MINI REVIEW

BJC

British Journal of Cancer (2013) 109, 1394–1402 | doi: 10.1038/bjc.2013.496

Keywords: DNA methylation; microarray; processing; analysis; bioconductor and R packages

Review of processing and analysis methods for DNA methylation array data

C S Wilhelm-Benartzi[1,8], D C Koestler[2,8], M R Karagas[2], J M Flanagan[1], B C Christensen[2,3], K T Kelsey[4,5], C J Marsit[2,3], E A Houseman[6] and R Brown*[,1,7]

[1]*Epigenetics Unit, Division of Cancer, Department of Surgery and Cancer, Faculty of Medicine, Ovarian Cancer Action Research Centre, Imperial College London, 4th floor IRDB, Hammersmith Campus, Du Cane Road, London W12 0NN, UK;* [2]*Section of Biostatistics and Epidemiology, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA;* [3]*Department of Pharmacology and Toxicology, Geisel School of Medicine at Dartmouth College, Hanover, NH 03755, USA;* [4]*Department of Pathology and Laboratory Medicine, Brown University, Providence, RI, USA;* [5]*Department of Epidemiology, Brown University, Providence, RI, USA;* [6]*Department of Public Health, Oregon State University, Corvallis, OR, USA and* [7]*Section of Molecular Pathology, Institute for Cancer Research, Sutton, UK*

The promise of epigenome-wide association studies and cancer-specific somatic DNA methylation changes in improving our understanding of cancer, coupled with the decreasing cost and increasing coverage of DNA methylation microarrays, has brought about a surge in the use of these technologies. Here, we aim to provide both a review of issues encountered in the processing and analysis of array-based DNA methylation data and a summary of the advantages of recent approaches proposed for handling those issues, focusing on approaches publicly available in open-source environments such as R and Bioconductor. We hope that the processing tools and analysis flowchart described herein will facilitate researchers to effectively use these powerful DNA methylation array-based platforms, thereby advancing our understanding of human health and disease.

Epigenetic mechanisms associated with DNA methylation of cytosine residues at CpG dinucleotides have a central role in normal human development and disease (Baylin and Jones, 2011). Advancements in high-throughput assessment of DNA methylation using microarrays or second-generation sequencing-based approaches have enabled the quantitative profiling of DNA methylation of CpG loci throughout the genome. As well as profiling the methylome of tumour compared with normal tissue, this has ushered in the era of epigenome-wide association studies, analogous to the genome-wide association studies, aimed at understanding the epigenetic basis of complex diseases such as cancer. The promise of methylation profiling in improving our understanding of cancer coupled with the current trend of decreasing cost and increasing coverage of DNA methylation microarrays has brought about a surge in the use of these technologies.

Here, we aim to provide both a review of issues encountered in the processing and analysis of array-based DNA methylation data and a summary of recent approaches proposed for handling those issues. Excellent reviews in the field of epigenetics and technical aspects of array-based assessment of DNA methylation are available, although this is a constantly developing research area (Laird, 2010; Petronis, 2010; Baylin and Jones, 2011; Rakyan *et al*, 2011; Bock, 2012). We seek to update perspectives on statistical issues that arise in the processing and analysis of array-based DNA methylation data (Siegmund, 2011), highlighting more recent methods proposed for this purpose. The subheadings shown in Figure 1 form the basis for the topics highlighted in this review. Our aim is to help researchers understand the growing body of statistical methods for array-based DNA methylation data, focusing on those freely available in open-source environments such as R or Bioconductor (Table 1). For this review, we chose to focus on Illumina's BeadArray assays; however, many of the general considerations described here are applicable to other array technologies. We also aim to counter some of the perceived limitations of these arrays, that is, there are too many 'false

*Correspondence: Professor R Brown; E-mail: b.brown@imperial.ac.uk
[8]These authors contributed equally to this work.

Received 24 April 2013; revised 23 July 2013; accepted 30 July 2013; published online 27 August 2013

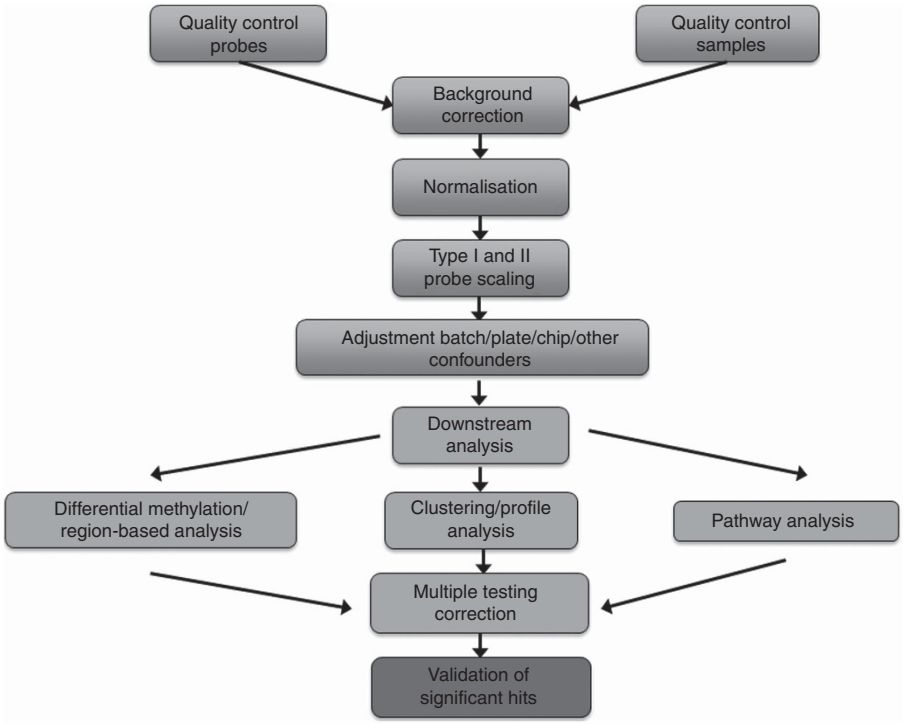

Figure 1. Methylation array data processing and analysis pipeline.

