epialleleR: an R/BioC package for sensitive allele-specific methylation analysis in NGS data

Oleksii Nikolaienko[1,*], Per Eystein Lønning[1,2] and Stian Knappskog[1,2]

[1] K. G. Jebsen Center for Genome-Directed Cancer Therapy, Department of Clinical Science, University of Bergen, Bergen, 5021, Norway
[2] Department of Oncology, Haukeland University Hospital, Bergen, 5021, Norway

* To whom correspondence should be addressed. Tel: +47 559 76 444; Email: oleksii.nikolaienko@uib.no

ABSTRACT

Constitutional epigenetic silencing of tumour suppressor genes has been detected in a small number of cancer patients. Recent findings have indicated that low-level mosaic methylation of the *BRCA1* gene promoter occurs in 5–8% of healthy individuals and is associated with a significantly elevated risk of breast and ovarian cancer. This raises the question of whether similar mosaic constitutional methylation may occur in other tumour suppressor genes as well, potentially being a significant contributor to cancer burden. While this suggests a new area for translational research, detection of low-level mosaic epigenetic events requires highly sensitive and robust methodology for methylation analysis. We here present epialleleR, a computational framework for sensitive detection, quantification and visualisation of low-frequency, mosaic epimutations in methylation sequencing data. Analysing simulated and real data sets, we provide in-depth assessments of epialleleR performance, and show that linkage to epihaplotype data allows epialleleR to detect low-frequency methylation events with a sensitivity superior to other tools commonly applied for methylation assessment. The epialleleR is freely available at https://github.com/BBCG/epialleleR and https://bioconductor.org/packages/epialleleR/ as an open source R/Bioconductor package.

INTRODUCTION

Cancer is a major health threat and cause of death worldwide. While the minority of cases are due to highly penetrant germline pathogenic variants (inherited cancers), the majority are considered sporadic cancers with no known germline genetic component.

In addition to genetic aberrations like single nucleotide variants, indels, copy number alterations and rearrangements, cancers are known to harbour epigenetic disturbances, leading to aberrant transcriptional up- and downregulation. Epigenetic aberrations may arise during different stages of carcinogenesis as somatic epimutations (mirroring somatic mutations), or in utero as constitutional normal tissue epimutations. Constitutional epimutations are coined primary epimutations if they are not associated with any cis-acting genetic alteration, or secondary epimutations, in case they are associated with a genetic variant (1). While secondary epimutations in the *MLH1* gene have been related to inherited colorectal cancers (2), and secondary epimutations in *BRCA1* – to ovarian cancer (3), such epimutations have been detected in a few cancer-prone families only. As for primary epimutations, several studies on large cohorts (4, 5) have linked low-level mosaic epimutations to breast and/or ovarian cancer risk. Research and interest in this field, however, has been limited by the fact that all these studies were conducted on patients already diagnosed with their cancers,

questioning whether normal tissue methylation in these patients may be a cancer-initiating event or a secondary effect of the disease itself. Recently we found low-frequency, mosaic methylation within the *BRCA1* gene promoter in healthy women to be associated with a significantly elevated risk for subsequent high-grade ovarian as well as triple-negative breast cancer, in a population based prospective cohort (6). This finding raises the provoking question of whether similar low-level mosaic methylation may affect other tumour suppressor genes, and be associated with an elevated risk of other cancer forms as well. While this opens a new research area related to cancer risk, there are technical issues to account for, as the low frequency of such mosaic epimutations limits the amplitude of observed changes in methylation. Thus, to explore such hypotheses, there is a need for robust and sensitive methylation detection techniques.

Currently, the most widely used methods for DNA methylation profiling are tiling arrays (such as Illumina HumanMethylation450 or HumanMethylationEPIC) and a variety of methylation sequencing techniques (for details see (7)). These methods have different pros and cons: arrays allow genome-wide assessment at a reduced cost, while the sequencing provides additional information on haplotype specificity of DNA methylation. The typical bioinformatic workflows designed to analyse both types of data usually result in sets of beta values (ratio of a count of methylated cytosines to the total sum of methylated and unmethylated bases) for each genomic position covered (8–10). While this approach is suitable for addressing large differences in DNA methylation profiles between two sets of case and control samples, it lacks sensitivity for low-level mosaic methylation detection, as the detection is hindered by sometimes much more common biological variation (11, 12) or technical artefacts (13, 14). Moreover, the lack of haplotype linkage makes such analysis difficult in tiling array-based data sets, and therefore requires nontrivial approaches (15). In contrast, analysis of NGS-based data can provide much higher sensitivity when the base-resolution methylation data is combined with information on allelic belongingness (epihaplotype linkage).

Here, we present a computational framework for sensitive detection and quantification of low-frequency, mosaic epimutations in methylation sequencing data. The provided methods can be used for the discovery of low-frequency epialleles connected to the disease risk (6), as well as for purposes allowing less sensitivity, such as assessments related to treatment response (16, 17), or to the development of treatment resistance (18). Importantly, the framework also allows to connect DNA methylation status with potential underlying cis-factors, such as single-nucleotide variation or mutations within the immediate proximity.

MATERIAL AND METHODS

Next-generation sequencing

White blood cells (WBC) DNA samples from anonymized males (n=88) (19, 20) and human HCT116 DKO methylated DNA control sample (Zymo Research, cat.no. D5014-2) were bisulfite converted, and five DNA fragments, representing promoter regions of five established tumour suppressor genes, were amplified using custom set of primers (*MLH1*, chr3:36993123–36993500; *CDKN2A*, chr9:21974554–21974921; *MGMT*, chr10:129467118–129467477; *CDH1*, chr16:68737102–

68737469; *BRCA1*, chr17:43125171–43125550), indexed, and sequenced similarly to as previously described (6) (GSE201688). The resulting average coverage was 5000x–50000x per amplicon.

