3-16-2010

ACCURATE GENOME-SCALE PERCENTAGE DNA METHYLATION ESTIMATES FROM MICROARRAY DATA

Martin J. Aryee
Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, aryee@jhu.edu

Zhijin Wu
Department of Community Health, Section of Biostatistics, Brown University

Christine Ladd-Acosta
Center for Epigenetics and Department of Medicine. Johns Hopkins School of Medicine

Brian Herb
Center for Epigenetics and Department of Medicine. Johns Hopkins School of Medicine

Andrew P. Feinberg
Center for Epigenetics and Department of Medicine. Johns Hopkins School of Medicine

See next page for additional authors

Authors
Martin J. Aryee, Zhijin Wu, Christine Ladd-Acosta, Brian Herb, Andrew P. Feinberg, Srinivasan Yegnasurbramanian, and Rafael A. Irizarry

Biostatistics (2010), **0**, 0, *pp.* 1–30
doi:10.1093/biostatistics/

Accurate genome-scale percentage DNA methylation estimates from microarray data

MARTIN J ARYEE*

Sidney Kimmel Comprehensive Cancer Center,
Johns Hopkins University, Baltimore, Maryland, USA
Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
aryee@jhu.edu

ZHIJIN WU

Department of Community Health, Section of Biostatistics,
Brown University, Providence, Rhode Island, USA

CHRISTINE LADD-ACOSTA
BRIAN HERB
ANDREW P FEINBERG

Center for Epigenetics and Department of Medicine,
Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

SRINIVASAN YEGNASUBRAMANIAN

Sidney Kimmel Comprehensive Cancer Center,
Johns Hopkins University, Baltimore, Maryland, USA

RAFAEL IRIZARRY

Department of Biostatistics,
Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
rafa@jhu.edu

SUMMARY

DNA methylation is a key regulator of gene function in a multitude of both normal and abnormal biological processes, but tools to elucidate its roles on a genome-wide scale are still in their infancy. Methylation

*To whom correspondence should be addressed.

sensitive restriction enzymes and microarrays provide a potential high-throughput, low-cost platform to allow methylation profiling. However, accurate absolute methylation estimates have been elusive due to systematic errors and unwanted variability. Previous microarray pre-processing procedures, mostly developed for expression arrays, fail to adequately normalize methylation-related data since they rely on key assumptions that are violated in the case of DNA methylation. We develop a normalization strategy tailored to DNA methylation data and an empirical Bayes percentage methylation estimator, that together yield accurate and precise absolute methylation estimates that can be compared across samples. We illustrate the method on data generated to detect methylation difference between tissues, and between normal and tumor colon samples.

Keywords: Epigenetics, DNA methylation; Microarray.

1. INTRODUCTION

DNA methylation is a chemical modification of DNA that plays a key role in regulation of gene expression. As an *epigenetic* mark, it encodes an additional layer of heritable information on top of DNA without changing the underlying genetic sequence. While all cell types in an organism share nearly the same genome sequence, their DNA methylation patterns can be markedly different (Song *et al.* , 2005; Meissner *et al.* , 2008). DNA methylation marks help encode tissue-specific transcriptional programs in diverse cell types and allows these gene expression patterns to be passed down to daughter cells. Chemically, DNA methylation involves the modification of a cytosine (C) base to form methyl-cytosine. These methylation marks are recognized by specialized proteins that bind the methylated DNA and inhibit the expression of neighboring genes (Bird, 2002). In adult cells of mammals, this modification occurs almost exclusively at cytosines that are immediately followed by a guanine (G) in the 5' to 3' direction, denoted *CpG*.

[FIGURE 1 ABOUT HERE]

The health implications of deciphering the DNA methylation code have recently received much attention both in the scientific literature and in the media (PBS, 2007; Cloud, 2010; SchüBeler, 2009). Smoking in young boys, for example, has been found to cause epigenetic changes that predispose their future offspring to obesity (Pembrey *et al.* , 2006). Similarly, a related study has shown that memory improvements in mice can be passed from one generation to the next via epigenetics (Arai *et al.* , 2009). Work in the rapidly evolving field of stem cell biology has shown that DNA methylation can contribute to the cellular memory mechanism used by stem cells to retain their pluripotent state during repeated cell divisions (Sen *et al.* , 2010). In cancer biology, it is clear that aberrations in DNA methylation almost universally accompany the initiation and progression of cancers (Jones & Baylin, 2002; Feinberg *et al.* , 2006). Much of the excitement surrounding epigenetics relates to the promise of therapies that alter the epigenetic code, activating or silencing disease-related genes. While the majority of such treatments are still hypothetical or experimental, two epigenetic drugs that reactivate tumor suppressor genes by removing methylation marks have recently received FDA approval (Sharma *et al.* , 2010; Kaminskas *et al.* , 2005). These studies and therapies highlight the medical promise of mapping and understanding the role of DNA methylation.

Fully describing the methylation profile of a given cell requires measuring the methylation state of every CpG. However, current practical laboratory protocols do not permit single cell methylation measurements and, because studied cell populations are known to be heterogeneous, methylation measurements are expected to be quantitative. Therefore, for any given cell type we aim to measure the percentage of methylated cells at each CpG site. These measurements can be made by treating DNA with sodium bisulfite, which selectively converts unmethylated cytosine (C) to uracil (U) while leaving methylated C as is, followed by DNA amplification and sequencing (Clark *et al.* , 2006; Frommer *et al.* , 1992). However,

this procedure, although considered a gold standard, comes at significant cost when applied genome-wide due to the amount of sequencing coverage required. Therefore, this technology is not yet suitable for affordable genome-wide measurements. The methods presented in this paper are motivated by the demand for high-throughput measurements necessary to construct genome-wide methylation profiles.

Recent advances in microarray technology and laboratory protocols provide an alternate high-throughput platform for assessing DNA methylation. Since methylation of adjacent cytosines in small regions of a few hundred base pairs is known to be highly correlated (Eckhardt *et al.* , 2006), lower resolution strategies based on methylated DNA enrichment provide a cost-effective alternative to bisulfite sequencing. A common approach involves the use of *methylation-sensitive restriction enzymes*, proteins that selectively cut strands of DNA at either methylated or unmethylated CpG sites (Khulan *et al.* , 2006; Ordway *et al.* , 2006). Following restriction-enzyme treatment methylated and unmethylated DNA fractions are separated based on fragment size. Coupled with DNA-detection microarrays, this strategy can provide accurate genome-wide methylation profiles. In this paper we focus on a protocol using the restriction enzyme McrBC which selectively cuts methylated DNA (Sutherland *et al.* , 1992; Ordway *et al.* , 2006).

