13 DNA Methylation Arrays: Methods and Analysis

N.P. Thorne, J.C. Marioni, V. Rakyan, A.E.K. Ibrahim, C. Massie, C. Curtis, J.D. Brenton, A. Murrell, and S. Tavaré

CONTENTS

13.1 INTRODUCTION

Over the last decade, microarrays have become a fundamental tool in biological research laboratories throughout the world. During this time, methods for performing microarray experiments have improved and expanded rapidly, creating an enormous demand for evaluation and comparison of emerging and existing technologies. Importantly, the responsibility for doing this lies as much with the data analyst as the data generator. Such evaluations are difficult since they are influenced by many factors, both financial and scientific. They require a good understanding of both the biological underpinnings of new array technologies and their applications, as well as the statistical issues involved when analyzing the resulting data. To date, there have been many empirical comparisons of technologies for expression array profiling, but newer applications are still lagging in this respect. With the recent growth in interest in applying microarrays to study a different aspect of the genome, namely the epigenome, this problem has again come to the fore. While there are many publications exploring the biology of DNA methylation and the epigenome, and a large number of articles describing the development of approaches for studying DNA methylation, there are few articles that address the analytic issues involved in these new experiments. This chapter aims to address this problem. It is aimed at the biologist who wants to understand the limitations in analyzing data obtained from different DNA methylation arrays, and the computational biologist wanting an entry point into this new and exciting area.

13.2 MAMMALIAN DNA METHYLATION

Mammalian DNA methylation describes a chemical modification that predominantly affects the **Q1** cytosine base of CG dinucleotides (Figure 13.1a) [1,2], commonly represented as CpG (the p indicates the phosphodiester bond that forms the backbone of the DNA strand). A CpG found on the sense strand of the DNA duplex will have a CpG in the reverse sense on the opposite strand (Figure 13.1b). DNA methylation of a CpG covalently adds a methyl group to the 5th carbon position on the cytosine base. In lower organisms, such as plants and *Escherichia coli*, methylation can also target other bases, including adenine [3]. CpGs are statistically underrepresented in the human genome [4] and are associated with repetitive DNA sequences including centromeric repeats, retroviral elements, and retrotransposons [5,6].

Methylation in promoter regions and other regulatory sequences can prevent transcription and these regions are often heavily methylated, suggesting that CpG methylation may have evolved as a defense mechanism to silence viral DNA [2]. However, CpG-rich sequences in actively transcribed gene-rich regions are mostly unmethylated and resistant to changes in methylation [6–8]. By convention, these regions are known as CpG islands and are often associated with gene promoters and regulatory regions [6,9]. CpG islands are defined by criteria including the length of the region, GC content, and CpG density [4,10].

Methylation patterns in the genome can be maintained through cell division and replication [2,11]. However, methylation may be dynamic, and the pattern and density of methylation in areas of active transcription may change during development to control key genes in a temporal or tissue-specific manner [2]. Regions that are normally methylated and become less methylated are referred to as hypomethylated and those that become methylated are called hypermethylated. The regulation of DNA methylation is closely associated with other covalent modifications of the histone proteins on which DNA is assembled to form chromatin [2]. These protein modifications include acetylation

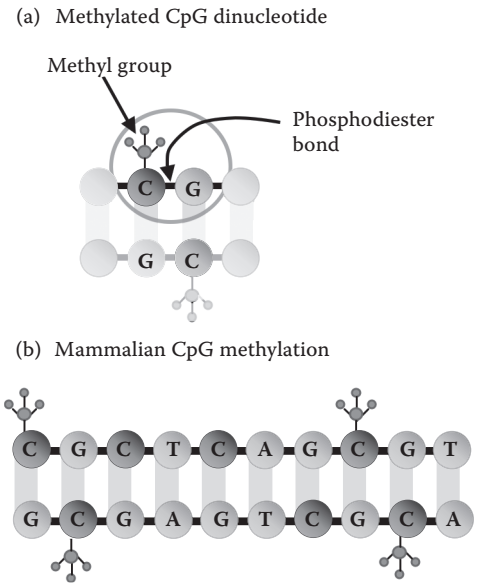


FIGURE 13.1 Illustration of (a) a methylated CpG dinucleotide. The cytosine and guanine bases are joined by a phosphodiester bond and a methyl group has been added to the cytosine. (b) gives a detailed illustration of double-stranded DNA with methylated CpGs in positive and negative strands.

(transcriptionally activating) and methylation (transcriptionally repressive). It remains unclear whether DNA methylation is a consequence or a cause of histone modification [2,11]. As there are known mechanisms for maintenance of methylated DNA during replication, it is plausible that histone methylation is maintained secondarily to DNA methylation [2].

Another generally accepted notion of DNA methylation is that it will spread locally [12,13] that is, once a region starts to become methylated, all CpGs within the region will become methylated (Figure 13.2a). This is consistent with the concept of a CpG island with boundaries defined by some signal in the DNA sequence. Using this principle, regions where a minority of CpGs are methylated would be called unmethylated since the region's methylation status is judged as a whole (see Figure 13.2b). Despite this, it is possible that small blocks of methylated (or unmethylated) regions may exist within a given CpG island. Furthermore, it is believed that certain CpGs within a region may be more important than others (i.e., some CpGs may be held under tighter evolutionary control [4]). Indeed, because mutational repair of methylated CpGs is harder than for nonmethylated CpGs, CpGs would tend to be lost through selection. This perhaps explains the lower than expected number of CpGs found throughout the mammalian genome.

Methylation of CpG islands within the promoters and body of a gene can lead to transcriptional silencing, while a lack of methylation may permit active transcription of the associated gene [14]. This regulation of gene expression is thought to occur as a result of conformational changes in the chromatin structure, altered binding capacity of transcription factors to methylated motifs in the promoter, and other effects altering regulatory elements such as enhancer and repressor sites [15].

A historical role for epigenetics has been in cancer research, especially in the search for abnormally hypomethylated oncogenes or hypermethylated tumor suppressor genes (i.e., genes promoting cancer that have become activated through hypomethylation, and genes suppressing cancer that have been deactivated through hypermethylation) [11,16,17]. Most CpG islands are usually unmethylated but, in cancer, promoter-associated CpG islands of certain genes can be hypermethylated [19]. Many of these hypermethylated genes are specific to certain cancers, suggesting that their aberrant methylation may be important [11,18]. Consequently, understanding the epigenome will

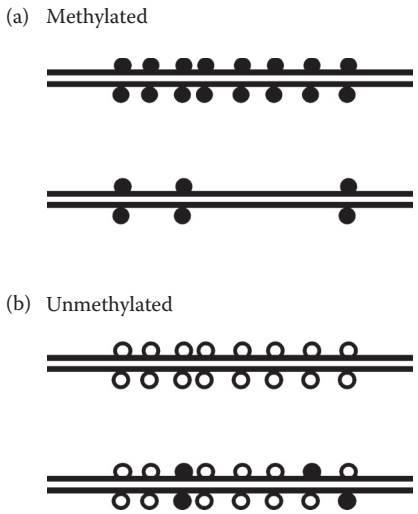


FIGURE 13.2 Two double-stranded DNA fragments that are both (a) methylated and (b) unmethylated. Filled circles represent methylated CpGs, whereas open circles represent unmethylated CpG sites. Two hemimethylated CpG sites are shown in the lower fragment in (b).

allow the control mechanisms of gene transcription to be better modeled, and, as a result, this area of research is growing rapidly. Moreover, DNA methylation is a potentially reversible modification and demethylating chemotherapies are being developed and considered for cancer treatment [6,11]. Further, sites of differential methylation between cell types are not limited to promoter regions. They have been found in exons, introns, enhancer sites, and intergenic regions—suggesting they might regulate miRNAs, reverse strand transcripts, or alternative splicing [2].

Unlike aberrant hypermethylation, cancer-related hypomethylation occurs on a more global scale [11]. Seemingly, indiscriminant hypomethylation occurs throughout the genome of cancer cells, affecting vast amounts of non-CpG island DNA that is normally methylated. Moreover, pervasive hypomethylation is also found in premalignant neoplastic cells, implying that epigenetic changes may constitute the earliest steps of tumorigenesis [19]. While the mechanism and role of global hypomethylation is not well understood, it is observed widely and undoubtedly plays a role in cancer initiation [2]. One hypothesis proposes that it unlocks normally silent repetitive elements, activating transposons that promote genomic rearrangements or interfere with normal transcriptional regulation in the tissue [20].

Of course, to study epigenetic changes in cancer, a basic understanding of DNA methylation in normal tissue is also required. It had been assumed that DNA methylation played a key role in gene switching events during development, but this view is currently being questioned and many classical notions of the epigenome are being scrutinized. Much of the controversy stems from inconsistent results from experiments involving knockout mouse models of genes that encode DNA methylation maintenance proteins [2]. Problems arise since changes in growth conditions and other environmental factors can directly alter epigenetic states [21,22]; experiments with cell lines have also encountered this limitation. Therefore, studies investigating epigenetic mechanisms must be highly controlled, carefully planned, and cautiously interpreted.

13.2.1 Measuring DNA Methylation

The ability to measure the extent of methylation at every CpG would mostly improve our understanding of the effects of DNA methylation. However, such precise measurements are currently

possible only with low-throughput technologies and are therefore limited to small portions of the genome. Despite this, these methods have been employed extensively in molecular biology research and DNA methylation patterns associated with some genes (typically candidate genes identified by other studies) have been studied extensively.

In recent years, efforts have been made to develop high-throughput, whole-genome approaches for measuring DNA methylation [23]. These have emerged in light of the continued evolution of microarray technologies for expression [24], copy number [25], SNP, and ChIP profiling [33]. Current approaches for DNA methylation arrays rely on one of the following principles: **Q2**

- *Enrichment*: Beginning with fragmented genomic DNA, the first approach enriches or separates fragments that are methylated from those that are not; one or both fractions are then hybridized to an array. Methods for enriching methylated sequences typically employ either methylation-sensitive restriction enzyme digestion or methyl-cytosine antibody precipitation.
- *Bisulfite conversion*: The second approach is much like SNP detection with microarrays. Probes are designed to discriminate target sequences containing methylated CpGs from those with unmethylated CpGs. This discrimination is possible because of the base conversion of unmethylated cytosine to uracil that occurs after bisulfite treatment of the DNA.

There are many variations on both approaches, all with specific advantages and disadvantages. Significant limitations of each approach are related to the quality and type of arrays used, including probe design and density across the genome. The data obtained inherit all biases, sources of variability, and limitations associated with a given approach. Since technologies and approaches for measuring DNA methylation are varied and still evolving, there is no consensus for statistical analyses. However, common themes arise, such as normalization issues, the effect of CpG and GC content of probes and genomic regions of interest, amplification biases, and enzyme and enrichment method efficiencies. Finally, we note that short-read resequencing of bisulfite-treated DNA provides another approach to measuring DNA methylation on a genome-wide scale (e.g., [26]).

13.2.2 Chapter Aims

Lately, the field of epigenetics has been growing at a phenomenal rate due principally to advances in technology that are enabling high-resolution, high-throughput quantitation of DNA methylation; the rest of this chapter reviews the microarray approaches that have been adapted for this purpose. Emphasis is given to the design of array-based DNA methylation experiments and their ability to answer different types of epigenetic questions, as well as the normalization and analysis considerations involved. The chapter will consider the main approaches to array-based DNA methylation assays including the platform, array, and probe design choices for each. Finally, a review of methods for validating array-based results will be given.

13.3 EXPERIMENTAL DESIGN CONSIDERATIONS

When designing microarray experiments, many factors need to be considered to ensure that the biological question of interest has the greatest chance of being answered. For DNA methylation arrays, the most important factors to acknowledge are the limitations a given approach has on the user's ability to answer this question.

13.3.1 What Samples to Hybridize?

DNA methylation array experiments are typically 2-color hybridizations. However, while design issues for two-color array-based experiments performed using other technologies typically revolve around

which mRNA or DNA samples to compare on an array, DNA methylation experiments are much more involved. The first decision to be made is the number of samples to hybridize to an array.

For a given sample, three fractions can be obtained (Figure 13.3a). The first is simply the non-enriched *input* sample, the second is the fraction enriched for methylated sequences, and the third is the fraction enriched for unmethylated sequences. Methods of enriching for methylated or unmethylated sequences are described in Section 13.4.

For within-array comparisons between samples (Figure 13.3b), either (i) methylated fractions or (ii) unmethylated fractions can be compared (*direct comparison*). Alternatively (Figure 13.3c), methylated or unmethylated fractions from two samples may each be compared to a common reference fraction (*indirect comparison*). This design choice is used in the method of differential methylation hybridization (DMH) (see Section 13.4 [27–32]). Two limitations of this approach are that the methylation status of the reference sample is usually unknown and, when comparing methylation between samples, it is often hard to find an appropriate common reference. This makes interpretation difficult, particularly when the extent, rather than simply the direction, of change in methylation is of interest.

Single-sample approaches (Figure 13.3d) comparing the (i) methylated or (ii) unmethylated enriched fraction to the unenriched input fraction avoid the problems associated with two-sample approaches [33,34]. Additionally, the log-ratio data obtained from such experiments are potentially easier to interpret. However, it has been shown that methylated and unmethylated sequences are not equally detectable for this design (Figure 13.4b)—the dynamic range of the log-ratios is restricted, and (theoretically) positive values are not possible since there is no enrichment for both methylated and unmethylated fractions in the hybridization.

