Chapter 13

Evaluation of MeDIP-Chip in the Context of Whole-Genome Bisulfite Sequencing (WGBS-Seq) in *Arabidopsis*

René Wardenaar, Haiyin Liu, Vincent Colot, Maria Colomé-Tatché, and Frank Johannes

Abstract

Studies of DNA methylation in *Arabidopsis* have rapidly advanced from the analysis of a single reference accession to investigations of large populations. The goal of emerging population studies is to detect differentially methylated regions (DMRs) at the genome-wide scale, and to relate this variation to gene expression and phenotypic diversity.

Whole-genome bisulfite sequencing (WGBS-seq) has established itself as a gold standard in DNA methylation analysis due to its high accuracy and single cytosine measurement resolution. However, scaling up the use of this technology for large population studies is currently not only cost prohibitive but also poses nontrivial bioinformatic challenges. If the end-point of the study is to detect DMRs at the level of several hundred base pairs rather than at the level of single cytosines, low-resolution array-based methods, such as MeDIP-chip, may be entirely sufficient. However, the trade-off between measurement accuracy and experimental/analytical practicality needs to be weighted carefully. To help make such experimental choices, we conducted a side-by-side comparison between the popular dual-channel MeDIP-chip Nimblegen technology and Illumina WGBS-seq in two independent *Arabidopsis* lines.

Our analysis shows that MeDIP-chip performs reasonably well in detecting DNA methylation at probe-level resolution, yielding a genome-wide combined false-positive and false-negative rate of about 0.21. However, detection can be susceptible to strong signal distortions resulting from a combination of dye bias and the CG content of effectively unmethylated genomic regions. We show that these issues can be easily bypassed by taking appropriate data preparation steps and applying suitable analysis tools.

We conclude that MeDIP-chip is a reasonable alternative to WGBS-seq in emerging *Arabidopsis* population epigenetic studies.

Key words DNA methylation, MeDIP-chip, Whole-genome bisulfite sequencing, Dye bias, *Arabidopsis*, Population epigenetics, Epigenomics

1 Introduction

DNA methylation is an epigenetic modification that involves the addition of a methyl group to the five position of the cytosine pyrimidine ring. In most animals and plants, this modification has a central role in the regulation of gene expression and in the

Tin-Lap Lee and Alfred Chun Shui Luk (eds.), *Tiling Arrays: Methods and Protocols*, Methods in Molecular Biology, vol. 1067,
DOI 10.1007/978-1-62703-607-8_13,

silencing of transposable elements [1]. Because of its important biological functions, there have been substantial efforts to characterize the complete DNA methylomes of various organisms [2].

In the model plant *Arabidopsis*, the first whole-genome DNA methylation analysis used single-channel Affymetrix and dual-channel Nimblegen tiling arrays [3, 4]. These studies relied on methylation-dependent immunoprecipitation techniques followed by hybridization to high-density microarray chips (MeDIP-chip), and achieved a resolution of 35 and 220 bp, respectively. This work was instrumental in providing the first picture of the distribution of DNA methylation and its relationship with known sequence annotation in this species. A more detailed view was later obtained by several studies employing Illumina whole-genome bisulfite sequencing (WGBS-seq [5, 6]). WGBS-seq combines bisulfite conversion of DNA with next-generation sequencing (NGS) technologies and provides single cytosine resolution.

While the above-mentioned studies focused on the DNA methylome of a single reference plant, more recent work has begun to document interindividual variation in DNA methylation in large populations. The ultimate aim is to relate this type of epigenetic variation to phenotypic diversity, and to ask broader questions about the role of epigenetics in adaptive evolution. A first step in this direction was recently taken by Schmitz et al. [7] and Becker et al. [8]. These authors performed WGBS-seq on 8–12 *Arabidopsis* lines and quantified the frequency and distribution of single methylation polymorphisms (SMPs) as well as differentially methylated regions (DMRs). These experiments generated roughly 200–500 GB of data and required construction of extensive data pipelines. Scaling up the use of WGBS-seq to even larger samples poses nontrivial bioinformatic challenges that range from data storage to downstream computation analysis. These challenges can hinder the routine application of this technology for future population epigenetic studies.

A viable alternative is to restrict DNA methylation analysis to the detection of DMRs, which typically range between 10 and 1,000 bp in length. In *Arabidopsis* as in other species DMRs appear to be functionally more important than SMPs [7–10] and appear to be a suitable unit of analysis. Focusing on DMRs has the important advantage that array-based measurement technologies, such as MeDIP-chip, could be employed in place of WGBS-seq because they provide sufficient resolution. The use of array-based methods can substantially reduce the bioinformatic resources required to perform population epigenetic studies. Nonetheless, loss of measurement accuracy resulting from hybridization compared to sequencing may present a significant drawback which not all researchers are willing to accept. Furthermore, it should be noted that unlike WGBS-seq, MeDIP-chip does not allow distinguishing between CG, CHG, and CHH methylation, a point which may be

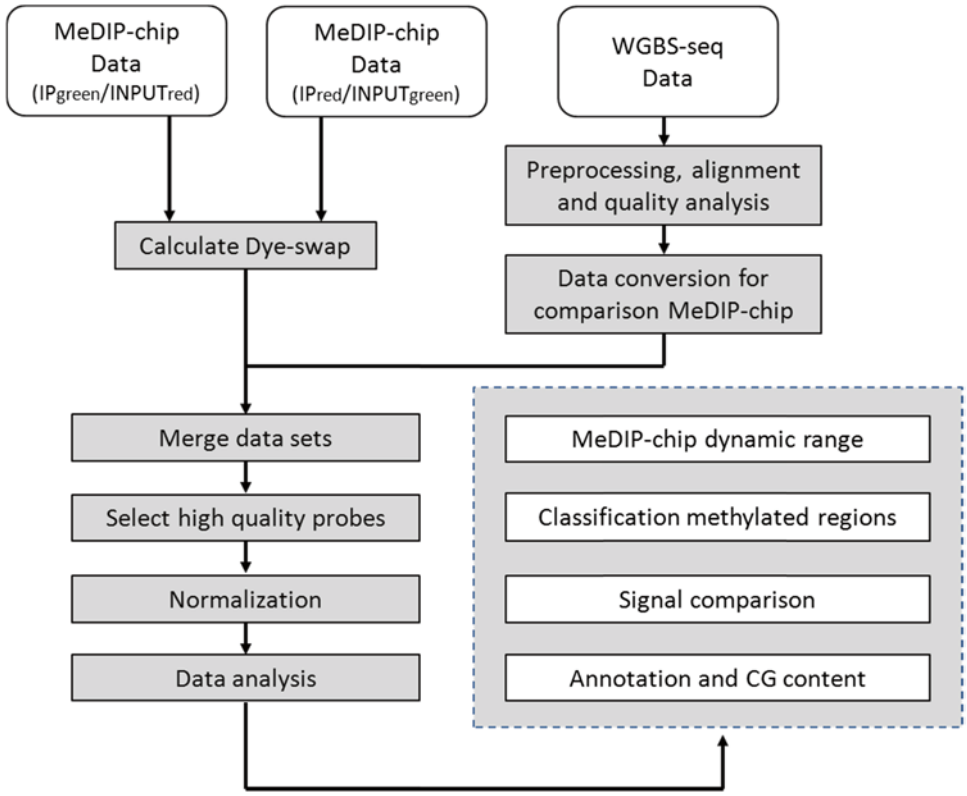

Fig. 1 Workflow evaluation MeDIP-chip

important to consider in some instances. Hence, the trade-off between loss of measurement accuracy and experimental/analytical practicality needs to be weighted carefully. To help make such experimental choices we conducted a side-by-side comparison between the popular dual-channel MeDIP-chip Nimblegen technology and Illumina WGBS-seq. The workflow shown in Fig. 1 serves as an outline of this chapter.

Our analysis shows that the dual-channel MeDIP-chip technology performs reasonably well in detection of probe-level DNA methylation, which is approximately the minimum resolution required for DMR detection. We estimate that MeDIP-chip yields a combined false-positive and false-negative rate of 0.21 genome-wide. However, we also find that detection can be critically dependent on prior data preparation steps, signal distortions arising from dye biases, and the statistical method used for detection. Based on our results, we make several simple but important recommendations regarding the experimental implementation of MeDIP-chip for population epigenetic studies (*see* **Note 1–3**).

