DNA methylome analysis using short bisulfite sequencing data

Felix Krueger[1,3], Benjamin Kreck[2,3], Andre Franke[2] & Simon R Andrews[1]

Bisulfite conversion of genomic DNA combined with next-generation sequencing (BS-seq) is widely used to measure the methylation state of a whole genome, the methylome, at single-base resolution. However, analysis of BS-seq data still poses a considerable challenge. Here we summarize the challenges of BS-seq mapping as they apply to both base and color-space data. We also explore the effect of sequencing errors and contaminants on inferred methylation levels and recommend the most appropriate way to analyze this type of data.

DNA methylation involves the addition of a methyl group to the C5 carbon residue (5mC) of cytosines by DNA methyltransferases[1,2]. DNA methylation is an important epigenetic mechanism used by higher eukaryotes and is involved in several key physiological processes, including regulation of gene expression, X-chromosome inactivation, imprinting, and silencing of germline-specific genes and repetitive elements[3]. Patterns of methylation are stably maintained through somatic cell division and can be inherited across generations. These patterns are sometimes perturbed in important human diseases, such as imprinting disorders and cancer[3–5]. Understanding how methylation patterns are established and maintained is therefore of great importance.

The sequence context in which a cytosine occurs is a key factor in determining the regulation of its methylation. Cytosines that occur as part of a C-G dinucleotide (CpG) are often highly methylated (~60–80% in mammals[2]) and are regulated differently to cytosines in other contexts. CpG methylation usually occurs on both DNA strands[1] to maintain methylation at CpGs during DNA replication. In contrast, non-CpG methylation must be re-established *de novo* after each cell division. Although it is present at considerable levels during early development or in pluripotent cell types[6–8], most non-CpG cytosines are generally unmethylated in differentiated tissues (~0.3–3% in mammals[2]).

As methylated cytosines are susceptible to spontaneous conversion into thymines through chemical deamination, they tend to be generally underrepresented in the genome[9,10], and they are often grouped in dense patches termed CpG islands. These islands tend to be unmethylated in the germline and are consequently less vulnerable to spontaneous deamination[11]. CpG islands are frequently associated with promoters, and the regulation of promoter methylation has been shown to affect the expression of the corresponding transcripts. Traditionally, CpG islands were defined using sequence-composition analysis[12–14] which predicted that the mouse genome contained a substantially lower number of CpG islands than the human genome. However, a recent report demonstrated that the occurrence of functional CpG islands is in fact quite similar in the two organisms[15].

Measuring methylation

Several methods exist for measuring DNA methylation at specific genomic loci, and these have been reviewed recently[16,17]. They range from methylated DNA immunoprecipitation or methyl binding protein enrichment of methylated fragments[18–20] to digestion with methylation-sensitive restriction enzymes[21] and bisulfite modification of DNA[22]. Comparisons of these methods showed that they all can be used to produce accurate DNA methylation data[21,23,24].

During bisulfite sequencing the treatment of DNA with sodium bisulfite converts cytosines into uracils, whereas methylcytosines remain unmodified. Uracils are read as thymines by DNA polymerase, so

[1]Bioinformatics Group, The Babraham Institute, Cambridge, UK. [2]Institute of Clinical Molecular Biology, Christian Albrechts University, Kiel, Germany. [3]These authors contributed equally to this work. Correspondence should be addressed to A.F. (a.franke@mucosa.de) or S.R.A. (simon.andrews@babraham.ac.uk).

PUBLISHED ONLINE 30 JANUARY 2012; DOI:10.1038/NMETH.1828

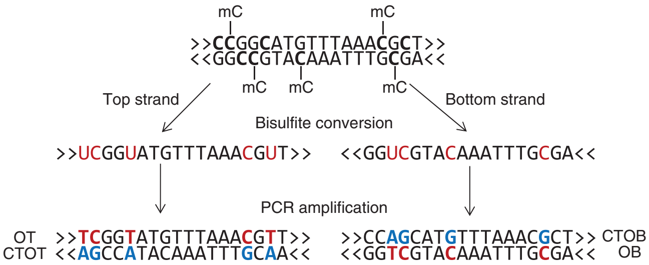

Figure 1 | Effect of bisulfite treatment of DNA. Bisulfite conversion of genomic DNA and subsequent PCR amplification gives rise to two PCR products and up to four potentially different DNA fragments for any given locus. (Hydroxy)methylated cytosine residues are resistant to bisulfite conversion and can be used as a readout of the DNA methylation state. mC, 5-methylcytosine; hmC, 5-hydroxymethylcytosine; OT, original top strand; CTOT, strand complementary to the original top strand; OB, original bottom strand; and CTOB, strand complementary to the original bottom strand.