Bioinformatic and statistical analyses

Massive parallel sequencing (NGS) reads were mapped/aligned to the GRCh38 human reference genome and the methylation was called using Illumina DRAGEN Bio-IT Platform (v3.9.5) with the following parameters: --methylation-mapping-implementation single-pass, --enable-methylation-calling true, --methylation-generate-cytosine-report false, --methylation-protocol non-directional, --enable-sort false, unless stated otherwise. R software environment for statistical computing (v4.1.2) was used for all downstream statistical analyses.

Cytosine reporting accuracy comparison

Four sets of paired-end sequencing reads (151bp, 50 million read pairs each set) were simulated using Sherman Bisulfite FastQ Read Simulator (21) with the following options: --length 151, --number_of_seqs 50000000, --paired_end, --minfrag 70, --maxfrag 400, --CG_conversion 0, --CH_conversion 99.5 and varying sequencing error rate (--error_rate parameter) of 0%, 0.1%, 0.3% or 0.6%. Human chromosome 19 sequence (GRCh38.p13 NC_000019.10, 58617616 bp, 1105620 forward strand CpGs) was used as a reference genome for read simulation and mapping/alignment due to its highest CpG content across all human chromosomes (22) and in order to maintain optimal balance of analysis speed and base coverage. Each set of reads was then duplicated, all read1 cytosines (C) and read2 guanines (G) in any context in the duplicate sets were replaced with thymines (T) and adenines (A), respectively. Then, duplicate sets were merged with original sets resulting in four sets of reads 100 million pairs each, with the cytosine conversion rate of exactly 50% and about 99.75% in CpG and non-CpG contexts, respectively.

The mapping and alignment of simulated reads was performed using Illumina DRAGEN Bio-IT platform v3.9.5 with the following modification in parameters: --methylation-protocol directional. Methylation reporting was done as described below (Speed comparison section).

Sensitivity comparison on admixed samples

In order to simulate variable methylation levels while maintaining biological heterogeneity of the samples, we selected ten male DNA NGS samples with the lowest methylation levels across all assayed regions, then admixed varying fractions of reads from two random samples and additionally “spiked” certain number of fully methylated reads from methylated DNA control sample. This resulted in 150 samples containing 0%, 0.01%, 0.03%, 0.1%, 0.3% or 1% of methylated reads per sample (25 samples per every category).

Read mapping, alignment, methylation calling, and generation of genome-wide cytosine reports was performed using Illumina DRAGEN Bio-IT platform as described above. Variant epiallele frequency calling was performed using epialleleR::generateCytosineReport with the following parameters: min.mapq=0, min.baseq=0, nthreads=4, threshold.reads=TRUE, report.context="CG".

Barnes-Hut t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis was performed using R package Rtsne v0.15 (23) and matrices of beta or VEF values for all genomic positions of

CpGs with the coverage of at least 1000x and available values for all analysed samples (number of CpGs, n=138).

Sensitivity comparison to methylation array data

Eight additional WBC DNA NGS samples from anonymized males carrying hypermethylated alleles in at least one of the assayed regions were selected, and variant epiallele frequency calling was performed using epialleleR::generateCytosineReport with the following parameters: min.mapq=30, min.baseq=20, nthreads=4, threshold.reads=TRUE, report.context="CG". The same DNA samples were also bisulfite converted using the Zymo EZ DNA Methylation Kit (Zymo Research, cat.no. D5001), and genome-wide methylation levels were assessed using Illumina HumanMethylationEPIC tiling arrays according to manufacturer's instructions. Resulting IDAT files were processed (normalized and annotated) with the minfi Bioconductor package (24) using the preprocessQuantile method with outlier thresholding enabled (GSE201689). For direct comparison, only the CpGs that are covered in all samples by both tiling arrays (p-value of 0) and targeted sequencing (minimum sequencing coverage of 5000x) were retained (*MLH1*, n=10; *CDKN2A*, n=2; *MGMT*, n=4; *CDH1*, n=7; *BRCA1*, n=14). Pairwise sample comparison was performed using t-test with Holm adjustment for multiple comparisons.

Speed comparison

Comparison of processing speed was performed on 29 BAM files containing paired-end alignments and methylation calls derived from bisulfite sequencing of human WBC DNA samples prepared using the following assays: A) amplicon-based sequencing of promoter regions of *BRCA1* gene (n=10 files, 0.12–0.33 million read pairs per file, average coverage of ~20,000x) (6); B) genome-wide capture-based bisulfite sequencing of promoter regions of 283 tumour suppressor genes (n=10 files, 1.11–2.31 million read pairs per file, average coverage of ~60x; and n=3 files, 51.4–73.4 million read pairs per file, average coverage of ~1000x) (25); C) whole-genome bisulfite sequencing (n=6 files, 497–723 million read pairs per file, average coverage of ~60x; epialleleR and Illumina DRAGEN Bio-IT Platform only). The two former data sets (A and B) were generated in-house, while the latter data (C) were obtained from NCBI Sequence Read Archive (GEO/SRA samples GSM3683953/SRX6640720, GSM3683958/SRX6640725, GSM3683965/SRX6640732, GSM3683951/SRX6640718, GSM3683955/SRX6640722, and GSM3683962/SRX6640729) and reported elsewhere (26).

Processing times to produce conventional cytosine reports were recorded as following: Bismark CX methylation reports were created using Bismark v0.22.3 (8) with the following parameters: command bismark_methylation_extractor, --paired-end, --no_overlap, --comprehensive, --gzip, --mbias_off --parallel 8, --cytosine_report, --CX, --buffer_size 64G. As parallel processing was requested, Bismark used up to 24 cores for some of its subtasks.

methylKit CX methylation reports were created using R/BioC package methylKit v1.20.0 (27) with the following parameters: function methylKit::processBismarkAln, minqual=0, mincov=0, save.context=c("CpG","CHG","CHH"), nolap=TRUE and location pointing to the location of a BAM file. Parallel processing is currently not available for methylKit::processBismarkAln.

epialleleR CX methylation reports were created using R/BioC package epialleleR v1.3.5 with the following parameters: function epialleleR::generateCytosineReport, min.mapq=0, min.baseq=0, nthreads=4 (number of HTSlib decompression threads), threshold.reads=FALSE, report.context="CX" and bam pointing to the location of a BAM file. epialleleR methods currently run in a single-threaded mode only but can benefit from additional BAM decompression threads provided by HTSlib.

