Review of processing and analysis methods for DNA methylation array data

Charlotte S. Wilhelm-Benartzi[1#], Devin C. Koestler[2#], Margaret R. Karagas[2], James M. Flanagan[1], Brock C. Christensen[23], Karl T. Kelsey[4], Carmen J. Marsit[23], E. Andres Houseman[5], Robert Brown[16]*

1 Epigenetics Unit, Department of Surgery and Cancer, Ovarian Cancer Action Research Centre, Imperial College London, 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; Department of Epidemiology; Brown University; Providence, RI USA

5 Department of Public Health, Oregon State University, Corvallis, Oregon

6 Section of Molecular Pathology, Institute for Cancer Research, Sutton, UK.

#Authors contributed equally to work

Running Title: Processing and analysis of DNA methylation data

Word count: 4089, **Number of tables**: 1, **Number of figures**: 2, **Abstract**: 118 words

*To whom correspondence should be addressed:

Professor Robert Brown

Epigenetics Unit, Division of Cancer, Department of Surgery and Cancer,

Faculty of Medicine, Imperial College London

4th floor IRDB, Hammersmith Campus, Du Cane Road, London W12 0NN, UK

Phone: 020 75941804; Fax: 020 75942129; email: b.brown@imperial.ac.uk

Abstract:

The promise of epigenome-wide association studies (EWAS) and cancer specific somatic 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, as well as to summarize advantages of recent approaches proposed for handling those issues; focusing on approaches publicly available in open-source environments such as R and Bioconductor. The processing tools and analysis flowchart described we hope 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 play 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 to normal tissue, this has ushered in the era of epigenome-wide association studies (EWAS), analogous to the genome-wide association studies (GWAS), 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, as well as to summarize recent approaches proposed for handling those issues. Excellent reviews of the field of epigenetics and technical aspects of array-based assessment of DNA methylation are available, although this is a constantly developing research area (Petronis, 2010, Rakyan et al., 2011, Laird, 2010, Baylin and Jones, 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 sub-headings shown in Figure 1 form the basis for the topics highlighted in this review. Our goal 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, i.e., that there are too many "false positives" in analyzing micro-array 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 help understanding of the pathogenesis of human disease.

Illumina bead-array technology for methylation:

Illumina adapted its BeadArray technology for genotyping to recognize bisulfite-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 summarized by the average signal for methylated (M) and unmethylated (U) alleles, and are used to compute the *Beta* value, where:

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

A Beta 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 HumanMethylation27, and the Infinium HD 450K Methylation array, which all use two fluorescent dye colors, but differ in the chemistries used to recognize the bisulfite-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 for 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 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), the R-package *HumMethQCReport* (Mancuso et al., 2011) also provides these plots . Figure 2 shows hybridization and bisulfite conversion plots for 450K data in the green channel. While 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 if they are beyond the expected range (e.g., median ± 3 standard deviations) of the signal intensities across all samples.

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, 2013) and *MethyLumi* (Davis, 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 pre-specified 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 (Touleimat and Tost, 2012, Davis, 2011).

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, more influence in 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 a SNP (Bock, 2012). 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 (Wang et al., 2012, Touleimat and Tost, 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. While 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 which can help identify mislabelled samples, as implemented in wateRmelon (Pidsley et al., 2013).

Background correction:

Background correction is platform specific, helps to remove non-specific signal from total signal and corrects for between-array artifacts. While 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 P, 2008), and furthermore *Limma* (Wettenhall and Smyth, 2004) and *MethyLumi* (Davis, 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, 2013) as a background estimation method; however, the authors acknowledge that this method may result in differing values compared to those estimated via *GenomeStudio*.

Normalization:

Normalization concerns the removal of sources of experimental artifacts, random noise, 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 normalization exist: (1) between array normalization, removing technical artifacts between samples on different arrays, and (2) within array normalization, correcting for intensity-related dye biases (Siegmund, 2011).

Due to the features of DNA methylation, there is a lack of consensus regarding the optimal approach for normalization 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 normalization 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 normalization method for the 450K assay (Illumina, 2008) which is also used in *MethyLumi* (Davis, 2011) and *Minfi* (Hansen, 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 non-genomic or of poor quality.

Quantile normalization is one of the most commonly used normalization techniques. LOESS normalization is an intensity-dependent normalization 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 normalization (Laird, 2010) assume similar total signal across samples and can therefore remove true biological signal, due to 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 normalization cannot be applied indiscriminantly across probe types.

