# Chapter 2
# DNA Methylation and Cell-Type Distribution

**E. Andrés Houseman**

**Abstract** Epigenetic processes form the principal mechanisms by which cell differentiation occurs. Consequently, DNA methylation measurements are strongly influenced by the DNA methylation profiles of constituent cell types as well as by their mixing proportions, raising the potential for confounding of direct molecular associations at single CpG dinucleotides by associations between overall cell-type distribution with phenotype or exposure. In this chapter we review the literature on epigenetics and cell mixture; we then present techniques for deconvolution of DNA methylation measurements, either in the presence or in the absence of reference data. Finally, we present several data analysis examples.

**Keywords** Cell composition • Confounding • DMP • DMR • Immune • Mediation

## 2.1 Introduction

In the last decade, numerous published articles have demonstrated associations between DNA methylation profiles and disease or exposure phenotypes. For example, DNA methylation profiles measured in blood have been shown to correlate with ovarian cancer (Teschendorff et al. 2009), bladder cancer (Marsit et al. 2011), cardiovascular disease (Kim et al. 2010), obesity (Dick et al. 2014), and environmental exposures (Kile et al. 2014; Koestler et al. 2013a; Joubert et al. 2012). These associations have led to an interest in *epigenome-wide association studies* (EWAS), which aim to investigate associations between DNA methylation and health or exposure phenotypes across the genome (Rakyan et al. 2011a). Many of these epidemiologic studies have employed the Infinium platforms by Illumina, Inc. (San Diego, CA): the older HumanMethylation27 (27K) platform, which interrogates

E.A. Houseman (✉)
School of Biological and Population Health Sciences, College of Public Health and Human Sciences, Corvallis, OR, USA
e-mail: andres.houseman@oregonstate.edu



A.E. Teschendorff (eds.), *Computational and Statistical Epigenomics*, Translational Bioinformatics 7, DOI 10.1007/978-94-017-9927-0_2

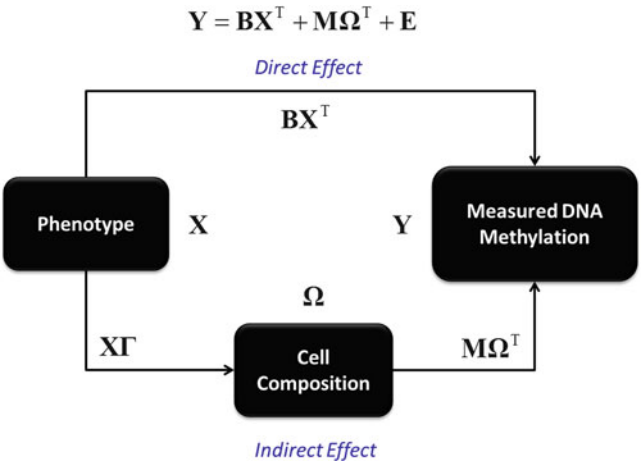


**Fig. 2.1** Mediation by cell composition

27,578 CpG loci, and the newer HumanMethylation450 (450K) platform, which interrogates 485,412 CpG loci. Both of these platforms measure locus-specific DNA methylation on an *average beta* scale, which is confined to the unit interval [0, 1] and roughly represents the fraction of methylated molecules in the given sample at the genomic position represented by the locus.

However, DNA methylation, associated with chromatin alterations, is partially responsible for coordination of gene expression in individual cells (Ji et al. 2010; Khavari et al. 2010; Natoli 2010). Consequently, normal tissue differentiation and cellular lineage is regulated by epigenetic mechanisms (Khavari et al. 2010), and DNA methylation shows substantial variation across tissue types (Christensen et al. 2009) as well as individual cell types, particularly distinct types of leukocytes (Ji et al. 2010). This understanding has led to a search for *differentially methylated regions* (DMRs) that distinguish specific cell lineages with high sensitivity and specificity (Baron et al. 2006). Figure 2.1 illustrates the consequence of heterogeneity in DNA methylation profile across cell types as it pertains to epidemiologic analysis of DNA methylation. In particular, DNA methylation measured in a tissue sample will be influenced both by cellular heterogeneity and by direct locus-specific phenotype effects. If the phenotype alters the composition of cells in the sample, then the *total effect* of phenotype on measured DNA methylation will be partially mediated by effects of phenotype on cell composition. For example, if a phenotype alters the immune system, then DNA methylation measured in blood will register both the indirect effects of the phenotype on the immune system as well as any direct effect not mediated by cell composition. When the direct effects are of principal interest in a study, then the cell-composition effects will represent a confound of the direct effects if they are not taken into account. This issue has been highlighted in numerous recent publications (Jaffe and Irizarry 2014; Koestler et al. 2012; Langevin et al. 2012, 2014; Li et al. 2014).

## 2.2 Fundamental Concepts

Much has been written about mediation and confounding, which are interrelated but distinct concepts (Robins and Greenland 1992; Pearl 2009; VanderWeele 2009). However, linear analysis is sufficient to untangle direct and mediated effects when (1) there is no modification of the effect of the independent variable (phenotype) on dependent variable (DNA methylation) by the mediator (cell composition) and (2) errors in the measurement of mediator (cell composition) and dependent variable (DNA methylation) are uncorrelated. Under these assumptions, several techniques are currently available for analyzing DNA methylation data while accounting for cellular heterogeneity. All of them assume essentially the following linear model for $m$ CpG loci measured on $n$ subjects:

$$\mathbf{Y} = \mathbf{B}\mathbf{X}^{\mathrm{T}} + \mathbf{M}\boldsymbol{\Omega}^{\mathrm{T}} + \mathbf{E}, \tag{2.1}$$

where $\mathbf{Y}$ is an $m \times n$ matrix of average beta values, $\mathbf{X}$ is an $n \times d$ design matrix of phenotype variables and potential confounders (for a total of $d$ covariates including an intercept), $\mathbf{B}$ is the $m \times d$ matrix of regression coefficients representing direct effects, $\mathbf{M}\boldsymbol{\Omega}^{\mathrm{T}}$ represents a linear mixture effect, with $\mathbf{M}$ an $m \times k$ matrix representing $m$ CpG-specific methylation states for $k$ cell types, $\boldsymbol{\Omega}$ is an $n \times k$ matrix representing subject-specific cell-type distributions (each row representing the cell-type proportions for a given subject), and $\mathbf{E}$ is an $m \times n$ matrix of errors with $E(\mathbf{E}) = \mathbf{0}_{m \times n}$. Note that the value $k$ is assumed to be known in advance. Note also that the entries of $\mathbf{Y}$, of $\mathbf{M}$, and of $\boldsymbol{\Omega}$ are assumed to lie in the unit interval, and that the rows of $\boldsymbol{\Omega}$ sum to one. In addition, we assume $\boldsymbol{\Omega}$ is a random variable that is potentially associated with $\mathbf{X}$. Although a Dirichlet model would most appropriately model the rows of $\boldsymbol{\Omega}$, we assume the following linear model as a computationally efficient approximation:

$$\boldsymbol{\Omega} = \mathbf{X}\boldsymbol{\Gamma} + \boldsymbol{\Xi}, \tag{2.2}$$

where $\boldsymbol{\Gamma}$ is a $d \times k$ matrix of covariate effects upon cell proportion and $\boldsymbol{\Xi}$ is an $n \times k$ error matrix. Figure 2.1 depicts these quantities in the context of mediation. Note that Eq. (2.1) explicitly omits interaction between $\mathbf{X}$ and $\boldsymbol{\Omega}$. With the additional assumption that $\mathbf{E}$ and $\boldsymbol{\Xi}$ are independent (and independent of $\mathbf{X}$), linear regression is sufficient for studying the mediation of phenotype effects on DNA methylation by cell composition. In particular, substituting (2.2) in (2.1),

$$\mathbf{Y} = \mathbf{B}\mathbf{X}^{\mathrm{T}} + \mathbf{M}\boldsymbol{\Omega}^{\mathrm{T}} + \mathbf{E} = \left(\mathbf{B} + \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}\right)\mathbf{X}^{\mathrm{T}} + \left(\mathbf{M}\boldsymbol{\Xi} + \mathbf{E}\right), \tag{2.3}$$

the total effect of $\mathbf{X}$ upon $\mathbf{Y}$ is $E\left(\mathbf{Y}\middle|\mathbf{X}\right) = \left(\mathbf{B} + \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}\right)\mathbf{X}^{\mathrm{T}}$, the direct effect is $\mathbf{B}\mathbf{X}^{\mathrm{T}}$, and the mediated, or *cell-composition effect*, is $\boldsymbol{\Delta}\mathbf{X}^{\mathrm{T}}$, where $\Delta = \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}$. Note that the error term for the total effects model is $\mathbf{M}\boldsymbol{\Xi} + \mathbf{E}$, which includes a term that depends on the cell-type-specific coefficient matrix $\mathbf{M}$.

In the remainder of this chapter, we present methods for estimating the total, direct, and cell-composition effects. We present both *reference-based* methods, that is, those relying on the availability of an external reference data set for estimating the matrix $\mathbf{M}$, and *reference-free* methods, those that do not require such reference data and treat $\mathbf{M}$ as essentially unknown.

## 2.3 Reference-Based Methods

When $\mathbf{\Omega}$ is known through explicitly measured cell counts, then it can be absorbed into the covariate matrix after deleting one of the cell types (in order to circumvent over-parameterization of the design matrix); subsequently, simple linear model methods such as limma (Smyth 2004) can be employed for analysis. For example, when a single cell type is being analyzed, $\mathbf{\Omega} = \mathbf{1}_n$, and cell type can effectively be ignored. Examples of single-cell-type studies include an analysis of DNA methylation associations with diabetes in CD14+ monocytes (Rakyan et al. 2011b) as well as associations between DNA methylation and autism in ectodermal cells (Berko et al. 2014). Alternatively, leukocyte counts may be available through standard complete blood count (CBC) methods and converted to proportions to obtain $\mathbf{\Omega}$, although standard methods will typically provide only coarse categories, for example, grouping all lymphocyte types together. Generally, finely differentiated cell counts can be obtained using cell sorting methods such as fluorescence-activated cell sorting (FACS) or magnetic-activated cell sorting (MACS). DNA methylation in a community cohort was characterized for peripheral blood mononuclear cells (PBMCs), accompanied by CBC counts (Lam et al. 2012). Another recent example demonstrated associations of DNA methylation with depression in postmortem brains using proportions of neuron and glial cells (Guintivano et al. 2013). Note that some mRNA expression analyses of blood have incorporated FACS measurements of individual leukocyte counts (Shen-Orr et al. 2010), but to date there are no major analyses of DNA methylation data in whole blood or PBMCs that have incorporated comprehensive FACS or MACS counts.

In many studies, it may be infeasible to obtain direct measures of cell counts. Fortunately, DNA methylation measurements themselves may be used to obtain approximate cell proportion estimates, as long as a reference data set is available for measuring the cell-type-specific mean methylation for a set of CpG loci that differentiate the types with a high degree of sensitivity and specificity. We have referred to such loci as *pseudo*-DMRs, since they are single locus markers rather than regions, although they are also commonly known as differentially methylated *positions* (DMPs). Interest in the detection of DMRs and DMPs for specific types of leukocytes has arisen from the study of tumor infiltration by lymphocytes (Accomando et al. 2012; Wiencke et al. 2012); this, in turn has led to more comprehensive characterization of genome-wide DNA methylation profiles for major types of leukocytes. Existing reference sets include an Infinium 27K data set (Houseman et al. 2012) as well as an Infinium 450K data set (Reinius et al.

2012). These data sets can be deployed to obtain estimates $\widehat{\boldsymbol{\Omega}}$ of cell proportions, as Houseman et al. (2012) have shown. The method is briefly described as follows.