Bismark, methylKit and epialleleR were tested on the workstation equipped with AMD EPYC 7742 64-core processor, 512GB of memory and the Red Hat Enterprise Linux Server release 7.9 (Developer Toolset 6, GCC v6.3.1), with BAM files retrieved from high-speed (10Gbps) network accessible storage.

DRAGEN CX methylation reports were created using Illumina DRAGEN Bio-IT platform v3.9.5 (Intel Xeon Gold 6126 48-core processor, 256GB of memory and CentOS Linux release 7.5.1804) with the following parameters: --methylation-generate-cytosine-reports=true, --enable-sort=false, --enable-duplicate-marking=false, --methylation-report-only=true and --bam-input pointing to the location of a BAM file. Default number of threads (up to 24) were used for data processing using DRAGEN; BAM files were accessed from local, high-speed NVMe solid state disk.

For Bismark and DRAGEN, elapsed time measurements were stably reproducible, thus processing time was recorded only once for each file. For methylKit and epialleleR, the tests were run five times in sequential random order by means of R package microbenchmark v1.4.9, and the average time was used in comparison to mitigate variability in processing time measurements.

RESULTS

epialleleR implementation

The presence of methylated *BRCA1* alleles in normal tissue (WBC) has been shown qualitatively for 5–8% of adult women (4). However, the associated quantitative changes in DNA methylation at the level of individual CpGs depend on the frequency of methylated epialleles and are often small, and therefore indistinguishable from the background methylation level due to inherent biological (potentially spurious single-base methylation events) and technical (sequencing errors) variance (11). Methylation statuses of neighbouring CpGs are often concordant (28), and such extended epigenetic changes are often associated with a gene expression silencing (29). Therefore, we focused on quantification of methylation events that span over several CpGs, i.e. detection of epihaplotypes. This is possible in NGS-based data sets, while it is not in array-based data where methylation information of different CpG cannot be connected to each other as in haplotype data. Despite the trivial nature of the task and an increasing demand for such analyses, no suitable generic solution was publicly available. We therefore implemented a solution using R software environment for statistical computing (30), a de facto standard for scientific data analysis. The implemented solution, epialleleR, loads methylation call strings and short sequence reads from supplied binary sequence alignment/map (BAM) file, optionally thresholds read pairs according to their average methylation level, and produces methylation reports for individual cytosines as well as genomic regions of interest (Fig. 1). Increasing scale and depth of methylation sequencing experiments impose a requirement on the speed of data processing. Therefore, all time-consuming subtasks were implemented using optimised C/C++

subroutines and, whenever possible, linked to HTSlib, unified C library for high-throughput sequencing data processing (31). The R package epialleleR is freely available at the Bioconductor package repository.

Accuracy analyses

To test the accuracy of methylation reporting for epialleleR versus three other commonly used tools, we simulated large sets of paired-end bisulfite sequencing reads (2x151bp, 100 million read pairs covering human chromosome 19). In contrast to real datasets, simulated data allows to calculate "ground truth" methylation levels for unbiased comparison. Simulation parameters were selected to obtain exact methylation levels of 50% for cytosines in CpG context (n=2211240) and methylation level of approximately 0.25% (bisulfite conversion rate of ~99.75%) for cytosines in CHG and CHH contexts (n=6593900 and 19210572, respectively). In addition to endogenous deamination events (11), bisulfite treatment-induced changes (13) and variation in conversion rates (32), sequencing itself can introduce errors that vary in range depending on assay type and sequencing technology (14). Therefore, we introduced variable level of artificial sequencing errors (0%, 0.1%, 0.3% or 0.6%) and evaluated their effect on the accuracy of reported methylation metrics, applying a selected set of methods (for comparison see Table 1). Analysis revealed that reported values were close to their theoretical expectations for all methods, with the epialleleR maintaining the smallest deviance of reported versus expected methylation beta values for all samples with sequencing errors introduced, possibly owing to read quality-assisted merging of paired reads (Table 2, further details in Supplementary table 1).

Sensitivity analyses

Concordantly methylated alleles (genomic regions with spatially extended / haplotype-specific methylation) are thought to possess high biological importance (6, 18). Spontaneous 5-methyl cytosine (5mC) deamination, sequencing errors, as well as genuine single-nucleotide methylation/demethylation events affect observed background methylation level and can therefore hinder the detection of low-frequency hyper- or hypomethylated alleles. Differences in experimental conditions provide an additional level of variability which can sometimes be tackled by normalisation during postprocessing (24). In contrast to the DNA methylation analysis using tiling arrays (such as Illumina HumanMethylation450 and HumanMethylationEPIC) which report average methylation values at the level of individual cytosines only, next-generation sequencing provides an additional data dimension by linking methylation levels of individual nucleotides within a genomic region covered by a sequencing read (epihaplotypes). However, this information is lost when methylation is assessed and reported without accounting for its allelic distribution. To evaluate the sensitivity of detection for low-frequency monoallelic hypermethylation events in next-generation sequencing data, we simulated an extended set of samples using real, amplicon-based bisulfite sequencing data for human WBC and methylation control DNA samples (the amplicons covered promoter regions of the tumour suppressors *MLH1*, *CDKN2A*, *MGMT*, *CDH1*, and *BRCA1*). Combining real WBC DNA bisulfite sequencing data allowed to introduce sample-to-sample variability although maintaining biologically relevant background methylation levels across sequenced regions, while admixing fully methylated

reads simulated low-frequency, concordant methylation events. Conventional cytosine reports as well as hypermethylated variant epiallele frequency (VEF) reports were prepared and used for unsupervised clustering of samples. Despite quite low overall methylation level of amplified regions (average beta value of 0.014, median of 0.005), analysis based on beta values was not able to discriminate between samples with 0.01%, 0.03%, 0.10%, 0.30% of methylated reads, or no methylated reads added (Fig. 2, left panels). On the other hand, VEF value-based t-SNE analysis resulted in spatially well-separated clusters that corresponded to each level of admixed methylated reads (Fig. 2, right panels).

