BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 13 2014, pages 1814–1822
doi:10.1093/bioinformatics/btu126

Genome analysis

Advance Access publication March 7, 2014

DMAP: differential methylation analysis package for RRBS and WGBS data

Peter A. Stockwell[1,*,†], Aniruddha Chatterjee[2,3,†], Euan J. Rodger[2] and Ian M. Morison[2]

[1]Department of Biochemistry, University of Otago, 710 Cumberland Street, Dunedin 9054, New Zealand, [2]Department of Pathology, Dunedin School of Medicine, University of Otago, 270 Great King Street, Dunedin 9054, New Zealand and [3]Gravida: National Centre for Growth and Development, 2-6 Park Ave, Grafton, Auckland 1142, New Zealand

Associate Editor: Inanc Birol

ABSTRACT

Motivation: The rapid development of high-throughput sequencing technologies has enabled epigeneticists to quantify DNA methylation on a massive scale. Progressive increase in sequencing capacity present challenges in terms of processing analysis and the interpretation of the large amount of data; investigating differential methylation between genome-scale data from multiple samples highlights this challenge.

Results: We have developed a differential methylation analysis package (DMAP) to generate coverage-filtered reference methylomes and to identify differentially methylated regions across multiple samples from reduced representation bisulphite sequencing and whole genome bisulphite sequencing experiments. We introduce a novel fragment-based approach for investigating DNA methylation patterns for reduced representation bisulphite sequencing data. Further, DMAP provides the identity of gene and CpG features and distances to the differentially methylated regions in a format that is easily analyzed with limited bioinformatics knowledge.

Availability and implementation: The software has been implemented in C and has been written to ensure portability between different platforms. The source code and documentation is freely available (DMAP: as compressed TAR archive folder) from http://biochem.otago.ac.nz/research/databases-software/. Two test datasets are also available for download from the Web site. Test dataset 1 contains reads from chromosome 1 of a patient and a control, which is used for comparative analysis in the current article. Test dataset 2 contains reads from a part of chromosome 21 of three disease and three control samples for testing the operation of DMAP, especially for the analysis of variance. Example commands for the analyses are included.

Contact: peter.stockwell@otago.ac.nz or aniruddha.chatterjee@otago.ac.nz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 14, 2013; revised on January 28, 2014; accepted on March 2, 2014

1 INTRODUCTION

DNA methylation is arguably the most stable epigenetic mark that plays a key role in regulating development and disease (Baylin and Bestor, 2002; Law and Jacobsen, 2010). One of the most fundamental challenges for epigeneticists is to identify DNA methylation differences between genomes. For instance, differential methylation between diseased and normal samples, interindividual variation within a population, differences between tissues or species and so on are of biological and clinical relevance.

The rapid improvement in next-generation sequencing technologies now provides opportunities to interrogate DNA methylation at single base resolution with high coverage across multiple samples. Bisulphite treatment converts unmethylated cytosines to uracils (and ultimately to thymine after amplification), although leaving methylated cytosines unchanged. Therefore, bisulphite treatment combined with next-generation sequencing (BS-Seq) has become a preferred method to generate base-resolution DNA methylation maps. Because whole-genome bisulphite sequencing (WGBS) is still expensive and generates challenging amounts of raw data, reduced representation bisulphite sequencing (RRBS) provides a cost-effective alternative for whole-genome methylation sequencing. RRBS has been widely used by several groups worldwide to interrogate functionally important genomic regions at high-sequencing coverage and sensitivity (Baranzini *et al.*, 2010; Bock *et al.*, 2011; Chatterjee *et al.*, 2012; Gertz *et al.*, 2011; Gu *et al.*, 2010; Smallwood *et al.*, 2011; Steine *et al.*, 2011; Xi *et al.*, 2012).

During the past few years, several alignment tools have been developed to cope with asymmetric mapping issues of bisulphite converted sequenced reads and to map millions of reads with reasonable speed to the reference genome. Some of these aligners are RMAP (Smith *et al.*, 2009), BS Seeker (Chen *et al.*, 2010), Bismark (Krueger and Andrews, 2011), RRBSMAP (Xi *et al.*, 2012), BatMeth (Lim *et al.*, 2012) and PASS-bis (Campagna *et al.*, 2013). Recent comparative analyses have improved our understanding of the efficiency, accuracy and algorithm of these aligners (Chatterjee *et al.*, 2012; Kunde-Ramamoorthy *et al.*, 2014). Additionally, tools have been developed for generating methylation calls and visualization. Integrated Genome Viewer (Thorvaldsdottir *et al.*, 2013) and MethVisual (Sun *et al.*, 2013) allow visualization of sequenced reads and regional analysis. BiQ Analyzer HT allows site-specific DNA methylation analysis (Schmieder and Edwards, 2011), and SAAP-RRBS can perform alignment, methylation calls, annotation of CpG sites and visualization (Ziller *et al.*, 2013).

methylKit (Akalin *et al.*, 2012a), an R package, enables detection of differentially methylated CpG sites (DMCs). methylKit

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

applies a Fisher's exact test or logistic regression to calculate *P*-values that are adjusted to q-values for multiple test correction using a SLIM approach (Wang *et al.*, 2011). For WGBS data, BSmooth performs local averaging and sample-wise smoothing of methylation values after alignment and methylation estimates by read position (Hansen *et al.*, 2012). BSmooth applies numerous CpG-wise t-tests, and based on a t-test threshold, differentially methylated regions (DMRs) are defined. BiSeq, an R package, performs smoothing of methylation data within CpG clusters considering spatial dependence (Hebestreit *et al.*, 2013). Differential methylation is then detected in CpG clusters, the false discovery rate is controlled and finally DMR boundaries are defined.