amplifying bisulfite-treated DNA by PCR yields products in which unmethylated cytosines appear as thymines. By comparing the modified DNA with the original sequence, the methylation state of the original DNA can therefore be inferred. Bisulfite treatment of 5-hydroxymethylcytosine (5hmC) yields a similar intermediate to 5mC, meaning that BS-seq can be used to detect whether a position is (hydroxy-) methylated but not to determine the exact type of modification[21,25] (**Fig. 1**). This limitation does not apply to antibody-based techniques, which can be used to specifically enrich 5hmC[26–28].

Capillary electrophoresis–based bisulfite sequencing was considered the gold standard for methylation analysis because of its clear readout and single-base resolution[22], but it could only be applied to relatively small regions. New sequencing technologies mean that BS-seq is now a viable option for the sequencing of entire mammalian methylomes[6–8,29–32] (**Supplementary Table 1**).

For researchers primarily interested in CpG island methylation, the cost of bisulfite sequencing can be reduced by enriching CpG-dense regions by digesting genomic DNA with a methylation-insensitive restriction enzyme containing a C-G as part of its recognition site and selecting short fragments[6,30,33]. Even though the selected fragments are used to interrogate only a few percent of the genome, these data are informative for the majority of CpG islands. This approach, termed reduced representation BS-seq (RRBS), has been extensively described and compared to other techniques[23,33–35], and several genome-wide methylation maps based on RRBS have been reported[6,30].

In this Review we provide an overview of the computational analysis of bisulfite sequencing data. We highlight points to consider when designing a BS-seq experiment and point out pitfalls that can occur during the initial analysis. We also discuss different alignment strategies and their implementation by current bioinformatic tools. In particular, we present the main differences between the analysis of base space (Illumina) and color space (SOLiD, Applied Biosystems) BS-seq data.

Challenges of BS-seq data mapping

As the methylation state of bisulfite-treated DNA must be inferred by comparison to an unmodified reference sequence, a correct alignment is of critical importance. This is challenging because the aligned sequences do not exactly match the reference, and the complexity of the libraries is reduced. Also, as cytosine methylation is not symmetrical, the two strands of DNA in the reference genome must be considered separately. A single site can have a different methylation state in different cells. Thus, when sequencing cell mixtures or tissue fractions, the percentage of methylation at each site needs to be determined[36].

When performing an alignment one must discriminate between different types of bisulfite-treated DNA libraries (for a schematic drawing, see ref. 16). In the first, termed directional libraries, adapters are attached to the DNA fragments such that only the original top or bottom strands will be sequenced[7,30]. Alternatively, all four DNA strands that arise through bisulfite treatment and subsequent PCR amplification can be sequenced with the same frequency in nondirectional libraries[32,37,38]. BS-seq mapping may therefore require up to four different strand alignments to be analyzed for each sequence. Because of the complexity of BS-seq alignments, standard sequence alignment software cannot be used. However, several different tools for BS-seq analysis have been developed.

Base-space BS-seq data alignments

Methylation-'aware' alignment tools consider both cytosine and thymine as potential matches to a genomic cytosine. This strategy provides the highest possible mapping efficiency (high sensitivity) because it makes optimal use of the information present in the reads. However, a drawback of this technique is that methylated sequences will be aligned with greater efficiency because they carry more information than their unmethylated counterparts, leading this type of aligner to overestimate methylation levels.

