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

Adam Nunn[1,2,†], Christian Otto[1,†], Peter F. Stadler[2] & David Langenberger[1]

[1]*ecSeq Bioinformatics GmbH, Leipzig, 04103, Germany and*

[2]*Institut für Informatik, Universität Leipzig, Leipzig, 04107, Germany*

†Authors contributed equally to this publication

Abstract

Calling germline SNP variants from bisulfite-converted sequencing data poses a challenge for conventional tools, 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 for resolving the interaction between genetic and epigenetic effects when elucidating the DNA methylome. The confounding effect of bisulfite conversion can be resolved by observing differences in allele counts on a per-strand basis. Herein, we present a computational pre-processing approach for adapting such data, thus enabling downstream analysis in this way using conventional variant calling software such as GATK or Freebayes.

Introduction

DNA methylation is among the most-studied of 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 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), which occurs typically in a CG sequence context in eukaryotes[10] but is also prevalent in CHG and CHH contexts 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 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 guanosine during replication, which in turn pairs with thymine in the amplified product of the original sequence 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 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 site by considering the proportion of cyto-

sine matches to thymine mismatches. Evidently, any thymine mismatches arising instead as a result of natural mutation are obscured by the bisulfite-conversion procedure and risk being mistaken as unmethylated cytosines.

Previous attempts to resolve such confounding positions in the genome have resulted in the development of specialised software such as BISCUIT[13], Bis-SNP[14], BS-SNPer[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 tools, 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 bisulfite sequencing data for a broad range of purposes.

The method involves a simple double-masking procedure which manipulates specific nucleotides and base quality (BQ) scores on alignments from bisulfite sequencing data (Fig. 1), prior to variant calling. This imposes an indirect strand-specificity on potential variant calls 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. For validation, SNPs derived from 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.

Results

In both test species, precision-sensitivity of the SNPs derived from equivalent WGBS data is demonstrably improved 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 Figure S1, S2), 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 ($\sim$99%) removes the variation caused by differences during sequencing, but not alignment. The resulting ROC-like curves demonstrate a comparable level of sensitivity (i.e. true positives) in both WGS and WGBS data following variant calling with Platypus, Freebayes and GATK3.8 UnifiedGenotyper (Fig. 3; Supplementary Figure S2), however there is a drop in precision caused in each case by an influx of false positives. Platypus performs better here with assembly-mode disabled, despite an optimised precision level of 0.9436 with WGS and 0.8991 with WGBS data, although following the double-masking procedure it can only run appropriately in assembly mode with BQ$\geq$0 (`--minBaseQual=0`). For each tool the proportional deviation of false positives relative to WGS data show similar profiles when partitioned by substitution context (Supplementary Figure S3), and the majority of false positives occur

in homozygous-reference positions in each of Freebayes (73.4%), GATK3.8 (90.8%) and Platypus (60.5%). While it is indeed possible that low quality variants arise in some loci due to reduced coverage, we surmise that many of these calls pertain also to differences that arise during alignment. The remaining false positives typically comprise true variants which have been assigned an incorrect genotype (e.g. heterozygous called as homozygous-alternative), which represent 4.2%, 2.9% and 4.6% of the total calls made by Freebayes, GATK3.8 and Platypus, respectively.

The indirect strand-specificity imposed on potential variant calls by the double-masking procedure is reflected by a reduction in the available sequencing depth required to make confident calls for potential N>T and N>A polymorphisms, in comparison to WGS data, which manifests predominately as a relative decrease in variant confidence metrics (i.e. QUAL and QualByDepth; described in Supplementary Table S1) on equivalent true positive SNPs (Fig. 4). 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 simulated WGBS data. There is also a minor increase in overall strand bias, as measured with the StrandOddsRatio (SOR) metric in 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 *in silico* bisulfite conversion, while the total false positive variants increased from 80,528 (6.81%) to 143,745 (10.24%).

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 possible 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 for example 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 presented 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 a threshold on base quality during genotype likelihood estimation. This can sometimes be circumvented, as demonstrated herein with Platypus, by allowing even a base quality of zero during assembly and then 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 during genotyping.

