorporate purity in a multiplicate framework, which may lead to biased results in many cases.

2.3 Proposed model formulation

We proposed the following linear regression model for DM analysis between two groups of tumor samples. Let Z be a vector by concatenating n_0 normal samples, n_1 tumor samples of subtype 1 and n_2 tumor samples of subtype 2. For simplicity of notation, we drop the subscript i in the following derivations. Denote all observed data as $Z = \left[X_1, \ldots, X_{n_0}, Y'_{1,1}, \ldots, Y'_{1,n_1}, Y'_{2,1}, \ldots, Y'_{2,n_2}\right]^T$, the observed data can be described as a linear model: $Z = W\beta + \varepsilon$, where

$$W = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & \lambda_{1,1} & 0 \\ \vdots & \vdots & \vdots \\ 1 & \lambda_{1,n_1} & 0 \\ 1 & \lambda_{2,1} & \lambda_{2,1} \\ \vdots & \vdots & \vdots \\ 1 & \lambda_{2,n_2} & \lambda_{2,n_2} \end{bmatrix}, \ \beta = \begin{bmatrix} m \\ \mu_1 \\ b \end{bmatrix} \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_0} \\ \varepsilon_{n_0+1} \\ \vdots \\ \varepsilon_{n_0+n_1} \\ \vdots \\ \varepsilon_{n_0+n_1+n_2} \end{bmatrix}.$$

Our framework is flexible to detect DM sites even if normal samples are not available (by removing normal-corresponding rows from W, as demonstrated in later part). And we also generalized our model to other comparison problems, e.g. comparison between two cancer types where their associated normal tissues are different (by adding one column to W, as demonstrated in Supplementary Material S1). This framework is also general enough to include our previous method (Zheng *et al.*, 2017) as a special case where there are only normal and tumor samples from one cancer type.

The estimated parameters from GLS for regression coefficients and covariance matrix are obtained as

$$\hat{\beta} = \left(W^T W\right)^{-1} W^T Z \triangleq HZ, \ \text{where } H = \left(W^T W\right)^{-1} W^T,$$

and

$$\text{var}(\hat{\beta}) = H\text{var}(Z)H^T.$$

The variance of Z is $\text{var}(Z) = \begin{bmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{bmatrix}$, where

$$\Sigma_1 = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} \sigma_1'^2 & 0 & 0 & 0 \\ 0 & \sigma_1'^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_1'^2 \end{bmatrix} \text{ and } \Sigma_3 = \begin{bmatrix} \sigma_2'^2 & 0 & 0 & 0 \\ 0 & \sigma_2'^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_2'^2 \end{bmatrix}.$$

Given estimated $\hat{\beta}$, regression residuals are $\hat{\varepsilon} = Z - W\hat{\beta}$, then the estimated values of σ^2, $\sigma_1'^2$ and $\sigma_2'^2$ are

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n_0} \hat{\varepsilon}_j^2}{n_0 - 3}, \quad \hat{\sigma}_1'^2 = \frac{\sum_{j=n_0+1}^{n_0+n_1} \hat{\varepsilon}_j^2}{n_1 - 3} \text{ and } \hat{\sigma}_2'^2 = \frac{\sum_{j=n_0+n_1+1}^{n_0+n_1+n_2} \hat{\varepsilon}_j^2}{n_2 - 3}.$$

We applied a shrinkage estimator on the estimated subtypes/normal variances to obtain $\hat{\sigma}^2$, $\hat{\sigma}_1'^2$ and $\hat{\sigma}_2'^2$, which was also utilized by (Park and Wu, 2016). The estimated variance of β can be obtained by dividing H into three parts and plugging estimated values of σ^2, $\sigma_1'^2$ and $\sigma_2'^2$ to equation $\text{var}(\hat{\beta}) = H\text{var}(Z)H^T$ as

$$\text{var}(\hat{\beta}) = H\text{var}(Z)H^T = [H_1 \ H_2 \ H_3] \begin{bmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{bmatrix} \begin{bmatrix} H_1^T \\ H_2^T \\ H_3^T \end{bmatrix}$$

$$= H_1\Sigma_1H_1^T + H_2\Sigma_2H_2^T + H_3\Sigma_3H_3^T.$$

For differential methylation between tumor subtypes, the Wald test statistics are calculated as

$$t = \frac{\hat{\beta}_{[3]}}{\sqrt{\text{var}(\hat{\beta})_{[3,3]}}},$$

where $\hat{\beta}_{[3]}$ is the third item of $\hat{\beta}$, and $\sqrt{\text{var}(\hat{\beta})_{[3,3]}}$ is the [3, 3] element of the matrix $\sqrt{\text{var}(\hat{\beta})}$. The Wald statistics follow a t distribution with $n_0 + n_1 + n_2 - 3$ degrees of freedom, and it is possible to calculate nominal P-values. Based on the calculated P-values at each CpG site, false discovery rate (FDR) can be estimated using established procedures such as Benjamini–Hochberg's method (Benjamini and Hochberg, 1995).

