COMPUTATIONAL TOOLS

Analysing and interpreting DNA methylation data

Christoph Bock

Abstract | DNA methylation is an epigenetic mark that has suspected regulatory roles in a broad range of biological processes and diseases. The technology is now available for studying DNA methylation genome-wide, at a high resolution and in a large number of samples. This Review discusses relevant concepts, computational methods and software tools for analysing and interpreting DNA methylation data. It focuses not only on the bioinformatic challenges of large epigenome-mapping projects and epigenome-wide association studies but also highlights software tools that make genome-wide DNA methylation mapping more accessible for laboratories with limited bioinformatics experience.

Biomarkers
Molecular assays that predict a clinical phenotype, such as disease status or response to a drug.

Reference epigenomes
Publicly available epigenome maps that comprise multiple epigenetic marks for the same cell type (for example, DNA methylation, several histone modifications and non-coding RNA expression).

Epigenome-wide association study
(EWAS). A study design that involves measuring an epigenetic mark in cases and controls to identify disease-associated differences.

CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria; Department of Laboratory Medicine, Medical University of Vienna, 1090 Vienna, Austria; and Max Planck Institute for Informatics, 66123 Saarbrücken, Germany.
e-mail: cbock@cemm.oeaw.ac.at
doi:10.1038/nrg3273

DNA methylation is the only epigenetic mark for which a detailed mechanism of mitotic inheritance has been described[1]. In vertebrates, the most common form of DNA methylation is 5-methylcytosine (5mC), which affects 70 to 80% of CpGs in the human genome[2]. High levels of 5mC in CpG-rich promoter regions are strongly associated with transcriptional repression, whereas CpG-poor genomic regions exhibit a more complex and context-dependent relationship between DNA methylation and transcriptional activity[3]. DNA methylation has been extensively studied for its role in various biological processes — including genomic imprinting[4], transposable elements silencing[5], stem cell differentiation[6], embryonic development[7] and inflammation[8] — and characteristic changes in DNA methylation have been reported for cancer[9] and several other diseases[10]. Given the delicate balance between stability and plasticity of DNA methylation patterns, it has been suggested that DNA methylation may provide a lifetime record of environmental exposures and a useful source of biomarkers for risk stratification and disease diagnostics[11–13].

Recent advances in next-generation sequencing and microarray technology make it possible to map DNA methylation genome-wide, at a high resolution and in a large number of samples[14]. These new methods create ample opportunities for epigenome research, but they also pose substantial challenges in terms of data processing, statistical analysis and biological interpretation of observed differences[15]. The scale of the bioinformatic challenges is best demonstrated by the goal of the International Human Epigenome Consortium to establish reference epigenomes for 1,000 biomedically relevant human cell populations[16]. To achieve this goal, it will be necessary to align in the order of 1 trillion sequencing reads to the human genome, to identify cell-type-specific DNA methylation patterns through hundreds of comparisons and to make the results accessible through user-friendly epigenome browsers and Web-based analysis tools. Additional challenges arise from the recent surge of interest in epigenetic epidemiology and the epigenetic basis of human diseases[17], and this has resulted in substantial efforts mapping DNA methylation in large case–control cohorts[18]. Such an epigenome-wide association study (EWAS) requires robust normalization algorithms for microarray-based DNA methylation data and rigorous statistical tests for identifying differentially methylated regions (DMRs) between cases and controls. Furthermore, as the technologies for DNA methylation mapping are becoming increasingly available in non-specialist laboratories, there is a growing need for easy-to-use software tools that facilitate the handling and analysis of large DNA methylation data sets.

This Review discusses relevant concepts, computational methods and software tools for analysing and interpreting DNA methylation data. The first section outlines essential steps of data processing and quality control as well as the transformation of raw sequencing reads and microarray data into accurate DNA methylation maps. The second section reviews tools for visualizing DNA methylation data and statistical methods for identifying sample-specific differences in DNA methylation. The third section summarizes computational methods that assist with the validation of differences in

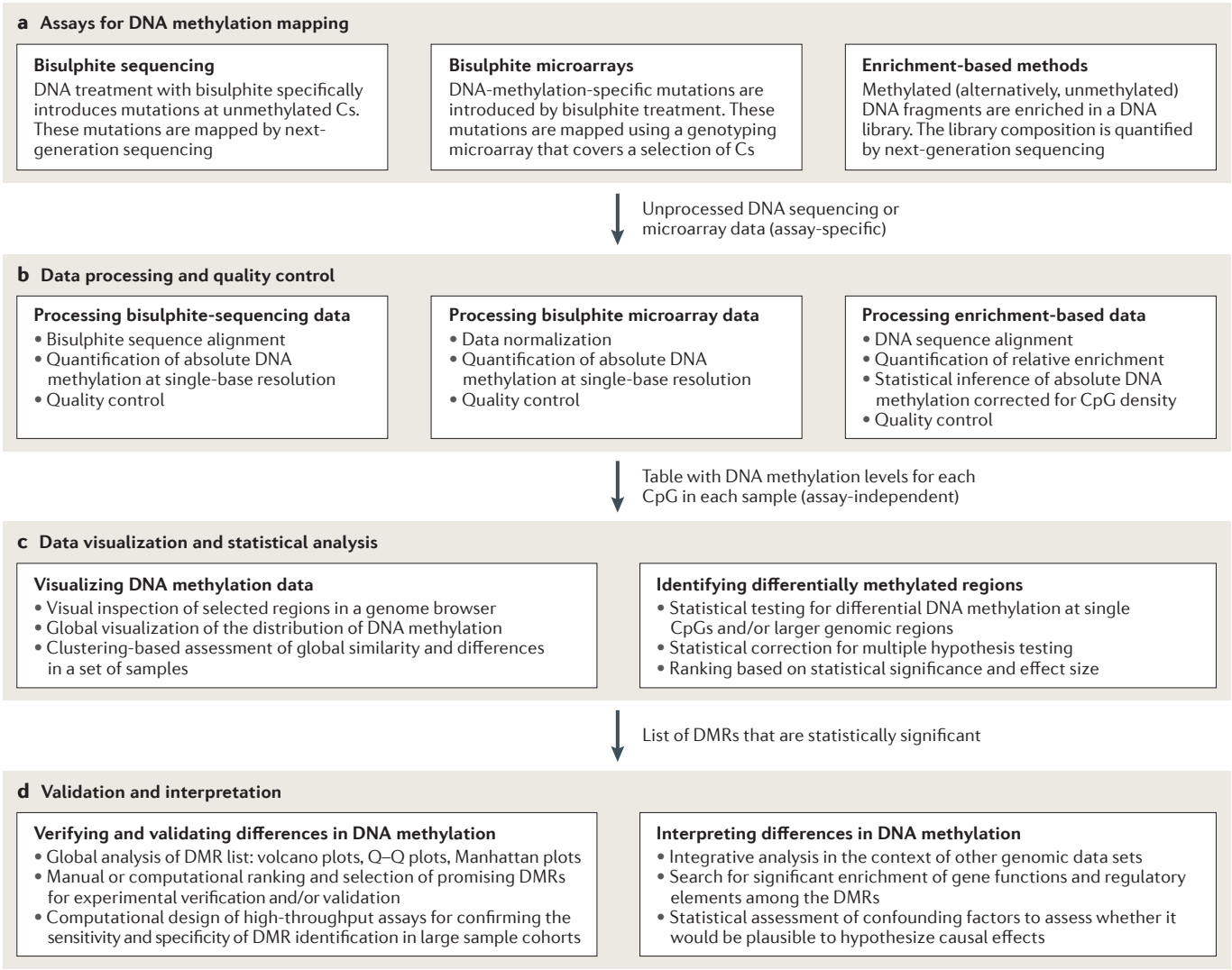

Figure 1 | **Workflow for analysing and interpreting DNA methylation data.** **a** | Genome-wide DNA methylation is mapped with one of the three most commonly used assays, resulting in methylation-specific DNA sequencing or microarray data. **b** | These raw data are processed and quality-controlled using assay-specific algorithms and software. The main result of data normalization is an assay-independent CpG methylation table that contains absolute DNA methylation levels (β-values) for all covered CpGs. **c** | Data visualization and statistical analysis identifies relevant associations and derives a list of differentially methylated regions (DMRs) between cases and controls. **d** | The resulting DMR list is validated both computationally and experimentally, and biological interpretation is assisted by computational tools. (Note that the separation of the analysis workflow into four subsequent steps constitutes a conceptual simplification, and there are a number of reasons why a specific study may need to deviate from this approach.)

Differentially methylated regions
(DMRs). Genomic regions that exhibit statistically significant differences in DNA methylation between sample groups.

Bisulphite
Bisulphite ions (HSO_3^-) selectively deaminate unmethylated but not methylated Cs, giving rise to Us, which are replaced by Ts during subsequent PCR amplification.

DNA methylation and their interpretation in a broader biological context. The final section outlines emerging trends in the analysis and interpretation of DNA methylation data. The structure of this Review follows the flow of a typical DNA methylation mapping study, as illustrated in FIG. 1, and a list of the described software tools is available from TABLE 1.

Data processing and quality control

Various experimental methods have been developed for genome-wide DNA methylation mapping, each with their own advantages and challenges[14,19,20]. In this Review, we focus on the three most popular approaches: bisulphite sequencing, bisulphite microarrays and enrichment-based methods (FIG. 1a; BOX 1). These three approaches pose distinct computational challenges during data processing and quality control, as outlined below.