It is important to consider that, unlike similar 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 it does mean that the methylation level itself must be evaluated independently. This is akin to several variant-independent approaches such as MethylDackel[22] 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.

Methods

The double-masking procedure involves two steps which are performed *in silico*. First, specific nucleotides in bisulfite contexts are converted to the corresponding reference base in order to obscure likely methylation sites from consideration as potential variants. Second, any given nucleotide which may potentially have arisen due to bisulfite conversion is assigned a base quality (BQ) score of 0. In single-end sequencing, the procedure applies in C>T context on Watson(+) strand alignments and G>A context on Crick(-) strand alignments. In paired-end sequencing, FW+ and RC-

alignments follow C>T context whereas FW- and RC+ alignments follow G>A context.

Validation datasets. High-quality reference variant datasets for human (NA12878) and *A. thaliana* (Cvi-0) accessions were obtained from Genome in a Bottle (GIAB) and the 1001 genomes project, respectively. Equivalent WGBS data were obtained from the NCBI Sequence Read Archive under accessions SRX3161707 and SRX248646. The original whole genome sequencing (WGS) data for *A. thaliana* Cvi-0 was also obtained, under accession SRX972441. Trimmed reads from this accession were subject to *in silico* bisulfite treatment ($\sim$99% conversion rate) with a custom in-house python script to generate a corresponding, simulated WGBS dataset.

Read processing and alignment. Reads were assessed with FastQC[23] v0.11.8 and, where appropriate, trimming performed with cutadapt[24] v2.5. WGS alignments were carried out with BWA[25] v0.7.17-r1188, and WGBS alignments with BWA-meth[26] v0.2.2. Read groups were merged with SAMtools[27] v1.9, where appropriate, and PCR duplicates subsequently marked with Picard[28] MarkDuplicates v2.21.1.

Variant calling. Following the double-masking procedure, variants were called using GATK[17] v3.8 UnifiedGenotyper, Freebayes[18] v1.3.1-dirty, and Platypus[19] v0.8.1.2, in all cases 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$\geq$0. For comparison, variants from the original bisulfite alignments were called also with BISCUIT[13] v0.3.16.20200420, Bis-SNP[14] v1.0.1, BS-SNPer[15] v1.1 and MethylExtract[16] v1.9.1. Default parameter settings were used, with the exception of minimum MAPQ and BQ thresholds which in all cases were set both to 1. The resulting variant

calls were normalised, decomposed and otherwise processed for comparison to the high-quality reference data using BCFtools[27] v1.9.

Benchmarking. Benchmarking itself was performed with vcfeval of RTG Tools[29] v3.11, which compares both the substitution context and estimated genotype of baseline variants from the truth set to each set of calls from bisulfite data in order to evaluate true positives, false positives and false negatives, 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.

1. Siegfried, Z., Eden, S., Mendelsohn, M., Feng, X., Tsuberi, B.Z. & Cedar, H. DNA methylation represses transcription in vivo. *Nature genetics* **22**, 203-206 (1999).

2. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y. & Pritchard, J.K. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome biology* **12**, R10 (2011).

3. Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E. & Ecker, J.R. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* **126**, 1189-1201 (2006).

4. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. & Garg, K. The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

5. Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M. & Jacobsen, S.E. Genome-wide hi-c analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. *Molecular cell* **55**, 694707 (2014).

6. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425432 (2007).

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* **461**, 427430 (2009).

8. Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. & Kakutani, T. Bursts of retrotransposition reproduced in Arabidopsis. *Nature* **461**, 423426 (2009).

9. Chen, K., Zhao, B.S. & He, C. Nucleic acid modifications in regulation of gene expression. *Cell chemical biology* **23**, 74-85 (2016).

10. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B. & Ecker, J.R. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).

11. Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. & Jacobsen, S.E. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452**, 215219 (2008).

12. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. & Ecker, J.R. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523536 (2008).