Alternatively, in unbiased approaches usually any residual cytosines in the BS-seq read and all cytosines in the reference genome are converted into thymines before the alignment is performed[7,30]. This means that the read sequence to be aligned is unaffected by its methylation state. It also means that there will be an exact match between the converted read and converted genome sequence so that standard sequence alignment tools can be used to perform the mapping[39,40]. This approach, however, comes at the cost of slightly reduced mapping efficiencies (**Fig. 2a**).

BS-seq in color space

In contrast to the intuitive base-space sequence generated by Illumina sequencers, SOLiD sequencing (Applied Biosystems) encodes its reads in color space such that each color resembles the transition from one base to the next[41]. Single-nucleotide polymorphisms (SNPs) can be called with high confidence because they will result in two adjacent color changes, whereas technical errors are indicated by a single color change (**Supplementary Fig. 1a,b**). Owing to the way color-space encoding works, residual cytosines are correctly converted into thymines in the bisulfite reads *in silico* before the mapping only if the reads are completely error-free. A single measurement error in the read would lead to incorrect conversions throughout the rest of the read (**Supplementary Fig. 1c**). As a consequence, the *in silico* cytosine to thymine conversion, which guarantees unbiased alignments, should not be performed on color-space datasets.

Current tools to align color space BS-seq data to a reference genome either use methylation-aware alignments (SOCS-B[42]), which can be computationally intensive for complex genomes,

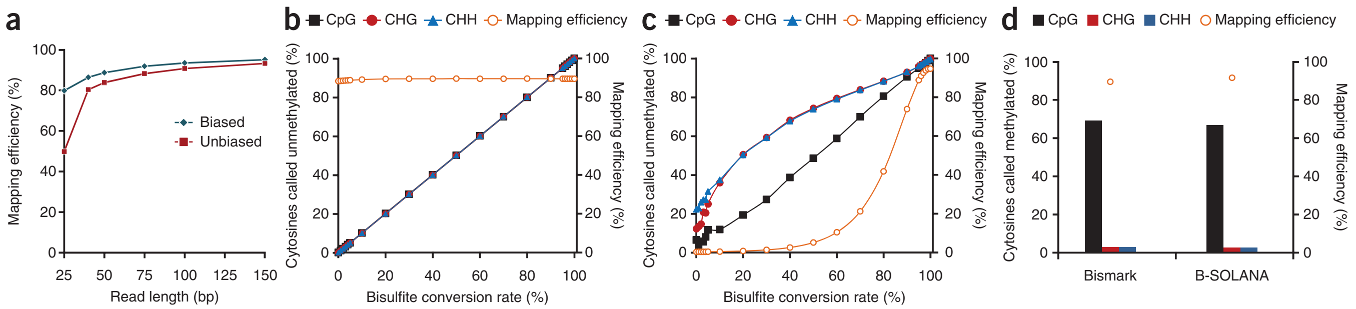

Figure 2 | Performance and accuracy of unbiased base-space and color-space BS-seq alignment tools. (**a**) A total of 10^6 random mouse genomic sequences of different lengths were aligned to the mouse genome (NCBIM37) with Bowtie as an example of methylation-aware mapping (biased) or with Bismark as an example of unbiased mapping (unbiased). Non-unique alignments were discarded. (**b**,**c**) A total of 10^6 random mouse base-space (Bismark; **b**) or human color-space (B-SOLANA; **c**) reads (75 base pairs) were simulated with different rates of bisulfite conversion (context is indicated) and aligned to the mouse (NCBIM37) or human (NCBI37) genomes. Bismark accurately detected various simulated methylation levels at a constant mapping efficiency. Alignment of color-space reads with B-SOLANA was efficient, and methylation calls were accurate only when methylation in non-CpG context was fairly low (ideally less than 5%). H (in CHG and CHH) stands for C, T or A. (**d**) Reads as in **b**,**c** were simulated with typical mammalian methylation levels (CpG context, 70%; CHG and CHH context, 3%) using Sherman (http://www.bioinformatics.bbsrc.ac.uk/projects/sherman/).

or align reads to *in silico*–converted versions of the reference genome, with bisulfite-induced mismatches are treated as normal mismatches. As different levels of methylation can result in increased numbers of mismatches to the reference genome, this approach is, however, not free of bias. One possibility to reduce mapping bias is to apply different *in silico* conversions to the reference genome and determine best alignments from the combined set of results of different mapping runs. This approach, however, requires prior knowledge of the methylation characteristics of the organism to be analyzed.