Lumi (Du P, 2008), also used in *HumMethQCReport* (Mancuso et al., 2011), offers an alternative to quantile normalization through a robust spline normalization, which is designed to normalize variance-stabilized data by combining features of both quantile and LOESS normalization (Du P, 2008). Another approach, subset quantile normalization (Wu and Aryee, 2010), normalizes the data based 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 normalization using a weighted P-spline intensity-dependent normalization technique and can be applied to two color arrays. A more recent method (Sun et al., 2011b), which we describe in more detail below, performs both normalization and batch effect correction. A comparison of different normalization pipelines for Illumina 450K data can be found in two recent publications (Pidsley et al., 2013, Marabita 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 to type II probes (21%) (Bibikova et al., 2011). Moreover, compared to Infinium I probes, the range of beta values obtained from the Infinium II probes is smaller; additionally, 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 due to the higher range of type I probes (Maksimovic et al., 2012). Additionally, region-based analyses assume that probes within those regions are comparable; potentially untenable due to 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 (Touleimat and Tost, 2012, Teschendorff et al., 2012, Pan et al., 2012).

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

The second, Subset-quantile normalization (Touleimat and Tost, 2012), uses the genomic location of CpGs to create probe subgroups through which they apply subset quantile-normalization. The reference quantiles used in this approach are based on type I probes with significant detection p-values (Touleimat and Tost, 2012).

Finally, the Beta mixture quantile dilation normalization method, implemented in the wateRmelon package (Pidsley et al., 2013), uses quantiles to normalize the type II probe values into a distribution comparable to the type I probes using a beta-mixture model fit to the type I and type II probes separately, then transforms the probabilities of class membership of the type II probes into quantiles (Beta values) using the parameters of the beta distributions of the type I distribution(Teschendorff et al., 2012). This method uses a three-state beta 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., 2012). 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 analyzing 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 (Sun et al., 2011b, Leek et al., 2010). Normalization has been shown to reduce some component of batch effects, though not all (Sun et al., 2011b, Leek et al., 2010, Teschendorff et al., 2009). 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 to 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 requires a transformation of methylation data, which follows the Beta 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 which does, they should only be used for a locus by locus analysis.

The array literature indicates that array position effects may also exist (van Eijk et al., 2012), so 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 randomization, 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. SVA estimates the source of batch effects directly from 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, that 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:

1. Methylation status:

Average Beta 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)

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

A normalized M value near 0 signifies a semi-methylated locus, a positive M value indicate that more molecules are methylated than unmethylated, while 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 Beta and M value is captured by (Du et al., 2010):

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

2. 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 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 (Touleimat and Tost, 2012, Bibikova et al., 2011) 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 a Fisher's exact test; other packages include (Kilaru et al., 2012, Wang et al., 2012, Barfield et al., 2012, Wettenhall and Smyth, 2004) , some allowing for the adjustment of potential confounders (Kilaru et al., 2012, Wang et al., 2012, Barfield et al., 2012). *Minfi* (Hansen, 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 less than 10 , *Minfi* (Hansen, 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 since, 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 employ 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 based on 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 to 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. While 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".

While 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, depend largely on the tissue being sampled (Houseman et al., 2012, Teschendorff et al., 2009). 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 consisting of 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, (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. Epigenetics In press).

3. Clustering/ Profile analysis:

Clustering refers to the grouping of objects into clusters, such that the objects within the same cluster are more similar compared to objects in different clusters. Due 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, Cluster Analysis of Genomic Data). Two disadvantages of K means are that it requires the pre-specification of the number of classes, which is not often known; furthermore, K-means creates clusters based only on the first moment, 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 principal component analysis (PCA) which is a latent variable method often applied as a dimension reduction procedure and used for detection of batch effects (Jolliffe, 2002). PCA was first applied to genome-wide Infinium HumanMethylation27 DNA methylation data in (Teschendorff et al., 2009). PCA 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 dataset (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 Random Matrix Theory (Plerou et al., 2002) which can be used to estimate the number of number of significant PCA components that are subsequently examined for their association with study-specific characteristics. RMT estimates the number of significant components of a data covariance matrix by comparing the statistics of the observed eigenvalues obtained from PCA, to 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, due to 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 P, 2008),

HumMeth27QCReport (Mancuso et al., 2011) and *methylkit* (Akalin et al., 2012) all provide hierarchical clustering and PCA options using normalized M values.