Table 1. R/Bioconductor packages for the processing and analysis of array-based DNA methylation data

DNA methylation processing/analysis step	R/Bioconductor packages
Quality control samples	IMA, HumMethQCReport, methylkit, MethyLumi, preprocessing and analysis pipeline, minfi
Quality control probes	IMA, HumMethQCReport, lumi, LumiWCluster, preprocessing and analysis pipeline, wateRmelon
Background correction	Limma, lumi, MethyLumi, minfi, preprocessing and analysis pipeline
Normalisation	Combat[a], HumMethQCReport, lumi, minfi, TurboNorm, MethyLumi, wateRmelon
Type 1 and 2 probe scaling	IMA, minfi, wateRmelon
Batch/plate/chip/confounder adjustment	Combat[a], CpGassoc, ISVA, MethLAB
Data dimension reduction	MethyLumi
Differential methylation analysis/region-based analysis	CpGAssoc, IMA, limma, methylkit, MethLAB, MethVisual, minfi, EVORA
Clustering/profile analysis	Lumi, ISVA, HumMeth27QCReport, methylkit, RPMM, SS-RPMM[b]
Multiple testing correction	CpGAssoc, methylkit, MethLAB, NHMMfdr

[a]Freely available for download: http://www.bu.edu/jlab/wp-assets/ComBat/Abstract.html.
[b]Freely available for download: http://bio-epi.hitchcock.org/faculty/koestler.html.

positives' in analysing microarray data (Ioannidis, 2007). We present the viewpoint that appropriate experimental design and downstream data processing and analysis pipelines will enable DNA methylation to be appropriately analysed and will help in understanding the pathogenesis of human disease.

ILLUMINA BEADARRAY TECHNOLOGY FOR METHYLATION

Illumina adapted its BeadArray technology for genotyping to recognise bisulphite-converted DNA for the interrogation of DNA methylation (Bibikova *et al*, 2011). The Illumina BeadArray assays use oligonucleotides conjugated to bead types to measure specific target sequences, measuring multiple beads per bead type. The bead types are summarised by the average signal for methylated (M) and unmethylated (U) alleles, and are used to compute the β-value, where

$$\beta = \frac{\mathrm{Max}(M,0)}{\mathrm{Max}(M,0)+\mathrm{Max}(U,0)+100}$$

A β-value of 0 equates to an unmethylated CpG site and 1 to a fully methylated CpG site. Illumina has developed three platforms for array-based assessment of DNA methylation: GoldenGate, Infinium Human Methylation27 and the Infinium HD 450K methylation array, which all use two fluorescent dye colours but differ in the chemistries used to recognise the bisulphite-converted sequence; however, we will focus on the Infinium arrays for the rest

of this work, as the GoldenGate array has been phased out from production. Furthermore, Illumina has developed their GenomeStudio software (Bibikova *et al*, 2011), which enables basic data analysis; however, for more in-depth analysis, many tools have been developed, as we will discuss below.

QUALITY CONTROL OF SAMPLES

The Infinium arrays include several control probes for determining the data quality, including sample-independent and -dependent controls (Illumina, 2011). To detect poorly performing samples in Illumina arrays, diagnostic plots of control probes in GenomeStudio are often used (Bibikova *et al*, 2011), and the R-package HumMethQCReport (Mancuso *et al*, 2011) also provides these plots. Figure 2 shows hybridisation and bisulphite conversion plots for 450K data in the green channel. Although the sample-independent and -dependent controls can be visually inspected to identify poor performing samples, an alternative approach involves using the raw signal intensities of the control probes and determining whether they are beyond the expected range (e.g., median ± 3 s.d.) of the signal intensities across all samples.

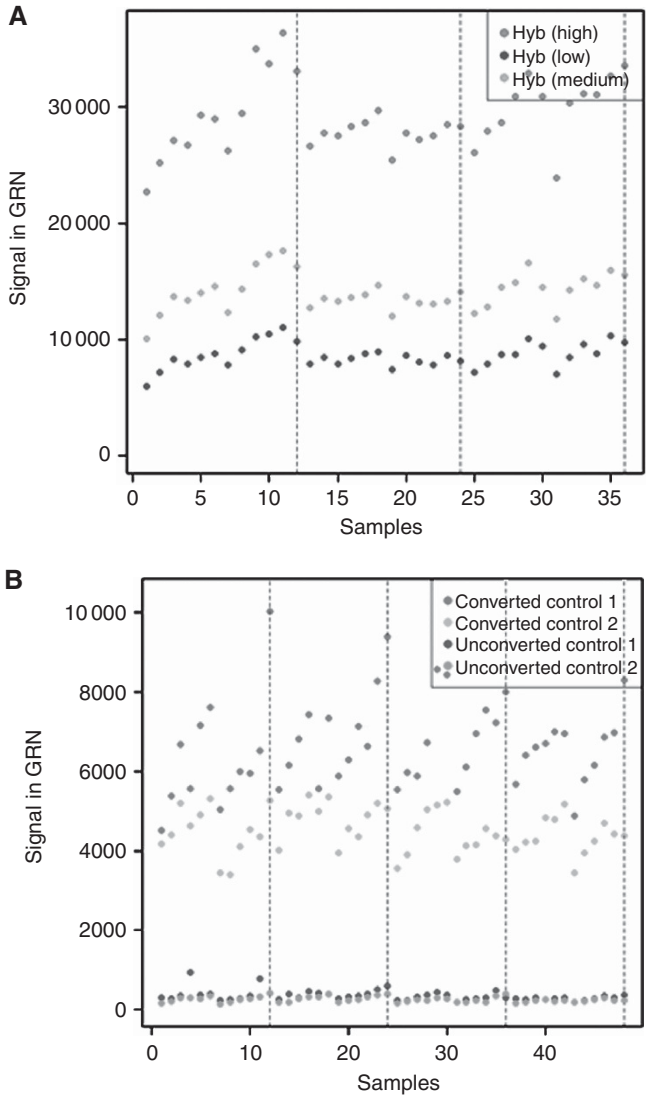

Figure 2. Quality control example from GenomeStudio 450K data. (**A**) Hybridisation quality control plot in the green channel. (**B**) Bisulphite conversion quality control plot in the green channel. In this example, the separation between high and low values indicates that hybridisation worked well. Furthermore, bisulphite conversion also performed well as converted controls have a higher signal than unconverted controls.