Suppose $S$ is an ordered set of DMP loci for distinguishing $k$ cell types, $\mathbf{y}_l^{(S)}$ is a DNA methylation measurement on the set $S$ for a purified sample of type $l \in \{1, \ldots, k\}$, and $E\left(\mathbf{y}_l^{(S)}\right) = \boldsymbol{\mu}_l^{(S)}$ for a vector $\boldsymbol{\mu}_l^{(S)}$ whose elements fall in the unit interval. If $\mathbf{M}^{(S)} = \left[\boldsymbol{\mu}_1^{(S)}, \ldots, \boldsymbol{\mu}_k^{(S)}\right]$ and $\mathbf{y}_*^{(S)}$ is a vector of DNA measurements on $S$ for a heterogeneous tissue sample of mixed cell types, type $l$ representing proportion $\omega_l \geq 0$ of the tissue sample $\left(\sum_{l=1}^{k} \omega_l \leq 1\right)$, then $E\left(\mathbf{y}_*^{(S)}\right) = \mathbf{M}^{(S)}\boldsymbol{\omega}$, where $\boldsymbol{\omega}^{\mathrm{T}} = [\omega_1, \ldots, \omega_k]$. It follows that $\boldsymbol{\omega}$ can be estimated by minimizing the quantity $\left\|\mathbf{y}_*^{(S)} - \mathbf{M}^{(S)}\boldsymbol{\omega}\right\|^2$; although this problem is easily solved by computing the least squares estimator for $\boldsymbol{\omega}$, slightly better results can be obtained by imposing the natural constraints $\omega_l \geq 0$ and $\sum_{l=1}^{k} \omega_l \leq 1$ onto the solution space. Quadratic programming (Goldfarb and Idnani 1983) can easily be employed to obtain an estimate $\widehat{\boldsymbol{\omega}}$ that obeys these constraints. This *cellular deconvolution* method was initially shown to work well in recovering proportions of artificial blood mixtures (Houseman et al. 2012). Subsequent validation studies have demonstrated acceptable performance of cellular deconvolution of DNA methylation data. Comparing estimated proportions of monocytes within PBMC samples (which lack granulocytes) obtained from a community cohort (Lam et al. 2012) to their corresponding CBC-derived quantities, Koestler et al. (2013b) measured a root-mean-square-error (RMSE) of approximately 5 percentage points (Koestler et al. 2013b). In a comprehensive analysis of six donor blood samples with counts measured using three distinct FACS techniques, Accomando et al. (2014) estimated a RMSE of about 3.0–4.3 percentage points for six distinct leukocyte subtypes; when compared with each other, the FACS methods produced RMSE values of approximately 2 percentage points (i.e., only slightly smaller magnitude) (Accomando et al. 2014). First popularized in a study of rheumatoid arthritis (Liu et al. 2013), the method has become a widely adopted method for estimating cell proportions when individual count data are unavailable.

The method is available in the R/Bioconductor package *minfi* (function *EstimateCellCounts*). The *minfi* library also supports mutual normalization of reference and target data sets, which leads to some improvement in the estimation of cell proportions. The R/bioconductor package *FlowSorted.Blood.450k* encapsulates the 450K leukocyte reference data set published by Reinius et al. (2012); a 27K leukocyte reference data set is available on Gene Expression Omnibus (GEO), accession number GSE39981.

Note that $\mathbf{M}^{(S)}$ should represent a reasonably exhaustive characterization of the cell types comprising the tissue to be analyzed, in that the $k$ profiled types represent the major portion of each sample (Houseman et al. 2012). Under these circumstances, the sum $\sum_{l=1}^{k} \omega_l$ will typically lie close to 1 for each sample.

Consequently, when incorporated into the design matrix $\mathbf{X}$ of Eq. (2.1) for data analysis, the matrix $\widehat{\mathbf{\Omega}}$ derived from these measures should omit one of the types, otherwise the resulting design matrix will exhibit poor conditioning and lead to unstable estimates. For example, in analyzing whole blood, the granulocyte proportion might be omitted, and in analyzing PBMC samples, the monocyte proportion might be omitted.

Note also that the cell-composition term $\mathbf{M}\mathbf{\Omega}^{\mathrm{T}}$ in Eq. (2.1) entails a linear mixing assumption that is most plausible for measurement scales which correspond to fractions of cells or molecules. Consequently, cellular deconvolution should always be performed on the average beta scale instead of a popular alternative, the *M-value* scale obtained by logit-transforming the average beta. In addition, genome-wide application of Eq. (2.1) is likely to produce slightly better fit to data when beta values are used instead of *M*-values. However, use of average beta values in regression analysis is complicated slightly by the non-normal nature of the error term. For mid-range values, beta values and *M*-value will covary in an approximately linear fashion, so that both scales will return similar results for loci that exhibit great sensitivity to cell composition (i.e., DMPs). An alternative to applying Eq. (2.1) directly is to remove the cell-composition effects on the beta scale before implementing genome-wide regression analysis on the *M*-value scale. This strategy is consistent with *removal of unwanted variability* (RUV) (Gagnon-Bartsch and Speed 2012; Jaffe and Irizarry 2014). In this approach, $\widehat{\mathbf{M}}$ is obtained by fitting the genome-wide DNA methylation data to the equation $\mathbf{Y} = \mathbf{M}\widehat{\mathbf{\Omega}}^{\mathrm{T}} + \mathbf{E}$, each column $\mathbf{y}$ of $\mathbf{Y}$ is adjusted for cell composition via $\mathbf{y}^{(adj)} \leftarrow \mathbf{y} - \widehat{\mathbf{M}}\,(\widehat{\boldsymbol{\omega}} - \overline{\boldsymbol{\omega}})$ (where $\widehat{\boldsymbol{\omega}}$ is the corresponding column of $\widehat{\mathbf{\Omega}}$, $\overline{\boldsymbol{\omega}} = n^{-1}\widehat{\mathbf{\Omega}}^{\mathrm{T}}\mathbf{1}_n$ is the average cell proportion profile), and each resulting adjusted value is logit-transformed to an *M*-value, $m_j^{(adj)} \leftarrow \log_2\left(\max\left\{y_j^{(adj)}, \varepsilon\right\}\right) - \log_2\left(\max\left\{1 - y_j^{(adj)}, \varepsilon\right\}\right)$, with $\varepsilon$ a small value chosen to avoid infinite *M*-values. Note that centering $\widehat{\boldsymbol{\omega}}$ by $\overline{\boldsymbol{\omega}}$ is necessary to avoid a non-negligible proportion of adjusted values $y_j^{(adj)}$ lying outside the unit interval, as the resulting values of $y_j^{(adj)}$ will be centered around the average DNA methylation value.

Finally, we note that associations between $\mathbf{X}$ and $\mathbf{\Omega}$ may be of scientific interest. Analysis is straightforward when $\mathbf{\Omega}$ is measured directly. However, when $\mathbf{\Omega}$ is estimated via cellular deconvolution, it is desirable to account for all sources of variability, including the contribution of measurement error from the reference data set. Houseman et al. (2012) describe a comprehensive method for conducting such analysis.

## 2.4 Reference-Free Methods

Although the method of Houseman et al. (2012) provides an algorithm for estimating cell proportions $\mathbf{\Omega}$ from DNA methylation data, it requires the existence of a reference data set. To date, such data sets exist only for blood (Accomando et al.