Based on the above framework, we can propose a model to identify DM between tumor subtypes without using the data from normal samples. Following the notations above, the observed data can be expressed in following regression form:

$$\begin{bmatrix} Y'_{1,1} \\ Y'_{1,2} \\ \vdots \\ Y'_{1,n_1} \\ Y'_{2,1} \\ Y'_{2,2} \\ \vdots \\ Y'_{2,n_2} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{1,1} & 0 \\ 1 & \lambda_{1,2} & 0 \\ \vdots & \vdots & \vdots \\ 1 & \lambda_{1,n_1} & 0 \\ 1 & \lambda_{2,1} & \lambda_{2,1} \\ 1 & \lambda_{2,2} & \lambda_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & \lambda_{2,n_2} & \lambda_{2,n_2} \end{bmatrix} \begin{bmatrix} m \\ \mu_1 \\ b \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n_1} \\ \varepsilon_{n_1+1} \\ \varepsilon_{n_1+2} \\ \vdots \\ \varepsilon_{n_1+n_2} \end{bmatrix}.$$

Calculation details are similar, except now we have $\text{var}(Z) = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$, where $\Sigma_1 = \begin{bmatrix} \sigma_1'^2 & 0 & 0 & 0 \\ 0 & \sigma_1'^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_1'^2 \end{bmatrix}$ and

$$\Sigma_2 = \begin{bmatrix} \sigma_2'^2 & 0 & 0 & 0 \\ 0 & \sigma_2'^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_2'^2 \end{bmatrix},$$ so $\text{var}(\hat{\beta}) = H_1\Sigma_1H_1^T + H_2\Sigma_2H_2^T$. Given estimated $\hat{\beta}$, regression residuals are $\hat{\varepsilon} = Z - W\hat{\beta}$, then the estimated values of $\sigma_1'^2$ and $\sigma_2'^2$ are

$$\hat{\sigma}_1'^2 = \frac{\sum_{j=1}^{n_1} \hat{\varepsilon}_j^2}{n_1 - 3} \text{ and } \hat{\sigma}_2'^2 = \frac{\sum_{j=n_1+1}^{n_1+n_2} \hat{\varepsilon}_j^2}{n_2 - 3}.$$

The Wald test statistics are calculated as

$$t = \frac{\hat{\beta}_{[3]}}{\sqrt{\text{var}(\hat{\beta})_{[3,\ 3]}}}.$$

The Wald test statistics follow a t distribution with $n_1 + n_2 - 3$ degrees of freedom, and the P-values can be obtained accordingly. Benjamini–Hochberg's method is applied on P-values to obtain FDRs.

We termed the proposed method as *InfiniumDM_ctl* when normal controls are available, and *InfiniumDM* when without normal controls. The method has been implemented as a function in the R package InfiniumPurify, which is freely available from GitHub.

3 Results

3.1 Tumor purities vary significantly among different cancer subtypes

We first examined the distributions of tumor purities of different subtypes by taking breast cancer samples (BRCA) from TCGA as an example. BRCA samples can be categorized into five intrinsic subtypes, i.e. basal, her2, luminal A and B and normal-like, based on expression profiles of 50 marker genes (PAM50) (Berger *et al.*, 2018). Previous studies found different subtypes show distinct clinical outcomes including recurrent risk, response to hormonal and chemical therapies (Liu *et al.*, 2016). Another cancer subtyping strategy which gains broad interest was the consensus Non-negative Matrix Factorization (cNMF). It groups samples based on a small set of metagenes (which defined as positive combination of original genes) rather than raw genes to avoid the curse of dimensionality and then uses consensus strategy to assemble different clustering results. We downloaded the cNMF clustering results for BRCA tumor samples using DNA methylation 450k array data at $k = 5$ from FIREHOSE of the Broad institute (https://gdac.broadinstitute.org/).

We examined two measurements of tumor purity for TCGA tumor samples. The first is by InfiniumPurify, which infers tumor purities from Infinium 450k array data with informative DM CpG sites using a density estimation algorithm (Zheng *et al.*, 2017). The second measurement is consensus purity estimate (CPE), which takes the median of purity estimates from four methods after normalization (Aran *et al.*, 2015). As shown in Figure 1, tumor purities estimated from InfiniumPurify vary significantly between different subtypes of BRCA samples for both cNMF (Fig. 1A) and PAM50 clusters (Fig. 1B) ($P = 5.04\text{e–}98$ and 4.8e–04 by F test). The respective tendency and P-values are similar for CPE (Fig. 1C and D) ($P = 3.12\text{e–}88$ and 9.57e–04 by F test) and other available purity estimates (Supplementary Fig. S1). Therefore, it is necessary for us