In addition to non-parametric techniques for clustering or profile analysis, Houseman et al developed a Recursive-Partitioning-Mixture Model (RPMM), an unsupervised, model-based, hierarchical clustering methodology for array-based DNA methylation data. RPMM assumes a beta mixture model to split samples between subgroups and provides an estimate for the number of clusters; furthermore is computationally efficient relative to the standard finite mixture model approach (Houseman et al., 2008). Due 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).

Semi-supervised 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. Semi-supervised 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, 2004). SPCA 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). SS-Clust's main disadvantage is that it requires pre-specification of the number of clusters; moreover, SPCA inherits the interpretability issues characteristic of PCA. Semi-supervised RPMM (Koestler et al., 2010) has been shown to outperform SS-Clust and SPCA under certain circumstances and does not require the pre-specification 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 to SVM and SPCA in (Zhuang et al., 2012) and shown to be far superior.

4. Pathway Analysis

Many researchers use pathway analysis to characterize 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; while 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 upon where they exist in relation to a gene or if they are mapped to any gene at all. Since 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, 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 (Kilaru et al., 2012, Wang et al., 2012, Barfield 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 bisulfite sequencing based methods such as pyrosequencing (Ammerpohl et al., 2009) and Epityper (Laird, 2010) in order 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 or GEO (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:

Due 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.

Reference List:

ADORJAN, P., DISTLER, J., LIPSCHER, E., MODEL, F., MULLER, J., PELET, C., BRAUN, A., FLORL, A. R., GUTIG, D., GRABS, G., HOWE, A., KURSAR, M., LESCHE, R., LEU, E., LEWIN, A., MAIER, S., MULLER, V., OTTO, T., SCHOLZ, C., SCHULZ, W. A., SEIFERT, H. H., 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, F. E., FIGUEROA, M. E., MELNICK, A. & MASON, C. E. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol,* 13, R87.

ALTSHULER, D. M., GIBBS, R. A., PELTONEN, L., DERMITZAKIS, E., SCHAFFNER, S. F., YU, F., BONNEN, P. E., DE BAKKER, P. I., DELOUKAS, P., GABRIEL, S. B., GWILLIAM, R., HUNT, S., INOUYE, M., JIA, X., PALOTIE, A., PARKIN, M., WHITTAKER, P., CHANG, K., HAWES, A., LEWIS, L. R., REN, Y., WHEELER, D., MUZNY, D. M., BARNES, C., DARVISHI, K., HURLES, M., KORN, J. M., KRISTIANSSON, K., LEE, C., MCCARROL, S. A., NEMESH, J., KEINAN, A., MONTGOMERY, S. B., POLLACK, S., PRICE, A. L., SORANZO, N., GONZAGA-JAUREGUI, C., ANTTILA, V., BRODEUR, W., DALY, M. J., LESLIE, S., MCVEAN, G., MOUTSIANAS, L., NGUYEN, H., ZHANG, Q., GHORI, M. J., MCGINNIS, R., MCLAREN, W., TAKEUCHI, F., GROSSMAN, S. R., SHLYAKHTER, I., HOSTETTER, E. B., SABETI, P. C., ADEBAMOWO, C. A., FOSTER, M. W., GORDON, D. R., LICINIO, J., MANCA, M. C., MARSHALL, P. A., MATSUDA, I., NGARE, D., WANG, V. O., REDDY, D., ROTIMI, C. N., ROYAL, C. D., SHARP, R. R., ZENG, C., BROOKS, L. D. & MCEWEN, J. E. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature,* 467, 52-8.

AMMERPOHL, O., MARTIN-SUBERO, J. I., RICHTER, J., VATER, I. & SIEBERT, R. 2009. Hunting for the 5th base: Techniques for analyzing DNA methylation. *Biochim Biophys Acta,* 1790, 847-62.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25, 25-9.

BAIR, E., TIBSHIRANI R. 2004. Semi-supervised methods to predict patient survival from gene expression data. . *PLoS Biol.,* 2.

BARFIELD, R. T., KILARU, V., SMITH, A. K. & CONNEELY, K. N. 2012. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics,* 28, 1280-1.