Processing bisulphite-sequencing data. As a result of DNA treatment with the bisulphite chemical, the vast majority of unmethylated Cs appears as Ts among the sequencing reads, whereas methylated Cs are largely protected from bisulphite-induced conversion. To calculate absolute DNA methylation levels from bisulphite-sequencing data, sequencing reads are aligned to the

Table 1 | **Software tools for the analysis and interpretation of DNA methylation data**

Software	Description	URL	Refs
Processing bisulphite-sequencing data			
B-SOLANA	Bisulphite aligner for processing bisulphite-sequencing data obtained in the two-base encoding of ABI SOLiD sequencers	http://code.google.com/p/bsolana	40
Bismark	Probably the most widely used three-letter bisulphite aligner; supports both Bowtie (fast, gap-free alignment) and Bowtie 2.0 (sensitive, gapped alignment)	http://www.bioinformatics.babraham.ac.uk/projects/bismark	28
Bis-SNP	Variant caller for inferring DNA methylation levels and genomic variants from bisulphite-sequencing reads that have been aligned by other tools	http://epigenome.usc.edu/publicationdata/bissnp2011	35
BRAT	Highly configurable and well-documented three-letter bisulphite aligner	http://compbio.cs.ucr.edu/brat	29,30
BS-Seeker	Basic three-letter bisulphite aligner based on Bowtie	http://pellegrini.mcdb.ucla.edu/BS_Seeker/BS_Seeker.html	31
BSMAP	Probably the most widely used wild-card bisulphite aligner	http://code.google.com/p/bsmap	21
GSNAP	Wild-card bisulphite aligner included in a widely used general-purpose alignment tool	http://share.gene.com/gmap	22
Last	Recent and well-validated wild-card bisulphite aligner included in a general-purpose alignment tool	http://last.cbrc.jp	23
MethylCoder	Three-letter bisulphite aligner that can be used with either Bowtie (high speed) or GSNAP (high sensitivity)	https://github.com/brentp/methylcode	32
Pash	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://brl.bcm.tmc.edu/pash	24
RMAP	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://www.cmb.usc.edu/people/andrewds/rmap	25
RRBSMAP	Variant of BSMAP that is specialized on reduced-representation bisulphite sequencing (RRBS) data	http://rrbsmap.computational-epigenetics.org	26
segemehl	Wild-card bisulphite aligner included in a general-purpose alignment tool	http://www.bioinf.uni-leipzig.de/Software/segemehl	27
Processing bisulphite microarray data			
ComBat	R script for correcting known or suspected batch effects using an empirical Bayes method	http://www.bu.edu/jlab/wp-assets/ComBat	52
Illumina BeadScan	Machine control and image processing software for Illumina Infinium microarray scanners	http://www.illumina.com/support/array/array_instruments/beadarray_reader.ilmn	
Illumina GenomeStudio	Graphical tool for data normalization, analysis and visualization of Illumina Infinium microarrays (and other genomic data types)	http://www.illumina.com/software/genomestudio_software.ilmn	
isva	R package for batch effect correction using an algorithm that is based on singular value decomposition	http://cran.r-project.org/web/packages/isva	50
methylumi	R/Bioconductor package for Infinium data normalization and general data handling	http://www.bioconductor.org/packages/release/bioc/html/methylumi.html	
minfi	R/Bioconductor package for Infinium data normalization, analysis and visualization	http://www.bioconductor.org/packages/release/bioc/html/minfi.html	
RnBeads	R package providing a software pipeline for Infinium data normalization, quality control, exploratory visualization and differentially methylated region (DMR) identification	http://rnbeads.computational-epigenetics.org	
SVA	R/Bioconductor package for correcting batch effects that are directly inferred from the data using surrogate variable estimation	http://www.bioconductor.org/packages/release/bioc/html/sva.html	53

Absolute DNA methylation levels
Percentage of methylated alleles for a given C; this value is always binary (0% or 100%) for single alleles but can take any value between 0% and 100% when averaging over many cells.

positions in the reference genome from which they were most likely to be derived, and the percentage of Cs and Ts are determined among all reads aligned to each C in the genomic DNA sequence.

The alignment of bisulphite-sequencing reads needs to account for the selective depletion of unmethylated Cs, but otherwise it can be carried out with short-read aligners that are similar to those used for chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq) or genome-resequencing data. Two alternative approaches have been developed (FIG. 2). Wild-card aligners (such as, BSMAP[21], GSNAP[22], Last[23], Pash[24], RMAP[25], RRBSMAP[26] and segemehl[27]) replace Cs in the genomic DNA sequence by the wild-card letter Y, which matches both Cs and Ts in the read sequence, or they modify the alignment scoring matrix in such a way that mismatches between Cs in the genomic DNA sequence and Ts in the read sequence are not penalized. By contrast, three-letter aligners (such as Bismark[28], BRAT[29,30], BS-Seeker[31] and

Table 1 (cont.) | **Software tools for the analysis and interpretation of DNA methylation data**

Software	Description	URL	Refs
Processing enrichment-based data			
BATMAN	Command-line tool for methylated DNA immunoprecipitation (MeDIP) data normalization	http://td-blade.gurdon.cam.ac.uk/software/batman	56
Bowtie	General-purpose aligner based on the Burrows–Wheeler transform	http://bowtie-bio.sourceforge.net	33
BWA	General-purpose aligner based on the Burrows–Wheeler transform	http://bio-bwa.sourceforge.net	55
MEDIPS	User-friendly R package for MeDIP data normalization	http://medips.molgen.mpg.de	58
MEDME	R package for MeDIP data normalization	http://espresso.med.yale.edu/medme	57
MeDUSA	Command-line software pipeline for MeDIP read alignment, data normalization, quality control and DMR identification	http://www2.cancer.ucl.ac.uk/medicalgenomics/medusa	60
MetMap	Command-line tool for normalization of DNA methylation data obtained using restriction enzymes that specifically cut unmethylated DNA	http://www.cs.berkeley.edu/~meromit/MetMap.html	62
MeQA	Command-line software pipeline for MeDIP read alignment, data normalization and quality control	http://life.tongji.edu.cn/meqa	59
Repitools	R/Bioconductor package for quality control and visualization of enrichment-based DNA methylation data	http://www.bioconductor.org/packages/release/bioc/html/Repitools.html	61
Visualizing DNA methylation data			
bigWig/bigBed tools	Command-line tools for preparing genome browser tracks in a format that allows efficient visualization over the Internet	http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64	69
Ensembl	Widely used Web-based genome browser that includes a regulatory build with various epigenome data sets	http://www.ensembl.org	71
HilbertVis	Graphical tool and R package for visualizing genomic data as two-dimensional fingerprint diagrams	http://www.ebi.ac.uk/huber-srv/hilbert	77
IGB	Graphical genome browser that is run locally on the user's computer	http://bioviz.org/igb	74
IGV	Widely used graphical genome browser that is run locally on the user's computer	http://www.broadinstitute.org/igv	73
UCSC Genome Browser	Widely used Web-based genome browser hosting all ENCODE data	http://genome.ucsc.edu	70
WashU Epigenome Browser	Web-based genome browser focusing on the human epigenome	http://epigenomegateway.wustl.edu	72
Identifying differentially methylated regions			
dmrFinder	Function for DMR detection that is a part of the charm package in R/Bioconductor	http://www.bioconductor.org/packages/release/bioc/html/charm.html	96
IMA	R package for exploratory analysis and DMR detection based on normalized Infinium data	http://www.rforge.net/IMA	89
NHMMfdr	R package for estimating false discovery rates (FDRs) using an explicit model of dependence between statistical tests performed for neighbouring CpGs	http://www.unc.edu/~pfkuan/softwares.htm	98
QDMR	User-friendly software tool for DMR identification based on Shannon entropy	http://bioinfo.hrbmu.edu.cn/qdmr	91
qvalue	R/Bioconductor package for calculating *q* value estimates of false discovery rates	http://www.bioconductor.org/packages/release/bioc/html/qvalue.html	97
Verifying and validating differences in DNA methylation			
BiQ Analyzer HT	Graphical tool for analysing locus-specific high-throughput bisulphite-sequencing data	http://biqanalyzer.computational-epigenetics.org	108
CpGassoc	R package for visualization and analysis of DNA methylation data	http://genetics.emory.edu/conneely	101
MassArray	R/Bioconductor package for processing Sequenom EpiTYPER data	http://www.bioconductor.org/packages/release/bioc/html/MassArray.html	106
MethLAB	Graphical interface for the visualization and analysis methods implemented in CpGassoc	http://genetics.emory.edu/conneely/MethLAB	102
MethMarker	Graphical tool for validating DMRs and designing DNA methylation biomarkers	http://methmarker.mpi-inf.mpg.de	105
PRIMEGENS	Web-based tool for large-scale primer design: for example, in the context of locus-specific high-throughput bisulphite sequencing	http://primegens.org	107

Table 1 (cont.) | **Software tools for the analysis and interpretation of DNA methylation data**

Software	Description	URL	Refs
Interpreting differences in DNA methylation			
AnnotationModules	Web-based tool for enrichment analysis based on genomic regions and using a diverse set of genome annotations	http://web.bioinformatics.cicbiogune.es/AM/AnnotationModules.php	117
EpiExplorer	Web-based tool for live exploration and interactive analysis of genomic region data in the context of public reference epigenome data sets	http://epiexplorer.mpi-inf.mpg.de	110
EpiGRAPH	Web-based tool for enrichment analysis based on genomic regions and using a diverse set of genome annotations	http://epigraph.mpi-inf.mpg.de	116
EVORA	R package for quantifying variation in DNA methylation as a cancer biomarker	http://cran.r-project.org/web/packages/evora	132
Galaxy	Widely used Web-based tool for genomic data processing and analysis	http://main.g2.bx.psu.edu	111
Genomic HyperBrowser	Web-based tool for statistical hypothesis testing based on genomic data sets	http://hyperbrowser.uio.no	112
GREAT	Web-based tool for Gene Ontology enrichment analysis based on genomic regions	http://great.stanford.edu	115

MethylCoder[32]) simplify bisulphite alignment by converting all Cs into Ts in the reads and for both strands of the genomic DNA sequence. This way, they can carry out the alignment exclusively on a three-letter alphabet (namely, A, G and T) using a standard aligner, such as Bowtie[33].

As illustrated in FIG. 2, wild-card aligners can be expected to achieve a higher genomic coverage but at the cost of introducing some bias towards increased DNA methylation levels. This is because three-letter aligners purge the remaining Cs from the bisulphite-sequencing reads and thereby decrease the sequence complexity, such that a larger percentage of reads is discarded owing to ambiguous alignment positions. By contrast, wild-card aligners are at an increased risk of introducing bias towards higher methylation levels because the extra Cs in a methylated sequencing read can raise the sequence complexity to a level that is sufficient for unique alignment to the genome, whereas the corresponding, unmethylated, T-containing read is discarded owing to non-unique alignment (see FIG. 2 for an example). Both of these issues are restricted to genomic regions that exhibit high levels of sequence identity with other parts of the genome (such as retrotransposons and segmental duplications) and become less relevant for longer sequencing reads; for this reason, other considerations, such as speed, memory consumption and user-friendliness, increasingly influence the selection of the most suitable bisulphite aligner.

After the bisulphite alignment has been completed, absolute DNA methylation levels are inferred from the frequency of Cs and Ts that align to each C in the genomic DNA sequence. Most researchers simply divide the number of observed Cs by the total number of Cs and Ts. However, experiences from genotype calling suggest that the accuracy can be improved by additional steps, such as local realignment, analysis of sequence quality scores and statistical modelling of allele distributions[34]. The Bis-SNP variant caller extends the well-validated Genome Analysis Toolkit (GATK) algorithm to bisulphite-sequencing data and thereby provides an important step in this direction[35]. Bis-SNP also removes a common error source in the analysis of DNA methylation data, as it can distinguish bisulphite-induced changes from genetic variants. This is possible because bisulphite-induced C-to-T variants exhibit a G on the opposing strand, whereas genetic C-to-T variants exhibit an A instead. Furthermore, Bis-SNP can directly infer accurate genotype information from bisulphite-sequencing data, and in combination with a prior report that describes the accurate identification of copy-number aberrations from bisulphite-sequencing data[36], these results suggest that bisulphite sequencing may be able to map genetic and epigenetic characteristics with acceptable accuracy in a single experiment.

