Published online 3 January 2014

Nucleic Acids Research, 2014, Vol. 42, No. 6 e43
doi:10.1093/nar/gkt1325

Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing

Govindarajan Kunde-Ramamoorthy[1], Cristian Coarfa[2], Eleonora Laritsky[1], Noah J. Kessler[3], R. Alan Harris[4], Mingchu Xu[4], Rui Chen[4], Lanlan Shen[1], Aleksandar Milosavljevic[4] and Robert A. Waterland[1,4,*]

[1]Department of Pediatrics, Baylor College of Medicine, USDA/ARS Children's Nutrition Research Center, Houston, TX 77030, USA, [2]Department of Molecular & Cell Biology, Baylor College of Medicine, Houston, TX 77030, USA, [3]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA and [4]Department of Molecular & Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

Received June 3, 2013; Revised November 4, 2013; Accepted November 29, 2013

ABSTRACT

Coupling bisulfite conversion with next-generation sequencing (Bisulfite-seq) enables genome-wide measurement of DNA methylation, but poses unique challenges for mapping. However, despite a proliferation of Bisulfite-seq mapping tools, no systematic comparison of their genomic coverage and quantitative accuracy has been reported. We sequenced bisulfite-converted DNA from two tissues from each of two healthy human adults and systematically compared five widely used Bisulfite-seq mapping algorithms: Bismark, BSMAP, Pash, BatMeth and BS Seeker. We evaluated their computational speed and genomic coverage and verified their percentage methylation estimates. With the exception of BatMeth, all mappers covered >70% of CpG sites genome-wide and yielded highly concordant estimates of percentage methylation ($r^2 \geq 0.95$). Fourfold variation in mapping time was found between BSMAP (fastest) and Pash (slowest). In each library, 8–12% of genomic regions covered by Bismark and Pash were not covered by BSMAP. An experiment using simulated reads confirmed that Pash has an exceptional ability to uniquely map reads in genomic regions of structural variation. Independent verification by bisulfite pyrosequencing generally confirmed the percentage methylation estimates by the mappers. Of these algorithms, Bismark provides an attractive combination of processing speed, genomic coverage and quantitative accuracy, whereas Pash offers considerably higher genomic coverage.

INTRODUCTION

DNA methylation, which occurs predominantly at cytosines within CpG dinucleotides in the mammalian genome, is an epigenetic mark fundamental to developmental processes including genomic imprinting, silencing of transposable elements and differentiation (1). Coupling bisulfite modification (2) with next-generation sequencing (Bisulfite-seq) provides information about cytosine methylation genome-wide at single-base resolution (3–5). Bisulfite modification deaminates unmethylated cytosines (i.e. most cytosines) to uracil, and these are subsequently converted to thymine during polymerase chain reaction amplification. The consequent reduced sequence complexity makes it challenging to map Bisulfite-seq reads to the reference genome using standard short read alignment tools (6). Additionally, the advent of Bisulfite-seq forces the question of what is the optimal resolution at which to study the methylome. Regional methylation changes encompassing several CpG sites may be more biologically meaningful than those occurring only at individual CpGs; further, it may often be impractical to perform analysis and validation at the level of individual CpG sites.

Several approaches have been developed to map Bisulfite-seq reads (6), including 'wild card' and 'three letter' aligning. There are two variations of the wild card approach; the first allows either Cs or Ts in reads to map to Cs in the reference genome (7). The second enumerates all C to T combinations for each k seed-length and then aligns by hashing and extension (8,9). In the three-letter approach, all Cs in both the reference genome and reads are converted to Ts, and mapping is performed using a seed and extend approach (10–12). Both strategies can use either gapped or ungapped alignment, depending on the underlying short read alignment tool. The gapped alignment method handles substitutions and small indels efficiently (13).

*To whom correspondence should be addressed. Tel: +1 713 798 0304; Fax: +1 713 798 7101; Email: waterland@bcm.edu

Bisulfite-seq mapping algorithms are mainly used to estimate percentage methylation at specific CpG sites (methylation calls), but also provide the ability to call single nucleotide and small indel variants (13) and copy number and structural variants (14). The analyses in this article focus exclusively on issues relevant to methylation calls; clearly, an algorithm's ability to map reads in various sequence contexts and make accurate methylation calls may have profound implications for the interpretation of Bisulfite-seq experiments. Descriptions of new Bisulfite-seq mapping tools typically compare global metrics such as proportion of uniquely mapped reads, global percentage methylation, computational requirements and running time. Benchmarking studies have been performed using real data downloaded from public databases [mainly human (3) or plant data (15)], simulated data (16) or combinations of both (10,12,17,18). However, none of these previous studies performed independent quantitative verification of methylation calls. The only previous comparison of Bisulfite-seq mapping algorithms using independently generated sequencing data involved a single reduced representation bisulfite sequencing data set obtained from one human sample (19). That study evaluated the mapping efficiency of Bismark, BSMAP and RMAPBS (7) as a function of read length and adaptor sequences. They also compared the total number of methylated CpG sites genome-wide, but did not compare overlap of CpG sites covered by the three mappers, or assess accuracy of methylation calls.

Hence, although proper analysis and interpretation of the increasing number of expensive Bisulfite-seq data sets critically depends on their performance, one may conclude that widely used mapping methods remain poorly characterized. To address this need, we selected three mapping algorithms for detailed comparison: Bismark (11), BSMAP (9) and Pash (8), which use Bowtie (20), SOAP (21) and in-house aligners, respectively. Bismark and BSMAP are the most widely used three-letter and wild card mapping algorithms, respectively (6). Pash performs a heuristic alignment of k-mer matches. To complement this main comparison, we also evaluated the performance of two additional mapping algorithms: BS Seeker (10) and BatMeth (17). We generated four human methylomes (representing two tissues from each of two individuals) and mapped the Bisulfite-seq reads independently using the five algorithms. Despite generally excellent concordance of the mapping results, our analysis (and subsequent independent verification by quantitative bisulfite pyrosequencing) highlights important differences among the mapping algorithms.

MATERIALS AND METHODS

Sample collection, Bisulfite-seq library preparation and sequencing

Two tissue samples, peripheral blood lymphocyte (PBL) and hair follicle (HF), from two healthy male adults (C01 and C02) were collected in accordance with institutional IRB regulations. HFs (30–50) were obtained by plucking scalp or facial hair from the same individuals. PBLs were isolated by Ficoll gradient centrifugation. Tissues were stored at −80°C until isolation of genomic DNA by proteinase-k digestion and phenol–chloroform extraction (22).

