# BayMeth: Improved DNA methylation estimation for affinity capture sequencing data using a flexible Bayesian approach

Andrea Riebler[1,2,*], Jenny Z. Song[3], Aaron L. Statham[3], Clare Stirzaker[3,4], Mirco Menigatti[5], Nadiya Mahmud[6], Charles A. Mein[6], Giancarlo Marra[5], Susan J. Clark[3,4], Mark D. Robinson[1,7,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[2]Institute of Social- and Preventive Medicine, University of Zurich, Hirschengraben 84, CH-8001 Zurich, Switzerland
[3]Epigenetics Laboratory, Cancer Research Program, Garvan Institute of Medical Research, Sydney 2010, New South Wales, Australia
[4]St Vincent's Clinical School, University of NSW, Sydney 2052, NSW, Australia
[5]Institute of Molecular Cancer Research, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[6]Genome Centre, Barts and the London, Queen Mary, University of London, Charterhouse Square, London EC1M 6BQ, United Kingdom
[7]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

Email: Andrea Riebler*- andrea.riebler@uzh.ch; Mark D. Robinson*- mark.robinson@imls.uzh.ch;

*Corresponding author

Version: August 21, 2012

## Abstract

**Background:** DNA methylation (DNAme) is a vital component of the epigenetic regulatory machinery and aberrations occur in many diseases, such as cancer and diabetes. In light of recent demethylating therapeutic agents, mapping and understanding DNAme profiles offers considerable promise for reversing the aberrant states. There are several approaches to analyze DNAme, which vary widely in cost, resolution and coverage. Affinity capture methods for DNAme (e.g. sequencing of methyl-binding domain captured regions, or methyl binding domain (MBD)-seq) strike a good balance between the high cost of whole genome bisulphite sequencing (WGBS) and the low coverage of methylation arrays. However, existing statistical methods to analyze these data are unable to differentiate between hypomethylation patterns and low capture efficiency, do not offer flexibility to correct for copy number variation (CNV), do not produce practical precision estimates and can suffer from long running times.

**Results:** We propose an empirical Bayes framework that uses a fully methylated (i.e. SssI treated) control sample to transform observed read densities into regional methylation estimates. In our model, inefficient

capture can readily be distinguished from low methylation levels, by means of larger posterior variances. Furthermore, we can integrate CNV by introducing an multiplicative offset into our Poisson model framework. Notably, our model offers analytic expressions for the mean and variance of the methylation level and thus is fast to compute. Our algorithm performs better in terms of bias, mean-squared error and coverage probabilities compared to existing approaches when applied to an human lung fibroblast (IMR-90) MBD-seq test dataset, where "true" methylation levels are available from WGBS. Directly integrating CNV improves estimation performance in a prostate cancer cell MBD-seq dataset.

**Conclusions:** Our model not only improves on existing methods, but flexibly allows explicit modeling of CNV, allows context-specific prior information and affords a computationally-efficient analytic estimator. Our method can be applied to methylated DNA affinity enrichment assays (e.g MBD-seq, MeDIP-seq) and a software implementation will be freely available in the Bioconductor `Repitools` package.

---

## Background

DNA methylation (DNAme) is [illegible]al component in the regulation of gene expression, is precisely controlled in development and is known to be aberrant in many diseases, such as cancer and diabetes [1]. In differentiated cells, DNAme occurs primarily in the CpG dinucleotide context. For CpG-island-associated promoters, increases in DNAme (i.e. hypermethylation) induce repression of transcription, while hypomethylated promoters are generally transcriptionally active. In cancer, tumor suppressor gene promoters are frequently hypermethylated, and therefore silenced, while hypomethylation can activate oncogenes, which collectively can drive disease progression [2, 3]. The detection and profiling of such abnormalities across cell types and patient cohorts is of great medical relevance, to both our basic understanding and translation to the clinic. Epigenetic patterns can be used as diagnostic markers, predictors of response to chemotherapy and for understanding mechanisms of disease progression [4–7]. Acquired epigenetic changes are potentially reversible, which provides important therapeutic opportunities; in fact, the US Food and Drug Administration has approved at least four epigenetic drugs while others are in late-stage clinical trials [6].

There are four classes of methods to profile DNAme genome-wide: chemical conversion, endonuclease

digestion, direct sequencing and affinity enrichment; combinations of techniques are also in use (e.g. reduced representation bisulphite sequencing (RRBS) [8]). For recent reviews of the available platforms, see [9–11]. Treatment of DNA with sodium bisulphite (BS) is the gold standard, giving a single-base readout that preserves methylated cytosines while unmethylated cytosines are converted to uracil. This approach can be coupled with high-throughput sequencing, e.g. whole genome bisulphite sequencing (WGBS), or a "genotyping" microarray (e.g. Illumina Human Methylation 450k array [12]). Because WGBS is genome-wide, it inefficiently reveals methylation status for low CpG-density regions [13] and is cost-limiting for larger cohorts; however, a tradeoff can be made between coverage and replication [14]. Meanwhile, Illumina arrays cover less than 2% of genomic CpG sites and enzymatic digestion approaches are limited by the location of specific sequences. Of course, there is considerable excitement surrounding third generation sequencing technologies that directly infer methylation status, but these are not yet readily available and generally offer lower throughput [15, 16].

An attractive alternative that seems to give a good tradeoff between cost and coverage, albeit at lower resolution, is affinity capture of methylated DNA, in the combination with high-throughput sequencing (e.g. MeDIP-seq [4, 17]). Using an antibody to 5-methylcytosine or a column affixed with methyl binding domain (MBD) proteins, subpopulations of methylated DNA can be captured (see Laird et al. [9]). Libraries of these fragments can be prepared, sequenced and mapped to a reference genome; with appropriate normalization, the density of mapped reads can be transformed to a quantitative readout of the regional methylation level. However, the capability of these procedures to interrogate a given genomic region is largely related to CpG-density, which influences the efficiency of capture and can differ from protocol to protocol [13]. Thus, statistical approaches are needed.

Several methods have been proposed to estimate DNAme from affinity-based DNAme data. MBD-isolated Genome Sequencing, a variant of MBD-seq, assumed a constant rate of reads genome-wide and used a single read density threshold to binarize as methylated or not [18]. State-of-the-art methods, such as Batman [17] and MEDIPS [19], build a linear model relating read density and CpG-density, which is then used to normalize the observed read densities. For MeDIP-seq data, both algorithms showed similar estimation performance [19], while MEDIPS is considerably more time-efficient. Recently, a tool called BALM used deep sequencing of MBD-captured populations and a bi-asymmetric-Laplace model. All methods, however, suffer from the same limitations: i) low capture efficiency cannot easily be distinguished from low methylation level; ii) other factors that directly affect read density, such as copy number variation (CNV), are not easily taken into account. For CNV correction, a few possibilities have emerged:

i) omit known regions of amplifications [4]; ii) adjust read densities manually by dividing by estimated copy number before running Batman [20]; or, iii) adjust using read density from an input sample [21].
We present a novel empirical Bayes model called BayMeth, based on the Poisson distribution, that explicitly models read densities of a fully methylated (e.g. SssI-treated DNA) control sample together those from a sample of interest. The model provides an understanding of where the assay can detect DNAme and allows direct integration of CNV and potentially other factors that affect read density. Notably, we derive an analytic expression for the mean methylation level and also for the variance. Credible intervals can be computed using numerical integration of the analytical posterior marginal distributions. Using MBD-seq on human lung fibroblast (IMR-90) DNA, where "true" methylation levels are available from WGBS, we show favorable performance against existing approaches in terms of bias, mean-squared error, Spearman correlation and coverage probabilities. In a second application to MBD-seq data on human prostate carcinoma (LNCaP) cells, we show that directly integrating CNV data provides additional performance gains.

## Results

### Statistical Model: BayMeth

DNAme data is obtained by MBD-seq or a similar assay. Let $y_{i1}$ and $y_{i2}$ denote the observed number of (uniquely) mapped reads for genomic regions $i = 1, \ldots, n$ for the sample of interest and the SssI control, respectively. Throughout this paper we use non-overlapping regions of 100bp, whereby only regions with at least 75% mappability and a CpG-density larger than zero are considered (see Methods). Let

$$y_{i1}|\mu_i, \lambda_i \sim \text{Poisson}\left(f \times \frac{\text{cn}_i}{\text{ccn}} \times \mu_i \times \lambda_i\right) \quad \text{, and} \quad y_{i2}|\lambda_i \sim \text{Poisson}(\lambda_i), \quad \text{with } \lambda_i > 0,\, 0 < \mu_i < 1. \tag{1}$$

Here, $\lambda_i$ denotes the region-specific read density at full methylation, $\mu_i$ the regional methylation level and $f > 0$ represents the (effective) relative sequencing depth between libraries (i.e. a normalization offset). Recently, an approximately linear relationship between the copy number state and the read density was shown [22]. In these situations, we can include an additional multiplicative offset $\frac{\text{cn}_i}{\text{ccn}}$ into our model formulation, where $\text{cn}_i$ denotes the copy number state at region $i$ and ccn is cell's most prominent CNV state (e.g. two in normal cells).