2 Data Sets and Data Preparation

In this chapter we consider DNA methylation data from two different epigenetic recombinant inbred lines (epiRILs; R60 and R202 [11]). The DNA methylomes of each epiRIL were measured using dual-channel Nimblegen MeDIP-chip and Illumina WGBS-seq [12]. This section provides a brief overview of these two measurement technologies as well as key data preparation steps.

2.1 MeDIP-Chip

Methylated DNA immunoprecipitation (MeDIP) is a large-scale purification technique used for the enrichment of methylated DNA fragments. This technique uses antibodies specific to methylated cytosines in order to separate methylated DNA fragments from unmethylated DNA fragments. Following this separation procedure the methylated fragments can either be hybridized to a tilling array (MeDIP-chip [13]) or sequenced (MeDIP-seq [14]) in order to assess the methylation status of the genome under consideration. The application of MeDIP-chip to the two epiRILs under consideration involved several experimental steps which are discussed in short in this paragraph. A more extensive description of the protocol used to obtain MeDIP-chip data described in this chapter is given by Cortijo et al. [15].

1. *Extraction and fragmentation*—DNA was extracted from aerial parts of 3-week-old *Arabidopsis* plants using a standard extraction kit (Qiagen DNeasy plant Maxi kit). Extracted DNA was fragmented using sonication. Sonication produced fragments with a size between 100 and 600 bp (verified with gel electrophoresis).
2. *Immunoprecipitation and amplification*—After this fragmentation step, the DNA was denatured and anti-5mC antibody was added to the IP DNA pool, which recognizes specifically methylated cytosines in single-stranded DNA. Magnetic beads containing binding sites for this antibody were then added to pull down (i.e., immunoprecipitate) methylated fragments. Following release of the antibody, both IP and input DNA fractions were amplified by PCR.
3. *Labeling and hybridization*—The IP and input fractions were differentially labeled with two fluorescent dyes (Cy3 and Cy5) and hybridized to Nimblegen whole-genome *Arabidopsis* tiling arrays (3×720 K array) containing 711,320 isothermal probes. Depending on the CG content, these probes range from 50 to 75 nucleotides in length and have an inter-probe spacing of about 110 base pairs on average.
4. *Scanning the tiling array*—After hybridization, the intensities of both dyes were obtained by scanning the tiling array. The scanner outputs two files with the raw intensities for each

probe on the tiling array: one file with the IP signals and the other file with the input signals.

5. *Signal calculation*—The IP and input signals were log transformed and subtracted from each other ($\log_2(\text{IP}) - \log_2(\text{input})$). Hence, probes with a low signal are from genomic regions that show low methylation levels, and those with a high signal are from genomic regions that show high methylation levels.

2.2 Whole-Genome Bisulfite Sequencing

WGBS-seq is a NGS technology used to determine the DNA methylation status of single cytosines. In the case of WGBS-seq, unlike other NGS technologies, the DNA is treated with sodium bisulfite before sequencing. Sodium bisulfite is a chemical compound that converts unmethylated cytosines into uracil [16, 17]. Knowing which cytosines have converted it is possible to determine which cytosines are methylated (not converted) and which ones are unmethylated (converted into U). After sequencing, the unmethylated cytosines appear as thymines. There are several ways of producing sequence libraries for WGBS-seq (*see* Ref. 18). The WGBS-seq data in this chapter was produced with a technique developed by Cokus et al. [6]. Here we describe in short the procedure.

1. *Extraction and fragmentation*—DNA was extracted from aerial parts of 3-week-old *Arabidopsis* plants using standard extraction kit (Qiagen DNeasy plant Maxi kit). Extracted DNA was fragmented by sonication.
2. *Adapter ligation and size selection*—A set of double-stranded adapter sequences was ligated to the fragmented DNA. These adapter sequences contained methylated adenine bases with DpnI restriction sites. The restriction sites are important for the removal of the first set of adapter sequences in one of the subsequent steps. Gel electrophoresis was used to obtain adapter-ligated fragments with an appropriate size.
3. *Bisulfite conversion and amplification*—After the addition of the adapter sequences the sodium bisulfite conversion is performed. During this step, unmethylated cytosines are changed into uracil. PCR was subsequently performed with the use of primers that were complementary to the converted adapter sequences.
4. *Removal of first adapter sequences and ligation sequencing adapters*—After the first PCR amplification step the first set of adapters was removed using DpnI restriction enzymes. A new set of sequencing adapters was subsequently ligated to BS-converted DNA fragments.
5. *Size selection and amplification*—Fragments with a size between 120 and 170 bp were selected with the use of gel electrophoresis, and a second and final PCR step was performed using

primers complementary to the sequencing adapters to yield a sequencing library.

6. *Sequencing*—Illumina sequencing technology (Illumina 1G Genome Analyzer) was used to produce read sequences with a length of 76 or 101 nt.

   After sequencing, the reads need to be mapped (or aligned) to a reference genome in order to infer the methylation status of the cytosines of the genome under consideration. However, mapping these converted sequences is not straightforward since unmethylated cytosines will result in mismatches with the reference genome (i.e., they appear as thymines). To circumvent this issue several programs have been developed that first convert the reference genome into a three-letter genome (i.e., all cytosines are changed into thymines; in silico [19]). The remaining cytosines of the read sequences also need to be changed into thymines before mapping. The in silico-treated read sequences are subsequently mapped to this three-letter reference genome. Once the mapping has been performed, methylation status can be inferred using the original sequence of the reference genome and the read. A thymine (read) mapped to a cytosine (reference) is an unmethylated cytosine. A cytosine mapped to a cytosine is a methylated cytosine. We utilized BS Seeker [20] for the mapping of the read sequences. BS Seeker is a python-based open-source mapping program for the alignment of bisulfite-treated sequences. The analysis includes preprocessing of the reads prior to alignment, the alignment itself, and quality analysis of the data. The steps are described further below. More details of the mapping of the reads can be found elsewhere [12].

7. *Removal of adapter parts*—When a DNA fragment is shorter than the read sequence a part of the adapter sequence will also be sequenced. The adapter sequence was added artificially and does not match with the reference genome. We therefore removed this part using a sliding window approach. The part that overlapped with the known adapter sequence was removed.

8. *Removal of short reads*—The removal of the adapter sequences resulted in some cases in reads with a length smaller than 30 nucleotides. These short reads are more difficult to map and were therefore removed. The proportion of short reads was in our case quite small and therefore the final read coverage was barely affected.

9. *Removal of duplicated reads*—Duplicated reads were removed in the final preprocessing step because they were likely produced during PCR amplification and were therefore not informative.

10. *Alignment to reference genome*—After these preprocessing steps the reads were mapped to a reference genome (TAIR 10)

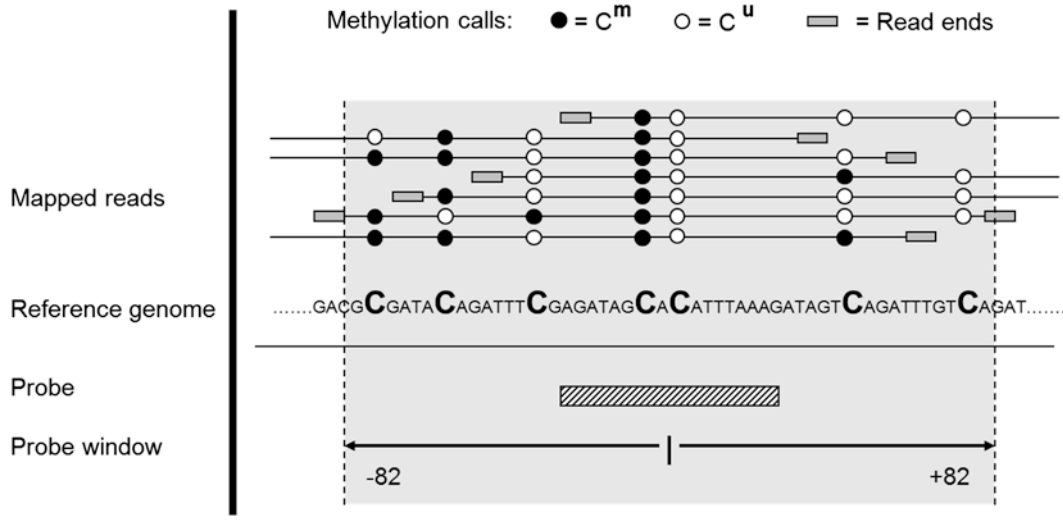

Fig. 2 Calculation of whole-genome bisulfite sequencing signals. The methylation calls within the window (*gray*) are used to calculate a normalized whole-genome bisulfite sequencing signal (see formula)

with the use of BS Seeker. After mapping the obtained average genome coverage was 29 and 27× for epiRIL 60 and 202, respectively (both strands combined).