Illumina libraries were generated according to the manufacturer's sample preparation protocol for genomic DNA. Approximately 1 μg of genomic DNA was fragmented to 200–500 bp and end-repaired. The 5′-ends of DNA fragments were phosphorylated, and a single adenine base was added to the 3′-end. Illumina adaptors were ligated to the genomic DNA.

Bisulfite modification was performed using the EZ DNA Methylation-Direct kit (Zymo Research) according to the manufacturer's instructions. The bisulfite-modified DNA was amplified by using adaptor-specific primers, and fragments of 200–500 bp were isolated by bead purification. The quantity and size distribution of sequencing libraries were determined using the Pico Green fluorescence assay and the Agilent 2100 Bioanalyzer, respectively. The DNA was sequenced on the Illumina HiSeq 2000 as 100-bp paired-end reads, following the manufacturer's protocols.

Reads quality control and mapping

For each library, the standard Illumina pipeline was used to perform base calling, and the results were generated in fastq format. Quality control of reads was accessed by running the FastQC program. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) helps to determine the best quality score and the right read length for trimming as the signals decline while Illumina cycles progress. Because the majority of the bases had quality scores $\geq$28, we decided to use a quality score filtering value $\geq$28 and read length $\geq$50 bp. The Cutadapt (23) program was used to trim adaptor sequences and perform quality score and read length filtering. This resulted in >85% of the reads for mapping and analysis.

QC-passed reads were mapped to the University of California, Santa Cruz (UCSC; http://genome.ucsc.edu) hg19 genome build using Pash 3.0, Bismark 0.7.4 (Bowtie 1 mode), BSMAP 2.6, BatMeth 1.04 and BS Seeker2. For all the mapping algorithms, we used default parameters as recommended by the authors except the mismatch parameter that was set to 7, as the libraries had longer read length. The uniquely mapped reads from each mapper were further processed using their respective post-processing scripts provided to estimate the percentage coverage and percentage methylation for CpG sites genome-wide. All the mapping was performed using a compute cluster with 36 nodes, each node containing 8 Intel Xeon E5540 CPUs and 24-GB RAM. The command lines used for all the mappers are provided in Supplementary Methods.

Data analysis

Identification of genome-wide CpG site coverage and percentage methylation

To identify the total number of CpG sites and their coordinates genome-wide (UCSC hg19), an in-house Perl program was used, which identified 28.2 million

CpG sites. Each CpG site was counted toward coverage if the read depth was ≥10, as the overall coverage of each library was ~26×. The total coverage and overlap of CpG sites between different mappers were calculated using an in-house Perl program. The percentage methylation scatter plots and Pearson correlations (r^2) were computed using the R package. Processed data with total number of reads and methylated reads for individual CpG sites are available in GEO (GSE44806).

Identification of 200-bp bins with two CpG sites coverage and percentage methylation

To logically identify an appropriate 'bin' size to interrogate DNA methylation, we divided the genome into five different bin sizes from 100 to 500 bp with an increment of 100 bp and computed the number of CpG sites in each bin and also the percentage of CpG site coverage genome-wide. This resulted in 6.2 million bins (200-bp bins with at least two CpG sites), which covered 85% of the total CpG sites and eliminated 60% of the reads, as these bins were present only in 40% of the genome. All bins containing <4 CpG sites were considered covered if at least 2 sites were covered by ≥10 reads, and those containing ≥4 CpG sites were considered covered if at least half of them were covered by ≥10 reads. The overlap of bins not covered by BSMAP or Bismark across four libraries (4-way Venn diagram) was generated using the 'VennDiagram' package (24) available in R. To test the significance of correlation between regions covered by all mappers and not covered by others, the absolute residuals of percentage methylation of individuals were compared using two-tailed t tests in R.

Characterization of genomic regions

To characterize the genomic features of regions covered by all mappers and not covered by others, we used DGV Struct Var, Segmental Dups and RepeatMasker tracks from the UCSC Genome Browser annotation database. The percentage overlap and the extent of overlap of bins with various genome features were computed using a combination of BEDtools (25) and in-house Perl scripts. The percentage nucleotide compositions for each library were computed using the Bioperl library (http://www.bioperl.org/wiki/Main_Page), in-house Perl scripts and R software. To test the significance of percentage nucleotide composition between regions covered by all mappers and other categories, for each library we computed the average percentage nucleotide composition and performed two-tailed paired *t*-tests ($n = 4$ libraries) with equal variance in R.

Bisulfite-seq read simulation and analysis

We simulated 10 million reads from hg19 using the software RMAP-bs (17). Reads were generated with length = 100 bp and allowing maximum three mismatches with bisulfite conversion efficiency of 99%. The 10 million reads were mapped to the hg19 build using Bismark, BSMAP and Pash mapping algorithms, allowing seven mismatches.

Quantitative verification of DNA methylation

To verify the accuracy of percentage methylation estimates of bins (200 bp) not covered by BSMAP, we designed 18 pyrosequencing assays (Supplementary Table S8) and performed site-specific analysis of CpG methylation. Bisulfite modification and pyrosequencing of the regions were performed as previously described (22). All pyrosequencing assays were first validated for quantitative accuracy by running methylation standards composed of known mixtures of completely methylated and unmethylated human genomic DNA (26).

Table 1. Comparison of mapping and post-processing times (in seconds) for 1 million reads

Algorithm	Mapping	Post-processing	Total
Bismark	1514	81	1595
BSMAP	800	1081	1881
Pash	3486	1504	4990
BS Seeker	1324	3867	5191
BatMeth	904	70	974

Post-processing times include time required to estimate percentage methylation at each methylated cytosine, but do not include time to load the reference genome into memory (required for BSMAP and Pash only).

RESULTS

CpG site level coverage and concordance of the mappers

Bisulfite-seq data sets were generated for PBL and HF DNA from each of two healthy males. We chose these two tissues because they represent two different germ layer lineages (mesoderm and ectoderm, respectively). We generated an average of 400 million 100-bp paired-end reads for each library, achieving 26 × average coverage per library; 95% of the reads were retained after adaptor trimming, quality score filtering (≥28) and read length filtering (≥50 bp) (27). Filtered reads were mapped to the human reference genome UCSC hg19 build. We mapped the reads as single-end reads, for greatest generalizability. We focused our main analysis on a comparison of Bismark, BSMAP and Pash. For each library, Bismark, BSMAP and Pash uniquely mapped 77–82%, 78–83% and 81–87% of the reads, respectively (Supplementary Table S1). Analysis of one library (Supplementary Table S2) indicated that >96% of reads were mapped by at least one mapper. A benchmarking analysis of mapping and post-processing 1 million reads on a single 8-core processor (Table 1) indicated that Bismark and BSMAP are substantially faster than Pash. All the algorithms include post-processing scripts to calculate coverage and percentage methylation at the CpG site level.