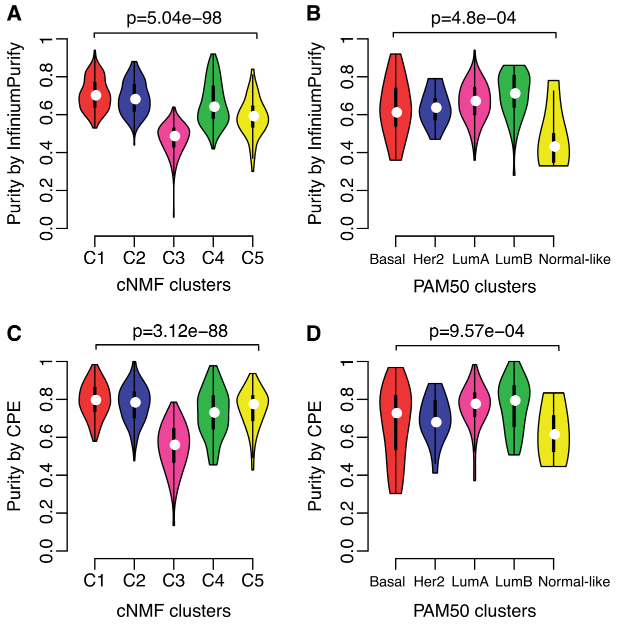

Fig. 1. Purity distributions for different BRCA subtypes. Purity distributions by InfiniumPurify for cNMF (**A**) and PAM50 (**B**) subtypes. Purity distributions by CPE for cNMF (**C**) and PAM50 (**D**) subtypes. *P*-values are obtained by F-test

to account for tumor purity in differential methylation analysis between different tumor subtypes.

Next, we aim to show that uneven levels of tumor purities between two sets of samples will undercut the differential methylation analysis if not properly adjusted for. We used methylation profiles of two cancer cell lines Mcf7 and T47dDm002p24h, and one normal cell line Hmec (all available from ENCODE) as reference to simulate tumor and normal samples (detailed simulation procedures are available in the first paragraph of Section 3.2). Since the reference is known, we used the methylation difference between two pure cell lines as criterion to define differential methylated CpG sites. Explicitly, if the absolute methylation difference of a CpG site between two cell lines is larger than 0.05, the CpG site is considered to be DM between these two cell lines and vice versa. As shown in Figure 2A, all CpG sites can be divided into three distinct groups according to their methylation levels in normal and two types of pure cancer cell lines. The first group consists of CpG sites showing no significantly methylated difference (methylation difference less than 0.05) between two types of cancer cells, as well as with normal cells. Obviously, this type of CpG site, which accounted for 37% of all CpG sites (Fig. 2B), will not be affected by the uneven levels of tumor purities. In other words, these CpG sites correspond to the true negative sites, for which the difference is always not significant between pure cancer cells and tumor subtypes. The second group consists of CpG sites that are also not DM between two cancer cell types, but at least one of them was DM with that in normal cells (type 2 sites). In such case, the uneven tumor purities between two types of tumor samples will end up with significant difference of observed methylation levels between two groups of tumor samples. An example CpG site for such case is shown in Figure 2C, where the mean methylation level of two types of cancer cells was almost the same (∼0.88), but significantly different from that in normal cells (∼0.14). If tumor purities of 2 sets of tumor samples are different, say 1 group has an averaged purity of 0.3 and the other is 0.6, then methylation levels between 2 groups of tumor samples will be quite different (around 0.38 versus 0.61). This type of CpG site accounts for 15% of total number of CpG sites (Fig. 2B), but contributes to the most of false positive CpG sites in DM analysis. In contrast, the third group only requires significant difference between two types of cancer cells, regardless of the difference between cancer and normal cells. This type of CpG site contains two scenarios, one is true positive sites, which are DM between pure cancer cells and the difference also holds after mixing with normal cells; the other is false negative sites, which are DM between pure cancer cells, but the difference is not significant after mixing the normal cells (Fig. 2D). Obviously, this type of CpG site constitutes the majority of all CpG sites, specifically, 48% of all CpG sites (Fig. 2B).

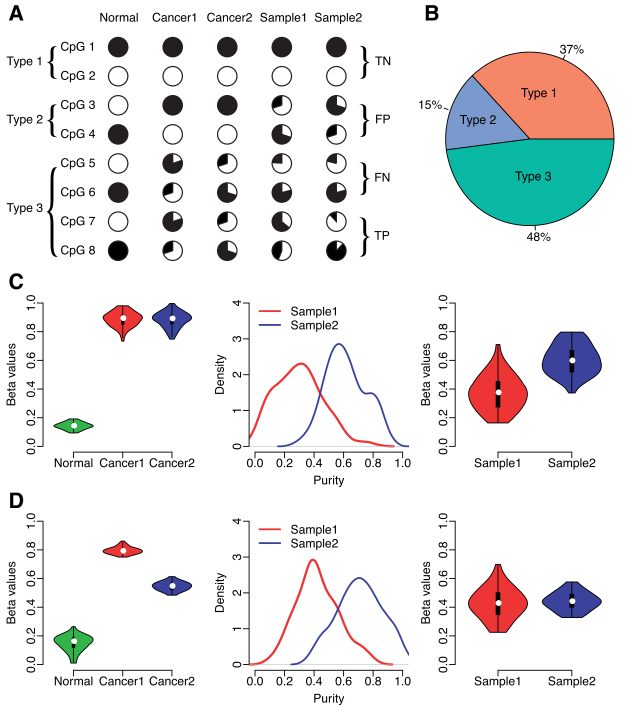

Fig. 2. Tumor purity affects differential methylation analysis between two sets of tumor samples. (**A**) Three types of CpG sites in DM analysis, where tumor samples are mixed from two cancer and one normal cell lines. (**B**) Proportion of three types of sites in simulation study. Examples of false positive (**C**) and false negative (**D**) sites in simulated data. Left panel shows methylation levels for pure normal and cancer cells. Middle panel shows tumor purity distributions in tumor samples. Right panel shows observed methylation levels in tumor samples