BS-seq data alignment tools

Several tools have been developed for the analysis of BS-seq datasets[17]. These not only differ considerably regarding their alignment speed, flexibility and ease of use but also in the information they report. Many older BS-seq data aligners only reported a bisulfite read mapping output, and the user had to extract methylation information from the alignments. More recent tools provide a comprehensive methylation output, which enables the end user to explore the biological effects of methylation more quickly[39,40,43]. Most recent tools, such as Bismark[40], BS-Seeker[39] or B-SOLANA[44], use existing short-read aligners (Bowtie[45] for the mentioned tools) and handle the requirements unique to BS-seq data analysis internally.

An example for a color-space alignment tool is B-SOLANA, with which reads are initially aligned to a reference genome in which all cytosines in non-CpG context had been *in silico*–converted into thymines and are then aligned to a second reference genome in which cytosines in all sequence contexts are converted into thymines. Unlike bisulfite alignments in base space (**Fig. 2b**), this method is not suited to accurate detection of arbitrary methylation levels in unknown samples because a high degree of methylation in the non-CpG context would produce too many mismatches in the alignment step (**Fig. 2c**). This would lead to a dramatic decrease in mapping efficiency and an apparent bias toward hypomethylation in the non-CpG context. However, for the majority of eukaryotic genomes with less than 5% methylation in non-CpG context[2], alignments can be generated efficiently and accurately (**Fig. 2d**).

In the rest of this Review we use Bismark to illustrate different aspects of BS-seq analysis. Bismark can accurately detect the simulated methylation state of cytosines in any sequence context while the mapping efficiency is completely unaffected (**Fig. 2b**). We summarize details of different software packages for BS-seq data analysis in **Table 1**.

Once a dataset consisting of best alignments has been determined based on predefined alignment criteria, the methylation state of positions involving cytosines in the reference sequence can be inferred. Then these methylation calls can serve to determine the ratios of methylated versus unmethylated cytosines at every position assayed. Later, analyses of the methylation data could include looking at minimum read depths, determining methylation states of individual cytosines or genomic features, or estimating cytosine-conversion errors or false discovery rates. The biological analysis of methylome data is manifold and beyond the scope of this review.

Factors affecting the accuracy of methylation calls

Two key factors are crucial when determining the methylation state of a read from a BS-seq experiment. First, the sequence of the read must be correct and derive entirely from a bisulfite-converted sequence in the original genome. Second, the read must be correctly mapped to the corresponding position of the targeted genome. Failure to meet either of these criteria will result in the generation of incorrect methylation calls and, in extreme cases, the noise from these miscalls can adversely affect the conclusions drawn from the whole experiment. If a base is misaligned or miscalled, then on average it will display a methylation rate of 50% because both cytosine and thymine are equally likely to be misplaced against a genomic cytosine. If the true methylation level is close to 0% or 100%, then a relatively small number of errors can disproportionally shift the predicted overall level of methylation.

Base-call qualities

In real data, the quality of base calls tends to fall as the length of the reads increases (**Supplementary Fig. 2a**). As base-call errors are random, the frequency for each base will tend toward 25% each at positions with high error rates. Another source of contamination that can lead to a change in base composition is the presence of (methylated) adaptor sequences, which we discuss below. Such deviations of the average nucleotide distribution toward later cycles of a library can usually be spotted in a base-composition analysis (**Supplementary Fig. 2b**). A tradeoff can be

