Nucleic Acids Research, 2025, **53**, gkaf970
https://doi.org/10.1093/nar/gkaf970
Computational Biology

Pipeline Olympics: continuable benchmarking of computational workflows for DNA methylation sequencing data against an experimental gold standard

Yu-Yu Lin[1], Kersten Breuer[1], Dieter Weichenhan[1], Pascal Lafrenz[1], Antonella Sarnataro[2], Agata Wilk[1], Maryna Chepeleva[3], Oliver Mücke[1], Maximilian Schönung[4,5], Franziska Petermann[6], Philip Reiner Kensche[7], Lena Weiser[7], Frank Thommen[7], Gideon Giacomelli[8], Karl Nordstroem[9,10], Edahi Gonzalez-Avalos[11,12], Angelika Merkel[13,14], Helene Kretzmer[15,16,17], Jonas Fischer[18], Stephen Krämer[4,19,20], Murat Iskar[21,22], Stephan Wolf[6], Ivo Buchhalter[7], Manel Esteller[14,23,24,25], Christian Lawerenz[8,26], Sven Twardziok[8], Marc Zapatka[21], Volker Hovestadt[27,28,29], Matthias Schlesner[19], Marcel H. Schulz[30,31], Steve Hoffmann[15,32], Clarissa Gerhauser[1], Jörn Walter[9], Mark Hartmann[4,5], Daniel B. Lipka[4,5,33,34], Yassen Assenov[35], Christoph Bock[36,37], Christoph Plass[1], Reka Toth[1,3,*,†], Pavlo Lutsik[1,2,*,†]

[1]Division of Cancer Epigenomics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[2]Department of Oncology, KU Leuven, 3000 Leuven, Belgium
[3]Multiomics Data Science Research Group, Department of Cancer Research, Luxembourg Institute of Health, 1445 Strassen, Luxembourg
[4]Section of Translational Cancer Epigenomics, Division of Translational Medical Oncology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[5]National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, 69120 Heidelberg, Germany
[6]NGS Core Facility, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[7]Omics-IT and Data Management Core Facility (ODCF), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[8]Berlin Institute of Health, University Hospital Charité, 10178 Berlin, Germany
[9]Department of EpiGenetics, Saarland University, 66123 Saarbrücken, Germany
[10]Present Address: Astra Zeneca, 431 83 Mölndal, Sweden
[11]Division of Signaling and Gene Expression, La Jolla Institute for Immunology, La Jolla, CA 92037, United States
[12]Present Address: Guardant Health, Redwood City, CA 94063, Unites States
[13]Bioinformatics Development and Statistical Genomics, Center for Genomic Regulation (CRG), Centro Nacional de Análisis Genómico (CNAG), 08028 Barcelona, Spain
[14]Josep Carreras Leukaemia Research Institute (IJC),08916 Badalona, Catalonia, Spain
[15]LIFE—Leipzig Research Center for Civilization Diseases, University Leipzig, 04103 Leipzig, Germany
[16]Max-Planck Institute for Molecular Genetics, 14195 Berlin, Germany
[17]Present address: Digital Health Cluster, Hasso Plattner Institute for Digital Engineering, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany
[18]Department for Computer Vision and Machine Learning, Max Planck Institute for Informatics, 66123 Saarbrücken, Germany
[19]Biomedical Informatics, Data Mining and Data Analytics, Faculty of Applied Computer Science and Medical Faculty, University of Augsburg, 86159 Augsburg, Germany
[20]Faculty of Biosciences, Heidelberg University, 69120 Heidelberg, Germany
[21]Division of Molecular Genetics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
[22]Present address: Friedriech Micher Institute, 4056 Basel, Switzerland
[23]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Catalonia, Spain
[24]Centro de Investigación Biomedica en Red Cancer (CIBERONC), 28029 Madrid, Spain
[25]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), 08036 Barcelona, Catalonia, Spain
[26]Present address: Steinbeis-Innovationszentrum Digital Health, 47799 Krefeld, Germany
[27]Broad Institute of MIT and Harvard, Cambridge, MA 02142, United States
[28]Department of Pediatric Oncology, Dana–Farber Cancer Institute, Boston, MA 02215, United States
[29]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA 02115, United States
[30]Institute for Computational Genomic Medicine, Goethe University, 60590 Frankfurt-am-Main, Germany
[31]German Center for Cardiovascular Research, 60590 Frankfurt-am-Mein, Germany

Received: December 14, 2024. **Revised:** August 5, 2025. **Accepted:** August 24, 2025

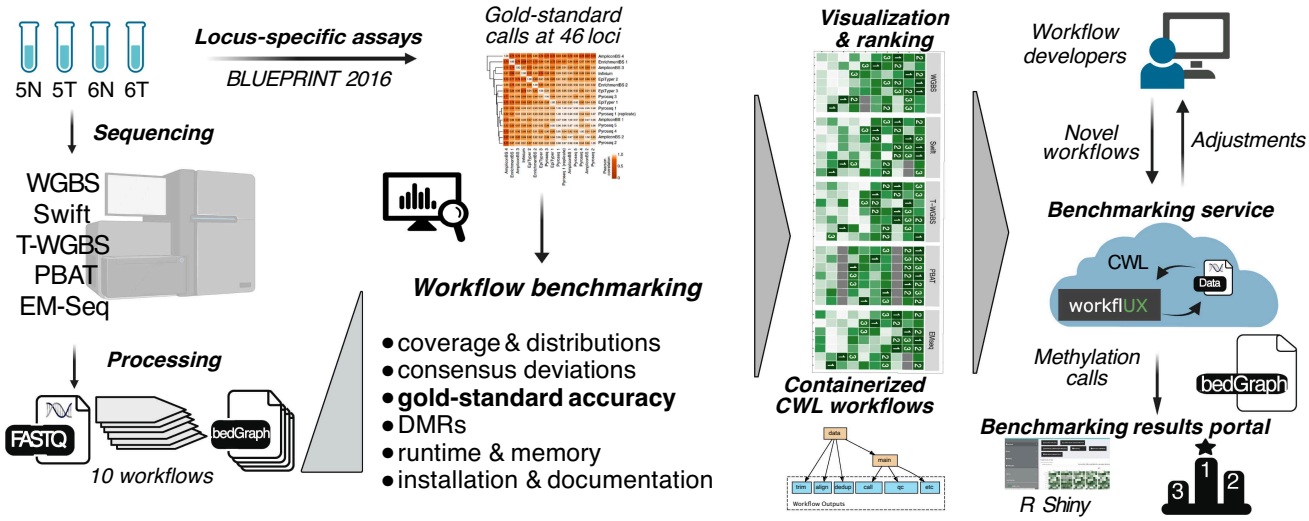
[32]Computational Biology Group, Leibniz Institute on Aging—Fritz Lipmann Institute and Friedrich Schiller, University of Jena, 07745 Jena, Germany
[33]Faculty of Medicine, Otto-von-Guericke-University, 39120 Magdeburg, Germany
[34]German Cancer Consortium (DKTK), 69120 Heidelberg, Germany
[35]Department of Mathematics and Technology, RheinAhrCampus Remagen, University of Applied Sciences Koblenz, 53424 Remagen, Germany
[36]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria
[37]Institute of Artificial Intelligence, Center for Medical Data Science, Medical University of Vienna, 1090 Vienna, Austria

*To whom correspondence should be addressed. Email: pavlo.lutsik@kuleuven.be
Correspondence may also be addressed to Reka Toth. Email: reka.toth@lih.lu
†The last two authors should be regarded as Joint Last Authors.

Abstract

DNA methylation is a widely studied epigenetic mark and a powerful biomarker of cell type, age, environmental exposures, and disease. Whole-genome sequencing following selective conversion of unmethylated cytosines into thymines via bisulfite treatment or enzymatic methods remains the reference method for DNA methylation profiling genome-wide. While numerous software tools facilitate processing of DNA methylation sequencing reads, a comprehensive benchmarking study has been lacking. In this study, we systematically compared complete computational workflows for processing DNA methylation sequencing data using a dedicated benchmarking dataset generated with five whole-genome profiling protocols. As an evaluation reference, we employed accurate locus-specific measurements from our previous benchmark of targeted DNA methylation assays. Based on this experimental gold-standard assessment and multiple performance metrics, we identified workflows that consistently demonstrated superior performance and revealed major workflow development trends. To ensure the long-term utility of our benchmark, we implemented an interactive workflow execution and data presentation platform, adaptable to user-defined criteria and readily expandable to future software.

Graphical abstract



Introduction

DNA methylation is a key epigenetic modification [1] that plays an essential role in development [2] and cell differentiation [3, 4] across many species including human. DNA methylation landscapes are altered in the course of mitotic divisions [5] and transition to cellular senescence [6], during aging [7, 8], as well as in pathological conditions including cancer [9–13] and other diseases [14, 15]. The higher stability of DNA methylation relative to gene expression, and its simpler analysis compared to other epigenomic marks contribute to the attractiveness of DNA methylation as an epigenetic biomarker of age [16], for early detection of cancer in liquid biopsies [17] and in forensic assays [18].

In eukaryotes, DNA methylation occurs predominantly at CpG dinucleotides. Numerous methods have been proposed to measure CpG methylation patterns, as extensively reviewed [19–22] and evaluated in dedicated benchmarking studies [23–29]. The most comprehensive is whole-genome bisulfite sequencing (WGBS), which provides a genome-wide, single-base pair, and single-strand resolution method based on the bisulfite conversion of unmethylated cytosines [30]. The Illumina Infinium microarrays [31, 32] as well as reduced representation bisulfite sequencing [33] provide additional genome-scale alternatives, measuring 2%–15% of the CpG sites. These methods have to be distinguished from targeted assays, such as amplicon bisulfite sequencing and bisulfite pyrosequencing [23, 34]. Furthermore, third-generation sequencing methods, such as nanopore or single-molecule real-time sequencing, are able to read out modified bases via direct sequencing of native DNA and are bisulfite-free [35].

Bisulfite treatment results in the chemical deamination of unmethylated cytosines and subsequently their change to

thymines. This induces DNA fragmentation and degradation, thus requiring high amounts of DNA input [26]. To overcome this issue, a variety of enhanced protocol variants for moderate- to low-input DNA amounts have been suggested, including tagmentation-based WGBS (T-WGBS) [36–38] and post-bisulfite adaptor tagging (PBAT) [39]. The former increases the efficiency of standard adaptor tagging, whereas the latter utilizes PBAT to avoid subsequent degradation of adaptor-tagged fragments. Finally, enzymatic methods such as enzymatic methyl-seq (EM-seq) replace bisulfite treatment with an enzymatic conversion step, which reduces DNA fragmentation and degradation [40]. In bisulfite sequencing protocols currently in use, the actual sequencing reads correspond to a converted version of the original top (OT) and original bottom (OB) strands. In some protocols, such as PBAT, the complementary strands, CTOT and CTOB, respectively, are predominantly sequenced [39]. These protocol-specific differences require special attention during data processing [29].

Analysis of bisulfite sequencing data generally includes four core steps: (i) read processing, including quality control and trimming; (ii) conversion-aware alignment; (iii) post-alignment processing or filtering; and (iv) calling of methylation states (and, optionally, of genotypes and structural variation). Many tools have been developed for each step, providing room for an overwhelming number of possible combinations and workflows. Read preprocessing includes basic quality control, such as read length and sequencing quality score distributions using e.g. the FastQC tool, standard read trimmers [41, 42], bisulfite alignment, alignment post-processing, as well as methylation calling and quantification. Methods to account for bisulfite conversion during the alignment include the no-cytosine three-letter alphabet [43–49], a wild card alignment [50–53], or a wild card-related approach that transforms the alignment into an asymmetric mapping problem [54]. In the three-letter approach, all cytosines in both the reference genome and the sequencing reads are converted to thymines, and mapped using a seed and extend approach [55]. Wild card aligners map both cytosines and thymines in the reads to cytosines in the reference genome. Post-processing includes filtering polymerase chain reaction (PCR) duplications with conventional tools as well as other quality filtering steps, e.g. filtering by alignment quality. Finally, calling and quantification of methylation states range from simple read count ratios [56, 57] to Bayesian model-based approaches [58] featuring local realignment [59]. Some methylation callers provide add-on functionalities, such as sequence variant calling [58–61].

Although numerous methods of profiling methylation have been proposed, no comprehensive attempt has been made at evaluating end-to-end data processing workflows. Previous benchmarks focused on a single processing task (typically the alignment), assessed relatively few methods, and lacked gold-standard control datasets [62–64].

Here, we present a comprehensive benchmark of data processing workflows for DNA methylation sequencing. Our study is based on gold-standard samples with highly accurate DNA methylation calls. We evaluated the workflows in the context of one standard, three low-input, and one EM-seq protocols. To simplify the choice of workflows for the users and enable seamless extension to future tools and workflows, we developed web resources for interactive data presentation and continuous and sustainable ("living") benchmarking.