To obtain high-quality DNA methylation data from the results of bisulphite sequencing, several technical details require careful attention (TABLE 2). First, most protocols include an enzymatic end repair step to restore double-stranded DNA after fragmentation, and this introduces constitutively unmethylated Cs at both ends of the DNA fragments. The magnitude of this effect can be assessed using the *M* bias plot and can be mitigated by discarding those positions among the sequencing reads that exhibit globally biased DNA methylation levels[37]. Second, a sizable fraction of DNA fragments is often shorter than the read length of the DNA-sequencing reaction. As a result, parts of the DNA-sequencing adaptor at the end of the fragments are being sequenced and can introduce methylated Cs into the sequencing reads. This effect can also be detected using *M* bias plots, and it is mitigated by trimming DNA sequences that align to the sequencing adaptor[38]. Third, some protocols carry out bisulphite treatment before adaptor ligation. In this case, half of the methylation-specific C-to-T conversions take the form of G-to-A substitutions[38] and require special handling during bisulphite alignment as well as during the inference of absolute DNA methylation levels. Fourth, DNA sequencers that use two-base encoding (such as

Sequence complexity
The diversity of the DNA sequence; it can be measured by the information content of the base composition.

Genotype calling
The determination of SNPs and other genetic variants in a given individual.

Genome Analysis Toolkit
(GATK). A widely used software tool for genotype calling based on next-generation sequencing data.

M bias plot
A quality-control diagram that plots mean DNA methylation levels for each position of the bisulphite-sequencing reads. Deviations from a horizontal line indicate biases.

Box 1 | **Methods for DNA methylation mapping**

Because DNA methylation patterns are erased by PCR amplification, current sequencing and microarray technologies cannot directly distinguish between methylated and unmethylated Cs. This limitation may eventually be overcome by single-molecule sequencing using nanopores or real-time monitoring of DNA polymerases[140,143]. But at the moment, an extra step is needed to convert DNA methylation information into readily assayable DNA sequence information. Various methods have been developed and are extensively reviewed[14,19,20] and benchmarked[64–66] elsewhere. Three approaches currently stand out as the most useful and popular: bisulphite sequencing, bisulphite microarrays and enrichment-based methods.

Bisulphite sequencing

Chemical treatment of the DNA with sodium bisulphite gives rise to methylation-specific sequence variants, which can be mapped and quantified by next-generation sequencing. Key advantages of this technology are its comprehensive genomic coverage, high quantitative accuracy and excellent reproducibility. Disadvantages include the high cost of whole-genome resequencing and the difficulty of discriminating between 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC)[144,145]. Bisulphite sequencing can be combined with enrichment strategies using restriction enzymes (in reduced-representation bisulphite sequencing (RRBS)[146]) or DNA fragment capture[147,148] to target bisulphite sequencing to a specific fraction of the genome and thereby to reduce the per-sample cost of sequencing.

Bisulphite microarrays

Bisulphite treatment in combination with specially designed genotyping microarrays makes it possible to measure DNA methylation levels at a preselected fraction of Cs throughout the genome. One key advantage of this approach is the (currently) lower per-sample cost compared with whole-genome bisulphite sequencing. Furthermore, bisulphite microarrays can be processed with the same infrastructure as genotyping microarrays, and the experimental procedures are somewhat easier and faster to carry out than the preparation of DNA sequencing libraries. Disadvantages include the lack of discrimination between 5mC and 5hmC, limited genomic coverage and high setup cost for designing custom microarrays. Bisulphite microarrays are commercially available only for the human genome.

Enrichment-based methods

DNA-methylation-specific antibodies, methyl-binding domain proteins or restriction enzymes are used to enrich for a fraction of highly methylated (or unmethylated) DNA fragments, and the enrichment of specific fragments is quantified by next-generation sequencing. The two key advantages of enrichment-based methods are the relatively low cost of achieving genome-wide coverage (albeit with a low statistical power in CpG-poor genomic regions[64]) and the ability to distinguish between different forms of DNA methylation: for example, using antibodies that specifically recognize 5hmC but not 5mC. However, these advantages come at the cost of relatively low resolution and high susceptibility to experimental biases.

ABI SOLiD) instead of the more common single-base encoding (such as Illumina, Ion Torrent and Roche 454) pose bioinformatic challenges that are only partially solved[37–40]. Fifth, double counting of the same DNA fragments must be avoided to prevent bias. It is therefore common practice to discard putative PCR duplicates by retaining only one read per start and end position in the reference genome and to trim the overlapping part of paired-end reads. Finally, the sensitivity and specificity of the bisulphite conversion should be monitored, ideally by control DNA with known levels of DNA methylation levels that is added before bisulphite treatment (for example, spike-in oligomers with 0%, 50% and 100% DNA methylation). Elevated levels of observed CpC methylation can also provide an indication of incomplete bisulphite conversion because CpC dinucleotides are rarely methylated in mammalian cells[41,42].

Processing bisulphite microarray data. Specialized microarrays have been developed for high-throughput profiling of bisulphite-converted DNA. These microarrays quantify the DNA methylation levels of a predefined subset of Cs, each of which is represented by dedicated probes on the microarray. The Illumina Infinium assay is by far the most widely used bisulphite microarray, and it has been the focus of substantial bioinformatic methods development. The latest version of the Infinium assay comprises slightly more than 450,000 covered CpGs[43]. It uses two different probe types with distinct experimental characteristics, thus requiring careful normalization to minimize technical bias[44–46]. The genomic coverage of the Infinium assay is more limited than that of most bisulphite-sequencing protocols (1.5% of CpGs in the human genome are present on the Infinium 450k microarray), but the compatibility with existing genotyping pipelines makes it an attractive assay for measuring DNA methylation in large sample cohorts.

Like other types of microarray (for example, for genotyping and transcription profiling), the bioinformatic processing of Infinium data comprises image processing and data normalization as its key steps. Image processing is almost always carried out using the vendor-provided Illumina BeadScan software (TABLE 1), whereas several options exist for normalizing the probe intensity data and for inferring absolute DNA methylation levels. The commercial Illumina GenomeStudio software (TABLE 1) provides a basic algorithm for signal normalization and background subtraction using positive and negative control probes, and a similar algorithm is implemented in R/Bioconductor[47] as a part of the open-source packages minfi and methylumi (TABLE 1). In addition, recent studies have shown that more sophisticated normalization steps can improve data quality and can reduce technical variation[44–46,48]. The SWAN method[45] for subset quantile normalization is now available as a part of the minfi package, and RnBeads (TABLE 1) provides a user-friendly pipeline for the normalization and quality control of Infinium data in R/Bioconductor. The main result of data normalization (FIG. 1b) is a table of β-values (and, optionally, *M* values) that serves as the starting point for further analyses (FIG. 1c). β-values are conceptually equivalent to the absolute DNA methylation levels calculated from bisulphite-sequencing data, whereas *M* values are logistically transformed β-values and exhibit a distribution that is better suited for use with some common statistical tests[49].

Despite the use of normalization algorithms for reducing technical artefacts, several sources of bias tend to persist in normalized Infinium data. Most importantly, it seems that batch effects are almost always present in large-scale Infinium data sets[50], and they can introduce severe bias during subsequent analysis steps if no adequate countermeasures are taken[51]. Batch effects are particularly problematic when there is strong confounding between sources of technical bias and the phenotype of interest (for example, if all cases are processed in one week, and all control samples are processed in another week). Such a setup makes it difficult to distinguish between undesirable technical variation and meaningful biological differences. It is therefore

R/Bioconductor
A powerful command-line tool for data processing, statistical analysis and visualization of biological data sets.

β-values
An alternative term for the absolute DNA methylation levels, which stems from the observation that the distribution of DNA methylation levels across the genome resembles a β-distribution.

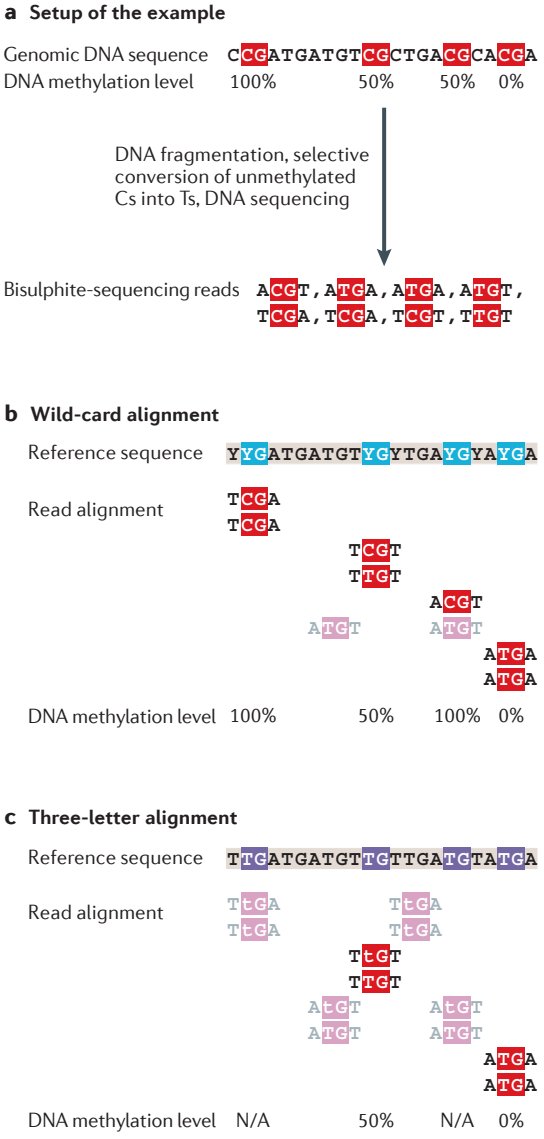

Figure 2 | **Two alternative strategies for bisulphite alignment. a** | An illustrative example of bisulphite sequencing for a DNA fragment with known DNA methylation levels at four CpGs and a total of eight bisulphite-sequencing reads. For easier visualization, the sequencing reads are four bases long (realistic numbers would be 50 to 200 bases), and the size of the genomic DNA sequence is just 23 bases (3 gigabases would be a realistic number for the human genome). **b** | Alignment of the bisulphite-sequencing reads (centre) to the reference sequence (top) using a wild-card aligner that tolerates zero mismatches and zero gaps. The aligner replaces each C in the reference sequence by the wild-card letter Y, which can match both C and T in the read sequences. Reads with more than one perfect alignment with the reference sequence are discarded (greyed out), and for each CpG in the genomic DNA sequence, the DNA methylation level (bottom) is calculated as the percentage of aligning Cs among all uniquely mapped reads. Note that the third CpG is incorrectly assigned a DNA methylation level of 100%, which is due to the fact that the unmethylated read was discarded as ambiguous, whereas the methylated read could be uniquely mapped. **c** | The same alignment carried out by a three-letter aligner, which also tolerates zero mismatches and zero gaps. The aligner replaces each C in the reference sequence by an upper-case T and each C in the sequencing reads by a lower-case t, with no distinction being made between upper-case T and lower-case t during the alignment. As a result of the reduced sequencing complexity with only three letters remaining, a larger number of reads align to more than one position in the reference sequence and are discarded. The three-letter alignment avoids incorrect results in this example, but it fails to provide any values for the first and third CpG. (As an alternative to discarding ambiguous reads, it is also possible to assign them randomly to one of the best-matching positions; in the current example, the wild-card alignment would provide correct results 50% of the time, whereas the three-letter alignment exhibits higher uncertainty and would be correct only 6.25% of the time.)