Other options for quality control of samples, which make use of detection *P*-values, are available in R and Bioconductor packages, such as the preprocessing and analysis pipeline (Touleimat and Tost, 2012), IMA (Wang *et al*, 2012), Minfi (Hansen and Aryee, 2013) and MethyLumi (Davis *et al*, 2011).

QUALITY CONTROL OF PROBES

Similar to sample quality control, it is customary to filter probes if a certain proportion of samples (i.e., >25%) have a detection *P*-value below a certain prespecified threshold (i.e., $P<0.05$) (Bibikova *et al*, 2011). In the IMA package (Wang *et al*, 2012), probes with missing values, those residing on the X chromosome, and those with a median detection *P*-value >0.05 across samples can be filtered out; other packages allowing such filtering include (Davis *et al*, 2011; Touleimat and Tost, 2012).

LumiWCluster (Kuan *et al*, 2010) includes a function for model-based clustering of methylation data using a weighted likelihood approach wherein higher-quality samples (i.e., those with a low median detection *P*-value) have larger weights and thus greater influence on the estimation of the mixture parameters for cluster inference. This approach avoids discarding probes, characteristic of hard-thresholding approaches, allowing the incorporation of all the data while accounting for the quality of individual observations.

A potential issue for quality control at the probe level stems from certain probes targeting CpG loci, which include single-nucleotide polymorphisms (SNPs) near or within the probe sequence or even in the target CpG dinucleotide; in fact, there may be up to 25% probes on the 450K array that are affected by an SNP (Liu *et al*, 2013). As methylation levels of a specific locus may be influenced by genotype (Dedeurwaerder *et al*, 2011a), investigators may want to remove those SNP-associated loci from their data, and several R packages have options for carrying this out (Touleimat and Tost, 2012; Wang *et al*, 2012). Genetic effects, however, should not be underestimated in methylation arrays. As was recently demonstrated in Fraser *et al* (2012), a large portion of population-specific DNA methylation levels may in fact be due to population-specific genetic variants, which are themselves affected by genetic or environmental interactions. Although rare SNPs are unlikely to affect methylation levels to a large extent, somatic mutations can impact methylation levels greatly, such as driver mutations in a tumour; hence, the importance of subsequent sequencing validation.

Additional probes that a researcher may want to remove from their data include the 'Chen probes'. This is evidenced in a recently published paper showing that there may be spurious cross-hybridisation of Infinium probes on the 450K array and further suggesting that cross-hybridisation to the sex chromosomes may account for the large gender effects that researchers have found on the autosomal chromosomes (Chen *et al*, 2013). Finally, a number of SNP probes are also included on the Infinium array that can help identify mislabelled samples, as implemented in wateRmelon (Pidsley *et al*, 2013).

BACKGROUND CORRECTION

Background correction is platform specific, helps to remove nonspecific signal from total signal and corrects for between-array artefacts. Although this can be performed using Illumina's GenomeStudio, several R packages contain background correction

functions. This includes the preprocessing and analysis pipeline for 450K data (Touleimat and Tost, 2012), providing background-level correction using lumi (Du, 2008), and furthermore Limma (Wettenhall and Smyth, 2004) and MethyLumi (Davis *et al*, 2011). Background can also be estimated by direct estimation from the density modes of the intensities measured by each probe. However, the latter has been shown to produce aberrant DNA methylation profiles, so using negative control probes may be preferred (Touleimat and Tost, 2012). One can also use Minfi (Hansen and Aryee, 2013) as a background estimation method; however, the authors acknowledge that this method may result in differing values compared with those estimated via GenomeStudio.

NORMALISATION

Normalisation concerns the removal of sources of experimental artefacts, random noise and technical and systematic variation caused by microarray technology, which, if left unaddressed, has the potential to mask true biological differences (Sun *et al*, 2011a). Two different types of normalisation exist: (1) between-array normalisation, removing technical artefacts between samples on different arrays, and (2) within-array normalisation, correcting for intensity-related dye biases (Siegmund, 2011).

Owing to the features of DNA methylation, there is a lack of consensus regarding the optimal approach for normalisation of methylation data. Specifically, there is an imbalance in methylation levels throughout the genome creating a skewness to the methylation log-ratio distribution; the degree of this skewness is dependent on the levels of methylation in particular samples (Siegmund, 2011). This imbalance is due to the non-random distribution of CpG sites throughout the genome and the link between CpG density and DNA methylation; for instance, CpG islands (CGI) are often unmethylated, whereas the opposite relationship is typically seen in non-CGIs in normal human cells (Baylin and Jones, 2011). Furthermore, total fluorescence signal is inversely related to DNA methylation levels (Siegmund, 2011). Many available normalisation methods were designed for gene expression array data and are based on assumptions that may not be appropriate for DNA methylation microarray data.

GenomeStudio provides an internal control normalisation method for the 450K assay (Illumina, 2008), which is also used in MethyLumi (Davis *et al*, 2011) and Minfi (Hansen and Aryee, 2013); by default, Genomestudio uses the first sample in the array as the reference and allows the user to reselect the reference sample as needed if the original sample is nongenomic or of poor quality.

Quantile normalisation is one of the most commonly used normalisation techniques. Locally weighted scatterplot smoothing (LOESS) normalisation is an intensity-dependent normalisation method that assumes independence between the difference in log fluorescence signals between two samples and the average of the log signals from the two dyes (Siegmund, 2011). Quantile and LOESS normalisation (Laird, 2010) assume similar total signal across samples and can therefore remove true biological signal, because of the nature of DNA methylation described above, and have assumptions unlikely to hold for methylation data. As the Infinium I and II probe types examine different subsets of the genome, described in detail below, quantile normalisation cannot be applied indiscriminantly across probe types.