Materials and methods

Workflow selection and deployment

We thoroughly reviewed the literature and publicly available software that falls within the scope of processing DNA methylation sequencing data (Supplementary Table S1). The main criteria to select the workflows for the benchmarking study were as follows: The latest update time was after 2020, and the number of citations per year was at least 10. In addition to the qualifying list, we added two recent workflows, *Biscuit* and *FAME*, and one well-established workflow frequently used by collaborators (*BAT*). The final selected workflows are the following: *BAT* [65], *Biscuit* [66], *Bismark* [45], *BSBolt* [67], *bwa-meth* [56], *FAME* [54], *gemBS* [58], *GSNAP* [68], *methylCtools* [57], and *methylpy* [69]. Of note, Bismark and *bwa-meth* workflows (with the exception of the read trimmer) are equivalent to the two variants of the Nextflow nf-core workflow *methylseq* [70]. Containerization and the Common Workflow Language (CWL) were used to enhance stability and reusability of all workflows. Each component of the workflow was implemented as a Docker container, and the pipeline was implemented in a CWL wrapper. The CWL ran on a standardized virtual machine (VM) equipped with the following specifications: CentOS 7.9, 2 × 14-core Intel Xeon E5-2660 v4 (56 threads), and 512 GB RAM. The processing times and maximum memory requirements were collected from the job notification reports of the IBM Spectrum LSF (load sharing facility) platform.

Comparison of installation options and documentation quality

We assessed the workflows with respect to the following criteria: installation instructions, main installation method, availability of a *conda* package, public container images, dependency management, quick start guide, availability of tutorials, command line reference, example dataset, workflow integration support, error handling, and community support. The ranking score scheme is given in Supplementary Table S1.

Sample acquisition

In our study, we obtained genomic DNA isolated from two pairs of fresh-frozen colon cancer tissue samples with adjacent normal from the BLUEPRINT technology benchmarking study (Patients 5 and 6) [23]. A detailed description of the samples and the DNA isolation protocol are available in the original study.

Library preparation and sequencing

We used five different whole-methylome sequencing protocols as outlined in Supplementary Table S2. Library preparation of T-WGBS and PBAT was performed at the Division of Cancer Epigenomics, while library preparation for WGBS, Swift, and EM-seq as well as the DNA sequencing were performed by the Genomics and Proteomics Core Facility at the German Cancer Research Center (DKFZ). All kits were used following the manufacturer's instructions, unless otherwise specified. It is important to note that the sequencing libraries were not prepared at the same time due to different protocols becoming available at a later time point, and this delay might have resulted in differences in sample degradation, especially affecting sample 6N in EM-seq.

Whole-genome bisulfite sequencing

WGBS was essentially performed as described by Lister *et al.* [30] using the EpiTekt Bisulfite Kit (Qiagen) and the TruSeq DNA Sample Prep Kit (Illumina). In brief, 2 μg genomic DNA from each sample was fragmented, end repaired, and 3′-dA tailed, and methylated adapters were appended by ligation. Size-selected ligation products were bisulfite treated, and sequencing libraries were generated with a low number of PCR cycles. Two lanes per sample were used on an Illumina HiSeq X Ten sequencing machine.

Swift bio Accel-NGS

The Swift protocol [71], as an alternative to PBAT, utilized Swift's proprietary Adaptase instead of random priming. Two hundred nanograms DNA was used to perform bisulfite treatment followed by library preparation using the Accel-NGS-Methyl-Seq Kit (Swift Bio). For sequencing, one lane on the HiSeq X Ten was used for each sample. Hereafter, we refer to this protocol as Swift.

Tagmentation-based whole-genome bisulfite sequencing

The details of the T-WGBS protocol have been described elsewhere [38]. Here, we used an input of 30 ng DNA. Four independent libraries were constructed for each sample to reduce the impact of the PCR amplification. All libraries and samples were equally allocated and sequenced on two HiSeq2000 (Illumina) lanes.

Post-bisulfite adaptor tagging with random priming

PBAT libraries were prepared as previously described [72] using a customized protocol for ultralow-input materials based on the single-cell bisulfite sequencing protocol [73]. In brief, 6 ng of purified DNA was subjected to bisulfite conversion, a single pre-amplification for 90 min at 37°C, adaptor tagging, and finally 14 cycles of PCR. Libraries were purified using double 0.7 × SPRISelect size selection kit application and sequenced on a HiSeq X Ten sequencer, applying 150 base pairs (bp) paired-end sequencing at the DKFZ NGS Core Facility in Heidelberg.

NEBNext Enzymatic Methyl-seq based on TET and APOBEC

Bisulfite-free library preparation was performed with the NEBNext Enzymatic Methyl-seq (EM-seq) Kit [40] using 50 ng of DNA. Each sample was sequenced using one lane on a HiSeq X Ten.

The detailed information about all read sets generated in this study including the list of FASTQ files, along with the total number of sequences and bases, read lengths, mean PHRED scores, non-CpG methylation levels, conversion rates, as measured by *Bismark* are given in Supplementary Table S2.

Protocol-specific differences in the processing workflows

All computational workflows consisted of the following steps: read preprocessing, alignment, post-processing, and methylation calling. A detailed description of the workflows, version numbers, and individual steps, and the protocol-specific parameter settings adapted for each protocol are described in the following sections, with detailed information listed in Supplementary Table S1.

Trimming

Following the manufacturer's recommendation, the random 3′-tails added by the Adaptase during the Swift protocol were removed. The removed segments were the last 15 bp of the R1 and the first 15 bp of the R2. In T-WGBS, the Tn5 transposase creates a short 9 bp gap at the 3′-end cutting site, which is subsequently repaired by DNA polymerase and DNA ligase. During this repair process, nine base pairs are incorporated at the end of the R2 using unmethylated cytosines. Therefore, these base pairs should not be used for methylation calling but can be used for alignment. The unmethylated regions were not considered during methylation calling in the workflows supporting this, like *Bismark*, *bwa-meth*, *GSNAP*, and *methylCtools*. For the others, additional hard trimming was applied to remove the base pairs before alignment.

Modified PBAT method used in this study is a nondirectional protocol that requires special handling of the reads. *BAT* works only with directional protocols; therefore, it was not run with our PBAT data. Although *gemBS* supports nondirectional protocols, we were faced with an error that we were unable to resolve with the help of the authors until the manuscript was written. Thus, *gemBS* was not run on PBAT data. During preprocessing, random hexamers (first and last 6 bps of R1 and R2), which were added during the two steps of random priming, were removed. *MethylCtools* provides a patch script to support PBAT, which determines the strands of read pairs R1 and R2 obtained from a nondirectional protocol (available from https://github.com/cimbusch/TWGBS). We ran this script on PBAT raw data to create the input for *methylCtools*.

Alignment

Bisulfite treatment of DNA, followed by PCR amplification, can produce four (bisulfite-converted) strands for a given locus. Depending on the adapter used, two different approaches were employed for the construction of BS-Seq libraries. In directional bisulfite sequencing, the library construction process selectively sequences either the OT strand or the OB strand of the DNA. Consequently, only reads aligned to OT and OB are considered, whereas alignments from the complementary strands (CTOT and CTOB) are typically not generated. In contrast, nondirectional bisulfite sequencing includes all four strands generated in the bisulfite-converted PCR process (OT, CTOT, OB, and CTOB) in the sequencing library with roughly equal likelihood, which means that alignments to all four strands are considered valid. When executing the analysis for PBAT, we configured the directional sequencing parameters for the four protocols using nondirectional sequencing. In all analyses, GRCh38 human genome assembly was used, which also included Phi X174 and Lambda phage sequences to facilitate the use of respective bisulfite conversion control spike-in in certain protocols.

Removal of PCR duplicates

All workflows, except *BAT*, included a duplicate removal step. The principle is to perform duplicate removal on the libraries independently. For WGBS, the two lanes sequencing the same library were aligned separately, merged, and deduplicated. For T-WGBS, which featured four independent libraries, the libraries were aligned and deduplicated individually.

Methylation calling

As mentioned in the "Trimming" section, some workflows have the possibility to ignore bases at the end of the reads during methylation calling. Apart from T-WGBS, other protocols might also benefit from such an approach. Ideally, we expect the probability of observing a methylated C to be constant across any given read. However, an increase or decrease in methylation level is often observed, especially at the end or at the beginning of the reads. This methylation bias or M-bias can lead to an overestimation of DNA methylation. Therefore, during methylation calling, the methylation status of the last several base pairs of the reads should be ignored. The exact length of the ignored region depends on the protocol and other technical parameters. Two workflows provide a function to determine the sequence affected by M-bias. The caller of *bwa-meth*, *MethylDackel* provides an automatic M-bias identification tool (*MethylDackel* mbias), whereas *Bismark* provides a function (coverage2cytosine) with R scripts to generate M-bias plots. Therefore, we applied sample-specific alignment trimming for these two workflows. The exact number of base pairs subjected to alignment trimming is listed in Supplementary Table S3. Other methylation callers do not specifically address the M-bias.

Statistical testing

In the genome-wide deviation analysis, we employed a paired *t*-test to determine whether the beta values of the protocol exhibited consistent underestimation tendencies. Similarly, for the deviation of preselected loci, we conducted a two-way analysis of variance (ANOVA) to assess the joint impact of samples and workflow procedures on accuracy. In the context of assessing the correlation between alignment rate and whole-genome deviation in PBAT, we utilize the "corr.test" package in R to calculate the Spearman correlation and test for its significance.

Benchmarking metrics

Methylation calls in .bed or .bedgraph format were imported using the R package *methrix* [74], giving *BSgenome.Hsapiens.UCSC.hg38* as a reference. All downstream analyses were performed on the summarized methylation and coverage matrices using R 4.1.

The area under the depth-versus-coverage dependency curves

The area under the curve was calculated based on the dependency of the coverage fraction at a given cutoff, where the x axis is the read coverage threshold, and the y axis is the cumulative fraction of covered genomic CpGs. The area under the curve (AUC) score only considers read coverage from 0 to 200 and is normalized to 1 by dividing it by 200.

Genome-wide deviation from data-driven consensus

We utilized a data-driven approach to create a consensus corridor for all the CpGs. This was achieved by employing three high-coverage protocols, including WGBS, Swift, and EM-seq. The consensus corridor was defined as the smallest region encompassing at least five measurements from each protocol. The mean absolute deviation from the pre-created reference among all CpGs determines the ranking of workflows.

Deviation from the consensus corridor of 46 preselected loci

To establish consensus methylation calls without the need of using simulated data, we preselected loci from a previous benchmarking study conducted by the BLUEPRINT Consortium [23]. In this study, 48 regions (16 mandatory and 32 recommended) were selected based on genome-wide methylation screen, and each selected region was analyzed using multiple technologies by different labs. We excluded two regions (recommended 29 and 30) due to a lack of data points ($\leq$3) leaving 46 regions, summarized in Supplementary Table S1. Following the original approach, we identified the consensus corridor as the narrowest interval, with measurements from three different technologies, adding a 5% flanking region. The absolute deviation was calculated for each method, workflow, and sample, based on each region. The difference $d_{w_i s_j r_k}$ was calculated as follows:

$$d_{w_i s_j r_k} = \begin{cases} 0, & m_{w_i s_j r_k} \in (C_u; C_l), \\ \min\left(abs(m_{w_i s_j r_k} - C_u); abs(m_{w_i s_j r_k} - C_l)\right), & m_{w_i s_j r_k} \notin (C_u; C_l), \end{cases}$$

where $m_{w_i s_j r_k}$ is the methylation level of the workflow i and sample j at the region k, C_u and C_l are the upper and the lower border of the consensus corridor. The mean absolute deviation from the consensus corridor of each protocol–workflow pair was then calculated as $\frac{\sum_{j=1,\ k=1}^{M,\ P} d_{s_j r_k}}{M \times P}$, where M and P are the number of samples and regions, respectively.

Differential methylation analysis

Combined metrics were established to assess the accuracy of differential methylation detection. First, it was treated as a classification problem and evaluated using the standard weighted area under the curve metric. The targeted methylation assay benchmarking study conducted by the BLUEPRINT Consortium also made sample-matched data from Illumina HumanMethylation450 arrays available. This dataset comprised six pairs of colon tumor–normal samples. Two of these pairs were included in our study; therefore, we used these data to estimate the accuracy of differential methylation calling on the sequencing-based results. Differential analysis of the arrays was performed using *limma* [75], as implemented in the *RnBeads* R package [76], in a paired setting. We used *DSS* [77] to analyze differential methylation of the sequencing-based data in a nonpaired setting because *DSS* required at least three samples for the paired mode. In both differential lists, a false discovery rate of 0.1 was used as the threshold to determine significant differences. The AUC score for the hypermethylation and hypomethylation events was calculated separately and combined into a weighted score with the number of events (see subsection on differential methylation scoring in the "Results" section below). The second part of the metrics is the Pearson correlation of delta beta values (beta-value difference between normal and tumor samples) from the array and sequencing. The beta value of the microarray was extracted using RnBeads, and the correlation between the delta values of the two patients was used as a metric.