Tiling arrays, such as Illumina HumanMethylationEPIC, is another widely used, amplification-free method to assess genome-wide DNA methylation for a reduced cost. In order to directly compare the sensitivities of targeted NGS and of the tiling arrays for the detection of low-frequency DNA methylation events, we employed both of the methods to analyse small set of samples (n=8) carrying low-frequency methylation in at least one of the assayed regions (promoter regions of *MLH1*, *CDKN2A*, *MGMT*, *CDH1*, and *BRCA1*). For unbiased comparison, we limited the corresponding data sets to the CpGs sufficiently covered by both techniques. Analysis revealed that VEF values of samples with many hypermethylated alleles (Fig. 3A) differ significantly from VEF values of samples with only a few or no hypermethylated alleles (Fig. 3B). In contrast, only a few significant differences remained when NGS beta values were used for sample comparison (Fig. 3C), while pairwise comparisons based on tiling array beta values did not reveal any significant differences between samples (Fig. 3D).

Sensitivity analyses

Methylation sequencing data produced by contemporary techniques varies in scale and depth and may contain several thousands to billions of single or paired-end reads. To analyse them efficiently, computational methods must be scalable and fast enough for as large as possible range of sample counts or data file sizes. We compared data processing speed for the selected software for methylation reporting across input data coming from various assays: amplicon-based (n=10 samples with a depth of coverage of ~20,000x), genome wide capture-based (n=10 with a depth of coverage of ~60x and n=3 with a depth of coverage of ~1000x) or whole-genome bisulfite sequencing (n=6 with a depth of coverage of ~60x). The obtained results confirm suitability of epialleleR for analysis of data sets of any depth and coverage (Fig. 4, Table 1).

DISCUSSION

While conflicting data has linked low-level mosaic primary constitutional epimutations to cancer risk for more than a decade (1), we have recently obtained firm evidence implicating primary epimutations within the *BRCA1* gene in an elevated risk of incident breast and ovarian cancer (6). The assumption that such epimutations may affect other tumour suppressor genes and, therefore, lead to other cancer forms (1), institutes a new research area with respect to cancer risk. Further, the findings of such epimutations in umbilical cord blood (4) indicate prenatal events of a yet unknown genesis. This creates the need for multidisciplinary studies on the mechanisms of these events and on their effects

in respect to cancer risk, as well as the need for ultrasensitive methods allowing sample assessment at a high scale.

Here, we present the details on a fast, accurate and sensitive method to detect, quantify and visualise epialleles in NGS data. The method shows its superiority versus conventional methods of methylation reporting in detection of low-frequency methylation events. We thoroughly tested it using bisulfite sequencing data; the method, however, can also be applied to analyse and compare data obtained using any methylation sequencing technique, as it is by design less susceptible to variations in conversion efficiency or sequencing quality. The implemented method is fully documented and can be easily used from within the R environment for statistical computing. With the method already revealing its suitability for detection of low-level mosaic methylation events in a large data set (6), we believe it constitutes an optimal tool for assessment of low-level mosaic methylation with respect to risk of cancer as well as other diseases of relevance.

AVAILABILITY

The epialleleR R/Bioconductor package is freely available at https://bioconductor.org/packages/epialleleR/ and https://github.com/BBCG/epialleleR. The R scripts used in this manuscript and the data underlying accuracy and sensitivity analyses are freely available at DataverseNO (https://doi.org/10.18710/2BQTJP). Sensitive data used for the processing speed assessment are available from the authors upon reasonable request (additional amplicon and capture-based bisulfite sequencing data).

ACCESSION NUMBERS

Public data for sensitivity analysis have been deposited at NCBI Gene Expression Omnibus under accession number GSE201690. Public whole-genome bisulfite sequencing data used for the processing speed assessment are available at NCBI Sequencing Read Archive under accession number SRP217135.

SUPPLEMENTARY DATA

Supplementary Data are available online.

FUNDING

This work was supported by the K.G.Jebsen foundation [grant number SKGJ-MED-020 to P.E.L.]; The Norwegian Cancer Society [grant number 190281-2017 to S.K.]; and The Norwegian Research Council [grant number 617344-1 to P.E.L.]. Funding for open access charge: The Norwegian Research Council.

CONFLICT OF INTEREST

P.E.L. has for other projects received research funding from AstraZeneca, Novartis, Pfizer, and Illumina, and honoraria through speaker's bureaux from AstraZeneca, Pierre-Fabre, Roche, AbbVie and Akademikonferens. He has participated in advisory boards for AstraZeneca, Laboratorios and Farmaceuticos Rovi. S.K. has

received research funding for other projects from AstraZeneca, Pfizer, and Illumina, and speaker's bureaux honoraria from AstraZeneca, Pfizer, and Pierre Fabre.

REFERENCES

1. Lønning,P.E., Eikesdal,H.P., Løes,I.M. and Knappskog,S. (2019) Constitutional Mosaic Epimutations – a hidden cause of cancer? *Cell Stress*, **3**, 118–135.

2. Hitchins,M.P., Wong,J.J.L., Suthers,G., Suter,C.M., Martin,D.I.K., Hawkins,N.J. and Ward,R.L. (2007) Inheritance of a cancer-associated MLH1 germ-line epimutation. *N Engl J Med*, **356**, 697–705.

3. Evans,D.G.R., van Veen,E.M., Byers,H.J., Wallace,A.J., Ellingford,J.M., Beaman,G., Santoyo-Lopez,J., Aitman,T.J., Eccles,D.M., Lalloo,F.I., *et al.* (2018) A Dominantly Inherited 5' UTR Variant Causing Methylation-Associated Silencing of BRCA1 as a Cause of Breast and Ovarian Cancer. *Am J Hum Genet*, **103**, 213–220.

4. Lønning,P.E., Berge,E.O., Bjørnslett,M., Minsaas,L., Chrisanthar,R., Høberg-Vetti,H., Dulary,C., Busato,F., Bjørneklett,S., Eriksen,C., *et al.* (2018) White Blood Cell BRCA1 Promoter Methylation Status and Ovarian Cancer Risk. *Ann Intern Med*, **168**, 326–334.