#### *Closed-form posterior methylation quantities*

In a Bayesian framework, prior distributions are assigned to all parameters. For the methylation level ($\mu_i$), we assume a uniform prior (i.e. a beta distribution with both parameters set to 1). Alternative prior

specifications, such as a mixture of beta distributions, are possible (see Discussion). For the region-specific density, we assume a gamma distribution, i.e. $\lambda_i \sim \text{Ga}(\alpha, \beta)$ using shape $\alpha > 0$ and rate $\beta > 0$ hyperparameters, which are determined in a CpG-dependent manner (see next Section). To make inferences for the regional methylation levels, $\mu_i$, we integrate out $\lambda_i$ from the joint posterior distribution:

$$\begin{aligned}\mathsf{p}(\mu_i|y_{i1}, y_{i2}) &= \int_0^\infty \mathsf{p}(\lambda_i, \mu_i|y_{i1}, y_{i2})d\lambda_i \\ &= \int_0^\infty \frac{\mathsf{p}(y_{i1}|\lambda_i, \mu_i)\,\mathsf{p}(y_{i2}|\lambda_i)\,\mathsf{p}(\lambda_i)\,\mathsf{p}(\mu_i)}{\mathsf{p}(y_{i1}, y_{i2})}d\lambda_i.\end{aligned}$$

Notably, $\mathsf{p}(y_{i1}, y_{i2})$ can be calculated analytically [23], so that the marginal posterior distribution

$$\mathsf{p}(\mu_i|y_{i1}, y_{i2}) = \frac{\mu_i^{y_{i1}}}{W}\left(1 - \frac{E(1-\mu_i)}{\beta + 1 + E}\right)^{-(\alpha + y_{i1} + y_{i2})}, \tag{2}$$

is given in closed form with $E = f \cdot \frac{\text{cn}_i}{\text{ccn}}$ and

$$W = \frac{1}{y_{i1}+1} \times {}_2F_1\left(y_{i1} + y_{i2} + \alpha, 1; y_{i1} + 2; \frac{E}{\beta + 1 + E}\right).$$

where ${}_2F_1(\,)$ is the Gauss hypergeometric function [24, page 558]. The posterior mean and the variance are analytically available (see Methods) and therefore straightforward to efficiently compute; credible intervals can be computed numerically from Equation (2). Estimates for bins with a CpG-density of zero are removed.

*Empirical Bayes for prior hyperparameter specification*

Our method takes advantage of the relationship between CpG-density and read depth to formulate a CpG-density-dependent prior distribution for $\lambda_i$. Taking CpG-density into account the prior should stabilize the methylation estimation procedure for low counts and in the presence of sampling variability (i.e. sampling a large population of DNA fragments). The hyperparameters $\alpha$ and $\beta$ of the gamma prior distribution are determined in a CpG-density-dependent manner using empirical Bayes. For each 100bp bin, we determined the weighted number of CpG dinucleotides within a window of 700bp (see Methods). Each region is classified based on its CpG-density into one of $K = 100$ non-overlapping CpG-density intervals (See x-axis tick marks in Supplementary Figure 1). Due to the small number of regions with extreme CpG-densities, the last interval width is larger.

For each class separately, we derive the values for $\alpha$ and $\beta$ under an empirical Bayes framework using maximum likelihood. Note that both read depths, from the SssI control and the sample of interest, are thereby taken into account, since $\lambda_i$ is a joint parameter affecting both. We end up with $K$ parameter sets

for shape $\alpha$ and rate $\beta$. Considering only the Poisson model for the SssI control above, we can derive the prior predictive distribution by integrating $\lambda_i$ out, resulting in a negative binomial distribution, as illustrated in Figure 1. Here, we show the SssI read densities by CpG-density, together with the predictive distribution evaluated at the center of the CpG-density classes.

### BayMeth improves estimation and provides realistic variability estimates

To take advantage of Lister et al. [25] single-base-resolution high-coverage methylome obtained by WGBS, we generated MBD-seq data using IMR-90 cells (see Methods), with affinity capture experiments run under the same conditions that we collected previously [2]. We applied our proposed model to IMR-90 and SssI MBD-seq data. The normalizing offset $f = 0.596$ is found based on calculating a scaling factor between highly methylated regions in IMR-90 relative to the SssI control (see Methods and Supplementary Figure 2). The prior parameters for the gamma distributions are determined by empirical Bayes for bins with at least 75% mappability, as discussed above (see Methods). We compared the results to those obtained by Batman [17] and MEDIPS [19]; BALM was tried (see Supplementary Figure XX) but not considered further due to poor performance. To provide plausible uncertainty estimates of the Bayesian approach Batman, we increased the default number of generated samples from 100 to 500; we only considered chromosome 7 since Batman is very computationally demanding. The WGBS data, here considered to be the "truth", are collapsed into 100bp bin estimates (see Methods) to match the estimates from MEDIPS, Batman and our approach. In total, chromosome 7 is comprised of 1588214 100bp bins: 814358 are excluded for lack of WGBS data, further 20510 bins are excluded due to no Batman estimates and then 194303 are excluded for low mappability; in total, algorithm comparisons are conducted on the remanding 559043 bins.

in process

The behavior of BayMeth and Batman is illustrated using an example region of chromosome 7 (see Figure 2A). WGBS levels, CpG-density and read counts per 100bp region of MBD-seq SssI and IMR-90 sample are shown. As expected, the number of reads in the SssI control is related to the CpG-density, whereas the read density in (MBD-seq) IMR-90 is modulated by both the region-specific density and the DNAme level. Regions lacking both IMR-90 and SssI reads suggest inefficient MBD-based affinity capture (e.g. region 'a'). Figure 2B shows posterior samples from Batman and inferred posterior distributions from BayMeth. For region 'a', Batman's posterior samples are concentrated between 0.7 and 1 (mean equal to 0.85). In contrast, BayMeth returns a mean methylation level of 0.49 together with a large 95% highest posterior density (HPD) interval $(0, 0.94)$, reflecting the uncertainty from having no SssI reads sampled.

For regions with no IMR-90 reads but efficient capture (e.g. region 'b'), both BayMeth and Batman provide sensible posterior marginal distributions and low DNAme estimates. If there are a small number of reads for IMR-90 with efficient capture (e.g. region 'c') the BayMeth posterior marginal is more disperse than Batman's, while both are close to zero. Region 'd' has a high number of reads for both samples and a true methylation level around 0.95. This level is covered by the 95% HPD region of BayMeth, while it lies outside of the density mass obtained by Batman overestimating this region.

Table 1 summarizes the estimation performance for chromosome 7 by means of mean bias (mean of differences between the posterior mean $\hat{\mu}_i$ and the true value $\mu_i$), MSE (mean of squared differences), Spearman correlation and mean Dawid-Sebastiani score (DSS):

$$\text{DSS} = \frac{1}{I}\sum_{i=1}^{I}\left(\frac{(\hat{\mu}_i - \mu_i)^2}{\sigma_i^2} + 2\log(\sigma_i)\right).$$

where $\hat{\mu}_i$ and $\sigma_i$ are the posterior mean and standard deviation, respectively, and $\mu_i$ is the WGBS methylation level. The DSS is a scoring rule that assesses both calibration and sharpness [27]. To account for uncertainty present in the WGBS estimates, we applied a threshold on the depth; we assess the performance using bins with at least 33 WGBS reads (unmethylated and methylated) corresponding to the 25% quantile of depth in the truth, which results in 414352 bins. Results are stratified into five groups according to depth in the SssI control, which should represent a surrogate of the capture efficiency. The first group $[0, 4]$ encompasses primarily low-CpG regions that are not well captured in MBD experiments, while the high $(27, 168]$ group represents primarily CpG island regions. On average, Batman tends to overestimate DNAme while MEDIPS tends to strongly underestimate. BayMeth, in contrast, is almost unbiased. The smaller bias in the point estimates obtained by BayMeth is also reflected in the MSE. For all methods, the MSE decreases with higher SssI depth, as expected due to the efficiency of capture. For all depth groups, BayMeth has the highest correlation with the WGBS estimates, which increases with higher SssI depth. Methylation estimates of the highest SssI depth group, namely $(27, 168]$, plotted for all methods against the "true" WGBS methylation levels are shown in Figure 3. Although the estimates of all methods are off the diagonal, BayMeth provides the most accurate point estimates. The overestimation of Batman and underestimation of MEDIPS is obvious, while the BayMeth errors vary almost symmetrically. Since MEDIPS does not provide uncertainty measures, no DSS scores are given in Table 1. DSS estimates are much higher for Batman than for BayMeth, which is a reflection of underestimated standard deviations by Batman (see Figure 2 B). In contrast, BayMeth tends to provide appropriate posterior standard deviations, which can result in negative DSS values.

To assess calibration, we computed coverage probabilities (frequency that the true value is captured within a credible interval). Stratified by the "true" WGBS methylation level, Figure 4 shows coverage probabilities at 80% and 95% level for regions deemed to be inside or outside a CpG island (Supplementary Figure 1). HPD intervals and quantile-based credible intervals (CI) are computed for BayMeth while only quantile-based CIs are available for Batman; coverage probabilities are not possible from MEDIPS output. As mentioned above, Batman has a tendency to underestimate the variance and results in lower coverage probabilities of the WGBS values; in contrast, BayMeth's coverage probabilities are much closer to the nominal levels and seem to be stable across the stratification.

### CNV-aware BayMeth improves DNAme estimation for prostate cancer cells

In the following, we illustrate the benefits of directly integrating CNV information into a cancer MBD-seq dataset. We apply our methodology to the autosomes of the LNCaP cell line. To motivate such an adjustment, Figure 5 shows the estimated copy number across chromosome 13 (using the PICNIC algorithm on Affymetrix genotyping arrays; see Methods), together with tiled MBD-seq read counts. Although read densities at a specific genomic region (again, 100bp non-overlapping bins) are influenced by a combination of effects (e.g. DNAme, CpG-density), a relationship between CNV and number of reads is clearly visible. In particular, a difference in read counts between regions with four copies and those with smaller copy numbers is apparent. We adjust for this bias through a multiplicative offset $\frac{\text{cn}_i}{\text{ccn}}$, where the prominent state is four copies (i.e. ccn = 4 in Equation (1)), as shown in Supplementary Figure 3. In addition, regions from this state ($\text{cn}_i = 4$) are used to determine the normalizing offset $f$ (here, estimated to be 0.712). The read depth stratified by copy number state together with mean and median estimates is shown in Supplementary Figure 4. In particular, for the three most frequent CNV states (2–4), read densities scale approximately linearly (with a slope of 1) with CNV, which justifies the structure of our multiplicative offset; copy-number offsets are shown in Table 2. Figure 6 shows the bias of DNAme point estimates of the different methods by integer CNV state (2–5); here, we used the Illumina HumanMethylation 450k array as the "true" methylation (see Methods), since methylation status should be unaffected by CNV [28]. Because CNV only affects MBD capture for methylated regions, we restrict this comparison to where the true methylation state is larger than 0.5 and we applied a threshold of 13 (median after excluding bins with a low depth of $[0, 4]$) to the number of reads in the SssI-control to select for regions where MBD-seq has good performance. Similar to the IMR-90 data, MEDIPS tends to underestimate, while Batman tends to overestimate. Using the standard normalization offset, BayMeth

provides biased estimates, predictably by CNV state. After including the copy-number-specific offset, these biases almost disappear. A scatterplot illustrating the benefits of including the copy-number-specific offset is shown in Figure 7 for copy number state two. In particular, bins that have been falsely underestimated (due to two copies instead of four) are corrected. Table 3 shows mean bias, MSE and Spearman correlation for the different approaches stratified by copy number state. In all measures, the adjusted version of BayMeth performs best. While the differences in the correlation estimates are small, clear advantages can be seen in terms of bias and MSE. In contrast to the other approaches, the performance estimates stay almost constant over the different copy number states and are close to zero.