In each of the four libraries, each of these mapping algorithms covered >70% of the 28.2 million CpG sites genome-wide with a read depth of ≥10 (Figure 1A and Supplementary Figure S1). More than 67% of all CpG sites genome-wide were covered by all three mappers in each library (Figure 1A and Supplementary Figure S1). Overall, the three mappers exhibited excellent concordance in CpG site-specific methylation calls ($r^2 \geq 0.95$) (Figure 1B–D). Contrary to the conjecture of Chatterjee

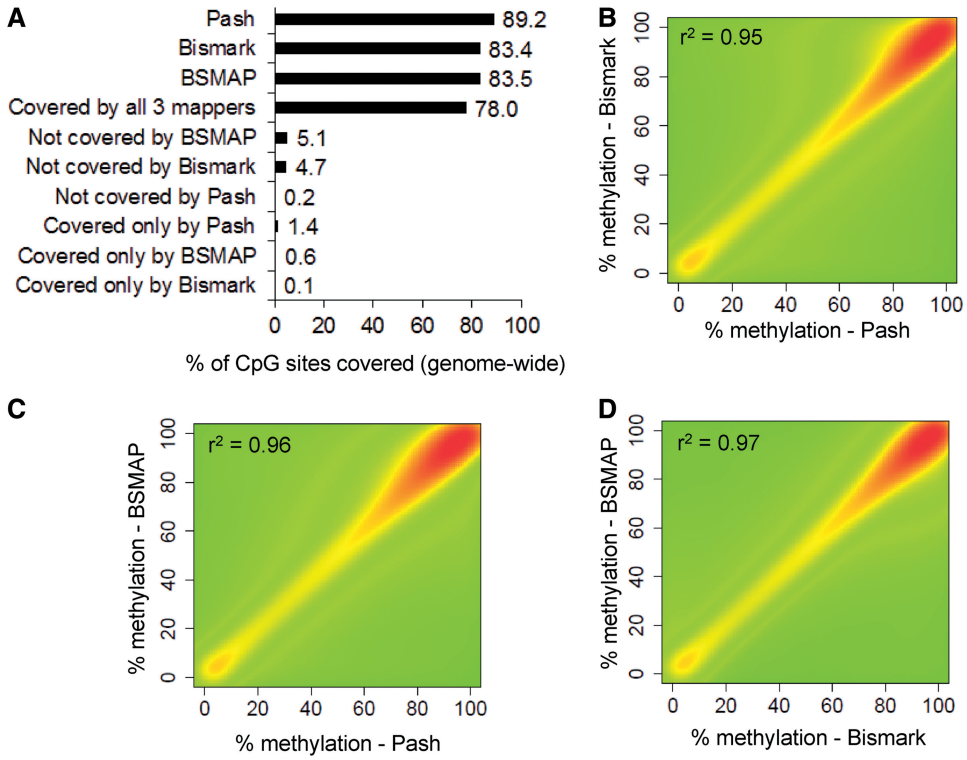

Figure 1. All three mappers provide excellent coverage and highly concordant estimates of CpG methylation genome-wide. (**A**) Percentage of CpG sites covered by Pash, Bismark and BSMAP, and the overlaps among them. Each mapping algorithm covers >80% of the CpG sites, and 78% are covered by all the three mapping algorithms. 'Not covered by BSMAP', for example, indicates the percentage of CpG sites that are covered by Pash and Bismark but not by BSMAP. Correlations of CpG site-specific percentage methylation calls among the different mapping algorithms are high: (**B**) Bismark versus Pash ($r^2 = 0.95$), (**C**) BSMAP versus Pash ($r^2 = 0.96$) and (**D**) BSMAP versus Bismark ($r^2 = 0.97$). Red, yellow and green indicate high, moderate and low densities, respectively. All data are for C01-HF library only, as an example.

et al. (19) that 'wild card' mappers may be biased toward highly methylated reads, average genome-wide DNA methylation estimates did not differ appreciably among the mappers (e.g. C01-HF methylation was 78.3%, 78.7% and 76.9% by Bismark, BSMAP and Pash, respectively).

We also mapped the reads from the four Bisulfite-seq libraries using BatMeth and BS Seeker. Although BatMeth was fast (Table 1), it uniquely mapped only ~50% of the reads in each library (Supplementary Table S1), which is comparable with the developers' results mapping Bisulfite-seq reads from an H1 cell line (17). Owing to its low mapping efficiency, we excluded BatMeth from further consideration. BS Seeker, which yielded average mapping speed but a long post-processing time (Table 1), uniquely mapped ~80% of the reads from each library (Supplementary Table S1), similar to the other three mapping algorithms. Compared with Bismark, BSMAP and Pash, BS Seeker mapping results were most concordant with those of Bismark. Of all the reads in multiple libraries mapped by BS Seeker, >98% were mapped to the same position by Bismark (Supplementary Table S3). Further, BS Seeker percentage methylation calls at individual CpG sites genome-wide were highly correlated with those of Bismark ($r^2 = 0.94$) (Supplementary Figure S2). Given that BS Seeker's mapping results are highly concordant with those of Bismark, for clarity and simplicity, we will focus subsequent analyses on Bismark, BSMAP and Pash.