3.2 Simulation

All simulations are based on DNA methylation 450k array data. We used methylation data from three cell lines Hmec, Mcf7 and T47dDm002p24h downloaded from ENCODE as reference to generate simulated data, where Hmec cell line is derived from normal breast cells, Mcf7 and T47dDm002p24h cell lines are derived from breast cancer cells. We first generated methylation profiles of a number of cells based on each individual cell line. Taken the normal breast endothelial cell line Hmec as an example, we used its methylation profile as baseline and added independent random noise of normal distribution $N(0, \sigma^2)$ where the SD σ is estimated from 96 normal breast samples from TCGA. Because the pure cancer samples are more heterogeneous than normal, we multiply the variances estimated from pure normal samples by two as the variances of pure cancer cells. Then the beta values for 2 sets of tumor samples were generated by mixing the simulated pure normal and cancer data with tumor purities generated from normal distributions with mean m ranges from 0.05 to 0.95, and SD 0.2, which are similar to the real data estimates.

Comprehensive simulation studies are conducted to evaluate the performance of our method as well as the three conventional methods from several different aspects. Our criterion to evaluate the methods is the accuracy of identifying true DM sites over non-DM sites. In detail, for a CpG site, if the absolute difference of the true mean methylation levels between two types of pure cancer samples is greater than a threshold, it is defined as a DM site. If the absolute difference of the true mean methylation levels is less than a threshold, it is deemed as a non-DM site. The receiver operating characteristic (ROC) curve was plotted for each simulation, and the area under the curve (AUC) was calculated. Higher AUC is expected from better method. We compared different methods, including minfi, limma and limma corrected by using tumor purity as an additive covariate (termed as limmaPurity hereafter) under different parameter settings. In each setting, the results presented in this section are averaged over 50 Monte Carlo datasets, expect for Figure 3A, which is only based on one dataset.

We first compared these algorithms under different DM cutoffs 0.05, 0.1, 0.15 and 0.2, which roughly provides proportion of DM sites at 47%, 37%, 31% and 26% of total number of CpG sites (349707). Figure 3A shows ROC curves and AUC values of thresholds 0.05 and 0.2 from different methods (results for thresholds 0.1 and 0.15 are shown in Supplementary Fig. S2), under the same sample size 50 and mean purity deviation 0.3. For all simulation scenarios, our method with control samples (InfiniumDM_ctl) provides the best results, followed by InfiniumDM. Limma and minfi present very similar performances (0.843 versus 0.845 at threshold 0.2), which is expected because they both used linear regression model. The accuracy of limmaPurity, which includes tumor purity as a covariate, has greatly improved performance compared with limma. In detail, the AUC values are 0.821 and 0.892 from limma and limmaPurity, while our proposed methods have AUC 0.941 for InfiniumDM and 0.968 for InfiniumDM_ctl when threshold is 0.05. The higher prediction accuracy of our method should attributed to our correct model hypothesis and statistically rigorous modeling of the purity effect. Moreover, the performance of all methods becomes better and the performance difference becomes smaller when the thresholds increase as expected, since increasing thresholds makes the differential signals easier to detect. We also compared the specificity of different methods under the same parameter settings. We found our proposed method achieved much higher specificity than other methods (Supplementary Fig. S3).

To investigate the impact of sample sizes on the performance of the different methods, we conducted another simulation study with the same settings as above using DM cutoff 0.05. Figure 3B shows the AUC values by the 5 methods under sample sizes 10, 50 and 100. InfiniumDM_ctl still provides the best accuracy among all methods, followed by InfiniumDM. The accuracies of limma and minfi are almost the same, on average about 0.81 for different sample sizes and limmaPuirty has better performance (~0.88). We also observe that the accuracies of limmaPurity and our proposed method have been greatly improved when sample size increases from 10 to 50. For example, the AUC increases from 0.84 to 0.89 for limmaPurity, and from 0.89 to 0.94 for InfiniumDM and from 0.92 to 0.96 for InfiniumDM_ctl. However, the performance from the 3 methods remained almost unchanged when the sample size increases from 50 to 100. These results indicate that 50 samples in each group are sufficient to achieve satisfactory performance under the current setting.

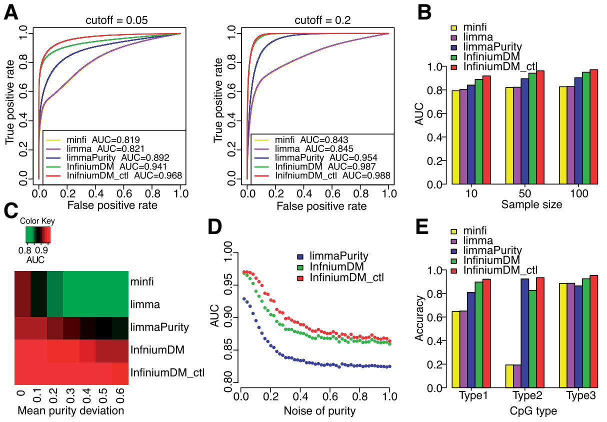

Fig. 3. DM detection accuracy from simulation under different scenarios. (**A**) ROC curves and AUC of the 5 methods including minfi, limma, limmaPurity, InfiniumDM and InfiniumDM_ctl on 2 simulated datasets with sample size 50 and cutoffs 0.05 (left panel), 0.2 (right panel). (**B**) Histogram of AUC by different sample sizes from the five methods. (**C**) Heatmap of AUC under different mean purity deviations. (**D**) AUC of different methods against noises of tumor purities. (**E**) DM calling accuracy of the five methods on three types of CpG sites