## Discussions and conclusions

DNA methylation plays a crucial role in various biological processes and is known to be aberrant in several human diseases, such as cancer. There are now a multitude of methylation profiling platforms, each with inherent advantages and disadvantages. Bisulfite-based approaches are considered the gold standard since they allow quantification at single-base resolution. However, applied genome-wide, this technique can be inefficient and expensive, in terms of CpGs covered per read or base sequenced [13]. On the other hand, affinity capture based approaches, such as MBD or MeDIP, combined with sequencing seem to provide a good compromise between cost and coverage, albeit at lower resolution; thus, we consider them to be an attractive alternative. Recently, affinity capture has been demonstrated using only hundreds of nanogramms of starting DNA, thus making these approaches applicable to a wider range of studies [29]. The key to our proposed method is the use of methylated DNA captured from a fully methylated SssI control; for future studies, we recommend such a sample should be collected under the same conditions used for the samples of interest. We used commercially available SssI-treated DNA [2] for the MBD-seq experiments and verified with the 450k platform that the overwhelming majority of CpG sites are indeed methylated (see Supplementary Figure 5); similarly, such a sample can be constructed directly and inexpensively [30]. Our proposed method, BayMeth, is a flexible empirical Bayes approach that transforms read densities into regional methylation estimates. Our model is based on a Poisson distribution and takes advantage of SssI control data in two ways: i) we model SssI data jointly with data from a sample of interest to preserve the linearity of the methylation estimation; ii) we explicitly get information about the region-specific read density as a function of CpG-density. Our method is similar in principle to MEDME, which was applied to fully methylated MeDIP microarray intensities [31]. However, our approach necessarily modifies assumptions for count data (i.e. read densities versus probe intensities) and is

effectively a *moderation* between the global fit that MEDME implements and a region-specific correction. We showed that BayMeth delivers improved performance against state-of-the-art techniques using IMR90 MBD-seq data, using two datasets where "true" methylation levels are available from WGBS or bisulphite-based methylation arrays. In general, MEDIPS grossly underestimates the methylation levels and does not offer variability estimates. Batman performs reasonably well, but our analyses suggest that variability estimates are generally underestimated. Our model performs best in point estimation and affords reasonably interval estimates. Notably, BayMeth offers analytic expressions for the posterior marginal distribution and the posterior mean and variance, avoiding computationally-expensive sampling algorithms (e.g. Batman). Furthermore, we can explicitly integrate existing CNV data, which offers improvement when applied to cancer datasets. CNV adjustments may be possible with existing approaches Batman or MEDIPS, based on ad-hoc transformations of the read counts (e.g. see [20]), but are not included within the model formulation. In contrast, our model preserves the count nature of the data.

A conceptual similar Bayesian hierarchical model, which involves MCMC sampling, has been proposed in the context of Methyl-Seq experiments, where methylation levels are derived based on enzymatic digestion using two enzymes [32]; a separate Poisson model is assumed for the tag counts of each enzyme. The models are linked through a shared parameter while one Poisson model contains a methylation level parameter $\mu$, assumed to be uniformly distributed *a priori*. In the two applications presented here, we also used a uniform prior for the methylation level. Of note, the analytical expressions for the mean, variance and posterior marginal distribution are still available using a prior based on a mixture of beta distributions (see Methods). Thus context-specific information, for example CpG-density or the position relative to transcriptional features, can be incorporated in the prior distribution for the methylation level. We have tried various weighted mixtures of two or three beta distributions that build in contextual information; however, these did not outperform the uniform prior.

To adjust the modeled mean for effects arising due to different library compositions in the SssI control and the sample of interest, we estimated a normalization offset. Furthermore, adjustments for CNV are included by a second multiplicative offset. In fact, this approach of using offsets to adjust the expected read density is quite general and could be extended beyond composition and CNV (e.g. see [22, 33]).

It is well known that methylation levels are dependent within neighboring regions. Thus, a potential improvement may involve modeling correlation between neighboring genomic bins. One approach might be Gaussian Markov random fields [34]; however, the analytical summaries are lost, so the the gain in performance may not justify the more complex model and associated computation cost.

# 1 Methods

## 1.1 MBD-seq on IMR-90, LNCaP and SssI DNA

We used LNCaP and SssI MBD-seq data and Affymetrix genotyping array data (LNCaP only) from Robinson et al. [26] and can be found at http://www.ncbi.nlm.nih.gov/geo under accession number GSE24546. Similarly, IMR-90 MBD-seq is available from GSE38679. Details of the DNA capture, preparation and sequencing can be found in Robinson et al. [2

### Calculation of CpG-density

CpG-density is defined to be a weighted count of CpG sites in a predefined region. We used the function `cpgDensityCalc` provided by the R-package Repitools [35] to get 100bp bin-specific CpG-density estimates using a linear weighting function and a window size of 700bp (since we expect fragments around 300bp).

### Calculation of mappability

Using Bowtie, all possible 36bp reads of the genome were mapped back against the hg18 reference, with no mismatches. At each base, a read can either unambiguously map or not. A mappability estimate gives the proportion of reads that can be mapped to a specific regions. To get bin-specific mappability estimates we used the function `mappabilityCalc` in the Repitools package [35]. In our analysis, a window of 500bp was used (250bp upstream and downstream from the center from each 100bp bin) and the percentage of mappable bases was computed.

### Derivation of region-specific methylation estimates from WGBS

In the Lister et al. IMR-90 WGBS data, the number of reads $r_j^+$ and $r_j^-$ overlaying a cytosine $j$ in the positive $(+)$ and negative strand $(-)$, respectively, is available. Furthermore, the number of these reads, $m_j^+$ and $m_j^-$, that contain a methylated cytosine, is known. A single-base methylation estimate can be obtained by $(m_j^+ + m_j^-)/(r_j^+ + r_j^+)$. To get a bin-specific methylation estimate all cytosines lying within a bin of interest $\mathcal{B}$ are taken into account:

$$\mu_{\mathcal{B}} = \frac{\sum_{j\in\mathcal{B}}(m_j^+ + m_j^-)}{\sum_{j\in\mathcal{B}}(r_j^+ + r_j^-)}.$$

Here, $\sum_{j\in\mathcal{B}}(r_j^+ + r_j^-)$ is termed depth.

### Derivation of region-specific methylation estimates from 450K arrays

First, the Illumina HumanMethylation450 methylation array was preprocessed using default parameter of the `minfi` package [36]; for each sample, a vector of *beta values*, one for each targeted CpG site representing methylation estimates are produced. To obtain (100bp) bin-specific methylation profiles, we averaged beta values from all CpG sites within 100bp (upstream and downstream; total window of 200bp) from the center of our 100bp bins.

### Determining the normalizing offset

The composition of a library influences the resulting read densities [37]. For example, the SssI control represents a more diverse set of DNA fragments since it captures the vast majority of CpG rich regions in the genome. Therefore, if the total sequencing depth were to be fixed, one would expect a relative undersampling of regions in SssI, compared to a sample of interest that is presumably largely unmethylated. To adjust the modeled mean (in the Poisson model) for these composition effects, we estimate a normalizing factor $f$ that accounts simultaneously for overall sequencing depth and composition. Supplementary Figure 2 shows an $M$ (log-ratio) versus $A$ (average-log-count) plot at 1,000,000 randomly chosen (100bp) bins for IMR-90 compared to the fully methylated control. A clear offset from zero is visible, where the distribution of $M$ values is skewed in the negative direction. The normalization offset $f$ is estimated as $f = 2^{median(M_{A>q})}$, with $q$ corresponding to a high (here, 0.998; more than 35000 points in both applications) quantile of $A$. In cancer samples where CNV are common, the normalization factor $f$ is calculated from bins that originate from the most prominent copy number state (e.g., ccn = 4 in LNCaP cells).

### Estimation of copy number

Copy number estimates were estimated from Affymetrix SNP6.0 genotyping array data by PICNIC [38], using default parameters. PICNIC is an algorithm based on a hidden Markov model to produce absolute allelic copy number segmentation.

**Empirical Bayes for prior specification**

For ease of readability let $E = f \times \frac{\mathrm{cn}_i}{\mathrm{ccn}}$. The joint marginal distribution of $y_{i1}, y_{i2}$ results as:

$$\begin{aligned}
p(y_{i1}, y_{i2}) &= \int\int p(y_{i1}|\mu_i, \lambda_i) p(y_{i_2}|\lambda_i) p(\lambda_i) p(\mu_i) d\lambda_i d\mu_i \\
&= \int_0^1 p(\mu_i) \left[ \int_0^\infty p(y_{i1}|\mu_i, \lambda_i) p(y_{i2}|\lambda_i) p(\lambda_i) d\lambda_i \right] d\mu_i \\
&= \int_0^1 p(\mu_i) \left[ \int_0^\infty \frac{(E\mu_i\lambda_i)^{y_{i1}} \lambda_i^{y_{i2}}}{y_{i1}! y_{i2}!} \exp\left(-E\mu_i\lambda_i\right) \times \exp(-\lambda_i) \times \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} \exp(-\beta\lambda_i) d\lambda_i \right] d\mu_i \\
&= \int_0^1 p(\mu_i) \left[ \frac{(E\mu_i)^{y_{i1}}}{y_{i1}! y_{i2}!} \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda_i^{y_{i1}+y_{i2}+\alpha-1} \exp\left(-(E\mu_i + 1 + \beta)\lambda_i\right) d\lambda_i \right] d\mu_i \\
&= \int_0^1 p(\mu_i) \left[ \frac{(E\mu_i)^{y_{i1}}}{y_{i1}! y_{i2}!} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(y_{i1} + y_{i2} + \alpha)}{(E\mu_i + 1 + \beta)^{y_{i1}+y_{i2}+\alpha}} \right] d\mu_i \\
&= \frac{E^{y_{i1}}}{y_{i1}! y_{i2}!} \frac{\beta^\alpha}{\Gamma(\alpha)} \Gamma(y_{i1} + y_{i2} + \alpha) \int_0^1 p(\mu_i) \frac{\mu_i^{y_{i1}}}{(E\mu_i + 1 + \beta)^{y_{i1}+y_{i2}+\alpha}} d\mu_i
\end{aligned}$$

Let $\mu_i \sim \mathrm{Be}(a, b)$, i.e. $p(\mu_i) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu_i^{a-1}(1-\mu_i)^{b-1}$, $a, b > 0$. (For a uniform distribution $a = b = 1$).