Compute runtime and memory usage

The DKFZ Scientific Computing Center allocated a few exclusive computing nodes for the project to ensure an identical and isolated environment. The computing nodes were equipped with 56 CPUs and 256 GB RAM. To estimate the CPU time and maximal RAM usage, we ran all protocol–workflow pairs on a normal sample of Patient 5 in this environment. The tasks

were submitted to the IBM Spectrum LSF platform. The values of "Run Time" and "Max Memory" in the job notification report were used as the metrics. Unfortunately, *Bismark* ran prohibitively slowly in this cluster environment with distributed network storage compared to single-node mode; therefore, we estimated the values through a down-sampling approach, including six subsamples ranging from 5% to 30%, with increments of 5% each. We processed these subsamples using *Bismark* and, based on the results, applied linear regression to estimate the overall execution time and memory usage.

Ranking

The rank average of all the measurements was used to summarize the results of the benchmarking study. For each metric, the rank scale ranged from 1 to 10, with 1 indicating the best. Each workflow has a rank score calculated by averaging all measurements across the five protocols.

$$S_w = \sum R_{w_i p_j m_k}.$$

The rank average is calculated as follows, where the $R_{w_i p_j m_k}$ is the rank of workflow i in protocol j, and metric k. If multiple workflows have the same rank average, then the average z-score for each metric is used as the secondary ranking basis.

Shiny app details

We utilized Shiny, an R package that simplifies the creation of interactive web applications and dashboards, to construct a rich data website to visualize and share the main findings of our benchmarking study, allowing users to customize how they access results according to their specific needs. We provided various statistics, such as methylation and coverage, along with detailed visualizations of 46 gold-standard loci and a customizable ranking table. The shiny application is available at https://compepigen.github.io/PipelineOlympics/shiny/.

workflUX server details

workflUX, formerly known as CWLab, is an open-source web application designed to streamline the deployment of big data workflows. Its standout features include platform versatility and seamless operation on Linux, MacOS, and Windows, ensuring compatibility with the preferred operating system. Furthermore, it offers support for containerization, including Docker, singularity, and udocker, enabling efficient dependency management and simplifying the workflow deployment process. workflUX seamlessly integrates with a range of CWL runners, such as cwltool, Toil, Cromwell, Reana, and CWLEXEC, empowering users to execute CWL workflows across various infrastructures, from single workstations to HPC clusters and cloud platforms. We implemented automated benchmarking using workflUX. Workflow developers are required to implement their workflows using CWL. The workflUX server utilizes downscaled versions of original datasets and related CWL job configurations to create an automated benchmarking service freely available at https://compepigen.github.io/PipelineOlympics/workflux/.

Results

Systematic review and selection of benchmarked software and workflows

We conducted a comprehensive literature search and reviewed published software tools for bisulfite sequencing data processing (Supplementary Table S1). We focused on complete workflows covering processing steps from raw reads to DNA methylation calls and excluded those that were not open source or not regularly maintained (see the "Materials and methods" section for details). Altogether, we included 10 workflows into our study: *BAT* [65], *Biscuit* [66], *Bismark* [45], *BSBolt* [67], *bwa-meth* [56], *FAME* [54], *gemBS* [58], *GSNAP* [68], *methylCtools* [57], and *methylpy* [69] (Fig. 1A). With this selection, we cover different approaches to bisulfite alignment (three-letter and wildcard) and DNA methylation calling (counting and Bayesian based).

Benchmarking study design and dataset

We selected tumor–normal sample pairs from two colon cancer patients originally used for benchmarking locus-specific methylation profiling technologies by the BLUEPRINT Consortium [23] (Fig. 1B). In this study, a few systematically selected genomic regions were profiled in a range of samples, including six colon tumor/adjacent normal sample pairs, and multiple labs performed 16 targeted DNA methylation assays, including AmpliconBS, EnrichmentBS, EpiTyper, Infinium, and Pyroseq. The combination of multiple targeted arrays and the collaboration of multiple labs resulted in highly accurate DNA methylation calls. This study established consensus corridors for the true DNA methylation levels at the assayed sites, which we use here as the gold-standard loci set for benchmarking. We sequenced all four samples using a representative set of five methyl-seq protocols: one standard (WGBS), three low input (Swift, T-WGBS, PBAT), and a bisulfite-free enzymatic protocol (EM-seq) (Supplementary Fig. S1). We obtained high-quality raw sequencing data ranging from 300 312 952 (PBAT, 6T) to 985 933 822 (WGBS, 6T) read pairs per sample (Supplementary Table S2). We implemented all selected workflows in CWL (https://commonwl.org [78]) and ran them on a dedicated VM in a fully controlled computational environment. The resulting methylation calls were aggregated and summarized using *methrix* [74] and evaluated using a range of criteria to establish a consistent ranking. Finally, we set up a cloud-based infrastructure to establish an environment that will enable researchers to continue benchmarking using their own workflows.

To understand the challenges each library preparation protocol poses for processing workflows, we first assessed the major properties of the generated data. Thus, we examined the data generated by each protocol to assess the extent of variation between sequencing protocols that might affect data quality and processing. We used median measurements from all workflows for visualization. All protocols generated high-quality reads, and quality trimming led to the loss of <2% of the data. Alignment rates were at least 92%, except for PBAT, where it dropped to 74%. Between 13% and 28% of the aligned reads were identified as PCR duplicates and removed (Supplementary Fig. S3 and Supplementary Table S3). As expected, the protocols showed significant variation in genome coverage, context preference, and DNA M-bias. WGBS exhibited the highest depth of CpG coverage with a median of >43 reads for all samples, whereas two low-input protocols, T-WGBS and PBAT, reached a median coverage of <13 reads (Fig. 2A, Supplementary Fig. S2A, and Supplementary Table S3). The sequencing depth was relatively uniform across the genome for most protocols and showed a characteristic bell-shaped distribution. In contrast, PBAT showed a positively

A

	BAT	Biscuit	Bismark	BSBolt	bwa-meth	FAME	gemBS	GSNAP	methylCtools	methylpy
Programming Language	Perl	C	Perl	Python	Python	C++	R, Python	C, Java, Perl	Python	Python
Repository / Deployment	GitHub, DC	GitHub, DC, bioconda	GitHub, bioconda	GitHub, bioconda	GitHub	GitHub	Github, SC, bioconda	Author website	GitHub	GitHub, bioconda
Functionality										
Non-directional	-	+	+	+	+**	+	+	+	+**	+
Adapter Trimming	TrimGalore	TrimGalore	TrimGalore	TrimGalore	Trimmomatic	TrimGalore	gemBS	TrimGalore	Trimmomatic	TrimGalore
BS-aligner	segemehl	Biscuit *	Bowtie2	BWA	BWA	FAME *	GEM3	GSNAP	BWA	Bowtie2
Alignment strategy	wildcard	3-letter	3-letter	3-letter	3-letter	wildcard	3-letter	wildcard	3-letter	3-letter
Duplicates remover	-	samblaster	Bismark	samtools	Picard	FAME	BScall	Picard	Picard	Picard
Methylation caller	segemehl	Biscuit	Bismark	BSBolt	MethylDackel	FAME	BScall	BisSNP	methylCtools	methylpy
Methylation calling strategy	count ratio	count ratio	count ratio	count ratio	count ratio	count ratio	Bayesian	Bayesian	count ratio	count ratio

* asymmetric mapping problem
** with a custom read-sorting heuristic

B

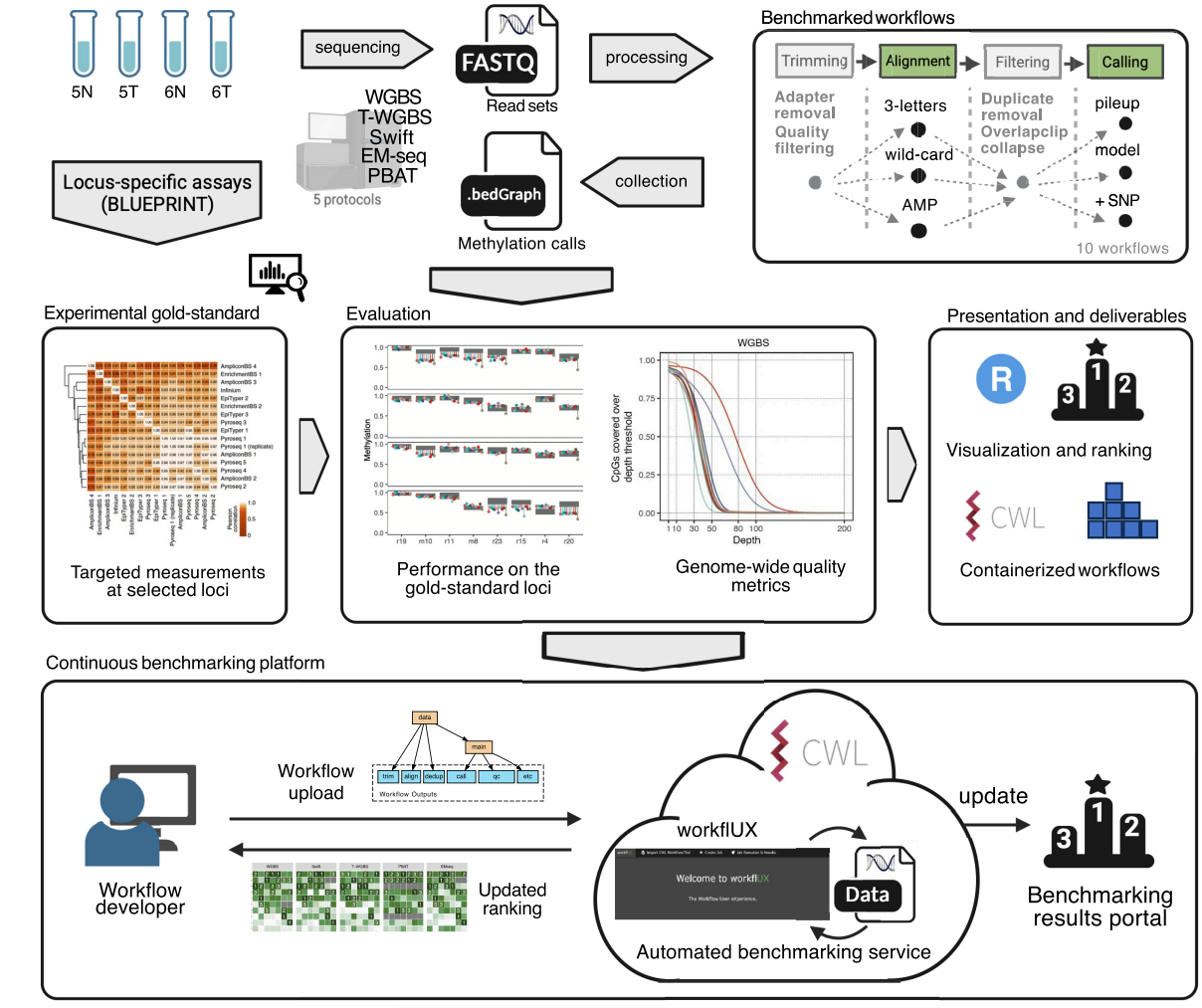


Figure 1. Workflow selection and study design. (**A**) Composition and main characteristics of the evaluated workflows. The criteria for choosing the workflows are explained in the "Workflow selection and deployment" section. "DC"/"SC" – Docker or Singularity container. (**B**) Schematic overview of the study design. Two pairs of colon cancer and adjacent normal samples were selected from [23] featuring consensus methylation corridors from multiple high-resolution methods at selected loci that served as the "gold-standard" measurements for our project. All samples were sequenced using five methyl-sequencing protocols, including one standard (WGBS), three low input (Swift, T-WGBS, PBAT) and one bisulfite free (EM-seq), and the data were processed using 10 selected workflows. CWL workflows and a Shiny-based portal for visualization are provided. To support the development of future workflows, we introduce a dedicated workflUX server to allow developers to execute their workflow on the selected datasets and compare the result with those presented in this study. The rankings will be updated on the open benchmarking service.

