DOI: 10.1093/gigascience/giad087
Research

epialleleR: an R/Bioconductor 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
*Correspondence address. Oleksii Nikolaienko, K. G. Jebsen Center for Genome-Directed Cancer Therapy, Department of Clinical Science, University of Bergen, Bergen 5021, Norway. Tel: +47 559 76 444; Email: oleksii.nikolaienko@uib.no

Abstract

Low-level mosaic epimutations within the *BRCA1* gene promoter occur in 5–8% of healthy individuals and are associated with a significantly elevated risk of breast and ovarian cancer. Similar events may also affect other tumor suppressor genes, potentially being a significant contributor to cancer burden. While this opens 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 visualization of mosaic epimutations in methylation sequencing data. Analyzing simulated and real data sets, we provide in-depth assessments of epialleleR performance and show that linkage to epihaplotype data is necessary to detect low-level methylation events. The epialleleR is freely available at https://github.com/BBCG/epialleleR and https://bioconductor.org/packages/epialleleR/ as an open-source R/Bioconductor package.

Keywords: epigenetics, DNA methylation, somatic mosaicism, epigenetic mosaicism, methylation sequencing

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 harbor epimutations [1, 2] (i.e., epigenetic disturbances) that lead to aberrant transcriptional up- and downregulation. Such aberrations are often studied at the level of cytosine DNA methylation. As typical promoters of active genes are hypomethylated, epimutations within such regions are manifested as DNA hypermethylation—the common mechanism of gene repression in cancer [3]. For example, aberrant DNA hypermethylation events (epimutations) within promoters of tumor suppressor genes *BRCA1*, *MGMT*, and *MLH1* were shown to be associated with downregulation of expression of these genes [4–6], and the presence of such epimutations further guides treatment strategies in clinical practice [7–9].

Epigenetic aberrations may arise during different stages of carcinogenesis as somatic epimutations (mirroring somatic mutations) or *in utero* (affecting several germline layers) as constitutional normal tissue epimutations. Several studies in large cohorts [10, 11] have linked constitutional (prenatal), mosaic (affecting a small subset of cells only) epimutations to breast and/or ovarian cancer risk. Research and interest in this field, however, have 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 frequent (occurring in >5% of healthy women) though low-level (down to 0.03% of affected alleles) mosaic epimutations within the *BRCA1* gene promoter to be associated with a significantly elevated risk for subsequent high-grade ovarian as well as triple-negative breast cancer, in a large, population-based prospective cohort [12]. This finding raises a provoking question of whether similar low-level mosaic epimutations may affect other tumor 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 epimutation detection techniques.

Currently, the most widely used methods for DNA methylation profiling are BeadChip arrays (such as Illumina HumanMethylation450 or HumanMethylationEPIC) and a variety of methylation sequencing techniques (for details see [13]). 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 analyze 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 [14–16]. While this approach is suitable for addressing large differences in DNA methylation profiles between 2 sets of samples (e.g., cases and controls), it lacks sensitivity for low-level mosaic epimutation detection, as the detection is hindered by sometimes much more common biological variation [17, 18] or technical artifacts [19, 20]. Moreover, the lack of haplotype linkage makes such analysis difficult in Bead-

Received: June 1, 2023. **Revised:** August 3, 2023. **Accepted:** September 26, 2023

Chip array-based datasets and therefore requires nontrivial approaches [21]. Gene promoter methylation present in a low fraction of molecules may be detected by conventional methylation-specific quantitative polymerase chain reaction (MS-qPCR), but the discrimination between methylated and unmethylated alleles is limited to the CpGs directly covered by the primers/probes [10]. In contrast to other methods, analysis of next-generation sequencing (NGS)-based data can provide much higher sensitivity when the base resolution methylation data are 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 (mitotically and/or meiotically heritable DNA methylation patterns [22]) connected to disease risk (as done previously in [12, 23]), as well as for purposes allowing less sensitivity, such as assessments related to treatment response [24, 25], or to the development of treatment resistance [26]. Importantly, the framework also allows one to connect DNA methylation status with potential underlying *cis*-factors, such as single-nucleotide variations or mutations within the immediate proximity.

The versatility of the framework makes it applicable for analysis of data from any methylation sequencing experiment, given that methylation in these data can be called at individual cytosine residues. Both single-end and paired-end sequencing alignment files can be used as an input, and in cases where methylation calls are not available, this framework allows one to call cytosine methylation and permanently store calls in a binary sequence alignment/map (BAM) file.

Similar to other tools that transform NGS reads into counts of bases or molecules, the framework is not designed to determine preanalytical bias, such as cell-type heterogeneity. Appropriate methods must be used to control confounders in the downstream analyses [27, 28].

Results

epialleleR implementation

The presence of hypermethylated *BRCA1* alleles (epimutations) in normal tissue (white blood cells [WBCs]) has been shown *qualitatively* for 5–8% of adult women [10]. However, the associated *quantitative* changes in DNA methylation at the level of individual CpGs are typically small (in most cases, the intraindividual frequency of epimutations is between 0.03% and 1% [12]) and therefore indistinguishable from the background methylation level due to inherent biological (potentially spurious single-base methylation events) and technical (sequencing errors) variance [17]. Methylation statuses of neighboring CpGs are often concordant [29], and such spatially extended epigenetic changes are often associated with a gene expression silencing [30]. Given the potential biological (gene inactivation) and clinical (cancer risk) importance of epimutations, we focused on quantification of hypermethylation events that span over several CpGs, accounting for both methylation status of individual CpGs within the sequence read as well as the average methylation level of the sequence read itself. This is possible in NGS-based data sets, while it is not in array-based data where methylation information of different CpGs cannot be connected to each other as in haplotype data.

As number of events that lead to variance in methylation (base deamination, random single-base methylation events, and sequencing errors) is limited at the level of individual reads (only a fraction of CpGs might be affected within the same read), the average methylation level of the read will be moderately affected by such events and can help distinguish hyper- from hypomethylated epialleles (where methylation statuses of the majority of CpGs are concordant and average methylation level is either close to 0% or 100%). We therefore hypothesized that thresholding sequence reads by their average methylation level will reduce the effect of biological and technical variance and facilitate the detection of infrequent hypermethylation events. As no suitable generic solution was publicly available, we implemented it using R software environment for statistical computing [31], a *de facto* standard for scientific data analysis. The implemented solution, epialleleR, loads methylation call strings and short sequence reads from the supplied BAM file, optionally thresholds read pairs according to their methylation properties, and produces methylation reports for individual cytosines as well as genomic regions of interest (Fig. 1A). During BAM loading, pairs of sequence reads and corresponding methylation call strings are merged according to Phred quality score values (i.e., base with the highest score is chosen) to preserve information of the highest quality. In contrast to approaches that involve simple trimming of overlapping parts of read2, the following approach might retain more information when higher-quality fragments of read2 (5′-end or middle) overlap with lower-quality fragments of read1 (3′-end). The optional thresholding defines a subpopulation of epialleles of interest and is based on the minimum number and the average methylation level of cytosines in various sequence contexts (e.g., CG, CHG, or CHH). The thresholding parameters are fully adjustable to target desired population of epialleles; their default values (minimum 2 CpG sites, minimum average methylation beta value of 0.5 for CpG sites, maximum average methylation beta value of 0.1 for non-CpG sites) performed well in the study linking mosaic *BRCA1* epimutations and cancer risk [12] and were used here in all downstream analyses.

The optional thresholding of sequence reads defines 2 modes of epialleleR (v.1.3.5, RRID:SCR_023913 [32]) function. Without thresholding, epialleleR produces conventional cytosine reports similar to the ones produced by other tools (e.g., Bismark [14]). In this case, methylation beta value for every genomic location is computed as a ratio of a number of methylated cytosines to the total number of methylated and unmethylated cytosines: $\beta = C/(C + T)$.

When read thresholding is performed (default mode of action), the level of methylation per every genomic position, denoted as a variant epiallele frequency (VEF), is calculated as a ratio of a number of methylated cytosines in read pairs passing the threshold (C^a) to total number of methylated and unmethylated cytosines in all read pairs: $VEF = C^a/(C + T)$ (see Fig. 1B for an example). When the report is prepared at a level of extended genomic regions rather than individual bases, VEF equals the ratio of a number of read pairs passing threshold (N^a) to the total number of read pairs (N) overlapping the region of interest: $VEF = N^a/N$. The term “variant epiallele” here represents a group of epialleles (i.e., individual methylation patterns) with similar methylation properties that is defined by thresholding; therefore, VEF effectively represents the frequency of this group of epialleles passing the threshold at the level of individual cytosines or extended genomic regions.

Methylation beta values (from conventional reporting) as well as VEF values (from default reporting mode with read thresholding) can be produced from any number of BAM files with no prior hypothesis, as long as experimental setup allows to call methylation on a per-base level. Both of these values effectively represent methylation levels per genomic position and, as such, can be directly used further as an input for other bioinformatic tools including, but not limited to, differential methylation analysis tools.