In contrast, when the methylated and unmethylated fractions from a single sample are compared within an array (Figure 13.3e), a wider range of log-ratio values is possible. In the MA-plot shown in Figure 13.4a, methylated sequences have positive (log-ratio, or M) values and unmethylated sequences have negative values. Sequences with values close to zero have ambiguous methylation status, something that can occur for a number of reasons: they may only be methylated in some of the sample (possibly due to tissue heterogeneity) or may not be fully methylated, and so can be enriched in both fractions hybridized to the array.

A limitation of this approach is that by enriching methylated and unmethylated fractions from a single sample, different systematic errors may be introduced in each channel. Such errors are difficult to identify and hard to account for in the analysis step since they are typically confounded with dye-biases and real methylation differences. Enzyme approaches are particularly susceptible to this since differing enzyme efficiencies occur. Additionally, the presence and frequency of enzyme recognition sites varies between sequences, which can introduce biases into the enrichment process. Thus, it is important to use bioinformatic methods to predict sequences that may be subject to such bias (i.e., those with few or no restriction sites for a given digestion enzyme) and toaccount for this in the analysis step.

13.3.2 Other Experimental Design Issues

In addition to the classical statistical notions that have been applied to microarray experimental design [35,36], appropriate experimental planning is equally critical to ensure that high-quality experiments are achieved, designs are robust, and all aspects of the experiment are meticulously recorded for quality assessment purposes. This is particularly important for DNA methylation arrays since they tend to be more complicated (i.e., involving multiple digestion or enrichment steps, purification, and amplification), which can lead to the introduction of systematic errors. Moreover, despite their high-throughput status, microarray experiments are expensive and time-consuming—consequently, these errors can be extremely costly. Another important factor to consider relates to the acquisition of samples. In particular, since tissue samples can be extremely heterogeneous, the investigator must be aware that the resulting data are based on averaging over cells that might possess quite different levels of methylation.

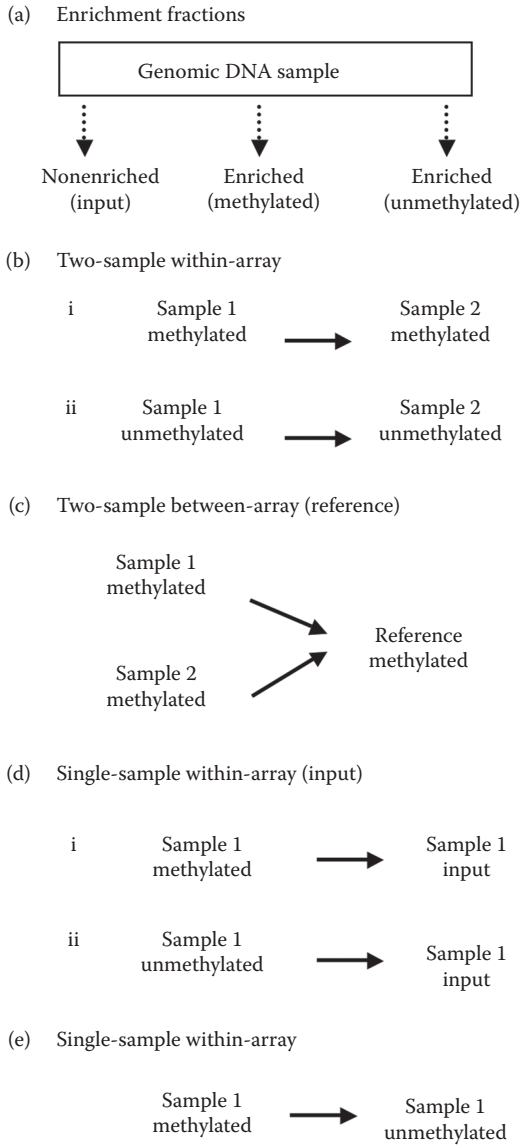


FIGURE 13.3 Fundamental design choices for two-color DNA methylation array experiments. Solid arrows indicate that target samples are cohybridized to an array. Shown in (a) are the three possible target fractions that can be obtained from a single genomic DNA sample; the input sample itself, a fraction enriched for methylated sequences from the input sample, or a fraction enriched for unmethylated sequences from the input sample. Given either the methylated or unmethylated fractions from two different samples, (b) illustrates the two-sample within-array designs. In (b), (i) compares methylated fractions from two samples and (ii) compares unmethylated fractions from two samples. These correspond to direct two-sample comparisons, whereas (c) is a two-sample between-array indirect comparison, commonly referred to as a reference design. Shown in (d) are single-sample within-array designs where the (i) methylated or (ii) unmethylated fraction is cohybridized to the input fraction from the same genomic sample. (e) shows the single-sample within-array design that directly compares the methylated and unmethylated fractions from a given sample.

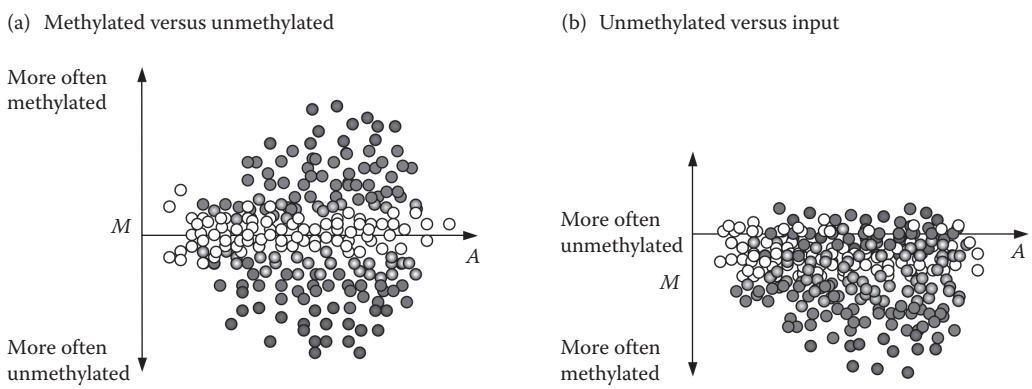


FIGURE 13.4 (See color insert following page XXX.) Illustrations of MA-plots for DNA methylation array data for experiments where (a) the methylated and unmethylated fractions from a single sample are cohybridized and (b) the unmethylated fraction from a single sample is cohybridized with the input (unenriched) fraction. *M* represents the log-ratio of the Cy5 and Cy3 channel intensities and *A* represents the log geometric mean of the Cy5 and Cy3 channel intensities. Dots represent *M* and *A* values for probes on a microarray. Methylation status is based on an average over the genomic DNA from all cells in the sample. Blue dots correspond to sequences that are mostly methylated in the sample, whereas red dots correspond to sequences that are mostly unmethylated in the sample. Unfilled circles correspond to sequences whose methylation status is ambiguous. Methylated and unmethylated sequences are more easily distinguished in (a) than (b). Theoretically, there should be no positive *M* values in (b).

13.4 DNA METHYLATION ARRAY APPROACHES

As described in Section 13.2.1, there are two main approaches for detecting DNA methylation using array technology: enrichment-based methods and schemes that rely on bisulfite conversion. The following sections give an overview of various methods that exploit these techniques to measure DNA methylation.

13.4.1 Restriction Endonuclease Enzymes

Restriction endonucleases cut double-stranded DNA by utilizing specific recognition sequences in the DNA (see Figure 13.5a–d for a description of the distinct cutting properties of different enzymes). Some restriction enzymes are methylation-sensitive: if their recognition sequence contains a CpG, methylation at this site can prevent endonuclease activity [37,38]. Similarly, other enzymes are methylation-dependent (Figure 13.5d). Additionally, many methylation-sensitive restriction enzymes (e.g., *HpaII* and *MspI*) also have nonsensitive isochizomers; these cut the same sequence and are therefore useful in control experiments. As a result of their ability to detect methylated or unmethylated DNA sequences, several low- and high-throughput techniques for assessing DNA methylation are based on methylation-sensitive restriction endonucleases.*

13.4.2 Methylation-Sensitive/Dependent Digestion with PCR Enrichment

A popular enzyme-based approach for DNA methylation arrays combines the following steps (Figure 13.6). The genomic DNA sample is first digested with a frequent cutting methylation-sensitive/dependent enzyme (Figure 13.6a). Since this enzyme's recognition site is small enough to be found regularly throughout the genome and does not contain a CpG, the resulting fragments are usually small enough to be subjected to polymerase chain reaction (PCR) amplification. The digested DNA

* A list of the canonical recognition sequences of methylation-sensitive enzymes is available at http://rebase.neb.com/rebase/rebms.html.

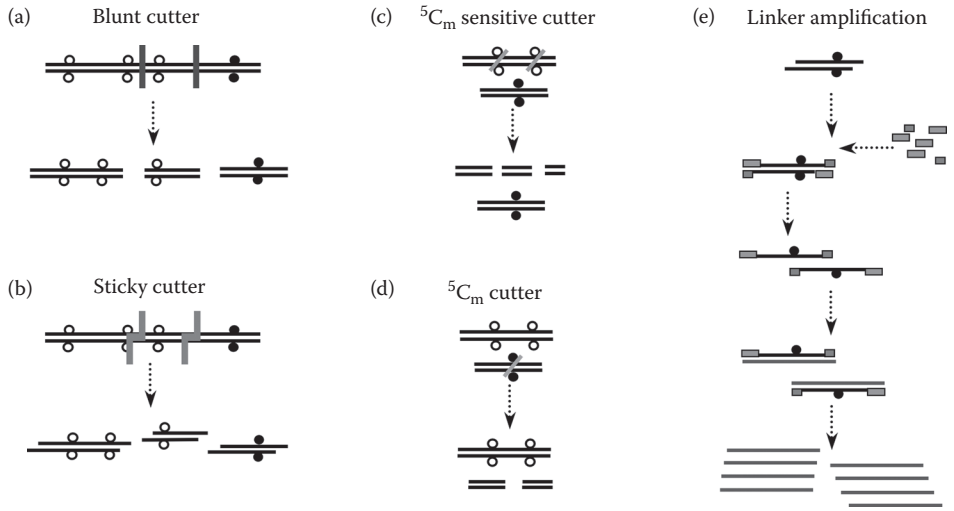


FIGURE 13.5 (See color insert following page XXX.) Restriction endonuclease enzymes cut double-stranded DNA at given recognition sites. These enzymes have different properties and activities, including (a) blunt cutting enzymes that leave no overhanging ends; (b) sticky cutters that leave overhanging ends; (c) methylation-sensitive enzymes (that only cut if their recognition site is unmethylated); and (d) methylation-dependent enzyme cutters. Linkers can be ligated to sticky or blunt ends and used to prime PCR amplification (e).

fragments are then ligated with linkers (Figure 13.5e); subsequent steps depend on the experiment's design (see Section 13.3).

If the design in Figure 13b(i) is employed, a methylation-sensitive enzyme is used and the process is repeated independently for a second sample (Figure 13.6b). During linker-mediated PCR the uncut fragments are amplified, leading to an enrichment of methylated sequences. The enriched fractions from each sample are then cohybridized to a microarray.

This method is generally known as DMH [27–32] and was developed to compare the methylation status of CpG islands in a test and reference sample. For the initial cutting step DMA uses *MseI* digestion; *MseI*'s recognition sequence (TTAA) is found frequently within bulk DNA, but rarely in CpG islands that therefore remain intact [10] after digestion. For the digestion step, DMH uses a combination of the methylation-sensitive enzymes, *BstUI*, *HhaI*, and *HpaII*. However, these enzymes are active under different conditions, which result in two separate digestion steps.

The next two methods start by splitting the fragmented sample into two. If the design described in Figure 13.3d is used, half the sample is set aside while the other half is digested using either a methylation-sensitive or methylation-dependent enzyme (Figures 13.6b,d or Figures 13.6c,d, respectively). The two fractions are then amplified separately using linker-mediated PCR before being cohybridized. This method is a modification of DMH that allows a sample's methylation status to be measured without using reference DNA [39]. Nouzova et al. [39] applied this method, using *MseI* for the initial digestion step, before creating the unmethylated fragment by digesting half the sample with *McrBC*, an enzyme that restricts methylated sequences and has a recognition sequence that is very frequent within CpG islands.

Finally, if the design in Figure 13.3e is used, one half of the sample is digested using a methylation-sensitive restriction enzyme while the other is digested with a methylation-dependent enzyme (Figures 13.6b,c). The two fractions are then amplified (using linker-mediated PCR) before being cohybridized. This approach was adopted by [33,34] and was motivated by a desire to compare methylated and unmethylated sequences within a single sample on an array. Like DMH and the method of Nouzova et al. [39], they begin by digesting genomic DNA with a frequent cutter, before digesting unmethylated sequences using a bioinformatically derived set of methylation-sensitive

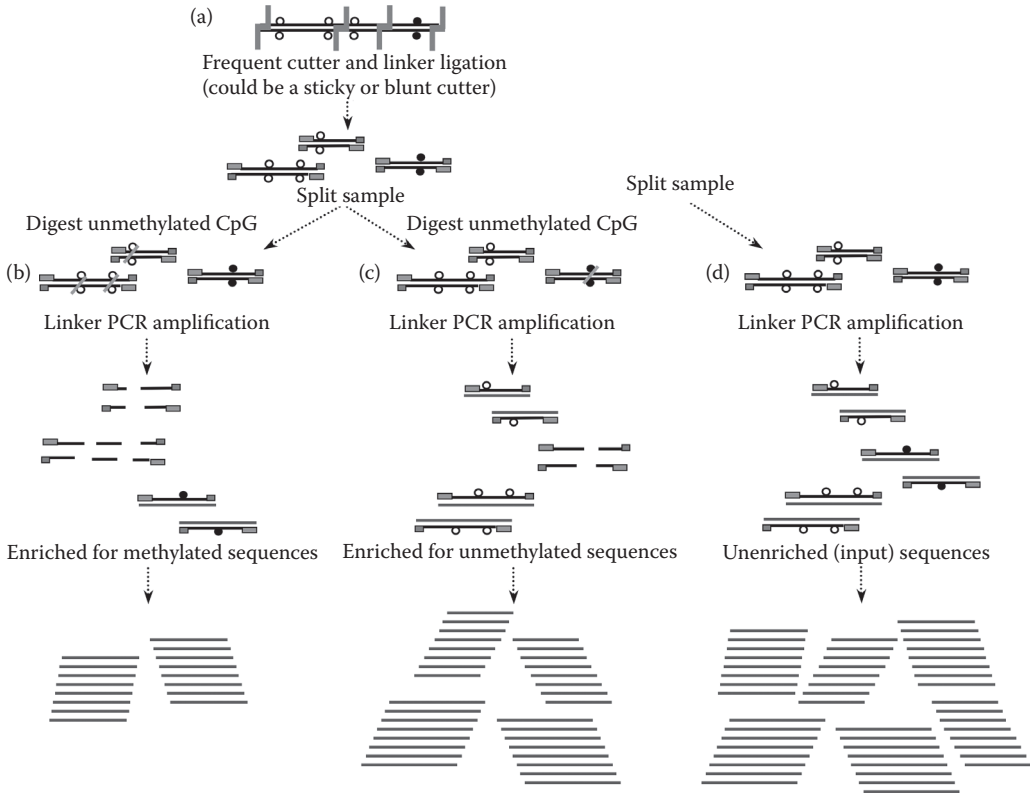


FIGURE 13.6 (See color insert following page XXX.) Enzyme approaches for obtaining methylated (b), unmethylated (c), and input (d) fractions from a single sample. Genomic DNA is cut with a frequent sticky cutter (a), linkers are ligated and the sample is split. Then separate fractions can be subjected to (b) unmethylated sites are digested such that intact sequences are amplified; (c) methylated sites are digested such that intact unmethylated sequences are amplified; and (d) the input sample (with ligated linkers) is amplified. Linker PCR amplification involves denaturation step to make single-stranded DNA that is used as a template for amplification. Any of the fractions from (b), (c) or (d) can then be cohybridized to an array.

enzymes that provide the maximum number of restriction sites within the predicted target fragments. In contrast, methylation-dependent digestion is performed using only a single enzyme.

13.4.3 Sticky Enzyme Approaches

Another method takes advantage of sticky cutting enzymes (Figure 13.5b) to separate methylated and unmethylated DNA; two variations on this approach are described in the following sections.

13.4.3.1 Enrichment of Consecutive Methylated Sites

Genomic DNA is digested with a methylation-sensitive blunt cutting enzyme [Figure 13.7b(i)] before a sticky cutting enzyme with the same recognition sequence is applied [Figure 13.7b(ii)]. If the recognition sequence for these enzymes is unmethylated, the DNA will be cut with blunt ends; the situation is reversed if the recognition sequence is methylated. Linkers are then ligated to the sticky ends [Figure 13.7b(iii)] and only fragments with sticky cuts at both ends are amplified [Figure 13.7b(iv)]. The amplified fragments correspond to sequences that contain consecutive methylated recognition sites.

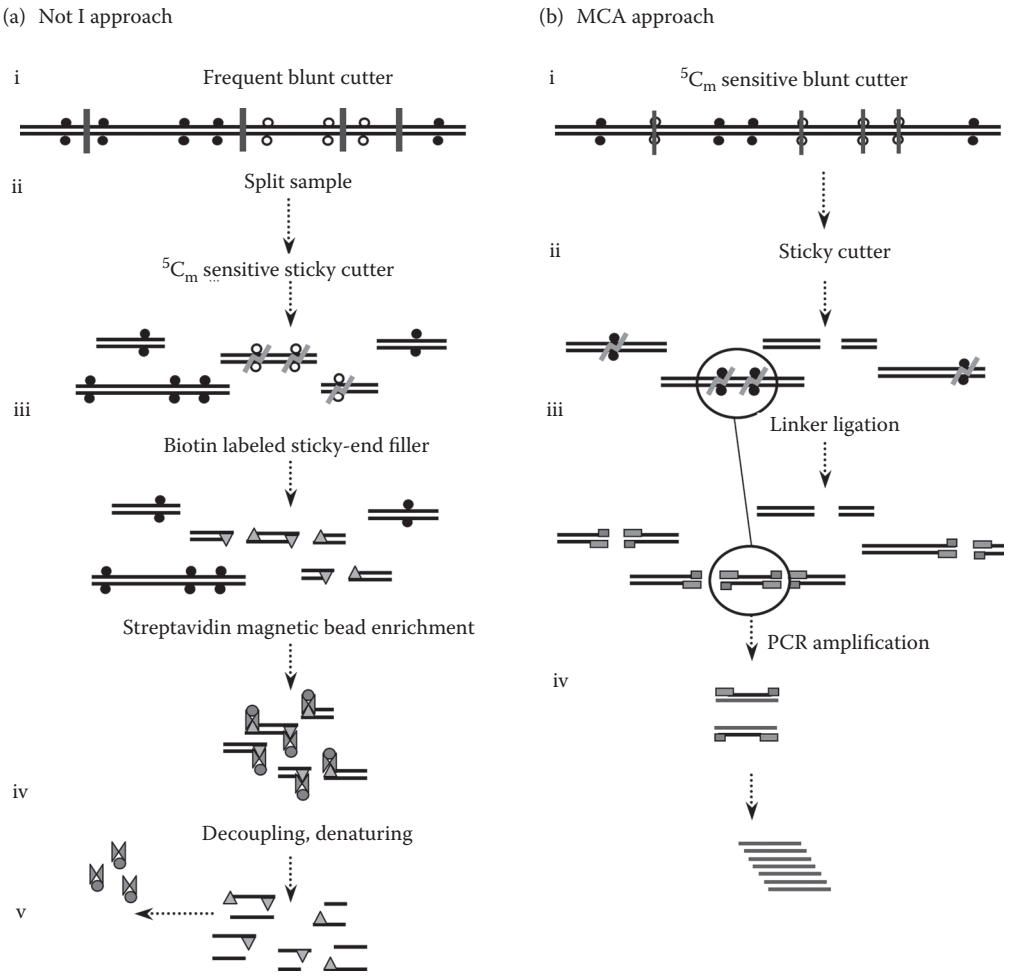


FIGURE 13.7 **(See color insert following page XXX.)** Enzyme approaches (a) that do not require amplification and (b) that require consecutive methylated recognition site for enrichment. In (a), genomic DNA is digested with a blunt cutter (i) and the sample is split (ii). One half is further digested with a methylation-sensitive sticky cutter (iii), the sticky ends are filled with biotin-labeled nucleotides and strong binding streptavidin-coated magnetic beads are added. Sequences bound to the beads are extracted (using a magnet) (iv) and a decoupling reaction is used to release the bound sequences (v). In (b), a methylation-sensitive blunt cutter digests the genomic DNA (i) and the sample is split (ii). One half is further digested with an enzyme [with the same recognition sequence as in (i)] whose activity is not inhibited by methylation (ii). After linker ligation (iii), only sequences with both 3′ and 5′ linkers are amplified (iv).

The method of methylated CpG island amplification (MCA) [40] uses this approach. It compares two samples by applying the methylation-sensitive blunt cutting restriction enzyme *Sma*I followed by the sticky cutting enzyme *Xma*I, both of which have the same recognition sequence. Historically, MCA was combined with a low-throughput method called representational difference analysis (RDA) [41]. One disadvantage of this approach is its dependence on an amplification step to enrich for sequences with consecutively methylated recognition sites. However, its biggest limitation is its reliance on the recognition sites (of the enzymes chosen) occurring frequently throughout the genome. If the gap between recognition sites is large compared to the size of a methylated region, the methylated sequence may not be detected. Hence, this method is only useful for detecting methylation in regions with closely spaced recognition sites.

13.4.3.2 Methylation-Sensitive Sticky Cut Enrichment

For this method, genomic DNA is digested with a frequent cutting blunt DNA restriction enzyme such as *EcoRV* (Figure 13.7a), along with a methylation-sensitive restriction enzyme (e.g., *NotI* or *BssHII*) that leaves sticky ends, resulting in an overhang on one strand of DNA. Consequently, unmethylated fragments of DNA (including those contained in unmethylated CpG islands) have the characteristic overhang seen in Figure 13.5b. The cleaved ends of these fragments are then filled with modified complementary nucleic acids [Figure 13.7a(iii)] that have been altered by labeling one of the nucleotides with a chemical, such as biotin-dNTP, that is used as a reporter. Biotin compounds form a strong ionic bond with streptavidin particles and hence, by coating magnetic beads with streptavidin, unmethylated fragments can be extracted from the digested DNA. Subsequently, the unmethylated fragments are removed from the magnetic beads using affinity purification before being cohybridized with reference genomic DNA. Ching et al. [42] applied this method using the enzyme, *NotI*.

An advantage of this approach is that the extracted unmethylated DNA does not have to be amplified before hybridization. However, blunt double-stranded cutting enzymes cut at a hexamer located randomly throughout the genome and, consequently, they are unable to discriminate between methylated and unmethylated regions. This could lead to small methylated sequences being included in the (supposedly) unmethylated DNA fragment. It is possible to use bioinformatic techniques to determine exactly where the enzyme will cut and so, in theory, this situation could be modeled. Another problem is that the ionic bond formed between biotin and streptavidin is extremely strong. To break this bond and extract the unmethylated fragment, the affinity purification requires the use of a low pH, a high temperature, and a strong denaturing agent (such as formamide). The subsequent purification of the extracted DNA prior to hybridization can result in a loss of yield.

13.4.4 Limitations of Enzyme-based Approaches

Methylation-sensitive restriction-based approach have several advantages. In particular, no base modification is required (unlike bisulfite modification methods) and they are relatively straightforward, specific, rapid, and inexpensive (certainly compared to HPLC/mass spectrometry). Furthermore, **Q1** sequence data from the Human Genome Project makes it possible to identify recognition sequences of methylation-sensitive restriction enzymes, allowing the prediction of restricted fragment sizes which enables identification of the best combination of enzymes for a particular assay. The main disadvantage of using a methylation-sensitive restriction method is the enzyme's inherent inability to digest completely the methylated sequences within the sample. For this reason, combined enzyme and antibody approaches have been suggested (Figure 13.8b). Additionally, the number of CpGs a single enzyme can assess depends upon its recognition sequence (Figure 13.9d), and the size of resulting fragments relative to the regions of methylated and unmethylated DNA (Figure 13.10b). We note that human sequence information, in conjunction with bioinformatics tools, allows the identification of restriction enzymes that can assess the methylation status of a particular CpG. Bioinformatic techniques can also be used to identify the best combination of enzymes for a particular method so that methylation levels for the largest number of loci possible can be assessed.

13.4.5 Methyl Antibody Approach

An alternative way of measuring DNA methylation on a genome-wide basis is to use a methyl antibody approach known as methylated DNA immunoprecipitation (MeDIP) [43]. As shown in Figure 13.8a, methylation-specific antibodies are used to enrich methylated fragments of the genome [43–45] and, by cohybridizing these with a reference sample, it is possible to identify methylated regions of the genome.

The method used to enrich the methylated fragment is analogous to the immunoprecipitation step in ChIP-chip experiments. After shearing the DNA (Figure 13.9a), a mouse monoclonal antibody against methylated cytosine is used to enrich for methylated fragments. DNA that has been sheared,

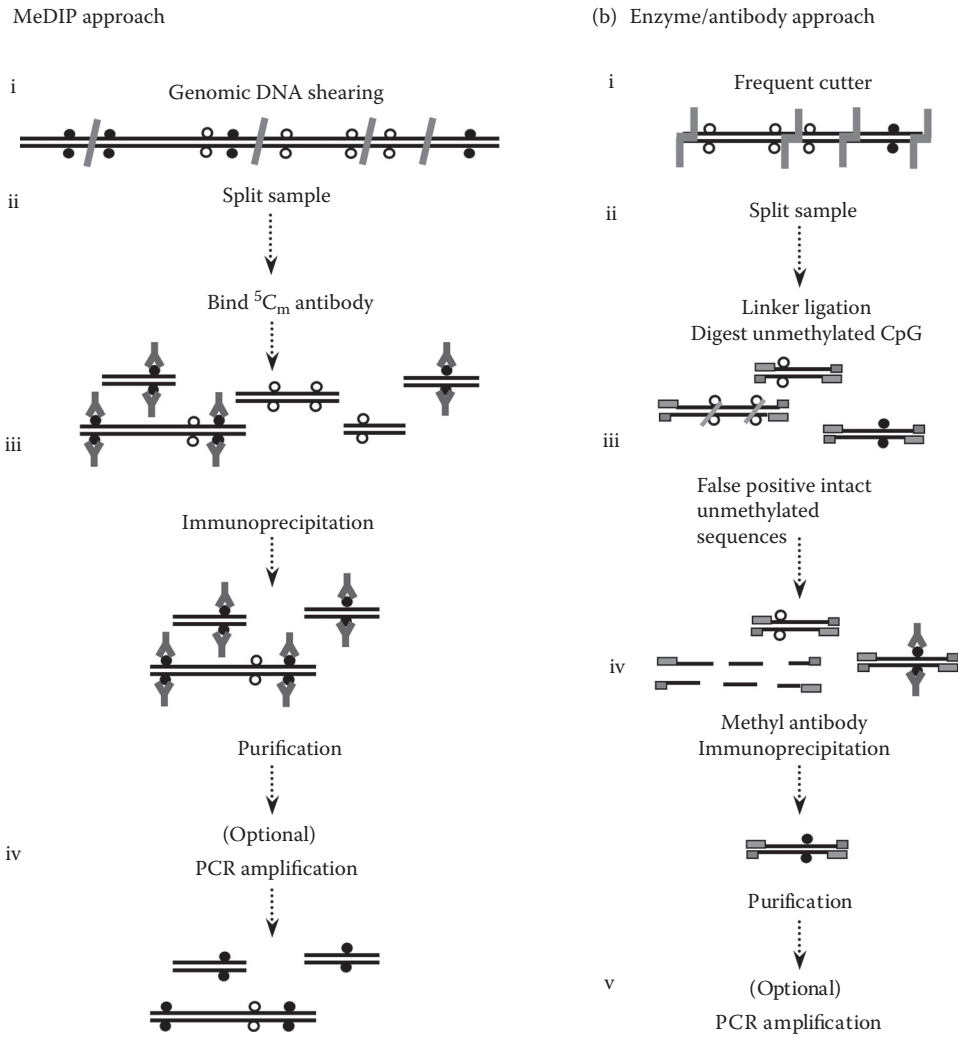


FIGURE 13.8 (See color insert following page XXX.) Approaches using (a) MeDIP antibodies and (b) enzyme digestion and antibodies. In (a), genomic DNA is sheared (i) and the sample is split (ii). Antibody proteins that bind to methylated CpGs are added to one half (iii) and sequences with bound antibodies are precipitated and this enriched sample is purified. PCR amplification may be used (iv) depending on the amount of sample required for hybridization. In (b), genomic DNA is digested with a sticky cutter (i), linkers are ligated and the sample is split (ii). One half is further digested with a methylation-sensitive enzyme leaving methylated sequences intact. Any unmethylated sequences that remain intact (escaping digestion) are subsequently removed using the MeDIP appraoch (iv). Finally, the sample is purified and (v) amplification may be used.

but not treated with the antibody, is used as the reference sample. After labeling the two samples with fluorescent dyes, they are hybridized to an array, and the data generated can be used to find methylated regions. The principle difference between ChIP-chip and MeDIP is that, in ChIP-chip experiments, the regions enriched for a particular protein tend to be symmetrical (Figure 13.10b) but, when MeDIP is performed, the level of methylation across a region can vary nonsymmetrically (Figure 13.10c). For example, a CpG island might display more methylation at the 5′ end relative to the level observed across the rest of the island. This means that analysis methods for ChIP-chip experiments may not be directly applied to data obtained using MeDIP; for more discussion of this, see Section 13.6.

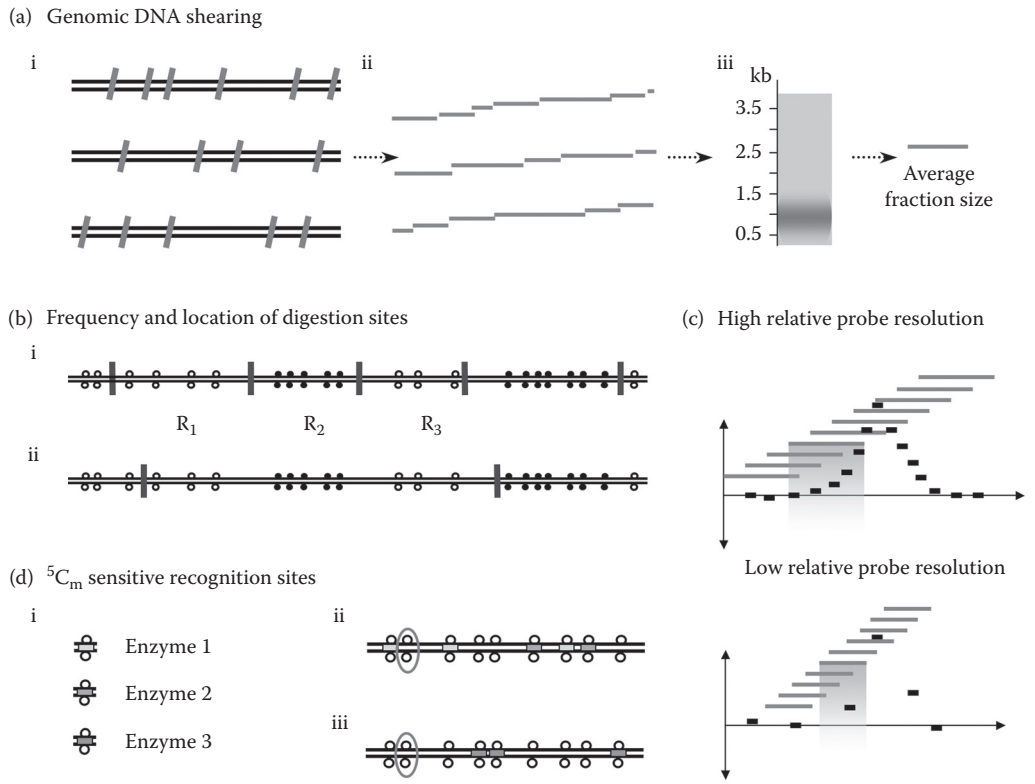


FIGURE 13.9 (See color insert following page XXX.) (a) Genomic DNA shearing; depicted are three genomes (i) which (after shearing) results in random overlapping fragments (cyan segments) (ii) of varying size depending on the amount of shearing applied. The overall size of the fragments should be confirmed by running the sample on an electrophoresis gel (iii). Bearing in mind the overlapping fragments resulting from genomic shearing, data obtained from consecutive probes along the genome (c) will be correlated. The extent of correlation between neighboring probes depends on the size of the average target fragments (cyan segments) relative to the probe resolution (spacing). In (c) where there is a high relative probe resolution, a single fragment may potentially bind to five neighboring probes, but where there is a low relative probe resolution, a given fragment may only bind to one probe. (b) If enzyme digestion is used to fragment the genome, infrequent cutting (ii) results in regions of altered methylation status (R1–R3) with ambiguous methylation. In contrast, where restriction sites are just as frequent as regions with altered methylation status (i), target fragments will be informative for methylation. (d) Often a combination of enzymes (i) is required for effective and informative digestion of DNA. For a given fragment, enzyme 1 has three recognition sites, (ii), enzyme 2 has two recognition sites (ii) and enzyme 3 has three recognition sites (iii). The combination of enzymes 1 and 2 results in five of the nine CpGs being included in a recognition site (ii). Q3

Antibody enrichment is known to be inefficient but importantly (and unlike restriction enzyme-based approaches) it is very specific. This means there is limited bias in the MeDIP enrichment but, if the amount of starting material is small, amplification may be required. Moreover, it is dose-dependent—the level of enrichment is positively correlated with the number of methylated cytosines (Figure 13.10c). However, there are also a number of drawbacks. In particular, the dose-dependency means that the CpG density of a region has to be considered in the analysis step to avoid the methylation of regions with low CpG content being underestimated. Additionally, it has been shown empirically that regions with a CpG density of <2% are not enriched efficiently. Notwithstanding these problems, MeDIP is a promising strategy for genome-wide methylation analysis as evidenced by a recent publication that used it to elucidate the genome-wide DNA methylation profile of *Arabidopsis*, the first high-density methylation profile of any genome [46].

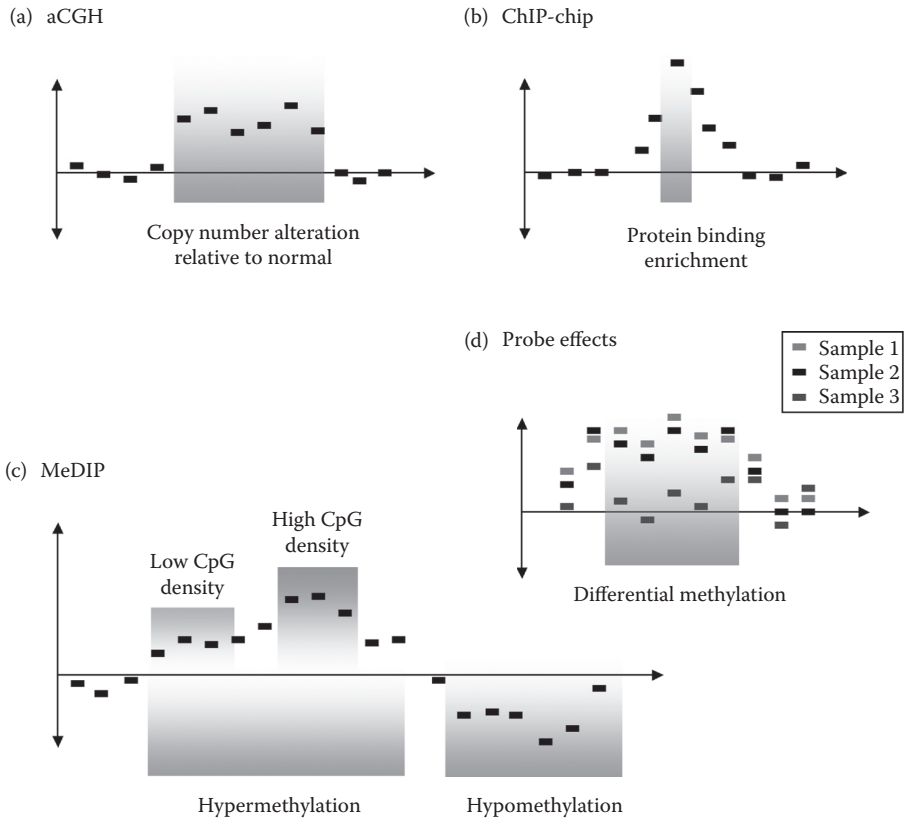


FIGURE 13.10 **(See color insert following page XXX.)** Log-ratio data for probes along the genome illustrating typical changes that are observed in (a) array CGH experiments, (b) ChIP-chip experiments, and (c) MeDIP methylation array experiments. Changes in log-ratio data are expected to appear like (a) gained or lost segments, (b) positive symmetric peaks, and (c) variably peaked regions (positive or negative, after normalization). In (c), the variable peak heights of a methylated region depend on the density of methylated CpGs. In addition to CpG density effects in MeDIP data, there will be (d) probe effects. Real changes in methylation (sample 3, blue data points) can be distinguished from changes due to probe effects when there are multiple samples. Similar probe effects are observed in samples 1, 2, and 3. However, sample 3 is differentially methylated across a five-probe region compared to the other samples. **Q3**

13.4.6 Detecting Methylated DNA by Methyl-CpG Binding Domain Proteins Affinity Purification

As described in Section 13.2, the methylation of CpG islands associated with genes can have a direct effect on gene expression by inhibiting the recruitment of certain transcription factors that are essential for gene expression [47]. However, accumulating evidence suggests that DNA methylation changes are associated with alterations in chromatin structure through post-translational modification of histones, or physical remodeling of chromatin structure [48,49]. Moreover, a functional link has been established between DNA methylation and chromatin structure, including a cumulative effect of both DNA methylation and chromatin modifications (e.g., histone deacetylation) on gene expression, implying a cooperative function [48]. A further link between DNA methylation and functional histone modifications is the family of DNA methyl-CpG binding domain proteins (MBD or MeCP) [31]. Certain members of this protein family bind specifically to symmetrically methylated CpGs and interact with large transcriptional repressor complexes that switch off gene expression

[49,50]. Therefore, the MBD family of proteins is thought to provide a functional link between DNA methylation and chromatin modifications that result in altered gene expression.

The specificity of certain MBD family members for methylated CpGs has been utilized in an alternative approach for identifying methylated regions. This technique, recently termed methylated-CpG island recovery assay (MIRA), involves the use of an immobilized MBD protein on a solid matrix over which the fragmented DNA sample of interest is passed, resulting in the affinity purification of methylated regions and the elution of unmethylated regions [10,51–54]. The isolated methylated DNA fragments can then be applied to genomic microarrays or analyzed by quantitative PCR. Several variations of this approach have been published using either the full-length MeCP2 protein, full-length MBD2, the MBD2/MBD3L1 complex, or the core MBD domain of MBD2 for affinity purification of methylated DNA [10,51–54]. Since full-length MBD proteins may have preference for certain sequences [55], it is likely that the use of the core MBD domain of MBD2 would provide the least biased approach. An alternative reagent has been recently described which may provide increased sensitivity and specificity of MIRA-like approaches, using an engineered poly-MBD protein [56].

MIRA has several advantages over other global methylation screening techniques. The MBD domains utilized in MIRA bind specifically to heavily methylated DNA, compared to restriction enzyme or methyl-cytosine antibody techniques which may also identify sites of single methylated CpGs. Further, the MBD proteins link DNA methylation, histone modification, and gene expression. Therefore, MBD purification approaches may isolate functionally relevant regions of the genome that are associated with the control of gene expression. However, these advantages also limit the types of sequences that can be isolated by MIRA—if a sequence is not sufficiently methylated, it may be missed by this technique.