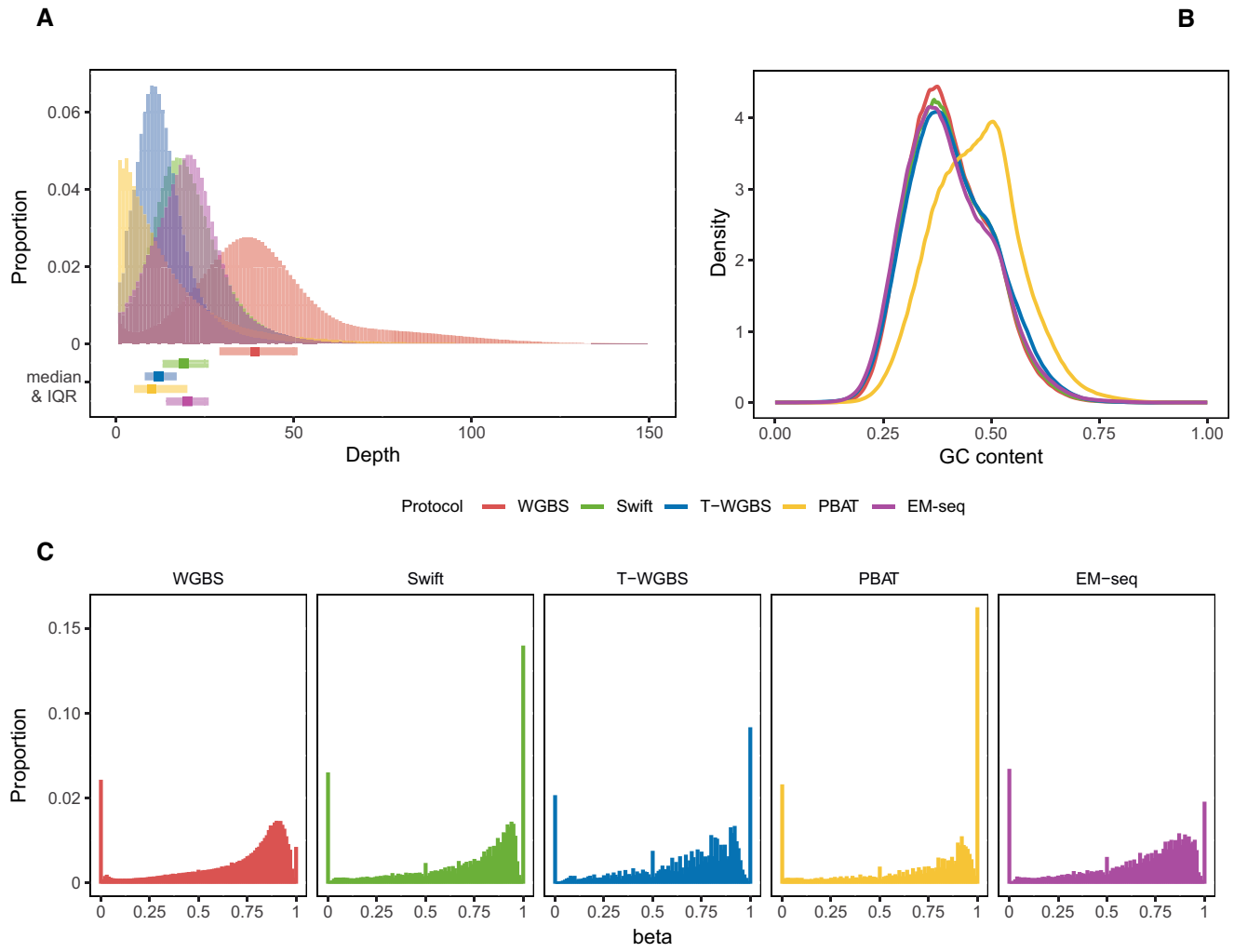


Figure 2. Overview of sequencing protocol differences. (**A**) Distribution of coverage depth for different protocols. The boxplots at the bottom of the histogram indicate the median and interquartile range values, offering insight into the central tendency and spread of the coverage distribution. (**B**) Distribution density of GC content for aligned reads. Note that not the read sequence itself but the reference genome sequence at the corresponding position was used for calculation. (**C**) Distribution of beta values resulting from the different protocols. CpGs covered by <10 reads were excluded. For panels (A) and (C), we filtered 1% of all CpG sites and used the results of all workflows for this same set of selected CpG sites to create the distribution histograms.

skewed distribution (Fig. 2A and Supplementary Fig. S2A) and a preference for GC-rich regions (Fig. 2B and Supplementary Fig. S2B). We excluded loci with a sequencing depth below 10 reads and explored the distribution of beta values, revealing strong variation among different protocols. Each protocol exhibited a broad mode at a beta value of ∼0.8 (Fig. 2C). Additionally, we found that, unlike other protocols, WGBS did not exhibit another sharp mode at beta-value = 1 corresponding to fully methylated sites. Absence of this peak was due to a systematically higher depth of coverage, as it accentuated when downsampling WGBS data to the genome-wide coverage level of PBAT (Supplementary Fig. S4). All protocols showed a protocol-specific methylation ratio shift at the end of the reads (M-bias), which required additional trimming or clipping (Supplementary Fig. S2C).

PBAT poses special challenges for data processing. For instance, it was shown that PBAT library preparation produces so-called chimeric reads containing sequences from two or more distinct genomic loci. [53]. The presence of such chimeric reads complicates proper alignment. This is primarily attributed to the more frequent occurrence of multiple alignment hits, resulting in higher ambiguity and reduced alignment rates. Furthermore, the fragments that were more likely to generate chimeric read pairs were not a random subset of all reads, possibly biasing methylation calling. Therefore, proper handling of chimeric reads not only improves coverage but also increases accuracy. Thus, we calculated the proportion of chimeric reads relative to the total mapped reads, considering reads that mapped to different chromosomes as chimeric (Supplementary Fig. S5). This allowed us to estimate the extent to which they occurred in different protocols. The results showed that the chimeric read proportion generated in PBAT was 6.64 times higher than in Swift and 10.33 times than in WGBS.

Genome-wide analysis of read coverage patterns identifies outlier workflows

We applied the 10 evaluated workflows to the data from five protocols, resulting in a total of 192 processing runs (*BAT*

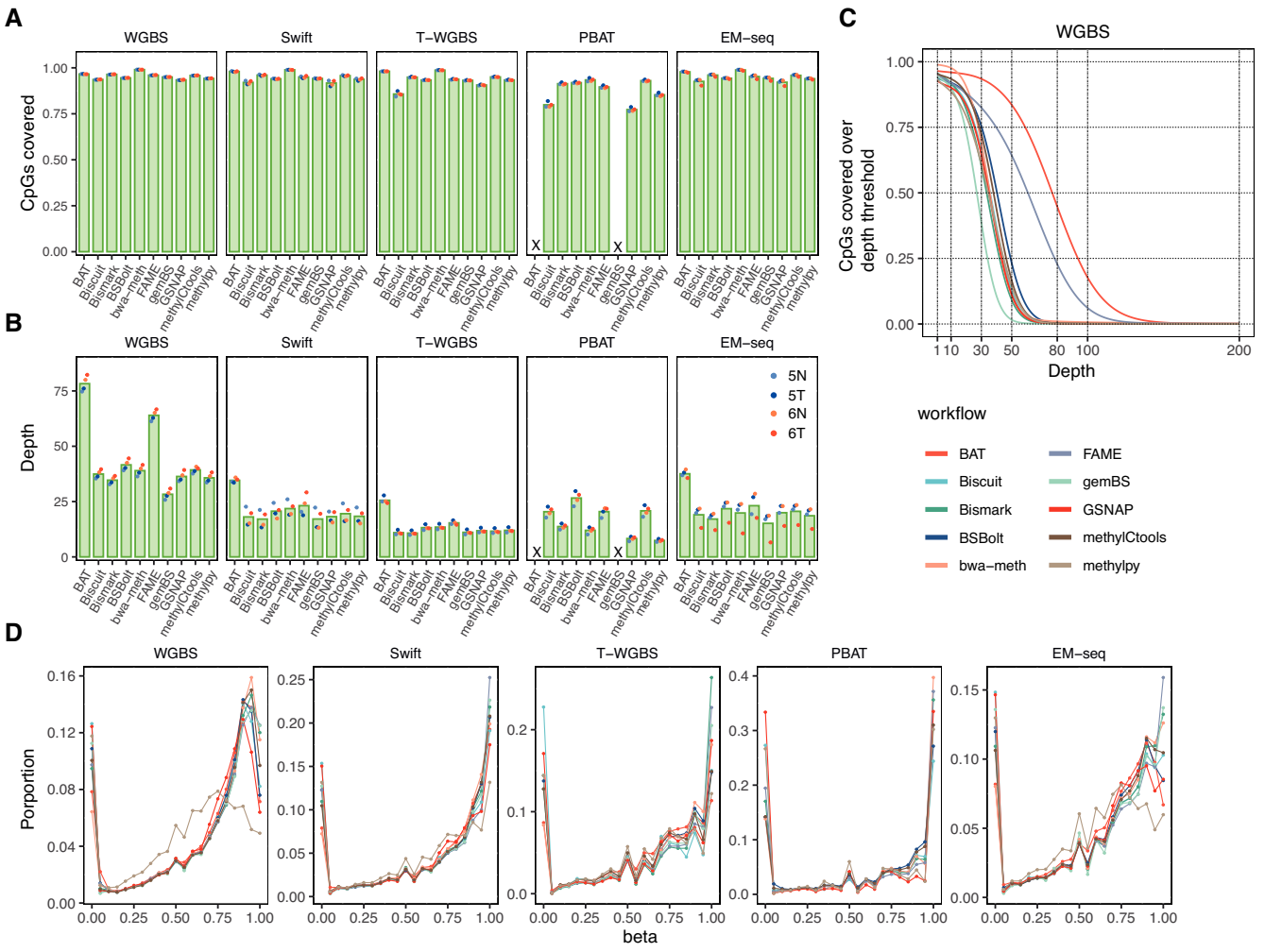


Figure 3. Genome-wide performance and data quality. (**A**) Median fraction of total genomic CpGs covered by at least one read in the output of each workflow. Individual samples represented by dots. A cross mark is used to indicate the absence of measurements for the corresponding workflow. (**B**) Average number of reads covering the CpG sites, with dots representing individual samples. A cross mark denotes the absence of measurements in the respective workflow. (**C**) Percentage of covered CpGs (*y* axis) with read depth below a threshold (*x* axis). It provides a practical method for determining the proper depth cutoff threshold. (**D**) Distribution of beta values returned by each evaluated workflow across all experimental protocols. The data points were sampled from all CpGs using a selection rate of 1/100. Curves show averaged distributions over all four samples, see Supplementary Fig. S7 for sample-wise plots.

did not support the nondirectional PBAT protocol and *gemBS* failed on PBAT data with an unresolvable error). A detailed summary of the data processing steps, with read counts after each step, is given in Supplementary Table S3. The workflows showed similar patterns of alignment rates for all input protocols, except for PBAT. Here, workflows lost 20%–77% of the reads during alignment, and even the better-performing ones, such as *BSBolt*, had <60% of the total reads remaining (Supplementary Fig. S3). The fraction of identified PCR duplicates was consistent across the workflows that included this step (Supplementary Table S3).

Although there were significant differences in the protocols and amount of input material used, the majority of the genomic CpG sites were covered with at least one read by most protocol–workflow combinations (Fig. 3A). We observed greater differences in low-input protocols, such as a lower percentage of CpGs covered in the T-WGBS using the *Biscuit* workflow. PBAT showed the largest discrepancies in the number of covered CpGs between workflows; however, the better performing ones, *BSBolt*, *Biscuit*, *FAME*, and *methylCtools*, reached genome-wide coverage other protocols.

We next assessed read coverage depth at individual CpGs, which is key for accurate estimation of methylation levels (Fig. 3B). The data revealed that processing with *BAT* and *FAME* generated the highest sequencing depth compared to all other protocols. Substantial variation was detected in PBAT, where the observed variation strongly correlated with the alignment rate (Supplementary Table S3). To ensure the statistical reliability of CpG-wise methylation levels and adequate support for methylation calls for downstream analyses, each CpG site must be covered by multiple reads. In practice, the sequencing depth threshold, is usually traded off against the genome-wide coverage to ensure a large proportion of CpG sites is retained. The depth-versus-coverage dependency curves for each workflow and sequencing protocol (Fig. 3C and Supplementary Fig. S6) revealed that most workflows showed comparable trends, with 75% of CpGs covered with at least 10 reads in WGBS, Swift, and EM-seq data (except for *gemBS* in EM-seq). In WGBS, two workflows (*BAT*

and *FAME*) showed increased read retention compared to all other workflows (Fig. 3C). In the case of *BAT*, this can be explained by the lack of duplicate removal. The high coverage in *FAME* data could be attributed to the double counting of CpG calls from the overlapping read mates. It was amplified in WGBS due to the shorter fragment sizes (Fig. 3C, Supplementary Fig. S4, and Supplementary Table S3). In T-WGBS, the samples had a lower depth, and all workflows, except for *Biscuit*, retained >75% of CpGs at the read coverage cutoff of 5. The largest differences between workflows were observed in PBAT, whereby only two workflows, *methylCtools* and *BSbolt*, retained at least 75% of CpGs with a coverage cutoff of 5, whereas two workflows, *methylpy* and *GSNAP*, lost half of the methylation sites with the same cutoff (Supplementary Fig. S6). To integrate coverage performance in the final evaluation, we introduced an area-under-the-curve metric integrating the depth and genome-wide breadth of coverage (see the "Materials and methods" section) as a quantitative measure of coverage retention.

Taken together, our evaluation of the global coverage metrics revealed surprisingly high variation in coverage depth across the workflows, even in the most deeply sequenced WGBS protocol, with several workflows being stark outliers. This highlighted substantial differences in read processing, especially related to alignment and PCR duplicate removal, potentially affecting downstream analyses.

Data-driven genome-wide methylation call consensus corridors elucidate workflow consistency

After evaluating genome-wide patterns of coverage and CpG retention, we asked how consistent the resulting DNA methylation calls and methylation level estimates were across workflows. We compared genome-wide methylation levels and observed lower (or in one case equal) methylation levels in tumor samples across all workflows and protocols (Supplementary Table S3), implying that, despite significant differences in effective coverage, all workflows could capture global methylation differences. Generally, the beta-value distributions were similar for all workflows except *methylpy*, with larger variations between protocols (Fig. 3D and Supplementary Fig. S7).