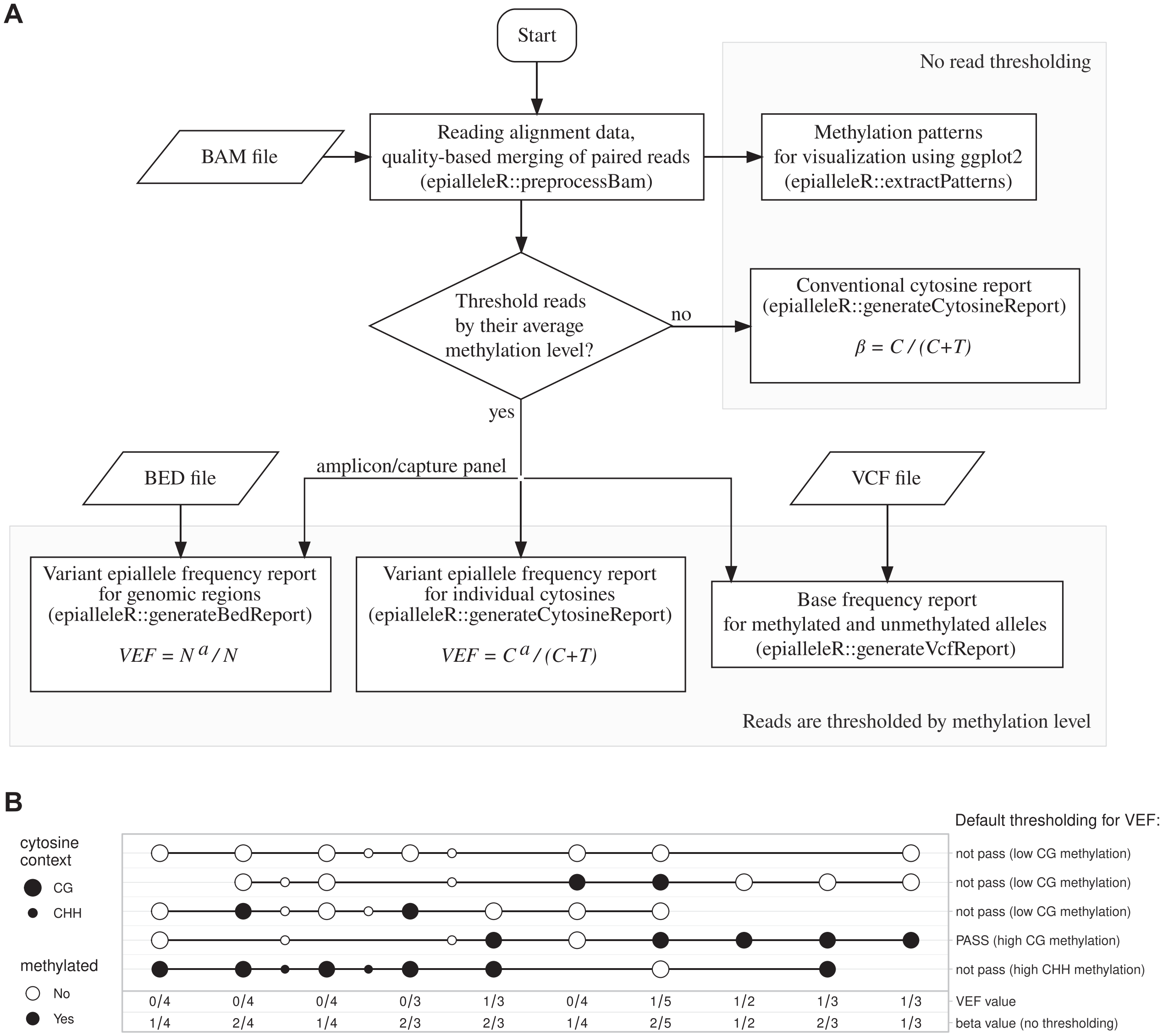
A
Start
No read thresholding
BAM file
Reading alignment data, quality-based merging of paired reads (epialleleR::preprocessBam)
Methylation patterns for visualization using ggplot2 (epialleleR::extractPatterns)
Threshold reads by their average methylation level?
no
Conventional cytosine report (epialleleR::generateCytosineReport)
β = C / (C+T)
yes
amplicon/capture panel
BED file
VCF file
Variant epiallele frequency report for genomic regions (epialleleR::generateBedReport)
VEF = N a / N
Variant epiallele frequency report for individual cytosines (epialleleR::generateCytosineReport)
VEF = C a / (C+T)
Base frequency report for methylated and unmethylated alleles (epialleleR::generateVcfReport)
Reads are thresholded by methylation level
B
cytosine context
CG
CHH
methylated
No
Yes
Default thresholding for VEF:
not pass (low CG methylation)
not pass (low CG methylation)
not pass (low CG methylation)
PASS (high CG methylation)
not pass (high CHH methylation)
0/4 0/4 0/4 0/3 1/3 0/4 1/5 1/2 1/3 1/3
VEF value
1/4 2/4 1/4 2/3 2/3 1/4 2/5 1/2 2/3 1/3
beta value (no thresholding)

Figure 1: (A) Flowchart of epialleleR package data-processing steps. The formulas using to calculate conventional beta as well as 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 at particular genomic position within read pairs passing a particular methylation threshold ($C^a \leq C$); *N*, total number of read pairs, mapped to a particular genomic region; N^a, number of mapped read pairs, passing a particular methylation threshold ($N^a \leq N$). (B) Schematic illustration of cytosine methylation (circles) within epialleles (horizontal lines) and results of thresholding by average read methylation level (labels on the right) using default parameters (i.e., at least 2 CpGs in CG context, at least 50% methylation within CG context, at most 10% methylation outside of CG context). These default thresholding parameters were chosen to detect hypermethylated alleles with biological relevance in tumor suppressor genes; detection of epimutations of a different nature may require adjustments to the default parameter values. Resulting per-cytosine beta and VEF values are given under each CpG (large circles). In the context of a typical CpG-rich regulatory region of an actively transcribed gene, the 3 hypomethylated epialleles on the top represent typically abundant scattered methylation or sequencing artifacts (only a minority of cytosines in CG context are called as methylated), the epiallele at the bottom represents the product of incomplete bisulfite conversion (cytosines in GG and non-CG contexts are methylated), and the second epiallele from the bottom represents a true biologically relevant epimutation (hypermethylation) that leads to gene silencing (majority of cytosines in CG context are methylated, while no methylation is detected in the non-CG context).

If methylation statuses of cytosines were not determined, epialleleR allows to create and store methylation calls, allowing analysis of BAM files created by various methylation sequencing alignment tools.

When optional data on single-nucleotide variants are provided (as a variant call format, VCF file, or a VCF object), epialleleR quantifies the balance or skewness of methylation between alleles, thereby enabling assessment of potential allele specificity of epimutations. In particular, this information is important for distinguishing epimutations that occurred through a single event followed by clonal expansion (e.g., prenatal epimutations that are present on the same allele in all affected cells, as in [12, 23]) from the ones that occurred in different cells independently and therefore present on both alleles. In some cases, allele specificity also allows to infer causality of epimutations in cancer development [23].

To provide a comprehensive range of means for epiallele analysis, the package also offers methods allowing visualization and characterization of *all* individual epialleles (methylation patterns) in a sample (see Fig. 1 and Supplementary figures for details). If

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

Method	Requires reference (genomic) sequence	Removes overlaps within read pairs	Outputs epiallele frequencies	Processing speed, read pairs per second
Bismark	yes (genome-wide cytosine reports) / no (bedGraph reports)	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

required, extracted patterns can include other, noncytosine bases of interest (e.g., single-nucleotide variations), which allows to connect methylation properties of epialleles with sequence features in proximity. During methylation pattern extraction, every epiallele is characterized by number of context sites and methylation level (average beta value) and is assigned with a unique identifier (Fowler-Noll-Vo FNV-1a non-cryptographic hash [33]) that solely depends on positions of included cytosine (and other optional) bases and their methylation states (or nucleotide symbols for optional bases), enabling to not only group epialleles by their methylation properties but also reliably and consistently track individual epialleles of high importance across different samples or even studies. The average beta values for all extracted patterns as well as patterns themselves can be explored to optimize thresholding parameters for a genomic region of interest.

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 optimized C/C++ subroutines and, whenever possible, linked to HTSlib, the unified C library for high-throughput sequencing data processing [34]. The R package epialleleR is freely available at the Bioconductor package repository [32].