The microarray measurements produced by the procedures described above present new statistical challenges. We have developed an empirical Bayes estimation strategy that, when combined with appropriately normalized McrBC-enriched DNA microarray data, produces accurate percentage methylation estimates with a mean error of 10% compared to bisulfite sequencing estimates. Since accurate methylation estimates are highly dependent on suitable pre-processing to remove systematic biases we also present a novel normalization strategy. In so doing, we demonstrate that well-established methods, developed largely in the context of gene expression analysis, are often inappropriate for DNA methylation data. While these methods have been widely and successfully used in a variety of other microarray applications,

certain key assumptions underlying the strategies are violated in the DNA methylation setting leading to inaccurate estimates.

This article is organized as follows. We begin with a description of our example datasets in Section 2. Section 3 lays out limitations of existing methods for pre-processing DNA methylation data. In Section 4 we describe our normalization strategy and the empirical Bayesian estimator of percentage methylation. Results demonstrating the utility of our methods are presented in Section 5. We conclude with a discussion in Section 6. Derivations and practical issues including microarray data quality control are included in the appendix.

2. Data description

2.1 *Microarray data*

DNA methylation assays typically involve random fragmentation of a DNA sample, followed by an enrichment step that selects for either methylated or unmethylated DNA. Prior to enrichment the DNA is split into two fractions one of which is left untreated. Methylation estimates are based on the relative quantities of DNA in the enriched fraction compared to the untreated (total input) fraction. In this paper we use two data sets generated using the restriction-enzyme McrBC. The first is derived from human liver, brain and spleen, with five samples from each tissue. The second consists of five normal colon and five colon cancer samples. The samples are described in Irizarry *et al.* (2009). Each sample was processed and hybridized to microarrays as detailed in Ordway *et al.* (2006) and Irizarry *et al.* (2008). Briefly, the assay involves enrichment of unmethylated DNA using the McrBC enzyme and the CHARM DNA methylation microarray (Irizarry *et al.* , 2008) available from Nimblegen. The CHARM microarray is a 2.1 million probe two-color array designed for maximal CpG coverage.

2.2 *Bisulfite-sequencing validation data*

We generated an independent verification data set using the same samples analyzed on the McrBC/CHARM microarrays. Ten regions containing an average of twelve CpGs and spanning an average of 220 base pairs were selected for methylation analysis by bisulfite treatment followed by sequencing. A total of 110 sequencing runs were performed with a subset of samples chosen for each region, generating percentage methylation estimates for each CpG.

3. MOTIVATION

A key prerequisite for estimation of absolute methylation levels from microarray data is pre-processing to accurately establish the baseline signal level associated with unmethylated regions. The basic measurement used to quantify methylation is the log-ratio (M) of intensities observed in the treated and control (total input) channels. Within-sample normalization aims to transform the log-ratios such that zero represents unmethylated regions and higher M-values represent more methylation. A plot of the microarray methylation log-ratio for probes from unmethylated regions reveals several problems with the raw signal (Figure 3a). First, the M-values are not centered on zero as is desirable for regions without methylation. Second, there is a strong sequence-dependent bias in the signal manifested as a "wave" in a plot of signal by genomic location. This wave is similar to that observed in array-CGH data (Marioni *et al.* , 2007), and can lead to both false negatives and false positives, particularly in calls of absolute methylation levels. The effect represents a sequence-specific bias as evidenced by its strong conservation across samples. Accurate assessment of absolute methylation levels is highly dependent on an analysis approach that adequately corrects for these biases. The Loess normalization strategy which has been widely and successfully used for other two-color microarray applications (Yang *et al.* , 2002) fails to adequately normalize the methylation

log-ratio (Figure 3b).

As with other microarray applications, the strong non-linear dependence of M on signal intensity is a significant source of bias. This effect is evident in a plot of M versus A , where A is the average of the two channel's log signals. The key Loess assumptions are that the majority of probes represent regions without signal (M=0), and that signal intensity-dependent deviations are an artifact. In applications such as differential gene expression analysis or transcription factor binding profiling these assumptions typically hold and Loess regression can effectively be used to estimate and remove signal intensity-dependent bias. In methylation experiments, however, both of these assumptions are violated. Since we expect many sites to be methylated, the average probe behavior no longer represents regions without signal. The baseline signal level estimated through Loess represents the mean methylation level rather than no methylation. If Loess normalization were used here, and we define M = 0 as the average value of unmethlyated sites, then one would incorrectly force M = 0 for many of the probes associated with methylation. This effect is clear from Figure 3b where virtually all unmethylated probes have M ¡ 0.

The second Loess assumption that the methylation log-ratio (M) should be independent of the average individual channel signal intensity (A) is also invalid in the methylation data setting. DNA methylation levels are related to CpG (and hence GC) density; CpG dense regions, referred to as CpG islands tend to be unmethylated, while lone CpGs are usually methylated. Since signal intensity is also known to be influenced by GC content the true M is related to A. Forcing the methylation level to be independent of A through Loess normalization thus actually introduces bias.

A further problem that remains unresolved after Loess normalization is the considerable variability between biological replicates. While established between-sample microarray normalization techniques such as quantile normalization (Bolstad *et al.* , 2003) are in many cases suitable for addressing this prob-

lem in DNA methylation data, we have identified important situations in which alternatives are necessary. Most methods were developed in the context of gene expression data and typically make the assumption that the overall amount of gene expression, and hence signal, should be the same across samples. While not strictly true, this assumption has proved reasonable for most gene expression data. In studies of DNA methylation, similarly, we have found this assumption reasonable for comparisons between normal tissues. In many situations of particular interest, however, such as comparisons between cancer and normal tissues, there may be significant global differences in methylation levels. Total methylation levels are known to vary significantly between, for example, cancer versus normal cells or stem cells versus differentiated cells (Jones & Baylin, 2007; Laurent *et al.* , 2010). As a result, normalizing such samples in a way that equalizes their total signal level may introduce bias. This is illustrated by a hierarchical clustering dendrogram of the colon tissue data set following quantile normalization (Figure 4b). One would expect the biological replicates within the cancer and normal tissue groups to cluster together, but the biological differences are obscured by technical artifacts.

Developments in laboratory tools have provided researchers with a promising platform for accurately and rapidly assaying DNA methylation. As described above, however, signal biases and limitations in current analytical methods present a barrier to making the most effective use of this promising technology.

4. METHODS

We present a two-component strategy for estimating absolute methylation levels. We first normalize the methylation log-ratios to remove systematic bias (Sections 4.1 and 4.2), and then transform the normalized log-ratios into percentage methylation estimates (Section 4.3). All normalization techniques depend on identifying individual features or overall array characteristics that can be assumed to be constant across

samples. Normalization transforms the data to equalize these features between samples. With this in mind we aim to select control probes to serve dual purposes during the normalization process: to set the zero-level for the methylation signal within each sample, and to reduce between sample technical variation.