A recent approach (COMPARE-MS) combined a restriction enzyme-based approach and MIRA to isolate methylated sequences [57]. This was reported to have increased sensitivity and to be applicable for high-throughput screening, suggesting that it may allow DNA methylation screening in the clinical setting.

13.4.7 Arrays for Bisulfite-Treated DNA

Most of the high-throughput methods discussed thus far have used enrichment-based approaches. An alternative is to apply bisulfite conversion in conjunction with arrays that look for converted unmethylated loci and unconverted methylated loci. Previously, this approach has been limited by difficulties in designing probes for converted DNA where there are effectively only three bases—this can lead to nonspecific hybridization. However, a SNP detection approach using beadarrays has been developed recently to detect specifically bisulfite converted loci [58].

For a particular locus of interest, four oligonucleotides (two allele-specific and two locus-specific) are used. The 3′ end of the allele-specific oligos are designed to hybridize with the bisulfite-modified DNA—one oligo will hybridize if the cytosine base has been converted to uracil, whereas the other will hybridize if no conversion has occurred. Moreover, a different PCR priming site is attached to the 5′ end of each oligo and these sites are fluorescently labeled with different dyes. The locus-specific oligos are designed so that the 5′ end has a locus-specific sequence, the 3′ end is a universal PCR priming site, and in the middle is an address that identifies the oligo with a genomic location. The sequences differ between the locus-specific oligos; for one oligo it is assumed that the CpG of interest is methylated and consequently all CpGs in the locus-specific sequence will also be methylated and vice versa. Subsequently, where there is allele-specific hybridization, a one-step primer extension is performed to ensure that perfect matching has occurred at the allele-specific methylation site. Next, the locus-specific oligos are ligated to their appropriate partner and the subsequent products amplified using PCR before being hybridized to a beadarray using standard techniques [59]. The methylation level can be measured by observing the amount of fluorescence emitted by the dyes attached to the primers of the allele-specific oligos.

The principal advantage and drawback of this method are related to the (very strong) assumption that all CpGs around the locus of interest share the same methylation status. While this means it is possible to have high confidence that any observed differences in methylation will be genuine, it is also known that (even locally) CpGs may not have the same methylation status. Consequently, if only one of the CpGs around the locus of interest has a different methylation status, the methylation status of the CpG of interest cannot be determined since neither allele-specific fragment can be amplified.

13.5 ARRAY CHOICES

Designing a microarray to investigate DNA methylation on a genome-wide level is a challenging problem. The investigator has to decide whether to use bacterial artificial chromosomes (BACs), CpG island clone libraries, or oligonucleotide (oligo) probes and, given this, what resolution (coverage) the probes should have and where they should be located on the genome.

One method is to use whole-genome BAC arrays originally designed for array comparative genomic hybridization (array CGH) studies [43]. BACs (~120 Kb long) are tiled across the genome and so all CpG islands (or other methylated regions) should be contained within, or straddle two or more, BACs depending upon the relative sizes of a given CpG island and the BACs covering that region. An advantage of this method is that BAC arrays already exist for many organisms and could be adapted easily for methylation studies. BACs also provide coverage in regions of the genome where it is difficult to design smaller probes due to repetitive sequence content. Additionally, because of the length of the probes, the noise associated with BAC array data is low relative to other microarray technologies. However, the length of a BAC relative to a CpG island means these arrays may not be sensitive enough to detect either small changes in methylation or small regions of methylation. In particular, many CpG islands (or other methylated regions) could be contained within a single BAC, resulting in data that are difficult to interpret. Further, even if only one methylated region is contained within a BAC, the large disparity in size between this region and the BAC may affect the hybridization and result in data where it is difficult to distinguish between noise and truly methylated regions.

An alternative approach is to use CpG island arrays, where the probes are taken from CpG island libraries [10]. In this case, there is no ambiguity about the location of the methylated regions—this is the principal benefit of this method. However, there are a number of problems; for example, it is assumed that all methylation of interest occurs in CpG islands—this is not necessarily so. Moreover, CpG island arrays often include probes that are not CpG islands or that are made up of repetitive genomic sequences; this can lead to an increase in the level of background noise due to cross-hybridization. Consequently, the downstream analysis of such arrays requires expert bioinformatic support.

Another method is to use oligo arrays. This area is developing rapidly but, as yet, no publications have described its application to genome-wide DNA methylation studies. However, it is likely that such publications will arise in the near future. Unlike BACs, oligos are small (generally between 25 and 70 bp in length) and are therefore not practical to generate an array where they are tiled across the whole genome. Consequently, when using oligo arrays to investigate DNA methylation, it is necessary to choose between a number of different layouts. One option is to space the oligos (approximately) evenly across the genome. Alternatively, promoter arrays or CpG island oligo arrays, where oligos are tiled within gene promoter regions or CpG islands, may be used [60]. Both of these methods have the advantage of providing more sensitive coverage than a BAC array [61]. One obvious advantage of promoter or CpG island arrays is their extremely high coverage in genomic regions where methylation changes might be expected to occur—depending upon the resolution, it may even be possible to discriminate between different levels of methylation within the same CpG island or promoter region. Of course, any of these methods run the risk of differential methylation occurring in regions where there are no probes. However, as technology develops, it ought to become

possible to design longer oligos (up to 200 bp in length) than are available at present [62]. Notwithstanding technical problems caused by repeat sequences, this should allow the design of whole-genome tiled oligo arrays that would be ideal for studying DNA methylation.

However, irrespective of the chosen layout, oligo arrays do have some drawbacks. In particular, since oligos are much shorter than BACs, the processed signal tends to be noisier. Indeed, when oligo arrays are used for array CGH experiments, it is generally necessary to average over a window of 3–5 consecutive probes to reduce the variability. Nevertheless, the effective resolution of oligo arrays may still be greater than BAC arrays, depending upon the density of oligos on a given array [61].

An additional difficulty when designing oligo arrays is the GC content of individual probes. While GC content (and correspondingly the probes' melting temperature) has been shown not to affect the assignation of probes as outliers [63], this is quite different from more subtle changes in the observed intensities that may be caused by GC-dependent hybridization efficiency biases. This effect would clearly be aggravated if a probe was located within a CpG island. Moreover, if a DNA amplification step is used, this may increase the GC-dependent bias depending upon the protocol used for target preparation (the amplification step is less commonly used in array CGH experiments, so this problem has not yet been thoroughly examined). It is worth noting that GC-dependent biases may be less of a problem if large profiling studies are being conducted, since cross-sample information can be used when calling differentially methylated regions (DMRs) in the analysis step. Of course, this assumes that much of the observed variation is systematic (i.e., the variation tends to be similar across arrays). Finally, we note that a probe's GC content is less of a problem in BAC arrays (in terms of hybridization bias) because of the probe's length; however it has still been observed to affect the quality of the generated data.

One other important problem, irrespective of the array design, is how target fractions are fragmented. In particular, the relationship between target fragments and the probes they hybridize to affect the interpretation of methylation measurements. A summary of the issues involved is given in Figure 13.11.

In summary, despite the problems mentioned above, it is likely that (as is already happening with array CGH experiments [61]), oligo arrays will supersede BAC arrays in DNA methylation studies. This is due principally to their superior resolution and their ability to better target small regions of the genome, such as CpG islands and other regulatory regions, where methylation is likely to occur.

13.6 DATA ANALYSIS ISSUES

Two major problems with the analysis of DNA methylation array data are (i) normalization and (ii) calling of methylation status at given genomic loci. Owing to the nature of methylation in the mammalian genome, global levels of methylation can differ radically between samples. Therefore, normalization of arrays within an experiment can be difficult—real differences might be normalized away. Calling methylation levels across the genome is also challenging but, given accurate normalization, differential methylation measures can be obtained. Finally, many data analysis issues are dictated by the experimental approach and the array and probe design.

13.6.1 Normalization Issues

Regardless of the approach, most microarray-based methods result in log-ratio data that are characteristically asymmetric. The skewness of the log-ratio distribution arises from a fundamental imbalance in methylation levels throughout the genome. In normal cells, there are generally more methylated than unmethylated sequences whereas, in cancer or diseased cells, the opposite situation can occur. The extent of the skewness is determined largely by the global levels of methylation in the samples studied. In addition to this skewness (which is specific to DNA methylation experiments), it is also recognized widely that dye-bias is a common problem in two-color microarray experiments that must be corrected.

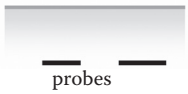
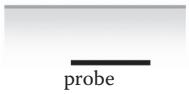
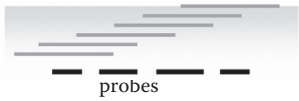
Many predicted target fragments ⊂ one probe	Predicted target fragments  probe	Ambiguous interpretation · Intensity at probes may be due to multiple predicted target fragments · gives information on methylation within and around probe
Predicted target fragments = probes	Predicted target fragments  probes	Clear interpretation · each probe hybridize single matching target fragment · gives information on methylation within each probe
One predicted target fragment > many probes	Predicted target fragment  probes	Ambiguous interpretation · multiple probes hybridize single predicted target fragment · gives information on methylation within and around probes
One predicted target fragment > one probe	Predicted target fragment  probe	Clear interpretation · single predicted target fragment hybridizes to single (smaller) probe · gives information on methylation at and around probe
Many random target fragments > many probes	Random target fragment  probes	Clear interpretation · random (overlapping) target fragments hybridize to multiple probes · gives (dependent) information on methylation within and around probes

FIGURE 13.11 **(See color insert following page XXX.)** Relationship between target fragments and probes in the interpretation of methylation measurements across probes. Fractionation of the genome by enzyme digestion results in predicted target fragments. Fractionation by shearing yields random (i.e., overlapping) target fragments. Target fragments that hybridize to a given probe determine the amount of methylation measured at that probe. **Q3**

Unfortunately, real differences in methylation between samples can be removed through inappropriate use of common normalization procedures. This fundamental problem has received little attention in the literature despite the fact that it can have a dramatic impact on the results. We now consider an example (Figure 13.12) of the type of problem that might arise if an inappropriate between-array method is used to remove the skewness. Two arrays are used: in the first, methylated sequences enriched from a normal sample are compared to unenriched sequences from the same sample, whereas, in the second, methylated sequences enriched from tumor tissue are compared to the corresponding unenriched sequences. The aim of the experiment is to compare DNA methylation in the tumor and normal samples. In array 2, many more negative log-ratios are obtained (corresponding to a large number of hypomethylated sequences), whereas in array 1 (the normal tissue array) more positive log-ratios are observed (corresponding to more methylated than unmethylated sequences). Assume for the moment that there is no dye-bias present in either of these arrays so that no correction for this is required (Figure 13.12a). Clearly, the median log-ratio for the first array is less than zero, whereas for the second array it is greater than zero. If a (between-array) median normalization were performed to make the two arrays "comparable" (i.e., both centred at zero), real differences in methylation between the tumor and normal array would be removed, as shown in Figure 13.12b. Consequently, even in the absence of dye-bias, performing between array normalization can be dangerous and should be performed only with great care.

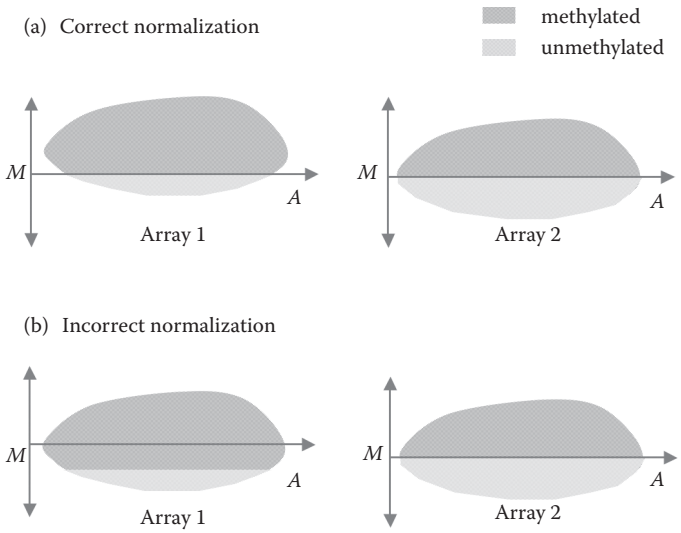


FIGURE 13.12 **(See color insert following page XXX.)** Diagram of MA-plots for two arrays after (a) correct normalization and (b) incorrect normalization. Shaded regions in the plots represent the areas of the plot where data would be observed. Blue areas represent data from truly methylated sequences. Orange areas represent data from truly unmethylated sequences. The sample hybridized to array 1 has considerably more methylated sequences than the sample hybridized to array 2. After correct normalization (a), methylated and unmethylated regions coincide with the $M = 0$ line for both arrays; positive M-values correspond to true methylation and negative M-values correspond to true unmethylation. After incorrect normalization, array 2 (which has been globally centerd) has some of the methylated region below the $M = 0$ line, falsely represented as unmethylated. **Q3**

13.6.2 Normalization Options

In an ideal experiment with no dye-bias (or other biases due, for example, to amplification or labeling), any global shift or skewness in log-ratios could be interpreted as a real difference in the proportion of methylated and unmethylated sequences present. However, dye-biases do occur and this variation is confounded with global changes in methylation. Consequently, it is difficult to assess whether a shift in log-ratios is due to dye-bias or real differences in global methylation. This is a particular problem in experiments that involve DNA methylation array profiling of samples with very different methylation profiles (e.g., experiments involving normal and tumor samples are very difficult to normalize). The one situation where dye-biases may be effectively normalized is where appropriate and reliable exogenous controls are available. Otherwise, it may be preferable to avoid dye-bias normalization. Alternatively, the investigator might normalize any obvious intensity-dependent effects without shifting the overall location of the log-ratios to zero. This might be achieved by applying a loess normalization followed by a global correction of the log-ratios back to their orginal median value. However, this approach does not seem very satisfactory. Methods such as VSN [64] or other affine transformations [65] may provide a normalization that is more robust to the asymmetry.

The practicality of using exogenous spiked controls depends on two main factors. The first is whether the intensity values of the controls cover sufficiently the range of possible intensities. This determines if an intensity-dependent normalization based on these controls is possible. The availability of such controls depends entirely on the number of different control probes on the array and whether the experiment is designed to use them in this way. The second, and probably more important, factor is whether the spiked material has been added in precisely the same amounts in both channels of the experiment—if not, a spike bias will be introduced into the normalization.

Spike biases can occur due to pipetting errors arising from trying to aliquot very small quantities of material—typically the amount of spiked material is very small relative to the amount of target material. To account for this problem, spiked material is generally diluted so that greater volumes are added, minimizing the chance of pipetting-based errors. However, there may be a constraint on the total volume of target material allowed (e.g., when a hybridization chamber is used). Additionally, it is possible that to cover the full range of intensities, spike controls at high concentrations may dominate the amplification step.

In expression array experiments, titration controls have also been used [36]. Here, all probes on the array are pooled, and a serial dilution of the pool is spotted onto the array, assuming that both samples hybridize equally to this pool, these controls can be used for normalization. A disadvantage of this approach is its presumption that these custom spots can be printed on the array. Additionally, the assumption that both target fractions will hybridize equally to the titration pool may not be valid for all samples and approaches.

For enzyme-based DNA methylation array approaches there exists a class of probes that are potentially useful for normalization. These so-called *uninformative probes* do not have restriction sites for the enzymes used to enrich for methylated or unmethylated fractions. The presence of such probes depends on the type of array being used. Probes on CpG island arrays are based on sequence libraries obtained by cutting the genome with *MseI*. Therefore, if *MseI* genomic digestion is used to fragment the genome, the probes will correspond to predicted target sequences. Target sequences without internal digestion sites should hybridize equally (to their corresponding probes) in both channels. If an oligo-based array is used, such uninformative probes may not be available. However, if the oligo probes are designed to be contained within predicted target sequences, there will be uninformative probes that may be used for normalization. Obviously, if the genomic DNA is sheared, rather than fragmented into predicted sequences using enzyme digestion, uninformative probes will not exist. Uninformative probes are also useful for quality assessment—the variability in their log-ratios reflect the inherent quality of the experiment and the log-intensities indicate the level at which single-copy sequences are detected (after amplification).

In experiments incorporating technical dye-swaps for every array, dye-bias self-correction may be sufficient. However, care must be taken to check the basic assumption that the dye-bias is the same between the dye-swap pair. If labeling occurs during the amplification step, dye-swaps are not true technical replicates since different amplification steps would have been used. In this case, dye-bias self-correction may not be appropriate since differences in amplification biases between arrays are confounded with dye effect. It might be hoped that normalization of a large reference design experiment would be helped by utilizing intelligently the fact that the reference channel is common to all arrays. However, as with using dye-swaps for self-normalization, it is important that the reference channel is indeed a technical replicate (i.e., same fragmentation, enzyme digestion or enrichment, amplification, and labeling steps) in each array.

DNA methylation experiments involving samples with similar global levels of methylation are easier to normalize. While the log-ratios for each array may still be considerably skewed, the important assumption is that the extent of skewness is roughly the same for all arrays within an experiment. Consequently, within-array intensity-dependent dye-bias normalization would modify the overall log-ratios for each array by the same amount. In other words, the arrays would be comparable to each other, but perhaps not to external experiments that include samples with significantly different global methylation levels. This approach to normalization and subsequent analysis of DNA methylation array data is thus restricted to homogeneous collections of samples.

13.6.3 Quality Assessment of DNA Methylation Arrays

Quality assessment is a critical step in the analysis of any microarray experiment. As for expression arrays, it is crucial to identify features that determine whether an experiment is of good quality. For example, one feature that might indicate a problem in the experiment is obvious dye-bias. One way of

exploring if this occurs is to compare plots of the red and green channel log-intensity distribution within and between arrays, as well as between different experimental batches. Other features of concern are spatial effects across the array or a low overall foreground-to-background signal intensity ratio.

In addition to examining these problems, another crucial quality control step is to generate MA-plots for every array and to compare the scatter before and after processing steps such as background correction and normalization. Caricatures of typical MA-plots observed for methylation experiments produced using different experimental designs and protocols are shown in Figure 13.4.

Given the extremely complicated experimental protocols described in Section 13.4 to effectively perform the quality assessment steps described above, it is crucial to have a sufficient number of replicate arrays, particularly technical replicates and dye-swaps. Besides being a very useful tool for determining the reliability of the data, replicate arrays also (for example) allow the investigator to discern what sort of MA-plot is associated with good quality data for a particular approach. Additionally, as more and more experiments are performed, the characteristics that are important in determining whether a dataset is of good quality will become apparent.

13.6.4 Analysis of Methylation Data

A key motivation for using DNA methylation arrays is to answer the question: how and where in the genome is my sample methylated? This involves trying to infer *absolute methylation* levels across the genome for a given sample. Alternatively, where a sample is compared to another (within an array), an equivalent aim would be to call changes in *relative methylation* levels across the genome. While single arrays can be used to determine relative methylation levels, multiple arrays are required to find DMRs of the genome (i.e., regions that are consistently differentially methylated).

In Section 13.4.2, we describe methods for finding DMR [i.e., regions of the genome that are differentially methylated *between* samples (arrays)], while in the following section we concentrate on methods for calling methylation *within* an array.

13.6.4.1 Calling Methylation within an Array

Low-throughput approaches (e.g., those based on bisulfite sequencing) often summarize results at individual CpGs as a percentage methylation measurement. Regardless of whether this is entirely meaningful (even from low-throughput data inferences) it is unlikely to be a feasible objective for array experiments.

Even ignoring problems caused by sample heterogeneity, a number of experimental factors make determining absolute methylation levels extremely difficult. For example, if the MeDIP approach has been used (Section 13.4.5), the number of binding sites of the methyl antibody in a particular region clearly impacts upon the amount of DNA amplified and this will affect the intensity (and hence the amount of methylation) observed. In particular, if two regions are fully methylated, but one has more binding sites, it will be enriched more efficiently and appear to be more methylated (Figure 13.10c). Consequently, it may not be biologically meaningful to state that methylation in one genomic region is greater than another. Similarly, for methylation-sensitive restriction enzyme-based approaches (including the NotI approach) the location of the restriction sites will determine how much sample DNA can be extracted and subsequently hybridized. Thus, if two regions are fully methylated but one contains a larger number of restriction sites, it might be more efficiently enriched and thus appear to be more methylated. One way to resolve this problem could be to locate an enzyme's restriction sites and factor this information into the analysis. However, given that hybridization biases will also occur, this is likely to be difficult. While bioinformatic approaches may be able to tackle this problem, it is not clear that they will be able to model other sources of bias that can impact upon the investigator's ability to measure absolute methylation. For example, will it be possible to model probe effects, enzyme or antibody efficiencies, restriction site or binding site frequencies, and amplification biases? These biases will undoubtedly be protocol (and even lab)

specific and very difficult to generalize. Moreover, very large experiments, involving extensive replication would probably be required to measure them.

Even if sample homogeneity could be assumed and the biases mentioned above were either nonexistent or removed, it is unlikely that regions 80% methylated could be distinguished with confidence from regions 90% methylated in the same sample—there would be considerable overlap in the data from both groups. However, it is realistic to expect (barring uninformative probes) that regions that are 10% methylated can be confidently distinguished from those that are 90% methylated. Despite appearing rather trivial, this represents a very meaningful biological difference, and is arguably more important than a measured (but not necessarily real) difference of 10% in methylation. To make this process more straightforward, probes in uninformative regions (i.e., having few restriction or binding sites, for example) can be identified bioinformatically and removed or downweighted. Moreover, probes that bind to fragments containing repetitive elements can also be considered for downweighting in statistical analyses.

Despite this, analyses that aim to attach an absolute methylation score (such as a percentage value) should proceed with caution for the reasons described above. However, calling changes in methylation levels across multiple probes is somewhat easier, particularly if the probe and target fragment resolution are high enough (Figure 13.9c). Analyses that average adjacent probes' log-ratios might allow better calling of true methylation changes in that region since many of the biases discussed above can essentially be averaged out. This may be acheived using a method as sophisticated as an HMM or as simple as a sliding window averaging technique. It is likely that HMM approaches might be problematic, not least because (i) many arrays will not have enough probes (relative to the size of methylated regions); and (ii) probes will probably be unequally spaced across the genome (i.e., there may be large gaps between probed promoter or CpG island regions). **Q1**

Instead of attempting to find absolute levels of methylation for each probe on an array, an alternative is to flag probes or regions where there is a difference in the relative methylation of the test sample to the reference sample; a number of analytic methods have been proposed for finding such regions.

Where microarrays have been used to investigate methylation at a genome-wide level [42,43], threshold-based approaches have been employed. This flags probes if their log-ratio is above or below a threshold generally derived from the standard or median absolute deviation of the log-ratios from all of the probes on the array. While this approach is simple to understand and implement, it fails to take account of all the information provided. It does not utilize the spatial dependency (i.e., probes that are genomically adjacent to each other are more likely to have the same methylation status than probes which are further apart) inherent in the data (Figure 13.9c). As the resolution of the arrays used for methylation analysis increases, it will become more important to take this into account; this will necessitate the development of model-based approaches.

Many analysis methods for finding copy number changes using array CGH experiments or regulatory elements/transcription factors using ChIP-chip techniques take spatial features into account [66–69] and, at first glance, some of these methods can be easily modified and applied to the problem at hand. However, there are a number of difficulties. The principle problem when modifying methods designed for analyzing array CGH data is that methylation status can change gradually over a region of the genome, whereas for array CGH data, it is assumed that copy number changes occur in steps (Figure 13.10a). Consequently, such methods might lead to an underestimation of the size of methylated regions (Figure 13.10c). (This will also be a problem for threshold-based approaches.) For ChIP-chip experiments on the other hand, analysis methods are designed to look for "bumps" (Figure 13.10b) which, superficially, seems more desirable. However, ChIP-chip analysis methods often assume that the bump is symmetrical and they are generally only interested in finding its centre—the peripheral region is of less interest. Additionally, ChIP-chip analysis techniques often rely on there being a large number of replica experiments [66]; this is unlikely to be the case for DNA methylation arrays due to the often large amounts of DNA required, the cost of such experiments, and the fact that this has not yet become a requirement in the literature.

Despite this, it seems likely that model-based approaches for tackling this problem will be developed; these approaches will depend on the technology used and the possibilities afforded by more repeated design experiments. However, it is difficult to speculate about the form this method will take. Moreover, until large datasets exist where the methylation status of the whole (or at least large parts of the) genome have been confirmed, it will not be possible to assess the efficacy of different methods.

13.6.4.2 Differential Methylation

All the methods described in the previous section can be thought of as "within-array" analysis. Another problem is how to combine information across arrays to explore whether regions of the genome are differentially methylated between samples. Methods for finding DMRs are heavily dependent upon the experimental design. For example, if the design illustrated in Figure 13.3c is employed, the same common reference sample must be used, otherwise cross-array inferences to find DMRs are effectively impossible. One of the major advantages of finding DMRs (rather than determining absolute or relative methylation levels within an array) is that a lot of the probe effects and other technical problems will be neutralized if we assume that the effects are the same for each array (Figure 13.10d). If we make this assumption, we can find DMRs by simply comparing the log-ratios from one array to another.

Of course, this also assumes that all of the arrays have been properly normalized. As discussed earlier, this is difficult; consequently, it may be hard to compare different arrays. In particular, even a slight difference in amplification efficiencies could result in subtle differences in the log-ratios which could lead to problems. Thus, developing methods for finding DMRs is more complicated than is apparent at a first glance. Because of this, and the lack of datasets where such methods can be tested, this remains an open and interesting research question.

13.7 ENZYME APPROACHES AND GENOMIC COPY NUMBER EFFECT

When using microarrays (or other technologies) to examine the methylation status of genomic regions, it is important to consider the number of copies of the genome present since this could effect the amount of methylation observed. For example, suppose we are interested in the methylation status of the same genomic region in two individuals, one of whom has two copies of the region and the other has three (or more) copies. We also assume that each strand of DNA in this region is methylated to the same extent for both patients. In this case, when the methylation status of this region is measured, it will appear as if more methylation occurs in the second individual relative to the first due to the additional number of copies of the genome. We note that this confounding will occur only if one of the experimental designs illustrated in Figures 13.3b, c or e are used. By hybridizing the test fraction alongside the input fraction, as shown in Figure 13.3d, the effect of copy number is neutralized. However, when the experimental designs described in Figures 13.3b, c, or d are used to properly assess methylation, it is also necessary to have a good understanding of the number of copies of the genome that might be present. This will be the case when methylation is examined using either one- or two-channel arrays.