Reporting accuracy analyses

First, we sought to validate the accuracy of methylation reporting by epialleleR in its conventional mode (no read thresholding) as compared with 3 other commonly used tools for which read thresholding is not available: Bismark [14], methylKit [35], and Illumina DRAGEN Bio-IT Platform. For this purpose, we simulated large sets of paired-end bisulfite sequencing reads (2 × 151bp, 100 million read pairs covering human chromosome 19). In contrast to real datasets, simulated data allow to calculate "ground-truth" methylation levels for unbiased comparison. Simulation parameters were selected to obtain exact methylation levels of 50% for cytosines in the CG context ($n = 2{,}211{,}240$) and methylation level of approximately 0.25% (bisulfite conversion rate of ∼99.75%) for cytosines in the CHG and CHH contexts ($n = 6{,}593{,}900$ and 19,210,572, respectively). In addition to endogenous deamination events [17], bisulfite treatment-induced changes [19], and variation in conversion rates [36], sequencing itself can introduce errors that vary in range depending on assay type and sequencing technology [20]. 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 on exactly the same task (BAM file to cytosine report) revealed that reported values were close to their theoretical expectations for all methods, with epialleleR being the least affected by sequencing errors, that is, 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 S1).

Of note, epialleleR does not require reference sequence in order to determine the correct sequence context of cytosine bases. All observed contexts for every genomic position are counted, and the most frequent context is assumed to be correct and therefore reported. This approach allows reporting of methylation events within *de novo* (not present in the reference genome) contexts, being at the same time not affected by sequencing errors that change sequence context of cytosine bases (Supplementary Table S1).

Sensitivity analyses

Concordantly methylated alleles (alleles with most of their CpGs having the same methylation status) may possess high biological importance [12, 23, 26]. Spontaneous 5-methyl cytosine (5mC) deamination, sequencing errors, and 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 normalization during postprocessing [37]. In contrast to the DNA methylation analysis using BeadChip 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 WBCs ($n = 10$ with almost no hypermethylated alleles, as described in Materials and Methods) and fully methylated control DNA samples. 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. The amplicons used covered promoter regions of the tumor suppressors *MLH1*, *CDKN2A*, *MGMT*, *CDH1*, and *BRCA1*. The distributions of per-read beta values (Supplementary Fig. S1) and methylation patterns (Supplementary Fig. S2) of admixed samples show the expected abundance of hypermethylated (average $\beta \geq 0.5$) alleles and confirm their high similarity to the real samples

Table 2: Selected accuracy metrics (average beta values and their variance) of cytosine methylation reporting. Average reported beta values that are closest to the expected beta values (0.0025 for cytosines in the CHG/CHH contexts and 0.5 for cytosines inthe CG context) and lowest variance values are shown in bold.

		CHH		CHG		CG	
Sequencing error rate	**Method**	**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**

(Supplementary Figs. S3 and S4). Conventional cytosine reports (no read thresholding) as well as VEF reports (with read thresholding) were prepared and used for unsupervised clustering of samples and differentially methylated region (DMR) discovery. Despite quite a low overall methylation level of amplified regions (average beta value of 0.014, median of 0.005; Fig. 2A), t-distributed stochastic neighbor embedding (t-SNE) analysis based on beta values was not able to discriminate between samples with 0.01%, 0.03%, 0.10%, and 0.30% of methylated reads or no methylated reads added (Fig. 2B, left panel). 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. 2B, right panel). Intergroup DMR discovery based on beta values (Fig. 2C, left panel) showed fewer number of regions found as well as higher associated false discovery rate (FDR), while discovery based on VEF values resulted in all 5 possible regions identified for all possible intergroup comparisons as well as generally lower associated FDR. When each sample with admixed methylated reads was compared against the group of samples without admixed methylated reads, recall metrics for differential (by DMRcate [38]; Fig. 2D) or aberrant (by ramr [21]; Fig. 2E) methylation analysis were notably higher for analyses based on VEF values (Fig. 2D, E, right panels) in comparison with analyses based on beta values (Fig. 2D, E, left panels). This shows that VEF values are more valuable for detection and analysis of low-frequency (≤1%) hypermethylation events than methylation beta values.

BeadChip arrays, such as Illumina HumanMethylationEPIC, are 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 BeadChip arrays for the detection of low-frequency DNA methylation events, we employed both of the methods to analyze 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*). Sample distributions of per-read beta values (Supplementary Fig. S3) and methylation patterns (Supplementary Fig. S4) show that these samples indeed contain varying frequencies of hypermethylated (average $\beta \geq 0.5$) alleles. For unbiased comparison, we limited the corresponding data sets to the CpGs assayed and sufficiently covered by both techniques. Analysis revealed that VEF values of samples with many hypermethylated alleles (e.g., A26 and A45 for *BRCA1*; as apparent from Fig. 3A) differ significantly (Fig. 3B) from VEF values of samples with only a few or no hypermethylated alleles (e.g., A02 or A05 for *BRCA1*; Fig. 3A). When VEF values were used for identification of aberrantly methylated regions (AMRs) or DMRs by ramr [21] or DMRcate [38], respectively, the significant regions found correlated well with the notable presence of hypermethylated alleles. Of note, slightly inferior performance of DMRcate is probably due to the fact that for some of the genomic regions, too many samples in this subset simultaneously contained hypermethylated epialleles. When DMRcate was used for the same purpose on an extended set of sequenced samples ($n = 18$, containing $n = 10$ samples characterized by the absence of hypermethylated alleles that were used to create the admixed sample set), its performance in identification of hypermethylated epiallele-containing samples was higher (Supplementary Fig. S5).

In contrast, only a few significant differences remained when NGS beta values were used for sample comparison (Fig. 3C), while pairwise comparisons based on BeadChip array beta values did not reveal any significant differences between samples (Fig. 3D). The search for aberrantly or differentially methylated regions using either NGS or array beta values did not result in identification of such regions in relevant (according to methylation patterns or beta value densities) samples. Generally higher beta values of BeadChip array as compared to NGS beta values likely mask subtle changes in methylation caused by the presence of infrequent hypermethylated alleles and hinder the detection of differences between samples.

Several scores to describe and quantify variability in DNA methylation in sequencing reads (within-sample heterogeneity [WSH]) have been proposed [39]. In order to assess WSH, we evaluated the difference in combinatorial entropy between each pair of samples using methclone [40] (Supplementary Fig. S6A). The largest (by absolute value) reported difference in combinatorial entropy of −2.59 between any pair of samples confirms a high similarity between sample methylation profiles, of note, being much smaller than cutoffs for epiallele shifts between samples analyzed

