Methy1Extract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data [version 2; peer review: 3 approved]

Guillermo Barturen¹,², Antonio Rueda¹,², José L. Oliver¹,², Michael Hackenberg¹,²

¹Dpto. de Genética, Facultad de Ciencias, Universidad de Granada, Granada, 18071, Spain
²Lab. de Bioinformática, Inst. de Biotecnología, Centro de Investigación Biomédica, Granada, 18016, Spain

Abstract

Whole genome methylation profiling at a single cytosine resolution is now feasible due to the advent of high-throughput sequencing techniques together with bisulfite treatment of the DNA. To obtain the methylation value of each individual cytosine, the bisulfite-treated sequence reads are first aligned to a reference genome, and then the profiling of the methylation levels is done from the alignments. A huge effort has been made to quickly and correctly align the reads and many different algorithms and programs to do this have been created. However, the second step is just as crucial and non-trivial, but much less attention has been paid to the final inference of the methylation states. Important error sources do exist, such as sequencing errors, bisulfite failure, clonal reads, and single nucleotide variants.

We developed Methy1Extract, a user friendly tool to: i) generate high quality, whole genome methylation maps and ii) detect sequence variation within the same sample preparation. The program is implemented into a single script and takes into account all major error sources. Methy1Extract detects variation (SNVs – Single Nucleotide Variants) in a similar way to VarScan, a very sensitive method extensively used in SNV and genotype calling based on non-bisulfite-treated reads. The usefulness of Methy1Extract is shown by means of extensive benchmarking based on artificial bisulfite-treated reads and a comparison to a recently published method, called Bis-SNP. Methy1Extract is able to detect SNVs within High-Throughput Sequencing experiments of bisulfite treated DNA at the same time as it generates high quality methylation maps. This simultaneous detection of DNA methylation and sequence variation is crucial for many downstream analyses, for example when deciphering the impact of SNVs on differential methylation. An exclusive feature of Methy1Extract, in comparison with existing software, is the possibility to assess the bisulfite failure in a statistical way. The source code, tutorial and artificial bisulfite datasets are available at http://bioinfo2.ugr.es/MethylExtract/ and http://sourceforge.net/projects/methylextract/, and also permanently accessible from 10.5281/zenodo.7144.
Corresponding authors: José L. Oliver (oliver@ugr.es), Michael Hackenberg (hackenberg@ugr.es)

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Spanish Government [BIO2008-01353 to JLO and BIO2010-20219 to MH], and Basque country ‘AE’ grant (GB).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2014 Barturen G et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

How to cite this article: Barturen G, Rueda A, Oliver JL and Hackenberg M. MethylExtract: High-Quality methylation maps and SNV calling from whole genome bisulfite sequencing data [version 2; peer review: 3 approved] F1000Research 2014, 2:217

https://doi.org/10.12688/f1000research.2-217.v2

Amendments from Version 1

This new version comprises several changes

1) The MethylExtract software was updated to version 1.5 including several important changes: i) compatibility to all Perl versions, ii) BAM files can be read directly (needs samtools installed), iii) several FLAG values can be given in order to suite for paired-end design

2) Several new benchmarking experiments were carried out like suggested by the referees: i) comparison of methylation profiling between MethylExtract and Bis-SNP using relaxed criteria (methylation values are considered as correct if they deviate only by 10% and 20% respectively from the real value), ii) analysis of artificial BS sequencing data for 5x and 35x coverage, iii) new runtime comparison which is based on the exactly same input files, iv) brief and descriptive comparison of results obtained from "real world" data (see comments to the referees)

3) The tutorial was completely revised including a new figure explaining how MethylExtract treats and indicates variant positions in the output files.

See referee reports

Introduction

DNA methylation at the cytosine carbon 5 position (5meC) is an important epigenetic mark in eukaryotic cells that is predominantly found in CpG or CpHpG (H = A,C,T) sequence contexts. Epigenetic modifications at the DNA level play important roles in embryonic development, transcription, chromosome stability, genomic imprinting, and in the silencing of transposons in plants. Furthermore, aberrant methylation is involved in the appearance of several disorders as cancer, immunodeficiency or centromere instability. The methylation pattern along the genome sequence carries biologically relevant information. For example: methylated promoter regions are generally associated with silenced transcription and DNA methylation in the gene body of transcribed genes is often increased. Given these findings, the generation of high quality whole genome methylation maps at a single cytosine resolution is an important step towards the understanding of how DNA methylation is involved in the regulation of gene expression or the generation of a pathologic phenotype. In addition, methylation maps may provide new insights into how the methylation patterns themselves are established.

Several high-throughput techniques have been developed able to generate whole genome methylation maps. In general, the techniques consist of a methylation-sensitive pre-treatment and a read-out step. The pre-treatments generally consist of digestion by methylation-sensitive endonucleases, methyl-sensitive immunoprecipitation or bisulfite conversion, while the read-out of the methylation information is done by hybridization, amplification or sequencing. Recently, several promising techniques have been developed that link the bisulfite conversion with High-Throughput Sequencing (MethylIC-Seq, BS-Seq or RRBS). Briefly, the bisulfite treatment converts un-methylated cytosines into uracil (converted to thymine after PCR amplification) while leaving methylcytosines unconverted. After sequencing the bisulfite-treated genomic DNA, the methylation state can be recovered from the sequence alignments.

Therefore, the methylation profiling from High-Throughput Bisulfite Sequencing data can be divided into two steps: the alignment of the reads, and the read-out of the methylation levels from the alignment. The alignment of bisulfite-treated reads is highly non-trivial due to the reduced sequence complexity given that all cytosines except methylcytosines are converted to thymines. This challenge has been extensively addressed over the last years and several algorithms are available that either align the reads in a 3-letter space or adapt the alignment scoring matrix in order to account for the C/T conversions. Among these algorithms are BSAMP, Bismark, MethylCoder, NGSmethPipe, BS Seeker, Last and BRAT-BWP. Note that some of these tools are not just alignment programs but can, in addition, perform the profiling of the methylation levels such as Bismark and MethylCoder. After alignment, the methylation states can be recovered: C/T mismatches indicate un-methylated cytosines while C/C matches reveal methylcytosines. However, several error sources--like sequencing errors, clonal reads, sequence variation, bisulfite failure and mis-alignments--can lead to a wrong inference of the methylation levels.

For example, C→T or T→C (on converted cytosines) sequencing errors would be incorrectly interpreted as un-methylated or methylated respectively biasing the results towards lower or higher methylation levels. On the other side, bisulfite failures bias the methylation levels only to higher levels; un-methylated cytosines are not converted and therefore detected as methylcytosines. The existence of sequence variation is another important error source that was traditionally disregarded in the data analysis of whole genome bisulfite sequencing (WGBS) experiments. A C/T SNV would be interpreted as un-methylated cytosine. Given that over two thirds of all Single Nucleotide Polymorphisms (SNPs) occur in a CpG context, having two alleles: C/T or G/A, sequence variation needs to be addressed as an important error source. A C/T SNP manifests on the complementary DNA strand as an adenine, while bisulfite deamination does not affect the guanine on the complementary strand (see Figure 1). This fact allows in principle to distinguish between sequence variation and bisulfite conversion and therefore to i) avoid wrong inference of the methylation state due to sequence variation and ii) detect sequence variation in the same sample preparation as the methylation levels. Profiling the methylation levels and the genotype of the sample from one experiment will be a very important step towards “putting the DNA back into methylation”, as the impact and importance of certain DNA sequences on the methylation levels have been recently demonstrated. To our knowledge, the first program that performed a threshold-based detection of sequence variation in bisulfite sequencing experiments was NGSmethPipe. This program detects sequence variation mainly to avoid wrong inference of the methylation level reporting those genome positions in the output. Only recently, the first state-of-the-art SNP calling algorithm based on the Genome Analysis Toolkit (GATK) was implemented to detect both methylation levels and sequence variation at high precision in a single experiment (Bis-SNP). Here we present MethylExtract, a multi-threaded tool for methylation profiling and sequence variation detection from alignments in standard BAM/SAM format. The tool is able to generate high quality methylation maps taking into account SNVs, putative bisulfite failures, reducing also the contribution of sequencing errors by means of the base quality PHRED score. In addition, it detects
sequence variation based on VarScan methodology\textsuperscript{28} reporting all detected SNVs in VCF format\textsuperscript{29}. Therefore, from a single sequencing experiment, MethylExtract obtains both the methylation levels and the sequence variation, which will increase the reliability of downstream analyses\textsuperscript{23}. We confirm its usefulness using extensive artificial BS data and a comparison to Bis-SNP. We show that while its SNV-calling performance is slightly less specific but more sensitive compared to Bis-SNP, MethylExtract performs better in methylation profiling, is easier to use and over twice as fast on a typical whole genome experiment.

Implementation
Scope and workflow
MethylExtract is implemented in Perl and consists of one main script and two auxiliary scripts that are exclusively dedicated to the statistical assessment of the bisulfite error. In general, the program takes standard BAM/SAM file format as input (previously aligned reads) and performs methylation profiling and SNV calling taking into account several error sources like sequencing errors, clonal reads and bisulfite failures. MethylExtract writes two output files. First, the methylation information for each cytosine including the coordinates, sequence context (CG, CHG, CHH), number of methylcytosines, read coverage and mean base quality (PHRED) score. The second output file reports the sequence variation in standard VCF format\textsuperscript{28}.

Frequently, whole genome bisulfite experiments include the estimation of the bisulfite conversion rate through a completely unmethylated genome (lambda phage for example). If the bisulfite conversion rate is known, statistical tests can be applied to infer whether an observed methylation level might be only due to failures of bisulfite conversion. The two auxiliary scripts allow i) estimating the bisulfite conversion rate by mapping the bisulfite-treated reads from the unmethylated genome only and ii) to apply a binomial statistics based test to infer the probability that the “real” methylation value lies within a given interval of the observed value.

Duplicated reads
The PCR step can lead to duplicated (clonal) reads, thus causing a bias in the read coverage. This bias might lead to incorrect inference at positions with allele-specific methylation (genetic imprinting), sequence variation, hemi-methylation, sequencing errors, bisulfite failure or those that are heterogeneous over the cell population. Frequently, the start coordinates of the alignments are used to eliminate duplicates like in SAMtools\textsuperscript{25}, adding a criterion to keep the best read among the duplicates. However, those approaches do not take into account that at a heterozygous locus two reads with the same start coordinate could represent two different alleles, thus not being clonal reads. The same applies for loci with genetic imprinting. The PCR step can lead to duplicated (clonal) reads, thus causing a bias in the read coverage. This bias might lead to incorrect inference at positions with allele-specific methylation (genetic imprinting), sequence variation, hemi-methylation, sequencing errors, bisulfite failure or those that are heterogeneous over the cell population. Frequently, the start coordinates of the alignments are used to eliminate duplicates like in SAMtools\textsuperscript{25}, adding a criterion to keep the best read among the duplicates. However, those approaches do not take into account that at a heterozygous locus two reads with the same start coordinate could represent two different alleles, thus not being clonal reads. The same applies for loci with genetic imprinting.

Figure 1. SNV detection in bisulfite converted reads. Sequence variation can be detected for a cytosine position analyzing the nucleotide frequency at the same position but on the complementary strand. Bisulfite conversion does not affect the guanine on the complementary strand, therefore the presence of any other base (H=A,C,T) might indicate the existence of an SNV. The figure illustrates three different situations: (a) a methylated cytosine in a CpG context without sequence variation (all reads that map to the position independently of the strand carry a cytosine in the corresponding position), (b) a heterozygous SNV (genotype C/T, SNV detected on the ‘+’ strand) and (c) a homozygous SNV (genotype T/T, SNV detected on the ‘−’ strand). The example in b) shows a heterozygous SNV: the 6 reads with A/G mismatch from a total of 11 reads mapping the position indicate a heterozygous variation. Furthermore, we can conclude that the cytosine allele is methylated (7 reads with C/C matches to the ‘−’ strand). The case illustrated in part c), shows 12 reads that show C/T mismatch (‘+’ strand in blue in the upper part). Without looking at the complementary strand, the inference would be a completely unmethylated cytosine. However, the 11 reads that map to the complementary strand show an A/G mismatch at the corresponding position (we would expect guanines in the case of bisulfite conversion). Note that on bisulfite treated datasets only G/A mapped on the ‘+’ strand and C/T on the ‘−’ strand (referred to the ‘+’ strand) can be used for SNV calling purposes. The figure was generated using the UCSC Genome Browser\textsuperscript{41}. 
the existence of sequence variation or putative clonal reads with a sequencing error in at least one read (disregarding mis-alignments). To restrict the impact of sequencing errors we used only the seed region of the read, i.e. the region with the highest quality. The seed is defined as those nucleotides at the 5’ end of the read (first 26 nt by default) that have a higher PHRED score than ‘minQ’.

Note that the two types of methods, the ones that use only the coordinates and our method using the coordinates and the sequence, have advantages and disadvantages. If the sequence differences are considered, biological meaningful information like sequence variation, genetic imprinting or hemi-methylation is maintained; however, our approach will be vulnerable to sequencing errors and bisulfite errors. The default option is to not perform the detection of duplicated reads, and thus any of the publically available tools can be used optionally to remove clonal reads prior to run MethylExtract.

5’ end trimming
The first nucleotides can be removed from the 5’ end of the read (3 bp for the MspI restriction sites of non-directional reduced representation bisulfite sequencing (RRBS) protocol), as also implemented by Bismark14.

Eliminating reads with putatively high bisulfite conversion failure
The bisulfite conversion error probability of un-methylated cytosines is usually below 1% in modern protocols. However, even for such low values, some positions could be incorrectly profiled, i.e. some methylated cytosines are actually unmethylated. MethylExtract implements a method proposed by Lister et al.10 to detect those reads with a high number of un-converted cytosines. By default, it eliminates reads with at least 90% of (presumably) unconverted cytosines in non-CpG contexts (Lister et al. used ≥ 3 methylated non-CpG cytosines). The default threshold is very conservative and only a rather small fraction of reads will be eliminated. Caution is needed if the user knows that the analyzed species (plants) or tissues (e.g. embryonic stem cells) contain an elevated number of unmethylated cytosines in non-CpG contexts. In those cases, this step should be better skipped as otherwise a bias will be introduced into the analysis.

Controlling sequencing errors
Sequencing errors are another important cause of incorrect methylation profiling (and SNV calling). The contribution of the individual bases can be controlled by means of the assigned PHRED score (i.e. an upper limit of sequencing error contribution to the wrongly inferred methylation states). For example, when setting PHRED score ≥ 20, thus accepting bases with a probability < 0.01 to be incorrectly called, the contribution of sequencing errors to the overall error would be less than 1%. By default, MethylExtract sets the minimum PHRED score to 20 (‘minQ’ parameter) which is then used for both methylation profiling and SNV calling (see below on the determination of the default values).

SNVs detection
SNVs are the most disregarded error source in the analysis of whole genome bisulfite sequencing data. Most tools would interpret a C to T substitution as an unmethylated cytosine, although a certain number of them are actually SNVs, and therefore this inference would be wrong. A C/T SNV manifests on the complementary DNA strand as an adenine, while bisulfite deamination does not affect the guanine on the complementary strand10 (Figure 1). The SNVs detection algorithm implemented in MethylExtract is an adaptation of the widely used varScan algorithm35. The main difference compared to SNV calling from non-bisulfite-treated DNA is the reduced amount of sequence information that can be used to detect sequence variation. The bisulfite treatment converts the unmethylated cytosines into thymines, and therefore, at cytosine positions nucleotides that might result from the bisulfite conversion cannot be used to detect sequence variation. For adenine and thymine, both strands can be used like in re-sequencing experiments. The algorithm works as follows: i) filter out positions that are covered by fewer reads than the minimum read depth (‘minDepthSNV’) – by default ‘minDepthSNV’ is set to 1, thus analyzing all positions that are covered by at least one read; ii) calculate the nucleotide frequencies including all base calls that pass the minimum PHRED score threshold (‘minQ’); iii) discard nucleotides with frequencies below a given threshold (‘varFraction’); iv) calculate a p-value for the variant positions (more than two nucleotides above ‘varFraction’) by means of Fisher’s exact test, v) only those positions with a p-value below a given threshold are considered as SNVs (‘maxPval’), and v) the two nucleotides with the highest frequencies are determined as the putative genotype of the sample at this position. Detected sequence variation is reported in VCF format, which can be used as input for SNP-annotation programs12 or VCFtools29.

Statistical assessment of the bisulfite conversion error
Bisulfite conversion failure has been addressed using binomial statistics for the two possible outcomes; methylated and un-methylated13. However, intermediate biologically meaningful states exist like allele specific methylation (with expected methylation levels of 0.5, if both homologous chromosomes have the same sequencing depth), or the reported partial methylation levels10. Therefore, we developed a statistical test for the methylation levels and not for the methylation state previously proposed1034. To apply this test, the user needs to know the bisulfite conversion rate obtained in the experiment. This rate needs to be established using an unmethylated genome (lambda phage, chloroplast, etc). We supply two additional scripts to i) estimate the bisulfite conversion rate using the appropriate experimental data, and ii) associate a p-value, based on binomial statistics, to each of the extracted methylation levels, as well as a procedure to control the false discovery rate15.

In order to calculate a p-value for a given methylation level, we first need to select an interval as we want to calculate the probability that the real methylation level lies within an interval of the observed methylation level. Once the interval is fixed, we can calculate the number of false unmethylated cytosines that would not change the methylation level, e.g. the methylation level would stay within the error interval.

Once we have detected the maximum number of false unmethylated cytosines that would maintain the methylation level within the error interval, we can calculate the p-value by means of the binomial distribution:
\[
p - \text{value} = 1 - \sum_{k=0}^{mc} \binom{mc}{k} p^k (1 - p)^{mc-k}
\]

being: \( p \) the bisulfite error rate, \( mc \) the number of observed methylcytosines at a given position and \( fmc \) the maximum number of allowed false methylcytosines. The \( p - \text{value} \) corresponds then to the probability to find more than \( fmc \) false methylcytosines at this position, e.g. the probability that the real methylation level lies outside the defined error interval.

To illustrate the method, let’s assume that we have a position that is covered by 21 reads with 17 methylcytosines. In this situation, we would have a methylation level of 0.81. If we fix the error interval at 0.1, we could accept up to 2 false methylcytosines. For two false methylcytosines, the methylation level would be \((17 - 2) / 21 = 0.714\) which lies within the error interval of \(0.81 - 0.1 < 0.714\) while 3 false methylcytosines would lead to a methylation level of 0.67 which lies outside the tolerated error interval. Note that the coverage depth of the position (number of reads) does not appear in the equation, but it does to calculate the maximum number of false methylcytosines. In this way, a higher coverage will lead to a higher number of allowed false methylcytosines and therefore to smaller \( p - \text{values} \). Finally, we implemented the Benjamini–Hochberg step-up procedure\(^3\) to control for the false discovery rate in multiple testing. This step can be optionally activated by the user.

**Results**

**General comparison to other available tools**

*MethylExtract* is currently one of the programs with most implemented features related to quality control. Together with *Bis-SNP* it is the only program that detects sequence variation, both to avoid incorrect methylation profiling and to assess the genotype of the used sample. Table 1 shows a comparison of the main features of all programs that allow methylation profiling from aligned reads. Apart from the used method to call the sequence variation, another important difference between *MethylExtract* and *Bis-SNP* is the number of scripts involved to run a full analysis. *Bis-SNP* requires the execution of: i) 3 scripts to sort, add read group tags (required by *GATK* tools) and mark duplicates, ii) 4 scripts to realign the reads and recalibrate the base quality score, iii) 2 scripts to obtain and sort the SNVs and number of methylcytosines and iv) an additional script to calculate the methylation levels on a standard format. In summary, *Bis-SNP* needs 10 different scripts to process reads from bisulfite-treated experiments. On the other hand, *MethylExtract* unifies all analysis steps into a single program which makes it especially suitable for users without a bioinformatics background. Another feature that is currently unique to *MethylExtract* is the possibility to assess the bisulfite failure in a statistical way. In order to achieve this, *MethylExtract* provides an auxiliary script to estimate the bisulfite conversion rate, and a second script that calculates the probability that the observed methylation level lies outside the selected interval of the real methylation level due to bisulfite conversion failures.

**Impact of SNVs on methylation levels**

As mentioned above, sequence variants can lead to incorrect inference of methylation values. Figure 2 illustrates the impact of C/T variation on the methylation values (C/(C+T) ratio) within CpGs contexts. Around 470,000 SNVs within CpG contexts (affecting to 2.08% of the CpG contexts on the genome) covered by at least 10 reads have been detected by *MethylExtract* in Lister’s H1 dataset\(^4\). *Figure 2* shows the methylation levels for non-variant positions (both alleles coincide with the reference) and for the variant sites, both in homozygosis and heterozygosis. The observed distribution of the methylation levels without variation has two maxima close to 0 and 1, which is similar to previous studies\(^4\). However, for heterozygous positions detected by *MethylExtract*, the methylation levels present a local maximum at approximately 0.5 (the T allele

---

**Table 1. Comparison of MethylExtract with different programs for methylation profiling and SNV calling.**

<table>
<thead>
<tr>
<th>FEATURES/SOFTWARE</th>
<th>MethylCoder</th>
<th>BS SEEKER</th>
<th>BRAT-BW</th>
<th>BSMAP/RRBSMAP</th>
<th>Bismark</th>
<th>Bis-SNP</th>
<th>MethylExtract</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input formats</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>Sam</td>
<td>Sam</td>
<td>Bam</td>
<td>Sam/Bam</td>
</tr>
<tr>
<td>5’ Trim</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Bisulfite failure</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Minimum depth</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Base call errors</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SNVs calling</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Methylation output formats</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>bed</td>
<td>vcf, bed, wig</td>
<td>* , bed, wig</td>
</tr>
<tr>
<td>Variation output formats</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>vcf</td>
<td>vcf</td>
<td></td>
</tr>
</tbody>
</table>

* Input formats: input formats used by each software. 5’ trim: allows the trimming of the 5’ end of the reads. Bisulfite failure: implementation of a step to discard reads where the bisulfite might have failed converting the un-methylated cytosines. Minimum depth: allows the user to discard positions with low coverage. Base call errors: discards positions that do not exceed a given minimum PHRED score value. SNVs calling: detects variation that can lead to wrong methylation levels or context estimation. Methylation output formats: available formats for the methylation results. Variation output formats: output formats for the sequence variation results. The asterisk (*) represents a non-standard input or output format, or the impossibility of extracting the methylation ratios from other alignment tools. The dash (-) represents the inexistence of SNV output format, because the software does not allow to detect them.
on one of the parental chromosomes biases the methylation levels to intermediate values, if the C allele is methylated) and a peak at 0 (if the C allele is un-methylated). Finally, for the homozygous positions where both chromosomes present the T allele, most of the methylation values are exactly 0. However, we know that no cytosine exist at those locus in the analyzed sample and therefore these values are incorrect and should be eliminated from the analysis. The incorrectly inferred methylation values for variant positions, both in homozygosis and heterozygosis, stress the need to detect and remove them from the analysis. For example, a CpG position with C→T SNV on both homologous chromosomes is eliminated by MethylExtract, as actually at this position no CpG exists in the sample. Furthermore, MethylExtract outputs the detected genotype of all profiled positions and therefore heterozygotic loci can be detected easily by the user and treated apart if wished.

Methylation profiling and SNV calling quality
MethylExtract implements several quality controls and is among the programs with most implemented features. Main features of MethylExtract are compared in Table 1 to a number of other, widely used programs. The implementation was validated in several ways. MethylExtract takes aligned reads as input and therefore we first compared the methylation profiling quality achieved on artificial bisulfite data when using two different tools for aligning bisulfite-treated reads: NGSmethPipe and Bismark. Next, we quantify the correctly profiled methylation levels and SNVs as a function of the main quality parameters using NGSmethPipe as aligner. Finally the predictive power of MethylExtract to detect methylation levels and sequence variation was compared to Bis-SNP, both in terms of sensitivity and positive predictive value as it was proposed for data-sets for which the number of true negatives tend to be much higher than false positives.

Generation of artificial BS data. For all further comparisons we will use artificial bisulfite data. The usage of this kind of data for benchmarking has the advantage that the true methylation levels and genotypes are known for each position, which is not true when using other experimental methods like microarrays as a golden standard. Artificial sequencing data has been used before in other studies assessing the SNV prediction quality of different algorithms.

Figure 2. Distribution of C/(C+T) ratios for cytosines within the CpG context in the H1 cell line. C/(C+T) values for cytosines at non-variant and variant (homo- and heterozygotic) positions were shown. The minimum read coverage was set to 10 reads.
To generate the artificial bisulfite data we used DNemulator\textsuperscript{18}. We obtained two datasets from the human contig GL000022.1 (11.2Mb), one with all CpGs completely methylated, and the other one with all CpGs completely unmethylated. DNemulator allows also simulating the genotypes of a diploid genome by introducing the sequence variation from a set of confirmed SNPs (dbSNP135)\textsuperscript{39}. Finally, we simulate a bisulfite conversion rate of 99%. The read quality scores are taken from real experimental data (Lister’s H1 dataset\textsuperscript{30}). All together, we generated artificial bisulfite sequencing datasets at two different coverages; 15x and 20x which corresponds to the coverage usually achieved in whole genome bisulfite sequencing experiments.

**MethylExtract with NGSmethPipe and Bismark input.** NGSmethPipe\textsuperscript{16} is a tool to align bisulfite-treated reads which was developed by our group. It is based on the Bowtie aligner and uses a 3-letter alphabet to map the bisulfite-treated reads. The program implements a preprocessing to improve the mapping accuracy\textsuperscript{18} and an alignment seed extension in order to increase the number of mapped reads. We launched both, NGSmethPipe and the well-established Bismark tool with default options to obtain the SAM/BAM input. Next we used MethylExtract on both input files to obtain the number of covered CpGs and the number of correctly recovered methylation values. Note that we know the correct methylation value for each position due to the use of artificial bisulfite data. A position is considered as correctly profiled, only if the obtained methylation value is identical to the real value. Figure 3 shows the result of this comparison. It shows that the obtained CpG coverage and number of correctly profiled positions is nearly identical both as a function of read coverage (15x and 20x) and for the methylated and unmethylated input data. The only remarkable difference is that NGSmethPipe leads to a slightly higher CpG coverage at 20x for both data sets. Nevertheless, the main conclusion is that MethylExtract yields nearly identical results for input sets obtained from NGSmethPipe and Bismark.

**Analysis of the MethylExtract quality parameters.** Next, we aimed to assess the impact of certain quality parameters implemented in

![Figure 3. CpGs methylation profiling comparison for alignment methods.](image-url)

The results obtained from MethylExtract (correctly profiled methylation values and CpG coverage) using two bisulfite short read aligners, NGSmethPipe and Bismark are compared. The results are nearly independent of the used alignment algorithm.
MethylExtract on the methylation profiling and SNV calling capacity. To detect sequence variation, MethylExtract relies on two main parameters, i) the relative nucleotide frequencies (‘varFraction’) and ii) the corresponding p-value. The ‘varFraction’ parameter determines if a position shows putatively variation: the position is analyzed only if at least one nucleotide that differs from the reference sequence has relative frequencies higher than ‘varFraction’. Only for these positions the corresponding p-value is calculated by means of a Fisher exact test. Figure 4 shows the impact of these parameters on the prediction sensitivity (Sn) and positive predictive value (PPV). Sequence variation is best detected by setting the ‘varFraction’ threshold close to 0.1 (yielding around 91% Sn and only 2% of false positives at a statistical significance of 0.05). If the ‘varFraction’ threshold is increased further, the probability to eliminate heterozygous loci increases steadily for positions with high bias in the read coverage between the two homologous chromosomes. If the p-value threshold is set to 0.01, a small increase in positive predictive value (PPV) is observed, but it causes a strong decrease in sensitivity. Therefore, we determined a ‘varFraction’ of 0.1 and a p-value threshold of 0.05 as the best (default) parameters to detect sequence variations.

The minimum base quality (‘minQ’) and the coverage depth (‘minDepthMeth’ for the methylation profiling) thresholds might be also important parameters to control de quality of methylation profiling and SNV calling. To analyze the impact of the minimum PHRED score parameter (‘minQ’) we fix the minimum read coverage (‘minDepthMeth’) in 3, as suggested by Laurent et al. \cite{40}, ‘varFraction’ = 0.1 and ‘maxPval’ = 0.05 (default values derived above). Figure 5 shows the fraction of correctly profiled methylation values and the PPV for SNVs. It can be seen that the correctly profiled positions increase approximately 31% (from 68% to 99%) and the SNVs around 71% (27% – 98%), when the minimum PHRED score is increased from 0 (all base calls are accepted) to 30 (0.001 error probability). The major difference between the methylated and un-methylated datasets is observed for the profiling of the

**Figure 4.** MethylExtract SNV calling as a function of the minimum relative nucleotide frequency (‘varFraction’). The figures show the sensitivity (Sn) and the positive predictive value (PPV) for SNV detection using two different p-value thresholds. The graphs are based on the methylated (top) and un-methylated (bottom) artificial bisulfite datasets at a mean 20x read coverage.
methylation level for which the percentage increases only from approximately 52% to 86%. The simulated bisulfite conversion failures will affect mainly un-methylated positions which can explain the observed differences. These results confirm that the ‘minQ’ threshold is critical to obtain high quality methylation profiling and genotyping results. The default value was set to 20 as higher values will lead to a coverage reduction compromising the SNV calling sensitivity.

Comparison with Bis-SNP
The comparison between MethylExtract and Bis-SNP needs to be based on identical alignment input files in BAM/SAM format. We obtained these files in a two-step process: First, we trim the input reads as it was done by Lister et al.\textsuperscript{30} and second, we align the bisulfite treated reads to the reference genome using Bismark\textsuperscript{36} with default parameters. Note that we based this comparison on Bismark, as the realignment and recalibration steps implemented in Bis-SNP require the read mapping quality, which is currently not available in NGSmethPipe.

Both methods were used with default parameters. We first compared the detection of sequence variation (SNVs) in terms of Sn and PPV. Figure 6, shows that in general Bis-SNP is more specific (between 1.9% and 3.9% higher PPV), being MethylExtract more sensitive (between 1% and 3.1% higher Sn). This trend can be seen for both artificial bisulfite datasets as well as for both read coverages. However, when comparing the fraction of correctly recovered methylation values, drastic differences can be seen (Figure 7).

---

**Figure 5.** MethylExtract SNV calling and methylation profiling as a function of the base quality. Both graphs show the positive predictive value (PPV) for SNV calling and the fraction of correctly profiled CpG methylation values (methylation profiling) as a function of the minimum base quality (PHRED score parameter ‘minQ’). The graphs are based on the methylated (top) and un-methylated (bottom) artificial bisulfite datasets at a mean 20x read coverage. Y-axis represents SNV PPV, Fraction of correct methylation values and CpG coverage. All of them vary between 0 to 1 therefore being represented together.
Furthermore, when the criteria for correctly profiled methylation values are relaxed, MethylExtract still yields higher fractions than Bis-SNP (Supplementary Figure 1). While Bis-SNP yields a slightly higher number of covered positions (Fraction of covered CpGs), MethylExtract is more specific. In all four comparisons using the stringent criteria (no deviation from the real methylation values is allowed), MethylExtract yields over 20% more correctly profiled positions compared to Bis-SNP. One explanation for this difference might be the PHRED score quality threshold implemented in MethylExtract.

Runtime comparison to Bis-SNP
As mentioned before, only Bis-SNP and MethylExtract perform the detection of SNVs which constitutes an additional CPU demanding task. Therefore, we only compared these two programs in terms of CPU time using a reduced Lister’s H1 dataset on a 24 core Intel(R) Xeon(R) CPU X5650 2.67GHz machine. Available memory is crucial for both methods. In order to not bias the comparison, we limited the available memory to 15GB for both programs allowing up to 15 threads. Both programs were tested using a 11GB BAM input file. After aligning with Bismark, we carried out the entire process for both tools (from the aligned reads to the methylation and SNV profiling). MethylExtract needed 6 hours 2 minutes to process the entire dataset including the sorting by coordinates and the removal of duplicated reads. Bis-SNP spend 2 hours 47 minutes sorting the file and removing putative clonal reads, 9 hours and 36 minutes realigning and recalibrating the reads, and 15 hours 54 minutes genotyping and retrieving the methylation levels. Therefore, it seems that MethylExtract is notably faster than Bis-SNP (approximately 4.5 times on this whole genome data set).

Conclusions
We present a user-friendly tool for methylation profiling and SNV calling in whole genome bisulfite sequencing experiments. MethylExtract takes standardized input formats (BAM/SAM) and writes out likewise broadly used file formats like WIG, BED and VCF. To show its usefulness, we compared it to Bis-SNP, a recently published method that is very similar in scope. Although Bis-SNP is

Figure 6. Comparison of SNV calling between MethylExtract and Bis-SNP. The top graph shows the sensitivity (Sn) and the bottom graph the specificity (PPV) obtained for the methylated and un-methylated artificial bisulfite datasets at two different mean coverages (5x, 15x, 20x and 35x).
more specific (less false positive predictions) in the detection of
SNVs, MethylExtract is more sensitive (higher number of recovered
SNVs). However, the main advantages of MethylExtract when
compared to Bis-SNP seem to rely in the higher percentage of cor-
rectly profiled methylation values, as it reaches values over 20%
higher compared to Bis-SNP. Other aspects that favor MethylExtract
are its user-friendliness (everything is implemented into one script)
and the run-time in comparison to Bis-SNP (over 4 times faster in a
whole genome bisulfite sequencing experiment).

Availability and requirements
MethylExtract is freely available. The source code, the tutorial
and artificial bisulfite datasets can be downloaded from the page
http://bioinfo2.ugr.es/MethylExtract/ and are also permanently

List of abbreviations used
5meC: DNA methylation at cytosine carbon 5 position; SNV; Single
Nucleotide Variation; WGBS: whole genome bisulfite sequencing;
SNP: Single Nucleotide Polymorphism; PPV: positive predictive
value; PHRED score: the quality score to each base call assigned
by the program PHRED; SAM format: Sequence Alignment/Map
format used for storing large nucleotide sequence alignments; BAM
format: the compressed binary version of the SAM format.

Figure 7. Comparison of CpG methylation values between MethylExtract and Bis-SNP. Both methods are compared in terms of fraction
of correctly profiled CpG methylation values (top) and the fraction of recovered CpG positions (bottom).

Author contributions
GB wrote the code and carried out the experiments, AR helped with
the benchmark experiments, and MH, JLO and GB designed the
software and wrote the manuscript. All the authors critically read
and approved the final version.

Competing interests
No competing interests were disclosed.

Grant information
This work was supported by the Spanish Government [BIO2008-
01353 to JLO and BIO2010-20219 to MH], and Basque country
‘AE’ grant (GB).

The funders had no role in study design, data collection and analysis,
decision to publish, or preparation of the manuscript.
Supplementary Figure 1. Methylation profiling comparison *MethylExtract* and *Bis-SNP* using relaxed criterion. Both methods are compared in terms of fraction of correctly profiled CpG methylation values. The upper part of the graph shows the result allowing up to 10% deviation from the real methylation values, while the lower part shows the outcome increasing this range to 20%. The analyses were done for unmethylated and methylated datasets at four different coverages (5x, 15x, 20x and 35x).

**References**


Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 22 May 2014

https://doi.org/10.5256/f1000research.3889.r3824

© 2014 Stadler M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Michael B. Stadler
Friedrich-Miescher Institute for Biomedical Research, Basel, Switzerland

The incompatibility with older Perl versions has been fixed, and the script (v1.7) runs now on RHEL6 (Perl v5.10.1).

I agree that a ROC analysis for the results presented in Fig. 6 would require subsampling of C’s to achieve a better balance between methylated and non-methylated states. It is unfortunate that the authors choose not to do that, because it limits the value of this analysis, however it should not preclude indexing of the article.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 01 April 2014

https://doi.org/10.5256/f1000research.3889.r3823

© 2014 Krueger F. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Felix Krueger
Babraham Bioinformatics, Babraham Institute, Babraham, UK

All my comments were already satisfactorily, or will be addressed in an upcoming version of MethylExtract.

Competing Interests: No competing interests were disclosed.
I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 27 January 2014

https://doi.org/10.5256/f1000research.2616.r3164

© 2014 Krueger F. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Felix Krueger
Babraham Bioinformatics, Babraham Institute, Babraham, UK

General comments
MethylExtract is a new tool that implements several QC related steps on already aligned Bisulfite-Seq data. This includes the deduplication of clonal reads, filtering based on base call quality, identification of potential bisulfite conversion errors and most notably the detection of single nucleotide variants (SNVs) that would affect methylation calls, the latter of which is until now only being performed by Bis-SNP.

After reading the documentation it was straightforward to set MethylExtract off, and it ran to completion in an acceptable time frame (~1d 7h for one lane of HiSeq data aligned against the human genome using Bismark). I liked that the deduplication tool that would potentially allow reads aligning to the same position in the genome to pass if they contain SNVs and thus originate from two distinct alleles (even though it is probably questionable whether one would expect to see this sort of reads a lot for large eukaryotic genomes at moderate read depth such as 15x).

Even though I’ve got a few questions about how certain things are handled or documented in the current implementation of MethylExtract, I was quite impressed at how easy it was to get hold of SNV information using just one command. I am sure MethylExtract will prove a useful tool in the genuine analysis of bisulfite data.

Handling paired-end alignments
The only time I found mention of paired-end reads was in the explanation of the deduplication option ‘delDup’ in the MethylExtract Manual. One can further see that the options ‘tagW’ and ‘tagC’ for FLAG values of reads aligning to the Watson and Crick strands default to 0 and 16, respectively, which are standard FLAG values for forward or reverse mapping single-end reads. If MethylExtract currently only handles single-end experiments this is fair enough, even though it would be a serious limitation seeing that most data we generate is in fact paired-end data. Paired-end reads will have at least four different FLAG values, so it is unclear if and how one would specify these? Would the SNV detection would still work with paired-end reads? Are there options to remove overlapping parts in the reads? In any case, I think paired-end reads should be documented better, both in the manuscript and the manual.

Handling SNV positions
If I understood it correctly, CpG positions with a homozygous C to T SNV are eliminated by MethylExtract before reaching the final methylation output, but does this also happen for heterozygous loci? The manual mentions: "(v) SNV can be detected and removed" - can they or are they?

It seems that MethylExtract reports only a single position (presumably the most 5' one?) for cytosines in CG or CHG context (this could also be mentioned in the manual). Are symmetrical cytosine positions completely eliminated from the output if SNVs are detected on at least one strand? Similarly, is the context of a cytosine determined purely by the genomic sequence or would a homozygous SNV effectively also change a C's context (e.g. would an A to G SNV from 'CAG' to 'CGG' change the context from CHG to CG?). I would find it useful to read some more information on how these cases are dealt with, maybe in the User Guide? Finally, to make it a bit more user-friendly I would welcome an option to include or exclude positions from the output specifically that were detected as homozygous or heterozygous SNVs.

**M-bias**

MethylExtract addresses several important aspects affecting the accuracy of BS-Seq experiments, however it doesn't mention the issue of methylation bias (M-bias) in the reads (described in Hansen et al., 2012). M-bias may have several causes, such as 5' bisulfite conversion failure (described in BSeQC - Lin et al., 2013), fill-in bias during library construction which is especially pronounced for paired-end reads (as an example see read 2 in this report: http://www.bioinformatics.babraham.ac.uk/projects/bismark/PE_report.html) or other technical reasons (e.g. PBAT libraries, an example report is available here: http://www.bioinformatics.babraham.ac.uk/projects/bismark/SE_report.html). Are there any plans to take M-bias into consideration in future versions of MethylExtract?

**5' end trimming**

Both the manuscript and the MethylExtract manual mention that there is an option to ignore bases at the 5’ end of reads, e.g. “3 bp for the MspI restriction sites of reduced representation bisulfite sequencing (RRBS) protocol”. While the option is certainly useful, e.g. for the removal of M-bias in the reads, it should be noted that for standard (= directional) RRBS libraries the first 3bp should reflect the true genomic methylation state of the MspI site and do not have to be ignored. The situation is somewhat different for non-directional reads or when reading through filled-in positions at the 3' end of reads. We have tried to illustrate this in a bit more detail in this brief RRBS guide (http://www.bioinformatics.babraham.ac.uk/projects/bismark/RRBS_Guide.pdf).

**Very minor**

The version of MethylExtract hosted on zenodo.org which I downloaded first was outdated (v1.3) and failed to run at some point. The latest version from http://bioinfo2.ugr.es/MethylExtract/ (v1.4) worked fine.

**Competing Interests:** No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The mapping and calling of cytosine methylation in whole genome bisulfite sequencing is a challenging task. Following sequence alignment, the localisation and scoring of reliable and quantitative positional methylation information requires a number of control functions including the detection and high quality scoring of SNVs. So far BiSNP has been used as the major tool for these tasks. The MethylExtract tool now offers a (slightly) improved software suite compiling state of the art (BiSNP-like) features with additional QCs. One advantage of MethylExtract is that the package can be executed in a single PERL script. Compared to BiSNP, MethylExtract reduces the error rate of false SNV calling by including an optimized PHRED score and controlling for bisulfite conversion error rates. MethylExtract accepts SAM and BAM alignment files and creates an independent SNV output file (VCF). With SAM as an input MethylExtract runs about 2x faster compared to BiSNP. A performance test on 15x and 20x artificial test alignment sets shows a better performance of MethylExtract in comparison to BiSNP but only with respect to specificity, while BiSNP has the better sensitivity. Unfortunately the authors did not run a direct comparison on real datasets. Overall MethylExtract is a nice compilation of surely useful tools for a comprehensive and quality controlled WGBS data analysis. The key features of the package are nicely documented. I only have my doubts that conversion rate errors calculated on spiked in control DNA really generates a meaningful background correction.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The performance of MethylExtract is evaluated using simulated sequence data (completely methylated or completely unmethylated containing sequencing errors and SNVs) and compared to Bis-SNP, a conceptually similar tool that is based on the GATK variant calling package. The simulations cover most important aspects of the tool; however, the paper would benefit from extended simulations and a test on a real dataset. For example, the current evaluation does not cover Bis-seq datasets with very low or very high coverage or intermediate methylation levels (see minor issues below).

The paper is clearly written, and the conclusions are supported by the presented results.

**Major issues:**
- The **MethylExtract perl script** (version 1.3) did not run in my environment (RHEL 6, with Kernel 2.6.32-220.7.1.el6.x86_64, perl 5.10.1 built for x86_64-linux-thread-multi). Trying to run the main script resulted in a compilation error:

  ```
  >perl MethylExtract_1.3.pl
  Type of arg 1 to keys must be hash (not hash element) at MethylExtract_1.3.pl line 318, near "}
  Execution of MethylExtract_1.3.pl aborted due to compilation errors.
  ```

  It is possible that the problem lies in the combination of the script and the test environment. However, the test environment fulfills the stated requirements and dependencies.

- **Figure 6** and corresponding text: A single point comparison of Sn/PPV between MethylExtract and Bis-SNP, both with default parameters, is not very informative. Typically there is a trade-off between sensitivity and specificity which can be influenced by the choice of parameter values such as the score or P value cutoffs. It is possible that with slightly altered parameter values, the improved Sn/reduced PPV of MethylExtract compared to Bis-SNP turn into the opposite. The two tools should therefore be compared using varying parameter settings or cutoffs (altering the trade-off between Sn and PPV) and then relating the resulting specificity and sensitivity in an ROC analysis. The same applies in principle to the results presented in Figure 3.

**Minor issues:**
- **Simulations - readout:** The current evaluation of "correct methylation" requires exact identity of simulated and estimated methylation states. This criterium is very stringent yet may not be able to uncover systematic problems. In practice, a very small deviation from the true methylation level may be tolerable. For illustration: A tool that produces many incorrect values that are off only by a small amount may be preferable to a tool that produces fewer incorrect values that are several-fold off. I would suggest using a continuous measure of performance (e.g. the differences between true and estimated methylation levels) or to allow for a minimum deviation.

- **Simulations - methylation levels:** In the introduction, the authors point out the value of methylation levels as opposed to methylation states. Also, intermediate methylation is present in virtually all real world Bis-seq data sets. The simulations should take this into account and also contain C's with intermediate methylation levels (e.g. around 50% methylation).

- **Simulations - coverage:** The current simulations are performed at 15- and 20 fold coverage and the two yield very similar results. More informative differences in performance may be observed when simulating data at even lower (~5-fold) or higher (>30-fold) coverage, which are commonly found in published Bis-seq datasets.
- **Runtime comparison:** It's surprising that even though MethylExtract supports BAM input, SAM.gz was used for runtime measurements, while Bis-SNP was reading from BAM input. Unpacking of alignments from BAM files is CPU-intensive, and I wonder if MethylExtract would take more time if it was run on the same input as Bis-SNP.

- **Table 1:** MethylExtract is listed to support both SAM and BAM inputs. However, it does not directly read BAM files, but converts them to SAM using SAMtools. Using such a conversion, BSMAP and Bismark also support BAM input. I would suggest not to discriminate between SAM and BAM inputs in the table to avoid confusion based on this subtle difference.

- **Impact of parameter choice when analyzing real world data:** For some parameters (e.g. “duplicated reads filter” and “elimination of bisulfite conversion failure”), it is unclear how they would impact results in a real world analysis. A comparison of the results obtained on an experimental dataset with different parameter values could identify sensitive parameters and guide users when choosing parameters for their own analysis.

- **Figure 5:** The labels of the two y-axes are missing. In addition, CpG coverage (blue line) was probably scaled to be plotted on the same axis; if that is the case, it should be described in the legend and/or indicated in the plot.

**Competing Interests:** No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

### Comments on this article

**Version 1**

**Michael Hackenberg**, Universidad de Granada, Granada, Spain

First of all we want to thank all three referees, Dr. Michael Stadler, Dr. Jörn Walter and Dr. Felix Krueger, for their thorough reviewing of our manuscript. We really appreciate this effort which helped to improve the manuscript and the algorithm. Below we will respond point by point to all raised issues.

**Referee 1 (Michael Stadler)**

**Major issues**

1. **MethylExtract incompatibility with older Perl versions:**

   We were not aware that MethylExtract was incompatible with Perl versions prior to 5.14. We rewrote the affected code and now the program should work with any Perl version. The MethylExtract version (1.4) that avoids this incompatibility was released immediately after Dr. Stadler detected this problem (04/11/2013).
2. **Figure 6 and corresponding text**: "A single point comparison of Sn/PPV between MethylExtract and Bis-SNP, both with default parameters, is not very informative. Typically there is a trade-off between sensitivity and specificity which can be influenced by the choice of parameter values such as the score or P value cutoffs. It is possible that with slightly altered parameter values, the improved Sn/reduced PPV of MethylExtract compared to Bis-SNP turn into the opposite. The two tools should therefore be compared using varying parameter settings or cutoffs (altering the trade-off between Sn and PPV) and then relating the resulting specificity and sensitivity in an ROC analysis. The same applies in principle to the results presented in Figure 3."

In general we totally agree that the best way to analyze the impact of a parameter on the prediction quality is by means of a ROC curve. However, in this particular case a ROC curve is not very informative. This is due to the high number of true negatives (non-variant positions that correctly have not been called). In such a case, methods reach very high Sp, and the ROC curves overlap at the upper left part of the graphic. On the other hand, Sn and PPV cannot be used for a ROC curve as the number of true positives is used in both equations.

We decided therefore to compare the two methods using the default parameters. We think that this is useful as many users might not explore the whole parameter space on their data but use directly the default parameters.

Regarding figure 3: This figure shows the “fraction of correct methylation values”, as the obtained values can only be classified as TP (methylation values correctly inferred) or FP (methylation values incorrectly inferred) which impedes the calculation of a ROC curve.

ROC curve comparing Bis-SNP and MethylExtract performance on SNV detection. The sensitivity and Specificity have been calculated for the 20x coverage simulated dataset.

**Minor issues**

1. **Simulations - readout**: We included now two additional analyses: i) methylation values falling within a range of ±10% of the simulated level are considered as correct and ii) methylation values falling within a range of ±20% of the simulated level are considered as correct. As expected, Bis-SNP improves notably (MethylExtract had less space for improvement). Nevertheless, MethylExtract does still get a higher fraction of correct methylation values, especially at low coverage. These results can be seen at Supplementary Figure 1.

2. **Simulations - Methylation levels**: In principle, the main difference between the methylated and unmethylated set is the impact of the bisulfite failure (no impact on the completely methylated set). That is why we chose to simulate both extreme cases. The result for any intermediate percentage of methylated cytosines must lie in-between these two.

3. **Simulations - Coverage**: We now extended the analysis to 5x and 35x. The general behavior of the programs and the main conclusions drawn from the 15x and 20x experiments remain. However, at 5x the discussed differences between MethylExtract and Bis-SNP are more pronounced when compared to higher coverage.
4. **Runtime comparison:** As suggested, we repeated the runtime measurement using exactly the same input data (BAM). We used a smaller input file extracting 11GB out of the previously used. Briefly, the entire Bis-SNP process took 28 hours and 17 minutes, while MethylExtract finished in 6 hours and 2 minutes. These new results have been included in the runtime comparison section.

5. **Table 1:** Since MethylExtract 1.4, MethylExtract reads BAM files directly.

6. **Impact of parameter choice when analyzing real world data:** We agree that an analysis of the parameters using real data might be interesting. However, we think that this would be largely descriptive as it would be hard to determine which parameter setting yields better results (in absence of a golden standard). Therefore, we opted to use simulation in this manuscript as this is currently the only way to obtain a golden standard.

7. **Figure 5:** Indeed, PPV, fraction of correct methylation values and CpGs coverage are scaled to be plotted in the same axis. A new line has been added to the figure legend, in order to clarify it. (“Y-axis represents SNVs PPV, Fraction of correct methylation values and CpGs coverage (all of them vary between 0 to 1 and are scaled to be represented together)”)

**Referee 2 (Jörn Walter)**

1. "*Unfortunately the authors did not run a direct comparison on real datasets.*"

   In general we think that a comparison on real data would be rather descriptive as the real values (methylation and SNV) are not known for this type of data (see also the response to Dr. Stadler, point 6). However, we applied both tools to chr22 of H1 datasets from Lister *et al.*. We found that MethylExtract yields 8.93% un-methylated and 71.35% methylated CpGs while Bis-SNP obtains 8.44% and 72.10% respectively. Using a minimal coverage of 10 reads, MethylExtract predicts 47,360 SNVs in H1 while Bis-SNP reports 13,496. This corresponds to 1.4 SNV per 1kb for MethylExtract and 0.4 for Bis-SNP being the estimate of the 1000 Genome Project 1.3 for autosomal chromosomes.

2. "*I only have my doubts that conversion rate errors calculated on spiked in control DNA really generates a meaningful background correction*"

   The method implemented in one of the auxiliary scripts was first proposed and used by Lister *et al.*. In theory, it should estimate the conversion error rate correctly if the DNA sequence was really un-methylated.

**Referee 3 (Felix Krueger)**

1. **Handling paired-end alignments:** "*Paired-end reads will have at least four different FLAG values, so it is unclear if and how one would specify these? Would the SNV detection would still work with paired-end reads? In any case, I think paired-end reads should be documented better, both in the manuscript and the manual.*"

   Up to version 1.4, ‘tagW’ and ‘tagC’ options only accepted one FLAG. This limitation indeed excluded the use of bisulfite aligners that use more than 2 FLAG values for pair-end alignments (as Bismark). We removed this drawback in version (1.5) which accepts multiple FLAGs for Watson
and Crick aligned reads (comma-separated FLAGs). This not only improves the pair-end support, but will also allow the user to combine pair and single-end alignments during the methylation and variation profiling step. For example, the user will specify “tagW=99,147 tagC=83,163” for a pair-end reads analysis or “tagW=0,99,147 tagC=16,83,163” for a combined pair and single-end reads analysis (described more thoroughly in the manual). Thanks for pointing out this limitation, which really was a glitch of the software.

2. "Are there options to remove overlapping parts in the reads?"

This is one of the major new features that we want to include in future releases.

3. **Handling SNV positions:** “If I understood it correctly, CpG positions with a homozygous C to T SNV are eliminated by MethylExtract before reaching the final methylation output, but does this also happen for heterozygous loci?"

Yes, such a homozygous SNV will be eliminated from the CpGs output. However the position might appear in the overall output under its “real context”. Please see example 1 on figure 1 in the manual which we added to clarify these situations.

4. "The manual mentions: "(v) SNV can be detected and removed" - can they or are they?"

The sentence has been rewritten to avoid misunderstandings: “SNVs (single nucleotide variants) are detected (the methylation level will be reassigned to the real sequence context found in the sample).”

5. "It seems that MethylExtract reports only a single position (presumably the most 5' one?) for cytosines in CG or CHG context (this could also be mentioned in the manual)."

We rewrote the ‘Output Formats’ section: “POS à methylation context most 5’ position on the Watson strand”.

6. "Are symmetrical cytosine positions completely eliminated from the output if SNVs are detected on at least one strand? Similarly, is the context of a cytosine determined purely by the genomic sequence or would a homozygous SNV effectively also change a C's context (e.g. would an A to G SNV from ‘CAG’ to ‘CGG’ change the context from CHG to CG?). I would find it useful to read some more information on how these cases are dealt with, maybe in the User Guide?"

Yes, all the homozygous SNVs detected will modify the methylation context. CpGs, CpHpGs or CpHpHs methylation levels with SNVs will be included in their “real contexts”. We have included a new section at the end of the manual “How MethylExtract manages SNVs within methylation contexts”, in order to clarify how MethylExtract assigns the methylation context in these cases.

7. "Finally, to make it a bit more user-friendly I would welcome an option to include or exclude positions from the output specifically that were detected as homozygous or heterozygous SNVs."

The methylation output files include the “real contexts” where the methylation has been measured (third column of the output files). For example: YG for C/T SNV in the CG output file, CWG for A/T
SNV in the CHG output file or CG for a CG context without SNVs. We will include the options to filter for real context in the next released version.

8. **M-bias:**

MethylExtract does not implement yet an automatically cutoff for the M-bias (as BSeQC does). The 5' trimming can be used for that purpose, but this automatically cutoff is another major feature that will be include in future versions.

9. **5' end trimming:**

We’ve rewritten it in the manuscript and in the manual, to avoid confusions.

10. **Zenodo version:**

Zenodo version of the software will be updated with the new version of the manuscript.

**Competing Interests:** No competing interests were disclosed.