As discussed in Section 3.1, uneven levels of tumor purity between two sets of samples can seriously undercut the differential methylation analysis if not properly adjusted for, we conducted an additional simulation study to examine the effect of purity differences between two types of tumor samples. We set the mean tumor purities in 1 group as 0.3, and the purity differences between 2 groups range from low (0) to high (0.6) with other parameter settings fixed. Figure 3C shows the average AUC estimates over 50 Monte Carlo datasets. It shows that the mean difference is closely related with the detection performance for minfi, limma and limmaPurity. In detail, the AUC values of the three methods gradually decrease with the increase of the mean purity difference, but our proposed method is robust against the change of purity difference. For example, when the mean purity difference is 0 (i.e. the tumor purities of the 2 groups were almost the same), the AUC values of limma and minfi are almost the same (~0.91) and AUC of limmaPurity is around 0.93. However, when the mean purity difference increases to 0.3, the AUC values of limma, minfi and limmaPurity decrease to 0.82, 0.82 and 0.9, respectively. When the mean purity difference continues to increase until 0.6, the AUC values of limma and minfi decrease to 0.79, the AUC of limmaPurity decreases to 0.87, while the AUC values of our methods are almost unchanged across all scenarios. Overall, InfiniumDM_ctl and InfiniumDM consistently outperform existing methods, with an average AUC around 0.95 for most scenarios.

Among all scenarios, we found that the reliability of tumor purity has vital impact on the detection accuracy of DM sites. We compared the AUC values of limmaPurity, InfiniumDM and InfiniumDM_ctl, using true tumor purities by adding different levels of noise as inputs (Gaussian distribution with mean 0 and SDs ranging between 0 and 1, stepped by 0.02). The DM cutoff is set as 0.05, the mean tumor purity of one group is 0.3 and the other group is 0.6. It is shown that the AUC values of all three methods drop with the increase of SD (Fig. 3D). Our methods still outperform limmaPurity (around 0.86 over 0.82), indicating that our proposed methods are more robust even estimated tumor purities are biased.

We also investigated the performance of our algorithms on three types of CpG sites. We conducted simulation studies with sample size 50, and the mean difference of tumor purity between 2 types of tumor samples is roughly 0.3. Here, we also used 0.05 as a threshold to divide all CpG sites into 3 groups according to the methylation levels in normal and 2 types of pure cancer cells. As methylation levels in two pure cancer cells are known for each CpG site, we defined the accuracy as the percentage of correctly predicted sites. CpG sites with FDR lower than 0.05 are defined as DM sites between 2 types of tumor samples. Figure 3E shows the accuracies for the three types of CpG sites from different methods. For type 1 sites, the accuracy of InfiniumDM_ctl is 0.92, followed by InfiniumDM with accuracy of 0.9 and limmaPurity shows a higher accuracy of 0.81 compared with limma and minfi (~0.65). In contrast, limma and minfi achieve the worst performance for the type 2 sites (~0.19). This is because limma and minfi ignore tumor purity in detection of DM sites, thus the uneven tumor purities between two sets of samples could seriously bias the analysis of this type of sites and cause false positive findings. While limmaPurity and InfiniumDM_ctl have much better accuracy compared with limma and minfi. For type 3 CpG sites, we still observed that the proposed method achieves the best performances (~0.95 versus 0.88). These results confirm the favorable performance of InfiniumDM in differential methylation analysis between tumor samples while considering tumor purity, over other existing and commonly used differential methylation analysis tools.

Finally, we also conducted a series of simulation studies by using more stringent threshold for DM analysis. The simulated data are generated as the same procedure as described above. The only difference is that the true DM sites are defined as those having minfi P-values less than 0.05 and absolute methylation difference greater than 0.05. As shown in Supplementary Figure S4, the proposed method demonstrates consistent better accuracies under all simulated parameter settings, i.e. DM cutoff, sample sizes, mean purity deviations and purity estimation biases. These results further confirm the favorable performance of InfiniumDM in differential methylation analysis between tumor samples.

3.3 Real data analyses on BRCA samples

We further applied our method to BRCA samples from the TCGA database. We downloaded level 3 DNA methylation 450k array data of breast cancer from TCGA (https://portal.gdc.cancer.gov/), and PAM50 subtyping information from Broad GDAC Firehose (https://gdac.broadinstitute.org/). Note that only a part of TCGA tumor samples have subtyping information, we obtained in total 214 BRCA samples with annotated PAM50 subtypes. We then applied the proposed method to identify DM CpG sites between any two subtypes. To evaluate the performance of InfiniumDM_ctl, 13 normal controls are included. In consistent with the simulation study, we also compared our method with minfi, limma and limmaPurity.