13. Wanding Zhou. biscuit-0.1.3. (2016) doi:10.5281/zenodo.48262

14. Liu, Y., Siegmund, K.D., Laird, P.W. & Berman, B.P. Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology* **13**, R61 (2012).

15. 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* **31**, 4006-4008 (2015).

16. Barturen, G., Rueda, A., Oliver, J.L. & Hackenberg, M. MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research* **2**, (2013).

17. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303 (2010).

18. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).

19. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., WGS500 Consortium, Wilkie, A.O.M., McVean, G. & Lunter, G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* **46**, 912-918 (2014).

20. Zook, J.M., 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. *Nature biotechnology* **32**, 246-251 (2014).

21. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. *Cell* **166**, 481491 (2016).

22. Ryan, D.P. MethylDackel: A (mostly) universal methylation extractor for BS-seq experiments. Available at https://github.com/dpryan79/MethylDackel (2020).

23. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).

24. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 1012 (2011).

25. Li, H. & Durbin, R. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

26. Pedersen, B.S., Eyring, K., De, S., Yang, I.V. & Schwartz D.A. Fast and accurate alignment of long bisulfite-seq reads. Preprint at https://arxiv.org/abs/1401.1129 (2014).

27. 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* **25**, 2078-2079 (2009).

28. Broad Institute. Picard Toolkit. Available at http://broadinstitute.github.io/picard (2019).

29. Cleary, J.G., Braithwaite, R., Gaastra, K., Hilbush, B.S., Inglis, S., Irvine, S.A., Jackson, A., Littin, R., Rathod, M., Ware, D., Zook, J.M., Trigg, L. & De La Vega, F.M. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. Preprint at https://www.biorxiv.org/content/10.1101/023754v2 (2015).

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.

Funding The European Training Network "EpiDiverse" received funding from the EU Horizon 2020 program under Marie Skłodowska-Curie grant agreement No 764965.

Availability The method has been implemented and is available at https://github.com/EpiDiverse/SNP

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

Correspondence Correspondence and requests for materials should be addressed to David Langenberger (email: david.langenberger@ecseq.com).

FIGURES

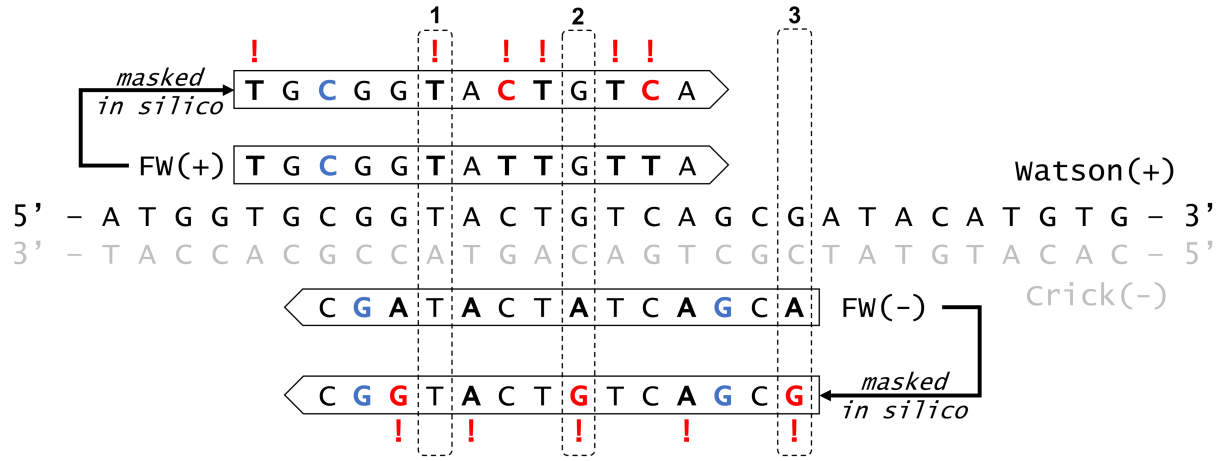

Figure 1: Overview of the double-masking procedure. The central sequence represents the reference genome, with example alignments (+FW and -FW) adjacent to each originating strand. Black, emboldened nucleotides potentially arise from bisulfite treatment. Blue colouring indicates 5mC/5hmC. Red colouring 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.