2012; Houseman et al. 2012; Reinius et al. 2012) and, to a limited extent, brain tissue (Guintivano et al. 2013). However, other tissues are of interest in EWAS. For example, population-based studies of DNA methylation have been published with DNA methylation measured in placenta (Banister et al. 2011; Suter et al. 2011; Wilhelm-Benartzi et al. 2012), umbilical cord tissue (Teh et al. 2014), and (with sparser arrays) buccal swabs (Breton et al. 2009; Kaminsky et al. 2009); no reference sets currently exist for these and other tissues of interest (e.g., adipose tissue).

To circumvent this problem, Houseman et al. (2014) propose a method for approximating the 2012 method. This method also assumes Eq. (2.1), but treats the matrix $\mathbf{M}$ as unknown. The method works by first fitting the model for total effects,

$$\mathbf{Y} = \mathbf{B}^*\mathbf{X}^{\mathrm{T}} + \mathbf{E}^*$$

where $\mathbf{B}^* = \mathbf{B} + \mathbf{M}\boldsymbol{\Gamma}^{\mathrm{T}}$ and $\mathbf{E}^* = \mathbf{M}\boldsymbol{\Xi}^{\mathrm{T}} + \mathbf{E}$, as evident from Eq. (2.3). Note that

$$\mathbf{R} = \left[\mathbf{B}^*, \mathbf{E}^*\right] = \mathbf{M}\left[\boldsymbol{\Gamma}^{\mathrm{T}}, \boldsymbol{\Xi}^{\mathrm{T}}\right] + [\mathbf{B}, \mathbf{E}]\,. \tag{2.4}$$

With $k$, the number of assumed cell types, chosen in advance by prior biological knowledge or using a method for estimating the number of factors in a factor-analytic model [e.g., using random matrix theory (Teschendorff et al. 2011)], the method associates the largest $k$ singular values of $\mathbf{R}$ with cell-composition effects. Specifically, applying a singular value decomposition (SVD) to $\widehat{\mathbf{R}} = \left[\widehat{\mathbf{B}}^*, \widehat{\mathbf{E}}^*\right]$, $\widehat{\mathbf{R}} = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^{\mathrm{T}} + \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^{\mathrm{T}}$, where $\mathbf{U}_1$ is an orthogonal $m \times k$ matrix, $\mathbf{U}_2$ is an orthogonal $m \times (n-k)$ matrix, $\mathbf{U}_1^T\mathbf{U}_2 = \mathbf{0}_{k\times(n-k)}$, $\mathbf{V}_1$ is an orthogonal $n \times k$ matrix, $\mathbf{V}_2$ is an orthogonal $n \times (n-k)$ matrix, $\boldsymbol{\Lambda}_1$ is a diagonal $k \times k$ matrix, $\boldsymbol{\Lambda}_2$ is a diagonal $(n-k)\times(n-k)$ matrix, and the two terms separate the $k$ largest singular values from the $n-k$ smallest ones (i.e., every diagonal element of $\boldsymbol{\Lambda}_1$ is larger than every diagonal element of $\boldsymbol{\Lambda}_2$), it is assumed that $\mathbf{M}\left[\boldsymbol{\Gamma}^{\mathrm{T}}, \boldsymbol{\Xi}^{\mathrm{T}}\right] = \mathbf{U}_1\boldsymbol{\Lambda}_1\mathbf{V}_1^{\mathrm{T}}$ and $[\mathbf{B}, \mathbf{E}] = \mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^{T}$. Note that the two terms on the right hand side of Eq. (2.4) must be orthogonal in order for this identity to hold; to ensure orthogonality it is sufficient to assume $\mathbf{M}^T\mathbf{E} = \mathbf{0}_{k\times n}$ and $\mathbf{M}^T\mathbf{B} = \mathbf{0}_{k\times d}$. The former condition is an essential assumption entailed by the linear regression represented by Eq. (2.1); the latter assumption, that "indirect" effects lie in a space orthogonal to the cell-type-specific profiles, represents an unverifiable biological condition also necessary for the deconvolution method of Houseman et al. (2012), although the Supplement of the 2012 paper argues that orthogonality will approximately hold if the effects in $\mathbf{B}$ are relatively sparse. Note also that association of the $k$ largest singular values with cell-composition effects represents another biological assumption, that the cell-composition effects will dominate the linear associations evident in the array. Under the assumptions just described, $\widehat{\mathbf{B}}$ is obtained by selecting the first $d$ columns of $\mathbf{U}_2\boldsymbol{\Lambda}_2\mathbf{V}_2^{\mathrm{T}}$. Note that $\widehat{\Delta} = \widehat{\mathbf{B}}^* - \widehat{\mathbf{B}}$ represents the matrix of coefficients that explain the cell-mediated associations between $\mathbf{X}$ and $\mathbf{Y}$, which may be of interest in some studies.

Houseman et al. (2014) also propose a method for generating bootstrap samples from the sampling distribution of $\widehat{\mathbf{B}}^*$ and $\widehat{\mathbf{B}}$, from which standard errors for $\widehat{\mathbf{B}}^*$, $\widehat{\mathbf{B}}$, and $\widehat{\mathbf{\Delta}}$ can be estimated. Briefly, the method generates a bootstrap sample $\mathbf{Y}^{(b)}$ of DNA methylation average beta values as $\mathbf{Y}^{(b)} = \widehat{\mathbf{B}}^*\mathbf{X}^T + \mathbf{E}^{(b)}$, where $\widehat{\mathbf{B}}^*$ is the estimated coefficient of total effects and $\mathbf{E}^{(b)}$ is a bootstrap error matrix constructed element-wise as $e_{ij}^{(b)} = q_{ij}^{(b)}\sqrt{\widehat{\mu}_{ij}\left(1-\widehat{\mu}_{ij}\right)}$, where $\widehat{\mu}_{ij}$ is the element of $\widehat{\mathbf{B}}^*\mathbf{X}^T$ corresponding to the $i^{\text{th}}$ column and $j^{\text{th}}$ row, $q_{ij}^{(b)}$ is the element of the matrix obtained by sampling with replacement from the columns of $\mathbf{Q}$ , each of whose elements $q_{ij}$ were obtained from $\widehat{\mathbf{E}}^* = \left(\widehat{e}_{ij}\right)$ and $\widehat{\mathbf{B}}^*\mathbf{X}^T$ as $q_{ij} = \widehat{e}_{ij}^*/\sqrt{\widehat{\mu}_{ij}\left(1-\widehat{\mu}_{ij}\right)}$. The method factors the error $\mathbf{E}^*$ element-wise as the product of a mean-dependent scaling factor $\sqrt{\widehat{\mu}_{ij}\left(1-\widehat{\mu}_{ij}\right)}$ and a "dispersion" value $q_{ij}$; this strategy respects the approximate beta distribution of $\mathbf{Y}$, while simultaneously preserving correlation across the rows (CpGs). The estimation method, as well as its corresponding bootstrap generation procedure, is publicly available in an R package entitled *RefFreeEWAS*.