Figure 4A shows the numbers of significant DM sites (defined as FDR < 0.01) in above between-subtype comparisons. On average, our methods output much more CpG sites than limmaPurity, limma and minfi. Note that the number of significant DM sites obtained by InfiniumDM_ctl is slightly less than that by InfiniumDM (67387 versus 68398). This may be possibly because normal tissues from clinical practices are also not pure but consist of different types of cells. The heterogeneity of normal samples would increase estimation variance and undermine detection power.

We next looked at the consistency of the detected DM CpG sites from different methods. We selected top 1000 CpG sites (ranked by q-values) detected from different methods, then a Venn diagram for luminal B-basal comparison is shown in Figure 4B (results for the rest comparisons are shown in Supplementary Fig. S5). The overlapped number is 634 for luminal B-basal comparison, and the average overlapped number is 377 for all 6 pairwise comparisons. We also noticed that, although both InfiniumDM and limmaPurity incorporate purity into the linear model, their overlap is not very high (449 on average). In contrast, the number of overlaps is 855 between InfiniumDM and InfiniumDM_ctl. This observation is related with how tumor purity is adjusted in the limmaPurity model, and again promotes the use of multiplicative formulations instead of additive models.

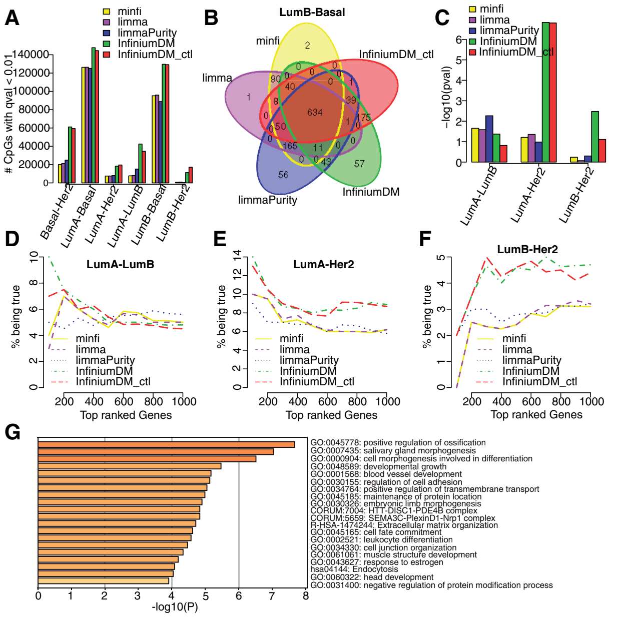

Fig. 4. Differential methylation results for BRCA samples from the TCGA database. (**A**) Numbers of differential methylated CpG sites (FDR < 0.01) by different DM calling methods. (**B**) Venn diagram of top 1000 DM CpG sites (ranked by q-values) between luminal B and basal subtypes by different methods. (**C**) Overlaps between predicted and true DM genes for luminal A-versus-luminal B, luminal A-versus-her2 and luminal B-versus-her2 by different methods. True DM genes are defined as genes with minfi *P*-value smaller than 0.05 between breast cancer cell lines with annotated PAM50 subtypes using RRBS data. Predicted DM genes are defined as genes with *P*-value smaller than 0.05 by different methods using Infinium 450k array data. (**D–F**) TDRs for different methods in the above three between-subtype comparisons. (**G**) Functional enrichment of DM genes (detected from top 1000 DM sites) by InfiniumDM between basal and her2 subtypes

The major difficulty to evaluate the performance of a DM calling method for real tumor samples is the lack of ground truth, i.e. true differential methylated CpG sites. We therefore resorted to an indirect strategy as follows. We downloaded reduced representation bisulfite sequencing (RRBS) data for 44 breast cell lines from the CCLE database (https://portals.broadinstitute.org/ccle), the PAM50 subtyping annotations of 84 breast cancer cell lines from Dai *et al.* (Dai *et al.*, 2017). Eventually we obtained 24 breast cancer cell lines with both RRBS and subtyping information, including 11 cell lines of luminal A subtype, 5 cell lines of luminal B subtype and 8 cell lines of her2 subtype. The RRBS data downloaded from CCLE have summarized methylation of CpG sites into gene levels. To make a fair comparison, we calculated the gene-level DNA methylation in Infinium 450k array data by averaging the beta values of CpG sites within promoter regions (1 kb upstream of TSS) (Fan *et al.*, 2016; Zheng *et al.*, 2017). Despite the differences between the DM using single CpG site versus gene context, we believe such benchmark still provides useful information in evaluating the proposed methods. Then, DM genes are defined as *P*-values less than 0.05 by different methods. For any pair of subtypes, we used DM genes from breast cancer cell lines as standard to evaluate the performance of the five methods. The true DM genes are defined as the ones with minfi *P*-values smaller than 0.05.

We first examined the overlapping DM genes detected by different methods in the above three pairwise subtype comparisons. Fisher's exact test is applied to check the significance of overlap between detected and true DM genes from different methods. It is clear that InfiniumDM and InfiniumDM_ctl show much smaller *P*-values compared with other methods in two of three between-subtype comparisons (Fig. 4C). Taken luminal A-versus-her2 as an example, the *P*-values by InfiniumDM_ctl and InfiniumDM are 1.57e−7 and 1.48e−7, whereas the *P*-values are 0.062, 0.044 and 0.1 by minfi, limma and limmaPurity, respectively. Even though limmaPurity obtained the smallest *P*-value (~0.005) compared with other methods in luminal A-luminal B comparison, InfiniumDM still has significant overlaps between detected DM genes and true DM genes (*P*-value is 0.04). Overall, our proposed method provides better detection accuracy than existing methods.