11. *Determination conversion rate*—One important step in the quality analysis of the data is to determine the (bisulfite) conversion rate. The conversion rate, which is the percentage of unmethylated cytosines that effectively changed into uracil, was determined for both epiRILs after mapping. The conversion rates were determined with the information of reads that were mapped to chloroplast DNA. The chloroplast DNA is known to be unmethylated and therefore any detected methylated cytosine is considered to be a non-converted unmethylated cytosine. Both epiRILs showed a conversion rate above 99 % which indicates that the data is of good quality.

2.3 Data Conversion and Normalization

One of the major differences between MeDIP-chip and WGBS-seq is mapping resolution. WGBS-seq can interrogate the methylation status of individual cytosines while MeDIP-chip achieves a resolution of about 165 bps. In order to facilitate a meaningful comparison between the two technologies, we converted the WGBS-seq data into a format that is comparable to that of MeDIP-chip. To achieve this we calculated the proportion of methylation calls in windows of 165 bps centered at the probe sequence (Fig. 2). By methylation calls we mean the individual methylation calls of each read sequence. This results in a signal ranging from zero to one. This signal is afterwards normalized for the number of cytosines in the probe window. Let C_j^m denote the number of cytosines that have been called methylated in the jth window, C_j^u the number of

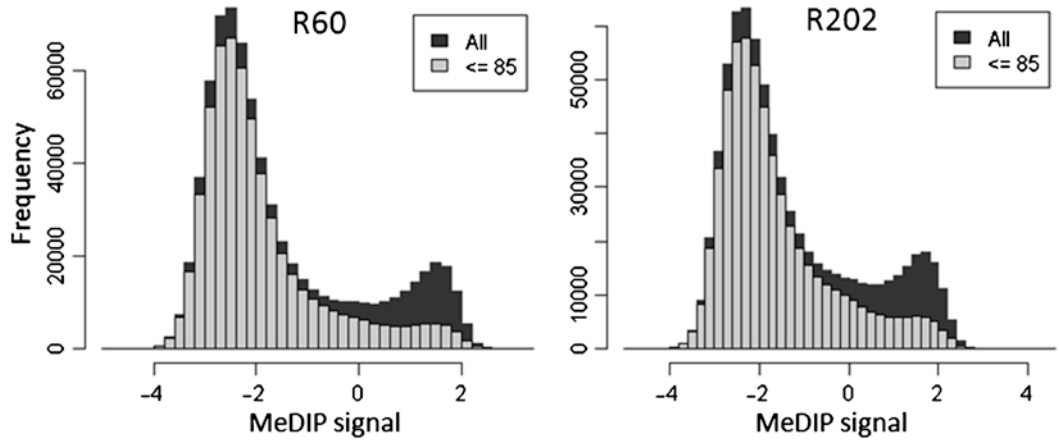

Fig. 3 Impact of removing probes with a high conservation score on probe signal distribution. Probes with a high conservation score (*dark gray*) typically show signal intensities similar to probes that correspond to genomic regions with high methylation levels

cytosines that have been called unmethylated in the jth window, R_j the total number of cytosines in the jth window according to the reference genome, and R_{max} the maximum number of cytosines across all windows. The converted and normalized WGBS-seq signal can be calculated as

$$\mathrm{WGBS}_{\mathrm{sig}}(j) = \frac{C_j^m}{C_j^m + C_j^u} \times \frac{R_j}{R_{\max}}.$$

We selected high-quality data for the analysis described in this chapter. In case of the WGBS-seq probe windows, we only selected windows with 35 or more cytosines, with at least half of the cytosines being covered by one or more read sequences. In case of the MeDIP-chip data we only selected probes with a conservation score smaller or equal to 85. The conservation score of a probe indicates the uniqueness of a probe sequence. These scores were obtained by performing a blast search. Scores are percentage of identity with the second best hit (score range 45–100). The best hit is with the genomic location for which the probe was designed. Probes with a high conservation score are more likely to cause cross-hybridization problems. As shown in Fig. 3, removal of probes with a high conservation score has a drastic impact on the MeDIP signal distribution, as they typically show signal intensities similar to probes that correspond to genomic regions with high methylation levels. *See* **Note 1** for recommendations concerning the quality of the data.

Analysis was performed on probes (i.e., probe windows) that were present in both data sets (MeDIP-chip and WGBS-seq signal

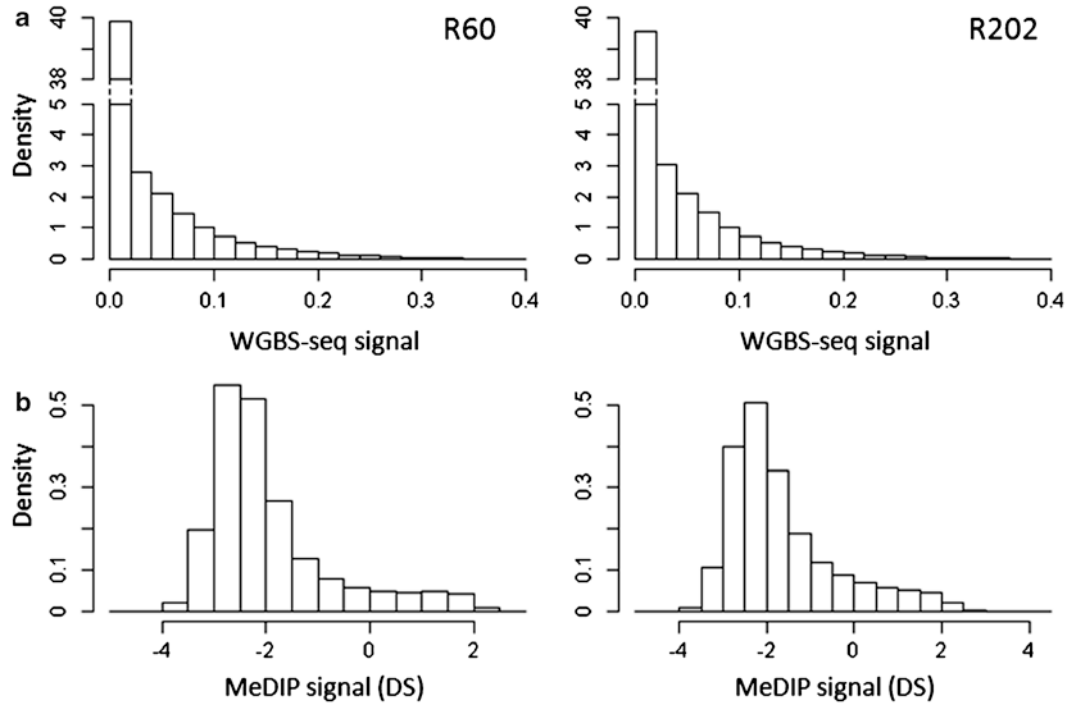

Fig. 4 Density histogram of WGBS-seq signals and MeDIP-chip signals following data conversion and normalization. (**a**) WGBS-seq signal distribution. (**b**) MeDIP-chip signal distribution

data). This yielded 551,688 and 550,676 high-quality probe windows in total, covering approximately 77.6 and 77.4 % of the genomes of R60 and R202, respectively. In case of the MeDIP-chip data a total of two dye-swap experiments were performed for each epiRIL. The log-transformed signals were also averaged over the two dye-swap experiments which resulted into three MeDIP-chip data sets:

- G/R data: IP labeled green (Cy3, G) and input labeled red (Cy5, R).
- R/G data: IP labeled red (Cy5, R) and input labeled green (Cy3, G).
- DS data: Average of G/R and R/G data (dye-swap data).