The 2014 method is similar to *surrogate variable analysis* (SVA) (Leek and Storey 2007; Teschendorff et al. 2011), which uses a factor-analytic decomposition similar to Eq. (2.1) but applies SVD or *independent components analysis* (ICA) to the error term $\hat{\mathbf{E}}^*$ rather than $\widehat{\mathbf{R}} = \left[\widehat{\mathbf{B}}^*, \widehat{\mathbf{E}}^*\right]$, thus potentially missing linear effects that are explicitly the result of cell composition. It is also similar in spirit to the recently published *Ewasher* method (Zou et al. 2014); this method models the phenotype as a function of methylation and potentially other confounding covariates, instead of modeling methylation as a function of phenotype and potential confounders. Specifically, the following model is assumed:

$$\mathbf{x} = \beta_j^{(Y)}\mathbf{y}_j + \mathbf{Z}^T\boldsymbol{\beta}_j^{(Z)} + m^{-1/2}\tilde{\mathbf{Y}}^T\mathbf{u} + \mathbf{e}_j, \tag{2.5}$$

where $\mathbf{x}$ is the $n \times 1$ matrix of subject phenotypes (dichotomous or continuous), $\mathbf{y}_j$ is the $n \times 1$ matrix of DNA methylation value measured for each at CpG $j$, $\mathbf{Z}$ is a $n \times d'$ matrix of potential confounders for each subject (including an intercept term), $\tilde{\mathbf{Y}}$ is the $m \times n$ matrix of standardized DNA methylation values obtained from $\mathbf{Y}$ by standardizing each row (CpG), $\mathbf{u}$ is an $m \times 1$ matrix of Gaussian random effects, each having variance $\sigma_u^2$ and uncorrelated across entries, $\mathbf{e}_j$ is an $n \times 1$ matrix of independent errors having variance $\sigma_{e,j}^2$, and $\beta_j^{(Y)}$ and $\boldsymbol{\beta}_j^{(Z)}$ are coefficients to be estimated. Estimation proceeds by considering the multivariate distribution of $\mathbf{x}$, whose variance–covariance matrix is $\mathbf{\Sigma}_x = m^{-1}\sigma_u^2\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}} + \sigma_{e,j}^2\mathbf{I}_{n\times n}$. Note that if $\tilde{\mathbf{Y}} = \tilde{\mathbf{M}}\mathbf{\Omega}^T$ captures the rescaled cell-composition effects, then $\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}} = \mathbf{\Omega}\tilde{\mathbf{M}}^T\tilde{\mathbf{M}}\mathbf{\Omega}^T$, which is essentially the contribution to $\mathbf{\Sigma}_x$ that would result from substituting the explicit cell-composition effect $\tilde{\mathbf{M}}\mathbf{\Omega}^T$ for $m^{-1/2}\tilde{\mathbf{Y}}^T\mathbf{u}$ in Eq. (2.5). Thus, the term $m^{-1/2}\tilde{\mathbf{Y}}^T\mathbf{u}$ captures cell-composition effects in a manner similar to the approach based on Eq. (2.1).

Note that these reference-free methods entail strong linearity assumptions and, ultimately, assumptions about the relationship between measured DNA methylation and the actual proportion of methylated cytosine molecules among the specific targeted loci. Consequently, the technical properties of the assay to be used should be considered carefully, and analysis should be preceded by the execution of a pre-processing pipeline that results in DNA measurements that are as comparable as possible across loci. For example, use of the popular 450K assay should entail proper normalization (Marabita et al. 2013), alignment of the biochemically distinct Type I and Type II probes (Dedeurwaerder et al. 2011; Teschendorff et al. 2013), and removal of loci whose probes contain common variants or cross-hybridize across the genome (Chen et al. 2013).

## 2.5 Data Examples

Several published analyses of DNA methylation data have employed the methods described above to adjust for heterogeneity in cell composition. Guintivano et al. (2014) incorporated blood count data to adjust for cellular heterogeneity in association between DNA methylation measured in blood and postpartum depression (Guintivano et al. 2014). Liu et al. (2013) published the first analysis that employed the Houseman et al. (2012) method of estimating cell proportions from DNA methylation data, demonstrating marked attenuation of significance in association of DNA methylation measured in blood with rheumatoid arthritis after adjusting for estimated cell proportions (Liu et al. 2013). Similarly, in a perinatal study of arsenic exposure in Bangladesh, Kile et al.(2014) demonstrated marked attenuation of significance in association of DNA methylation measured in cord blood with ingestion of inorganic arsenic via drinking water after adjusting for cell proportions, additionally suggesting that arsenic exposure could alter the proportion of CD4+ and CD8+ T lymphocytes (Kile et al. 2014). Koestler et al. (2013a, b) demonstrated association of cord blood methylation and urinary inorganic arsenic concentration after adjusting for cell proportion (Koestler et al. 2013a). Finally, Jaffe and Irizarry (2014) employed several methods including the Houseman et al. (2012) method to demonstrate that the commonly acknowledged association between age and DNA methylation can be explained in large part by age-related changes in cell composition (Jaffe and Irizarry 2014).

Using two data sets, we briefly compare and contrast some of the methods described in this chapter: the community cohort data published by Lam et al. (2012) and re-analyzed by Koestler et al. (2013a, b), and the rheumatoid arthritis data set published by Liu et al. (2013) and re-analyzed by Houseman et al. (2014) and Zou et al. (2014). See Houseman et al. (2014) for additional details.

For 26,486 autosomal CpG sites assayed by the 27K array, Fig. 2.2 shows quantile-quantile (QQ) plots on a logarithmic scale comparing a uniform distribution against nominal $p$-values obtained using several different methods: unadjusted ("Unadj", representing total effect $\mathbf{B}^*$), reference-based ["Ref", representing direct

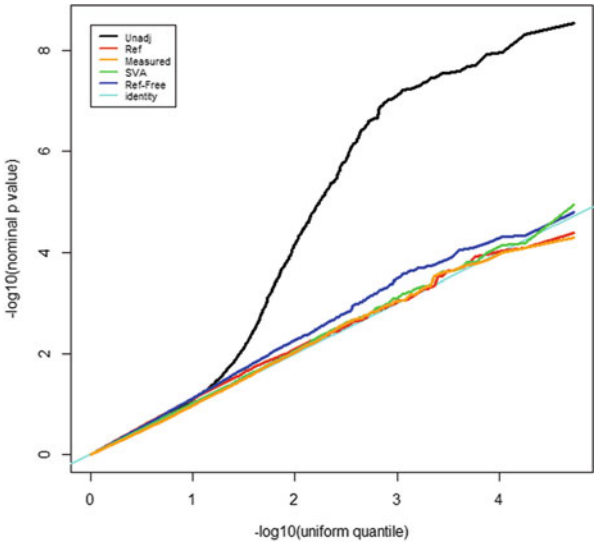


**Fig. 2.2** Analysis of DNA methylation and IL-6 response bioassay in a community cohort

effect **B** obtained by applying the method of (Houseman et al. 2012), to obtain cell proportion estimates $\widehat{\mathbf{\Omega}}$], a direct effect based on monocyte/lymphocyte proportions measured by CBC ("Measured"), a direct effect estimate based on SVA ("SVA") with $k = 11$ assumed surrogate variables, and a direct effect estimate based on the reference-free approach of Houseman et al. (2014) with $k = 10$ (see the original article for details on the choice of $k$). Each $p$-value represents significance of association between DNA methylation in PBMCs measured on an average beta scale and IL-6 response to phorbol-12-myristate-13-acetatein. All methods except the unadjusted method result in $p$-values that are effectively uniform (i.e., characteristic of a null effect). This suggests that there may be a strong total effect of the IL-6 phenotype on DNA methylation, but that this effect is explained by alterations in monocyte/lymphocyte proportions and accounted for using the reference-based and SVA methods. Note that Fig. 2.2 suggests a small number of CpGs with slightly elevated significance for the reference-free method; however, the distribution of $p$-values across the 26,486 CpGs is consistent with a uniform distribution, as Fig. 2.3 implies. Figure 2.3 shows the QQ plots for unadjusted and reference-free methods, superimposed upon 95 % probability bands representing their corresponding null distributions obtained from 1,000 bootstrap estimates using a method suggested in the supplementary material of Houseman et al. (2014). This plot suggests significant modification of total DNA methylation by the IL-6 phenotype, but no significant alteration after accounting for covariation in monocytes. Figure 2.4 compares significance of the $\widehat{\Delta}$ coefficients from the reference-free method with significance of the monocyte coefficients from the linear model using only the measured monocyte proportions. There is high concordance in significance between the two methods; by Fisher's exact test, concordance of $p$-values less than 0.001 is quite high (odds ratio = 47.5, 95 % confidence interval: 21.1–106, Fisher $p < 10^{-16}$). Thus, this

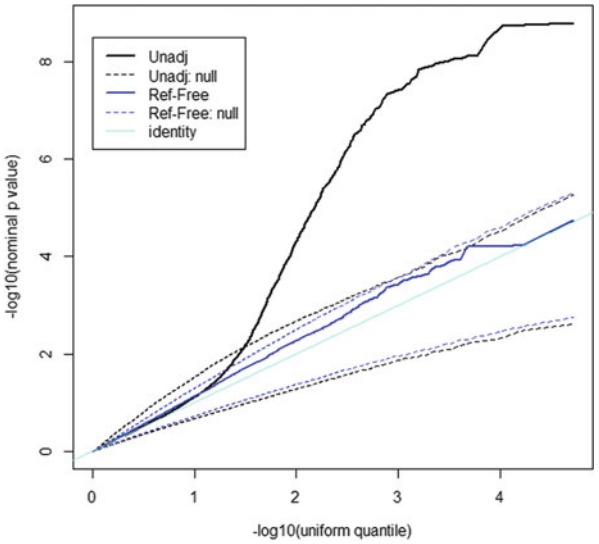


**Fig. 2.3** Analysis of DNA methylation and IL-6 response bioassay in a community cohort: comparison with bootstrap-based null sampling distribution

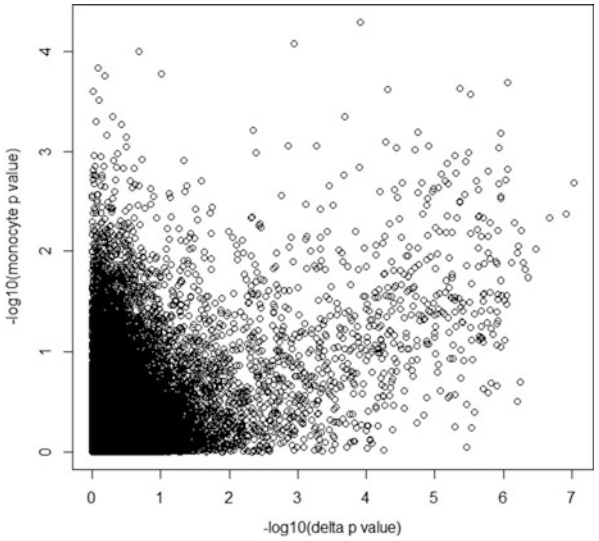


**Fig. 2.4** Analysis of DNA methylation and IL-6 response bioassay in a community cohort: comparison of significance of cell-composition effects from reference-free methods with significance of effects of known monocyte proportions

analysis demonstrates how $\Delta$ coefficients can be used to identify DMPs for distinct cell types within a sample. This strategy was used in a recent article evaluating the effect of cellular heterogeneity on breast tissue (Houseman and Ince 2014).