Then

$$\begin{aligned}
p(y_{i1}, y_{i2}) &= \frac{\Gamma(y_{i1} + y_{i2} + \alpha)}{\Gamma(\alpha) y_{i1}! y_{i2}!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} E^{y_{i1}} \beta^\alpha \int_0^1 \frac{\mu_i^{y_{i1}+a-1}(1-\mu_i)^{b-1}}{(E\mu_i + 1 + \beta)^{y_{i1}+y_{i2}+\alpha}} d\mu_i \\
&= \dots \\
&\overset{\star}{=} \frac{\Gamma(y_{i1} + y_{i2} + \alpha)}{\Gamma(\alpha) y_{i1}! y_{i2}!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} E^{y_{i1}} \frac{\beta^\alpha}{(\beta + 1 + E)^{y_{i1}+y_{i2}+\alpha}} \int_0^1 \frac{(1-t_i)^{y_{i1}+a-1} t_i^{b-1}}{\left(1 - \frac{E}{E+1+\beta} \cdot t_i\right)^{y_{i1}+y_{i2}+\alpha}} dt_i \qquad (3) \\
&= \frac{\Gamma(y_{i1} + y_{i2} + \alpha)}{\Gamma(\alpha) y_{i1}! y_{i2}!} \left(\frac{\beta}{\beta + 1 + E}\right)^\alpha \left(\frac{E}{\beta + 1 + E}\right)^{y_{i1}} \left(\frac{1}{\beta + 1 + E}\right)^{y_{i2}} \frac{\Gamma(a+b)\Gamma(y_{i1} + a)}{\Gamma(a)\Gamma(y_{i1} + a + b)} \times \\
&\qquad {}_2F_1\left(y_{i1} + y_{i2} + \alpha, b; y_{i1} + a + b; \frac{E}{\beta + 1 + E}\right).
\end{aligned}$$

In the step marked with $\star$ we substituted $(1 - \mu_i)$ with $t_i$, where $dt_i = -d\mu_i$, to get the desired form of the Gauss hypergeometric function (the limits of the integral stay thereby unchanged), which is defined by:

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-zt)^{-a} dt, \quad c > b > 0$$

where $|z| < 1$ is the radius of convergence [24, see page 558]. (Note, $|z| = |E/(\beta + 1 + E)| < 1$ in (3), so that convergence is granted). Model (3) is similar to the beta binomial (BB)/negative binomial (NB) model derived in [39] and [23].

Using a mixture of $M$ beta distributions as prior distribution for $\mu_i$, i.e. $\mu_i \sim \sum_{m=1}^M w_m \,\mathrm{Be}(a_m, b_m)$, where $0 \le w_m \le 1$, for all $m = 1, \dots, M$, and $\sum_{m=1}^M w_m = 1$ we get:

$$p(y_{i1}, y_{i2}) = \frac{\Gamma(y_{i1} + y_{i2} + \alpha)}{\Gamma(\alpha) y_{i1}! y_{i2}!} \left(\frac{\beta}{\beta + 1 + E}\right)^\alpha \left(\frac{E}{\beta + 1 + E}\right)^{y_{i1}} \left(\frac{1}{\beta + 1 + E}\right)^{y_{i2}} \times W \qquad (4)$$

with

$$W = \sum_{m=1}^{M} \left[ w_m \cdot \frac{\Gamma(a_m + b_m)\Gamma(y_{i1} + a_m)}{\Gamma(a_m)\Gamma(y_{i1} + a_m + b_m)} \times \, {}_2F_1 \left( y_{i1} + y_{i2} + \alpha, b_m; y_{i1} + a_m + b_m; \frac{E}{\beta + 1 + E} \right) \right].$$

Under the empirical Bayes approach, the parameters of (4) can be estimated using maximum likelihood. The parameters are thereby determined in a CpG-density-dependent manner. Each 100bp bin is classified based on its CpG-density into one of $K = 100$ non-overlapping CpG-density classes: $\mathcal{C}_1, \ldots, \mathcal{C}_K$. The class size $|\mathcal{C}_k|$, i.e. the number of 100bp bins in class $k$, is denoted by $n_k$. We derive for each class separately the set of prior parameters using empirical Bayes leading finally to $K$ parameter sets. The corresponding log likelihood function for class $k$ is given by

$$l(\mathbf{w}^{(k)}, \mathbf{a}^{(k)}, \mathbf{b}^{(k)}, \alpha^{(k)}, \beta^{(k)} | \mathbf{y}_1^{(k)}, \mathbf{y}_2^{(k)}) = \sum_{j=1}^{n_k} \log(p(y_{j1}^{(k)}, y_{j2}^{(k)} | \mathbf{w}^{(k)}, \mathbf{a}^{(k)}, \mathbf{b}^{(k)}, \alpha^{(k)}, \beta^{(k)}). \tag{5}$$

Here $\mathbf{y}_1^{(k)} = (y_{11}^{(k)}, \ldots, y_{n_k 1}^{(k)})$ and $\mathbf{y}_2^{(k)} = (y_{12}^{(k)}, \ldots, y_{n_k 2}^{(k)})$ denote the read counts of the bins contained in class $\mathcal{C}_k$. Further $\mathbf{w}^{(k)} = (w_1^{(k)}, \ldots, w_M^{(k)})$, $\mathbf{a}^{(k)} = (a_1^{(k)}, \ldots, a_M^{(k)})$, $\mathbf{b}^{(k)} = (b_1^{(k)}, \ldots, b_M^{(k)})$, $\alpha^{(k)}$, $\beta^{(k)}$ denote the parameters for CpG density class $k$. (Using a uniform prior for the methylation level $\mu_i$ only parameters $\alpha^{(k)}$ and $\beta^{(k)}$ are left). In Equation (5), we assume that genomic regions are independent. For a discussion of this assumption, see the Discussion Section.

**Derivation of the posterior marginal distribution for the methylation level.**

Our main interest lies in the marginal posterior distribution of the methylation level $\mu_i$

$$p(\mu_i | y_{i1}, y_{i2}) = \int_0^\infty p(\lambda_i, \mu_i | y_{i1}, y_{i2}) d\lambda_i,$$

where

$$\begin{aligned} p(\lambda_i, \mu_i | y_{i1}, y_{i2}) &= \frac{p(y_{i1}, y_{i2} | \lambda_i, \mu_i) p(\lambda_i, \mu_i)}{p(y_{i1}, y_{i2})} \\ &\overset{\text{cond.indep}}{=} \frac{p(y_{i1} | \lambda_i, \mu_i) p(y_{i2} | \lambda_i) p(\lambda_i) p(\mu_i)}{p(y_{i1}, y_{i2})} \\ &= \frac{\lambda_i^{y_{i1} + y_{i2} + \alpha - 1} \exp(-(E\mu_i + 1 + \beta)\lambda_i)(\beta + 1 + E)^{\alpha + y_{i1} + y_{i2}} p(\mu_i) \mu_i^{y_{i1}}}{\Gamma(y_{i1} + y_{i2} + \alpha) \times W} \end{aligned}$$

Thus:

$$p(\mu_i|y_{i1},y_{i2}) = \frac{\mu_i^{y_{i1}} p(\mu_i)(\beta+1+E)^{\alpha+y_{i1}+y_{i2}}}{\Gamma(y_{i1}+y_{i2}+\alpha)\times W}\int_0^\infty \lambda_i^{y_{i1}+y_{i2}+\alpha-1}\exp(-(E\mu_i+1+\beta)\lambda_i)d\lambda_i$$
$$= \frac{\mu_i^{y_{i1}} p(\mu_i)}{W}\left(1-\frac{E(1-\mu_i)}{\beta+1+E}\right)^{-(\alpha+y_{i1}+y_{i2})}$$

The mean of the marginal posterior of $\mu_i$ is given by:

$$\mathsf{E}(\mu_i|y_{i1},y_{i2}) = \frac{A}{W}$$

with

$$A = \sum_{m=1}^{M}\left[w_m\cdot\frac{\Gamma(a_m+b_m)\Gamma(y_{i1}+a_m{+}1)}{\Gamma(a_n)\Gamma(y_{i1}+a_m+b_m{+}1)}\times\ {}_2F_1\left(y_{i1}+y_{i2}+\alpha,b_m;y_{i1}+a_m+b_m{+}1;\frac{E}{\beta+1+E}\right)\right].$$