One way of determining copy number is to hybridize DNA sheared using *XbaI* (or another digestion enzyme) to the same array used to measure DNA methylation. This has the advantage that the use of array CGH for determining copy number is well known and the protocols/analysis techniques are well established. Additionally, it ensures that copy number can be detected for every probe on an array.

Alternatively, technology is being developed that enables the measurement of copy number, loss of heterozygosity (LOH), and DNA methylation using the same array [70]. After using *XbaI* to shear the DNA and a methylation restriction enzyme (such as *HpaII*) to enrich for methlyated fragments, this method (called MSNP) separates SNPs into three groups depending upon whether an *XbaI* (DNA) fragment contains a *HpaII* binding site and whether this binding site might have been eliminated/caused by the presence of a polymorphism. Subsequently, *XbaI* fragments not containing a

HpaII binding site are used to measure copy number, and fragments containing a *HpaII* binding site are used to measure methylation.

While this approach has the advantage that information about a number of different genetic features can be determined from the same array, it also has a number of drawbacks. In particular, it has been observed [33] that multiple methylation restriction enzymes have to be used for DNA methylation to be measured at a sufficient number of locations across the genome—this will significantly reduce the number of SNPs that can be used to determine copy number. Additionally, it is not possible to detect copy number at the same SNPs where methylation has been measured, which means that the resolution may be insufficient to analyse both methylation and copy number separately. Consequently, perhaps the best way of confidently determining copy number and DNA methylation level is to carry out two different hybridizations, one using DNA sheared using only *XbaI* (or a similar enzyme) and the other where the DNA has been treated in such a way that methylation can be detected.

13.8 VALIDATION CHOICES

Validating data obtained from microarray experiments is essential. To confirm that genes with known methylation levels are correctly identified, it is necessary to select a number of genes and compare the results obtained in the array experiment with methylation analyses performed using a different technology. To analyze a particular gene, knowledge of its structure and sequence is required since (in most cases) its methylation status will be examined in only a small region, usually in CpG islands or CpG-rich regions near the gene promoter. The examination of such regions is typically performed using methods based upon bisulfite conversion. Bisulfite sequence analysis is performed by treating DNA with sodium bisulfite which results in the deamination of unmethylated cytosine to uracil, while leaving methylated cytosine residues unchanged [71]. DNA sequencing can then be used to identify methylated cytosines with the exercise being reduced to differentiating between SNPs (cytosines vs thymines).

Variations of bisulfite sequence analyses offer the opportunity to examine a number of CpGs simultaneously and can be scaled up to assay multiple sample sets. In this section, we give a brief description of some analysis methods that use bisulfite conversion in conjunction with PCR to interrogate the methylation status of a small genomic region. For a more detailed review see [72].

A number of commercially available kits for bisulfite conversion are available and are continually being improved with regard to the yield and stability of bisulfite converted DNA, enabling longer amplicons (typically up to 700 bp) to be obtained from small amounts of starting material. However, the limiting step in bisulfite sequence analyses is the conversion process itself, which results in significant DNA degradation such that 84–96% of DNA is affected [73]. While numerous attempts have been made to optimize bisulfite treatment by striking a balance between achieving complete cytosine conversion and minimal DNA degradation [73–75], degradation remains an issue. Thus, it is important to determine the amount of degradation that occurs during specific reaction conditions and to consider the effect on the amplicon of interest, and a recent method towards this end has been described [76]. Fragmentation not only sets an empirical upper size limit on the PCR amplicon (~400–500 bp), but the longer the amplicon, the fewer intact templates there will likely be. In addition, bisulfite treatment results in reduced sequence diversity, generating AT-rich regions and long stretches of thymines, which can be difficult for polymerases to read. Thus, PCR amplification of bisulfite-treated DNA can be challenging and requires careful primer design to avoid mispriming and primer–dimer formation.

13.8.1 Bisulfite Sequencing

Conventional bisulfite sequencing consists of amplifying a specific region of interest and then sequencing the PCR products. To simplify the sequence analysis, PCR products are cloned into bacterial plasmids, and single clones, each representing one PCR amplicon, are sequenced. This

approach enables a number of adjacent CpGs (up to 50) to be analyzed on a single amplicon and, by analyzing multiple clones of the same sample, sample-specific profiles can be generated. For an application of bisulfite sequencing to the study of cell division, see Ref. [77]. This approach could be made allele-specific if combined with SNP detection (provided that the SNP is not masked by the bisulfite conversion), and is still the method of choice for examining imprinted genes. The disadvantage of this method is that cloning PCR products is time-consuming, and the cost of sequencing multiple clones of the same PCR reaction can be high.

Bioinformatics software that enables direct quantitative analyses from sequence traces is emerging [78]. These programs allow for direct sequencing of complex mixes of amplicons generated from a single PCR. An obvious advantage of sequencing over other methods is that single CpG profiles can be generated for a locus of interest. In addition, sequencing can be outsourced and thus expensive equipment need not be purchased.

13.8.2 Methylation-specific PCR/Quantitative Methylation-Dependent PCR (MethyLight)

Methylation-specific PCR (MSP), employs methylation-specific primers that exploit the sequence differences in methylated versus unmethylated bisulfite-treated DNA at a particular locus [79]. Parallel amplification reactions using unmethylation-specific primers should also be performed for each DNA sample. Thus, methylation is determined by the ability of specific primers to allow for amplification. The PCR products can be examined following nondenaturing polyacrylamide gel electrophoresis and ethidium bromide staining such that the presence of a band of the appropriate molecular weight indicates the methylation status of the allele in the original sample. Such products may be compared but, due to variations in PCR efficiency with different primers, quantitative interpretation should be cautious. Specificity can be improved by designing primers that anneal to multiple CpG sites or with the CpG dinucleotide at the 3′ end of the primer. This approach is most useful for querying densely methylated CpG islands.

A variation on this approach allows for more quantitative measures of methylation by employing real-time quantitative PCR and is referred to as the MethyLight method. As in MSP, methylation-specific primers are employed, but a methylation-specific fluorescence reporter probe that anneals to the amplified region of interest is also incorporated. Quantification is performed based on methylated reference sequences that are included on each plate to control for plate-to-plate variations and involves several optimization steps such as the generation of standard curves. Although these assays can be scaled up to quite high throughput, the number of CpGs that can be assayed depends on the probe and generally these are designed based on the assumption that all CpGs within the region queried share the same methylation status.

13.8.3 Combined Bisulfite Restriction Analysis

Combined bisulfite restriction analysis (COBRA) is a low-throughput method that determines the level of methylation at specific genomic loci. After DNA has been bisulfite-treated, the region of interest is amplified using primers that do not span CpG sites to generate an amplicon that contains a CG recognizing restriction site. Digestion of the PCR product with an appropriate restriction enzyme results in the digestion of only those products that have unmodified (i.e., methylated) cytosine in the CpG. For example, BstUI's recognition sequence, CGCG, if methylated would be unchanged after bisulfite modification. But if the recognition sequence were unmethylated, bisulfite modification would change it into TGTG and the PCR product would not be restricted [80].

A drawback of this method is that since a single restriction site is analyzed, if more than one CpG is present in the amplicon, these may or may not be identically methylated. Consequently, the PCR product could be a complex mixture of various amplicons which could impact upon the efficiency of the digestion step. Additionally, during the melting and annealing steps of PCR, heteroduplexes

of methylated and unmethylated PCR strands that are resistant to restriction digestion may form, and therefore give false-negative results. False-positive results can occur when there is incomplete bisulfite conversion.

13.8.4 Methylation-Sensitive Single Nucleotide Primer Extension

Methylation-sensitive single nucleotide primer extension (Ms-SNuPE) is an adaptation of a technique originally designed for the analysis of SNPs in the context of mutation detection [81] and for the quantification of allele-specific expression [82,83]. Essentially, this approach employs paired primer extensions such that the Ms-SNuPE primers anneal to the PCR-generated template and subsequently terminate 5′ the cytosine residue to be queried [84,85]. In this application, bisulfite-treated DNA is PCR amplified and the gel excised PCR product is incubated with the appropriate Ms-SNuPE primers, polymerase, and radiolabeled dNTPs. The incorporation of [^{32}P]dCTP or [^{32}P]dTTP is then assessed following denaturing polyacrylamide gel-electrophoresis and phosphorimaging, and is used to determine the relative amounts of methylated (C) versus unmethylated (T) cytosines at the original CpG site. Similarly, the complementary strand can be queried using primers designed to incorporate either [^{32}P]dATP or [^{32}P]dGTP. Nonradioactive fluorescent labeling and quantification schemes can also be adapted to this assay. This approach allows for the simultaneous analysis of several CpG dinucleotides in a single reaction and provides a quantitative readout of the ratio of methylated to unmethylated cytosines at a particular CpG site.

Adaptations of this approach employ matrix-assisted laser desorption ionization/time-of-flight (MALDI-TOF) mass spectrometry to discriminate between the two primer extension products based on the GOOD assay for SNP analysis.

13.8.5 Pyrosequencing

Pyrosequencing is a sequencing by synthesis method that offers rapid and accurate quantification of CpG methylation sites [86]. After bisulfite conversion and PCR amplification, a sequence-specific primer is hybridized to the strand to be interrogated. The nucleotides are dispensed sequentially according to the predicted sequence which is programed into the pyrosequencer (Biotage). Each time a nucleotide is incorporated into a sequence, pyrophosphate (PPi) is released, and this energy is used for the enzymatic conversion of luciferin to oxyluciferin. This generates light in proportion to the released PPi, which is captured on a CCD camera and recorded as a peak. Q1
Before the addition of the next nucleotide in the sequence, a nucleotide-degrading enzyme or apyrase such as uracil *N*-glycosylase (UNG) is employed to remove excess nucleotides. The result is synchronized nucleotide addition or real-time quantitative sequencing such that the amount of cytosine and thymine incorporation during extension can be used to quantitatively determine the C/T ratio at specific loci. At present this technology is expensive and still being improved. The potential advantages are its accuracy, high-throughput nature, and the minimization of variation between experiments.

13.9 CONCLUSIONS

Using high-throughput microarrays to interrogate DNA methylation on a genome-wide level is an exciting research area that could yield important insights into many biological problems. However, as described in this chapter, there are many outstanding technological and analytical issues that have to be resolved so that the investigator can have more confidence that the data obtained are of "good" quality. Determining data quality might be made easier by the Human Epigenome Project [22] which uses low-throughput technology to assess the methylation status of each base of the genome for a number of individuals—this will provide a rich test dataset on which the efficacy of

technologies and analysis methods might be assessed. Besides testing the performance of these methods, another important consideration will be how to improve the effective resolution of DNA methylation microarrays. This will be crucial if we want to combine data from DNA methylation arrays with other microarray technologies (e.g., expression, array CGH, ChIP-chip, or micro RNA), since such a comparison will be limited by the effective resolutions of the data obtained. Such combined analysis will be essential to improve our understanding of how different genetic phenomena interact and contribute to (for example) biological development and tumorigenesis.

REFERENCES

1. Doerfler, W. 1983. DNA methylation and gene activity. *Annu Rev Biochem*, 52, 93–124.
2. Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev*, 16, 6–21.
3. Finnegan, E. J., Peacock, W. J., and Dennis, E. S. 2000. DNA methylation, a key regulator of plant development and other processes. *Curr Opin Genet Dev*, 10, 217–223.
4. Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *J Mol Biol*, 196, 261–282.
5. Yoder, J. A., Walsh, C. P., and Bestor, T. H. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*, 13, 335–340.
6. Baylin, S. B. 2005. DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol*, 2, S4–S11.
7. Bird, A. P. 1986. CpG-rich islands and the function of DNA methylation. *Nature*, 321, 209–213.
8. Song, F., Smith, J. F., Kimura, M. T., Morrow, A. D., Matsuyama, T., Nagase, H., and Held, W. A. 2005. Association of tissue-specific differentially methylated regions (tdms) with differential gene expression. *Proc Natl Acad Sci U S A*, *102*, 3336–3341.
9. Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. 1992. CpG islands as gene markers in the human genome. *Genomics*, 13, 1095–1107.
10. Cross, S. H., Charlton, J. A., Nan, X., and Bird, A. P. 1994. Purification of CpG islands using a methylated DNA binding column. *Nat Genet*, 6, 236–244.
11. Laird, P. W. 2005. Cancer epigenetics. *Hum Mol Genet*, 14 Spec No 1, R65–R76.
12. Bestor, T. H. 1992. Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *EMBO J*, 11, 2611–2617.
13. Pradhan, S., Bacolla, A., Wells, R. D., and Roberts, R. J. 1999. Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *J Biol Chem*, 274, 33002–33010.
14. Robertson, K. D. 2005. DNA methylation and human disease. *Nat Rev Genet*, 6, 597–610.
15. Robertson, K. D. 2002. DNA methylation and chromatin—unraveling the tangled web. *Oncogene*, 21, 5361–5379.
16. Costello, J. F. and Plass, C. 2001. Methylation matters. *J Med Genet*, 38, 285–303.
17. Esteller, M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet*, 8, 286–298.
18. Esteller, M., Corn, P. G., Baylin, S. B., and Herman, J. G. 2001. A gene hypermethylation profile of human cancer. *Cancer Res*, 61, 3225–3229.
19. Feinberg, A. P., Ohlsson, R., and Henikoff, S. 2006. The epigenetic progenitor origin of human cancer. *Nat Rev Genet*, 7, 21–33.
20. Feinberg, A. P. and Vogelstein, B. 1983. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, 301, 89–92.
21. Paz, M. F., Fraga, M. F., Avila, S., Guo, M., Pollan, M., Herman, J. G., and Esteller, M. 2003. A systematic profile of DNA methylation in human cancer cell lines. *Cancer Res*, 63, 1114–1121.
22. Murrell, A., Rakyan, V. K., and Beck, S. 2005. From genome to epigenome. *Hum Mol Genet*, 14 Spec No 1, R3–R10.
23. Callinan, P. A. and Feinberg, A. P. 2006. The emerging science of epigenomics. *Hum Mol Genet*, 15 Spec No 1, R95–R101.
24. Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. 2006. Microarray data analysis: From disarray to consolidation and consensus. *Nat Rev Genet*, 7, 55–65.
25. Pinkel, D. and Albertson, D. G. 2005. Comparative genomic hybridization. *Annu Rev Genomics Hum Genet*, 6, 331–354.

