Biostatistics (2010), **1**, 1, *pp.* 1–11

doi:10.1093/biostatistics/bio000

Supplementary material to:

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

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

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

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

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

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

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

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

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

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

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

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. $p_0 = 1 - P(q_i > 0)$ is estimated by the fraction of probes for which $m_i < 0$. We assessed the sensitivity of percentage methylation estimates to p_0 estimation. Varying p_0 estimates across the range observed in the 25 samples (7%-17%) resulted in a maximum percentage methylation change of 4% suggesting robustness to p_0 estimation error.

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 *and others* (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 ($<$75% in this paper, for example) 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.

3. Background signal removal

Background signal is removed using a modified version of the Robust Multichip Average (RMA) convolution model (Irizarry *and others*, 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. We modify the RMA hyperparameter estimation procedure by taking advantage of anti-genomic background probes, available on most current array designs, to more accurately estimate the background component. In addition, we use GC-stratification to account for the dependence on background signal level with GC-content.

While removing background signal levels has the benefit of reducing bias this comes at the expense of increased variance (Scharpf *and others*, 2007). The increase in variance can potentially 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. 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 *and others*, 2008). This bias is largely believed to be the result of PCR amplification whose efficiency is size-dependent. Thompson *and others* (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 relationship is similar to that described previously for the samples with normal methylation levels

(Figure 1a), it does not hold for a fully methylated sample (Figure 1b). 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 2).

5.

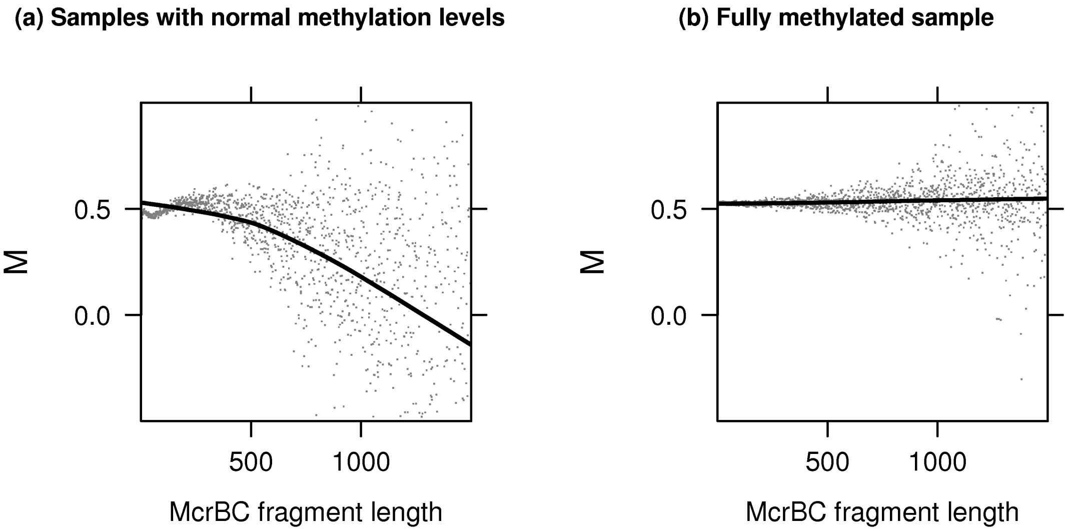

Fig. 1. The x-axis shows restriction fragment length and the y-axis shows the median enrichment log-ratio for (a) the tissue samples with normal methylation levels, and b) a fully methylated sample.

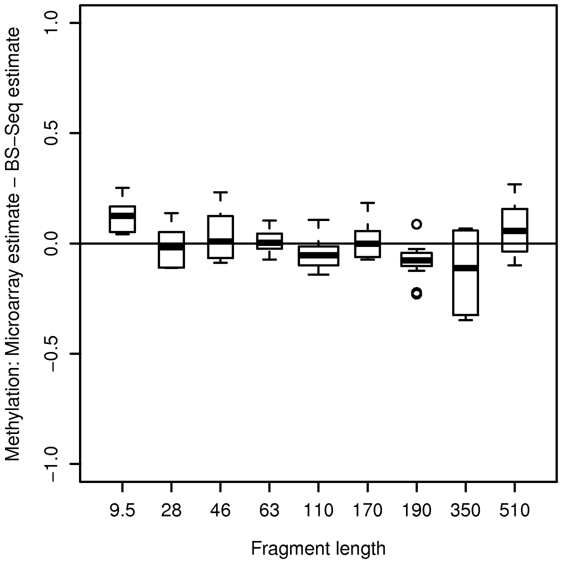

Fig. 2. DNA fragment length and CpG density do not bias the methylation estimate. The data represents the discrepancy between microarray percentage methylation estimates and an independent bisulfite sequencing verification data set.

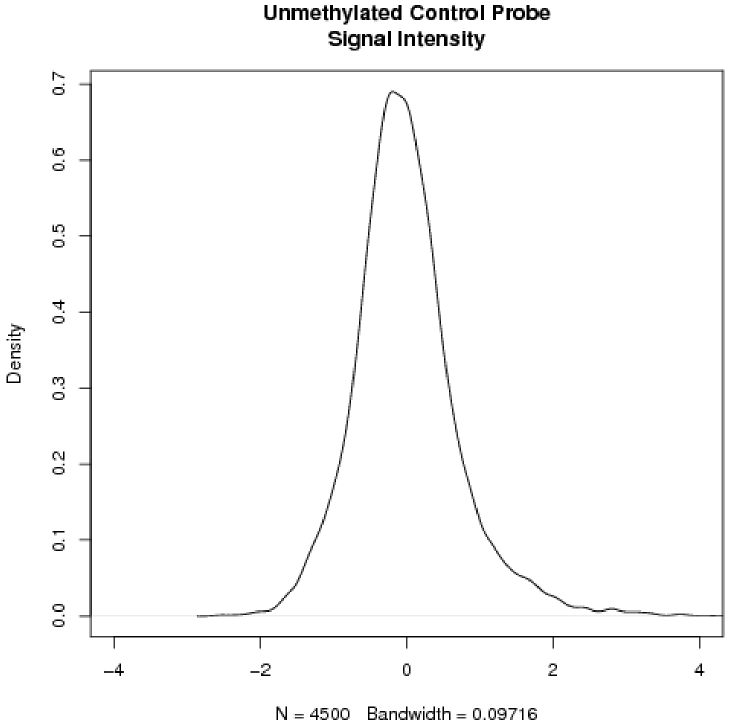

Fig. 3. The distribution of unmethylated control probe log-ratios is centered at 0 following pre-processing.

REFERENCES

IRIZARRY, RAFAEL A, HOBBS, BRIDGET, COLLIN, FRANCOIS, BEAZER-BARCLAY, YASMIN D, ANTONELLIS, KRISTEN J, SCHERF, UWE AND SPEED, TERENCE P. (2003, Aug). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)* **4**(2), 249–64.

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

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