Table 1 | Software packages for BS-seq analysis and their performance parameters

	Bismark	BRAT	BSMAP	BS-Seeker	MethylCoder	RMAP-BS	SOCS-B	B-SOLANA
Matching tool	Bowtie[44]	Reference hashing and wildcard matching	SOAP[48]	Bowtie[45]	Bowtie/GSNAP[44,49]	Wildcard/position-weight-matrix matching	Robin-Karp hashing	Bowtie[45]
Reference	40	43	50	39	51	52	42	44
Version	0.5.0	1.2.2	0.2.1	N/A	0.14.1	2.05	2.1	0.1.0
Language	Perl	C++	C++	Python	Python	C++	Perl	Python
Library type[a]	D and ND	D and ND[b]	D and ND	D and ND[c]	D and ND	D and ND[b]	D and ND	D
Sequencing technology	Base space	Base space	Base space	Base space	Base space	Base space	Color space	Color space
Sequencing mode	Single-end and paired-end	Single-end and paired-end	Single-end and paired-end	Single-end	Single-end and paired-end	Single-end	Single-end	Single-end
Best alignment criteria	Lowest number of non-BS[d] mismatches	Lowest number of non-BS mismatches	Lowest number of mismatches	Lowest number of mismatches	Lowest number of non-BS mismatches	Lowest number of mismatches	Lowest number of non-BS mismatches	Lowest number of mismatches
Output	Mapping output including methylation calls and extra tools	Mapping output and extra tools for methylation calls	Mapping output	Mapping output including methylation calls	Mapping output and methylation call output	Mapping output	Mapping output including methylation calls	Mapping output including methylation calls
Advantages	Unbiased mapping; performance[e]	Unbiased mapping; performance	–	Unbiased mapping; performance	Unbiased mapping; performance	–	Ignores bisulfite-induced color-space mismatches	Performance
Drawbacks	–	Inflexible parameters	Performance[e]; biased mapping	–	–	Unbiased mapping only in C-G context; biased mapping in non-CG context	Performance[e]	–

[a] D, directional library; ND, nondirectional library. [b] Requires two separate runs. [c] Requires presence of a tag sequence. [d] Non-BS, not bisulfite induced. [e] Performance here signifies run time on a reasonable time scale (that is, a few hours, as compared to several days or even weeks for the same technique with a human dataset).
–, not applicable.

made, in which longer reads increase coverage but also increase the number of incorrect methylation calls. Although it would be possible to weight a bisulfite methylation call based on the quality of the original base call, this is not currently done by any of the commonly used analysis tools and would only be of benefit for miscalled bases rather than misaligned reads.

To quantify the effect of miscalled bases, we simulated a 75–base-pair read dataset containing no methylation and added random miscalls at rates between 0.01% and 10% following an exponential decay model over the length of the sequence (**Supplementary Fig. 2c**). As the error rate increased, so did the apparent methylation level.

A way to counteract methylation miscalls or mismapping events as a consequence of base call errors in the reads, is to select strict alignment parameters. To demonstrate this, we simulated bisulfite reads with sequences carrying varying numbers of false base calls and aligned this dataset to the mouse reference genome using increasingly stringent cutoffs. Increasing the mapping stringency prevented sequences with several mismatches from aligning, thus reducing the number of erroneously inferred methylation states (**Supplementary Fig. 2d**) but at the cost of reduced mapping efficiency. A better way of decreasing methylation call errors from such poor quality data is to trim off low-quality base calls before read alignments are carried out. Such adaptive quality trimming can be performed with several publically available tools, such as cutadapt (http://code.google.com/p/cutadapt/), the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), PRINSEQ[46] (http://prinseq.sourceforge.net/), SolexaQA[47] (http://solexaqa.sourceforge.net/), Trimmomatic (http://www.usadellab.org/cms/index.php?page=trimmomatic) and others.

Sequencing into the adaptor

In many libraries, a proportion of reads will run through the insert and begin to sequence the adaptor on the 3′ end. Including adaptor sequence in a read will dramatically decrease the mapping efficiency of the read and will add a subset of random methylation calls.