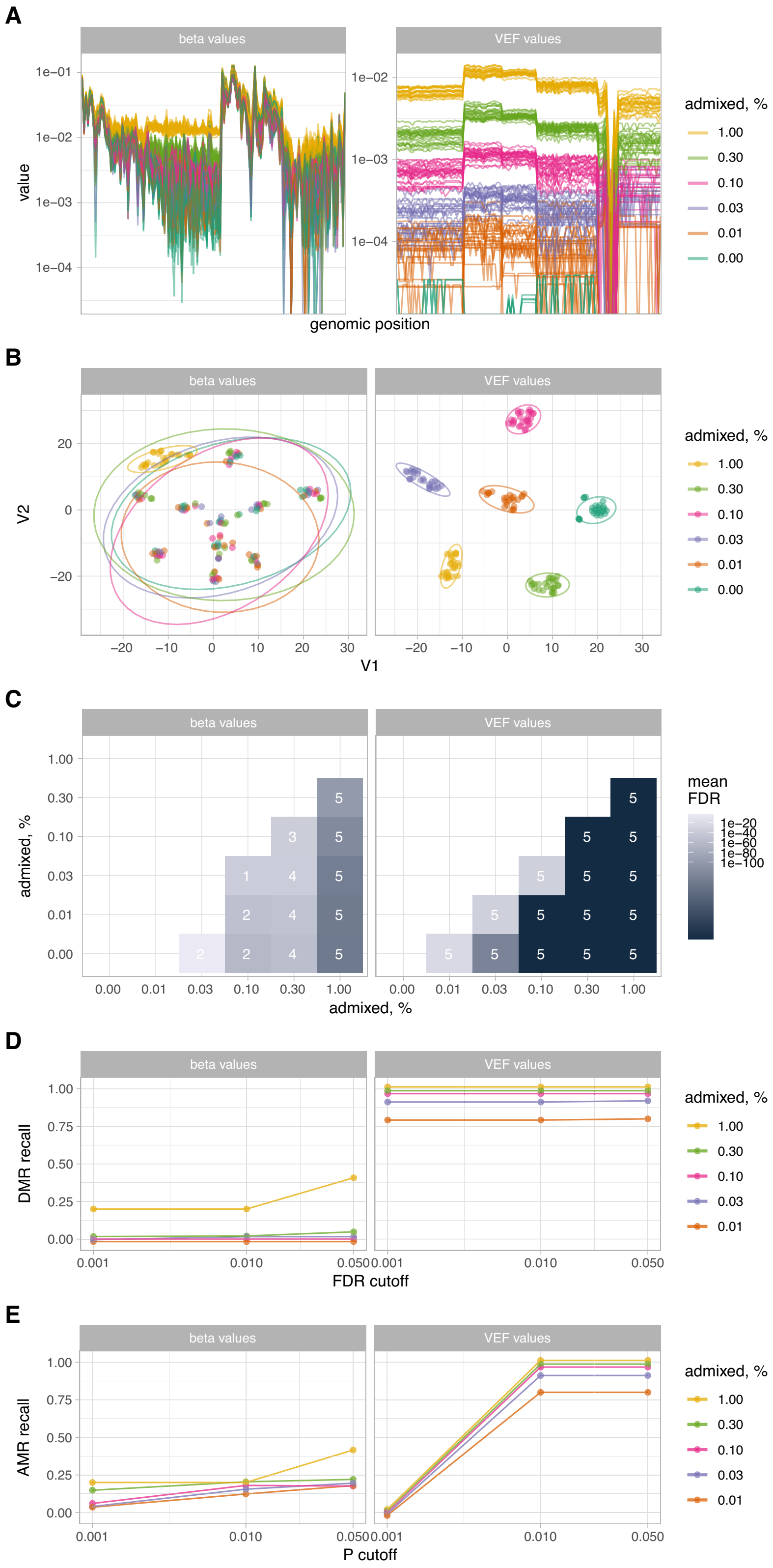

Figure 2: (A) Line plots of beta (left panel) and VEF (right panel) values for individual samples, color-coded according to the amount of admixed methylated reads. Each line represents a sample; y-axis, methylation value of all CpGs ($n = 138$) sorted by their genomic position (categorical x-axis). (B) Embedding plots for t-SNE analysis using beta (left panel) and VEF (right panel) values. Ellipses represent 95% confidence levels. (C) Heatmap of mean false discovery rate for differentially methylated regions (DMRs) identified by DMRcate. Labels indicate the number of DMRs found (of a total of 5 regions possible). (D) Recall rate for DMR identification using DMRcate for varying FDR cutoffs. (E) Recall rate for aberrantly methylated region (AMR) identification using ramr for varying *P* value cutoffs.

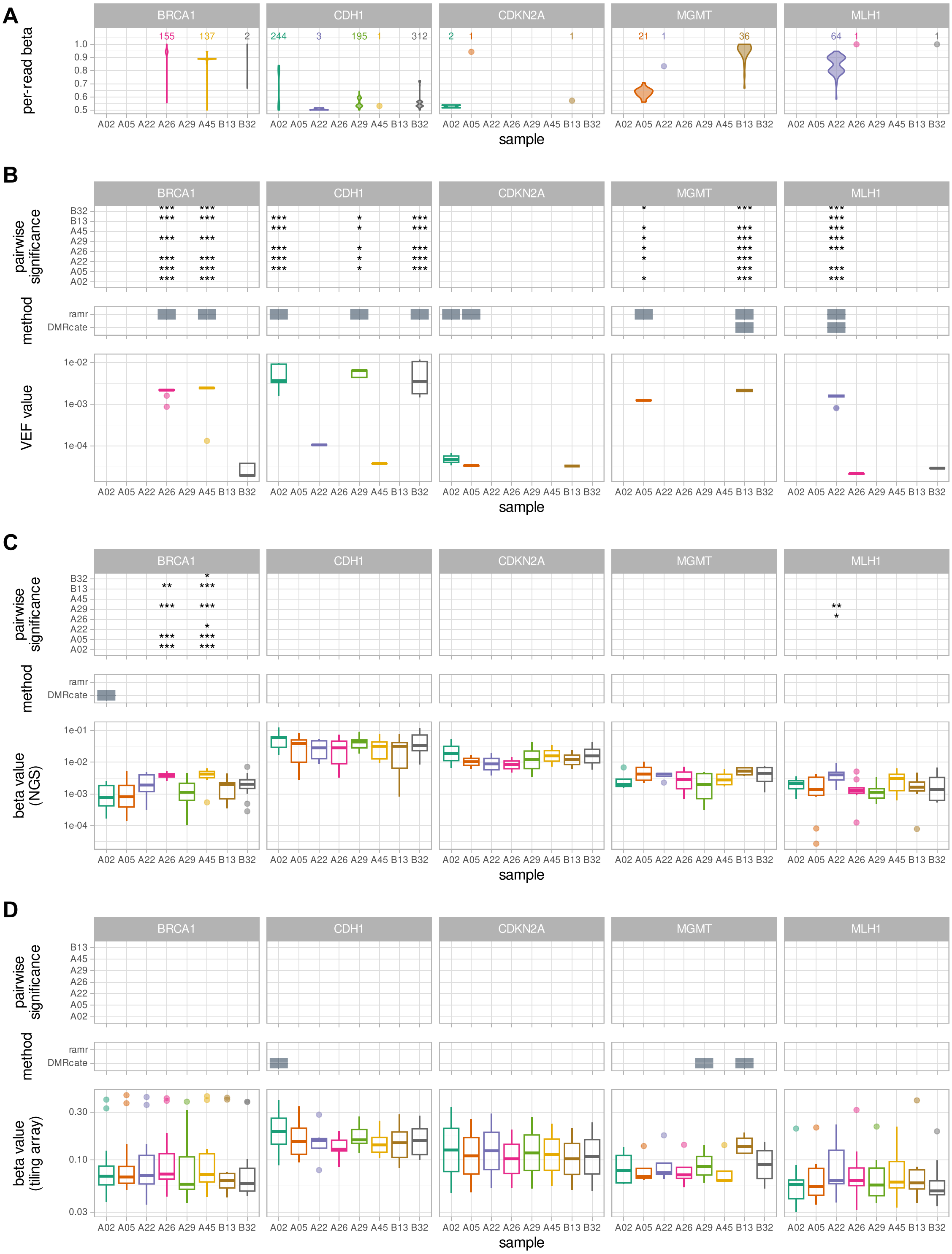

Figure 3: (A) Distribution of per-read beta values for NGS read pairs covering CpGs that are common for NGS and BeadChip array. For clarity, only the reads with average beta of at least 0.5 (i.e., representing hypermethylated epialleles) are included. Single observations are shown as dots; number of observations is given above. Complete density plots are provided in the Supplementary Fig. S3. Corresponding methylation patterns are provided in the Supplementary Fig. S4. (B) Lower panel: boxplots of NGS-derived VEF values for individual CpGs; middle panel: significant aberrantly or differentially methylated regions identified by ramr or DMRcate, respectively, based on VEF values; upper panel: significance levels from pairwise comparison of VEF values. (C) Lower panel: boxplots of NGS-derived beta values for individual CpGs; middle panel: significant aberrantly or differentially methylated regions identified by ramr or DMRcate, respectively, based on NGS-derived beta values; upper panel: significance levels from pairwise comparison of NGS-derived beta values. (D) Lower panel: boxplots of BeadChip array-derived beta values for individual CpGs; middle panel: significant aberrantly or differentially methylated regions identified by ramr or DMRcate, respectively, based on BeadChip array-derived beta values; upper panel: significance levels from pairwise comparison of BeadChip array-derived 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 corresponds 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 $P \geq 0.05$.

in [39] (−60 and lower). Further, we also calculated 4 additional heterogeneity scores: combinatorial entropy, epipolymorphism, fraction of discordant read pairs (FDRP), and proportion of discordant reads (PDR). The scores themselves (Supplementary Fig. S6B) and the levels of score-based pairwise significance between samples (Supplementary Fig. S6C) are not generally consistent with fractions of hypermethylated (average $\beta \geq 0.5$) alleles (Fig. 3A and Supplementary Fig. S3) or VEF values (Fig. 3B): for example, samples A26 and A45 have a notable fraction of hypermethylated reads in the *BRCA1* promoter region compared to other samples, although it is not reflected at the level of WSH scores. Importantly, WSH scores produced cannot be directly used as an input for DMR analysis tools, which are commonly employed to characterize exact differences in methylation between samples.

It is known that DNA methylation profiles of blood samples depend on the varying contribution of individual blood cell types [41, 42]. While we cannot exclude that hypermethylated alleles present in the samples analyzed here originate from a particular blood cell type, low-level, mosaic epimutations of at least *BRCA1* were previously shown to be independent of blood subfraction composition [10]. Of note, only 1 CpG (cg05785947 in *CDH1*) out of 37 used in NGS versus BeadChip array comparison here was found to be significantly differentially methylated between blood cell types of healthy males, and none of the CpGs were significantly differentially methylated between blood cell types of newborns.