Coverage of 200-bp bins containing at least two CpG sites

With the goal of drawing the most biologically meaningful comparisons among the mapping results, and to provide a basis for verification of DNA methylation estimates, we sought to identify a logical and appropriate approach to efficiently collapse the site-specific data into genomic regions while still maintaining sufficient resolution to discriminate between genomic regions with different methylation patterns. Focusing on 200-bp 'bins' containing at least two CpG sites covers 85% of CpG sites in the human genome (Supplementary Figure S3A), while eliminating from consideration the 60% of reads that map to CpG-poor regions (Supplementary Figure S3B) (with commensurate reduction in downstream computational requirements). In addition to efficiently covering the majority of CpG sites in the genome, 200-bp bins are biologically attractive, as they approximate nucleosomal resolution. In total, we identified 6.2 million 200-bp bins containing at least two CpG sites in the hg19 UCSC genome build (henceforth referred to as 'bins'). All bins containing <4 CpG sites were considered covered if at least two sites were covered by at least 10 reads, and those containing >4 CpG sites were considered covered if at least half of the CpG sites were covered by at least

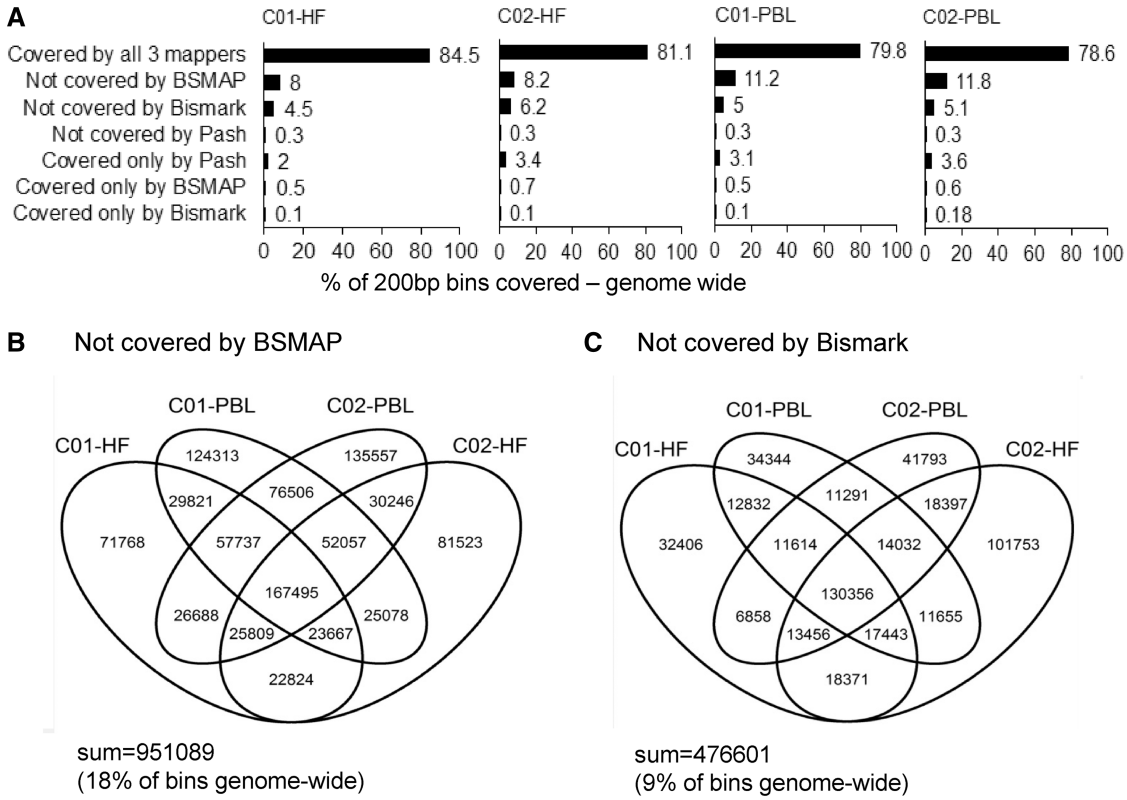

Figure 2. Genome-wide coverage of 200-bp bins containing ≥2 CpG sites by different mapping algorithms. (**A**) Percentage of bins covered by all mappers and not covered by individual mappers across all four libraries (C01-HF, C01-PBL, C02-HF and C02-PBL). More than 78% of the bins are covered by all three mappers in each library. (**B**) Comparing methylation across four libraries requires that each bin be covered in all four libraries. Fully 18% of bins are not covered by BSMAP in at least one library, and (**C**) 9% of bins are not covered by Bismark in at least one library.

10 reads. In each of the four libraries, >78% of bins were covered by all three mapping algorithms (Figure 2A). In all, 8–12% of bins were not covered by BSMAP, 5–6% were not covered by Bismark and <1% of the bins were not covered by Pash. Hence, although BSMAP mapped more reads than Bismark (Supplementary Table S1), more of the reads mapped by Bismark were within genomic regions containing CpG sites. Although failing to cover 5–10% of bins may not seem like a huge loss, we asked to what extent such losses would be compounded when performing comparisons across different libraries. Among the four libraries in our experiment, >18% of the bins were not covered by BSMAP, and 9% of bins were not covered by Bismark in at least one library (Figure 2B and C). Such losses will, of course, increase with the number of samples under comparison. Hence, the choice of mapping algorithm can have a substantial impact on the results of a comparative methylome analysis.

Choice of the mapping algorithm affects ability to detect interindividual and tissue-specific methylation differences

Bins that are covered by all mappers (Figure 3A) and those not covered by Bismark (Figure 3B) exhibited a high correlation ($r^2 > 0.9$) of average percentage methylation between individuals, whereas bins not covered by BSMAP (Figure 3C) showed a lower correlation ($r^2 = 0.84$, $P < 10^{-10}$ compared with those covered by all mappers). Bins covered by all mappers (Figure 3D) showed substantial tissue-specific variation in methylation calls between HF and PBL ($r^2 = 0.43$). Bins not covered by Bismark (Figure 3E) showed a slightly but significantly higher correlation between HF and PBL ($r^2 = 0.53$, $P < 10^{-10}$), indicating that these tend not to be regions of tissue-specific variation in DNA methylation. Remarkably, among bins not covered by BSMAP (Figure 3F), there was a significantly lower inter-tissue correlation ($r^2 = 0.27$, $P < 10^{-10}$), indicating that regions in which BSMAP fails to map also tend to be regions of tissue-specific variation. Hence, at least in the two individuals and two tissues we compared, regions that were mapped by Bismark and Pash, but not by BSMAP, were enriched for both interindividual and tissue-specific variation in DNA methylation. This suggests that the sequence characteristics that render certain genomic regions difficult for BSMAP to map are also associated with mechanisms of interindividual and tissue-specific epigenetic regulation.