Lumi (Du, 2008), also used in HumMethQCReport (Mancuso *et al*, 2011), offers an alternative to quantile normalisation through a robust spline normalisation, which is designed to normalise variance-stabilised data by combining features of both quantile and LOESS normalisation (Du, 2008). Another approach, subset quantile normalisation (Wu and Aryee, 2010), normalises the data on the basis of on a subset of negative control or CpG-free probes that are independent of DNA methylation but suffers the same issues as other quantile approaches. The TurboNorm R package (van Iterson *et al*, 2012) provides an alternative to LOESS normalisation using a weighted P-spline intensity-dependent normalisation technique and can be applied to two colour arrays. A more recent method (Johnson *et al,* 2007), which we describe in more detail below, performs both normalisation and batch-effect correction. A comparison of different normalisation pipelines for Illumina 450K data can be found in two recent publications (Marabita *et al*, 2013; Pidsley *et al*, 2013).

TYPE I AND II PROBE SCALING

Another potential methodological concern stems from the fact that the 450K array uses two different types of probes, prompting the recommendation of rescaling to make the probe distributions comparable (Bibikova *et al*, 2011). Specifically, the 450K array has 485 577 probes, of which 72% use the Infinium type II primer extension assay where the unmethylated (red channel) and methylated (green channel) signals are measured by a single bead (Bibikova *et al*, 2011). The remainder use the Infinium type I primer extension assay (also used in the 27K Infinium array) where the unmethylated and methylated signals are measured by different beads in the same colour channel (Bibikova *et al*, 2011). Importantly, the two probes differ in terms of CpG density, with more CpGs mapping to CpG islands for type I probes (57%) as compared with type II probes (21%) (Bibikova *et al*, 2011). Moreover, compared with Infinium I probes, the range of β-values obtained from the Infinium II probes is smaller; in addition, the Infinium II probes also appear to be less sensitive for the detection of extreme methylation values and display a greater variance between replicates (Dedeurwaerder *et al*, 2011a).

The divergence in the methylation distribution range has implications for statistical analysis of the array data. For example, in a supervised analysis of all probes, an enrichment bias towards type I probes may be created when ranking probes because of the higher range of type I probes (Maksimovic *et al*, 2012). In addition, region-based analyses assume that probes within those regions are comparable, potentially untenable because of the diverging chemistries on the 450K array (Maksimovic *et al*, 2012). Moreover, when performing profile analyses or clustering, the differing chemistries between the two probes types may drive the clustering solution.

Attempts have been made to use rescaling to 'repair' the divergence between these two types of probes. The first correction method proposed was peak-based correction (Dedeurwaerder *et al*, 2011a), implemented in IMA (Wang *et al*, 2012), wherein the Infinium II data is rescaled on the basis of the Infinium I data assuming a bimodal shape of the methylation density profiles. However, several researchers have noted that this method is sensitive to variation in the shape of DNA methylation density curves and does not work well when the density distribution does not exhibit well-defined peaks or modes (Pan *et al*, 2012; Touleimat and Tost, 2012; Teschendorff *et al,* 2013).

Three alternative approaches have been proposed recently to address the limitations of the peak base correction approach. The first, subset-quantile within-array normalization (Maksimovic *et al*, 2012), is available in Minfi. Subset-quantile within-array normalization determines an average quantile distribution using a subset of probes defined to be biologically similar on the basis of CpG content and allows the Infinium I and II probes to be normalised together (Maksimovic *et al*, 2012).

The second, subset quantile normalisation (Touleimat and Tost, 2012), uses the genomic location of CpGs to create probe subgroups through which they apply subset quantile normalisation.

The reference quantiles used in this approach are based on type I probes with significant detection *P*-values (Touleimat and Tost, 2012).

Finally, the β-mixture quantile dilation normalisation method, implemented in the wateRmelon package (Pidsley *et al*, 2013), uses quantiles to normalise the type II probe values into a distribution comparable to the type I probes using a β-mixture model fit to the type I and type II probes separately and then transforms the probabilities of class membership of the type II probes into quantiles (β-values) using the parameters of the β-distributions of the type I distribution (Teschendorff *et al,* 2013). This method uses a three-state β-mixture model but does not use fit to the middle 'hemimethylated' component in the normalisation; therefore, it does not require a trimodal distribution (Teschendorff *et al,* 2013). An advantage of BMIQ is that it avoids selecting subsets of probes matched for biological characteristics as done in the previous method and was found to be the best algorithm for reducing probe design bias in a recent paper (Marabita *et al*, 2013).

Rescaling using the methods mentioned above may be unnecessary when analysing 450K data on a CpG-by-CpG basis because the comparisons will be made at the individual probe level.

ADJUSTMENT BATCH/PLATE/CHIP/OTHER CONFOUNDERS

DNA methylation arrays are susceptible to batch effects: technical remnants that are not associated with the biological question but with unrelated factors such as laboratory conditions or experiment time (Leek *et al*, 2010; Sun *et al*, 2011b). Normalisation has been shown to reduce some component of batch effects, although not all (Teschendorff *et al*, 2009; Leek *et al*, 2010; Sun *et al*, 2011b). Sound study design is critical for proper evaluation of and correction for batch effects: for instance, samples from different study groups should be split randomly or equally into different batches (Johnson *et al*, 2007). By properly correcting for batch effects, one can combine data from multiple batches, enabling greater statistical power to measure a specific association of interest (Johnson *et al*, 2007).

Several methods have been proposed to adjust for batch effects. ComBat uses an empirical Bayes procedure for this (Johnson *et al*, 2007), is robust to outliers in small sample sizes and can adjust for other potential confounders along with batch (Sun *et al*, 2011b). However, this method can be computationally burdensome and was initially developed for gene expression data; therefore, it requires a transformation of methylation data, which follows the β-distribution, to satisfy the assumption of normality.