*Proof.*

$$\mathsf{E}(\mu_i|y_{i1},y_{i2}) = \int_0^1 \mu_i p(\mu_i|y_{i1},y_{i2})d\mu_i$$
$$= \frac{1}{W}\sum_{m=1}^{M}\left[\int_0^1 \frac{w_m\frac{\Gamma(a_m+b_m)}{\Gamma(a_m)\Gamma(b_m)}\mu_i^{a_m+y_{i1}}(1-\mu_i)^{b_m-1}}{\left(1-\frac{E(1-\mu_i)}{\beta+1+E}\right)^{\alpha+y_{i1}+y_{i2}}}d\mu_i\right],$$

where each integral can again be written in terms of the Gauss hypergeometric function:

$$\int_0^1 \frac{w_m\frac{\Gamma(a_m+b_m)}{\Gamma(a_m)\Gamma(b_m)}\mu_i^{a_m+y_{i1}}(1-\mu_i)^{b_m-1}}{\left(1-\frac{E(1-\mu_i)}{\beta+1+E}\right)^{\alpha+y_{i1}+y_{i2}}}d\mu_i$$
$$= \frac{w_m\Gamma(a_m+b_m)}{\Gamma(a_m)\Gamma(b_m)}\int_0^1 \frac{(1-t_i)^{a_m+y_{i1}}t_i^{b_m-1}}{\left(1-\frac{E}{\beta+1+E}t_i\right)^{\alpha+y_{i1}+y_{i2}}}dt_i$$
$$= \frac{w_m\Gamma(a_m+b_m)}{\Gamma(a_m)\Gamma(b_m)}\frac{\Gamma(b_m)\Gamma(y_{i1}+a_m+1)}{\Gamma(y_{i1}+a_m+b_m+1)}\ {}_2F_1\left(y_{i1}+y_{i2}+\alpha,b_m;y_{i1}+a_m+b_m+1;\frac{E}{\beta+1+E}\right)$$
$$= \frac{w_m\Gamma(a_m+b_m)\Gamma(y_{i1}+a_m+1)}{\Gamma(a_m)\Gamma(y_{i1}+a_m+b_m+1)}\ {}_2F_1\left(y_{i1}+y_{i2}+\alpha,b_m;y_{i1}+a_m+b_m+1;\frac{E}{\beta+1+E}\right).$$

□

The variance of the marginal posterior distribution of $\mu_i$ can be computed using the computational formula for the variance $\mathrm{Var}(\mu_i|y_{i1},y_{i2}) = \mathsf{E}(\mu_i^2|y_{i1},y_{i2}) - (\mathsf{E}(\mu_i|y_{i1},y_{i2}))^2$, where

$$\mathsf{E}(\mu_i^2|y_{i1},y_{i2}) = \frac{B}{W}$$

with

$$B = \sum_{m=1}^{M}\left[w_m\cdot\frac{\Gamma(a_m+b_m)\Gamma(y_{i1}+a_n{+}2)}{\Gamma(a_m)\Gamma(y_{i1}+a_m+b_m{+}2)}\times\ {}_2F_1\left(y_{i1}+y_{i2}+\alpha,b_m;y_{i1}+a_m+b_m{+}2;\frac{E}{\beta+1+E}\right)\right].$$

so that

$$\mathrm{Var}(\mu_i|y_{i1},y_{i2}) = \frac{B}{W} - \left(\frac{A}{W}\right)^2$$

### Details on Batman specifications

Batman is an algorithm implemented in JAVA and run from the command prompt. The original Batman can be downloaded from http://td-blade.gurdon.cam.ac.uk/software/batman/; we used an unreleased version "20090617" from Thomas Down; the commands used to run Batman are given in the Supplementary Material.

### Details on MEDIPS specifications

We used the R-Bioconductor MEDIPS version 1.4.0 and followed the available tutorial (medips.molgen.mpg.de/MEDIPS.1.0.0/MEDIPS.pdf from October 18, 2010); the detailed command sequence is given in the Supplementary Material. MEDIPS returns methylation estimates in the range from zero to 1000, which we rescaled to the interval $[0, 1]$. In our comparison, we used the absolute methylation score (AMS) provided by MEDIPS.

## Authors contributions

The statistical approach was conceived and developed by AR and MDR, with biological and technical insight from ALS, CS, MM, SJC and GM. Implementation and data analyses were done by AR with contributions from MDR. Data was collected by JZS, ALS, NM, CM and MM. AR and MDR wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

## Acknowledgements

AR gratefully acknowledges funding of the "Forschungskredit" and the URPP (University Research Priority Program in Systems Biology/Functional Genomics) grant of the University of Zurich. We thank Elena Zotenko, Marcel Coolen and Mattia Pelizzola for useful discussions on the experimental and computational strategies.

## Acronyms

**DNAme** DNA methylation

**RRBS** reduced representation bisulphite sequencing

**CNV** copy number variation

**WGBS** whole genome bisulphite sequencing

**BS** sodium bisulphite

**MBD** methyl binding domain

**IMR-90** human lung fibroblast

**LNCaP** human prostate carcinoma

**HPD** highest posterior density

## References


1. de Bustros A, Nelkin BD, Silverman A, Ehrlich G, Poiesz B, Baylin SB: **The short arm of chromosome 11 is a "hot spot" for hypermethylation in human neoplasia**. *Proceedings of the National Academy of Sciences* 1988, **85**(15):5693–5697.
2. Clark SJ, Melki J: **DNA methylation and gene silencing in cancer: which is the guilty party?** *Oncogene* 2002, **21**(35):5380–5387.
3. Esteller M: **Cancer epigenomics: DNA methylomes and histone-modification maps**. *Nature Reviews Genetics* 2007, **8**(4):286–298.
4. Ruike Y, Imanaka Y, Sato F, Shimizu K, Tsujimoto G: **Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing**. *BMC Genomics* 2010, **11**:137.
5. Stein RA: **Epigenetics—The link between infectious diseases and cancer**. *Journal of the American Medical Association* 2011, **305**(14):1484–1485.
6. Baylin SB, Jones PA: **A decade of exploring the cancer epigenome—biological and translational implications**. *Nature Reviews Cancer* 2011, **11**:726–734.
7. Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond**. *Nature Reviews Genetics* 2012, [http://dx.doi.org/10.1038/nrg3230].
8. Gu H, Bock C, Mikkelsen TS, Jager N, Smith ZD, Tomazou E, Gnirke A, Lander ES, Meissner A: **Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution**. *Nature Methods* 2010, **7**(2):133–136.
9. Laird PW: **Principles and challenges of genome-wide DNA methylation analysis**. *Nature Reviews Genetics* 2010, **11**(3):191–203.
10. Lister R, Ecker JR: **Finding the fifth base: genome-wide sequencing of cytosine methylation**. *Genome Research* 2009, **19**(6):959–966.
11. Kerick M, Fischer A, Schweiger MR: **Generation and Analysis of Genome-Wide DNA Methylation Maps**. In *Bioinformatics for High Throughput Sequencing*. Edited by Rodríguez-Ezpeleta N, Hackenberg M, Aransay AM, Springer New York 2012:151–167.
12. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R: **High density DNA methylation array with single CpG site resolution**. *Genomics* 2011, **98**:288–295.
13. Robinson MD, Statham AL, Speed TP, Clark SJ: **Protocol matters: which metylome are you actually studying?** *Epigenomics* 2010, **2**(4):587–598.
14. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA, Feinberg AP: **Increased methylation variation in epigenetic domains across cancer types**. *Nat. Genet.* 2011, **43**(8):768–775.
15. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H: **Continuous base identification for single-molecule nanopore DNA sequencing**. *Nature Nanotechnology* 2009, **4**(4):265–270.

16. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing**. *Nature Methods* 2010, **7**(6):461–465.

17. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, Tavaré S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis**. *Nature Biotechnology* 2008, **26**(7):779–785.

18. Serre D, Lee BH, Ting AH: **MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome**. *Nucleic Acids Research* 2010, **38**(2):391–399.

19. Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, Herwig R, Adjaye J: **Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage**. *Genome Research* 2010, **20**(10):1441–1450.

20. Feber A, Wilson GA, Zhang L, Presneau N, Idowu B, Down TA, Rakyan VK, Noon LA, Lloyd AC, Stupka E, Schiza V, Teschendorff AE, Schroth GP, Flanagan A, Beck S: **Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors**. *Genome Research* 2011, **21**(4):515–524.

21. Lan X, Adams C, Landers M, Dudas M, Krissinger D, Marnellos G, Bonneville R, Xu M, Wang J, Huang THM, Meredith G, Jin VX: **High Resolution Detection and Analysis of CpG Dinucleotides Methylation Using MBD-Seq Technology**. *PLoS ONE* 2011, **6**(7):e22226.

22. Robinson MD, Strbenac D, Stirzaker C, Statham AL, Song JZ, Speed TP, Clark SJ: **Copy-number-aware differential analysis of quantitative DNA sequencing data**. *Genome Research* 2012, [http://dx.doi.org/10.1101/gr.139055.112].

23. Fader PS, Hardie BGS: **A note on modelling underreported Poisson counts**. *Journal of Applied Statistics* 2000, **27**(8):953–964.

24. Abramowitz, Stegun A: *Handbook of Mathematical functions with Formulas, Graphs and Mathematical Tables*. New York: Dover Publications 1970.

25. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR: **Human DNA methylomes at base resolution show widespread epigenomic differences**. *Nature* 2009, **462**(7271):315–322.

26. Robinson MD, Stirzaker C, Statham AL, Coolen MW, Song JZ, Nair SS, Strbenac D, Speed TP, Clark SJ: **Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation**. *Genome Research* 2010, **20**(12):1719–1729.

27. Gneiting T, Raftery AE: **Strictly proper scoring rules, prediction, and estimation**. *Journal of the American Statistical Association* 2007, **102**(477):359–378.

28. Houseman EA, Christensen BC, Karagas MR, Wrensch MR, Nelson HH, Wiemels JL, Zheng S, Wiencke JK, Kelsey KT, Marsit CJ: **Copy number variation has little impact on bead-array-based measures of DNA methylation**. *Bioinformatics* 2009, **25**(16):1999–2005.

29. Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, Beck S, Butcher LM: **Methylome analysis using MeDIP-seq with low DNA concentrations**. *Nature Protocols* 2012, **7**(4):617–636.

30. Carvalho RH, Haberle V, Hou J, van Gent T, Thongjuea S, van Ijcken W, Kockx C, Brouwer R, Rijkers E, Sieuwerts A, Foekens J, van Vroonhoven M, Aerts J, Grosveld F, Lenhard B, Philipsen S: **Genome-wide DNA methylation profiling of non-small cell lung carcinomas**. *Epigenetics Chromatin* 2012, **5**:9.

31. Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, Molinaro AM: **MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment**. *Genome Research* 2008, **18**(10):1652–1659.

32. Wu G, Yi N, Absher D, Zhi D: **Statistical Quantification of Methylation Levels by Next-Generation Sequencing**. *PLoS ONE* 2011, **6**(6):e21034, [http://dx.doi.org/10.1371%2Fjournal.pone.0021034].