M values
Logistically transformed β-values. The transformation mitigates some statistical problems of the β-value (namely, limited value range and strongly bimodal distribution) at the cost of reduced biological interpretability.

Batch effects
Systematic biases in the data that are unrelated to the research question but that arise from undesirable (and often unrecognized) differences in sample handling.

Confounding
A nonrandom relationship between the phenotype of interest and external factors (for example, batch effects or population structure) that can give rise to spurious associations.

advisable to process samples in an order that minimizes confounding between potential sources of batch effects (for example, processing date and microarray batch) and the phenotype of interest (for example, cases versus controls) and to use tools for batch effect removal, which can substantially increase robustness and statistical power[50,52,53]. Other common biases in bisulphite microarray data include nonspecific binding of DNA fragments to multiple probes (which has been shown to cause false positives for sex-specific DNA methylation on the autosomes[54]) and the presence of genetic variants affecting probe binding or read-out. The impact of these technical issues can be minimized by removing all probes that exhibit a high sequence identity with multiple genomic regions as well as those overlapping with common genetic variants.

Processing enrichment-based data. Enrichment-based assays for DNA methylation mapping use various methods for enriching DNA in a methylation-specific manner. Methylated DNA can be enriched using methylation-specific antibodies (in methylated DNA immunoprecipitation coupled with high-throughput sequencing (MeDIP–seq)), methyl-CpG-binding domain (MBD) proteins (in MBD sequencing (MBD-seq)) or a restriction enzyme that specifically cuts methylated DNA (in methylation-dependent restriction enzyme sequencing (McrBC-seq)). Alternatively, unmethylated DNA can be enriched using restriction enzymes that specifically cut unmethylated DNA (for example, in HpaII tiny fragment enrichment by ligation-mediated PCR coupled with sequencing (HELP–seq)). Next-generation sequencing of the resulting DNA libraries counts the frequency of specific DNA fragments in each library and provides the raw data from which DNA methylation levels can be inferred. In contrast to bisulphite sequencing, the DNA methylation information is not contained in the read sequence but in the enrichment or depletion of sequencing reads that map to specific regions of the genome. As a result, enrichment-based methods require careful

Table 2 | **Quality control of DNA methylation data**

Category	Common problems and data quality issues	Background	Potential solutions
Study cohort	Genotyping data indicate systematic genetic differences (population structure) between cases and controls	DNA methylation levels are correlated with the DNA sequence and can differ significantly between haplotypes	Carry out stratified sampling to reduce the impact of population structure and/or to statistically correct for unavoidable population structure
Sample material	DNA quality, quantity and sample homogeneity are unsatisfactory or dissimilar between cases and cohorts	Standards for quality control of sample material have been established in the context of transcription analysis[152], and many of these points also apply to DNA methylation analysis	Bisulphite sequencing can provide more robust results than other technologies, especially when DNA amounts are small and of low quality. Bioinformatic analysis can, under some circumstances, correct for sample heterogeneity[130,153]
Sample handling	Evidence of sample mix-ups based on genotype, presence of the Y chromosome and/or X-chromosome inactivation status inferred from DNA methylation data	Sample mix-ups seem to occur with rates above 1% in essentially any large study. Genotype and sex can be inferred from sequencing and microarray data to identify and to resolve mix-ups	If genotype data are available for each individual, reconstruct the correct sample annotations by matching of unique genotypes. Otherwise, discard all potentially problematic samples or try to resolve mix-ups using quantitative trait locus (QTL) analysis[154]
Batch effects	Systematic differences between samples processed, for example, on different days, by different people or on different machines	Batch effects are hard to avoid in any genomic study, but they can in part be corrected using statistical methods, especially if potential sources of bias are documented and if there is little confounding between the batch effect and the phenotype of interest	ComBat can effectively correct for known or suspected batch effects[52], whereas the SVA algorithm tries to infer batch effects directly from the data[53]. Depending on the characteristics of the data set, one or the other approach may work better
Repetitive DNA	A large percentage of high-ranking differentially methylated regions (DMRs) overlap with repetitive regions	Repetitive DNA elements can cause cross-hybridization for microarray-based methods and inaccurate alignments for sequencing-based methods	By aligning all microarray probe sequences or sequencing reads to the reference genome and keeping track of multiple hits, it is possible to identify and flag problematic probes, reads and genomic regions[54,155,156]
Bisulphite conversion	Evidence of incomplete bisulphite conversion of unmethylated Cs (<99%), as indicated by unmethylated spike-in controls or by high levels of CpC methylation; alternatively, over-conversion of methylated Cs (>1%), as indicated by methylated spike-in controls	Although bisulphite conversion kits typically give rise to excellent results, difficult samples (for example, formalin-fixed, paraffin-embedded material) may require extensive optimization. Furthermore, overestimation or underestimation of DNA methylation levels may arise from sequencing into the constitutively methylated adaptor and from the use of constitutively unmethylated Cs during end repair	Extended or repeated bisulphite treatment increases the conversion rate of unmethylated Cs but at the cost of increased DNA degradation and over-conversion of methylated Cs[146,157,158]. Problems with adaptor sequences can be identified using FastQC and countered by aggressive adaptor trimming. Biases due to end repair can be identified using *M* bias plots[37] and can be resolved by trimming affected reads positions

handling of batch effects because any fluctuations in DNA-sequencing coverage will directly affect the DNA methylation measurement. Furthermore, to obtain absolute DNA methylation measurements, it is necessary to statistically correct for region-specific differences in CpG density.

The first step in the analysis of enrichment-based DNA methylation data is reference genome alignment, which can be done using a standard aligner, such as the BWA[55] or Bowtie[33]. On the basis of the alignment, relative enrichment scores are calculated by extending the sequencing reads to the estimated DNA fragment size and counting the number of unique reads that overlap with each CpG or with the genomic regions of interest (typically a tiling map of the reference genome). These enrichment scores are indicative of regional DNA methylation levels, but they are heavily confounded by the uneven distribution of CpGs throughout mammalian genomes. For example, a region with a high CpG density and moderate levels of DNA methylation can give rise to higher enrichment scores than a region with a low CpG density but with high levels of DNA methylation, and a region without any aligning reads can result either from the absence of DNA methylation or from difficulties in sequencing or aligning reads originating from this region. Several algorithms have been proposed for correcting this bias and for inferring absolute DNA methylation levels at a single-base resolution. The BATMAN algorithm[56] uses a Bayesian method, which provides accurate results but becomes impractically slow when applied to large data sets. The MEDME method uses a logistic regression model for data normalization, but it is rarely used owing to the need for calibration using a fully methylated reference sample[57]. The MEDIPS software combines concepts from BATMAN and MEDME into a data normalization and analysis pipeline that is sufficiently fast and easy to use to be practical for routine processing of MeDIP–seq and related data types[58]. Finally, MeQA[59] and MeDUSA[60] both provide convenient wrapper pipelines around BWA and MEDIPS, thus allowing semi-automated processing of enrichment-based data sets, and Repitools is a package for the R statistics software that facilitates quality control of enrichment-based data sets[61].

Enrichment scores
The relative enrichment of DNA fragments from a given genomic region compared to a control experiment (such as sequencing of unenriched DNA).

Tiling map
Segmentation of the genome into tiling windows of a fixed and typically small size (for example, 100 bases).

Logistic regression model
A type of regression model used for modelling the relationship between a binary outcome variable and one or more predictor variables.

Although the described normalization algorithms were originally developed for MeDIP–seq data, they are also useful for other methods that enrich for methylated DNA, including MBD-seq and McrBC-seq. By contrast, protocols that use restriction enzymes for enriching unmethylated DNA require a different approach. These restriction enzymes cleave unmethylated DNA at specific recognition sites (for example, unmethylated CCGG for the widely used HpaII restriction enzyme), resulting in fragment pools that depend at the same time on the genomic distribution of DNA methylation and on the genomic distribution of the targeted restriction sites. Bayesian networks have been used for inferring absolute DNA methylation levels for single CpGs from the complex fragment distributions[62], and a bioinformatic analysis pipeline has been described for processing data obtained with the HELP–seq assay[63]. Unfortunately, methods that enrich for unmethylated DNA were not included in any of the recent technology comparisons[64–66], but an earlier study based on tiling microarrays showed that their genomic coverage is fairly limited compared with alternative methods[67].

Data visualization and statistical analysis

For bisulphite-based protocols — and with appropriate normalization also for enrichment-based methods — the results of data processing can be summarized in a table containing absolute DNA methylation levels for each covered CpG in each analysed sample (FIG. 1b). This CpG methylation table constitutes the basis for subsequent analysis steps (FIG. 1c), and it decouples them from the experimental assay. The usefulness of separating data processing from data analysis is underlined by experiences from genome-wide association studies, which maintain standardized genotype tables in the variant call format (VCF) for analysing genotype data of various sources (for example, genotyping microarrays by different vendors or low-coverage genome sequencing). The definition of a similar 'methylation call format' (MCF) could facilitate the visualization and analysis of DNA methylation data using the methods outlined below.