Processing speed analyses

Methylation sequencing data produced by contemporary techniques vary in scale and depth and may contain several thousands to billions of single or paired-end reads. To analyze them efficiently, computational methods must be scalable and fast enough for as large as possible range of sample counts or data file sizes. Unfortunately, many academic tools use computationally complex algorithms that do not scale to contemporary tasks. We compared data-processing speed for epialleleR versus methylKit, Bismark, and DRAGEN Bio-IT Platform, performing exactly the same task (BAM file to cytosine report) of methylation reporting across input data coming from various assays: amplicon based ($n = 10$ samples with a depth of coverage of ∼20,000×), genome-wide capture based ($n = 10$ with a depth of coverage of ∼60× and $n = 3$ with a depth of coverage of ∼1,000×), or whole-genome bisulfite sequencing (WGBS, $n = 6$ with a depth of coverage of ∼60×). The obtained results confirm very efficient implementation of epialleleR and its suitability for analysis of datasets of any depth and coverage (Fig. 4, Table 1).

Discussion

While conflicting data have linked low-level mosaic primary constitutional epimutations to cancer risk for more than a decade [43], we have recently obtained firm evidence implicating primary epimutations within the *BRCA1* gene in an elevated risk of incident breast and ovarian cancer [12]. The assumption that such epimutations may affect other tumor suppressor genes and, therefore, lead to other cancer forms [43] institutes a new research area with respect to cancer risk. Further, the findings of such epimutations in umbilical cord blood [10, 23] 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 visualize epialleles in NGS data. The method shows its superiority versus conventional methods of methylation reporting, especially when applied for detection of low-frequency methylation events, as it is by design less susceptible to variations in conversion efficiency or sequencing quality. Although epialleleR is not a differential methylation analysis tool, its output can be directly used to group samples based on their methylation profiles (by applying a simple threshold as in [12, 23] or using unsupervised clustering), as well as an input for other differential/aberrant methylation analysis software (the latter is not possible for WSH analysis tools).

The default epialleleR parameters that were used for read thresholding in the present and linked studies [12, 23] are sought to be optimal for the detection of aberrant hypermethylation events within normally unmethylated genomic regions such as CpG-rich regulatory regions of tumor suppressor genes. If the nature of regions of interest deviates from the one described above, methylation characteristics can be explored using other epialleleR methods (e.g., extractPatterns), and thresholding parameters can be adjusted to detect desired methylation events.

We thoroughly tested epialleleR using bisulfite sequencing data; the method, however, can also be applied to analyze and compare data obtained using any methylation sequencing technique (reduced representation bisulfite sequencing [RRBS], oxidative bisulfite sequencing [oxBS-seq], and Tet-assistant bisulfite sequencing [TAB-seq]), as long as methylation in these data can be called at individual cytosine residues instead of being analyzed by comparing relative abundance of the fragments (such as for methylation sensitive restriction enzyme sequencing [MRE-seq] or methylated DNA immunoprecipitation sequencing [MeDIP-seq]).

The possibility to call cytosine methylation for alignment files created by different short sequence aligners and subtle though noticeable changes in cytosine reporting accuracy, together with immense speed gain, make epialleleR a method of choice not only for discovery of infrequent hypermethylated epialleles (as in [12, 23]) but also as a tool to produce conventional (no read thresholding) cytosine reports from any methylation sequencing alignment files.

The implemented method is fully documented and can be easily used from within the R environment for statistical computing. With the epialleleR already revealing its suitability for detection of low-level mosaic methylation events in a large dataset [12, 23], we believe it constitutes an optimal tool for assessment of low-level mosaic epimutations with respect to risk of cancer as well as other diseases of relevance.

Conclusions

Here, we present epialleleR, a very fast, accurate, and sensitive method to detect, quantify, and visualize epialleles in NGS data. Efficient implementation and improvements in cytosine reporting accuracy allow us to recommend epialleleR not only for analysis of methylation patterns and to enhance low-level differentially methylated region discovery but also as a conventional cytosine reporting tool for various kinds of methylation sequencing data. The epialleleR R/Bioconductor package is freely available at [44, 45].

Materials and Methods

Next-generation sequencing

WBC DNA samples from anonymized males ($n = 88$) [46, 47] and human HCT116 DKO methylated DNA control sample (Zymo Research, cat. D5014-2) were bisulfite converted, and 5 DNA

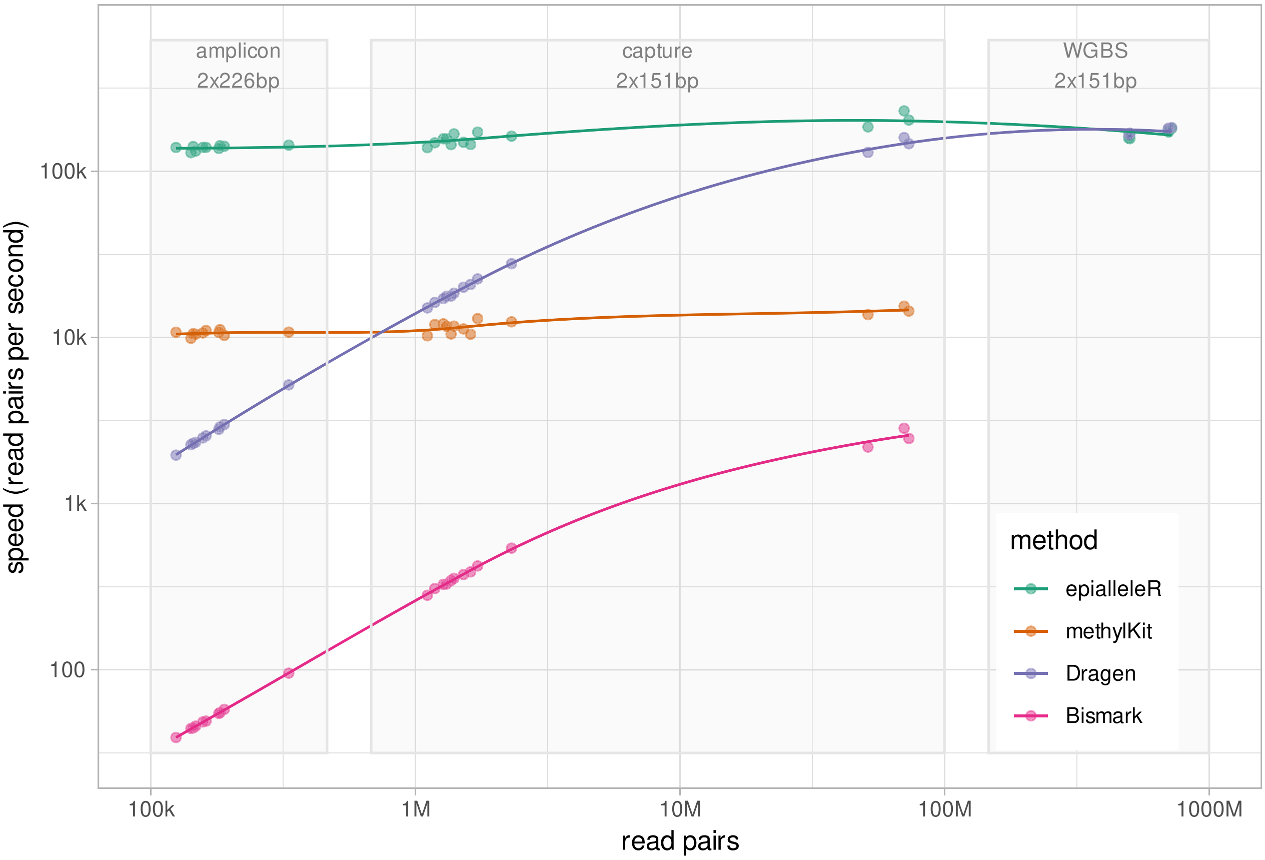

Figure 4: Data-processing speed (in read pairs per second) of epialleleR as compared to 3 other methods for methylation reporting (methylKit, Bismark, and DRAGEN Bio-IT Platform). Read count (in number of pairs) is given at x-axis; light gray boxes outline data obtained by targeted amplicon-based, genome-wide capture-based, or whole-genome bisulfite sequencing.

fragments, representing promoter regions of 5 established tumor suppressor genes, were amplified using a custom set of primers (GRCh38 assembly coordinates of assayed regions: *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 [12] (GSE201688). The resulting average coverage was 5,000× to 50,000× 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 nondirectional, and –enable-sort false, unless stated otherwise. R software environment for statistical computing (v4.1.2) was used for all downstream statistical analyses.

The frequency of hypermethylated alleles across assayed regions in $n = 88$ male WBC DNA samples was estimated using epialleleR::generateAmpliconReport with the following parameters: min.mapq=30, min.baseq=20, nthreads=4, threshold.reads=TRUE, report.context="CG," and bed.file pointing to a location of a BED (browser extensible data) file with genomic regions amplified (see amplicon coordinates above). Two sample subgroups ($n = 8$ and $n = 10$) were selected for sensitivity analyses based on the frequencies of hypermethylated alleles as explained below.