Finally, quantile normalization was applied to bring these three data sets to a common scale [21]. Figure 4 displays a density histogram of the WGBS-seq signal for the two epiRIL experiments (R60 and R202) and the MeDIP-chip signal (DS).

2.4 Software

The analysis described in this chapter was performed using R [22]. R is a command-line software environment for statistical computing and graphics. The Hidden Markov Model (HMM) for probe classification was programmed in C++ [15].

3 Results

3.1 Assessment of MeDIP-Chip Dynamic Range

Since the WGBS-seq signal provides a measure of the proportion of methylated cytosines in a given probe window, we were able to assess the dynamic range of the MeDIP-chip technology directly by empirical comparison. To achieve this we calculated the median MeDIP-chip signal for sliding windows along the entire WGBS-seq signal range (Window size: 0.05; step size: 0.01; Fig. 5a). We find that MeDIP-chip exhibits good sensitivity for low-to-intermediate methylation levels (WGBS-seq range: 0.00 to ~0.13). In this low-to-intermediate range, there is a nearly linear relationship between the WGBS-seq signal and the MeDIP median signal, but the MeDIP-chip sensitivity falls off quickly and saturates at a WGBS-seq signal value of about 0.28. Above this point, MeDIP-chip is effectively unable to differentiate between methylation levels. However, in R60 and R202 this saturation effect affects only a relatively small number of probe windows, 0.19 % (N = 1,056) and 0.22 % (N = 1,220) of all regions genome-wide, and 0.58 and 0.67 % of all methylated regions (*see* Table 1), respectively. Signal saturation should therefore not be a matter of great concern in the analysis of the *Arabidopsis* methylome. Similar conclusions can be reached when considering the MeDIP signal on its original scale (data not shown), rather than on the log-transformed scale, which indicates that saturation is not caused by scaling issues.

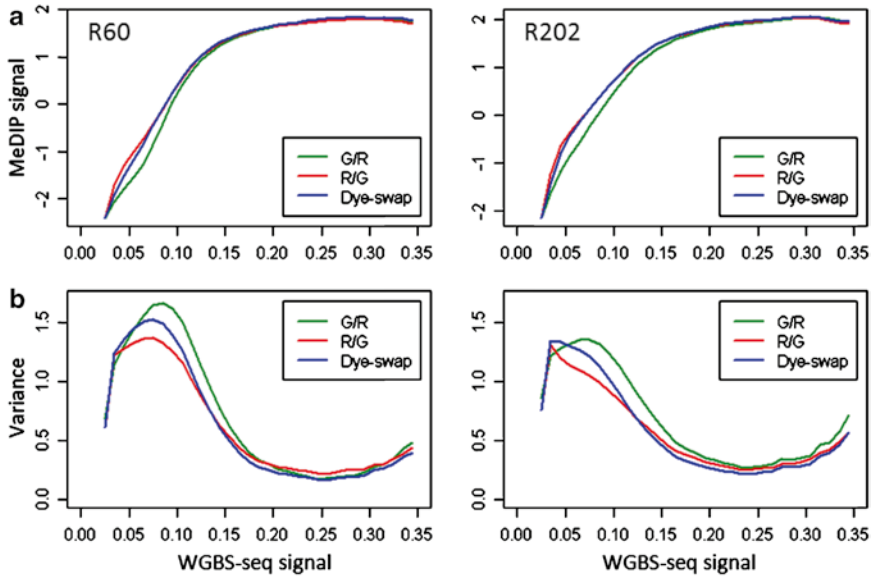

Fig. 5 Median and variance of the MeDIP signal along the entire WGBS-seq signal range. Median MeDIP signal (**a**) and variance MeDIP signal (**b**) for sliding window along the entire WGBS-seq signal range. The G/R data show less sensitivity for low-to-intermediate WGBS-seq signals and also show a higher variance compared to the R/G data

Table 1
Classification probe windows using WGBS-seq

	R60	R202
Classification cutoff	5.09E–3	7.03E–3
# of probe windows selected for analysis	551,688 (77.6)	550,676 (77.4)
# of unmethylated probe windows	369,358 (67.0)	367,526 (66.7)
# of methylated probe windows	182,330 (33.0)	183,150 (33.3)
Median signal, unmethylated windows	9.58E–4 (0.000–5.09E–3)	9.65E–4 (0.000–7.02E–3)
Median signal, methylated windows	3.32E–2 (5.09E–3–0.552)	3.33E–2 (7.03E–3–0.562)

Reported are the number of unmethylated and methylated windows using a genome-wide false-positive rate of 0.01

Our analysis also indicates that there are clear dye-related differences in the MeDIP median signal. The R/G data appears to respond more sensitively to changes in WGBS-seq methylation levels compared with the G/R and the DS data (Fig. 5a). While these dye differences disappear at the saturation point (WGBS-seq signal ~ 0.28), they are most prominent in the optimal dynamic range. This divergence is even more severe when we consider the MeDIP signal variance across the complete WGBS-seq range using the same sliding window approach (Fig. 5b). Ideally, the signal variance should be low and constant across methylation levels. Figure 5b illustrates that this is clearly not the case: the MeDIP signal variance is largest within the optimal dynamic range but decreases rapidly with increasing methylation levels. Notably, the G/R data displays a 1.20 (R60)- and a 1.28 (R202)-fold increase in signal variance (on average) relative to the R/G data suggesting that it is substantially noisier, and the dye-swap (DS) fails to correct this bias. This latter observation is contrary to what is typically seen in expression microarrays where dye-swaps have proved to be an effective strategy [23, 24]. *See* **Note 2** for recommendations concerning the labeling of the IP and input DNA.

3.2 Classification of Methylated Regions Using MeDIP-Chip

From a data analysis standpoint, the classification of individual probes according to their underlying methylation status (i.e., methylated or unmethylated) is critically dependent on the variance of the MeDIP signal. When signal variation is large, classification tends to be more difficult. Many statistical analysis methods have been proposed to minimize this problem and to facilitate accurate probe classification in the context of MeDIP-chip data (e.g., [15, 25–29]). The development of such methods continues to be an active area of research. Classification is particularly problematic in regions of the MeDIP signal distribution where signals