4.1 *Within sample normalization*

To address the limitations of Loess normalization we employ a method that uses knowledge of genome sequence, assay properties and DNA methylation patterns to select a subset of probes for which its assumptions do hold. By fitting a Loess regression to these control probes we obtain a valid correction curve that can be applied to the remaining probes. The key step in selecting control probes is to identify unmethylated regions of the genome. Since mammalian DNA is almost exclusively methylated at CpG sites, we can typically achieve this by selecting probes from CpG-free regions, guaranteeing a signal that represents unmethylated DNA. For these probes we expect both that M equal zero and that M be independent of A. Ideally, such control probes are included on the array by design. The CHARM microarray, for example, contains 4,500 probes located in CpG-free regions to be used for this purpose. Alternatively, a suitable subset of probes can be identified for many standard array designs, allowing the use of our method with a broad set of platforms.

In addition to setting the zero-level using the control probe Loess procedure, a simple scale normalization with minimal assumptions is useful to establish the signal level associated with fully methylated regions. We typically scale the methylation log-ratios such that the 99th percentile has an M-value corresponding to 99% percent methylation (see Section 4.3). This is roughly equivalent to the assumption that the top 1% of probes represent almost completely methylated regions.

Since methylation estimates are derived from the log-ratio of signal intensity in the two channels, the

presence of background signal biases these estimates towards zero. To address this, we implement background removal prior to the log-ratio calculation using the Robust Multichip Average (RMA) convolution model (Irizarry *et al.* , 2003). The RMA model assumes that the observed intensity is the sum of normally distributed background noise and the true signal, modeled as an exponential. The parameters of the normal and exponential components are empirically estimated using all probes jointly. While removing background signal levels has the benefit of reducing bias this comes at the expense of increased variance (Scharpf *et al.* , 2007). The increase in variance can lead to an inflated false positive rate when identifying methylated or differentially methylated regions in downstream analysis, but this is largely mitigated by taking variance estimates into account.

4.2 *Between sample normalization*

In many situations of particular interest in DNA methylation studies, such as comparisons between cancer and normal tissues, samples might be expected to exhibit significant global differences in methylation levels. In these cases we propose the use of subset quantile normalization (Wu, 2009), a modified version of quantile normalization that avoids assumptions about total signal level. It takes advantage of control probes representing regions known to exhibit the same behaviour across samples. While spike-in controls can serve this purpose it has recently been shown that negative control probes can also be used (Wu, 2009). Since negative controls by design measure non-biological signals they provide a good basis for assessing technical variation between arrays. We find that probes from non-CpG regions (Section 4.1) serve this purpose well. Such probes measure quantities that are not dependent on the methylation status of individual samples and technical variation due to probe effects can be expected to be constant across samples. Since the control probes will be used as reference points during between-sample normalization,

a further consideration is that their sequence characteristics should represent those of the array as a whole, particularly in terms of GC-content.

Due to significant probe effects (Wu, 2009), the negative control features from unmethylated regions span almost the entire dynamic range of signal within the enriched channel (Figure 2b). This is explained by the observation that differences in individual probe hybridization efficiencies and the effects of varying amounts of cross-hybridization are frequently of similar magnitude or greater than the biological differences of interest (Irizarry *et al.* , 2003; Johnson *et al.* , 2006; Li & Wong, 2001; Wu *et al.* , 2004).

The use of subset quantile normalization (Wu, 2009) allows us to take advantage of the facts that the individual negative control features should show the same behaviour across all samples, and that they also cover the dynamic range of the signal probes. In this approach, the control probes are used as "anchors" when normalizing the data. First an empirical reference distribution is created by quantile normalizing the control features. A target distribution is created from a weighted average of this empirical distribution and normal-mixture distribution to allow extrapolation beyond the range of control probe signals. We then map probes that fall in the q^{th} quantile of control probes on their array to the q^{th} quantile of the target distribution.

[FIGURE 2 ABOUT HERE]

We have adapted subset quantile normalization, originally proposed in the context of single-color microarray data, to the two color setting. Figure 2 motivates the use of subset quantile normalization on the enriched channel data rather than directly on the methylation log-ratios (M). On the M scale, where probe effects have mostly been cancelled out, the control probes span a smaller fraction of the signal probe dynamic range and are therefore less useful as normalization anchors. Prior to normalizing the enriched channel we establish a common baseline by first normalizing the total input channel, leaving

M-values unchanged. As the total input channel represents genomic DNA which can be assumed to be essentially identical in all samples, it is reasonable to make an even stronger assumption than that of quantile normalization, and instead set each total input probe to its median value across all samples.

4.3 *Percentage methylation estimates*

The final pre-processing step involves estimation of percentage methylation from the normalized data. The use of a Bayesian estimator ensures that the values are constrained to the appropriate range. The procedures described in Sections 4.1 and 4.2 serve to remove the major within- and between-sample biases. The total input and enriched channels for probe i can be then be represented as $I_i^{T*} = \phi_i \epsilon_i^T$ and $I_i^{E*} = (1 - p_i)\phi_i \epsilon_i^E$ respectively. p_i represents our primary quantity of interest, the percentage of methylated CpGs at a given locus i. ϕ_i captures the probe effect. Taking log-ratios of the observed intensities gives $m_i = q_i + e_i$, where $q_i = -\log(1 - p_i)$ and $e_i = \log \epsilon_i^T - \log \epsilon_i^E$. Larger values of q_i represent more methylation with zero representing no methylation. Pre-processing ensures that the error term is centered at 0. Examination of empirical distributions suggests it is reasonable to use a log-normal model for the error terms ϵ_i and an exponential model for q_i. To allow for the non-zero probability of no methylation ($q_i = 0$) we model q_i as a mixture of a point mass at 0 and $exp(\alpha)$. Under these conditions we are able to derive a closed form estimator of q_i. This model is similar to the Robust Multichip Average convolution model with the differences that the normal component, e_i, is centered at 0 and not truncated, and that the signal component, q_i, is modelled as a mixture of a point mass and an exponential distribution. We calculate the expected value of q_i given the observed log-ratio, m_i, and then solve for p_i.

$$E(q_i \mid m_i) = \left[a + \sigma \frac{\phi(a/\sigma)}{\Phi(a/\sigma)} \right] \cdot P(q_i > 0) \tag{4.1}$$

where $a = m - \alpha\sigma^2$. The derivation of equation 4.1 is given in the appendix.

5. RESULTS

5.1 *Within-sample normalization*

Figure 3c shows the methylation signal in unmethylated regions following our control probe Loess procedure. Data from the 25 individual samples is shown in grey with the median across samples in black. Unlike Loess normalization (Figure 3b), our procedure effectively establishes the baseline zero-level necessary to make accurate estimates of absolute methylation levels. In addition we find that over 80% of the variation due to the wave effect in McrBC/CHARM DNA methylation data is explained by a non-linear function of the individual channel intensities and can therefore be mitigated by our control probe Loess procedure.