Visualizing DNA methylation data. As the first step of DNA methylation analysis, it is useful to inspect a selection of genomic regions visually in a genome browser, including candidate genes for which biologically meaningful differences are suspected but also a set of randomly selected regions. To prepare for effective visual inspection, the CpG methylation tables derived during data processing (FIG. 1b) are converted into a file format that allows dynamic visualization of large data sets over the Internet. The bigBed format supports the visualization of single CpGs with a colour coding according to their DNA methylation levels[64]; alternatively, the bigWig format can represent DNA methylation levels of single CpGs by the heights of interspersed vertical bars[68]. These binary files are created from text-only BED and WIG files using specialized tools that reformat the data in a way that allows efficient extraction of region-specific information[69]. The resulting files are uploaded to a Web-accessible directory and imported into a Web-based genome browser for visualization (for example, UCSC Genome Browser[70], Ensembl[71] or WashU Human Epigenome Browser[72]). Alternatively, desktop genome browsers, such as IGV[73] and IGB[74], can be used for visual inspection of DNA methylation data. These locally running genome browsers tend to be somewhat faster but require the user to download and to import additional genome annotations that are available by default in Web-based genome browsers.

Complementarily to region-specific visualization with genome browsers, various types of diagrams can be used to obtain a more global view on DNA methylation data. Box plots and violin plots visualize the distribution of DNA methylation across the genome and can identify global changes (for example, genome-wide demethylation in differentiating erythrocytes[75] or widespread gain of DNA methylation in cultured cell lines[76]). The Hilbert curve method compresses genome-wide maps into compact, two-dimensional diagrams[77], and these are useful for detecting spatial patterns in the distribution of DNA methylation. Tree-like diagrams of DNA methylation across all classes of repeat elements in the reference genome highlight global trends in the epigenetic regulation of repetitive DNA[64]. Finally, scatter plots of DNA methylation levels provide a comprehensive picture of similarity and differences between sample pairs; because of their level of detail, such scatter plots constitute useful supplementary figures for convincing the reader of the quality and reproducibility of a data set[78,79]. All of these diagrams are fairly straightforward to generate using R/Bioconductor.

The most common goal of DNA methylation mapping is the identification of systematic differences between groups of samples, such as between patients afflicted by a disease and a healthy control group. Towards this goal, it is often useful to carry out hierarchical clustering on all relevant DNA methylation maps and to visualize the relative similarity among the samples by a clustering tree. For example, this approach successfully identified epigenetically and phenotypically distinct patient subsets in several cancers[80,81]. Most researchers carry out hierarchical clustering using standard methods that are available in their preferred statistical software package, such as 'hclust' and 'heatmap' in R. Moreover, specialized clustering algorithms have been developed[82–84] that account for the characteristic bimodal distribution of DNA methylation data and may give rise to more robust results than standard methods. Altogether, the described methods provide great flexibility for visualizing DNA methylation data, but in the absence of specialized graphical analysis software it is currently necessary to be familiar with a command-line tool — such as R/Bioconductor — to carry out effective visual analysis of DNA methylation data.

Identifying differentially methylated regions. After an initial analysis of global trends in a DNA methylation data set, the typical next step is the identification of differentially methylated regions (DMRs) that exhibit consistently different DNA methylation levels between sample groups (for example, cases versus controls).

CpG methylation table
A data table that contains DNA methylation levels (and, optionally, confidence scores) for each assayed CpG in each sample after normalization and quality control.

These DMRs can be as small as a single C or as large as an entire gene locus, depending on the biological question of interest and on the bioinformatic methods used for their identification. Although a single methylated CpG may occasionally be linked to gene expression regulation[85] and may affect disease risk[86,87], the vast majority of DMRs reported in the literature fall within a size range of a few hundred to a few thousand bases. This range coincides with typical sizes of gene-regulatory regions, and it is widely believed that DMRs can control cell-type-specific transcriptional repression of an associated gene[1,3,88].

DMR detection is commonly carried out on CpG methylation tables (FIG. 1c) — except for enrichment-based methods, which may profit from having the analysis carried out directly on read count data. In the most basic form of DMR detection, *t* tests or Wilcoxon rank-sum tests compare the DNA methylation levels of each C with sufficient data between two sample groups[89]. Several more advanced methods have been described that aim to improve DMR detection using mixture models[90], Shannon entropy[91], logistic *M* values[49], feature selection[92], stratification of *t* tests[93], aggregation of genomic regions by type[94], statistical correction for copy-number aberrations[95] or linear regression in combination with batch effect removal and peak detection[96]. Although the conceptual arguments behind each of these methods are plausible, in the absence of a systematic benchmarking study, it remains difficult to predict which methods will work best for real-world DNA methylation data sets.

Importantly, any statistical method that tests for differences in DNA methylation at a large number of genomic loci needs to correct for multiple hypothesis testing. This correction is almost exclusively done by controlling the false discovery rate (FDR). To that end, the distribution of uncorrected *P* values is analysed, and an FDR is inferred for each DMR. The inference is often made using the *q* value method available in R/Bioconductor[97]. Furthermore, an alternative method has recently been proposed to increase statistical power by modelling the dependence between statistical tests carried out for neighbouring CpGs[98]. Because of the large number of CpGs in the genome, only the strongest single-CpG differences tend to remain significant after multiple testing corrections. The result is often a high false-negative rate, especially when sample numbers and effect sizes are small.

Two complementary methods can be used to improve the statistical power for detecting weak differences in DNA methylation. First, statistical comparisons can be carried out on larger genomic regions rather than on single CpGs, such that neighbouring CpGs with similar differences in DNA methylation reinforce each other and give rise to more significant results (FIG. 3). This method can be applied to a genome-wide tiling map of the genome to support unconstrained genome-wide DMR discovery, or it can be focused on a preselected set of candidate genomic regions[64]; the latter approach substantially increases the statistical power for detecting those DMRs that are located among the candidate regions, but at the cost of missing out on DMRs elsewhere. Second, hierarchical models are useful for addressing the problem that small standard deviations frequently arise by chance and can result in highly significant but often spurious results[99]. By estimating the standard deviation of a given CpG or genomic region as the average of observed and expected values, more robust *P* values can be obtained for DNA methylation comparisons with many measurements and few samples per sample group.

False discovery rate
(FDR). A measure of significance that corrects for a large number of statistical tests being carried out on the same data set.

Effect size
A measure for the strength of association between two variables that provides important complementary information to *P* values and false discovery rates.

Validation and interpretation

Statistical comparison between sample groups typically results in a list of DMRs (FIG. 1c), which provides the basis for validation and biological interpretation of the observed differences (FIG. 1d). Importantly, statistical significance does not prove biological significance, and before proceeding to validation and interpretation, it is often useful to rank significant DMRs by a broader measure for the strength of association (for example, incorporating not only observed *P* values but also the relative and absolute differences in DNA methylation between sample groups). In a well-designed and successful study, it would be expected that experimental validation rates and biologically meaningful enrichment scores are highest for the first quartile of top-ranking DMRs and gradually decrease for the second, third and fourth quartiles of lower-ranking but still significant DMRs, thus providing a useful indicator of data quality and interpretability.

Verifying and validating differences in DNA methylation. After a ranked list of DMRs has been established, its accuracy and reproducibility should be confirmed by a combination of computational and experimental methods. As a first step, it is usually advisable to inspect the strongest DMRs manually in a genome browser and to look for warning signs that are indicative of technical artefacts. Common issues include excessive overlap of top-ranking DMRs with regions of repetitive DNA and unexplained clustering of DMRs in certain parts of the genome (TABLE 2). The manual inspection of individual DMRs should be complemented by quality-control plots visualizing global properties of the DMR list. For example, volcano plots show the relationship between statistical significance and the magnitude of differences in DNA methylation between sample groups, as in a recent study on EVI1-associated leukaemia[100]; Q–Q plots are useful for identifying global biases that can lead to inflated *P* values; and Manhattan plots demonstrate the distribution of DMRs across the genome[101,102].

When the computational analysis indicates that a data set is of sufficient quality to proceed, additional verification and/or validation experiments are useful for confirming the accuracy and reproducibility of the observed DMRs. In this context, verification refers to the experimental confirmation of DNA methylation measurements on the same set of samples using a different assay, which establishes the technical accuracy of the measurement; by contrast, validation is carried out in a new sample cohort with similar characteristics, and it aims to confirm the biological reproducibility of a DMR.