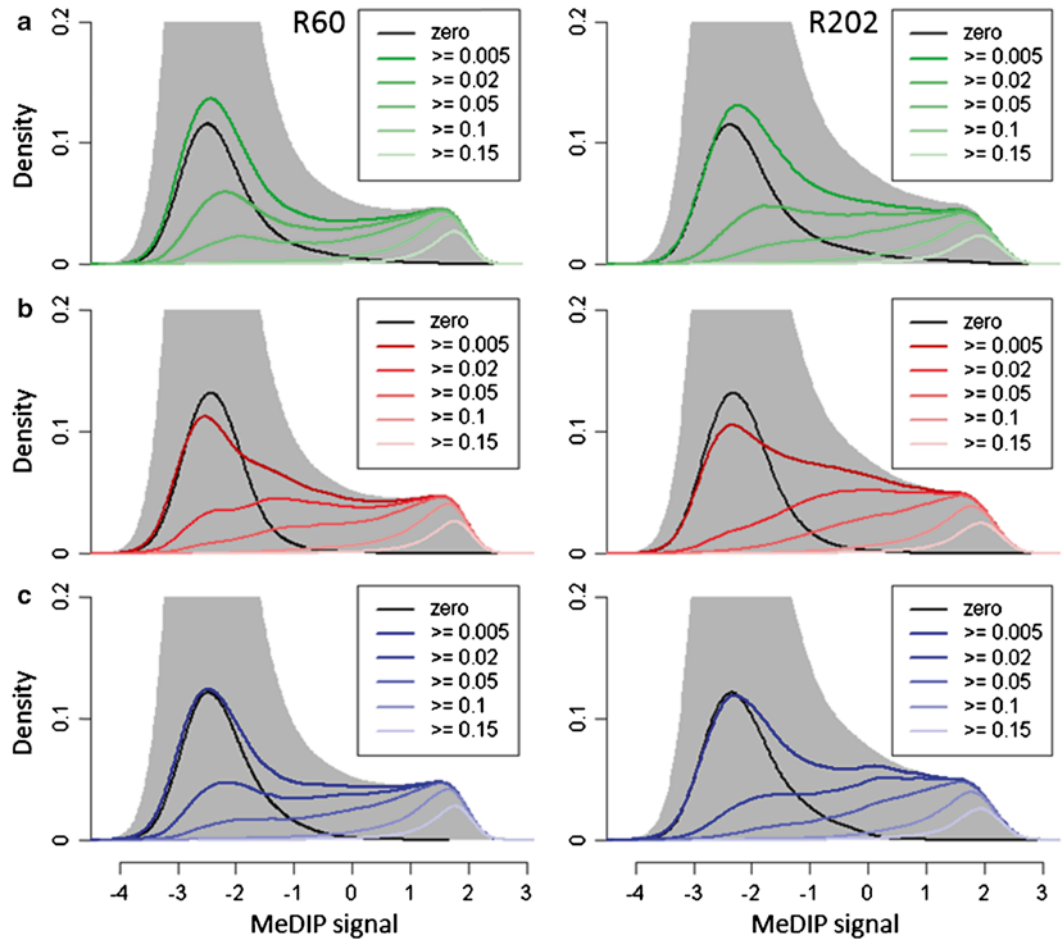

Fig. 6 Overlap unmethylated and methylated probe signal distributions. Shown are the MeDIP distributions that correspond to a certain WGBS-seq signal range. It becomes clear that there is a significant overlap between the MeDIP signal distributions with a low WGBS-seq signal (unmethylated; *black* is WGBS-seq signal of zero) and those with a high WGBS-seq signal (methylated). The G/R data (**a**) tends to have a higher overlap compared to the R/G data (**b**). The DS data (dye-swap; **c**) shows an average overlap result. It also shows a higher overlap compared to the R/G data

from unmethylated probes overlap with those from methylated probes (Fig. 6). In this case, the task is to find an informative cut-off that would minimize both false-positive and false-negative methylation calls.

3.2.1 Defining the "Gold Standard"

To illustrate this problem empirically we use the WGBS-seq signal to define high-confidence methylated probe windows based on a measurement error distribution. To achieve this, let the measurement error, y, of the jth probe window be given by

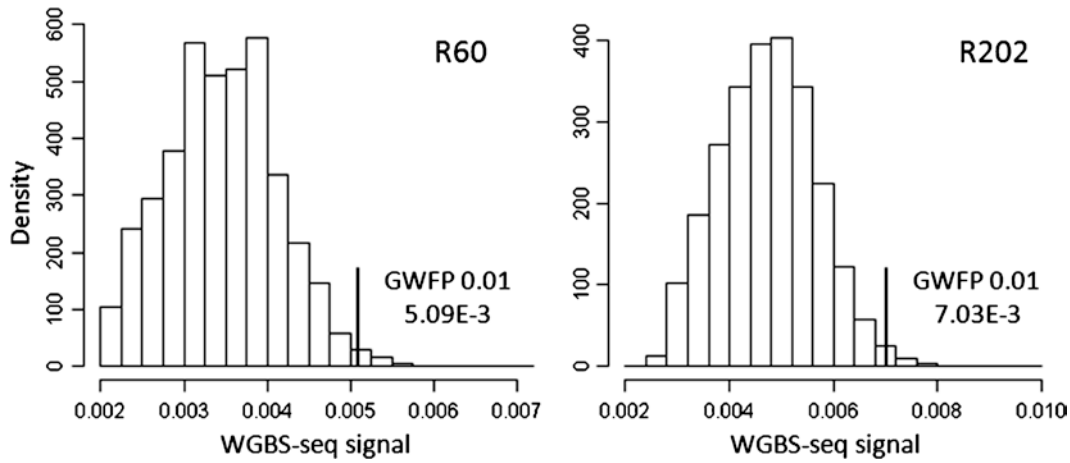

Fig. 7 Measurement error distributions. Shown are the measurement error distributions of both epiRILs and the signal cutoffs that correspond to a genome-wide false-positive rate of 0.01

$$y(j) = (1 - \mathrm{CR}) \times \frac{R_j}{R_{\max}},$$

where R_j is the number of cytosines in the jth probe window, $R_{\max}$ the maximum number of cytosines across all windows, and CR is the overall conversion rate. Furthermore, let us define the empirical density distribution of the error as $f(y)$ (Fig. 7). Using this distribution, the genome-wide false-positive (GWFP) rate can be calculated numerically using

$$\mathrm{GWFP} = 1 - \int_0^T f(y)\mathrm{d}y,$$

where T is the WGBS-seq signal threshold that is needed to meet a given GWFP level. We find that for GWFP = 0.01, the WGBS-seq threshold is approximately 5.09E−3 and 7.03E−3 for R60 and R202, respectively. Hence, we define probe window j as methylated if the WGBS-seq signal of that region is larger than the threshold T. At this threshold level, we find that 33.0 % (N = 182,330) and 33.3 % (N = 183,150) of all probe windows (genome-wide) can be confidently called methylated with this technology in R60 and R220, respectively (Table 1).

3.2.2 MeDIP Signal Classification Based on a Naïve Classifier

We use the WGBS-seq-derived classification to assess the problem of determining the methylation status of probes in the context of MeDIP-chip data. We first consider a naïve classifier which consists of a single MeDIP cutoff. According to this classifier, a probe is considered methylated if its signal exceeds the cutoff and as unmethylated if its signal falls below it. Comparing the

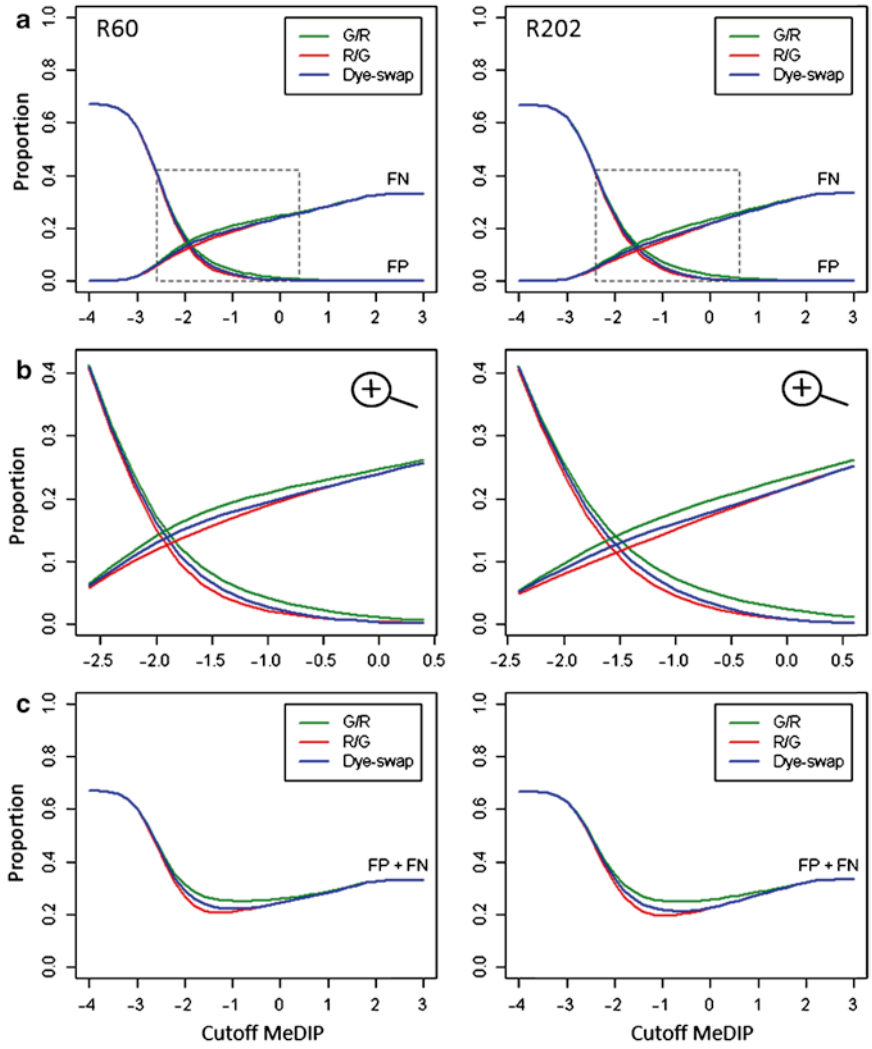

Fig. 8 False-positive and false-negative rates using a naïve classifier. Shown are the proportions of false-positive and false-negative probe classifications for different classification cutoffs of the MeDIP data (**a**) and (**b**) and the sum of the two (**c**). The G/R data shows a substantial higher proportion of FP and FN compared to the R/G data

resulting calls to those obtained from the WGBS-seq classification (Subheading 3.2.1) allows us to define the MeDIP false-positive and false-negative rates associated with the naïve classifier. Figure 8a, b shows the distribution of false-positive and false-negative rates for series of cutoffs across the entire MeDIP signal range (−4 till 3 with step size 0.2). This analysis shows that there is a considerable trade-off between minimizing false positives and false negatives while maximizing the total number of regions detected as methylated. We find that the optimal cutoff corresponds to an MeDIP signal of about−1.5 (Fig. 8), which yields a