5. Prajzendanc,K., Domagała,P., Hybiak,J., Ryś,J., Huzarski,T., Szwiec,M., Tomiczek-Szwiec,J., Redelbach,W., Sejda,A., Gronwald,J., *et al.* (2020) BRCA1 promoter methylation in peripheral blood is associated with the risk of triple-negative breast cancer. *International Journal of Cancer*, **146**, 1293–1298.

6. Lønning,P., Nikolaienko,O., Pan,K., Kurian,A., Eikesdal,H., Pettinger,M., Anderson,G., Prentice,R., Chlebowski,R. and Knappskog,S. Constitutional BRCA1 methylation and risk of incident triple-negative breast cancer and high-grade serous ovarian cancer. *In review*.

7. Sun,R. and Zhu,P. (2022) Advances in measuring DNA methylation. *Blood Sci*, **4**, 8–15.

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

9. Maksimovic,J., Phipson,B. and Oshlack,A. (2016) A cross-package Bioconductor workflow for analysing methylation array data. *F1000Res*, **5**, 1281.

10. Fortin,J.-P., Triche,T.J.,Jr and Hansen,K.D. (2017) Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, **33**, 558–560.

11. Youk,J., An,Y., Park,S., Lee,J.-K. and Ju,Y.S. (2020) The genome-wide landscape of C:G > T:A polymorphism at the CpG contexts in the human population. *BMC Genomics*, **21**, 270.

12. Gu,J., Stevens,M., Xing,X., Li,D., Zhang,B., Payton,J.E., Oltz,E.M., Jarvis,J.N., Jiang,K., Cicero,T., *et al.* (2016) Mapping of Variable DNA Methylation Across Multiple Cell Types Defines a Dynamic Regulatory Landscape of the Human Genome. *G3 (Bethesda)*, **6**, 973–986.

13. Kint,S., Spiegelaere,W.D., Kesel,J.D., Vandekerckhove,L. and Criekinge,W.V. (2018) Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. *PLOS ONE*, **13**, e0199091.

14. Stoler,N. and Nekrutenko,A. (2021) Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*, **3**, lqab019.

15. Nikolaienko,O., Lønning,P.E. and Knappskog,S. (2022) ramr: an R/Bioconductor package for detection of rare aberrantly methylated regions. *Bioinformatics*, **38**, 133–140.

16. Kondrashova,O., Topp,M., Nesic,K., Lieschke,E., Ho,G.-Y., Harrell,M.I., Zapparoli,G.V., Hadley,A., Holian,R., Boehm,E., *et al.* (2018) Methylation of all BRCA1 copies predicts response to the PARP inhibitor rucaparib in ovarian carcinoma. *Nat Commun*, **9**, 3970.

17. Nesic,K., Kondrashova,O., Hurley,R.M., McGehee,C.D., Vandenberg,C.J., Ho,G.-Y., Lieschke,E., Dall,G., Bound,N., Shield-Artin,K., *et al.* (2021) Acquired RAD51C Promoter Methylation Loss Causes PARP Inhibitor Resistance in High-Grade Serous Ovarian Carcinoma. *Cancer Res*, **81**, 4709–4722.

18. Hurley,R.M., McGehee,C.D., Nesic,K., Correia,C., Weiskittel,T.M., Kelly,R.L., Venkatachalam,A., Hou,X., Pathoulas,N.M., Meng,X.W., *et al.* (2021) Characterization of a RAD51C-silenced high-grade serous ovarian cancer model during development of PARP inhibitor resistance. *NAR Cancer*, **3**, zcab028.

19. Knappskog,S., Bjørnslett,M., Myklebust,L.M., Huijts,P.E.A., Vreeswijk,M.P., Edvardsen,H., Guo,Y., Zhang,X., Yang,M., Ylisaukko-Oja,S.K., *et al.* (2011) The MDM2 promoter SNP285C/309G haplotype diminishes Sp1 transcription factor binding and reduces risk for breast and ovarian cancer in Caucasians. *Cancer Cell*, **19**, 273–282.

20. Knappskog,S., Gansmo,L.B., Romundstad,P., Bjørnslett,M., Trovik,J., Sommerfelt-Pettersen,J., Løkkevik,E., Norwegian Breast Cancer Group trial NBCG VI, Tollenaar,R.A.E.M., Seynaeve,C., *et al.* (2012) MDM2 promoter SNP344T>A (rs1196333) status does not affect cancer risk. *PLoS One*, **7**, e36263.

21. Krueger,F. (2011) Sherman - bisulfite-treated Read FastQ Simulator.

22. Harris,R.A., Raveendran,M., Worley,K.C. and Rogers,J. (2020) Unusual sequence characteristics of human chromosome 19 are conserved across 11 nonhuman primates. *BMC Evol Biol*, **20**, 33.

23. Krijthe,J.H. (2015) Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation.

24. Aryee,M.J., Jaffe,A.E., Corrada-Bravo,H., Ladd-Acosta,C., Feinberg,A.P., Hansen,K.D. and Irizarry,R.A. (2014) Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.

25. Poduval,D.B., Ognedal,E., Sichmanova,Z., Valen,E., Iversen,G.T., Minsaas,L., Lønning,P.E. and Knappskog,S. (2020) Assessment of tumor suppressor promoter methylation in healthy individuals. *Clin Epigenetics*, **12**, 131.

26. Zhou,L., Ng,H.K., Drautz-Moses,D.I., Schuster,S.C., Beck,S., Kim,C., Chambers,J.C. and Loh,M. (2019) Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. *Sci Rep*, **9**, 10383.

27. Akalin,A., Kormaksson,M., Li,S., Garrett-Bakelman,F.E., Figueroa,M.E., Melnick,A. and Mason,C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*, **13**, R87.

28. Huh,I., Wu,X., Park,T. and Yi,S.V. (2019) Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species. *Briefings in Bioinformatics*, **20**, 33–46.