33. Hansen KD, Irizarry RA, Wu Z: **Removing technical variability in RNA-seq data using conditional quantile normalization**. *Biostatistics* 2012, **13**(2):204–216.

34. Rue H, Held L: *Gaussian Markov Random Fields: Theory and Applications*. London: Chapman & Hall/CRC Press 2005.

35. Statham AL, Strbenac D, Coolen MW, Stirzaker C, Clark SJ, Robinson MD: **Repitools: an R package for the analysis of enrichment-based epigenomic data**. *Bioinformatics* 2010, **26**(13):1662–1663.

36. Hansen KD, Aryee M: *minfi: Analyze Illumina's 450k methylation arrays*. [R package version 1.3.3].

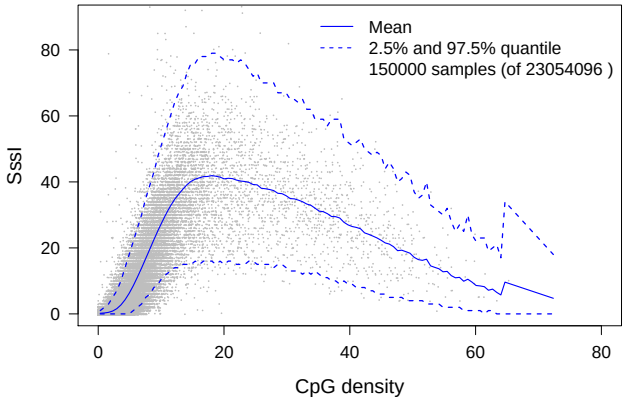
37. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biology* 2010, **11**(3):R25.

38. Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widaa S, Futreal PA, Stratton MR: **PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data**. *Biostatistics* 2010, **11**:164–175.

39. Schmittlein DC, Bemmaor AC, Morrison DG: **Why Does the NBD Model Work? Robustness in Representing Product Purchases, Brand Purchases and Imperfectly Recorded Purchases**. *Marketing Science* 1985, **4**(3):pp. 255–266, [http://www.jstor.org/stable/183907].

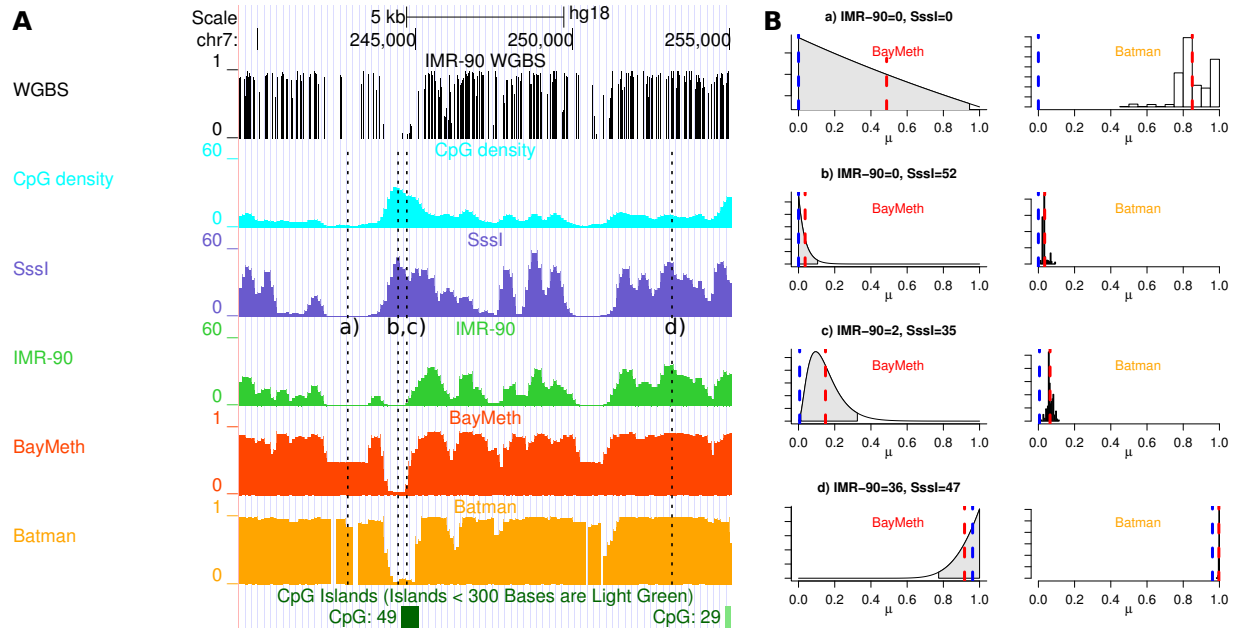
## Figures

**Figure 1 - SssI read depth versus CpG-density together with prior predictive distribution**




Mean, 2.5% and 97.5%-quantile of the prior predictive distribution of the SssI control data together with the read depth of 150000 randomly chosen 100bp regions. The parameters of this negative binomial distribution are derived using an empirical Bayes approach by maximizing the joint marginal distribution of the IMR-90 and SssI control counts stratified into 100 CpG-density groups. Only counts from bins with a mappability larger than 0.75 were considered.

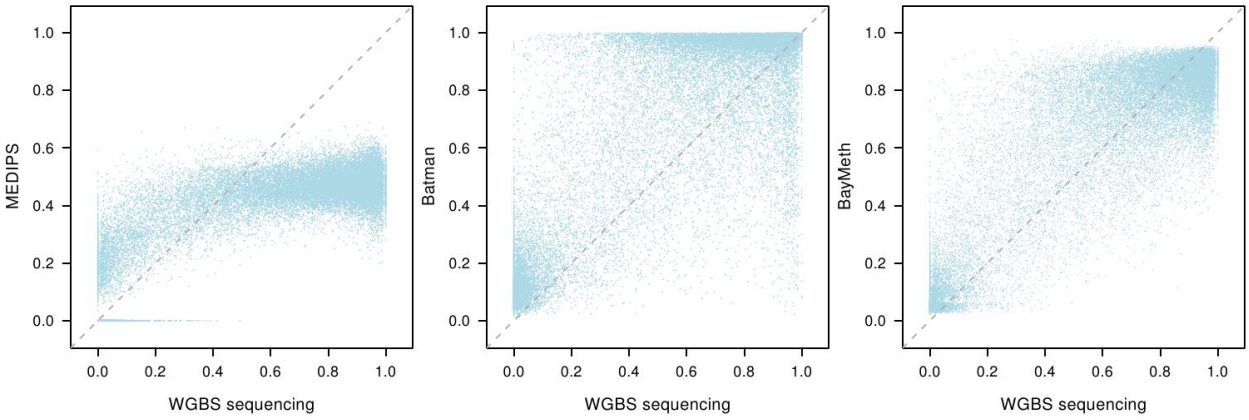
**Figure 2 - Example data tracks for IMR-90 chromosome 7**




Panel A: Shown are the WGBS methylome (black) per CpG-site as obtained by Lister and others [25]. CpG-density (light blue), and read counts for SssI-treated DNA (purple) and IMR-90 cells (green)

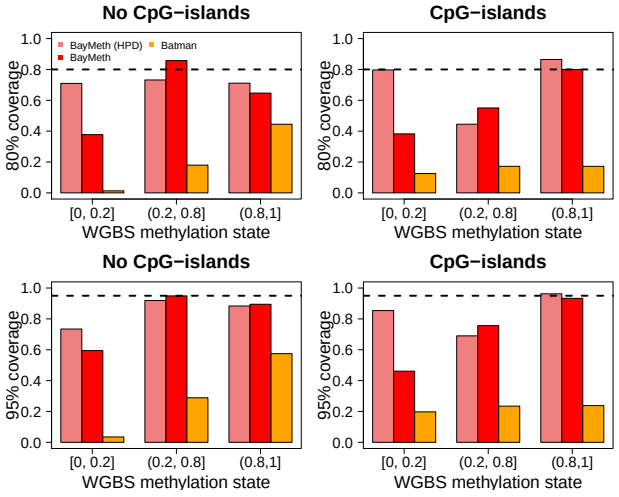
obtained by MBD-seq based on 100bp non-overlapping bins are shown. Methylation estimates for BayMeth (red) and Batman (orange) are provided. Panel B: For 4 specific bins of panel A (denoted a, b, c, d) detailed posterior information of BayMeth and Batman is provided. For BayMeth posterior marginals together with 95% HPD regions (grey-shaded) are shown. For Batman the posterior samples are plotted as histograms. For both approaches the posterior mean is indicated (red dashed line) together with the "true" WGBS derived methylation estimate (blue dashed line).

**Figure 3 - Regional methylation estimates for IMR-90 chromosome 7**



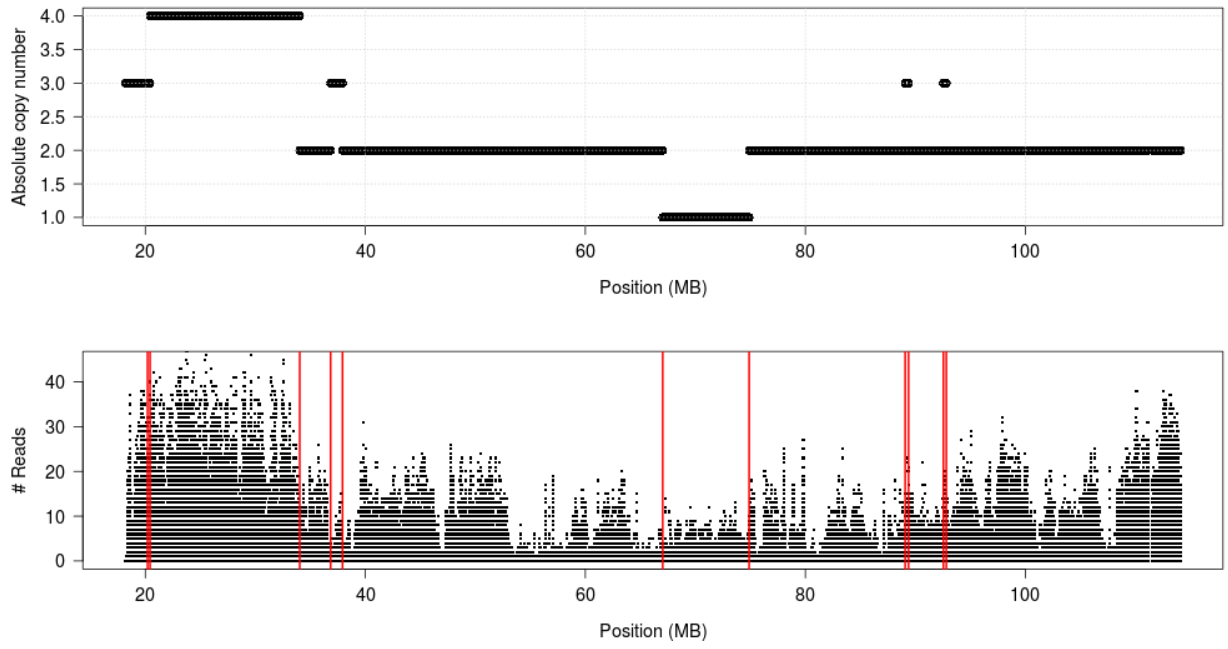

Regional DNAme estimates of MEDIPS, Batman and BayMeth, respectively, plotted against WGBS methylation levels for the 75% of bins with the largest depth in the truth (cutoff are 33 reads) where the depth in the SssI control is $(27, 168]$.