We simulated the addition of varying lengths of Illumina adaptor sequence onto a BS-seq library containing no base call errors and measured the effect on both mapping efficiency and methylation calls (**Supplementary Fig. 3a**). The mapping efficiency decreased steadily with increasing adaptor contamination, but methylation errors were tightly linked to the sequence of the adaptor. Each addition of a cytosine in the adaptor caused a dramatic spike in the observed level of methylation (data not shown). Nondirectional libraries are even more susceptible to adaptor contamination as the introduction of guanine or adenine into reads aligning to the complementary bisulfite strands can introduce additional errors. Appropriate steps to identify and remove adaptor contamination, such as *k*-mer analysis (**Supplementary Fig. 3b**) with tools such as FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and adaptor-trimming software (for example, cutadapt, the FASTX toolkit, Trimmomatic, FAR (http://sourceforge.net/projects/theflexibleadap/) and others), should therefore always be taken before read alignments are carried out.

When we introduced both base call quality degradation and adaptor contamination into simulated BS-seq data reads, we observed a greatly reduced mapping efficiency compared to perfect genomic sequences (**Supplementary Fig. 3c**). When we universally trimmed the same sequences to shorter read lengths,

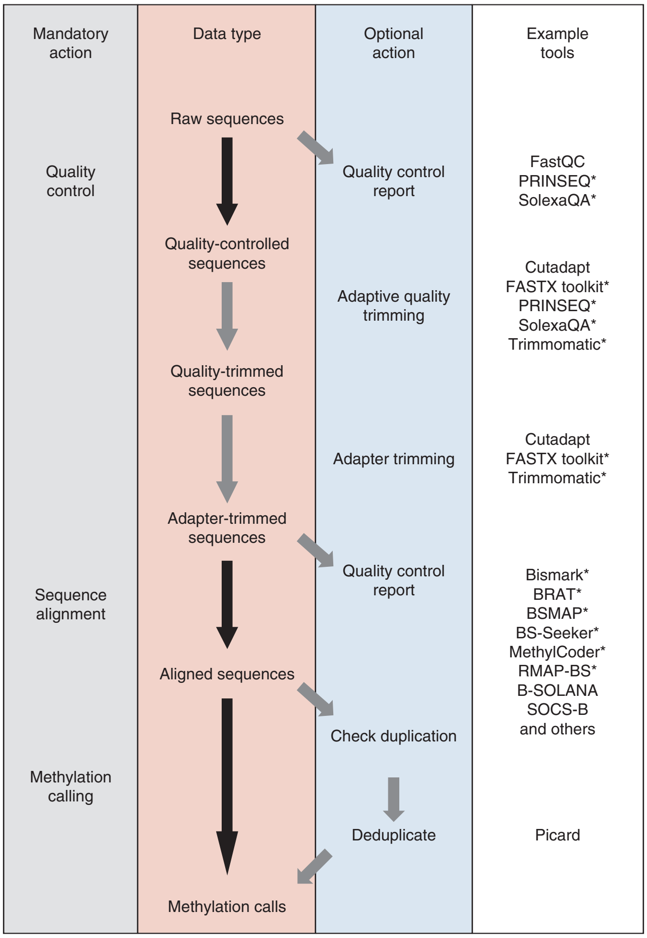

Figure 3 | Recommended workflow for the primary analysis of BS-seq data. Black arrows depict required steps, gray arrows indicate optional steps. *, only works with base-space data.

the mapping efficiency increased, reaching a maximum between 50 and 75 base pairs. This demonstrates that increasing the read length of bisulfite sequences does not necessarily translate into a linear increase in methylation information gained from an experiment. Similarly, paired-end reads do not automatically yield twice the amount of methylation data compared to single-end experiments because they result in a considerable amount of redundant methylation calls where both reads overlap.

Bisulfite conversion rate

In a BS-seq experiment we implicitly assume that all unmethylated cytosines are converted into thymines. However, this conversion may not run to completion. Incomplete conversion of unmethylated cytosines is indistinguishable from methylation and can thus introduce false positive methylation calls. In contrast, prolonged bisulfite treatment causes the sample to degrade in a way which enriches the small amount of remaining material for methylated reads.

Some studies have tried to avoid non-conversion errors by removing reads that exceeded an arbitrary threshold of methylation in a non-CpG context[7,30,37]; however, this procedure assumes very low methylation in a non-CpG context and hence introduces a potential bias against methylated reads. One option for estimating the bisulfite conversion rate is to use spike-in controls of nonnative DNA with a known methylation state. However, it should be kept in mind that such controls might not necessarily have the same conversion properties as the DNA sample to be analyzed.