26. Taylor, K., Kramer, R., Davis, J., Guo, J., Duff, D., Xu, D., Caldwell, C., and Shi, H. 2007. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res*, 67, 8511–8518.
27. Huang, T. H., Perry, M. R., and Laux, D. E. 1999. Methylation profiling of CpG islands in human breast cancer cells. *Hum Mol Genet*, 8, 459–470.
28. Wei, S. H., Chen, C.-M., Strathdee, G., Harnsomburana, J., Shyu, C.-R., Rahmatpanah, F., Shi, H., et al. 2002. Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers. *Clin Cancer Res*, 8, 2246–2252.
29. Yan, P. S., Perry, M. R., Laux, D. E., Asare, A. L., Caldwell, C. W., and Huang, T. H. 2000. CpG island arrays: An application toward deciphering epigenetic signatures of breast cancer. *Clin Cancer Res*, 6, 1432–1438.
30. Yan, P. S., Chen, C. M., Shi, H., Rahmatpanah, F., Wei, S. H., Caldwell, C. W., and Huang, T. H. 2001. Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res*, 61, 8375–8380.
31. Yan, P. S., Chen, C.-M., Shi, H., Rahmatpanah, F., Wei, S. H., and Huang, T. H.-M. 2002. Applications of CpG island microarrays for high-throughput analysis of DNA methylation. *J Nutr*, 132, 2430S–2434S.
32. Yan, P. S., Efferth, T., Chen, H.-L., Lin, J., Rodel, F., Fuzesi, L., and Huang, T. H.-M. 2002. Use of CpG island microarrays to identify colorectal tumors with a high degree of concurrent methylation. *Methods*, 27, 162–169.
33. Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., et al. 2006. Microarray-based DNA methylation profiling: Technology and applications. *Nucleic Acids Res*, 34, 528–542.
34. Ibrahim, A. E., Thorne, N. P., Baird, K., Barbosa-Morais, N. L., Tavaré, S., Collins, V. P., Wyllie, A. H., Arends, M. J., and Brenton, J. D. 2006. MMASS: An optimized array-based method for assessing CpG island methylation. *Nucleic Acids Res*. Epub ahead of print. **Q4**
35. Glonok, G. F. and Solomon, P. J. 2004. Factorial and time course designs for cDNA microarray experiments. *Biostatistics*, 5, 89–111.
36. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 4, e15.
37. McClelland, M., Nelson, M., and Raschke, E. 1994. Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. *Nucleic Acids Res*, 22(17), 3640–3659.
38. McClelland, M. 1983. The effect of site specific methylation on restriction endonuclease cleavage (update). *Nucleic Acids Res*, 11, r169–r173.
39. Nouzova, M., Holtan, N., Oshiro, M. M., Isett, R. B., Munoz-Rodriguez, J. L., List, A. F., Narro, M. L., Miller, S. J., Merchant, N. C., and Futscher, B. W. 2004. Epigenomic changes during leukemia cell differentiation: Analysis of histone acetylation and cytosine methylation using CpG island microarrays. *J Pharmacol Exp Ther*, 311, 968–981.
40. Toyota, M., Ho, C., Ahuja, N., Jair, K. W., Li, Q., Ohe-Toyota, M., Baylin, S. B., and Issa, J. P. 1999. Identification of differentially methylated sequences in colorectal cancer by methylated CpG island amplification. *Cancer Res*, 59, 2307–2312.
41. Lisitsyn, N., Lisitsyn, N., and Wigler, M. 1993. Cloning the differences between two complex genomes. *Science*, 259, 946–951.
42. Ching, T. T., Maunakea, A. K., Jun, P., Hong, C., Zardo, G., Pinkel, D., Albertson, D. G., et al. 2005. Epigenome analyses using BAC microarrays identify evolutionary conservation of tissue-specific methylation of SHANK3. *Nat Genet*, 37, 645–651.
43. Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., and 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.
44. Mukhopadhyay, R., Yu, W., Whitehead, J., Xu, J., Lezcano, M., Pack, S., Kanduri, C., et al. 2004. The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res*, 14, 1594–1602.
45. Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., et al. 2006. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat Genet*, 38, 149–153.
46. Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., Henderson, I. R., et al. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, 126, 1189–1201.

47. Inamdar, N. M., Ehrlich, K. C., and Ehrlich, M. 1991. CpG methylation inhibits binding of several sequence-specific DNA-binding proteins from pea, wheat, soybean and cauliflower. *Plant Mol Biol*, 17, 111–123.
48. Cameron, E. E., Bachman, K. E., Myohanen, S., Herman, J. G., and Baylin, S. B. 1999. Synergy of demethylation and histone deacetylase inhibition in the re-expression of genes silenced in cancer. *Nat Genet*, 21, 103–107.
49. Hendrich, B. and Bird, A. 1998. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol*, 18, 6538–6547.
50 Zhang, Y., Ng, H. H., Erdjument-Bromage, H., Tempst, P., Bird, A., and Reinberg, D. 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev*, 13, 1924–1935.
51. Shiraishi, M., Chuu, Y. H., and Sekiya, T. 1999. Isolation of DNA fragments associated with methylated CpG islands in human adenocarcinomas of the lung using a methylated DNA binding column and denaturing gradient gel electrophoresis. *Proc Natl Acad Sci U S A*, 96, 2913–2918.
52. Brock, G. J., Huang, T. H., Chen, C. M., and Johnson, K. J. 2001. A novel technique for the identification of CpG islands exhibiting altered methylation patterns (ICEAMP). *Nucleic Acids Res*, 29, E123.
53. Rauch, T. and Pfeifer, G. P. 2005. Methylated-CpG island recovery assay: A new technique for the rapid detection of methylated-CpG islands in cancer. *Lab Invest*, 85, 1172–1180.
54. Rauch, T., Li, H.,Wu, X., and Pfeifer, G. P. 2006. MIRA-assisted microarray analysis, a new technology for the determination of DNA methylation patterns, identifies frequent methylation of homeodomain-containing genes in lung cancer cells. *Cancer Res*, 66, 7939–7947.
55. Klose, R. J., Sarraf, S. A., Schmiedeberg, L., McDermott, S. M., Stancheva, I., and Bird, A. P. 2005. DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol Cell*, 19, 667–678.
56. Jorgensen, H. F., Adie, K., Chaubert, P., and Bird, A. P. 2006. Engineering a high-affinity methyl-CpG-binding protein. *Nucleic Acids Res*, 34, e96.
57 Yegnasubramanian, S., Lin, X., Haffner, M. C., DeMarzo, A. M., and Nelson, W. G. 2006. Combination of methylated-DNA precipitation and methylation-sensitive restriction enzymes (COMPARE-MS) for the rapid, sensitive and quantitative detection of DNA methylation. *Nucleic Acids Res*, 34, e19.
58. Bibikova, M., Lin, Z., Zhou, L., Chudin, E., Garcia, E. W., Wu, B., Doucet, D., et al. 2006. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res*, 16, 383–393.
59. Gunderson, K. L., Kruglyak, S., Graige, M. S., Garcia, F., Kermani, B. G., Zhao, C., Che, D., et al. 2004. Decoding randomly ordered DNA arrays. *Genome Res*, 14, 870–877.
60. Weber, M., Hellmann, I., Stadler, M., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, 39, 457–466.
61. Ylstra, B., van den Ijessel, P., Carvalho, B., Brakenhoff, R. H., and Meijer, G. A. 2006. BAC to the future! or oligonucleotides: A perspective for micro array comparative genomic hybridisation (array CGH). *Nucleic Acids Res*, 34, 445–450.
62. Egeland, R. D. and Southern, E. M. 2005. Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. *Nucleic Acids Res*, 33, e125.
63. Leiske, D. L., Karimpour-Fard, A., Hume, P. S., Fairbanks, B. D., and Gill, R. T. 2006. A comparison of alternative 60-mer probe designs in an in-situ synthesized oligonucleotide microarray. *BMC Genomics*, 7, 72.
64. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. 2002. Variance stabilization applied to micrarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18, S96–S104.
65. Bengtsson, H. and Hossjer, O. 2006. Methodological study of affine transformations of gene expression data with proposed robust non-parametric multi-dimensional normalization method. *BMC Bioinformatics*, 7, 100.
66. Buck, M. J. and Lieb, J. D. 2004. ChIP-chip: Considerations for the design, analysis and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83, 349–360.
67. Marioni, J. C., Thorne, N. P., and Tavaré, S. 2006. BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, 22, 1144–1146.
68. Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. 2004. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5, 557–572.
69. Qi, A., Rolfe, P. A., MacIsaac, K., Gerber, G. K., Pokholok, D., Zeitlinger, J., Danford, T., et al. 2006. High-resolution computational models of genome binding events. *Nat Biotechnol*, 24, 963–970.

70. Yuan, E., Haghighi, F., White, S., Costa, R., McMinn, J., Chun, K., Minden, M., and Tycko, B. 2006. A single nucleotide polymorphism chip-based method for combined genetic and epigenetic profiling: Validation in decitabine therapy and tumor/normal comparisons. *Cancer Res*, 66, 3443–3451.
71. Frommer, M., McDonald, L., Millar, D., Collis, C., Watt, F., Grigg, G., Molloy, P., and Paul, C. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A*, 89, 1827–1831.
72. Fraga, M. F. and Esteller, M. 2002. DNA methylation: A profile of methods and applications. *Biotechniques*, 33, 632–649.
73. Grunau, C., Clark, S., and Rosenthal, A. 2001. Bisulfite genomic sequencing: Systematic investigation of critical experimental parameters. *Nucleic Acids Res*, 29, E65–E65. **Q5**
74. Raizis, A., Schmitt, F., and Jost, J. 1995. A bisulfite method of 5-methylcytosine mapping that minimizes template degradation. *Anal Biochem*, 226, 161–166.
75. Olek, A., Oswald, J., and Walter, J. 1996. A modified and improved method for bisulphite based cytosine methylation analysis. *Nucleic Acids Res*, 24, 5064–5066.
76. Ehrich, M., Zoll, S., Sur, S., and van den Boom, D. 2007. A new method for accurate assessment of DNA quality after bisulfite treatment. *Nucleic Acids Res*, 35, e29.
77. Shibata, D. and Tavaré, S. 2006. Counting divisions in a human somatic cell tree: How, what and why? *Cell Cycle*, 5, 610–614.
78. Lewin, J., Schmitt, A. O., Adorjan, P., Hildmann, T., and Piepenbrock, C. 2004. Quantitiative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics*, 20, 3005–3012.
79. Herman, J., Graff, J., Myöhänen, S., Nelkin, B., and Baylin, S. 1996. Methylation-specific PCR: A novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A*, 93, 9821–9826.
80. Xiong, Z. and Laird, P. W. (1997). COBRA: A sensitive and quantitative DNA methylation assay. *Nucleic Acids Res*, 25, 2532–2534.
81. Kuppuswamy, M., Hoffmann, J., Kasper, C., Spitzer, S., Groce, S., and Bajaj, S. 1991. Single nucleotide primer extension to detect genetic diseases: Experimental application to hemophilia B (factor IX) and cystic fibrosis genes. *Proc Natl Acad Sci U S A*, 88, 1143–1147.
82. Singer-Sam, J., LeBon, J., Dai, A., and Riggs, A. 1992. A sensitive, quantitative assay for measurement of allele-specific transcripts differing by a single nucleotide. *PCR Methods Appl*, 1, 160–163.
83. Szabó, P. and Mann, J. 1995. Allele-specific expression and total expression levels of imprinted genes during early mouse development: Implications for imprinting mechanisms. *Genes Dev*, 9, 3097–3108.
84. Gonzalgo, M. and Jones, P. 1997. Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Res*, 25, 2529–2531.
85. Gonzalgo, M. and Liang, G. 2007. Methylation-sensitive single-nucleotide primer extension (Ms-SNuPE) for quantitative measurement of DNA methylation. *Nat Protoc*, 2, 1931–1936.
86. Tost, J. and Gut, I. 2007. DNA methylation analysis by pyrosequencing. *Nat Protoc*, 2, 2265–2275.
87. Hoheisel, J. D. 2006. Microarray technology: Beyond transcript profiling and genotype analysis. *Nat Rev Genet*, 7, 200–210.