**Figure 4 - Coverage probabilities stratified by CpG island status and true methylation level**



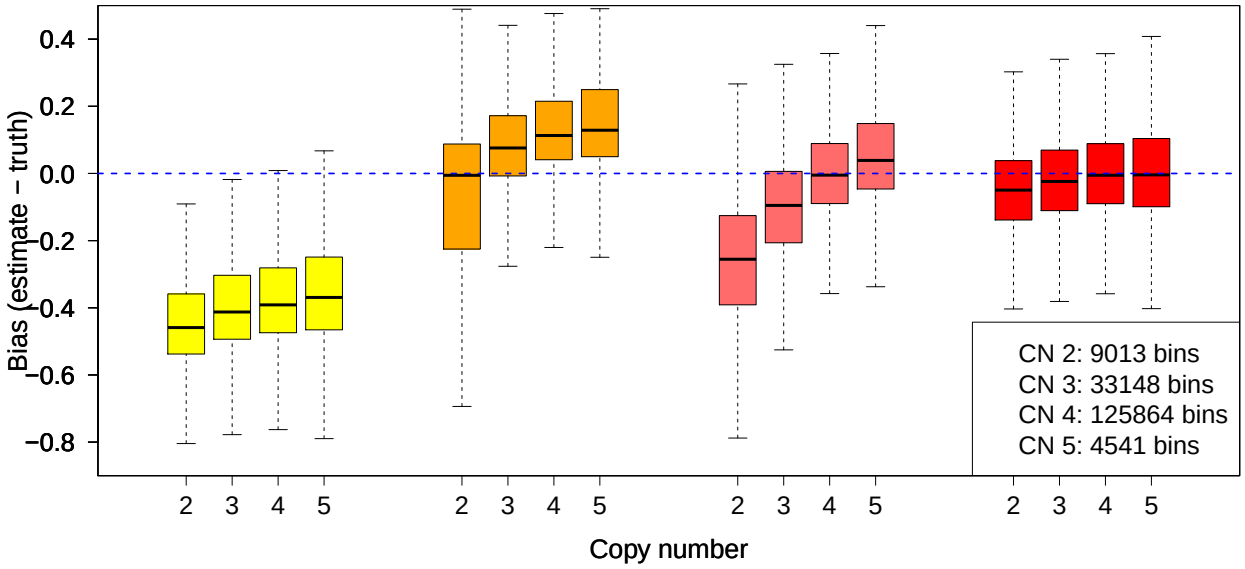
Coverage probabilities (frequency in which the true value is within a predefined credible interval) at 80% and 95% level are shown for the 75% of bins with the largest depth in the truth (cutoff are 33 reads). HPD intervals (light red) and intervals based on quantiles (red) are used for BayMeth. For Batman only quantile-based intervals (orange) are available, while MEDIPS does not return any uncertainty estimates. The nominal coverage value is indicated (black dashed line) as a reference. Genomic regions are stratified by CpG-density using the threshold of 12.29 which separates CpG islands from non-CpG islands, compare Supplementary Figure 1. Further stratification by the true methylation level as derived from WGBS [25] is provided.

**Figure 5 - Relation between copy number state and regional affinity enrichment**



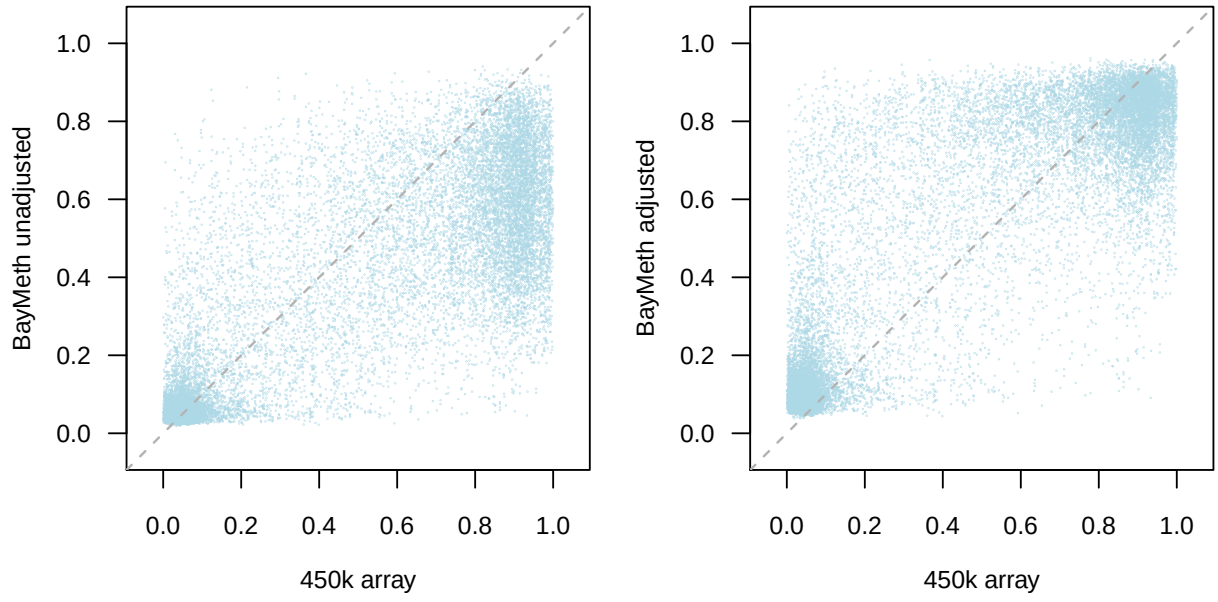

Top: Copy number estimates of LNCaP cell line obtained by the PICNIC [38] algorithm for 100bp bins across human chromosome 13 with a mappability of at least 75%. Bottom: Read counts of affinity capture sequencing data for the same bins.

**Figure 6 - Bias of LNCaP methylation estimates compared to 450k array beta values**




Boxplot of bias (Estimated methylation level - 450K array beta value) for MEDIPS (yellow), Batman (orange), unadjusted BayMeth (light red) and adjusted BayMeth (red) stratified by the most prominent copy number. (Outliers are not shown.) The results are shown genome-wide for 100bp bins with at least 75% mappability and where the true methylation estimate is larger than 0.5. A threshold of 13 is applied for the depth of SssI. The blue dashed line indicates a bias of zero.

**Figure 7 - Effect of adjusting for CNV in LNCaP cell line**



Methylation estimates for copy number state two derived by BayMeth compared to 450k array beta values. A threshold of 13 is applied for the depth of SssI, which leads to 61696 100bp-bins. Left: Without adjustments for CNV. Right: With adjustment for CNV.

# Tables

## 1.2 Table 1 - Performance assessment for IMR-90 analysis (chromosome 7)

Results are shown for bins with a truth depth larger than the 25% quantile (cutoff are 33 reads), stratified into five groups by SssI depth. Shown are the number of bins per group, mean bias, MSE, Spearman correlation, DSS and coverage probabilities at 80% and 95% level.

| SssI depth | #Bins | Method | Bias | MSE | Cor | DSS | 80%-coverage quantile (HPD) | 95%-coverage quantile (HPD) |
|---|---|---|---|---|---|---|---|---|
| $[0, 4]$ | 305638 | BayMeth | -0.04 | 0.08 | 0.36 | -1.53 | 0.70 (0.72) | 0.89 (0.89) |
| | | Batman | 0.22 | 0.14 | 0.31 | 325.42 | 0.30 | 0.43 |
| | | MEDIPS | -0.38 | 0.26 | 0.29 | — | — | — |
| $(4, 7]$ | 22196 | BayMeth | 0.05 | 0.05 | 0.65 | -1.98 | 0.75 (0.72) | 0.88 (0.88) |
| | | Batman | 0.16 | 0.07 | 0.61 | 3726.37 | 0.23 | 0.34 |
| | | MEDIPS | -0.23 | 0.11 | 0.45 | — | — | — |
| $(7, 14]$ | 28871 | BayMeth | 0.06 | 0.04 | 0.69 | -1.97 | 0.73 (0.70) | 0.86 (0.86) |
| | | Batman | 0.16 | 0.07 | 0.65 | 5721.26 | 0.19 | 0.28 |
| | | MEDIPS | -0.21 | 0.10 | 0.49 | — | — | — |
| $(14, 27]$ | 28928 | BayMeth | 0.05 | 0.03 | 0.76 | -2.28 | 0.70 (0.70) | 0.83 (0.86) |
| | | Batman | 0.15 | 0.06 | 0.73 | 2158.72 | 0.15 | 0.23 |
| | | MEDIPS | -0.20 | 0.09 | 0.59 | — | — | — |
| $(27, 168]$ | 28719 | BayMeth | 0.02 | 0.03 | 0.78 | -2.56 | 0.64 (0.72) | 0.78 (0.86) |
| | | Batman | 0.11 | 0.05 | 0.75 | 13558.87 | 0.13 | 0.20 |
| | | MEDIPS | -0.22 | 0.10 | 0.67 | — | — | — |