Other R packages exist to adjust for batch effects. MethLAB (Kilaru *et al*, 2012) and CpG assoc (Barfield *et al*, 2012) allow the adjustment for batch using a mixed-effects model framework. However, because these methods do not directly adjust the data, unlike ComBat, they should be used only for a locus-by-locus analysis.

The array literature indicates that array position effects may also exist (van Eijk *et al*, 2012), and thus new batch correction techniques may be needed to take those into account. When phenotype distribution is heterogeneous across chips, which can occur in small samples even after randomisation, methods such as ComBAT can fail; in this case, linear mixed-effects models treating chip effects as random is an alternative.

However, in certain cases, the true sources of batch effects or confounding are unknown or cannot be adequately modelled statistically (Leek *et al*, 2010). In such cases, two methods, surrogate variable analysis (SVA) (Leek and Storey, 2007) and independent surrogate variable analysis (ISVA) (Teschendorff *et al*, 2011), also available as the ISVA R package, are very useful. Surrogate variable analysis estimates the source of batch effects directly from the array data, and variables estimated with SVA (SVs) can then be included into the statistical model as covariates (Leek and Storey, 2007). A modified version of SVA, ISVA, identifies features correlating with the phenotype of interest in the presence of potential or unknown confounding factors, which are modelled as statistically independent surrogate variables or ISVs (Teschendorff *et al*, 2011). This method could also be used for batch effects by constructing ISVs that are associated with these as potential confounders and including them in the analytical model. A problem with this technique occurs when the ISVs correlate both with the phenotype of interest and with the potential confounders, making model covariate selection difficult. Furthermore, ISVA and SVA do not directly adjust the methylation data, like ComBAT does, which may be problematic if the analytical goal is clustering. One could, however, fit a model with the estimated SVs or ISVs and compute the residuals for subsequent analyses.

DOWNSTREAM ANALYSIS

Methylation status. Average β or the β-value is a commonly used metric to denote the level or percentage of methylation for an interrogated locus. Investigators also use the *M*-value, or log ratio, to measure methylation (Du *et al*, 2010):

$$M = \log_2 \frac{\mathrm{Max}(M,0)}{\mathrm{Max}(U,0)}$$

A normalised *M*-value near 0 signifies a semimethylated locus, a positive *M*-value indicates that more molecules are methylated than unmethylated, whereas negative *M*-values have the opposite interpretation (Du *et al*, 2010). An *M*-value is attractive in that it can been used in many statistical models derived for expression arrays that assume normality (Du *et al*, 2010). However, β-values are much more biologically interpretable than their counterpart; furthermore, a recent paper found supervised principal components analysis (SPCA), as described below, to work better in the context of β-values as opposed to *M*-values (Zhuang *et al*, 2012). The relationship between the β- and *M*-value is captured by (Du *et al*, 2010):

$$M = \log_2 \frac{\beta}{1-\beta}$$

Differential methylation/region-based analysis. Locus-by-locus analyses examine the relationship between a phenotype of interest and methylation of individual CpG sites across the genome, seeking to find differentially methylated sites. Differential methylation analysis aims to determine methylation differences between the specific groups (such as cases and controls), such as probe-wise or locus-specific methylation differences; the two terminologies are therefore equivalent when at the individual locus level. A very simple example is Delta B (Bibikova *et al*, 2011; Touleimat and Tost, 2012), where a difference is applied to two groups' methylation medians for each CpG locus; if the absolute value of the difference in medians across samples of each group is higher than 0.2, then that locus is considered to be differentially methylated. This 0.2 threshold corresponds to the recommended difference in methylation between samples that can be detected with 99% confidence (Bibikova *et al*, 2011). MethVisual (Zackay and Steinhoff, 2010) tests whether each CpG site has independent membership between two groups using Fisher's exact test; other packages include (Wettenhall and Smyth, 2004; Barfield *et al*, 2012; Kilaru *et al*, 2012; Wang *et al*, 2012), some allowing for the adjustment of potential confounders (Barfield *et al*, 2012; Kilaru *et al*, 2012; Wang *et al*, 2012). Minfi (Hansen and Aryee, 2013) uses linear regression and an F-test to test for a univariate association between the methylation of individual loci and continuous or

categorical phenotypes, respectively. When sample sizes are <10, Minfi (Hansen and Aryee, 2013) has options for using limma (Wettenhall and Smyth, 2004). Specifically, limma uses an empirical Bayes moderated *t*-test, computed for each probe, which is similar to a *t*-test, except that the standard errors have been shrunk towards a common value. *M*-values should be used in these cases as, being based on a Bayesian Gaussian model, they will rely much more heavily on the Gaussianity assumption (Zhuang *et al*, 2012). The IMA package (Wang *et al*, 2012) allows site (methylation locus)-specific and region (all loci in a gene)-specific differential methylation analysis using Student's *t*-test and empirical Bayes statistics. For region analysis, IMA will compute the mean, median or Tukey's biweight robust average for the loci within that region and create an index (Wang *et al*, 2012). Methylkit (Akalin *et al*, 2012) allows for analysis at the site or regional level using logistic regression or Fisher's exact test. With multiple samples per group, methylkit will preferentially use logistic regression, enabling also the inclusion of potential confounders (Akalin *et al*, 2012); to get stable estimates of the regression coefficients in logistic regression about 10 events per variable are necessary (Peduzzi *et al*, 1996).

Differential methylation analysis can also be performed by measuring variability between methylation loci as opposed to using statistical tests on the basis of differences in mean methylation (Xu *et al*, 2013). This is available in the EVORA package, allowing an investigator to use differential variability in methylation of CpGs and to then associate them with a phenotype of interest, such as cancer status (Teschendorff and Widschwendter, 2012; Xu *et al*, 2013).

As noted in several recent works, nearby CpG loci tend to have methylation levels that are highly correlated (Leek *et al*, 2010). As a result, statistical analyses that assume independence may be problematic. Methods are being developed to deal with this potential problem and include bump-hunting techniques (Leek and Storey, 2007), which take into account CpG proximity and borrow strength across neighbouring probes. Although these approaches were originally developed for CHARM assays, they may be adapted to the less-dense 450K array, pending careful attention to the tuning parameters for defining a 'region'.