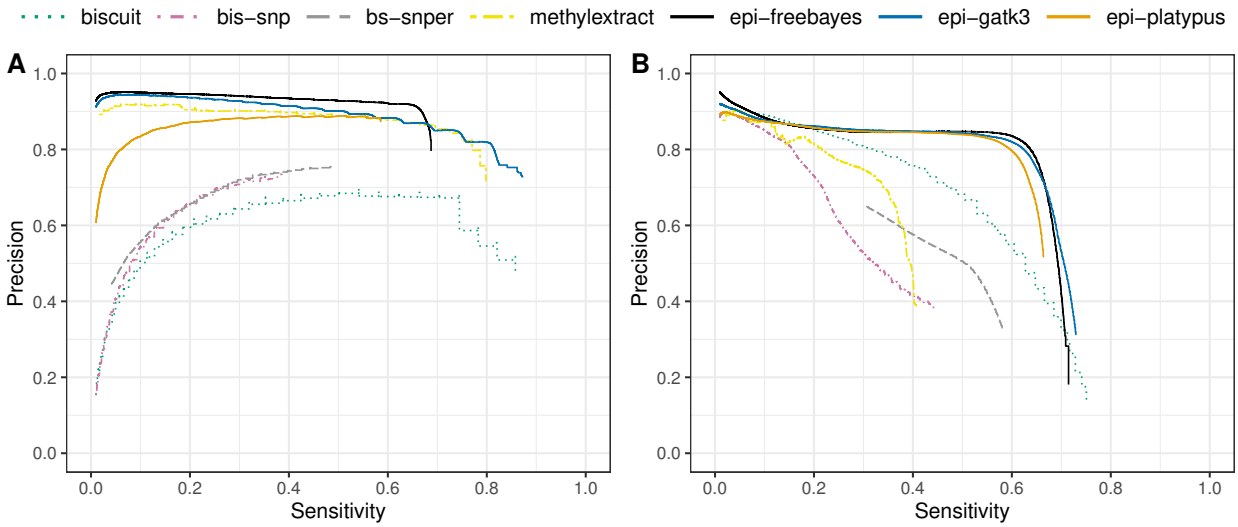

Figure 2: Precision-sensitivity plots derived from established benchmark datasets for **(A)** *A. thaliana* and **(B)** human. Software with the epi- prefix are intended for standard sequencing data but in this case run after preprocessing with the double-masking procedure. With variant quality (QUAL) as a qualifier, epi-freebayes, epi-gatk3 and epi-platypus perform consistently high in terms of the optimal balance of true and false positives, with F1 scores of 0.7670, 0.8117 and 0.7026 in *A. thaliana* and 0.7104, 0.7017 and 0.6854 in human datasets, respectively. Epi-platypus suffers a similar downward trend as other tools in the precision of high-quality variants in *A. thaliana*.