## 1.3 Table 2 - Copy number specific offset

Copy number specific offsets defined as $f \times \frac{\text{cn}_i}{\text{ccn}}$ derived for 100bp non-overlapping bins of LNCaP autosomes, which have a mappability of at least 75%. Note, that $f$ is only derived based on bins with the most common copy number state four.

| Copy number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Combined offset | 0.178 | 0.356 | 0.534 | 0.712 | 0.890 | 1.068 | 1.246 | 1.424 |

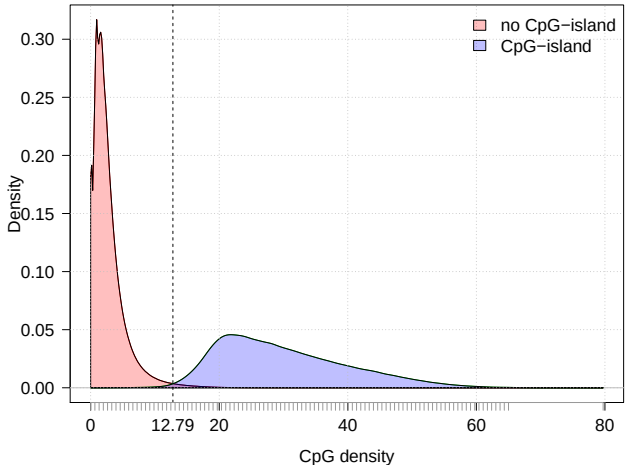
## 1.4 Table 3 - Performance assessment for LNCaP analysis by copy number

Results are shown for 100bp-bins with a mappability of at least 0.75 stratified into the four most frequent copy number states. A threshold of 13 is applied for the depth of the SssI-control. Shown are the number of bins per copy number state, mean bias, MSE and Spearman correlation.

| Copy number | #Bins | Method | Bias | MSE | Cor |
|---|---|---|---|---|---|
| 2 | 18011 | BayMeth (adj) | 0.04 | 0.04 | 0.78 |
| | | BayMeth (unadj) | -0.12 | 0.06 | 0.78 |
| | | Batman | 0.03 | 0.06 | 0.74 |
| | | MEDIPS | -0.23 | 0.11 | 0.76 |
| 3 | 65982 | BayMeth (adj) | 0.05 | 0.04 | 0.80 |
| | | BayMeth (unadj) | -0.02 | 0.04 | 0.80 |
| | | Batman | 0.11 | 0.06 | 0.77 |
| | | MEDIPS | -0.19 | 0.09 | 0.76 |
| 4 | 256078 | BayMeth (adj) | 0.05 | 0.04 | 0.81 |
| | | BayMeth (unadj) | 0.05 | 0.04 | 0.81 |
| | | Batman | 0.16 | 0.08 | 0.79 |
| | | MEDIPS | -0.17 | 0.09 | 0.76 |
| 5 | 11790 | BayMeth (adj) | 0.04 | 0.03 | 0.83 |
| | | BayMeth (unadj) | 0.08 | 0.04 | 0.83 |
| | | Batman | 0.18 | 0.08 | 0.80 |
| | | MEDIPS | -0.12 | 0.07 | 0.80 |

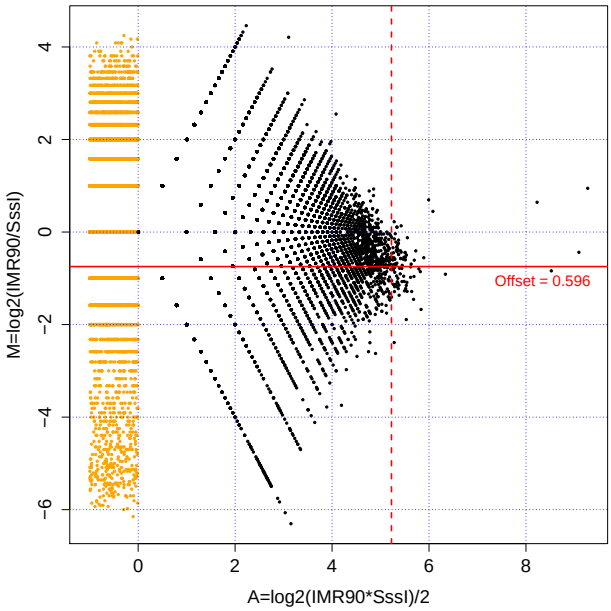
## Supplementary Figures

### 1.5 Figure 1 - CpG-density stratified by CpG island status



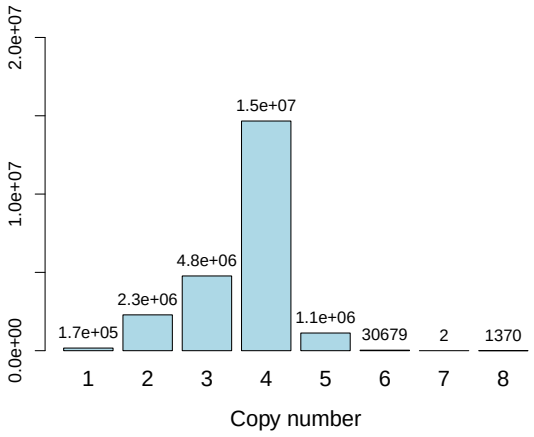

Genome-wide CpG-density for bins with a mappability larger than 75% stratified by CpG island status as extracted from the cpgIslandExt-table of the UCSC genome browser. The vertical line marks the intersection of both densities. The grey tick-marks along the x-axis illustrate the CpG-density classes used for the empirical Bayes approach in the IMR-90 application.

### 1.6 Figure 2 - Normalizing offset



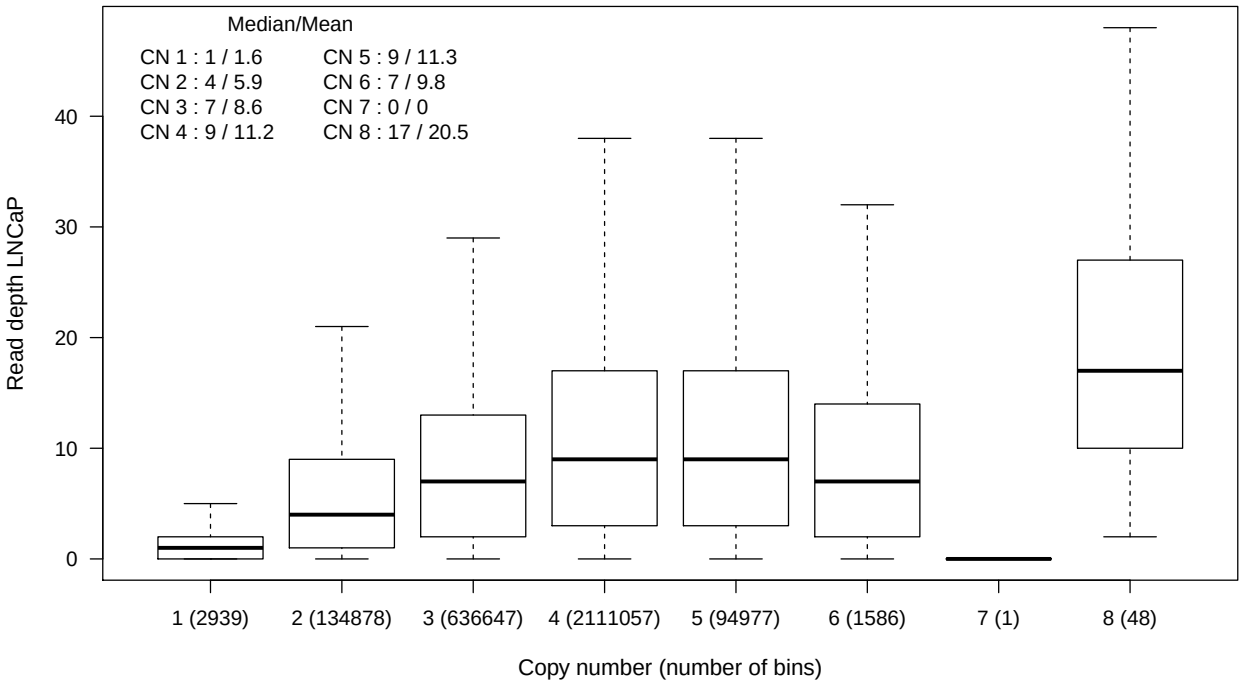

Log-fold change ($M$) versus log-concentration ($A$) illustrated for 50000 randomly chosen bins. The red dotted line shows the 0.998 quantile $q$ of $A$ determined from all bins. The red straight line shows the estimated normalization offset $f = 2^{median(M_{A>q})}$. A 'smear' of yellow points at a low $A$ value represents counts that are low in either of the two samples.

### 1.7 Figure 3 - Copy number frequencies for LNCaP



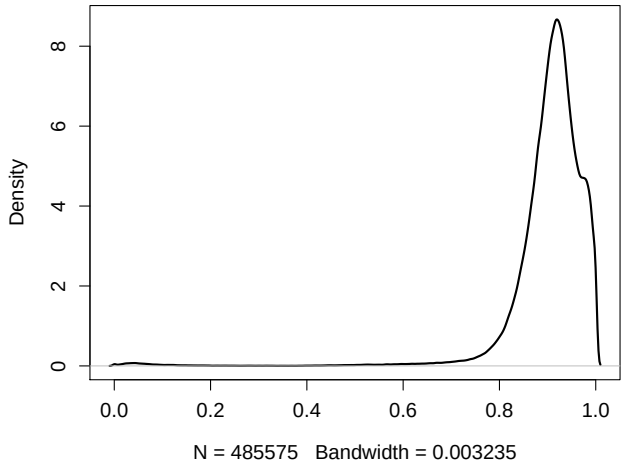

Copy number frequencies in LNCaP for 100bp-bins with a mappability larger than 0.75.

### 1.8 Figure 4 - Read depth of LNCaP by copy number




Read depth stratified by copy number is shown for 100bp-bins with a mappability larger than 0.75 and with a SssI depth larger than four. Median and mean read depth are given per copy number state.

### 1.9 Figure 5 - Distribution of estimated methylation levels for SssI sample using Illumina HumanMethylation450 arrays




Density plot of 450k beta values.