To quantitatively assess the similarity of the methylation calls between workflows, we defined a discrepancy score as the mean of absolute pairwise methylation difference between two methylation call vectors (Fig. 4A). As expected, the workflows showed the highest similarity in high-coverage WGBS protocol, with *methylpy* being a single outlier, and an increase in discrepancy on data from low-input protocols and more shallow sequencing depth, with PBAT showing the largest differences (Fig. 4A). To compare workflows at the genome-wide level, we introduced a data-driven consensus corridor for each CpG site by taking the smallest range covered by at least five workflows per protocol. We included WGBS, Swift, and EM-seq results in the calculations and excluded T-WGBS and PBAT, due to overall significantly lower coverage (Fig. 4B). We ranked the workflows based on the proportion of measurements that fell inside the consensus corridor (Fig. 4C). We used the sum of genome-wide deviations from the consensus corridor as the primary metric to evaluate the efficacy of genome-wide methylome profiling (Fig. 4D). Protocols with higher coverage exhibited superior performance in terms of high accuracy (resulting in measurements closer to zero) and low variability (with minor variations observed within the protocol). We also examined the deviation with respect to genomic annotations and concluded that for T-WGBS and PBAT, genic regions exhibited lower deviations than intergenic regions (Supplementary Fig. S8). However, since this effect was not observed in PCR-free WGBS, it might also be an artefact of extensive PCR amplification cycles in these protocols.

In summary, for WGBS, Swift, and EM-seq, most of evaluated workflows demonstrated high accuracy, with only *methylpy* exhibiting a slight outlier trend. Workflow performance on PBAT data showed the highest variability. In this scenario, *Biscuit*, *BSBolt*, and *methylCtools* achieved the most favorable results. Notably, all three workflows showed high alignment rates for PBAT data (Supplementary Fig. S3), while the ones with higher deviation scores, *bwa-meth* and *methylpy*, were characterized by a much lower alignment rate. This might be explained by workflow differences in handling PBAT chimeric reads. The alignment rate showed low and statistically non-significant correlation with the deviation score in PBAT (average of sample-wise Spearman = −0.339; Supplementary Fig. S9).

Collectively, the performance of methylation calls between workflows strongly depends on the characteristics of the data and on the specific protocol used to generate it. Technical differences between the workflows did not have major impact upon datasets with higher sequencing depths (WGBS, Swift, EM-seq). In contrast, on PBAT data the methylation calls were less consistent, implying that low coverage and technical challenges amplify even minor differences between workflows. Therefore, proper selection of data processing workflows is particularly important for low-input protocols.

Workflow accuracy evaluation against the experimental reference pinpoints workflow-specific pitfalls

We sought to obtain objective estimates of methylation call accuracy by utilizing the 46 preselected loci from the multimethod, multicenter BLUEPRINT study from which a consensus corridor containing the most likely true methylation values was derived (Supplementary Table S4; see the "Materials and methods" section for details) [23]. We used these consensus corridors as ground-truth methylation measurements. Upon initial inspection, we observed that for 45%–52% of the loci, methylation values returned by workflows lay within the consensus corridors across all protocols. Certain loci, such as m3 and m4, displayed stronger deviations that were sample-specific and affected all workflows (Fig. 5A and Supplementary Figs S10–S15).

To summarize the estimates, we calculated the deviation from the consensus corridor at each locus (Fig. 5B and Supplementary Fig. S16). We did not observe strong systematic differences between workflows. Most protocols and workflows, especially WGBS, tended to report lower methylation levels compared to the consensus measurement across several targeted assays with complementary strengths (Fig. 5B). On the other hand, the extent of the deviation was highly sample specific. We observed a tendency that the lowly methylated regions ($ß < 0.2$) were more accurately measured, whereas the accuracy decreased for the highly methylated ($ß \geq 0.8$) and especially the intermediately methylated ($0.2 \leq ß < 0.8$) regions (Fig. 5B and Supplementary Fig. S16). We calculated the accuracy of the workflow as the mean absolute de-

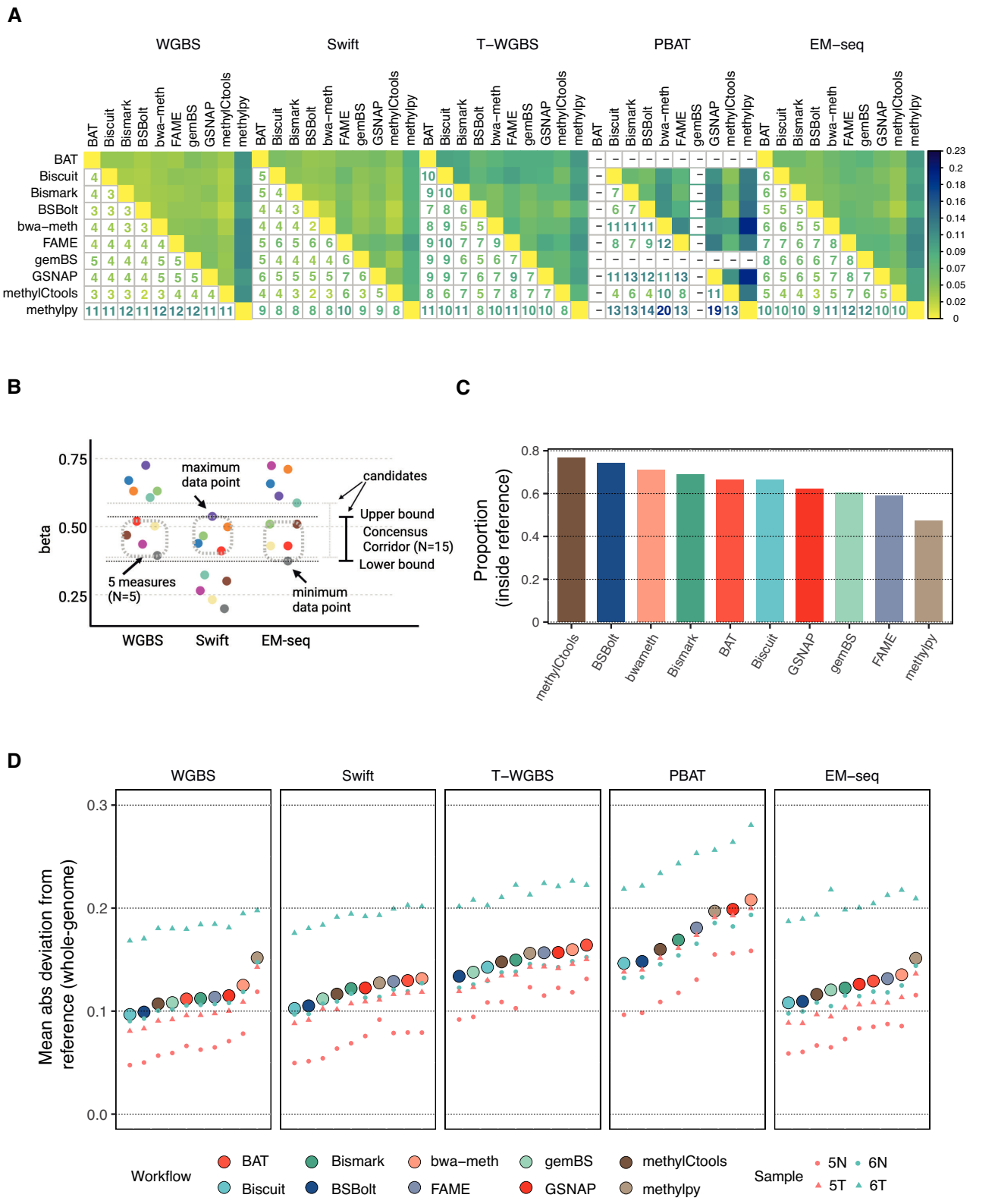


Figure 4. Genome-wide deviation and methylome similarity among workflows. (**A**) Methylome dissimilarity between workflows in each protocol. The numbers represent the discrepancy score, defined as mean pairwise difference of beta values across all CpGs multiplied by 100. Smaller values indicate higher levels of similarity between methylomes from different workflows. (**B**) Definition of genome-wide consensus corridors for all CpGs. The consensus corridor is defined as the smallest region encompassing at least five measurements from each of the three high-coverage protocols, WGBS, Swift, and EM-seq. (**C**) Fraction of measurements that falls within the all-protocol consensus corridor for each workflow. (**D**) Genome-wide mean absolute deviation from the border of the consensus corridors. Within a protocol, the workflows are sorted in ascending order.

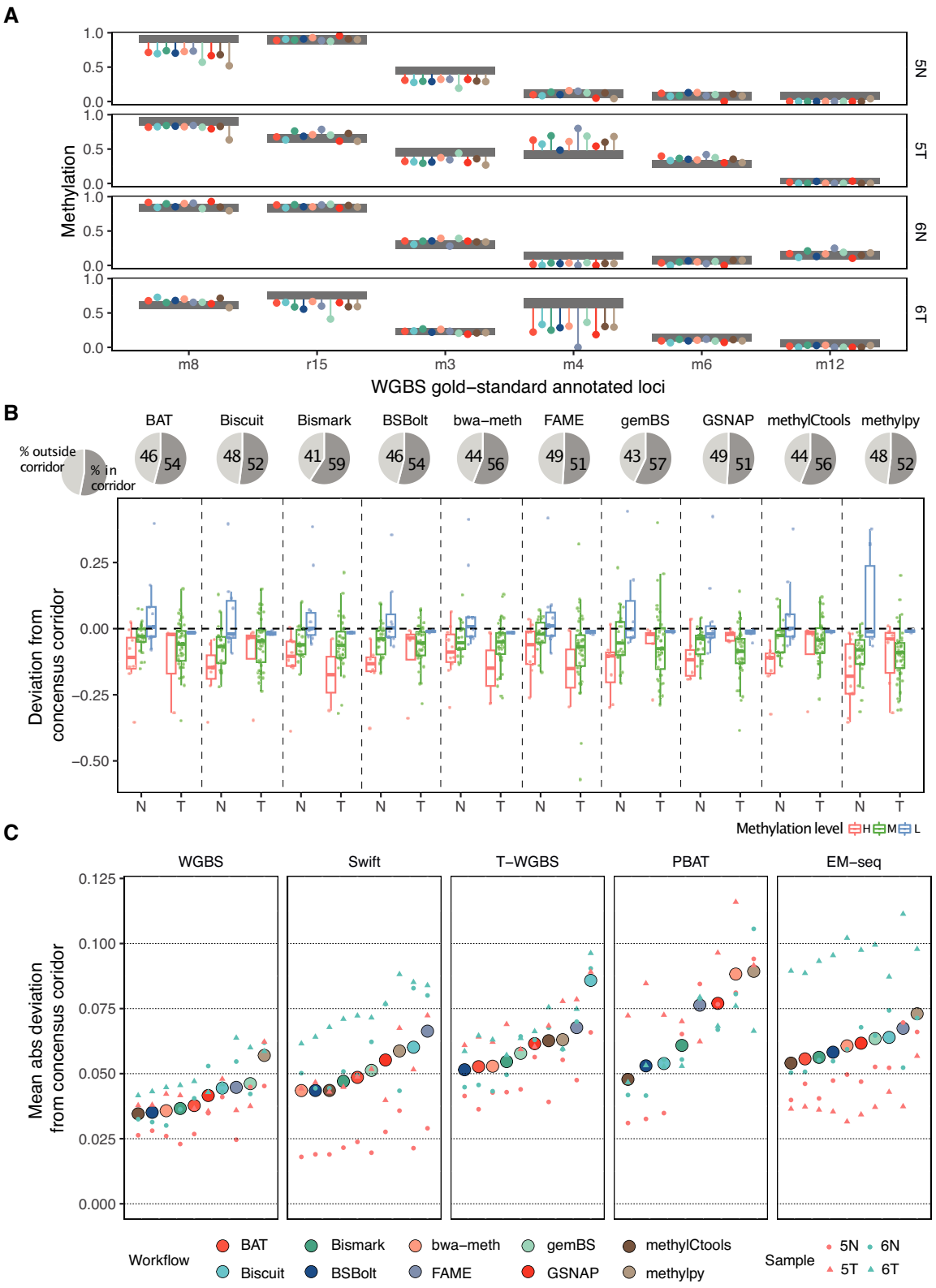


Figure 5. Assessment of methylation call accuracy based on the experimental gold standard. (**A**) Deviation from the gold-standard consensus corridors for six selected loci. Several example loci were selected here to represent high (left), intermediate, and low (right) methylation levels. Gray boxes represent the consensus corridors (see the "Materials and methods" section). The dots show the measured beta values for each workflow, while the lines depict their deviation from the consensus corridors. (**B**) Deviation from the consensus corridors of WGBS for all samples combined. The pie charts at the top show the proportion of sites outside/within the consensus corridor. The box plot below shows the distribution of deviation excluding the data inside of the consensus corridor, and N/T indicate for normal/tumor, respectively. (**C**) Mean absolute deviation by protocols and workflows. Within a protocol, the workflows were sorted ascendingly. The deviation is the average of four samples and the deviation of four samples was labeled on the same vertical line.

viation across the 46 loci (Fig. 5C). The effect of the workflow was statistically significant only for T-WGBS and PBAT (two-way ANOVA analysis, the *P*-value for T-WGBS and PBAT are 6.45×10^{-4} and 1.2×10^{-3}, respectively, WGBS 0.845, Swift 0.141, EM-seq 0.839) when adjusted for sample-wise differences.