Table 2
Number of false-positive and false-negative probe classification

	R60			R202		
	FP	FN	FP + FN	FP	FN	FP + FN
G/R—naïve cutoff	17,973	120,335	138,308 (0.251)	25,945	111,320	137,265 (0.249)
G/R—HMM	12,796	114,197	126,993 (0.230)	12,385	110,766	123,151 (0.224)
R/G—naïve cutoff	16,763	97,826	114,589 (0.208)	24,932	83,325	108,257 (0.197)
R/G—HMM	11,564	108,238	119,802 (0.217)	13,985	101,413	115,398 (0.210)
DS—naïve cutoff	14,901	107,635	122,536 (0.222)	15,852	101,227	117,079 (0.213)
DS—HMM	12,905	107,769	120,674 (0.219)	15,179	100,496	115,675 (0.210)

Reported are the number of false-positive and false-negative probe classifications using either a naïve cutoff or a Hidden Markov Model. Numbers of the naïve cutoff are based on the most optimal cutoff in Fig. 8c (smallest number of FP + FN)

combined false-positive and false-negative rate of about 0.23. Consistent with the dye-effects illustrated in our assessment of the dynamic range (Subheading 3.1), the combined false-positive and false-negative rates show substantial dye-dependence, with genome-wide rates of about 0.25, 0.20, and 0.22 for the G/R, R/G, and the DS data, respectively (Fig. 8; Table 2). Hence, the G/R data yields the highest misclassification rate which is likely caused by its high signal variance. Again, the dye-swap does not correct this problem (Fig. 8). *See* **Note 2** for recommendations concerning the labeling of the IP and input DNA.

3.2.3 MeDIP Signal Classification Based on Hidden Markov Model

This dye bias can be partially alleviated if one considers, instead of the above naïve classifier, a more sophisticated statistical classification approach based on HMMs (Fig. 9, Table 2). We have recently proposed an HMM for the analysis of MeDIP-chip [15]. This model has been shown to outperform alternative methods in terms of speed, sensitivity, and specificity [28]. An important aspect of this model is, as with HMMs in general, that it borrows signal information from immediately surrounding probes, and therefore significantly reduces measurement noise. This leads to a more robust inference of the underlying methylation status of a given probe window and makes methylation analysis less susceptible to dye effects. Figure 9 illustrates this point clearly; it shows that the HMM analysis of the R/G, G/R, and DS data results in much smaller misclassification differences between these data sets in terms of overall false-positive and false-negative rates (about 0.23, 0.21, and 0.21 for the G/R, R/G, and the DS data, respectively; Table 2). *See* **Note 3** for recommendations concerning the

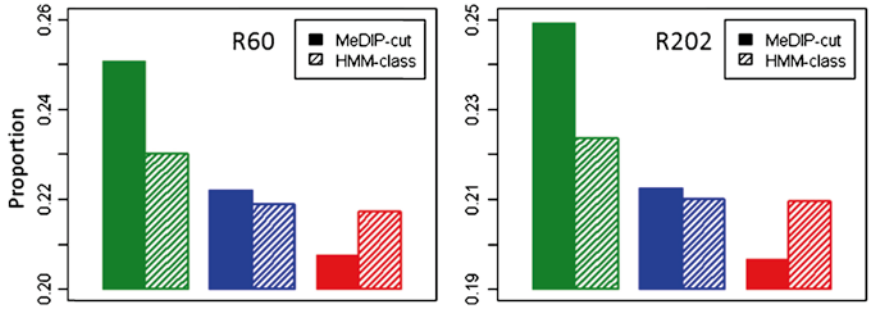

Fig. 9 Comparison performance of a naïve classifier and a Hidden Markov Model. Shown are the proportion of misclassified probes (FP + FN) obtained using the most optimal MeDIP classification cutoff (MeDIP-cut; smallest number in Fig. 8c) or the HMM classification (HMM-class). The color of each bar corresponds to the three data sets (*green*: G/R data; *red*: R/G data; *blue*: DS data)

analysis of the data. Nonetheless, despite this improvement, dye-related differences, particularly in the G/R data, do persist and continue to affect our ability to infer the correct methylation status of a given genomic region (Fig. 9). It is therefore of interest to identify and characterize the sources of this bias in the MeDIP-chip data.

3.3 Dye Bias in MeDIP-Chip Is Associated with Low Methylation Levels and CG Content

To explore the source of the observed dye bias, we start by plotting the two dye combinations (G/R and G/R) in a scatter plot (Fig. 10a). We find that there is a subset of the probes that shows a relatively higher signal for the G/R data compared to the R/G data. Inspection of the WGBS-seq signal corresponding to these probes indicates that the methylation level of these probe windows should be low (high density of probes around zero). This expectation is indeed reflected in the R/G signal (Fig. 10c), but not in the G/R signal which seems to be vastly exaggerated (Fig. 10b). Since the dye-swap signal yields only an average of the R/G and G/R data it cannot correct this bias (Fig. 10d).

Annotation analysis of this subset of probes shows that they contain a high proportion of transposon and intergenic sequences relative to genic sequences (Fig. 11a). In *Arabidopsis*, it is well known that genes have a higher CG percentage compared to transposons (Fig. 12). This raises the question whether CG content may be a key contributor to the observed dye bias.

In order to explore this possibility more generally, we calculated the CG content of the probe window for each probe on the tiling array and examined its relationship with signal intensity in the G/R, R/G, and DS data sets. For clarity we restricted our analysis to probes that were unmethylated according to WGBS-seq (*see* Table 1). In this way we could rule out any trends arising from differences in methylation levels. Our analysis shows that the signal intensity of unmethylated regions in the G/R data is subject