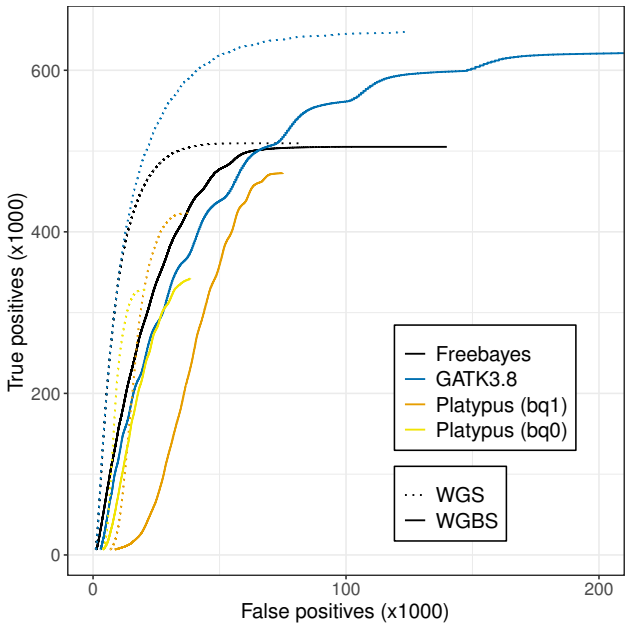

Figure 3: ROC-like plot comparing SNPs derived from real data (WGS) to those from equivalent data after *in silico* bisulfite conversion and preprocessing with the double-masking procedure (WGBS), in *A. thaliana* (Cvi-0). Platypus is run in both standard mode (bq1) and assembly mode (bq0). With variant quality (QUAL) as a qualifier, Platypus (bq0) performs most similarly to the equivalent WGS data and returns the highest overall precision of 0.8991, albeit with the lowest optimised F1 score of 0.6449, whereas GATK3.8 (UnifiedGenotyper) demonstrates the best overall balance of true and false positives on WGBS data with an optimised F1 score of 0.8508.

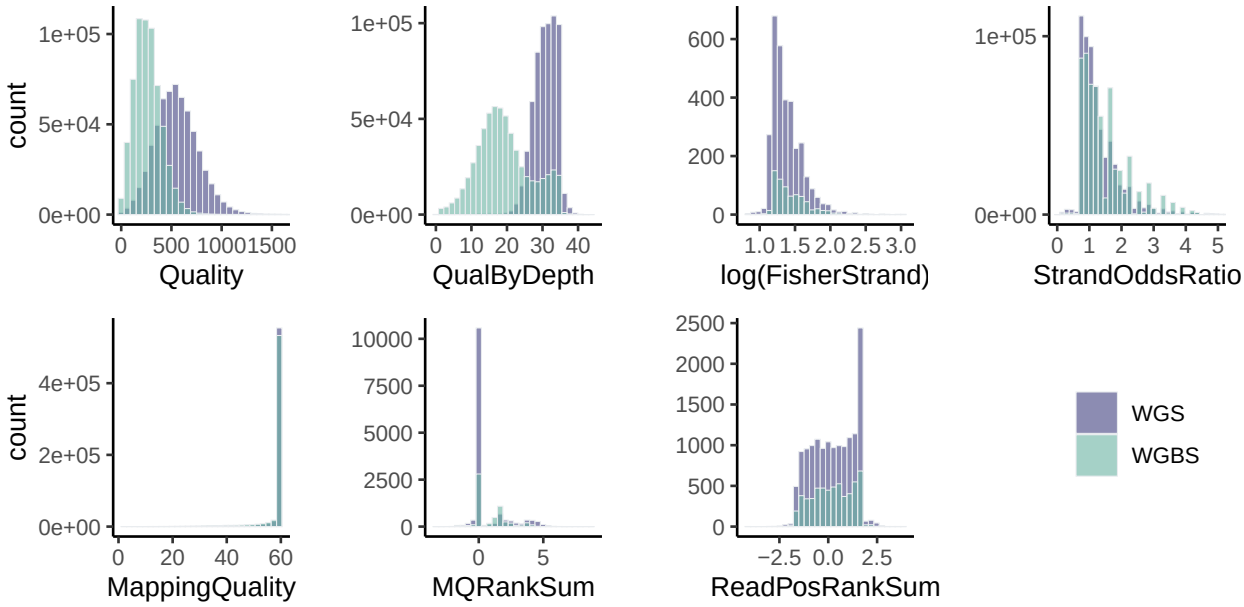

Figure 4: Quality control metric comparisons between SNPs derived from real data (WGS) and those derived from equivalent data after *in silico* bisulfite conversion (WGBS), in *A. thaliana* (Cvi-0). Current definitions for each metric according to GATK are given in Supplementary Table S1. Distributions are derived from the intersection of true positive calls made in each case by GATK3.8 UnifiedGenotyper, and each metric is considered specifically for hard-filtering purposes in GATK best-practices. MQRankSum and ReadPosRankSum are only evaluated for a subset of calls, and in both datasets the vast majority of calls achieve a FisherStrand score of 0 (not shown).