Characterization of genomic regions differentially covered by the mappers

We sought to identify genomic features that potentially explain the differential mapping characteristics of the three algorithms. Relative to bins covered by all three mappers, those not covered by BSMAP were found to

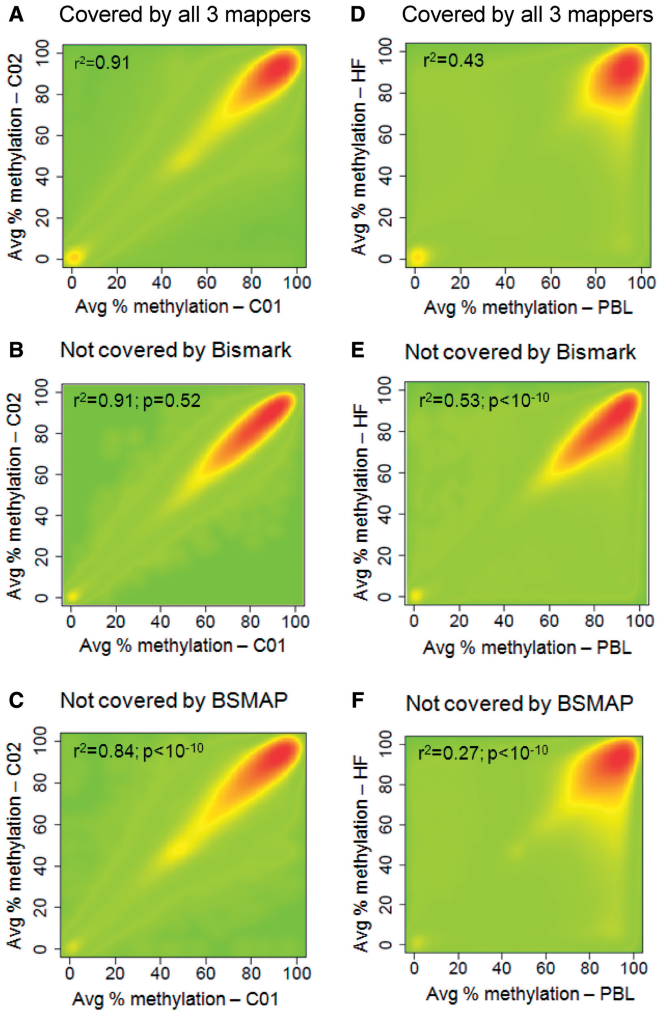

Figure 3. Evaluation of interindividual and tissue-specific variation of percentage methylation according to different mapping algorithms. (**A–C**) Correlation of percentage methylation across individuals, according to mapping category. (A) Average percentage methylation (per bin, across all mappers) is highly concordant in individual 2 (C02) versus individual 1 (C01) ($r^2 = 0.91$). (B) For bins not covered by Bismark, interindividual correlation ($r^2 = 0.91$) is comparable with that across all mappers ($P = 0.52$). [Note: Statistical significance is indicating whether the correlation shown is different from that in (A).] (C) For bins not covered by BSMAP, interindividual correlation is reduced [$r^2 = 0.84$; significantly lower than in bins covered by all mappers ($P < 10^{-10}$)]. (**D–F**) Correlation of percentage methylation between tissues (PBL versus HF) according to different mapping algorithms. (D) Bins covered by all mappers show substantial tissue-specific variation ($r^2 = 0.43$). (E) For bins not covered by Bismark, inter-tissue correlation is significantly higher ($r^2 = 0.53$, $P < 10^{-10}$ relative to those covered by all mappers) [Statistical significance is indicating whether the correlation shown is different from that in (D)]. (F) For bins not covered by BSMAP, inter-tissue correlation is significantly lower ($r^2 = 0.27$, $P < 10^{-10}$ relative to those covered by all mappers).

be similar in terms of proportion overlapping with structural variations, segmental duplications and repetitive elements (Figure 4A). Conversely, bins covered by Pash and BSMAP, and those covered only by Pash, were highly enriched for structural variations and segmental duplications (Figure 4A). For bins that do overlap with these features, the extent of overlap is essentially 100% (Supplementary Figure S4). Although bins covered differentially by the mappers showed a similar high prevalence of repetitive elements (Figure 4A), SINE elements were enriched in bins not covered by Bismark and depleted in bins not covered by BSMAP (Supplementary Figure S5). We next evaluated nucleotide composition of the bins not covered by different mappers (complete details in Supplementary Table S4). Bins covered by all three mappers showed equal percentage composition (~25% each) of A, T, G and C (Figure 4B), comparable with that in bins not covered by Bismark (Figure 4C). However, bins not covered by BSMAP were highly enriched in T ($P = 1.65 \times 10^{-5}$) and depleted in G ($P = 7.9 \times 10^{-6}$; Figure 4D); a similar but less dramatic pattern was found in regions covered only by Pash (Figure 4E). Together, these data indicate that regions called as uniquely mapped by BSMAP and Pash, but not by Bismark, are largely associated with structural variations and segmental duplications. Regions mapped by Bismark and Pash, but not by BSMAP, on the other hand, are characterized most strikingly by low G content.

Mapping simulated reads confirms unique ability of Pash

Our analysis (Figure 4A) suggests that BSMAP and Pash can uniquely map reads within regions of structural variation and segmental duplication that are not covered by Bismark. Because we performed mapping at relatively low stringency (allowing up to seven mismatches), an alternative explanation is that some of the 'unique' mappings by BSMAP and Pash are incorrect. Of all reads mapped, >95% included only two or fewer mismatches (Supplementary Table S5), suggesting that our low stringency did not have a dramatic effect on mapping accuracy. To test this directly, we performed a mapping experiment using simulated reads. We simulated 10 million 100-bp reads from hg19, and mapped them twice by each of Bismark, BSMAP and Pash, allowing up to three or seven mismatches, respectively. The number of mismatches allowed had essentially no effect on mapping efficiency (Supplementary Table S6), so we focused on the results based on up to seven mismatches (the same stringency we used in mapping the libraries). Pash mapped fewer of the simulated reads overall (7.8 M versus 9.3 M for the other two mappers) but nearly the same number within the 200-bp bins (i.e. regions containing most of the CpG sites) (Supplementary Table S7). More than 99.5% of the mappings by Bismark and BSMAP were accurate, compared with ~95% for Pash (Supplementary Table S7). Nonetheless, when we characterized the genomic features of the regions mapped differentially by the three algorithms, we obtained results strikingly similar to those in Figure 4A. Bins covered only by Pash were twice as likely to overlap with structural variation, and 20 times as likely to overlap with segmental duplication in particular, compared with bins covered by all three mappers (Supplementary Figure S6). However, the simulation experiment did not confirm a special ability of BSMAP to map in these regions (compare 'not covered by Bismark' in Figure 4A and Supplementary Figure S6).