Next, we evaluated the accuracy of predicted DM genes by different methods using the DM genes from cell lines comparisons as benchmark. Figures 4D–F show the True Discovery Rate (TDR) curves for three between-subtype comparisons. We observed higher performance of InfiniumDM_ctl and InfiniumDM at all top-ranking genes than other three methods. For example, for luminal A-her2 comparison, among top 200 DM detections, 10.5% of them are true DM genes for InfiniumDM and InfiniumDM_ctl, whereas the percentages are 9.5%, 9.5% and 7% by minfi, limma and limmaPurity, respectively. Overall, the TDRs from minfi and limma are very similar, limmaPurity provides better accuracy compared with them, InfiniumDM_ctl and InfiniumDM obtain the best accuracy among all methods in the three comparisons. However, we also noticed that the overall TDRs by all methods, including existing and our proposed ones, are quite low (Fig. 4D–F). This is due to data type differences between the mixed data (obtained from TCGA tumor tissues with DNA methylation 450k array) and benchmark data (24 CCLE cell lines with RRBS). These two data are actually different in a few aspects, i.e. they cover different CpG sites, and the selected cell lines may not fully capture the sample characteristic of real tumor samples. In a word, our current gene-based DM is more like a silver benchmark rather than a gold standard.

Finally, we looked at the functional enrichment of the identified DM genes. For each between-subtype comparison, we first identified top 1000 DM CpG sites by each method, and then mapped these CpG sites to genes, respectively. The obtained genes are then input to Metascape (Zhou *et al.*, 2019) to explore their functional enrichments. We only take basal-her2 comparison as an example, the rest between-subtype comparisons are shown in Supplementary Figures S6–11. We got 23 enriched terms (q-value < 0.05) that are detected only by InfiniumDM (q-values by minfi, limma and limmaPurity all exceed 0.05). A list of all 23 terms and their q-values by different methods is provided in Supplementary Table S1. If focusing on the top 20 most enriched terms by InfiniumDM (Fig. 4G), we found 10 terms that are only detected by InfiniumDM compared with the other 3 methods and many of them are associated with breast cancer development, response to chemical therapy and subtyping according to the literature. For example, the GO term 0030155 (regulation of cell adhesion) is reported to control the tumorigenicity of cells of basal subtype and block cancer progression (Chekhun *et al.*, 2013). GO: 0001568 (blood vessel development, q-value 7.82e–3) is associated with basal subtype (Bujor *et al.*, 2018), and can be used for stratification of patients who might benefit from therapies targeting angiogenesis. GO: 0043627 (response to estrogen, q-value is 2.68e–2) is known as a gold standard for breast tumor subtyping (Rieger *et al.*, 2010). We also conducted further enrichment analysis using the exact location of top 1000 CpG sites (ranked by q-values) from different methods for basal-versus-her2 comparison, instead of genes, using EnrichR (Kuleshov *et al.*, 2016). Compared with existing methods limma, limmaPurity and minfi, our proposed methods InfiniumDM and InfiniumDM_ctl identified much more pathway terms that were directly related with cancer (Supplementary Fig. S12). For example, 'Basal cell carcinoma' and 'Melanogenesis' are exclusively identified by InfiniumDM and InfiniumDM_ctl. Among the existing methods, limmaPurity is the only one that identifies some cancer-related terms, such as 'Gastric cancer', 'Breast cancer' and 'Pathways in cancer'. However, these terms rank even higher in our proposed methods, suggesting that InfiniumDM and InfiniumDM_ctl can identify more relevant results than existing methods.

In conclusion, real data analyses in this section demonstrate that our proposed method can provide more sensitive, accurate and biologically meaningful results compared with other methods serving similar purpose.

4 Discussion

In this paper, we comprehensively investigated the impact of tumor purity in differential methylation analysis between two groups of tumor samples from the same cancer type. We found that uneven distributions of tumor purities between two groups of tumor samples will lead to both false positive and negative DM sites if not properly accounted for. We showed, through rigorous statistical formulation and simulation studies, that the common methods for differential analysis with the consideration of tumor purity by using purity as an additive covariate is not appropriate that it fails to incorporate purity in a multiplicate framework. Thus, we proposed a method using a generalized least square model to account for tumor purity and detect DM sites between two groups of tumor samples. Our model adopts a multiplicative formulation to account for tumor purity in detecting differential methylation, thus eliminates bias from a naïve model by only including tumor purity as an additive covariate. Simulation and real data analyses demonstrate that our approach provides more accurate and thus favorable results.

The proposed model actually represents a wide class of differential analysis methods, and can be easily extended to other related problems. For example, we can generalize the proposed model to differential methylation analysis between two distinct cancer types where the normal contaminations are allowed to be different (Supplementary Material S1). Moreover, the hypothesis testing is also very flexible. We can test whether the means of normal tissues in two groups are the same, whether the changes between cancer and normal tissue in two groups are the same and whether the means of the pure cancer tissues are the same. For all these tests, we have performed simulations and also obtained reasonable results (Supplementary Figs S13–S15).

