Mining cancer methylomes: prospects and challenges

Clare Stirzaker[1,2*], **Phillippa C. Taberlay**[1,2*], **Aaron L. Statham**[1], **and Susan J. Clark**[1,2]

[1] Epigenetics Program, Garvan Institute of Medical Research, The Kinghorn Cancer Centre, Sydney 2010, NSW, Australia
[2] St Vincent's Clinical School, University of NSW, Sydney 2010, NSW, Australia

There are over 28 million CpG sites in the human genome. Assessing the methylation status of each of these sites will be required to understand fully the role of DNA methylation in health and disease. Genome-wide analysis, using arrays and high-throughput sequencing, has enabled assessment of large fractions of the methylome, but each protocol comes with unique advantages and disadvantages. Notably, except for whole-genome bisulfite sequencing, most commonly used genome-wide methods detect <5% of all CpG sites. Here, we discuss approaches for methylome studies and compare genome coverage of promoters, genes, and intergenic regions, and capacity to quantitate individual CpG methylation states. Finally, we examine the extent of published cancer methylomes that have been generated using genome-wide approaches.

DNA methylation and (de)regulation of the epigenome

Epigenetic regulation (see Glossary) of normal cellular processes is typically driven in a cell type-dependent manner, requiring a complex interplay between different layers of epigenetic information, including DNA methylation, nucleosome positions, histone modifications, and expression of noncoding RNA. Several epigenetic mechanisms help establish and consolidate the correct higher-order chromatin structures and gene-expression patterns during differentiation and development. Of these, DNA methylation is the best-studied epigenetic modification in mammals. Precise DNA methylation patterns are established during embryonic development and are mitotically inherited through multiple cellular divisions. DNA methylation is necessary for normal cell development [1,2], underpinning X chromosome inactivation [3,4], control of some tissue-specific gene expression, and regulation of imprinted alleles [2,5,6], with widespread effects on cellular growth and genomic stability [7–9].

DNA methylation in mammalian cells is characterized by the addition of a methyl group at the carbon-5 position of cytosine residues within CpG dinucleotides through the action of DNA methyltransferase enzymes, forming 5-methylcytosine (5MeC) [10]. There are approximately 28 million CpG sites in the genome, but these are not evenly distributed; in fact, the bulk of the genome is depleted of CpG sites with less than one quarter of the expected frequency. By contrast, clusters of CpG sites occur at the expected frequency, termed 'CpG islands', and these commonly span promoters of house-keeping genes. Promoter CpG islands typically remain unmethylated in normal cells and are associated with active gene expression during differentiation (CpG island, promoter; Figure 1). By contrast, methylated CpG island promoters are associated with gene repression. Regions of intermediate CpG densities also exist across the genome, often in the body of genes. Unlike CpG island promoters, extensive exonic or genic methylation is typically associated with active gene expression (genic; Figure 1). CpG island 'shores' are regions of comparatively low CpG density, located approximately 2 kb from CpG islands [11]. Shores also exhibit tissue- and cancer-specific differential methylation and are associated with gene repression [12]. Beyond CpG islands and shores, the remainder of genome displays a lower than expected frequency of CpG sites and is typically methylated in normal cells (intergenic; Figure 1). This includes CpG-poor promoters and distal enhancers that regulate tissue-specific genes (tissue specific; Figure 1).

Despite extensive knowledge of DNA methylation events, the underlying biology largely remains an enigma, particularly the mechanism by which it is altered in diseased states, such as cancer. Normal epigenetic processes are disrupted during the initiation and progression of

Glossary

Bisulfite genomic sequencing: sequencing of bisulfite-treated DNA allowing resolution of the methylation state of every cytosine in the target sequence, at single-molecule resolution. This is considered the 'gold standard' for DNA methylation analysis.
Bisulfite modification: exploits the different sensitivities of cytosine and 5-meC to deamination by bisulfite under acidic conditions in which cytosine undergoes conversion to uracil, whereas 5-meC remains unreactive.
Cancer methylome: the map of DNA methylation across a cancer cell genome.
DNA methylation: the addition of a methyl CH_3 group to the cytosine base at the carbon 5 position (5-meC) in DNA; found primarily in the context of CpG dinucleotides in eukaryotes.
Epigenetic mechanisms: the mechanisms that govern the role of epigenetics in gene expression without changing the underlying DNA sequence; include chromatin structure, histone modifications, nucleosome positioning, and DNA methylation.
Epigenetics: the study of changes in gene expression or phenotype of a cell, caused by mechanisms other than changes in the underlying DNA sequence.
Methylome: the genome-wide map of DNA methylation.

Corresponding author: Clark, S.J. (s.clark@garvan.org.au).
Keywords: cancer methylome; epigenetics; DNA methylation.
*Equal first authors.

0168-9525/$ – see front matter
http://dx.doi.org/10.1016/j.tig.2013.11.004

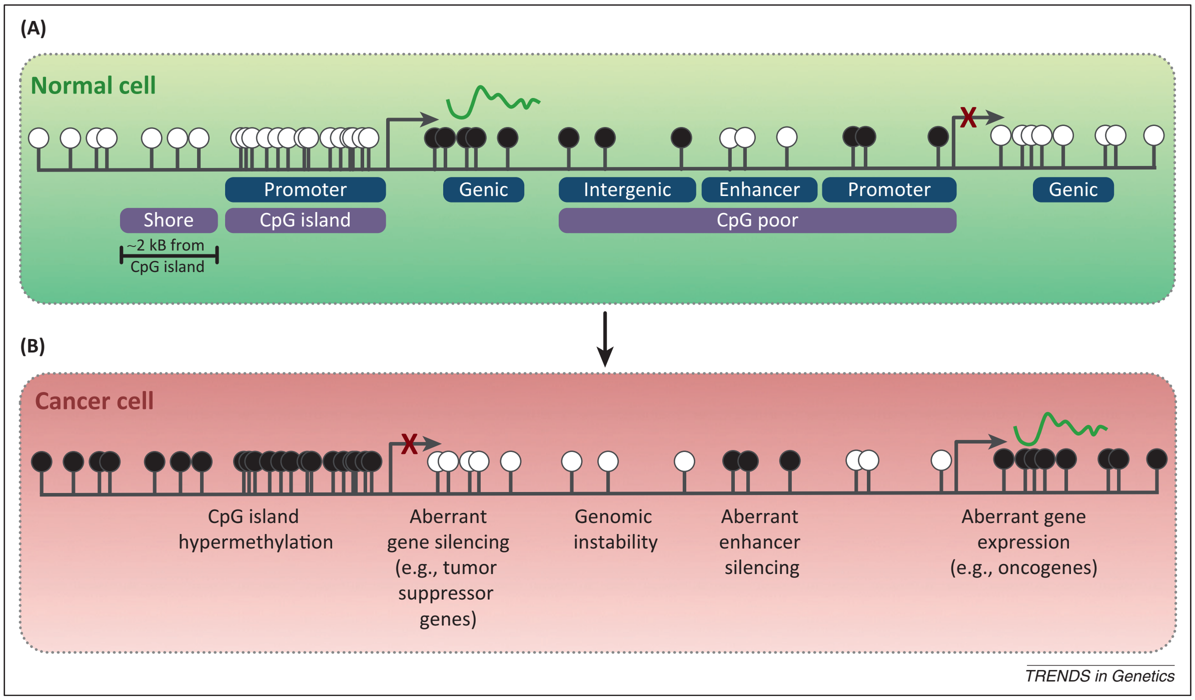


Figure 1. DNA methylation and (de)regulation of the genome. A schematic representation of the methylome and a summary of major changes that occur in cancer cells. CpG islands are often associated with gene promoters and are resistant to DNA methylation in normal cells **(A)** (green). Gene expression can occur, and is highly correlated with high levels of gene body (genic) methylation. CpG-poor regions (intergenic), with the exception of enhancers, are typically methylated in normal cells. Similarly, CpG-poor promoters are silenced by DNA methylation and exhibit a closed chromatin structure unless gene expression is required (tissue specific). In cancer cells **(B)**, CpG islands are prone to DNA hypermethylation, which results in aberrant gene silencing (e.g., of tumor suppressor genes). Concomitant hypomethylation of intergenic regions and CpG-poor promoters contributes to genomic instability and aberrant gene expression (e.g., of oncogenes), respectively. White circle, unmethylated CpG; black circle, methylated CpG.

cancer, including global changes in DNA methylation patterns [13]. CpG island hypermethylation is common and often associated with the silencing of tumor suppressor genes and downstream signaling pathways [13–16] (Figure 1). Whereas CpG islands become susceptible to DNA methyltransferase activity, CpG-poor regions undergo hypomethylation during transformation, resulting in an overall decrease in total genomic 5MeC in cancer cells [13,14,16] (Figure 1). The exception includes CpG-poor, distal enhancers that are unmethylated in normal cells but often gain methylation [17,18] in cancer cells (Figure 1). Global hypomethylation in cancer is thought to contribute to genomic instability and aberrant expression of some oncogenes, such as *MYC* [19] (Figure 1), which results in deregulation of cellular processes.

The opportunity now exists to provide more comprehensive maps of cancer DNA methylomes using whole genome-based technologies [20–25]. These technologies will help provide greater insight into the underlying mechanism and location of cancer-specific methylation changes at individual CpG residues and may aid in further identification of potential epigenetic-based cancer biomarkers.

Genome-wide methylome technologies

DNA methylation analyses were initially restricted to relatively localized CpG-rich regions of the genome, but several methods have now been developed to map DNA methylation on a genomic scale. Here, we describe four different genome-wide approaches (summarized in Figure 2): whole-genome bisulfite sequencing (WGBS); methyl-binding domain capture sequencing (MBDCap-Seq); reduced-representation-bisulfite-sequencing (RRBS); and Infinium HumanMethylation450 BeadChips (HM450, Illumina). We discuss some of the requirements, merits, and challenges that should be considered when choosing a methylome technology to ensure that it will be informative.

Whole-genome bisulfite sequencing

Bisulfite-sequencing, which was developed in 1992–1994 by Frommer and Clark [26,27], is considered the 'gold standard' for DNA methylation analyses because CpG methylation can be measured at single-base resolution. DNA is treated with sodium bisulfite to convert cytosine to uracil, which is converted to thymine after PCR amplification, whereas 5MeC residues are not converted and remain as cytosines [27]. Clonal sequencing of bisulfite-converted PCR products from a single genomic region have typified the approach until recently; however, the development of high-throughput sequencing now facilitates the generation of genome-wide, single-base resolution DNA methylation maps from bisulfite-converted DNA (Figure 2). To perform WGBS, genomic DNA (1–5 μg) is sheared and ligated to methylated adaptors before size selection and bisulfite conversion, followed by library construction and high-throughput sequencing (Figure 2). More than 500 million paired-end reads are required to achieve approximately 30-fold coverage of the 28 217 009 CpG sites on autosomes and sex chromosomes; typically approximately 95% of all CpG sites in the genome can be assessed using WBGS. The first methylome was generated from the *Arabidopsis thaliana* genome in 2008 [28,29], and the first human methylomes of embryonic stem cells and IMR90 fibroblasts were

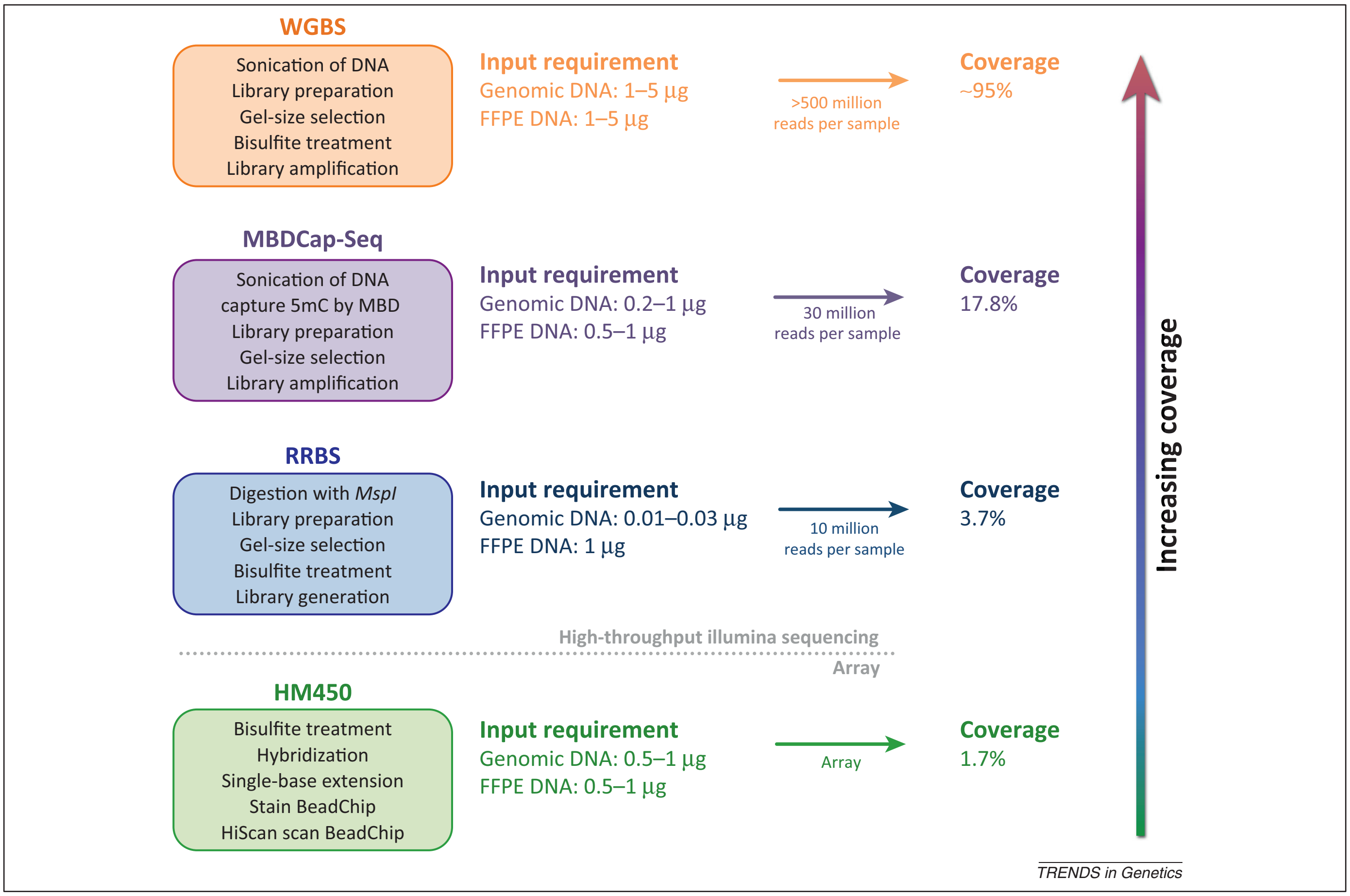


Figure 2. Summary of techniques to interrogate whole-genome DNA methylation. The figure compares the maximum coverage of whole-genome bisulfite sequencing (WGBS, orange; most genomic coverage), MBD capture sequencing (MBDCap-Seq, purple), reduced representation bisulfite sequencing (RRBS, blue), and HumanMethylation450 BeadChip (HM450, green) assays for measuring genome-wide DNA methylation. A summary of the standard workflow for each method is shown (colored boxes, left). The amount of genomic DNA or formalin-fixed paraffin-embedded tissue (FFPE) needed to perform each technique reliably ranges from 0.01 ug (RRBS) to 5 ug (WGBS), which may influence platform selection. The minimum number of unique sequencing reads varies from 10 million reads (RRBS) to >500 million reads (WGBS), whereas the HM450 platform utilizes array technology. Therefore, the cost of each technique is approximately proportional to the amount of data needed to analyze reliably the data, and the coverage of the genome [range, 1.7% (HM450) – ~95% (WBGS)].

reported by Lister *et al.* in 2009 [22]. To date, relatively few WGBS human cancer [30–32] or related [33] methylomes have been generated, likely due to the overall cost of the assay, technical expertise, and downstream computational requirements.

WGBS has the advantage of providing single-nucleotide resolution and whole-genome coverage. However, it typically requires relatively large quantities of DNA (1–5 ug) and accurate interpretation requires computational expertise. Commercial bisulfite conversion reagents exist in kit form; yet, standard WGBS protocol/s or library preparation methods are just beginning to emerge. Sequencing providers are performing WGBS using customized in-house methods, but the technique currently is not particularly amenable to high-throughput use, particularly in a clinical setting, partly due to the extensive hands-on and depth of sequencing required. Finally, the bioinformatics requirements for data interpretation present additional challenges. Initial WGBS studies relied upon in-house adaptations of genome sequencing pipelines to bisulfite data and unpublished bespoke analysis pipelines [22,34–36]; however, public tools for the analysis of WGBS data are being developed as the technique becomes more accessible.

Enrichment-based technologies

Genome-wide affinity-based methods rely on enrichment of methylated regions, followed by microarray hybridization or next-generation sequencing (Figure 2). Two of the common enrichment approaches include methyl-DNA immunoprecipitation (MeDIP), which uses a monoclonal antibody specific for 5-methylcytosine [37] and affinity capture with MBDCap proteins [38,39]. Both MeDIP and MBDCap can be combined with next-generation sequencing (MeDIP-Seq and MBDCap-Seq). However, due to bias in the different capture technologies, distinctive genomic regions are commonly interrogated [40]. MeDIP is based on immunoprecipitation of single-stranded DNA fragments and targets methylated regions of low CpG density (e.g., intergenic regions). By contrast, the MBD-based strategy captures double-stranded methylated DNA fragments and favors enrichment of CpG-dense regions (e.g., CpG islands) [41].

Here, we highlight MBDCap-Seq as one of the most widely used capture approaches. The workflow for MBDCap-Seq exhibits similarities to WGBS, but is devoid of a bisulfite conversion step (Figure 2). To perform MBDCap-Seq, genomic DNA (0.2–1 µg) is sonicated before capturing methylated DNA with MBD protein coupled to streptavidin beads. Following capture, the bound methylated DNA can be eluted as a single fraction or in a step-wise elution series to enrich different CpG densities. Enriched DNA is then subjected to library preparation and high-throughput sequencing (Figure 2). Although the method is more efficient with amounts of >~0.2-μg DNA from fresh-frozen tissue, genomic DNA preparations for cancer methylomes can also be isolated from formaldehyde-fixed paraffin embedded tissue (FFPET), and is amenable to MBDCap-seq using as little as approximately 0.5 μg of DNA. Approximately 30 million single-end reads are required for accurate interpretation of data. MBDCap-Seq performed on fully methylated DNA can yield approximately 18% coverage of the genome because it captures approximately 5 million methylated CpG sites (Figure 2).

MBDCap-seq is a simple approach that does not require bisulfite conversion and can be used to identify differentially methylated regions [40,41]. However, a notable disadvantage of MBDCap-Seq is that it does not provide single-nucleotide resolution. Rather, it identifies regions containing multiple methylated CpG sites typically at CpG-rich regions in a readout similar to chromatin immunoprecipitation (ChIP-Seq). Furthermore, MBDCap-Seq is only marginally quantitative because the number of reads mapping to a particular region of the genome depends on the density of methylated CpG sites [41].

Reduced representative bisulfite sequencing

RRBS is an efficient and high-throughput technique used to analyze methylation profiles at a single-nucleotide level from regions of high CpG content (e.g., CpG islands), but does not interrogate intergenic or lowly methylated regions of the genome (Figure 2) [24,42]. RRBS relies first on the digestion of genomic DNA (0.01–0.03 μg) with a methylation-insensitive restriction enzyme, such as MspI (C′CGG), that selects genomic regions with moderate to high CpG density, such as CpG islands, followed by DNA size fractionation (Figure 2). This ‘reduced representation’ of the genome is sequenced similarly to WGBS to generate a single-base pair resolution DNA methylation map [24,42]. A minimum of approximately 10 million sequencing reads are required for the downstream analysis of RRBS data sets, leading to approximately 3.7% actual coverage of CpG dinucleotides genome-wide or approximately 1 million CpG sites.

One of the main advantages of RRBS is that it is more cost-effective than WGBS, because it targets bisulfite sequencing to an enriched population of the genome, while retaining single-nucleotide resolution. RRBS data are restricted to regions with moderate to high CpG density, and are enriched for promoter-associated CpG islands. However, RRBS interrogates only <4% of the approximately 28 million CpG dinucleotides distributed throughout the human genome. Thus, a lack of coverage at intergenic and distal regulatory elements is a potential disadvantage of the method. In addition, although RRBS data can be processed using similar WGBS pipelines (e.g., [43,44]) data analysis requires a similar level of expertise and, hence, involves similar challenges.

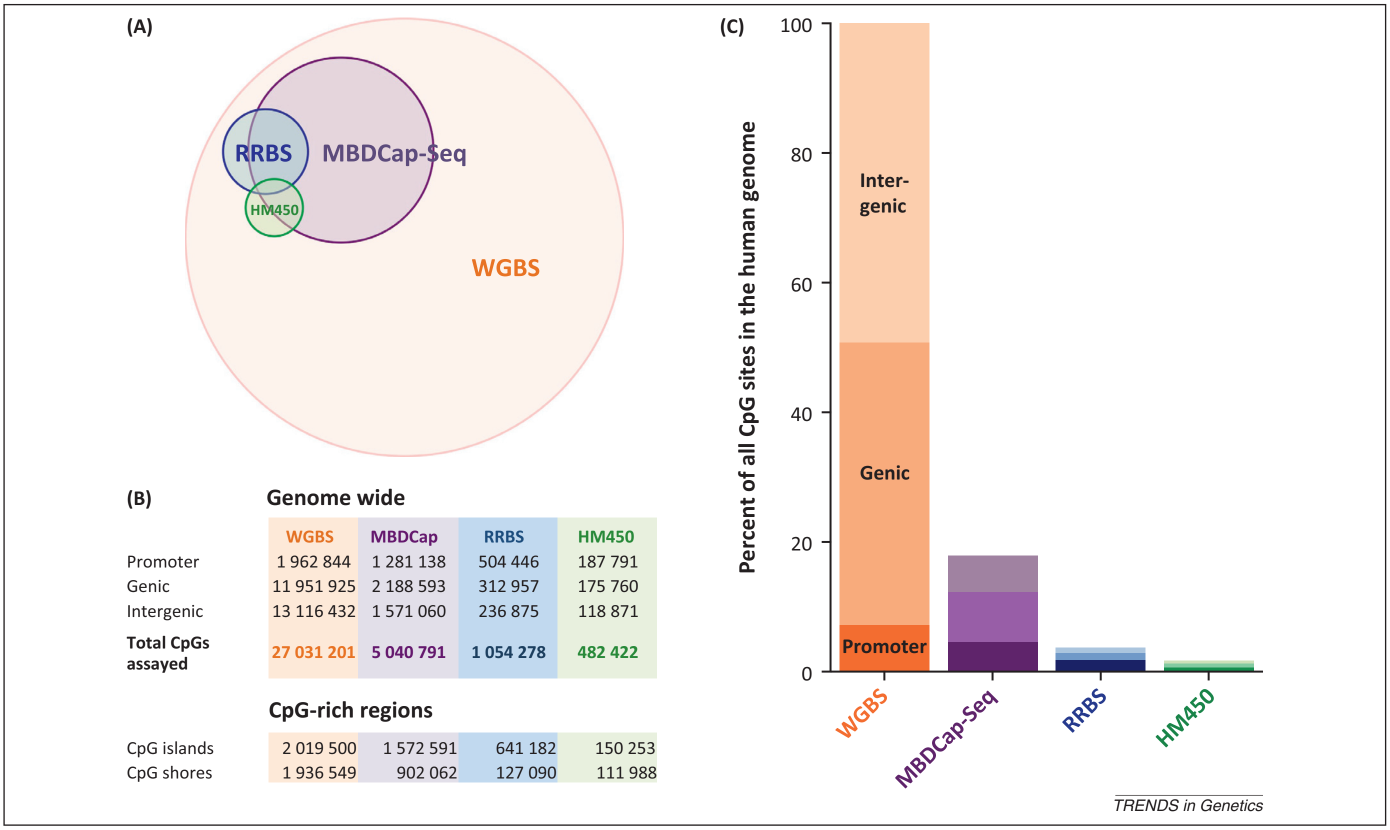
Infinium HumanMethylation450 BeadChip

The HM450 is an attractive option for genome-wide DNA methylation analyses in a variety of cell types. It is suitable for clinical samples, including FFPE tissue, it requires little starting material (approximately 0.5 μg), is cost effective, and can be used in a high-throughput manner. The technology is distinct from the other methylation technologies described above, in that it does not depend on capture or enrichment, or use of restriction enzymes or high-throughput sequencing for data generation (Figure 2). The HM450 protocol begins with the bisulfite conversion of genomic DNA (0.5–1 μg) (Figure 2). Converted genomic DNA is hybridized to arrays that contain predesigned probes to distinguish chemically methylated (cytosine) and unmethylated (converted to uracil). A single-base extension step incorporates a labeled nucleotide that is fluorescently stained. Scanning of the array detects the ratio of fluorescent signal arising from the unmethylated probe compared with the methylated probe, allowing the level of methylation to be determined (Figure 2).

The HM450 BeadChip interrogates 482 422 cytosines across the human genome, which represents only approximately 1.7% of all CpG sites in the human genome (Figure 2), substantially less than other methods. However, these sites are enriched for CpG (99.3%) residues and almost half (>41%, approximately 197 790 CpG sites) of the probes on the array cover intergenic regions, such as bioinformatically predicted enhancers, DNase I hypersensitive sites, and validated differentially methylated regions (DMRs) [45,46]. HM450 can be performed on both fresh-frozen and FFPE DNA, and methods are now being optimized to enable smaller amounts (0.2 μg) to be profiled efficiently [47]. Therefore, HM450 has become the method of choice for genome-wide DNA methylation analyses of profile large cohorts, because it requires a low amount of input material and it is cost effective. However, when using HM450 BeadChip technology, there are also some issues to consider. First, the design is heavily biased due to preselection and inclusion of probes that interrogate only certain CpG sites that have been previously identified in methylation-based assays and, therefore, the design is not hypothesis neutral. Second, it is assumed that CpG sites located adjacent to those interrogated by the probes will be similarly un/methylated, which is known as the ‘co-methylation assumption’ [48]. Finally, there are behavioral differences between the two types of probe design on the array, and the filtering of probes may be affected by single nucleotide polymorphisms, which need to be factored in to the data analysis pipelines [49].

Comparison of genome-wide coverage

The major advantage of WGBS is that, in theory, the methylation state of almost every single CpG dinucleotide (total 28 217 009) in the genome can be determined at single molecule resolution (Figure 3A,B). By contrast, with MBDCap-Seq, RRBS, and HM450, there is substantially less coverage with approximately 5 040 790, approximately 1 054 280, and 482 422 individual CpG sites, respectively, interrogated (Figure 3A,B). Notably, only a proportion of CpG sites are commonly interrogated by all three techniques (Figure 3A). MBDCap-Seq has greater



Genome wide

	WGBS	MBDCap	RRBS	HM450
Promoter	1 962 844	1 281 138	504 446	187 791
Genic	11 951 925	2 188 593	312 957	175 760
Intergenic	13 116 432	1 571 060	236 875	118 871
Total CpGs assayed	27 031 201	5 040 791	1 054 278	482 422

CpG-rich regions

	WGBS	MBDCap	RRBS	HM450
CpG islands	2 019 500	1 572 591	641 182	150 253
CpG shores	1 936 549	902 062	127 090	111 988

Figure 3. Proportion of promoters, genic, and intergenic regions interrogated by each technique. **(A)** The overlap and relative proportions of whole-genome bisulfite sequencing (WGBS), MBD capture sequencing (MBDCap-Seq), reduced representation bisulfite sequencing (RRBS), and HumanMethylation450 BeadChip (HM450) is plotted in a Venn diagram. **(B)** The total number of CpG dinucleotides (based on fully methylated DNA) covered by each technique is shown, and ranges from 482 422 (RRBS) to 27 031 201 (WBGS). Each CpG site is located in a promoter or genic or intergenic region of the genome and the distribution of these sites is detailed in the upper panel. The number of CpG sites covered by each technique that overlap CpG-rich regions (CpG island and CpG shores) is also shown in the lower panel. **(C)** WGBS covers approximately 95% of all CpG sites in the genome, most of which are located in intergenic or genic regions (approximately 12 million in each category) and the remainder in promoters (approximately 2 million). By contrast, HM450 interrogates the DNA methylation state of approximately 120 000 intergenic, approximately 170 000 genic, and approximately 180 000 promoter CpG sites. Data are expressed as a percentage of all CpG sites in the human genome.

coverage of promoter (approximately 1 281 140) and CpG island (approximately 1 572 590) CpG sites, as well as greater regional coverage of intergenic regions (approximately 1 571 060) and shores (approximately 902 060), compared with RRBS and HM450 arrays (Figure 3B). Moreover, when the genome is sorted into functional categories (promoter, genic, or intergenic; Figure 3B), it becomes clear that each technique, except for WGBS, is biased for different regulatory regions of the genome (Figure 3B,C). For example, MBDCap-Seq interrogates 1 572 591 CpG island sites (approximately 31% of all CpG sites assayed using MBDCap-Seq) compared with WGBS, which interrogates all 2 019 500 CpG island sites in the genome (approximately 7.5% of all CpG sites assayed using WGBS). Although RRBS covers less than 5% of all CpG sites in the human genome (Figure 3B), it enriches for regions of the genome that have a high CpG content and of the more than approximately 1 million CpG sites interrogated, almost 50% (504 446) are within promoter regions and 641 182 CpG sites are within CpG islands. Although HM450 arrays cover the fewest number of CpG sites (Figure 3B,C), the arrays provide good coverage of methylation at CpG island promoters. Nonetheless, WGBS is the only method to date that best represents regions of lower CpG density, such as intergenic 'gene deserts', partially methylated domains, and distal regulatory elements (e.g., enhancers) that potentially facilitates control of tissue-specific expression and noncoding RNA expression, which are commonly deregulated in cancer.

Comparison of DNA methylation data output

Consistent with variations in genomic coverage, the data output of the genome-wide DNA methylation approaches differs considerably (summarized in Figure 4). We have used *CAV1* and *GSTP1* gene promoters to illustrate the differences in methylation signal and coverage across CpG island gene promoters and adjacent intergenic and genic regions (Figure 4A,B). With the exception of MBDCap-Seq, WGBS, RRBS, and HM450 all measure both unmethylated and methylated cytosines at single CpG sites and, therefore, are fully quantitative, but the accuracy depends on coverage. Notable is the explicit detail of CpG methylation in WGBS data (Figure 4C,D). With sufficient sequencing depth, individual WGBS and RRBS sequencing reads allow the separation of DNA methylation data for each strand, the detection of cytosine methylation in a non-CpG context [22], heterogeneous patterns, and allele-specific DNA methylation (Figure 4C,D). Current bisulfite-based methodologies cannot distinguish between 5mC and other novel structurally similar DNA modifications that have recently been discovered, including 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC). This

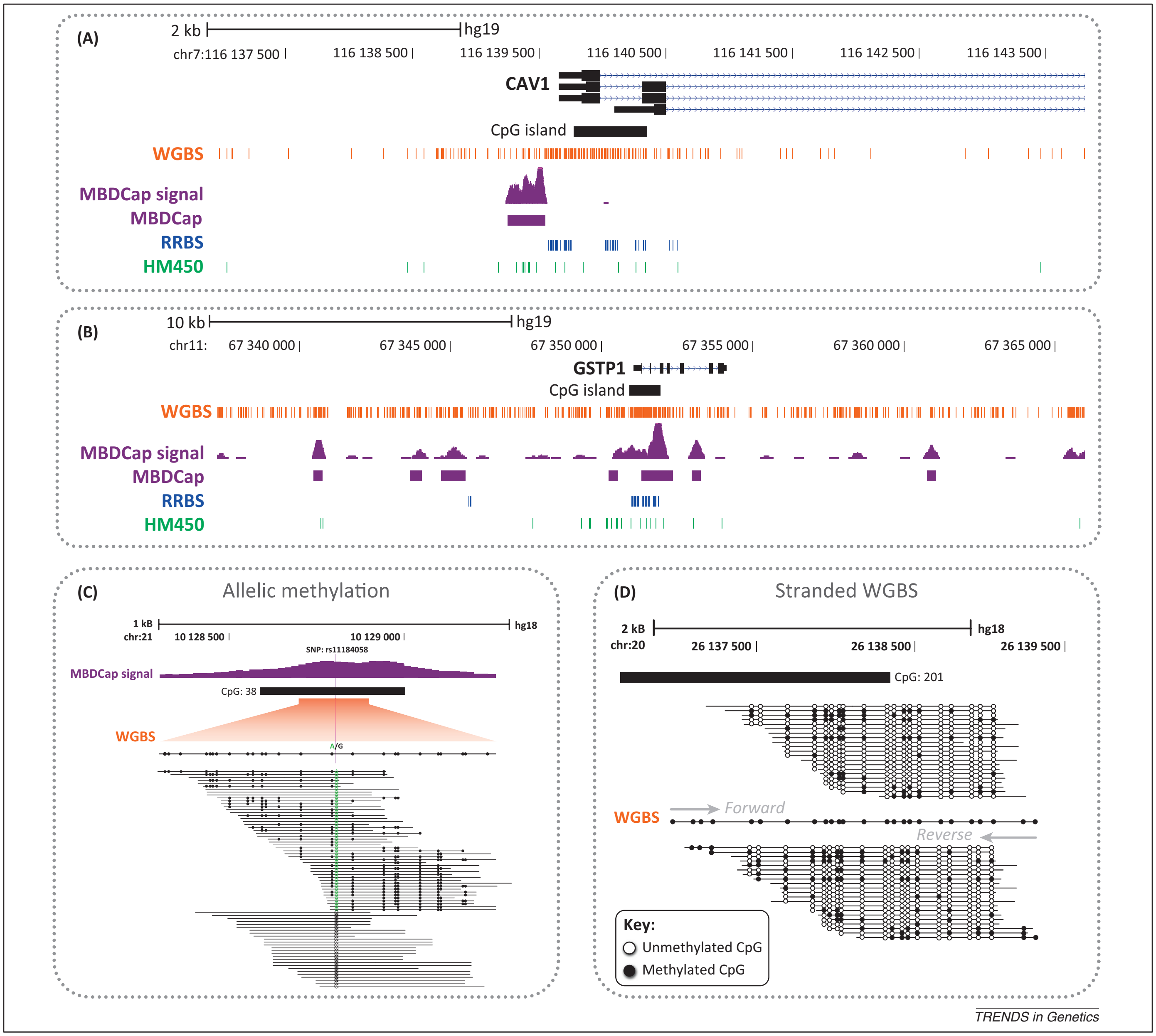
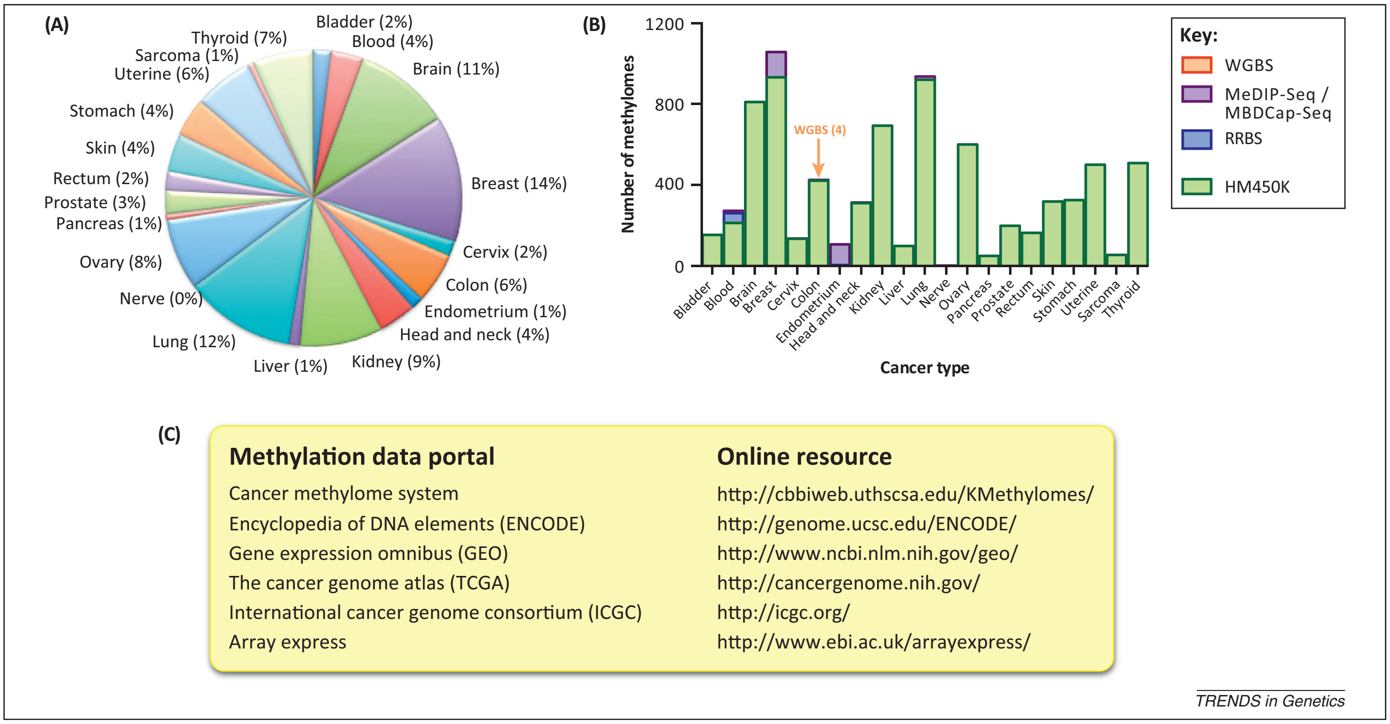


Figure 4. Comparison of DNA methylation approaches. The signal from MBD capture sequencing (MBDCap-Seq) is averaged and not fully quantitative, whereas explicit detail can be viewed in whole-genome bisulfite sequencing (WGBS) data. **(A,B)** Screenshots of glutathione-*S*-transferase P1 gene (*GSTP1*) and caveolin 1 (*CAV1*) gene promoters and adjacent intergenic and genic regions show that reduced representation bisulfite sequencing (RRBS) and HumanMethylation450 BeadChip (HM450) largely capture CpG sites surrounding promoters, but few CpG sites in the genic and intergenic regions of these genes. MBDCap is not fully quantitative and relies on accurate analysis and interpretation of the raw signal. **(C)** Individual sequencing reads allow the separation of DNA methylation data by genomic sequence (e.g., single nucleotide polymorphisms; SNP), demonstrating the phenomenon of allele-specific DNA methylation. **(D)** Heterogeneous methylation (defined as either sporadic methylation within an individual DNA molecule or differential levels of methylation between individual DNA molecules) can be observed in WGBS data, as can the unique information obtained from the forward- and reverse-sequencing strands.

may have consequences for data interpretation, potentially leading to an overestimation of DNA methylation levels. However, innovative detection methods are being developed, such as those that allow specific detection of 5mC and 5hmC [50,51], which opens up future possibilities to develop whole-genome approaches to assess all methylation modifications simultaneously.

Bioinformatics

A particular challenge of any genome-wide approach is the downstream computational requirements for obtaining meaningful outcomes. The main disadvantages of WGBS at the present time are the onerous computational resources needed for read alignment [43,52–54], and the current need to develop custom bioinformatics scripts. Additionally, WGBS studies to date have performed few, if any, replicates, which severely limits statistical power and the ability to distinguish actual alterations from biological variability [55]. RRBS also requires bioinformatics expertise for analysis; however, the greatly reduced amount of data produced per experiment requires comparatively modest computational resources [56]. MBDCap-Seq requires less bioinformatics expertise and can be analyzed using established algorithms [41,57]. However,



Methylation data portal	Online resource
Cancer methylome system	http://cbbiweb.uthscsa.edu/KMethylomes/
Encyclopedia of DNA elements (ENCODE)	http://genome.ucsc.edu/ENCODE/
Gene expression omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo/
The cancer genome atlas (TCGA)	http://cancergenome.nih.gov/
International cancer genome consortium (ICGC)	http://icgc.org/
Array express	http://www.ebi.ac.uk/arrayexpress/

Figure 5. Cancer methylomes. **(A)** Cancer methylomes of at least 21 broad cancer types have been completed, representing >8000 individual data sets. Most DNA methylomes have been produced for breast (approximately 14% of all methylomes), lung (approximately 12% of all methylomes), and brain tumors (approximately 11% of all methylomes). Data from rare cancers are also beginning to be performed (nerve; one methylome, approximately 0.01% of all methylomes). The data are expressed as percentage of all methylomes produced, regardless of tumor origin, and show a wide distribution of methylomes across a broad range of cancer types. **(B)** We compared the techniques used to generate each methylome. HumanMethylation450 BeadChip (HM450; green) clearly dominates as the method of choice for high-throughput methylation studies. Currently, only four whole-genome bisulfite sequencing (WGBS; orange) data sets have been produced (colon). MBD capture sequencing (MBDCap-Seq; purple) has been used to measure DNA methylation in blood, brain, endometrial, and lung tumors, whereas the use of reduced representation bisulfite sequencing (RRBS; blue) has been limited to blood cancer. **(C)** Key online resources for accessing publicly available methylation data are summarized.

MBDCap-Seq is not fully quantitative and, therefore, relies on accurate analysis and interpretation of the raw signal. In particular, failure to control for copy-number alterations can lead to inaccuracies in methylation measurements, an issue that affects cancer samples [57]. More mature bioinformatics analysis pipelines exist for HM450 [58–60], and these pipelines already include normalization measures to analyze data [49,59], meaning that these arrays may be the most accessible genome-wide DNA methylation assay.

Sequencing coverage

The amount of sequencing needed to yield meaningful results differs substantially between techniques. The main disadvantage of WGBS at the present time is the cost of sequencing, which requires >500 million reads (100 bp paired-end) per sample (approximately 30 x coverage), or approximately three sequencing lanes on the Illumina HiSeq. At a 'shallow' sequencing depth (1–5 x coverage), regions of high and low average methylation can be quantitated, whereas at a 'deep' sequencing depth (30 x), individual CpG sites can be accurately quantitated. With sufficient coverage, it is possible to apply adaptations of genomic variant detection algorithms [61] to interrogate the genotype and methylation status of the samples simultaneously, enabling applications such as the assessment of allele-specific methylation (Figure 4C). By contrast, MBDCap-Seq only requires short read chemistry (50 bp single-end) and a relatively 'shallow' sequencing depth (approximately 30 million reads per sample), allowing six samples to be multiplexed per HiSeq lane. However, RRBS requires only 10 million reads per sample (Figure 2). Notably, the sequencing depth required correlates with the genome coverage capability of each approach. HM450 arrays do not rely on high-throughput sequencing for data generation.

Summary of cancer methylome studies

To date, approximately 8000 cancer methylomes have been generated (Figure 5A). Most major cancers have at least one representative methylome, with no one type being overrepresented as a proportion of all methylomes available (Figure 5A). However, it is clear that the HM450 arrays dominate studies investigating cancer methylomes (Figure 5B). Indeed, the Cancer Genome Atlas consortium (TCGA; http://cancergenome.nih.gov) is a portal for understanding the genomic basis of more than 200 human cancer types. Among the massive data sets that are accessible to all researchers, TCGA has profiled the DNA methylome in approximately 7500 samples using the HM450 methodology [62–66]. These data sets largely comprise the newer, HM450 array. To date, only two deeply sequenced WGBS of primary tumors have been completed [30,32], three shallowly sequenced WGBS tumors (all colon; Figure 5B) [31] and approximately 55 RRBS analyses, of which most investigate primary blood cancers (Figure 5B). However, the limited number WGBS cancer methylomes is likely to change drastically as the cost of the technology and ease of bioinformatics analyses improves. A summary of DNA methylation data portals is shown in Figure 5C.

What have we learnt from cancer methylome studies?

The development of next-generation sequencing technologies and ability to map the changes in DNA methylation across many cancer types has led to huge advances in knowledge. DNA methylation studies have revealed that changes are not restricted to CpG island promoters, but occur genome wide, including genic and intergenic regions. The intergenic space is vast and houses distal regulatory elements, including enhancers and noncoding RNA genes, and is a frequent site for the mutation hotspots in cancer [67,68]. It is now clear that DNA methylation in distal regulatory regions is also associated with transcriptional regulation. Methylation in genic or exonic regions is also associated with changing levels of transcription, where high methylation occurs in active genes and lower methylation in repressed genes [22]. Somatic mutations in noncoding regions add another dimension to the complexity of deregulation of the cancer epigenome, given that mutation hotspots can be caused by DNA methylation [69,70] and that genetic mutations can be strongly associated with changes in methylation patterns [9,66,71–74].

Cancer methylomes now face finer interpretation as we try to understand architectural differences, such as long-range epigenetic silencing (LRES; [75]) or long-range epigenetic activation (LREA; [76]), as well as discrete changes, such as atypical DNA methylation at localized CpG sites, partially methylated domains (PMDs; [30,31,77]) and DMRs [31] that may be responsible for disabling or enabling key gene regulatory elements. The identification of DNA methylation valleys (DMVs) in embryonic stem cells points to novel genomic features that may also be evident in tumor methylomes [36]. Altered cancer methylomes are commonly associated with changes in transcriptional output and altered genomic stability. Indeed, cancer cells undergo a multitude of step-wise and cumulative methylation changes that impinge on crucial biological pathways that potentially influence proliferation rates, response to extracellular signals, and the response to DNA damage.

Yet, not all aberrant DNA methylation changes drive disease. It is, and will be, important to distinguish driver from passenger roles [78], which will enable an even more precise stratification of cancer subtypes [66,79] and personalized therapeutic programs [9,80]. One of the first studies investigating the role of DNA methylation drivers and passengers demonstrated that cancer cells are potentially addicted to the modified epigenome [78]. Future analyses will reveal the specific DNA methylation signatures that are either associated or drive the survival capacity of cancer cells. However, distinct methylation patterns are being used to classify distinct subtypes [81–84]. For example, the CpG Island Methylator Phenotype (CIMP), first described in colorectal cancer [85] and evident in many other cancer types [86], indicates that DNA methylation is potentially useful for disease classification. In fact, CIMP has recently been reported to be associated with underlying genetic mutations, such as somatic isocitrate dehydrogenase-1 (IDH1) mutations and mutations in ten-eleven translocation (TET) methylcytosine dioxygenase-2 (TET2).

Advances in genome-wide DNA methylation technology have also enabled new strategies for the identification of early novel diagnostic and prognostic cancer biomarkers [87,88]. Already, the measurement of promoter hypermethylation of individual genes has been successfully implemented in the clinic. For example, the glutathione-*S*-transferase P1 gene (*GSTP1*) gene is methylated in >90% of prostate cancers [89] and Septin 9 (*SEPT9*) is hypermethylated in colorectal cancer; both are currently being used for early cancer detection in tissue samples and body fluids [90]. Moreover, promoter hypermethylation of the *MGMT* DNA-repair gene is a clear predictor of tumor responsiveness to alkylating agents in patients with glioblastoma [91,92]. These examples highlight the promise of translating epigenetic markers into a clinical setting, especially given that the deregulation of cellular epigenetic patterns is an early event in carcinogenesis.

Concluding remarks and future perspectives

The advent of genome-wide approaches to map the cancer methylome, and the ability to identify differentially methylated loci, is leading to the development of panels of biomarkers that increase the specificity and sensitivity for improved diagnostic potential [93,94]. In cancer treatment, one of the major challenges is to stratify tumor types, because most cancer subtypes do not behave as a single entity in response to current therapies. The ability to identify epigenetic events associated with survival from archival cancer samples is revealing epigenetic prognostic signatures that can be used to cluster subtypes upon diagnosis to enable better treatment options. The future production of cancer methylomes, especially with detailed information of the approximately 28 million CpG sites in each different cancer cell type will further advance understanding of the role of DNA methylation in epigenetic-based molecular function and disease progression. Ultimately, however, the choice of which whole-genome methylation approach to use will depend on the quantity and quality of DNA available, accessibility to next-generation sequencing, bioinformatics expertise, cost, and, finally, consideration of the question being asked and the required coverage of the genome.

Acknowledgments

We thank Elena Zotenko for the MBDCap-Seq coverage analysis and members of the Clark Laboratory for helpful discussions and careful reading of the manuscript. P.C.T. is a Cancer Institute NSW Career Development Fellow. S.J.C. is a National Health and Medical Research Council (NH&MRC) Senior Principal Research Fellow. This work was further supported by NH&MRC Project Grants (to S.J.C., P.C.T., and C.S.).

References

1 Bird, A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature* 321, 209–213
2 Li, E. *et al.* (1993) Role for DNA methylation in genomic imprinting. *Nature* 366, 362–365
3 Mohandas, T. *et al.* (1981) Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* 211, 393–396
4 Gartler, S.M. and Riggs, A.D. (1983) Mammalian X-chromosome inactivation. *Annu. Rev. Genet.* 17, 155–190
5 Swain, J.L. *et al.* (1987) Parental legacy determines methylation and expression of an autosomal transgene: a molecular mechanism for parental imprinting. *Cell* 50, 719–727

6 Reik, W. *et al.* (1987) Genomic imprinting determines methylation of parental alleles in transgenic mice. *Nature* 328, 248–251

7 Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.* 16, 6–21

8 Jones, P.A. and Takai, D. (2001) The role of DNA methylation in mammalian epigenetics. *Science* 293, 1068–1070

9 Taberlay, P.C. and Jones, P.A. (2011) DNA methylation and cancer. *Prog. Drug Res.* 67, 1–23

10 Bestor, T.H. (1988) Cloning of a mammalian DNA methyltransferase. *Gene* 74, 9–12

11 Irizarry, R.A. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* 41, 178–186

12 Doi, A. *et al.* (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* 41, 1350–1353

13 Jones, P.A. and Baylin, S.B. (2007) The epigenomics of cancer. *Cell* 128, 683–692

14 Baylin, S.B. and Jones, P.A. (2011) A decade of exploring the cancer epigenome: biological and translational implications. *Nat. Rev. Cancer* 11, 726–734

15 Herman, J.G. and Baylin, S.B. (2003) Gene silencing in cancer in association with promoter hypermethylation. *N. Engl. J. Med.* 349, 2042–2054

16 Jones, P.A. and Baylin, S.B. (2002) The fundamental role of epigenetic events in cancer. *Nat. Rev. Genet.* 3, 415–428

17 Aran, D. *et al.* (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* 14, R21

18 Akhtar-Zaidi, B. *et al.* (2012) Epigenomic enhancer profiling defines a signature of colon cancer. *Science* 336, 736–739

19 Laird, P.W. and Jaenisch, R. (1994) DNA methylation and cancer. *Hum. Mol. Genet.* 3, 1487–1495

20 Bock, C. *et al.* (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* 28, 1106–1114

21 Harris, R.A. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* 28, 1097–1105

22 Lister, R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322

23 Meissner, A. *et al.* (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* 33, 5868–5877

24 Meissner, A. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454, 766–770

25 Smith, Z.D. *et al.* (2009) High-throughput bisulfite sequencing in mammalian genomes. *Methods* 48, 226–232

26 Frommer, M. *et al.* (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

27 Clark, S.J. *et al.* (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* 22, 2990–2997

28 Lister, R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133, 523–536

29 Cokus, S.J. *et al.* (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452, 215–219

30 Berman, B.P. *et al.* (2012) Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet.* 44, 40–46

31 Hansen, K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* 43, 768–775

32 Ziller, M.J. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500, 477–481

33 Heyn, H. *et al.* (2012) Distinct DNA methylomes of newborns and centenarians. *Proc. Natl. Acad. Sci. U.S.A.* 109, 10522–10527

34 Hon, G.C. *et al.* (2012) Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* 22, 246–258

35 Lister, R. *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471, 68–73

36 Xie, W. *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153, 1134–1148

37 Weber, M. *et al.* (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* 37, 853–862

38 Serre, D. *et al.* (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.* 38, 391–399

39 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

40 Nair, S.S. *et al.* (2011) Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics* 6, 34–44

41 Robinson, M.D. *et al.* (2010) Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. *Genome Res.* 20, 1719–1729

42 Gu, H. *et al.* (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods* 7, 133–136

43 Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572

44 Hebestreit, K. *et al.* (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* 29, 1647–1653

45 Bibikova, M. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98, 288–295

46 Sandoval, J. *et al.* (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 6, 692–702

47 Hosein, A.N. *et al.* (2012) The use of the Illumina FFPE Restoration Protocol to obtain suitable quality DNA for SNP-based CGH: a pilot study. *Hered. Cancer Clin. Pract.* 10, A85

48 Eckhardt, F. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* 38, 1378–1385

49 Pidsley, R. *et al.* (2013) A data–driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 14, 293

50 Booth, M.J. *et al.* (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* 336, 934–937

51 Yu, M. *et al.* (2012) Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat. Protoc.* 7, 2159–2170

52 Chen, P.Y. *et al.* (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* 11, 203

53 Xi, Y. *et al.* (2012) RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics* 28, 430–432

54 Chatterjee, A. *et al.* (2012) Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Res.* 40, e79

55 Hansen, K.D. *et al.* (2011) Sequencing technology does not eliminate biological variability. *Nat. Biotechnol.* 29, 572–573

56 Wang, T. *et al.* (2013) RRBS-Analyser: a comprehensive web server for reduced representation bisulfite sequencing data analysis. *Hum. Mutat.* 34, 1606–1610

57 Robinson, M.D. *et al.* (2012) Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Res.* 22, 2489–2496

58 Hansen, K.D. *et al.* (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* 13, R83

59 Maksimovic, J. *et al.* (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* 13, R44

60 Touleimat, N. and Tost, J. (2012) Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 4, 325–341

61 Liu, Y. *et al.* (2012) Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* 13, R61

62 Cancer Genome Atlas Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615

63 Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70

64 Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337

65 Larman, T.C. *et al.* (2012) Spectrum of somatic mitochondrial mutations in five cancers. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14087–14091

66 Noushmehr, H. *et al.* (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17, 510–522

67 Hodis, E. *et al.* (2012) A landscape of driver mutations in melanoma. *Cell* 150, 251–263

68 Barbieri, C.E. *et al.* (2013) The mutational landscape of prostate cancer. *Eur. Urol.* 64, 567–576

69 Cohen, N.M. *et al.* (2011) Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell* 145, 773–786

70 Wang, R.Y. *et al.* (1982) Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. *Biochim. Biophys. Acta* 697, 371–377

71 Olivier, M. *et al.* (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* 2, a001008

72 Shen, H. and Laird, P.W. (2013) Interplay between the cancer genome and epigenome. *Cell* 153, 38–55

73 Weisenberger, D.J. *et al.* (2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat. Genet.* 38, 787–793

74 You, J.S. and Jones, P.A. (2012) Cancer genetics and epigenetics: two sides of the same coin? *Cancer Cell* 22, 9–20

75 Coolen, M.W. *et al.* (2010) Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. *Nat. Cell Biol.* 12, 235–246

76 Bert, S.A. *et al.* (2013) Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* 23, 9–22

77 Raddatz, G. *et al.* (2012) Dnmt3a protects active chromosome domains against cancer-associated hypomethylation. *PLoS Genet.* 8, e1003146

78 De Carvalho, D.D. *et al.* (2012) DNA methylation screening identifies driver epigenetic events of cancer cell survival. *Cancer Cell* 21, 655–667

79 Figueroa, M.E. *et al.* (2010) DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* 17, 13–27

80 Kelly, T.K. *et al.* (2010) Epigenetic modifications as therapeutic targets. *Nat. Biotechnol.* 28, 1069–1078

81 Dedeurwaerder, S. and Fuks, F. (2012) DNA methylation markers for breast cancer prognosis: unmasking the immune component. *Oncoimmunology* 1, 962–964

82 Fackler, M.J. *et al.* (2011) Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence. *Cancer Res.* 71, 6195–6207

83 Fang, F. *et al.* (2011) Breast cancer methylomes establish an epigenomic foundation for metastasis. *Sci. Transl. Med.* 3, 75ra25

84 Szyf, M. (2012) DNA methylation signatures for breast cancer classification and prognosis. *Genome Med.* 4, 26

85 Toyota, M. *et al.* (1999) CpG island methylator phenotype in colorectal cancer. *Proc. Natl. Acad. Sci. U.S.A.* 96, 8681–8686

86 Hughes, L.A. *et al.* (2013) The CpG island methylator phenotype: what's in a name? *Cancer Res.* 73, 5858–5868

87 Sandoval, J. *et al.* (2013) A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J. Clin. Oncol.* 31, 4140–4147

88 Gyparaki, M.T. *et al.* (2013) DNA methylation biomarkers as diagnostic and prognostic tools in colorectal cancer. *J. Mol. Med.* 91, 1249–1256

89 Jeronimo, C. *et al.* (2011) Epigenetics in prostate cancer: biologic and clinical relevance. *Eur. Urol.* 60, 753–766

90 Ladabaum, U. *et al.* (2013) Colorectal cancer screening with blood-based biomarkers: cost-effectiveness of methylated septin 9 DNA versus current strategies. *Cancer Epidemiol. Biomarkers Prev.* 22, 1567–1576

91 Hegi, M.E. *et al.* (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* 352, 997–1003

92 Esteller, M. *et al.* (2000) Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* 343, 1350–1354

93 Devaney, J. *et al.* (2011) Epigenetic deregulation across chromosome 2q14.2 differentiates normal from prostate cancer and provides a regional panel of novel DNA methylation cancer biomarkers. *Cancer Epidemiol. Biomarkers Prev.* 20, 148–159

94 Mikeska, T. *et al.* (2012) DNA methylation biomarkers in cancer: progress towards clinical implementation. *Expert Rev. Mol. Diagn.* 12, 473–487