In contrast to alignment and visualization tools, the number of software packages available to detect DMRs is small. Some tools offer detection of DMCs and some DMRs and most of them are hardwired, i.e. they provide little flexibility in the choice of analysis parameters such as coverage criteria, DMR length and the type of statistical test. Further, most available tools do not provide options for further analysis of DMRs with respect to its genomic position such as the transcription start sites, genes, CpG features and so on.

Here, we describe differential methylation analysis package (DMAP), a pipeline that can directly import the output from any bisulphite aligner in Sequence Alignment/Map (SAM) format and identify differential methylation. We have primarily designed the package to handle data from RRBS experiments (which uses 40–220 bp MspI digested genomic fragments), but it can be used effectively to investigate WGBS data for any eukaryotic genome as well. A suite of statistical tests is included in DMAP [Chi-square test, Fisher's exact test and analysis of variance (ANOVA)] to identify methylation differences between different groups and conditions. For RRBS, we introduce a novel approach of identifying differential methylation based on MspI fragments [differentially methylated fragment (DMF)]. Further, DMAP provides genomic relationship information (nearest gene, exon, introns and CpG features) for each DMR or DMF.

2 METHODS AND ALGORITHMS

2.1 DMAP package and input data

DMAP contains two main programs. (i) *diffmeth:* The input files to *diffmeth* are either SAM files from Bismark alignment (Krueger and Andrews, 2011) or the older native format produced by the Bismark methylation_extractor program, comprising a single line for each mapped CpG giving the chromosome, the CpG position and the methylation status (+/−). Alternatively, if other aligners (such as BSMAP and RMAPBS) are used, then the files (BED file or text files) can be processed by the *rmapbscpg2* ancillary program before analysis with *diffmeth*. By default, *diffmeth* does not impose any *P*-value cutoff for identifying DMR; it returns a *P*-value for each investigated region/fragment to allow user-specified threshold *P*-values and independent application of multiple test corrections.

(ii) The final output file from *diffmeth* program can then be used in the second main program of DMAP, *identgeneloc,* to identify proximal genes and features (transcription start sites, exons/introns, etc.), relationship to CpG features (CpG island core/shore/shelf) and distances from each feature (Fig. 1). This operation is performed by a command-line program, which reads genomic feature table information and relates candidate regions from the previous step to annotated features. The application uses code originally developed in another context (Jacobs *et al.*, 2009) and is capable of parsing feature table information from GenBank, EMBL, GTF, GFF3 and SeqMonk feature files, although the latter has been extensively tested. If SeqMonk feature table information is used (-Q switch), then it is possible to specify biotype for a gene (e.g. protein coding, pseudo-gene, miRNA). Supplementary Information 2 and the program document contain a user guide to set up the software and a step-by-step instruction manual for operation of the analysis pipeline.

2.2 Units of DNA methylation analysis

2.2.1 DMC approach Differential methylation patterns can be investigated in several ways. One of the approaches is to analyze each CpG site (with adequate coverage) in each sample and then to identify DMCs. DMAP permits the user to interrogate the methylated (represented as + sign) and unmethylated (represented as – sign) counts for single CpG sites (e.g. Table 4) in the datasets, but does not have options for detecting DMCs. methylKit, an R package, uses a single CpG approach and provides options for detecting DMCs in RRBS and BS-Seq data (Akalin *et al.*, 2012a). However, in WGBS or RRBS protocols, millions of CpG sites are investigated (e.g. in humans, WGBS covers ~30 million and RRBS covers ~4 million CpG sites). The investigation of a large number of CpG sites greatly enhances the false discovery rate. Variation at single sites is greater than that of a contig of sites because the relatively lower coverage per site increases the sampling variation (Ehrlich and Lacey, 2013). A DMC approach is perhaps more useful when a small number of CpG sites are analyzed.

2.2.2 DMR approach Use of a fixed or sliding window (typically 1000 bp length) as a unit of methylation analysis is another common approach for detecting DMRs (Bock *et al.*, 2012; Li *et al.*, 2010). DMAP includes options for investigating differential methylation on a user-specified tiled window of any length. Although the tiled DMR approach is well-suited for WGBS, for RRBS, where only 2.5% of the genome is sequenced, the majority of the windows will be empty or have partial inclusion of fragments. Further, if a small region is variably/differentially methylated between individuals, use of a 1000 bp or longer window might dilute this variation (Ehrlich and Lacey, 2013) and therefore might be not be detected if large window size is used.

2.3 Implementing MspI fragments as a unit of analysis for RRBS (DMF approach)

For RRBS, we introduced a new MspI fragment-based approach for investigating DNA methylation. This approach is conceptually similar to the DMR approach, but instead of fixed-length windows, MspI-digested fragments of 40–220 bp lengths were used as the unit of analysis. After Bismark alignment, the methylation_extractor program returns information for each mapped CpG site, its genomic position and methylation status. *diffmeth*