End repair

It is crucial that the DNA methylation state of each fragment is not artificially modified before treatment with bisulfite because any amplification by a polymerase will erase any methylation marks that were present. In RRBS experiments, for instance, each fragment is generated by the digestion of a genome with a restriction endonuclease. The most commonly used enzyme for this type of library is MspI, which, upon cleavage, leaves a 5′ C-G overhang on the ends of each fragment[33]. To allow the addition of sequencing adapters, the overhangs are end-repaired using either methylated or unmethylated cytosines. These filled-in bases will align perfectly against the reference genome but will not maintain the original methylation state, and care must therefore be taken to exclude these bases from methylation calling. This problem only affects the 3′ end of reads when the read length is longer than the fragment to be sequenced (RRBS is probably more affected as fragments are usually size-selected to be 40 to 220 base pairs[24,33]). In such cases, reads should be screened for the occurrence of cytosine residues or a second MspI site just before reading into a potential adaptor contamination toward the 3′ end and trimmed back until the modified bases have been removed. In addition, paired-end or nondirectional RRBS libraries may also contain reads originating from filled-in MspI sites at the beginning of the reads, which consequently need to be excluded from downstream analysis[38].

Single-nucleotide variants

Any SNPs that are a cytosine in the reference genome but a thymine in the experimental sample would appear as consistent calls of unmethylated cytosines. Such errors would be impossible to detect from the quality of the reads because the base calls would be good and only the isolated nature of the effect seen might suggest that it is a technical rather than a biological effect. Both BS-seq alignments and methylation calls assume that the genomic reference sequence that reads are compared to is correct. Thus, if no SNP information is available, one has to expect a certain extent of systematic errors. These effects could be minimized by integrating available genomic-variation data, for example, from SNP databases into the reference sequence before bisulfite alignments are carried out or by using nucleotide information of the opposing genomic strand.

Conclusions

The primary analysis of BS-seq data should always start with a thorough assessment of the raw sequence data. Reads with low base call qualities or adaptor contamination should be identified and trimmed rigorously, even if this entails the risk of losing a few base pairs of real data, because the gain of confidence in correct alignments and methylation calls outweighs this minor data loss. Considering that mapping efficiencies of BS-seq and standard genomic reads converge quickly for read lengths greater than 40 base pairs (**Fig. 1a**), single-end reads of 50–75 base pairs seem to offer a reasonable compromise, providing good mapping capacity without running into problems associated with longer read lengths.

For base-space data, it is critical not to tolerate a high level of non-bisulfite mismatches (mismatches not induced by bisulfite treatment—that is, all mismatches other than (unmethylated) C-to-T mismatches) during the alignments because this allows reads to align to incorrect positions in the genome, resulting in false methylation calls. This can become especially relevant for reads originating from highly repetitive sites or from regions that are not yet part of the genome assembly. Stringent alignments are equally important for color-space data, but because color-space mapping approaches work differently, the strategy may have to be adapted to the individual needs of specific tools.

Many of the aspects of BS-seq discussed here, taken on their own, do not seem to have particularly drastic effects. Their combination, however, could easily lead to several million false methylation calls, which might have profound effects on the biological conclusions drawn from an experiment. Additional attention should be paid to reducing the number of artifacts that can only be spotted after the alignments have been performed. Duplicate reads or regions displaying abnormally high read coverage should be excluded from the analysis because they can comprise a sizeable proportion of the experiment and thereby introduce considerable bias.

Given one has a choice before starting an experiment, it is currently most convenient to opt for a platform generating base-space data because it can measure methylation over a wide dynamic range of methylation levels in any cytosine context with equal efficiencies, and most available tools are tailored to this kind of data. For small genomes or genomes fulfilling certain criteria regarding their methylation state, there are now also good tools available to handle color-space reads.