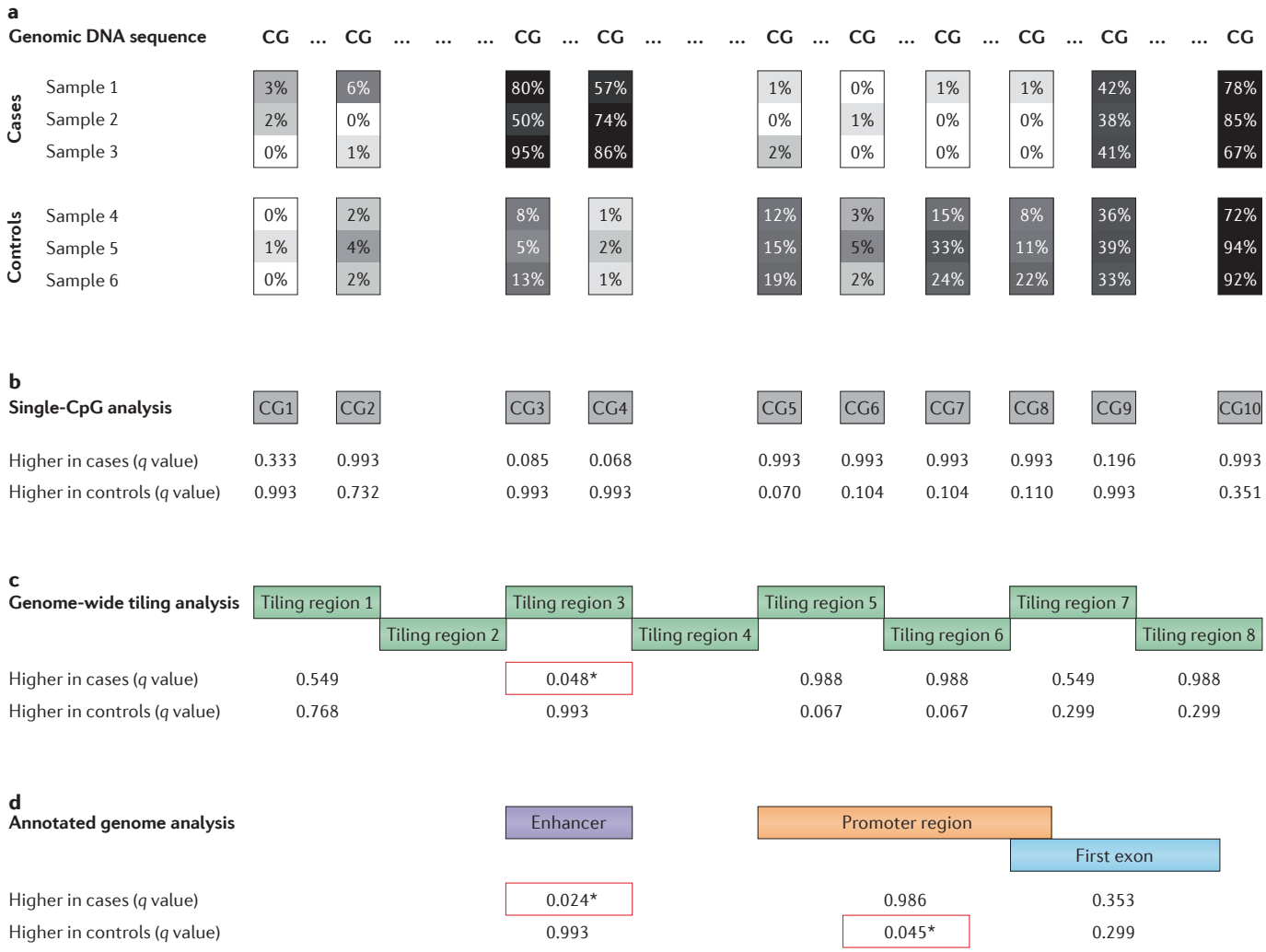

Figure 3 | **Effective identification of differentially methylated regions in a highly annotated genome.** **a** | An illustrative example of differences in DNA methylation within the promoter region of a gene and at an upstream enhancer. For easier visualization, DNA methylation data are shown for only three cases and three controls (a realistic number would be hundreds of samples) and for ten CpGs in total (dozens to hundreds of CpGs are realistic numbers for a typical promoter region). **b** | When the DNA methylation levels between cases and controls are compared at the resolution of single CpGs, all multiple-testing-corrected *q* values exceed 0.05 and are therefore considered to be insignificant. **c** | When combining statistical evidence from neighbouring CpGs over a fixed distance (tiling regions highlighted in green), one region is identified as being significantly more highly methylated among the cases compared to the controls (*q* value = 0.048). **d** | When combining statistical evidence across all CpGs that can be assigned to the same functional element on the basis of external genome annotation data, two regions are identified as being differentially methylated: the upstream enhancer (highlighted in purple) is significantly more highly methylated in the cases (*q* value = 0.024), and the promoter region (in orange) is significantly more highly methylated in the controls (*q* value = 0.045). The figure is based on the following statistical methods. Differences in DNA methylation at single CpGs (in **b**) are identified by unpaired, one-sided *t* tests, which assess whether or not the DNA methylation levels at the specific CpG are significantly higher among the cases than among the controls, and vice versa. The reason for using two separate one-sided tests lies in the ability to combine their results as described below; nevertheless, one two-sided test works equally well if no combination of *P* values is intended. For the tiling region analysis (in **c**), the locus is segmented into equally spaced regions, and the statistical significance for each of these regions is assessed using a generalization of Fisher's method[151]. This method combines the *P* values of all single CpGs that fall into the region while accounting for linear correlations between neighbouring CpGs (which are estimated to be at or below 0.8 on the basis of empirical observations for genome-wide bisulphite-sequencing data). The annotated genome analysis (in **d**) uses external genome annotation data to focus the statistical analysis on those combinations of CpGs that are likely to work together as an epigenetic switch: for example, by deactivating a known promoter or enhancer element. In all three cases, *q* values are calculated as estimates of the multiple-testing-corrected false discovery rate (FDR)[97], and a *q* value of 0.05 is used as the significance threshold for each direction of the comparison. Note that in this example the analysis of tiling regions increases the statistical power because neighbouring CpGs exhibit correlated changes in DNA methylation, and the incorporation of genome annotation data leads to further improvements, because the CpGs in the enhancer as well as those in the promoter exhibit a coordinated switch of their DNA methylation levels.

Box 2 | **Accessing public reference epigenome data sets**

Several large-scale initiatives are currently producing DNA methylation maps for a broad range of cell types and diseases. The International Human Epigenome Consortium (IHEC) is establishing comprehensive epigenome maps for 1,000 biomedically relevant human cell populations[16] — a task that has been distributed across multiple contributing projects. For example, the US Reference Epigenome Mapping Centers (REMC) project focuses on stem cells and primary tissue samples from healthy donors[149], the European BLUEPRINT project analyses various blood cell types and their associated diseases[150], and the German DEEP project investigates cell types that are relevant for metabolic and inflammatory diseases. All IHEC data are being distributed via the global bioinformatic hubs of the US National Center for Biotechnology Information (with its Gene Expression Omnibus (GEO) database) and the European Bioinformatics Institute (EBI; with its European Genome–Phenome Archive (EGA) database). Additional portals hosting IHEC data include the Epigenome Atlas at the Baylor College of Medicine, the Roadmap Epigenomics Visualization Hub at Washington University and the Ensembl Regulatory Build at the EBI. Complementarily to the focus of the IHEC on primary cell types, the Encyclopedia of DNA Elements (ENCODE) project provides extensive epigenome data for cultured cell lines. These data are available through the US National Institutes of Health GEO database, and they can also be browsed and downloaded from the ENCODE portal at the University of California, Santa Cruz. Finally, the International Cancer Genome Consortium (ICGC) establishes not only genome and transcriptome profiles but also epigenome maps of 50 different cancer types. Aggregated data are freely accessible from the ICGC Data Portal, whereas an application is required to access raw sequencing data and genotype information of individual patients.

Verification and validation are usually done using locus-specific DNA methylation assays[103,104], thereby reducing the cost of studying large validation cohorts. Software tools are available to support assay design and data analysis for the most widely used protocols. For example, MethMarker[105] provides a graphical user interface for designing and validating locus-specific DNA methylation assays based on bisulphite pyrosequencing, combined bisulphite restriction analysis (COBRA), methylation-specific PCR (MSP) and MethyLight. The MassArray R package supports assay design and data analysis for the mass-spectrometry-based Sequenom EpiTYPER assay[106], and PRIMEGENS[107], in combination with BiQ Analyzer HT[108], provides user-friendly support for deep bisulphite sequencing of selected loci, which is currently the validation method with the highest quantitative accuracy[109].

If DNA methylation mapping is carried out with the sole purpose of identifying a few interesting DMRs, then it is usually sufficient to validate the reproducibility of hand-picked DMRs in a second sample cohort. However, if the goal is to validate a large list of DMRs as a community resource, it is necessary to select DMRs randomly for validation to avoid biases due to manual selection of validation candidates. Whenever possible, researchers should report not only *P* values but also cross-validated sensitivity and specificity values to support the strength of the validated association. A carefully conducted validation study can make technical verification in the original cohort dispensable, especially when working with well-established DNA methylation assays.

Interpreting differences in DNA methylation. Although the biological conclusions drawn from a genome-wide DNA methylation data set ultimately depend on the researcher's understanding of the investigated phenotype, the process of data interpretation can be assisted by computational tools. For example, the Web-based EpiExplorer software facilitates hypothesis generation by allowing live exploration and interactive analysis of DMR lists in the context of public reference epigenome data sets[110], and genome analysis tools such as Galaxy[111] and the Genomic HyperBrowser[112] simplify the comparison of DMR data with other genomic data sets that are available online. Furthermore, researchers frequently apply gene set enrichment and pathway analysis tools[113] to identify biologically meaningful trends in lists of DMRs, as shown by a recent study of shared DNA methylation patterns in ageing and in bladder cancer[114]. Because most enrichment analysis tools require gene names as an input, DMRs need to be mapped to neighbouring genes before the enrichment analysis is carried out. Alternatively, we can use the GREAT Web server[115], which internally maps genomic regions to genes and statistically controls for the fact that genes differ in size and in their relative distance to each other.

Given that the correlation between promoter-associated DNA methylation and gene expression differences is modest in magnitude (Pearson correlation coefficients of around −0.3 are commonly observed[37,79]), gene-based enrichment analysis should be complemented with an enrichment analysis that is done directly on genomic regions[116,117]. For example, using a collection of genomic region sets obtained from Cistrome[118] and other public data sources, it has recently been shown that DMRs associated with *in vivo* differentiation of adult stem cells are much more strongly enriched for cell-type-specific transcription factor binding than for any functional gene annotations[79]. Despite the effectiveness of such analyses for identifying significant associations between DMRs and other types of genomic regions, it is important to keep in mind that correlation does not necessarily imply any causal relationship. Most DMRs exhibit moderate CpG densities[119] and therefore tend to co-locate with genomic regions that fall into a similar range of CpG densities, including weak CpG islands[120], CpG island shores[119] and other putative gene-regulatory elements. The risk of over-interpreting spurious associations can be reduced by selecting suitable control sets (for example, DMRs for other comparisons within the same study) and by using statistical models that explicitly control for CpG density as a confounding factor.

The interpretation of DNA methylation maps is further complicated by diverse sources of biological variation. First, inter-individual variation in DNA methylation is common among healthy individuals and appears to be heavily influenced by underlying genetic differences[121–125]. Before we can hope to correct for this source of variation statistically, it will be necessary to establish a comprehensive map of genetic quantitative trait loci (QTLs) that affect DNA methylation[126–128]. Second, different cell types in the same tissue or organ tend to exhibit highly characteristic DNA methylation profiles[79,129], which can lead to spurious DMRs if the cell composition is different between cases and controls. To address this issue, an algorithm has been developed for inferring the cell composition of whole blood on the

Bisulphite pyrosequencing
A locus-specific method for accurate quantification of DNA methylation levels at a small number of CpGs in many samples.

Combined bisulphite restriction analysis
(COBRA). A method that combines bisulphite treatment with sequence-specific restriction enzymes for locus-specific analysis of DNA methylation.

Methylation-specific PCR
(MSP). A method for highly sensitive detection of locus-specific DNA methylation using PCR amplification of bisulphite-converted DNA.

MethyLight
A variant of methylation-specific PCR that is highly quantitative and practical for measuring locus-specific DNA methylation levels in many samples.

EpiTYPER
An assay for measuring locus-specific DNA methylation in many samples on the basis of a combination of bisulphite treatment and mass spectrometry.

Cross-validation
A method for estimating the predictive power of a differentially methylated region or biomarker by carrying out training and validation on different portions of the same data set.

Quantitative trait loci
(QTLs). Genomic regions that control a phenotype of interest, such as the DNA methylation levels of another genomic region.