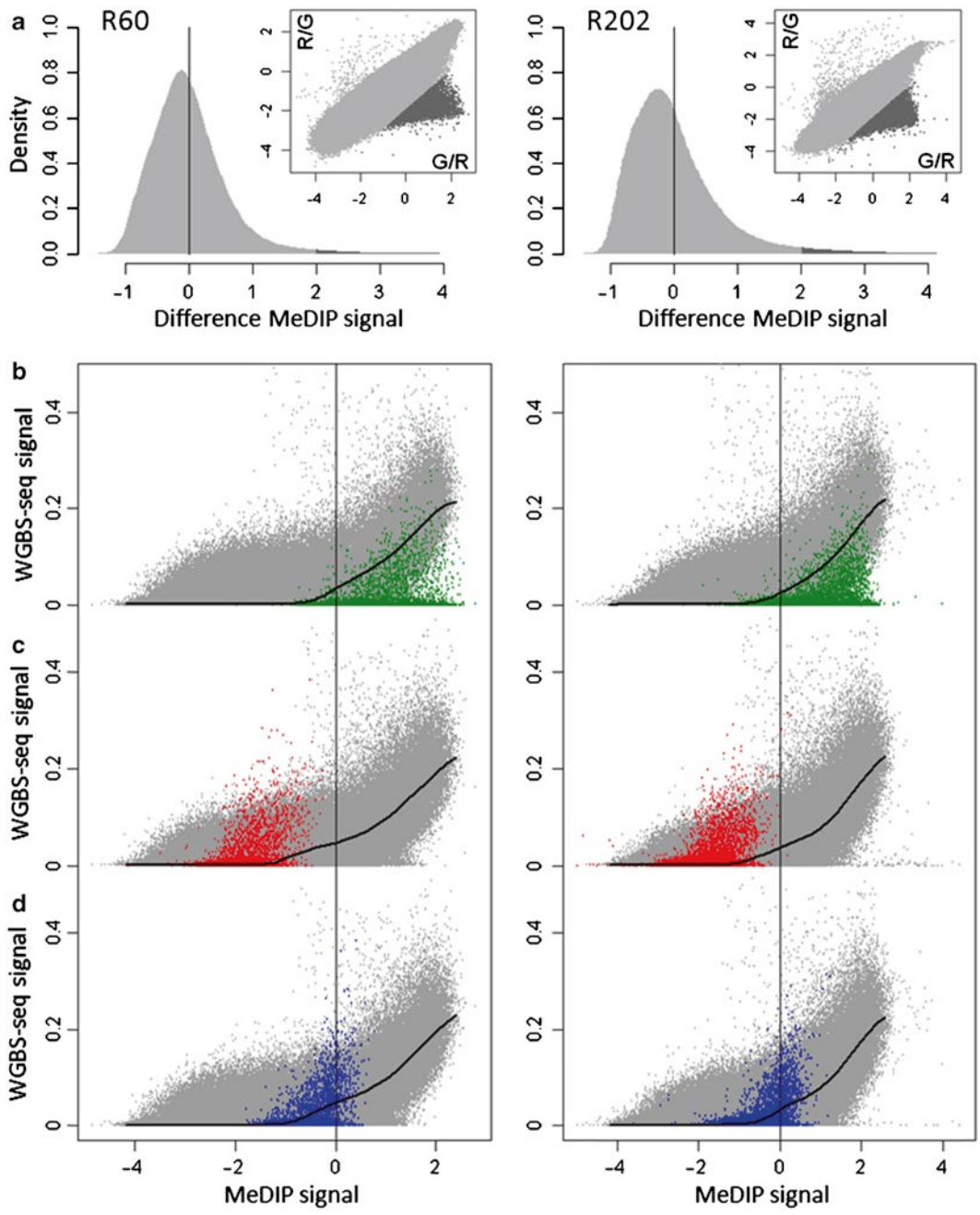

Fig. 10 Inspection of non-correlating probes using WGBS-seq signal data. (**a**) Shown is a density plot of the difference of both dye combinations. The inset shows a scatter plot of both dye combinations. Probes with a signal difference higher than two are indicated with dark gray. (**b**)–(**d**) Scatter plots of WGBS-seq data (*y*-axis) and each of the three MeDIP data sets (*x*-axis). The non-correlating probes (*dark gray* in panel **a**) are highlighted according to the color assigned to each data set (*green*: G/R data; *red*: R/G data; *blue*: DS data). The *black line* shows the median WGBS-seq signal for sliding windows along the entire MeDIP signal range

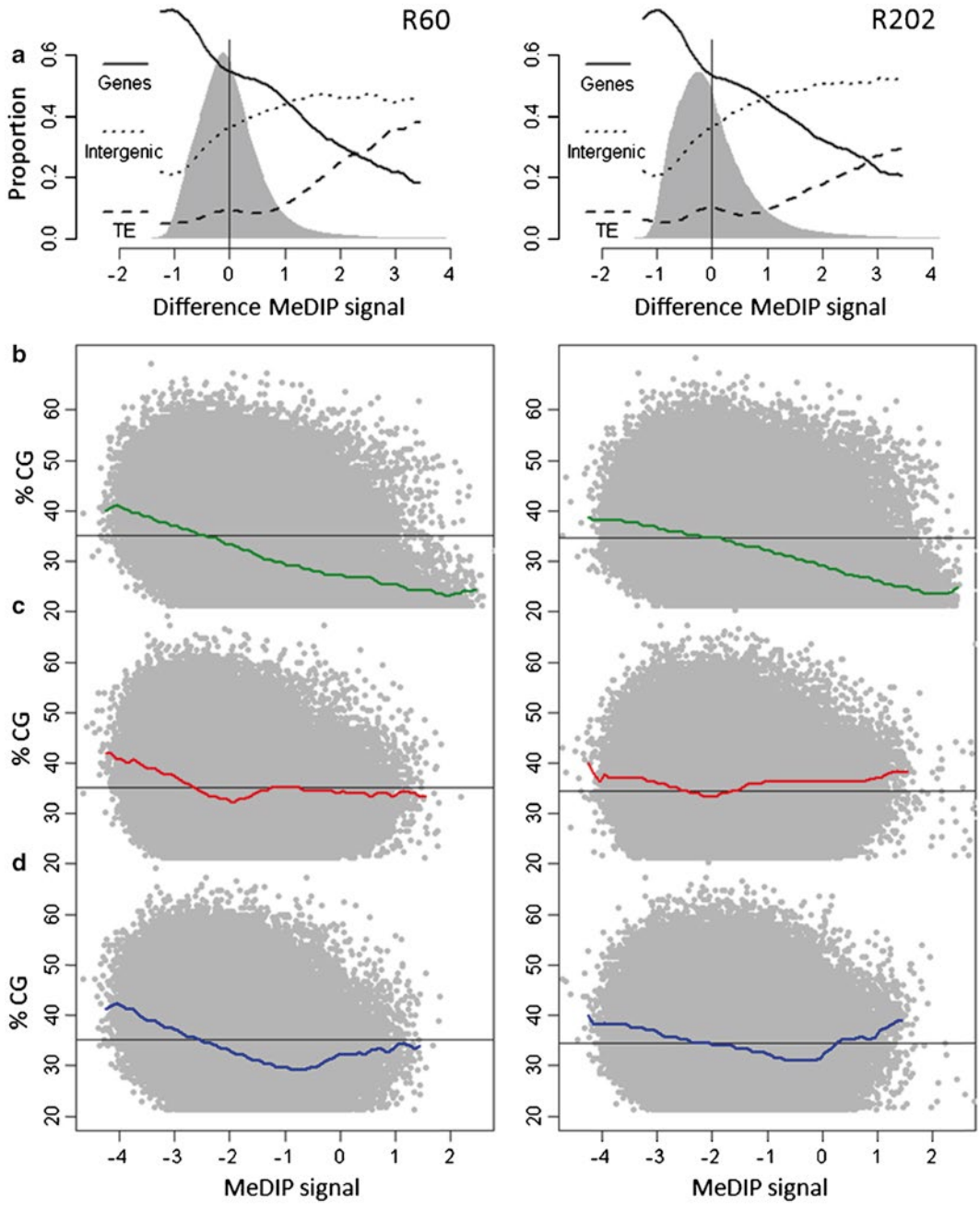

Fig. 11 Annotation analysis and CG bias of unmethylated probes. (**a**) Shown is the proportion of genic, transposable element (TE) and intergenic probes along the entire MeDIP difference range (G/R–R/G). (**b**)–(**d**) Scatter plots of the CG percentage of each probe window (*y*-axis) and the MeDIP signal of each of the three MeDIP data sets (*x*-axis). The *trend lines* show a clear negative relationship between CG content and signal intensity. The color of each line corresponds to the three data sets (*green*: G/R data; *red*: R/G data; *blue*: DS data)

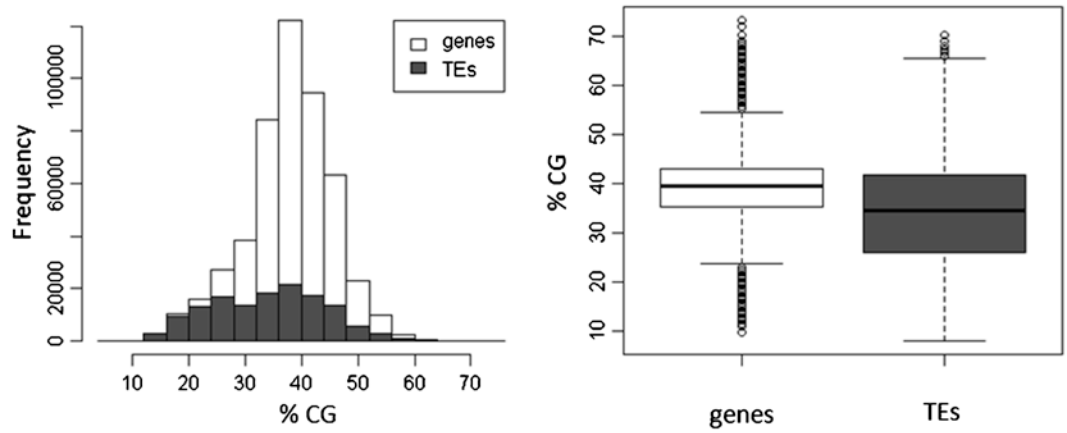

Fig. 12 CG content of genes and transposons. Shown are the CG content distributions of genes and transposable elements (TEs)

to strong dye biases (Fig. 11b). We find a clear negative linear relationship between CG content and signal intensity; that is, signal intensity is highest for probes with low CG content and lowest for probes with high CG content. By contrast, CG content appears to have little influence on the signal intensity in the R/G data (Fig. 11c), and the DS displays intermediate levels of CG bias (Fig. 11d). *See* **Note 2** for recommendations concerning the labeling of the IP and input DNA.