Similar to the genome-wide consistency analysis, we observed that *methylpy* tended to underestimate methylation at specific loci (Fig. 5B and Supplementary Fig. S16), a pattern inconsistent with other established workflows. We investigated this divergence by using selected gold-standard loci. To illustrate this, aligned reads for *BSBolt* and *methylpy* at locus r23 in the normal sample from Patient 6 were examined using the Integrative Genomics Viewer browser. Despite sharing the same alignment, these two tools produced different methylation results (Supplementary Fig. S17). Using the Multiple Sequence Alignment tool Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/), we observed that the sequences from the *methylpy* intermediate alignment and the final alignment were different from the raw reads and alignment of *BSBolt* (Supplementary Fig. S17). This inconsistency appears to be the result of an error in the simulation of the bisulfite conversion. *Methylpy* employs a three-letter bisulfite alignment strategy that conducts *in silico* conversion of reads prior to alignment with the three-letter genomic reference. In the methylation calling step, the converted reads must be restored to their original sequences. Our observations indicated a systematic error or bug at this specific step, resulting in unrestored bases. This provided one explanation for the outlier calls produced by *methylpy*.

Taken together, the analysis of methylation calls at gold-standard loci allowed us to obtain realistic estimates of workflow accuracy and helped us objectively rank the workflows. Furthermore, exemplary deeper investigation of outliers allowed to find specific pitfalls of individual workflows, demonstrating the potential usefulness of our results for workflow debugging.

The choice of processing workflow affects downstream differential methylation analysis

Depending on the analytical objective, the absolute level of methylation might not be as important as the identification of differentially methylated loci (DMLs) or regions (DMRs) between experimental conditions and subgroups. To address this in our study, we devised a combined strategy to assess the workflow differential methylation impact that combines two differential methylation-centered metrics (Fig. 6A). Based on the tumor–normal pairs included in our study, we investigated whether various workflows affected the accuracy of DMR identification with a standard DMR caller. As an external reference, we used Illumina HumanMethylation450 array data for all six tumor–normal pairs of the original study [23] to increase statistical power compared to our two sample pairs. Because the array covers fewer CpG sites than genome-wide sequencing, we restricted our comparison to CpG sites covered by the array. We applied the same DML identification procedure (see the "Materials and methods" section) to all workflows and calculated AUC scores to evaluate the performance of the workflows (Fig. 6B). As a second metric for differential analysis, we used the correlation between tumor–normal beta-value differences obtained from sequencing and microarray data (Fig. 6C).

All workflows demonstrated the capability of achieving an AUC value of at least 0.68. However, when considering correlation to Infinium array data, only in the case of WGBS the correlation coefficient surpassed 0.5, possibly indicating the increase of consistency with higher sequencing depth. Nonetheless, these two metrics showed a high degree of consistency across various protocols, both in terms of ranking and differences between workflows (Fig. 6D). To integrate both assessments, we established a composite metric by combining the weighted AUC and the correlation to evaluate the effectiveness of DMR identification. The final DMR performance metric was determined by averaging the rankings of the workflow based on the two metrics.

Overall, we conclude that, although inferior to the effect of the sequencing depth, the workflow choice did have a measurable impact upon the differential methylation analysis. Given that the experimental protocol and depth of sequencing are often predetermined by sample availability and sequencing resources available, workflow users should carefully consider their data processing strategy.

Workflows show drastic differences in computational performance

Computational efficiency is one of the major software selection criteria in practice. To help users identify workflows that fit their available computational resources, we measured the run time and maximal memory usage of each workflow (Fig. 7A and B). All workflows implement support for parallel computations. Therefore, we allocated all the resources of the computing node when measuring the resource usage. We observed substantial variation between workflows in terms of both runtime and memory requirements, regardless of the protocols. The running time varied between two extremes: 4 h for *gemBS* and 14 days for *GSNAP*. As expected, the extremely deeply sequenced WGBS protocol had the longest running time in most workflows. Excluding WGBS, PBAT required a slightly longer run time than other protocols. We suspect that handling four different strands in PBAT leads to higher computational load. Overall, *gemBS*, *FAME*, and *Biscuit* were the fastest workflows, ranking among the top three across all tested scenarios. Operating memory usage ranged between 16 GB for *Bismark* and 319 GB for *GSNAP*. *Bismark* had the consistently best memory footprint in all tests, followed by *FAME* and *methylpy*. Notably, the tools with perfect memory footprint oftentimes showed very long runtime and vice versa, indicating that the availability of respective resources can be decisive for the choice of workflow.

Final performance ranking and recommendations

To offer guidance in selecting a bisulfite sequencing workflow, we combined all evaluation results to create an overall performance ranking. We first ranked the workflows based on each evaluation criterion. This ranking reflects the performance of each workflow in this specific category (see the "Materials and methods" section for details). The final ranks were then calculated by averaging the individual ranks (Supplementary Table S5). Ranking was derived across all protocols considering six metrics (Fig. 8A) as well as individually for each protocol (Fig. 8B).

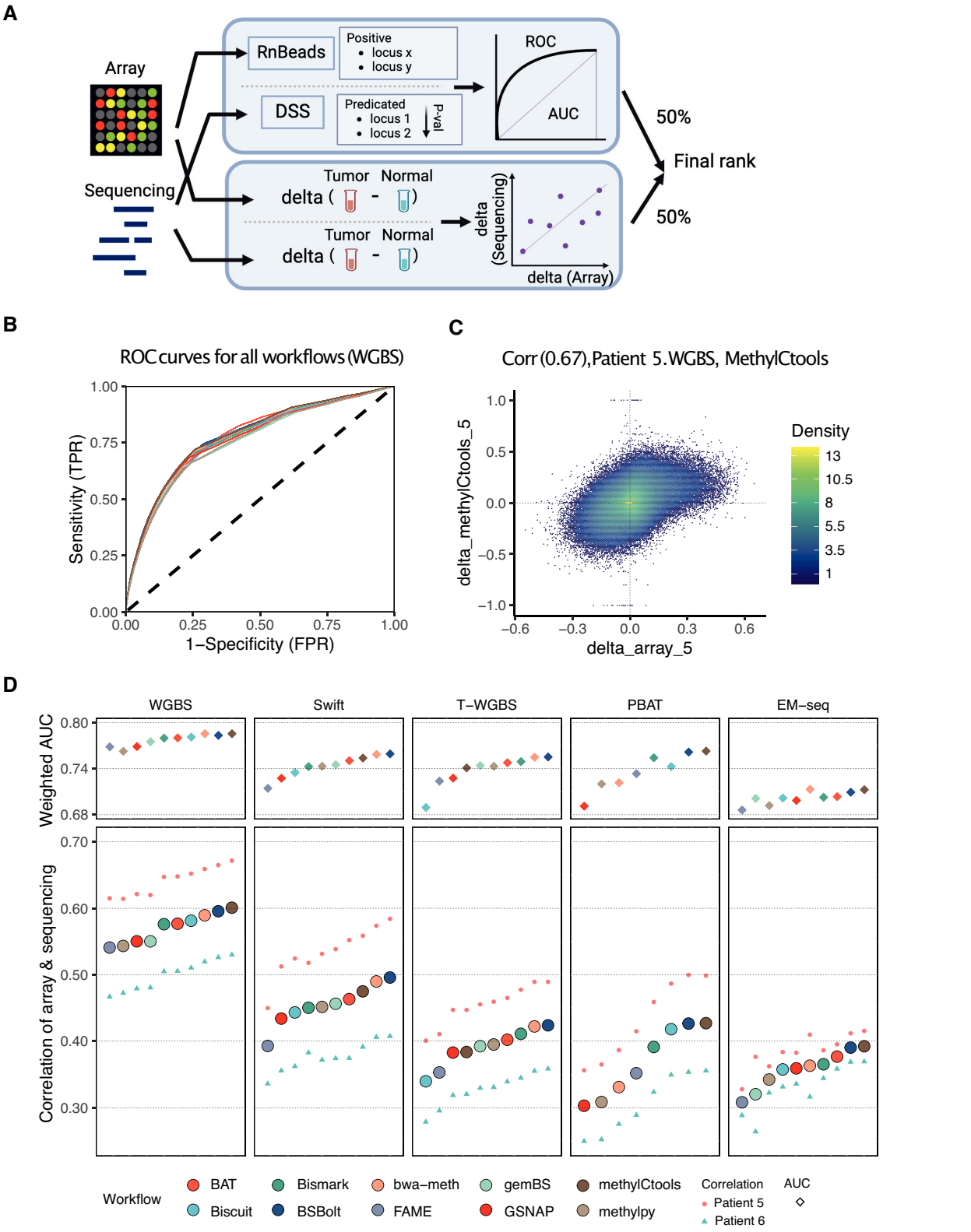


Figure 6. Effects of data processing on differential methylation analysis. (**A**) Flowchart of the differential methylation analysis. Two types of performance measurements were employed. First, we took the DMLs as based on Illumina HumanMethylation450 arrays for six colon cancer tumor–normal pairs as ground truth and compared them to the DMLs acquired on the sequencing data by each workflow. After applying the same filtering rules, the predicting power was estimated by ROC curves. Second, pairwise differences between tumor and normal samples were calculated both for array data as well as for sequencing. The correlation between these differences served as the second metric. The final result was determined by averaging the ranks from two metrics. (**B**) ROC curves for all workflows, based on WGBS data. Color code as in panel (D). (**C**) An example of 2D density correlation plot of methylation delta beta (tumor–normal) in microarray and sequencing data for Patient 5, WGBS protocol, and *methylCtools* workflow. (**D**) Weighted AUC and correlation of delta beta. The upper plot shows the AUC score, and the lower plot shows the correlation of delta. Within each protocol, the workflows were sorted in ascending order based on their correlation values.

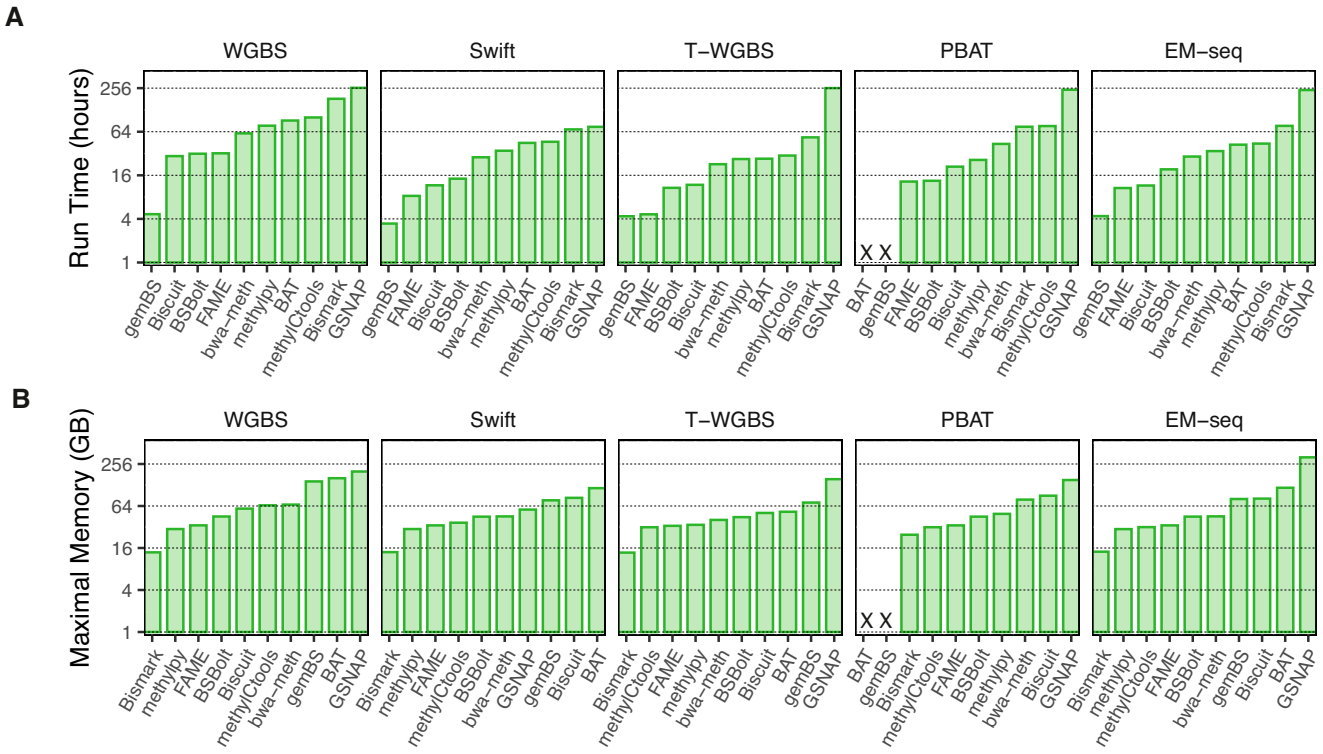


Figure 7. Computing resource requirements. (**A**) Runtime in hours for processing sample 5N. (**B**) Peak operating memory usage in gigabytes (GB) for processing sample 5N.

According to our results, *BSBolt* ranked first among all five sequencing protocols. Following this, *methylCtools* achieved second place in three protocols and third place in one protocol. The subsequent rankings were held by *bwa-meth*, *BAT*, *Bismark*, and *Biscuit*, each of which secured the top three positions in at least one of the five protocols. The final summarized ranking considers all metrics with equal weights. This is usually not the case in real-world studies; for example, small differences in coverage depth are not important if the workflow is accurate. Therefore, we created an interactive interface that presents the results of our study and helps customize the workflow ranking. This tool is available at https://compepigen.github.io/PipelineOlympics/shiny.

In addition to the workflow performance, we systematically investigated the simplicity and diversity of installation and the documentation quality according to a number of criteria (Supplementary Table S6). We also derived an integrated installation simplicity and documentation quality score and ranked the workflows accordingly. *Biscuit*, *methylpy*, and *Bismark* scored highest in this comparison (Supplementary Table S6).

Continuous benchmarking platform for bisulfite sequencing workflows

As our final step, we aimed to turn our evaluation of 10 workflows into a continuous ("living") benchmarking study. For this, we chose to provide a web service for workflow developers allowing them quickly assess quality and overcome usual challenges, such as the availability of performance evaluation datasets, reliable metrics, and convenient execution environments. We implemented our continuous benchmarking platform using German public de.NBI IAAS cloud. As portal to our platform, we employed workflUX (https://github.com/workflux/workflUX), a user-friendly web service we earlier developed for running workflows implemented CWL. In brief, future developers can implement their workflows, or their simple single-command wrappers, in CWL and upload them into worklfUX for evaluation. Continuous benchmarking processes these uploads using downscaled versions of our benchmarking datasets and compares the results to 10 existing workflows in our study, offering valuable feedback for future improvements. Usage instructions can be found on the website for reference This service is deployed on the deNBI.cloud and will be made available to developers at no cost at https://compepigen.github.io/PipelineOlympics/workflux.

Discussion

Here, we present the results of our systematic benchmarking to provide an overview of the available complete workflows for bisulfite sequencing data and thoroughly evaluate their performance. We included regularly maintained, open-source workflows that cover different existing approaches and chose 10 to be included in our study (Fig. 1A and Supplementary Table S1). Since different library preparation protocols have different technical and methodological aspects, we included five approaches: the originally developed WGBS, bisulfite-free (EM-seq) and low-input protocols (T-WGBS, Swift, PBAT). In addition, we present a dynamically extendable framework to include any number of additional tools, aiming to help both developers and users to evaluate their workflows of choice.

The main challenge of benchmarking studies is the lack of established ground truth measurements. Simulated data often do not capture all known and latent sources of bias and noise,

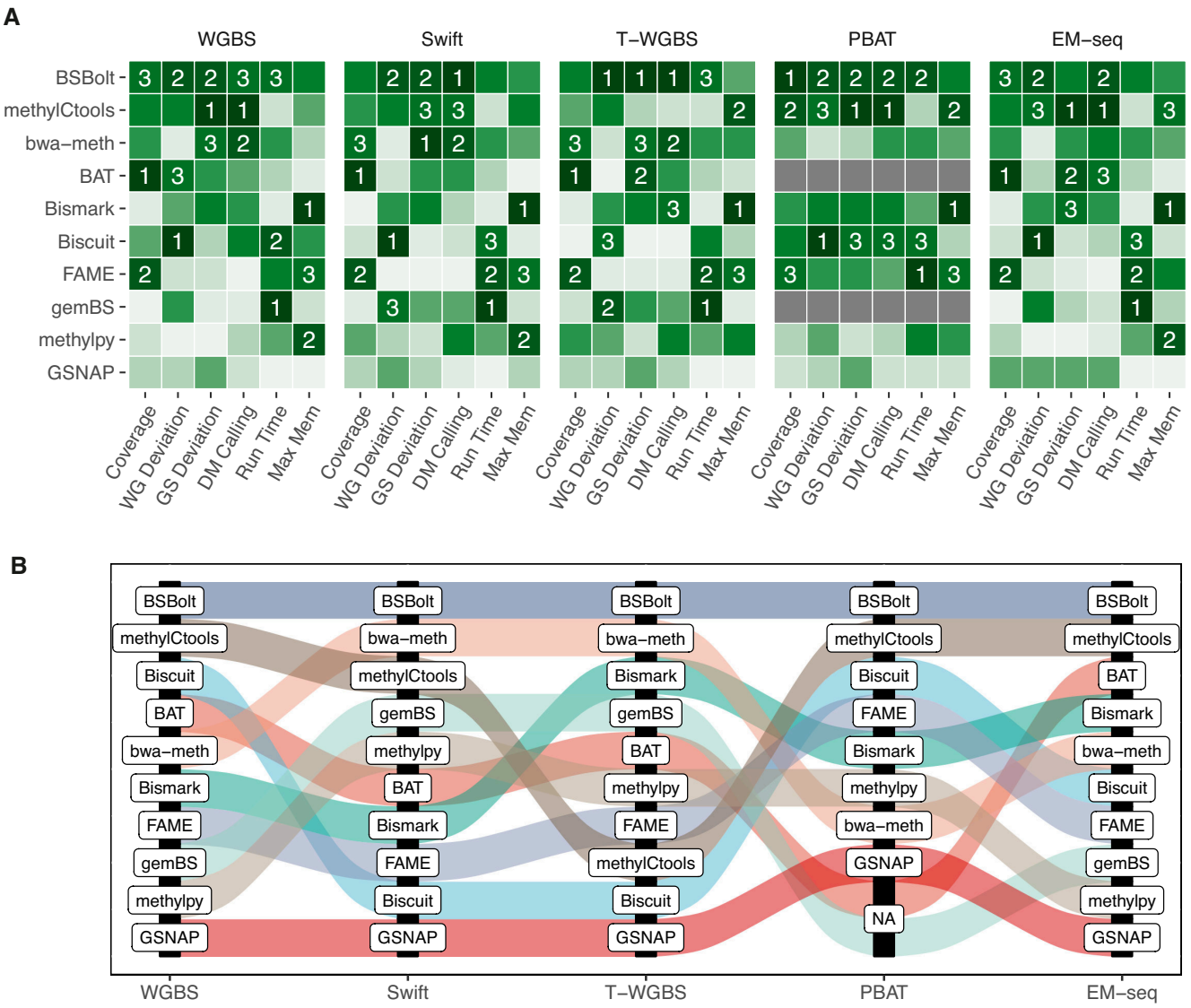


Figure 8. Final performance ranking. (**A**) Summarized results of the benchmarking study. Color scale reflects the rank of the workflow for each metric (1 to 10, 1 is the best; dark green to light). Six metrics were incorporated as follows. (i) "Coverage": area under the curve of the "%CpGs covered over coverage threshold" plot (Fig. 3C). (ii) "WG deviation" represents the deviation across the whole genome (Fig. 4C), while (iii) "GS Deviation": the deviation observed at the 46 gold-standard loci (Fig. 5C). (iv) "DM calling": the average of weighted AUC and the correlation of array and sequencing (Fig. 6C). (v) "Run Time": execution time, (vi) "Max Mem": the peak memory usage (Fig. 7A and B). The workflows were ordered by the rank average of all measurements and numbers mark the first three workflows by comparison. (**B**) Alluvial plot shows the rankings for each protocol.

whereas, for real-world data, the true underlying methylation levels are unknown. To tackle this problem, we used DNA samples from an earlier technology benchmarking study, providing highly accurate consensus DNA methylation measurements for a selection of CpG sites in tumor and normal samples. Using several genome-wide and locus-specific metrics, we objectively ranked the performance of data processing workflows in a protocol-specific manner.

Our overview of DNA methylation data generated using different protocols revealed significant protocol-specific challenges, despite the overall consistency of the generated methylation data. In particular, we identified a nonuniform coverage with a bias toward GC-rich regions in the ultralow-input PBAT protocol, causing a noticeable shift in the methylation ratio. Although not in frequent use anymore, PBAT is related to modern low-input and single-cell bisulfite sequencing protocols. Our findings emphasize the need for careful workflow selection for each type of sequencing data, particularly for the low-input protocols.

The integration of all our benchmarks into the final decathlon-style ranking showed a largely consistent workflow performance pattern across the considered experimental protocols and allowed the identification of three major performance groups. *BSBolt*, *bwa-meth*, and *methylCtools* showed reproducibly high performance for almost all criteria and protocols. Although the size of our workflow set does not allow for a rigorous statistical evaluation, it is worth noting that all three best-performing tools use a reduced alphabet (three-letter) bisulfite alignment strategy with the Burrows-Wheeler Aligner (BWA) as their underlying alignment engine. In the mid-performing group, *BAT*, *Bismark*, *gemBS*, and *FAME* showed good performance, particularly for runtime and memory usage, whereas *methylpy* and *GSNAP* performed worse in many metrics. *Biscuit* showed the most variable perfor-

mance across experimental protocols, yet performed very well in PBAT, which is most similar to single-cell bisulfite sequencing protocols. Notably, we detected notable differences in CPU time and observed the trend that the workflows showing high levels of runtime and memory usage efficiency—*gemBS*, *FAME*, *Biscuit* and *Bismark*—tended to score lower in the accuracy benchmarks, revealing two diverging pathways in the evolution of workflows. Last but not least, the ranking of workflows with respect to deployment and documentation quality was not associated with the actual workflow performance, revealing that some workflow development paid more attention to these aspects.

Our results show that proper workflow selection is more important with low-depth, low-input protocols, since low sequencing depth amplifies the methodological and implementation-level pitfalls of each workflow. The final choice of workflow depends on individual priorities and can be affected by multiple factors. Therefore, we developed an interactive web application that visualizes the results and adapts ranking to individual importance weights.

Given the constantly accelerating influx of new bioinformatics software, comprehensive and objective benchmarking is becoming crucial. New software tools appear regularly and there are many possible combinations of tools throughout the workflow. Therefore, we introduced a dynamically expandable benchmarking platform allowing for easy addition of new tools and workflows, with the main goal of helping future developers optimize their tools. Taken together, our study helps a broad range of users to choose the proper workflows adjusted for their needs and allows expansion to upcoming data processing tools and workflows, establishing the concept of dynamic benchmarking. This can serve as an example for benchmarking software for other data types, helping to increase the quality and reproducibility of data analysis.

Acknowledgements

We are grateful to Ingrid Scholz for her help with data import and to Matthias Bieg and Charles Imbusch for sharing the implementation of the *methylCtools* workflow for PBAT data. Some of the figures are created using Biorender.com.

Author contributions: P.L.,R.T., Y.A., C.P. and C.B. initiated the study and acquired funding. Y.L. implemented of all workflows on local compute infrastructure, performed data processing, analysis, interpretation, created all figures and tables and wrote the manuscript with P.L and R.T. M.E. and C.B provided samples for the study. D.W., O.M., M.Schoe., M.H., F.P. performed sequencing experiments supervised by S.W., C.G., D.B.L and C.P. L.W., G.G., F.T. and I.B. managed sequencing data and computing infrastructure. K.B., A.W., K.N., S.K., P.K., J.F., E.A.H., H.K., S.H., A.M., M.H.S., A.S. supported workflow implementation and bioinformatic analyses. Y.L., P.Lafr., M.C. developed web-services supported by S.T. and C.L. C.P., C.B., Y.A., J.W., C.G., M.H., D.B.L, V.H., M.Z., M.I., E.A.G., S.H., M.Schl., M.H.S. supported data interpretation. P.L. and R.T. jointly supervised the study. All authors read, edited and approved the final text of the manuscript.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This study was supported by grants from the German Ministry of Education and Science (BMBF) for the consortium BSmadeEZ (031L0162B to P.L., R.T., and Y.A. and 031L0162A to C.L.) and from Horizon Europe project EOSC4Cancer (101058427). P.L. was supported by the BMBF-funded German Network for Bioinformatics Infrastructure (de.NBI) within its partner project de.NBI-epi/Heidelberg (031L0101A). de.NBI also provides computational resources and hosts web services. P.L. and C.P. received funding from the German Cancer Aid (DKH) for the project CO-CLL (70113869). P.L. was furthermore supported by a BOFZAP starting grant (STG/22/024) from KU Leuven. D.B.L. received funding from the Wilhelm Sander-Stiftung (2022.010.1) and from the BMBF (HEROES-AYA consortium, subproject 3, 01KD2207A). This research has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 824110—EASI-Genomics (to M.Schle. and S.K.). Additional support to M.H.S. by the Cardio-Pulmonary Institute (CPI) (EXC 2026, 390649896) and the German Center for Cardiovascular Research (D ZHK) (81Z0200101). M.Schoe. was supported by the Joachim Herz Foundation (Add-on Fellowships for Interdisciplinary Life Science). Funding to pay the Open Access publication charges for this article was provided by Luxembourg Institute of Health (to R.T.) and KU Leuven BOFZAP starting grant (to P.L.).

Data availability

We have uploaded the raw sequencing data to European Genome–Phenome Archive under the accession number EGAS50000000541. The CWL versions of all evaluated workflows with containerization details, as well as all analysis R scripts, are publicly available at https://github.com/CompEpigen/PipelineOlympics and deposited via Zenodo under http://doi.org/10.5281/zenodo.16573651.

References

1. Schubeler D. Function and information content of DNA methylation. *Nature* 2015;**517**:321–6. https://doi.org/10.1038/nature14192
2. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol* 2019;**20**:590–607. https://doi.org/10.1038/s41580-019-0159-6
3. Jones PA, Taylor SM. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* 1980;**20**:85–93. https://doi.org/10.1016/0092-8674(80)90237-8
4. Suelves M, Carrio E, Nunez-Alvarez Y *et al.* DNA methylation dynamics in cellular commitment and differentiation. *Brief Funct Genomics* 2016;**15**:443–53.
5. Yang Z, Wong A, Kuh D *et al.* Correlation of an epigenetic mitotic clock with cancer risk. *Genome Biol* 2016;**17**:205. https://doi.org/10.1186/s13059-016-1064-3
6. Cruickshanks HA, McBryan T, Nelson DM *et al.* Senescent cells harbour features of the cancer epigenome. *Nat Cell Biol* 2013;**15**:1495–506. https://doi.org/10.1038/ncb2879

7. Locke WJ, Guanzon D, Ma C *et al.* DNA methylation cancer biomarkers: translation to the Clinic. *Front Genet* 2019;**10**:1150. https://doi.org/10.3389/fgene.2019.01150
8. Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. *Trends Genet* 2021;**37**:1012–27. https://doi.org/10.1016/j.tig.2021.05.002
9. Cunningham JM, Christensen ER, Tester DJ *et al.* Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res* 1998;**58**:3455–60.
10. Capper D, Jones DTW, Sill M *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* 2018;**555**:469–74. https://doi.org/10.1038/nature26000
11. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* 2003;**349**:2042–54. https://doi.org/10.1056/NEJMra023075
12. Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene* 2002;**21**:5400–13. https://doi.org/10.1038/sj.onc.1205651
13. Oakes CC, Seifert M, Assenov Y *et al.* DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia. *Nat Genet* 2016;**48**:253–64. https://doi.org/10.1038/ng.3488
14. Villicana S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol* 2021;**22**:127. https://doi.org/10.1186/s13059-021-02347-6
15. Ligthart S, Marzi C, Aslibekyan S *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol* 2016;**17**:255. https://doi.org/10.1186/s13059-016-1119-5
16. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;**14**:R115. https://doi.org/10.1186/gb-2013-14-10-r115
17. Luo H, Wei W, Ye Z *et al.* Liquid biopsy of methylation biomarkers in cell-free DNA. *Trends Mol Med* 2021;**27**:482–500. https://doi.org/10.1016/j.molmed.2020.12.011
18. Kader F, Ghai M. DNA methylation and application in forensic sciences. *Forensic Sci Int* 2015;**249**:255–65. https://doi.org/10.1016/j.forsciint.2015.01.037
19. Bibikova M, Fan JB. Genome-wide DNA methylation profiling. *Wiley Interdiscip Rev Syst Biol Med* 2010;**2**:210–23. https://doi.org/10.1002/wsbm.35
20. Beck S, Rakyan VK. The methylome: approaches for global DNA methylation profiling. *Trends Genet* 2008;**24**:231–7. https://doi.org/10.1016/j.tig.2008.01.006
21. Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* 2014;**15**:647–61. https://doi.org/10.1038/nrg3772
22. Barros-Silva D, Marques CJ, Henrique R *et al.* Profiling DNA methylation based on next-generation sequencing approaches: new insights and clinical applications. *Genes* 2018;**9**:429. https://doi.org/10.3390/genes9090429
23. consortium BLUEPRINT. Quantitative comparison of DNA methylation assays for biomarker development and clinical applications. *Nat Biotechnol* 2016;**34**:726–37.
24. Foox J, Nordlund J, Lalancette C *et al.* The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biol* 2021;**22**:332. https://doi.org/10.1186/s13059-021-02529-2
25. Harris RA, Wang T, Coarfa C *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 2010;**28**:1097–105. https://doi.org/10.1038/nbt.1682
26. Olova N, Krueger F, Andrews S *et al.* Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. *Genome Biol* 2018;**19**:33. https://doi.org/10.1186/s13059-018-1408-2
27. 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
28. Bock C, Tomazou EM, Brinkman AB *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol* 2010;**28**:1106–14. https://doi.org/10.1038/nbt.1681
29. Morrison J, Koeman JM, Johnson BK *et al.* Evaluation of whole-genome DNA methylation sequencing library preparation protocols. *Epigenetics Chromatin* 2021;**14**:28. https://doi.org/10.1186/s13072-021-00401-y
30. Lister R, Pelizzola M, Dowen RH *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**:315–22. https://doi.org/10.1038/nature08514
31. Bibikova M, Lin ZW, Zhou LX *et al.* High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* 2006;**16**:383–93. https://doi.org/10.1101/gr.4410706
32. Pidsley R, Zotenko E, Peters TJ *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* 2016;**17**:208. https://doi.org/10.1186/s13059-016-1066-1
33. Meissner A, Mikkelsen TS, Gu HC *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;**454**:766–70. https://doi.org/10.1038/nature07107
34. Tanic M, Moghul I, Rodney S *et al.* Comparison and imputation-aided integration of five commercial platforms for targeted DNA methylome analysis. *Nat Biotechnol* 2022;**40**:1478–87. https://doi.org/10.1038/s41587-022-01336-9
35. Liu T, Conesa A. Profiling the epigenome using long-read sequencing. *Nat Genet* 2025;**57**:27–41. https://doi.org/10.1038/s41588-024-02038-5
36. Adey A, Shendure J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res* 2012;**22**:1139–43. https://doi.org/10.1101/gr.136242.111
37. Wang Q, Gu L, Adey A *et al.* Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc* 2013;**8**:2022–32. https://doi.org/10.1038/nprot.2013.118
38. Weichenhan D, Wang Q, Adey A *et al.* Tagmentation-based library preparation for low DNA input whole genome bisulfite sequencing. *Methods Mol Biol* 2018;**1708**:105–22.
39. Miura F, Enomoto Y, Dairiki R *et al.* Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res* 2012;**40**: e136. https://doi.org/10.1093/nar/gks454
40. Vaisvila R, Ponnaluri VKC, Sun Z *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res* 2021;**31**:1280–9. https://doi.org/10.1101/gr.266551.120
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20. https://doi.org/10.1093/bioinformatics/btu170
42. Chen SF, Zhou YQ, Chen YR *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90. https://doi.org/10.1093/bioinformatics/bty560
43. Chen PY, Cokus SJ, Pellegrini M. BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinf* 2010;**11**:203. https://doi.org/10.1186/1471-2105-11-203
44. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;**13**:R83. https://doi.org/10.1186/gb-2012-13-10-r83
45. 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
46. Lim JQ, Tennakoon C, Li GL *et al.* BatMeth: improved mapper for bisulfite sequencing reads on DNA methylation. *Genome Biol* 2012;**13**:R82. https://doi.org/10.1186/gb-2012-13-10-r82
47. Marco-Sola S, Sammeth M, Guigo R *et al.* The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* 2012;**9**:1185–8. https://doi.org/10.1038/nmeth.2221

48. Pedersen B, Hsieh TF, Ibarra C *et al.* MethylCoder: software pipeline for bisulfite-treated sequences. *Bioinformatics* 2011;**27**:2435–6. https://doi.org/10.1093/bioinformatics/btr394
49. Ryan DP, Ehninger D. Bison: bisulfite alignment on nodes of a cluster. *BMC Bioinf* 2014;**15**:337. https://doi.org/10.1186/1471-2105-15-337
50. Coarfa C, Yu FL, Miller CA *et al.* Pash 3.0: a versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. *BMC Bioinf* 2010;**11**:572. https://doi.org/10.1186/1471-2105-11-572
51. Li YX, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinf* 2009;**10**:232.
52. Smith AD, Chung WY, Hodges E *et al.* Updates to the RMAP short-read mapping software. *Bioinformatics* 2009;**25**:2841–2. https://doi.org/10.1093/bioinformatics/btp533
53. Wu P, Gao Y, Guo W *et al.* Using local alignment to enhance single-cell bisulfite sequencing data efficiency. *Bioinformatics* 2019;**35**:3273–8. https://doi.org/10.1093/bioinformatics/btz125
54. Fischer J, Schulz MH. Efficiently quantifying DNA methylation for bulk- and single-cell bisulfite data. *Bioinformatics* 2023;**39**:btad386. https://doi.org/10.1093/bioinformatics/btad386
55. Bock C. Analysing and interpreting DNA methylation data. *Nat Rev Genet* 2012;**13**:705–19. https://doi.org/10.1038/nrg3273
56. Pedersen BS, Eyring K, De S *et al.* Fast and accurate alignment of long bisulfite-seq reads. arXiv, https://doi.org/10.48550/arXiv.1401.1129, 6 January 2014, preprint: not peer reviewed.
57. Hovestadt V, Jones DT, Picelli S *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* 2014;**510**:537–41. https://doi.org/10.1038/nature13268
58. Merkel A, Fernandez-Callejo M, Casals E *et al.* gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. *Bioinformatics* 2019;**35**:737–42. https://doi.org/10.1093/bioinformatics/bty690
59. Liu YP, Siegmund KD, Laird PW *et al.* Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol* 2012;**13**:R61. https://doi.org/10.1186/gb-2012-13-7-r61
60. Barturen G, Rueda A, Oliver JL *et al.* MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Res* 2013;**2**:217. https://doi.org/10.12688/f1000research.2-217.v1
61. Gao S, Zou D, Mao L *et al.* BS-SNPer: SNP calling in bisulfite-seq data. *Bioinformatics* 2015;**31**:4006–8. https://doi.org/10.1093/bioinformatics/btv507
62. Kunde-Ramamoorthy G, Coarfa C, Laritsky E *et al.* Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Res* 2014;**42**:e43. https://doi.org/10.1093/nar/gkt1325
63. Sun XW, Han Y, Zhou LY *et al.* A comprehensive evaluation of alignment software for reduced representation bisulfite sequencing data. *Bioinformatics* 2018;**34**:2715–23. https://doi.org/10.1093/bioinformatics/bty174
64. Tran H, Porter J, Sun MA *et al.* Objective and comprehensive evaluation of bisulfite short read mapping tools. *Adv Bioinformatics* 2014;**2014**:472045.
65. Kretzmer H, Otto C, Hoffmann S. BAT: bisulfite Analysis Toolkit: BAT is a toolkit to analyze DNA methylation sequencing data accurately and reproducibly. It covers standard processing and analysis steps from raw read mapping up to annotation data integration and calculation of correlating DMRs. *F1000Res* 2017;**6**:1490.
66. Zhou W, Johnson BK, Morrison J *et al.* BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. *Nucleic Acids Res* 2024;**52**:e32. https://doi.org/10.1093/nar/gkae097
67. Farrell C, Thompson M, Tosevska A *et al.* BiSulfite Bolt: a bisulfite sequencing analysis platform. *Gigascience* 2021;**10**:giab033. https://doi.org/10.1093/gigascience/giab033
68. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010;**26**:873–81. https://doi.org/10.1093/bioinformatics/btq057
69. Schultz MD, He Y, Whitaker JW *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* 2015;**523**:212–6. https://doi.org/10.1038/nature14465
70. Ewels P, Hüther P, Miller E *et al.* nf-core/methylseq: Huggy mollusc. v3.0.0 ed. Zenodo. 6 January 2024. https://zenodo.org/records/10463781
71. Schumacher CK, L, Cunningham K. NGS library preparation for balanced, comprehensive methylome coverage from low-input quantities. *Nat Methods* 2015;**12**:v–vi. https://doi.org/10.1038/nmeth.f.386
72. Hey J, Halperin C, Hartmann M *et al.* DNA methylation landscape of tumor-associated macrophages reveals pathways, transcription factors and prognostic value relevant to triple-negative breast cancer patients. *Intl J Cancer* 2023;**152**:1226–42. https://doi.org/10.1002/ijc.34364
73. Clark SJ, Smallwood SA, Lee HJ *et al.* Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc* 2017;**12**:534–47. https://doi.org/10.1038/nprot.2016.187
74. Mayakonda A, Schonung M, Hey J *et al.* Methrix: an R/bioconductor package for systematic aggregation and analysis of bisulfite sequencing data. *Bioinformatics* 2020;**36**:5524–25.
75. Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47. https://doi.org/10.1093/nar/gkv007
76. Muller F, Scherer M, Assenov Y *et al.* RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol* 2019;**20**:55. https://doi.org/10.1186/s13059-019-1664-9
77. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* 2016;**32**:1446–53. https://doi.org/10.1093/bioinformatics/btw026
78. Crusoe MR, Abeln S, Iosup A *et al.* Methods included: standardizing computational reuse and portability with the Common Workflow Language. *Commun ACM* 2022;**65**:54–63. https://doi.org/10.1145/3486897

Received: December 14, 2024. **Revised:** August 5, 2025. **Accepted:** August 24, 2025