Although the above methods have proved successful in identifying individual CpG sites that associate with some phenotype/exposure of interest, the extent to which the methylation of these sites reflect true changes to the methylome or represent heterogeneity in underlying cell-type distributions depends largely on the tissue being sampled (Teschendorff *et al*, 2009; Houseman *et al*, 2012). We recently developed a set of statistical methods that exploit the use of leukocyte-specific DMRs for inferring changes in cell mixture proportions based solely on peripheral blood profiles of DNA methylation (Houseman *et al*, 2012). Under certain constraints, this approach can be used to approximate the underlying distribution of cell proportions among samples comprising a heterogeneous mixture of cell populations with distinct DNA methylation profiles (Houseman *et al*, 2012). This method has recently been used for predicting cell-type proportions, which were then subsequently added as additional covariate terms in a differential methylation analysis of rheumatoid arthritis cases/controls (Liu *et al*, 2013). Furthermore, the methods of Houseman *et al* (2012) were recently validated using a publicly available data set (Lam *et al*, 2012) that consisted of both PBMC-derived DNA methylation profiles and complete blood cell (CBC) counts for 94 healthy, non-diseased adult subjects (Koestler *et al*, 2013).

Clustering/profile analysis. Clustering refers to the grouping of objects into clusters, such that the objects within the same cluster are more similar compared with objects in different clusters. Owing to the interest in identifying molecular subtypes in the context of cancer, clustering has become a staple technique in the analysis of array-based DNA methylation data.

Two very well-known non-hierarchical methods used to cluster DNA methylation include K-means and K-medoids, also known as partitioning around medoids or PAM (Pollard *et al*, 2005). Two disadvantages of K-means are that it requires the prespecification of the number of classes, which is not often known; furthermore, K-means create clusters based only on the first moment, which is problematic in cases where the variance of a specific probe contains biologically important information. Another commonly used method to detect patterns in methylation data is PCA, which is a latent variable method often applied as a dimension reduction procedure and used for the detection of batch effects (Jolliffe, 2002). Principal component analysis was first applied to genome-wide Infinium HumanMethylation27 DNA methylation data as shown in Teschendorff *et al* (2009). Principal component analysis is used to develop a smaller number of artificial variables, called principal components, which account for most of the variance in the observed variables of a data set (Jolliffe, 2002); usually only the first few components are kept as potential predictors for statistical modelling (Jolliffe, 2002). However, additional principal components may be of biological significance as shown in Teschendorff *et al* (2009). A method to estimate the number of significant PCA components is available in the ISVA package (Teschendorff *et al*, 2011). This algorithm is based on the Random Matrix Theory (Plerou *et al*, 2002), which can be used to estimate the number of significant PCA components that are subsequently examined for their association with study-specific characteristics. Random Matrix Theory estimates the number of significant components of a data covariance matrix by comparing the statistics of the observed eigenvalues obtained from PCA with those obtained from a random matrix. The main disadvantage with PCA lies in the poor interpretability of the resulting principal components and the requirement of a large sample size in order to obtain reliable results.

Another well-known clustering method is hierarchical clustering, which builds a binary tree by successively merging similar samples or probes based on a measure of similarity (Eisen, 1998). However, because of its unsupervised nature, this form of clustering may or may not predict a phenotype of interest, as it does not use data beyond methylation to form clusters. Lumi (Du, 2008), HumMeth27QCReport (Mancuso *et al*, 2011) and methylkit (Akalin *et al*, 2012) all provide hierarchical clustering and PCA options using normalised *M*-values.

In addition to non-parametric techniques for clustering or profile analysis, Houseman *et al* (2008) developed a recursive-partitioning mixture model (RPMM), an unsupervised, model-based, hierarchical clustering methodology for array-based DNA methylation data. Recursive-partitioning mixture model assumes a β-mixture model to split samples between subgroups and provides an estimate for the number of clusters; furthermore, it is computationally efficient relative to the standard finite mixture model approach (Houseman *et al*, 2008). Owing to the inherent correlation in the methylation status of nearby CpG sites, there have also been efforts to incorporate correlation structures based on the proximity of CpGs in the context RPMM (Leek *et al*, 2010).

Semisupervised methods use both array-based genomic data and clinical data for identifying profiles that are associated with a clinical variable of interest, such as survival. Semisupervised clustering (SS-Clust) begins by identifying a set of genes that correlate with a phenotype of interest, followed by unsupervised clustering of samples based on the set of genes (Bair and Tibshirani, 2004). Supervised principal components analysis uses a similar methodology to SS-Clust, but replaces unsupervised clustering with PCA, providing a 'risk score' for each patient, which is then used as a continuous predictor of survival (Jolliffe, 2002). Semisupervised clustering's main disadvantage is that it

requires prespecification of the number of clusters; moreover, SPCA inherits the interpretability issues characteristic of PCA. Semisupervised RPMM (Koestler *et al*, 2010) has been shown to outperform SS-Clust and SPCA under certain circumstances and does not require the prespecification of the number of clusters.

One of the first attempts to discover novel tumour classes through profiling of methylation data involved a supervised method called support vector machine (SVM) including a cross-validation method to evaluate its prediction performance (Adorjan *et al*, 2002). This approach was initially very computationally intensive but was a precursor to other profile analysis methods. Another method, Elastic net, is a shrinkage and selection method, which produces a sparse model with good prediction accuracy, while encouraging a grouping effect (Zou and Hastie, 2005); this algorithm is now being widely used on all types of omics data (Barretina *et al*, 2012; Hannum *et al*, 2013) and was compared with SVM and SPCA in Zhuang *et al*, 2012 and shown to be far superior.