Finally, the essence of the data modeling and statistical inference of our method can be applied to other platforms, and even other types of genomics data. In addition, our model can be further extended to incorporate more cell type information from cancer studies. Although difference between cancer and normal cells contributes the most and biggest differences in cancer data, the distinctions of different normal or cancer cell types can be further included in our framework. The assessment of these extensions is beyond the scope of this study. We will continue the exploration in our future works.

Acknowledgement

The authors thank Dr Hao Wu at Emory University for helpful discussions and suggestions.

Funding

This work was partly supported by the National Key R&D Program of China (2018YFA0900600 to X.Z.), National Natural Science Foundation of China (61902061 to W.Z., 61572327 and 61972257 to X.Z.), Natural Science Foundation of Shanghai (19ZR1437000 to X.Z.) and Science and Technology Research Project of Jiangxi Education Department (GJJ170445 to W.Z.).

Conflict of Interest: none declared.

References

Ahn,J. *et al.* (2013) DeMix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, **29**, 1865–1871.

Ahuja,N. *et al.* (2016) Epigenetic therapeutics: a new weapon in the war against cancer. *Annu. Rev. Med.*, **67**, 73–89.

Akalin,A. *et al.* (2012) MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.*, **13**, R87–R89.

Aran,D. *et al.* (2015) Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, **6**, 8971.

Aryee,M.J. *et al.* (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

Bao,L. *et al.* (2014) AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next generation sequencing data. *Bioinformatics*, **30**, 1056–1063.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Berger,A.C. *et al.* (2018) A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*, **33**, 690–705.

Bujor,I.S. *et al.* (2018) Evaluation of vascular proliferation in molecular subtypes of breast cancer. *In Vivo*, **32**, 79–83.

Burbee,D. *et al.* (2001) Epigenetic inactivation of RASSF1A in lung and breast cancers and malignant phenotype suppression. *J. Natl. Cancer Inst.*, **93**, 691–699.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Chekhun,S.V. *et al.* (2013) Expression of biomarkers related to cell adhesion, metastasis and invasion of breast cancer cell lines of different molecular subtype. *Exp. Oncol.*, **35**, 174–179.

Dai,X. *et al.* (2017) Breast cancer cell line classification and its relevance with breast tumor subtyping. *J. Cancer*, **8**, 3131–3141.

Fan,S. *et al.* (2016) Computationally expanding infinium HumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis. *Bioinformatics*, **32**, 1773–1778.

Feng,H. *et al.* (2014) A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.*, **42**, e69.

Hakkinen,A. *et al.* (2018) Identifying differentially methylated sites in samples with varying tumor purity. *Bioinformatics*, **34**, 3078–3085.

Hansen,K.D. *et al.* (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83–R10.

Hebestreit,K. *et al.* (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, **29**, 1647–1653.

Jaffe,A.E. and Irizarry,R.A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.*, **15**, R31–R39.

Kuan,P.F. *et al.* (2010) A statistical framework for Illumina DNA methylation arrays. *Bioinformatics*, **26**, 2849–2855.

Kuleshov,M.V. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.

Li,Z. *et al.* (2019) Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics*, **35**, 3898–3905.

Liu,M.C. *et al.* (2016) PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline- and taxane-based chemotherapy: correlative analysis of C9741 (Alliance). *NPJ Breast Cancer*, 2, 15023.

Morris,T. *et al.* (2014) ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics*, **30**, 428–430.

Park,Y. *et al.* (2014) MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, **30**, 2414–2422.

Park,Y and Wu,H. (2016) Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, **32**, 1446–1453.

Peters,T.J. *et al.* (2015) *De novo* identification of differentially methylated regions in the human genome. *Epigenet. Chromatin*., **8**, 6.

Raddatz,G. *et al.* (2012) Dnmt3a protects active chromosome domains against cancer-associated hypomethylation. *PLoS Genet*., **8**, e1003146.

Rieger,M.E. *et al.* (2010) The embryonic transcription cofactor LBH is a direct target of the Wnt signaling pathway in epithelial development and in aggressive basal subtype breast cancers. *Mol. Cell. Biol.*, **30**, 4267–4279.

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Robinson,M.D. *et al.* (2010) EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Sun,D. *et al.* (2014) MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.*, **15**, R38.

Warden,C. *et al.* (2013) COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Res.*, **41**, e117.

Wu,H. *et al.* (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.*, 43, e141.

Yoon,J. *et al.* (2001) Hypermethylation of the CpG island of the RASSF1A gene in ovarian and renal cell carcinomas. *Int. J. Cancer*, **94**, 212–217.

Yoshihara,K. *et al.* (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.

Zhang,W. *et al.* (2017) Accounting for tumor purity improves cancer subtype classification from DNA methylation data. *Bioinformatics*, **33**, 2651–2657.

Zheng,S.C. *et al.* (2018) Identification of differentially methylated cell types in epigenome-wide association studies. *Nat. Methods*, **15**, 1059–1066.

Zheng,X. *et al.* (2017) Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol.*, **18**, 17.

Zhou,Y. *et al.* (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.*, **10**, 1523.