BARRETINA, J., CAPONIGRO, G., STRANSKY, N., VENKATESAN, K., MARGOLIN, A. A., KIM, S., WILSON, C. J., LEHAR, J., KRYUKOV, G. V., SONKIN, D., REDDY, A., LIU, M., MURRAY, L., BERGER, M. F., MONAHAN, J. E., MORAIS, P., MELTZER, J., KOREJWA, A., JANE-VALBUENA, J., MAPA, F. A., THIBAULT, J., BRIC-FURLONG, E., RAMAN, P., SHIPWAY, A., ENGELS, I. H., CHENG, J., YU, G. K., YU, J., ASPESI, P., JR., DE SILVA, M., JAGTAP, K., JONES, M. D., WANG, L., HATTON, C., PALESCANDOLO, E., GUPTA, S., MAHAN, S., SOUGNEZ, C., ONOFRIO, R. C., LIEFELD, T., MACCONAILL, L., WINCKLER, W., REICH, M., LI, N., MESIROV, J. P., GABRIEL, S. B., GETZ, G., ARDLIE, K., CHAN, V., MYER, V. E., WEBER, B. L., PORTER, J., WARMUTH, M., FINAN, P., HARRIS, J. L., MEYERSON, M., GOLUB, T. R., MORRISSEY, M. P., SELLERS, W. R., SCHLEGEL,

R. & GARRAWAY, L. A. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature,* 483, 603-7.

BAYLIN, S. B. & JONES, P. A. 2011. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer,* 11, 726-34.

BENJAMINI, Y., DRAI, D., ELMER, G., KAFKAFI, N. & GOLANI, I. 2001. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res,* 125, 279-84.

BIBIKOVA, M., BARNES, B., TSAN, C., HO, V., KLOTZLE, B., LE, J. M., DELANO, D., ZHANG, L., SCHROTH, G. P., GUNDERSON, K. L., FAN, J. B. & SHEN, R. 2011. High density DNA methylation array with single CpG site resolution. *Genomics,* 98, 288-95.

BOCK, C. 2012. Analysing and interpreting DNA methylation data. *Nat Rev Genet,* 13, 705-19.

CHEN, Y. A., LEMIRE, M., CHOUFANI, S., BUTCHER, D. T., GRAFODATSKAYA, D., ZANKE, B. W., GALLINGER, S., HUDSON, T. J. & WEKSBERG, R. 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics,* 8, 203-9.

DAVIS, S., P. DU, S. BILKE, T. TRICHE, M. BOOTWALLA. 2011. Methylumi: for handling Illumina DNA methylation data

Bioconductor [Online].

DEDEURWAERDER, S., DEFRANCE, M., CALONNE, E., DENIS, H., SOTIRIOU, C. & FUKS, F. 2011a. Evaluation of the Infinium Methylation 450K technology. *Epigenomics,* 3, 771-84.

DEDEURWAERDER, S., DESMEDT, C., CALONNE, E., SINGHAL, S. K., HAIBE-KAINS, B., DEFRANCE, M., MICHIELS, S., VOLKMAR, M., DEPLUS, R., LUCIANI, J., LALLEMAND, F., LARSIMONT, D., TOUSSAINT, J., HAUSSY, S., ROTHE, F., ROUAS, G., METZGER, O., MAJJAJ, S., SAINI, K., PUTMANS, P., HAMES, G., VAN BAREN, N., COULIE, P. G., PICCART, M., SOTIRIOU, C. & FUKS, F. 2011b. DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol Med,* 3, 726-41.

DU P, K. W., LIN SM 2008. lumi: a pipeline for processing Illumina microarray. *Bioinformatics,* 24, 1547–1548.

DU, P., ZHANG, X., HUANG, C. C., JAFARI, N., KIBBE, W. A., HOU, L. & LIN, S. M. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics,* 11, 587.

EDGAR, R., DOMRACHEV, M. & LASH, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res,* 30, 207-10.

EISEN, M. 1998. Cluster analysis and display of genome-wide expression patterns. . *Proc. Natl Acad. Sci. USA* 95, 14863-14868.

FRASER, H. B., LAM, L. L., NEUMANN, S. M. & KOBOR, M. S. 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, J. B., 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, 359-67.

HANSEN, K. D., ARYEE M. 2013. minfi: Analyze Illumina's 450k methylation arrays. *Bioconductor* [Online].

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* 6 65–70.

HOUSEMAN, E. A., ACCOMANDO, W. P., KOESTLER, D. C., CHRISTENSEN, B. C., MARSIT, C. J., NELSON, H. H., WIENCKE, J. K. & KELSEY, K. T. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics,* 13, 86.