Cytosine reporting accuracy comparison

Four sets of paired-end sequencing reads (151 bp, 50 million read pairs each set) were simulated using Sherman Bisulfite FastQ Read Simulator (RRID:SCR_001294) [48] 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%. The quality scores of these simulated sequences followed an exponential decay curve, which resulted in a higher number of base errors toward the 3′-end of the read (as seen in real data). Human chromosome 19 sequence (GRCh38.p13 NC_000019.10, 58,617,616 bp, 1,105,620 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 [49] and in order to maintain optimal balance of analysis speed and base coverage. Each set of reads was then duplicated, and 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 (i.e., unmethylated reads) were merged with original sets (i.e., methylated reads), resulting in 4 sets of reads 100 million pairs each, with the cytosine conversion rate of exactly 50% and about 99.75% in CG and non-CG contexts, respectively.

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

Sensitivity comparison on admixed samples

In order to simulate variable methylation levels while maintaining biological heterogeneity of the samples, we selected 10 male DNA NGS samples with the lowest frequency of hypermethylated alleles across all assayed regions, then admixed varying fractions of reads from 2 random samples and additionally "spiked" a certain number of fully methylated reads from the 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 were performed using the Illumina DRAGEN Bio-IT Platform as described above. VEF calling was performed using epialleleR::generateCytosineReport with the following parameters: min.mapq=0, min.baseq=0, nthreads=4, threshold.reads=TRUE, and report.context="CG."

Methylation patterns and per-read beta values for all samples were extracted using epialleleR::extractPatterns with the following parameters: min.mapq=30, min.baseq=20, nthreads=4, clip.patterns=FALSE, and bed.file pointing to a location of the BED file with genomic regions amplified (see amplicon coordinates above).

Barnes–Hut t-SNE analysis was performed using R package Rtsne v0.15 [50] and matrices of beta or VEF values for all genomic positions of CpGs with the coverage of at least 1,000× and available values for all analyzed samples (total number of CpGs, n = 138; *MLH1*, n = 20; *CDKN2A*, n = 35; *MGMT*, n = 33; *CDH1*, n = 32; *BRCA1*, n = 18).

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 VEF calling was performed using epialleleR::generateCytosineReport with the following parameters: min.mapq=30, min.baseq=20, nthreads=4, threshold.reads=TRUE, and report.context="CG." The same DNA samples were also bisulfite converted using the Zymo EZ DNA Methylation Kit (Zymo Research, cat. D5001), and genome-wide methylation levels were assessed using Illumina HumanMethylationEPIC BeadChip arrays according to the manufacturer's instructions. Resulting IDAT files were processed (normalized and annotated) with the minfi Bioconductor package [37] using the preprocessQuantile method with outlier thresholding enabled (GSE201689). For direct comparison, only the CpGs that are covered in all samples by both BeadChip arrays (P value of 0) and targeted sequencing (minimum sequencing coverage of 5,000×) were retained (*MLH1*, n = 10; *CDKN2A*, n = 2; *MGMT*, n = 4; *CDH1*, n = 7; *BRCA1*, n = 14). Pairwise sample comparison was performed using a t-test with Holm adjustment for multiple comparisons.

The sets of CpGs that are differentially methylated between cell blood types were reported previously: DNA methylation profiles for 6 blood cell types from 6 males [41, 51] and DNA methylation profiles for 7 blood cell types from cord blood of 104 newborns [42, 52]. CpG-level differential methylation analysis P values were Holm-adjusted, and the ones that remained significant (adjusted $P \leq 0.05$; n = 73,629 of total 456,655 for male blood data set; n = 221,246 of total 429,794 for newborn cord blood data set) were checked for overlap with the set of CpGs analyzed in this study (n = 35 CpGs of total n = 37 were present in each of male/newborn datasets).

Differential methylation analysis

DMRs were called using R package DMRcate (v2.12.0) with the following parameters: lambda=1000, min.cpgs=2, and pcutoff="fdr" [38]. AMRs were called using R package ramr (v1.6.0) with the following parameters: ramr.method="beta," min.cpgs=2,and merge.window=500 [21]. To enable maximum likelihood estimation of beta distribution parameters, all zeros were replaced with minimum double values (2.26e-308).

For intergroup DMR discovery in admixed samples, pairwise comparison of sample groups defined by the number of admixed reads was performed (n = 25 samples in each group) using the default level of the FDR cutoff (equals 0.05). For DMR discovery in real samples, as DMRcate methods require 2 classes/categories for comparison, every real sample from the test dataset was tested against all the other samples using the default FDR cutoff value.

To assess DMR (or AMR) recall metrics in admixed samples, every sample with admixed reads was compared using DMRcate (or ramr) to the group of 25 samples without admixed reads at a varying level of FDR (or P value) cutoff of 0.05, 0.01, or 0.001. As the admixed reads covered all 5 assayed regions, only the total number of real positive (P) regions (equals 5 for each comparison), the number of true-positive (TP) regions, and the number of false-negative (FN) regions (FN = P – TP) were known, while the numbers of true-negative (TN) or false-positive (FP) regions were undefined. Therefore, recall, or true positive rate (TPR = TP/P), was chosen as a sensitivity metric.

Within-sample heterogeneity

Estimation of WSH was performed on 8 samples used in the sensitivity comparison between array- and NGS-based methylation profiling. Difference in entropy was evaluated using methclone (v1) [40] with a distance cutoff of 500 and minimum read coverage of 1,000 for every pair of samples. As methclone outputs values for multiple genomic regions, the minimum value (representing absolute largest difference) was selected and used further. Entropy, epipolymorphism, FDRP, and PDR were evaluated using R package WSH (v0.1.6) [39] with the following options: mapq.filter=30, window.size=500, and bam.file pointing to a location of the BAM file. Due to exponential complexity of the FDRP calculation, option max.reads was set to 100 for FDRP calculation and to 1e+06 otherwise. Pairwise sample score comparison was performed using a t-test with Holm adjustment for multiple comparisons.

Processing 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 the *BRCA1* gene (n = 10 files, 0.12–0.33 million read pairs per file, average coverage of ∼20,000×) [12], (B) genome-wide capture-based bisulfite sequencing of promoter regions of 283 tumor suppressor genes (n = 10 files, 1.11–2.31 million read pairs per file, average coverage of ∼60×, and n = 3 files, 51.4–73.4 million read pairs per file, average coverage of ∼1,000×) [53, 54]. and (C) whole-genome bisulfite sequencing (n = 6 files, 497–723 million read pairs per file, average coverage of ∼60×; epialleleR and Illumina DRAGEN

Bio-IT Platform only). The 2 former data sets (A and B) were generated in house and described previously, 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 [55].

Processing times to produce conventional cytosine reports were recorded as following:

Bismark CX methylation reports were created using Bismark v0.22.3 (RRID:SCR_005604) [14] with the following parameters: command bismark_methylation_extractor, –paired-end, –no_overlap, –comprehensive, –gzip, –mbias_off –parallel 8, –cytosine_report, –CX, and –buffer_size 64G. Genome-wide cytosine methylation report but not bedGraph report was chosen in order to obtain results of highest quality (not affected by sequencing errors). As parallel processing was requested, Bismark used up to 24 cores for some of its subtasks.

methylKit CX methylation reports were created using R/Bioconductor package methylKit v1.20.0 (RRID:SCR_005177) [35] 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/Bioconductor 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.

Illumina DRAGEN is a hardware solution that relies on the presence of the FPGA accelerator card, which precludes DRAGEN software execution on other platforms. At the same time, outdated software development tools available at DRAGEN (GCC v4.8.5, R v3.6.0) impede installation of third-party software and R/Bioconductor packages and may potentially affect their performance. Therefore, testing of methylation reporting tools was carried out in 2 different settings.

Bismark, methylKit, and epialleleR were tested on the workstation equipped with an AMD EPYC 7742 64-core processor, 512 GB 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 (10 Gbps) 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, 256 GB 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) was used for data processing using DRAGEN; BAM files were accessed from a local, high-speed NVMe solid state disk.

For Bismark and DRAGEN, elapsed time measurements were stably reproducible, and thus processing time was recorded only once for each file. For methylKit and epialleleR, the tests were run 5 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.

Availability of Source Code and Requirements

The epialleleR R/Bioconductor package (biotools:epialleleR, RRID:SCR_023913) is freely available at https://bioconductor.org/packages/epialleleR/ and https://github.com/BBCG/epialleleR. The R scripts used in this manuscript are freely available at DataverseNO (https://doi.org/10.18710/2BQTJP).

Project name: epialleleR
Project homepage: https://github.com/BBCG/epialleleR
Bioconductor: https://bioconductor.org/packages/epialleleR/
Operating system: Linux, macOS, Windows
Programming language: R, C, C++
Other requirements: C++17, GNU make
License: Artistic-2.0
biotools: epialleleR
RRID: SCR_023913

Version 1.3.5 of the epialleleR R/Bioconductor package was used [32]. A previous version of this article was deposited in bioRxiv (doi: 10.1101/2022.06.30.498213) and the epialleleR has been applied in [12, 23] with data available at NCBI Gene Expression Omnibus under accession number GSE243966.