Mendelian randomization
Epidemiological method for assessing the causal role of an exposure for a phenotype of interest, using genetic variants that are affected neither by the exposure nor by the phenotype.

basis of aggregated DNA methylation maps[130]. Third, variation in DNA methylation within a tissue appears to be elevated in cancer samples[37,131,132], which may have important implications for cancer therapy[133,134]. Fourth, allele-specific DNA methylation is widespread in mammalian genomes, and bioinformatic methods have been developed for identifying affected genomic regions on the basis of genetic differences between alleles[135] or statistical evidence for a mixture of two allelic distributions underlying the observed DNA methylation levels[136,137]. Finally, the interpretation of EWAS results is complicated by the hybrid role of DNA methylation combining aspects of a genotype (that is, a heritable mark that can influence disease risk) with aspects of a cellular phenotype (that is, a molecular mechanism that can be affected by diseases), which makes it difficult to distinguish whether DNA methylation differences influence the phenotype of interest or vice versa. The concept of two-step Mendelian randomization has been suggested as a potential solution for this problem[138], but its validation in practice is still pending and will probably require large sample sets.

Future directions

The experimental and bioinformatic methods are now available for applying genome-wide DNA methylation mapping to a broad range of phenotypes, diseases and biological questions. Over the coming years, we will witness a surge of epigenomes being generated by individual laboratories and by public consortia (BOX 2). DNA methylation mapping will often complement genome sequencing and transcriptome mapping, as in the studies of the International Cancer Genome Consortium[139]. Furthermore, comprehensive mapping of histone modifications, nucleosome positioning, transcription factor binding and chromosomal organization will be carried out in those cell types that can be obtained in sufficient quantities, but for practical reasons they will be done at a much smaller scale than we will observe for DNA methylation mapping. These developments pose major challenges for computational epigenetics. Most importantly, we need to establish powerful and efficient tools for integrative data analysis to distil biologically relevant findings from the steeply growing heap of epigenome data. These tools will routinely combine user-generated data sets for a specific research question with publicly available reference epigenomes, they will detect common pathways as well as sample-specific deviations, and they will help users to derive biologically relevant hypotheses for experimental follow-up. Large parts of the bioinformatic analysis infrastructure will be operated using Web-based tools, but there will also be increasing pressure to ensure data privacy as DNA methylation mapping follows personal genome sequencing into the clinic.

Despite unprecedented progress in our ability to map and to analyse DNA methylation, we are far from understanding the causes and consequences of the differences in DNA methylation that we observe, and a host of new questions is already on the horizon. For example, what will we see when nanopore sequencing enables us to observe the chemical modifications of the DNA throughout the genome[140]? What is the functional relevance of the recently discovered variants of DNA methylation, including 5-hydroxymethylcytosine (5hmC), 5-formylcytosine and 5-carboxylcytosine[141]? Will hypervariable DNA methylation emerge as a new hallmark of cancer[133,134]? Can we predict the future behaviour of cells on the basis of their epigenomes? How can we design epigenetic combination therapies that will make a difference in the clinic[142]? It is likely that a comprehensive effort in functional epigenomics will be required to answer these questions and to interpret the large-scale epigenome maps that are currently being generated. These are exciting times for computational epigenetics!

1. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
2. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.* **10**, 2709–2721 (1982).
3. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Rev. Genet.* **13**, 484–492 (2012).
4. Barlow, D. P. Genomic imprinting: a mammalian epigenetic discovery model. *Annu. Rev. Genet.* **45**, 379–403 (2011).
5. Bestor, T. H. The host defence function of genomic methylation patterns. *Novartis Found. Symp.* **214**, 187–199 (1998).
6. Meissner, A. Epigenetic modifications in pluripotent and differentiated cells. *Nature Biotech.* **28**, 1079–1088 (2010).
7. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).
8. Martin, M. & Herceg, Z. From hepatitis to hepatocellular carcinoma: a proposed model for cross-talk between inflammation and epigenetic mechanisms. *Genome Med.* **4**, 8 (2012).
9. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. *Nature Rev. Cancer* **11**, 726–734 (2011).
10. Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**, 433–440 (2007).
11. Walker, C. L. & Ho, S. M. Developmental reprogramming of cancer susceptibility. *Nature Rev. Cancer* **12**, 479–486 (2012).
12. Laird, P. W. The power and the promise of DNA methylation markers. *Nature Rev. Cancer* **3**, 253–266 (2003).
13. Bock, C. Epigenetic biomarker development. *Epigenomics* **1**, 99–110 (2009).
14. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Rev. Genet.* **11**, 191–203 (2010).
15. Bock, C. & Lengauer, T. Computational epigenetics. *Bioinformatics* **24**, 1–10 (2008).
16. Satterlee, J. S., Schübeler, D. & Ng, H. H. Tackling the epigenome: challenges and opportunities for collaboration. *Nature Biotech.* **28**, 1039–1044 (2010).
17. Foley, D. L. *et al.* Prospects for epigenetic epidemiology. *Am. J. Epidemiol.* **169**, 389–400 (2009).
18. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nature Rev. Genet.* **12**, 529–541 (2011).
This Review describes the planning and execution of an EWAS for common diseases.
19. Robinson, M. D., Statham, A. L., Speed, T. P. & Clark, S. J. Protocol matters: which methylome are you actually studying? *Epigenomics* **2**, 587–598 (2010).
20. Beck, S. Taking the measure of the methylome. *Nature Biotech.* **28**, 1026–1028 (2010).
21. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
22. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
23. Frith, M. C., Mori, R. & Asai, K. A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Res.* **40**, e100 (2012).
24. Coarfa, C. *et al.* Pash 3.0: a versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. *BMC Bioinformatics* **11**, 572 (2011).
25. Smith, A. D. *et al.* Updates to the RMAP short-read mapping software. *Bioinformatics* **25**, 2841–2842 (2009).
26. Xi, Y. *et al.* RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinformatics* **28**, 430–432 (2012).
27. Otto, C., Stadler, P. F. & Hoffmann, S. Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics* **28**,1698–1704 (2012).
28. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
29. Harris, E. Y., Ponts, N., Levchuk, A., Roch, K. L. & Lonardi, S. BRAT: bisulfite-treated reads analysis tool. *Bioinformatics* **26**, 572–573 (2010).
30. Harris, E. Y., Ponts, N., Le Roch, K. G. & Lonardi, S. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. *Bioinformatics* **28**, 1795–1796 (2012).
31. Chen, P. Y., Cokus, S. J. & Pellegrini, M. BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* **11**, 203 (2010).
32. Pedersen, B., Hsieh, T. F., Ibarra, C. & Fischer, R. L. MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics* **27**, 2435–2436 (2011).

33. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
34. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Rev. Genet.* **12**, 443–451 (2011).
35. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* **13**, R61 (2012).
36. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS ONE* **6**, e16327 (2011).
37. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nature Genet.* **43**, 768–775 (2011).
This reference describes a comprehensive and well-documented analysis of a cancer-specific DNA methylation data set.
38. Krueger, F., Kreck, B., Franke, A. & Andrews, S. R. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods* **9**, 145–151 (2012).
39. Chung, C. A. High-throughput sequencing of the methylome using two-base encoding. *Methods Mol. Biol.* **910**, 71–86 (2012).
40. Kreck, B. *et al.* B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics* **28**, 428–429 (2012).
41. Ramsahoye, B. H. *et al.* Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl Acad. Sci. USA* **97**, 5237–5242 (2000).
42. Ziller, M. J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet.* **7**, e1002389 (2011).
43. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
44. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**, 771–784 (2011).
This study carried out an independent empirical evaluation of the Illumina Infinium 450k assay.
45. Makismovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).
46. Touleimat, N. & Tost, J. Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325–341 (2012).
47. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
48. Wang, D. *et al.* Comparison of different normalization assumptions for analyses of DNA methylation data from the cancer genome. *Gene* **506**, 36–42 (2012).
49. Du, P. *et al.* Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
50. Teschendorff, A. E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**, 1496–1505 (2011).
51. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Rev. Genet.* **11**, 733–739 (2010).
This Perspectives article highlights the prevalence of batch effects in genomic data and suggests ways of addressing this problem.
52. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
53. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
54. Chen, Y. A. *et al.* Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. *Genomics* **97**, 214–222 (2011).
55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
56. Down, T. A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotech.* **26**, 779–785 (2008).
57. Pelizzola, M. *et al.* MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Res.* **18**, 1652–1659 (2008).
58. Chavez, L. *et al.* Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res.* **20**, 1441–1450 (2010).
59. Huang, J. *et al.* MeQA: a pipeline for MeDIP-seq data quality assessment and analysis. *Bioinformatics* **28**, 587–588 (2012).
60. Wilson, G. *et al.* Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers. *GigaScience* **1**, 3 (2012).
61. Statham, A. L. *et al.* Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics* **26**, 1662–1663 (2010).
62. Singer, M. *et al.* MetMap enables genome-scale Methyltyping for determining methylation states in populations. *PLoS Comput. Biol.* **6**, e1000888 (2010).
63. Jing, Q., McLellan, A., Greally, J. M. & Suzuki, M. Automated computational analysis of genome-wide DNA methylation profiling data from HELP-tagging assays. *Methods Mol. Biol.* **815**, 79–87 (2012).
64. Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature Biotech.* **28**, 1106–1114 (2010).
65. Harris, R. A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotech.* **28**, 1097–1105 (2010).
66. Robinson, M. D. *et al.* Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation. *Genome Res.* **20**, 1719–1729 (2010).
References 64, 65 and 66 carried out an empirical benchmarking of widely used methods for genome-wide DNA methylation mapping.
67. Irizarry, R. A. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* **18**, 780–790 (2008).
68. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
69. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).
70. Karolchik, D. *et al.* The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* **36**, D773–D779 (2008).
71. Flicek, P. *et al.* Ensembl 2008. *Nucleic Acids Res.* **36**, D707–D714 (2008).
72. Zhou, X. *et al.* The Human Epigenome Browser at Washington University. *Nature Methods* **8**, 989–990 (2011).
This reference describes a useful Web-based software tool for visualization and graphical analysis of human reference epigenomes.
73. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotech.* **29**, 24–26 (2011).
74. Nicol, J. W., Helt, G. A., Blanchard, S. G. Jr, Raja, A. & Loraine, A. E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730–2731 (2009).
75. Shearstone, J. R. *et al.* Global DNA demethylation during mouse erythropoiesis *in vivo*. *Science* **334**, 799–802 (2011).
76. Smiraglia, D. J. *et al.* Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum. Mol. Genet.* **10**, 1413–1419 (2001).
77. Anders, S. Visualization of genomic data with the Hilbert curve. *Bioinformatics* **25**, 1231–1235 (2009).
78. Bock, C. *et al.* Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell* **144**, 439–452 (2011).
79. Bock, C. *et al.* DNA methylation dynamics during *in vivo* differentiation of blood and skin stem cells. *Mol. Cell* **47**, 633–647 (2012).
80. Figueroa, M. E. *et al.* Leukemic *IDH1* and *IDH2* mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* **18**, 553–567 (2010).
81. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010).
82. Houseman, E. A. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of β distributions. *BMC Bioinformatics* **9**, 365 (2008).
83. Marjoram, P., Chang, J., Laird, P. W. & Siegmund, K. D. Cluster analysis for DNA methylation profiles having a detection threshold. *BMC Bioinformatics* **7**, 361 (2006).
84. Siegmund, K. D., Laird, P. W. & Laird-Offringa, I. A. A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics* **20**, 1896–1904 (2004).
85. Xu, J. *et al.* Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proc. Natl Acad. Sci. USA* **104**, 12377–12382 (2007).
86. Raval, A. *et al.* Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell* **129**, 879–890 (2007).
87. Moser, D. *et al.* Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site. *Neuropsychopharmacology* **34**, 458–467 (2008).
88. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
89. Wang, D. *et al.* IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28**, 729–730 (2012).
90. Wang, S. Method to detect differentially methylated loci with case-control designs using Illumina arrays. *Genet. Epidemiol.* **35**, 686–694 (2011).
91. Zhang, Y. *et al.* QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res.* **39**, e58 (2011).
92. Zhuang, J., Widschwendter, M. & Teschendorff, A. E. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics* **13**, 59 (2012).
93. Chen, Z., Liu, Q. & Nadarajah, S. A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data. *Bioinformatics* **28**, 1109–1113 (2012).
94. Poage, G. M. *et al.* Identification of an epigenetic profile classifier that is associated with survival in head and neck cancer. *Cancer Res.* **72**, 2728–2737 (2012).
This study demonstrates how the aggregation by genomic sequence features can reduce the multiple-testing burden and increase the power of a small and otherwise underpowered EWAS.
95. Robinson, M. D. *et al.* Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Res.* 9 Aug 2012 (doi:10.1101/gr.139055.112).
96. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200–209 (2012).
This study describes a flexible workflow for identifying DMRs in a statistically sound manner.
97. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
98. Kuan, P. F. & Chiang, D. Y. Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. *Biometrics* 19 Jan 2012 (doi:10.1111/j.1541-0420.2011.01730.x).
99. Ji, H. & Liu, X. S. Analyzing 'omics data using hierarchical models. *Nature Biotech.* **28**, 337–340 (2010).
100. Lugthart, S. *et al.* Aberrant DNA hypermethylation signature in acute myeloid leukemia directed by EVI1. *Blood* **117**, 234–241 (2011).
101. Barfield, R. T., Kilaru, V., Smith, A. K. & Conneely, K. N. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics* **28**, 1280–1281 (2012).
102. Kilaru, V., Barfield, R. T., Schroeder, J. W., Smith, A. K. & Conneely, K. N. MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics* **7**, 225–229 (2012).
103. Kristensen, L. S. & Hansen, L. L. PCR-based methods for detecting single-locus DNA methylation biomarkers in cancer diagnostics, prognostics, and response to treatment. *Clin. Chem.* **55**, 1471–1483 (2009).