HOUSEMAN, E. A., CHRISTENSEN, B. C., YEH, R. F., MARSIT, C. J., KARAGAS, M. R., WRENSCH, M., NELSON, H. H., WIEMELS, J., ZHENG, S., WIENCKE, J. K. & KELSEY, K. T. 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 Bioinformatics,* 9, 365.

ILLUMINA 2008. Infinium Assay Methylation Protocol Guide.

ILLUMINA. 2011. GenomeStudio/BeadStudio software Methylation Module.

IOANNIDIS, J. P. 2007. Why most published research findings are false: author's reply to Goodman and Greenland. *PLoS Med,* 4, e215.

JOHNSON, W. E., LI, C. & RABINOVIC, A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics,* 8, 118-27.

JOLLIFFE, I. T. 2002. *Principal Component Analysis,* New York.

KILARU, V., BARFIELD, R. T., SCHROEDER, J. W., SMITH, A. K. & CONNEELY, K. N. 2012. MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics,* 7, 225-9.

KOESTLER, D.C., CHRISTENSEN, B.C., KARAGAS, M.R., MARSIT, C.J., LANGEVIN, S.M., KELSEY, K.T., WIENCKE, J.K., HOUSEMAN, E.A. 2013. Blood-based profiles of DNA methylation predict the underlying distribution of cell types. *Epigenetics [IN PRESS].*

KOESTLER, D.C., MARSIT, C.J., CHRISTENSEN, B.C., KARAGAS, M.R., BUENO, R., SUGARBAKER, D.J., , KELSEY, K.T., HOUSEMAN, E.A. 2010. Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics, 26(20), 2578-85.*

KUAN, P. F., WANG, S., ZHOU, X. & CHU, H. 2010. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics,* 26, 2849-55.

LAIRD, P. W. 2010. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet,* 11, 191-203.

LAM, L. L., EMBERLY, E., FRASER, H. B., NEUMANN, S. M., CHEN, E., MILLER, G. E. & KOBOR, M. S. 2012. Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci U S A,* 109 Suppl 2, 17253-60.

LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. & IRIZARRY, R. A. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet,* 11, 733-9.

LEEK, J. T. & STOREY, J. D. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet,* 3, 1724-35.

LIU, Y., ARYEE, M. J., PADYUKOV, L., FALLIN, M. D., HESSELBERG, E., RUNARSSON, A., REINIUS, L., ACEVEDO, N., TAUB, M., RONNINGER, M., SHCHETYNSKY, K., SCHEYNIUS, A., KERE, J., ALFREDSSON, L., KLARESKOG, L., EKSTROM, T. J. & FEINBERG, A. P. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol,* 31, 142-7.

MAKSIMOVIC, J., GORDON, L. & OSHLACK, A. 2012. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol,* 13, R44.

MANCUSO, F. M., MONTFORT, M., CARRERAS, A., ALIBES, 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, M. E., RUHRMANN, S., FAGERSTROM-BILLAI, F., JAGODIC, M., SUNDBERG, C. J., EKSTROM, T. J., TESCHENDORFF, A. E., TEGNER, J. & GOMEZ-CABRERO, D. 2013. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics,* 8, 333-46.

PAN, H., CHEN, L., DOGRA, S., TEH, A. L., TAN, J. H., LIM, Y. I., LIM, Y. C., JIN, S., LEE, Y. K., NG, P. Y., ONG, M. L., BARTON, S., CHONG, Y. S., MEANEY, M. J., GLUCKMAN, P. D., STUNKEL, W., DING, C. & HOLBROOK, J. D. 2012. Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics,* 7, 1173-87.

PEDUZZI, P., CONCATO, J., KEMPER, E., HOLFORD, T. R. & FEINSTEIN, A. R. 1996. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol,* 49, 1373-9.

PETRONIS, A. 2010. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature,* 465, 721-7.