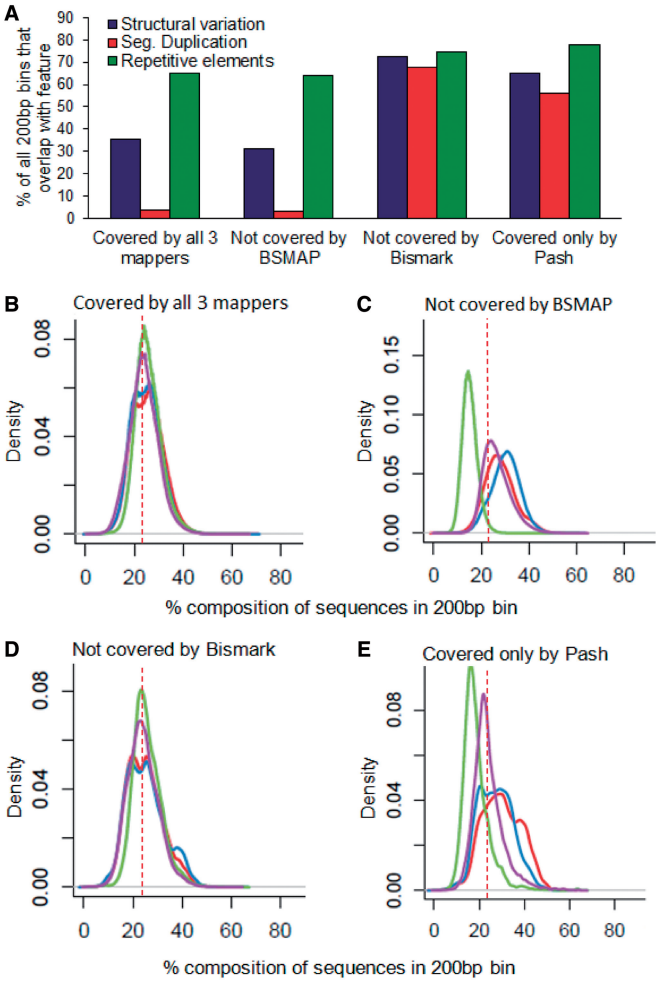

Figure 4. Characterization of genomic regions differently covered by the three mapping algorithms (all four libraries combined). (**A**) Percentage of covered 200-bp bins overlapping with different genomic features, by mapper category. Compared with regions covered by all three mappers, those not covered by Bismark and covered only by Pash are highly enriched for overlap with segmental duplications and structural variations. In regions covered by all mappers (**B**) and in those not covered by Bismark (**C**), percentage nucleotide compositions are all equal. (A: red, T: blue, G: green, C: purple.) (**D**) Regions not covered by BSMAP are enriched for 'T' and depleted of 'G' nucleotides. (**E**) Regions covered only by Pash have an under-representation of 'G' nucleotides.

Quantitative verification of regions by pyrosequencing

In bins that are covered by some mappers but not by others, there are two potential scenarios. The methylation calls may be reliable, or some mappers may map more promiscuously than others, providing erroneous methylation calls. Without independent quantitation of DNA methylation in the sequenced samples, it is impossible to distinguish between these two possibilities. Therefore, we verified regional DNA methylation estimates by quantitative bisulfite pyrosequencing. Because most of the regions not mapped by Bismark or Pash overlap with structural variations and segmental duplications (Figure 4A), we were unable to design pyrosequencing assays for these regions. Therefore, our verification focused on regions not covered by BSMAP. We designed 18 pyrosequencing assays by selecting among bins showing low, medium and high levels of methylation, as well as those showing tissue-specific variation. All pyrosequencing assays were first validated for quantitative accuracy by running methylation standards composed of known mixtures of completely methylated and unmethylated human genomic DNA (26,28). Of 18 assays designed, 4 were found to be unreliable. Among the remaining 14, the percentage methylation estimates by Bismark and Pash agreed remarkably well with the pyrosequencing data (Figure 5), with just a few regions exhibiting modestly overestimated (Figure 5D, E and K) or underestimated methylation calls (Figure 5F, G, and J). These data clearly indicate that in the bins that were not mapped by BSMAP, Bismark and Pash were able to map correctly and provide reliable estimates of DNA methylation.

DISCUSSION

This is the first study comparing several Bisulfite-seq mapping algorithms on a large data set across multiple sequencing libraries with biologically meaningful sample variation. Our goal was not to perform a comprehensive comparison of the many published mapping algorithms, but rather to determine the extent to which mapping characteristics of several commonly used algorithms may affect experimental outcomes. Previous studies reporting performance of Bisulfite-seq mapping algorithms (8–11,16,17) used only a small number of real public data (2–15 million reads) or simulated data (1 million reads) to compare mapping efficiency. The only other mapper comparison study using independently generated sequence data (19) was limited to reduced representation bisulfite sequencing and focused mainly on the performance of mapping efficiency as a function of read length and coverage. Those authors generated independent data, but did not verify the percentage methylation levels quantitatively. They observed that all the aligners covered >80% of the CpG sites within the reduced representation (RR) genome, consistent with our findings genome-wide.

Although Bisulfite-seq offers single CpG resolution, most whole-genome Bisulfite-seq studies analyze the data on a lower level of resolution, using, e.g a 1-kb sliding window with a 100-bp step (3) or a 2- or 5-kb tiling window (5,29). However, no previous studies have attempted to identify an optimal resolution based on an integrated analysis of CpG density and genomic content across various bin sizes. The approach we propose here, selecting 200-bp bins with at least two CpG sites, combines excellent genomic coverage (85% of CpG sites genome-wide) and high resolution, while reducing downstream computational requirements by eliminating from consideration 60% of the reads mapping into CpG-poor regions (Supplementary Figure S3). This approach, if adopted broadly, could simplify direct comparisons among various Bisulfite-seq studies. Moreover, these highly informative regions could also provide a basis for Bisulfite-seq-targeted enrichment reagents (30), substantially decreasing sequencing requirements.