Additional Files

Supplementary Fig. S1. Scaled density of per-read beta values from all admixed samples combined, split by the level of admixed reads and genomic region of interest. The y-axis is scaled to 1 and limited to 0.015. The hypermethylated ($\beta \geq 0.5$) reads are increasingly apparent on the right sides of plots in accordance with an increase in admixed methylated reads.

Supplementary Fig. S2. Methylation patterns from all admixed samples combined, split by the level of admixed reads and genomic region of interest. Lines depict patterns, and open and closed circles depict unmethylated and methylated cytosines, respectively. Numbers on the right of every pattern indicate how many times each pattern occurs for every given gene/sample set. Due to very high number of methylation pattern types, only the most abundant pattern (if any) is shown for each range of average beta value: [0,0.2), [0.2,0.4), [0.4,0.6), [0.6,0.8), [0.8,1]. The hypermethylated ($\beta \geq 0.5$) patterns are increasingly apparent at the top of plots in accordance with an increase in admixed methylated reads.

Supplementary Fig. S3. Scaled density of per-read beta values from $n = 8$ real samples used to compare sensitivity of methylation profiling by NGS and array, split by sample and genomic region of interest. The y-axis is scaled to 1 and limited to 0.015. The population of hypermethylated ($\beta \geq 0.5$) reads that are apparent in the Fig. 3A in the main text are pointed to by black arrows.

Supplementary Fig. S4. Methylation patterns from $n = 8$ real samples used to compare sensitivity of methylation profiling by NGS and array, split by sample and genomic region of interest. Lines depict patterns, and open and closed circles depict unmethylated and methylated cytosines, respectively. Numbers on the right of every pattern indicate how many times each pattern occurs for every given gene/sample set. Due to very high number of methylation pattern types, only the most abundant pattern (if any) is shown for each range of average beta value: [0,0.2), [0.2,0.4), [0.4,0.6), [0.6,0.8), [0.8,1]. The hypermethylated ($\beta \geq 0.5$) patterns represent hypermethylated epialleles that are present in certain samples/regions, as shown in Fig. 3A in the main text and Supplementary Fig. S3.

Supplementary Fig. S5. (A) Distribution of per-read beta values for NGS reads covering CpGs that are common for NGS and BeadChip 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: boxplots of NGS-derived VEF values for individual CpGs; upper panel: significant aberrantly or differentially methylated regions identified by ramr or DMRcate, respectively, based on VEF values. (C) Lower panel: boxplots of NGS-derived beta values for individual CpGs; upper panel: significant aberrantly or differentially methylated regions identified by ramr or DMRcate, respectively, based on NGS-derived beta values. (B, C) The lower and upper hinges of boxes correspond to the first (Q1) and third (Q3) quartiles; the bar in the middle corresponds to the median value; the upper and lower whisker extend to Q3 + 1.5 ∗ IQR and Q1 − 1.5 ∗ IQR, respectively, while the values outside this range (outliers) are plotted as dots. Zero values are not plotted. The coloring is preserved for $n = 8$ samples used in Fig. 3 of the main text. The $n = 10$ samples used to create admixed samples and not included in Fig. 3 are plotted in light gray.

Supplementary Fig. S6. (A) Heatmap of minimum (equals absolute largest) difference in combinatorial entropy for all pairs of samples, split by genomic region of interest. (B) Heatmap of combinatorial entropy, epipolymorphism, fraction of discordant read pairs (FDRP), and proportion of discordant reads (PDR), split by genomic region of interest. (C) Heatmap of P values for pairwise comparison of samples using WSH scores, split by score and genomic region of interest. $***P < 0.001$, $**P < 0.01$, $*P < 0.05$, blank $P \geq 0.05$.

Supplementary Table S1. Complete metrics of cytosine reporting across all 3 possible cytosine genomic contexts (CHH, CHG, and CpG), obtained using simulated chr19 reads with varying sequencing error rate by selected tools. “reported,” number of cytosines present in the cytosine reports; “valid context,” number of cytosines for which genomic context was correctly identified; “invalid context,” number of cytosines for which genomic context was incorrectly identified; “not covered,” number of cytosines not present in a cytosine report; “mean coverage,” average coverage of all cytosines in this context; “mean” and “variance,” average value and variance for beta values of all cytosines in this context; “is 0.5,” number of cytosines with beta value of exactly 0.5 (ground truth for this dataset); “is not 0.5,” number cytosines with beta value not equal to 0.5.

Abbreviations

AMR: aberrantly methylated region; BAM: binary sequence alignment/map; BED: browser extensible data; CpG: cytosine followed by a guanine; DMR: differentially methylated region; FDR: false discovery rate; FDRP: fraction of discordant read pairs; IQR: interquartile range; NGS: next-generation sequencing; PDR: proportion of discordant reads; t-SNE: t-distributed stochastic neighbor embedding; VCF: variant call format; VEF: variant epiallele frequency; WBC: white blood cell; WGBS: whole-genome bisulfite sequencing; WSH: within-sample heterogeneity.

Authors’ Contributions

Conceived the project: O.N., P.E.L., S.K. Supervised the project: P.E.L., S.K. Conceived, designed, and implemented the software and the analysis pipeline: O.N. Wrote the paper: O.N., P.E.L., S.K. All authors read and approved the final manuscript.

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.

Data Availability

The data underlying accuracy and sensitivity analyses are freely available at DataverseNO [56]. Sensitive data used for the processing speed assessment are available from the authors in accordance with study protocols.

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.

Competing Interests

P.E.L. has received research funding for other projects from AstraZeneca, Novartis, Pfizer, and Illumina and honoraria through speaker’s bureaus 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 bureau honoraria from AstraZeneca, Pfizer, Novartis, and Pierre Fabre.

Ethics Approval and Consent to Participate

Ethics approvals and other relevant information for patient-generated data used in speed assessment were included and described in previous studies [12, 46, 53, 54]. All analyses of biomaterial were approved by Regional Ethics Committees for medical research, and all samples were collected after written informed consent from the sample donors (REK-vest Norway reference numbers: 3.2008.1932, 2015/1493, and 2018/1566).

References

1. Horsthemke B. Epimutations in human disease. Curr Top Microbiol Immunol 2006;310:45–59. https://doi.org/10.1007/3-540-31181-5_4.
2. Oey H, Whitelaw E. On the meaning of the word “epimutation.” Trends Genet 2014;30:519–20. https://doi.org/10.1016/j.tig.2014.08.005.
3. Kazanets A, Shorstova T, Hilmi K, et al. Epigenetic silencing of tumor suppressor genes: paradigms, puzzles, and potential. Biochim Biophys Acta 2016;1865:275–88. https://doi.org/10.1016/j.bbcan.2016.04.001.
4. Esteller M, Silva JM, Dominguez G, et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. J Natl Cancer Inst 2000;92:564–9. https://doi.org/10.1093/jnci/92.7.564.
5. Toffolatti L, Scquizzato E, Cavallin S, et al. MGMT promoter methylation and correlation with protein expression in primary central nervous system lymphoma. Virchows Arch 2014;465:579–86. https://doi.org/10.1007/s00428-014-1622-6.
6. Simpkins SB, Bocker T, Swisher EM, et al. MLH1 promoter methylation and gene silencing is the primary cause of microsatel-

lite instability in sporadic endometrial cancers. Hum Mol Genet 1999;8:661–6. https://doi.org/10.1093/hmg/8.4.661.