PIDSLEY, R., CC, Y. W., VOLTA, M., LUNNON, K., MILL, J. & SCHALKWYK, L. C. 2013. A data-driven approach to preprocessing Illumina 450K methylation array data. ***BMC Genomics,*** **14, 293.**

PLEROU, V., GOPIKRISHNAN, P., ROSENOW, B., AMARAL, L. A., GUHR, T. & STANLEY, H. E. 2002. Random matrix approach to cross correlations in financial data. ***Phys Rev E Stat Nonlin Soft Matter Phys,*** **65, 066126.**

POLLARD, K. S., M. J. VAN DER LAAN Cluster Analysis of Genomic Data. 2005. ***Bioinformatics and Computational Biology Solutions Using R and Bioconductor*****.**

RAKYAN, V. K., DOWN, T. A., BALDING, D. J. & BECK, S. 2011. Epigenome-wide association studies for common human diseases. ***Nat Rev Genet,*** **12, 529-41.**

SIEGMUND, K. D. 2011. Statistical approaches for the analysis of DNA methylation microarray data. ***Hum Genet,*** **129, 585-95.**

SUN, S., HUANG, Y. W., YAN, P. S., HUANG, T. H. & LIN, S. 2011a. Preprocessing differential methylation hybridization microarray data. ***BioData Min,*** **4, 13.**

SUN, Z., CHAI, H. S., WU, Y., WHITE, W. M., DONKENA, K. V., KLEIN, C. J., GAROVIC, V. D., THERNEAU, T. M. & KOCHER, J. P. 2011b. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. ***BMC Med Genomics,*** **4, 84.**

TESCHENDORFF, A. E., MARABITA, F., LECHNER, M., BARTLETT, T., TEGNER, J., GOMEZ-CABRERO, D. & BECK, S. 2012. A Beta-Mixture Quantile Normalisation method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. ***Bioinformatics*****.**

TESCHENDORFF, A. E., MENON, U., GENTRY-MAHARAJ, A., RAMUS, S. J., GAYTHER, S. A., APOSTOLIDOU, S., JONES, A., LECHNER, M., BECK, S., JACOBS, I. J. & WIDSCHWENDTER, M. 2009. An epigenetic signature in peripheral blood predicts active ovarian cancer. ***PLoS One,*** **4, e8274.**

TESCHENDORFF, A. E. & WIDSCHWENDTER, M. 2012. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. ***Bioinformatics,*** **28, 1487-94.**

TESCHENDORFF, A. E., ZHUANG, J. & WIDSCHWENDTER, M. 2011. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. ***Bioinformatics,*** **27, 1496-505.**

THOMAS, P. D., CAMPBELL, M. J., 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, 2129-41.**

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, 325-41.**

TRICHE, T. J., JR., WEISENBERGER, D. J., VAN DEN BERG, D., LAIRD, P. W. & SIEGMUND, K. D. 2013. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res,* 41, e90.

VAN EIJK, K. R., DE JONG, S., BOKS, M. P., LANGEVELD, T., COLAS, F., VELDINK, J. H., DE KOVEL, C. G., JANSON, E., STRENGMAN, E., LANGFELDER, P., KAHN, R. S., VAN DEN BERG, L. H., HORVATH, S. & OPHOFF, R. A. 2012. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics,* 13, 636.

VAN ITERSON, M., DUIJKERS, F. A., MEIJERINK, J. P., ADMIRAAL, P., VAN OMMEN, G. J., BOER, J. M., VAN NOESEL, M. M. & MENEZES, R. X. 2012. A novel and fast normalization method for high-density arrays. *Stat Appl Genet Mol Biol,* 11.

WANG, D., YAN, L., HU, Q., SUCHESTON, L. E., HIGGINS, M. J., AMBROSONE, C. B., JOHNSON, C. S., SMIRAGLIA, D. J. & LIU, S. 2012. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics,* 28, 729-30.

WEST, J., BECK, S., WANG, X. & TESCHENDORFF, A. E. 2013. An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Sci Rep,* 3, 1630.

WETTENHALL, J. M. & SMYTH, G. K. 2004. limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics,* 20, 3705-6.

WU, Z. & ARYEE, M. J. 2010. Subset quantile normalization using negative control features. *J Comput Biol,* 17, 1385-95.

XU, X., SU, S., BARNES, V. A., 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.

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, A. E. 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. B,* 67, 301–320.

Funding:

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 U.S. NIH grants (R01 CA121147, R01 CA078609, and R01 CA100679. MRK is funded by P20 ES018175, R01 CA57494 and EPA RD83459901.

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
Normalization	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

Figure Legends:

Figure 1: Methylation array data processing and analysis pipeline. Abbreviations: QC= Quality Control

Figure 2: **Quality Control Example from GenomeStudio 450K data** A. Hybridization quality control plot in the green channel. B. Bisulfite conversion quality control plot in the green channel. In

this example, the separation between high and low values indicates that hybridization worked well. Furthermore, bisulfite conversion also performed well as converted controls have a higher signal than unconverted controls.