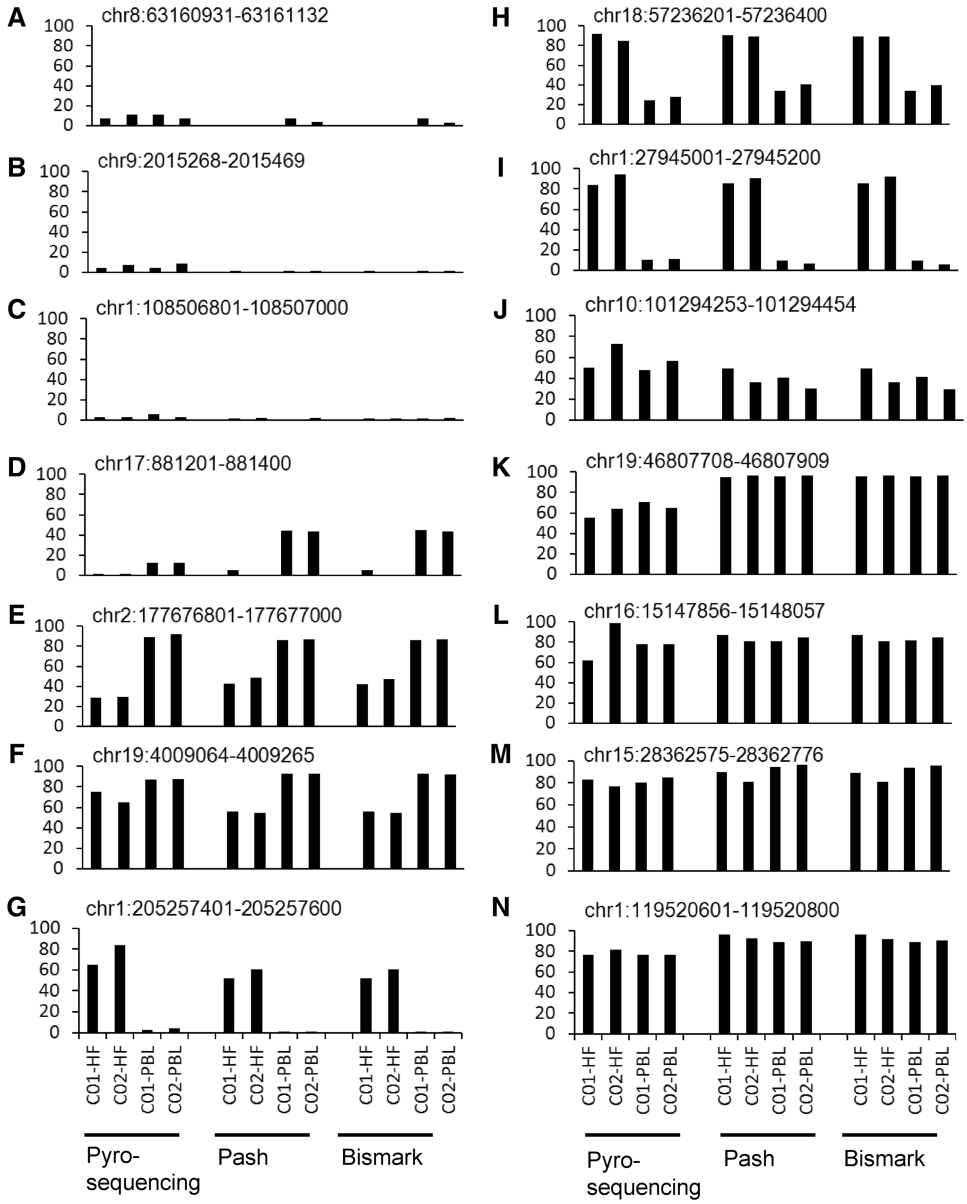

Figure 5. Verification of percentage methylation by quantitative bisulfite pyrosequencing in bins not covered by BSMAP. (**A–C**) Regions in which Bismark and Pash found low percentage methylation in all four libraries. (**D–F**) Regions in which Bismark and Pash found tissue-specific variation (i.e. low in HF and higher in PBL). (**G–I**) Regions in which Bismark and Pash found tissue-specific variation (i.e. high in HF and lower in PBL). (**J–N**) Regions showing medium to high percentage methylation across all four libraries. Overall, the percentage methylation measured by quantitative pyrosequencing compared favorably with the estimates obtained by Bisulfite-seq.

The goal of most Bisulfite-seq experiments is to draw comparisons across multiple samples. Our mapper comparison study is the first to encompass multiple Bisulfite-seq libraries, providing the unique opportunity to quantify the degree of mapping losses as the sample number increases. Our analysis showed that in each library, up to 10% of bins were covered by Bismark and Pash but not by BSMAP; however, when attempting to draw comparisons across just four libraries, this loss of data escalated to 18% of bins genome-wide (Figure 2B). Interestingly, these regions are significantly enriched for both interindividual (Figure 3C) and tissue-specific variation (Figure 3F). This indicates that genomic regions in which BSMAP fails to map may be of particular interest with respect to biologically meaningful variation in DNA methylation. Notably, these regions do not overlap with structural variation and segmental duplication (Figure 4A), but tend to be low in G and rich in T bases (Figure 4D). Hence, these regions are likely difficult to map because bisulfite conversion introduces additional T bases, leading to severely decreased sequence complexity. Subtle distinctions in the alignment strategies of the mappers likely explain the differences in performance. Both Bowtie1 and SOAP align sequences using a primary seed at the start of the sequence. Therefore, potential differences in performance between Bismark and BSMAP likely stem from how the

underlying mappers are used. BSMAP considers multiple seeds for T-rich sites and might not initiate mappings in complex regions, whereas in the case of Bismark, the mapping is done on the three-letter alphabet, exploring all possible mappings. Pash first identifies good anchors (perhaps without T-converted base pairs) throughout a read and then extends the alignment to the entire read. However, BSMAP attempts to seed an alignment only at the beginning of each read, and uses heuristics that abort searches in highly ambiguous regions to maximize mapping speed. This 'primary seeding' could explain the lower coverage we obtained using BSMAP. By generating our own Bisulfite-seq libraries, we had the ability to follow up with quantitative verification. Our quantitative bisulfite pyrosequencing showed excellent agreement with Bisulfite-seq percentage methylation estimates (Figure 5), demonstrating that regions in which BSMAP failed to map are mapped correctly by the other mappers.

Regions covered by BSMAP and Pash, but not by Bismark, showed a high prevalence of overlap with structural variation, particularly segmental duplications (Figure 4A). (However, owing to their non-uniqueness, verification of methylation calls in these regions by pyrosequencing was not feasible.) Based on their ability to tolerate mismatches and find key anchoring base pairs, 'wild card' algorithms should in theory outperform their 'three-letter' counterparts in ambiguous regions such as segmental duplications and repeats. Moreover, Burrows–Wheeler-based aligners such as Bowtie (which Bismark incorporates) achieve high speeds by relying on genomic indices with lossy compression, and thus could ignore potential unique alignments. The performance of Pash in mapping duplicative regions containing only small amounts of unique sequence similarity has been extensively validated (8). Here, using simulated reads (in which we know whence the reads originated), we confirmed that Pash has an exceptional ability to map Bisulfite-seq reads in regions of structural variation (Supplementary Figure S6). However, the simulation did not confirm this ability for BSMAP. Although Pash and BSMAP are both using the 'wild-card' strategy, they use different approaches. Pash explores the various k-mers in the reads, oblivious to the complexity of the genomic sequences, whereas BSMAP hashes multiple seeds for genomic locations, and might not index for read mapping regions of high complexity (e.g. with a large number of CGs and potential C/T ambiguities in the sequenced reads). It should be noted that our comparison was focused on the human genome and included only 100-bp reads mapped as single-end reads. Comparison results may be different in other species, for other read lengths or using paired-end mapping.

CONCLUSIONS

In summary, we have completed a comparative analysis of five read mappers for methylome mapping. Although methylation calls derived using these widely used mappers are highly concordant, we identified significant and important differences in their performance in specific types of genomic regions. In particular, Bismark provides an attractive combination of processing speed, genomic coverage and quantitative accuracy, whereas Pash, although computationally more demanding, offers considerably higher genomic coverage owing to its ability to map within regions of structural variation. BS Seeker yields mapping results similar to those of Bismark, but provides more detailed information in post processing, which may be attractive to some users. We hope our results will help investigators select the Bisulfite-seq mapping algorithm with optimal performance characteristics for their project, and provide useful guidance toward the development of the next generation of mapping tools.

ACCESSION NUMBERS

The raw sequence reads for all four libraries have been deposited in fastq format in GEO [GSE44806].

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The National Institutes of Health Roadmap Epigenomics Program [U01DA025956 to A.M. and R.A.W.]; National Institutes of Health—NIDDK [1R01DK081557]; United States Department of Agriculture [CRIS 6250-51000-055 to R.A.W.]. Funding for open access charge: National Institutes of Health Roadmap Epigenomics Program [U01 DA025956].

Conflict of interest statement. A.M. receives royalties from and participates in the commercial licensing of the Pash program.

REFERENCES

1. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.
2. Clark,S.J., Harrison,J., Paul,C.L. and Frommer,M. (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.*, **22**, 2990–2997.
3. Lister,R., Pelizzola,M., Dowen,R.H., Hawkins,R.D., Hon,G., Tonti-Filippini,J., Nery,J.R., Lee,L., Ye,Z., Ngo,Q.M. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.
4. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
5. Seisenberger,S., Andrews,S., Krueger,F., Arand,J., Walter,J., Santos,F., Popp,C., Thienpont,B., Dean,W. and Reik,W. (2012) The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell*, **48**, 849–862.
6. Bock,C. (2012) Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.*, **13**, 705–719.
7. Smith,A.D., Chung,W.Y., Hodges,E., Kendall,J., Hannon,G., Hicks,J., Xuan,Z. and Zhang,M.Q. (2009) Updates to the RMAP short-read mapping software. *Bioinformatics*, **25**, 2841–2842.
8. Coarfa,C., Yu,F., Miller,C.A., Chen,Z., Harris,R.A. and Milosavljevic,A. (2010) 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.
9. Xi,Y. and Li,W. (2009) BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.
10. Chen,P.Y., Cokus,S.J. and Pellegrini,M. (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, **11**, 203.
11. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
12. Pedersen,B., Hsieh,T.F., Ibarra,C. and Fischer,R.L. (2011) MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics*, **27**, 2435–2436.
13. Coarfa,C. and Milosavljevic,A. (2008) Pash 2.0: scaleable sequence anchoring for next-generation sequencing technologies. *Pacific Symposium on Biocomputing*, **13**, 102–113.
14. Miller,C.A., Hampton,O., Coarfa,C. and Milosavljevic,A. (2011) ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**, e16327.
15. Cokus,S.J., Feng,S., Zhang,X., Chen,Z., Merriman,B., Haudenschild,C.D., Pradhan,S., Nelson,S.F., Pellegrini,M. and Jacobsen,S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.
16. Campagna,D., Telatin,A., Forcato,C., Vitulo,N. and Valle,G. (2013) PASS-bis: a bisulfite aligner suitable for whole methylome analysis of Illumina and SOLiD reads. *Bioinformatics*, **29**, 268–270.
17. Lim,J.Q., Tennakoon,C., Li,G., Wong,E., Ruan,Y., Wei,C.L. and Sung,W.K. (2012) BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. *Genome Biol.*, **13**, R82.
18. Otto,C., Stadler,P.F. and Hoffmann,S. (2012) Fast and sensitive mapping of bisulfite-treated sequencing data. *Bioinformatics*, **28**, 1698–1704.
19. Chatterjee,A., Stockwell,P.A., Rodger,E.J. and Morison,I.M. (2012) Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Res.*, **40**, e79.
20. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
21. Li,R., Li,Y., Kristiansen,K. and Wang,J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.
22. Waterland,R.A., Kellermayer,R., Laritsky,E., Rayco-Solon,P., Harris,R.A., Travisano,M., Zhang,W., Torskaya,M.S., Zhang,J., Shen,L. *et al.* (2010) Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet.*, **6**, e1001252.
23. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.
24. Chen,H. and Boutros,P.C. (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, **12**, 35.
25. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
26. Shen,L., Guo,Y., Chen,X., Ahmed,S. and Issa,J.P. (2007) Optimizing annealing temperature overcomes bias in bisulfite PCR methylation analysis. *Biotechniques*, **42**, 48–58.
27. Krueger,F., Kreck,B., Franke,A. and Andrews,S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.
28. Harris,R.A., Wang,T., Coarfa,C., Nagarajan,R.P., Hong,C., Downey,S.L., Johnson,B.E., Fouse,S.D., Delaney,A., Zhao,Y. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.
29. Hansen,K.D., Langmead,B. and Irizarry,R.A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, **13**, R83.
30. Ivanov,M., Kals,M., Kacevska,M., Metspalu,A., Ingelman-Sundberg,M. and Milani,L. (2013) In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. *Nucleic Acids Res.*, **41**, e72.