The best practices recommended here (**Fig. 3**) apply to data generated on current sequencing platforms. It will be interesting to see whether forthcoming single-molecule sequencing technologies will be able to live up to their promise to revolutionize the way in which methylation is measured.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS
This work was funded by the Biotechnology and Biological Sciences Research Council, UK. A.F. and B.K. received infrastructure support from the Deutsche Forschungsgemeinschaft Excellence Cluster 'Inflammation at Interfaces'.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Law, J.A. & Jacobsen, S.E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
2. Pelizzola, M. & Ecker, J.R. The DNA methylome. *FEBS Lett.* **585**, 1994–2000 (2010).
3. Robertson, K.D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).
4. Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat. Genet.* **41**, 1350–1353 (2009).
5. Esteller, M. Epigenetics in cancer. *N. Engl. J. Med.* **358**, 1148–1159 (2008).
6. 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).
7. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009). **This was the first human methylome analyzed at single-base resolution using whole-genome bisulfite next-generation sequencing.**
8. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).
9. Bird, A.P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
10. Coulondre, C., Miller, J.H., Farabaugh, P.J. & Gilbert, W. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775–780 (1978).
11. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).
12. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
14. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
15. Illingworth, R.S. *et al.* Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134 (2010).
16. Lister, R. & Ecker, J.R. Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.* **19**, 959–966 (2009).
17. Laird, P.W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
18. Down, T.A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* **26**, 779–785 (2008).
19. Jacinto, F.V., Ballestar, E. & Esteller, M. Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques* **44**, 35–39 (2008).
20. Serre, D., Lee, B.H. & Ting, A.H. MBD-isolated Genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res.* **38**, 391–399 (2010).
21. Li, N. *et al.* Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods* **52**, 203–212 (2010).
22. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* **89**, 1827–1831 (1992).
23. Bock, C. *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* **28**, 1106–1114 (2010).
24. Harris, R.A. *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.* **28**, 1097–1105 (2010). **A detailed comparison of different sequencing-based technologies to analyze DNA methylation genome-wide.**
25. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* **5**, e8888 (2010).
26. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398–402 (2011).
27. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394–397 (2011).
28. Song, C.X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.* **29**, 68–72 (2011).
29. Li, Y. *et al.* The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* **8**, e1000533 (2010).
30. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008). **This study reported the first genome-wide DNA methylation in mouse cells generated by RRBS.**
31. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. USA* **107**, 8689–8694 (2010).
32. Popp, C. *et al.* Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101–1105 (2010).
33. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **6**, 468–481 (2011).
34. Gu, H. *et al.* Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods* **7**, 133–136 (2010).

35. Smith, Z.D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods* **48**, 226–232 (2009).
36. Song, F. *et al.* Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. USA* **102**, 3336–3341 (2005).
37. Cokus, S.J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
This study reported a methylome of *Arabidopsis thaliana* at single-base resolution generated via a nondirectional bisulfite sequencing library.
38. Smallwood, S.A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat Genet.* **43**, 811–814 (2011).
39. Chen, P.Y., Cokus, S.J. & Pellegrini, M.B.S. Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics* **11**, 203 (2010).
40. Krueger, F. & Andrews, S.R. Bismark: A flexible aligner and methylation caller for Bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
41. McKernan, K.J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
42. Ondov, B.D. *et al.* An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System. *Bioinformatics* **26**, 1901–1902 (2010).
43. Harris, E.Y., Ponts, N., Levchuk, A., Roch, K.L. & Lonardi, S. BRAT: bisulfite-treated reads analysis tool. *Bioinformatics* **26**, 572–573 (2010).
44. Kreck, B. *et al.* B-SOLANA: An approach for the analysis of two-base encoding bisulfite sequencing data. *Bioinformatics* published online, doi:10.1093/bioinformatics/btr660 (6 December 2011).
45. 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).
46. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
47. Cox, M.P., Peterson, D.A., Biggs, P.J. & Solexa, Q.A. At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**, 485 (2010).
48. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
49. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
50. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
51. Pedersen, B., Hsieh, T.F., Ibarra, C. & Fischer, R.L. MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics* **27**, 2435–2436 (2011).
52. Smith, A.D. *et al.* Updates to the RMAP short-read mapping software. *Bioinformatics* **25**, 2841–2842 (2009).