104. Sepulveda, A. R. *et al.* CpG methylation analysis-current status of clinical assays and potential applications in molecular diagnostics: a report of the Association for Molecular Pathology. *J. Mol. Diagn.* **11**, 266–278 (2009).
105. Schüffler, P., Mikeska, T., Waha, A., Lengauer, T. & Bock, C. MethMarker: user-friendly design and optimization of gene-specific DNA methylation assays. *Genome Biol.* **10**, R105 (2009).
106. Thompson, R. F., Suzuki, M., Lau, K. W. & Greally, J. M. A pipeline for the quantitative analysis of CG dinucleotide methylation using mass spectrometry. *Bioinformatics* **25**, 2164–2170 (2009).
107. Srivastava, G. P., Guo, J., Shi, H. & Xu, D. PRIMEGENS-v2: genome-wide primer design for analyzing DNA methylation patterns of CpG islands. *Bioinformatics* **24**, 1837–1842 (2008).
108. Lutsik, P. *et al.* BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Res.* **39**, W551–W556 (2011).
109. Potapova, A. *et al.* Systematic cross-validation of 454 sequencing and pyrosequencing for the exact quantification of DNA methylation patterns with single CpG resolution. *BMC Biotechnol.* **11**, 6 (2011).
110. Halachev, K., Bast, H., Albrecht, F., Lengauer, T. & Bock, C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. *Genome Biol.* (in the press).
This reference describes a Web-based software tool for interactive exploration and biological hypothesis generation based on epigenome data.
111. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
112. Sandve, G. K. *et al.* The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol.* **11**, R121 (2010).
113. Nam, D. & Kim, S. Y. Gene-set approach for expression pattern analysis. *Brief Bioinform.* **9**, 189–197 (2008).
114. Marsit, C. J. *et al.* DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. *J. Clin. Oncol.* **29**, 1133–1139 (2011).
115. McLean, C. Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nature Biotech.* **28**, 495–501 (2010).
116. Bock, C., Halachev, K., Büch, J. & Lengauer, T. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi-) genomic data. *Genome Biol.* **10**, R14 (2009).
117. Hackenberg, M. & Matthiesen, R. Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. *Bioinformatics* **24**, 1386–1393 (2008).
118. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
119. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genet.* **41**, 178–186 (2009).
120. Bock, C., Walter, J., Paulsen, M. & Lengauer, T. CpG island mapping by epigenome prediction. *PLoS Comput. Biol.* **3**, e110 (2007).
121. Bjornsson, H. T. *et al.* Intra-individual change over time in DNA methylation with familial clustering. *JAMA* **299**, 2877–2883 (2008).
122. Bock, C., Walter, J., Paulsen, M. & Lengauer, T. Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res.* **36**, e55 (2008).
123. Gertz, J. *et al.* Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet.* **7**, e1002228 (2011).
124. Heijmans, B. T., Kremer, D., Tobi, E. W., Boomsma, D. I. & Slagboom, P. E. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human *IGF2/H19* locus. *Hum. Mol. Genet.* **16**, 547–554 (2007).
125. Xie, H. *et al.* Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.* **39**, 4099–4108 (2011).
126. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).
127. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
128. Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86**, 411–419 (2010).
References 125, 126 and 127 describe the systematic identification of genetic variants that are associated with DNA methylation levels across the genome.
129. Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338–342 (2010).
130. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
This reference describes a computational method for inferring the cellular composition of heterogeneous tissues from aggregate DNA methylation data.
131. Jaffe, A. E., Feinberg, A. P., Irizarry, R. A. & Leek, J. T. Significance analysis and statistical dissection of variably methylated regions. *Biostatistics* **13**, 166–178 (2012).
132. Teschendorff, A. E. & Widschwendter, M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics* **28**, 1487–1494 (2012).
133. Feinberg, A. P. & Irizarry, R. A. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl Acad. Sci. USA* **107**, S1757–S1764 (2010).
134. Pujadas, E. & Feinberg, A. P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148**, 1123–1131 (2012).
135. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816–831 (2012).
136. Fang, F. *et al.* Genomic landscape of human allele-specific DNA methylation. *Proc. Natl Acad. Sci. USA* **109**, 7332–7337 (2012).
137. Peng, Q. & Ecker, J. R. Detection of allele-specific methylation through a generalized heterogeneous epigenome model. *Bioinformatics* **28**, i163–i171 (2012).
138. Relton, C. L. & Davey Smith, G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int. J. Epidemiol.* **41**, 161–176 (2012).
139. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
140. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnol.* **4**, 265–270 (2009).
141. Wu, H. & Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev.* **25**, 2436–2452 (2011).
142. Bock, C. & Lengauer, T. Managing drug resistance in cancer: lessons from HIV therapy. *Nature Rev. Cancer* **12**, 494–501 (2012).
143. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7**, 461–465 (2010).
144. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
145. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* **5**, e8888 (2010).
146. Gu, H. *et al.* Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature Methods* **7**, 133–136 (2010).
147. Diep, D. *et al.* Library-free methylation sequencing with bisulfite padlock probes. *Nature Methods* **9**, 270–272 (2012).
148. Lee, E. J. *et al.* Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. *Nucleic Acids Res.* **39**, e127 (2011).
149. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotech.* **28**, 1045–1048 (2010).
150. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotech.* **30**, 224–226 (2012).
151. Morton, B. B. A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987–992 (1975).
152. Allison, D. B., Cui, X., Page, G. P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nature Rev. Genet.* **7**, 55–65 (2006).
153. Shen-Orr, S. S. *et al.* Cell type-specific gene expression differences in complex tissues. *Nature Methods* **7**, 287–289 (2010).
154. Westra, H. J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104–2111 (2011).
155. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144–2146 (2011).
156. Zhang, X., Mu, W. & Zhang, W. On the analysis of the Illumina 450k array data: probes ambiguously mapped to the human genome. *Front. Genet.* **3**, 73 (2012).
157. Ehrich, M., Zoll, S., Sur, S. & van den Boom, D. A new method for accurate assessment of DNA quality after bisulfite treatment. *Nucleic Acids Res.* **35**, e29 (2007).
158. Warnecke, P. M. *et al.* Identification and resolution of artifacts in bisulfite sequencing. *Methods* **27**, 101–107 (2002).

Acknowledgements
The author would like to thank S. Beck, M. Esteller, T. Lengauer, A. Meissner, H. Stunnenberg and J. Walter for helpful discussions and all past and present collaborators for sharing their ideas, data and insights.

Competing interests statement
The author declares no competing financial interests.

FURTHER INFORMATION
Christoph Bock's homepage: http://www.medical-epigenomics.org
CeMM Research Center for Molecular Medicine: http://www.cemm.oeaw.ac.at
ENCODE Project at UCSC: http://genome.ucsc.edu/ENCODE
Ensembl Regulatory Build: http://www.ensembl.org/info/docs/funcgen/index.html
European BLUEPRINT Epigenome Mapping Consortium: http://www.blueprint-epigenome.eu
European Genome–Phenome Archive (EGA): https://www.ebi.ac.uk/ega
Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo
Human Epigenome Atlas: http://www.epigenomeatlas.org
ICGC Data Portal: http://dcc.icgc.org/web
International Human Epigenome Consortium: http://www.ihec-epigenomes.org
***Nature Reviews Genetics* Series on Computational tools:** http://www.nature.com/nrg/series/computational
Roadmap Epigenomics Visualization Hub: http://genomebrowser.wustl.edu
US NIH Roadmap Epigenomics Mapping Consortium: http://www.roadmapepigenomics.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF