Nunn *et al. BMC Genomics* (2022) 23:477
https://doi.org/10.1186/s12864-022-08691-6

BMC Genomics

SOFTWARE Open Access

Manipulating base quality scores enables variant calling from bisulfite sequencing alignments using conventional bayesian approaches

Adam Nunn[1,2†], Christian Otto[1†], Mario Fasold[1], Peter F Stadler[2,3,4,5,6,7] and David Langenberger[1*]

Abstract

Background: Calling germline SNP variants from bisulfite-converted sequencing data poses a challenge for conventional software, which have no inherent capability to dissociate true polymorphisms from artificial mutations induced by the chemical treatment. Nevertheless, SNP data is desirable both for genotyping and to understand the DNA methylome in the context of the genetic background. The confounding effect of bisulfite conversion however can be conceptually resolved by observing differences in allele counts on a per-strand basis, whereby artificial mutations are reflected by non-complementary base pairs.

Results: Herein, we present a computational pre-processing approach for adapting sequence alignment data, thus indirectly enabling downstream analysis on a per-strand basis using conventional variant calling software such as GATK or Freebayes. In comparison to specialised tools, the method represents a marked improvement in precision-sensitivity based on high-quality, published benchmark datasets for both human and model plant variants.

Conclusion: The presented "double-masking" procedure represents an open source, easy-to-use method to facilitate accurate variant calling using conventional software, thus negating any dependency on specialised tools and mitigating the need to generate additional, conventional sequencing libraries alongside bisulfite sequencing experiments. The method is available at https://github.com/bio15anu/revelio and an implementation with Freebayes is available at https://github.com/EpiDiverse/SNP

Keywords: Bisulfite sequencing, Genetic variant, SNP, Genotype, DNA methylation, Epigenetics, Epigenomics, Benchmarking

Background

DNA methylation is among the most-studied of the molecular mechanisms involved in epigenetics, and has been associated for example with changes in gene expression [1–3], chromosome interactions [4, 5], and genome stability through the repression of transposable elements [6–8]. It is a base modification most often characterised by the addition of a methyl group to a cytosine nucleotide [9], to form 5-methylcytosine (5mC) or one of its derivatives e.g. 5-hydroxymethylcytosine (5hmC). Cytosine methylation occurs typically in a CG sequence context in eukaryotes [10] but is also prevalent in CHG and CHH contexts (where H is any base but G) in plants [11].

Following the emergence of next-generation sequencing (NGS) technologies, library preparation protocols such as BS-seq [11] and MethylC-seq [12] were devised

†Adam Nunn and Christian Otto contributed equally to this work.
*Correspondence: david.langenberger@ecseq.com
[1]ecSeq Bioinformatics GmbH, Sternwartenstraße 29, 04103 Leipzig, Germany
Full list of author information is available at the end of the article

which facilitate the nucleotide-resolution analysis of DNA methylation patterns through the chemical treatment of sample DNA with sodium bisulfite. The treatment catalyses the deamination of unmethylated cytosines to uracil, while methylated cytosines remain unaffected, to produce non-complementary, single-stranded (ss)DNA. As these strands then undergo PCR, uracil pairs with adenosine rather than the original guanosine during replication, which in turn pairs with thymine in the final, amplified product in place of the original cytosine. The resulting paired-end libraries therefore contain four distinct read-types: the forward (FW) and reverse complement (RC) of the bisulfite-converted sequence from the original Watson(+) strand, and the forward and reverse complement of the bisulfite-converted sequence from the original complementary Crick(-) strand. Mapping such reads to the known genome requires specialised software, but when performed successfully can reveal the underlying extent of DNA methylation over each potential 5mC site by considering the proportion of cytosine matches to thymine mismatches. Evidently, any thymine mismatches arising instead as a result of natural mutation are obscured by bisulfite conversion and risk being mistaken as unmethylated cytosines.

Previous attempts to resolve such confounding positions in the genome, to determine both the correct methylation level and reveal single nucleotide polymorphisms (SNPs), have resulted in the development of specialised software such as BISCUIT (https://github.com/huishenlab/biscuit), Bis-SNP [13], BS-SNPer [14], gemBS [15] and MethylExtract [16]. Each case combines methylation calling and variant calling into a single, concurrent analysis to produce output in a custom variant call format (VCF). No single approach however considers the variant calling itself as a primary, independent outcome. Users looking additionally to leverage SNP data for e.g. genotyping or purposes unrelated to DNA methylation are therefore limited by the scope and rationale behind the development of existing tools. Instead, the present application aims to abstract variant calling as a standalone objective in order to facilitate analysis with conventional software, such as GATK [17], Freebayes [18], or Platypus [19], thereby optimising precision-sensitivity during SNP discovery and allowing users to make the most out of their bisulfite sequencing data for a broader range of purposes.

Under a simple Bayesian framework to variant calling, the conditional probability of observing the true genotype G given the variants observed in the sequencing data D can be represented for example by equation (1), which formulates the problem as the derivation of a prior estimate of the genotype $P(G)$ and the likelihood of observing the data $P(D|G)$.

$$P(G|D) = \frac{P(G)P(D|G)}{\sum_i P(G_i)P(D|G_i)} \quad (1)$$

Given that NGS data is seldom error-free, even the simplest model will typically incorporate base quality (BQ) information directly into the Bayesian inference of genotypes as a fundamental scaling factor for the data likelihood estimation. The BQ score itself is a phred-based quality value which denotes on each position the estimated probability that the base caller identified the correct nucleotide during sequencing. In the context of variant calling from bisulfite-treated NGS data, any potential nucleotide conversions present in the resulting sequencing reads can, in principle, be considered analogous to zero-quality base calls. Leveraging this mechanism imposes an indirect strand-specificity on potential variants which cannot otherwise be dissociated from the effect of bisulfite conversion, dictating that they be informed only by opposite-strand alignments where the original, complementary nucleotide is hence unaffected by the treatment.

Implementation

The method presented herein involves a simple "double-masking" procedure which manipulates specific nucleotides and BQ scores on alignments from bisulfite sequencing libraries (Fig. 1), with the formal procedure on individual alignments described in Algorithm 1. It involves two steps which are performed in silico. First, specific nucleotides in bisulfite contexts are converted to the corresponding reference base, in order to prevent any preselection of sites which are informed exclusively by the artificial bisulfite treatment. This circumvents what can potentially be millions of positions from even being considered by the variant caller as candidate variants for analysis, thus reducing valuable analysis time and conserving computational resources. Second, any given nucleotide which may potentially have arisen due to bisulfite conversion is assigned a BQ score of 0. This drives the variant caller to make the correct decision in regards to genotype on positions where there is real evidence of a SNP. As the procedure is informed by decisions made during alignment, it behaves in exactly the same manner and is applicable to both directional and non-directional sequencing libraries. In paired-end sequencing, the procedure applies in a C>T context on mate 1 alignments to the Watson strand (FW+) and mate 2 alignments to the Crick strand (RC-), whereas mate 1 alignments to the Crick strand (FW-) and mate 2 alignments to the Watson strand (RC+) follow G>A context. Reads obtained from single-end sequencing behave in equivalent manner to mate 1 in paired-end sequencing.

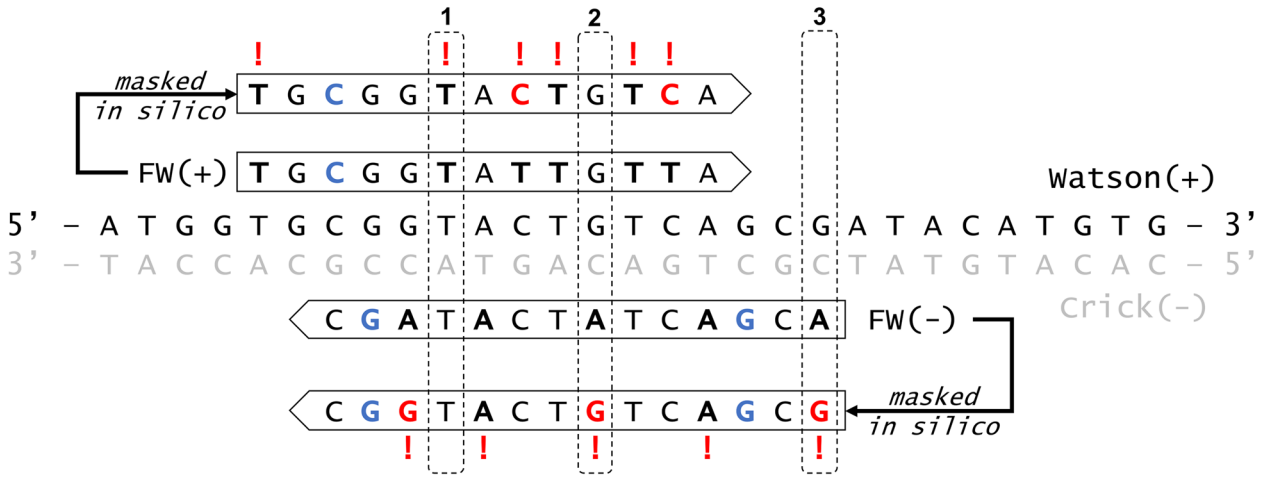

Fig. 1 An overview of the double-masking procedure. The central sequence represents the reference genome, with example alignments (+FW and -FW) adjacent to each originating strand. Black, emboldened nucleotides potentially arise from bisulfite treatment. Blue colouring indicates 5mC/5hmC. Red colouring represents in silico nucleotide manipulation, and corresponding base quality manipulations are indicated with an exclamation mark. In example **(1)** the variant caller is informed only by the -FW alignment, and in **(2)** only by the +FW alignment. As there is no equivalent Watson(+) alignment in **(3)** it is impossible to determine whether the apparent G>A polymorphism arises from bisulfite or a natural mutation

In contrast to previous approaches with bisulfite data, the method is applied as a pre-processing step prior to variant calling, thereby facilitating interoperability with conventional, state-of-the-art variant calling software. For validation, SNPs derived from published, experimental whole genome bisulfite sequencing (WGBS) data in human (NA12878) and *Arabidopsis thaliana* (Cvi-0) accessions are compared to high-quality variant standards and high-confidence regions obtained from the NIST Genome in a Bottle initiative [20] and the 1001 genomes project [21], respectively. The method presented herein has been implemented as a standalone python script available at https://github.com/bio15anu/revelio, which is intended to be adapted and "plugged-in" to any variant pipeline working with bisulfite data so that the user can choose whichever alignment and variant calling software best suits their purposes. An open-source example of a working pipeline for whole genome data is available at https://github.com/EpiDiverse/SNP, which is itself a branch of the EpiDiverse Toolkit [22]. The presented software is also implemented by epiGBS2 in the analysis of reduced-representation bisulfite data [23].

Validation datasets

All datasets analysed in this study are derived from published, public domain resources. High-quality reference variant datasets for human (NA12878) and *A. thaliana* (Cvi-0) accessions were obtained from Genome in a Bottle (GIAB) (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/) and the 1001 genomes project (https://1001genomes.org/data/GMI-MPI/releases/v3.1/), respectively. The corresponding reference genomes GRCh38 (GCF_000001405.26) and TAIR10 (GCF_000001735.3) were obtained from NCBI. Equivalent WGBS data were obtained from the NCBI Sequence Read Archive under accessions SRX3161707 (paired-end, ~46X) and SRX248646 (single-end, ~34X). Please refer to Suzuki et al. [24] and 1001 genomes [21] for further technical specifications regarding these datasets. The original whole genome sequencing (WGS) data for *A. thaliana* Cvi-0 was also obtained, under accession SRX972441 (paired-end, ~62X). Both trimmed reads and alignments from this accession were subject individually to in silico bisulfite treatment (~99% conversion rate), using custom in-house python scripts, to generate corresponding, simulated WGBS datasets.

Read processing and alignment

Reads were assessed with FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and, where appropriate, trimming performed with cutadapt v2.5 [25]. WGS alignments were carried out with BWA v0.7.17-r1188 [26], and WGBS alignments with BWA-meth v0.2.2 [27]. Read groups were merged with SAMtools v1.9 [28], where appropriate, and PCR duplicates subsequently marked with Picard MarkDuplicates v2.21.1 (http://broadinstitute.github.io/picard).

Variant calling

Following the double-masking procedure, variants were called using GATK v3.8 UnifiedGenotyper [17], Freebayes v1.3.1-dirty [18], and Platypus v0.8.1.2 [19], in all cases

Algorithm 1 The double-masking procedure as performed on each alignment.

procedure DOUBLEMASKING(M, W, S) ▷ boolean tests M for Mate 1 and W for Watson strand, and a set S of all aligned base pairs
  $A \leftarrow \varnothing$
  if ($M = true$ **and** $W = true$) **or** ($M = false$ **and** $W = false$) **then**
    $CT \leftarrow true$ ▷ bisulfite conversion in C>T context
  else
    $CT \leftarrow false$ ▷ bisulfite conversion in G>A context
  end if
  for all $U \in S$ **do**
    $(U_0, U_1, U_2) \leftarrow U$ ▷ each aligned pair U is a subset containing the corresponding reference base, query base, and query base quality, respectively
    if $CT = true$ **and** $U_0 = cytosine$ **and** $U_1 = thymine$ **then**
      $U_1 \leftarrow cytosine$
      $U_2 \leftarrow 0$
    else if $CT = false$ **and** $U_0 = guanine$ **and** $U_1 = adenine$ **then**
      $U_1 \leftarrow guanine$
      $U_2 \leftarrow 0$
    else if ($CT = true$ **and** $U_1 = thymine$) **or** ($CT = false$ **and** $U_1 = adenine$) **then**
      $U_2 \leftarrow 0$
    end if
    $A \leftarrow A \cup \{U\}$ ▷ modified or unmodified pair is added to a new alignment set
  end for
  return A
end procedure

with a hard filter of 1 on both minimum mapping quality (MAPQ) and BQ. Variants were called in addition using Platypus on assembly-mode with BQ≥0. For comparison, variants from the original bisulfite alignments were called also with BISCUIT v0.3.16.20200420 (https://github.com/huishenlab/biscuit), Bis-SNP v1.0.1 [13], BS-SNPer v1.1 [14] and MethylExtract v1.9.1 [16]. Default/recommended parameter settings were used, with the exception of minimum MAPQ and BQ thresholds which in all cases were set both to 1. Please refer to Supplementary Table S1 for the complete command line in each case. The resulting variant calls were normalised, decomposed and otherwise processed for comparison to the high-quality reference data using BCFtools v1.9 [28].

Benchmarking

Benchmarking itself was performed with vcfeval of RTG Tools v3.11 [29], which compares both the substitution context and estimated genotype of baseline variants from the truth set to each set of calls from bisulfite data. True positives, false positives and false negatives are evaluated in response to varying common filtering thresholds such as sequencing depth (DP), quality (QUAL) and genotype quality (GQ). Variants must occur with both the same substitution context and genotype in order to be evaluated as a true positive. Sensitivity refers to the true positives as a fraction of the truth set positives, whereas precision refers to the true positives as a fraction of the discovered variants (Supplementary Table S2). The F1 score reflects the balance of precision and sensitivity via the harmonic mean of both measures, and can be optimised relative to each filter by taking the maximum value in response to varying the relevant threshold.

Results

In benchmark data sets for both test species, precision-sensitivity of the SNPs derived from WGBS data is demonstrably improved following double-masking in comparison to existing methods (Fig. 2). Notably, common filtering metrics such as variant quality (QUAL) and genotype quality (GQ) behave as could be expected in conventional sequencing data (Fig. 2; Supplementary Fig. S1), facilitating in many cases the use of established best-practice criteria for selecting high-confidence calls. Additional comparison of SNPs derived from real WGS data (*A. thaliana*; accession Cvi-0) and equivalent WGBS data, following in silico bisulfite conversion (~99%) of sequencing reads, removes the variation caused by differences during sequencing, but not alignment. The resulting ROC-like curves demonstrate a comparable level of sensitivity (i.e. true positives) in both WGS and WGBS data following variant calling with Platypus, Freebayes and

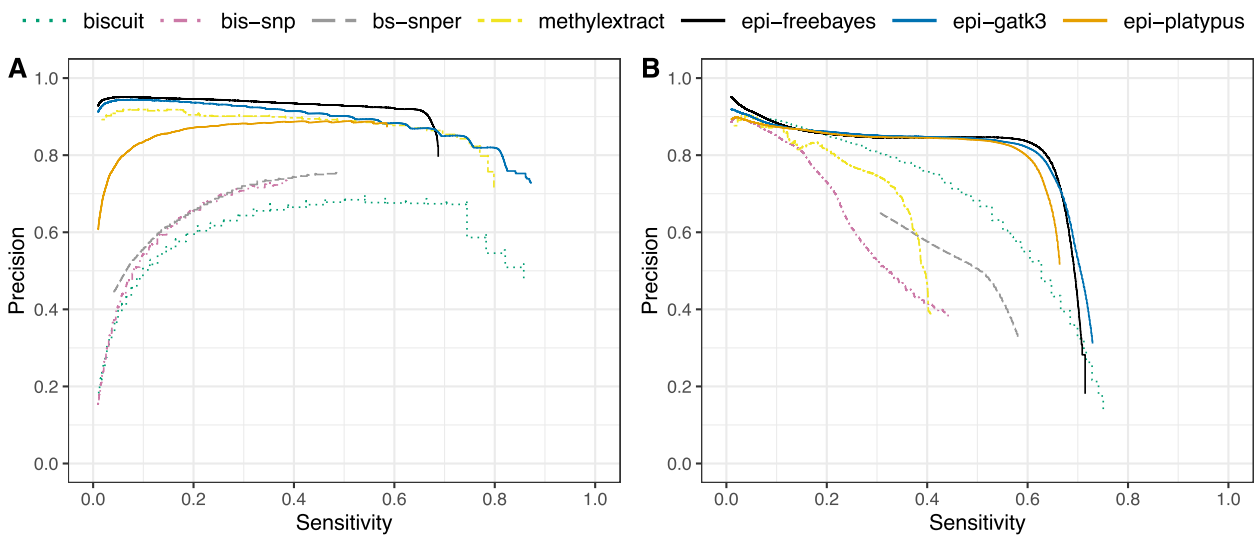

Fig. 2 Precision-sensitivity of variants called in real data. In response to an increasing variant quality (QUAL) threshold, SNPs derived from published WGBS data are compared to those derived from established benchmark datasets for **A** *A. thaliana* (Cvi-0) and **B** human (NA12878). Software with the epi- prefix are intended for conventional DNA sequencing libraries but in this case run after preprocessing with the double-masking procedure. True and false positives are evaluated based on both the substitution context and the estimated genotype

GATK3.8 UnifiedGenotyper (Fig. 3), however there is a drop in precision driven in each case by an influx of false positives. When in silico bisulfite conversion is instead applied directly to the WGS alignments, thus eliminating variation due to the alignment of bisulfite-treated reads, the differences in false positives are reduced for each tool. All software demonstrate an appreciable performance, with GATK3.8 achieving the highest raw number of both true and false positives, followed by Freebayes and then Platypus, for both WGS and WGBS data. The total number of false positives derived from in silico WGBS alignments however represent only 1.0%, 3.8% and 4.3% of the total, unfiltered calls for those same tools respectively, when discounting the fraction shared in the equivalent WGS data.

The overall balance between precision and sensitivity can be evaluated using the harmonic mean, to denote the F1 score, which can be compared between differ-

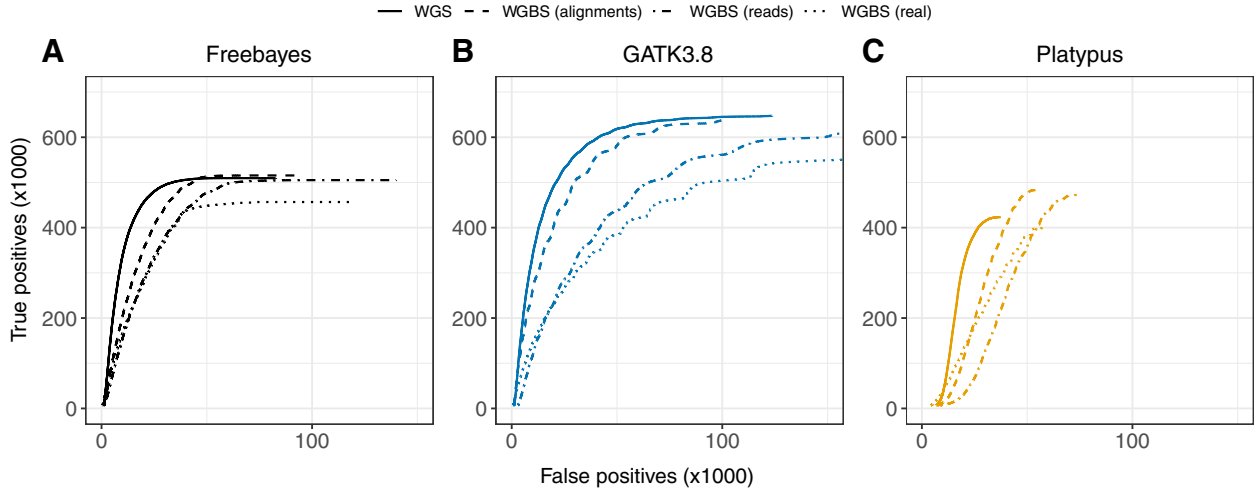

Fig. 3 ROC-like comparisons in real and simulated data. In response to an increasing variant quality (QUAL) threshold, SNPs derived from real WGS data are compared to those derived from equivalent WGBS data after in silico bisulfite conversion of either reads or alignments, followed by preprocessing with the double-masking procedure, in *A. thaliana* (Cvi-0). The real WGBS dataset from Figure 2A is also displayed alongside in each panel for comparison. Panels show results from conventional software **A** Freebayes, **B** GATK3.8 and **C** Platypus (default mode). True and false positives are evaluated based on both the substitution context and the estimated genotype

ent software and data types (Table 1). With in silico WGBS reads, the optimal F1 scores for GATK3.8, Freebayes and Platypus were identified at 0.8508, 0.8039 and 0.7709, respectively, with a corresponding QUAL threshold of 80, 41 and 27. The overall best-performing tool was therefore GATK3.8, achieving 0.8685 sensitivity and 0.8338 precision at the optimal level, followed by Freebayes with 0.7335 sensitivity but a higher precision of 0.8894. Freebayes performed more similarly between the in silico WGBS reads and the real WGBS dataset, however, suggesting it may account better for differences in library composition and layout. Platypus performs better overall in default mode, despite an optimal precision level of 0.9436 for WGS and 0.8991 for WGBS data with assembly-mode enabled (not shown). The reduced overall performance due to lower sensitivity may in-part arise due to the need to set a pre-emptive threshold for Platypus at BQ≥0 (`-minBaseQual=0`), following the double-masking procedure, to avoid over-filtering regions during local assembly.

When considering only those variants called by GATK3.8 UnifiedGenotyper, the relative fraction of true and false positive variants shared between each dataset, before and after filtering according to GATK best-practices (described in Supplementary Table S3), helps to further decompose the factors mainly responsible for the differences observed with WGS and WGBS data (Fig. 4). For example, among the unfiltered true positives the majority of variants are similar and shared between all datasets, with a smaller, secondary, sub-fraction shared only among the real WGS data and both simulated WGBS datasets (paired-end, ∼62X). After filtering, the number of true positive variants are reduced mainly in the real WGBS dataset (single-end, ∼34X), suggesting that variable sequencing library composition is driving these differences. Upon further inspection, the filter on StrandOddsRatio (SOR) appeared to be excluding the majority of true positive variants filtered out in the real WGBS data, likely as a result of an indirect strand-specificity imposed on potential variant calls by the double-masking procedure. When filtering the true positives in the same manner from the real WGBS dataset in the NA12878 human line (Fig. 2B; paired-end, ∼46X), however, these variants were only reduced by ∼13%. With some low-coverage libraries it might therefore be prudent to relax the SOR filter when seeking to obtain confident calls from WGBS data. The false positives, on the other hand, are reflected primarily in the real WGBS dataset and the artificial dataset simulated from real WGS reads (subsequently aligned as a WGBS library). Here, it is the variant confidence metrics (i.e. QUAL and QualByDepth) which are driving the differences after filtering. Taken together this further suggests that the influx of false positives relative to real WGS data are driven primarily by differences in both alignment and library composition, both of which have a direct influence on variant calling.

Table 1 Optimised F1 scores in *A. thaliana* (Cvi-0). In comparison to the reference SNPs obtained from 1001 genomes consortium data, scores are derived when using real WGS and WGBS data, alongside in silico WGBS data derived from the WGS reads and alignments, respectively

	Real data		in silico	
	WGS	WGBS	reads	alignments
GATK3.8	0.9189	0.8177	0.8508	0.9069
Freebayes	0.8247	0.7670	0.8039	0.8247
Platypus (default)	0.7423	0.7026	0.7709	0.7935
Platypus (assembly)	0.6378	0.5980	0.6449	0.6509

This indirect strand-specificity imposed on potential variant calls by the double-masking procedure can be expected to reduce the available sequencing depth required to make confident calls for potential polymorphisms involving thymine, in comparison to WGS data. In the equivalent, in silico WGBS library derived from WGS reads, this would seem to manifest predominately as a relative decrease in variant confidence metrics on true positive SNPs (Supplementary Fig. S2). The number of true positive variants that would fail the recommended hard-filtering thresholds (QUAL<30 or QD<2.0), however, increased only from 1,730 (<0.27%) in WGS data to 9,762 (<1.55%) in the in silico WGBS data. In this simulated, paired-end library there is only a minor increase in overall strand bias, as measured with the SOR metric in GATK3.8 UnifiedGenotyper, where true positive variants that would fail the recommended hard-filtering threshold (SOR>3) increased from 18,045 (2.79%) in WGS data to 31,487 (5.0%) with simulated WGBS data. All together the number of true positive variants lost after hard-filtering increased from 30,858 (4.77%) to 56,695 (9.0%) due to the in silico bisulfite conversion, while the total false positive variants increased from 80,528 (6.81%) to 143,745 (10.24%).

Between all selected variant callers, the proportional deviation of false positives from in silico WGBS reads, relative to WGS data, show similar profiles when partitioned by substitution context (Supplementary Fig. S3). A total of 92.3%, 77.3% and 72.8% of the total false positives here occur in positions which are homozygous-reference in the truth set for each of GATK3.8, Freebayes, and Platypus, respectively, after filtering those shared in the equivalent WGS data. These positions represent 12.0%, 5.6% and 5.6% of the total, unfiltered calls made by each tool. The remaining false positives typically comprise true variants which have been assigned an incorrect genotype (e.g. homozygous-alternative called as heterozygous), representing 2.9%, 4.2% and 4.6% of the total, unfiltered calls.

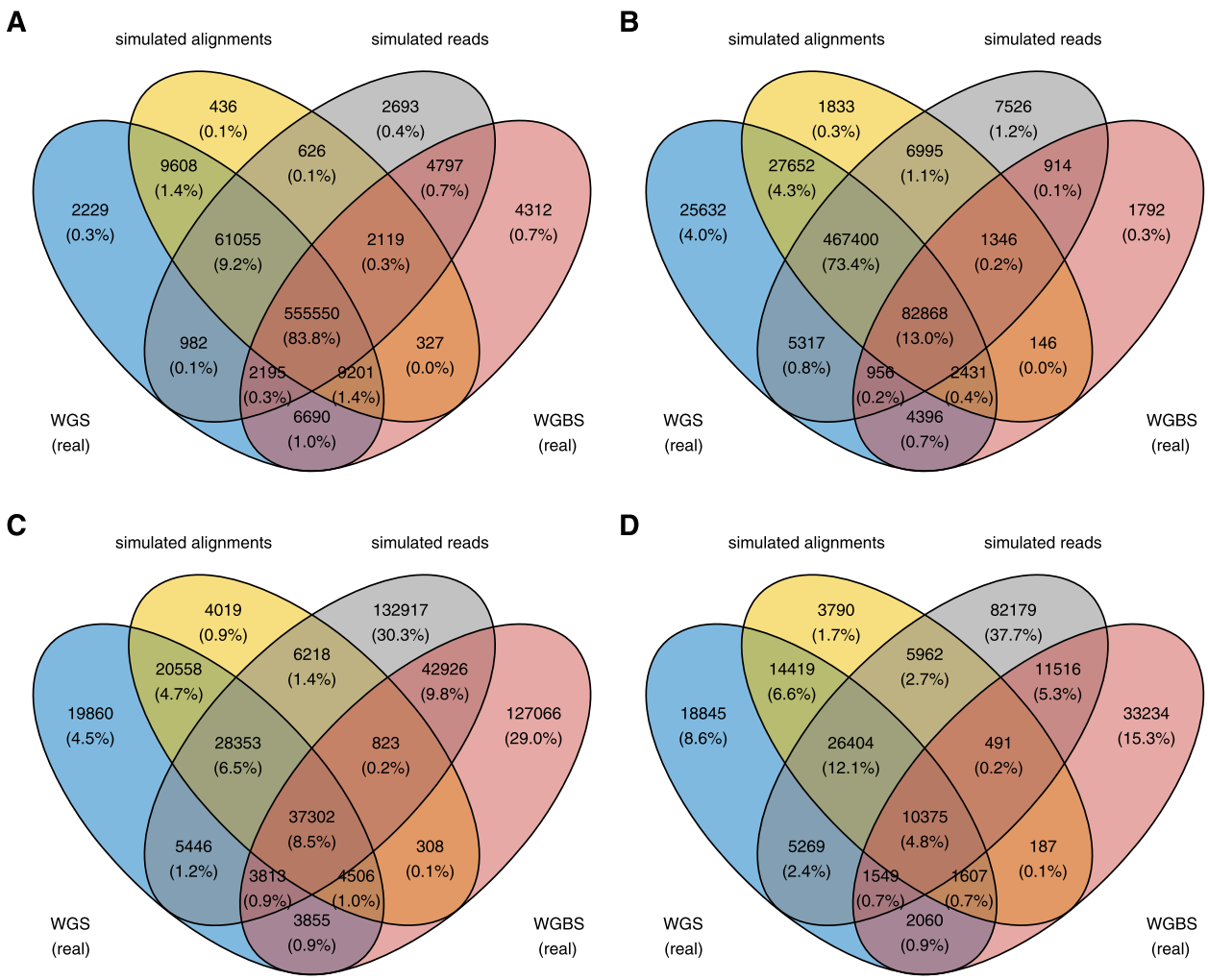

Fig. 4 The shared fraction of true and false positive variants in real and simulated data for *A. thaliana* (Cvi-0), following analysis with GATK UnifiedGenotyper. Distinct WGBS datasets were simulated from both the real WGS alignments and the real WGS reads, separately. The panels denote **A** true positives, before and **B** after filtering, according to recommended hard-filter thresholds in GATK best-practices, and **C** false positives, also before and **D** after filtering. The thresholds chosen for filtering are further described in Supplementary Table S3

Many of these cases suffer a low GQ likely as a consequence of reduced sequencing depth by limiting calls in bisulfite contexts to opposite-strand alignments. Such positions are also considered among the false negatives, alongside the fraction of true SNPs which are not called at all from bisulfite data. When considering the sequencing depth distribution of false negatives from in silico WGBS alignments, discounting those shared in the WGS data, there is a peak at ∼4-5x in addition to a larger peak which correlates with the distribution for the true positives at ∼18-20x (not shown). Accounting for a minimum per-position sequencing depth of ∼7-10x should generally therefore be enough to make a successful call, disregarding differences due to WGBS alignment or significant deviations from typical sequencing biases (e.g. strand bias). More generally, aiming for a genome-wide coverage of at least ∼40X, using a paired-end, directional library, would appear to be the optimal recommendation for analysis based on the complete results of this study.

Discussion

Conventional germline variant callers can be broadly categorised as alignment-based, such as GATK3.8 UnifiedGenotyper, or haplotype-based, such as Freebayes and Platypus. Both strategies are concerned with correctly identifying variants at a given locus and inferring probabilistic genotype likelihoods based on allelic count differences, however they differ in their consideration of proximal variants to establish phase. Whilst UnifiedGenotyper considers precise alignment information in a position-specific, independent manner, Freebayes considers the literal sequence of each overlapping read to obtain the

context of local phasing and derive longer haplotypes for genotyping. Some modern variant callers, including for example Platypus and GATK HaplotypeCaller, expand upon the haplotype-based approach by incorporating local assembly to aid in resolving potential indels. Bisulfite sequencing data can be made conceptually compatible with each of these described approaches, following pre-processing with the double-masking procedure, with the caveat that the chosen software for calling variants handles base quality specifically during the estimation of genotype likelihoods, ideally with an option for hard-filtering. Local assembly presents an added difficulty in that base quality is often considered additionally for read trimming during construction of De Bruijn graphs, e.g. in determination of "ActiveRegions" in GATK HaplotypeCaller, and is typically codependent on the same parameter used for setting its threshold during Bayesian inference. This can sometimes be circumvented, as demonstrated herein with Platypus, by allowing even a base quality of zero during local assembly before relying on the genotype likelihood model to weight such positions appropriately during variant calling, but such a case is not ideal. If masked nucleotides are allowed to be included in the model for deriving genotype likelihoods then the allelic balance on each variant will skew towards any mutations arising from bisulfite conversion, leading to a greater incidence of false positives.

To the best of our knowledge, the software chosen for comparison during this benchmark analysis represent almost the full extent of available, bisulfite-aware variant callers. In one instance a tool had to be omitted for both reasons of compatibility and because we were unable to run the variant calling aspect outside the context of a larger pipeline. gemBS [15] is a pipeline suite which includes mapping, quality control, variant calling and extraction of methylation values. Attempts to run just the variant calling aspect (bs_call) using the standard alignment files generated in this study were unsuccessful, meaning we had to re-run the mapping too with gemBS, thus introducing a discrepancy in comparison to other tools. Furthermore, the variant output was returned in a custom, non-standard VCF format which made it very difficult to separate sequence variants from methylated sites in a manner which was also conducive to a fair, systematic comparison with the other variant calling software. These results were thus omitted so as not to disadvantage gemBS under an experimental design which may simply not be elaborate enough in this case for a fair and robust evaluation of its performance.

Finally, it is important to consider that, unlike most other bisulfite-aware tools, variant calling with the presented approach is almost completely dissociated from the influence of cytosine methylation. The advantage of this is an improved sensitivity for high-confidence variants with fewer false positives, whilst preserving the underlying model of selected tools, but the methylation level itself must be evaluated independently. This is akin to several variant-independent approaches such as MethylDackel (https://github.com/dpryan79/MethylDackel) and GATK MethylationTypeCaller which are commonly used to estimate the methylation level without knowledge of the underlying SNPs. In combination with the presented approach it would be feasible to derive accurate variant-adjusted methylation calls, or even allele-specific methylation without the need for a corresponding genotype dataset obtained by conventional DNA sequencing.

Conclusion

The double-masking procedure facilitates sensitive and accurate variant calling directly from bisulfite sequencing data using software intended for conventional DNA sequencing libraries. The procedure can be readily adapted to existing software pipelines and does not necessitate any additional understanding of customised VCF files. Given sufficient sequencing depth, accurate alignment with minimal deviation from expected sequencing biases, and an appropriate level of filtering based on variant quality metrics, the SNPs derived from WGBS data are comparable to those from WGS data. The method presents a viable, alternative strategy to those who would otherwise need to sequence corresponding libraries of each type in order to better understand the role of DNA methylation in the context of the genetic background.

Availability and requirements

Project name: Revelio
Project home page: http://github.com/bio15anu/revelio
Operating system(s): Linux, MacOS
Programming language: Python (3.8.5)
Other requirements: pysam library (0.16.0.1)
License: MIT license
Any restrictions to use by non-academics: none

Abbreviations

5mC: 5-methylcytosine; 5hmC: 5-hydroxymethylcytosine; NGS: Next generation sequencing; PCR: Polymerase chain reaction; WGS: Whole genome sequencing; WGBS: Whole genome bisulfite sequencing; SNP: Single nucleotide polymorphism; VCF: Variant call format; BQ: Base quality; MAPQ: Mapping quality; QUAL: Variant quality; GQ: Genotype quality; ROC: Receiver operating characteristic; NIST: National Institute of Standards and Technology; GIAB: Genome in a Bottle; GATK: Genome Analysis ToolKit

Acknowledgements

We would like to thank all the members of the EpiDiverse Consortium for their active and invaluable support in discussing, developing and performing these analyses. Special thanks to Morgane Van Antro and Samar Fatma for their assistance in optimising the benchmarking procedure.

Authors' contributions

CO, PFS and DL identified the gap to be addressed by a new method, and all authors were involved in discussions leading to the conception of the presented method. AN and CO further developed the method and conceived the subsequent benchmark analysis. AN and MF implemented a working

software pipeline. AN performed the benchmarking under the oversight of CO, and all authors interpreted the results. AN drafted the manuscript, to which all authors contributed revisions. All authors read and approved the final manuscript.

Funding
The European Training Network "EpiDiverse" received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
All datasets used and/or analysed during the current study are derived from published, public domain resources. The reference genomes GRCh38 (GCF_000001405.26) and TAIR10 (GCF_000001735.3) were obtained from the National Center for Biotechnology Information (NCBI). Corresponding WGBS data from human (NA12878) and *A. thaliana* (Cvi-0) were obtained from the NCBI Sequence Read Archive under accessions SRX3161707 and SRX248646, respectively. The original WGS data for Cvi-0 was also obtained, under accession SRX972441, and the in silico data simulated from this library are available from the corresponding author on reasonable request. The human reference genome `GRCh38_no_alt_plus_hs38d1_analysis_set` and the high-quality benchmark VCF data from GIAB and 1001 genomes were obtained from the following URLs: https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/references/GRCh38/; https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv4.2.1/; https://1001genomes.org/data/GMI-MPI/releases/v3.1/.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]ecSeq Bioinformatics GmbH, Sternwartenstraße 29, 04103 Leipzig, Germany. [2]Department of Computer Science, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany. [3]Interdisciplinary Center for Bioinformatics; German Center for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig; Competence Center for Scalable Data Services and Solutions; Leipzig Research Center for Civilization Diseases; and Leipzig Research Center for Civilization Diseases (LIFE), University of Leipzig, 04109 Leipzig, Germany. [4]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany. [5]Institute for Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria. [6]Facultad de Ciencias, Universidad National de Colombia, Sede Bogotá, Colombia. [7]Santa Fe Institute, Santa Fe, USA.

Received: 24 March 2022 Accepted: 11 June 2022
Published online: 28 June 2022

References

1. Siegfried Z, Eden S, Mendelsohn M, Feng X, Tsuberi BZ, Cedar H. Dna methylation represses transcription in vivo. Nat Genet. 1999;22:203–6.
2. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. Genome Biol. 2011;12:10.
3. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR. Genome-wide high-resolution mapping and functional analysis of dna methylation in arabidopsis. Cell. 2006;126:1189–201.
4. Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, Jacobsen SE. Genome-wide hi-c analyses in wild-type and mutants reveal high-resolution chromatin interactions in arabidopsis. Mol Cell. 2014;55:694–707.
5. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
6. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007;447:425–32.
7. Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O. Selective epigenetic control of retrotransposition in arabidopsis. Nature. 2009;461:427–30.
8. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T. Bursts of retrotransposition reproduced in arabidopsis. Nature. 2009;461:423–6.
9. Chen K, Zhao BS, He C. Nucleic acid modifications in regulation of gene expression. Cell Chem Biol. 2016;23:74–85.
10. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human dna methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.
11. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. Nature. 2008;452:215–9.
12. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in arabidopsis. Cell. 2008;133:523–36.
13. Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-snp: combined dna methylation and snp calling for bisulfite-seq data. Genome Biol. 2012;13:61.
14. Gao S, Zou D, Mao L, Liu H, Song P, Chen Y, Zhao S, Gao C, Li X, Gao Z, Fang X. Bs-snper: Snp calling in bisulfite-seq data. Bioinformatics. 2015;31:4006–8.
15. Merkel A, Fernández-Callejo M, Casals E, Marco-Sola S, Schuyler R, Gut IG, Heath SC. gembs: high throughput processing for dna methylation data from bisulfite sequencing. Bioinformatics. 2019;35:737–42.
16. Barturen G, Rueda A, Oliver JL, Hackenberg M. Methylextract: high-quality methylation maps and snv calling from whole genome bisulfite sequencing data. F1000Research. 2013;2:217.
17. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Res. 2010;20:1297–303.
18. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012. https://arxiv.org/abs/1207.3907.
19. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, WGS500 Consortium, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912–8.
20. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. Nat Biotechnol. 2014;32:246–51.
21. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell. 2016;166:481–91.
22. Nunn A, Can SN, Otto C, Fasold M, Díez Rodríguez B, Fernández-Pozo N, Rensing SA, Stadler PF, Langenberger D. Epidiverse toolkit: a pipeline suite for the analysis of bisulfite sequencing data in ecological plant epigenetics. NAR Genomics Bioinforma. 2021;3:106.
23. Gawehns F, Postuma M, van Antro M, Nunn A, Sepers B, Fatma S, van Gurp TP, Wagemaker NCAM, Mateman C, Milanovic-Ivanovic S, Grosse I, van Oers K, Vergeer P, Verhoeven KJF. epigbs2: Improvements and evaluation of highly multiplexed, epigbs-based reduced representation bisulfite sequencing. Mol Ecol Resour. 2022;3:106.
24. Suzuki M, Liao W, Wos F, Johnston AD, DeGrazia J, Ishii J, Bloom T, Zody MC, Germer S, Greally JM. Whole-genome bisulfite sequencing with improved accuracy and cost. Genome Res. 2018;28:1364–71.
25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–12.
26. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25:1754–60.
27. Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. arXiv. 2014. https://arxiv.org/abs/1401.1129.
28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. Bioinformatics. 2009;25:2078–9.

29. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, Zook JM, Trigg L, De La Vega FM. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. bioRxiv. 2015. https://www.biorxiv.org/content/10.1101/023754.

Publisher's Note