[FIGURE 3 ABOUT HERE]

5.2 *Between-sample normalization*

As might be expected, when comparing samples with significantly different overall levels of DNA methylation we find that quantile normalization introduces biases that can obscure true biological differences. This is evident in a hierarchical clustering dendrogram of the colon tissue data set (Figure 4). Panel (a) shows samples clustered following subset quantile normalization. The biological differences between cancer and normal tissue clearly divides the samples into two groups. On the other hand, the sample clustering breaks down following quantile normalization (panel b), suggesting that artifacts have been introduced, obscuring the biological differences.

[FIGURE 4 ABOUT HERE]

As a second metric of ability to retain biological variation while suppressing technical variation we also examined the behaviour of the top 10,000 most variable probes by calculating F-statistics for group differences. Given that we expect true differences between tumor and non-tumor tissue, larger F-statistics are desirable and indicative of better ability to detect between-group differences. The mean F-statistics following quantile and subset quantile normalization are 5.2 and 7.6 respectively, suggesting that subset quantile normalization retains considerably more real biological signal while reducing technical variability. On the other hand, when comparing F-statistics for the normal tissue data set where the samples have similar global methylation levels, full quantile normalization achieves greater between-group separability.

5.3 *Validating microarray percentage methylation estimates*

We compared array-derived estimates of percentage methylation with an independent bisulfite sequencing data set targeting 10 regions. We summarized the sequencing reactions by the median percentage methylation across the CpGs in the region for a total of 110 methylation estimates across the 25 samples. Corresponding microarray estimates are derived from the median of the 2-8 probes in the region.

[FIGURE 5 ABOUT HERE]

Figure 5 plots array-derived percentage methylation estimates against the bisulfite sequencing data. The correlation between the microarray estimates and bisulfite sequencing data is high at 93% with an average discrepancy of 10%. These results suggest that the microarray-derived estimates are a good proxy for bisulfite sequencing data.

[FIGURE 6 ABOUT HERE]

Our results also highlight the importance of background subtraction when estimating methylation levels. Figure 6 shows the error in microarray estimates of percentage methylation made with and without

background removal, as compared to the bisulfite sequencing verification data. Only when background removal is used as part of the pre-processing procedure are the microarray-derived estimates centered on the gold-standard methylation values.

6. Discussion & Conclusion

The strategy presented here involves a computationally efficient closed-form Bayesian estimator of percentage methylation coupled with normalization methods tailored to DNA methylation microarray data. We demonstrate that the technique, together with data generated using the McrBC methylation-sensitive restriction enzyme and the CHARM DNA methylation microarray, achieves a high-degree of correlation with bisulfite sequencing data with an average discrepancy of 10%.

Both the within-sample (between-channel) and between-sample normalization methods hinge on identifying suitable control probes from unmethylated regions. Since mammalian cells are almost exclusively methylated at CpG sites, this can typically be achieved by identifying stretches of CpG-free DNA. Depending on the properties of the methylation assay it may be possible to relax this CpG-free requirement. Since most methylation-sensitive restriction enzymes only recognize CpGs when flanked by specific bases, other CpGs are essentially invisible to the enzyme and need not be excluded when selecting control regions. Choosing suitable control probes is slightly more complicated in systems where cells may have significant levels of non-CpG methylation, as has been demonstrated in human stem cells (Lister *et al.* , 2009; Ramsahoye *et al.* , 2000). One solution is to choose to study only CpG sites through the use of a CpG-specific enrichment strategy. In this case, non-CpG methylation is undetectable and the standard control probe selection procedure can be applied.

Between-sample normalization is complicated by the possibility of significantly different levels of

total DNA methylation between samples. Such comparisons, such as between cancer and normal cells, are often of particular interest from a DNA methylation perspective. Our results suggest that in situations where we have strong a priori reason to believe that global methylation differences exist, subset quantile normalization is superior to quantile normalization since it avoids the assumption of equality in global methylation levels. When this assumption is significantly violated, quantile normalization introduces significant bias that may mask the underlying biological signal. In situations where overall methylation levels are not drastically different, on the other hand, the stronger assumptions of quantile normalization are to be preferred. Since subset quantile normalization makes weaker assumptions about the data, it therefore has less ability to correct large between-sample biases. In effect, successful use of subset quantile normalization is dependent on high quality data to a greater extent than quantile normalization.

As genomics seeks to unravel the diverse methylomes represented across cell types, it will be essential to have accurate and affordable high-throughput methods to query methylation. The analytical methods presented here will help provide a cost-effective means to globally profile DNA methylation by leveraging what we already know about genome structure and methylation patterns. Efforts to map out developmental and disease pathways will be fortified by a better understanding of genomic methylation patterns.

7. FUNDING

This work was supported by: NIH/NCI grant P50CA58236, Department of Defense Prostate Cancer Research Program grant PC073533, and the Maryland Stem Cell Research Fund (MSCRFE_0102-00). *Conflict of Interest:* None declared.

REFERENCES

ARAI, JUNKO A, LI, SHAOMIN, HARTLEY, DEAN M, & FEIG, LARRY A. 2009. Transgenerational rescue of a genetic defect in long-term potentiation and memory formation by juvenile enrichment. *J Neurosci*, **29**(5), 1496–502.

BIRD, ADRIAN. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev*, **16**(1), 6–21.

BOLSTAD, B M, IRIZARRY, R A, ASTRAND, M, & SPEED, T P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**(2), 185–93.

CLARK, SUSAN J, STATHAM, AARON, STIRZAKER, CLARE, MOLLOY, PETER L, & FROMMER, MARIANNE. 2006. DNA methylation: bisulphite modification and analysis. *Nat Protoc*, **1**(5), 2353–64.

CLOUD, J. 2010. Why Your DNA Isn't Your Destiny. *Time Magazine*.

ECKHARDT, FLORIAN, LEWIN, JOERN, CORTESE, RENE, RAKYAN, VARDHMAN K, ATTWOOD, JOHN, BURGER, MATTHIAS, BURTON, JOHN, COX, TONY V, DAVIES, ROB, DOWN, THOMAS A, HAEFLIGER, CAROLINA, HORTON, ROGER, HOWE, KEVIN, JACKSON, DAVID K, KUNDE, JAN, KOENIG, CHRISTOPH, LIDDLE, JENNIFER, NIBLETT, DAVID, OTTO, THOMAS, PETTETT, ROGER, SEEMANN, STEFANIE, THOMPSON, CHRISTIAN, WEST, TONY, ROGERS, JANE, OLEK, ALEX, BERLIN, KURT, & BECK, STEPHAN. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, **38**(12), 1378.

FEINBERG, ANDREW P, OHLSSON, ROLF, & HENIKOFF, STEVEN. 2006. The epigenetic progenitor origin of human cancer. *Nature Reviews Genetics*, **7**(1), 21.

FROMMER, M, MCDONALD, L E, MILLAR, D S, COLLIS, C M, WATT, F, GRIGG, G W, MOLLOY, P L, & PAUL, C L. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA*, **89**(5), 1827–31.