7. Veeck J, Ropero S, Setien F, et al. BRCA1 CpG island hypermethylation predicts sensitivity to poly(adenosine diphosphate)-ribose polymerase inhibitors. J Clin Oncol 2010;28:e563–4. https://doi.org/10.1200/JCO.2010.30.1010.
8. Yu W, Zhang L, Wei Q, et al. O6-Methylguanine-DNA methyltransferase (MGMT): challenges and new opportunities in glioma chemotherapy. Front Oncol 2019;9:1547. https://doi.org/10.3389/fonc.2019.01547.
9. Guastadisegni C, Colafranceschi M, Ottini L, et al. Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. Eur J Cancer 2010;46:2788–98. https://doi.org/10.1016/j.ejca.2010.05.009.
10. Lønning PE, Berge EO, Bjørnslett M, et al. White blood cell BRCA1 promoter methylation status and ovarian cancer risk. Ann Intern Med 2018;168:326. https://doi.org/10.7326/M17-0101.
11. Prajzendanc K, Domagała P, Hybiak J, et al. BRCA1 promoter methylation in peripheral blood is associated with the risk of triple-negative breast cancer. Int J Cancer 2020;146:1293–8. https://doi.org/10.1002/ijc.32655.
12. Lønning PE, Nikolaienko O, Pan K, et al. Constitutional BRCA1 methylation and risk of incident triple-negative breast cancer and high-grade serous ovarian cancer. JAMA Oncol 2022;8:1579. https://doi.org/10.1001/jamaoncol.2022.3846.
13. Sun R, Zhu P. Advances in measuring DNA methylation. Blood Sci 2022;4:8–15. https://doi.org/10.1097/BS9.0000000000000098.
14. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics 2011;27:1571–2. https://doi.org/10.1093/bioinformatics/btr167.
15. Maksimovic J, Phipson B, Oshlack A. A cross-package bioconductor workflow for analysing methylation array data. F1000Res 2016;5:1281. https://doi.org/10.12688/f1000research.8839.3.
16. Fortin J-P, Triche TJ Jr, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics 2017;33:558–60. https://doi.org/10.1093/bioinformatics/btw691.
17. Youk J, An Y, Park S, et al. The genome-wide landscape of C:g >T:a polymorphism at the CpG contexts in the human population. BMC Genomics 2020;21:270. https://doi.org/10.1186/s12864-020-6674-1.
18. Gu J, Stevens M, Xing X, et al. Mapping of variable DNA methylation across multiple cell types defines a dynamic regulatory landscape of the human genome. G3 (Bethesda) 2016;6:973–86. https://doi.org/10.1534/g3.115.025437.
19. Kint S, Spiegelaere WD, Kesel JD, et al. Evaluation of bisulfite kits for DNA methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR. PLoS One 2018;13:e0199091. https://doi.org/10.1371/journal.pone.0199091.
20. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinformatics 2021;3. https://doi.org/10.1093/nargab/lqab019.
21. Nikolaienko O, Lønning PE, Knappskog S. ramr: an R/bioconductor package for detection of rare aberrantly methylated regions. Bioinformatics 2021;38:133–40. https://doi.org/10.1093/bioinformatics/btab586.
22. Hofmeister BT, Lee K, Rohr NA, et al. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. Genome Biol 2017;18:155. https://doi.org/10.1186/s13059-017-1288-x.
23. Nikolaienko O, Eikesdal HP, Gilje B, et al. Prenatal BRCA1 epimutations contribute significantly to triple-negative breast cancer development. medRxiv. https://doi.org/10.1101/2023.05.14.23289949.
24. Kondrashova O, Topp M, Nesic K, et al. Methylation of all BRCA1 copies predicts response to the PARP inhibitor rucaparib in ovarian carcinoma. Nat Commun 2018;9:3970. https://doi.org/10.1038/s41467-018-05564-z.
25. Nesic K, Kondrashova O, Hurley RM, et al. Acquired RAD51C promoter methylation loss causes PARP inhibitor resistance in high-grade serous ovarian carcinoma. Cancer Res 2021;81:4709–22. https://doi.org/10.1158/0008-5472.CAN-21-0774.
26. Hurley RM, McGehee CD, Nesic K, et al. Characterization of a RAD51C-silenced high-grade serous ovarian cancer model during development of PARP inhibitor resistance. NAR Cancer 2021;3. https://doi.org/10.1093/narcan/zcab028.
27. Qi L, Teschendorff AE. Cell-type heterogeneity: why we should adjust for it in epigenome and biomarker studies. Clin Epigenet 2022;14:31. https://doi.org/10.1186/s13148-022-01253-3.
28. Liang L, Cookson WOC. Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies. Hum Mol Genet 2014;23:R83–8. https://doi.org/10.1093/hmg/ddu284.
29. Huh I, Wu X, Park T, et al. Detecting differential DNA methylation from sequencing of bisulfite converted DNA of diverse species. Briefings Bioinf 2019;20:33–46. https://doi.org/10.1093/bib/bbx077.
30. Anastasiadi D, Esteve-Codina A, Piferrer F. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. Epigenetics Chromatin 2018;11:37. https://doi.org/10.1186/s13072-018-0205-1.
31. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2023.
32. Nikolaienko O. epialleleR: fast, epiallele-aware methylation caller and reporter. https://doi.org/10.18129/B9.bioc.epialleleR. Accessed 8 October 2023.
33. Fowler G, Noll LC, Vo K-P, et al. The FNV Non-Cryptographic Hash Algorithm. Internet Engineering Task Force. Report No. draft-eastlake-fnv-20. https://datatracker.ietf.org/doc/draft-eastlake-fnv/20/. Accessed 8 October 2023.
34. Bonfield JK, Marshall J, Danecek P, et al. HTSlib: c library for reading/writing high-throughput sequencing data. Gigascience 2021;10:giab007. https://doi.org/10.1093/gigascience/giab007.
35. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 2012;13:R87. https://doi.org/10.1186/gb-2012-13-10-r87.
36. Sun Z, Vaisvila R, Hussong L-M, et al. 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 2021;31:291–300. https://doi.org/10.1101/gr.265306.120.
37. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30:1363–9. https://doi.org/10.1093/bioinformatics/btu049.
38. Peters TJ, Buckley MJ, Chen Y, et al. Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate. Nucleic Acids Res 2021;49:e109. https://doi.org/10.1093/nar/gkab637.
39. Scherer M, Nebel A, Franke A, et al. Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. Nucleic Acids Res 2020;48:e46. https://doi.org/10.1093/nar/gkaa120.

40. Li S, Garrett-Bakelman F, Perl AE, et al. Dynamic evolution of clonal epialleles revealed by methclone. Genome Biol 2014;15:472. https://doi.org/10.1186/s13059-014-0472-5.
41. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One 2012;7:e41361. https://doi.org/10.1371/journal.pone.0041361.
42. Bakulski KM, Feinberg JI, Andrews SV, et al. DNA methylation of cord blood cell types: applications for mixed cell birth studies. Epigenetics 2016;11:354–62. https://doi.org/10.1080/15592294.2016.1161875.
43. Lønning PE, Eikesdal HP, Løes IM, et al. Constitutional mosaic epimutations—a hidden cause of cancer? Cell Stress 2019;3:118–35. https://doi.org/10.15698/cst2019.04.183.
44. Nikolaienko O. "epialleleR: an R/BioC package for sensitive allele-specific methylation analysis in NGS data. Bioconductor. 2022. https://doi.org/doi:10.18129/B9.bioc.epialleleR. Accessed 8 October 2023.
45. epialleleR GitHub. https://github.com/BBCG/epialleleR. Accessed 8 October 2023.
46. Knappskog S, Bjørnslett M, Myklebust LM, et al. The MDM2 promoter SNP285C/309G haplotype diminishes Sp1 transcription factor binding and reduces risk for breast and ovarian cancer in Caucasians. Cancer Cell 2011;19:273–82. https://doi.org/10.1016/j.ccr.2010.12.019.
47. Knappskog S, Gansmo LB, Romundstad P, et al. MDM2 promoter SNP344T>A (rs1196333) status does not affect cancer risk. PLoS One 2012;7:e36263. https://doi.org/10.1371/journal.pone.0036263.
48. Krueger F. Sherman—bisulfite-treated Read FastQ Simulator. https://www.bioinformatics.babraham.ac.uk/projects/sherman/. Accessed 8 October 2023.
49. Harris RA, Raveendran M, Worley KC, et al. Unusual sequence characteristics of human chromosome 19 are conserved across 11 nonhuman primates. BMC Evol Biol 2020;20:33. https://doi.org/10.1186/s12862-020-1595-9.
50. Krijthe JH. Rtsne: T-Distributed Stochastic Neighbor Embedding Using Barnes-Hut Implementation. https://github.com/jkrijthe/Rtsne. Accessed 8 October 2023.
51. Jaffe AE. FlowSorted.Blood.450k: Illumina HumanMethylation Data on Sorted Blood Cell Populations. Bioconductor. 2023. https://doi.org/doi:10.18129/B9.bioc.FlowSorted.Blood.450k
52. Andrews SV, Bakulski KM. FlowSorted.CordBlood.450k: Illumina 450k Data on Sorted Cord Blood Cells. Bioconductor. 2023. https://doi.org/doi:10.18129/B9.bioc.FlowSorted.CordBlood.450k.
53. Poduval DB, Ognedal E, Sichmanova Z, et al. Assessment of tumor suppressor promoter methylation in healthy individuals. Clin Epigenetics 2020;12:131. https://doi.org/10.1186/s13148-020-00920-7.
54. Eikesdal HP, Yndestad S, Elzawahry A, et al. Olaparib monotherapy as primary treatment in unselected triple negative breast cancer. Ann Oncol 2021;32:240–9. https://doi.org/10.1016/j.annonc.2020.11.009.
55. Zhou L, Ng HK, Drautz-Moses DI, et al. Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. Sci Rep 2019;9:10383. https://doi.org/10.1038/s41598-019-46875-5.
56. Nikolaienko O. Replication data for: "epialleleR: an R/BioC package for quantifying and analysing low-frequency DNA methylation." 2022. DataverseNO, V3. https://doi.org/10.18710/2BQTJP.

Received: June 1, 2023. **Revised:** August 3, 2023. **Accepted:** September 26, 2023