29. Anastasiadi,D., Esteve-Codina,A. and Piferrer,F. (2018) Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. *Epigenetics & Chromatin*, **11**, 37.

30. R Core Team (2022) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

31. Bonfield,J.K., Marshall,J., Danecek,P., Li,H., Ohan,V., Whitwham,A., Keane,T. and Davies,R.M. (2021) HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, **10**, giab007.

32. Sun,Z., Vaisvila,R., Hussong,L.-M., Yan,B., Baum,C., Saleh,L., Samaranayake,M., Guan,S., Dai,N., Corrêa,I.R., *et al.* (2021) Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res*, 10.1101/gr.265306.120.

TABLE AND FIGURES LEGENDS

Table 1. Selected characteristics of software/hardware solutions for methylation reporting.

method	requires reference (genomic) sequence	removes overlaps within read pairs	outputs epiallele frequencies	processing speed, read pairs per second
Bismark	yes	yes (trims read2)	no	40–2,800
methylKit	no	yes (trims read2)	no	9,900–15,400
DRAGEN	yes	yes (trims read2)	no	2,000–183,000
epialleleR	no	yes (base with the highest quality is chosen)	yes	129,000–231,000

Table 2. Selected accuracy metrics (average beta values and their variance) of methylation reporting. Average beta values that are closest to the expected beta values and lowest variance values are shown in bold.

sequencing error rate	method	CHH		CHG		CpG	
		mean	variance	mean	variance	mean	variance
0.00%	DRAGEN	**0.002501**	1.28E-05	**0.002500**	1.27E-05	0.499997	5.96E-07
	Bismark	0.002501	1.34E-05	0.002500	1.33E-05	**0.499997**	**5.92E-07**
	methylKit	0.002503	**1.28E-05**	0.002501	**1.27E-05**	0.499997	5.97E-07
	epialleleR	**0.002501**	1.28E-05	**0.002500**	1.27E-05	0.499997	5.96E-07
0.10%	DRAGEN	0.002661	1.37E-05	0.002662	1.36E-05	0.499835	1.73E-06
	Bismark	0.002646	1.42E-05	0.002648	1.41E-05	0.499850	1.69E-06
	methylKit	0.002662	1.37E-05	0.002664	1.36E-05	0.499835	1.75E-06
	epialleleR	**0.002624**	**1.35E-05**	**0.002626**	**1.34E-05**	**0.499872**	**1.54E-06**

0.30%	DRAGEN	0.002977	1.53E-05	0.002980	1.53E-05	0.499504	3.22E-06
	Bismark	0.002928	1.57E-05	0.002931	1.57E-05	0.499554	3.03E-06
	methylKit	0.002978	1.52E-05	0.002982	1.52E-05	0.499506	3.21E-06
	epialleleR	**0.002857**	**1.47E-05**	**0.002860**	**1.47E-05**	**0.499628**	**2.59E-06**
0.60%	DRAGEN	0.003498	1.79E-05	0.003506	1.78E-05	0.498942	1.29E-05
	Bismark	0.003393	1.81E-05	0.003402	1.81E-05	0.499051	1.25E-05
	methylKit	0.003497	1.78E-05	0.003501	1.78E-05	0.498952	1.28E-05
	epialleleR	**0.003237**	**1.66E-05**	**0.003241**	**1.65E-05**	**0.499222**	**1.14E-05**

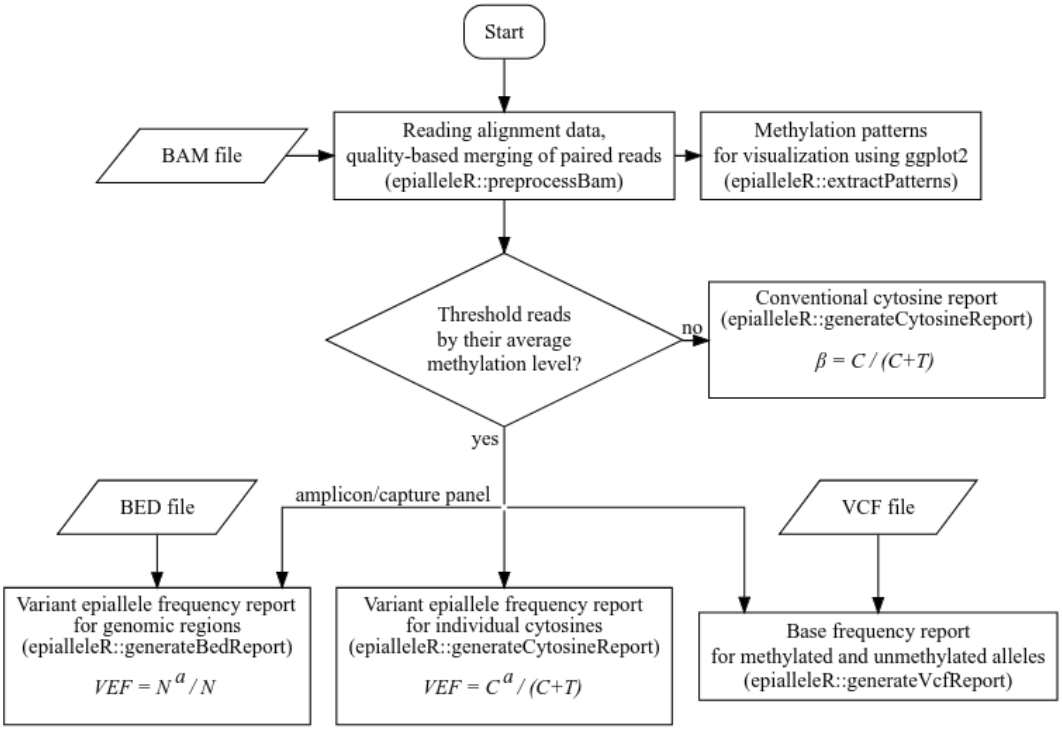

Fig. 1. Flowchart of epialleleR package data processing steps. The formulas using to calculate conventional beta as well as variant epiallele frequency (VEF) values are given in boxes. C and T, total number of cytosines and thymines at particular genomic position, respectively; C^a, number of cytosines in reads passing a particular methylation threshold; N, total number of reads, mapped to a particular genomic region; N^a, number of mapped reads, passing a particular methylation threshold.

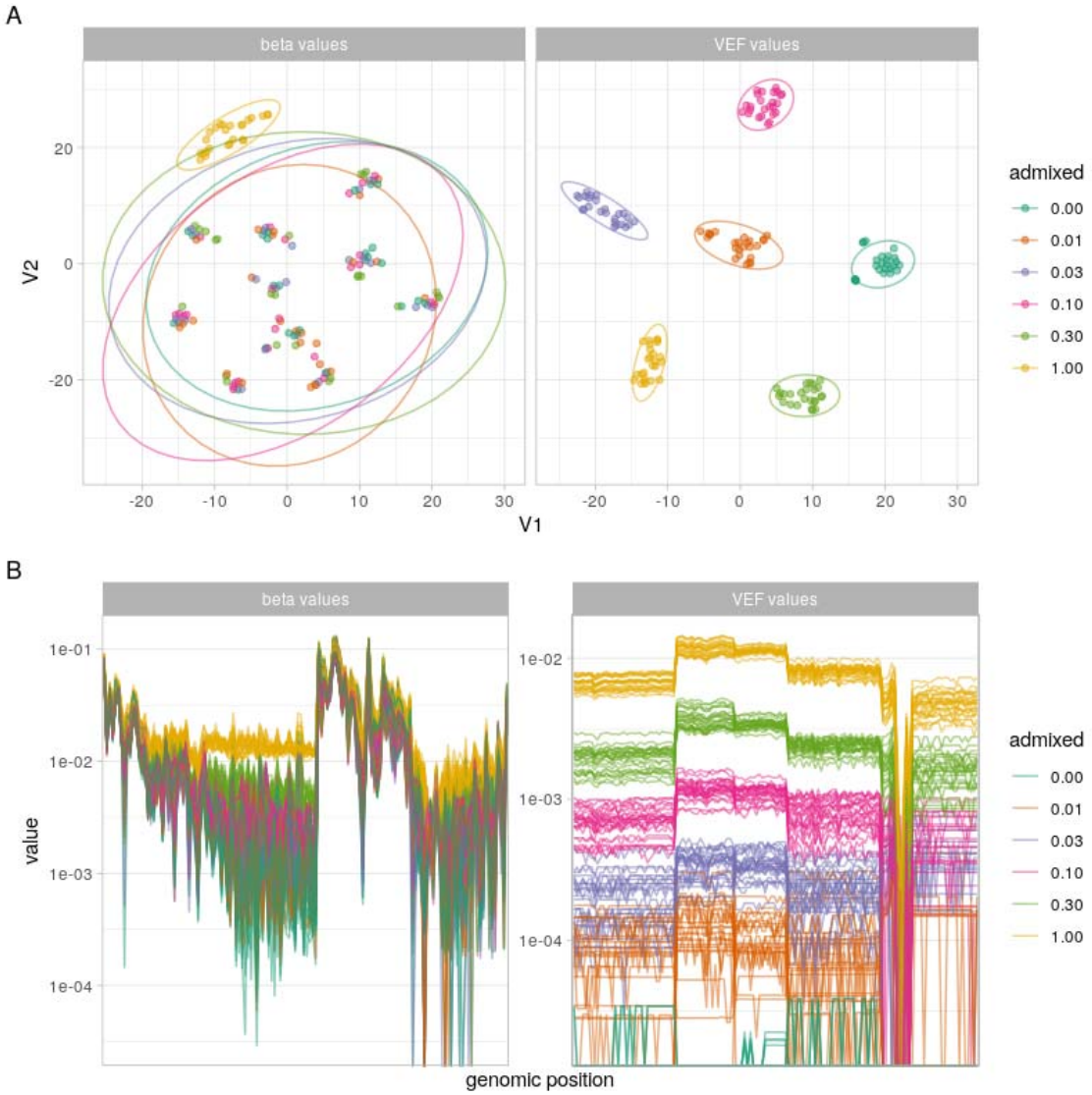

Figure 2. (A) Line plots of beta (left panel) and VEF (right panel) values for individual samples, colour-coded and grouped according to the amount of admixed methylated reads. (B) Embedding plots for t-SNE analysis using beta (left panel) and VEF (right panel) values. Ellipses represent 95% confidence levels.