Pathway analysis. Many researchers use pathway analysis to characterise the function of the gene in which the individual or group of loci are found. Several software packages do this; however, we focus on two freely available resources that can also be used in R. The Gene Ontology (GO) provides a very detailed representation of functional relationships between biological processes, molecular function and cellular components across eukaryotic biology (Ashburner *et al*, 2000). Another resource that borrows heavily from GO is PANTHER (Thomas *et al*, 2003), which relates protein sequence relationships to functional relationships. However, many commonly used pathway analysis methods are based on gene expression correlation or protein–protein interaction; although pathway perturbations are likely to be evident in expression changes across all genes of a pathway, a single well-placed alteration of DNA methylation, acting as an epigenetic switch, may alter all downstream mRNA expression. In light of this, sensitivity for detecting significant pathways is lower for DNA methylation than it might be for mRNA expression. In addition, unlike mRNA expression, CpGs have different implications for expression depending on where they exist in relation to a gene or if they are mapped to any gene at all. As the 450K array has great heterogeneity with respect to the CpG representation by gene region, there is the potential for pathway analysis on 450K data to be biased by CpG selection. In addition, as genes are not equally covered throughout the array through the number of probes in their specific regions, this may further bias this analysis. Therefore, in using such approaches, we recommend stratification by gene region (e.g., promoter) to decrease the potential for bias. Once a specific region has been chosen, then pathway analysis, GSEA, or integration with interaction networks could be a fruitful procedure, as recently demonstrated in Dedeurwaerder *et al* (2011b) and West *et al* (2013).

MULTIPLE TESTING CORRECTION

Once the analysis has identified top hits, multiple testing correction is necessary to reduce the likelihood of identifying false-positive loci by adjusting statistical confidence measures by the number of tests performed. Bonferroni correction consists of multiplying each probability by the total number of tests performed; this controls the family-wise error rate (Holm, 1979).

A less-conservative, widely used, approach involves controlling the FDR (q-value) or the expected proportion of false discoveries among the discoveries; this also uses a sequential P-value method (Benjamini *et al*, 2001); several R packages allow for the adjustment of the FDR (Barfield *et al*, 2012; Kilaru *et al*, 2012; Wang *et al*, 2012). All of the aforementioned methods assume statistical independence of the multiple tests, which can be violated when tests exhibit strong correlations (as mentioned above); furthermore, q-values imply subsequent validation in an independent sample, which may not occur. A potential solution to this independence assumption is with the use of permutation testing in which the phenotype of interest is randomly re-assigned, and the data reanalysed. CpG assoc provides a permutation testing option to obtain empirical P-values (Barfield *et al*, 2012).

VALIDATION OF SIGNIFICANT HITS

The final step in the proper processing and analysis of DNA methylation arrays is validation of significant hits by an independent experimental approach or data resource. The gold standard is bisulphite sequencing-based methods, such as pyrosequencing (Ammerpohl *et al*, 2009) and Epityper (Laird, 2010), to provide high-throughput quantitation (Siegmund, 2011). Another valuable resource for validation (and exploration) of DNA methylation array data is publicly available repositories such as the Gene Expression Omnibus (Edgar *et al*, 2002). Finally, with the availability of data resources such as the above and HAPMAP (Altshuler *et al*, 2010), researchers can now integrate their methylation array data with these resources, to help further understand molecular and genomic profiles that contribute to outcomes of interest such as cancer risk.

CONCLUSIONS

Owing to the plethora and complexity of methods for array processing and analysis, described above, and to the multitude of researchers using DNA methylation arrays, there is a need to create a protocol of good practice to ensure that study results are of the highest quality possible. Just as gold standard laboratory methods are crucial to the generation of quality biological data, gold standard processing and analytical methods are equally as important. Through the proper use of the processing and analysis flowchart described above, we hope that potential users will best harness these powerful array-based tools, which will in turn lead to rapid discoveries in human health and disease.

ACKNOWLEDGEMENTS

This work was supported by Cancer Research UK program A6689 (JMF, CWB and RB). JMF is funded by Breast Cancer Campaign; RB is funded by Ovarian Cancer Action; CJM and EAH are funded by NIMH R01 MH094609; KTK is funded by the US NIH Grants (R01 CA121147, R01 CA078609 and R01 CA100679); and MRK is funded by P20 ES018175, R01 CA57494 and EPA RD83459901.

REFERENCES

Adorjan P, Distler J, Lipscher E, Model F, Muller J, Pelet C, Braun A, Florl AR, Gutig D, Grabs G, Howe A, Kursar M, Lesche R, Leu E, Lewin A, Maier S, Muller V, Otto T, Scholz C, Schulz WA, Seifert HH, Schwope I, Ziebarth H, Berlin K, Piepenbrock C, Olek A (2002) Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res* **30**: e21.

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE (2012) MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**: R87.

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K,

Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**: 52–58.

Ammerpohl O, Martín-Subero JI, Richter J, Vater I, Siebert R (2009) Hunting for the 5th base: techniques for analyzing DNA methylation. *Biochim Biophys Acta* **1790**: 847–862.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29.

Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**: E108.

Barfield RT, Kilaru V, Smith AK, Conneely KN (2012) CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics* **28**: 1280–1281.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jané-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi Jr P, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**: 603–607.

Baylin SB, Jones PA (2011) A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* **11**: 726–734.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**: 279–284.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R (2011) High density DNA methylation array with single CpG site resolution. *Genomics* **98**: 288–295.

Bock C (2012) Analysing and interpreting DNA methylation data. *Nat Rev Genet* **13**: 705–719.

Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**: 203–209.

Davis S, Du P, Bilke S, Triche T, Bootwalla M (2011) MethyLumi: for handling Illumina DNA methylation data Bioconductor (Online). Available at http://www.bioconductor.org/packages/2.12/bioc/html/methylumi.html.

Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F (2011a) Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**: 771–784.

Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, Michiels S, Volkmar M, Deplus R, Luciani J, Lallemand F, Larsimont D, Toussaint J, Haussy S, Rothé F, Rouas G, Metzger O, Majjaj S, Saini K, Putmans P, Hames G, van Baren N, Coulie PG, Piccart M, Sotiriou C, Fuks F (2011b) DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol Med* **3**: 726–741.

Du P, Kibbe WA, Lin SM (2008) Lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**: 1547–1548.

Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinform* **11**: 587.

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**: 207–210.