IRIZARRY, RAFAEL, HOBBS, BRIDGET, COLLIN, FRANCOIS, BEAZER-BARCLAY, YASMIN, ANTONELLIS, KRISTEN, SCHERF, UWE, & SPEED, TERENCE. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**(2), 249.

IRIZARRY, RAFAEL A, LADD-ACOSTA, CHRISTINE, CARVALHO, BENILTON, WU, HAO, BRANDENBURG,

SHERI A, JEDDELOH, JEFFREY A, WEN, BO, & FEINBERG, ANDREW P. 2008. Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.*, **18**(5), 780–90.

IRIZARRY, RAFAEL A, LADD-ACOSTA, CHRISTINE, WEN, BO, WU, ZHIJIN, MONTANO, CAROLINA, ONYANGO, PATRICK, CUI, HENGMI, GABO, KEVIN, RONGIONE, MICHAEL, WEBSTER, MAREE, JI, HONG, POTASH, JAMES, SABUNCIYAN, SARVEN, & FEINBERG, ANDREW P. 2009. Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat Genet*, **41**(2), 178.

JOHNSON, W EVAN, LI, WEI, MEYER, CLIFFORD A, GOTTARDO, RAPHAEL, CARROLL, JASON S, BROWN, MYLES, & LIU, X SHIRLEY. 2006. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci USA*, **103**(33), 12457–62.

JONES, PETER A, & BAYLIN, STEPHEN B. 2002. The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, **3**(6), 415.

JONES, PETER A, & BAYLIN, STEPHEN B. 2007. The epigenomics of cancer. *Cell*, **128**(4), 683–92.

KAMINSKAS, EDVARDAS, FARRELL, ANN, ABRAHAM, SOPHIA, BAIRD, AMY, HSIEH, LI-SHAN, LEE, SHWU-LUAN, LEIGHTON, JOHN K, PATEL, HASMUKH, RAHMAN, ATIQUR, SRIDHARA, RAJESHWARA, WANG, YONG-CHENG, PAZDUR, RICHARD, & FDA. 2005. Approval summary: azacitidine for treatment of myelodysplastic syndrome subtypes. *Clin. Cancer Res.*, **11**(10), 3604–8.

KHULAN, BATBAYAR, THOMPSON, REID F, YE, KENNY, FAZZARI, MELISSA J, SUZUKI, MASAKO, STASIEK, EDYTA, FIGUEROA, MARIA E, GLASS, JACOB L, CHEN, QUAN, MONTAGNA, CRISTINA, HATCHWELL, ELI, SELZER, REBECCA R, RICHMOND, TODD A, GREEN, ROLAND D, MELNICK, ARI, & GREALLY, JOHN M. 2006. Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res.*, **16**(8), 1046–55.

LAURENT, LOUISE, WONG, ELEANOR, LI, GUOLIANG, HUYNH, TIEN, TSIRIGOS, ARISTOTELIS, ONG, CHIN THING, LOW, HWEE MENG, SUNG, KEN WING KIN, RIGOUTSOS, ISIDORE, LORING, JEANNE, & WEI, CHIA-LIN. 2010. Dynamic changes in the human methylome during differentiation. *Genome Res.*, Feb.

LI, C, & WONG, W H. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and

outlier detection. *Proc Natl Acad Sci USA*, **98**(1), 31–6.

LISTER, RYAN, PELIZZOLA, MATTIA, DOWEN, ROBERT H, HAWKINS, R DAVID, HON, GARY, TONTI-FILIPPINI, JULIAN, NERY, JOSEPH R, LEE, LEONARD, YE, ZHEN, NGO, QUE-MINH, EDSALL, LEE, ANTOSIEWICZ-BOURGET, JESSICA, STEWART, RON, RUOTTI, VICTOR, MILLAR, A HARVEY, THOMSON, JAMES A, REN, BING, & ECKER, JOSEPH R. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, Jan, 1–8.

MARIONI, JOHN C, THORNE, NATALIE P, VALSESIA, ARMAND, FITZGERALD, TOMAS, REDON, RICHARD, FIEGLER, HEIKE, ANDREWS, T DANIEL, STRANGER, BARBARA E, LYNCH, ANDREW G, DERMITZAKIS, EMMANOUIL T, CARTER, NIGEL P, TAVARÉ, SIMON, & HURLES, MATTHEW E. 2007. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.*, **8**(10), R228.

MEISSNER, ALEXANDER, MIKKELSEN, TARJEI S, GU, HONGCANG, WERNIG, MARIUS, HANNA, JACOB, SIVACHENKO, ANDREY, ZHANG, XIAOLAN, BERNSTEIN, BRADLEY E, NUSBAUM, CHAD, JAFFE, DAVID B, GNIRKE, ANDREAS, JAENISCH, RUDOLF, & LANDER, ERIC S. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**(7205), 766.

ORDWAY, J, BEDELL, J, CITEK, R, NUNBERG, A, GARRIDO, A, KENDALL, R, STEVENS, J, CAO, D, DOERGE, R, KORSHUNOVA, Y, HOLEMON, H, MCPHERSON, J, LAKEY, N, LEON, J, MARTIENSSEN, R, & JEDDELOH, J. 2006. Comprehensive DNA methylation profiling in a human cancer genome identifies novel epigenetic targets. *Carcinogenesis*, **27**(12), 2409.

PBS. 2007. Ghost in Your Genes. *NOVA*.

PEMBREY, MARCUS E, BYGREN, LARS OLOV, KAATI, GUNNAR, EDVINSSON, SÖREN, NORTHSTONE, KATE, SJÖSTRÖM, MICHAEL, GOLDING, JEAN, & TEAM, ALSPAC STUDY. 2006. Sex-specific, male-line transgenerational responses in humans. *Eur J Hum Genet*, **14**(2), 159–66.

RAMSAHOYE, BERNARD H, BINISZKIEWICZ, DETLEV, LYKO, FRANK, CLARK, VICTORIA, BIRD, ADRIAN P, & JAENISCH, RUDOLF. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci USA*, **97**(10), 5237.

SCHARPF, ROBERT B, IACOBUZIO-DONAHUE, CHRISTINE A, SNEDDON, JULIE B, & PARMIGIANI, GIOVANNI. 2007. When should one subtract background fluorescence in 2-color microarrays? *Biostatistics (Oxford, England)*, **8**(4), 695–707.

SCHÜBELER, D. 2009. Epigenomics: Methylation matters. *Nature*, **462**(7271), 296.

SEN, GEORGE L, REUTER, JASON A, WEBSTER, DANIEL E, ZHU, LILLY, & KHAVARI, PAUL A. 2010. DNMT1 maintains progenitor function in self-renewing somatic tissue. *Nature*, **463**(7280), 563.

SHARMA, SHIKHAR, KELLY, THERESA, & JONES, PETER. 2010. Epigenetics in cancer. *Carcinogenesis*, **31**(1), 27.

SONG, FEI, SMITH, JOSEPH F, KIMURA, MAKOTO T, MORROW, ARLENE D, MATSUYAMA, TOMOKI, NAGASE, HIROKI, & HELD, WILLIAM A. 2005. Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc Natl Acad Sci USA*, **102**(9), 3336–41.

SUTHERLAND, E, COE, L, & RALEIGH, E A. 1992. McrBC: a multisubunit GTP-dependent restriction endonuclease. *J Mol Biol*, **225**(2), 327–48.

THOMPSON, REID F, REIMERS, MARK, KHULAN, BATBAYAR, GISSOT, MATHIEU, RICHMOND, TODD A, CHEN, QUAN, ZHENG, XIN, KIM, KAMI, & GREALLY, JOHN M. 2008. An analytical pipeline for genomic representations used for cytosine methylation studies. *Bioinformatics*, **24**(9), 1161–7.

WU, ZHIJIN. 2009. Subset Quantile Normalization using Negative Control Features. *Johns Hopkins University, Dept. of Biostatistics Working Papers*.

WU, ZHIJIN, IRIZARRY, RAFAEL A, GENTLEMAN, ROBERT, MARTINEZ-MURILLO, FRANCISCO, & SPENCER, FORREST. 2004. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat Assoc*, **99**(468), 909–917.

YANG, YEE HWA, DUDOIT, SANDRINE, LUU, PERCY, LIN, DAVID M, PENG, VIVIAN, NGAI, JOHN, & SPEED, TERENCE P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**(4), e15.

8. APPENDIX

8.1 *Derivation of equation 4.1*

We derive the posterior expectation of q_i given the observed log-ratio m_i. q_i is modelled as $exp(\alpha)$ and e_i as $N(0, \sigma^2)$ where α and σ^2 are estimated empirically from the data. We drop the i subscript for convenience.

$$E(q \mid m) = 0 \cdot P(q = 0) + E(q \mid q > 0, m) \cdot P(q > 0)$$

The joint distribution of q and e is given by

$$f_{q,e|q>0}(q, e) = \alpha \exp(-\alpha q) \frac{1}{\sigma} \phi(\frac{e}{\sigma})$$

Since the Jacobian of the transformation is 1 the joint distribution of q and m is simply

$$f_{q,m|q>0}(q, m) = \alpha \exp(-\alpha q) \frac{1}{\sigma} \phi(\frac{q - m}{\sigma})$$

The marginal distribution of m is given by

$$f_{m|q>0}(m) = \int_0^\infty f_{q,m|q>0}(q, m) dq$$

Substituting $w = (q - m)/\sigma$ we have

$$\begin{aligned} f_{m|q>0}(m) = \textstyle\int_{-m/\sigma}^{\infty} \alpha \exp(-\alpha(\sigma w + m))\Phi(w)\, dw & \\ = \quad \alpha \exp(-\alpha m) \int_{-m/\sigma}^{\infty} \exp(-\alpha\sigma w) \frac{1}{\sqrt{2\pi}} \exp(-\frac{w^2}{2}) dw & \\ = \quad \alpha \exp(-\alpha m) \exp(\frac{\alpha^2\sigma^2}{2}) \int_{-m/\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(w+\sigma\alpha)^2}{2} dw & \end{aligned}$$

By substituting $z = w + \sigma\alpha$ we obtain the right hand side:

$$\begin{aligned} & \alpha \exp(-\alpha m) \exp(\frac{\alpha^2\sigma^2}{2}) \int_{-m/\sigma+\sigma\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) dz \\ = \alpha \exp & \left(\frac{\alpha^2\sigma^2}{2} - \alpha m \right) \Phi \left(\frac{m - \alpha\sigma^2}{\sigma} \right) \end{aligned}$$

We can then write the posterior distribution of q given the data m,

$$\begin{aligned} f_{q|q>0,m}(q \mid m) &= \frac{f_{q,m|q>0}(q,m)}{f_{m|q>0}(m)} \\ &= \frac{\alpha \exp(-\alpha q) \frac{1}{\sigma} \phi(\frac{q-m}{\sigma})}{\alpha \exp\left(\frac{\alpha^2\sigma^2}{2} - \alpha m\right) \Phi\left(\frac{m-\alpha\sigma^2}{\sigma}\right)} \\ &= \frac{\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(q - (m - \alpha\sigma^2)^2)}{\Phi(\frac{m-\alpha\sigma^2}{\sigma})} \\ &= \frac{\frac{1}{\sigma}\phi(\frac{q-a}{\sigma})}{\Phi(\frac{a}{\sigma})} \end{aligned}$$

where $a = m - \alpha\sigma^2$. By substituting $z = (q - a)/\sigma$ we obtain the posterior expectation, conditional on $q > 0$:

$$
\begin{aligned}
E(q \mid q > 0, m) &= \frac{1}{\Phi(\frac{a}{\sigma})} \int_0^\infty \frac{q}{\sigma} \phi\left(\frac{q-a}{\sigma}\right) dq \\
&= \frac{1}{\Phi(\frac{a}{\sigma})} \int_{-a/\sigma}^\infty (\sigma z + a)\phi(z) dz \\
&= \frac{1}{\Phi(\frac{a}{\sigma})} a \int_{-a/\sigma}^\infty \phi(z) dz + \frac{1}{\Phi(\frac{a}{\sigma})} \sigma \int_{-a/\sigma}^\infty z\phi(z) dz \\
&= \frac{1}{\Phi(\frac{a}{\sigma})} \left[a\Phi\left(\frac{a}{\sigma}\right) + \sigma\phi\left(\frac{a}{\sigma}\right) \right] \\
&= a + \sigma \frac{\phi(\frac{a}{\sigma})}{\Phi(\frac{a}{\sigma})}
\end{aligned}
$$

Thus

$$
E(q \mid m) = \left[a + \sigma \frac{\phi(\frac{a}{\sigma})}{\Phi(\frac{a}{\sigma})} \right] \cdot P(q > 0)
$$

$P(q_i = 0)$ is estimated by the fraction of $m_i < 0$. The parameters α and σ^2 are estimated as in the RMA convolution model, but with GC-stratification and with the restriction that the normal component be centered at 0.

8.2 *Microarray data quality assessment*

Data quality metrics provide a useful tool for identifying outlier probes or entire arrays that should be considered for exclusion from the analysis. Methylation levels are estimated by comparing the treated (enriched) channel to the untreated (total input) channel. In the case of the McrBC approach for instance, methylation levels can be estimated from the amount of depletion in the treated channel compared to the untreated channel. As a result, the range of measurable methylation (the dynamic range) is determined in large part by the quality of the untreated channel signal. Since the untreated channel measures total

DNA, all probes are expected to record a high signal. Similar to the approach of Thompson *et al.* (2008) we assess the quality of the untreated channel signal by comparing these probes to the signal from the background probes that measure cross-hybridization and scanner optical noise. We define a probe's quality score as its percentile rank among those background probes with the same GC-content. Probes with consistently low scores can be flagged for exclusion from the analysis. Similarly, the array quality score, defined as the mean probe score, is a useful metric for identifying outlier arrays to be removed.

A heatmap plot of probe intensity by physical location is a second useful tool for identifying hybridization problems. Since probes are typically located randomly across an array, we do not expect any spatial bias in signal strength. Both channels should show uniform signal intensity over the physical array. This is particularly useful for the enriched channel where we cannot compare probes to the background level since low intensity is indicative of methylation.

8.3 *CpG density / Fragment length bias*

Enrichment of methylated DNA by restriction enzyme based approaches has been reported to be dependent on the digested fragment length (Thompson *et al.* , 2008). This bias is largely believed to be the result of PCR amplification whose efficiency is size-dependent. Thompson *et al.* (2008) present a normalization scheme to adjust for this bias. A second contributing factor to the dependence may be a true relationship between methylation levels and fragment length. This is reasonable given that restriction fragment length is dependent on CpG density which is known to be a determinant of methylation levels. To isolate the effect of these factors we generated a fully methylated sample by in-vitro treatment with Sss1 methylase. We examined the effect of fragment length in the context of the CHARM assay by plotting the median enrichment log-ratio by fragment length, as estimated using McrBC recognition sites. While the relation-

ship is similar to that described previously for the samples with normal methylation levels (Figure 7a), it does not hold for a fully methylated sample (Figure 7b). This suggests that, in the context of this assay, the methylation log-ratio need not be corrected for fragment length biases. Further evidence for lack of bias is provided by comparing the methylation log-ratios to independent sequencing verification data, where we observe no relationship between error and fragment length (Figure 8).

[FIGURE 7 ABOUT HERE]

[FIGURE 8 ABOUT HERE]

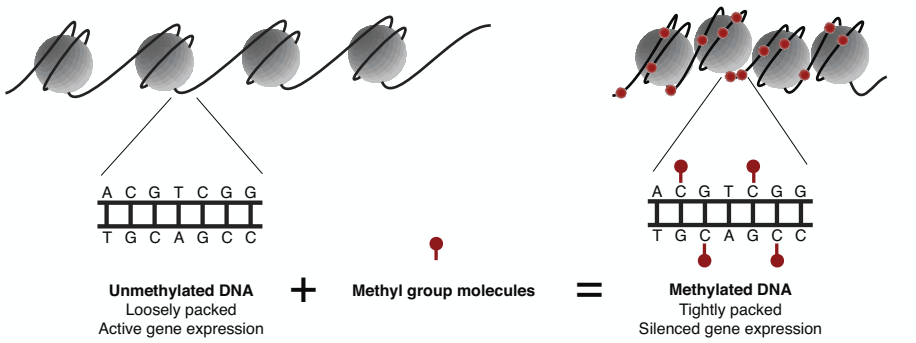

Fig. 1. DNA (black strand) is wrapped around histone proteins (grey spheres). Unmethylated DNA (left) tends to be loosely packed. Genes in such regions are accessible to the cell's transcriptional machinery and can be expressed. DNA methylation involves the addition of methyl group molecules (red circles) to cytosine bases. Highly methylated DNA (right) is tightly packed resulting in silenced gene expression.

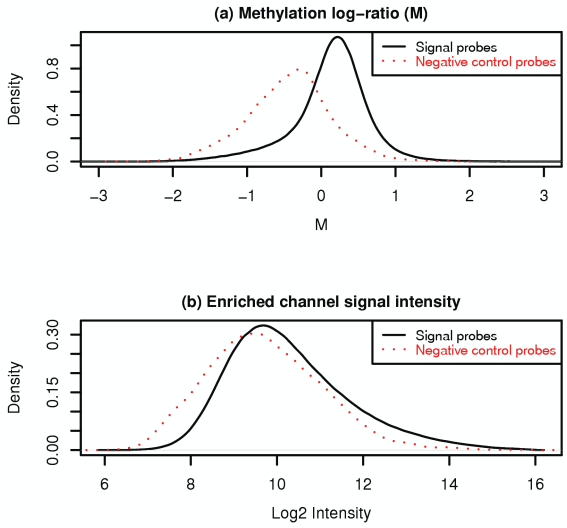

Fig. 2. While negative control features representing unmethylated regions have lower signals than the signal probes on the M (log-ratio) scale (a), they span almost the entire dynamic range of signal in the enriched channel (b) as a result of probe effects.

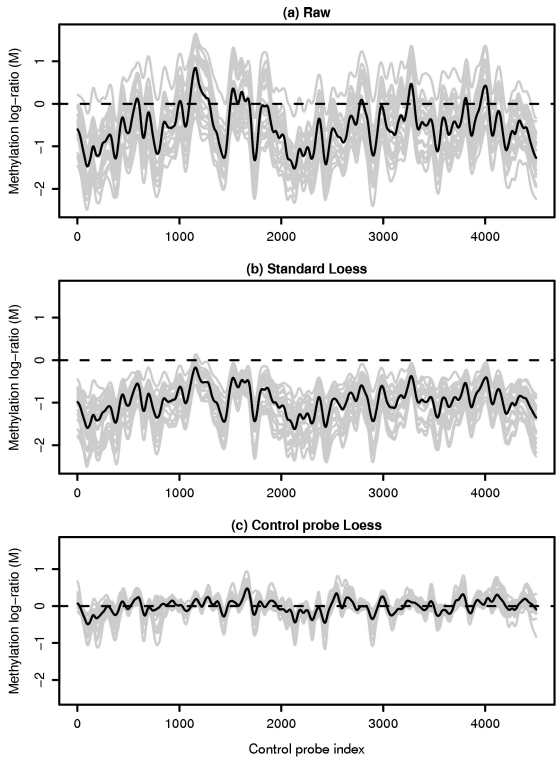

Fig. 3. Methylation log ratio across 30 unmethylated control regions in a CHARM microarray. The light grey lines show individual sample profiles while the dark lines represent the median signal across samples, clearly showing strong conservation of the 'wave' artifact between samples. Neither the raw data (a) nor the standard Loess normalized (b) signals are zero-centered as is desirable for unmethylated regions. Control-probe Loess normalization (c) achieves both a mean-zero signal for unmethylated regions and an 80% reduction in variation compared to the raw signal.

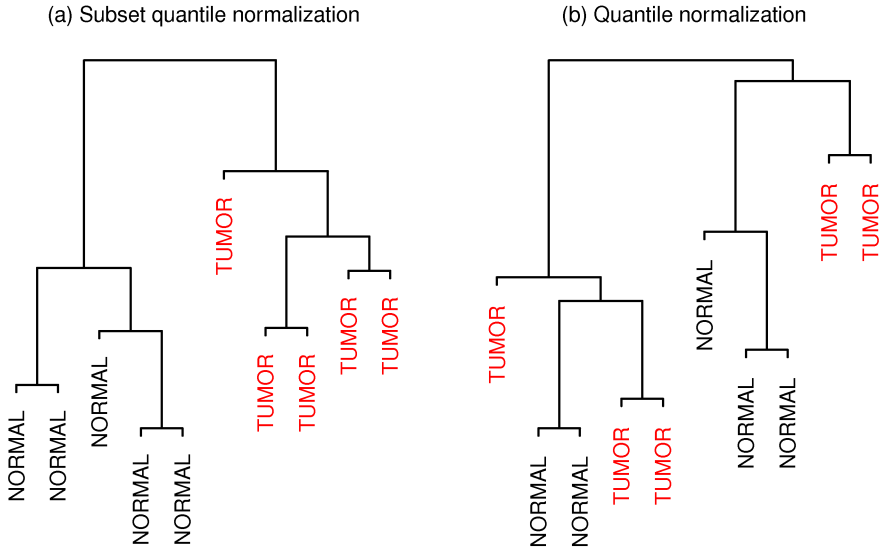

Fig. 4. Hierarchical clustering dendrogram of 5 normal colon and 5 colon tumor samples following (a) subset quantile normalization, and (b) quantile normalization. Subset quantile normalization results in perfect group separation. The top 10,000 most variable probes are used in each case.

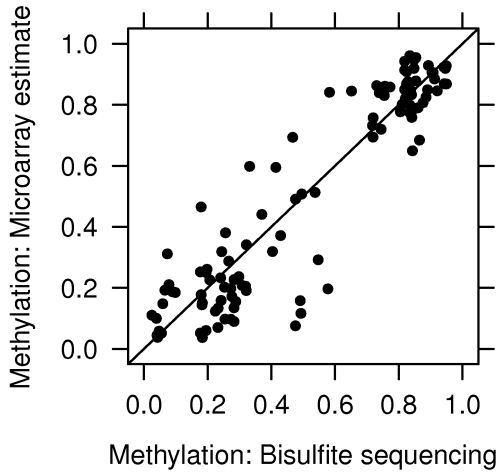

Fig. 5. Percentage methylation estimates. The y-axis shows microarray DNA methylation estimates derived from the median of the probes in each validation region. The x-axis shows methylation from an independent gold-standard validation data set obtained by bisulfite treatment and sequencing. The mean difference between microarray and gold-standard estimates is 10%.

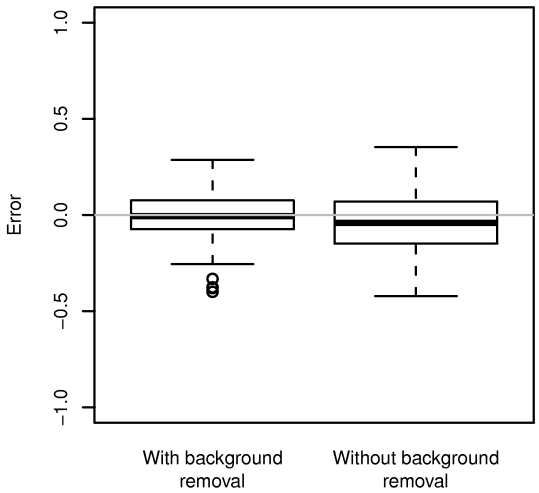

Fig. 6. Error in microarray estimates of percentage methylation with and without background removal. Bisulfite sequencing was used as the gold standard measurement.

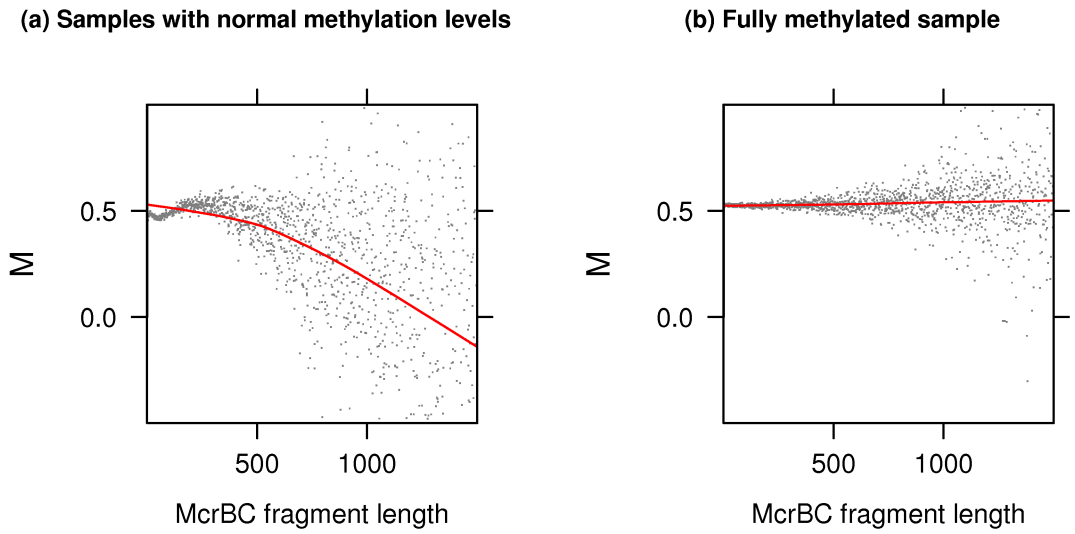

Fig. 7. The x-axis shows restriction fragment length and the y-axis shows the median enrichment log-ratio for (a) the tissue samples with normal methylation levels, and b) a fully methylated sample.

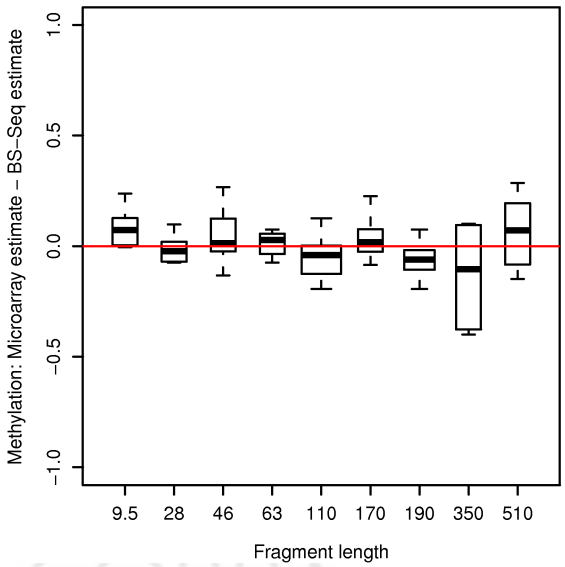

Fig. 8. DNA fragment length and CpG density do not bias the methylation estimate. The data represents the discrepancy between microarray percentage methylation estimates and an independent bisulfite sequencing verification data set.