Royce et al. [30] considered normalizing tiling array signals for the CG content of probes. While this procedure may work for some applications, such as transcription factor binding data (ChIP-chip), its application to gene expression tiling arrays has been shown to lead to overnormalization and hence to a loss of signal information [31]. Overnormalization is expected to be even more drastic in MeDIP-chip data where CG content is correlated with DNA methylation levels. Correcting for CG content, in this case, will reduce signal intensities arising from probes with true positive methylation measurements. A simple solution to bypass these issues is to work exclusively with R/G data where CG bias appears to be minimal (Fig. 11c, *see* **Note 2**).

4 Concluding Remark

Although the mapping resolution of MeDIP-chip (~165 bps) is much lower than the single cytosine measurements that can be achieved with WGBS-seq, this array-based technology provides a level of resolution that should be sufficient for the detection of most functionally important differentially methylated regions. MeDIP-chip requires fewer bioinformatic resources and therefore

scales more easily to large samples. Provided several experimental and data preparation steps are followed (*see* **Note 1–3**), MeDIP-chip presents a viable alternative to WGBS-seq in future population epigenetic studies.

5 Notes

1. *Recommendations for data preparation*: Prior to MeDIP-chip analysis, potentially cross-hybridizing probes should be removed. They typically show signal intensities similar to probes that correspond to genomic regions with high methylation levels (Fig. 3). Failure to remove cross-hybridizing probes can therefore result in the detection of a large number of false positives. However, removal of these probes will result in loss of measurement coverage; but this drawback is no different from sequencing-based approaches where short reads that do not map uniquely are usually excluded.
2. *Recommendations for dye-labeling*: Dye-related biases can pose serious concerns in dual-channel MeDIP-chip. Labeling the immunoprecipitate (IP) DNA with Cy3 (green) and the control DNA (input) with Cy5 (red) (i.e., G/R data) introduces strong signal distortions that significantly compromise measurement accuracies. These biases are particularly pronounced in genomic regions with low methylation and low CG content. This signal bias disappears when the opposite labeling strategy is employed (labeling IP with Cy5 and input with Cy3, i.e., R/G data). As a result of the G/R dye bias, dye-swap experiments in MeDIP-chip always perform worse than the R/G data alone, despite the fact that DS consists of twice as much data. Hence, despite its routine use in expression micro-array studies, we do not recommend the use of dye-swaps in dual-channel MeDIP-chip. This means that experimental costs can be reduced by a factor of two without loss of measurement information.
3. *Recommendations for data analysis*: The classification of probes as methylated or unmethylated requires a sound statistical approach. The best methods for MeDIP-chip are variants of HMMs [28]. The assumptions of HMMs are fundamentally consistent with the data properties arising from MeDIP-chip experiments. These assumptions are the following: (1) A probe signal is a noisy proxy for an underlying (unobserved) methylation state and (2) methylation states are spatially correlated along the genome owing to the array design and the propensity of DNA methylation to occur in clusters. Our application of a recent HMM designed for *Arabidopsis* MeDIP-chip [15] resulted in a genome-wide false-positive rate of about 0.02 and

false-negative rate of about 0.19 (a combined rate of 0.21) for the R/G data. This relatively high false-negative rate implies that the application of this HMM misses regions with low methylation levels. Less customized methods may yield even higher misclassifications. In population studies, this limitation will restrict the calling of DMRs to clear methylation differences between individuals (e.g., no methylation versus high methylation) and will likely fail to detect more subtle DMRs (e.g., no methylation versus low methylation). One way to improve this situation is to consider at least one additional technical R/G replicate.

Acknowledgements

This work was supported by grants from the Netherlands Organization for Scientific Research (NWO) (to F.J. and M.C.-T) and the Netherlands Bioinformatics Centre (NBIC) (to R.W.). Work in the Colot lab is supported in part by the European Union Network of Excellence EpigeneSys.

References

1. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11:204–220
2. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet 11:191–203
3. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell 126:1189–1201
4. Zilberman D, Gehring M, Tran RK et al (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 39:61–69
5. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523–536
6. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD et al (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 452:215–219
7. Schmitz RJ, Schultz MD, Lewsey MG et al (2011) Transgenerational epigenetic instability is a source of novel methylation variants. Science 334:369–373
8. Becker C, Hagmann J, Müller J et al (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature 480:245–249
9. Boyes J, Bird A (1992) Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. EMBO J 11:327–333
10. Lorincz MC, Schübeler D, Hutchinson SR, Dickerson DR, Groudine M (2002) DNA methylation density influences the stability of an epigenetic imprint and Dnmt3a/b-independent de novo methylation. Mol Cell Biol 22:7572–7580
11. Johannes F, Porcher E, Teixeira F, Saliba-Colombani V, Simon M, Agier N et al (2009) Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet 5:e1000530
12. Colomé-Tatché M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A et al (2012) Features of the *Arabidopsis* recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc Natl Acad Sci USA 109:16240–16245
13. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA

methylation in normal and transformed human cells. Nat Genet 37:853–862
14. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E et al (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 26:779–785
15. Cortijo S, Wardenaar R, Colomé-Tatché M, Johannes F, Roudier F, Colot V (2012) Genome-wide analysis of DNA methylation in *Arabidopsis* using MeDIP-chip. Methods Mol Biol 15:2930–2939
16. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW et al (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci USA 89:1827–1831
17. Clark SJ, Harrison J, Paul CL, Frommer M (1994) High sensitivity mapping of methylated cytosines. Nucleic Acids Res 22:2990–2997
18. Lister R, Ecker JR (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res 19:959–966
19. Krueger F, Kreck B, Franke A, Andrews SR (2012) DNA methylome analysis using short bisulfite sequencing data. Nat Methods 9:145–151
20. Chen PY, Cokus SJ, Pellegrini M (2010) BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics 11:203
21. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193
22. R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, ISBN: 3-900051-07-0. http://www.R-project.org
23. Dobbin KK, Kawasaki ES, Petersen DW, Simon RM (2005) Characterizing dye bias in microarray experiments. Bioinformatics 21:2430–2437
24. Dombkowski AA, Thibodeau BJ, Starcevic SL, Novak RF (2004) Gene-specific dye bias in microarray reference designs. FEBS Lett 560:120–124
25. Martin-Magniette ML, Mary-Huard T, Bérard C, Robin S (2008) ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. Bioinformatics 24:i181–i186
26. Andrews S (2007) ChIPmonk: software for viewing and analysing ChIP-on-chip data. BMC Syst Biol 1(Suppl 1):P80
27. Johannes F, Wardenaar R, Colomé-Tatché M, Mousson F, de Graaf P, Mokry M et al (2010) Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. Bioinformatics 26:1000–1006
28. Seifert M, Cortijo S, Colomé-Tatché M, Johannes F, Roudier F, Colot V (2012) MeDIP-HMM: genome-wide identification of distinct DNA methylation states from high-density tiling arrays. Bioinformatics 28:2930–2939
29. Li W, Meyer CA, Liu XS (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. Bioinformatics 21(Suppl 1):i274–i282
30. Royce TE, Rozowsky JS, Gerstein MB (2007) Assessing the need for sequence-based normalization in tiling microarray experiments. Bioinformatics 23:988–997
31. Gilbert D, Rechtsteiner A (2009) Comments on sequence normalization of tiling array expression. Bioinformatics 25:2171–2173