Eisen M (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95**: 14863–14868.

Fraser HB, Lam LL, Neumann SM, Kobor MS (2012) Population-specificity of human DNA methylation. *Genome Biol* **13**: R8.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* **49**: 359–367.

Hansen KD, Aryee M (2013) Minfi: Analyze Illumina's 450K methylation arrays. Bioconductor (Online). Available at http://bioconductor.org/packages/2.12/bioc/vignettes/minfi/inst/doc/minfi.pdf.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* **6**: 65–70.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform* **13**: 86.

Houseman EA, Christensen BC, Yeh RF, Marsit CJ, Karagas MR, Wrensch M, Nelson HH, Wiemels J, Zheng S, Wiencke JK, Kelsey KT (2008) Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinform* **9**: 365.

Illumina (2008) *Infinium Assay Methylation Protocol Guide*. Illumina: San Diego, CA, USA.

Illumina (2011) *GenomeStudio/BeadStudio Software Methylation Module*.

Ioannidis JP (2007) Why most published research findings are false: author's reply to Goodman and Greenland. *PLoS Med* **4**: e215.

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**: 118–127.

Jolliffe IT (2002) *Principal Component Analysis*. New York, NY, USA.

Kilaru V, Barfield RT, Schroeder JW, Smith AK, Conneely KN (2012) Methlab: a graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics* **7**: 225–229.

Koestler DC, Christensen B, Karagas MR, Marsit CJ, Langevin SM, Kelsey KT, Wiencke JK, Houseman EA (2013) Blood-based profiles of DNA methylation predict the underlying distribution of cell types. *Epigenetics* **8**(8): 816–826.

Koestler DC, Marsit CJ, Christensen BC, Karagas MR, Bueno R, Sugarbaker DJ, Kelsey EA, Houseman KT (2010) Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics* **26**(20): 2578–2585.

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

Laird PW (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* **11**: 191–203.

Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kobor MS (2012) Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci USA* **109**(Suppl 2): 17253–17260.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**: 733–739.

Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**: 1724–1735.

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, Ekström TJ, Feinberg AP (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* **31**: 142–147.

Maksimovic J, Gordon L, Oshlack A (2012) SWAN: subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**: R44.

Mancuso FM, Montfort M, Carreras A, Alibés A, Roma G (2011) HumMeth27QCReport: an R package for quality control and primary analysis of Illumina Infinium methylation data. *BMC Res Notes* **4**: 546.

Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, Sundberg CJ, Ekström TJ, Teschendorff AE, Tegnér J, Gomez-Cabrero D (2013) An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics* **8**: 333–346.

Pan H, Chen L, Dogra S, Teh AL, Tan JH, Lim YI, Lim YC, Jin S, Lee YK, Ng PY, Ong ML, Barton S, Chong YS, Meaney MJ, Gluckman PD, Stunkel W, Ding C, Holbrook JD (2012) Measuring the methylome in

clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics* **7**: 1173–1187.
Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* **49**: 1373–1379.
Petronis A (2010) Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* **465**: 721–727.
Pidsley R, Wong Y, Volta CC, Lunnon M, Mill K, Schalkwyk LC. J (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genom* **14**: 293.
Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Guhr T, Stanley HE (2002) Random matrix approach to cross correlations in financial data. *Phys Rev E* **65**: 066126.
Pollard KS, Van Der Laan MJ. Cluster Analysis of Genomic Data (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer: Berlin, Germany.
Rakyan VK, Down TA, Balding DJ, Beck S (2011) Epigenome-wide association studies for common human diseases. *Nat Rev Genet* **12**: 529–541.
Siegmund KD (2011) Statistical approaches for the analysis of DNA methylation microarray data. *Hum Genet* **129**: 585–595.
Sun S, Huang YW, Yan PS, Huang TH, Lin S (2011a) Preprocessing differential methylation hybridization microarray data. *BioData Min* **4**: 13.
Sun Z, Chai HS, Wu Y, White WM, Donkena KV, Klein CJ, Garovic VD, Therneau TM, Kocher JP (2011b) Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Med Genom* **4**: 84.
Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S (2013) A Beta-Mixture Quantile Normalisation method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. *Bioinformatics* **29**(2): 189–196.
Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, Jones A, Lechner M, Beck S, Jacobs IJ, Widschwendter M (2009) An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One* **4**: e8274.
Teschendorff AE, Widschwendter M (2012) Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**: 1487–1494.
Teschendorff AE, Zhuang J, Widschwendter M (2011) Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**: 1496–1505.
Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**: 2129–2141.
Touleimat N, Tost J (2012) Complete pipeline for Infinium(R) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**: 325–341.
van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, de Kovel CG, Janson E, Strengman E, Langfelder P, Kahn RS, van den Berg LH, Horvath S, Ophoff RA (2012) Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genom* **13**: 636.
van Iterson M, Duijkers FA, Meijerink JP, Admiraal P, van Ommen GJ, Boer JM, van Noesel MM, Menezes RX (2012) A novel and fast normalization method for high-density arrays. *Stat Appl Genet Mol Biol* **11**(4): Article 5. 10.1515/1544-6115.5.
Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28**: 729–730.
West J, Beck S, Wang X, Teschendorff AE (2013) An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Sci Rep* **3**: 1630.
Wettenhall JM, Smyth GK (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* **20**: 3705–3706.
Wu Z, Aryee MJ (2010) Subset quantile normalization using negative control features. *J Comput Biol* **17**: 1385–1395.
Xu X, Su S, Barnes VA, De Miguel C, Pollock J, Ownby D, Shi H, Zhu H, Snieder H, Wang X (2013) A genome-wide methylation study on obesity: differential variability and differential methylation. *Epigenetics* **8**(5): 522–533.
Zackay A, Steinhoff C (2010) MethVisual – visualization and exploratory statistical analysis of DNA methylation profiles from bisulfite sequencing. *BMC Res Notes* **3**: 337.
Zhuang J, Widschwendter M, Teschendorff AE. (2012) A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics* **13**: 59.
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Statist Soc Ser B* **67**: 301–320.