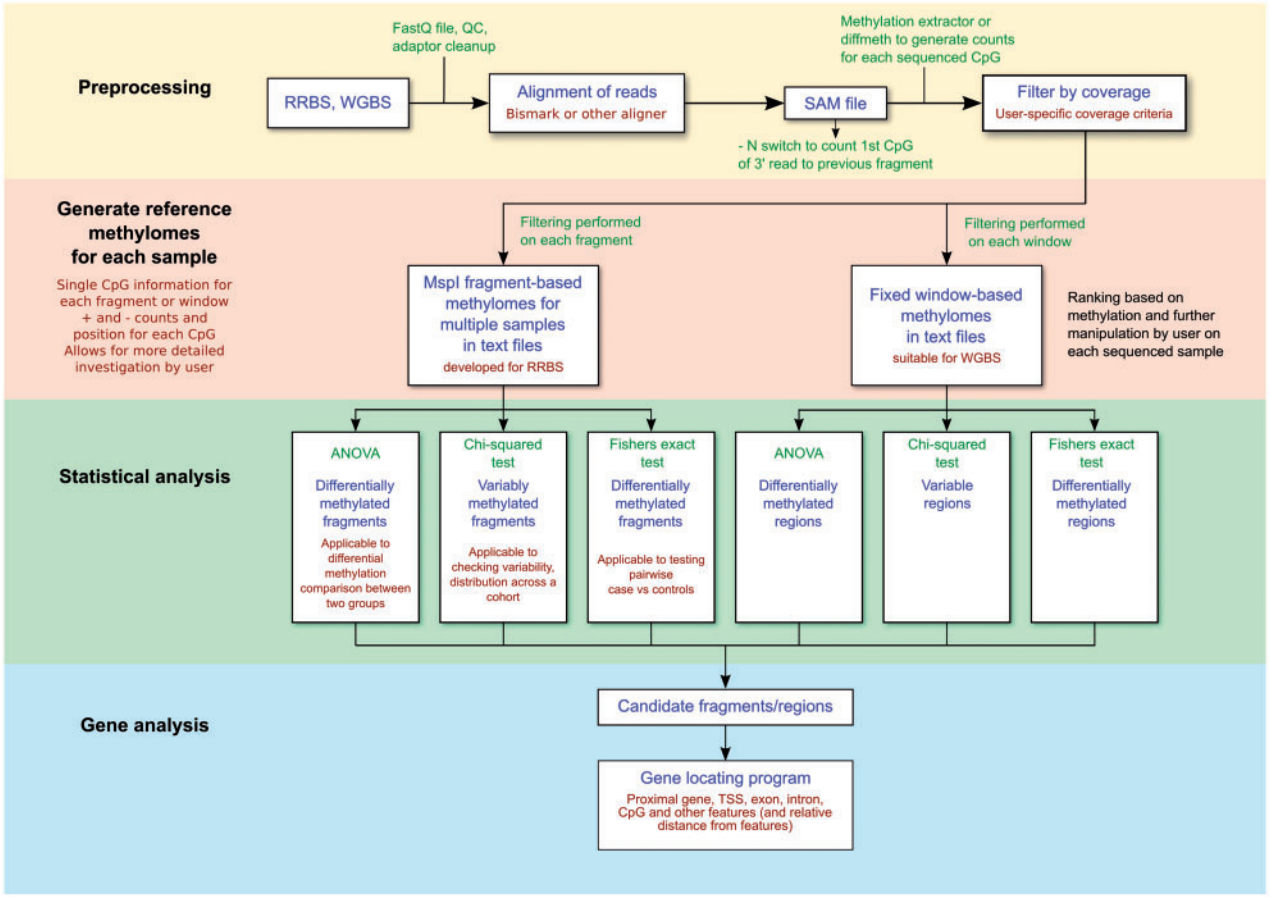

Fig. 1. Flow diagram of options and analysis units in DMAP

scans the genomic sequence of each chromosome for MspI cleavage sites and CpGs, building a list of those conforming to the required size range of 40–220 bp. For each sample, the methylated and unmethylated CpGs are checked to find the fragment on the list (if any) to which the CpG positions match and the methylated and unmethylated counts are incremented. This information was used to calculate coverage (+ and − counts, where + is a methylated CpG and − is an unmethylated CpG) and quantify the methylation of each fragment. Then based on a coverage cutoff (different cutoff criteria for filtering fragments can be applied; see details in documentation), fragments with low sequencing coverage were discarded, and a list of fragments and their methylation status were produced for each sample. Appropriate statistical tests and a *P*-value can then be applied to identify DMFs. The sequenced reads in RRBS come from MspI-digested fragments, and as one CpG site is included in the recognition site (C'CGG) of the enzyme, each fragment will contain at least one valid analyzable CpG, irrespective of the CpG density of the region. DMAP provides flexibility in the choice of coverage criteria to include fragments from both CpG-poor and CpG-rich regions (see documentation for setting coverage threshold).

While implementing the fragment-based approach, adjacent MspI fragments posed a computational challenge. If two fragments are adjacent, then the methylation counts of the mapped reads, which start from the junctional MspI site, could be counted in either of the fragments, as in the genome they share the same location. However, for an accurate coverage calculation for the fragments, the methylation counts from the CpG at the junction of two adjacent fragments are attributed to the fragment from which they were derived (see Supplementary Information 1, section 7, for detailed demonstration of this behavior).

To ensure correct attribution of junctional CpG methylation, we added a feature to DMAP that uses the data from the SAM files (with the –N switch) to retain the identity of the strand onto which a sequenced read is mapped. Then for the reverse strand-mapped reads, the program identifies the fragment to which these reads were mapped and accumulates the CpG information for that fragment, including the first CpG site of the read but not for the succeeding adjacent fragment. In the reference genome, the last CpG site of an MspI fragment and the first CpG site of the adjacent fragment are the same, but after this correction, CpG information from the overlapping reads was counted under the fragment from which the reads were derived.

Non-specific activity of the MspI enzyme and partially degraded DNA could result in sequenced reads without a MspI start site. For RRBS, 29% non-MspI sequenced reads has been reported (Akalin *et al.*, 2012b). However, we observed

a lower prevalence of non-MspI reads in our test RRBS libraries (median percentage of non-MspI reads = 8.1; $n = 11$). Presence of non-MspI reads might affect unique alignment but does not influence DMAP. DMAP calculates methylated and unmethylated counts for each CpG site in a fragment; therefore, a fragment will qualify for subsequent analysis if sufficient reads had uniquely mapped to it, irrespective of the base composition of sequenced reads (i.e. MspI or non-MspI start of a read).

In a MspI-fragment, the CpG sites are likely to show more similarities in terms of read coverage and DNA methylation levels compared with the CpG sites in a large DMR. Further, as a fragment is a contig of sites, it will decrease the sampling and technical variation to a greater extent. Therefore, we believe the fragment-based approach is the intermediate approach to DMR/DMC approaches and perhaps better suited for RRBS data.

2.4 Statistical tests to identify DMF and DMR

The Fisher's exact test over a sliding window with a specified fold difference is a widely used approach for assessing differential methylation between two samples or groups of samples (Bock *et al.*, 2010; Glastad *et al.*, 2013; Gu *et al.*, 2010; Ivanov *et al.*, 2013; Li *et al.*, 2010). For RRBS data analysis, for a given MspI fragment, Fisher's exact tests can be performed between multiple samples by giving a continuous list of SAM files in the command in *diffmeth* tool of DMAP package. In this case, the probability of multiple pairwise tests will be given and the lowest probability taken to indicate the extent of differential methylation, although this may obscure a number of insignificant differences between other samples.

To investigate the extent of interindividual variability in DNA methylation across multiple samples, we have used a Chi-squared test. To perform this test, a list of SAM files should be provided as an input to *diffmeth*. Various thresholds can be applied to restrict the tests to fragments and samples that meet criteria for CpG number, density of CpG mapping and fold difference. For WGBS data, instead of fragments, tiled windows of defined length can be investigated in a similar way. Output from this part of DMAP consists of a line for each qualifying fragment or region giving the chromosome number, region start and end positions, length, CpG count, coverage, the probability and the type of statistic applied (Fig. 1 and Table 3).

Classifying subjects into treatment or disease versus control groups is a usual task in DNA methylation analysis. To compare methylation between two different groups, one strategy is to concatenate the CpG position files or SAM files for each group and perform a pairwise comparison. However, this might lead to significant data loss while comparing multiple samples. A better strategy is implemented in DMAP using ANOVA and the F ratio test to determine the significance of methylation differences between the groups in relation to the residual variation within each group. ANOVA runs allow SAM files to be assigned to either of two groups, generating an F (1, n) value where n depends on the number of qualifying individuals for each region or fragment. The statistical significance of the F statistic is estimated using a continued fraction iterative method (Press *et al.*, 1993).

3 RESULTS AND OUTPUTS

3.1 Reference methylome

DMAP can produce reference methylomes for individual samples after filtering regions with a specified coverage criterion [for example, reference methylomes can be based on MspI fragments (for RRBS) or user-specified tiled windows]. Table 1 shows the *diffmeth* output from a fragment-based methylome for a human RRBS library generated from peripheral blood. A similar output based on 1000 bp tiled windows is shown in Table 2 for a different region of the same dataset. These outputs are produced as text files, which can be easily subjected to further analysis.

3.2 Differential methylation analysis

While running differential methylation analysis, DMAP produces a list of analyzed regions showing corresponding *P*-values and the details (such as name of the test, degrees of freedom if applicable) of the statistical test applied. By default, DMAP does not impose any cutoff value to detect DMF or DMR. Users can specify statistics and *P*-value cutoff to set a threshold for calling DMF or DMR. This provides flexibility to the users and options to apply multiple test corrections methods (e.g. Bonferroni, false discovery rate or Holms methods) and set stringent *P*-value cutoffs for detecting differential methylation. Table 3 shows an example of candidate fragments after a test has been performed on five human peripheral blood RRBS samples for differential methylation analysis using the χ^2 statistic. A similar analysis can be performed using tiled windows.

3.3 Single CpG investigation

DMAP does not allow detection of differential methylation at single CpG sites; however, if investigation of each CpG is sought, the *diffmeth* program of DMAP can produce + (methylated) and − (unmethylated) counts for each CpG site within a fragment or tiled window for each sample. Table 4 provides an example of single CpG counts of an MspI fragment in chromosome 1, which contained eight CpG sites for five RRBS samples as produced by *diffmeth*. Alternatively, single CpG site differential methylation can be performed using methylKit (Akalin *et al.*, 2012a).

Table 1. MspI fragment based methylome for RRBS

Chromosome number	Start	End	Length	CpGs	+ and − hits	%Methylation
1	863 942	864 129	188	7	60+/10−	85.71
1	864 313	864 414	102	3	14+/50−	21.88
1	875 309	875 363	55	3	10+/96−	9.43
1	877 737	877 866	130	13	0+/216−	0
1	879 180	879 369	190	8	72+/8−	90

Note: The CpGs column indicates the number of unique CpGs in the fragment. + (methylated) and − (unmethylated) hits gives the total number of counts in the fragment, and the % methylation was calculated from these counts by the diffmeth program in the DMAP package.

Table 2. One thousand base pair tiled window-based reference methylome

Chromosome number	Start	End	Length	CpGs	+ and − hits	+-hits/CpG	%Methylation
1	1 100 001	1 101 000	1000	63	492+/174−	10.57	73.87
1	1 115 001	1 116 000	1000	53	609+/107−	13.51	85.06
1	1 146 001	1 147 000	1000	36	353+/37−	10.83	90.51
1	1 243 001	1 244 000	1000	167	17+/2066−	12.47	0.82
1	1 244 001	1 245 000	1000	115	68+/1605−	14.55	4.06

Note: The CpGs column indicates the number of unique CpGs in each window of 1000 bp. The + (methylated) and − (unmethylated) hits gives the total number of counts in the window. +-hits/CpG represents the average coverage per CpG in each window. The % methylation was calculated from these counts by the diffmeth program in the DMAP package.

Table 3. Candidate regions after differential methylation analysis[a]

Chromosome number	Start	End	Len	CpGs	Total hits	Pr	Test
1	10 497	10 588	92	8	14 639	5.55E-16	Chi_188.83_9df
1	662 657	662 705	49	5	1988	0.03161	Chi_18.32_9df
1	805 467	805 521	55	10	9339	2.22E-16	Chi_151.46_10df
1	839 516	839 591	76	4	2719	1.44E-15	Chi_226.68_10df
1	845 847	845 934	88	5	1871	0.1048	Chi_15.82_10df

[a]Chi-squared test was performed on RRBS outputs from 11 individuals taking MspI fragment as a unit of analysis. Note, degrees of freedom is not always 10 (n − 1), as some samples had insufficient coverage for some fragments.

Table 4. Single CpG counts for multiple individuals for a fragment[a]

Sample	CpG sites							
	10 497	10 525	10 542	10 563	10 571	10 577	10 579	10 589
1	88+	275+	271+	237+	278+	180+	155+	166+
	18−	16−	21−	55−	11−	111−	137−	18−
2	38+	69+	70+	66+	72+	48+	48+	22+
	3−	4−	4−	7−	1−	25−	25−	9−
3	129+	463+	467+	448+	474+	264+	242+	285+
	26−	16−	15−	33−	7−	62−	83−	32−
4	95+	276+	277+	268+	283+	225+	196+	174+
	5−	12−	11−	19−	6−	65−	94−	13−
5	95+	276+	277+	268+	283+	225+	196+	174+
	5−	12−	11−	19−	6−	65−	94−	13−

Note: Fragment details: #Chr 1; Start 10 497 bp; End 10 588 bp; Length 92 bp; CpGs 8.

3.4 Gene and feature identification

The *identgeneloc* program of DMAP relates each DMF or DMR (or any regions of interest from reference methylomes) to the nearest gene by comparing the genomic coordinates of the start and the end of the DMF or DMR with the coordinates of the gene and gives relative distances from the transcription start site. *identgeneloc* considers the sense of the gene (5′ or 3′) and relates the DMF or DMR with respect to the upstream region of the gene. The program includes options for users to impose distance limits on how far valid genes can lie from the fragment. *identgeneloc* can also provide CpG features (CpG island, shore or core) for a DMF or DMR. Further, for a region internal to a gene, an option is included to return information on whether the fragment is located on an exon, intron or spans over intron/exon or exon/intron boundaries. Table 5 provides an example of an *identgeneloc* output, showing candidate DMFs, generated from a human RRBS library (annotation source: SeqMonk feature table file). The output is a tab-delimited text file.

3.5 Comparison with other tools

We performed a comparative performance analysis between DMAP, methylKit and BiSeq using test dataset 1, which is available at http://biochem.otago.ac.nz/research/databases-software/.

Table 5. Output from gene locating operation in DMAP

Chromosome number	Start	End	Unique CpG	*P*	Chi value	Gene distance	Gene feature	CpG feature	Strand	Gene
1	10 497	10 588	8	5.55E-16	Chi_188.8381_9df	19779	—	CpGI_shore	5′	MIR1302-11
1	805 467	805 521	10	2.22E-16	Chi_151.4682_10df	−6815	on intron	CpGI_shore	3′	FAM41C
1	839 516	839 591	4	1.44E-15	Chi_226.6835_10df	7225	—	CpGI_shore	5′	RP11-5407.1
1	870 573	870 636	6	1.03E-09	Chi_62.8695_10df	−10375	on intron	—	5′	SAMD11
1	896 009	896 063	11	1.07E-08	Chi_55.2993_9df	−95	on exon	CpGI_core	5′	KLHL17
1	909 381	909 461	6	0	Chi_115.3572_8df	−7583	exon intron boundary	—	5′	PLEKHN1
1	911 470	911 539	7	2.21E-12	Chi_74.2590_9df	−6026	on exon	CpGI_shore	3′	C1orf170
1	911 540	911 600	4	0	Chi_109.2673_8df	−5956	on exon	CpGI_shore	3′	C1orf170
1	911 995	912 069	6	1.55E-15	Chi_597.1420_10df	−5501	exon intron boundary	CpGI_shore	3′	C1orf170

Note: The *identgenloc* here provides data of chromosome, length of the region, number of CpG sites contained within the region, *P*-value, statistical test applied, distance in relation to the gene (calculated from the start of the gene), relationship with the gene (e.g. upstream, exon, intron), CpG feature relation, strand and the name of the associated gene. The − in the gene distance column indicates the region is inside the gene body. The output is a tab-delimited text file, suitable for importing into Microsoft Excel. Gene distance is calculated from the start of the gene.

Test dataset 1 contains two SAM files with uniquely aligned reads for chromosome 1 from a control (2 851 855 reads) and a disease sample (1 122 068 reads). A pairwise test was performed on CpG sites or regions common between both samples. DMAP completed the differential methylation analysis in <3.1 min, which was >5 times faster than methylKit and >10 times faster than BiSeq (Table 6). The main reason for the faster operation of DMAP is that after the alignment—using the SAM files—it is possible to filter fragments (or tiled windows) by coverage, set fold methylation difference criteria and perform a statistical test in one step with a single command. Further, because DMAP is written in C it runs as compiled machine code, and therefore, it executes efficiently. After differential methylation analysis, a second operation of DMAP (performed by *identgeneloc*) produces gene and CpG feature of the candidate regions. *identgeneloc* took 1.2 min to complete the second operation for 9362 investigated fragments in chromosome 1 between the disease and control from test dataset 1.

Analysis by methylKit (version 0.5.7) resulted in 935 DMCs (filtered by CpG coverage of ≥10, DESTRAND = TRUE, q-value of <0.01 and percent methylation difference >25%). BiSeq (version 1.2.4) investigated 165378 clusters and resulted in 402 DMR entries (criteria of analysis was min.sites = 20, quantile (totalReads (rrbs.clust.unlim) [ind.cov],0.9, minDiff = 0.25, max.dist = 100). Of the BiSeq DMRs, 131 of them were single CpGs. The other 271 regions varied in length; the median length of the DMRs was 18.5 bp and the largest DMR 781 bp. Using the same dataset, DMAP was used to investigate 9362 common fragments (coverage filter: ≥2 CpGs in a fragment having coverage of ≥10; the F2 t10 switch in DMAP) containing 78 318 CpG sites. DMAP identified 367 significant DMFs (basic cutoff = 0.05, adjusted *P*-value cutoff = 5.34×10^{-6} after Bonferroni correction). The 367 DMFs contained 3215 CpG sites (Table 6).

When we compared the co-ordinates of the DMCs derived from methylKit, we found that 318 of 935 DMCs from methylKit overlapped with the CpGs contained within the 367 DMFs identified by DMAP. Similarly, 190 DMCs from methylKit overlapped with CpGs within the 402 DMRs identified by BiSeq. Also, 439 of 1827 CpG sites in BiSeq-identified DMRs overlapped with 3215 CpG sites in the DMAP-identified DMFs. Overall, 157 CpG sites were identified by all three tools. To investigate if the tools performed better in identifying any particular regions of the genome compared with the others, we mapped the genomic locations of DMCs, DMFs and DMRs identified by methylKit, DMAP and BiSeq respectively (Fig. 2). Using SeqMonk feature table information (based on Ensembl annotation), each region was related to its nearest protein coding gene and distances from the start of the gene was calculated. DMAP identified higher proportion of promoter associated regions and lower intronic regions compared to methylKit and BiSeq. However, methylKit identified a higher proportion of CpG sites that were 5 kb or further apart from the transcription start site (TSS > 5 kb) compared to the regions identified by DMAP and BiSeq. All three tools identified similar levels of exonic differential methylation. Further, BiSeq and DMAP identified similar proportion of exon-intron junction DMRs. Exon-intron junction identification is not possible with methylKit as it identifies DMCs. The CpGs that were common between all the three tools were more prevalent in far upstream of the genes (TSS > 5 kb) and relatively lower in the promoter and introns.

4 DISCUSSION

A higher number of DMCs is expected compared to the number of DMFs/DMRs, because each DMF/DMR will contain several differentially methylated CpGs. DMF/DMR will also contain CpG sites that individually will not qualify as differentially methylated. Further, there will be several independent DMCs that will not form part of a DMF or DMR. This explains the finding of higher DMCs identified by methylKit and lower DMR/DMFs identified by BiSeq and DMAP. A frame-to-frame comparison of differential methylation patterns between these tools is not possible because each of them uses a different unit of analysis to determine differential methylation. The

Table 6. Comparative differential methylation analysis between DMAP, methylKit and BiSeq[a]

Tool	Number of DMCs	Number of DMRs	Number of DMFs	Number of CpG sites contained in the DMFs	Time taken to perform analysis[a]
DMAP	—		367	3215	3.1 min
methylKit (version 0.5.7)	935		—	935	18 min[b]
BiSeq (version 1.2.4)	131[c]	271	—	1827	25 min[d]

[a]This analysis was performed on a Mac Pro with 64 bit duo quad core Intel Xeon processors and with 22GB RAM running MacOS 10.7.
[b]MethylKit produces a CpG.txt file from each SAM file and accepts that as an input to create R object; this step is relatively time-consuming.
[c]BiSeq produced a mix of 131 DMCs and 271 DMRs of variable length.
[d]The predictMeth step was relatively time-consuming for BiSeq.

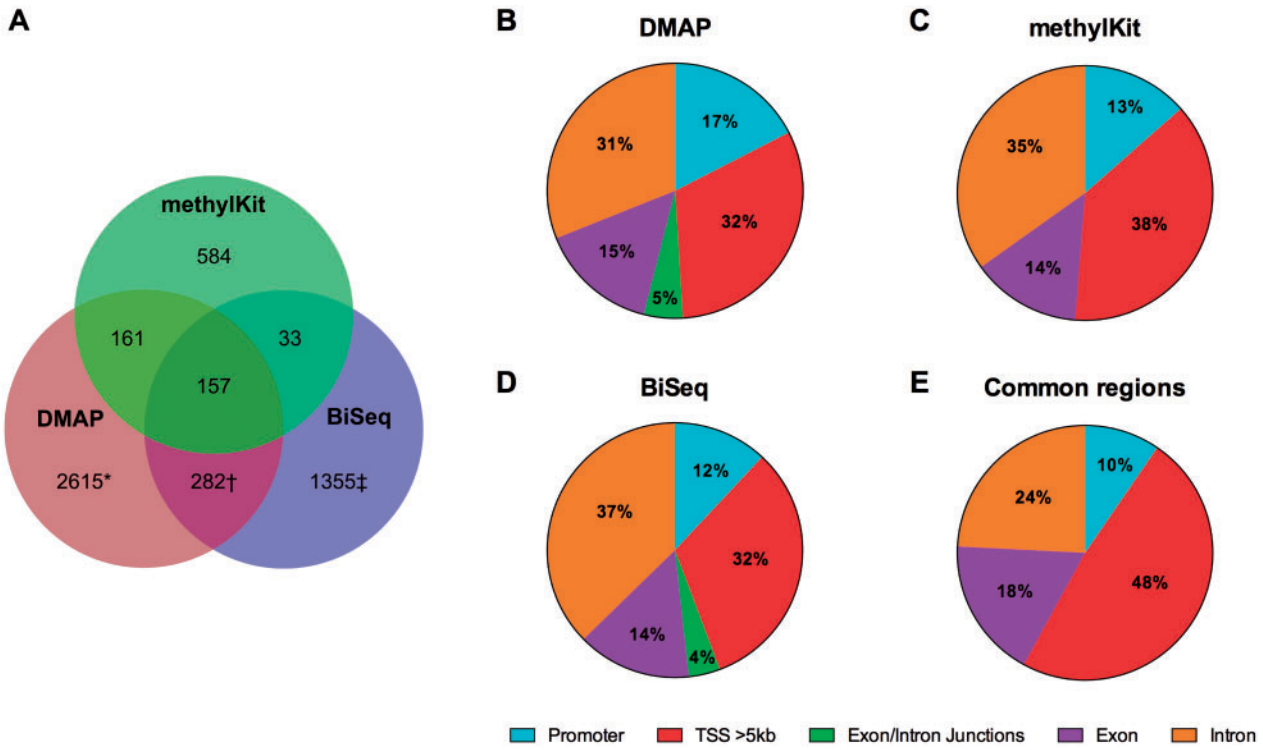

Fig. 2. Overlap and genomic locations of the regions or CpG sites identified by methylKit, DMAP and BiSeq. (**A**) Overlapping CpG sites between methylKit, DMAP and BiSeq. *For DMAP, 3215 CpG sites are within the 367 DMFs, and ‡ for BiSeq, 1867 CpG sites are within the 402 DMRs were included. † indicates the overlap between the 3215 sites of DMAP versus 1867 sites of BiSeq. (**B**) Genomic locations of the 367 DMFs of DMAP. (**C**) Genomic locations of the 935 DMCs identified by methylKit. (**D**) Genomic locations of the 402 DMRs identified by BiSeq. (**E**) Genomic locations of 157 CpG sites that were common to all three programs. Promoters were defined as regions 0–5 kb upstream from the start of the gene

fragment-based approach is specific to DMAP whereas methylKit detects DMCs and BiSeq detects DMRs of variable length. Each program uses different statistical tests (see Table 7), different parameters for defining the unit of DNA methylation analysis (i.e., DMR boundaries or DMCs), and different criteria for including CpGs based on coverage. methylKit and DMAP investigate differential methylation in the 4 million CpG sites or 647626 MspI fragments (in a 40–220 bp human RRBS genome), respectively. However, BiSeq considers the spatial arrangement of CpG sites in the genome and defines CpG clusters by specifying a number of frequently covered CpG sites (option: min.sites) that are close to each other (option: max.dist) and uses these clusters for subsequent analysis. Imposing flexible criteria for defining CpG clusters (i.e., lowering min.sites value) in analysis would result in higher number analyzable CpG clusters and higher number of DMRs but might enhance the chances of false discovery. Further, for this analysis we did not specify any DMR cut-off length in BiSeq. Therefore DMCs were also detected in BiSeq as we wanted to include all the possible regions for a fair comparison with methylKit. During the review process, we found that BiSeq version 1.2.0 and 1.2.3 contained an erroneous readBismark function, which produced a very high number of DMRs for our analysis (as a result of false methylation calls), the function was corrected in version 1.2.4. Therefore,

Table 7. Comparison of analysis features of DMAP, methylKit, BiSeq and SAAP-RRBS

Features	DMAP	methylKit	BiSeq	SAAP-RRBS
Program	Command line tool, written in C	R package	R package	R package
Alignment[a]	No	No	No	Yes
Single/paired end	Accepts both	Accepts both	Accepts both	Single-end only
Differential methylation analysis	Yes	Yes	Yes	No
Statistical method	Fisher's exact Chi-squared test and ANOVA[b]	Fisher's exact or logistic regression[c] to calculate *P*-values. *P*-values were adjusted to q values	Beta regression model	—
DMC approach	No[d]	Yes	No[e]	—
DMR approach	Yes	No	Yes	—
DMF approach	Yes	No	No	—
Gene features of each candidate	Yes	No	No	—
CpG features of each candidate	Yes	No	No	—

[a]Supports Bismark alignment files.
[b]Multiple test correction can be applied post hoc after the differential methylation analysis.
[c]The choice of statistical test will depend on number of samples in each condition.
[d]Although does not support DMC approach but provides single CpG counts for each base in the whole RRBS dataset.
[e]BiSeq produces a mix of DMCs and DMRs if a DMR length is not specified.

it is advisable to use several tools in combination to ensure detection of sensible detection of differentially methylated regions for biological interpretation.

In terms of coverage, methylKit called the CpG sites with 10 or more reads from the aligned SAM files. In DMAP, fragments having two CpG sites with 10 or more reads were included for analysis. In contrast, BiSeq uses a quantile approach for smoothing methylation data where a higher weighting is given to CpG sites with high coverage and sites with an unusually high coverage are excluded (for example, to the 90% quantile in this analysis). These differences could account for some of the variation observed in the comparative differential methylation analysis performed here.

We present DMAP, an analysis package that filters and processes aligned bisulphite sequenced data to generate comprehensive reference methylomes (tile based and fragment based) with flexibility for users. From SAM files, DMAP provides statistically significant DMRs and relates them to genes and CpGs. Statistical approaches for the analysis of genome-wide methylation data are not yet well characterized. A caveat to the use of statistical tests for fragment or window-based approaches is that methylation values for CpGs within a sequenced read are likely to be correlated, and thus statistical significance can be overestimated. Therefore, further work is needed to devise better statistical methods for accurate detection of differential methylation.

Aside from some awk scripts, DMAP is written in C and executes efficiently. In our test runs, the *diffmeth* program was able to produce a list of candidate regions (while processing 11 human RRBS samples) in 4 h. The output from this step was processed in 20 min by *identgeneloc* to produce gene features of the candidate regions. Although the package was initially developed for the human genome, the code was modified to work with any eukaryotic genome. Optionally, DMAP has no expectation of an X and Y chromosome and can work with any number of autosomal chromosomes. We tested the package with zebrafish genome (Zv9 assembly), which has 25 chromosomes (and no X and Y chromosome), and all the features described in this article worked successfully (Chatterjee *et al*., 2013). To our knowledge, DMAP is the first tool that accepts unsorted raw SAM alignment files as an input, detects DMR or DMF, provides information and distances of nearest genes and CpG features in relation to each DMF or DMR. The outputs (exported to text files) are relatively easy for bench scientists without bioinformatics expertise to analyze and use with other tools.

ACKNOWLEDGEMENTS

The authors thank New Zealand Genomics Limited (Centre for Innovation, 87 St David Street, Dunedin 9016, New Zealand) for providing support to Dr Peter Stockwell. The authors are grateful to Dr Mik Black, Department of Biochemistry, for his advice in statistical analysis. The authors thank Dr Simon Andrews and Dr Felix Krueger, Babraham Bioinformatics, Cambridge, UK, for their help during the development and trouble shooting of some aspects of DMAP.

Funding: This work was supported by Gravida: National Centre for Growth and Development and Health Research Council (HRC) (09/085D), New Zealand.

Conflict of interest: none declared.

REFERENCES

Akalin,A. *et al*. (2012a) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*., **13**, R87.

Akalin,A. *et al.* (2012b) Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet.*, **8**, e1002781.

Baranzini,S.E. *et al.* (2010) Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, **464**, 1351–1356.

Baylin,S. and Bestor,T.H. (2002) Altered methylation patterns in cancer cell genomes: cause or consequence? *Cancer Cell*, **1**, 299–305.

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

Bock,C. *et al.* (2011) Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, **144**, 439–452.

Bock,C. *et al.* (2012) DNA methylation dynamics during *in vivo* differentiation of blood and skin stem cells. *Mol. Cell*, **47**, 633–647.

Campagna,D. *et al.* (2013) PASS-bis: a bisulfite aligner suitable for whole methylome analysis of Illumina and SOLiD reads. *Bioinformatics*, **29**, 268–270.

Chatterjee,A. *et al.* (2012) Technical considerations for reduced representation bisulfite sequencing with multiplexed libraries. *J. Biomed. Biotechnol.*, **2012**, 741542.

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

Chatterjee,A. *et al.* (2013) Mapping the zebrafish brain methylome using reduced representation bisulfite sequencing. *Epigenetics*, **8**, 979–989.

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

Ehrlich,M. and Lacey,M. (2013) DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. *Epigenomics*, **5**, 553–568.

Gertz,J. *et al.* (2011) Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet.*, **7**, e1002228.

Glastad,K.M. *et al.* (2013) Evidence of a conserved functional role for DNA methylation in termites. *Insect Mol. Biol.*, **22**, 143–154.

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

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

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

Ivanov,M. *et al.* (2013) In-solution hybrid capture of bisulfite-converted DNA for targeted bisulfite sequencing of 174 ADME genes. *Nucleic Acids Res.*, **41**, e72.

Jacobs,G.H. *et al.* (2009) Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res.*, **37**, D72–D76.

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

Kunde-Ramamoorthy,G. *et al.* (2014) Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Res.* [Epub ahead of print, doi:10.1093/nar/gkt1325, January 3, 2014].

Law,J.A. and Jacobsen,S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.*, **11**, 204–220.

Li,Y. *et al.* (2010) The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.*, **8**, e1000533.

Lim,J.Q. *et al.* (2012) BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. *Genome Biol.*, **13**, R82.

Press,W.H. *et al.* (1993) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK.

Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Smallwood,S.A. *et al.* (2011) Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat. Genet.*, **43**, 811–814.

Smith,A.D. *et al.* (2009) Updates to the RMAP short-read mapping software. *Bioinformatics*, **25**, 2841–2842.

Steine,E.J. *et al.* (2011) Genes methylated by DNA methyltransferase 3b are similar in mouse intestine and human colon cancer. *J. Clin. Invest.*, **121**, 1748–1752.

Sun,S. *et al.* (2013) MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment. *BMC Bioinformatics*, **14**, 259.

Thorvaldsdottir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform*, **14**, 178–192.

Wang,H.Q. *et al.* (2011) SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, **27**, 225–231.

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

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