For 384,410 autosomal CpG sites assayed by the 450K array and having probes free of common variants, Fig. 2.5 shows QQ plots on a logarithmic scale comparing a uniform distribution against nominal $p$-values obtained using the same methods as for Fig. 2.2, except for the "Measured" method since measured cell counts were unavailable for this data set. Additionally, for SVA, $k = 53$ surrogate variables were

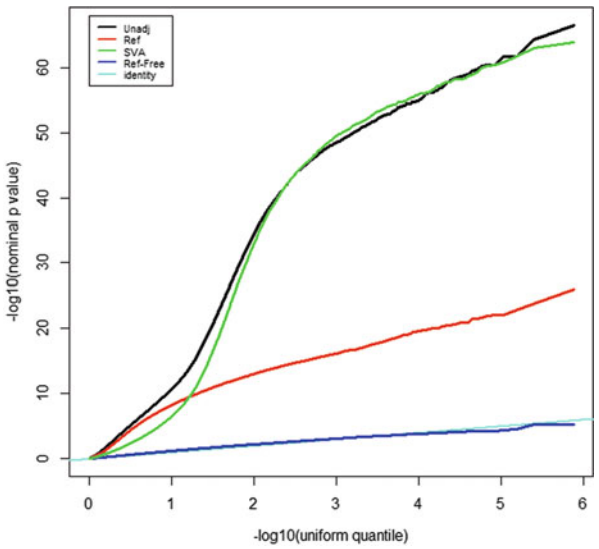


**Fig. 2.5** Analysis of DNA methylation and rheumatoid arthritis

assumed, and for the reference-based method, $k = 37$ cell types were assumed; these values were based on application of appropriate dimension-estimating algorithms (Houseman et al. 2014). Each $p$-value represents significance of association between rheumatoid arthritis case status and DNA methylation in whole blood measured on an average beta scale. The unadjusted and SVA-adjusted methods result in QQ plots reflecting strong significance; the QQ plot from the reference-based approach reflects attenuated but still moderately strong significance; and the reference-free approach reflects null association. As previously suggested (Houseman et al. 2014), the reference-free approach may be capturing subtle shifts in proportions of cell types not profiled in the reference data set used for the reference-based adjustment. Note that while SVA was adequate for cell-composition adjustment in the previous analysis, it was insufficient for the present one.

## 2.6 Conclusions

Heterogeneity in cell type is an important consideration in the analysis of DNA methylation measured from complex tissues. In many applications, the phenotype of interest may alter the composition of cell types within the target tissue, thus altering DNA methylation profile independently of specific molecular alterations that are not mediated by cell type. Therefore, proportions of each cell type should be included in models for phenotypic effects of DNA methylation. In the best-case scenario, proportions of each cell type will be available for each sample. However, since the cell sorting techniques necessary for measuring these proportions can be costly,

many studies lack these measurements. In such a situation, the cell proportions can be estimated directly from DNA methylation data if a reference data set exists for the cell types that constitute the target tissue. If no such reference data set exists, recently published reference-free methods can be used to account for cellular heterogeneity when estimating phenotype associations with DNA methylation, although more work is needed to validate these new methods.

## References

Accomando WP, Wiencke JK, Houseman EA, Butler RA, Zheng S, Nelson HH, Kelsey KT. Decreased NK cells in patients with head and neck cancer determined in archival DNA. Clin Cancer Res. 2012;18(22):6147–54. doi:10.1158/1078-0432.CCR-12-1008.

Accomando WP, Wiencke JK, Houseman EA, Nelson HH, Kelsey KT. Quantitative reconstruction of leukocyte subsets using DNA methylation. Genome Biol. 2014;15(3):R50. doi:10.1186/gb-2014-15-3-r50.

Banister CE, Koestler DC, Maccani MA, Padbury JF, Houseman EA, Marsit CJ. Infant growth restriction is associated with distinct patterns of DNA methylation in human placentas. Epigenetics. 2011;6(7):920–7. doi:10.4161/epi.6.7.16079.

Baron U, Türbachova I, Hellwag A, Eckhardt F, Berlin K, Hoffmuller U, Gardina P, Olek S. Research paper DNA methylation analysis as a tool for cell typing. Epigenetics. 2006;1(1):55–60.

Berko ER, Suzuki M, Beren F, Lemetre C, Alaimo CM, Calder RB, Ballaban-Gil K, Gounder B, Kampf K, Kirschen J, Maqbool SB, Momin Z, Reynolds DM, Russo N, Shulman L, Stasiek E, Tozour J, Valicenti-McDermott M, Wang S, Abrahams BS, Hargitai J, Inbar D, Zhang Z, Buxbaum JD, Molholm S, Foxe JJ, Marion RW, Auton A, Greally JM. Mosaic epigenetic dysregulation of ectodermal cells in autism spectrum disorder. PLoS Genet. 2014;10(5):e1004402. doi:10.1371/journal.pgen.1004402.

Breton CV, Byun HM, Wenten M, Pan F, Yang A, Gilliland FD. Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation. Am J Respir Crit Care Med. 2009;180(5):462–7. doi:10.1164/rccm.200901-0135OC.

Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203–9. doi:10.4161/epi.23470.

Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R, Sugarbaker DJ, Yeh RF, Wiencke JK, Kelsey KT. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5(8):e1000602. doi:10.1371/journal.pgen.1000602.

Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450 K technology. Epigenomics. 2011;3(6):771–84. doi:10.2217/epi.11.105.

Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, Meduri E, Morange PE, Gagnon F, Grallert H, Waldenberger M, Peters A, Erdmann J, Hengstenberg C, Cambien F, Goodall AH, Ouwehand WH, Schunkert H, Thompson JR, Spector TD, Gieger C, Tregouet DA, Deloukas P, Samani NJ. DNA methylation and body-mass index: a genome-wide analysis. Lancet. 2014;383(9933):1990–8. doi:10.1016/S0140-6736(13)62674-4.

Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539–52. doi:10.1093/biostatistics/kxr034.

Goldfarb D, Idnani A. A numerically stable dual method for solving strictly convex quadratic programs. Math Program. 1983;27(1):1–33.

Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics. 2013;8(3):290–302. doi:10.4161/epi.23924.

Guintivano J, Arad M, Gould TD, Payne JL, Kaminsky ZA. Antenatal prediction of postpartum depression with blood DNA methylation biomarkers. Mol Psychiatry. 2014;19(5):560–7. doi:10.1038/mp.2013.62.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86. doi:10.1186/1471-2105-13-86.

Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014. doi:10.1093/bioinformatics/btu029.

Houseman EA, Ince TA. Normal cell-type epigenetics and breast cancer classification: a case study of cell mixture–adjusted analysis of DNA methylation data from tumors. Cancer Inform. 2014;13 Suppl 4:53.

Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014;15(2):R31.

Ji H, Ehrlich LI, Seita J, Murakami P, Doi A, Lindau P, Lee H, Aryee MJ, Irizarry RA, Kim K, Rossi DJ, Inlay MA, Serwold T, Karsunky H, Ho L, Daley GQ, Weissman IL, Feinberg AP. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature. 2010;467(7313):338–42. doi:10.1038/nature09367.

Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, Huang Z, Hoyo C, Midttun O, Cupul-Uicab LA, Ueland PM, Wu MC, Nystad W, Bell DA, Peddada SD, London SJ. 450 K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2012;120(10):1425–31. doi:10.1289/ehp.1205412.

Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A. DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet. 2009;41(2):240–5. doi:10.1038/ng.286.

Khavari DA, Sen GL, Rinn JL. DNA methylation and epigenetic control of cellular differentiation. Cell Cycle. 2010;9(19):3880–3.

Kile ML, Houseman EA, Baccarelli AA, Quamruzzaman Q, Rahman M, Mostofa G, Cardenas A, Wright RO, Christiani DC. Effect of prenatal arsenic exposure on DNA methylation and leukocyte subpopulations in cord blood. Epigenetics. 2014;9(5):774–82. doi:10.4161/epi.28153.

Kim M, Long TI, Arakawa K, Wang R, Yu MC, Laird PW. DNA methylation as a biomarker for cardiovascular disease risk. PLoS One. 2010;5(3):e9692. doi:10.1371/journal.pone.0009692.

Koestler DC, Marsit CJ, Christensen BC, Accomando W, Langevin SM, Houseman EA, Nelson HH, Karagas MR, Wiencke JK, Kelsey KT. Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. Cancer Epidemiol Biomarkers Prev. 2012;21(8):1293–302. doi:10.1158/1055-9965.EPI-12-0361.

Koestler DC, Avissar-Whiting M, Houseman EA, Karagas MR, Marsit CJ. Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. Environ Health Perspect. 2013a;121(8):971–7. doi:10.1289/ehp.1205925.

Koestler DC, Christensen B, Karagas MR, Marsit CJ, Langevin SM, Kelsey KT, Wiencke JK, Houseman EA. Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. Epigenetics. 2013b;8(8):816–26. doi:10.4161/epi.25430.

Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kobor MS. Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci U S A. 2012;109 Suppl 2:17253–60. doi:10.1073/pnas.1121249109.

Langevin SM, Koestler DC, Christensen BC, Butler RA, Wiencke JK, Nelson HH, Houseman EA, Marsit CJ, Kelsey KT. Peripheral blood DNA methylation profiles are indicative of head and neck squamous cell carcinoma: an epigenome-wide association study. Epigenetics. 2012;7(3):291–9. doi:10.4161/epi.7.3.19134.

Langevin SM, Houseman EA, Accomando WP, Koestler DC, Christensen BC, Nelson HH, Karagas MR, Marsit CJ, Wiencke JK, Kelsey KT. Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. Epigenetics. 2014;9(6):884–95.

Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35. doi:10.1371/journal.pgen.0030161.

Li H, Zheng T, Chen B, Hong G, Zhang W, Shi T, Li S, Ao L, Wang C, Guo Z. Similar blood-borne DNA methylation alterations in cancer and inflammatory diseases determined by subpopulation shifts in peripheral leukocytes. Br J Cancer. 2014. doi:10.1038/bjc.2014.347.

Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, Alfredsson L, Klareskog L, Ekstrom TJ, Feinberg AP. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31(2):142–7. doi:10.1038/nbt.2487.

Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, Sundberg CJ, Ekstrom TJ, Teschendorff AE, Tegner J, Gomez-Cabrero D. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8(3):333–46. doi:10.4161/epi.24008.

Marsit CJ, Koestler DC, Christensen BC, Karagas MR, Houseman EA, Kelsey KT. DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. J Clin Oncol. 2011;29(9):1133–9. doi:10.1200/JCO.2010.31.3577.

Natoli G. Maintaining cell identity through global control of genomic organization. Immunity. 2010;33(1):12–24. doi:10.1016/j.immuni.2010.07.006.

Pearl J. Causal inference in statistics: an overview. Stat Surv. 2009;3:96–146.

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011a;12(8):529–41. doi:10.1038/nrg3000.

Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, Daunay A, Busato F, Mein CA, Manfras B, Dias KR, Bell CG, Tost J, Boehm BO, Beck S, Leslie RD. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet. 2011b;7(9):e1002300. doi:10.1371/journal.pgen.1002300.

Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, Söderhäll C, Scheynius A, Kere J. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7(7):e41361.

Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992;3(2):143–55.

Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. Cell type-specific gene expression differences in complex tissues. Nat Methods. 2010;7(4):287–9. doi:10.1038/nmeth.1439.

Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):3.

Suter M, Ma J, Harris A, Patterson L, Brown KA, Shope C, Showalter L, Abramovici A, Aagaard-Tillery KM. Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. Epigenetics. 2011;6(11):1284–94. doi:10.4161/epi.6.11.17819.

Teh AL, Pan H, Chen L, Ong ML, Dogra S, Wong J, MacIsaac JL, Mah SM, McEwen LM, Saw SM, Godfrey KM, Chong YS, Kwek K, Kwoh CK, Soh SE, Chong MF, Barton S, Karnani N, Cheong CY, Buschdorf JP, Stunkel W, Kobor MS, Meaney MJ, Gluckman PD, Holbrook JD. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res. 2014. doi:10.1101/gr.171439.113.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, Jones A, Lechner M, Beck S, Jacobs IJ, Widschwendter M. An epigenetic signature in peripheral blood predicts active ovarian cancer. PLoS One. 2009;4(12):e8274. doi:10.1371/journal.pone.0008274.

Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics. 2011;27(11):1496–505. doi:10.1093/bioinformatics/btr171.

Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29(2):189–96.

VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology. 2009;20(1):18–26. doi:10.1097/EDE.0b013e31818f69ce.

Wiencke JK, Accomando WP, Zheng S, Patoka J, Dou X, Phillips JJ, Hsuang G, Christensen BC, Houseman EA, Koestler DC, Bracci P, Wiemels JL, Wrensch M, Nelson HH, Kelsey KT. Epigenetic biomarkers of T-cells in human glioma. Epigenetics. 2012;7(12):1391–402. doi:10.4161/epi.22675.

Wilhelm-Benartzi CS, Houseman EA, Maccani MA, Poage GM, Koestler DC, Langevin SM, Gagne LA, Banister CE, Padbury JF, Marsit CJ. In utero exposures, infant growth, and DNA methylation of repetitive elements and developmentally related genes in human placenta. Environ Health Perspect. 2012;120(2):296–302. doi:10.1289/ehp.1103927.

Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. Epigenome-wide association studies without the need for cell-type composition. Nat Methods. 2014;11(3):309–11. doi:10.1038/nmeth.2815.

http://www.springer.com/978-94-017-9926-3

Computational and Statistical Epigenomics
Teschendorff, A.E. (Ed.)
2015, V, 217 p. 42 illus., 41 illus. in color., Hardcover
ISBN: 978-94-017-9926-3