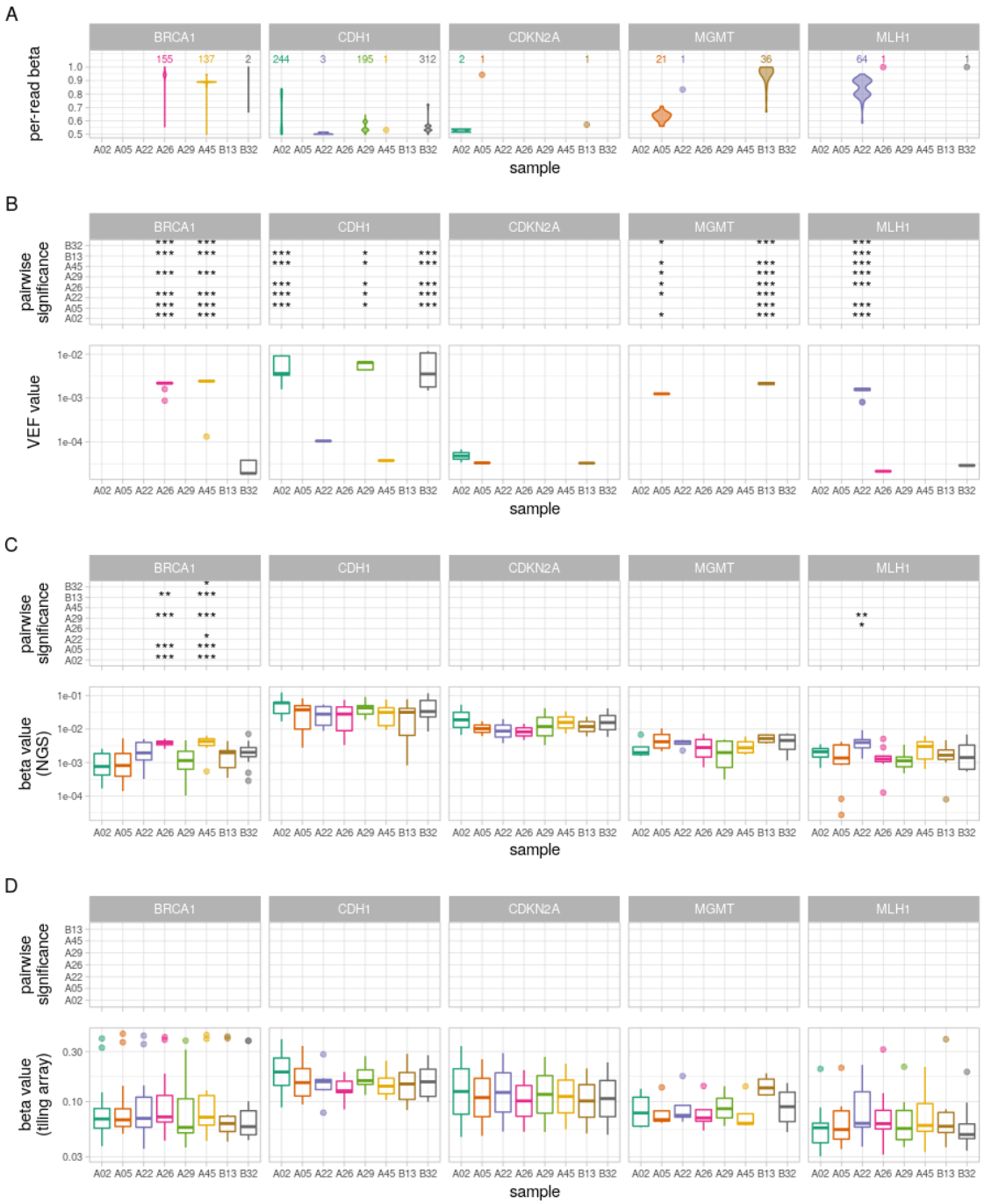

Figure 3. (A) Distribution of per-read beta values for NGS reads covering CpGs that are common for NGS and tiling array. For clarity, only the reads with average beta of at least 0.5 (i.e., representing hypermethylated epialleles) are included. Single observations are shown as dots, number of observations is given above. (B) Lower panel: box plots of NGS-derived VEF values for individual CpGs; upper panel: significance levels from pairwise comparison of VEF values. (C) Lower panel: box plots of NGS-derived beta values for individual CpGs; upper panel: significance levels from pairwise comparison of beta values. (D) Lower panel: box plots of tiling array-derived beta values for individual CpGs; upper panel: significance levels from pairwise comparison of beta values. (B–D) The lower and

upper hinges of boxes correspond to the first (Q_1) and third (Q_3) quartiles; the bar in the middle correspond to the median value; the upper and lower whisker extend to Q_3+1.5*IQR and Q_1-1.5*IQR, respectively, while the values outside this range (outliers) are plotted as dots. Zero values are not plotted. *** p<0.001, ** p<0.01, * p<0.05, blank p≥0.05.

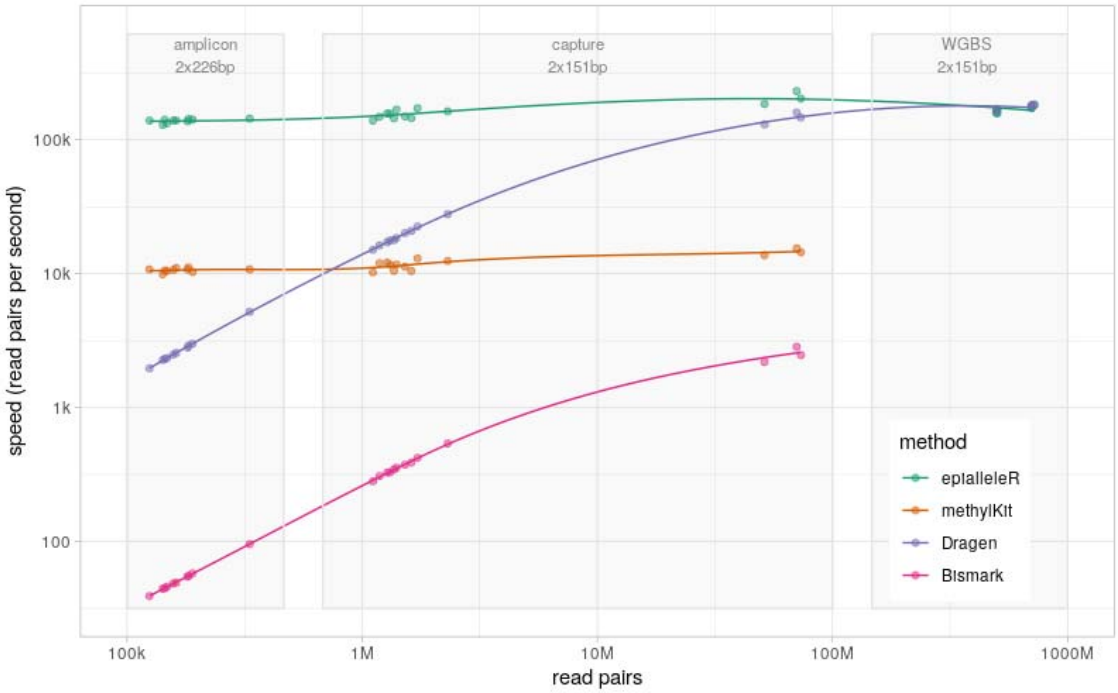

Figure 4. Data processing speed (in read pairs per second) of methods for methylation reporting. Read count (in number of pairs) is given at x-axis; light grey boxes outline data obtained by targeted amplicon-based, genome-wide capture-based, or whole-genome bisulfite sequencing.